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

Konstantinos Mamis[1] and 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

*Email: ibozic@uw.edu

**Abstract**
Cell-free DNA (cfDNA) is a promising biomarker for cancer detection. However, cfDNA dynamics in the presence of cancer 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 from healthy tissue due to cancer may lead to the larger observed multiplicative increases in cfDNA levels.

## MAIN TEXT

### 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 (3-6), especially in advanced and metastatic stages (7). Furthermore, Stroun *et al.* (8) 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 (9-11). 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 (12). 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 from Cohen *et al.* (13), 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.* (12) 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.* (13) report the cfDNA concentrations in plasma of 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 (Materials and Methods, Table S1); we refer to this portion of cfDNA as circulating mutant DNA (cmDNA). We first observe that the 631 healthy individuals (Materials and Methods) can be divided into two groups: 572 (90.6%) that have cmDNA detected and have median cfDNA concentration 2.83ng/mL, and 59 (9.4%) that have no mutations detected in plasma and exhibit significantly (54 times) lower cfDNA concentrations (median 0.05ng/mL) (Fig. S1). 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.

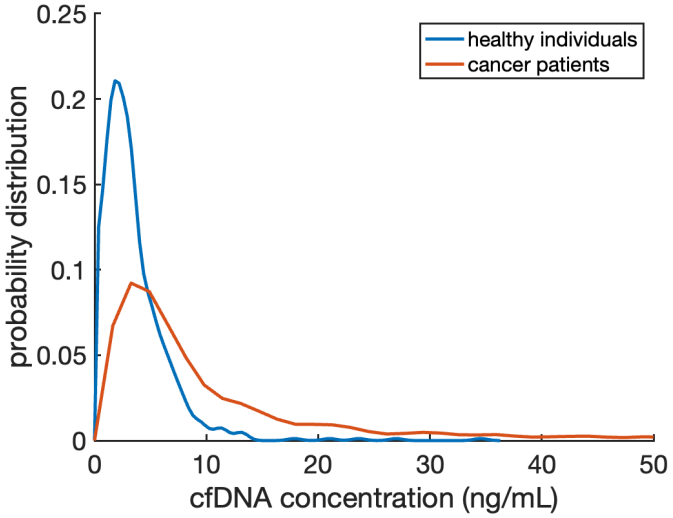


**Figure 1. Probability distribution of cfDNA concentrations in healthy individuals and patients with stage I-III cancer.** Probability density functions for the cfDNA concentration in healthy individuals with mutations detected in plasma (*N*=572) and patients with lung, breast, colorectal, pancreatic, ovarian, esophageal, stomach and liver cancer in stages I-III (*N*=952), as reported in Cohen *et al.* (2018) (13).

**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 (7, 11). 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 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 is plotted on a logarithmic *x*-scale (Fig. 2).

The multiplicative increase in cfDNA concentration 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, Fig. S2, Table S2, Materials and Methods). 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

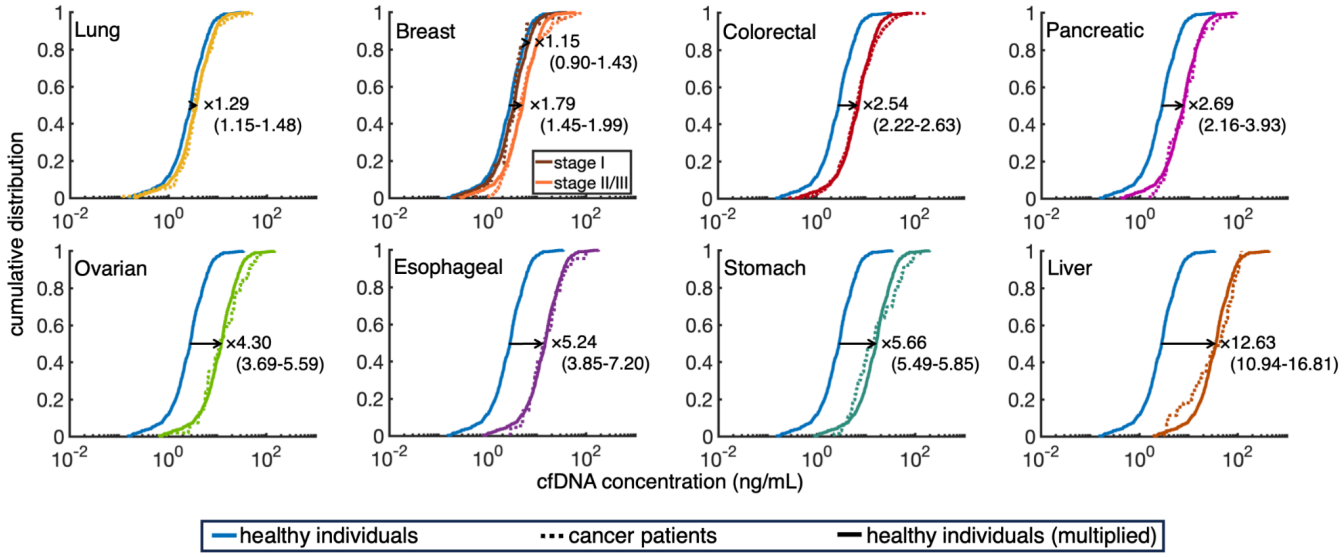


**Figure 2. Quantifying the multiplicative increase in cfDNA concentration in the presence of cancer.** Cumulative distribution functions for the cfDNA concentration in healthy individuals with mutations detected in plasma (*N*=572) and patients with lung (*N*=103), stage I breast (*N*=32), stage II/III breast (*N*=177), colorectal (*N*=385), pancreatic (*N*=47), ovarian (*N*=54), esophageal (*N*=44), stomach (*N*=66) and liver (*N*=44) cancer. 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.

To determine if a multiplicative increase in cfDNA concentration is also present in the earliest cancer cases, we calculate the multiplicative factor 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 a multiplicative factor of 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 (Fig. S3). These ranges for multiplicative factors overlap with the ranges calculated when all stage I-III cancer patients were 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. We find that the mutant allele frequency (MAF) for the mutations detected in plasma of cancer patients of all types is low - median MAF is 0.06%, and 88.2% of mutations detected in plasma have MAF below 1%. Combined with the fact that 82% of cancer patients' tumors harbor mutations (at high frequency) that can be detected by the liquid biopsy panel (13), this indicates that early-stage tumors in all cancer types analyzed do not contribute significantly to cfDNA concentration; most of cfDNA originates from non-tumor (healthy) tissue.

**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 we use (13), both the top mutation in plasma and the mutations detected in tumor are reported for 645 of the 959 cancer patients (Materials and Methods). Comparing the cmDNA concentration in healthy individuals and in cancer patients in which cmDNA detected in plasma is not detected in tumor, 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, Fig. S4, Table S3, 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.

