Elucidating tumor evolutionary patterns using high-depth molecular data

Caitlin Guccione
University of Rhode Island
DIMACS REU at Rutgers University & Rutgers Cancer Institute of New Jersey

Dr.Hossien Khiabanian
Rutgers Cancer Institute of New Jersey
Assistant Professor

Cancer is the second leading cause of death in the United States[1] and yet it only has two main treatments, radiation and chemotherapy. A more efficient way to eliminate cancerous cells is with a targeted approach[2]. In order to create more effective precision medication, there exists a need to understand how cancer develops and and to determine which cancerous mutations are most frequent in patients. The optimal way to answer these questions is by sequencing cancer tumors and tracking mutations over time with the help of mathematical trees. We use two genetic distances, Hamming and Nei to help structure the trees. We conclude that Nei's distance does a better job of accurately reflecting the changes in mutations over time and thus can be used in the future to track the evolution of cancerous cells.

I. INTRODUCTION

The second leading cause of death in the United States is cancer[1] and yet it only has two main treatments, radiation and chemotherapy. Although chemotherapy is somewhat effective at removing cancer, it kills hundreds of healthy cells in the process. A more effective way of killing cancerous cells is through targeted therapy or a medication that goes after a single specific mutation[2]. Currently, there are very few targeted therapy medications available that only help a select number of patients. Targeted medications take a large amount of scientific research to develop because they involve finding and focusing on a very specific type of cell which is a lot more challenging for a medication to achieve.

In order to cure more patients at a faster rate and in a less damaging way, common mutations need to be located. This way scientists can focus on finding targeted therapy treatment for these common mutations as opposed other less frequent ones. Another way to improve cancer treatments is to understand the evolution of cancer including growth rate and the factors that affect it. This would allow doctors to be one step ahead and therefore know which cancerous mutations may develop at which rates to prevent relapse before it occurs. Targeted medication can be used earlier in the process or more aggressive chemotherapy may be applied to mutations that would otherwise replicate rapidly.

In this study, we tracked the appearance, disappearance, and change in the amount of certain cancerous mutations. The best way to track these mutations among patients is with the help of high-throughput sequencing[3]. Our study uses, high-throughput sequencing to read the DNA from a tumor and determine which parts are cancerous, what percentage of the tumor is cancerous, and what types of cancerous mutations occurred. The output reveals which mutations are present and how prevalent they are in the sample.

These results can then be displayed in a mathematical model called a phylogenetic tree[4]. This visual representation of the patient data is very similar to a tree found in the graph theory portion of mathematics because it too has no cycles and only one path from every vertex. However, it differs due to the fact that phylogenetic trees may have more edges than vertices in order to more accurately display the data. The appearance of a phylogenetic tree and the relationship between trees can be calculated since the connection between vertices and the length of the edges are all based on mathematical calculations. We mainly use a neighbor-joining method in our tree construction. This simply means the tree is created from a distance matrix, which calculates the variation between two samples. There are multiple ways to determine exactly how different two samples may be from each other which will be discussed later in the paper. Regardless of the method used, once the distances are placed into the matrix, an algorithm which repeatedly finds the most similar samples is applied until all samples are used and a tree is created[5][6]. The end result is a tree with branches representing the distance between mutations and an overall picture of the changes in mutations over time. Figure 1 shows how different types of trees reflex different evolutionary paths.

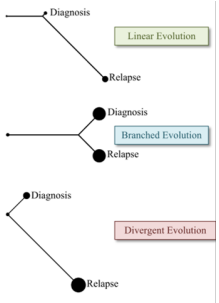

FIG. 1. Phylogentic Tree Relationships

In this study we use high-throughput sequencing data from patients at Rutgers Cancer Institute of New Jersey to construct phylogenetic trees and evaluate them. The main goal is to look for patterns among trees that may lead to answers about cancer's evolution and track which mutations come up most frequently among patients.

II. MATERIALS AND METHODS:

A. Sequencing Data

After a cancerous tumor has been sequenced, all relevant information needs to be pulled from the high-throughput sequencing lab's output[3]. A few demographic features such as gender, age, and type of cancer are found in sequencing output and may be relevant for future work in analyzing cancer growth but, are not our focus. The important data pulled includes the purity of the sample, the allele frequency, location and depth of all cancerous mutations. The purity of the sample is simply the amount of the sample that was found to contain cancerous mutations. The allele frequency reveals out of all cells in the tumor, what percentage contain that particular cancerous mutation. The location of the mutation is given by the gene and amino acid where the mutation occurred, along with the depth representing how confident we are in the allele frequency of the given mutation.

