University of Waterloo
Faculty of Mathematics

# Optimization of Circulating Tumor Fraction Prediction for Multi-Cancer Early Detection using Next-Generation Sequencing Data

Geneseeq Technology Inc.

Simon Hampton
1B Honours Mathematics
July 15th, 2024

# Executive Summary

This report, titled "Optimization of multi-cancer early detection pipeline to predict circulating tumor fraction from next-generation sequencing data," provides a pipeline that estimates circulating tumor allele fraction using next-generation sequencing data. The primary objective of this study is to provide a reproducible pipeline of a process outlined in a study done by Jamshidi and colleagues (2022).

The report is structured to offer a detailed mathematical summary of circulating tumor allele frequency (cTAF) prediction from a blood sample, evaluate its accuracy on different cTAF concentrations, and provide the code for future recreation of the results. The methodology is tested two different ways. One of which is constructing samples by subsampling normal control and tumor tissue, combining the results into a singular binary alignment map (BAM) file where the cTAF is known. The other is comparing the cTAF calculation to the previously used method; average variant allele frequency. Using both of these testing methods ensures a thorough and accurate prediction.

Key findings from the report include that a Bayesian model using a Poisson distribution may be better for estimating circulating tumor fraction in cell-free DNA (cfDNA) than the methods used previously. This can prove useful for evaluating machine learning models that perform multi-cancer early detection (MCED).

The report also offers recommendations for how to use or improve the findings. These recommendations are based largely around fine-tuning the procedure to improve accuracy.
In conclusion, this report provides valuable findings and practical recommendations that can be utilized by future oncologists attempting to improve MCED models.

# Table of Contents

# List of Tables and Figures

# 1. Introduction

Multi-cancer early detection (MCED) is a range of diagnostic tests whose objective is to use one minimally invasive test to potentially identify multiple types of cancer. A paper written by Hubbell and colleagues claims that adding an MCED test to standard cancer care "could reduce 5-year cancer mortality by 39% in those intercepted" (Hubbell et al., 2021). There have been recent advancements in cell-free DNA (cfDNA)-based MCED tests, also known as liquid biopsies. CfDNA is fragmented DNA that cells release to the blood, most commonly through apoptosis, cell death. Cancerous cells also release DNA into the blood flow, and this cancerous cell-free DNA is referred to as circulating tumor DNA (ctDNA).

Several tests are being conducted to create machine learning models capable of identifying cancer types from sequencing data, but none are currently being used clinically. Cancer signal as a unit must be determined to compare such models to each other. Currently, it is possible to pathologically determine the circulating tumor DNA fraction in the blood, but it is extremely resource-intensive and not realistic to repeat extensively for many models. This motivates the approximation of circulating tumor DNA fraction from a blood sample, which is much more easily accessible, though less accurate. Currently, average variant allele frequency (VAF) is being used to determine the ctDNA content in blood, but this is inaccurate due to outliers having a large impact on results. The goal of this report is to provide a more accurate approximation of ctDNA content in blood. This can assist future bioinformaticians in determining the limit of detection of MCED machine learning models.

Liquid biopsy MCED as a methodology has the potential to be groundbreaking because of the additional data encoded in cfDNA reads. Compared to alternative DNA sequencing methods, next-generation sequencing (NGS) can pick up lower frequency mutations which are usually present in the case of a tumor, making it possible to identify mutated DNA fragments (Heitzer et al., 2018). Liquid biopsy MCED tests are still under development and have not yet reached widespread use but have potential to have clinical success in the future for two main reasons. Firstly, it is generally more feasible than traditional methods such as X-rays or MRIs. Blood-based tests are less costly, more widely available, and is a procedure that almost any trained healthcare professional can perform. Additionally, sequencing cfDNA in a blood sample can be more accurate at detecting tumors in the early stages. According to Oxford Academic, "liquid biopsies may be able to uncover cancer lymph node invasion and distant metastasis prior to detection with imaging or biopsy" (Foser et al., 2024). The future of oncology seems to be trending towards using each of these methods in conjunction but for early detection, liquid biopsies surely have the edge.

