Characterizing Noise in Whole Genome Cell-Free DNA Analysis

by

Shayan Chowdhury

Landau Lab, New York Genome Center/Weill Cornell Medical Center

Stuyvesant High School

Columbia ID: C004098257

DOB: 11/14/2002

Abstract

In the context of precision medicine cancer treatment, liquid biopsy is emerging as a revolutionary and critical new tool. In precision medicine, physician's decisions are guided by the genetic profile of a patient and their tumor. Liquid biopsy is the sequencing and analysis of cell-free circulating DNA (cfDNA). This DNA is not bound to a particular cell, rather is found in various bodily fluids including the bloodstream. In cancer patients, cfDNA partially consists of circulating tumor DNA (ctDNA), tumor DNA that has been shed into the blood. ctDNA in the blood is a promising biomarker in the blood for detection of cancer below the surface of clinical detection levels. This applies both for early detection of cancer and for "surveillance": tracking the response of patients to treatments and by extension, for identifying patients with high risk of treatment failure. However, current cfDNA protocols have a lower limit of detection far above the clinical need. These methods lack both the sensitivity and specificity to be applicable in the clinical setting yet due to various sources of error in DNA extraction, sequencing, and alysis protocols. These errors increase false positive detection, which in turn make the detection of trace amounts of tumor DNA within the vast set of cfDNA impossible.

We focused our analysis on various forms of errors, classifying the possible errors leading to noise into four broad categories. These are general and often overlapping, but help to describe the sources of noise in cell-free DNA analysis, which must be filtered out in order to improve detection methods in ctDNA. In order to determine if these noise detections were random or had some inherent bias, we set up a mathematical model and a synthetic experiment to confidently identify and predict which sites would be found outside of our random expectation. We compared this to our experimental data and found very prominent bias in noise detections for specific sites. Using this knowledge, we analyzed differences between "clean" and "noisy" sites, which are sites that rarely carry errors or often carry errors, respectively. We examined the features of noisy sites and propose future directions towards the the development of a neural network framework to train a model based on these features to predict a 'noise score' for a given site.

Intro & Background

Precision medicine is the approach of selectively tailoring treatments to a specific patient's disease. In the context of cancer, precision medicine is emerging as a critical tool for clinicians. The genetic profile of a patient and their tumor can be used to guide and inform treatment. (Ignatiadis et al., 2014)

The traditional method for evaluating tumor biology is tissue biopsy, which involves surgically removing a sample of tumorous cells and evaluating the sample's histological or genetic profile. However, biopsies are inherently invasive procedures that are expensive and often come with high risks to patient health. In some cases, a biopsy might not be an option for the patient due to either inaccessibility of the tumor or other health-related conditions. Similarly, repeated use of imaging techniques such as CT scans can pose a risk to patient health due to radiation exposure (Johnson et al., 2014). Additionally, although these techniques do give a holistic representation of the patient's disease progression, they do not provide clinicians with insight into the underlying biology of cancer. To gain insight to a patient's response to treatment and track the evolution of their tumor over time, imaging and tissue biopsy be impractical to carry out tissue biopsies on a recurrent basis.

A relatively recent development in precision medicine to track tumor evolution is liquid biopsy. Liquid biopsy is the sequencing and analysis of tumor information detectable in the blood and other body fluids through non-invasive means. Upon cell death, tumors shed this information in the form of cell-free circulating tumor DNA (ctDNA), which is not bound to a particular cell but can be found in various bodily fluids including the bloodstream (Ma et al., 2015). This process holds great promise for the detection and analysis ctDNA for precision medicine treatment of cancer.

The field of liquid biopsy is broad, with a wide array of biomarkers used to detect and measure tumor presence in a patient's bloodstream. In some rare cancers, specific protein biomarkers have been benchmarked, for example the PSA test used in prostate cancer (Saini 2016). However, in most cases, such proteins do not exist. Much ongoing research focuses on developing tests for DNA, RNA, or circulating tumor cells.

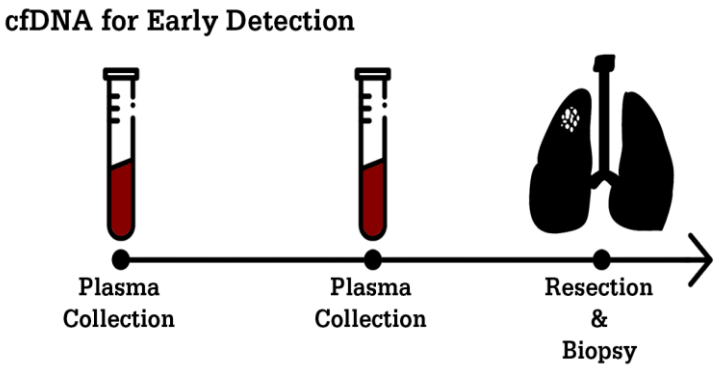


Figure 1 | Procedure for cfDNA for Early Detection
Early detection refers to finding cancer in a patient as soon as possible, in some cases even before it would be able to be detected on a CT scan. First, blood plasma is collected from patients in a non-invasive manner. If the plasma shows indications of ctDNA, then a thorough biopsy would be taken, to confirm the suspicions of the patient having a tumor. In the plasma collection process, the higher sensitivity has the potential to help create enormous breakthroughs in clinical outcomes by detecting cancer much earlier.