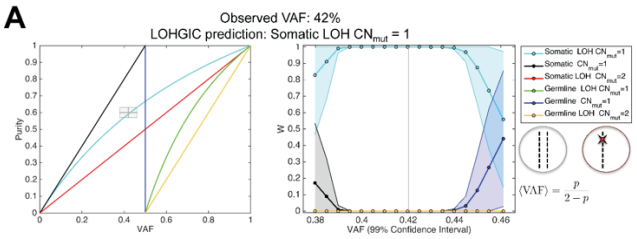

FIG. 2. LOHGIC Sample Output

B. LOHGIC

Along with the information gathered from the sequencing data, the type of mutation that occurred may also be calculated using LOHGIC algorithm[7]. Figure 2 shows an example of LOHGIC's output. This software takes the purity, allele frequency, and depth of any mutation and outputs the type of mutation that occurred and how confident it is with its selection. There are three main parts to a mutation which include, germline vs. somatic, the number of copies of a gene (Y), and the number of cancerous copies (CNmut). Figure 3 shows different types of mutations and what they mean. This information is crucial because it affects the treatment of the patient. For example, germline mutations are ones the patient is born with and thus are present in every cell in the body, including non-cancerous ones and can be passed onto their children. As opposed to somatic mutations, which just occur in that particular cancerous cell. Germline mutations require a much more aggressive approach to treatment since they are more likely to appear all over the body with the mutation being at least partially present in every cell. Another important value to consider from LOHGIC's output is the difference between Y and CNmut. This reveals if the cell has any working alleles left or all wild-type working copies of a gene are deleted. Clearly if the entire cell has gone cancerous, this will also require a more aggressive approach than if just one allele is cancerous.

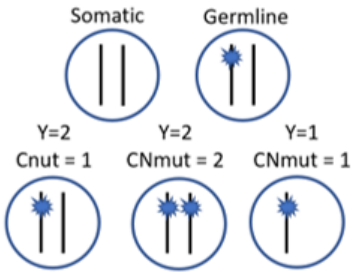

FIG. 3. Types of Mutations

C. Cancer Cell Frequency

In terms of interpreting that data, it may seem like using the allele frequency may be the best way to measure the impact or amount of each mutation, but in reality, the allele frequency is heavily skewed by the purity of the sample. This can be seen clearly in figure 4. A better representation of the data is by using cancer cell frequency, $< CCF >$. This measure takes the total number of cells with a certain mutation out of the total number of cancerous cells as opposed to taking the number of mutated alleles out of all alleles. The CCF is less likely to be influenced by shifts in purity and also gives cells that only have one mutated allele an equal weight of cancerous cells that have two affected alleles. Along with calculating CCF, we also added error bars based on the depth of the allele frequency inputted. This way the outputted data will have the error that comes with using sequencing data providing a measure of accuracy.

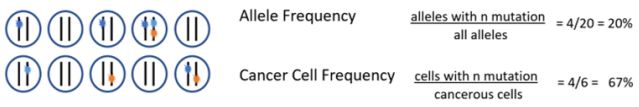

FIG. 4. Allele Freq. vs. CCF

D. Errors within Purity

There is discrepancy in the data at times, particularly when it comes to purity. The purity of the sample is found by dying the sample and then attempting to distinguish between cancerous and noncancerous cells underneath a microscope. As a result, some samples have two purities, one of which is 30% and the other 80%. As seen in figure 5, purity has a dramatic effect on the CCF calculation.

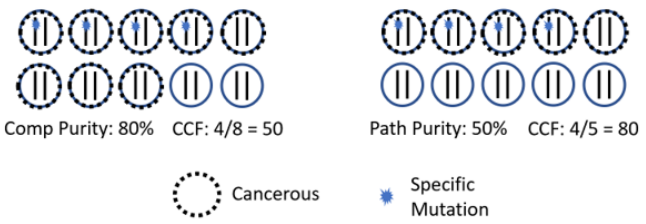

FIG. 5. Error within Purity

III. RESULTS

A. Cleaning Up Purity

As shown figure 5, an inaccurate purity leads to a misleading CCF. Before calculating and comparing CCFs across samples, the purity needs to be accurate. So, a purity cleaning algorithm was designed to help determine the correct purity.

