CS 262: Cancer Genomics Lecture

Anvita Gupta

February 11, 2016

Introduction to Cancer

- All of us start from one cell, called the zygote. This zygote then divides into two more cells, which then divide again, eventually producing the billions of cells that make up our bodies.
- The cells in our body are still constantly dividing– there are 100 billion cell divisions going on every day in our bodies! Some cells divide more than others (ex. skin cells divide more rapidly than heart cells, which don't divide at all)
- The process of cell division is called **mitosis**. In Mitosis, DNA is duplicated and one copy is given to each progeny cell.
- Cell division usually occurs properly, but sometimes there are "bugs" and the DNA is not copied properly. Errors in DNA replication are known as **mutations**.
- Sometimes these mutations lead the cell to undergo apoptosis, or cell death. Often these mutations are silent. Occasionally, these mutations are beneficial to the cell.
- **Cancer** is a disease that occurs because mutations lead certain cells to divide much more quickly than other cells in the body, leading to the formation of a tumor. Cancer is the leading cause of mortality for people under 65.
- In this way, Cancer can be thought of as a disease of the genome– caused by a single mutation or by a series of mutiple mutations.
- Studying the DNA of cancer cells can help researchers understand what has gone wrong in the cell and what processes have been affected.
- In the process of **Angiogenesis**, a cancer tumor forms blood vessels in order to gain nutrients and survive.
- A Tumor becomes most dangerous through the process of **Metastasis**, in which cancer cells spread to other parts of the body through the bloodstream.
- There are several main factors leading to cancer:

  - Genetics
  - Viruses
  - Exposure to radiation
  - Toxins
  - Diet rich in fat and low in vegetables

- Treatment Options:
  - Surgery
  - Radiotherapy– prevent cells from multiplying
  - Chemotherapy– medicine to prevent cancer cells from multiplying (attack cells that divide rapidly)
  - Note: Chemotherapy and Radiotherapy also act on healthy cells, because they act on any cells that are dividing quickly (including hair and skin).

Cancer and DNA Sequencing

- Why is so Cancer so Hard to Cure?
  - There is no one type of cancer: the mutations that lead to lung cancer are not necessarily the mutations that lead to stomach caner
  - Even 2 tumors in same person can have different DNA mutations!
  - Furthermore, even a single tumor can have many types of mutations. This is called **Intra-tumor heterogeneity**
  - Can think of mutations in a tumor as a tree: start with one, proliferation occurs, more and more types of mutations accumulate. (Figure 1)

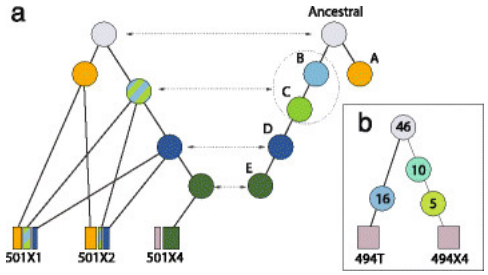

Figure 1: Example of tree of mutations accumulated over time, created from breast cancer xenoengraftment data. Generated by LiCHeE (explained below).

- Good News! Due to advancement of sequencing technology, can sequence cancer cells at very cheap cost. This allows us to attack specific mutations in a tumor.

- Multi-Sample Sequencing
  - Helps us understand the "tree" of Cancer mutations (Fig 1)
  - This, in turn, helps us understand what the **driver** mutations were of the tumor, and differentiate them from silent **passenger mutations**
  - Cancer cells from the primary tumor location are sequenced at different time points, to give us an idea of how the tumor has evolved.
- DNA Sequencing Machines provide us with a set of many short reads (100 bp long) which we can align to a reference genome to understand variations of the genome with respect to a reference.
- The Realignment Step fixes some of the local alignments of the short reads to the reference genome, and filters out noise.
- Hard to align genome to standard reference– get more mutations than expect because of natural variations in the human genome
- Can align to reference genomes from different populations, thus making the alignment better and accounting for natural variations.
- There are issues with finding mutations using this approach: namely Mapping Errors, Sequencing Errors (technology). Cancer cell samples specifically might also be contaminated with normal cells, reducing the amount of reads we have of the cancer cell mutation.
- Question: Maybe important mutation only occurs in 1% of cells– how to filter this important mutation from noise?
- We are searching for two main types of single nucleotide variant mutations:
  - Single Nucleotide Polymorphisms (SNPs)- inherited from parents, so every single cell in body would have those variations. These are known as *germline mutations*.
  - Somatic mutations (SSNVs)- acquired during lifetime, not present in every single cell of body
  - Cancer is driven by many somatic mutations
  - To determine whether mutation is somatic or not, need to compare to results we get from normal cells in the body

LICHeE: Fast and Scalable Inferences of Cancer Lineages

- Goal: Understand intratumor mutations (identify subclonal tumor populations)
- Input Data: Matrix where every row represents a mutation that occured in the tumor