The clinical applications of these tests can be classified into two distinct classes. One of such applications involving ctDNA in the clinical setting involves detecting cancerous tumors in early-stage patients with improved risk assessment (Figure 1). In this setting, ctDNA tests can be used for at-risk patient populations to detect cancer earlier, and lead to better outcomes. There is a distinct need for screening in these populations; the United States Preventative Service Taskforce concluded from a study that all individuals from age 55-80 with an extensive smoking history should be screened by CT due to their high risk for cancer and the improved outcomes of early detection. However, the American Academy of Family Physicians warns that due to the unknown risk of exposure to radiation associated with CT they cannot conclusively support this recommendation (AAFP 2013). Therefore, there is a strong need for alternative screening methods without medical risks to the wellbeing of healthy individuals. For instance, almost half of all patients with stage I cancer, which is almost always curable through surgery alone, exhibited detectable levels of ctDNA and more than two-thirds of patients with stage III, which is also curable in many cases did the same as well (Bettegowda et al., 2014). These results prove to be already extremely promising in early detection and further advancements in detecting ctDNA could completely change the convention of cancer detection (Babayan et al., 2018).

ctDNA can also prove to be a promising biomarker in tracking the response of patients to treatments and by extension, identifying patients with a high risk of treatment failure. Disease recurrence is common, found in up to 40% of patients, with many being incurable, especially in tumors that progress earlier. ctDNA shows promise as a reliable biomarker to detect diseases below the surface of clinical detection levels and offers further long-term survival potential. Patients whose tumor cells at the point of dying release their DNA into the bloodstream as detectable ctDNA had a risk ratio of 228:1 compared to patients with undetectable ctDNA for disease progression (Roschewski et al., 2015).

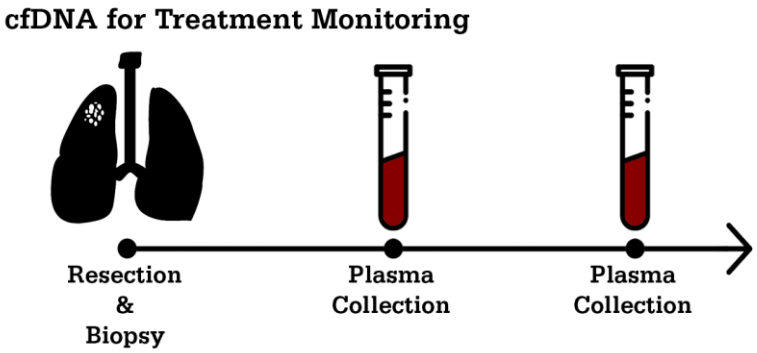
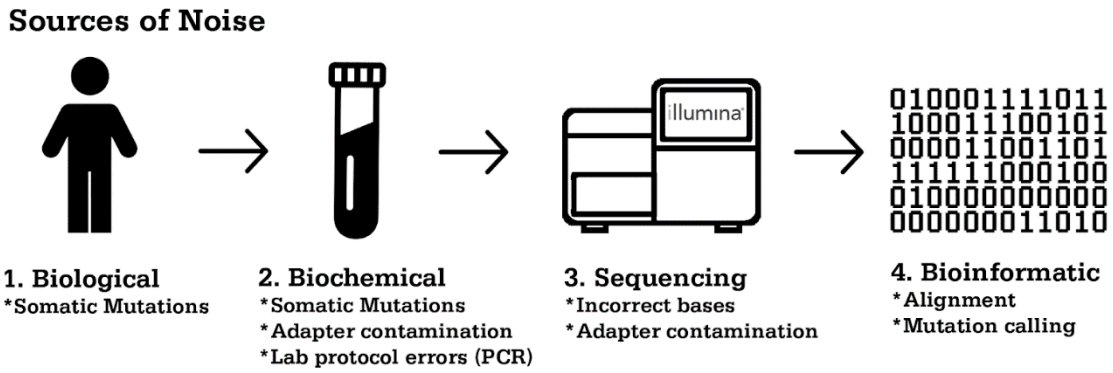


Figure 2 | Procedure for cfDNA for Treatment Monitoring or "Surveillance"
This schematic shows doctors using cfDNA to monitor a patient that is already being treated. A possible alternative to this would be taking many CT scans of patients. However, this is first of all unsafe due to the radiation it subjects onto people, resulting in the procedure needed a large grace period before repetitions. A non-invasive blood test as liquid biopsy gives clinicians treating patients a way to see if patients are responding to treatment or if disease is recurring. To find traces of patient A's tumor in patient A's blood, clinicians could search patient A's cfDNA for single reads of mutations that match mutations found in patient A's tumor.

