

## **ABSTRACT**

Mutations-spontaneous changes in DNA sequences-are fundamental to evolution, and contribute to everything from disease susceptibility in people to agriculture.

Population geneticists are tasked with understanding mutations - historically categorizing them as either beneficial, harmful, or neutral based on how they affect fitness. However, our understanding of these categories is often limited by incomplete data and an inability to travel back in time to track mutations. Over four chapters, I will explore how large datasets can enhance our understanding of these three fundamental mutation categories. First, I will use 27 terabases of gene expression data from 300 studies to investigate whether rarely expressed genes accumulate harmful mutations at higher rates than constitutively expressed genes. Next, I will leverage about 205 terabases of DNA-sequencing data and use k-mers (DNA subsequences) to measure neutral variation in 112 natural plant species. The results suggest that current methods for estimating diversity reliant on reference genomes underestimate genetic variation. The following chapter expands upon the idea of using k-mers to measure genetic diversity, numerically and analytically exploring the relationship between k-mer diversity and nucleotide diversity. Overall, k-mer diversity scales linearly with nucleotide diversity and we showcase the use of bloom filters to decrease the memory burden of the k-mer diversity calculations. Finally, I will use 250,000 evolutionary simulations to train machine learning models to infer the time it took for a fixed beneficial mutation to spread. Overall, large datasets like these hold many opportunities to revisit old biological questions and further our understanding of mutation trajectories.