An evaluation of the genome alignment landscape

Alexandre Fonseca

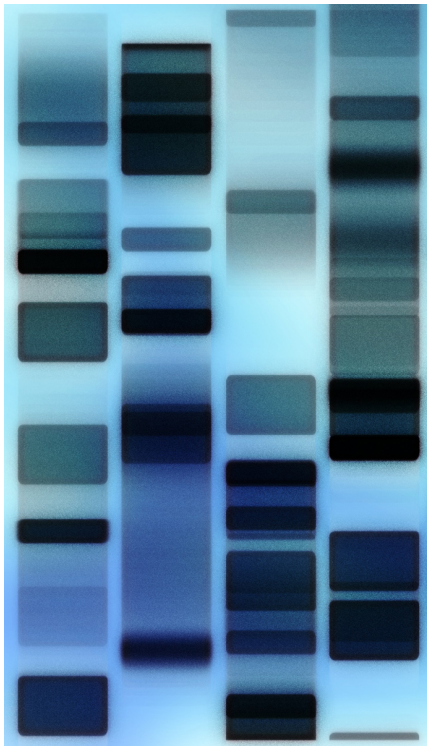
KTH Royal Institute of Technology

December 16, 2013

Table of Contents

What is genetic research?

- The study of an individual's genome.
  - DNA & RNA.
- Allows identification of:
  - Particular traits and characteristics.
  - Anomalous mutations.
- Applicability: agriculture, medicine, ...
- Genome analysis pipeline.



Motivation

- Next Generation Sequencing (NGS):
  - Parallelization of the sequencing process.
  - Each run produces thousands of small reads.
  - >400x more throughput than 1^{st} generation.
- Alignment:
  - Matching reads to correct segments of a reference genome.
  - Most complex and time-consuming task.
  - Most implementations still very centralized.
    - Parallelization at the thread/processor level.
    - Hard to scale to handle increasing amounts of data.
    - Could we leverage the power of the cloud?

Objective of the project

- Evaluate and compare different sequence aligners:
  - Alignment duration.
  - Alignment accuracy.
  - Scalability.
- Centralized Aligners:
  - Bowtie1 - 1.0.0 (April 9^{th}, 2013).
  - BWA - 0.5.10 (November 13^{th}, 2013)
  - Bowtie2 - 2.1.0 (February 21^{st}, 2013).
- Distributed Aligners:
  - Crossbow - 1.2.1 (May 30^{th}, 2013)
  - SEAL - 0.3.2 (February 7^{th}, 2013)

Table of Contents

Hardware

- Evaluated on the 7-node SICS cluster. Each node:
  - 2x 6-core AMD Opteron 24355 CPUs.
  - 32 GB of RAM.
  - 1TB of disk.
- Node interconnection: 1Gbps full duplex Ethernet.

Software

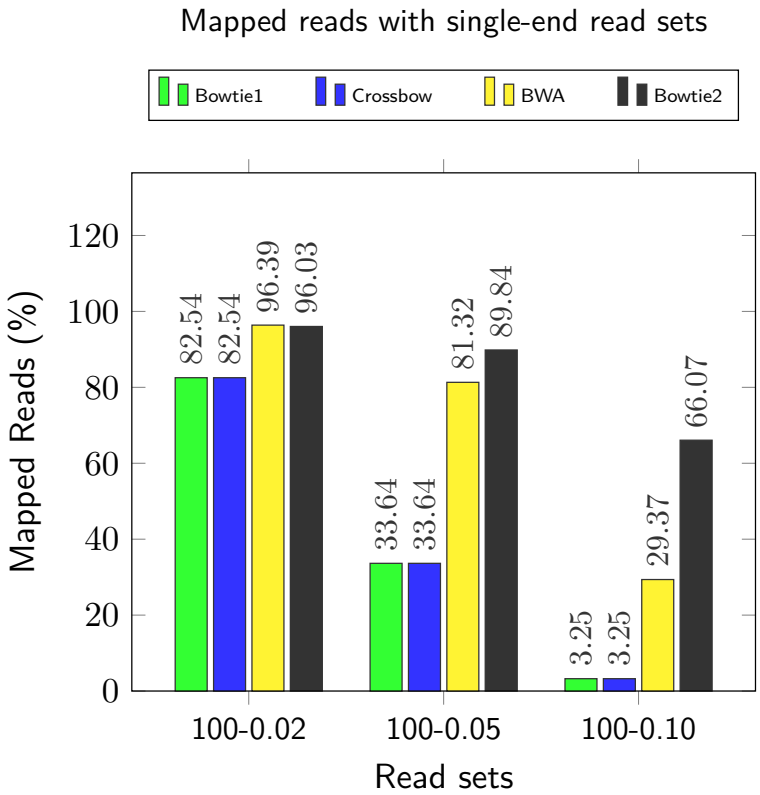
- Shared Hadoop 2.2.0 installation:
  - 5 NodeManagers on nodes 1-5.
  - ResourceManager on node 6 and NameNode on node 7.
  - 16GB of RAM and 12 cores available to each NodeManager.
  - Maximum memory usage per container: 8GB.
- Additional software:
  - FastXToolkit - 0.0.13
  - PicardTools - 1.101
  - SAMTools - 0.1.19
  - SRAToolkit - 2.3.3
  - WGSim - `https://github.com/lh3/wgsim` (a12da33)

