Early-stage cancer results in a multiplicative increase in cell-free DNA originating from healthy tissue

Konstantinos Mamis[1], Ivana Bozic[1,2]*

[1]Department of Applied Mathematics, University of Washington, Seattle WA, USA
[2]Herbold Computational Biology Program, Fred Hutchinson Cancer Center, Seattle WA, USA

ORCID: 0000-0001-9772-810X (KM), 0000-0003-2761-825X (IB)

*Corresponding author
Email: ibozic@uw.edu (IB)

Abstract

Cell-free DNA is a promising biomarker for cancer detection. However, the sources of elevated cell-free DNA (cfDNA) in patients with early-stage cancer, and the mechanisms by which cfDNA is shed into, and subsequently cleared from the circulation are still poorly understood. Leveraging a rich dataset of cfDNA in healthy individuals and early-stage cancer patients, we find a multiplicative increase in cfDNA concentration in the presence of cancer. This increase is cancer type-specific, ranging from a ~1.3-fold increase in lung cancer, to a ~12-fold increase in liver cancer, and does not originate from tumor, but from healthy tissue. Employing an additional dataset reporting the tissue of origin of cfDNA, we observe a significant increase in the correlation between cfDNA originating from leukocytes and from non-leukocyte sources in cancer patients. Introducing a mathematical model for cfDNA dynamics, we find that the observed correlation can be explained by a saturation mechanism in cfDNA clearance. Saturation in clearance implies that smaller increases in cfDNA shedding may lead to proportionally larger increases in cfDNA levels. Our findings quantify cfDNA dynamics in patients with cancer, with implications for improving the accuracy of liquid biopsies for early cancer detection.

Introduction

Successful detection of cancer in early stages results in more effective treatment, and significantly improved survival prognosis. Identifying biomarkers that can reliably indicate the presence of early-stage cancers is one of the major challenges in oncology [1]. One such promising biomarker is cell-free DNA (cfDNA), composed of the DNA fragments not bound to cells, that can be detected in the circulation. The presence of cfDNA in the blood was first discovered by Mandel and Metais [2] in the 40's; since then, increased levels of cfDNA concentrations have been reported in cancer patients [3–5], especially in advanced and metastatic stages [6]. Furthermore, Stroun *et al.* [7] uncovered in the 80's that cfDNA from cancer patients harbored mutations that were also present in patients' tumors.

These discoveries motivated further research on detecting cancer by performing "liquid biopsies" on cfDNA, as a less intrusive alternative to the biopsy on the tumor itself. However, the mechanisms by which DNA is shed into, and subsequently cleared from the circulation are not fully understood [8–10].

The cfDNA clearance occurs primarily in the liver (70%-90%), mediated by Kupffer cells, with additional contributions from kidneys and the spleen [11]. Clearance is rapid, with reported cfDNA half-life estimates ranging from approximately 15 minutes to 2.5 hours [12–15]. In settings of impaired cfDNA clearance, such as sepsis [16] or, in the cancer context, the pharmacologic attenuation of cfDNA clearance in tumor-bearing mice [17], cfDNA levels rise without a significant change in composition (tissue-of-origin in sepsis; tumor fraction in mouse experiments), indicating that clearance acts non-selectively on cfDNA. Importantly, clearance capacity is finite and becomes saturated at high concentrations, as demonstrated by dose-dependent hepatic uptake of radiolabeled mononucleosomes in mice [18]. These mechanisms are critical for interpreting cfDNA dynamics under cancer conditions, yet their quantitative impact remains incompletely characterized.

Recent work demonstrates that the majority of cfDNA in cancer patients with highly elevated cfDNA originates from leukocytes, and that the presence of cancer likely results in a systemic increase in cfDNA concentration in these patients [19]. Whether a similar systemic increase due to cancer is present in all cancer patients is currently not known, as many patients, in particular those with early-stage cancer, do not exhibit elevated cfDNA levels.

In this setting, several questions must be answered in order to better understand the mechanics of DNA circulating in plasma, which could lead to improving the accuracy of liquid biopsy screening tests. In what way, and how much does cancer presence affect the total cfDNA concentration? Is this effect cancer type-specific, and does it differ between early cancer stages? In the present study, we address these questions through a novel analysis of a large cfDNA dataset from Cohen *et al.* [20], and quantify the differences in cfDNA dynamics between healthy individuals and patients with stage I-III ovarian, liver, stomach, pancreatic, esophageal, colorectal, lung, and breast cancer. Employing an additional dataset from Mattox *et al.* [19] reporting the tissue of origin for cfDNA in healthy individuals and pancreatic and ovarian cancer patients, we are able to provide insight regarding the mechanisms by which cfDNA is cleared from circulation, and estimate the increases in cfDNA shedding from leukocytes and non-leukocyte sources due to cancer.

Results

Cohen et al. [20] reported cfDNA concentrations in plasma for healthy individuals and cancer patients, as well as the concentration and mutant allele frequency (MAF) of DNA fragments in plasma harboring mutations in cancer driver genes; we refer to this portion of cfDNA as circulating mutant DNA (cmDNA) (Materials and Methods, Table S1). Among the healthy individuals included in our analysis, cmDNA was

detected in 90.6% of samples, with a median cfDNA concentration of 2.83 ng/mL, whereas samples without detectable mutations showed much lower cfDNA (median 0.05 ng/mL; S1 Fig.). As we aim to quantify and compare the effect of cancer on both cfDNA and cmDNA concentrations, we focus on the 90.6% of healthy individuals with cmDNA detected in plasma. Full exclusion criteria and cohort sizes for healthy individuals and cancer patients are provided in Materials and Methods.

Multiplicative increase in cfDNA originating from healthy tissue

Patients with late-stage cancer typically have cfDNA concentrations that are significantly higher than those in healthy individuals [6,10]. However, cfDNA concentrations in patients with early-stage cancer often overlap with those found in healthy individuals (Fig. 1). Performing the Kolmogorov-Smirnov (K-S) statistical test to determine the similarity of probability distributions of cfDNA concentration in healthy individuals and cancer patients, we find that the presence of stage I-III lung, breast, colorectal, pancreatic, ovarian, esophageal, stomach and liver cancer leads to a multiplicative increase in cfDNA concentration compared to healthy individuals (Fig. 2). In other words, we find that the entire dataset of cfDNA concentrations is multiplied by a positive factor α in patients with cancer compared to healthy individuals (Kolmogorov-Smirnov test at 5% significance level, Materials and Methods). This effect can be seen as a horizontal "shift" or translation when the cumulative distribution function (CDF) is plotted on a logarithmic x-scale (Fig. 2).

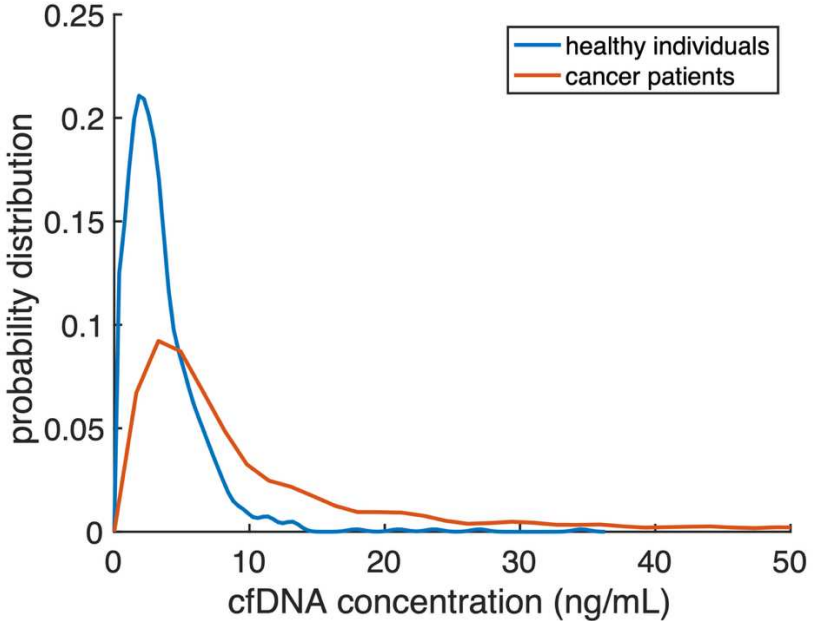


Figure 1. Probability distribution of cfDNA concentrations in healthy individuals and patients with stage I-III cancer. Probability density functions for the cfDNA concentration in healthy individuals with mutations detected in plasma (N=572) and patients with lung, breast, colorectal, pancreatic, ovarian, esophageal, stomach and liver cancer in stages I-III (N=952), as reported in Cohen *et al.* (2018) [20].

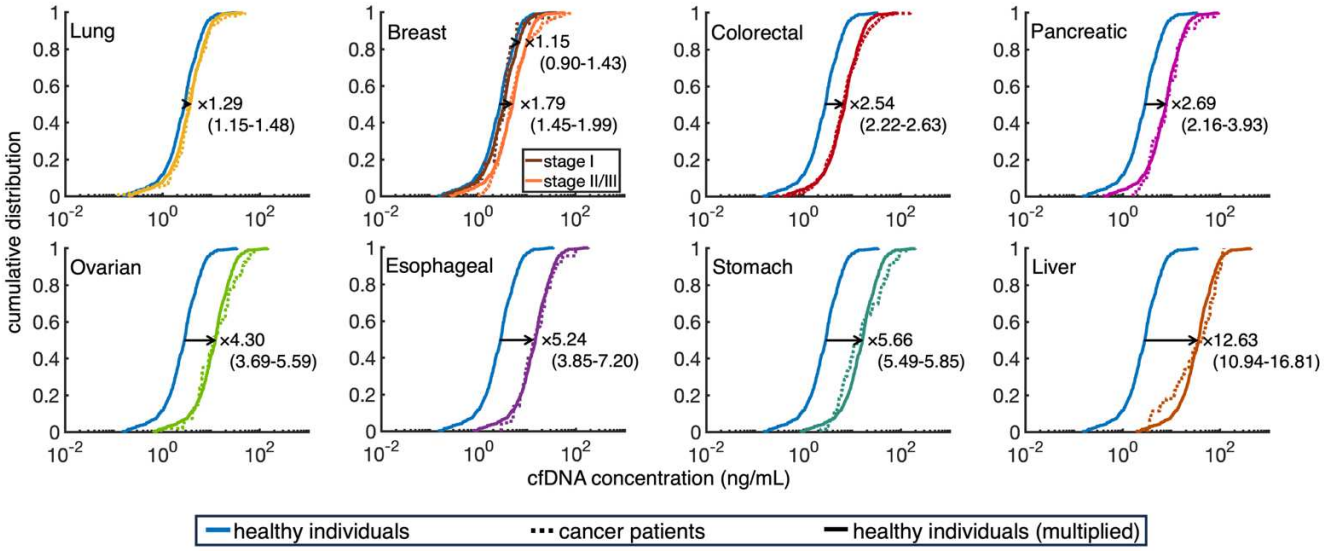


Figure 2. Quantifying the multiplicative increase α in cfDNA concentration in the presence of cancer. Cumulative distribution functions for the cfDNA concentration in healthy individuals with mutations detected in plasma (N=572) and patients with lung (N=103), stage I breast (N=32), stage II/III breast (N=177), colorectal (N=385), pancreatic (N=47), ovarian (N=54), esophageal (N=44), stomach (N=66) and liver (N=44) cancer. The value and range shown on each panel represent the multiplicative increase α and its range of values not rejected by the K-S test at the 5% significance level. Median empirical cfDNA concentrations (corresponding to 0.5 value of cumulative distribution) are 2.83 ng/mL for healthy individuals, and 3.31, 3.42, 4.60, 6.62, 7.76, 11.18, 13.97, 14.14, and 42.27 ng/mL for patients with lung, stage I breast, stage II/III breast, colorectal, pancreatic, ovarian, esophageal, stomach and liver cancer, respectively.

The multiplicative increase α is specific to each cancer type: 1.29 (value range not rejected by K-S test at 5% significance level: 1.15-1.48) for lung cancer, 1.15 (0.9-1.43) for stage I breast cancer, 1.79 (1.45-1.99) for stage II/III breast cancer, 2.54 (2.22-2.63) for colorectal, 2.69 (2.16-3.93) for pancreatic, 4.3 (3.69-5.59) for ovarian, 5.24 (3.85-7.2) for esophageal, 5.66 (5.49-5.85) for stomach, and 12.63 (10.94-16.81) for liver cancer (Fig. 1, S2 Fig., S2 Table, Materials and Methods). Additionally, as the visual counterpart of the K-S test, we construct the Probability–Probability (P–P) plots, which compare the CDFs of two datasets (Materials and Methods). For each cancer type, the P–P plots show strong concordance between the empirical CDFs for cfDNA concentrations in cancer patients and in healthy individuals multiplied by α (Fig. S2). P-P plots for stomach and liver cancer diverge slightly from the diagonal due to greater differences at the distribution tails between cancer and scaled healthy CDFs (Fig. 2). P-P plots amplify tail effects, but the overall agreement supports the validity of the shift.

There are no statistically significant differences in cfDNA concentration between stages I to III, for all cancer types except for breast and liver cancer; however, the data for stage I pancreatic, esophageal and liver cancer is limited (2, 5 and 5 patients respectively). We found no significant differences in cfDNA concentrations between sexes. Although the median age of healthy individuals was younger than that of cancer patients (36.5 years compared to 57–68 years; Table S1b), there was no statistical difference in cfDNA levels between the 288 healthy individuals under 40 years and the 263 healthy individuals over 50 years of age (K-S test $p > 0.05$, S3 Figure). We selected the < 40 y and > 50 y age groups of healthy individuals to reflect the previously reported median ages (36.5 y in healthy individuals and 57-68 y in

cancer patients) and to test whether an age gap of that size is associated with significantly different cfDNA levels within healthy individuals. The intermediate 40-50 y group is small (n = 21) and was not analyzed separately. Furthermore, we grouped all healthy individuals into ≤ 40 y and > 40 y age groups; again, the Kolmogorov-Smirnov test showed no significant differences in cfDNA levels between these two age groups.

To determine if a multiplicative increase in cfDNA concentration is also present in the earliest cancer cases, we calculate α between cfDNA concentrations in healthy individuals and cancer patients, restricting the cancer data to only stage I cases. We do not perform the calculation for pancreatic, esophageal and liver cancer, due to insufficient number of patients in stage I. Breast cancer patients have already been analyzed separately in stage I and stage II-III due to the significant difference in cfDNA concentrations between these groups. We find that α is 1.33 (1.04-1.71) for lung cancer, 2.39 (2.02-2.88) for colorectal cancer, 6.99 (3.64-14.82) for ovarian cancer, and 4.83 (2.27-6.38) for stomach cancer (S4 Fig.). These ranges for α overlap with the ranges calculated when all stage I-III cancer patients were considered. We note that, for ovarian cancer, the multiplicative factor is 6.99 (3.64–14.82) for stage I compared to 4.30 (3.69–5.59) for stages I–III. The apparently higher factor for stage I ovarian cancer reflects the small sample size (N = 9), where a few patients with elevated cfDNA disproportionately influence the estimate, resulting in a substantially wider range. Importantly, the ranges for stage I overlap with those for stages I–III, demonstrating consistency of the multiplicative factor between stage I and all stages combined. Additionally, cfDNA concentrations do not differ significantly between stage I and all stages combined in all four cancer types considered.

We next asked if the source of the increased cfDNA concentration in early-stage cancer patients is the cancer itself or if the increase is originating from the healthy tissue. To address this, we analyzed 484 patients from the Cohen et al. [20] dataset whose tumors harbored at least one high-frequency mutation (MAF ≥30%). MAF in plasma reflects both the clonality of the mutation in its tissue of origin and that tissue's contribution to total cfDNA. For example, if a mutation has a MAF in the tumor of 30%, and if 50% of cfDNA originates from the tumor, then, on average, the mutation should have MAF around 15% in plasma. Even if the biopsy overrepresents the mutation compared to the entire tumor -for example, MAF of 30% in the biopsy versus 10% in the tumor- if 50% of cfDNA originates from the tumor, the expected plasma MAF would still be around 5%. The exact numbers may be different in some cases due to differences in copy number between cancer and healthy tissue, but are expected to exhibit similar behavior. We recognize that clearance dynamics could influence absolute cfDNA levels; however, experimental evidence indicates that impaired clearance increases cfDNA globally without altering its composition [16,17], suggesting that clearance differences do not systematically bias the relative contribution of tumor versus healthy tissue to plasma cfDNA. Therefore, if most of cfDNA originates from the tumor, the high-MAF mutations in the tumor should appear in plasma with significant MAF (≫1%). However, in these 484 patients, only 3.1% of top plasma mutations in cancer patients have MAF above 10%, 5.8% above 5%, and 13.4% above 1% (S5 Fig., red line). These findings indicate that, even when tumors carry high-clonality mutations, most cfDNA in plasma does not originate from the tumor but from non-tumor (healthy) tissue.

