How to Multiple Align Huge Number of Pyrosequencing Reads

Fahad Saeed and Ashfaq Khokhar

Department of Electrical and Computer Engineering, University of Illinois at Chicago

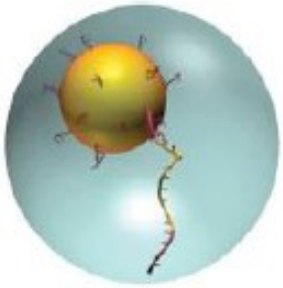
Osvaldo Zagordi, and Niko Beerenwinkel

Department of Biosystems Science and Engineering, ETH Zurich

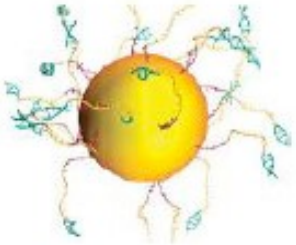
Overview

- Genome Sequencing with Pyrosequencing (454 GS20 Platform)
- Need for Multiple Alignments
- Problem Statement and Assumptions
- Algorithmic approach
- Quality Results
- Ongoing works
- Conclusions and Future work

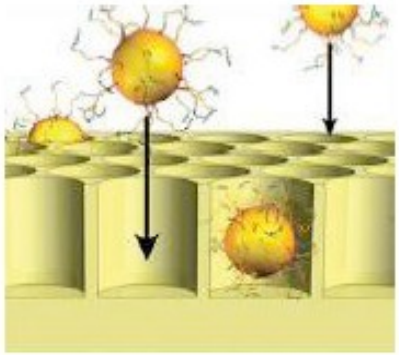
A look at the 454 sequencing technology



DNA fragments are attached to the beads

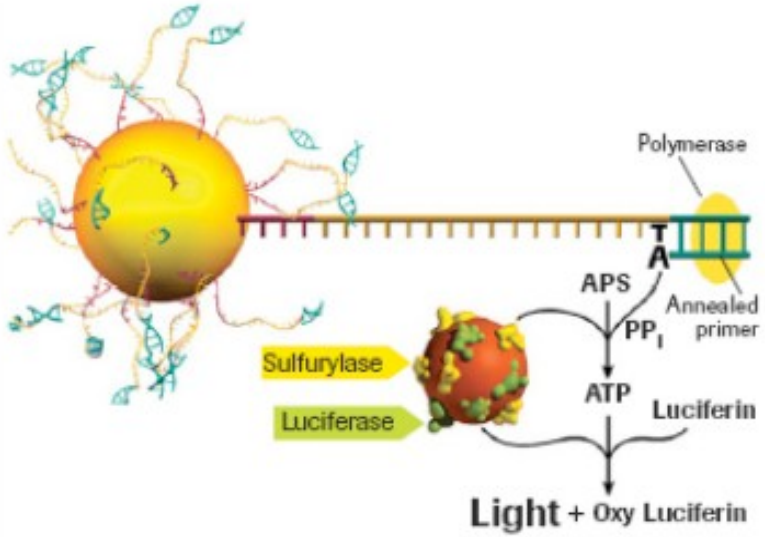


fragments are "amplified"



single beads in ~100.000 wells

DNA is read by detecting light emission associated to base incorporation



Genome Sequencing with pyrosequencing

- Fast and Cheap Sequencing: an important goal
- Pyrosequencing: attractive alternative to current sequencing techniques