However, detection of ctDNA is limited using existing technologies. Current state of the art methods do not yet have the sensitivity and specificity to be useful in the clinical setting. This is caused in part by errors in cfDNA protocols, which create a lower limit of detection that is far above the clinical need. These errors increase false positives, which in turn make the detection of trace amounts of tumor DNA impossible. We classify errors into four broad categories. These are general and often overlapping, but help to describe the sources of noise in cell-free DNA analysis.



First, there are errors in bioinformatic analysis of genetic data. These errors can occur in various parts of the pipeline, including during alignment and in mutation calling. Since the underlying nature of genetic information analysis involves large amounts of data, in many cases heuristic approaches that are needed for analysis necessarily lead to error (Saeys et al., 2007).

Second, are sequencing errors. When a fragment of DNA is being sequenced it can undergo errors such as reading incorrect bases or contamination from adapters attached to each fragment for sequencing. Duplex sequencing, UMIs, and base confidence scores are sometimes used to reduce this errors type of error. Existing deep learning methods have been used to learn patterns of error in sequencing, particularly for application in cell-free DNA (Kothen-Hill et al., 2018). Such methods identify specific mutations that are likely to be errors and remove them from the data.

Third are biochemical errors. PCR errors, sample contamination, or in-vitro DNA damage can cause random or systematic errors in sequencing data. Similar to sequencing errors, methods such as consensus UMI tests can be used to correct biochemical errors.

Lastly, are biological errors. A source of noise derived from pure biological error includes random mutations over age. As we age, our normal tissue progressively undergo. One example was a recent study looking at TP53 mutations (Salk et al., 2018). TP53 is a classical tumor mutation, studied broadly across many cancer types. However, low-frequency TP53 mutations were found to exist in healthy tissues from individuals of all ages, from infants to elders and found commonly throughout healthy tissues. The findings show that these mutations are also found to progressively increase in abundance with age. Due to its nonrandom and positively selected trends in mutations over time, TP53 mutations are found to be similar to that of cancer mutations. Similar work has been done in other diseases, such as the finding of Clonal

Hematopoiesis of Indeterminate Potential (CHIP), which showed massive clonal selection in B cells of patients that do not have the disease (Steensma 2015). Therefore, even without bioinformatic or sequencing errors, true biological mutations in healthy individuals create a further layer of noise that presents a challenge to sensitive detection in circulating tumor DNA within cell-free DNA.

We set out to define if patterns of errors in cell-free cirulating DNA were random or systematic, and characterize the noise found in cfDNA sequencing.

Methods

Synthetic Methodology

First, we wanted to identify if the noise detected in whole genome sequencing of cell-free DNA was random or systematic. We compiled a list of sites using tissue biopsies from 12 lung cancer patients. These specific mutations constituted sites that could be searched for in cfDNA, particularly in the disease monitoring context of the patient from which the tumor was resected. Therefore, all subsequent analysis was done using these sites. In total, there were 585,517 mutations called.

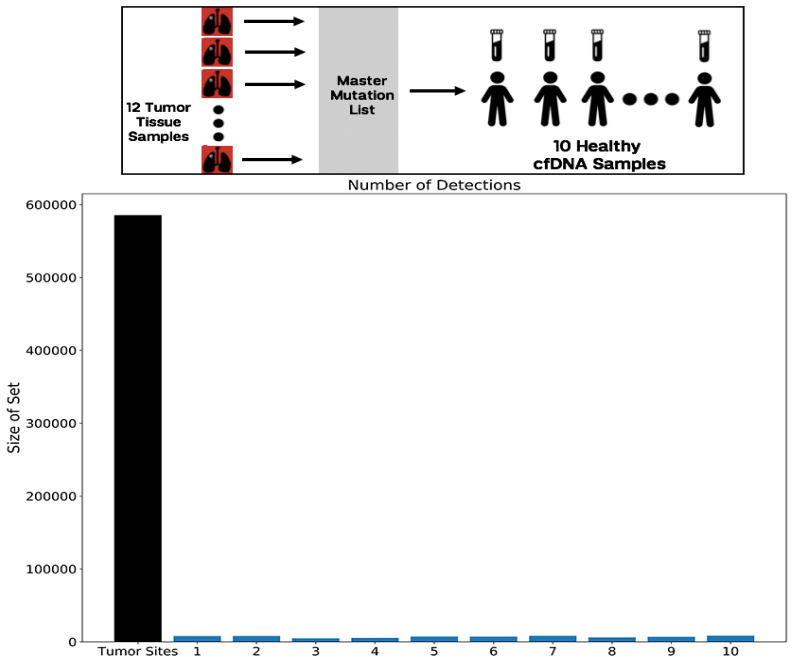


Figure 4 | Collection of tissue samples from patients and the number of detections in the master tumor sites file in relation to the healthy plasmas/ error sets

Tissue samples were collected from 12 patients with lung cancer via a biopsy and from 10 healthy patients via liquid biopsy. Each Healthy Plasma/Error Set is an experimentally detected subset of the Tumor Sites shown on the leftmost column. This data consists of sites that were detected in the blood of healthy individuals which represents sites with errors in sequencing or other analyses. We assessed the likelihood of randomly selected sites in tumor sites appearing in multiple noise sets by comparing relative sizes of each of the sets to the much larger tumor sites set. To put into perspective, the tumor sites set was approximately 100 times larger than each of the error sets. Thus, we hypothesized that the assessed probability of a site appearing in multiple error sets would be very low.

