**Problem Description**

- Genome assembly is a fundamental problem in the field of bioinformatics.
- The goal lies in reconstructing an unknown genome from short DNA fragments obtained from it.
- With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours leading to TB/day data accumulation.
- Research focus: Implement new scalable methods for performing extreme-scale genome assembly suited for microbial genomes and other complex eukaryotic genomes.

**De Novo Genome Assembly**

**Input:** Reads that may contain errors

- Extract k-mers (ks3): chop reads into k-mers, process to remove errors

- De Bruijn graph (DBG): construct graph where, nodes=k-mers, edges= (k-1)-mer overlap.

**Output:**

- Traverse graph to enumerate long contiguous genomic regions or contigs

**K-mer Frequency**

Frequency of k-mers in read datasets depends on:

a) Sequencing coverage (C)  

b) Error rate

**Speedup of FastEtch and Bi-FastEtch on C. elegans (50x, k=32)**

**FastEtch: Experiments and Results**

- Experiments conducted on a single 128GB DDR4 memory node of NERSC Cori

**Distributed-memory method:** PaKman (In Progress)

**Objectives:**

- Implementation of a fully-distributed (MPI + OpenMP) extreme-scale genome assembler aiming to tackle the assembly of large genomes

**Essential Contributions:**

1) Handle I/O overheads at the time of reading and distributing reads dataset  
2) Load balance the distribution of k-mers  
3) Novel data-structure to minimize inter-process communication at the time of contig generation  
4) Demonstration on shared and distributed memory systems

**PaKman preliminaries**

- Strong scaling results on NERSC Cori for dataset: C.elegans (100x), size=11GB, with # of distinct k-mers: 1,585,416,564

**References**


**Acknowledgements**

This research was supported in part by U.S. DOE grant DE-SC-0006516 and U.S. DOE under Contract No. DE-AC02-05CH11231.