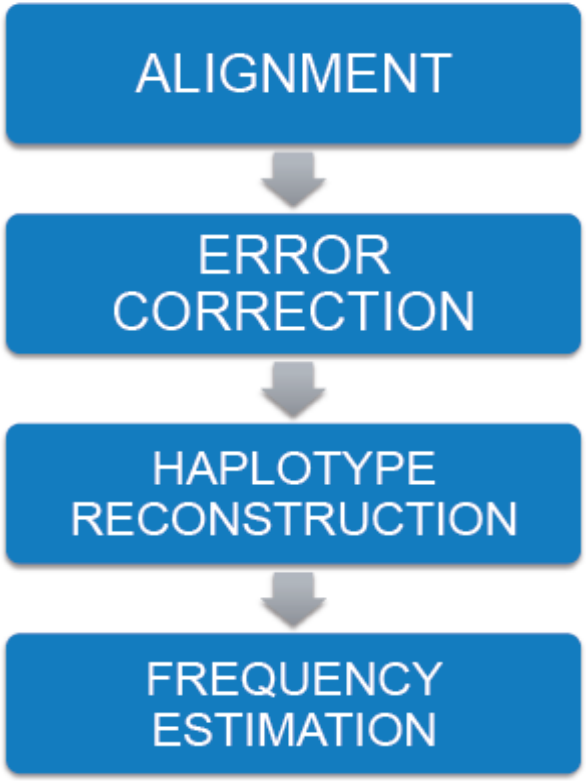
	Sanger	454/Roche
bps per run	~10^5	~10^8
read length	700-1000	~400
cost per run	~1000 $	~15000 $
cost per Mbp	10K $	100 $
accuracy	high	low (in-dels)

Key Issues with pyrosequencing

- Read Length: on average 250-400bp
- Orientation: Original and complement (Ignored for this work)
- Errors
  - Insertions: ~36%
  - Deletions: ~27%
  - Ambiguous Bases: ~21%
  - Substitutions: ~16%