Due to the extremely low variant allele fraction of tumor DNA in total cfDNA found in circulation, mutation calling in cell-free DNA is done with just a *single* mutated read, as compared to tumor tissue samples where mutations are called by finding sites at which the tumor sample has *many* mutated reads. We checked the mutations found in the 12 lung cancer patients against whole genome sequencing data of 10 healthy patients. A mutation was called at any site where the normal cfDNA contained the mutations using the same a single read. All of the mutations detected could be defined as ‘noise’,

since it is known (with near certainty) that they did not originate from a tumor. In each of the healthy patients, we detected between 4805 and 8599 of the normal sites.

Each set of sites detected in a healthy patient's cfDNA was a subset of the full 585,517 total tumor derived mutations. In order to determine if these noise detections were random or had some inherent bias, we set up a mathematical model to predict the number of normal subsets a site would be detected in. This mathematical model is essentially a sampling problem, with the total tumor sites being a large pool from which we are drawing subsets.

Essentially, we are trying to detect the tumor sites from the biopsy within the midst of all the cfDNA found in the liquid biopsy, which we would record as ctDNA. However, the ratio of ctDNA to cfDNA is very low, leading to many healthy sites being recorded, even in some cases as tumor sites. These sites are referred to as 'noisy' because they are detected as false positive data. In other words, these sites are detected as cancerous despite no cancer actually developing and in some cases, some have the characteristics of healthy mutations found in a subset of healthy patient sites.

Mathematical Model

Before conducting analysis using patient data, we made a mathematical model that implements a nested binomial distribution to find the distribution of probabilities for finding a number of given sites within a number of control sets. In general terms, a binomial distribution takes in two parameters and returns the discrete probability distribution of the number of 'successes' in a sequence of trials and a given constant probability of 'success' for one trial. A trivial example of this is a fair coin toss, where the probabilities of getting either heads or tails are the same throughout. To elaborate, we define an arbitrary variable, X as the number of heads (in this case 3) we want to find the probability for from flipping a fair coin 5 times. The probability of getting three heads out of five tosses is the combination of all possible ways to get the three heads (TTHHH, HTHTH, HHHTT, etc.) or C(5, 3), multiplied by the probability of getting heads 5 consecutive times, which is the inverse of 2 to the power of 5, or 1/32.

$$\text{X} = \text{number of heads from flipping a } \textbf{fair} \text{ coin 5 times}$$

$$P(X = 3) = C(5, 3)\frac{1}{2^5} = \binom{5}{3}\frac{1}{32} = \frac{10}{32}$$

Based on this definition and understanding of binomial distribution, we can see this in the context of our experiment:

Probability distribution of m sites intersecting k controls

Given

t = size of tumor sites set (585517 in this case)

n = number of control sets (10 in this case)

k $\in [0, 10]$ (number of control sets that each site intersects with)

m : number of sites intersected through k controls

p : probability of a site being in each normal (inverse of normal size)

$$P(m \text{ sites}, k \text{ controls}) = \binom{t}{m} q^m (1-q)^{t-m}$$

q: probability distribution for each site in k of the controls

$$q(1 \text{ site}, k \text{ controls}]) = \binom{n}{k} p^k (1-p)^{n-k}$$

Here, P is the probability distribution of finding *m* sites within *k* control sets. One thing to immediately note is that P is dependent on q, which is the probability distribution of finding one given site in *k* control sets. Since P is looking at an *m* number of sites, it needs to account for the probability distributions for each site in the set of *m*, which is why q is needed as a nested binomial distribution in this case. However, one major oversight of this model is that in reality, not all the sizes of the normal sets are the size, which by extension means that the probability of a site being in one normal (or the probability of 'success') is NOT the same across trials or normal sets. To account for a set of normal set sizes, our model would be much more complicated and impractical to write out a more complicated formula involving all combinatorial possibilities for each probability (*p* value).

Thus, we decided to form our very own synthetic experiment to form a more accurate statistical model for the distribution of sites across the normal sets, this time accounting for the relative sizes of each set. Instead of one nested binomial draw, as shown in the model above, in our synthetic model we do a *k* number of separate Bernoulli trials, a trial that only returns binary output in the form of 'success' or 'failure', or in our case, whether a site is in a normal set or not.

Experimental v Synthetic

Using the sizes of the healthy plasma sets, we created a synthetic protocol in which 12 random subsets of synthetic tumor sites as "error sets." For confidence, this was repeated for 10 iterations.