We observe that the ranges of multiplicative factor values for cfDNA and cmDNA from healthy tissue are overlapping for each of the cancer types (Fig. S5), except for breast cancer where ranges are comparable yet non-overlapping. We found no significant differences in cmDNA concentrations between cancer stages or sexes.

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.* (13) which we analyze in this work were generated using the SafeSeqS approach for PCR, developed to reduce the introduction of artifactual mutations (14). In subsequent work (15) 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 (Fig. S5, Table S3, Materials and Methods).

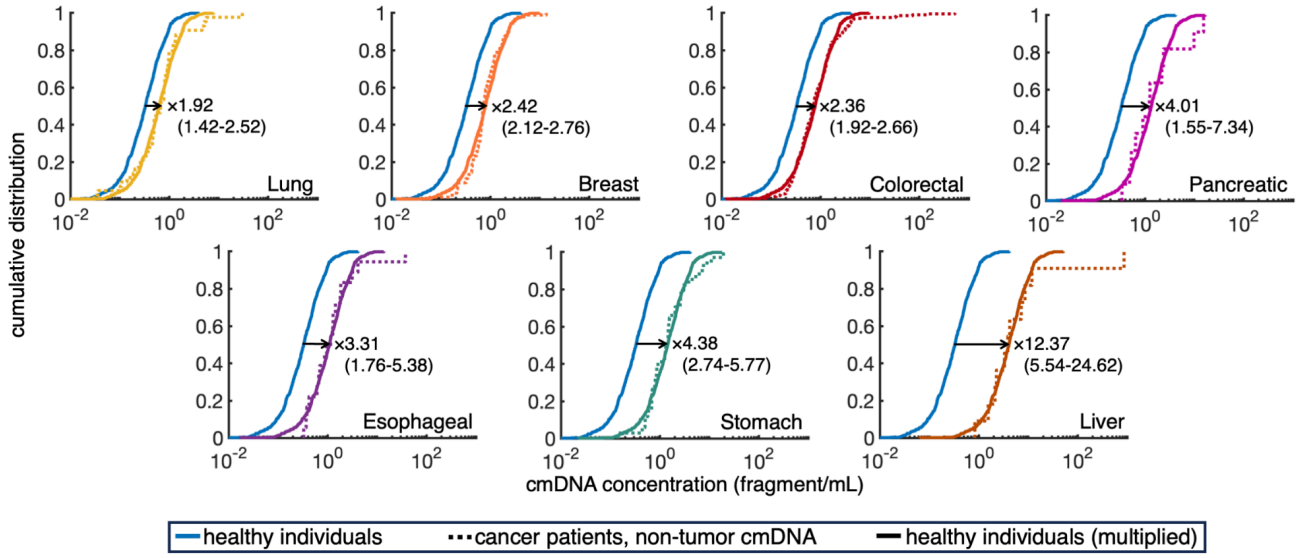


**Figure 3. Mutations detected in plasma of cancer patients but not detected in tumor exhibit a multiplicative increase in concentration compared to mutations detected in plasma of healthy individuals.** Cumulative distribution functions for the concentration of cmDNA found in plasma of healthy individuals (*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. Median empirical cmDNA concentrations (corresponding to 0.5 value of cumulative distribution) are 0.33 fragment/mL for healthy individuals, and 0.67, 0.73, 0.72, 1.23, 1.05, 1.39, and 3.85 fragment/mL for patients with lung, breast, colorectal, pancreatic, esophageal, stomach and liver cancer, respectively.

**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.* (12) 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.* (12) 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.* (16) and Moss *et al.* (17), 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 a factor 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.* (12) and Cohen *et al.* (13) 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.* (12).

In healthy individuals from Mattox *et al.* (12), 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.* (16) is used. Use of the reference deconvolution matrix by Moss *et al.* (17) 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 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, Supplementary 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).

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

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

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 $\varepsilon = \ln(2)/t_{1/2}$. Assuming that cfDNA is in equilibrium, with cfDNA shed being equal to cfDNA cleared, the cfDNA concentrations under the linear model (1) are

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

Under Eqs. (2) of the linear clearance model, the increases in cfDNA concentrations from leukocytes and non-leukocyte sources, $C_l$ and $C_{nl}$, so may originate from corresponding increases in cfDNA shedding rates, or from decreases in clearance rate ε. From the data by Mattox *et al.* (12), 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 determine if the correlation in the multiplicative increases in cfDNA shedding in the presence of cancer increased correlation between cfDNA concentrations from leukocytes and non-leukocytes, we allow the multiplicative factors for leukocyte and non-leukocyte cfDNA to vary uniformly from patient to patient, while also being perfectly correlated. For each realization, we multiply the leukocyte and non-leukocyte cfDNA concentrations for every healthy individual by the varying multiplicative factors, and then calculate the correlation between the cfDNA concentrations.

We consider the following ranges of values for multiplicative factors: (3.15-4.45) and (6.24-9.94) for pancreatic cancer, and (2.03-3.37), (2.39-4.49) for ovarian cancer. We then perform the random selection of multiplicative factors from the said intervals 500,0000 times. The resulting correlation remains low: average correlation coefficient -0.07 with standard deviation 0.01 for pancreatic cancer, and -0.03$\pm$0.04 for ovarian cancer. The results regarding cfDNA correlation remain practically unaltered for the case of multiplicative factors that are uncorrelated or with correlation coefficient 0.5. We conclude that the linear clearance model cannot easily account for observed behaviors of cfDNA dynamics.

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

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

with $v$ and $k$ being positive constants. In Eq. (3c) for the total cfDNA concentration, clearance term is of Michaelis-Menten type: 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 (18, 19). 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:

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

where $\tilde{S}_l$, $\tilde{S}_{nl}$ are the shedding rates normalized by the maximum clearance rate; $\tilde{S}_l = S_l/v$, $\tilde{S}_{nl} = S_{nl}/v$. In order for model (3) to predict finite cfDNA concentrations, the total shedding rate $S_{tot}$ should always be less than the maximum clearance rate $v$. 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 Supplementary Material).

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 $\widetilde{S}_l$, $\widetilde{S}_{nl}$ 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.* (12) 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.* (12).

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.* (12) 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. 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.* (12) for different deconvolution methods.

**ROC curves for cancer detection using cfDNA levels.** Finally, in order to determine the implications of multiplicative cfDNA increase in cancer detection, we determined the Receiver Operating Characteristic (ROC) curves, for each cancer type individually (Fig. S6). ROC curves are calculated from the logistic regression fit to cfDNA concentration data in healthy individuals and patients with stage I-III cancer. We observe that, in cancer types with high multiplicative factors for cfDNA concentrations, such as in ovarian, esophageal, stomach and liver cancer, the area under curve (AUC) approaches 1, being equal or greater than 0.9. This implies that for cancer types with high multiplicative cfDNA factors, cfDNA concentration can be used as an indication of tumor presence.

