Direct detection of early-stage cancers using circulating tumor DNA

Jillian Phallen,1 Mark Sausen,2 Vilmos Adleff,1 Alessandro Leal,1 Carolyn Hruban,1 James White,1 Valsamo Anagnostou,1 Jacob Fiksel,1 Stephen Cristiano,1 Eniko Papp,1,* Savannah Speir,1 Thomas Reinert,3 Mai-Britt Worm Orntoft,3 Brian D. Woodward,4 Derek Murphy,2 Sonya Parpart-Li,2 David Riley,2 Monica Nesselbush,2 Naomi Sengamalay,2 Andrew Georgiades,2 Qing Kay Li,1 Moges Rørbæk Madsen,5 Frank Viborg Mortensen,6 Joost Huiskens,7,8 Cornelis Punt,9 Nicole van Grieken,9 Remond Fijneman,10 Gerrit Meijer,10 Hatim Husain,4 Robert B. Scharpf,1 Luis A. Diaz Jr.,1† Siân Jones,2 Sam Angiuoli,2 Torben Ørntoft,3 Hans Jørgen Nielsen,11 Claus Lindbjerg Andersen,3 Victor E. Velculescu1†

Early detection and intervention are likely to be the most effective means for reducing morbidity and mortality of human cancer. However, development of methods for noninvasive detection of early-stage tumors has remained a challenge. We have developed an approach called targeted error correction sequencing (TEC-Seq) that allows ultrasensitive direct evaluation of sequence changes in circulating cell-free DNA using massively parallel sequencing. We have used this approach to examine 58 cancer-related genes encompassing 81 kb. Analysis of plasma from 44 healthy individuals identified genomic changes related to clonal hematopoiesis in 16% of asymptomatic individuals but no alterations in driver genes related to solid cancers. Evaluation of 200 patients with colorectal, breast, lung, or ovarian cancer detected somatic mutations in the plasma of 71, 59, 59, and 68%, respectively, of patients with stage I or II disease. Analyses of mutations in the circulation revealed high concordance with alterations in the tumors of these patients. In patients with resectable colorectal cancers, higher amounts of preoperative circulating tumor DNA were associated with disease recurrence and decreased overall survival. These analyses provide a broadly applicable approach for noninvasive detection of early-stage tumors that may be useful for screening and management of patients with cancer.

INTRODUCTION

More than 14 million individuals are newly diagnosed with cancer worldwide each year, with the majority having invasive or metastatic disease (1). It is well established that much of the morbidity and mortality in human cancer is related to the late diagnosis of this disease, where surgical and pharmacologic therapies are less effective (2). Unfortunately, clinically proven biomarkers that can be used to broadly diagnose and guide patient management early in the course of disease are not available. Serum-based protein biomarkers such as cancer antigen 125, carcinoembryonic antigen (CEA), prostate-specific antigen, and cancer antigen 19-9 are commonly used for monitoring cancer patients, but because these proteins are also found in the serum of individuals without cancer, they are typically not useful for disease diagnosis (3–7). Other approaches for early detection of cancer, such as stool-based molecular tests or colonoscopies, are limited to individual tumor types and have challenges in patient compliance (8, 9). Currently, no widely applicable biomarkers have been developed for broad detection of human cancer.

The development of noninvasive liquid biopsy methods based on the analysis of cell-free DNA (cfDNA) provides the opportunity for a new generation of diagnostic approaches. Although cfDNA in the circulation was first described more than 50 years ago (10), abnormalities in cancer patients were observed only decades later (11, 12) and showed that such individuals have higher amounts of cfDNA. In patients with cancer, a fraction of cfDNA is tumor-derived and is termed circulating tumor DNA (ctDNA). In principle, analysis of ctDNA has the advantage of identifying alterations that are specific to the tumor. The application of next-generation sequencing (NGS), together with advanced computational methods, has recently allowed ctDNA-based tumor genotyping in a variety of cancer types (13–22). However, these approaches have largely been applied in patients with late-stage cancers or have used tumor tissue sequencing to guide mutational analyses in the blood.

Here, we have developed an ultrasensitive approach for direct analysis of sequence alterations in commonly altered cancer genes in cfDNA without prior knowledge of alterations in the tumor. The sensitivity and specificity of the methodology were evaluated in a clinically relevant cohort of healthy individuals and those with early-stage disease in four common cancers. We identified sequence alterations in cell proliferation genes in individuals without cancer, established the sensitivity of the approach for detecting tumor-specific alterations in the plasma of cancer patients, evaluated concordance between plasma and tumor samples from the same patients, and showed that the amounts of ctDNA can serve as a predictive marker of patient outcome. Overall, these analyses provide information on the potential use and limitations...
of large-scale mutation-based measurements of ctDNA for early diagnosis in common cancers.

RESULTS
Targeted error correction sequencing
We developed a methodology for comprehensive analysis of sequence alterations in driver genes that are commonly mutated in colorectal, lung, ovarian, breast, and other cancers. Similar to targeted analyses of cancer tissues (23), we first selected genes that were frequently mutated in these tumors and focused our analyses on either the entire coding regions or the most highly mutated exons of these genes. An analysis of the frequency of these alterations in the COSMIC database of somatic mutations in cancer (24) revealed that more than three quarters of patients would be expected to have at least one mutation in 55 genes among the intended cancers as well as other common tumor types (Table 1 and table S1). We hypothesized that a larger panel of genes would increase the probability of detecting at least one gene alteration in the plasma from any given cancer patient. Because alterations in the blood have previously been reported in healthy individuals, we examined three additional genes as well as specific sequence positions in three genes of the 55-gene panel (table S1) that were known to be somatically altered in clonal hematopoietic expansion, myelodysplasia, or other hematological malignancies (25–27).

