IPDPS - HiCOMB 2010

Ninth IEEE International Workshop on High Performance Computational Biology

# Exploring Parallelism in Short Sequence Mapping Using Burrows-Wheeler Transform

Doruk Bozdag[1], Ayat Hatem[1,2], Umit V. Catalyurek[1,2]

[1] Department of Biomedical Informatics

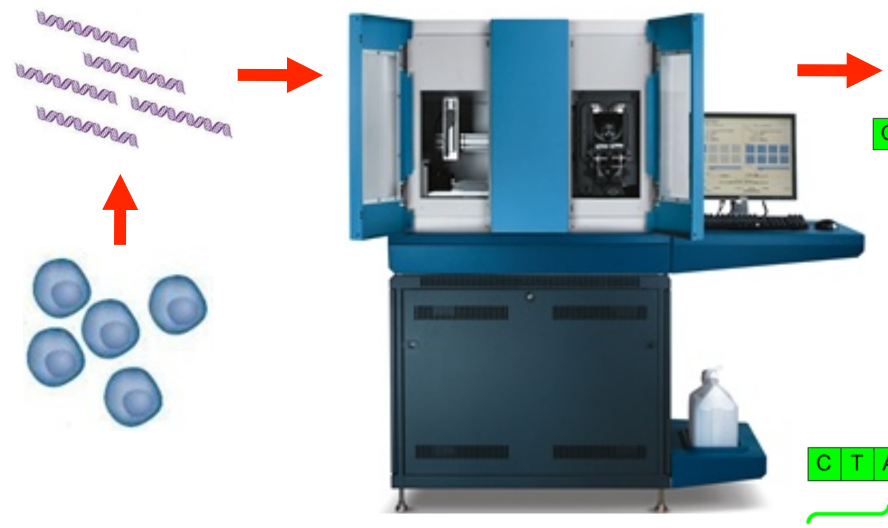
[2] Department of Electrical and Computer Engineering

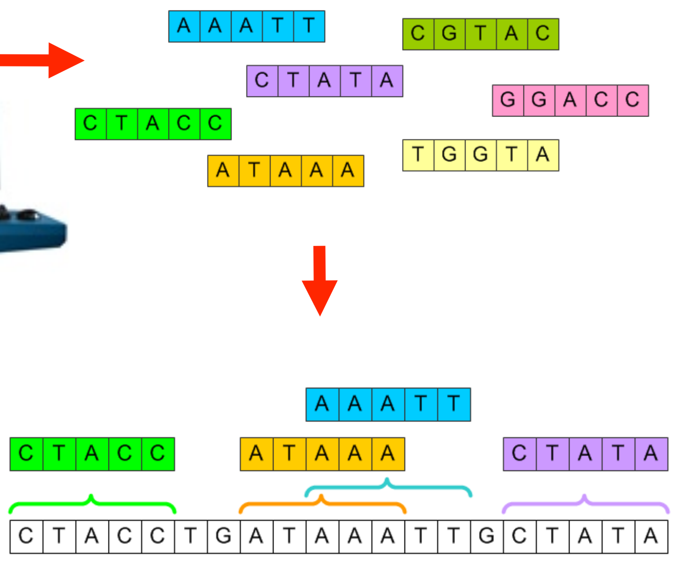
The Ohio State University

## Outline

- Motivation
- Burrows-Wheeler Transform
- Parallelization strategies
- Experimental results
- Conclusion and future work

## Motivation





**ANALYSIS**

**SEQUENCING**

- High throughput sequencing instruments (SOLiD, Solexa, 454) can sequence more than 1 billion bases a day
    - Hundreds of millions of 35-50 base reads

**MAPPING**

- Map reads to a reference genome efficiently
    - Human genome: 3Gb
- Sequential mapping takes about a day
    - **Need fast, parallel algorithms that can handle mismatches**

## Our Ultimate Goal

- Develop generic parallelization framework
  - Identify limitations due to the application scenarios and tools
- Find the “right” tool for the given problem
  - Find the best way to parallelize for a given tool and scenario
  - Quality vs Runtime tradeoffs
- This work is a second step towards that goal [Bozdag IPDPS’09]

