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. For example, somatic mutations are thought to cause many forms of cancer. However, little to no work has been done on normal somatic mutation patterns in mammals. Without this knowledge, we are unable to fully appreciate the extent that cancers exhibit abnormal mutation patterns. Thus, there is need for basic research on somatic mutation in normal tissues.

Cells of different tissues have different properties. One of these properties that could significantly impact the patterns of somatic mutation is the tissue renewal rate, as cells that divide more frequently should have more mutations due to replication errors. Additionally, each cell of a tissue express different genes than cells of another tissue. When a gene is not expressed, its DNA is sometimes methylated, which can cause mutations. Furthermore, cells are known to preferentially repair segments of DNA that are actively transcribed. This means that different cells from different tissue types should exhibit different patterns of mutation. However, the extent of these differences is currently unknown. I will investigate how these properties of different tissues affect the rate and distribution of mutations in the genome using a mouse model.

Hypothesis 1: Mutation rates vary in different 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 represent each of the germ layers and contain many epithelial tissues, which have high renewal rates[1]. Most of these tissues have been successfully collected and used for cell fate mapping using a small subset of genomic markers[2] and should therefore be suitable for finding mutations and provide a sufficient number of mutations for analysis.

I will do whole-genome Illumina sequencing on each of my samples then use the Genome Analysis Toolkit (GATK) to identify potential mutations. This works by comparing the obtained sequencing data and comparing it to the mouse reference genome. However, as sequencing errors are common, the software must determine if observed differences are true mutations or sequencing errors. Most variant callers are designed to identify germline mutations, and most designed for use on somatic tissue assume the sample is from a cancer tumor. This may impact the quality of mutations identified by the GATK.

Therefore, I will also use the De Bruijn graph mutation caller DiscoSNP++[3] to identify mutations. Rather than comparing to the reference genome, a De Bruijn graph mutation caller finds places where the sequences of different samples overlap except by one nucleotide, which is the location of the putative mutation. By comparing the results of these two distinct methods along with the results across replicates, I will filter out false positives.

To further improve the mutation calls, I will develop a method to jointly determine the relationships of the tissues and detect variants. I will use the high-confidence mutations obtained by the methods above to determine the developmental relationships of the tissues, then use these relationships to remove putative mutants that are unlikely to be real because they conflict with these relationships. I will then have a set of high-quality mutations that I will use to find which tissues have higher mutation rates. I have partially implemented such a method to detect somatic mutations in the leaves of a *Eucalyptus* tree, where the

relationships between the leaves are apparent. Further development of this method will enable me to distribute it as a resource usable by anyone studying somatic mutation.

Anticipated Results: Epithelial cells have the highest renewal rates[1] and therefore undergo more cell divisions, so they should have undergone more cycles of DNA replication. This provides more opportunities for mutations to be introduced. Therefore, I expect these tissues to have a higher mutational load than the other tissues. Thus, my hypothesis will be supported if I consistently observe higher numbers of mutations in epithelial tissues.

Hypothesis 2: The set of genes specifically expressed in one tissue will be less 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 described above 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 less than is expected by chance. To rule out the effect of gene function, I will also determine whether genes that perform a class of functions are disproportionally mutated.

Anticipated Results: Methylation is an important factor for suppressing gene expression, but sometimes causes mutations. Simultaneously, transcription-coupled repair enhances the rate of DNA repair in highly-expressed regions[4]. Therefore, in samples of a particular tissue I expect to see fewer somatic mutations in windows containing genes specifically expressed in that tissue. If this is the case, my hypothesis will be supported.

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 in normal tissue, how mutation rates vary within an individual, and how gene expression does or does not affect somatic mutation.

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 experiment will provide a comparative tool to greatly aid our understanding of how cancer deviates from normal patterns of somatic mutation. It will also be useful for developmental biologists to understand the genetic relationships between different tissues. All data generated by the project will be deposited in a repository for public use and all computational tools developed for the project will be open source and freely available for modification and use by others. This project will give me the opportunity to train an undergraduate data scientist to understand how to use and develop bioinformatic tools. I will also discuss this project at a future ASU Night of the Open Door outreach event to educate adults and children on somatic mutation and its relation to cancer. I will additionally be better able to give insight on mutation to secondary school students through ASU's Ask a Biologist program, where these students submit questions to be answered by an expert.

References: [1]Frank S.A. *Dynamics of Cancer: Incidence, Inheritance, and Evolution* (2007) [2]Salipante S.J., Horwitz M.S., *Proc Natl Acad Sci USA,* (2006) [3]Uricaru R. *et al. Nucleic Acids Res,* (2015) [4]Hanawalt P.C., Spivak G., *Nat. Rev. Mol. Cell Biol,* (2008)