## 2. Literature Review and Background Information

A research paper published by Jamshidi et al. (2022) outlines the importance of evaluating the clinical limit of detection of MCED tests. Limit of detection (LOD) refers to the smallest amount of a substance that can be distinguished with confidence from noise.

Two terms are defined that are the focus of the report. Circulating tumor frequency (cTF) is a measure of the percentage of ctDNA that is tumor-derived and circulating tumor allele frequency (cTAF) is the percentage of ctDNA that contains specific tumor mutations. They are distinct metrics as not every sequencing read in a tumor contains a mutation, so one can approximate cTAF by multiplying cTF by the median variant allele frequency (VAF) found in a tumor tissue sample. Previously, to determine LOD of MCED tests, the mean or median VAF was used, but these metrics each have their own errors, making their evaluation of MCED tests inaccurate. If mean VAF is used, it becomes largely affected by outliers, and if median VAF is used, the result is unreliable when the number of mutations in a sample is low, an occurrence that is often the case. Both metrics are also impacted by factors such as heterogeneity, copy number and imbalance sequencing, which is undesirable.

It is extremely important to approximate cfDNA tumor fraction because it provides a metric over which different classification machine learning models for MCED can be compared to each other. If there is an accurate prediction of cTAF, different models can have their predictions run on data with smaller and smaller cTAF concentrations to find the LOD of each model. The goal of this report is to produce a pipeline that can optimize their approximation model to calculate cTF and cTAF. If the findings are that cTAF is an effective metric for LOD, this can be beneficial to the evaluation of MCED machine learning models.

## 3. Materials and Methods

The method requires variant calling format (VCF) files from both a blood and tumor tissue sample from the same patient. Specifically, it requires the number of reference alleles and alternate alleles for each section of the sequenced DNA. The depth at each section is obtained by adding together the number of reference alleles and alternate alleles, and the variant allele fraction (VAF) is obtained by dividing alternate alleles by the depth. These variables are all used in the calculation of cTF and cTAF, outlined under "Calculation of cTF and cTAF" in the paper written by Jamshidi et. al. (2022). Circulating tumor fraction and circulating tumor allele fraction are calculated as follows:

**Poisson distribution:** P(X = k) is a theoretical probability from the Poisson distribution which predicts the likelihood of obtaining k mutants if the true parameters are $\lambda$. In this case, it represents the odds of obtaining the cTF $k$ if $\lambda$ is the ground truth cTF.

$$f(k;\lambda) = \mathrm{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$\lambda = Depth \times cTF \times VAF_{tumour}$$

**Likelihood**: probability of observing the specific alternate allele counts (AD) given a cancer tumor fraction (cTF).

$$P(counts|cTF) = \prod \mathrm{Poisson}(\mathrm{count}_i; \lambda_i)$$

Where $\lambda_i = \mathrm{depth}_i \times cTF \times \mathrm{VAF}_i$ ($\mathrm{depth}_i$ is the sequencing depth at location i, and $\mathrm{VAF}_i$ is the allele frequency in the tumor at location i).

**Prior**: the uniform prior distribution is used as it reflects a lack of prior knowledge about cTF, treating all values from 0 to 1 as equally likely.

$$\text{Prior}(cTF) = \text{Uniform}(cTF)$$

**Posterior**: Combines the prior and the likelihood to update beliefs about the model parameters after considering the data. It is calculated as the product of the likelihood and the prior.

$$\text{Posterior}(cTF) \propto \text{Likelihood}(cTF) \times \text{Prior}(cTF)$$

**Normalization:**

$$\text{Normalization Constant} = \int_0^1 \text{Posterior}(cTF)d(cTF)$$

Where $d(cTF)$ represents integrating with respect to cTF

$$\text{Normalized Posterior}(cTF) = \frac{\text{Posterior}(cTF)}{\text{Normalization Constant}}$$