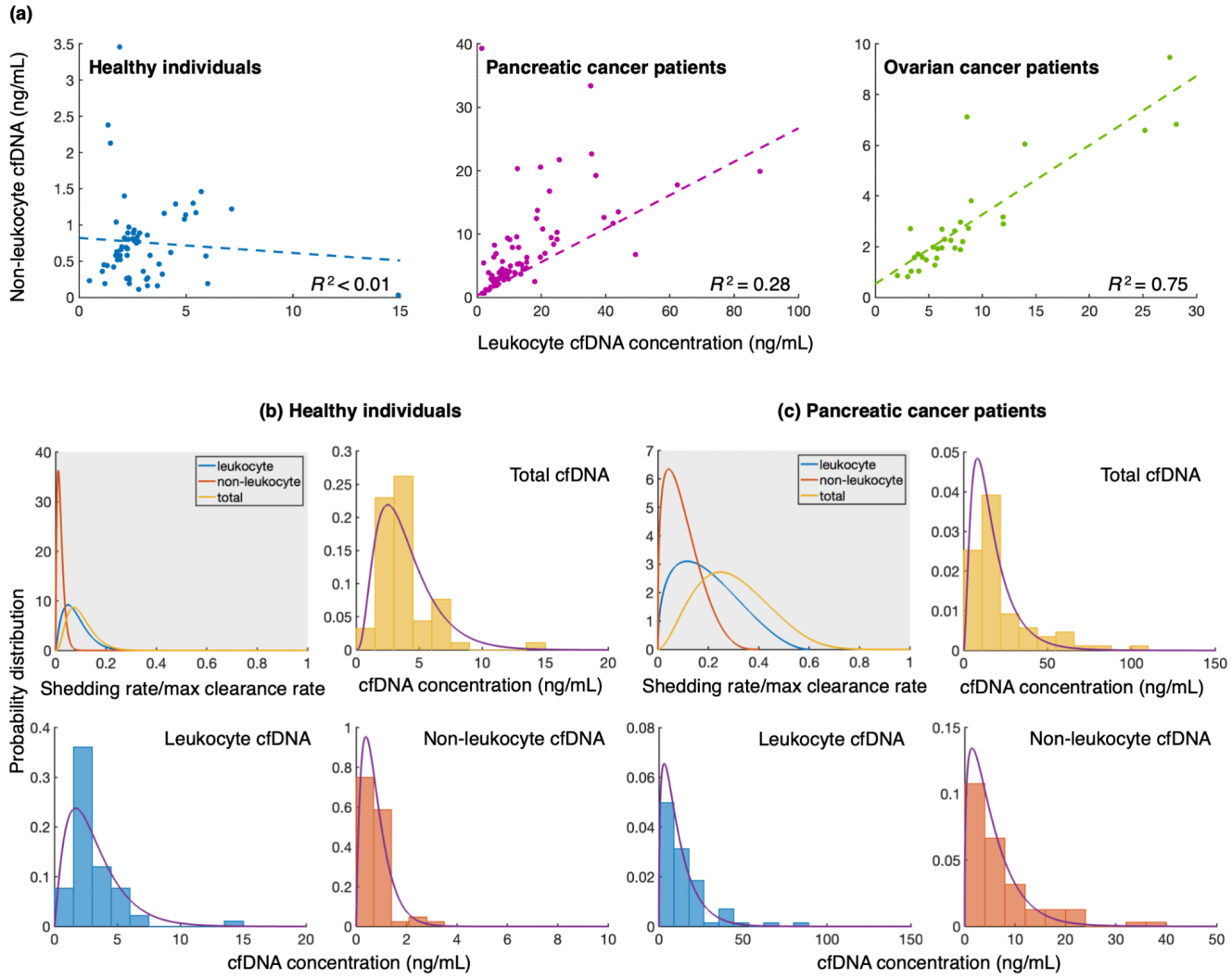


**Figure 4. Correlation and distributions of the leukocyte, non-leukocyte and total cfDNA concentrations. (a)** Correlation between leukocyte and non-leukocyte concentrations, excluding outliers with high cfDNA concentration. Dots represent the pair of leukocyte and non-leukocyte cfDNA concentration for normal individuals and pancreatic and ovarian cancer patients from Mattox *et al.* (12) under a quadratic programming deconvolution algorithm employing Sun *et al.* (16) reference matrix. **(b), (c)** Gray panel: Estimated leukocyte, non-leukocyte and total cfDNA shedding rates normalized by the maximum clearance rate $v$, considered for the case of healthy individuals (b) and pancreatic cancer patients (c). 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.* (12), under the same deconvolution method as for (a).

## 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 ~30% in lung to a 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 the cmDNA of each cancer patient as non-tumor or tumor cmDNA, by checking whether the mutation detected in plasma is among the mutations detected in tumor. We note that a mutation detected in plasma but not in tumor may originate from the tumor if the tumor biopsy has missed to identify its presence. However, we do not expect this to have a significant effect on our conclusions as we observe a statistically significant difference in mutant allele frequency of plasma mutations in cancer patients, based on their origin (detected or not detected in tumor). Furthermore, for all cancer types, the distribution of MAF of cmDNA not detected in the tumor stays close to the distribution of cmDNA MAF in healthy individuals (Fig. S7). This is also reflected in the median cmDNA MAF, which for healthy individuals is 0.04%, for cmDNA not detected in tumor lies between 0.03% and 0.06%, while for cmDNA detected in the tumor it is an order of magnitude higher, 0.28% to 0.90%. Thus, without being able to rule out with certainty a possible miscategorization as tumor or non-tumor cmDNA for a particular cancer patient, we conclude that this distinction in MAFs indicates that the cmDNA shedding cell population is different in healthy individuals/cancer patients with cmDNA not detected in the tumor, and cmDNA in cancer patients shed from the tumor.

We also analyze an additional dataset (12) 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. 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) (20) 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) (13). 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), Fig. S8); 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 (21). Cohen *et al.* (2017) (20) 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) (13), 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) (20) is 1.45 (1.30-1.59) (Fig. S9). This factor is ~2 times lower than the multiplicative factor for pancreatic cancer in Cohen *et al.* (2018) (13) data, as expected.

## Materials and Methods

*Patient data*

We perform a statistical analysis of the concentration (ng/mL) of cfDNA detected in plasma for 812 healthy individuals and for: 104 patients with lung cancer, 209 with breast cancer, 388 with colorectal cancer, 93 with pancreatic cancer, 54 with ovarian cancer, 45 with esophageal cancer, 68 with stomach cancer, and 44 with liver cancer reported by Cohen *et al.* (13). All cancer cases reported are in stages I to III. The dataset includes samples from 181 healthy individuals and 46 pancreatic cancer evaluated in a previous study (15). In order to ensure that all samples were analyzed following the same process, we exclude these samples from the present work (Table S1).