With the synthetic error sets, we ran analyses to find the frequencies of each tumor site in each of the sets. We evaluated each site in the master tumor sites set and checked how many

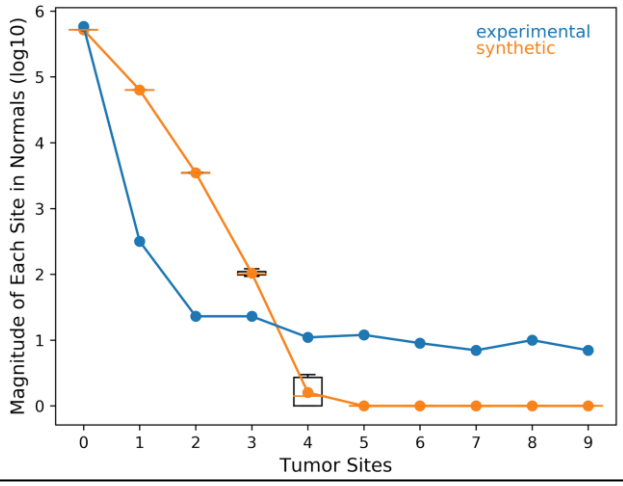


Figure 5 | Procedure for cfDNA for Treatment Monitoring or "Surveillance"
This figure was generated from two sets of data as signified in the legend at the top right. The blue "experimental" data is data from healthy control patients and the number of detections through each set and those that follow it. On the other hand, the orange plot highlights the synthetic error data generated from the mathematical model of the synthetic protocol. The *y*-axis signifies the magnitude of tumor sites which exist in *x* number of normal sets, which is interchangeable with "error set."

error sets the site appeared in. Through this, we calculated the frequencies of each tumor site in each subsequent error set. Afterward, these frequencies were intersected into one comprehensive noise set that detailed the frequencies of tumor sites across multiple sets. We noted that after increasing our number of sets to detect in, the frequencies of tumor sites across those decreased exponentially. In other words, the likelihood of tumor sites appearing in multiple sets diminished drastically. In the figure above, it's quite notable that in our synthetic model, there were no tumor sites that appeared in 5 error sets, yet the experimental model states a different story. It illustrates that there were sites at least to the order of magnitude higher than 1 that appeared in 5 healthy patients. As evident in the figure, there are sites that appear in all 12 healthy patients.

This analysis is done in order to identify and remove sites that may act as "noise" or insignificant data. This is done to improve the validity and credibility for ultimately selecting sites that may be in fact linked directly to cancer and not simply a false positive.

The observed frequency of sites overlapping throughout random error sets decreases and likelihood of sites appearing in all 10 of the random error sets would be virtually impossible (as illustrated by the plot of synthetic frequencies). Although the experimental data plot follows its synthetic counterpart initially, it quickly diverges with enormous differences being observed in tumor site frequencies in 5 to 9 error sets.

We note that the experimental findings show more sites appearing in multiple error sets. This disparity between the synthetic and experimental findings is due to certain biases in normal sets which makes specific sites in the genome more error prone than others.

Comparing "Clean" and "Noisy" Sites

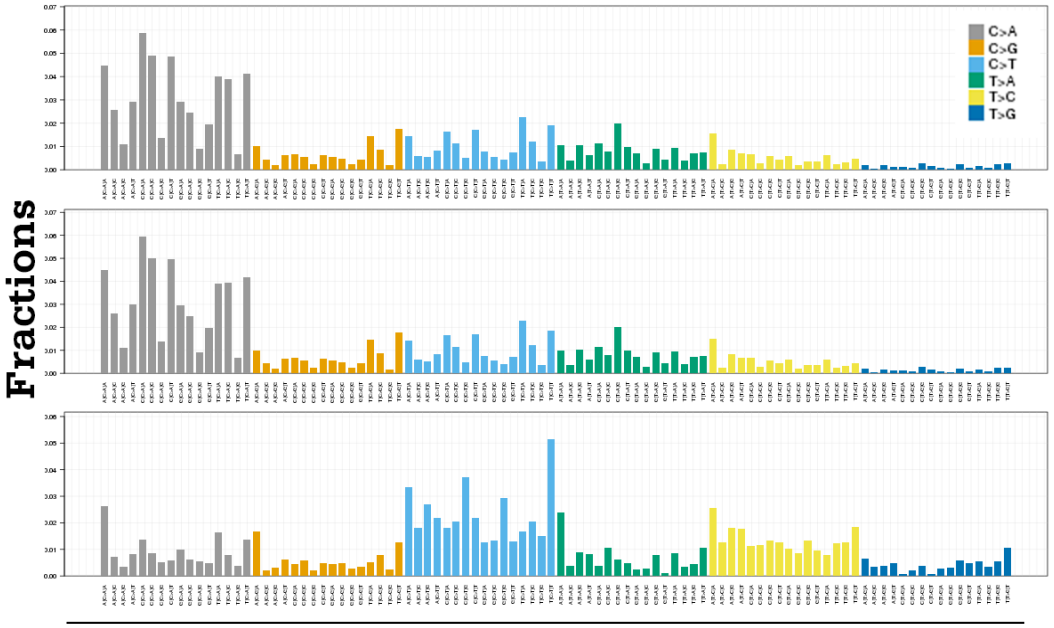


Figure 6 | Trinucleotide Context of Mutations by Base Change
The top panel represents the trinucleotide context of the full list of mutations found in 12 tumors (N=585517). The middle panel shows the trinucleotide context of sites that found in 0 or 1 normals (N=574393), which reflects the overall pattern of the sample. The bottom panel represents sites found in 8, 9, or 10 normal cfDNA samples (N=2549). This trinucleotide context is distinct from that found in the overall sample, with a much stronger bias towards C>T mutations.

