Comparing activity levels of primary tumors and metastases using circulating tumor DNA

Executive abstract

Quantitatively comparing activity levels of primary tumors and metastases is unmet clinical need. The abilities of tumors shedding circulating tumor DNA (ctDNA) could serve as a marker for their activity levels. I built a hierarchical linear model using mutation allele frequencies in plasma, primary tumors, and metastases to estimate the contributions of primary and metastatic lesions to ctDNA. My results showed that in all patients included in this study, metastases had only around 51% probability releasing more ctDNA than primary tumors, indicating almost equal activity levels. Therefore, primary and metastatic lesions should be considered equally when planning treatment.

Introduction

Circulating tumor DNA (ctDNA) is a fraction of cell-free DNA (cfDNA) in the bloodstream or other bodily fluids that originate from active malignant tumors. Levels of ctDNA are typically very low and correlate with progressive activities of tumors. Dormant tumors release very few ctDNA, which is often below the limit of detection of routine diagnostic techniques. While progressing tumors actively shed ctDNA, comparatively much easier to detect. A number of studies have revealed that ctDNA levels may indicate the response of tumors to certain types of therapies and predict the risk of disease progression.

In patients with metastatic diseases, there may exist different therapeutic options for primary and metastatic lesions. Therefore, quantitatively assessing the relative activity levels of primary tumors and metastases is of clinical significance for optimal treatment planning and/or disease management. However, despite the technical advances in measuring the overall ctDNA levels, it remains an unmet clinical need to quantify the contributions of each malignancy to shedding ctDNA, which underlie their activity levels, due to the lack of an established analytical method.

In this small project, I aim to build a Bayesian model to estimate the proportions of ctDNA that are shed by primary tumors and metastases in a simple scenario where unifocal lung cancer patients developed mono-metastases (in other words, each patient has only one primary lesion and one metastatic lesion). Specifically, for a mutation, given its allelic frequencies in primary tumor tissue (AF_P), metastasis tissue (AF_M), and plasma sample (AF_{ctDNA}), we assume that a small proportion (α) of total plasma DNA is from the primary tumor and that another small proportion (β) is from the metastasis. Ideally, without any noise and system error, we will have $AF_{ctDNA} = \alpha \cdot AF_P + \beta \cdot AF_M$ (*). The comparison of proportional ctDNA shedding, α and β, by a primary tumor and its paired metastasis, respectively, could serve as an approximation to the comparison of actual activity levels. Through statistical modeling, I wish to answer the question – what is the probability that metastases shed more ctDNA than primary tumors, $P(\beta > \alpha)$? (Step1)

* $(\alpha + \beta)$ is the proportion of ctDNA in total plasma DNA (cfDNA). As cfDNA consists mostly of DNA molecules from normal tissues, the absolute values of α and β are expected at a small magnitude.

Data

I collected data of 202 mutations from 15 stage-IV lung cancer patients pathologically confirmed with a mono metastasis to the liver from an internal database. Each row in the dataset represented a mutation. Four variables were included – the patient ID (ID), the allele frequency in plasma (AFct), the allele frequency in primary tumor (AFp), and the allele frequency in metastasis (AFm). As introduced above, AFct may be linearly related with AFp and AFm. Given dozens of observations of the four variables, we were able to estimate the slope coefficients representing the ctDNA contributions by primary tumors and metastases, which provided insight to answering my question on the probability of metastases releasing more ctDNA than primary tumors.

AFp/AFm and AFct data were from genomic sequencing of tissue and plasma samples, respectively. A challenge in the data was some missing values caused by poor sequencing quality of tissue samples. Mutation records with missing values were removed. The final data input into the model included 157 mutations from 10 patients. (Step2)

The scatter plots showed that AFct was likely linearly correlated with both AFp and AFm (**Figure 1**). (Step3)