Cohen *et al.* (13) also report the concentration (fragment/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.* (13) 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 in plasma, mutations found in tumor are also reported for 336 of the patients with colorectal, 70 with lung, 116 with breast, 15 with pancreatic, 7 with ovarian, 36 with esophageal, 47 with stomach, and 24 with liver cancer.

*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 (23), 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.

*Multiplicative factors for cfDNA/cmDNA*

We determine the value of multiplicative factor $\alpha$ for which we obtain the minimum Kolmogorov distance between the empirical CDF for $\alpha$ times the sample of cfDNA concentration in healthy patients 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 $\alpha$, and the $p$-value of K-S test between the sample of cfDNA concentration in healthy patients with mutations detected in plasma multiplied by $\alpha$, and the cfDNA concentration in patients of each cancer type (Table S2).

We also report the value range of the multiplicative factor $\alpha$, 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 patients multiplied by the respective factor come from the same distribution. We also depict the P-P plots as the visual counterpart of the K-S test (22), 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. In the P-P plots for each cancer type, we observe good agreement between the empirical CDFs for cfDNA concentrations in cancer patients and in healthy individuals multiplied by the cancer type-specific multiplicative factor $\alpha$ (Fig. S2).

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. S4. 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 (15).

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

For the dataset of healthy individuals from Mattox *et al.* (12), we estimate parameter $k$ of the saturation model, 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 $\widetilde{S}_l$, $\widetilde{S}_{nl}$ from leukocytes and non-leukocyte sources in the absence of cancer. First, we deduce 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, Supplementary Material):

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

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

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

From Eqs. (5), (6a), and for a mean total cfDNA concentration $m_{tot}^C = 3.8 ng/mL$ reported in healthy individuals by Mattox *et al.* (12), we estimate model parameter $k$ to $34.2 ng/mL$. 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 $\widetilde{S}_l$, $\widetilde{S}_{nl}$ 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, Supplementary Material):

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

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, 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

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

where $\mathrm{B}(\alpha,\beta)$ is the Beta function, and $c = 0.6$ for the leukocyte, and $c = 0.4$ for the non-leukocyte shedding rate. We choose the ratio of leukocyte to non-leukocyte maximum shedding rates from the ratio of leukocyte to non-leukocyte cell turnover rates, ~1.5 (23). The rescaled Beta distribution parameters α, β are determined via the mean $m^S$ and coefficient of variation $CV^S$ of each normalized shedding rate

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

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 $C_l$, $C_{nl}$:

$$f_{C_lC_{nl}}(x,y) = \frac{k}{(k+x+y)^3} f_{\tilde{S}}\left(\frac{x}{k+x+y};\alpha_l,\beta_l,c_l\right) f_{\tilde{S}}\left(\frac{y}{k+x+y};\alpha_{nl},\beta_{nl},c_{nl}\right). \tag{10}$$

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.

**References**


1. D. Crosby, S. Bhatia, K. M. Brindle, L. M. Coussens, C. Dive, M. Emberton, S. Esener, R. C. Fitzgerald, S. S. Gambhir, P. Kuhn, T. R. Rebbeck, S. Balasubramanian, Early detection of cancer. *Science* **375,** eaay9040 (2022).
2. P. Mandel, P. Metais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C. R. Seances Soc. Biol. Fil.* **142,** 241–3 (1948).
3. O. B. Alese, N. Cook, A. Ortega-Franco, M. B. Ulanja, L. Tan, J. Tie, Circulating Tumor DNA: An Emerging Tool in Gastrointestinal Cancers. *ASCO Educational Book* **42,** 279–298 (2022).
4. A. J. Bronkhorst, V. Ungerer, S. Holdenrieder. The emerging role of cell-free DNA as a molecular marker for cancer management. *Biomol. Detect. Quantif.* **17,** 100087 (2019).

5. F. Diehl, K. Schmidt, M. A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokoll, S. A. Szabo, K. W. Kinzler, B. Vogelstein, L. A. Diaz Jr., Circulating mutant DNA to assess tumor dynamics. *Nat. Med.,* **14,** 985–990 (2008).
6. L. Roncucci, M. Ponz De Leon, A. Scalmati, G. Malagoli, S. Pratissoli, M. Perini, N. J. Chahin, The influence of age on colonic epithelial cell proliferation. *Cancer* **62,** 2373–2377 (1988).
7. C. Bettegowda, M. Sausen, R. J. Leary, I. Kinde, Y. Wang, N. Agrawal, B. R. Bartlett, H. Wang, B. Luber, R. M. Alani, E. S. Antonarakis, N. S. Azad, A. Bardelli, H. Brem, J. L. Cameron, C. C. Lee, L. A. Fecher, G. L. Gallia, P. Gibbs, D. Le, R. L. Giuntoli, M. Goggins, M. D. Hogarty, M. Holdhoff, S.-M. Hong, Y. Jiao, H. H. Juhl, J. J. Kim, G. Siravegna, D. A. Laheru, C. Lauricella, M. Lim, E. J. Lipson, S. K. N. Marie, G. J. Netto, K. S. Oliner, A. Olivi, L. Olsson, G. J. Riggins, A. Sartore-Bianchi, K. Schmidt, l-M. Shih, S. M. Oba-Shinjo, S. Siena, D. Theodorescu, J. Tie, T. T. Harkins, S. Veronese, T-L Wang, J. D. Weingart, C. L. Wolfgang, L. D. Wood, D. Xing, R. H. Hruban, J. Wu, P. J. Allen, C. M. Schmidt, M. A. Choti, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, N. Papadopoulos, L. A. Diaz Jr, Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **6,** (2014).
8. M. Stroun, P. Anker, P. Maurice, J. Lyautey, C. Lederrey, M. Beljanski, Neoplastic Characteristics of the DNA Found in the Plasma of Cancer Patients. *Oncology* **46,** 318–322 (1989).
9. A. Kustanovich, R. Schwartz, T. Peretz, A. Grinshpun. Life and death of circulating cell-free DNA. *Cancer Biol. Ther.* **20,** 1057–1067 (2019).
10. G. Siravegna, S. Marsoni, S. Siena, A. Bardelli. Integrating liquid biopsies into the management of cancer. *Nat. Rev. Clin. Oncol.* **14,** 531–548 (2017).
11. F. Diehl, M. Li, D. Dressman, Y. He, D. Shen, S. Szabo, L. A. Diaz Jr, S. N. Goodman, K. A. David, H. Juhl, K. W. Kinzler, B. Vogelstein. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. USA* **102,** 16368–16373 (2005).
12. A. K. Mattox, C. Douville, Y. Wang, M. Popoli, J. Ptak, N. Silliman, L. Dobbyn, J. Schaefer, S. Lu, A. H. Pearlman, J. D. Cohen, J. Tie, P. Gibbs, K. Lahouel, C. Bettegowda, R. H. Hruban, C. Tomasetti, P. Jiang, K.C. Allen Chan, Y. M. D. Lo, N. Papadopoulos, K. W. Kinzler, B. Vogelstein. The Origin of Highly Elevated Cell-Free DNA in Healthy Individuals and Patients with Pancreatic, Colorectal, Lung, or Ovarian Cancer. *Cancer Discov.* **13,** 2166–2179 (2023).
13. J. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, C. Douville, A. Javed, F. Wong, A. Mattox, R. Hruban, C. Wolfgang, M. Goggins, M. Dal Molin, T. Wang, R. Roden, A. Klein, J. Ptak, L. Dobbyn, J. Schaefer, N. Silliman, M. Popoli, J. Vogelstein, J. Browne, R. Schoen, R. Brand, H. Wong, A. Mansfield, J. Jen, S. Hanash, M. Falconi, P. Allen, S. Zhou, C. Bettegowda, L. Diaz Jr , C. Tomasetti, K. Kinzler, B. Vogelstein, A. Lennon, N. Papadopoulos. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359,** 926–930 (2018).
14. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108,** 9530–9535 (2011).
15. J. Cohen, C. Douville, J. Dudley, B. Mog, M. Popoli, J. Ptak, L. Dobbyn, N. Silliman, J. Schaefer, J. Tie, P. Gibbs, C. Tomasetti, N. Papadopoulos, K. Kinzler, B. Vogelstein, Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. *Nat. Biotechnol.* **39,** 1220–1227 (2021).
16. K. Sun, P. Jiang, K. Chan, J. Wong, Y. Cheng, R. Liang, W. Chan, E. Ma, S. Chan, S. Cheng, R. Chan, Y. Tong, S. Ng, R. Wong, D. Hui, T. Leung, T. Leung, P. Lai, R. Chiu, Y. Lo, Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments, *Proc. Natl. Acad. Sci. USA* **112**, E5503–E5512 (2015).