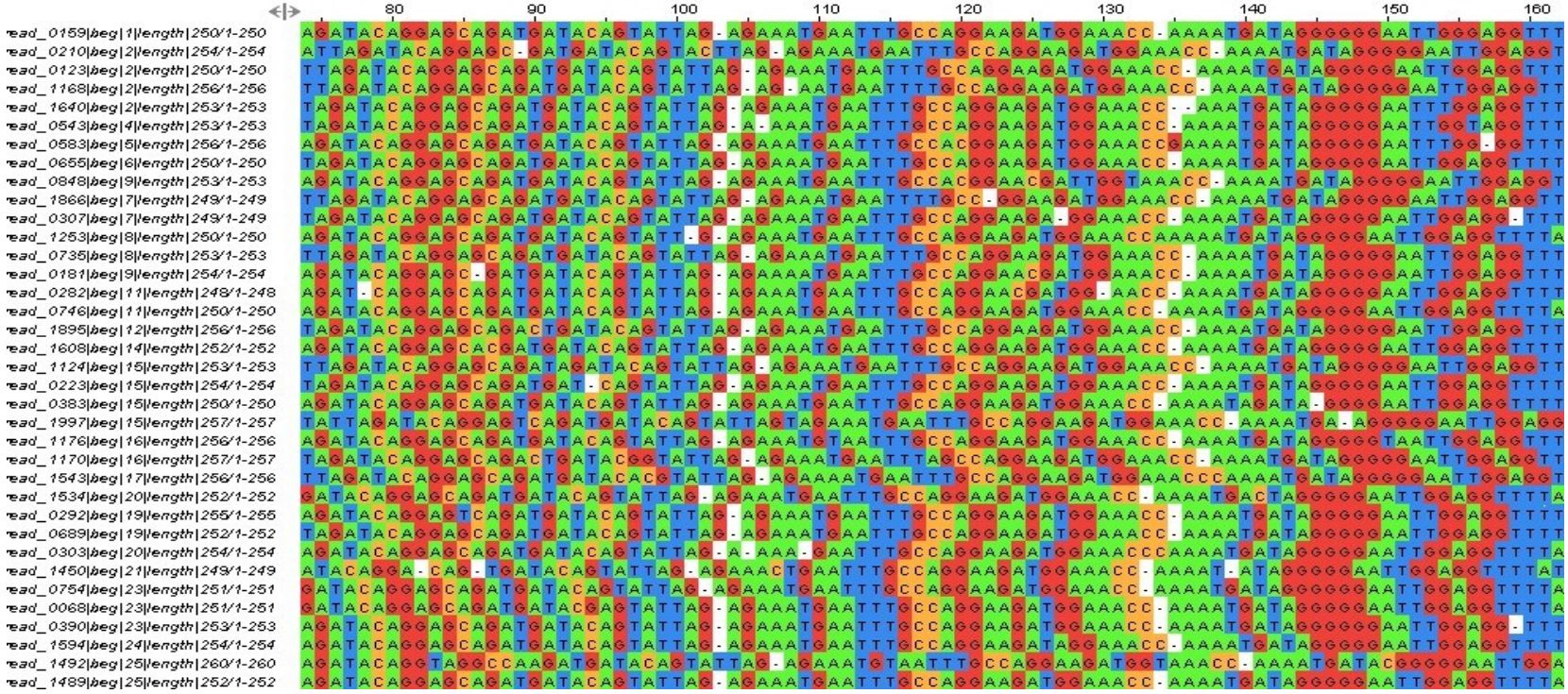
Need for Alignments

- Required as a preprocessing step for number of procedures.
- For example, from reads to haplotypes:



Pairwise alignment

- If reference genome is available, then pairwise alignments can be done
- But don't work :
  - Primarily because of insertions due to pyrosequencing



Issues with Multiple Alignments of short pyrosequencing reads

- Huge number of reads: ~100,000 reads in a single run
- Out of box alignment software are not feasible :
- Because of high complexity. Most are of the order of $O(N^4)$ e.g. Muscle, Clustalw
- Do not give accurate alignments for reads.
  - Primary reason: Do not take into account the position of the reads w.r.t. genome.