We then set out to analyze the fundamental differences between sites detected in many normal samples as compared to sites that were detected in zero or very few normal samples. First, we examined the trinucleotide profiles of detected sites. The trinucleotide context, sometimes called the mutational signature, has been shown to underlie various biochemical or biological processes, including the tobacco signature that underlies lung cancer (Alexandrov 2012). We found a distinct mutational signature in sites that were detected across many normals, finding a pervasive C>T pattern. Subsequent work would be needed to determine if this is associated with biological, biochemical, sequencing or bioinformatics errors in our methodology.

Then, we wanted to determine potential regional patterns of errors across the genome. We checked each mutation site from the master tumor list against references pulled from the UCSC genome browser. We checked CGI (C-G Islands), LTR (Long Terminal Repeats), Exons, Intergenic Regions, B and T cell enhancers, and TSS (transcription start sites). The extensive data for this can be found in the supplemental. We found that each of these regions had a different noise profile. This suggests that biological sources of noise are potentially a source of error. Of note, B and T cell enhancers are biologically specific to blood cells. It is known that non-tumor cfDNA is predominantly made from blood cell DNA. Therefore, it is necessary to further investigate the potential impact of C.H.I.P. on noise rates in ctDNA detection.

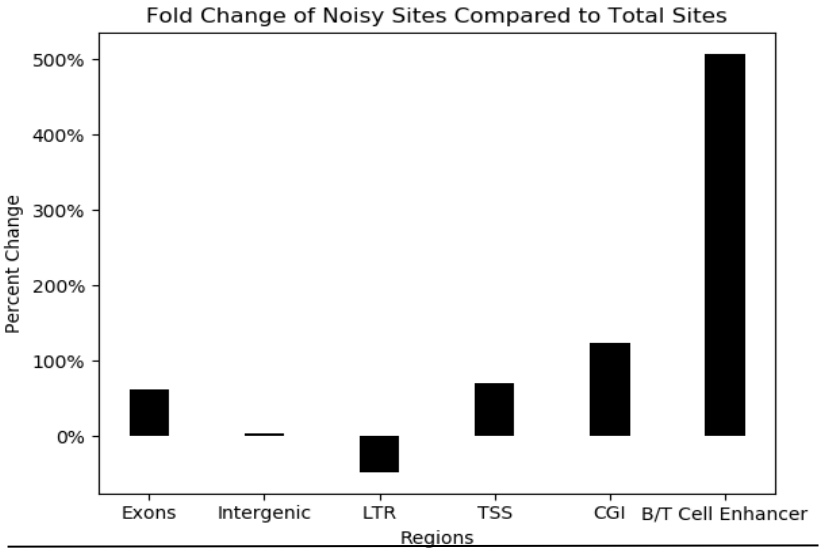
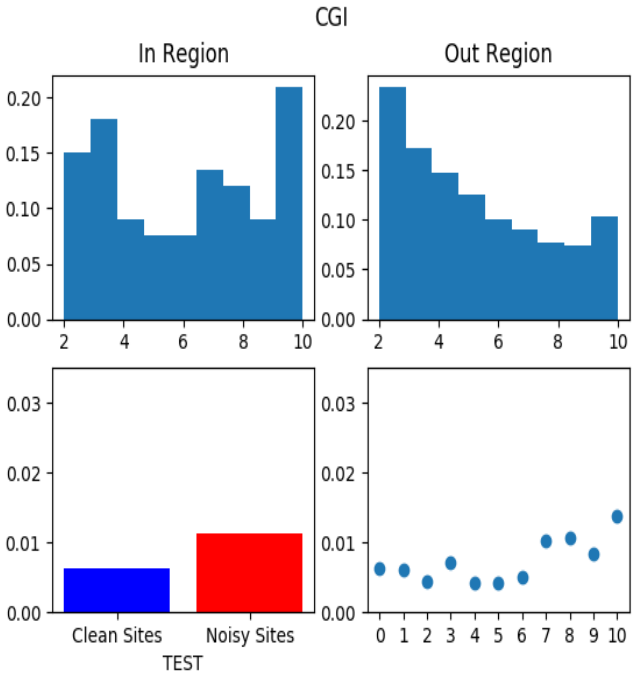