17. J. Moss, J. Magenheim, D. Neiman, H. Zemmour, N. Loyfer, A. Korach, Y. Samet, M. Maoz, H. Druid, P. Arner, K. Fu, E. Kiss, K. Spalding, G. Landesberg, A. Zick, A. Grinshpun, A. Shapiro, M. Grompe, A. Wittenberg, B. Glaser, R. Shemer, T. Kaplan, Y. Dor, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease, *Nat Commun* **9**, 5068 (2018).
18. D. S. C. Han, Y. M. D. Lo, The Nexus of cfDNA and Nuclease Biology, *Trends Genet.* **37,** 758–770 (2021).
19. D. Lo, J. Zhang, T. Leung, T. Lau, A. Chang, M. Hjelm, Rapid clearance of fetal DNA from maternal plasma, *Am. J. Hum. Genet.* **64,** 218–224 (1999).
20. J. Cohen, A. Javed, C. Thoburn, F. Wong, J. Tie, P. Gibbs, C. Schmidt, M. Yip-Schneider, P. Allen, M. Schattner, R. Brand, A. Singhi, G. Petersen, S. Hong, S. Kim, M. Falconi, C. Doglioni, M. Weiss, N. Ahuja, J.He, M. Makary, A. Maitra, S. Hanash, M. Dal Molin, Y. Wang, L. Li, J. Ptak, L. Dobbyn, J. Schaefer, N. Silliman, M. Popoli, M. Goggins, R. Hruban, C. Wolfgang, A. Klein, C. Tomasetti, N. Papadopoulos, K. Kinzler, B. Vogelstein, A. Lennon, Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proc. Natl. Acad. Sci. USA* **114,** 10202–10207 (2017).
21. K. Prakash, S. Aggarwal, S. Bhardwaj, G. Ramakrishna, C. K. Pandey, Serial perioperative cell-free DNA levels in donors and recipients undergoing living donor liver transplantation. *Acta Anaesthesiol. Scand.* **61,** 1084–1094 (2017).
22. NIST/SEMATECH. e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/. (2023).
23. R. Sender, R. Milo, The distribution of cellular turnover in the human body, *Nat. Med.,* **27**, 45–48, (2021).

**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).

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

**Competing interests**: The authors declare no competing interests.

**Data and materials availability:** The data analyzed in this study were obtained from Supplementary Materials of
Cohen *et al.* (2018) [13] https://doi.org/10.1126/science.aar3247
Cohen *et al.* (2017) [21] https://doi.org/10.1073/pnas.1704961114
Mattox *et al.* (2023) [12] https://doi.org/10.1158/2159-8290.CD-21-1252

## SUPPLEMENTARY MATERIALS

### Supplementary Text

Correlation coefficients between leukocyte and non-leukocyte cfDNA

For the dataset and the four different deconvolution methods from Mattox *et al.* (12), 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.* (16) QP | Sun *et al.* (16) NNLS | Moss *et al.* (17) QP | Moss *et al.* (17) 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.* (16) QP | Sun *et al.* (16) NNLS | Moss *et al.* (17) QP | Moss *et al.* (17) 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.* (16) QP | Sun *et al.* (16) NNLS | Moss *et al.* (17) QP | Moss *et al.* (17) 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.* (16) QP | Sun *et al.* (16) NNLS | Moss *et al.* (17) QP | Moss *et al.* (17) 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.

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 $k$,

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

with $E[\bullet]$ being the expectation operator. The Pearson correlation coefficient ρ and the coefficients of variation $CV_l^C$ and $CV_{nl}^C$ are $k$ −independent:

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

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

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

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

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

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

We choose $(x, y) = (\tilde{S}_l, \tilde{S}_{nl})$, and $\left(x_0, y_0\right) = (m_l^S, m_{nl}^S)$ with $m_l^S = E\left[\tilde{S}_l\right]$, $m_{nl}^S = E\left[\tilde{S}_{nl}\right]$ being the mean values, and take the expected value in both sides of Eq. (8)

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

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

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

In the right-hand side of Eq. (S9), the terms containing the first-order partial derivatives of $f$ are equal to zero, since $E\left[\tilde{S}_l - m_l^S\right] = 0$, $E\left[\tilde{S}_{nl} - m_{nl}^S\right] = 0$. 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

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

where $M_l^{(n)} = E\left[\left(\widetilde{S}_l - m_l^S\right)^n\right]$, $M_{nl}^{(n)} = E\left[\left(\widetilde{S}_{nl} - m_{nl}^S\right)^n\right]$ denote the centered moments of $\widetilde{S}_l$, $\widetilde{S}_{nl}$ 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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

