Genome Polymorphism Detection Through Relaxed de Brujin Graph Construction

THE TEAM

GROUP - 13

G.Venkata Vignesh

AM.EN.U4AIE21074

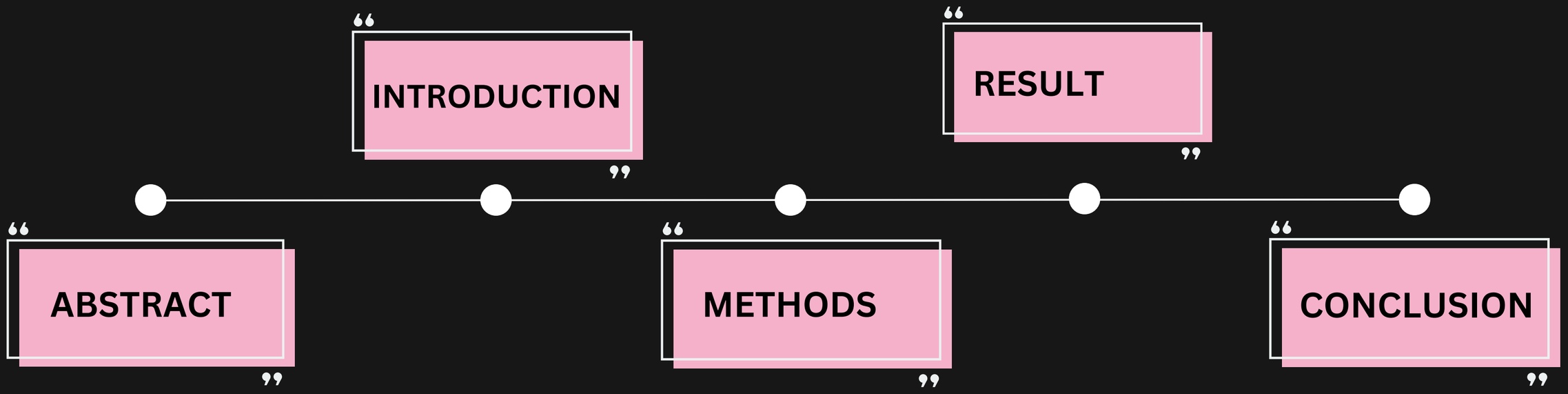
V.Praneeth

AM.EN.U4AIE21066

C.Guru Aswini Dath

AM.EN.U4AIE21023

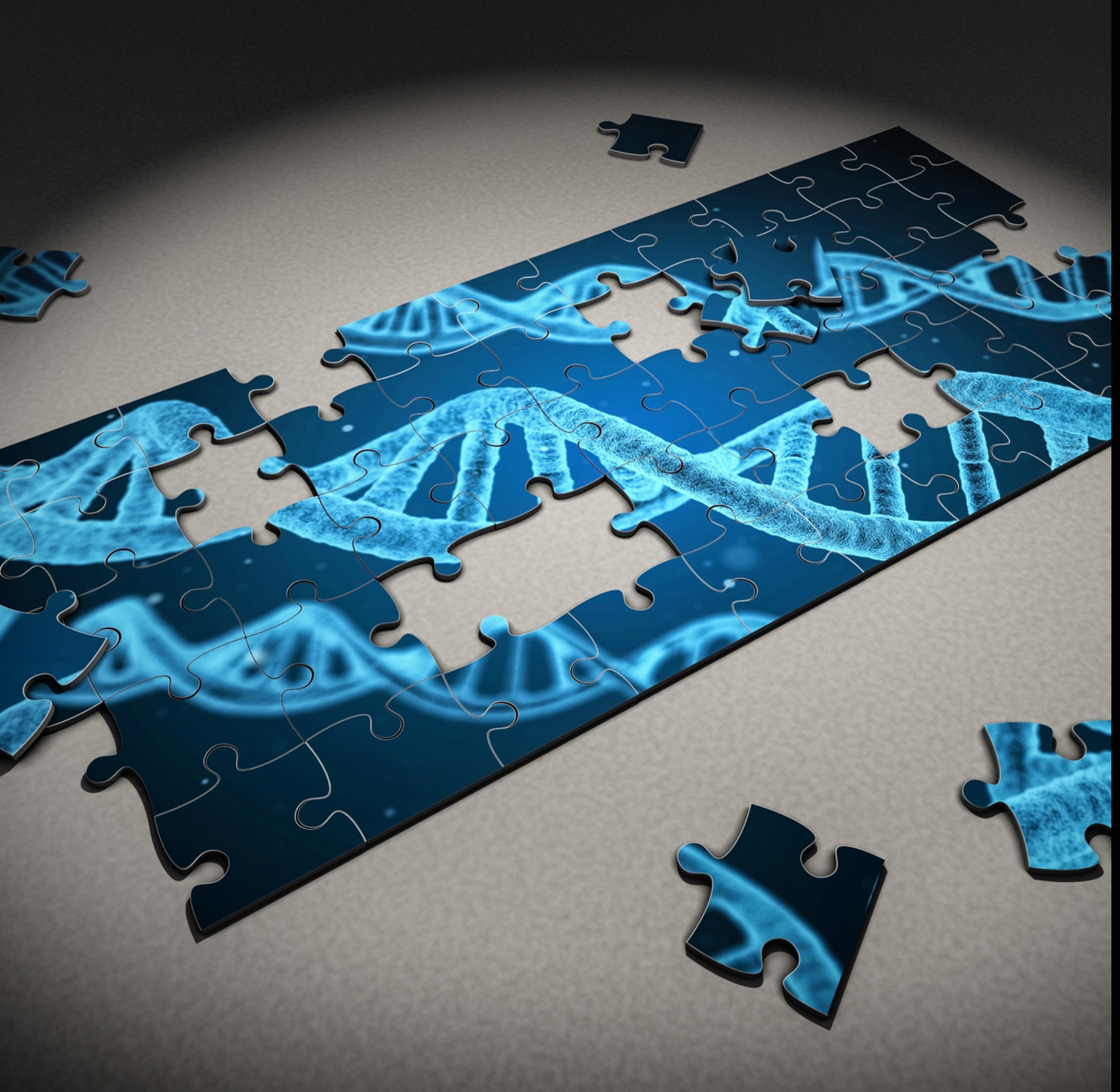
TIMELINE



ABSTRACT

- Comparing genomes to identify polymorphisms is a difficult task, especially beyond single nucleotide polymorphisms.
- Polymorphism detection is important in diseaseassociation studies as well as in phylogenetic tree reconstruction.
- We present a method for identifying polymorphisms in genomes by using a modified version de bruijin graphs.
- Now a days, data structures are widely used in genome assembly from Next Generation Sequencing.