The expectation that a mutation with high MAF in the tumor should be present in plasma with significant MAF (≫1%), assuming that most of cfDNA originates from the tumor, holds even if the mutation was missed by the tumor biopsy, provided it is theoretically detectable in plasma (i.e. included in the mutational panel). For that reason, we also examined the distribution of plasma mutation MAFs for the 162 cancer patients whose tumors harbored only mutations with MAF < 30% (S5 Fig., blue line). The results were similar to the distribution of plasma MAFs for patients with tumor MAF≥30%; in fact, the two groups were indistinguishable, as confirmed by a Kolmogorov–Smirnov test ($p > 0.05$).

Multiplicative increase in non-tumor circulating mutant DNA

We next hypothesized that, since cfDNA is mainly shed from non-tumor sources and exhibits a multiplicative increase in the presence of cancer, then a similar increase should also be observed in circulating mutant DNA (cmDNA) that does not originate from tumor. In the dataset reported by Cohen et al. [20], which includes a total of 959 cancer patients with stage I–III disease across eight cancer types, information on both the top mutation detected in plasma and the mutations identified in tumor is available for 645 patients (Materials and Methods, S1c Table). The cmDNA concentration reported by Cohen et al. [20] corresponds to the top plasma mutation. When the top mutation detected in plasma is not found among the mutations reported in the tumor biopsy, we conclude that this mutation most likely does not originate from the tumor, but rather from healthy tissue. Comparing the cmDNA concentration in healthy individuals and in cancer patients in which cmDNA detected in plasma is not detected in tumor (S1c Table), we find increases by a multiplicative factor for every cancer type: 1.92 (1.42-2.52) for lung, 2.42 (2.12-2.76) for breast, 2.36 (1.92-2.66) for colorectal, 4.01 (1.55-7.34) for pancreatic, 3.31 (1.76-5.38) for esophageal, 4.38 (2.74-5.77) for stomach, and 12.37 (5.54-24.62) for liver cancer (Fig. 3, S6 Fig., S3 Table, Materials and Methods). We do not calculate the multiplicative factor for ovarian cancer, since the dataset reported only one ovarian cancer patient with cmDNA not originating from the tumor.

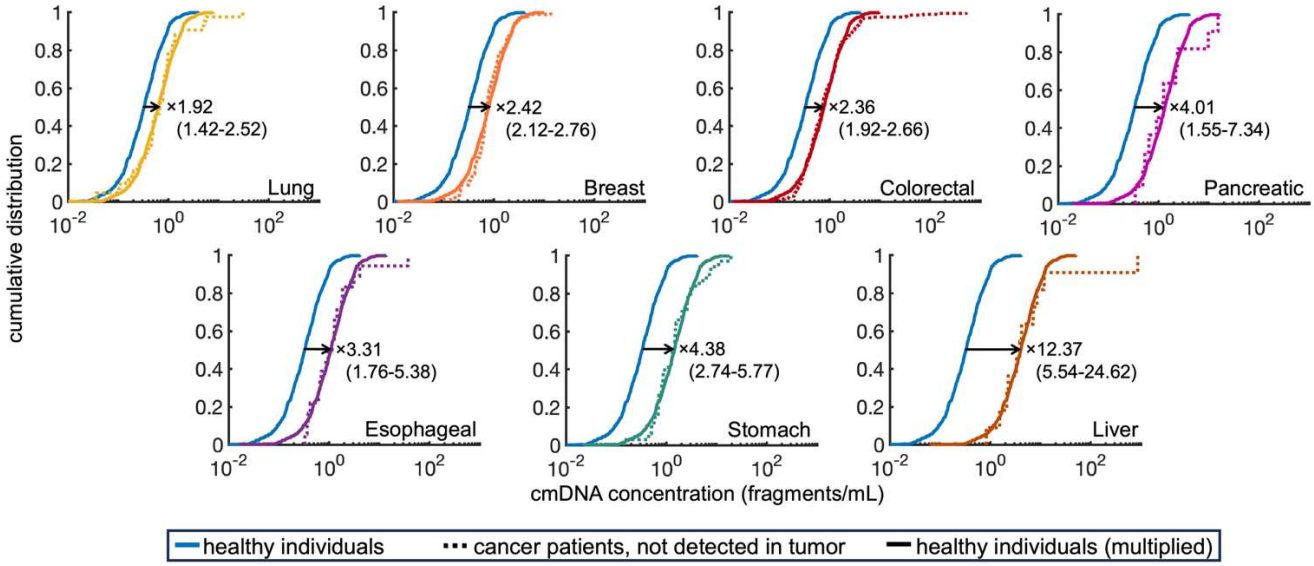


Figure 3. Mutations detected in plasma of cancer patients but not detected in tumor exhibit a multiplicative increase in concentration compared to mutations detected in plasma of healthy individuals. Cumulative distribution functions for the concentration of cmDNA found in plasma of healthy individuals (N=572) and patients with lung (N=43), breast (N=90), colorectal (N=205), pancreatic (N=11), esophageal (N=18), stomach (N=34) and liver (N=11) cancer in which the mutation in plasma was not detected in tumor. The value and range shown on each panel represent the multiplicative increase α and its range of values not rejected by the K-S test at the 5% significance level. Median empirical cmDNA concentrations (corresponding to 0.5 value of cumulative distribution) are 0.33 fragments/mL for healthy individuals, and 0.67, 0.73, 0.72, 1.23, 1.05, 1.39, and 3.85 fragments/mL for patients with lung, breast, colorectal, pancreatic, esophageal, stomach and liver cancer, respectively.

Mutant DNA in plasma may originate from healthy or tumor tissue; however, parallel sequencing polymerase chain reaction (PCR) methods may introduce artifactual mutations with low MAFs. Circulating mutant DNA data from Cohen *et al.* [20] which we analyze in this work were generated using the SafeSeqS approach for PCR, developed to reduce the introduction of artifactual mutations [21]. In subsequent work [22], Cohen *et al.* report that the artifactual mutations in plasma analyzed by SafeSeqS method appeared with a MAF lower than 0.013%. Hence, we perform the cmDNA analysis again, disregarding the mutations with low MAFs; the results remain similar to the results obtained for all mutations (S7 Fig., S3 Table, Materials and Methods).

Insight into cfDNA dynamics

The multiplicative increase in cfDNA levels observed between healthy and cancer patients can be the result of increased cfDNA shedding from non-tumor sources, a slower rate of cfDNA clearance, or a combination of both. In order to provide insight regarding the cfDNA dynamics and the cause of cfDNA increase in early-stage cancer, we utilize data from Mattox *et al.* [19] regarding the cfDNA tissue of origin for 64 healthy individuals, 79 patients with stage I-III pancreatic cancer, and 32 patients with stage I-III ovarian cancer. The identification of cfDNA tissue of origin in Mattox *et al.* [19] was based on the deconvolution of the cfDNA methylation profile using two different reference matrices for the tissue-specific methylation regions, as determined by Sun *et al.* [23] and Moss *et al.* [24], and two different deconvolution algorithms (quadratic programming or non-linear least squares). This results in four different cfDNA compositions reported for each individual from this dataset.

First, for this dataset, we perform the same statistical analysis to determine the multiplicative factor between healthy and cancer patients, finding an α of 4.4 (3.8-4.81) and 2.5 (2.1-3.33) for pancreatic and ovarian cancer respectively. The ranges of multiplicative factor values as calculated from the datasets of Mattox *et al.* [19] and Cohen *et al.* [20] are overlapping for pancreatic cancer; for ovarian cancer they are comparable yet non-overlapping, with a lower multiplicative factor calculated from the dataset of Mattox *et al.* [19].

In healthy individuals from Mattox *et al.* [19], most cfDNA originates from leukocytes; the different deconvolution methods employed report a median 66.8%-78% of cfDNA from leukocytes. In pancreatic and ovarian cancer patients, the cfDNA percentage from leukocytes is similar (median 65.1%-70.8% and 70.3%-75.5% respectively), in cases where the reference deconvolution matrix by Sun *et al.* [23] is used. Use of the reference deconvolution matrix by Moss *et al.* [24] results in decreased percentages of cfDNA from leukocytes (~53% in pancreatic cancer and 52%-58.5% in ovarian cancer patients).

From this dataset, we also calculate the Pearson correlation coefficient for cfDNA concentrations from leukocytes and non-leukocyte sources, after we exclude outliers (three outliers with high cfDNA concentrations from healthy individuals and one from ovarian cancer patients, S1 Text). Note that the exclusion of outliers does not have a significant effect on the multiplicative factors calculated above. Correlation coefficient in healthy individuals is calculated from -0.29 to 0.20, depending on the cfDNA deconvolution method, 0.31-0.58 in pancreatic cancer patients, and 0.32-0.87 in ovarian cancer patients. Thus, apart from the multiplicative increase in total cfDNA concentration in cancer patients, we also observe an increase in the correlation between cfDNA concentrations originating from leukocytes and non-leukocyte sources, from weakly (anti-)correlated in healthy individuals, to significant positive correlation in cancer patients (Fig. 4a).

Stratifying by stage, we find correlations of 0.57–0.80 for stage I (N=13), 0.29–0.58 for stage II (N=60), and 0.33–0.51 for stage III (N=6) in pancreatic cancer, and 0–0.75 for stage I (N=8), 0.46–0.89 for stage II (N=5), and 0.33–0.88 for stage III (N=18) in ovarian cancer. These results indicate no consistent increase in correlation with stage; in fact, pancreatic cancer shows the highest correlation in stage I. Because stage-specific sample sizes are small, correlation estimates may exhibit high variability. To mitigate this,

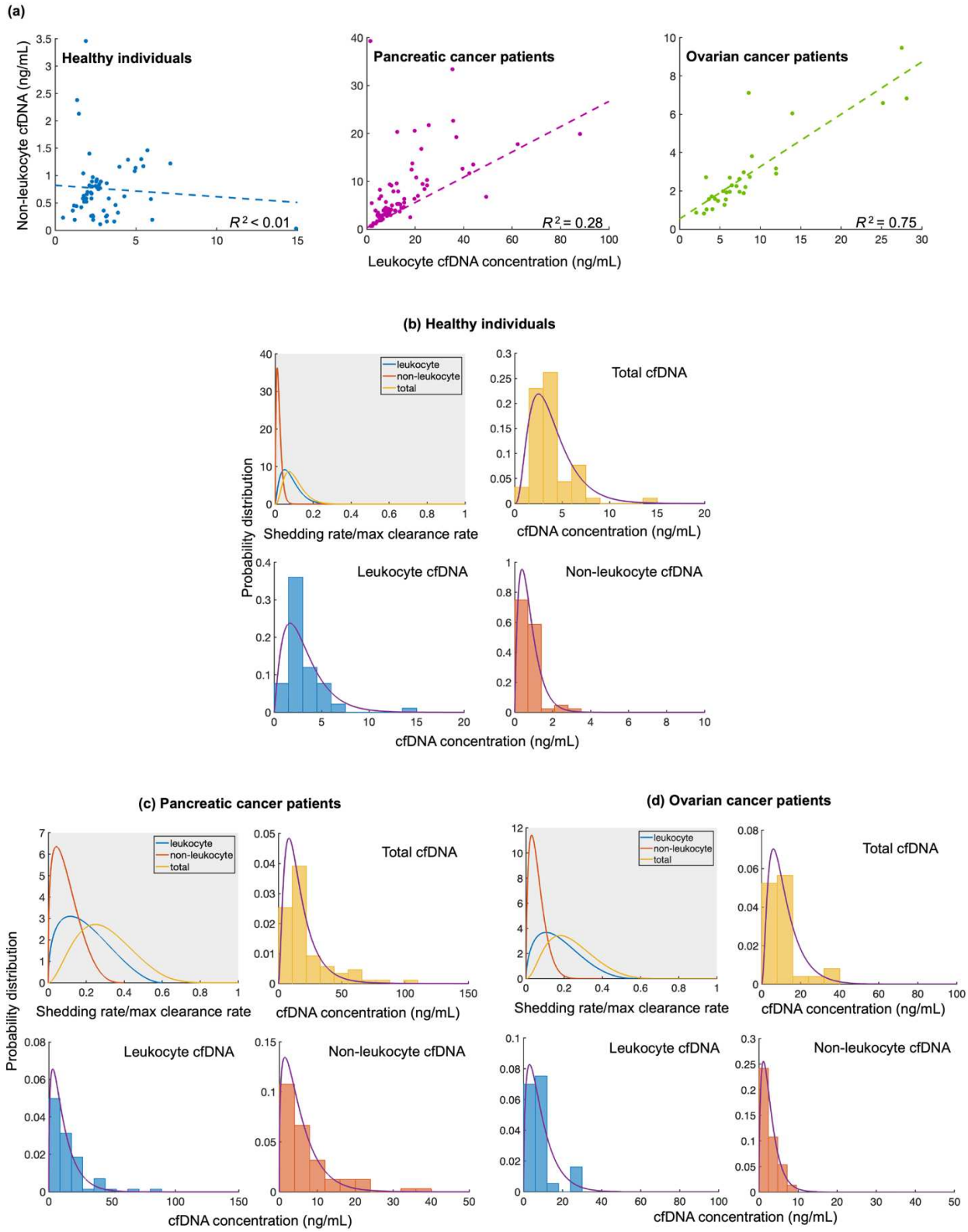


Figure 4. Correlation and distributions of the leukocyte, non-leukocyte and total cfDNA concentrations. (a) Correlation between leukocyte and non-leukocyte concentrations, excluding outliers with high cfDNA concentration. Dots represent the pair of leukocyte and non-leukocyte cfDNA

concentration for normal individuals and pancreatic and ovarian cancer patients from Mattox *et al.* [19] under a quadratic programming deconvolution algorithm employing Sun *et al.* [23] reference matrix. **(b), (c), (d)** Gray panel: Estimated leukocyte $\tilde{S}_l$, non-leukocyte $\tilde{S}_{nl}$ and total cfDNA $\tilde{S}_{tot}$ shedding rates normalized by the maximum clearance rate; $\tilde{S}_l = S_l/v$, $\tilde{S}_{nl} = S_{nl}/v$, $\tilde{S}_{tot} = S_{tot}/v$, considered for the case of healthy individuals (b), pancreatic cancer patients (c), and ovarian cancer patients (d). White panels: Saturation model results for the distribution of leukocyte, non-leukocyte and total cfDNA concentrations are depicted by purple curves. Histograms represent data for cfDNA concentrations from Mattox *et al.* [19], under the same deconvolution method as for (a).

we performed a two-dimensional Kolmogorov–Smirnov (2D KS) test (Materials and Methods) on paired leukocyte and non-leukocyte cfDNA concentrations for each cancer type and deconvolution method. For samples of the same cancer type and deconvolution method but of different stages, the 2D KS test did not reject the null hypothesis that they originate from the same joint distribution, suggesting that cfDNA composition does not significantly change across stages I-III in pancreatic and ovarian cancer.

We formulate a mathematical model for cfDNA dynamics in order to obtain insight into the changes in cfDNA shedding or clearance that could lead to the observed multiplicative increase in total cfDNA concentration. The cfDNA dynamics model should also be able to explain the increase in correlation between cfDNA concentrations originating from leukocytes and non-leukocyte sources observed in cancer patients.

We first consider a model for cfDNA dynamics with linear clearance

$$C_l' = S_l - \varepsilon C_l,\ C_{nl}' = S_{nl} - \varepsilon C_{nl},\ C_{tot}' = S_{tot} - \varepsilon C_{tot}, \tag{1a,b,c}$$

where C_l, C_{nl} are the cfDNA concentrations and S_l, S_{nl} are the rates of cfDNA shedding from leukocytes and non-leukocyte sources respectively. $C_{tot} = C_l + C_{nl}$ and $S_{tot} = S_l + S_{nl}$ are the total cfDNA concentration and the total cfDNA shedding rate. In Eqs. (1), ε denotes the rate of cfDNA clearance; it is related to cfDNA half-life $t_{1/2}$ as $\varepsilon = ln\ (2)/t_{1/2}$. We assume cfDNA dynamics reach a quasi-steady state between shedding and clearance. This assumption is supported by the rapid cfDNA clearance (half-life 15 minutes to 2.5 hours [5,6,13–15]), compared to tumor growth and therapy-driven changes in shedding rates, which occur over days to weeks, as reported in longitudinal studies of circulating tumor DNA (ctDNA) dynamics during cancer progression and therapy response [25–28]. Assuming quasi-steady state is reasonable for modeling purposes and has also been adopted in previous mechanistic studies [13,27]. Moreover, as noted earlier, two-dimensional Kolmogorov–Smirnov tests on the Mattox et al. [19] dataset indicate no significant differences in cfDNA levels or composition from leukocyte and non-leukocyte sources across stages I–III for pancreatic and ovarian cancer. Given the rapid clearance and the absence of stage-specific variation, it is reasonable to assume that cfDNA levels are at quasi-steady state. Under this assumption, the cfDNA concentrations calculated from the linear model (1) are

$$C_l = S_l/\varepsilon,\ C_{nl} = S_{nl}/\varepsilon,\ C_{tot} = S_{tot}/\varepsilon. \tag{2a,b,c}$$

Under Eqs. (2) of the linear clearance model, the observed increases in cfDNA concentrations from leukocytes (C_l) and non-leukocyte sources (C_{nl}) may originate from corresponding increases in cfDNA shedding rates or from a decrease in the clearance rate ε. From the data by Mattox *et al.* [19], we calculate the magnitude of the multiplicative increases in the distributions of leukocyte and non-leukocyte cfDNA: 3.73 (3.15-4.45) and 7.99 (6.24-9.94) for pancreatic cancer, and 2.53 (2.03-3.37), 3.29 (2.39-4.49) for ovarian cancer. A decrease in cfDNA clearance rate should in principle increase both leukocyte and non-leukocyte cfDNA by the same factor. In pancreatic cancer, the multiplicative increase in cfDNA from non-leukocyte sources is significantly higher than from leukocytes. We deduce that the increased cfDNA

concentration in the presence of pancreatic cancer does not originate from decreased cfDNA clearance only, but stems at least partly from increased cfDNA shedding from non-leukocyte and potentially also leukocyte sources.