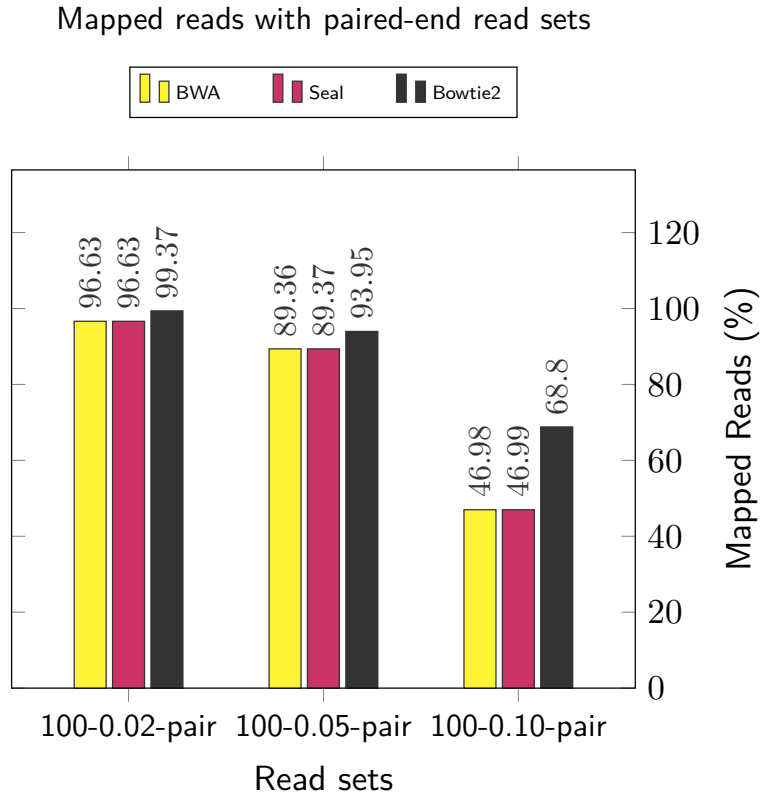
Inputs

- hg-19 reference human genome.
- Single-ended and paired-ended sets of reads.
  - Bowtie1 and Crossbow only suppport single-ended.
  - SEAL only supports paired-ended.
- Two categories of read sets:
  - Simulated - Sampled from hg-19.
    - 900k 100 base-pair reads.
    - 2%, 5% and 10% error rates.
    - 0.09% SNP mutation rate.
    - 0.01% indel mutation rate.
    - 200MB (x2).
  - Real read sets - From the NA12878 individual.
    - Chromosome 4 - 156M 101bp reads - 35.5GB (x2).
    - Chromosome 20 - 51M 101bp reads - 12GB (x2).
    - Chromosome 21 - 31M 101bp reads - 7GB (x2).