By calculating the derivative of ρ with respect to $m_l^S$ we get

$$\frac{\partial\rho}{\partial m_l^S} = \frac{\left(1-m_{tot}^S\right)m_l^S\left(m_{nl}^S\right)^3\left(CV_l^S\right)^2\left(CV_{nl}^S\right)^2}{\sqrt{\left(m_l^S\right)^2\left(1-m_{nl}^S\right)^2\left(CV_l^S\right)^2+\left(m_l^S\right)^2\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2}\sqrt{\left[\left(m_l^S\right)^2\left(m_{nl}^S\right)^2\left(CV_l^S\right)^2+\left(1-m_l^S\right)^2\left(m_{nl}^S\right)^2\left(CV_{nl}^S\right)^2\right]^3}} > 0$$

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 $m_l^S, m_{nl}^S \ll 1$: $1 - m_l^S \approx 1, 1 - m_{nl}^S \approx 1$, $\left(m_l^S\right)^2 \approx 0$, $\left(m_{nl}^S\right)^2 \approx 0$. Thus, Eqs. (S11) for mean cfDNA concentrations are further simplified to

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

From Eqs. (S14) we also deduce that

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

*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:*

$$\rho \approx m_l^S + m_{nl}^S = m_{tot}^S. \quad \text{(S19)}$$

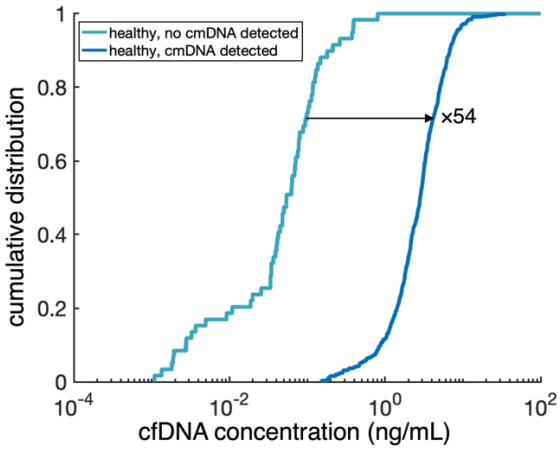


**Figure S1. Multiplicative factor effect in cfDNA concentration between healthy patients with or without cmDNA detected.** Empirical cumulative distribution functions for the cfDNA concentration in healthy individuals with no mutations detected in liquid biopsies (9.4% of healthy individuals) and healthy individuals with mutations detected in plasma (90.6% of healthy individuals). By performing the K-S test, we observe that cfDNA concentrations in healthy individuals with no mutations detected are consistent with the cfDNA concentrations in healthy individuals with mutations detected in plasma, multiplied by 1/54 (*p*-value: 0.19).

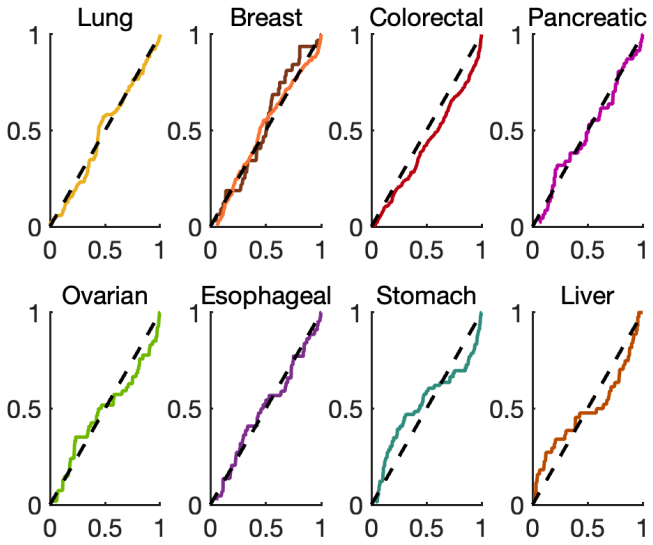


**Figure S2. P-P plots for the multiplicative factor effect in cfDNA concentration between healthy and cancer patients.** In each panel, we depict the probability pairs calculated from the empirical CDF for cfDNA concentration of cancer patients of a particular type, and the empirical CDF of cfDNA concentration in healthy patients multiplied by the respective factor from Table S2. In breast cancer panel, we depict both the P-P plot for stage I (brown curve) and stage II-III (orange curve).

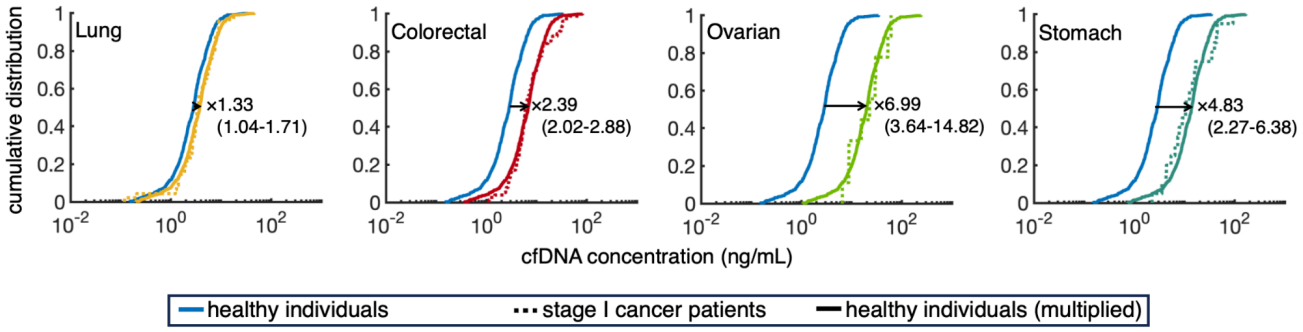


**Figure S3. Quantifying the multiplicative increase in cfDNA concentration in the presence of stage I cancer.** Cumulative distribution functions for the cfDNA concentration in healthy individuals with mutations detected in plasma (*N*=572) and patients with stage I lung (*N*=46), colorectal (*N*=76), ovarian (*N*=9) and stomach (*N*=20) cancer. Median empirical cfDNA concentrations (corresponding to 0.5 value of cumulative distribution) are 2.83 ng/mL for healthy individuals, and 3.45, 6.07, 22.83 and 10.60 ng/mL for patients with stage I lung, colorectal and stomach cancer, respectively.