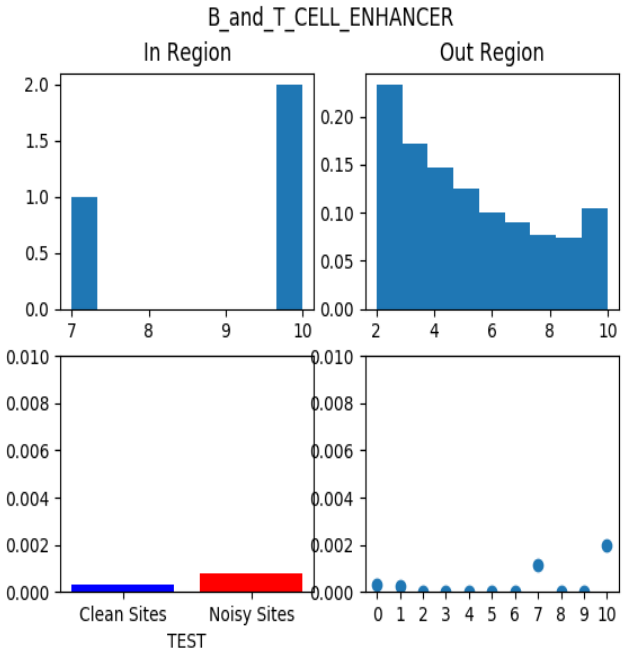
Figure 7 | Fold change of noisy compared to total detections We compared the percent composition of several genomic features in "noisy" sites as compared to their percent composition in all sites. Notably, promotors for B and T cell enhancers were far more represented in noisy sites. Biological phenomena, such as Clonal Hematopoiesis, may be a unique source of noise that would limit sensitivity and specificity unless properly filtered.

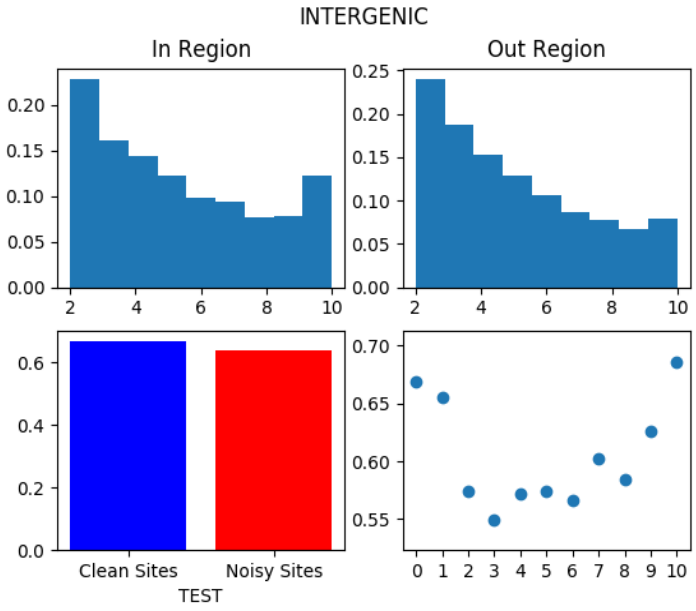
Discussion

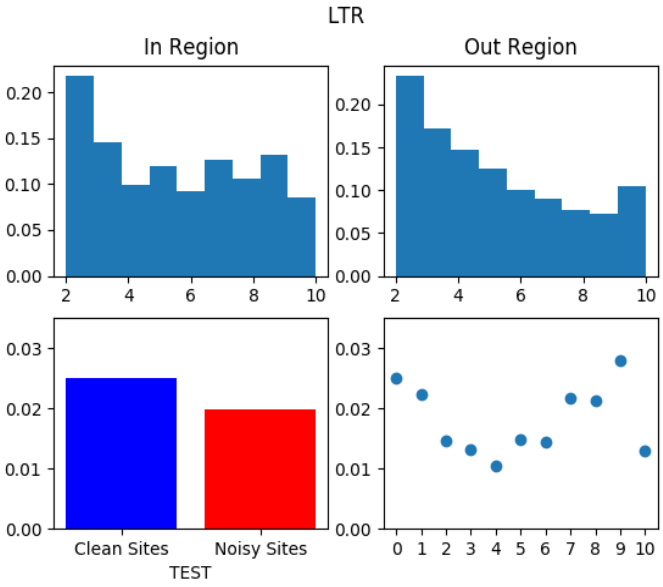
Sensitive mutation calling fundamentally requires a noise model in order to remove noise and decrease lower limits of detection (Gerstung et al., 2014). Here, we have begun to characterize noise patterns found in whole genome sequencing of cell-free DNA. We classify specific mutations in fragments detected through liquid biopsy as either noisy or clean and use mathematical modeling and simulation to confidently identify which sites are outside our random expectation. Although such a process seems to be viable and with relative accuracy, conducting analyses on often such large volumes of data that can be derived from liquid biopsy methods can be incredibly computationally expensive as well time consuming. Not to mention, the accuracy of some of the features explored here may be susceptible to errors themselves, especially since not all the manually selected features may have may be representative of the entire population of error-prone sites. As such, utilizing this method of manual algorithmic and computational analyses of such data may lead to false positives and other errors in detection. A second possible approach would be using deep learning methods to classify specific mutations, based on their loci or prevalence in other sites as either being significant or erroneous. A similar approach has been used to refind lists of mutations found in primary tumors (Ainscough 2018). Similar methods have been trained to refine lists of mutations. Essentially, a neural network would be trained with data derived from other biopsy methods to form a comprehensive understanding of site characteristics in hopes of ultimately being able to derive a function capable of returning results acceptable to a degree when given specific inputs. The method of learning for the network would probably be supervised due to the fact that our expected results would be within a discrete set of answers: error or possibly tumorous. Alternatively, we could measure the degree to which a site is actually cancerous (using a float point value) in which case a linear regression model would be needed. There are varying ways that a neural network can be trained to detect errors which are mostly determined by the outputs and how the learning method needs to be tweaked to prevent bias in the form of overfitting or underfitting.