Detection of sequence alterations using conventional NGS is limited to a relatively high fraction of mutant to wild-type DNA (>1%) and, as such, is typically not useful for analyses of ctDNA, which may be present in minute amounts in the blood. Although methods have been developed for analysis of ctDNA in late-stage cancer patients (13–21), no method has been systematically applied for analysis of early-stage disease. We developed a custom capture and sequencing approach called targeted error correction sequencing (TEC-Seq) to allow sensitive and specific detection of low-abundance sequence alterations using NGS (Fig. 1). This methodology is based on targeted capture of multiple regions of the genome and deep sequencing (~30,000x) of DNA fragments. The 58 genes analyzed in this study comprised 50,930 captured bases. Specific steps were performed for analysis of rare tumor-specific alterations in DNA molecules and for elimination of potential amplification, sequencing, and contamination errors as well as other sources of alterations in the blood. These included (i) optimizing library generation and capture for conversion of cfDNA for subsequent analyses; (ii) maximizing representation of unique cfDNA molecules analyzed using mapping positions and a small number of prespecified barcodes; (iii) redundant sequencing, where multiple identical DNA molecules are generated and sequenced and any sequence changes are reconciled; (iv) filtering of mapping and sequencing artifacts; and (v) identifying and removing germline and hematopoietic cell proliferation alterations.

Conceptually, the number of genome equivalents analyzed provides a lower limit of detection for any genomic analysis. A high sensitivity approach would aim to maximize the number of unique molecules assessed while allowing for a broad and facile analysis in a range that is above the actual number of fragments present in a biologic sample. We optimized methodologies for extraction and conversion of cfDNA to genomic libraries. Initially, we considered using the start and end genome mapping positions of paired-end sequenced fragments as “endogenous barcodes” to distinguish between individual molecules. However, Monte Carlo simulations suggested that the tight size distribution of cfDNA molecules observed in the plasma would result in a smaller number of possible end mapping combinations and therefore underestimate the true complexity of cfDNA in the circulation (fig. S1). To extend the complexity of endogenous barcodes, we introduced a limited set of sequence indices as “exogenous barcodes” in the initial steps of library generation. Kinde et al. (28) reported the use of a large number of random exogenous barcodes as unique identifiers for analysis of rare mutations in DNA populations. However, simulations with a relatively small number of long prespecified exogenous barcodes (4–16) suggested that these, in combination with endogenous barcodes, would be sufficient to distinguish among different cfDNA molecules in the plasma from a typical blood draw (fig. S1). Extending the number of barcodes substantially beyond this number has the theoretical disadvantage of misassignment among barcodes through sequencing errors and primer dimers that can form during library formation.

We first evaluated the characteristics of the TEC-Seq approach for detecting known tumor-specific alterations from a mixture of DNA from tumor cell lines at different dilutions (ranging from 100 to 0.1%) with unrelated wild-type DNA. Libraries with eight exogenous barcodes were sequenced with an average of ~32,224 sequence reads at each position among the 58 genes analyzed (table S2). We designed thresholds that were expected to identify >99% of alterations with a mutant allele fraction of 0.5% at the anticipated sequencing depth. Alterations were considered if they were present in all copies of multiple sequences of each DNA molecule with identical endogenous and exogenous barcodes and were not removed by additional error filtering steps. Hotspot alterations at positions previously observed to be frequently altered in cancer patients were evaluated with more sensitive thresholds because the a priori probability that these alterations were tumor-derived is higher than that of other alterations. Alterations present in common germline variant databases or in 25% or greater of reads were considered germline and removed from further analysis, unless the mutations were identical to known hotspot alterations or represented truncating mutations in common tumor suppressor genes. Analysis of the altered positions in the dilution samples revealed high concordance to the expected fraction of mutant molecules (r = 0.93; P < 0.0001, Pearson correlation; fig. S2 and table S2), as well as high sensitivity and specificity. The analytical sensitivity was 97.4% overall and 100 and 89% for detecting mutations present at 0.2 and 0.1%, respectively, using minimum thresholds of 0.05% in hotspot positions and 0.1% at all other locations. No false positives were detected over the 80,930 bases analyzed in 38 dilution analyses, resulting in less than one error in 3 million bases sequenced (error rate of <3.3 × 10−7 false-positive mutation calls per base; specificity, >99.9999%; table S2).

Table 1. Cancer cases containing alterations in driver genes.

<table>
<thead>
<tr>
<th>Tissue type</th>
<th>Cases in COSMIC</th>
<th>Detectable cases*</th>
<th>Detectable fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>1002</td>
<td>719</td>
<td>72%</td>
</tr>
<tr>
<td>Colorectal</td>
<td>1248</td>
<td>1071</td>
<td>86%</td>
</tr>
<tr>
<td>Lung</td>
<td>1198</td>
<td>932</td>
<td>78%</td>
</tr>
<tr>
<td>Ovarian</td>
<td>647</td>
<td>524</td>
<td>81%</td>
</tr>
</tbody>
</table>

*Detectable cases indicate those with at least one alteration in the cancer driver genes analyzed (table S1).
Evaluation of plasma from healthy individuals

We used TEC-Seq to examine plasma specimens from 44 healthy individuals (tables S3 and S4). These individuals were not known to have cancer and provided their blood samples as part of a routine cancer screening visit (colonoscopy or Papanicolaou test). Samples were processed within 2 hours from collection and centrifuged twice at high speed to ensure that cells and cellular debris were removed and that only cfDNA was analyzed. From the ~4 ml of plasma obtained from each individual, we generated TEC-Seq libraries and sequenced these to ~30,000-fold coverage. Through these analyses, no mutations were observed in the cancer driver genes analyzed in our panel, consistent with the estimated specificity observed in our dilution analyses. Alignment to the reference genome of multiple distinct molecules containing identical redundant changes is used to identify bona fide alterations.