INTRODUCTION

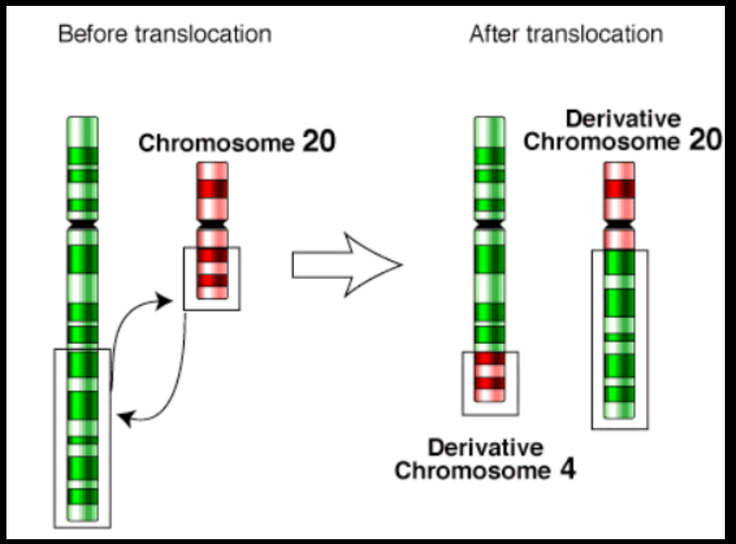


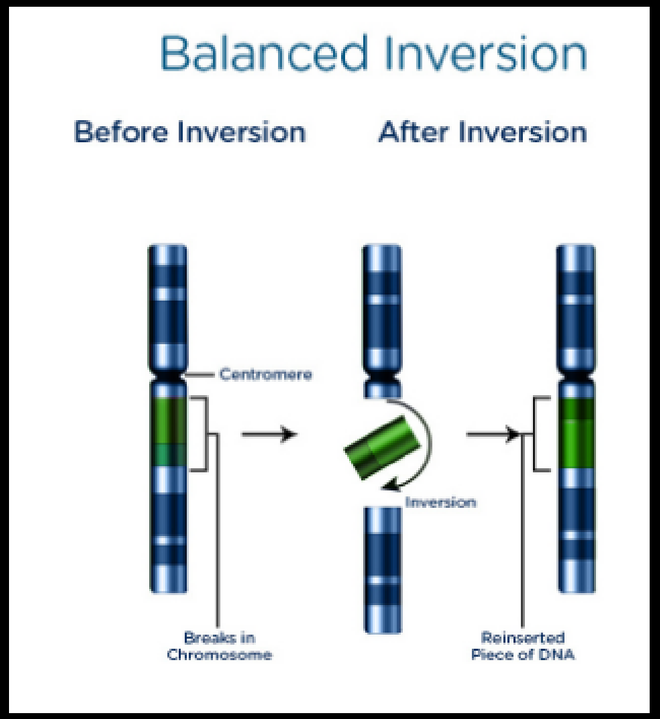
- Genomic variation such as translocations and inversions have been shown to cause many human diseases.

- Translocations have been shown to be the cause of several different types of cancer, such as Burkitt’s lymphoma and acute promyelocytic leukemia.

TRANSLOCATION:

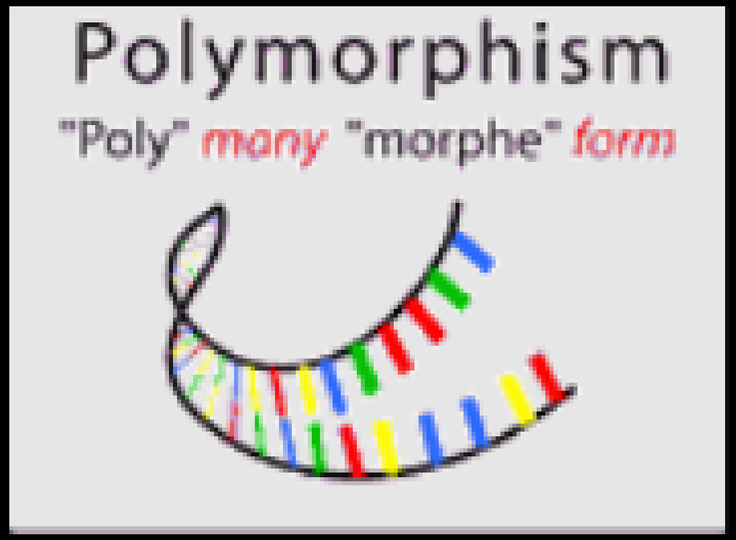
A genetic change in which a piece of one chromosome breaks off and attaches to another chromosome.





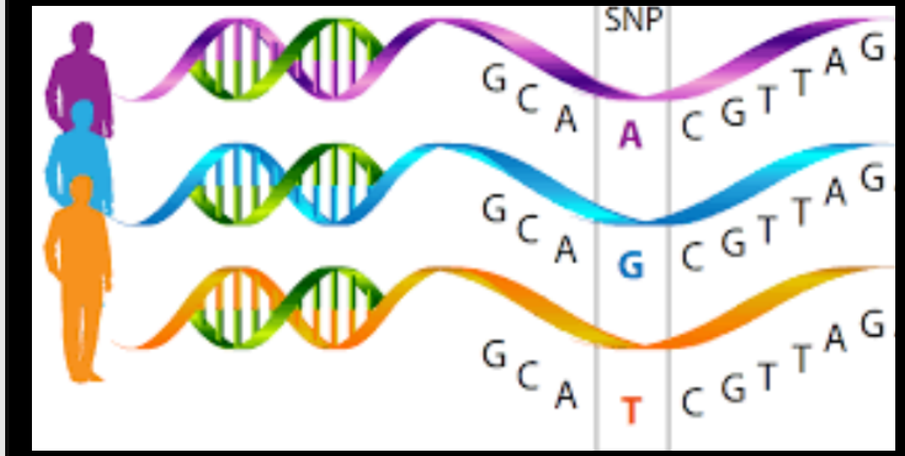
INVERSIONS:

An inversion is a chromosome rearrangement in which a segment of a chromosome becomes inverted within its original position.



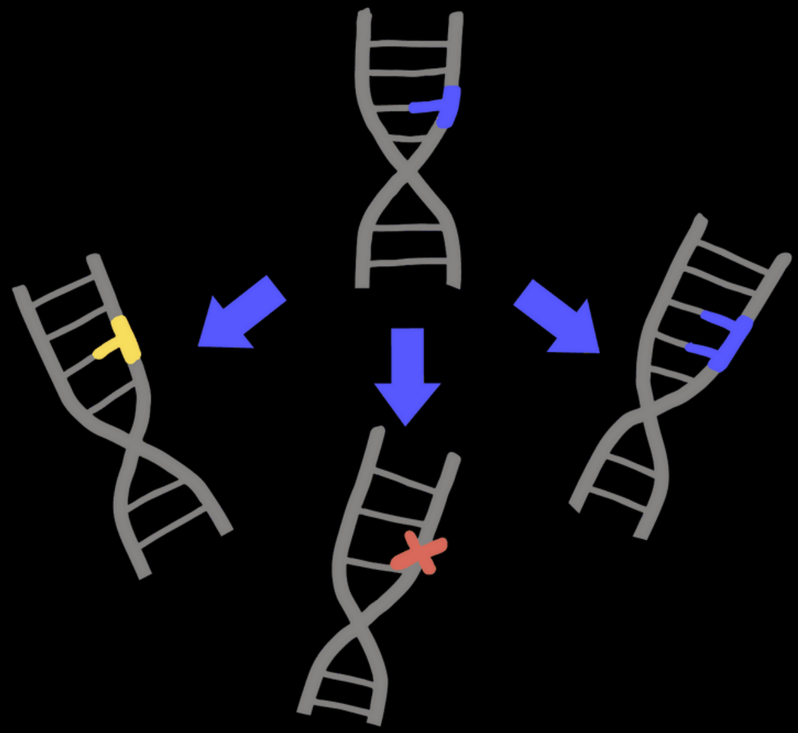
POLYMORPHISM

It is the presence of two or more variant forms of a specific DNA sequence that can occur among different individuals or populations.



SINGLE NUCLEOTIDE POLYMORPHISMS

- DNA sequence variation that occur when a single nucleotide differs between members of biological species.
- This may be subtitution,insertion or deletion.



Mutation

It is defined as varation in the DNA sequece leading to delection,insertion,Duplication,substit ution of an nuceliotes.