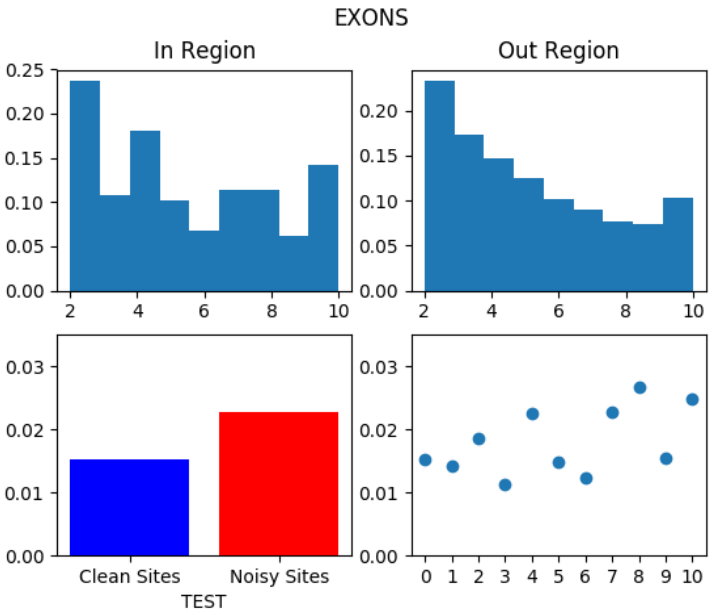
SUPPLEMENTARY FIGURES










EXONS
In Region
Out Region
0.25
0.20
0.15
0.10
0.05
0.00
2
4
6
8
10
0.03
0.02
0.01
0.00
Clean Sites
Noisy Sites
TEST
0 1 2 3 4 5 6 7 8 9 10

REFERENCES:

M. Ignatiadis, S.-J. Dawson; Circulating tumor cells and circulating tumor DNA for precision medicine: dream or reality?, *Annals of Oncology*, Volume 25, Issue 12, 1 December 2014, Pages 2304–2313, https://doi.org/10.1093/annonc/mdu480

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S. M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. A., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., Shih, l. M., Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T. L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N., … Diaz, L. A. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. Science translational medicine, 6(224), 224ra24.

Ma M, Zhu H, Zhang C, Sun X, Gao X, Chen G. "Liquid biopsy"-ctDNA detection with great potential and challenges. Ann Transl Med. 2015;3(16):235.

Roschewski M, Dunleavy K, Pittaluga S, et al. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B-cell lymphoma: a correlative biomarker study. Lancet Oncol. 2015;16(5):541-9.

Yvan Saeys, Iñaki Inza, Pedro Larrañaga; A review of feature selection techniques in bioinformatics, *Bioinformatics*, Volume 23, Issue 19, 1 October 2007, Pages 2507–2517, https://doi.org/10.1093/bioinformatics/btm344

Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics. 2014;30(9):1198-204.

Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Horvat, R., . . . Risques, R. A. (2018). Ultra-sensitive sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan. doi:10.1101/457291

Babayan, A., & Pantel, K. (2018). Advances in liquid biopsy approaches for early detection and monitoring of cancer. *Genome Medicine,10*(1). doi:10.1186/s13073-018-0533-6

Kothen-Hill ST, Zviran A, Schulman R, Maloney D, Huang KY, Omans N, Liao W, Robine N, &Landau DA. *Deep learning mutation prediction enables early stage lung cancer detection in liquie biopsy.* Sixth International Conference on Learning Representations, Workshop Track, 2018.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. **Signatures of mutational processes in human cancer**. Nature, 500(7463):415–421, 2013.

(2016). PSA and beyond: alternative prostate cancer biomarkers. Cellular oncology (Dordrecht), 39(2), 97-106.

Lung Cancer - Clinical Preventive Service Recommendations. (2016, January 28). Retrieved from https://www.aafp.org/patient-care/clinical-recommendations/all/lung-cancer.html

Final Update Summary: Lung Cancer: Screening. U.S. Preventive Services Task Force. July 2015. https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/lung-cancer-screening

Johnson, J. N., Hornik, C. P., Li, J. S., Benjamin, D. K., Yoshizumi, T. T., Reiman, R. E., . . . Hill, K. D. (2014). Cumulative Radiation Exposure and Cancer Risk Estimation in Children With Heart Disease. *Circulation, 130*(2), 161-167. doi:10.1161/circulationaha.113.005425

Kamal, Y., Cheng, C., Frost, H. R., & Amos, C. I. (2018). Predictors of disease aggressiveness influence outcome from immunotherapy treatment in renal clear cell carcinoma. *OncoImmunology*, 1-12. doi:10.1080/2162402x.2018.1500106

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., . . . Griffith, O. L. (2018, November 05). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. Retrieved from https://www.nature.com/articles/s41588-018-0257-y