- Every column represents the frequency of this mutation in the particular tumor sample (total number of reads with this mutation / total number of reads)
- Goal is to construct a hierarchy of mutations according to how the mutation occurred during the tumor progression
- Thus, can infer composition of a given sample– how many subpopulations it contains, etc.
- Using the frequencies of variants, we can reconstruct a history of mutation development, with all mutations stemming from a single parent mutation
- The different cells with different progressions of tumors can be identified morphologically under a microscope from samples of cells taken from different parts of the tumor
- Prefer to have as many samples as possible, since if the sample size is small, then the subcounts of a mutation might be so small that we will miss out on diversity if we take only a few samples (analogy- unlikely to find a rare bird in a forest if only take a few samples of birds)
- In addition, we want to have as many reads as possible covering each position (very high coverage) to help eliminate noise.
- **Steps of LICHeE (Figure 2):**
  - i Grouping and Clustering SSNVs by VAF (Variant-allele frequency)
  - ii Evolutionary Constraint Network Construction
  - iii Lineage Tree Search and Ranking

- Main assumption: Perfect Phylogeny Model Assumption
  - Mutations do not recur independently in different cells, or 2 cells w same mutation must have inherited it from a common ancestral cell
  - Couldn't have been case that 2 cells got same mutation by chance
  - This isn't always true in practice- but simplifies our model
- Constraint 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.This is because if M1 is parent, then every cell that has M2 will have M1 also, so M1 cannot be present in a smaller set of samples than M2 if it is a parent.
- Constraint 2: If Mutation 1 (M1) has a lower frequency than Mutation 2 (M2), then Mutation 1 could not have been an ancestor of mutation 2. This is because, if M1 was a parent with a lower frequency, then there are cells that have M2 and not M1, which is impossible is M1 is a parent.

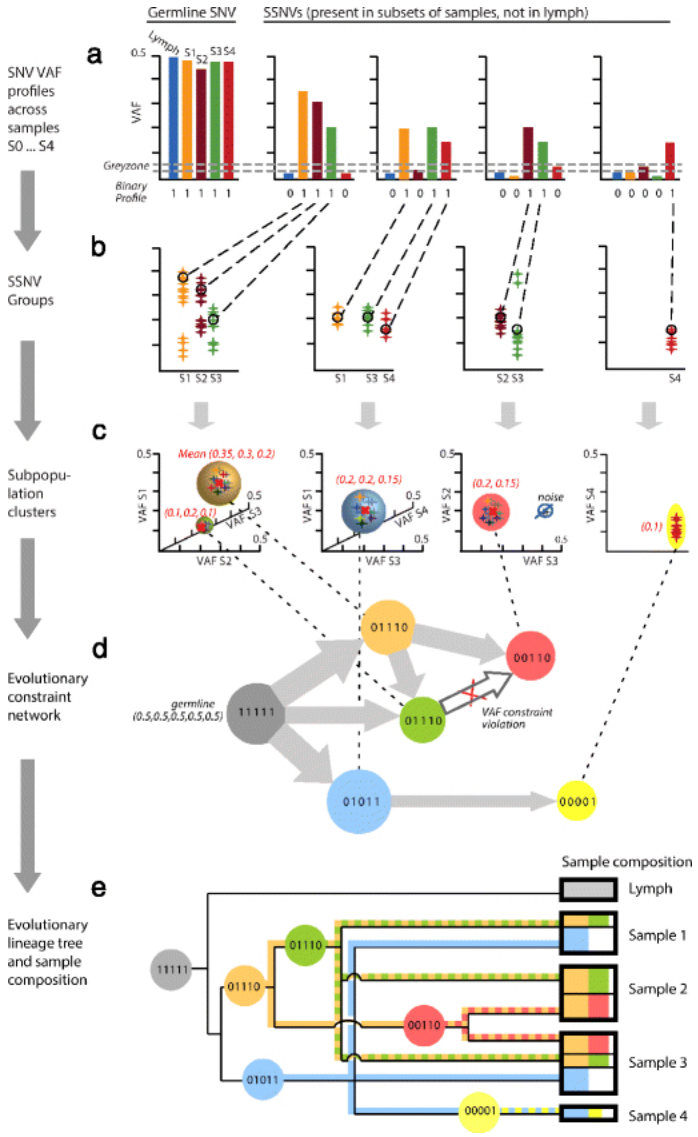

Figure 2: Steps of LICHeE algorithm.