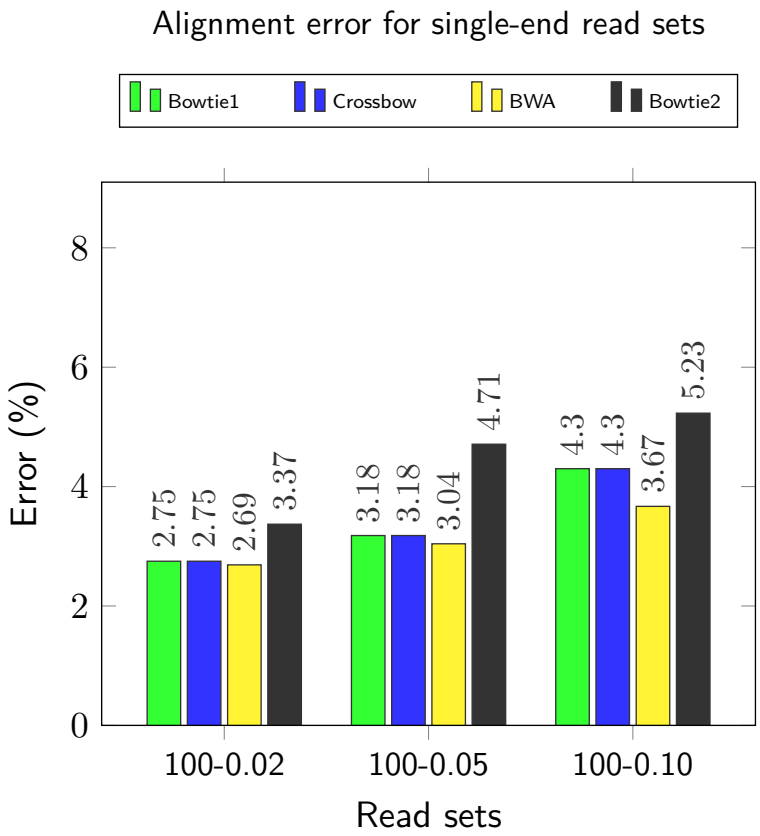
Table of Contents

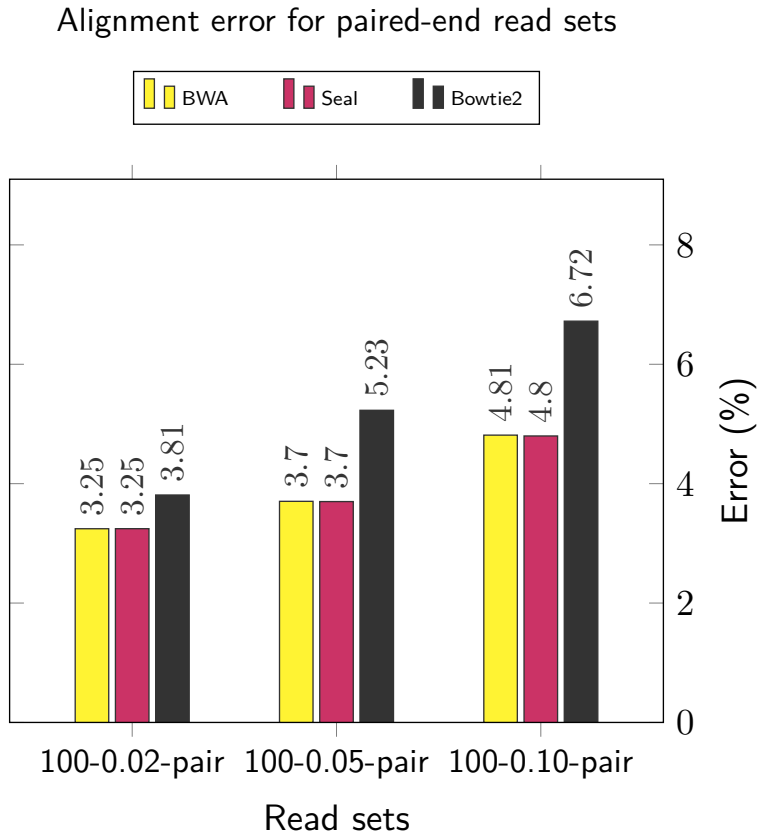
Mapped Reads



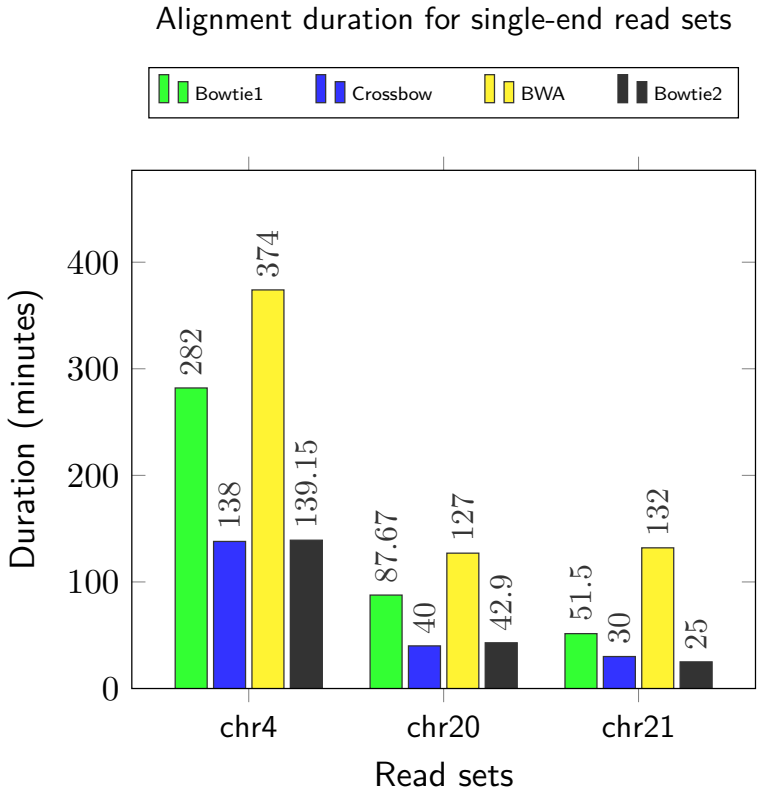


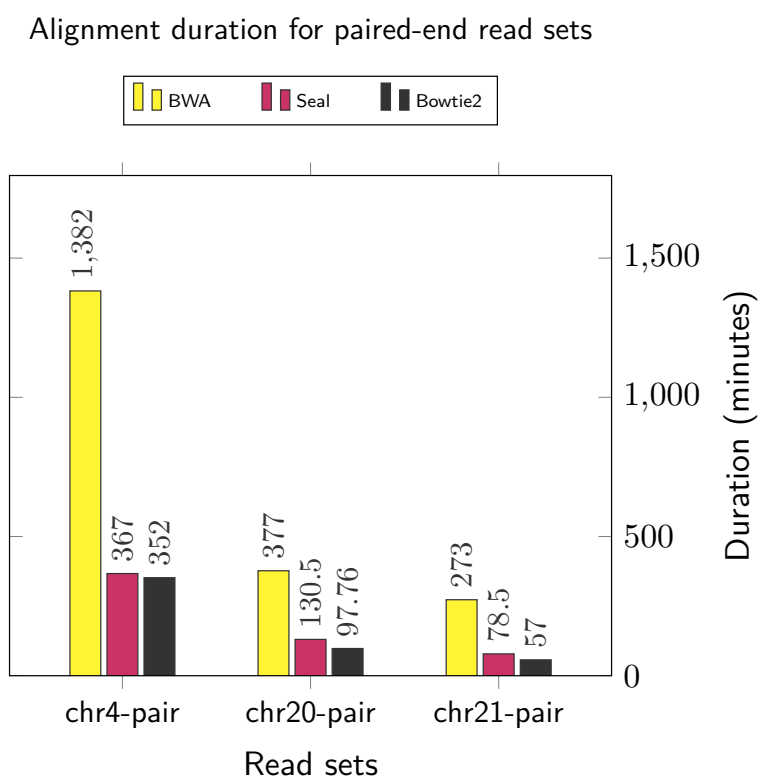
Error


Alignment error for single-end read sets
Bowtie1
Crossbow
BWA
Bowtie2
Error (%)
0
2
4
6
8
2.75
2.75
2.69
3.37
3.18
3.18
3.04
4.71
4.3
4.3
3.67
5.23
100-0.02
100-0.05
100-0.10
Read sets


Alignment error for paired-end read sets
BWA
Seal
Bowtie2
Error (%)
0
2
4
6
8
3.25
3.25
3.81
3.7
3.7
5.23
4.81
4.8
6.72
100-0.02-pair
100-0.05-pair
100-0.10-pair
Read sets

Duration


Alignment duration for single-end read sets
Bowtie1
Crossbow
BWA
Bowtie2
Duration (minutes)
0
100
200
300
400
282
138
374
139.15
87.67
40
127
42.9
51.5
30
132
25
chr4
chr20
chr21
Read sets


Alignment duration for paired-end read sets
BWA
Seal
Bowtie2
Duration (minutes)
0
500
1,000
1,500
1,382
367
352
377
130.5
97.76
273
78.5
57
chr4-pair
chr20-pair
chr21-pair
Read sets