Fig. 1. Schematic of the TEC-Seq method. cfDNA is extracted from the blood and converted to a genomic library through ligation of a pool containing a small number of dual-index barcode adapters. The resulting cfDNA library is captured and redundantly sequenced to produce multiple duplicates of each DNA fragment. Sequence reconciliation among duplicate fragments identifies alterations present in identical DNA molecules with the same start and end position and exogenous barcodes. Alignment to the reference genome of multiple distinct molecules

Analysis of plasma from patients with cancer

We next analyzed plasma samples from 194 patients with breast cancer (n = 45), colorectal cancer (n = 42), lung cancer (n = 65), and ovarian cancer (n = 42). The cohort consisted of untreated patients who had localized or metastatic disease, with most of the patients diagnosed at stages I and II (table S3). We found that the concentration of cfDNA in plasma from cancer patients was ~29 ng/ml, significantly higher than that observed in healthy individuals (average of 7 ng/ml; P = 0.001, unpaired t test; Fig. 3A). In the colorectal cancer cohort, where a larger number of later-stage patients were analyzed, we found that samples from patients with metastatic disease had higher concentration of cfDNA than those from patients with earlier stages of disease (average of 66 ng/ml for stage IV patients versus 21 ng/ml for stages I to III; P = 0.006, unpaired t test; Fig. 3B).

We examined the cfDNA from these patients using the TEC-Seq approach. Of the 194 patients analyzed, more than three quarters of colorectal cancer patients, two-thirds of ovarian cancer patients, and most of the lung and breast cancer patients had detectable alterations in driver genes (Table 2). These detection rates were higher in some cases than the theoretical estimates for these cancer types (Tables 1 and 2). More than three quarters of patients with advanced disease (stages III and IV) and 62% of patients with localized disease (stages I and II) were detected among all tumor types (Table 2). The amounts of ctDNA varied among cancer types, with breast cancer having the lowest mutant allele fraction (P = 0.028, unpaired t test; Fig. 3C). Similar to observations of cfDNA, the amounts of ctDNA were higher in metastatic disease compared to earlier-stage disease among all cancer types (P < 0.0001, unpaired t test; Figs. 3D and 4). Eighty of 128 detected cases had at least one alteration in a gene hotspot position (Fig. 4). The affected genes and distribution of alterations for each tumor type were similar to common driver gene alterations that have previously been reported in these cancers (fig. S3). On average, 2.1 alterations, including 0.9 changes at hotspot positions, were observed in each patient with detectable ctDNA, with lung and colorectal cancers having a higher

All of the alterations were identified in DNA methyltransferase 3a (DNMT3A), a gene that is clonally altered under preleukemic conditions and myelodysplasia (25–27). Three of the mutations were predicted to result in the R882C change previously observed in clonal hematopoiesis, but other alterations have not been previously reported. These mutations were identified at mutant allele fractions of 0.16 to 5.3%, substantially lower than previous observations in blood cells of healthy individuals (25–27). Our analyses suggest that a higher fraction of asymptomatic individuals may harbor such somatic alterations than had been previously reported through cellular analyses of these genes in the blood.
number of alterations per case (Fig. 4). By limiting analysis only to a specific set of hotspot variants as others have reported (21), the fraction of cases detected was reduced to 56% of those identified by TEC-Seq. These observations highlight the benefit of analyzing a broader panel of driver gene regions to increase the possibility of detecting tumor-specific alterations in the plasma.

Comparison of mutations in plasma with those in matched tumor and blood cells
Of the 194 patients in our study, 152 cases had matched tumor and normal tissues that we analyzed using an independent targeted NGS approach (tables S3 and S4). We examined these cases to determine whether the mutations identified in the plasma were tumor-specific or may have originated during blood cell expansion. The plasma analyses performed using TEC-Seq were performed separately and did not rely on any knowledge of alterations identified through these parallel tissue analyses.

We detected 87 changes in the circulation of 194 patients at allele fractions >25% and considered these to be likely germline variants. Analysis of 63 of these variants in the available corresponding blood cells identified all these changes to be germline (table S6). These observations suggested that cfDNA can be used to accurately identify...
with cancer, regardless of stage. Analysis of matched white blood cells
from individuals with alterations in these genes identified the corresponding
mutation in most of the cases, consistent with the notion that the altera-
tions in cfDNA originated from these cells (table S5).

After accounting for blood cell proliferation and germline altera-
tions, we identified 313 candidate tumor-specific changes in the plasma
samples from 128 of the 194 patients analyzed. We further evaluated
216 of these alterations in 100 patients where matched tumor tissue
and blood cells were available. We found that 155 of the 216 (72%) altera-
tions were identical in both plasma and tumor samples (Fig. 5). Among
stage III and IV patients, 65 of 84 (77%) variants were concordant,
whereas for early-stage patients, 90 of 132 (68%) alterations were
concordant. In line with these observations, we found that 70 of
the 75 (93%) alterations with a mutant allele fraction >1% in the plasma
were detected in the tumor tissue of the same individual. Overall, 82
of the 100 (82%) patients had at least one alteration observed in the
circulation that was identical to that in the tumor specimen.