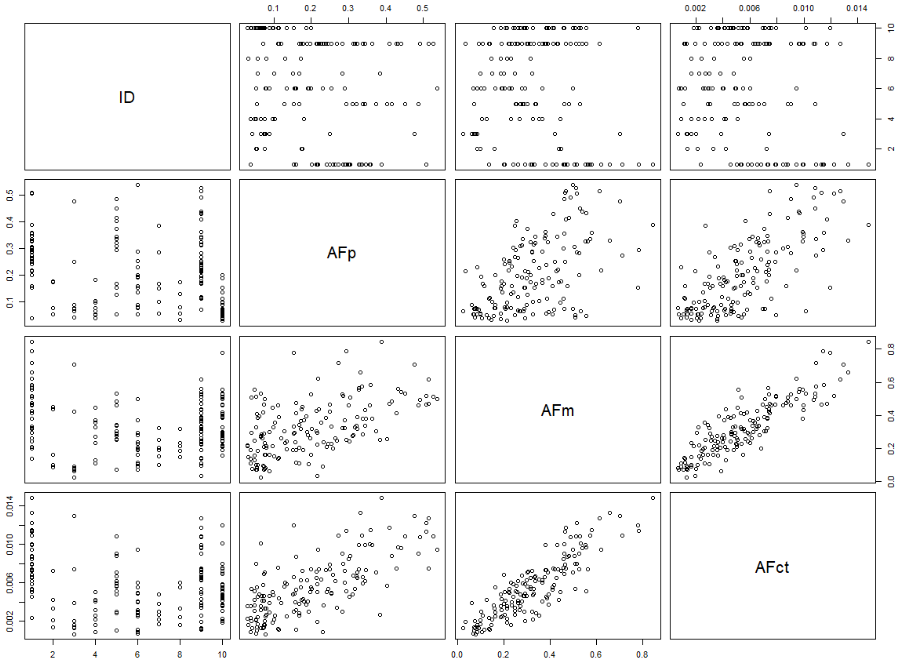


Figure 1. Correlations between four variables

Model

Given the prior knowledge that the response variable AFct linearly correlated with AFp and AFm, I postulated a hierarchical linear model to estimate the ctDNA contents originating from primary and metastatic lesions. Considering the individual variance, tumors in different patients may have different ctDNA-shedding abilities underlying different activity levels. I gave different α and β values for each patient. However, since the patient data had been controlled for the most important clinicopathological features including the cancer type, metastatic site, and pathological stage, different non-negative α_{ID} values across patients may come from a common exponential distribution $exp(\lambda_\alpha)$. The same applied to non-negative β_{ID} values. While for the intercept γ, considering that ctDNA sequencing noises mainly came from random PCR/sequencing/system errors, I did not give it patient-specific values.

The parameters α_{ID} and β_{ID} can be used to calculate the probability P(β > α) in each patient and over all 10 patients. The parameters μ_α and μ_β can be used for general comparisons.

The full hierarchical specification is shown below:

$$AFct_i \mid ID_i, AFp_i, AFm_i, \underset{\sim}{\alpha}, \underset{\sim}{\beta}, \gamma, \sigma^2 \overset{ind}{\sim} N\left(\gamma + \alpha_{ID_i} AFp_i + \beta_{ID_i} AFm_i, \sigma^2\right) \quad i = 1, \ldots, 157;\ ID_i\ in\ \{1, \ldots, 10\}$$

$$\gamma \sim N(0, 10^{-4})$$

$$\sigma^2 \sim Gamma(1,1)$$

$$\alpha_{ID} \mid \lambda_\alpha \overset{iid}{\sim} exp(\lambda_\alpha) \quad ID = 1, \ldots, 10$$

$$\beta_{ID} \mid \lambda_\beta \overset{iid}{\sim} exp(\lambda_\beta) \quad ID = 1, \ldots, 10$$

$$\lambda_\alpha \sim Gamma(2, 10^{-3}/2)$$
$$\lambda_\beta \sim Gamma(2, 10^{-3}/2)$$

Based on the magnitudes of variables AFct (~10^{-3}), AFp (~10^{-1}), and AFm (~10^{-1}), I assigned λ_α and λ_β with exponential priors with mean at 10^{-3} and γ centered at 0. (Step 4)

I fit the model using R and JAGS. Trace plots showed no long-term trends. The Gelman and Rubin diagnostic generated scale reduction factors that were very close to 1 for all parameters. The autocorrelation diagnostic showed no apparent autocorrelation. These indicate that the MCMC has probably converged.

