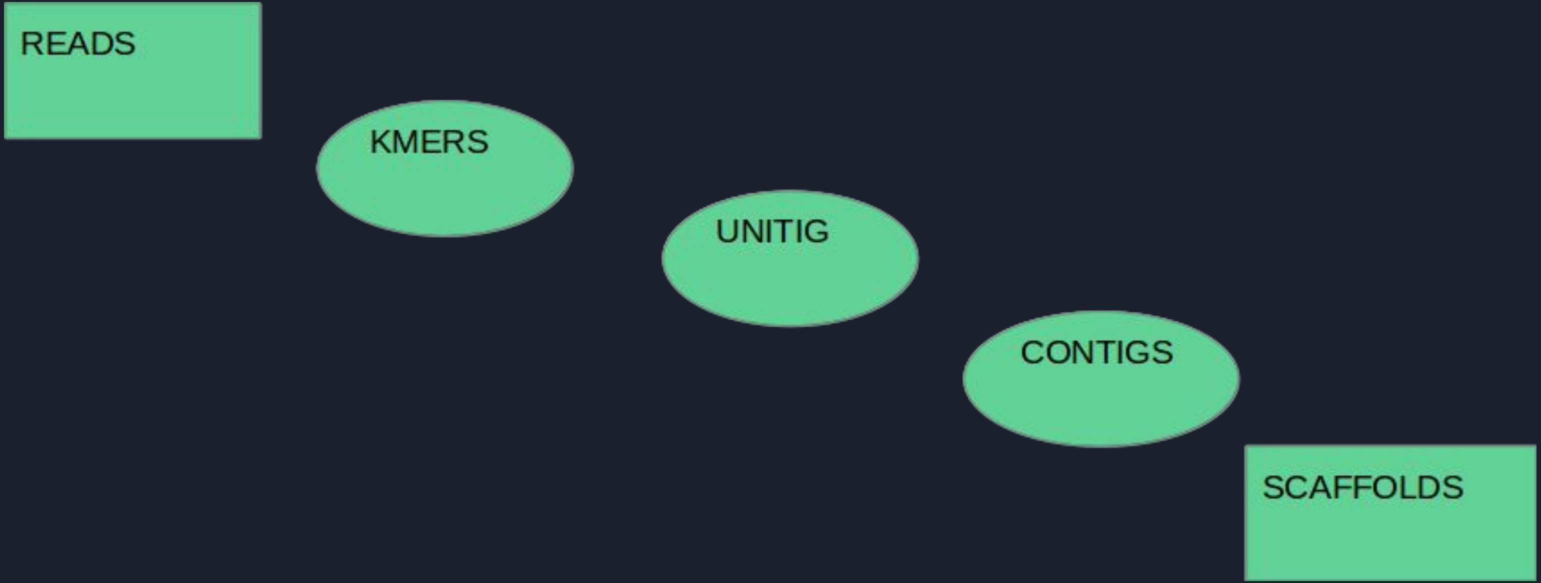
Genome Assembler

Harshita Goyal 2014B5A70779P

Assembly Process



READS

KMERS

UNITIG

CONTIGS

SCAFFOLDS

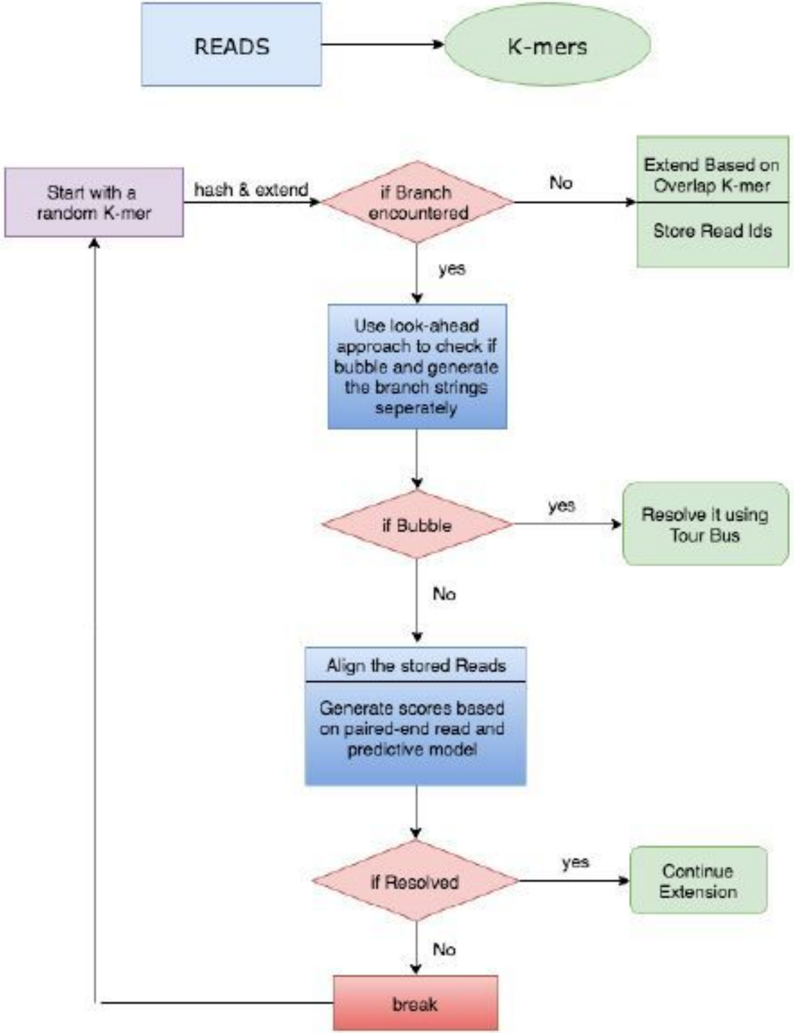
Algorithm

- Preprocessing :
    - Correct or eliminate erroneous reads.
    - Distributing data(reads).
- Parallel reads
    - Each node will construct K-mers .
    - M-mers will be extracted from the k-mers and binning based on the extracted m-mers will be done.
- Synchronization
    - Nodes will communicate (broadcast) to merge , balance and distribute the bins.
    - An index of which nodes have which bins will be present in every node.
- Unitig formation
    - Maximal length contigs within each m-mer bin and node is done.
- Branch resolution
    - While contig formation , if the m-mer bin changes ; this information will be stored in a buffer and batch communication will be done.
    - Contigs will be stored in union- find structure

M-mer formation

CGTTGATCAATTTG	Read
CGTTGATC	M-mer : rev_ comp (CGTT) = AACG
GTTGATCA	M-mer : rev_ comp (TGAT) = ATCA
GATCAATT	M-mer : AATT
ATCAATTT	M-mer : rev_ comp (ATTT) = AAAT

Unitig to Contig


READS
K-mers
Start with a random K-mer
hash & extend
if Branch encountered
No
Extend Based on Overlap K-mer
Store Read Ids
yes
Use look-ahead approach to check if bubble and generate the branch strings seperately
if Bubble
yes
Resolve it using Tour Bus
No
Align the stored Reads
Generate scores based on paired-end read and predictive model
if Resolved
yes