## Short Sequence Mapping Tools

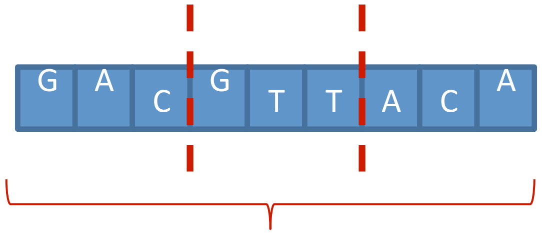
- Many tools have been developed:
  - MapReads, MAQ, RMAP, SHRiMP, ZOOM, mrFAST, SOCS, PASS
- State of the art tools:
  - BWA, Bowtie, and SOAPv2
  - All of them are based on the Burrows-Wheeler Transform
  - Two step mapping approach:
    - Build the index for the reference genome
    - Map reads to the index

## Burrows-Wheeler Transform

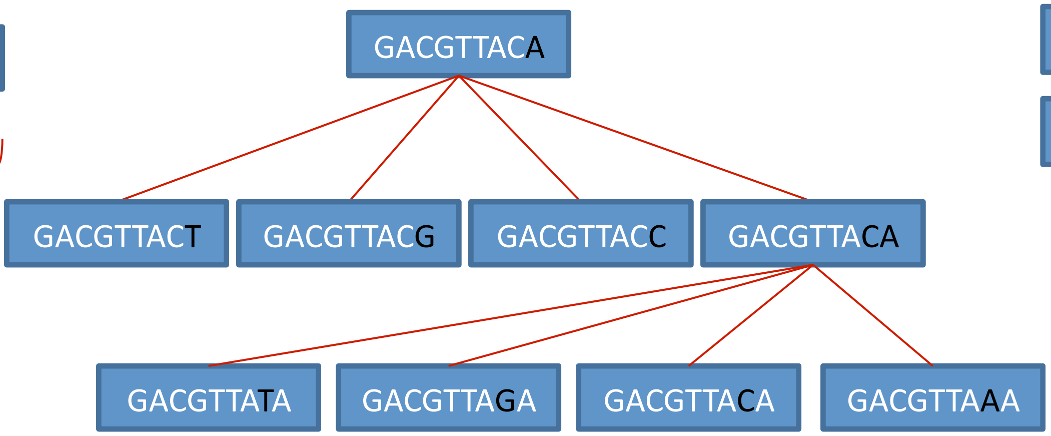
- The Burrows-Wheeler transform (BWT) of a text T is a reversible permutation of the characters in that text
- Designed originally for data compression
- Used by data indexing techniques due to its efficiency
- BWT-based index can be searched in a small memory footprint
- Exact string matching algorithm has been developed by [FOCS 2000] to search through BWT-based index

## Inexact Matching

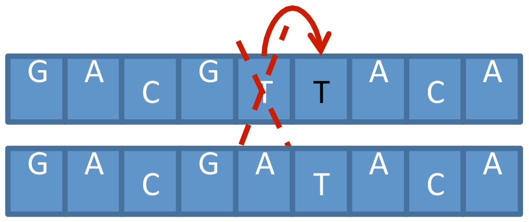
- BWA, SOAP, and Bowtie use an exact matching algorithm based on the BWT-index
- Each one provides a different method to handle inexact matches