The effective sizes for α_{ID} and β_{ID} (ID in {1,…,10}) were around 2000 (range: 1702.480 – 2724.457) over 5000*3 iterations. λ_α and λ_β had effective sizes of 1553.063 and 1533.561, respectively. Those for γ and σ^2 were both over 14500, suggesting that these parameters mixed well.

The residuals showed no obvious trends or outliers. The qqnorm plot appeared fairly linear, suggesting these residuals were approximately coming from a normal distribution. The model assumption of normal distribution was met. (Step 5 & 6)

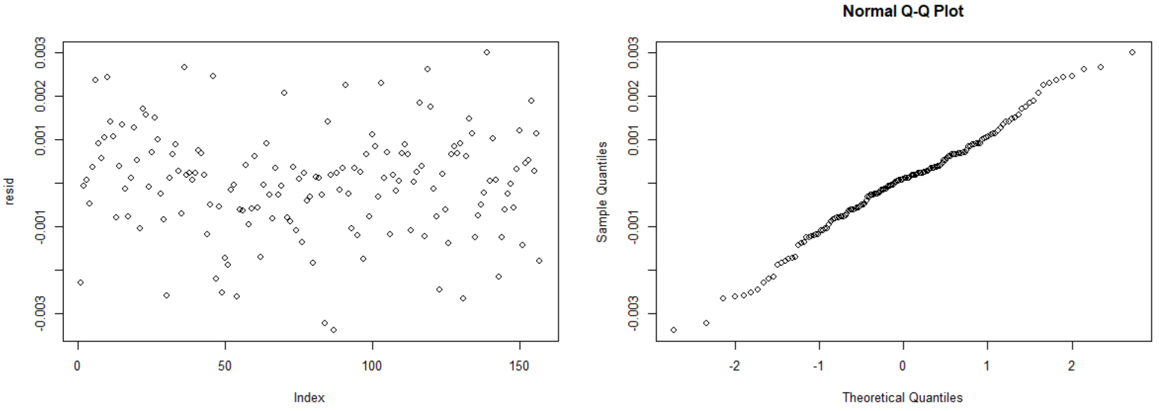


Figure 2. Residual analysis

The current model was adequate to answer my question though, I tested one alternative model with the same assumption of normal distribution but given independent distributions for coefficients (α and β) in each patient, instead of two common exponential distributions exp(λ_α) and exp(λ_β). The DIC value for the alternative model was 394.2, greater than that for the origin model (393.1). Therefore, the original model was the better for answering the question. (Step 7)