- Example: IF M1 occurs w/ freq 0.2, and M2 occurs w/ freq 0.4, then M1 can't be parent of M2.
- However, if M1 occurs with freq of 0.5 and M2 occurs w/ freq of 0.4, then M1 could be a parent of M2. This means some cells have both M1 and M2, but some cells have just M1 and not M2.
- **NOTE**: Have to assume that M1 and M2 did not occur at the same location

- Constraint 3: Looking at 3 mutations– M1, M2, M3. If freq(M2+M3) ¿ freq (M1), then M1 not a parent of M2 and M3, because M1 would have to be contained in all cells with M2 and M3. However, since the sum of frequencies of M2 + M3 is higher than M1, there are some cells with M2 or M3 but not M1, so the frequency of the child would be higher than that of the parents, which isn ot allowed.
- However, M1 could be a parent to just M2 or M3.
- Because the pathways are divergent, a cell can't have both M2 and M3 if they both descended from M1.
- LiCHEE finds all lineage trees that satisfy the above 3 constraints.
- Reconstructing Evolutionary Constraint Network:
  - Construct a DAG where one each node of the graph represents a mutation or a group of mutations that occurs in the same subset of samples w the same frequency
  - Have edges to other nodes if constraints 1 and 2 are satisfied.
- Why is this a DAG? Why can't we have cycles?
  - We need to encode all possible precedence relationships between each pair of nodes
  - There are no cycles in the graph, as we cannot have freq(A) freq(B) and freq(B) freq(A).The edges in the graph must always point in the direction of lower frequencies, so once a path goes down, it cannot come back up to create a cycle.
  - The True lineage tree will be embedded in this network (subset of edges of this graph)
  - We assume that the ordering of the nodes does not violate the perfect phylogeny model, so the edges present in the graph are present in that order in the true lineage tree.
  - Finally, we search the DAG for all spanning trees that satisfy constraint 3 also
  - Spanning tree: has all nodes of our network (each node can have only 1 parent node– some subset of the edges)
  - **NOTE**: We always assume the root of our tree is a normal cell
- Overview of LiCHEE from start to finish:

  i Grouping/Clustering mutations by Presence across samples and VAF similarity
  - See **Figure 2a**
  - Binary Profiling: Is this frequency high enough?

- Assign 1 if frequency above a threshold, and designate as 0 if below the threshold (might be noise/frequency error).
- When deciding for mutations that are not clear: place mutations in a gray zone and allow points in 0 or 1 to "pull" points in the gray zone into a definite zone. This methodology is known as a mean-set cover formulation, and tries to minimize the total number of different profiles in the sample (heuristic algorithm).
- If mutation is present with high frequency in all samples, including lymph, designate that mutation as germline. This is because lymph tissue is normal (non-cancerous), so any variation in lymph is considered to be a germline mutation.
- Then Group each mutation according to its profile into SSNV groups.
- Cluster each group based on VAF, as in **Figure 2b and 2c**
- Don't know how many clusters a priori- cluster using Gaussian Mixture Models or some other algorithm
- Must allow for some buffer in clustering due to sequencing errors

ii Evolutionary Constraint Network Construction

- **Figure 2d**
- Each cluster is node in graph.
- Edge created if mutation's frequency means it can legally be child of another mutation (VAF of child is less than parent, with some error margin for noise).
- The tree created satisfies Constraint 1 and 2 only. All valid lineage trees are embedded in this graph.

iii Lineage Tree Searching and Ranking

- Goal to find a spanning tree in network that satisfies constraint 3– sum of frequencies of all children less than frequency of parents
- Looks for all the spanning trees first (Stops expanding the tree once Constraint #3 violated, and backtracks).
- Runtime proportional to number of spanning trees in the graph and the total size of the graph (O(trees+edges+nodes))
- Ranks trees by minimizing the total error score.
- Construct spanning tree of this graph with different samples as leaves.
- Illustration in **Figure 2e**
- Connect each sample to the last node on the path of mutation nodes present in the sample. In other words, the last node on all the paths leading to a particular sample.
- Each path to different destination is another subclone: 2 paths to Sample 1 means that Sample 1 has 2 different lineages
- Can calculate frequency of cells with a certain mutation by the frequency of last nodes

  - Note: These constraints don't declare only get 1 tree! Don't observe full tree- so some things could be moving around

- **Results**
  - Shows composition of samples into subclones
  - Can also dynamically update tree on fly and normalize differences
  - Data: CcRCC Study (Renal Cell Carcinoma Study)
  - In this study, researchers manually reconstructed the trees (used traditional phylogeny methods like Maximum Parsimony)
  - Trees LiCHEE Constructed were basically identical and generated automatically
  - Researchers pregenerated clusters for mutations and decided binary patterns (whether in dominant cluster or minor cluster)- not good strategy, because just because a mutation occurs w/ one frequency in one sample, doesn't mean it occurs with that frequency in all samples
  - Specialized cancer phylogeny means were able to find more heterogeneous mutations
  - LICHeE was also validated by simulating cancer evolution in silico; can sample from these generated trees to see what would expect in some tumor (in terms of number of mutations, frequencies, etc.) This is an additional way of validating our predicted results.