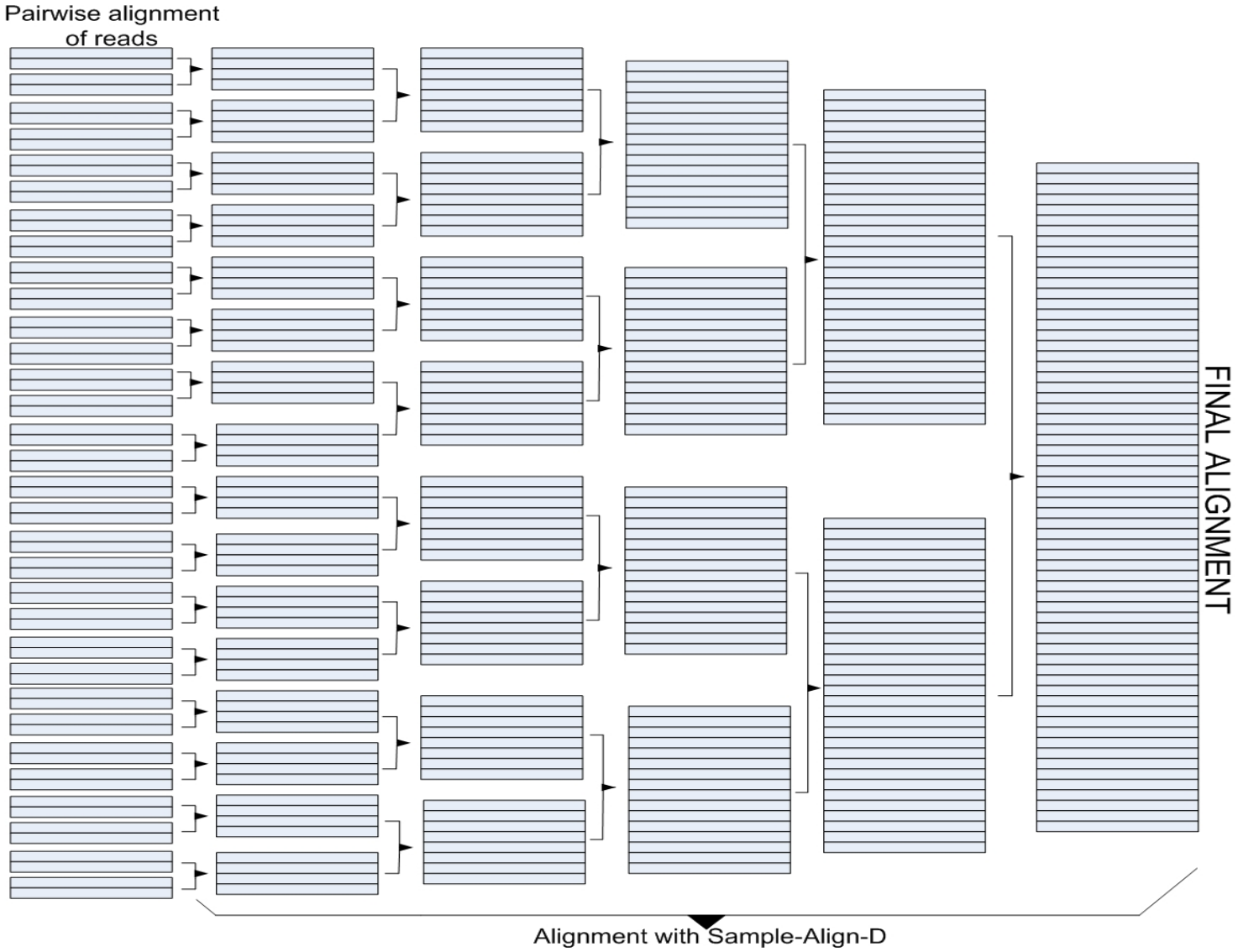
Pyro-Align Algorithm

- Based on the parallel framework of Sample-Align-D algorithm
  - Sample-Align-D gives super-linear speed-ups on multiprocessors; hence gives speed advantages when used on sequential machines
- Assumptions:
  - The original genome or the wildtype is available
  - The reads are in ‘forward’ orientation

Pyro-Align Algorithm sketch

- Align each read to the reference genome, using semi-global alignment.
  - This will place the reads in correct positions
- Do a Hierarchical Progressive alignment:
  - Reorder the reads using the starting positions of the reads to 'generate' a guidance tree.
  - Do pairwise alignments according to the tree.
  - Do profile-profile alignments in a hierarchical fashion

Hierarchical Progressive alignment



Pyro-Align Algorithm Complexity Analysis

- Semi-global alignment = $O(N L_R L_G)$
- Clustering & Reordering = $O(N+NL_G)$
- Pairwisealignments = $O(N L^2_R)$
- Profile-Profile alignments = $O(N \log N L^2_G)$
- Total Asymptotic Complexity= $O(N \log N L^2_G + N L^2_R)$
- Where:
- N= number of reads
- L_R= Average length of reads