We next investigated if the linear model of cfDNA dynamics can account for the observed increase in correlation between leukocyte and non-leukocyte cfDNA in cancer patients. We first note that a decrease in cfDNA clearance cannot explain increased correlation under the linear model (1). On the other hand, increases in cfDNA shedding from leukocyte and non-leukocyte sources may be correlated as they may both be affected by the presence of cancer. To capture this, we allow the multiplicative factors for leukocyte and non-leukocyte cfDNA shedding to vary across patients according to uniform distributions over the ranges of values determined previously from the KS test. We further assume that the multiplicative factors are perfectly linearly correlated, so that cancer presence increases shedding from both sources simultaneously. We provide an analytical calculation of the Pearson correlation coefficient for leukocyte and non-leukocyte cfDNA concentrations in cancer under the linear clearance model after applying multiplicative shedding factors to cfDNA concentrations from healthy individuals (Supplementary Text S2). From it we obtain the result that, even with variability in shedding increases across patients and perfect correlation between the shedding factors, the linear clearance model predicts the same correlation coefficient (from -0.29 to 0.20) between leukocyte and non-leukocyte cfDNA in cancer patients as in healthy individuals, for the four deconvolution methods. Therefore, the linear clearance model cannot easily account for the observed increase in cfDNA correlation in cancer patients.

As the linear model (1) is insufficient to explain the correlation observed in data, we consider a model with saturation in cfDNA clearance:

$$C_l' = S_l - \frac{vC_l}{k+C_{tot}},\ C_{nl}' = S_{nl} - \frac{vC_{nl}}{k+C_{tot}},\ C_{tot}' = S_{tot} - \frac{vC_{tot}}{k+C_{tot}}, \tag{3a,b,c}$$

with v and k being positive constants. In Eq. (3c) for the total cfDNA concentration, clearance term is of Michaelis-Menten type [29]: for low cfDNA concentrations, the clearance term is approximately linear with respect to cfDNA concentration, with clearance rate v/k; as the total cfDNA concentration increases, the clearance term approaches v, which is the maximum cfDNA clearance rate. The Michaelis-Menten clearance term is able to model the finite capacity of Kupffer liver cells where most of cfDNA (70%-90%) is cleared from circulation [11,12]. In Eqs. (3a,b), the clearance rate for leukocyte and non-leukocyte cfDNA is the same and equal to the clearance rate $v/(k + C_{tot})$ for the total cfDNA. This is based on the assumption that all cfDNA fragments are well mixed in blood circulation, and that all undergo the same clearance process, regardless of their tissue of origin. From Eqs. (3), the equilibrium values of leukocyte, non-leukocyte and total cfDNA concentrations in plasma are given by:

$$C_l = \frac{k\tilde{S}_l}{1-\tilde{S}_{tot}},\ C_{nl} = \frac{k\tilde{S}_{nl}}{1-\tilde{S}_{tot}},\ C_{tot} = \frac{k\tilde{S}_{tot}}{1-\tilde{S}_{tot}}, \tag{4a,b,c}$$

where $\tilde{S}_l$, $\tilde{S}_{nl}$, $\tilde{S}_{tot}$ are the leukocyte, non-leukocyte and total shedding rates normalized by the maximum clearance rate; $\tilde{S}_l = S_l/v$, $\tilde{S}_{nl} = S_{nl}/v$, $\tilde{S}_{tot} = S_{tot}/v$. In our analysis, we impose that the distribution of the total normalized shedding rate $\tilde{S}_{tot}$ remains below 1. This reflects the physiological requirement that total cfDNA shedding should not exceed the maximum clearance capacity, as doing so would lead to unbounded cfDNA accumulation in circulation. Furthermore, normalization provides a dimensionless scale (0–1), which enables direct comparison between shedding distributions in healthy and cancer conditions and facilitates clear interpretation and quantitative assessment of how cfDNA shedding from healthy tissue shifts in the presence of stage I-III cancer (Fig. 4). Under the saturation in clearance model (3), the correlation between leukocyte and non-leukocyte cfDNA concentrations is positive and increases as the average normalized shedding rate from leukocytes and/or non-leukocyte sources is increased (Eq. S13 in S3 Text).

To illustrate the ability of the saturation model (3) to explain the observed cfDNA dynamics, we determine model parameters that allow good agreement with observed data. First, by considering a weak correlation between leukocyte and non-leukocyte cfDNA concentrations in healthy individuals, we can estimate the model parameter k, as well as the mean values and correlations of variation (CV -standard deviation over average) of normalized cfDNA shedding rates from leukocytes and non-leukocyte sources (Materials and Methods). We choose rescaled Beta distributions to model the normalized shedding rates, in order to ensure that the maximum total shedding rate is less than maximum clearance rate. We estimate the values of distribution parameters using the estimated mean values and coefficients of variation for the normalized shedding rates from leukocytes and non-leukocyte sources in healthy individuals. Substituting the estimate for k and the fitted distributions for $\tilde{S}_l$, $\tilde{S}_{nl}$ into Eqs. (4), we derive distributions for leukocyte, non-leukocyte and total cfDNA concentrations (Materials and Methods) that are in agreement with the data from Mattox *et al.* [19] for healthy individuals (K-S test p-value>0.05 for all distributions, Fig.4b). Furthermore, the correlation between leukocyte and non-leukocyte cfDNA concentrations predicted by the model is weak, ~0.1, in agreement with the correlation -0.29 to 0.20 calculated for healthy individuals from Mattox *et al.* [19].

By considering a 2.5-fold increase in the average leukocyte shedding rate and a 5-fold increase in the average non-leukocyte shedding rate, saturation model (3) predicts distributions for leukocyte, non-leukocyte and total cfDNA concentrations that are in agreement with the data for pancreatic cancer patients (K-S test p-value>0.05 for all distributions, Fig.4c). Furthermore, it predicts a correlation coefficient 0.46 between leukocyte and non-leukocyte cfDNA concentrations, which is close to the median of the value range 0.31-0.58 calculated for pancreatic cancer patients in Mattox *et al.* [19] for different deconvolution methods. The aforementioned increases in the leukocyte and non-leukocyte shedding rates add up to a 3-fold increase in the average total shedding rate, and they account for a 4.4, a 3.6 and a 7.1-fold increase in the average total, leukocyte, and non-leukocyte cfDNA concentrations respectively. The resulting increases in cfDNA concentrations are greater than the respective increases in shedding rates, due to the saturation effect.

Similarly, by considering 2.2-fold and 3-fold increases in the average leukocyte and non-leukocyte shedding rates respectively, we obtain from model (3) distributions for all cfDNA concentrations that are consistent with the data for ovarian cancer patients (Fig. 4d). The correlation coefficient calculated from the model is 0.32, which is at the lower end of the value range 0.32-0.87 calculated for ovarian cancer patients in Mattox *et al.* [19] for different deconvolution methods.

ROC curves for cancer detection using cfDNA level

Finally, in order to determine the implications of multiplicative cfDNA increase in cancer detection, we constructed the Receiver Operating Characteristic (ROC) curves for each cancer type from [20] individually (S8 Fig., S4 Table). ROC curves were generated using cfDNA concentration as the sole predictor, with thresholds scanned across its range, and constructed using 10-fold cross-validation with bootstrap resampling (Materials and Methods). Performance estimates were stable, as indicated by narrow confidence intervals obtained. We observe that, in cancer types with high multiplicative factors for cfDNA concentrations, such as ovarian, esophageal, stomach, and liver cancer, the area under the curve (AUC) approaches 1, being equal to or greater than 0.9. For lung, colorectal and stomach cancers where sufficient numbers of stage I cases were available, ROC curves and performance metrics were generally similar to that of the full cohort, indicating that cfDNA concentration retains its predictive value in early-stage disease. For ovarian cancer, AUC improved when considering only stage I cases, but this result should be interpreted with caution due to the small sample size (9 patients). Across all cancer types and stages, negative predictive value (NPV) remained consistently high (>0.85), underscoring the reliability of cfDNA concentration for ruling out disease even in early-stage cancers.

Discussion

Our results indicate that cfDNA shed from healthy tissue exhibits a multiplicative shift towards higher concentrations in the presence of early-stage cancer; this increase is quantifiable and ranges from 1.3-fold in lung cancer to 12-fold increase in liver cancer. The agreement between the cancer-type specific multiplicative factors for total cfDNA concentration, and the concentration of circulating mutant DNA not originating from tumor, demonstrates the consistency of our finding.

We categorized cfDNA mutations in cancer patients as tumor or non-tumor derived based on whether the mutation detected in plasma was also detected in the tumor. Mutations detected in plasma but not in the tumor could still originate from the tumor if the biopsy missed regions harboring these mutations, and large-scale cfDNA clearance could further reduce plasma MAF relative to tumor MAF. While these possibilities cannot be excluded for individual cases, their overall impact appears limited:

When comparing plasma MAF distributions across three groups -mutations detected in plasma of healthy individuals, mutations detected in plasma but not in the tumor of cancer patients, and mutations detected in both plasma and tumor- we find that mutations not detected in the tumor cluster closely with mutations from healthy individuals (Kolmogorov–Smirnov test, $p > 0.05$; S9 Fig.). This concordance suggests that most plasma mutations not detected in the tumor indeed originate from non-tumor, healthy tissue.

Published data reinforce this interpretation: tumor-derived ctDNA in stage I–III cancers typically appears at 0.3–0.9% MAF [6,30], matching the median 0.3% MAF of the group of tumor-detected mutations from Cohen et al. [20]. On the other hand, cfDNA related to clonal hematopoiesis of indeterminate potential (CHIP), a major source of plasma mutations from healthy tissue, has MAF <0.1% [31,32]. In the Cohen et al. dataset [20], 92.5% of plasma mutation MAFs in healthy individuals and 87.4% of MAFs of plasma mutations not detected in the tumor of cancer patients were below 0.1%.

After we observed the multiplicative increase in cfDNA between healthy individuals and cancer patients, we asked about its potential diagnostic implications. Thus, we assessed the predictive performance of cfDNA concentration through receiver operating characteristic curves. For comparison purposes, we relate our findings to CancerSEEK, a multi-analyte blood test that integrates circulating-tumor DNA mutations and eight protein biomarkers [20]. CancerSEEK achieved an overall AUC of 91% (90–92%) when distinguishing cancer patients from healthy controls across the eight cancer types; at a fixed specificity >99%, the test reported a median sensitivity of 70% overall and 43% for stage I cancers, with variation by cancer type (e.g., 98% for ovarian, 100% for liver, and 33% for breast). In our analysis, cfDNA levels in plasma as the sole predictor produced AUC values above 0.90 for ovarian (0.901), stomach (0.918), esophageal (0.941), and liver cancers (0.951). For cancer types with sufficient number of stage I cases, we observed no major differences in sensitivity and specificity between all stages combined and stage I, suggesting that cfDNA concentration retains its discriminative ability even at early stages. Notably, our analysis shows that cfDNA concentration alone has limited predictive value in lung cancer. This is consistent with findings from the CancerSEEK study, which likewise reported lower sensitivity for lung cancer compared with most other cancer types. Differences in absolute specificity between CancerSEEK and our analysis reflect the fact that CancerSEEK optimized for ultra-high specificity, whereas our cutoff cfDNA levels balanced sensitivity and specificity. Overall, this comparison highlights that cfDNA concentration alone, without considering the mutations it harbors, can provide substantial diagnostic information, underscoring its potential as a simple and cost-effective component of liquid biopsy strategies for early cancer detection.

We also analyze an additional dataset [19] that reports the tissue of origin of cfDNA in healthy individuals and patients with pancreatic and ovarian cancer. From this dataset, we also observe another effect of cancer on cfDNA: a significant increase in the correlation between the cfDNA concentrations originating from leukocytes and non-leukocyte sources. Introducing simple models for cfDNA dynamics, we demonstrate

that the linear clearance model, under which the increase in cfDNA levels originates from either a proportional increase in shedding rate or decrease in clearance rate, is insufficient to explain the observed increase in correlation. Rather, we show that the correlation between leukocyte and non-leukocyte cfDNA concentrations can be a manifestation of a saturation mechanism in cfDNA clearance. The saturation in clearance model is able to explain both the multiplicative increase in total cfDNA concentration and the increase in correlation, by considering increases in shedding rates that are smaller than the observed increases in cfDNA concentrations. The saturation effect is biologically plausible since cfDNA clearance occurs primarily via Kupffer cells in the liver that have a finite capacity [11,12]. For the cases of pancreatic and ovarian cancer where cfDNA composition was available [19], we observed a higher increase in cfDNA shedding from non-leukocyte sources compared to leukocytes. This disproportionate increase may reflect cancer-induced tissue stress and collateral damage in epithelial and stromal compartments, which are likely more vulnerable to inflammatory signaling and metabolic dysregulation than immune cells. These processes, including apoptosis, necrosis, and activation of stress-response pathways such as autophagy and DNA damage response, are documented in epithelial cells during cancer progression and can occur even in early-stage disease as part of adaptation to oncogenic and inflammatory stress [33–38]. We also note that, although NETosis can contribute to cfDNA release from leukocytes, its role is primarily associated with advanced cancer and metastatic processes rather than early-stage disease [39].

Highlighting the role of saturation effects in cfDNA clearance could contribute to existing mechanistic models, which often assume linear clearance [13,27]. This is relevant not only for early-stage cancer modeling, but also for monitoring and estimating tumor burden in later stages, where cfDNA kinetics inform treatment response and residual disease assessment.

We note that the reported cfDNA concentrations may differ significantly depending on the method used for the analysis of the samples. Employing a different methodology, Cohen *et al.* (2017) [40] report a median cfDNA concentration of 4.25ng/mL in healthy individuals, which is significantly different than the 2.83ng/mL reported in Cohen *et al.* (2018) [20]. However, we find that the multiplicative factor effect is also present between cfDNA concentrations from samples of healthy individuals from the two studies (multiplicative factor value 1.57 (1.34-1.79), S10 Fig.); this finding can be proven useful when cfDNA data from different studies are combined.

Finally, the cancer-specific multiplicative factor may be affected by the effects of therapy or anesthesia. Anesthesia has been reported to reduce the concentration of cfDNA ~2.5-fold [41]. Cohen *et al.* (2017) [40] report cfDNA concentrations from 182 healthy individuals and 221 patients with stage I-III pancreatic cancer. As many pancreatic cancer patients in this study had their blood sample taken after administration of anesthesia, we expect that the multiplicative factor for pancreatic cancer is up to 2.5-fold lower compared to Cohen *et al.* (2018) [20], which excluded patients where anesthesia was known to be administered prior to sample collection. Indeed, we find that the multiplicative factor for pancreatic cancer from Cohen *et al.* (2017) [40] is 1.45 (1.30-1.59) (S11 Fig.). This factor is ~2 times lower than the multiplicative factor for pancreatic cancer in Cohen *et al.* (2018) [20] data, as expected.