Continue Extension
No
break

Branch Resolution

- From the unitig terminal k-mer, the next possible signatures can be found.
- Such queries can be kept in a buffer (per node) to optimize communication cost.
- Paired end and ML model will be used for branch resolution.
- A union- find structure will be maintained in all nodes. It will store information regarding all merges in all the nodes.

Features

Example: 9 contigs with the lengths 2,3,4,5,6,7,8,9,and 10 ; sum = 54 ; half of the sum = 27, and the size of the genome also happens to be 54. 50% of this assembly.

N50: the sequence length of the shortest contig at 50% of the total genome length.

Eg. 10 + 9 + 8 = 27 (half the length of the sequence). Therefore, N50=8.

L50 : smallest number of contigs whose length sum makes up half of genome size.

Eg. L50 = 3

N90: the sequence length of the shortest contig at 90% of the total genome length.

Eg. 10 + 9 + 8 +7+6+5+4= 49 . Therefore, N90=4 .

Additional Features

- Paired end information
- GC bias
- Repetition features
- NGXX values

Current problems

- Whether the local unitigs will produce good enough assemblies?
- How to distribute the bins (with load balancing) , without global information of binning’?
- Whether batch union-find updates are possible?
- Repeats have not been handled till now.

Inadequacies in the model

- Erroneous reads are not handled.
- Scaffolding is not being done.
- Bubble resolution will be done by the standard approach.
- K-mer and m-mer size is fixed.