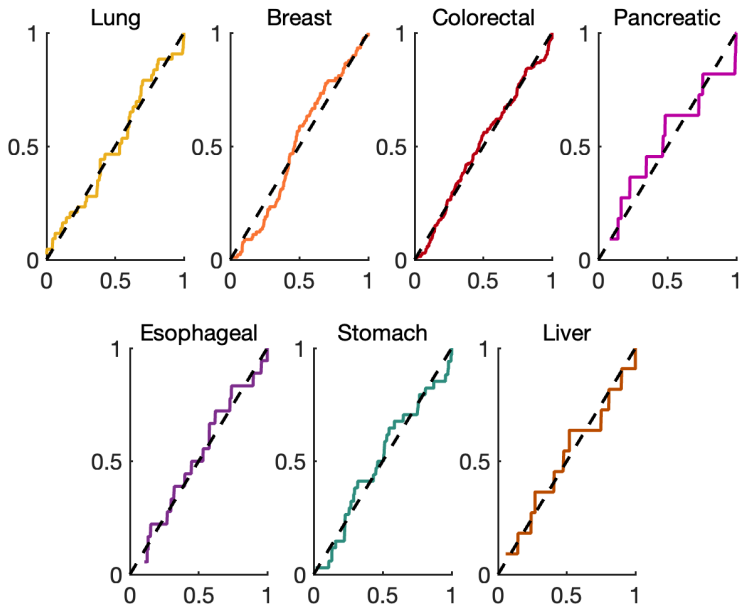


**Figure S4. P-P plots for the multiplicative factor effect in cmDNA concentration between healthy and cancer patients.** In each panel, we depict the probability pairs calculated from the empirical CDF for cmDNA concentration shed from non-tumor tissue of cancer patients of a particular type, and the empirical CDF of cmDNA concentration in healthy patients multiplied by the respective factor from Table S3 (values for all MAF). Ovarian cancer is not depicted due to insufficient data.

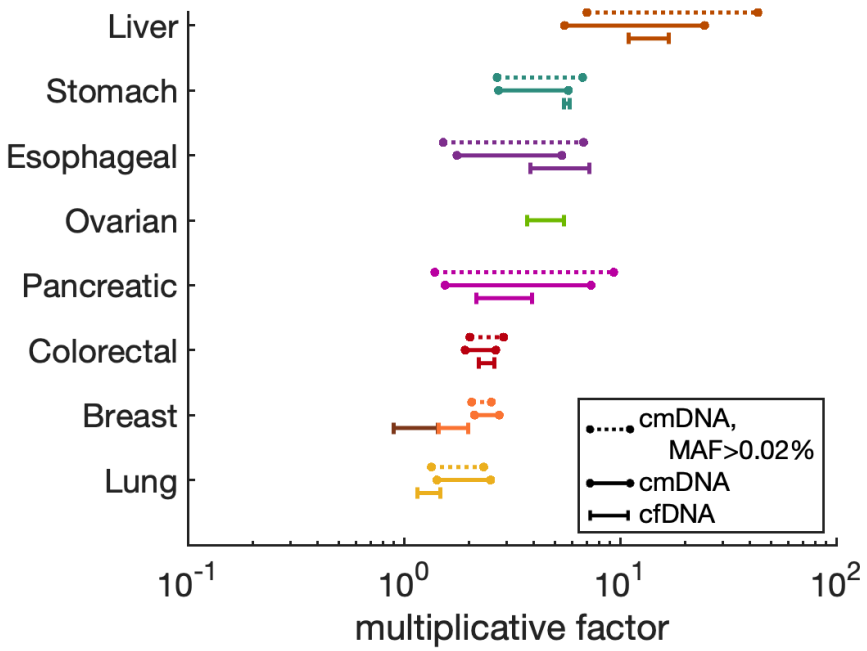


**Figure S5. Multiplicative factor ranges for each cancer type, quantifying the increase in concentrations of total cfDNA in plasma, and of cmDNA not shed from cancer tissue, in the presence of cancer.** For each cancer type, we depict the ranges of multiplicative factor not rejected by K-S test at 5% significance level for cfDNA and cmDNA concentrations. In the case of breast cancer, we depict the range of cfDNA multiplicative factor for both stage I (brown curve) and stage II-III (orange curve).

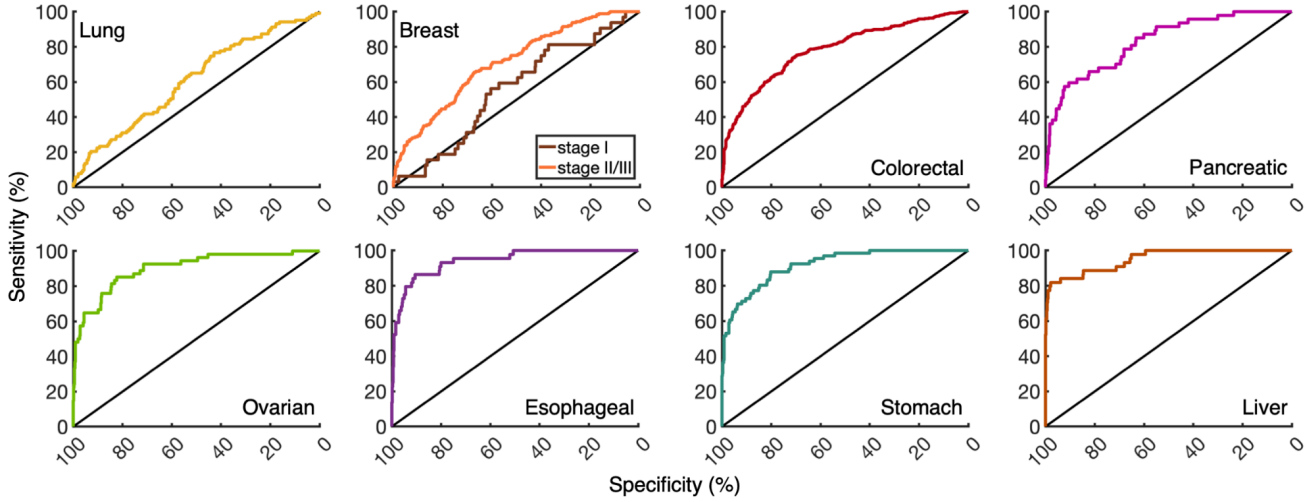


**Figure S6. Receiver Operating Characteristic (ROC) curves for each cancer type.** ROC curves for the cancer detection based on the individual's cfDNA concentration, for each cancer type. ROC curves are calculated from the logistic regression fit to cfDNA concentration data in healthy individuals and patients with stage I-III cancer. The number of healthy individuals and cancer patients is the same as in Fig. 2. The area under curve (AUC) for every ROC is 0.61, 0.54, 0.71, 0.78, 0.83, 0.90, 0.94, 0.92, 0.95 for lung, breast stage I, breast stage II/III, colorectal, pancreatic, ovarian, esophageal, stomach and liver cancer respectively.