Materials and Methods

Patient data

We analyzed cfDNA concentrations (ng/mL) in plasma from healthy individuals and patients with stage I–III cancer reported by Cohen *et al.* [20]. The original dataset included 812 healthy individuals and 959 cancer patients across eight cancer types (lung, breast, colorectal, pancreatic, ovarian, esophageal, stomach, and liver). To ensure methodological consistency, we excluded 181 healthy individuals and 46 pancreatic cancer patients that were processed in a previous study using a different protocol [22], leaving 631 healthy individuals and 913 cancer patients for analysis (Table S1a). Among the 631 healthy individuals, 572

(90.6%) had circulating mutant DNA (cmDNA) detected in plasma and were included in comparative analyses, while 59 (9.4%) without detectable mutations were excluded from analyses involving cmDNA.
Cohen *et al.* [20] also report the concentration (fragments/mL) and MAF% of the circulating mutant DNA (cmDNA) of the top DNA mutation detected in plasma for the same population of healthy individuals and cancer patients. The top mutation is chosen as the mutation with the highest omega score -a measure of statistical significance. Omega score is calculated in Cohen *et al.* [20] for every mutation in each sample based on the comparison between the mutation MAF and the reference MAF distributions among the healthy individuals and cancer patients in training sets. The panel for mutation detection from plasma includes 61 amplicons, each querying an average of 33 base pairs within one of the 16 cancer driver genes. Genes included in the panel and the number of amplicons for each gene are: *NRAS* (2), *CTNNB1* (2), *PIK3CA* (4), *FBXW7* (4), *APC* (2), *EGFR* (1), *BRAF* (1), *CDKN2* (2), *PTEN* (4), *FGFR2* (1), *HRAS* (1), *KRAS* (3), *AKT1* (1), *TP53* (31), *PPP2R1A* (1), *GNAS* (1). Apart from the top mutation detected in plasma, mutations found in tumor are also reported for 70 patients with lung cancer, 116 with breast, 336 with colorectal, 15 with pancreatic, 7 with ovarian, 36 with esophageal, 47 with stomach, and 24 with liver cancer. For these patients, we summarize in Table S1c the number of cases in which the top mutation in plasma was also detected in the tumor or not.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov (K-S) test is a nonparametric test for continuous one-dimensional probability distributions that we employ to check whether two datasets may come from the same distribution [42], Sec. 1.3.5.16. More specifically, in the case of two datasets, we use the Kolmogorov-Smirnov test to check whether the multiplication of all elements of one dataset by the same deterministic factor may result in the two datasets originating from the same distribution.
For the Mattox et al. [19] dataset, which includes paired leukocyte and non-leukocyte cfDNA concentrations are available, we also performed a two-dimensional Kolmogorov–Smirnov (2D KS) test on two-dimensional samples, each one consisting of leukocyte cfDNA concentration and non-leukocyte cfDNA concentration for every cancer patient. The test was used to compare the joint distributions of these paired concentrations across cancer stages. The 2D KS test is a nonparametric method that checks whether two 2D samples originate from the same bivariate distribution [43].

Multiplicative factors for cfDNA/cmDNA

We determine the value of multiplicative factor α for which we obtain the minimum Kolmogorov distance between the empirical CDF for α times the sample of cfDNA concentration in healthy individuals with mutations detected in plasma, and the empirical CDF for the sample of cfDNA concentration in patients of each cancer type. We report the value of the multiplicative factor α, and the p-value of K-S test between the sample of cfDNA concentration in healthy individuals with mutations detected in plasma multiplied by α, and the cfDNA concentration in patients of each cancer type (Table S2).
We also report the value range of the multiplicative factor α, for which the K-S test does not reject, at 5% significance level, the hypothesis that cfDNA data for cancer patients and the cfDNA data for healthy individuals multiplied by the respective factor come from the same distribution. We also depict the P-P plots [42], Sec. 1.3.3.22. P-P plots are constructed by plotting the values of the empirical cumulative distribution functions (CDFs) of the two datasets for the same value of cfDNA concentration as a pair on the (x,y) plane. If all pairs stay close to the diagonal line from (0,0) to (1,1), this means that the two datasets follow the same distribution.
We calculate the multiplicative factors α between the cmDNA concentrations reported in healthy individuals, and the concentrations of cmDNA not shed from tumor in patients of the various cancer types.

The top DNA mutation in plasma was not detected in the tumor for 43 patients with lung, 90 with breast, 205 with colorectal, 11 with pancreatic, 1 with ovarian, 18 with esophageal, 34 with stomach and 11 with liver cancer. The value of a resulting in the least Kolmogorov distance between the empirical CDFs, along with the 5% significance level ranges, are shown in Table S3; the respective P-P plots, which also stay close to the diagonal, are shown in Fig. S6. We exclude ovarian cancer from cmDNA concentration analysis, due to insufficient data. We also calculate the multiplicative factors α for cmDNA concentrations, excluding the samples with MAF less than i) 0.01%, ii) 0.015%, and iii) 0.02% (Table S3), since mutations detected in such low MAFs may be artifactual [22].

Estimation of saturation model parameters and cfDNA shedding rates in healthy individuals

For the dataset of healthy individuals from Mattox *et al.* [19], we estimate parameter k of the cfDNA dynamics model (3) with saturation in clearance, as well as the averages m_l^S, m_{nl}^S and coefficients of variation (standard deviation over average) CV_l^S, CV_{nl}^S of normalized shedding rates $\tilde{S}_l, \tilde{S}_{nl}$ from leukocytes and non-leukocyte sources in the absence of cancer. First, we deduce from data a low correlation coefficient $\rho = 0.1$ between leukocyte and non-leukocyte concentrations in healthy individuals. Low correlation is predicted by the saturation model (3) for small average values of the normalized shedding rates, $m_l^S, m_{nl}^S \ll 1$, where Pearson's correlation coefficient ρ between leukocyte and non-leukocyte cfDNA concentrations is approximated by the average total shedding rate (Eq. S19 in S3 Text):

$$\rho \cong m_l^S + m_{nl}^S \,. \tag{5}$$

Furthermore, the average total cfDNA concentration, and the CV of leukocyte cfDNA concentration are expressed as (Eqs. S14c, S16a in S3 Text):

$$m_{tot}^C = m_l^C + m_{nl}^C \cong \frac{k(m_l^S + m_{nl}^S)}{1 - m_l^S - m_{nl}^S}, \;\; CV_l^C \cong \frac{CV_l^S}{1 - m_l^S - m_{nl}^S}. \tag{6a,b}$$

From Eqs. (5), (6a), and for a mean total cfDNA concentration $m_{tot}^C = 3.8ng/mL$ reported in healthy individuals by Mattox *et al.* [19], we estimate model parameter k to $34.2ng/mL$. Additionally, the ratios of the averages m_l^S, m_{nl}^S, and the coefficients of variation CV_l^S, CV_{nl}^S for shedding rates $\tilde{S}_l$, $\tilde{S}_{nl}$ are approximated by the ratio of averages m_l^C, m_{nl}^C and CVs CV_l^C, CV_{nl}^C of the leukocyte and non-leukocyte cfDNA concentrations (Eqs. S15, S17 in S3 Text):

$$\frac{m_l^S}{m_{nl}^S} \cong \frac{m_l^C}{m_{nl}^C}, \;\; \frac{CV_l^S}{CV_{nl}^S} \cong \frac{CV_l^C}{CV_{nl}^C}. \tag{7a,b}$$

From Eqs. (5) and (7a), and for a reported ${m_l}^C/{m_{nl}}^C = 4$ from the dataset of healthy individuals, we estimate the leukocyte and non-leukocyte average normalized shedding rates to $m_l^S = 0.08$ and $m_{nl}^S = 0.02$. Last, from Eqs. (6b), (7b) and for ${CV_l}^C/{CV_{nl}}^C = 0.9$, $CV_l^C = 0.68$ for healthy individuals as calculated from the Mattox *et al.* [19] dataset, we estimate $CV_l^S = 0.61$ and $CV_{nl}^S = 0.68$ for the normalized shedding rates.

We use the estimated mean values and CVs to determine distributions for the normalized shedding rates. In order to ensure that the total normalized shedding rate is less than 1, we choose the rescaled Beta distribution for each of the normalized shedding rates

$$f_{\tilde{S}}(x; a, b, c) = \frac{1}{c}\frac{(x/c)^{a-1}(1-x/c)^{b-1}}{B(a,b)}, \;\; x \in (0, c), \tag{8}$$

where $B(a,b)$ is the Beta function. In Eq. (8), the parameter c represents the maximum normalized shedding rate for each source. We set $c = 0.6$ for leukocytes and $c = 0.4$ for non-leukocytes to reflect that, since cfDNA shedding occurs primarily during cell turnover [8,11], the ratio of leukocyte to non-leukocyte maximum shedding rates should approximate the ratio of their cell turnover rates, which is ~1.5 [44]. The rescaled Beta distribution parameters a, b are determined via the mean m^S and coefficient of variation CV^S of each normalized shedding rate

$$a = \frac{m^S}{c}\left(\frac{c-m^S}{m^S(CV^S)^2} - 1\right),\ b = \left(1 - \frac{m^S}{c}\right)\left(\frac{c-m^S}{m^S(CV^S)^2} - 1\right). \tag{9a,b}$$

Assuming that the normalized shedding rates from leukocytes and non-leukocyte sources are independent, we determine via Eqs. (4a,b) the joint distribution of leukocyte and non-leukocyte cfDNA concentrations C_l, C_{nl}:

$$f_{C_l C_{nl}}(x,y) = \frac{k}{(k+x+y)^3} f_{\tilde{S}}\left(\frac{x}{k+x+y}; a_l, b_l, c_l\right) f_{\tilde{S}}\left(\frac{y}{k+x+y}; a_{nl}, b_{nl}, c_{nl}\right). \tag{10}$$

From the joint distribution (10), we determine the distributions for leukocyte, non-leukocyte and total cfDNA concentrations, as well as the correlation coefficient between leukocyte and non-leukocyte cfDNA concentrations.

Construction of ROC curves

We computed Receiver Operating Characteristic (ROC) curves using cfDNA concentration as the predictor [45]. For each cancer type, the dataset was partitioned into $k = 10$ folds using cross-validation. In each iteration, 9 folds were used for training and 1 fold for testing. ROC curves were generated by plotting $\text{Sensitivity} \times 100\%$ versus $(1 - \text{Specificity}) \times 100\%$ across all possible thresholds, where:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},\ \ \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

with TP, TN, FP, and FN denoting true positives, true negatives, false positives, and false negatives, respectively. Overall performance was summarized by the Area Under the Curve (AUC) and its 95% confidence interval, estimated using 1000 bootstrap resamples [46]. To quantify the variability in ROC curves arising from cross-validation fold partitioning, the 10-fold results were resampled with replacement for the 1000 bootstrap iterations. We created a grid of 200 fixed specificity values from 0 to 1, indexed by $j = 1, \ldots, 200$. For each bootstrap iteration, we computed the mean sensitivity s_j^* across folds at each specificity value from the grid. At every grid point j, we derived empirical distributions from the bootstrap sensitivity samples. The 95% confidence bands were then obtained as the 2.5th and 97.5th percentiles of these distributions:

$$\text{CI}_{95\%}(s_j) = \left[\text{Percentile}_{2.5}(s_j^*), \text{Percentile}_{97.5}(s_j^*)\right].$$

The optimal cutoff cfDNA level was selected by scanning all possible thresholds on the ROC curve and choosing the one that maximizes Youden's J statistic, that balances sensitivity and specificity [45]:

$$\text{J} = \text{Sensitivity} + \text{Specificity} - 1$$

This threshold value was applied to the test fold to compute the positive predictive value (PPV), and negative predictive value (NPV):

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},\ \ \text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

PPV represents the probability that a subject with a positive test truly has the disease, while NPV represents the probability that a subject with a negative test truly does not have the disease. PPV and NPV values reflect the prevalence of each cancer type in our study cohort. To quantify uncertainty in performance

metrics we applied the same bootstrap resampling strategy used for AUC estimation. For each of the 1000 bootstrap iterations, the cross-validation folds were resampled with replacement, the optimal cutoff was re-estimated, and the metrics were recomputed. This generated empirical distributions for Sensitivity, Specificity, PPV, and NPV. The 95% confidence intervals for each metric were obtained as the 2.5th and 97.5th percentiles of the corresponding bootstrap distributions:

$$\mathrm{CI}_{95\%}(m) = [\mathrm{Percentile}_{2.5}(m^*), \mathrm{Percentile}_{97.5}(m^*)],$$

where $m \in \{\mathrm{Sensitivity}, \mathrm{Specificity}, \mathrm{PPV}, \mathrm{NPV}\}$ and m^* denotes bootstrap samples of metric m.

Acknowledgements
The authors thank Bert Vogelstein for helpful comments on the manuscript.

Funding
National Science Foundation grant DMS-2045166 (KM, IB)
Pacific Institute for the Mathematical Sciences (KM).

Competing interests
The authors declare no competing interests.

Author contributions
KM and IB designed and performed research and wrote the manuscript.

Artificial Intelligence declaration
No AI technologies have been used in the creation of this manuscript.

Data and materials availability
The data analyzed in this study were obtained from Supplementary Materials of
Cohen *et al.* (2018) [20] https://doi.org/10.1126/science.aar3247
Cohen *et al.* (2017) [40] https://doi.org/10.1073/pnas.1704961114
Mattox *et al.* (2023) [19] https://doi.org/10.1158/2159-8290.CD-21-1252
MATLAB code used to generate the results and figures of this work is publicly available at:
https://doi.org/10.5281/zenodo.19026910

References

[1] Crosby D, Bhatia S, Brindle KM, et al. Early detection of cancer. Science 2022;375. https://doi.org/10.1126/SCIENCE.AAY9040.

[2] Mandel P, Métais P. Les acides nucleiques du plasma sanguin chez l' homme. C R Seances Soc Biol Fil 1948.

[3] Alese OB, Cook N, Ortega-Franco A, et al. Circulating Tumor DNA: An Emerging Tool in Gastrointestinal Cancers. Am Soc Clin Oncol Educ Book 2022;42:279–98. https://doi.org/10.1200/EDBK_349143.

[4] Bronkhorst AJ, Ungerer V, Holdenrieder S. The emerging role of cell-free DNA as a molecular marker for cancer management. Biomol Detect Quantif 2019;17. https://doi.org/10.1016/j.bdq.2019.100087.

[5] Diehl F, Schmidt K, Choti MA, et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 2008;14:985–90. https://doi.org/10.1038/NM.1789.

[6] Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6. https://doi.org/10.1126/SCITRANSLMED.3007094.

[7] Stroun M, Anker P, Maurice P, et al. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology 1989;46:318–22. https://doi.org/10.1159/000226740.

[8] Kustanovich A, Schwartz R, Peretz T, et al. Life and death of circulating cell-free DNA. Cancer Biol Ther 2019;20:1057–67. https://doi.org/10.1080/15384047.2019.1598759.

[9] Siravegna G, Marsoni S, Siena S, et al. Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol 2017;14:531–48. https://doi.org/10.1038/NRCLINONC.2017.14.

[10] Diehl F, Li M, Dressman D, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci U S A 2005;102:16368–73. https://doi.org/10.1073/PNAS.0507904102.

[11] Han DSC, Lo YMD. The Nexus of cfDNA and Nuclease Biology. Trends in Genetics 2021;37:758–70. https://doi.org/10.1016/j.tig.2021.04.005.

[12] Dennis Lo YM, Zhang J, Leung TN, et al. Rapid clearance of fetal DNA from maternal plasma. Am J Hum Genet 1999;64:218–24. https://doi.org/10.1086/302205.

[13] Avanzini S, Kurtz DM, Chabon JJ, et al. A mathematical model of ctDNA shedding predicts tumor detection size. Sci Adv 2020;6. https://doi.org/10.1126/SCIADV.ABC4308.

[14] Yamamoto R, Asano H, Tamaki R, et al. Dynamics and Half-Life of Cell-Free DNA After Exercise: Insights from a Fragment Size-Specific Measurement Approach. Diagnostics (Basel) 2025;15. https://doi.org/10.3390/DIAGNOSTICS15010109.

[15] Sender R, Noor E, Milo R, et al. What fraction of cellular DNA turnover becomes cfDNA? Elife 2024;12. https://doi.org/10.7554/ELIFE.89321.

[16] Cano-Gamez K, Maclean P, Inoue M, et al. The circulating cell-free DNA landscape in sepsis is dominated by impaired liver clearance. Cell Genomics 2025;5. https://doi.org/10.1016/j.xgen.2025.100971.

[17] Martin-Alonso C, Tabrizi S, Xiong K, et al. Priming agents transiently reduce the clearance of cell-free DNA to improve liquid biopsies. Science (1979) 2024;383:1–10. https://doi.org/10.1126/SCIENCE.ADF2341.

[18] Gauthier VJ, Tyler LN, Mannik M. Blood clearance kinetics and liver uptake of mononucleosomes in mice. J Immunol 1996;156:1151–6.

[19] Mattox AK, Douville C, Wang Y, et al. The Origin of Highly Elevated Cell-Free DNA in Healthy Individuals and Patients with Pancreatic, Colorectal, Lung, or Ovarian Cancer. Cancer Discov 2023;13:2166–79. https://doi.org/10.1158/2159-8290.CD-21-1252.

[20] Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359:926–30. https://doi.org/10.1126/SCIENCE.AAR3247.

[21] Kinde I, Wu J, Papadopoulos N, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 2011;108:9530–5. https://doi.org/10.1073/PNAS.1105422108;ISSUE:ISSUE:DOI.

