Cancer Genomics

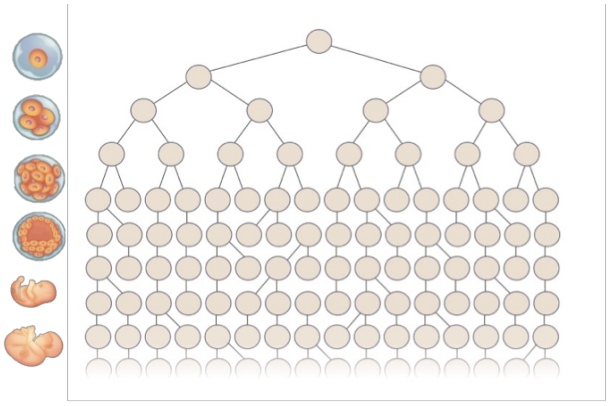
Lecturer: Victoria Popic

February 5, 2015

Scribe: Melissa Johnson

How can we use computer science to learn more about how cancer develops? Research in this area is exploding because sequencing has become cheap and some of the data has become publicly available.

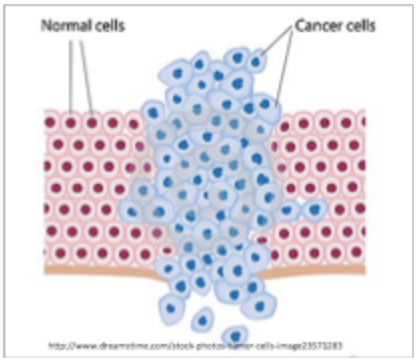
Mutations



Source: Dorna Kashef-Haghighi

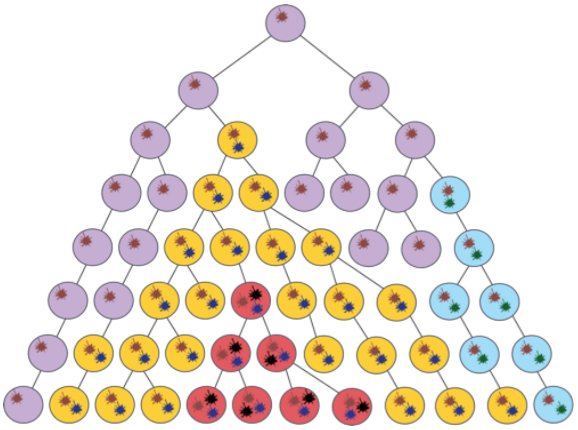
Life starts with one cell, a zygote. This cell goes through many rounds of cell division and differentiation to make up and renew all necessary human body cells. Sometimes when cells divide, they gain **mutations**. Most of these mutations will not have any side effects, or will cause death only for the harboring cells thus not affecting other cells in the body.

However, sometimes these mutations are highly malignant, and do cause problems. This is especially the case when a mutation affects the ability of a cell to divide and grow, or the mechanism that controls cell death. This type of mutation can result in an uncontrollable division and proliferation of one cell. And this is when a tumor forms.
The tumor can be a result of one mutation, or an accumulation of mutations over time.



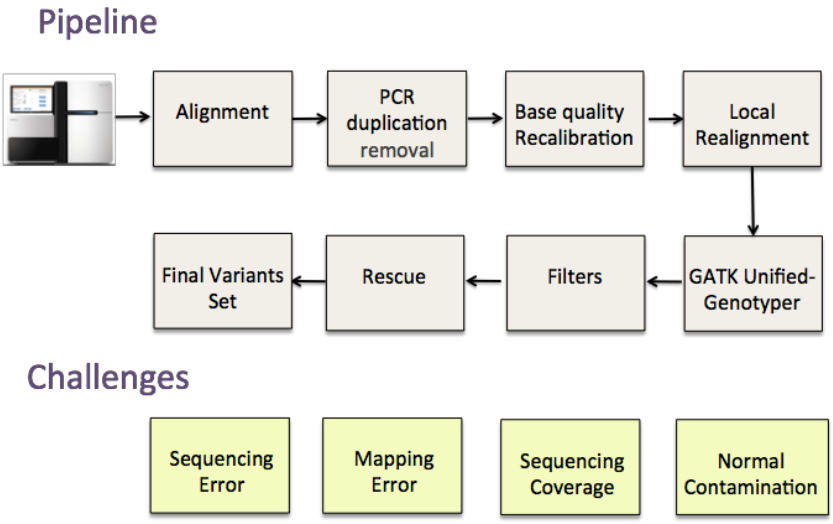
As the tumor grows, cells accumulate more and more mutations and become increasingly more malignant. At some point, some cells can become aggressive enough that they then metastasize to other parts of the body. So cancer is **inherently** a disease of the **genome**. And it's a **process** that develops over a long time. In order to find a cure for cancer or prevent it at an early stage, we have to understand this process at the **genetic** level.

