Background: Somatic mutations occur in non-sex-cells and are not passed to offspring. While these mutations do not directly affect the evolution of a population, they can cause diseases in an individual. Some cancers are caused by somatic mutations. For example, studies have shown a dramatic drop in lung cancer risk after smoking cessation[1], indicating that the management of somatic mutation is important for reducing cancer risk. Somatic mutations have been studied in the context of disease, but little to no work has been done on somatic mutation patterns in normal, non-pathogenic tissue in mammals. Without this knowledge, we are unable to fully appreciate whether cancers exhibit abnormal mutation patterns. Another important factor for understanding somatic mutations in healthy tissue is the type of genes mutated. When a gene is not expressed, its DNA may be supercoiled[2], which could offer protection from mutagens and degredation. Each cell of a tissue expresses different genes than cells of another tissue. It is currently unknown how these expression differences affect which genes are mutated in a tissue. I will investigate whether there are differences in the somatic mutation rate of different tissues.

Hypothesis 1: Mutation rates vary in different mammalian tissues, and tissues with higher renewal rates will have more mutations.

Methods: To address this hypothesis, I will work with mice labs at ASU to collect cerebellum, cerebral cortex, kidney, heart, lung, stomach, uteral, parotid gland, prostate, skin, and liver tissue samples in triplicate from each of several mice. These tissues largely represent epithelial tissues which have high renewal rates[3]. Most of these tissues have been successfully collected and used for cell fate mapping using a small subset of genomic markers[4] and should therefore be suitable for finding mutations and provide a sufficient number of mutations for analysis. Even if the sample size is insufficient for tissue-level specificity, it will be possible to pool similar tissues to improve statistical power.

I will do whole-genome Illumina sequencing on each of my samples then use a De Bruijn graph mutation caller called DiscoSNP++[5] to identify mutations in each of the samples. During DNA sequencing, the DNA is randomly shredded. The sequence of each small fragment, called a read, is generated. A De Bruijn graph is a way to put these reads back together to find the sequence of the original DNA. If multiple samples are used, it also finds places where these reads overlap except by one nucleotide. These locations are mutations or sequencing errors.

Most methods require comparison to a reference genome and will assume that if only a small portion of reads in a sample differ from the reference, it is an error and not a mutation. Since De Bruijn graphs don't use a reference genome for comparison, it reduces false negatives from samples containing low numbers of mutations. This is important, as I expect a fairly low number of mutants in each sample. This comes at a cost of an increased false positive rate. However, I will use the replicates I collected to reduce the number of false positives. I will then compare the number of mutations appearing in each tissue type to find which samples have higher mutation rates. I will also use a traditional reference-based approach to compare the De Bruijn graph and reference-based methods.

I have applied a similar method to 24 samples from a *Eucalyptus* tree to successfuly detect a large number of somatic mutations in the leaves of the plant. Further development of the software capable of performing this filtration will enable me to deploy it as a computational tool usable by any scientist doing mutation studies with replicates.

Anticipated Results: I expect that epithelial tissues will have higher mutational load than neuronal tissues. Since these cells have the highest renewal rates[3] and therefore undergo more cell divisions, they should have undergone more DNA duplications, providing more opportunities to obtain mutations. Thus, my hypothesis will be supported if I consistently observe higher numbers of mutations in epithelial tissues. Furthermore, it will be interesting to identify how the number of mutations change after normalizing by the average renewal rate of the tissue.

Hypothesis 2: The set of genes specific to a tissue will be more mutated in samples of that tissue than in samples of other tissues.

Methods: I will use the data obtained in the mutation rate variation experiment to identify the genes mutated in each tissue. As mutations will be rare, I will map the mutations found to 10 megabase windows on the mouse reference genome. Then, I will identify the genes in these windows.

The gene ontology is a controlled vocabulary used to describe the functions, associations, and cellular location of a particular gene product. I will associate each window with the gene ontology terms of the genes in that window and use a clustering algorithm to determine if the windows with mutations are annotated as specific to the tissue of the sample or not.To rule out the effect of gene function, I will also see if genes that perform a class of functions are disproportionally mutated.

Anticipated Results: My hypothesis is supported if I observe that windows containing genes specific to a particular tissue have more mutations in that tissue than in other tissues. Patterns in the function of genes mutated may help explain these differences.

Intellectual Merit: How rates of somatic mutation differ between different tissues is poorly understood. However, diseases such as cancer often arise by somatic mutation. This project represents a first step in more clearly understanding somatic mutation, how mutation rates vary within an individual, and how this variation contributes to variation in cancer incidence.

Broader Impacts: Understanding the link between background somatic mutation rate and cancer incidence can be clinically useful for diagnosing, preventing, and treating the disease. This knowledge will help clarify theories of the evolution and maintenance of multicellularity. I will also generate a tissue phylogeny, useful for developmental biologists. Additionally, all data generated by the project will be deposited in a repository for public use and all computational analyses and tools developed for the project will be open source and freely available for modification and use. This will aid others who attempt to pursue similar analyses or wish to develop similar tools. This project will give me the opportunity to train an undergraduate data scientist to understand how to hand and analyze large amounts of data and develop computational tools for the biology community. I will also discuss this project at a future Night of the Open Door event to educate adults and children on somatic mutation and its relation to cancer. This will help people understand the biology behind why smoking cessation reduces cancer risk. I will additionally be better able to give insight on mutation to primary and secondary school students through the Ask a Biologist program, where these students submit questions to be answered by an expert.

References: [1]Miller Y.E *Am J Respir Cell Mol Biol* (2005) [2]Gilbert N., Allan J., *Curr Opin Genet Dev* (2014) [3]Frank S.A. *Dynamics of Cancer: Incidence, Inheritance, and Evolution* (2007) [4]Salipante S.J., Horwitz M.S., *Proc Natl Acad Sci USA,* (2006) [5]Uricaru R. *et al. Nucleic Acids Res,* (2015)