[22] Cohen JD, Douville C, Dudley JC, et al. Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. Nat Biotechnol 2021;39:1220–7. https://doi.org/10.1038/S41587-021-00900-Z.

[23] Sun K, Jiang P, Chan KCA, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A 2015;112:E5503–12. https://doi.org/10.1073/PNAS.1508736112;ISSUE:ISSUE:DOI.

[24] Moss J, Magenheim J, Neiman D, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun 2018;9. https://doi.org/10.1038/S41467-018-07466-6.

[25] Stejskal P, Goodarzi H, Srovnal J, et al. Circulating tumor nucleic acids: biology, release mechanisms, and clinical relevance. Mol Cancer 2023;22:15. https://doi.org/10.1186/s12943-022-01710-w.

[26] Thompson JC, Scholes DG, Carpenter EL, et al. Molecular response assessment using circulating tumor DNA (ctDNA) in advanced solid tumors. Br J Cancer 2023;129:1893–902. https://doi.org/10.1038/S41416-023-02445-1.

[27] Li A, Lou E, Leder K, et al. Early Circulating Tumor DNA Kinetics as a Dynamic Biomarker of Cancer Treatment Response. JCO Clin Cancer Inform 2025;9. https://doi.org/10.1200/CCI-24-00160.

[28] Leite da Silva LF, Saldanha EF, de Menezes SA, et al. Plasma ctDNA kinetics as a predictor of systemic therapy response for advanced non-small cell lung cancer: a systematic review and meta-analysis. Oncologist 2025;30. https://doi.org/10.1093/ONCOLO/OYAE344.

[29] Keener James, Sneyd James. Mathematical physiology. I, Cellular physiology. 2nd ed. Springer-Verlag New York; 2009.

[30] Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017;9. https://doi.org/10.1126/SCITRANSLMED.AAN2415.

[31] Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med 2014;371:2488–98. https://doi.org/10.1056/NEJMOA1408617.

[32] Bolton KL, Ptashkin RN, Gao T, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet 2020;52:1219–26. https://doi.org/10.1038/S41588-020-00710-0.

[33] Ma Y, Kroemer G. The cancer-immune dialogue in the context of stress. Nat Rev Immunol 2024;24:264–81. https://doi.org/10.1038/S41577-023-00949-8.

[34] Altea-Manzano P, Decker-Farrell A, Janowitz T, et al. Metabolic interplays between the tumour and the host shape the tumour macroenvironment. Nat Rev Cancer 2025;25:274–92. https://doi.org/10.1038/S41568-024-00786-4.

[35] Haase K, Jamal-Hanjani M. Clues to why the weight-loss condition cachexia arises when cancer occurs. Nature 2025;646:805–6. https://doi.org/10.1038/D41586-025-02594-8.

[36] Davidson BA, Miranda AX, Reed SC, et al. An in vitro CRISPR screen of cell-free DNA identifies apoptosis as the primary mediator of cell-free DNA release. Commun Biol 2024;7. https://doi.org/10.1038/S42003-024-06129-1.

[37] Galluzzi L, Vitale I, Aaronson SA, et al. Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. Cell Death Differ 2018;25:486–541. https://doi.org/10.1038/S41418-017-0012-4.

[38] Roos WP, Thomas AD, Kaina B. DNA damage and the balance between survival and death in cancer biology. Nat Rev Cancer 2016;16:20–33. https://doi.org/10.1038/NRC.2015.2.

[39] Papayannopoulos V. Neutrophil extracellular traps in immunity and disease. Nat Rev Immunol 2018;18:134–47. https://doi.org/10.1038/NRI.2017.105.

[40] Cohen JD, Javed AA, Thoburn C, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proceedings of the National Academy of Sciences 2017;114:10202–7. https://doi.org/10.1073/PNAS.1704961114.

[41] Prakash K, Aggarwal S, Bhardwaj S, et al. Serial perioperative cell-free DNA levels in donors and recipients undergoing living donor liver transplantation. Acta Anaesthesiol Scand 2017;61:1084–94. https://doi.org/10.1111/AAS.12947.

[42] NIST/SEMATECH. e-Handbook of Statistical Methods n.d. http://www.itl.nist.gov/div898/handbook/ Sections 1.3.3.22 Probability Plot, and 1.3.5.16 Kolmogorov-Smirnov Goodness-of-Fit Test.

[43] Peacock JA. Two-dimensional goodness-of-fit testing in astronomy. Mon Not R Astron Soc 1983;202:615–27. https://doi.org/10.1093/MNRAS/202.3.615.

[44] Sender R, Milo R. The distribution of cellular turnover in the human body. Nat Med 2021;27:45–8. https://doi.org/10.1038/S41591-020-01182-9.

[45] Zhou X-H, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd ed. Hoboken, NJ: Wiley; 2011.

[46] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC; 1994.

Electronic Supplementary Material for "Early-stage cancer results in a multiplicative increase in cell-free DNA originating from healthy tissue"

by Konstantinos Mamis and Ivana Bozic

S1 Text. Correlation coefficients between leukocyte and non-leukocyte cfDNA

For the dataset and the four different deconvolution methods from Mattox *et al.* [19], we calculate the Pearson and Spearman correlation coefficients between the cfDNA concentrations from leukocytes and non-leukocyte sources, for healthy individuals

Correlation coefficients for healthy individuals

	Sun *et al.* [23] QP	Sun *et al.* [23] NNLS	Moss *et al.* [24] QP	Moss *et al.* [24] NNLS
Pearson	0.84	0.94	0.51	0.31
Spearman	0.20	0.50	0.02	-0.25

From the discrepancy between Pearson and Spearman coefficients, we deduce that Pearson coefficient is affected by the presence of outliers; by excluding the three outliers with significantly higher cfDNA concentrations compared to the rest of the dataset, we calculate

Correlation coefficients for healthy individuals (excluding the three outliers)

	Sun *et al.* [23] QP	Sun *et al.* [23] NNLS	Moss *et al.* [24] QP	Moss *et al.* [24] NNLS
Pearson	-0.07	0.20	-0.16	-0.29
Spearman	0.08	0.42	-0.13	-0.42

After excluding the three outliers, Pearson and Spearman coefficients are in better agreement, showing a weak correlation between leukocyte and non-leukocyte cfDNA concentrations, especially for the deconvolution methods using Sun et al. reference matrix. Calculating the correlation coefficients for the pancreatic and ovarian cancer patients excluding the possible outliers we obtain:

Correlation coefficients for pancreatic cancer patients

	Sun *et al.* [23] QP	Sun *et al.* [23] NNLS	Moss *et al.* [24] QP	Moss *et al.* [24] NNLS
Pearson	0.53	0.58	0.31	0.39
Spearman	0.72	0.73	0.47	0.44

Correlation coefficients for ovarian cancer patients (excluding the one outlier)

	Sun *et al.* [23] QP	Sun *et al.* [23] NNLS	Moss *et al.* [24] QP	Moss *et al.* [24] NNLS
Pearson	0.87	0.87	0.32	0.60
Spearman	0.85	0.85	0.18	0.29

We observe that both in pancreatic and ovarian cancer, the leukocyte and non-leukocyte cfDNA concentrations are strongly correlated. In our analysis we will employ the Pearson correlation coefficient since its calculation from both datasets and analytic distributions is straightforward.

S2 Text. Correlation coefficient in cfDNA composition for the linear clearance model

The cfDNA concentration from leukocyte (C_l) and non-leukocyte sources (C_{nl}) in healthy individuals are random quantities, with mean values, variances and covariance

$$m_l = E[C_l],\ \ \sigma_l^2 = E[C_l^2] - (E[C_l])^2,\ \ m_{nl} = E[C_{nl}],\ \ \sigma_{nl}^2 = E[C_{nl}^2] - (E[C_{nl}])^2,$$
$$Cov(C_l, C_{nl}) = E[C_l C_{nl}] - E[C_l]E[C_{nl}],$$

with $E[\cdot]$ being the expectation operator. Under the linear clearance model, $C_l = S_l/\varepsilon$, $C_{nl} = S_{nl}/\varepsilon$, $C_{tot} = S_{tot}/\varepsilon$, multiplicative increases by factors α_l, α_{nl} in leukocyte and non-leukocyte cfDNA shedding rates due to cancer presence result in a multiplicative increase of leukocyte and non-leukocyte cfDNA levels by the same factors

$$\alpha_l C_l,\ \ \alpha_{nl} C_{nl}.$$

The multiplicative factors α_l, α_{nl} in cancer are uncorrelated with the cfDNA levels C_l, C_{nl} in healthy individuals. We assume that the multiplicative factors α_l, α_{nl} vary across patients and follow uniform distributions over the ranges (c_l, d_l) and (c_{nl}, d_{nl}) respectively. These ranges are selected so that, when the healthy leukocyte and non-leukocyte cfDNA concentrations are multiplied by factors from these ranges, the resulting distributions satisfy the Kolmogorov–Smirnov (KS) test at the 5% significance level when compared to the corresponding cfDNA concentrations in cancer patients. Last, we assume that the multiplicative factors α_l, α_{nl} are perfectly linearly correlated, meaning that cancer presence increases cfDNA shedding from both leukocyte and non-leukocyte sources simultaneously:

$$\alpha_{nl} = p\alpha_l + q.$$

The constants p, q are determined as

$$p = \frac{d_{nl} - c_{nl}}{d_l - c_l},\quad q = c_{nl} - c_l,$$

so that $\alpha_{nl} \in (c_{nl}, d_{nl})$ for $\alpha_l \in (c_l, d_l)$.

To calculate the Pearson correlation coefficient between leukocyte $\alpha_l C_l$ and non-leukocyte $\alpha_{nl} C_{nl}$ cfDNA concentrations in cancer, we calculate their covariance

$$
\begin{aligned}
Cov(\alpha_l C_l, \alpha_{nl} C_{nl}) &= E[\alpha_l C_l \alpha_{nl} C_{nl}] - E[\alpha_l C_l] E[\alpha_{nl} C_{nl}] \\
&= E[\alpha_l C_l (p\alpha_l + q) C_{nl}] - E[\alpha_l C_l] E[(p\alpha_l + q) C_{nl}] \\
&= (pVar(\alpha_l) + p(E[\alpha_l])^2 + qE[\alpha_l]) Cov(C_l, C_{nl}),
\end{aligned}
$$

where $Var(\alpha_l)$ is the variance of the multiplicative factor of cfDNA levels from leukocytes. We also calculate the variances

$$Var(\alpha_l C_l) = E[\alpha_l^2] E[C_l^2] - (E[\alpha_l])^2 m_l^2 = Var(\alpha_l)\sigma_l^2 + Var(\alpha_l) m_l^2 + (E[\alpha_l])^2 \sigma_l^2$$

$$Var(\alpha_{nl} C_{nl}) = Var(\alpha_{nl}) E[C_{nl}^2] + (E[\alpha_{nl}])^2 \sigma_{nl}^2 = p^2 Var(\alpha_l)\sigma_{nl}^2 + p^2 Var(\alpha_l) m_{nl}^2 + (pE[\alpha_l] + q)^2 \sigma_{nl}^2$$

The Pearson correlation coefficient for $\alpha_l C_l$, $\alpha_{nl} C_{nl}$ reads

$$
\begin{aligned}
\rho &= \frac{Cov(\alpha_l C_l, \alpha_{nl} C_{nl})}{\sqrt{Var[\alpha_l C_l]}\sqrt{Var[\alpha_{nl} C_{nl}]}} = \\
&= \frac{\left(Var(\alpha_l) + (E[\alpha_l])^2 + \frac{q}{p} E[\alpha_l]\right) Cov(C_l, C_{nl})}{\sqrt{Var(\alpha_l)\sigma_l^2 + Var(\alpha_l) m_l^2 + (E[\alpha_l])^2 \sigma_l^2}\sqrt{Var(\alpha_l)\sigma_{nl}^2 + Var(\alpha_l) m_{nl}^2 + \left(E[\alpha_l] + \frac{q}{p}\right)^2 \sigma_{nl}^2}} = \\
&= \frac{Var(\alpha_l) + (E[\alpha_l])^2 + \frac{q}{p} E[\alpha_l]}{\sqrt{Var(\alpha_l)\left(1 + \frac{m_l^2}{\sigma_l^2}\right) + (E[\alpha_l])^2}\sqrt{Var(\alpha_l)\left(1 + \frac{m_{nl}^2}{\sigma_{nl}^2}\right) + \left(E[\alpha_l] + \frac{q}{p}\right)^2}} \rho_{\text{healthy}},
\end{aligned}
$$

where $\rho_{\text{healthy}} = Cov(C_l, C_{nl})/(\sigma_l^2 \sigma_{nl}^2)$ is the Pearson correlation coefficient between leukocyte and non-leukocyte cfDNA in healthy individuals. We have the ranges $(c_l, d_l) = (3.15, 4.45)$, $(c_{nl}, d_{nl}) = (6.24, 9.94)$ for pancreatic cancer and $(c_l, d_l) = (2.03, 3.37)$, $(c_{nl}, d_{nl}) = (2.39, 4.49)$ for ovarian cancer. Since we consider α_l, α_{nl} to be uniformly distributed, we calculate the means and variances that appear in the above relation for correlation coefficient

$$E[\alpha_l] = \frac{c_l + d_l}{2} = 3.8 \text{ for pancreatic cancer, and } 2.7 \text{ for ovarian cancer}$$

$$Var[\alpha_l] = \frac{(d_l - c_l)^2}{12} = 0.14 \text{ for pancreatic cancer, and } 0.15 \text{ for ovarian cancer}$$

also

$$\frac{q}{p} = 1.09 \text{ for pancreatic cancer, and } 0.23 \text{ for ovarian cancer}$$

and from the data of healthy individuals, for the four deconvolution methods, we calculate

$$\frac{m_l^2}{\sigma_l^2} = 2.16, 2.18, 1.40, 1.28 \text{ and } \frac{m_{nl}^2}{\sigma_{nl}^2} = 1.74, 2.62, 1.96, 1.22.$$

We observe that, for both cancer types, the mean dominates, $E[\alpha_l] \gg Var[\alpha_l]$, and also m_l^2/σ_l^2 and m_{nl}^2/σ_{nl}^2 are of order one, so they do not offset the dominance. Thus, in the ratio multiplying ρ_{healthy}, the terms dependent on $Var[\alpha_l]$ can be omitted:

$$\frac{Var(\alpha_l) + (E[\alpha_l])^2 + \frac{q}{p} E[\alpha_l]}{\sqrt{Var(\alpha_l)\left(1 + \frac{m_l^2}{\sigma_l^2}\right) + (E[\alpha_l])^2}\sqrt{Var(\alpha_l)\left(1 + \frac{m_{nl}^2}{\sigma_{nl}^2}\right) + \left(E[\alpha_l] + \frac{q}{p}\right)^2}} \approx \frac{E[\alpha_l]\left(E[\alpha_l] + \frac{q}{p}\right)}{E[\alpha_l]\left(E[\alpha_l] + \frac{q}{p}\right)} = 1,$$

resulting in $\rho \approx \rho_{\text{healthy}}$ for both pancreatic and ovarian cancers. Indeed, by substituting the values of all parameters, for each cancer type separately and for each of the deconvolution methods, we get ρ from -0.29 to 0.20, effectively coinciding with the Pearson correlation coefficient between leukocyte and non-leukocyte cfDNA concentrations in healthy individuals.

S3 Text. Moment approximations for the model with saturation in clearance

For the saturation in clearance model,

$$C_l' = S_l - \frac{vC_l}{k+C_{tot}},\ C_{nl}' = S_{nl} - \frac{vC_{nl}}{k+C_{tot}} \qquad \text{(S1a,b)}$$

the equilibrium leukocyte, non-leukocyte and total cfDNA concentrations C_l, C_{nl} and $C_{tot} = C_l + C_{nl}$ are

$$C_l = \frac{kS_l}{v-S_{tot}},\ C_{nl} = \frac{kS_{nl}}{v-S_{tot}},\ C_{tot} = \frac{kS_{tot}}{v-S_{tot}}, \qquad \text{(S2a,b)}$$

where S_l, S_{nl} and $S_{tot} = S_l + S_{nl}$ are the leukocyte, non-leukocyte and total cfDNA shedding rates, v is the maximum cfDNA clearance rate, and k is the Michaelis-Menten parameter. By introducing the shedding rates normalized by the maximum clearance rate, $\tilde{S}_l = \frac{S_l}{v}$, $\tilde{S}_{nl} = \frac{S_{nl}}{v}$, $\tilde{S}_{tot} = \frac{S_{tot}}{v}$ we express Eqs. (S2a,b) equivalently as

$$C_l = \frac{k\tilde{S}_l}{1-\tilde{S}_{tot}},\ C_{nl} = \frac{k\tilde{S}_{nl}}{1-\tilde{S}_{tot}},\ C_{tot} = \frac{k\tilde{S}_{tot}}{1-\tilde{S}_{tot}}. \qquad \text{(S3a,b,c)}$$

From Eqs. (S3a,b) we deduce that the saturation effect does not depend on the value of maximum clearance rate v per se; rather, it depends on total shedding rate as a percentage of v, i.e. $\tilde{S}_{tot}$: For $\tilde{S}_{tot} \ll 1$, the saturation effect is negligible, with the concentrations approaching

$$C_l = \frac{k}{v}S_l,\ C_{nl} = \frac{k}{v}S_{nl},\ C_{tot} = \frac{k}{v}S_{tot}, \qquad \text{(S4a,b,c)}$$

which are the equilibrium concentrations for the cfDNA model with linear clearance

$$C_l' = S_l - \varepsilon C_l,\ C_{nl}' = S_{nl} - \varepsilon C_{nl}, \qquad \text{(S5a,b,c)}$$

and clearance rate $\varepsilon = v/k$. Saturation in model (S1a,b) becomes more pronounced as the total shedding rate approaches the maximum clearance rate.