To evaluate reproducibility of the approach between separate blood
draws in the same patients, we assessed six late-stage patients with lung
cancer where blood was obtained early during the course of treatment.
These patients were undergoing treatment but were observed to have
progressive or stable disease. Despite the difference in time between
the blood draws, we found that 90% of the alterations observed in the
second blood draw were present at the time of the first blood draw (17 of
19 alterations), with one patient having no alterations at both time
points (fig. S4). All alterations present with a mutant allele fraction
≥1% were observed at both time points.

In a subset of colorectal cancer patients, we evaluated whether the
observations we detected in the plasma could be independently con-
firmed using droplet digital polymerase chain reaction (ddPCR), a method
that is highly sensitive for detection of single-base substitutions (29).
We examined six driver alterations detected in the plasma: two that
were also detected in matched tumors and four that were absent. Five of
the six driver alterations were detected in the plasma by ddPCR at levels
similar to those observed by TEC-Seq (fig. S5A). Those not detected in
tumors by targeted sequencing were similarly not identified through
ddPCR approaches. We also evaluated 10 mutations that corresponded
to the most common changes in KRAS, PIK3CA, and BRAF that we
detected in these tumors but were not present in the plasma of these
patients. Although we confirmed that these alterations were in the tumors
of these patients, we found that those not detected by TEC-Seq analyses
remained undetected by ddPCR in the plasma, presumably because the
amounts of ctDNA corresponding to these alterations were extremely
low in these patients (fig. S5B).

To assess the possibility that tumor heterogeneity may be respon-
sible for the apparent lack of concordance between specific alterations
in the plasma and those in the tumor, we analyzed multiple tumor sites
from colorectal cancer patient CGCRC307 using ddPCR. We character-
ized 10 different regions of the tumor as well as a subsequent metastatic
site for an R201C alteration in the GNAS gene that we detected in the
plasma but not in the tumor of this patient. Although we found a BRAF
V600E alteration in all samples analyzed, the GNAS R201C substitution
was not detected in the original tumor biopsy but was detected as a
subclonal change in only a portion of the primary tumor, suggesting that
it developed later in tumorigenesis (fig. S6). The GNAS R201C change
identified had been previously reported in colorectal cancers (30) and
has been shown to promote intestinal tumorigenesis through activation
of both Wnt and ERK pathways (31). Consistent with this notion, we
found the GNAS alteration to be clonal in the metastatic lesion that was
identified 2 years after the primary tumor in this patient (fig. S6). These

---

Table 2. Cancer patients detected using TEC-Seq. NA, not applicable.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Patients (n)</th>
<th>Patients with ctDNA alterations (n)</th>
<th>Fraction of patients with ctDNA alterations (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorectal</td>
<td>I</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>I-IV</td>
<td>42</td>
<td>35</td>
</tr>
<tr>
<td>Lung</td>
<td>I</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>32</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>I-IV</td>
<td>71</td>
<td>44</td>
</tr>
<tr>
<td>Ovarian</td>
<td>I</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>I-IV</td>
<td>42</td>
<td>30</td>
</tr>
<tr>
<td>Breast</td>
<td>I</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>29</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>I-IV</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>All</td>
<td>I and II</td>
<td>138</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>III and IV</td>
<td>62</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>I-IV</td>
<td>200</td>
<td>134</td>
</tr>
</tbody>
</table>

---

Similar to our observations in healthy individuals, we identified
alterations in DNMT3A and five other genes involved in blood cell
proliferation in the plasma of cancer patients (table S5). The fraction
of patients with detectable changes in these genes correlated with age,
as previously observed (P = 0.013, unpaired t test) (25–27). Unlike tumor-
specific alterations, the allele fractions of blood cell proliferation altera-
tions in cfDNA were similar among healthy individuals and patients
with cancer, regardless of stage. Analysis of matched white blood cells

---

Fig. 4. ctDNA in patients with breast, colorectal, lung, and ovarian cancer. Patients (n = 194) are each represented by a tick mark. (Left) Bar chart shows the number of alterations detected for each case. (Middle) Stage, cancer type, and histopathological subtype are represented by colored vertical bars. (Right) Mutant allele fractions for each alteration detected per patient are indicated with an “x” at the mean. Alterations are colored on the basis of hotspot status and whether any alterations were detected in the case.
results suggest that plasma alterations not detected in the matched tumor specimens may represent bona fide somatic mutations in ctDNA derived from heterogeneous primary or occult lesions.

**ctDNA and disease progression**

Tumor-specific markers may be useful for evaluating disease progression. In colorectal cancer, CEA is commonly used to monitor patients after therapy to determine recurrence or progressive disease (7, 32). Of the 29 colorectal cancer patients for whom CEA values were available, all 10 cases with CEA concentrations >5 ng/ml had detectable ctDNA (tables S3 and S6). However, among the 19 patients with negative or borderline CEA results, 13 had detectable ctDNA, including patients of all stages (tables S7 and S8). There was no significant correlation between ctDNA and CEA concentrations (Pearson correlation coefficient = −0.017; \( P = 0.93 \)).

We next examined whether preoperative ctDNA analyses may be related to disease recurrence and survival after surgical resection. We hypothesized that elevated amounts of ctDNA were more likely to be associated with large primary lesions that were incompletely resected or with occult metastases. A total of 31 colorectal cancer patients had potentially curative resections, including 8 stage I, 9 stage II, 10 stage III, and 4 stage IV patients with liver-only metastases. For these patients, the median mutant allele fraction was 0.21%. However, several patients had mutant allele fractions >3 median absolute deviations from the median mutant allele fraction, or >2%. As predicted, we found that high amounts of ctDNA correlated with poor prognosis (fig. S7). Patients with increased ctDNA had a shorter progression-free survival (PFS) and overall survival (OS) compared to those with lower ctDNA amounts (\( P < 0.0001 \) for PFS and OS, log-rank test; Fig. 6, A and B). The prognostic value for PFS was statistically significant in multivariate models, adjusted for stage as a categorical covariate [hazard ratio, 36.3; 95% confidence interval (CI), 2.8 to 471.1; \( P = 0.006 \), Cox proportional hazards model].