B. Purity Cleaning algorithm

For each sample, there is only one correct purity since purity is defined as the percentage of cancerous cells. The problem is how do we find the correct purity based on the data and mutations given. In order to simplify the model, we currently assume that all mutations are clonal or occur in every cancer cell. This is not true for all datasets and will be improved upon in the future, but for simplification, it was assumed for this algorithm. The reason we use only clonal mutations is due to the fact that their CCF should be equal to one. This way the CCF can be set to one and the calculations on purity can be traced back using allele frequency, Y, and CNmut values. The algorithm assumes that no purity is given and follows 5 basic steps:

1. Calculate the purity for each model

Due to work done with LOHGIC[7], the expected variant allele frequency, $< VAF >$, can be calculated for each model based on the purity, Y, and CNmut values. The equations can be found in figure 6. Since the VAF, Y and CNmut value is known for each mutation, the purity value can be solved for. Because we are not sure on exactly which out of the models shown in figure 6 the mutation may be, the purity is calculated for all possible models depending on Y.

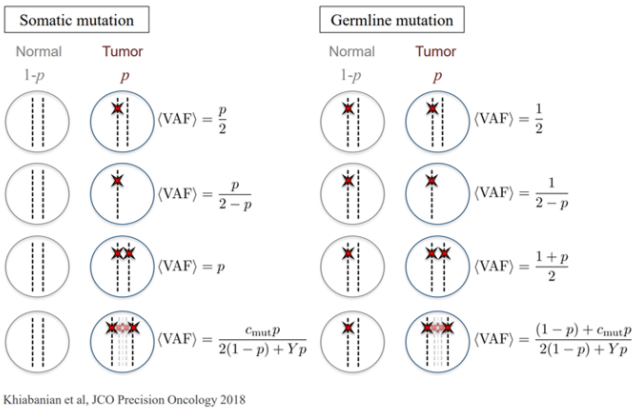

FIG. 6. Somatic vs. Germline

2. Calculate the CCF for each model

Once the purity is known for each model and mutation, the CCF may be calculated following the algorithm explained in the section titled Cancer Cell Frequency and using the following equation:

$$CCF = \frac{\text{Cells with } n \text{ mutation}}{\text{Cancerous Cells}}$$

3. Find the weight of each model

Any mutation may have any of the mutation types in figure 6, LOHGIC gives us an idea of which models are most likely and the probability that such model is correct. These probabilities are known as each model's weight.

4. Using the following Eqaution:

$$\sum_i \sum_j Wij(CCFij - 1)^2$$

Where i= number of mutations and j= possible models based on Y and W is the weights from LOHGIC. This is a least squared equation that assumes the CCF should be as close to 1 as possible, hence why only

clonal mutations are ideal for this algorithm. It then takes the weight of the model into account and sums up all possible models to output a final purity estimate.

C. Sequencing Data Simulation

In order to ensure that our purity cleaning algorithm was correct, a sample sequencing data simulation was created. This program chooses a random purity that is hidden in a separate file from fabricated data. It creates data based on the models in figure 6, choosing a random Y value when necessary. CNmut is also randomly selected based upon the Y value. A bit of noise is added to the allele frequency using a binomial distribution based on a random depth since real data would not be totally clean. In addition, the possibility of getting a few subclonal mutations was also added. As a result fabricated data would be as close to actual data as possible without making the majority of mutations subclonal.

IV. ANALYSIS

A. Genetic Distance Types

Each mutation found inside sequencing data contains a large amount of information. We have further processed the data to get CCF. With a large amount of data now available, a new challenge is discovering which parts accurately measure the change in cancerous mutations over time. The goal of forming trees is to find the genetic distance or the degree of separation, between mutations. Luckily, this is a common problem for biologists tracing through evolution and finding species. Therefore, there are a variety of statistical methods that can be applied to find an accurate measurement. Below we discuss a few genetic distance measurements and their advantages and disadvantages.[8]

B. Hamming Distance

This measurement is perhaps the simplest way to trace mutations over time. It first finds all mutations found in all samples. In our case, all samples refer to mutations from the same patients at three different dates. Once this list is complete, each time is measured by adding one for every mutation present and adding nothing if the mutation is not present. Figure 7 displays how the table and tree would look using Hamming distance. This genetic distance is helpful because it can be simply calculated, but it's not incredibly useful because it ignores how much of each mutation is present in each sample. For example, if a mutation is present in 1% of the sample it has the same effect on the tree as a mutation that's present in 80% of the sample.