As the tumor grows, there is differentiation within the tumor itself as the cells accumulate more and more mutations (similar to the process of differentiation that occurs as a zygote grows into an organism).



Variant Calling

With today's technology, we can analyze and sequence genomes, and figure out what the modifications were that happened and led to cancer. You would find a lot of differences that aren't actually relevant, but some of them are. We can use **variant calling** to figure out how the DNA differs from a reference sequence (using sequencing technology).



Trying to find mutations in DNA can result in very noisy results, and we can get a lot of errors. Sequencing error, mapping error, non-uniform sequencing coverage and normal contamination in cancer lesions can all result in false positive or false negative variant calls.

We can filter and recalibrate our results to improve them. For instance, we can use local realignment to shift reads around a bit based on other reads that map to that location. We can use the results of other reads to figure out if the read I'm a little bit uncertain about is where it should be. After all these steps, we should get a good set of candidates for variants.

This process is a general one, not just for cancer sequencing. Cancer sequencing is more specific and in some ways harder (especially because it may be that only a small number of cells in your sample have the variant), but the general process is the same.

Single Nucelotide Variants

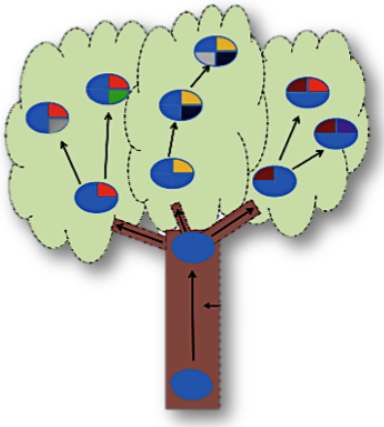
There are two types of variants one can find when comparing a cancer sample to a reference genome:

- **Germline variants** (SNPs) - germline variants are inherited from the parents. These are pretty common (you would expect an individual difference about every 1000 positions) and they are present in all body cells.
- **Somatic variants** (SSNVs) - variants that are acquired, not inherited. Some may be cancer-causing. These variants will only be present in some of the cells of the body.

To distinguish between germline and somatic variants, all cancer studies also sequence a sample from normal cells of the same patient. The variants that are present in both cancer and normal are then labeled as germline variants. In our projects we rely on both types of mutations to analyze cancer genomes.

Intra-Tumor Heterogeneity

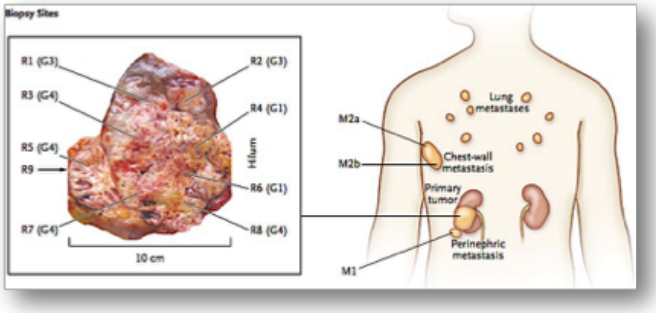
A tumor is basically a mixture of different cells with different mutational signatures. Each cells have potentially different mutations that they harbor. Some of the cells will have mutations in common, but as the cancer cells keep dividing, they generate more and more mutations. So, some of the cells in the sample will have a mutation, and some won't. This complicates the analysis of the cancer genome significantly.



Branched Tree Evolution Model

We can think of these variants as falling on a tree. The trunk represents the mutation that is transferred to all of the tumor cells. But, as the cells differentiate, we get branches of mutations that occur in some tumor cells but not all of them.

Multi-Sample Sequencing



In order to learn more about what kinds of mutations have occurred to form a tumor, we use multi-sample sequencing: we collect multiple samples of the tumor in order to compare the variants found in each. There are several different ways we can do multi-sample sequencing:

- Collect samples from different regions of the mass
- Collect from the tumor and the metastasis sites
- Collect from the same tumor over time

We use absence of mutations from certain samples to understand the progression of the evolutionary tree.

LICHeE: Lineage Inference for Cancer Heterogeneity and Evolution

- LICHeE aims to reconstruct the evolutionary tree of a particular tumor.
- Figure out which mutations happen in the parent cells, which happen later on.
- Should be **fast** and **scalable**.