These same predictions were observed in patients with resectable stage I to III disease (\( P = 0.0006 \) for PFS and \( P < 0.0001 \) for OS, log-rank test; Fig. 6, C and D). We also evaluated other thresholds of increased amounts of ctDNA and found that these were statistically significantly associated with worse outcome (\( P = 0.008 \) for 0.5% mutant allele fraction and \( P = 0.0001 \) for 1% mutant allele fraction, log-rank test). In addition, we found that considering ctDNA amounts as a continuous variable correlated with outcome (hazard ratio, 1.13; 95% CI, 1.03 to 1.24; \( P = 0.01 \) for PFS and OS, Cox univariate test). Together, these results indicate that liquid biopsy analyses offer both a quantitative and qualitative assessment of disease progression. Although previous analyses have found a limited association between preoperative CEA concentrations and OS (7, 32), CEA concentrations among our patients were not associated with disease outcome (\( P = 0.75 \) for PFS and \( P = 0.73 \) for OS, log-rank test; fig. S8). These analyses from a limited and heterogeneous cohort of patients suggest that preoperative ctDNA amounts may provide a useful marker of disease outcome in operable colorectal cancer.

**DISCUSSION**

These analyses provide an approach for noninvasive direct detection of patients with early-stage disease across common cancer types. A conceptual benefit of this approach is that detectable alterations in ctDNA are, by definition, clonal and therefore indicate an underlying population of cells with identical somatic mutations. This high degree of specificity is one of the potential benefits of ctDNA detection compared to other blood-based biomarkers, which may be increased in other normal tissues in patients without cancer.
Although ctDNA analyses have raised the possibility of direct detection of patients with early-stage disease (13, 33), the de novo identification of somatic alterations has remained a major challenge for the development of early detection approaches. The analytical performance characteristics of the TEC-Seq method suggest that it may be suitable for such analyses. Other methods have been used for analyses of cfDNA in late-stage cancer patients (13–22), but the specificity and sensitivity of these methods may limit their applicability for detection of early-stage disease. A variety of experimental and bioinformatic aspects may contribute to the high specificity of the TEC-Seq method compared to previous approaches, including deep sequencing (>30,000-fold coverage), use of a small number of adaptors with long prespecified barcodes, and multiple bioinformatic filtering steps comprising error correction, removal of repetitive sequences and mapping artifacts, and identification and removal of germline and hematopoietic sequences.

Using the TEC-Seq approach, no tumor-derived alterations were identified in the plasma of the healthy individuals in our study. Although the average age of the healthy cohort was younger than the cancer patients analyzed, this corresponds to an age at which cancer screening may be initiated. Likewise, the concordance between liquid and tumor biopsies was high and suggested that liquid biopsies may have advantages for detection of heterogeneous tumor-specific alterations that may be missed by tissue biopsies. In the colorectal cancer case analyzed through multiple tissue biopsies, we showed that heterogeneous alterations appeared to have lower amounts of ctDNA and may explain the wide range of mutant allele fractions in ctDNA in the same individuals. One concern is that clonal hematopoietic changes may be confounded with heterogeneous tumor-specific mutations (25, 27) and lead to over-
MATERIALS AND METHODS

Study design

This study presents a prospective analysis of cfDNA using an ultrasensitive sequencing and analysis platform to detect somatic sequence alterations in early-stage cancers. We analyzed 250 plasma samples from 244 individuals, including 44 healthy individuals and 200 patients with colorectal (n = 42), lung (n = 71), ovarian (n = 42), or breast (n = 45) cancer over a range of stages, with most patients exhibiting localized disease. We estimated that analysis of at least 42 patients for each tumor type would provide a 96% power to detect 50% of cases with a 95% CI of 35 to 65%. We evaluated the sensitivity and specificity of the TEC-Seq method to detect cfDNA in early-stage patients without previous knowledge of alterations in their tumors. We detected sequence alterations in hematopoietic expansion genes in healthy individuals, established the sensitivity of the approach for detecting tumor-specific alterations in the blood of cancer patients, evaluated concordance between alterations identified in cfDNA and tumor samples from the same patients, and assessed whether preoperative cfDNA can serve as a marker of patient outcome.

Patient and sample characteristics

Plasma samples from healthy individuals and plasma and tissue samples from patients with breast, lung, ovarian, and colorectal cancers were obtained from ILSBio/Bioreclamation, Aarhus University, the Academic Medical Center of the University of Amsterdam, and University of California, San Diego. All samples were obtained under Institutional Review Board–approved protocols with informed consent for research use at participating institutions.

Plasma samples from healthy individuals were obtained at the time of routine screening, including for colonoscopies or Pap smears. Individuals were considered healthy if they had no previous history of cancer and negative screening results. Plasma samples from individuals with colorectal, lung, ovarian, or breast cancer were obtained at the time of diagnosis, before tumor resection. Serially collected plasma samples from lung cancer patients were collected over a course of treatment, with the exception of stage IV colorectal cancer patients whenever available. Tumor specimens were obtained from ILSBio/Bioreclamation, Aarhus University, the Academic Medical Center of the University of Amsterdam, and University of California, San Diego. All samples were obtained under Institutional Review Board–approved protocols with informed consent for research use at participating institutions.