SOAP:
Split read into 3 fragments for hits that allow two mismatches



BWA:
Enumerate all possible strings

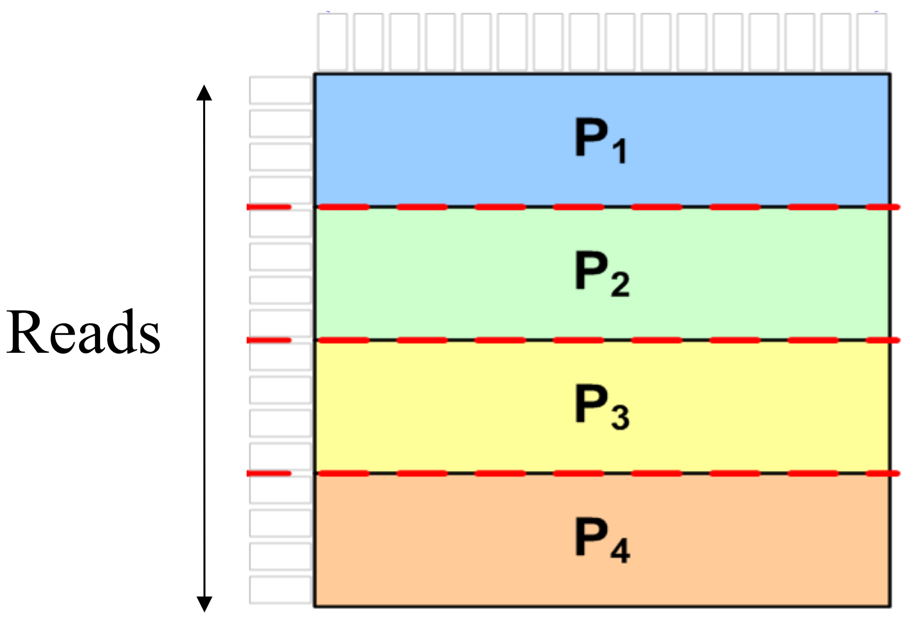


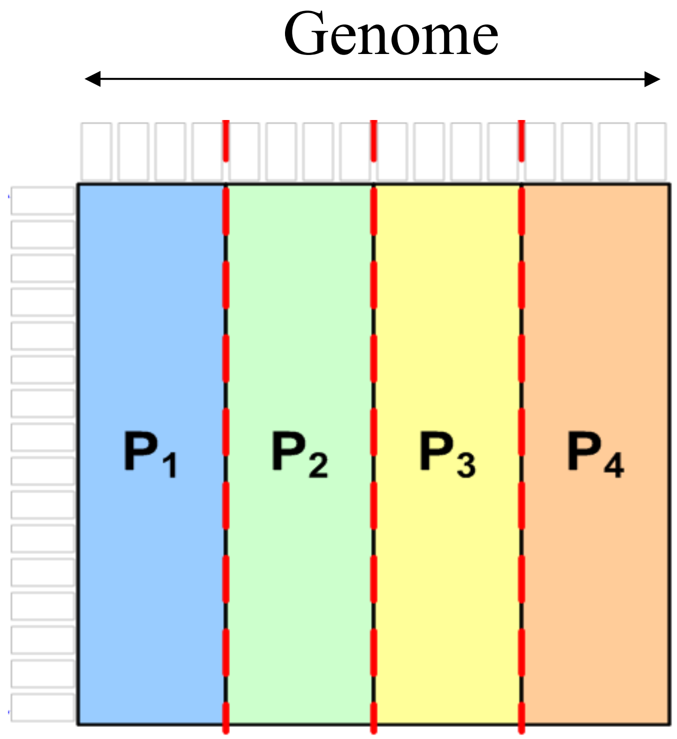
Bowtie:
Match not found at T, backtrack, change, find exact match

# Short Sequence Mapping

- Quality of mapping depends on different factors
- Improved quality $\longrightarrow$ Increased computational cost
- Tools provide different options to compromise quality to limit the computational cost
- Solution: parallel processing strategies

# Parallelization Strategies





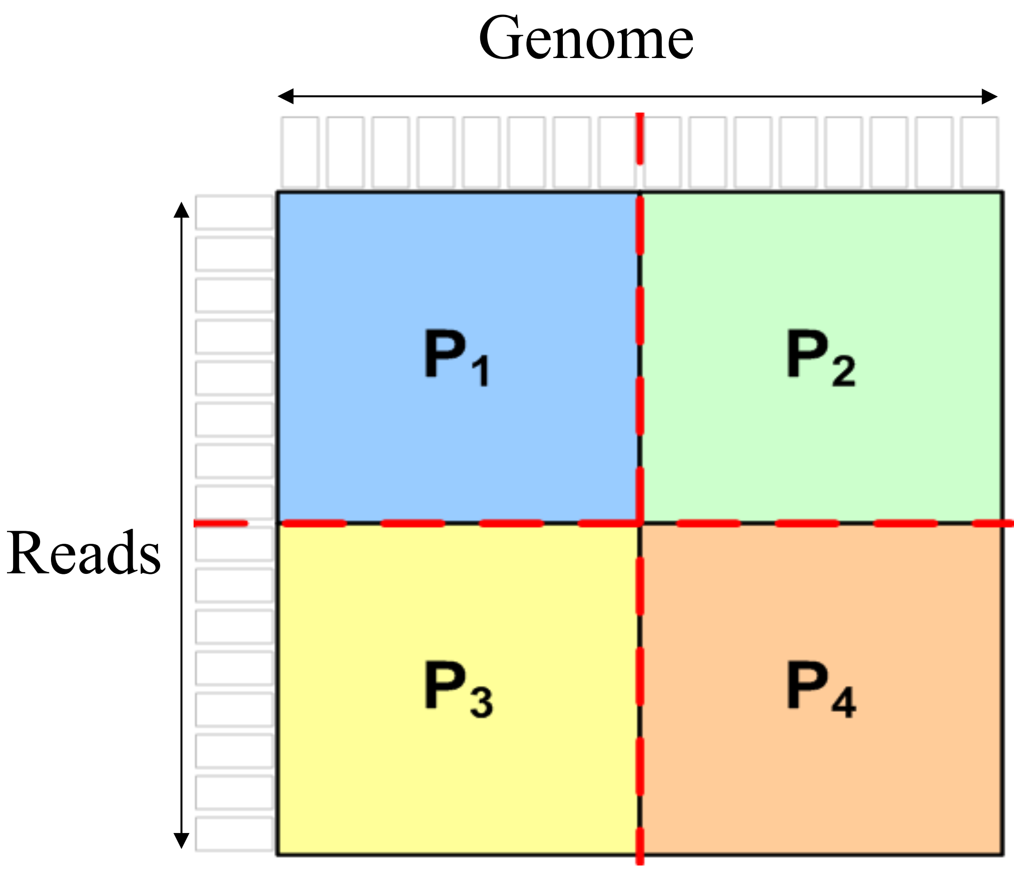
- Partition reads into NR parts
- Mapping very large number of reads to a small genome