**Estimate cTF:**

$$\text{cTF}_{\text{estimate}} = F(cTF)^{-1}(0.5)$$

Where

$$F(cTF) = \int_0^{cTF} \text{Normalized Posterior}(t), dt$$

$$F^{-1}(p) = inf cTF : F(cTF) \geq p$$

**Calculate cTAF:**

$$\text{cTAF} = \text{cTF}_{\text{estimate}} \times \text{median}(\{\text{tumor allele frequencies}\})$$

A Python program following this outlined process was created and included for reference in Appendix A. Since some of the numbers become too small for the Python processor to deal with, a logarithmic scale is used. The code reads in a TSV file with mutation calling data, which

includes depth of sequencing, reference allele counts, and alternate allele counts for both cancerous tissue samples and normal control plasma samples. Since the purpose of the code is to predict the value of an unknown variable, testing it is not as straightforward. To check the accuracy of the predictions, two methods are used and are outlined in the following paragraphs and used in the program.

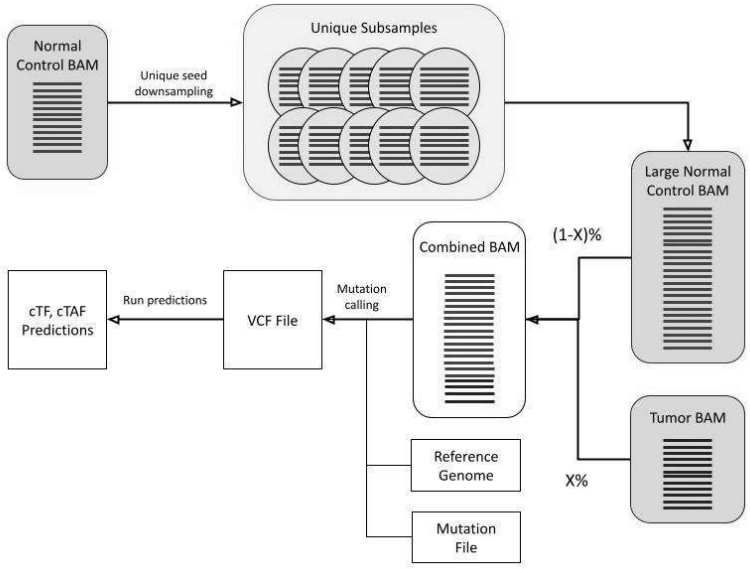
**Method 1: Simulated Plasma BAM File from Tissue BAM and Normal BAM**

A blood sample was simulated using a healthy tissue sample and a tumor tissue sample so that the ground truth cTAF is known in advance and can be compared to the prediction output of the program. DNA from the blood is virtually indistinguishable from tissue DNA, so this is both appropriate and feasible (Li et al., 2023). Data was taken from real healthy tissue samples making up the normal control BAM file with around five million reads after de-duplication and recalibration. To have the most precise measurements, more data than this is needed in order to increase the mean depth of the reads. To do this, the data was subsampled in multiple unique groups, each containing 50% of the reads in the original file. Then, to up sample, the files were combined into one BAM file, referred to as “large normal control”. Finally, the tumor BAM files were down sampled to obtain the desired tumor purity when merged with the large normal control file.

Tissue data from tumors is also taken, with the tumor BAM file containing around fourteen million reads and 60 unique cancer mutations. This data is down sampled to match the required cTAF concentration when added to the large normal control. The cTAF concentrations tested are 1%, 2%, 5%, 7%, 10%, 20%, 30%, 40% and 50%. Five trials are performed for each concentration with unique subsamples of both files each time. Mutation calling is then performed

on the combined BAM files to obtain the number of reference alleles, alternate alleles, and the depth at that location for each gene in the simulated sample.

**Figure 1: Flowchart for One Trial of Method 1**




Note: This process using a normal control BAM file and tumor BAM is repeated five times for each percentage for the nine percentages. Created using Google Drawings.

This information is summarized in Figure 1, with deliberate ratios between reads from large normal control and tumor put forth into the combined BAM according to the desired concentrations. The variable X represents the fraction of the combined BAM composed of reads from the tumor BAM. This process is repeated five times per percent purity for the nine purities mentioned above.