Input: SSNV multi-sample variant allele frequencies (VAFs)

Algorithm steps:

1. Grouping and clustering SSNVs
2. Evolutionary Constraint Network Construction
3. Lineage Tree Search and Ranking

Perfect Phylogeny Model: Constraints

We use a **perfect phylogeny model** to impose constraints on the algorithm. Basically we assume that mutations do not recur independently in different cells; cells sharing the same mutation must have inherited it from a common ancestral cell. It's possible, but very unlikely that mutation happened in same place, in different cells, independently. The much more likely scenario is that it was inherited.

If this is the assumption, you can derive a few constraints:

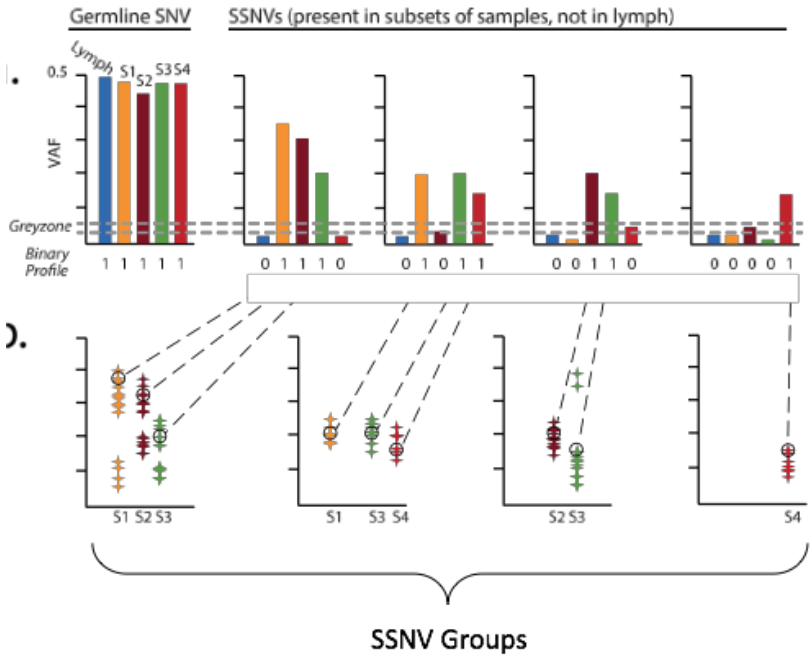
1. A mutation present in a given set of samples cannot be a successor of a mutation present in a smaller subset of these samples
2. A mutation cannot have a VAF higher than that of its predecessor mutation (except due to CNVs)
3. The sum of the VAFs of mutations disjointly present in distinct subclones cannot exceed the VAF of a common predecessor mutation present in these subclones

Our goal is to find all lineage trees that satisfy the above three constraints. The constraints are per sample, that is, they should hold within a sample.

Step 1: Grouping and clustering SSNVs:

We group and cluster SSNVs by presence patterns across samples and VAF similarity.

Presence patterns across samples:



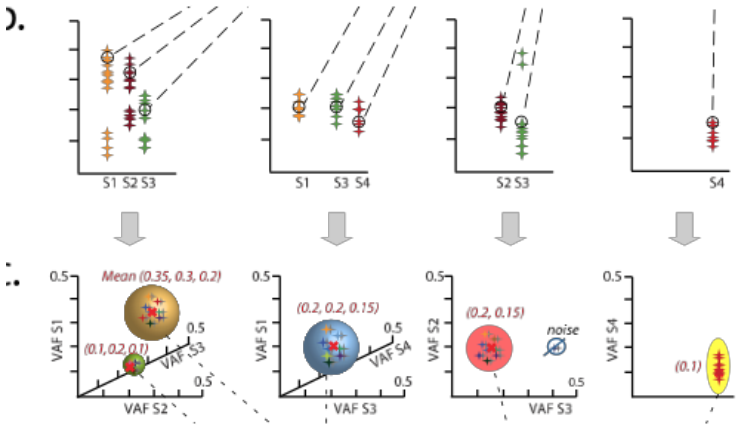
We assign 1 if present, 0 if absent to a mutation in each of the samples. We use a hard threshold/cutoff for simplicity. If the frequency of a mutation is below this cutoff, we say it is absent. This is because we might have sequencing errors; there's a lot of noise.

In the above charts, the first mutation is present in very high frequencies in all samples, including LYMPH, which is normal tissue (non-cancerous). So, any variant present in Lymph is not cancer because it is present in germline.

We then graph what is not present in lymph, and group mutations based on their profiles (SSNV groups).