- Partition genome into NG parts
- Mapping a small number of reads to a large genome

## Parallelization Strategies

- Partition reads and genome
- Deciding number of read parts (NR) and genome parts (NG) depends on the number of reads and size of the genome

- Two main application scenarios
  - **Index the genome each time for matching:**
    - **Partition reads and genome**
  - Index the genome once:
    - Partition reads only

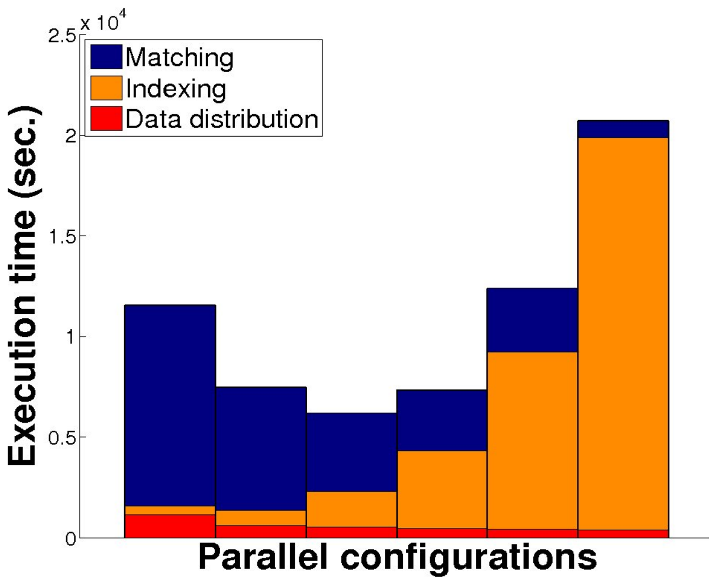


## Experimental Setup

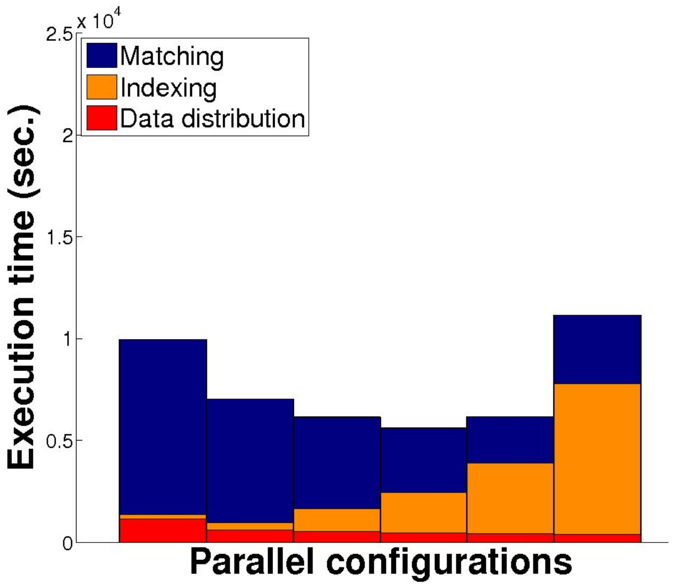
- Compared Bowtie v0.10.1, BWA v0.5.0, SOAP v2.20
- Experiments on 32-node dual 2.4 GHz Opteron cluster with 8GB of memory per node
- Used three reference genomes: human (3.1 Gbp), zebrafish (1.5 Gbp) and lancelet (0.9 Gbp)
- Reads:
  - Real data from a single run of SOLiD system of length 50bp
  - Synthetic data generated by wgsim of length 70bp
    - Wgsim tool is a part of SAMtools package