**Method 2: Mean VAF from Real Plasma Sample**

The second validation method involves a comparison to the previous standard method for determining cTAF. As mentioned previously, average VAF is the current standard for

determining the LOD of a MCED model. It is inaccurate due to outliers having a large impact on the results but should still approximately equal average VAF. The calculation for average VAF is:

$$A_{VAF} = \frac{\sum_{i=0}^{k-1} \frac{alt_i}{depth_i}}{k}$$

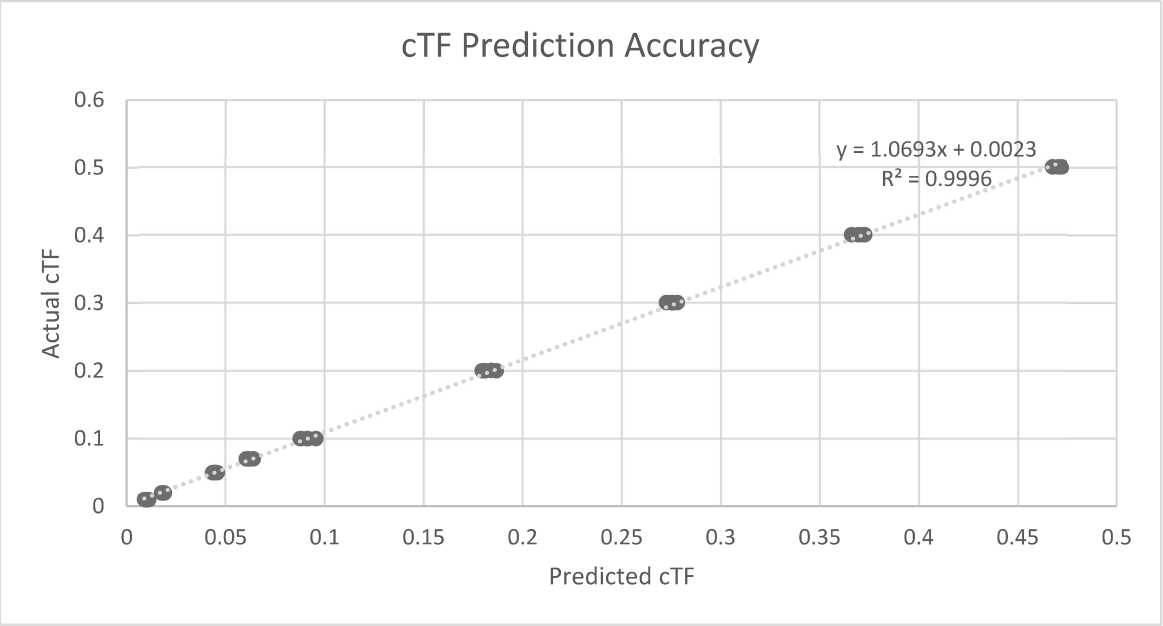
Where $alt_i$ represents the alternate allele count at position $i$, $depth_i$ represents the depth at position $i$, and k represents the number of mutations present in the sample. The formula is intuitive as it calculates the sum of VAFs for each mutation and divides by the total number of mutations.

This method uses a different sample, which comes from a real plasma sample, contrary to Method 1, which constructs a simulated plasma sample from tumor tissue and tissue normal control. The comparison to average VAF can also be used to determine when average VAF is more unreliable if this calculation for cTAF is later proven to be accurate and precise.

## 4. Results and Analysis

For Method 1, two scatter plots are produced: one comparing predicted cTF with actual cTF, and another to visualize how large the percentage error is for varying magnitudes of actual cTF.