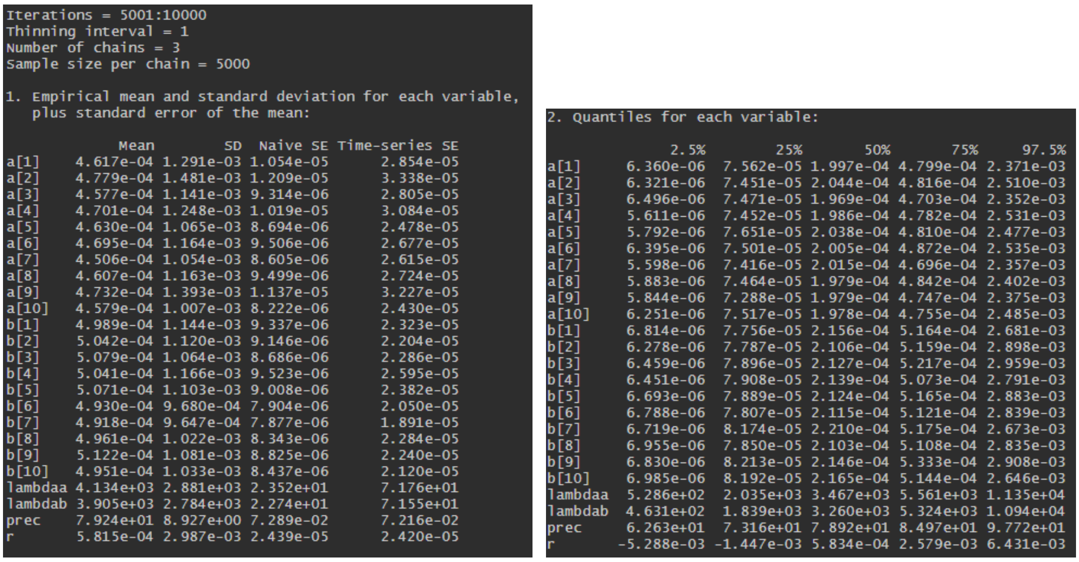
Results

The hierarchical model had 24 parameters (**Figure 3**).

α_k and β_k (k = 1,...,10) were the proportions of ctDNA contributions from the respective primary tumor and metastasis in patient k. α_k ranged from 4.506X10^{-4} to 4.779X10^{-4}, following an exponential distribution with λ_α = 4.134X10^3. β_k ranged from 4.918X10^{-4} to 5.122X10^{-4}, following an exponential distribution with λ_β = 3.905X10^3. γ was the intercept referring to the predicted mutation allele frequency in ctDNA ($AFct_{pred}$) when neither the primary tumor nor metastasis detected this mutation (AFp = 0 & AFm = 0), which indicated a background system error for this set of samples. γ = 5.815X10^{-4} on average account for 10.4% of the mutation allele frequencies detected in ctDNA. The precision for the normal distribution of AFct was 7.924X10^1.

The probability of $\beta > \alpha$ in each patient was slightly greater than 0.5 (**Figure 4**), with a median of 0.5145 (range: 0.5117 – 0.5209). This suggests that primary tumors and metastases in all patients shed almost equal proportions of ctDNA, exhibiting indistinguishable activity levels.