Sample preparation and NGS of cfDNA

Whole blood was collected in EDTA tubes and processed immediately or within 2 hours after storage at 4°C to separate plasma and cellular components by centrifugation at 800g for 10 min at 4°C. Plasma was centrifuged a second time at 18,000g at room temperature to remove any remaining cellular debris and stored at −80°C until the time of DNA extraction. DNA was isolated from plasma using the Qiagen Circulating Nucleic Acids Kit (Qiagen GmbH) and eluted in LoBind tubes (Eppendorf AG). Concentration and quality of cfDNA were assessed using the Bioanalyzer 2100 (Agilent Technologies).

TEC-Seq NGS cfDNA libraries were prepared from 5 to 250 ng of cfDNA. Genomic libraries were prepared using the NEBNext DNA Library Prep Kit for Illumina [New England Biolabs (NEB)] with four main modifications to the manufacturer’s guidelines: (i) The library purification steps used the on-bead Ampure XP approach to minimize sample loss during elution and tube transfer steps (36); (ii) NEBNext End Repair, A-tailing, and adapter ligation enzyme and buffer volumes were adjusted as appropriate to accommodate the on-bead Ampure XP purification strategy; (iii) a pool of eight unique Illumina dual index adapters with 8–base pair (bp) barcodes was used in the ligation reaction instead of the standard Illumina single or dual index adapters with 6- or 8-bp barcodes, respectively; and (iv) cfDNA libraries were amplified with Phusion Hot Start Polymerase. Incorporation of these modifications improved conversion efficiency from 13.4% before modifications to 34.1% in validation analyses of 38 cases incorporating these changes. Analysis of plasma samples from healthy individuals and cancer patients revealed a conversion efficiency of 40%, with a significant correlation between input DNA amount and the number of distinct molecules analyzed (Pearson correlation r = 0.55; 95% CI, 0.46 to 0.64; P < 0.0001; fig. S9).

Briefly, cfDNA was combined with End Repair Reaction Buffer (NEB) and End Repair Enzyme Mix (NEB) and incubated for 30 min at 20°C. The end-repair reaction was purified with Agencourt Ampure XP Beads (Beckman Coulter). A-tailing was performed by adding 6 μl of daT-A-Tailing Reaction Buffer (NEB) and 3.6 μl of Klenow (NEB) to the end-repaired cfDNA and incubating for 30 min at 37°C. A-tailed cfDNA was purified using Agencourt Ampure XP Buffer (Beckman Coulter). Adaptor oligonucleotides containing the TEC-Seq dual index pools and Quick T4 DNA Ligase (NEB) were mixed with A-tailed, on-bead cfDNA and incubated for 15 min at 20°C. Ligated cfDNA was purified with two rounds of Agencourt Ampure XP Buffer. The cfDNA library was amplified using Phusion Hot Start DNA polymerase (Thermo Fisher Scientific) and PCR primers published for the Nextera DNA Library Prep Kit: 5′-AATGATACGGCGACCACCGA-3′ and 5′-CAAGCAGAAGACGGCATACGAG-3′ (Illumina Inc.). For each genomic library, PCRs contained 2 μl of cfDNA library, 15.5 μl of H2O, 1.25 μl of dimethyl sulfoxide, 5.0 μl of 5X Phusion HF Buffer, 0.5 μl of dextrose nucleoside triphosphate (dNTP) mix containing 10 mM each dNTP (Life Technologies), 0.5 μl of each primer, and 0.25 μl of Phusion Hot Start Polymerase. The following PCR conditions were used: 98°C for 30 s; 12 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 30 s; and 72°C for 5 min. Purification of the amplified cfDNA library was performed using Agencourt Ampure XP Beads. Concentration and quality of cfDNA libraries were assessed using the Bioanalyzer 2100 (Agilent Technologies).

Targeted capture was performed using the Agilent SureSelect reagents and a custom set of hybridization probes targeting 58 genes (table S1) per the manufacturer’s guidelines. The captured library was amplified with Phusion Hot Start Polymerase (NEB). The concentration and quality of captured cfDNA libraries were assessed on the Bioanalyzer 2100 using the DNA 1000 Kit (Agilent Technologies). TEC-Seq libraries were sequenced using 100-bp paired-end runs on the Illumina HiSeq 2000/2500 (Illumina).

Sample preparation and NGS of tumor-normal pairs

Sample preparation, library construction, targeted capture, NGS, and bioinformatic analyses of tumor and normal samples were performed as previously described (23, 37). Briefly, DNA was extracted from matched FFPE or frozen tumor tissue and buffy coat samples using the Qiagen DNA FFPE Tissue Kit or Qiagen DNA Blood Mini Kit (Qiagen GmbH). Genomic DNA from tumor and normal samples was fragmented and used for Illumina TruSeq library construction.
Using Illumina CASAVA (Consensus Assessment of Sequence and Variation) software (version 1.8), including demultiplexing and masking of dual-index adapter sequences. Sequence reads were aligned against the human reference genome (version hg18 or hg19) using NovoAlign with additional realignment of select regions using the Needleman-Wunsch method (23). The positions of the alterations we have identified have not been affected by the different genome builds.

Next, candidate somatic mutations, consisting of point mutations, small insertions, and deletions, were identified using VariantDx (23) across the targeted regions of interest. VariantDx examined sequence alignments of cfDNA plasma samples while applying filters to exclude alignment and sequence artifacts. Specifically, an alignment filter was applied to exclude quality-failed reads, unpaired reads, and poorly mapped reads in the plasma. A base quality filter was applied to only include bases with a reported Phred quality score $\geq 30$.