GENOMIC POLYMORPHISMS

- GENOME WIDE ASSOCIATION STUDIES (GWAS)
- PHYLOGENETIC TREE RECONSTRUCTION

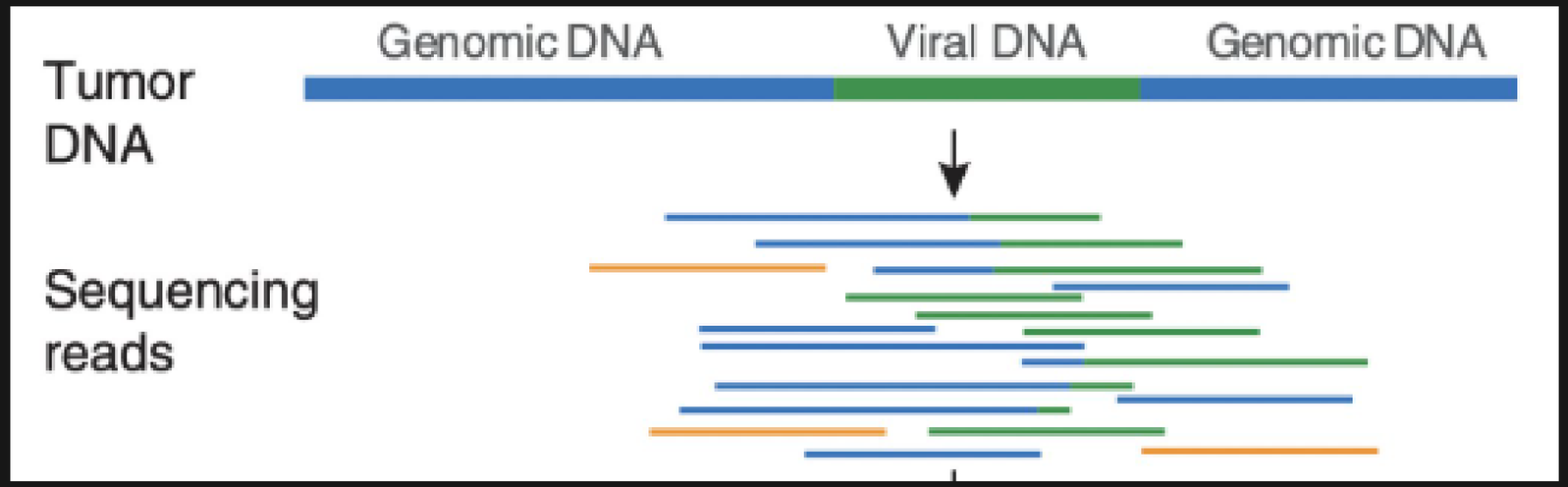
GWAS

- It is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait.

- More than 10% of reads were unmapped when mapping 12.2 million reads to the human genome using the popular Burrows-Wheeler Aligner.

