# Thesis Outline

Ethan Ashby

October 9 2020

## 1 Cancer: A Mutational Malady

This section is going to contain a summary of sources I've encountered on cancer genetics. It will predominantly focus on how mutation causes cancer, introduce somatic mutation, why study of mutation is important in the diagnosis and treatment of human cancer.
This section will outline the biological literature regarding the heterogeneity in mutational processes in cancer. I will spend significant time describing that frequencies and varieties of somatic mutation vary between cancer types.

## 2 The "Hidden Iceberg" of Rare Mutation and its application to cancer primary site classification

This section will consider mutational heterogeneity in the light of important clinical tasks. I will illustrate the importance of studying mutational heterogeneity through relevant clinical problems: the detection of cancer driver genes, clonal relatedness testing, liquid biopsy, and classifying cancers of unknown primary site. I will introduce this final project as my project of focus.
I will demonstrate that the preponderance of mutations are rare, and use this "hidden iceberg" of rare mutation to motivate the need to develop statistical methods to harness rare mutation data for clinical tasks in primary site classification.
This section will include review the intuition, proof, methodological details, and application of Smoothed Good-Turing Frequency Estimation to extract tissue-specific signals from the cancer genome.
This section will conclude by considering the limitations of previously research: chiefly, that only major cancer-associated genes were analyzed in this fashion.

# 3 Enhancing Tissue-Specific Signals and Harnessing More Mutation Data by Aggregation

I will introduce the dataset that I'm working with (TCGA non-hypermutated somatic variant mutation data from 6696 patients). I will also introduce these two problems:

- Extending this analysis to a larger fraction of the cancer exome
- Segmenting the genome to maximizing tissue specificity of mutation probabilities

I will pose thoughtful aggregation of mutation as a solution to both problems. By aggregating mutation data in gene groups, we can capture mutation from a larger fraction of the cancer exome, and enhance tissue specific signals of encoded by mutation probabilities.
I will end this section by posing the problem of how to thoughtfully aggregate mutations. I will say that we should aggregate mutation data in genes that show similar tissue-specific patterns of mutation.

# 4 An Overview of Statistical Divergence

In this section, I will review the literature related to statistical divergence. I will define statistical divergence, explain the intuition behind it, describe major families of divergence, and provide a survey of useful divergence measures.

# 5 Using Statistical Divergences to Measure Proximity between Cancer Genes

In this section, I will connect statistical divergence to the problem at hand. I will justify and select a suite of divergence/similarity measures to determine the proximity of cancer genes.
I will introduce my results, which will likely include a simulation study to compare the performance and select a/several divergence measures for further consideration. I will also apply these divergences to genes in TCGA data to generate a proximity matrix that I will display through heatmaps.

# 6 Clustering genes to maximize tissue specific signals

I will use the similarity matrix generated in the previous section and apply a (several?) clustering algorithms to actually generate gene groups with hopefully enriched tissue-specific signals. I will take care to carefully explain the

clustering approach. This section could potentially use a novel semi-supervised agglomerative hierarchical clustering approach that I concieved of to generate tissue-specific gene groups.

# 7 Conclusions

I will provide an overview of the problem, my approach, my results, and what the results of this project mean in relation to cancer genomics and the clinical task of cancer primary site classification.

# 8 The Tip of the Iceberg: Next Steps in Harnessing Rare Mutation for Primary Site Classification

I will describe next steps in this research. Chiefly application to both coding and noncoding regions in cancer using WGS data, and extending this analysis to mutation data generated from underrepresented racial/ethnic groups in genomics studies.

# 9 References

Watch out... There are gonna be a lot!