In our setting, we consider parameters v, k to be deterministic, while shedding rates S_l, S_{nl} are independent random variables whose sum is less than v. We note that since parameter k enters formulas (S3a,b) for cfDNA concentrations as a deterministic multiplicative factor, average cfDNA concentrations depend on k,

$$E[C_l] = kE\left[\frac{\tilde{S}_l}{1-\tilde{S}_{tot}}\right],\ E[C_{nl}] = kE\left[\frac{\tilde{S}_{nl}}{1-\tilde{S}_{tot}}\right],\ E[C_{tot}] = kE\left[\frac{\tilde{S}_{tot}}{1-\tilde{S}_{tot}}\right], \qquad \text{(S6a,b,c)}$$

The Pearson correlation coefficient ρ and the coefficients of variation CV_l^C and CV_{nl}^C are k −independent:

$$\rho = \frac{Cov[C_l, C_{nl}]}{\sqrt{Var[C_l]}\sqrt{Var[C_{nl}]}} = \frac{Cov\left[\frac{\tilde{S}_l}{1-\tilde{S}_{tot}}, \frac{\tilde{S}_{nl}}{1-\tilde{S}_{tot}}\right]}{\sqrt{Var\left[\frac{\tilde{S}_l}{1-\tilde{S}_{tot}}\right]}\sqrt{Var\left[\frac{\tilde{S}_{nl}}{1-\tilde{S}_{tot}}\right]}}$$

$$CV_l^C = \frac{\sqrt{Var[C_l]}}{E[C_l]} = \frac{\sqrt{Var\left[\frac{\tilde{S}_l}{1-\tilde{S}_{tot}}\right]}}{E\left[\frac{\tilde{S}_l}{1-\tilde{S}_{tot}}\right]}, \quad CV_{nl}^C = \frac{\sqrt{Var[C_{nl}]}}{E[C_{nl}]} = \frac{\sqrt{Var\left[\frac{\tilde{S}_{nl}}{1-\tilde{S}_{tot}}\right]}}{E\left[\frac{\tilde{S}_{nl}}{1-\tilde{S}_{tot}}\right]}. \tag{S7a,b,c}$$

We approximate the moments of C_l, C_{nl} by using Taylor expansions under the expectation operator, given that the moments of normalized shedding rates $\tilde{S}_l$, $\tilde{S}_{nl}$ are known: The Taylor expansion for a function $f(x,y)$ with two arguments is expressed as

$$f(x,y) = f(x_0,y_0) + \frac{\partial f(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial f(x_0,y_0)}{\partial y}(y-y_0) +$$

$$+\sum_{n=2}^{\infty}\frac{1}{n!}\frac{\partial^n f(x_0,y_0)}{\partial x^n}(x-x_0)^n + \sum_{n=2}^{\infty}\frac{1}{n!}\frac{\partial^n f(x_0,y_0)}{\partial y^n}(y-y_0)^n +$$

$$+\sum_{n_1+n_2\geq 2, n_1,n_2\neq 0}\frac{1}{n_1!\,n_2!}\frac{\partial^{n_1+n_2} f(x_0,y_0)}{\partial x^{n_1}\partial y^{n_2}}(x-x_0)^{n_1}(y-y_0)^{n_2}. \tag{S8}$$

We choose $(x,y) = (\tilde{S}_l, \tilde{S}_{nl})$, and $(x_0,y_0) = (m_l^S, m_{nl}^S)$ with $m_l^S = E[\tilde{S}_l]$, $m_{nl}^S = E[\tilde{S}_{nl}]$ being the mean values, and take the expected value in both sides of Eq. (S8)

$$E[f(\tilde{S}_l,\tilde{S}_{nl})] = f(m_l^S, m_{nl}^S) + \frac{\partial f(m_l^S, m_{nl}^S)}{\partial x}E[\tilde{S}_l - m_l^S] + \frac{\partial f(m_l^S, m_{nl}^S)}{\partial y}E[\tilde{S}_{nl} - m_{nl}^S] +$$

$$+\sum_{n=2}^{\infty}\frac{1}{n!}\frac{\partial^n f(m_l^S, m_{nl}^S)}{\partial x^n}E\left[\left(\tilde{S}_l - m_l^S\right)^n\right] + \sum_{n=2}^{\infty}\frac{1}{n!}\frac{\partial^n f(m_l^S, m_{nl}^S)}{\partial y^n}E\left[\left(\tilde{S}_{nl} - m_{nl}^S\right)^n\right] +$$

$$+\sum_{n_1+n_2\geq 2, n_1,n_2\neq 0}\frac{1}{n_1!\,n_2!}\frac{\partial^{n_1+n_2} f(m_l^S, m_{nl}^S)}{\partial x^{n_1}\partial y^{n_2}}E\left[\left(\tilde{S}_l - m_l^S\right)^{n_1}\left(\tilde{S}_{nl} - m_{nl}^S\right)^{n_2}\right] \tag{S9}$$

In the right-hand side of Eq. (S9), the terms containing the first-order partial derivatives of f are equal to zero, since $E[\tilde{S}_l - m_l^S] = 0$, $E[\tilde{S}_{nl} - m_{nl}^S] = 0$. Furthermore, by assuming that normalized shedding rates $\tilde{S}_l$, $\tilde{S}_{nl}$ are independent, all cross moments in the last summation in the right-hand side of Eq. (S9) zero. Thus, we simplify Eq. (S9) to

$$E[f(\tilde{S}_l,\tilde{S}_{nl})] = f(m_l^S, m_{nl}^S) + \sum_{n=2}^{\infty}\frac{1}{n!}\left[\frac{\partial^n f(m_l^S, m_{nl}^S)}{\partial x^n}M_l^{(n)} + \frac{\partial^n f(m_l^S, m_{nl}^S)}{\partial y^n}M_{nl}^{(n)}\right], \tag{S10}$$

where $M_l^{(n)} = E\left[\left(\tilde{S}_l - m_l^S\right)^n\right]$, $M_{nl}^{(n)} = E\left[\left(\tilde{S}_{nl} - m_{nl}^S\right)^n\right]$ denote the centered moments of $\tilde{S}_l$, $\tilde{S}_{nl}$ respectively. Using Eq. (S10), we calculate the series expansion for the first and second moments of C_l, C_{nl}, with $m_{tot}^S = m_l^S + m_{nl}^S$ being the mean normalized total shedding rate:

$$E[C_l] \cong \frac{km_l^S}{1-m_{tot}^S}\left\{1 + \sum_{n=2}^{\infty}\frac{(1-m_{nl}^S)\left(\frac{M_l^{(n)}}{m_l^S}\right) + M_{nl}^{(n)}}{(1-m_{tot}^S)^n}\right\}$$

$$E[C_{nl}] \cong \frac{k m_{nl}^S}{1-m_{tot}^S}\left\{1+\sum_{n=2}^{\infty}\frac{M_l^{(n)}+(1-m_l^S)\left(\frac{M_{nl}^{(n)}}{m_{nl}^S}\right)}{(1-m_{tot}^S)^n}\right\}$$

$$E[C_l^2] \cong \frac{k^2(m_l^S)^2}{(1-m_{tot}^S)^2}\left\{1+\sum_{n=2}^{\infty}\frac{(1-m_{nl}^S)[(n-1)(1-m_{nl}^S)+2m_l^S]\frac{M_l^{(n)}}{(m_l^S)^2}+(n+1)M_{nl}^{(n)}}{(1-m_{tot}^S)^n}\right\}$$

$$E[C_{nl}^2] \cong \frac{k^2(m_{nl}^S)^2}{(1-m_{tot}^S)^2}\left\{1+\sum_{n=2}^{\infty}\frac{(n+1)M_l^{(n)}+(1-m_l^S)[(n-1)(1-m_l^S)+2m_{nl}^S]\frac{M_{nl}^{(n)}}{(m_{nl}^S)^2}}{(1-\mu_{tot})^n}\right\}$$

$$E[C_l C_{nl}] \cong \frac{k^2 m_l^S m_{nl}^S}{(1-m_{tot}^S)^2}\left\{1+\sum_{n=2}^{\infty}\frac{[n(1-m_{nl}^S)+m_l^S]\frac{M_l^{(n)}}{m_l^S}+[n(1-m_l^S)+m_{nl}^S]\frac{M_{nl}^{(n)}}{m_{nl}^S}}{(1-m_{tot}^S)^n}\right\}$$

By keeping up to linear terms with respect to normalized shedding rates variances $M_l^{(2)}$, $M_{nl}^{(2)}$ in the series expression above, we approximate the mean values, and centered second-order moments for cfDNA concentrations

$$E[C_l] \cong k\left\{\frac{m_l^S}{1-m_{tot}^S}+\frac{(1-m_{nl}^S)(m_l^S)^2(CV_l^S)^2+m_l^S(m_{nl}^S)^2(CV_{nl}^S)^2}{(1-m_{tot}^S)^3}\right\} \tag{S11a}$$

$$E[C_{nl}] \cong k\left\{\frac{m_{nl}^S}{1-m_{tot}^S}+\frac{(m_l^S)^2 m_{nl}^S(CV_l^S)^2+(1-m_l^S)(m_{nl}^S)^2(CV_{nl}^S)^2}{(1-m_{tot}^S)^3}\right\} \tag{S11b}$$

$$E[C_{tot}] \cong k\left\{\frac{m_{tot}^S}{1-m_{tot}^S}+\frac{(m_l^S)^2(CV_l^S)^2+(m_{nl}^S)^2(CV_{nl}^S)^2}{(1-m_{tot}^S)^3}\right\} \tag{S11c}$$

$$Var[C_l] \cong \frac{k^2\left[(m_l^S)^2(1-m_{nl}^S)^2(CV_l^S)^2+(m_l^S)^2(m_{nl}^S)^2(CV_{nl}^S)^2\right]}{(1-m_{tot}^S)^4} \tag{S11d}$$

$$Var[C_{nl}] \cong \frac{k^2\left[(m_l^S)^2(m_{nl}^S)^2(CV_l^S)^2+(1-m_l^S)^2(m_{nl}^S)^2(CV_{nl}^S)^2\right]}{(1-m_{tot}^S)^4} \tag{S11e}$$

$$Cov[C_l C_{nl}] \cong \frac{k^2\left[m_{nl}^S(1-m_{nl}^S)(m_l^S)^2(CV_l^S)^2+m_l^S(1-m_l^S)(m_{nl}^S)^2(CV_{nl}^S)^2\right]}{(1-m_{tot}^S)^4}. \tag{S11f}$$

From Eqs. (S11) we approximate the coefficients of variation for the leukocyte and non-leukocyte cfDNA concentrations

$$CV_l^C \cong \frac{\sqrt{\left(1-m_{nl}^S\right)^2\left(CV_l^S\right)^2+\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}}{\left(1-m_{tot}^S\right)+\frac{\left(1-m_{nl}^S\right)m_l^S\left(CV_l^S\right)^2+\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}{1-m_{tot}^S}} \quad \text{(S12a)}$$

$$CV_{nl}^C \cong \frac{\sqrt{\left(m_l^S\right)^2\left(CV_l^S\right)^2+\left(1-m_l^S\right)^2\left(CV_{nl}^S\right)^2}}{\left(1-m_{tot}^S\right)+\frac{\left(m_l^S\right)^2\left(CV_l^S\right)^2+\left(1-m_l^S\right)m_{nl}^S\left(CV_{nl}^S\right)^2}{1-m_{tot}^S}} \quad \text{(S12b)}$$

and the Pearson's correlation coefficient between the leukocyte and non-leukocyte cfDNA concentrations

$$\rho \cong \frac{m_{nl}^S\left(1-m_{nl}^S\right)\left(m_l^S\right)^2\left(CV_l^S\right)^2+m_l^S\left(1-m_l^S\right)\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}{\sqrt{\left(m_l^S\right)^2\left(1-m_{nl}^S\right)^2\left(CV_l^S\right)^2+\left(m_l^S\right)^2\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}\sqrt{\left(m_l^S\right)^2\left(m_{nl}^S\right)^2\left(CV_l^S\right)^2+\left(1-m_l\right)^2\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}} \quad \text{(S13)}$$

By calculating the derivative of ρ with respect to ${m_l}^S$ we get $\partial\rho/\partial m_l^S > 0$. Since m_l^S and m_{nl}^S appear in the right-hand side of Eq. (S13) in symmetric positions, $\partial\rho/\partial m_{nl}^S > 0$ too. *Thus, the saturation in clearance model predicts increase in correlation between C_l and C_{nl} as the average normalized shedding rate from either leukocytes or non-leukocyte sources increases.*

Small shedding rates regime

For the case of small shedding rates $m_l^S, m_{nl}^S \ll 1$: $1-m_l^S \approx 1, 1-m_{nl}^S \approx 1$, $(m_l^S)^2 \approx 0$, $(m_{nl}^S)^2 \approx 0$. Thus, Eqs. (S11) for mean cfDNA concentrations are further simplified to

$$E[C_l] \approx \frac{km_l^S}{1-m_{tot}^S},\ E[C_{nl}] \approx \frac{km_{nl}^S}{1-m_{tot}^S},\ E[C_{tot}] \approx \frac{km_{tot}^S}{1-m_{tot}^S}. \quad \text{(S14a,b,c)}$$

From Eqs. (S14) we also deduce that

$$\frac{m_l^S}{m_{nl}^S} \approx \frac{E[C_l]}{E[C_{nl}]}, \quad \text{(S15)}$$

i.e. the ratio of average leukocyte to non-leukocyte shedding rates is approximated by the ratio of average leukocyte to non-leukocyte cfDNA concentrations.

Eqs. (S12) for the coefficients of variation are simplified to

$$CV_l^C \cong \frac{CV_l^S}{1-m_{tot}^S},\quad CV_{nl}^C \cong \frac{CV_{nl}^S}{1-m_{tot}^S}, \quad \text{(S16a,b)}$$

which implies that the saturation effect results in sightly increased coefficients of variation for cfDNA concentrations compared to the coefficients of variation of the respective shedding rates. From Eqs. (S16) we also deduce that

$$\frac{CV_l^S}{CV_{nl}^S} \approx \frac{CV_l^C}{CV_{nl}^C}, \quad \text{(S17)}$$

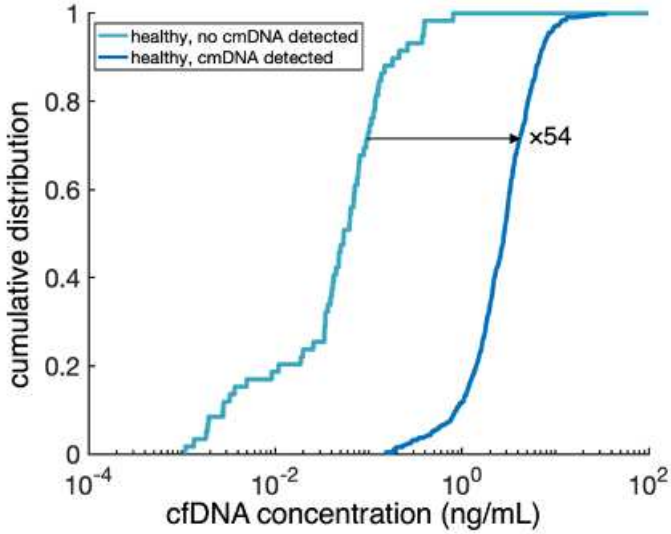
i.e. the ratio of the coefficients of variation for leukocyte to non-leukocyte shedding rates is approximated by the ratio of the coefficients of variation for leukocyte to non-leukocyte cfDNA concentrations.