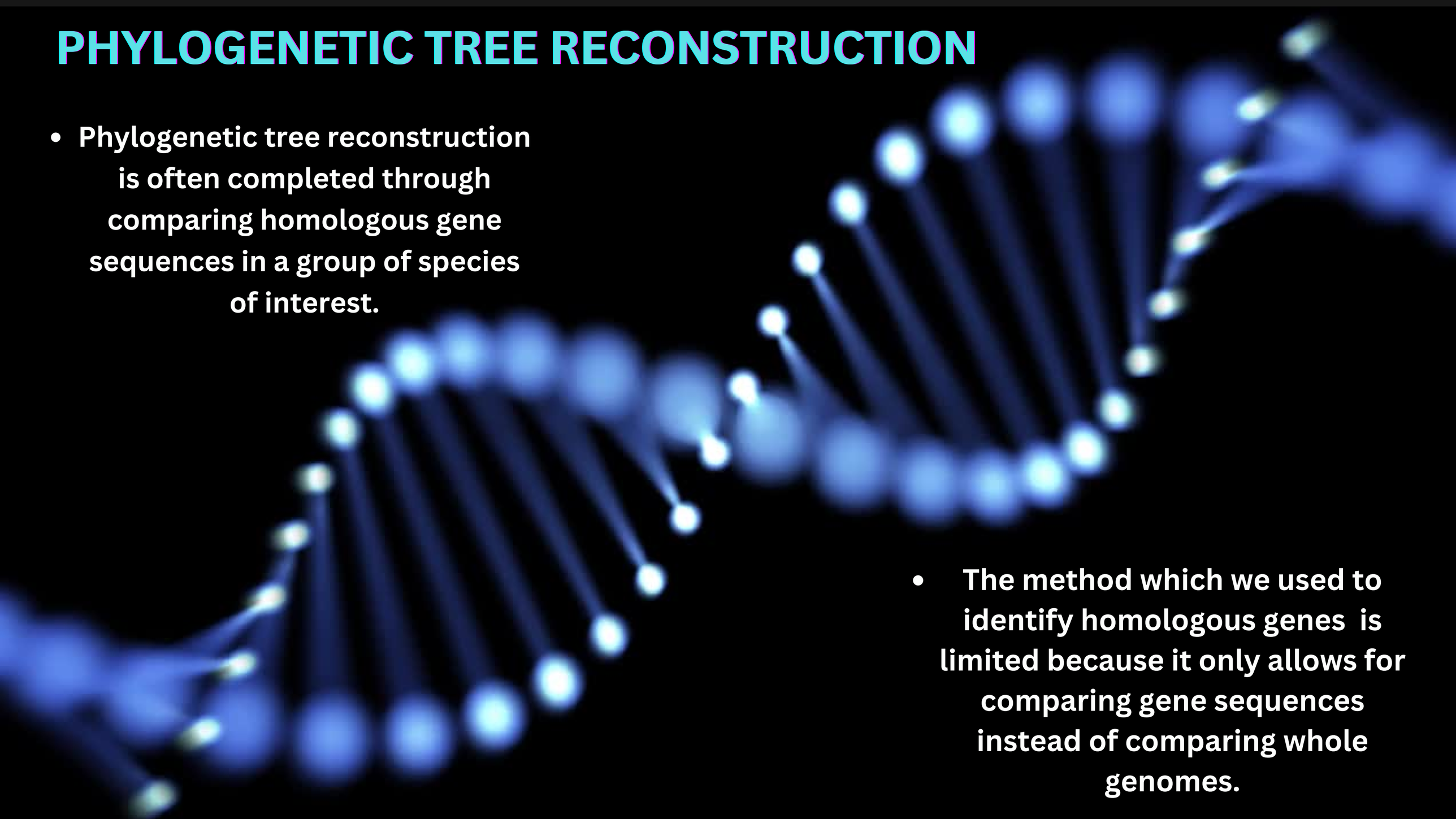
- Some of the reads will be left unmapped due to errors generated during sequencing. Other reads are left unmapped for unknown reasons. It may be that some unmapped reads vary significantly

- Unmapped reads may contain more interesting and novel biological information than mapped reads because these reads diverge enough from the reference genome to remain unmapped



PHYLOGENETIC TREE RECONSTRUCTION

- Phylogenetic tree reconstruction is often completed through comparing homologous gene sequences in a group of species of interest.



- The method which we used to identify homologous genes is limited because it only allows for comparing gene sequences instead of comparing whole genomes.

METHODS

De Brujin graph

Relaxed de Bruijn Graph

DE BRUJIN GRAPH

- De Bruijn graphs are usually representative of a single species and are commonly used for genome assembly.

- 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 .

Relaxed de Bruijn Graph

- The graph contains sequence information for multiple species.
- Kmers can occur multiple times in the grapH.
- relaxing these constraints on the de Bruijn graph, we are able to identify interesting genomic variation in a tractable amount of time and space.

Algorithm:

De Bruijn graph:

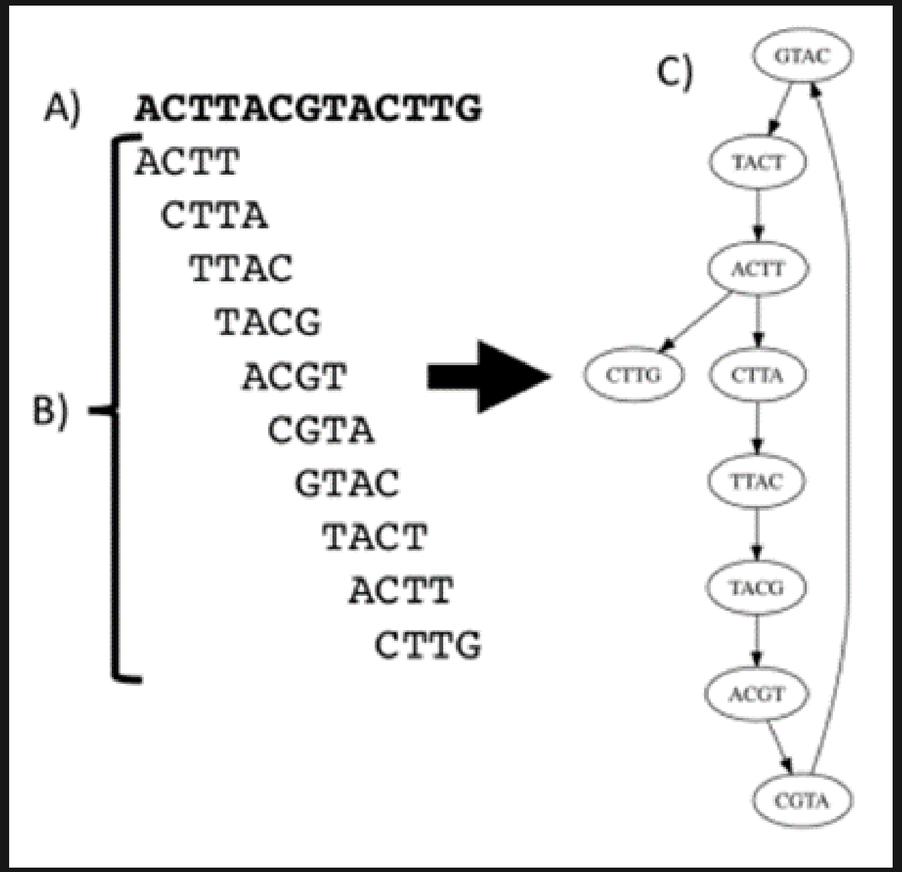
Code:

```
def Construct(seq,k):
    for i in range(0,len(seq)-k+1):
        kmer = seq[i:i+k]
        kmerk.append(kmer)
    return kmerk

seq=input("Enter the dna sequencing")
k=int(input("Enter the k value"))
Construct(seq,k)

for i in kmerk:
    su=i[-(k-1):]
    pre=i[:(k-1)]
    suffix.append(su)
    prefix.append(pre)
    graph.add_edge(su,pre)
```

Output:



Relaxed de Bruijn Graph:

Algorithm 1:

Algorithm 1 Initial relaxed de Bruijn graph construction.