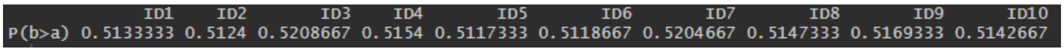
This model, however, is built on a very simplified scenario where each patient had only one primary and one metastatic lesions of the same patho-histology and similar patho-architectures (protected sensitive data, not used in the model but helped the initial data collection). Firstly, the model could not be applied to patients with more than two tumors. Secondly, real-world tumors may have varying sizes and clinicopathological characteristics, which could violate the assumption that ctDNA-shedding abilities of primary tumors in different patients follow a common distribution. Thirdly, the random error did not consider sequencing-depth associated error, which should be specified if the sequencing depth data is available. Fourthly, biologically, mutation allele frequencies should be adjusted by focal and global copy numbers, which data were not available in our project. Therefore, the model may have limited generalizability under real-world settings. (Step 8)



```
Iterations = 5001:10000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

             Mean        SD  Naive SE Time-series SE
a[1]    4.617e-04 1.291e-03 1.054e-05      2.854e-05
a[2]    4.779e-04 1.481e-03 1.209e-05      3.338e-05
a[3]    4.577e-04 1.141e-03 9.314e-06      2.805e-05
a[4]    4.701e-04 1.248e-03 1.019e-05      3.084e-05
a[5]    4.630e-04 1.065e-03 8.694e-06      2.478e-05
a[6]    4.695e-04 1.164e-03 9.506e-06      2.677e-05
a[7]    4.506e-04 1.054e-03 8.605e-06      2.615e-05
a[8]    4.607e-04 1.163e-03 9.499e-06      2.724e-05
a[9]    4.732e-04 1.393e-03 1.137e-05      3.227e-05
a[10]   4.579e-04 1.007e-03 8.222e-06      2.430e-05
b[1]    4.989e-04 1.144e-03 9.337e-06      2.323e-05
b[2]    5.042e-04 1.120e-03 9.146e-06      2.204e-05
b[3]    5.079e-04 1.064e-03 8.686e-06      2.286e-05
b[4]    5.041e-04 1.166e-03 9.523e-06      2.595e-05
b[5]    5.071e-04 1.103e-03 9.008e-06      2.382e-05
b[6]    4.930e-04 9.680e-04 7.904e-06      2.050e-05
b[7]    4.918e-04 9.647e-04 7.877e-06      1.891e-05
b[8]    4.961e-04 1.022e-03 8.343e-06      2.284e-05
b[9]    5.122e-04 1.081e-03 8.825e-06      2.240e-05
b[10]   4.951e-04 1.033e-03 8.437e-06      2.120e-05
lambdaa 4.134e+03 2.881e+03 2.352e+01      7.176e+01
lambdab 3.905e+03 2.784e+03 2.274e+01      7.155e+01
prec    7.924e+01 8.927e+00 7.289e-02      7.216e-02
r       5.815e-04 2.987e-03 2.439e-05      2.420e-05
```

```
2. Quantiles for each variable:

             2.5%        25%       50%       75%     97.5%
a[1]     6.360e-06  7.562e-05 1.997e-04 4.799e-04 2.371e-03
a[2]     6.321e-06  7.451e-05 2.044e-04 4.816e-04 2.510e-03
a[3]     6.496e-06  7.471e-05 1.969e-04 4.703e-04 2.352e-03
a[4]     5.611e-06  7.452e-05 1.986e-04 4.782e-04 2.531e-03
a[5]     5.792e-06  7.651e-05 2.038e-04 4.810e-04 2.477e-03
a[6]     6.395e-06  7.501e-05 2.005e-04 4.872e-04 2.535e-03
a[7]     5.598e-06  7.416e-05 2.015e-04 4.696e-04 2.357e-03
a[8]     5.883e-06  7.464e-05 1.979e-04 4.842e-04 2.402e-03
a[9]     5.844e-06  7.288e-05 1.979e-04 4.747e-04 2.375e-03
a[10]    6.251e-06  7.517e-05 1.978e-04 4.755e-04 2.485e-03
b[1]     6.814e-06  7.756e-05 2.156e-04 5.164e-04 2.681e-03
b[2]     6.278e-06  7.787e-05 2.106e-04 5.159e-04 2.898e-03
b[3]     6.459e-06  7.896e-05 2.127e-04 5.217e-04 2.959e-03
b[4]     6.451e-06  7.908e-05 2.139e-04 5.073e-04 2.791e-03
b[5]     6.693e-06  7.889e-05 2.124e-04 5.165e-04 2.883e-03
b[6]     6.788e-06  7.807e-05 2.115e-04 5.121e-04 2.839e-03
b[7]     6.719e-06  8.174e-05 2.210e-04 5.175e-04 2.673e-03
b[8]     6.955e-06  7.850e-05 2.103e-04 5.108e-04 2.835e-03
b[9]     6.830e-06  8.213e-05 2.146e-04 5.333e-04 2.908e-03
b[10]    6.985e-06  8.192e-05 2.165e-04 5.144e-04 2.646e-03
lambdaa  5.286e+02  2.035e+03 3.467e+03 5.561e+03 1.135e+04
lambdab  4.631e+02  1.839e+03 3.260e+03 5.324e+03 1.094e+04
prec     6.263e+01  7.316e+01 7.892e+01 8.497e+01 9.772e+01
r       -5.288e-03 -1.447e-03 5.834e-04 2.579e-03 6.431e-03
```

Figure 3. Posterior summaries of parameters



```
               ID1    ID2       ID3    ID4       ID5       ID6       ID7       ID8       ID9      ID10
P(b>a) 0.5133333 0.5124 0.5208667 0.5154 0.5117333 0.5118667 0.5204667 0.5147333 0.5169333 0.5142667
```

Figure 4. Posterior probability of β > α

Conclusions

For unifocal lung cancer patients that developed a mono liver metastasis, the primary tumor and metastasis showed almost equal ctDNA-shedding abilities (P(β > α) median at 0.5145), indicating similar activity levels. Therefore, primary and metastatic lesions should receive equal attention on during treatment planning and disease management.