# Experiments on Real Data

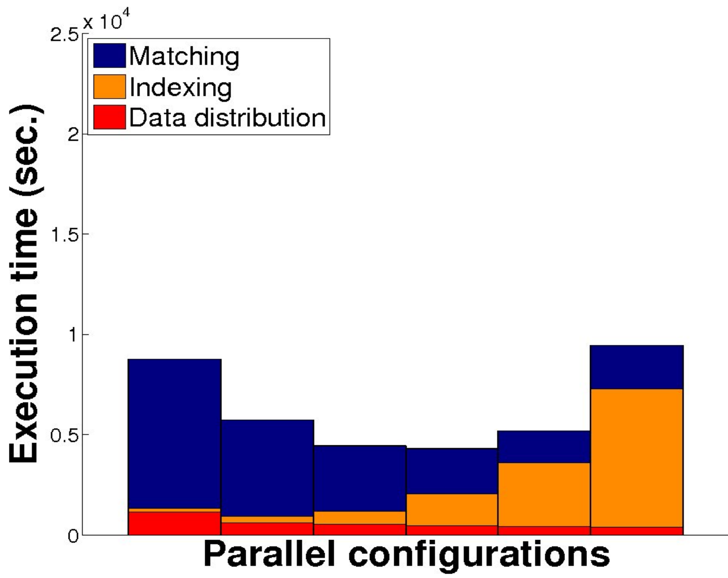
- Number of nodes: 32
- NR x NG: 1x32, 2x16, 4x8, 8x4, 16x2, 32x1
- G: Human. R: 130M
- Bowtie best configuration: 4x8. BWA and SOAP: 8x4