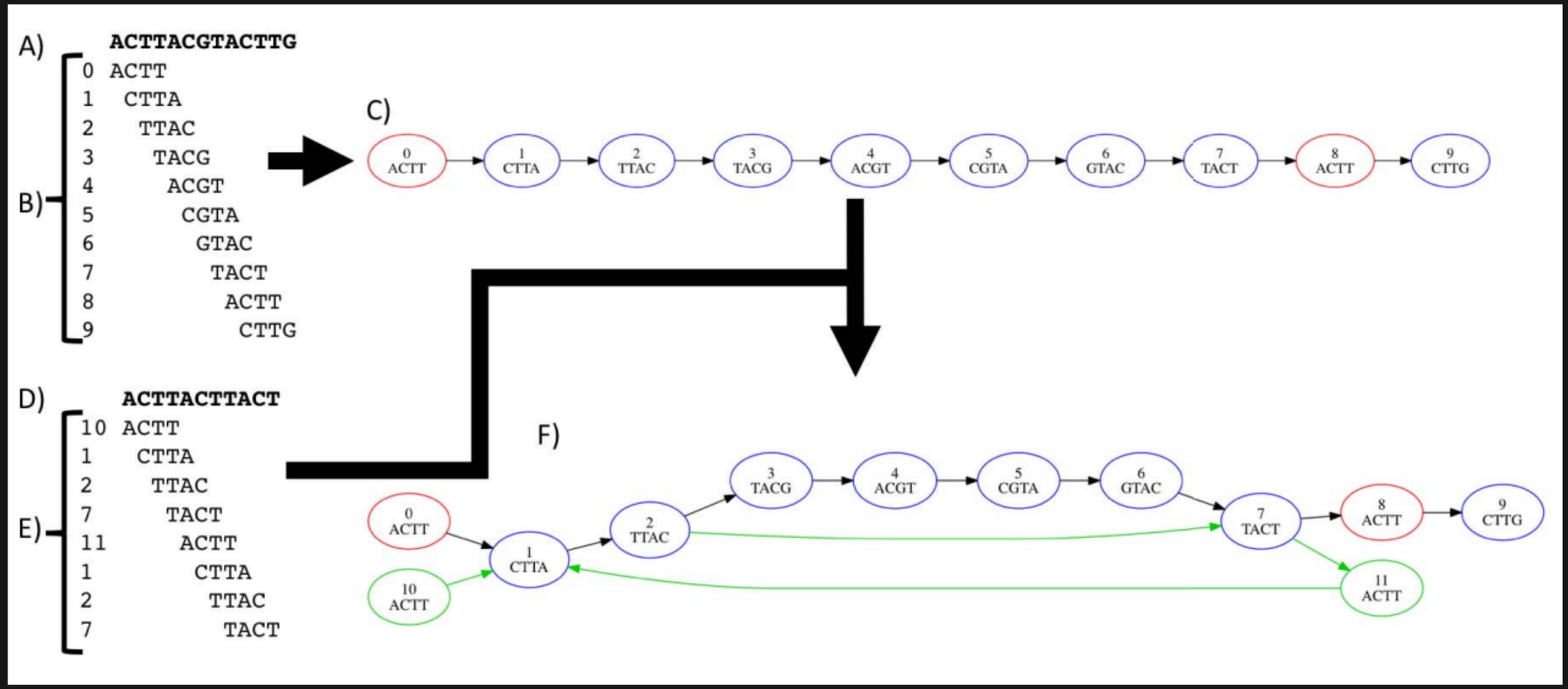
```
1:  procedure CONSTRUCT(seq, k)
2:      Input: DNA sequence seq, kmer length k
3:      Output: kmer counts occs,
                index counter curidx,
                relaxed de Bruijn graph g
                kmer-index reverse lookup table rlookup
4:      occs ← occurrences of each kmer
5:      curidx ← 0
6:      g ← an empty graph
7:      for each kmer kmer in seq do
8:          l ← prefix of kmer
9:          lidx ← curidx
10:         curidx ← curidx + 1
11:         occs[l] ← occrs[l] + 1
12:         rlookup[l] ← lidx
13:         r ← suffix of kmer
14:         ridx ← curidx
15:         curidx ← curidx + 1
16:         occs[r] ← occrs[r] + 1
17:         rlookup[r] ← ridx
18:         g.addedge(lidx, ridx)
19:     end for
20:     return occs, curidx, g
21: end procedure
```

Algorithm 2:

Algorithm 2 Appending new sequences after initial graph construction.

```
1:  procedure APPEND(g, occs, curidx, seq, k, rlookup)
2:      Input: initialized relaxed de Bruijn graph g
               kmer occurrence counter occs
               index counter curidx
               DNA sequence seq
               kmer length k
               kmer-index reverse lookup table rlookup
3:      for each kmer kmer in seq do
4:          l ← prefix of kmer
5:          if occs[l] == 1 then
6:              lidx ← rlookup[l]
7:          else
8:              lidx ← curidx
9:              curidx ← curidx + 1
10:         end if
11:         r ← suffix of kmer
12:         if occs[r] == 1 then
13:             lidx ← rlookup[r]
14:         else
15:             ridx ← curidx
16:             curidx ← curidx + 1
17:         end if
18:         g.addedge(lidx, ridx)
19:     end for
20: end procedure
```

Observation:



RESULT:

- In the simplified graph, the mutation, insertion and inversion form simple bubble structures in the graph while the translocation forms a much more complex structure.

- In these very ideal conditions the generated graph shows structures that could be used to generated a phylogenetic signal or for phenotype association with additional generated graphs from other individuals.

conclusion

- In this work, we have presented a method for constructing a unified, relaxed de Bruijn graph that contains more than one sequence source. The relaxed de Bruijn graph enables identification of graph structures that may be used as a signal for phylogenetic tree reconstruction or for use in association studies for phenotypes.