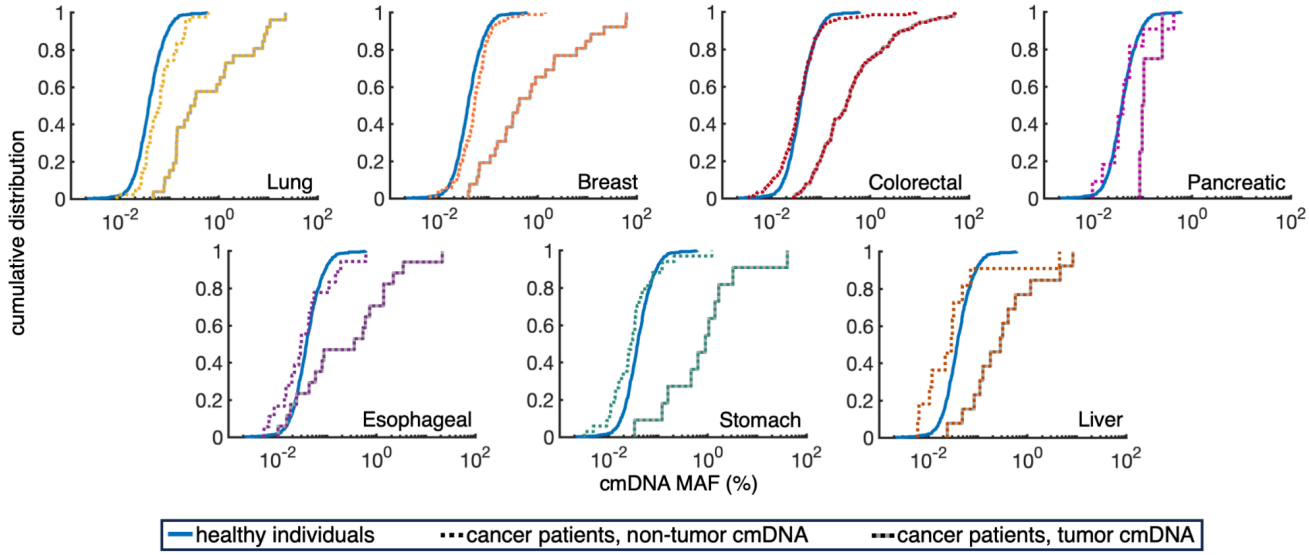


**Figure S7.** Cumulative distribution functions for the MAF percentages of mutations detected in plasma of healthy individuals (*N*=572) and patients with cmDNA not detected in tumor/detected in tumor with lung (*N*=43/26), breast (*N*=90/26), colorectal (*N*=205/129), pancreatic (*N*=11/4), esophageal (*N*=18/17), stomach (*N*=34/11) and liver (*N*=11/13) cancer. Median empirical MAF percentages (corresponding to 0.5 value of cumulative distribution) are 0.04% for healthy individuals, 0.06%, 0.05%, 0.04%, 0.04%, 0.03%, 0.03%, and 0.03% for cancer patients with non-tumor cmDNA, and 0.29%, 0.40%, 0.33%, 0.10%, 0.35%, 0.90% and 0.90% for patients with tumor cmDNA, with lung, breast, colorectal, pancreatic, esophageal, stomach and liver cancer, respectively.

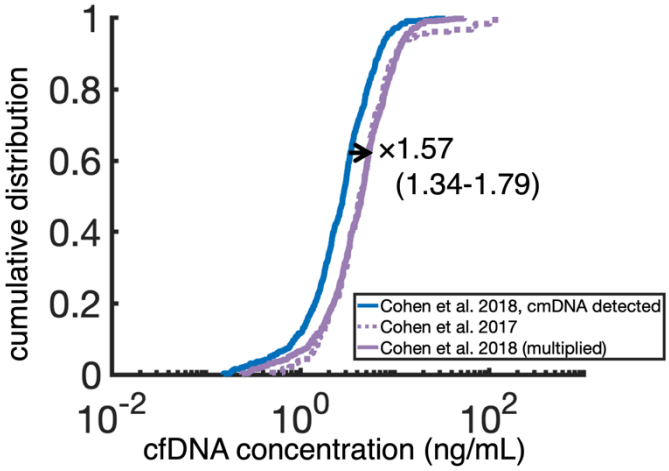


**Figure S8. Multiplicative factor effect in cfDNA concentration between healthy individuals reported between two studies by Cohen et al.** Empirical cumulative distribution functions for the cfDNA concentration in healthy individuals with mutations detected from the 2018 study by Cohen et al., and healthy individuals from the 2017 study by Cohen et al. By performing the K-S test, we observe that cfDNA concentrations in healthy individuals from the 2017 study are consistent with the cfDNA concentrations in healthy individuals from the 2018 study, multiplied by 1.57 (*p*-value: 0.95).

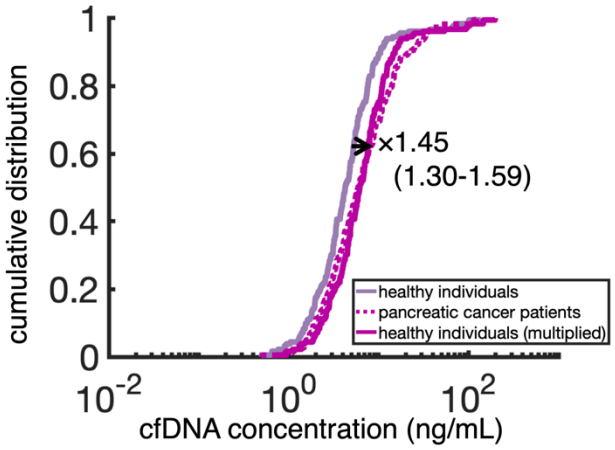


**Figure S9. Multiplicative factor effect in cfDNA concentration between healthy individuals and pancreatic cancer patients reported in Cohen et al. 2017.** Empirical cumulative distribution functions for the cfDNA concentration in 182 healthy individuals, and 221 pancreatic cancer patients from the 2017 study by Cohen et al. By performing the K-S test, we observe that cfDNA concentrations in healthy individuals are consistent with the cfDNA concentrations in pancreatic cancer patients, multiplied by 1.45 (*p*-value: 0.39).

| cfDNA dataset | data from [13] | exclude data included in [13] originating from [21] | exclude data with no cmDNA detected | data analyzed in Results | data from [21] 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 | - |

**Table S1.** Cell-free DNA data from Cohen *et al.* 2018 (13) and Cohen *et al.* 2017 (20) analyzed in the manuscript.

| 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 |

**Table S2. Multiplicative factors for cfDNA concentration, for each cancer type.** We report the *p*-value of the K-S test between cfDNA samples from patients of each cancer and the cfDNA samples from healthy individuals with cmDNA detected multiplied by the corresponding multiplicative factor. We also report the range of values for the multiplicative factor with K-S test *p*-values >0.05.

| 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 |

**Table S3. Multiplicative factors for cmDNA concentration not shed from tumor, for each cancer type.** We consider all samples, and samples with MAF >0.01%, 0.015%, 0.02%, and report the *p*-value of the K-S test between cmDNA samples from patients of each cancer and the cmDNA samples from healthy individuals multiplied by the corresponding multiplicative factor. We also report the range of values for the multiplicative factor with K-S test *p*-values >0.05.