Scalability

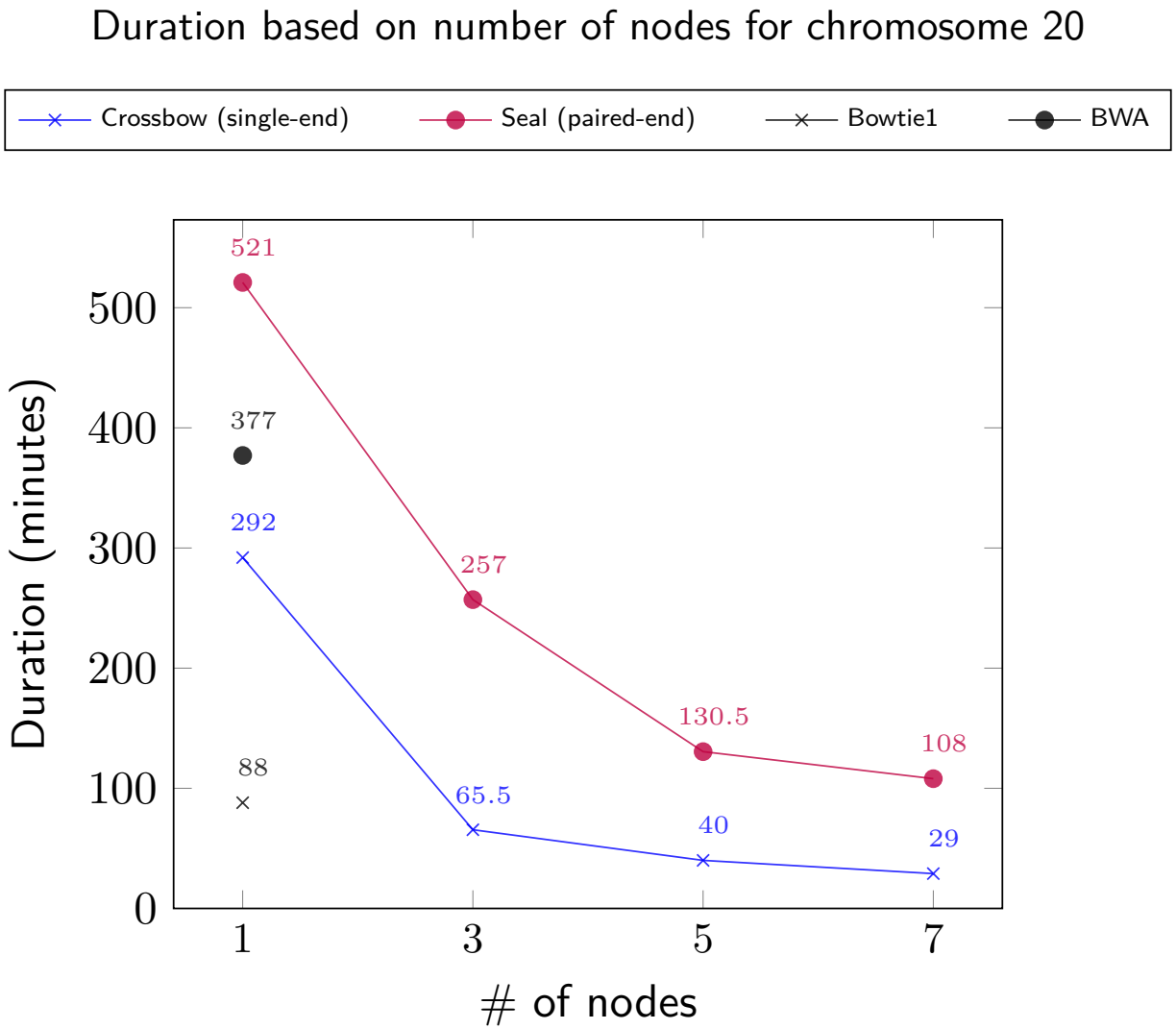


Table of Contents

Conclusion

- Distributing the alignment is feasible.
- More than 3x speedup with no accuracy impact.
- Different aligners target different optimization areas.
- Chance to improve with newer algorithms.

Conclusion

- Distributing the alignment is feasible.
- More than 3x speedup with no accuracy impact.
- Different aligners target different optimization areas.
- Chance to improve with newer algorithms.

Questions?