VAF-based clustering

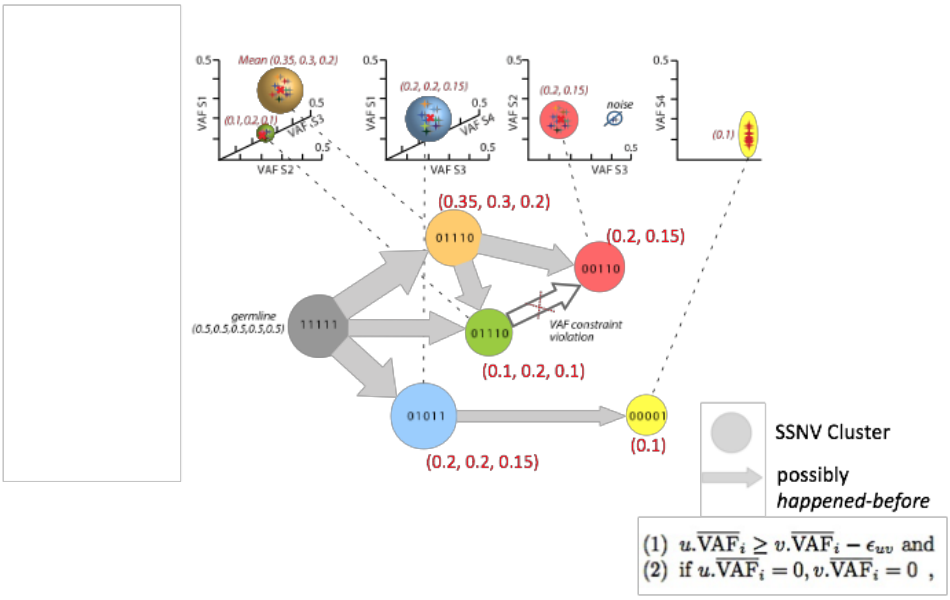
Next, take and cluster based on VAF with some standard clustering algorithm. We make clusters of mutations occurring with similar frequencies. Even if mutations co-occur in the exact same number of cells, there may be noise from sequencing errors, which is why we have to allow for some buffer.



Step 2: Evolutionary Constraint Network Construction

- encodes whether a given cluster of SSNVs could have preceded another
- valid lineage trees are embedded in this network

We can't do this with individual mutations (not enough data) but we can do this with groups of mutations. We can trim a lot of the search space because the valid lineage trees are embedded in this network. The network encodes whether one cluster and another cluster could have preceded each other, but isn't build adhering to our third constraint.