Bowtie



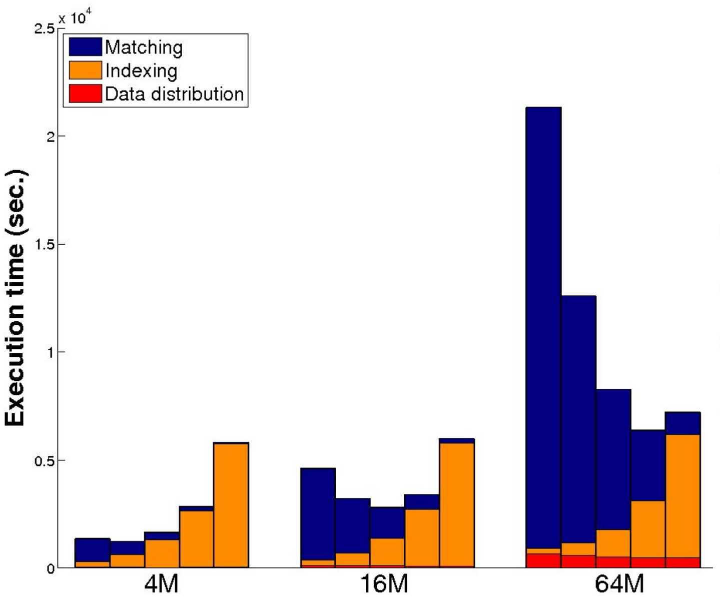
BWA



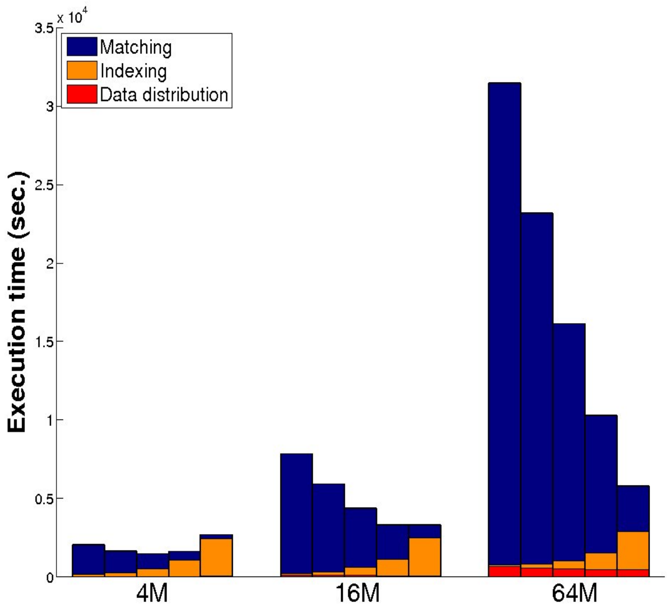
SOAP

# Experiments on Synthetic Data

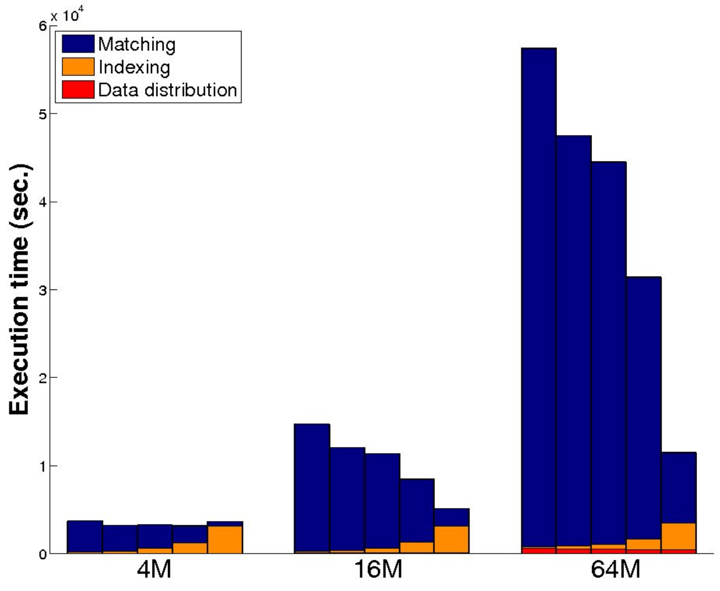
- Number of nodes: 16
- NR x NG: 1x16, 2x8, 4x4, 8x2, 16x1
- G: Zebrafish, R: 4M, 16M, 64M
- Matching time >> indexing time, larger NR is better



