**Introduction**

Polymorphism involves one of two or more variants of a particular DNA sequence. The most common type of polymorphism involves variation at a single base pair. Polymorphisms can also be much larger in size and involve long stretches of DNA. Called a single nucleotide polymorphism, or SNP (pronounced snip), scientists are studying how SNPs in the human genome correlate with disease, drug response, and other phenotypes.

Detecting polymorphisms in the genome is an important task for an individual specimen (disease association studies) and for a species as whole (phylogenetic tree reconstruction). Whether it be identifying single nucleotide polymorphisms (SNPs) in an individual compared to a reference genome or comparing different species, identifying polymorphic differences is a difficult task. Methods, however, are usually extremely conservative and only identify simple variation (SNPs, insertions, deletions) leaving more complex variation (translocations, inversions) unexamined.

Genomic variation such as translocations and inversions have been shown to cause many human diseases. Translocations have been shown to be the cause of several different types of cancer, such as Burkitt’s lymphoma and acute promyelocytic leukaemia . They have also been shown to be associated with schizophrenia . Studying and identifying different types of genomic polymorphisms could have impact on two very important fields in biology: genome wide association studies as well as phylogenetic tree reconstruction.

**Genome Wide Association Studies**

In genomics, a genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. GWA studies typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases but can equally be applied to any other genetic variants and any other organisms.

In GWAS, next-generation sequence (NGS) reads are mapped to a reference genome. Differences, commonly SNPs and indels, are then identified from the read mapping results. This method has helped identify and associate many mutations with different diseases. Read mapping, however, is a difficult task. More than 10% of reads were unmapped when mapping 12.2 million reads to the human genome using the popular Burrows-Wheeler Aligner . Some of the reads will be left unmapped due to errors generated during sequencing. Other reads are left unmapped for unknown reasons. It may be that some unmapped reads vary significantly from the reference genome making read mapping difficult. Mapped reads represent reads that are similar enough to the reference genome to be mapped with a given set of parameters. Unmapped reads may contain more interesting and novel biological information than mapped reads because these reads diverge enough from the reference genome to re- main unmapped. Harnessing unmapped reads enables more thorough analysis of how individuals within a species differ and how genomic rearrangements may affect phenotypes.

**Phylogenetic Tree Reconstruction**

Phylogenetic tree reconstruction is usually done by comparing homologous gene sequences in a group of species of interest. Identifying the homologous genes is a difficult task and is often a conservative process, allowing for only gene sequences that are very similar to be clustered together. This approach is limited because it only allows for comparing gene sequences instead of comparing whole genomes . Comparing the entire genome of one species to  
another is valuable to see if genomic rearrangements or other structural variations occurred to the genome. Accounting for these genomic variations may serve as a future phylogenetic signal in future phylogenetic tree reconstruction.

**Methods**

For utilizing unmapped reads and to compare whole genomes is to construct a relaxed de Bruijn graph that allows for more complex genomic variation to be observable.

**Standard de Bruijn Graph**

A standard de Bruijn graph is a graph structure that represents the genome of an organism. de Bruijn graphs are commonly used for genome assembly and usually representative of a single species .Beyond genome assembly, they have also been found to increase the percent mapped reads when mapping reads to a de Bruijn graph versus contigs. In a de Bruijn graph, each node represents a unique kmer. Edges in the graph represent kmer overlaps. The graph is usually constructed from NGS reads where reads are broken into kmers and used to populate the graph.

In graph theory, the standard de Bruijn graph is the graph obtained by taking all strings over any finite alphabet of length ℓℓ as vertices and adding edges between vertices that have an overlap of ℓ−1ℓ−1. In the following, we consider assembly using a slightly modified version of the standard de Bruijn graph from the L-spectrum of a genome.

**Relaxed de Bruijn Graph**

Our relaxed de Bruijn graph differs from a standard de Bruijn Graph is two major ways:

1) The graph contains sequence information for multiple species

2) Kmers can occur multiple times in the graph

By relaxing these constraints on the de Bruijn graph, we are able to identify interesting genomic variation in a tractable amount of time and space. Conceptually, this method can be thought of as merging two separate deBruijn graphs by exploiting uniquely occurring kmers in one sequence as anchor points to merge the graphs.