Project 3 (Medium):

Sequencing with Insertions

Rory Snively

The Problem

As if re-sequencing a genome weren't hard enough…

What you see:

AATCGGTCGTA GGCTGATGCTAGCTGATTCG

AATCGGTCGTACGGCTGATGCTAGCTGATTCG

What your computer sees:

AATCGGTCGTAGGCTGATGCTAGCTGATTCG

AATCGGTCGTACGGCTGATGCTAGCTGATTCG

What Does It All Mean?

- Some benchmarks:
  - N base pairs in the genome
  - M reads of length L
  - Up to T insertions per read

  A naïve solution would be O[M * N * (L ^ T)]

Baseline Method

- One base pair at a time…
  - With a 2 million base pair genome, each read takes on average 5.5 seconds to sequence.
  - Assuming linear growth (very generous), implies about 90 seconds per read on full-sized genome.
    - (That's almost 300 years to sequence the genome – with only single coverage!)

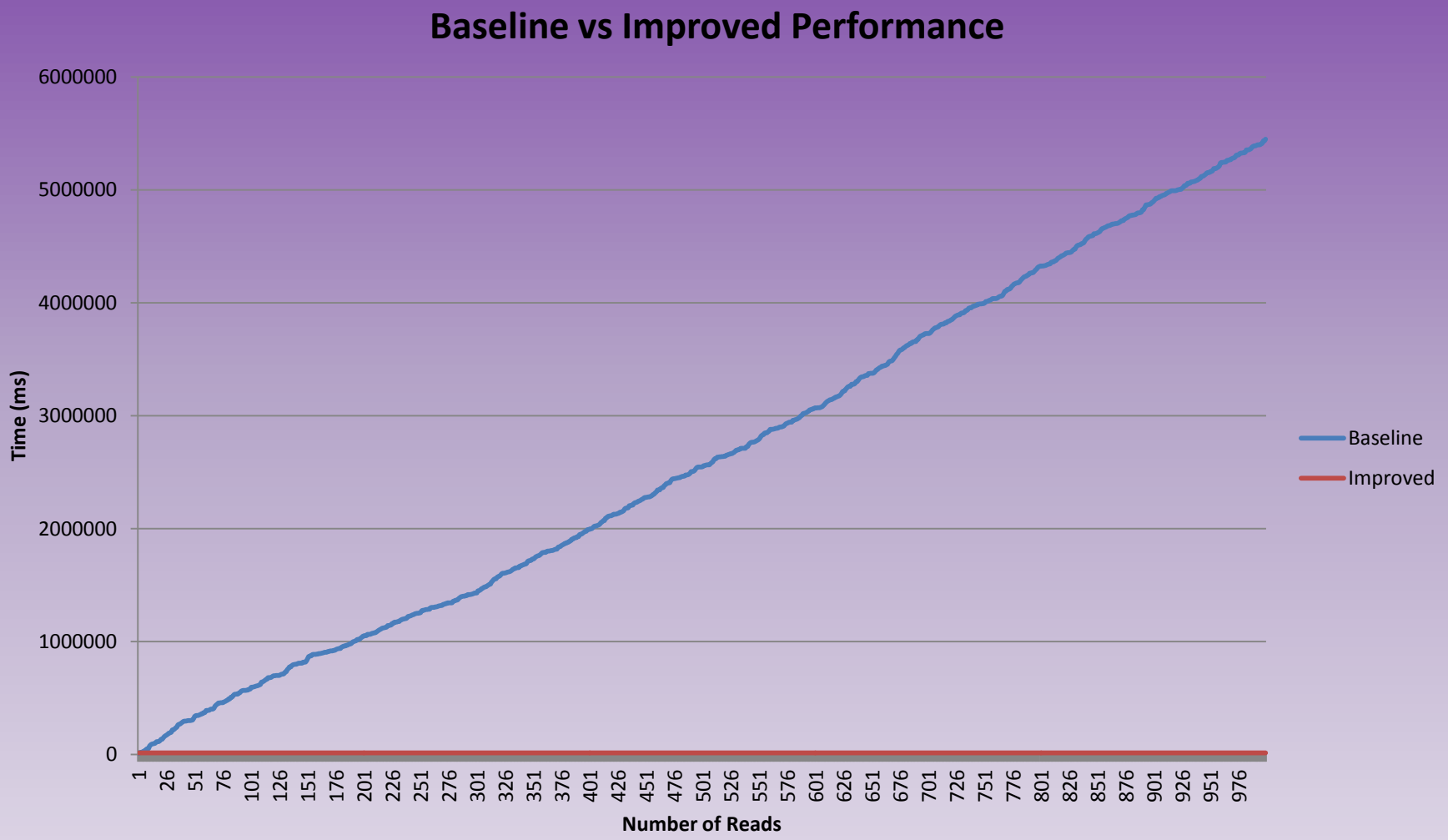
Improved Method

- Indexing the genome beforehand in linear time.
  - 2 million base pairs
  - Reads of size 30 (indexed in 10 bp chunks)
  - Takes about 16 seconds to index

Improved Method

- Looking up reads…
  - Allowing for 2 insertions
  - Break up reads into 3 chunks, and determine proper location
  - Lookups take on average 0.6 seconds
- Reassembling the genome
- Accuracy

Performance



Food for Thought

- Even the improved method takes time
- Each of our projects represent only a small space of the genetics problem
- Much work to be done!

Thank You!

Questions?