**Figure 2: Method 1 Actual cTF Plotted Against Calculated cTF**



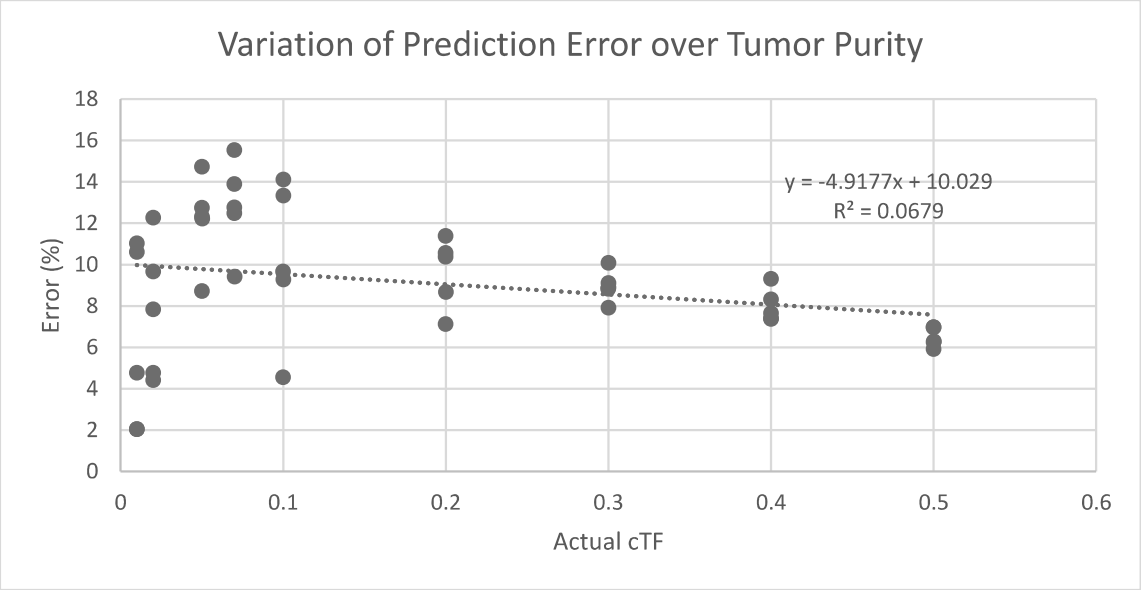

Note: Perfect predictions would result in a R-squared value of 1 and a slope of 1. The actual observations are extremely close to that target.

(S. Hampton, personal communication, June 10, 2024)

A very strong positive correlation can is observed between predicted and actual cTF, representing very accurate predictions. The R-squared value of 0.9996 is also an indication that the calculated cTF predictions are linear. It can also be observed that for higher tumor purities, the predictions are slightly low, indicating that the code is less reliable with higher concentrations. If more trials are done and similar trends noticed, it is possible that a bias can be introduced to make the predictions more accurate for future variations of the code. However, it is

less common to have these large cTF values in real samples, meaning that this error is impactful, but only occasionally.

**Figure 3: Method 1 Error Variation**




Note: Error percentage is calculated by dividing the absolute difference between Actual and Predicted cTF by Actual cTF.

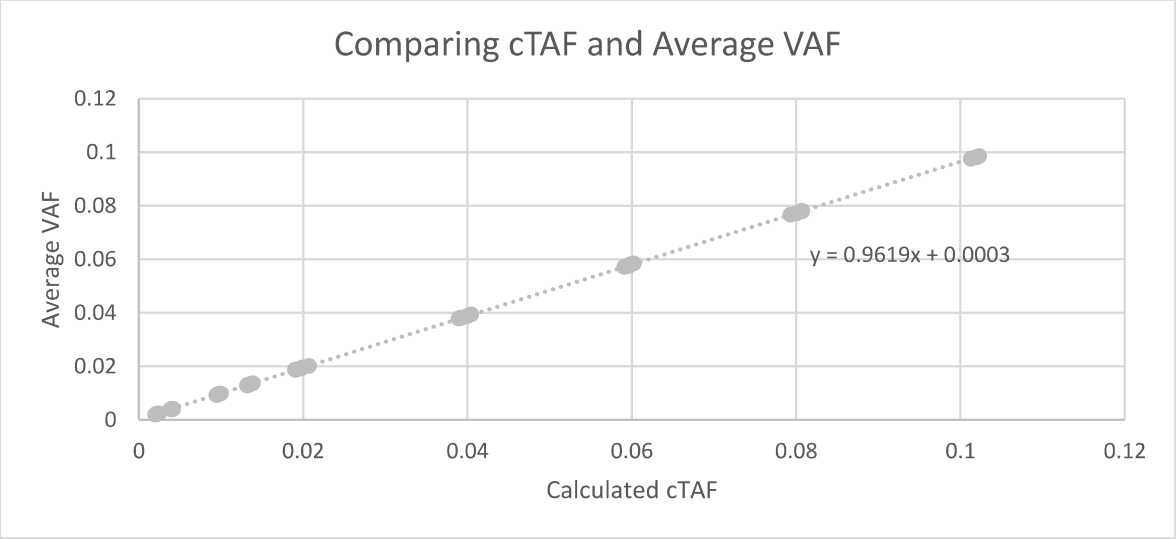
(S. Hampton, personal communication, June 12, 2024)