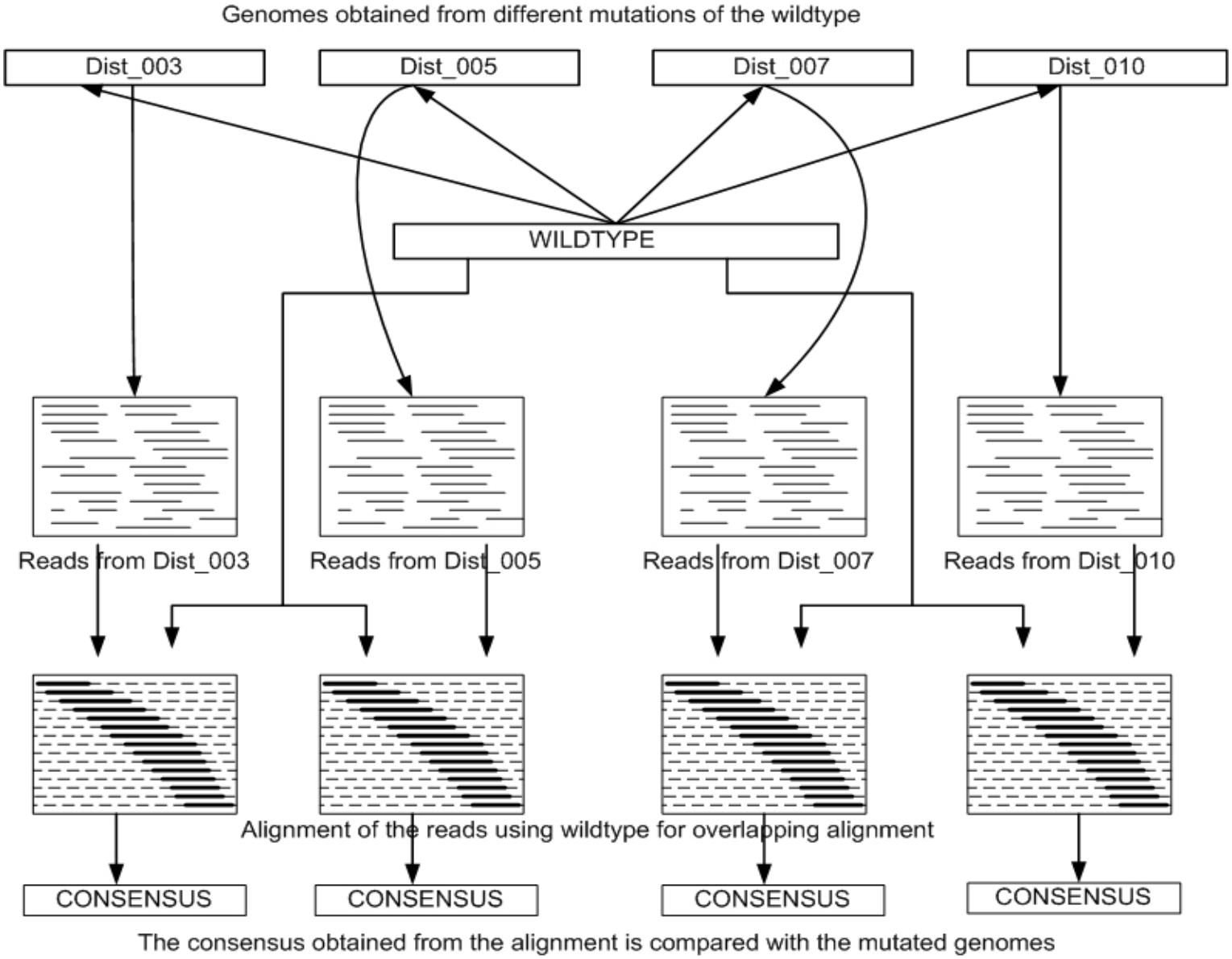
Quality Assessment

- Objective:
  - To access the quality of the alignments w.r.t the original genome
  - Ensure that the system is able to handle reads from multiple haplotypes
- Problems:
  - Ground truth for the alignments is not available(cannot be done with eyeballing due to large number of reads)
  - Standard benchmarks such as Balibase or Prefab cannot be used.