C. Nei's Distance

Nei's distance[8] is a measurement that takes mutations and genetic drift into account to form the following equation:

$$\frac{\sum_i Pi \cdot Qi + (1-Pi)\cdot(1-Qi)}{\sqrt{\sum_i P_i^2 + (1-P_i)^2} \cdot \sqrt{\sum_i Pi^2 + (1-Pi)^2}}$$

In this case, P and Q are the CCF's from two separate time points for every mutation found at ever time in the patient data. The reason the summation is removed and $(1-P_i)$ and $(1-Q_i)$ are added to this equation is due to the fact that we do not know how many mutations are possible on this particular site, but we do know that this mutation is either absent or present. Since CCF is usually not simply 100% or 0%, Nei's distance does a better job of taking how frequently mutations occur into account as opposed to the mutation just being simply present or absent. Not only is Nei's distance more accurate than Hamming distance, it has an advantage over other models for a few other reasons. First of all, Nei's distance is built off of a concept called the Kinship coefficient which measures the probability that two alleles from the same person are identical[8]. This concept can then be further translated into genetic distance or divergence between mutations. Secondly, Nei's distance has the advantage of, if the rate of mutations is constant, producing a linear output in reference to time since mutations naturally occur at a certain rate.

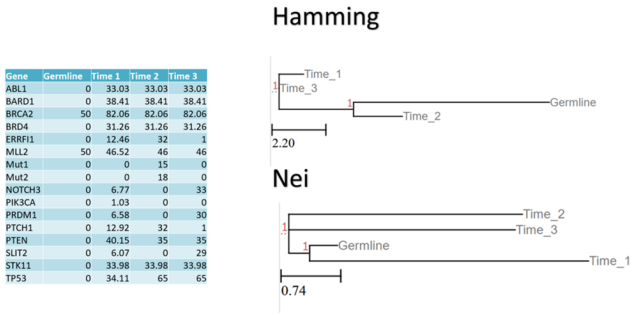

FIG. 7. Hamming vs. Nei Distances

D. Comparison

If we take a closer look at data shown in figure 7, Hamming and Nei create two drastically different trees. The data on the left side shows the location of the mutation and the allele frequency for each time, the right side shows the resulting trees. Hamming places germline closest to Time 2 because Time 2 has the largest amount of missing mutations making it appear similar to germline. In reality, if you take a closer look at the data, Time 1 is actually closer to germline because although it does have

most of the mutations present, a lot of its allele frequencies are very small. Thus, as discussed above it makes sense to use Nei's distance in future trees.

E. Future Works

In the future, there are a few minor adjustments that would make significant improvements in our research. The first would be finding an accurate way to add subclonal mutations. This involves coming up with an algorithm to do so and thus may take some time. Other ways to improve would be to find a way to add error[9] into either calculating Nei's distance or into the tree itself. Since we are using sequencing data, it is never going to be perfectly clean; therefore, we must show this in our results by also having a degree of error. Finally, after building these trees for a large number of patients, it would be helpful to find a way to track patterns within the trees in order to understand the evolution of cancer cells and determine which mutations occur most frequently. [5]

ACKNOWLEDGEMENTS

This research was conducted as part of a Research Experiences for Undergraduates at Rutgers University New Brunswick along with the Rutgers Cancer Center of New Jersey. The work was done under Dr. Hossien Khiabanian in Hossien Labs. This was funded by NSF grant CCF-1559855: REU Site: DIMACS. A speical thanks to DIMACS for hosting and Lazaros Gallos for directing.

REFERENCES

[1] S. R. L., M. K. D., and J. Ahmedin, CA: A Cancer Journal for Clinicians **68**, 7, https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21442.
[2] E. A. Ashley, Nature Genetics **17** (2016).
[3] J. D. M. Sara Goodwin and W. R. McCombie, Nature Genetics **17** (2016).
[4] R. Schwartz and A. A. Schäffer, Nature Genetics **18**, 213 (2017).
[5] D. A. B. Boc, A. and V. Makarenkov, Nucleic Acids Research **40(W1), W573-W579** (2012).
[6] F. S. Jaime Huerta-Cepas and P. Bork, Mol Biol Evol **doi: 10.1093/molbev/msw046** (2016).
[7] H. Khiabanian and K. M. Hirshfield, JCO Precision Oncology (2018).
[8] M. Nei, The American Naturalist **106**, 283 (1972).
[9] "A Summary of Error Propagation," `http://ipl.physics.harvard.edu/wp-uploads/2013/03/PS3_Error_Propagation_sp13.pdf`.