Graphing this relationship shows that the error relative to the magnitude of the tumor purity is not the largest, something that may have been wrongly inferred from Figure 1. Instead, the percent error is somewhat consistent throughout, which lines up with Figure 1's extremely high R-squared value. In fact, one of the lowest average errors is observed for the 50% purity plasma sample and error on average is lower for 1% and 2% also. The error relative to magnitude is largest around the 5% to 10% tumor purity range. In general, more clustered points in this type of chart could mean a systematic bias, so the general scattering of error points is a good sign. At low cTF concentrations, individual differences in alternate alleles and rounding have a larger impact on the predictions, so the resulting error would likely be more scattered compared to high

cTF concentrations. This is observed in Figure 3, and for high cTF concentrations, the error is less dispersed.

For Method 2, one scatter plot is created, plotting average VAF against cTAF predictions. Both metrics attempt at predicting the fraction of mutated tumor-derived DNA in the blood. The purpose of this is to identify the range(s) where the two metrics are the most dissimilar which can assist in determining the factors that may impact one or the other to give a suboptimal prediction. Circulating tumor fraction (cTF) and circulating tumor allele fraction (cTAF) are not the same since cTF accounts for only tumor-derived DNA, but since tumors also contain healthy reads, this fraction must be multiplied by the purity of the tumor to obtain cTAF. Tumor purity is simply the median fraction of reads in a tumor that have a mutation.

**Figure 4: Method 2 Comparison to Average VAF**




Note: Perfect predictions would result in a R-squared value of 1 and a slope of 1. The actual observations are extremely close to that target.

(S. Hampton, personal communication, June 10, 2024)

It is observable from this scatter plot that the two metrics have a very linear relationship, as expected. However, the slope (0.9619) is below one, meaning that calculated cTAF is

increasing at a greater rate than average VAF for higher concentrations of mutated tumor-derived reads. A prior hypothesis is that average VAF can be more easily affected by outliers. The trend observed in Figure 4 aligns with this since it is much more likely to have individual low-purity mutations in a high-purity sample than individual high-purity mutations in a low-purity sample. For this reason, the outliers in high purity samples are likely bringing down the average VAF, making it lower in comparison to calculated cTAF.

## 5. Discussion

The process for turning BAM files into combined subsampled files, and then performing mutation calling is included in Appendix B for reference. A Python script was used to combine and subsample the BAM files using command-line embedding, and Pysam in Python was used to perform the mutation calling, when given a file with the list of mutations present in the sample.

In summary, the findings presented in this report suggest that using a Poisson Distribution as part of a Bayesian model, as suggested by Jamshidi and colleagues (2022), can predict circulating tumor fraction consistently below 18% error in the trials done. When compared to average VAF, cTAF is likely better at approximating mutant tumor-derived read percentage from a plasma sample as average VAF is more prone to outliers at higher concentrations, which is where the main variation from cTAF occurs. The results that this report has shown are promising, but not conclusive. To obtain more conclusive results, the procedure must be performed on multiple samples with a wider range of percentage reads from the tumor.

The allele-specific copy number analysis tool “FACETS” tool can be used to obtain the most reliable results for validation. Alternatively, a pathologist can determine the ground truth tumor-derived mutant allele fraction, but this is extremely resource-intensive and not practical for evaluating machine learning models. If cTAF can approximate this value accurately, it is much more practical since it only requires a blood sample.