A mutation identified in cfDNA was considered a candidate somatic mutation only when (i) three distinct paired reads contained the mutation in the plasma (each redundantly sequenced at least three times) with a distribution of start and cycle positions when compared to the reference genome, and the number of distinct paired reads containing a particular mutation in the plasma was at least 0.1% of the total distinct read pairs; or (ii) four distinct paired reads contained the mutation in the plasma (each redundantly sequenced at least four times) with a distribution of start and cycle positions when compared to the reference genome, and the number of distinct paired reads containing a particular mutation in the plasma was at least 0.05% and less than 0.1% of the total distinct read pairs; and (iii) the mismatched base was not present in $\geq 1$% of the reads in a panel of unmatched normal samples and not present in a custom database of common germline variants derived from dbSNP (The Single Nucleotide Polymorphism Database).

Mutations arising from misplaced genome alignments, including paralogous sequences, were identified and excluded by searching the reference genome. Candidate somatic mutations were further filtered on the basis of gene annotation to identify those occurring in protein-coding regions. Functional consequences were predicted using snpEff and a custom database of CCDS (Consensus Coding Sequence), RefSeq, and Ensembl annotations using the latest transcript versions available on hg18 and hg19 from the University of California, Santa Cruz (https://genome.ucsc.edu/). Predictions were ordered to prefer transcripts with canonical start and stop codons and CCDS or RefSeq transcripts over Ensembl when available. Finally, mutations were filtered to exclude intronic and silent changes, while retaining mutations resulting in missense mutations, nonsense mutations, frameshifts, or splice site alterations.

Candidate alterations were defined as somatic hotspots if the nucleotide change and amino acid change were identical to an alteration observed in $\geq 20$ cancer cases reported in the COSMIC database. Alterations that were not hotspots were retained only if either (i) seven or more distinct paired reads contained the mutation in the plasma, and the number of distinct paired reads containing a particular mutation in the plasma was at least 0.1% and less than 0.2% of the total distinct read pairs, or (ii) six or more distinct paired reads contained the mutation in the plasma, and the number of distinct paired reads containing a particular mutation in the plasma was at least 0.2% of the total distinct read pairs.

Candidate mutations were further limited through identification and removal of common germline variants present in $\geq 25\%$ of reads or $<25\%$ of reads if the variant was recurrent and most of the alterations at that position had a mutant allele fraction of $\geq 25\%$ (table S6). Variants known to be at a somatic hotspot position or producing a truncating mutation in a tumor suppressor gene were not excluded as germline changes. Because of the high frequency of mutations in specific genes and the possible confounding between somatic and germline changes, we limited analyses in the APC gene to frameshift or nonsense mutations and in KRAS, HRAS, and NRAS to positions 12, 13, 61, and 146. Finally, we excluded hematopoietic expansion-related variants that have been previously described, including those in DNMT3A, IDH1, and IDH2 and specific alterations within ATM (residue 3008), GNAS (residue 202), or JAK2 (residue 617) (table S1) (25–27).

To evaluate the sensitivity of the TEC-Seq approach using dilutions of cell lines with known mutations, we used a mixture of cell lines obtained from American Type Culture Collection and combined in ratios to reflect the mutant allele frequency. The cell lines in the mutant pool included CCL-237, CRL-2158, CRL-2547, CRL-7585, CRL-9068, CRL-2177, CCL-231, CRL-2871, CRL-5908, CRL-5908, CCL-224, and CRL-5894. To evaluate sensitivity and specificity, we used dilutions of a cell line (CGBR4C, CRL-2338), which had been previously sequenced to examine both mutant and wild-type bases in the 58 genes in our panel (30). For analyses at all dilutions, we considered those alterations where the mutant allele fraction was expected to be at 0.1% or higher. To calculate the per-base error rate for conventional sequencing in samples from healthy individuals, we summed the number of false-positive calls at each genomic position and divided this by the total coverage at that base for the 44 healthy individuals. The upper limit of the per-base error rate of TEC-Seq was determined by assuming one alteration per base if no error was identified and dividing by the total coverage at each base for the 44 healthy individuals analyzed.

To compare the TEC-Seq bioinformatic approach to iDES-enhanced CAPP-Seq (cancer personalized profiling by deep sequencing), we used the bioinformatic components of iDES combined with the requirement of multiple distinct read families based on endogenous and exogenous barcodes (19, 21) (https://cappseq.stanford.edu/ides/).

Analyses of NGS data from tumor-normal pairs

Primary processing of NGS data from tumor-normal pairs and identification of putative somatic mutations were completed using Illumina CASAVA software (version 1.8) and VariantDx custom software, respectively, as previously described (23).

Statistical analyses

We used a variety of methods for determining significance. To test the linear association between expected and observed mutant allele fractions (fig. S2), we used Pearson’s product moment correlation coefficient. To quantify the difference in mean error rate by genomic position for conventional sequencing and TEC-Seq, we used a paired (by genomic position) $t$ test assuming equal variances. Differences in means of unpaired (independent) samples were tested using a two-sample $t$ test assuming equal variances (such as for comparisons involving the concentration of cfDNA in plasma between healthy and cancer...
To assess whether high mutual allele fractions are associated with patient outcomes, we defined patients with high mutual allele fractions as those with values >3 median absolute deviations from the median mutual allele fraction observed in 31 colorectal cancer patients analyzed. We used a median absolute deviation rather than an SD because the mutant allele fractions were skewed, and the median absolute deviation provides a more robust-to-outlier measure of the SD.

We compared PFS and OS between patients with low and high mutual allele fraction using the log-rank test in univariate analyses and the Cox proportional hazards in multivariate analyses (38, 39).

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/9/403/eaan2415/DC1

Fig. S1. Simulations using limited exogenous barcodes.