Bowtie



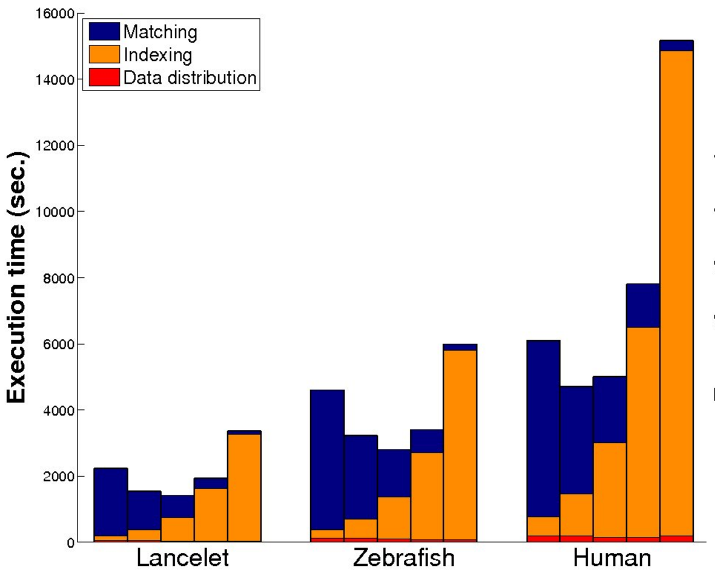
BWA



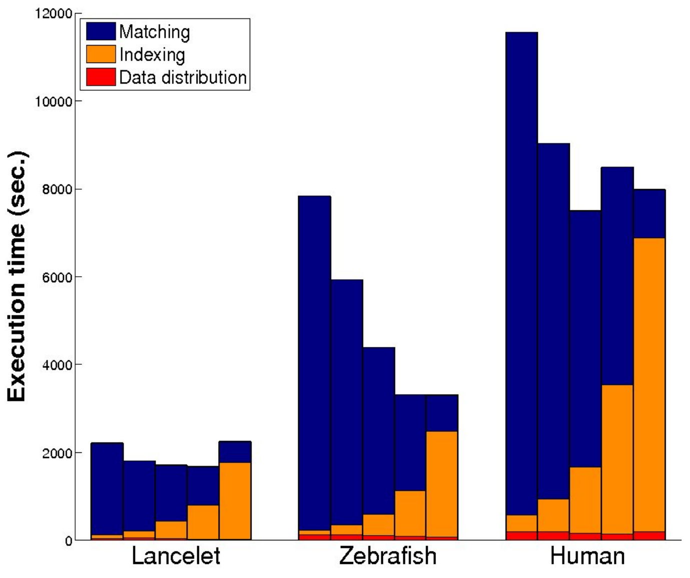
SOAP

# Experiments on Synthetic Data

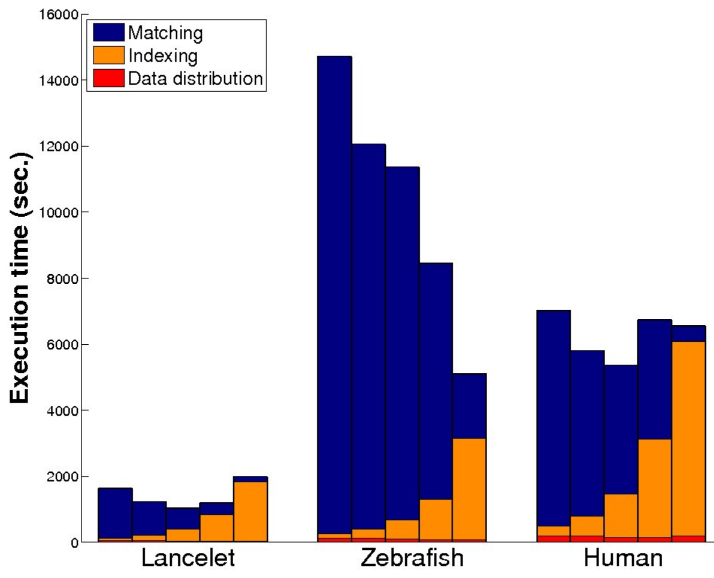
- Number of nodes: 16
- NR x NG: 1x16, 2x8, 4x4, 8x2, 16x1
- G: Lancelet, Zebrafish, Human. R: 16M
- Larger NG speeds up the indexing phase