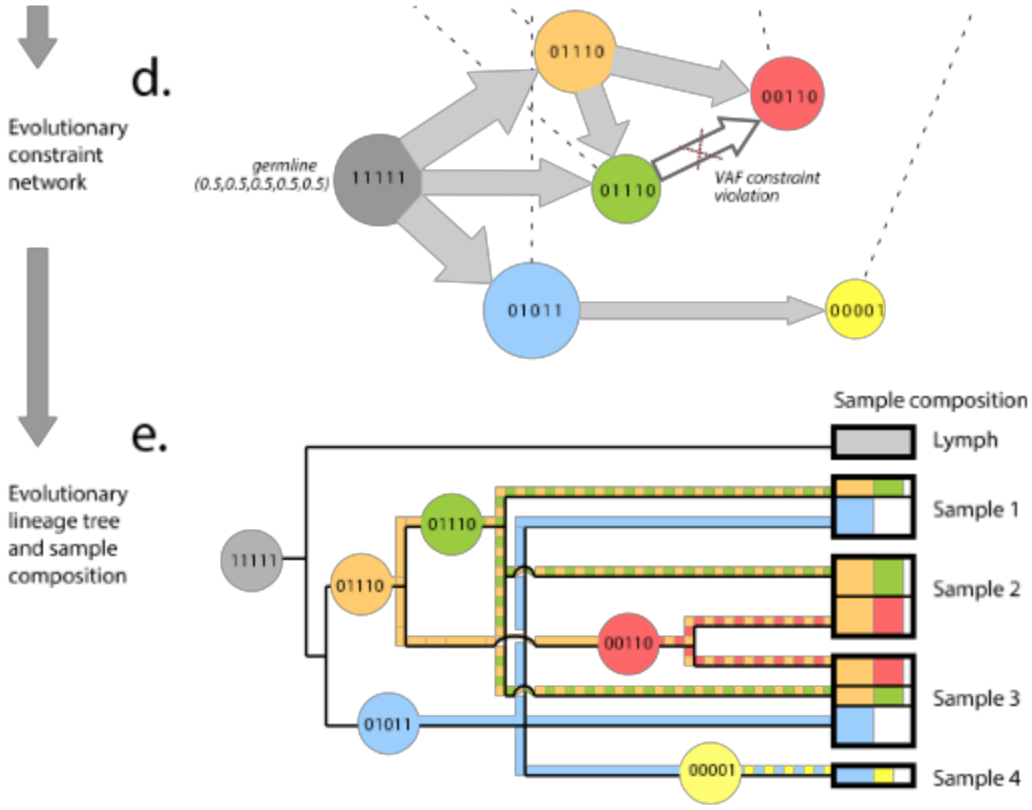


We construct the network by making each cluster a node and putting an edge between the nodes if the mutations in the first node could have been a parent of the third child (a happened before relationship).
For each sample, the VAF of the frequency of the parent is higher than the VAF of the child (with some error margin). If the variant is absent in the parent, it should be absent in the child.

Step 3: Lineage Tree Search and Ranking

—search for spanning trees satisfying VAF constraints within an error margin
—top tree minimizes the squared deviation from the cluster centroids.

- The search finds all spanning trees first: it's a tree, each node with only one parent. Runtime depends on number of embedded trees (O(nodes + edges + trees))
- Then, I apply the third constraint to each node of the tree (the sum of the children <= VAF of the parent), ignore it if it doesn't fit this constraint.
- If I already know I can't have an additional child without breaking third condition, I can terminate that branch early (slight optimization, but same runtime).



Searching the network, we get a tree. The samples are at the leaves of this tree.We connect each sample to the last node on the path of mutation nodes that are present in the sample. So, I basically connect a particular sample to the last node on all the paths that can go to the sample.

You can connect the sample to multiple nodes, which divides the sample into subclones. You can look at the relative frequency of the mutations to figure out the relative numbers of cells in each subclone in that sample. This gives us the signatures of each subclone.

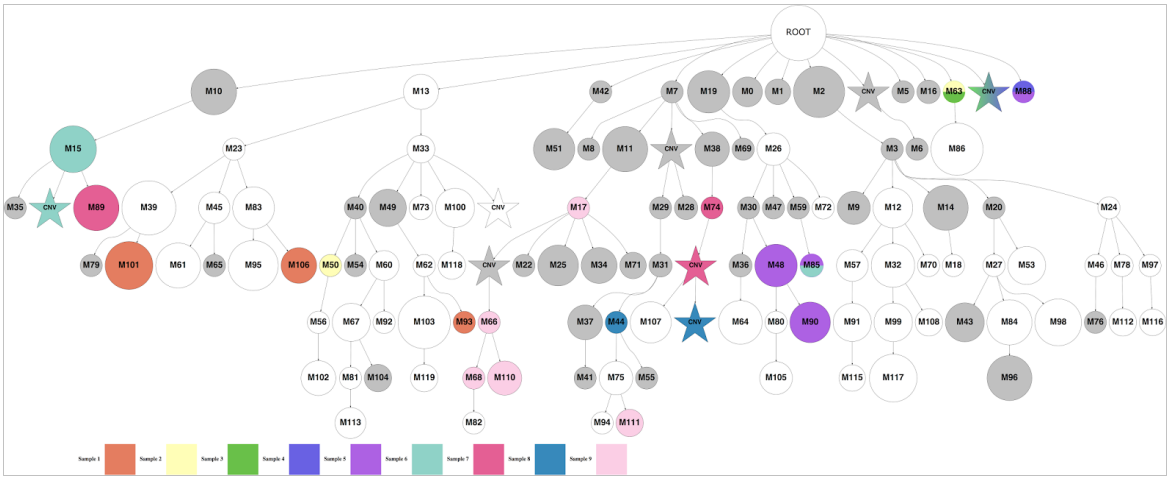
Results

- LICHeE Runtime is very fast in practice for the datasets we worked with.
- You can explore composition of each node, provide metadata
- You can see composition of each sample.
- Also allows you to remove nodes.

ccRCC Study by Gerlinger et. al (2014)

- 8 patients, 587 SNVs
- Studied particular places in genome they thought were of interest.
- High coverage to make sequencing more accurate
- Extracted samples from individual patients, looked at mutations.
- They ran and produced trees, and sometimes had interesting decompositions of subclones.
- They also did manual processing applying traditional phylogemony techniques. These results matched the tree exactly; they were very good results. The subclones found were also validated.

Simulations



- Implemented very simple cancer lineage simulator.
- Starts with a cell probability of live, die, mutate (divide, grow, or die)
- Select different samples of nodes (random or localized)
- Illustrates type of work we do to validate.