Eq. (S13) for the correlation coefficient is simplified to

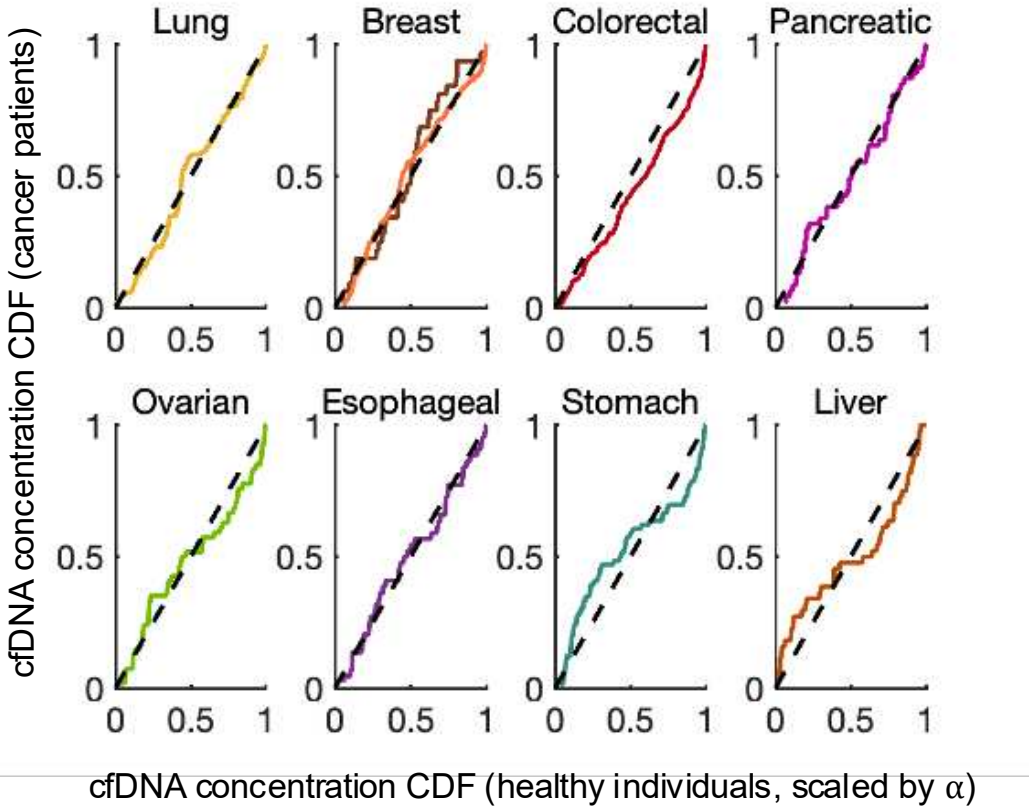
$$\rho \approx \frac{m_l^S\left(CV_l^S\right)^2+m_{nl}^S\left(CV_{nl}^S\right)^2}{CV_l^S CV_{nl}^S}. \quad \text{(S18)}$$

From Eq. (S18) we further deduce that, in the case where $CV_l^C \approx CV_{nl}^C$, and thus $CV_l^S \approx CV_{nl}^S$ from Eq. (S17), *correlation coefficient is approximated by the average total normalized shedding rate:*

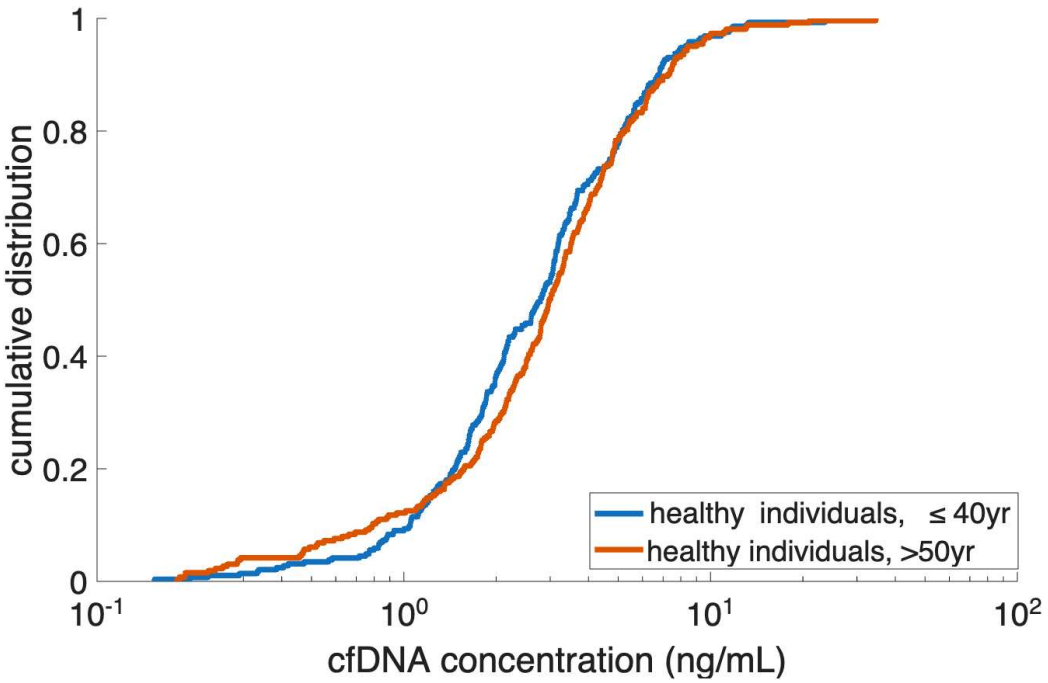
$$\rho \approx m_l^S + m_{nl}^S = m_{tot}^S. \tag{S19}$$



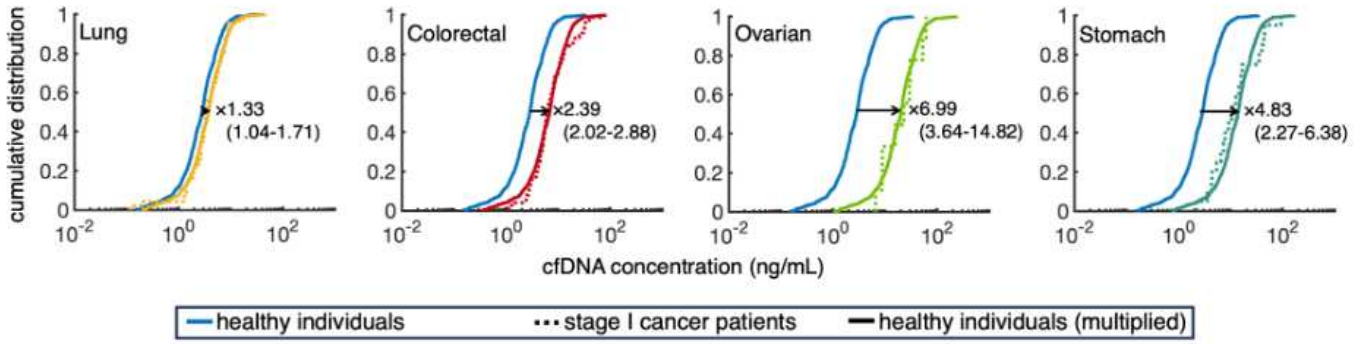
S1 Figure. Multiplicative factor effect in cfDNA concentration between the 572 healthy individuals with cmDNA detected and the 59 healthy individuals without cmDNA detected.



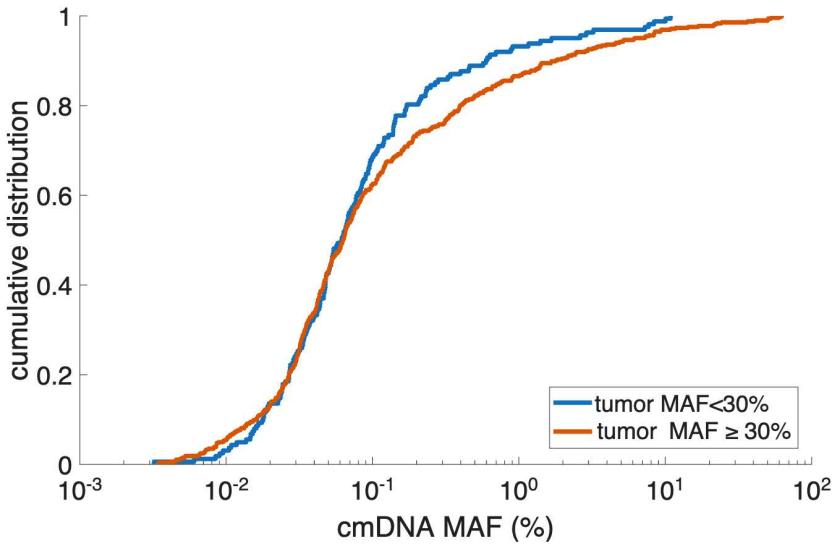
S2 Figure. P-P plots for the effect of multiplicative factor α in cfDNA concentration between the 572 healthy individuals and the 103 patients with lung cancer, 209 with breast cancer, 385 with colorectal cancer, 47 with pancreatic cancer, 54 with ovarian cancer, 44 with esophageal cancer, 66 with stomach cancer, and 44 with liver cancer.



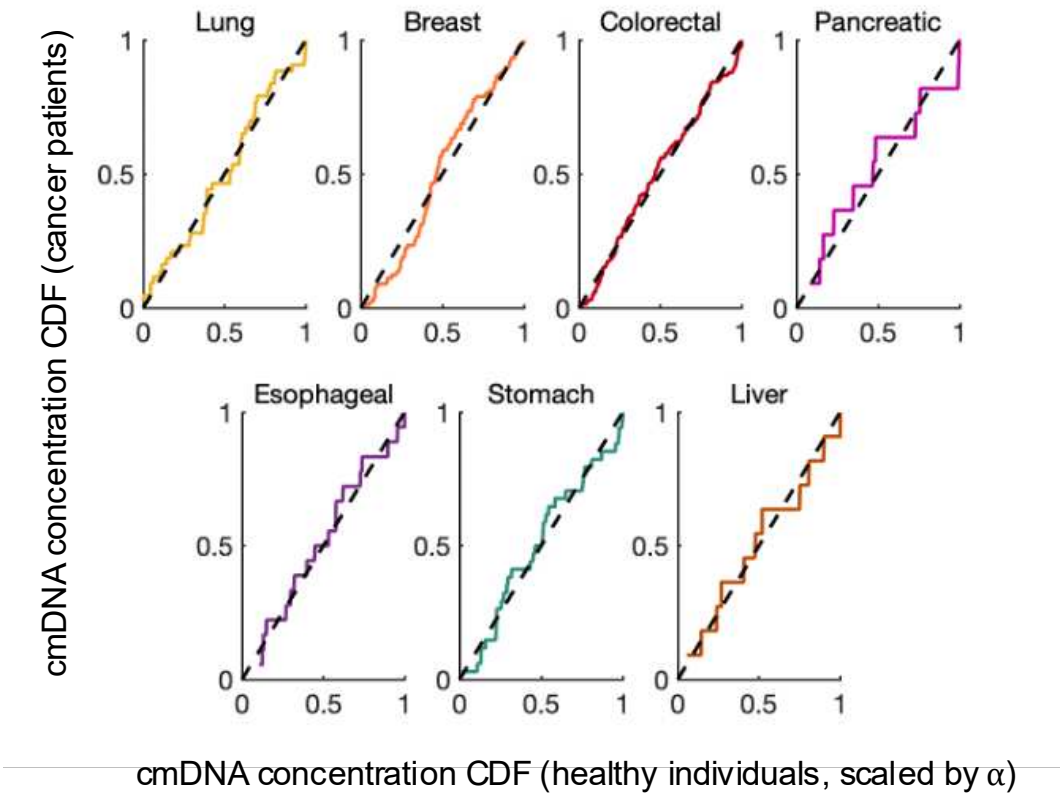
S3 Figure. Cumulative distribution function for the cfDNA concentrations of the 288 healthy individuals under 40 years, and of the 263 healthy individuals over 50 years of age.



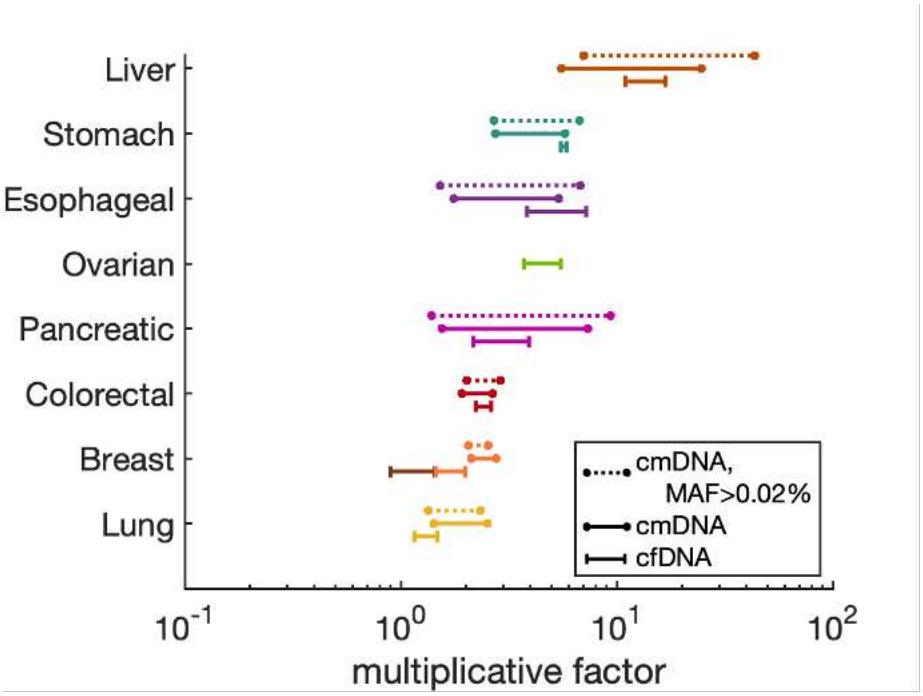
S4 Figure. Quantifying the multiplicative increase in cfDNA concentration between the 572 healthy individuals and the 46, 76, 9 and 20 patients with stage I lung, colorectal, ovarian and stomach cancer respectively.



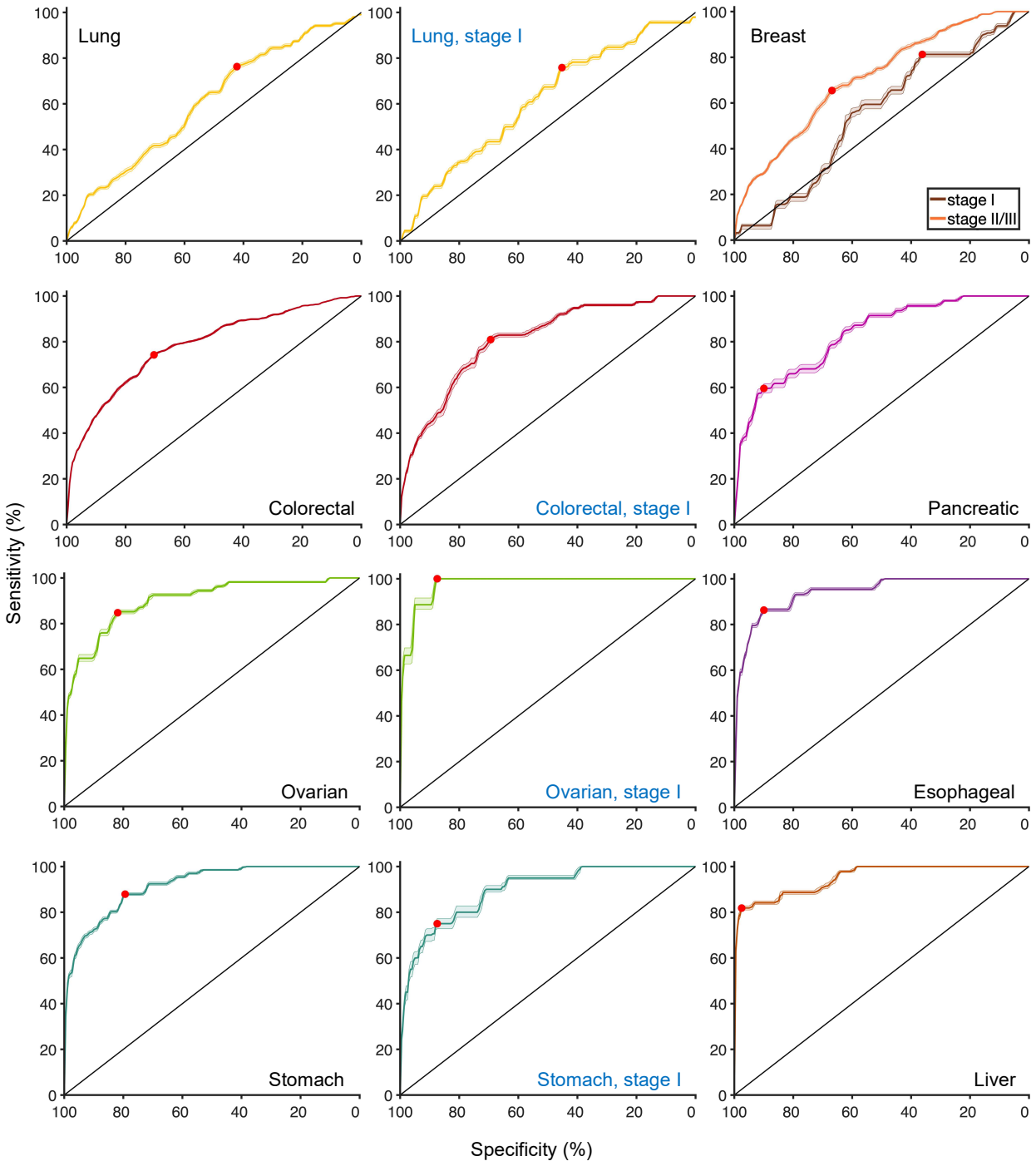
S5 Figure. Cumulative distribution functions for the MAF of the top plasma mutation for the 162 cancer patients with all mutations with MAF<30% in the tumor (blue line), and for the 484 cancer patients with at least one mutation with MAF≥30% in the tumor (red line).