## 6. Conclusions and Recommendations

Though the accuracy of this prediction has been tested to meet the required standards, there are still improvements that can be done to potentially increase the accuracy or reliability of the given cTAF prediction. This includes further testing that can be done if the resources are available and continuations of the research for future applications.

Firstly, when inserting values for cTF using the linear spacing function in numpy, 1000 points, evenly spaced, between 0 and 1 were used. In future implementations of the code, there is the possibility that one can use more points, or an uneven distribution of points, representing more likely cTF values to reduce unnecessary computation. This could even be used in conjunction with other techniques, such as calculating the mean VAF, and using such information to inject more relevant information about possible cTF values.

Further, a simulated in vitro validation method for the process can be implemented, offering a third method of validating the function's outputs. To perform this method, one must down sample a real cancer patient's plasma BAM file and combine the reads with normal control reads from either white blood cells or tissue. Then, the code can be run on the resulting VCF files produced and compared to mean and median VAF, like what was performed in Method 2. Adding a third method for checking outputs adds more reliability to the predictions and may assist in finding systematic biases or inconsistencies.

Finally, Method 1 can be improved by including more target percentages for cTF beyond the ones that are included. Doing this would fill in gaps in the error graph and produce a more accurate representation of where the largest systematic errors lie. This can also help evaluate

more accurately when average and median VAF fail to produce an accurate prediction by comparing those metrics to the predicted cTAF.

Overall, the findings indicate that cTAF is an accurate approximation for mutant tumor-derived allele fraction. However, the reliability has not yet been measured extensively, so more testing is required. Since the code used is all provided, the process used in this report can be relatively easily recreated for additional tests. Using cTAF will be important in determining the limit of detection of MCED prediction machine learning models, guiding researchers more accurately towards better prediction results.

## 7. References

Foser, S., Maiese, K., Digumarthy, S. R., Puig-Butille, J. A., & Rebhan, C. (2024). Looking to the future of early detection in cancer: Liquid Biopsies, imaging, and Artificial Intelligence. Clinical Chemistry, 70(1), 27–32. https://doi.org/10.1093/clinchem/hvad196

Heitzer, E., Haque, I. S., Roberts, C. E., & Speicher, M. R. (2018). Current and future perspectives of liquid biopsies in Genomics-Driven Oncology. Nature Reviews Genetics, 20(2), 71–88. https://doi.org/10.1038/s41576-018-0071-5

Hubbell, E., Clarke, C. A., Aravanis, A. M., & Berg, C. D. (2021). Modeled reductions in late-stage cancer with a multi-cancer early detection test. Cancer Epidemiology, Biomarkers & Prevention, 30(3), 460–468. https://doi.org/10.1158/1055-9965.epi-20-1134

Jamshidi, A., Liu, M. C., Klein, E. A., Venn, O., Hubbell, E., Beausang, J. F., Gross, S., Melton, C., Fields, A. P., Liu, Q., Zhang, N., Fung, E. T., Kurtzman, K. N., Amini, H., Betts, C., Civello, D., Freese, P., Calef, R., Davydov, K., … Swanton, C. (2022). Evaluation of cell-free DNA approaches for multi-cancer early detection. Cancer Cell, 40(12). https://doi.org/10.1016/j.ccell.2022.10.022

Li, S., Zeng, W., Ni, X., Liu, Q., Li, W., Stackpole, M. L., Zhou, Y., Gower, A., Krysan, K., Ahuja, P., Lu, D. S., Raman, S. S., Hsu, W., Aberle, D. R., Magyar, C. E., French, S. W., Han, S.-H. B., Garon, E. B., Agopian, V. G., … Zhou, X. J. (2023). Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring. Proceedings of the National Academy of Sciences, 120(28). https://doi.org/10.1073/pnas.2305236120

# Appendix:

Most essential pieces of the code used is in the following GitHub repository:

https://github.com/Simon-xP/ctf-ctaf-approximation