Fig. S2. Validation of the TEC-Seq approach.

Fig. S3. Mutation frequencies in cancer genes.

Fig. S4. cDNA mutant allele fractions in serial blood draws.

Fig. S5. Comparison of cDNA mutant allele fractions measured by TEC-Seq and ddPCR.

Fig. S6. cDNA and tumor heterogeneity.

Fig. S7. Preoperative cDNA mutant allele fractions in colorectal cancer patients.

Fig. S8. Preoperative CEA in colorectal cancer patients.

Fig. S9. Conversion efficiency of cDNA.

Table S1. Genes analyzed by TEC-Seq (provided in a separate Excel file).

Table S2. Summary of TEC-Seq validation (provided in a separate Excel file).

Table S3. Summary of patients analyzed (provided in a separate Excel file).

Table S4. Summary of genomic analyses (provided in a separate Excel file).

Table S5. Alterations in blood cell proliferation genes in healthy individuals and cancer patients (provided in a separate Excel file).

Table S6. Germline alterations identified in cDNA (provided in a separate Excel file).

Table S7. Somatic alterations detected in cfDNA of cancer patients (provided in a separate Excel file).

Fig. S10. Conversion efficiency of cDNA.

Fig. S11. Mutation frequencies in cancer genes.

Fig. S12. cDNA mutant allele fractions in serial blood draws.

Fig. S13. Comparison of cDNA mutant allele fractions measured by TEC-Seq and ddPCR.

Fig. S14. cDNA and tumor heterogeneity.

Fig. S15. Preoperative cDNA mutant allele fractions in colorectal cancer patients.

Fig. S16. Preoperative CEA in colorectal cancer patients.

Fig. S17. Conversion efficiency of cDNA.

Table S1. Genes analyzed by TEC-Seq (provided in a separate Excel file).

Table S2. Summary of TEC-Seq validation (provided in a separate Excel file).

Table S3. Summary of patients analyzed (provided in a separate Excel file).

Table S4. Summary of genomic analyses (provided in a separate Excel file).

Table S5. Alterations in blood cell proliferation genes in healthy individuals and cancer patients (provided in a separate Excel file).

Table S6. Germline alterations identified in cDNA (provided in a separate Excel file).

Table S7. Somatic alterations detected in cfDNA of cancer patients (provided in a separate Excel file).

REFERENCES AND NOTES


Acknowledgments: We thank members of our laboratories for critical review of the manuscript. Funding: This work was supported, in part, by the U.S. NIH (grants CA121113, CA006973, and CA180950), the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up to Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Commonwealth Foundation, the Cigarette Restitution Fund, the Burroughs Wellcome Fund, the Maryland Genetics, Epidemiology and Medicine Training Program, the International Association for the Study of Lung Cancer/Prevent Cancer Foundation, the Danish Council for Independent Research (11-05240), the Danish Council for Strategic Research (1309-00068), the Novo Nordisk Foundation (NNF14OC012747), and the Danish Cancer Society (R133-A520 and R40-A1965-11-52). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Author contributions: J.P. designed the study, performed experiments and analyses, and wrote the paper. M.S., V. Adleff, and R.B.S. performed experiments and analyses and wrote the paper. A.L., C.H., J.W., J.F., E.P., S., R.T.-M., B.W.O., B.D.W., D.M., S.P.-L., D.R., M.N., N.S., A.G., Q.L.K., M.R.M., F.V.M., J.H., C.P., N.v.G., F.H., S.J., S.A., T.D., and H.J.N. performed analyses. V. Anagnostou, S.C., G.M., and L.A.D. performed analyses and provided funding. C.L.A. performed analyses, provided funding, and wrote the paper. V.E.V. designed the study, performed analyses, provided funding, and wrote the paper. Competing interests: J.P. and V.E.V. are inventors on patent applications 62/501,686 and 62/516,009 submitted by Johns Hopkins University related to detection of cancer. M.S., L.A.D., and V.E.V. are inventors on a patent application 62/422,355 submitted by Personal Genome Diagnostics related to nonunique barcodes for genotyping. V.E.V. is a founder of the Personal Genome Diagnostics, is a member of its Scientific Advisory Board and Board of Directors, and owns a Personal Genome Diagnostics stock, which is subject to certain restrictions under university policy. V.E.V. is also on the Scientific Advisory Board for Ignyta. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies. L.A.D. is a founder of Personal Genome Diagnostics and PapGene and a stock owner for both entities, a member of the Personal Genome Diagnostics Board of Directors, and a consultant for Personal Genome Diagnostics, Merck, and Cell Design Labs. V. Adleff is a consultant for Personal Genome Diagnostics. Data and materials availability: Data have been deposited at the European Genome-phenome Archive, which is hosted at the European Bioinformatics Institute, under study accession EGA50000102577.

Submitted 16 March 2017
Resubmitted 13 May 2017
Accepted 22 July 2017
Published 16 August 2017
10.1126/scitranslmed.aan2415


Direct detection of early-stage cancers using circulating tumor DNA

Finding smaller needles in haystacks

The detection and analysis of cell-free DNA in patients’ blood are becoming increasingly accepted in oncology. However, this approach has generally been applied for the monitoring of patients with existing tumors. It has not been useful for early diagnosis of cancer because of insufficient sensitivity to detect really small tumors that only shed minute quantities of DNA into the blood, as well as difficulties with identifying cancer-associated genetic changes without knowing what mutations are present in the primary tumor. A method developed by Phallen et al., called targeted error correction sequencing, addresses both of these limitations and demonstrates the feasibility of detecting circulating cell-free DNA from many early tumors, suggesting its potential use for cancer screening.