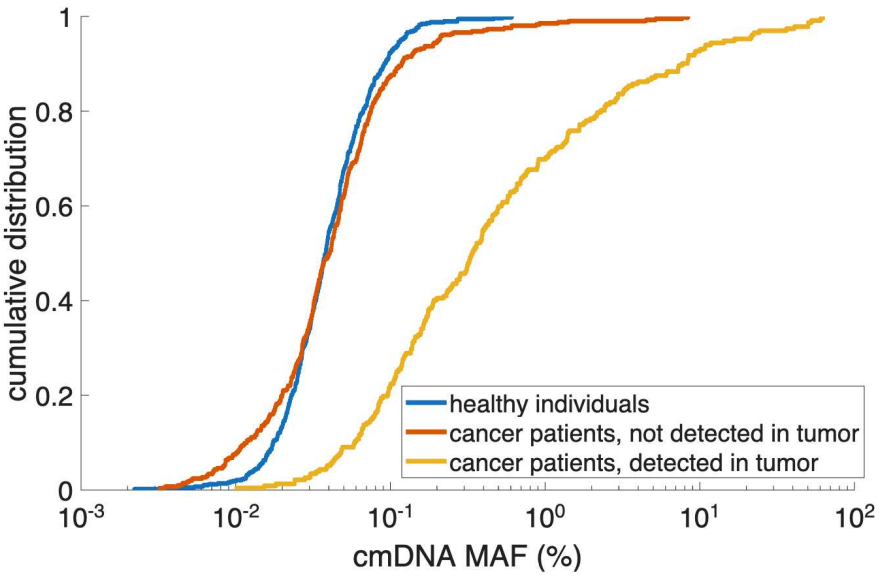
S6 Figure. P-P plots for the effect of multiplicative factor α in cmDNA concentration between the 572 healthy individuals and the 43, 90, 205, 11, 18, 34, 11 patients with lung, breast, colorectal, pancreatic, esophageal, stomach and liver cancer respectively, with the top plasma mutation not detected in the tumor.



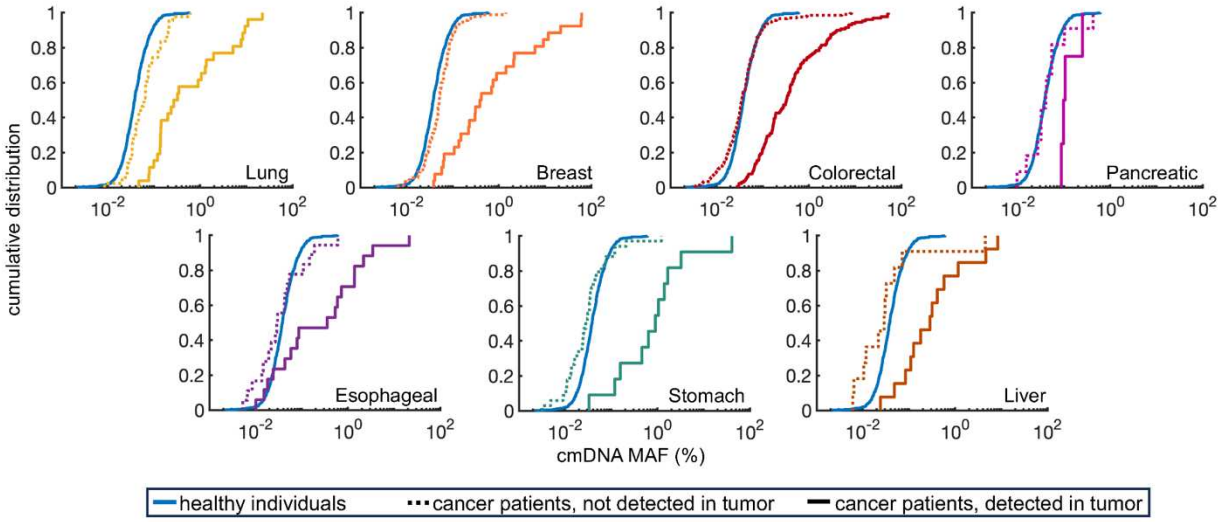
S7 Figure. Multiplicative factor ranges for each cancer type, quantifying the increase in concentrations of total cfDNA in plasma, and of cmDNA not shed from cancer tissue, in the presence of cancer.



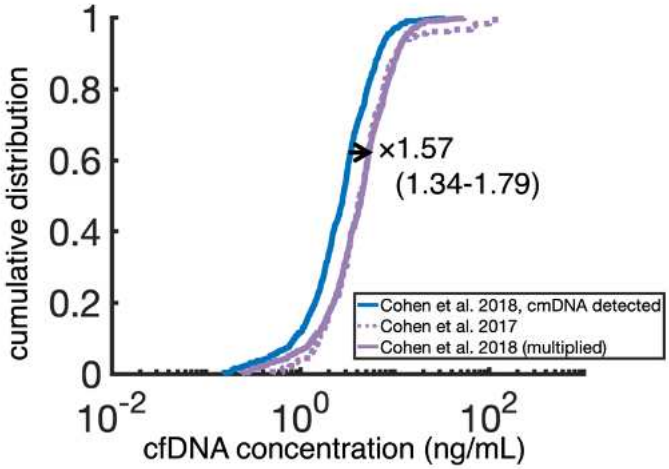
S8 Figure. Receiver Operating Characteristic (ROC) curves with 95% confidence intervals for cfDNA-based discrimination between cancer patients and healthy controls. Solid lines indicate mean ROC curves across 10-fold cross-validation; shaded regions represent bootstrap-derived confidence intervals. Red dots mark the optimal cutoff determined by Youden's J statistic on the mean ROC curve. Curves are shown for each cancer type (all stages) and for stage I subsets (lung, colorectal, ovarian, and stomach) to illustrate early detection performance. Pancreatic, esophageal, and liver cancers are not presented for stage I due to insufficient sample sizes.



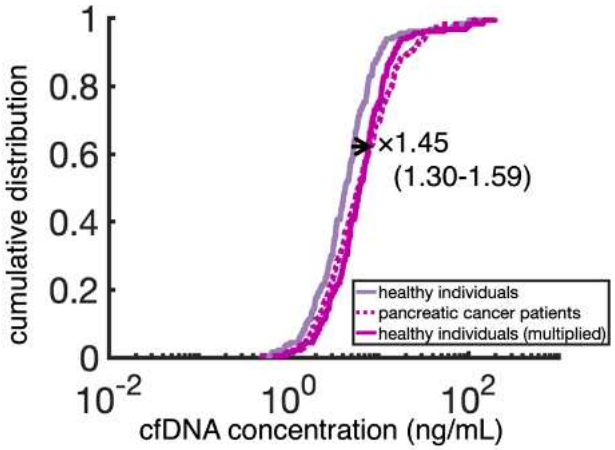
S9a Figure. Cumulative distribution function for the MAFs of top plasma mutation in 572 healthy individuals, 413 cancer patients whose top plasma mutation was not detected in tumor, and for 232 cancer patients whose top plasma mutation was detected in the tumor.



S9b Figure. Cumulative distribution function of plasma mutant allele frequencies (MAFs) for the top plasma mutation of three groups: healthy individuals (solid blue line), patients of each cancer type whose top plasma mutation was not detected in the tumor (dotted line), and patients of each cancer type whose top plasma mutation was detected in the tumor (solid line). Sample sizes for each cancer type: lung (detected/detected: 43/26), breast (90/26), colorectal (205/129), pancreatic (11/4), esophageal (18/17), stomach (34/11), and liver (11/13). Ovarian cancer is not depicted since in the dataset of 7 patients, there was only one with the top mutation not detected in tumor.



S10 Figure. Multiplicative factor effect in cfDNA concentration between the 572 and 182 healthy individuals reported in the two studies by Cohen et al. of 2018 and 2017 respectively.



S11 Figure. Multiplicative factor effect in cfDNA concentration between the 182 healthy individuals and the 221 pancreatic cancer patients reported in Cohen et al. 2017.

cfDNA dataset	data from [20]	exclude data included in [20] originating from [40]	exclude data with no cmDNA detected	data analyzed in Results	data from [40] analyzed separately in Discussion
healthy	812	181	59	572	182
lung cancer	104	-	1	103	-
breast cancer	209	-	-	209	-
colorectal cancer	388	-	3	385	-
pancreatic cancer	93	46	-	47	221
ovarian cancer	54	-	-	54	-
esophageal cancer	45	-	1	44	-
stomach cancer	68	-	2	66	-
liver cancer	44	-	-	44	-

S1a Table. Cell-free DNA data from Cohen *et al.* 2018 [20] and Cohen *et al.* 2017 [40] analyzed in the manuscript.

	cfDNA dataset analyzed in Results	Age in yrs median (range)	Sex M / F	Cancer Stage I / II / III
healthy	572	36.5 (17–87)	317 /255	-
lung cancer	103	68 (43–86)	62 / 41	46 / 27 / 30
breast cancer	209	57 (29–89)	5 / 204	32 / 114 / 63
colorectal cancer	385	65 (22–93)	210 / 175	76 / 191 / 118
pancreatic cancer	47	67 (44–86)	28 / 19	2 / 39 / 6
ovarian cancer	54	60 (37–81)	0 / 54	9 / 4 / 41
esophageal cancer	44	60 (30–82)	39 / 5	5 / 28 / 11
stomach cancer	66	65 (35–82)	51 / 15	20 / 29 / 17
liver cancer	44	62 (28–79)	34 / 10	5 / 19 / 20

S1b Table. Summary of cfDNA dataset analyzed in Results, including sample size, age distribution (median and range), sex distribution (M/F), and cancer stage proportions (I/II/III) for each cancer type and healthy individuals.

	Patients with mutations in tumor reported	Patients with no mutations detected in plasma	Top mutation in plasma not detected in tumor	Top mutation in plasma also detected in tumor
lung cancer	70	1	43	26
breast cancer	116	-	90	26
colorectal cancer	336	2	205	129
pancreatic cancer	15	-	11	4
ovarian cancer	7	-	1	6
esophageal cancer	36	1	18	17
stomach cancer	47	2	34	11
liver cancer	24	-	11	13

S1c Table. Summary of the cmDNA dataset. For each cancer type: the first column contains the number of patients whose mutations in tumor were reported; the remaining columns all pertain to these patients. The second column reports the number of patients with no plasma mutations detected (these patients are excluded from analysis in the Results). The last two columns report the number of patients for whom the top plasma mutation was not detected in the tumor versus also detected in the tumor.

Cancer	multiplicative factor	K-S *p*-value
Lung	1.29 (1.15-1.48)	0.62
Stage I Breast	1.15 (0.90-1.43)	0.78
Stage II/III Breast	1.79 (1.45-1.99)	0.48
Colorectal	2.54 (2.22-2.63)	0.30
Pancreatic	2.69 (2.16-3.93)	0.82
Ovary	4.30 (3.69, 5.59)	0.44
Esophageal	5.24 (3.85-7.20)	0.94
Stomach	5.66 (5.49-5.85)	0.10
Liver	12.63 (10.94-16.81)	0.28

S2 Table. Multiplicative factors for cfDNA concentration, for each cancer type.

Cancer	multiplicative factor	K-S *p*-value	multiplicative factor MAF >0.01%	K-S *p*-value	multiplicative factor MAF >0.015%	K-S *p*-value	multiplicative factor MAF >0.02%	K-S *p*-value
Lung	1.92 (1.42-2.52)	0.89	1.93 (1.45-2.52)	0.88	1.89 (1.41-2.45)	0.88	1.75 (1.34-2.34)	0.89
Breast	2.42 (2.12-2.76)	0.41	2.50 (2.22-2.79)	0.27	2.44 (2.16-2.66)	0.31	2.35 (2.06-2.54)	0.31
Colorectal	2.36 (1.92-2.66)	0.51	2.49 (1.98-2.70)	0.42	2.51 (2.00-2.74)	0.45	2.59 (2.02-2.89)	0.44
Pancreatic	4.01 (1.55-7.34)	0.86	4.01 (1.43-8.53)	0.82	4.01 (1.40-8.47)	0.82	4.72 (1.39-9.35)	0.78
Esophageal	3.31 (1.76-5.38)	0.99	3.36 (1.77-6.07)	0.99	3.64 (1.74-6.57)	0.99	3.47 (1.52-6.78)	0.99
Stomach	4.38 (2.74-5.77)	0.86	4.56 (2.99-6.25)	0.87	5.19 (2.98-7.70)	0.83	5.01 (2.70-6.71)	0.62
Liver	12.37 (5.54-24.62)	0.99	14.56 (6.88-37.14)	0.99	17.78 (7.19-46.06)	0.99	16.96 (7.03-43.43)	0.99

S3 Table. Multiplicative factors for cmDNA concentration not shed from tumor, for each cancer type.

Cancer Type	AUC (95% CI)	Optimal Cutoff	Sensitivity at Cutoff (95% CI)	Specificity at Cutoff (95% CI)	PPV at Cutoff (95% CI)	NPV at Cutoff (95% CI)
Lung	0.608 (0.594, 0.622)	2.457	0.763 (0.589, 0.778)	0.422 (0.408, 0.485)	0.179 (0.157, 0.200)	0.890 (0.862, 0.916)
Lung, stage I	0.616 (0.606, 0.633)	2.633	0.758 (0.535, 0.838)	0.452 (0.439, 0.526)	0.091 (0.081, 0.105)	0.949 (0.934, 0.965)
Breast, stage I	0.556 (0.524, 0.575)	2.455	0.813 (0.450, 0.833)	0.377 (0.362, 0.518)	0.060 (0.051, 0.068)	0.965 (0.945, 0.983)
Breast, stages I, II	0.710 (0.697, 0.723)	3.697	0.655 (0.575, 0.727)	0.668 (0.630, 0.705)	0.368 (0.320, 0.412)	0.860 (0.825, 0.893)
Colorectal	0.784 (0.778, 0.788)	4.105	0.742 (0.700, 0.766)	0.704 (0.676, 0.718)	0.629 (0.596, 0.669)	0.799 (0.775, 0.820)
Colorectal, stage I	0.809 (0.787, 0.828)	4.081	0.809 (0.657, 0.861)	0.693 (0.657, 0.739)	0.269 (0.222, 0.318)	0.960 (0.937, 0.980)
Pancreatic	0.830 (0.814, 0.851)	6.307	0.595 (0.420, 0.667)	0.899 (0.738, 0.902)	0.233 (0.158, 0.316)	0.961 (0.945, 0.977)
Ovarian	0.901 (0.890, 0.908)	5.453	0.848 (0.670, 0.880	0.819 (0.797, 0.869)	0.314 (0.267, 0.360)	0.980 (0.967, 0.990)
Ovarian, stage I	0.976 (0.968, 0.989)	6.776	1.000 (0.608, 1.000)	0.874 (0.862, 0.925)	0.099 (0.025, 0.174)	0.998 (0.994, 1.000)
Esophageal	0.941 (0.936, 0.950)	7.026	0.863 (0.748, 0.950)	0.899 (0.888, 0.932)	0.429 (0.373, 0.479)	0.987 (0.981, 0.993)
Stomach	0.918 (0.909, 0.928)	5.111	0.879 (0.795, 0.940)	0.794 (0.776, 0.834)	0.340 (0.317, 0.361)	0.981 (0.975, 0.989)
Stomach, stage I	0.896 (0.883, 0.930)	5.859	0.750 (0.550, 0.850)	0.874 (0.759, 0.872)	0.098 (0.054, 0.141)	0.985 (0.976, 0.995)
Liver	0.951 (0.944, 0.959)	11.519	0.818 (0.688, 0.885)	0.975 (0.968, 0.988)	0.791 (0.670, 0.893)	0.984 (0.977, 0.991)

S4 Table: ROC-based performance metrics for cfDNA-driven cancer detection across cancer types and stages. The table reports AUC values with 95% confidence intervals, optimal cutoff points, sensitivity and specificity at the cutoff (with 95% CIs), and predictive values (PPV and NPV) at the cutoff (with 95% CIs).