Quality Assessment

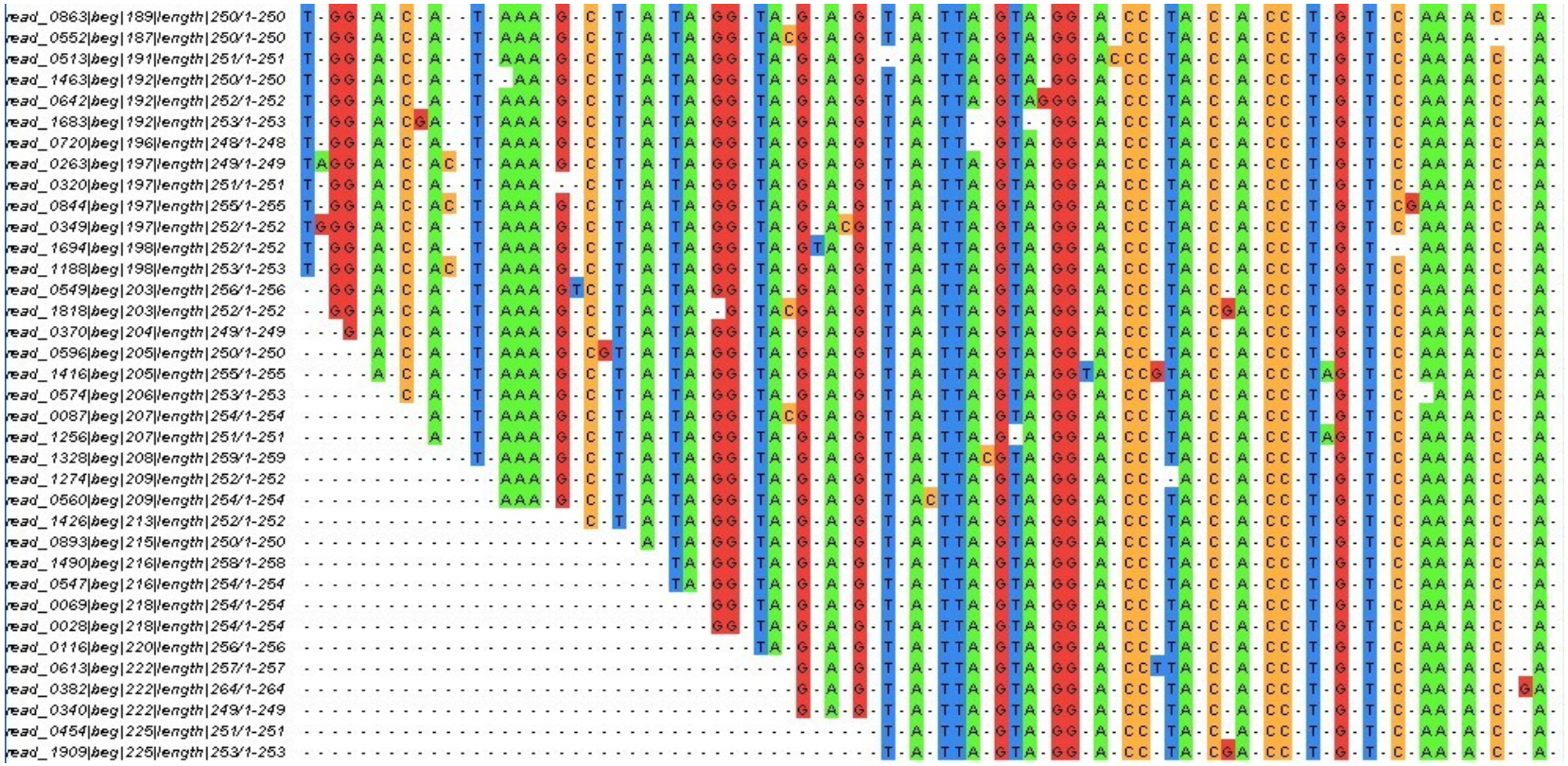
- We choose HIV pol gene with length of 1970bp as wildtype for the experiments
- Four sets of genomes from the wildtype are produced at 3%, 5%, 7% and 10% mutations.
- These mutated genomes are used to produce the reads.
- The reads are aligned and the consensus from the reads is compared with the reference genome.

Quality Assessment



Results

- Successful alignment of reads from mutated genome reads as well as the mixture.



Results

Results

Results

Ongoing Works—Sketch of parallelization of pyro-Align

- Proposed Approach:

➢ In parallel on multiple processors, align each read to the reference genome, using semi-global alignment.

➢ Do a Parallel Hierarchical Progressive alignment using sample-align-D

➢ We expect super-linear speedups for parallel pyro-align giving enormous advantage in terms of timing and memory

Conclusions and Future work

- A low complexity algorithm for aligning huge number of pyrosequencing reads is presented.
- Successfully aligned the reads from mutated and mixture of mutated genomes.
- Presented the quality assessment and compared with pairwise alignments
- We are working on parallelization of the algorithm

Acknowledgements

- We would like to thank the Beerenwinkel group for the support at Biosystems science and engineering dept at ETH Zurich.

- Thanks also goes to Nick Eriksson of the University of Chicago

- Special thanks to Tanya Berger Wolf of Laboratory for Computational Population Biology, UIC.

Questions ??