Bowtie



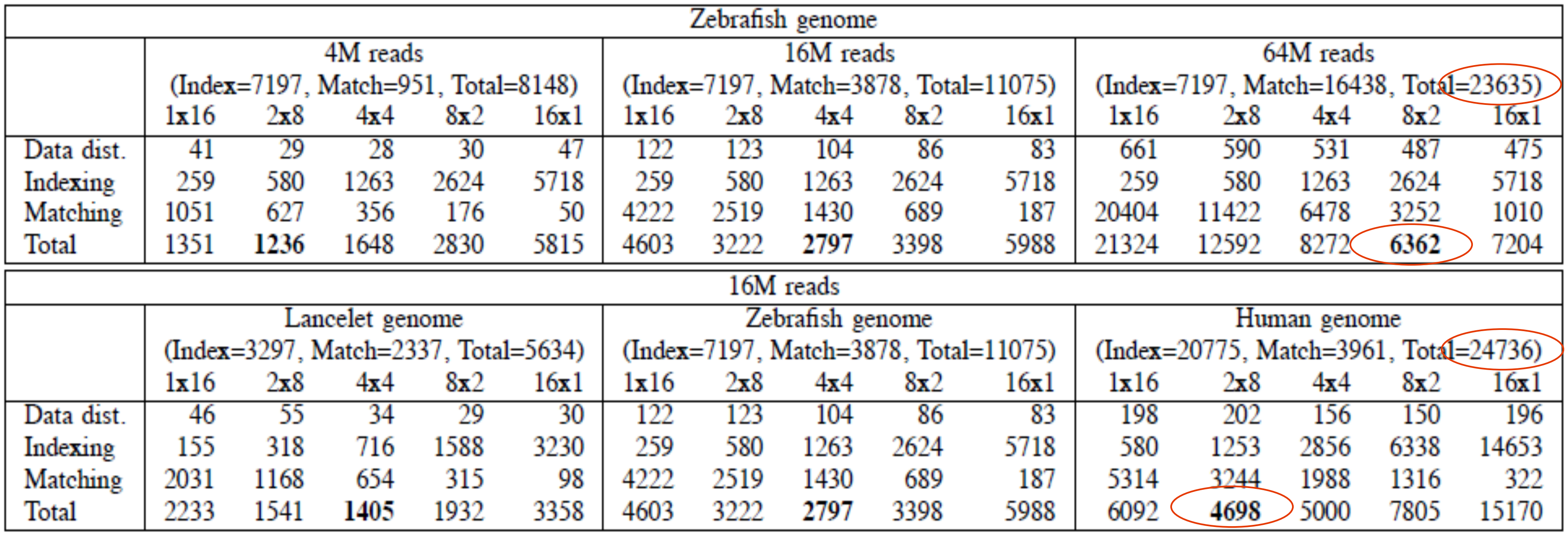
BWA



SOAP

# Experiments on Synthetic Data

- Best configuration: provided 2.2 to 10.7 speed up on 16 nodes



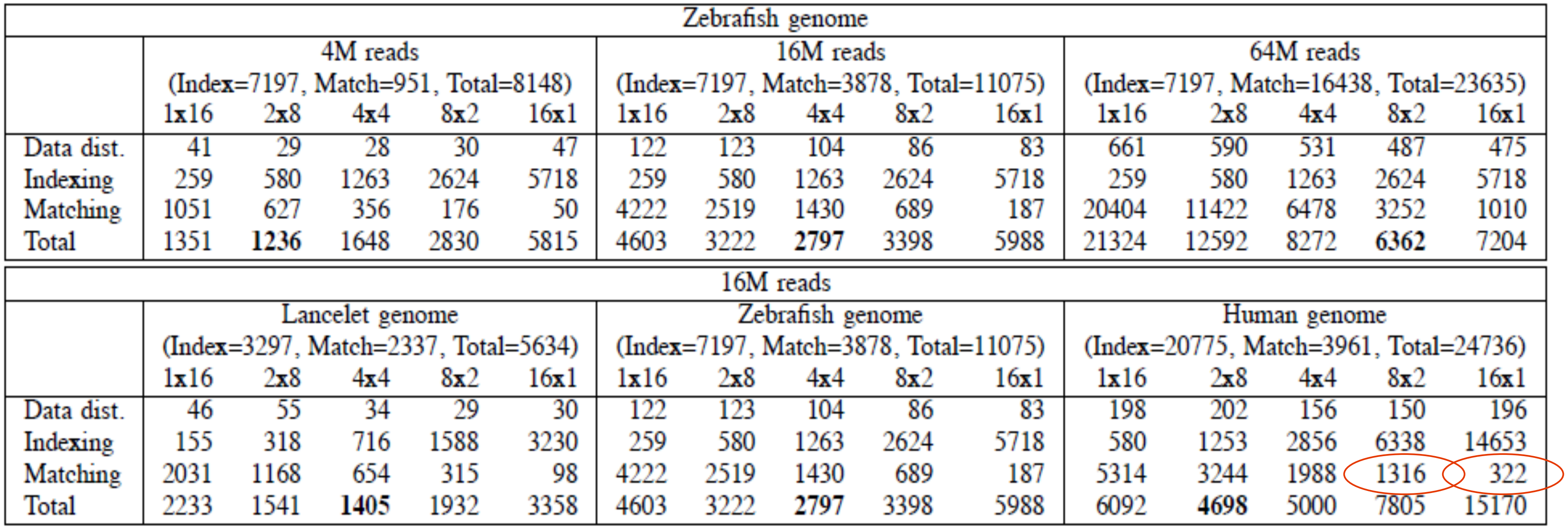
| Zebrafish genome | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4M reads (Index=7197, Match=951, Total=8148) | | | | | 16M reads (Index=7197, Match=3878, Total=11075) | | | | | 64M reads (Index=7197, Match=16438, Total=23635) | | | | |
| | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 |
| Data dist. | 41 | 29 | 28 | 30 | 47 | 122 | 123 | 104 | 86 | 83 | 661 | 590 | 531 | 487 | 475 |
| Indexing | 259 | 580 | 1263 | 2624 | 5718 | 259 | 580 | 1263 | 2624 | 5718 | 259 | 580 | 1263 | 2624 | 5718 |
| Matching | 1051 | 627 | 356 | 176 | 50 | 4222 | 2519 | 1430 | 689 | 187 | 20404 | 11422 | 6478 | 3252 | 1010 |
| Total | 1351 | **1236** | 1648 | 2830 | 5815 | 4603 | 3222 | **2797** | 3398 | 5988 | 21324 | 12592 | 8272 | **6362** | 7204 |

| 16M reads | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Lancelet genome (Index=3297, Match=2337, Total=5634) | | | | | Zebrafish genome (Index=7197, Match=3878, Total=11075) | | | | | Human genome (Index=20775, Match=3961, Total=24736) | | | | |
| | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 |
| Data dist. | 46 | 55 | 34 | 29 | 30 | 122 | 123 | 104 | 86 | 83 | 198 | 202 | 156 | 150 | 196 |
| Indexing | 155 | 318 | 716 | 1588 | 3230 | 259 | 580 | 1263 | 2624 | 5718 | 580 | 1253 | 2856 | 6338 | 14653 |
| Matching | 2031 | 1168 | 654 | 315 | 98 | 4222 | 2519 | 1430 | 689 | 187 | 5314 | 3244 | 1988 | 1316 | 322 |
| Total | 2233 | 1541 | **1405** | 1932 | 3358 | 4603 | 3222 | **2797** | 3398 | 5988 | 6092 | **4698** | 5000 | 7805 | 15170 |

Bowtie sequential and parallel running time

# Experiments on Synthetic Data

- Some unexpected results

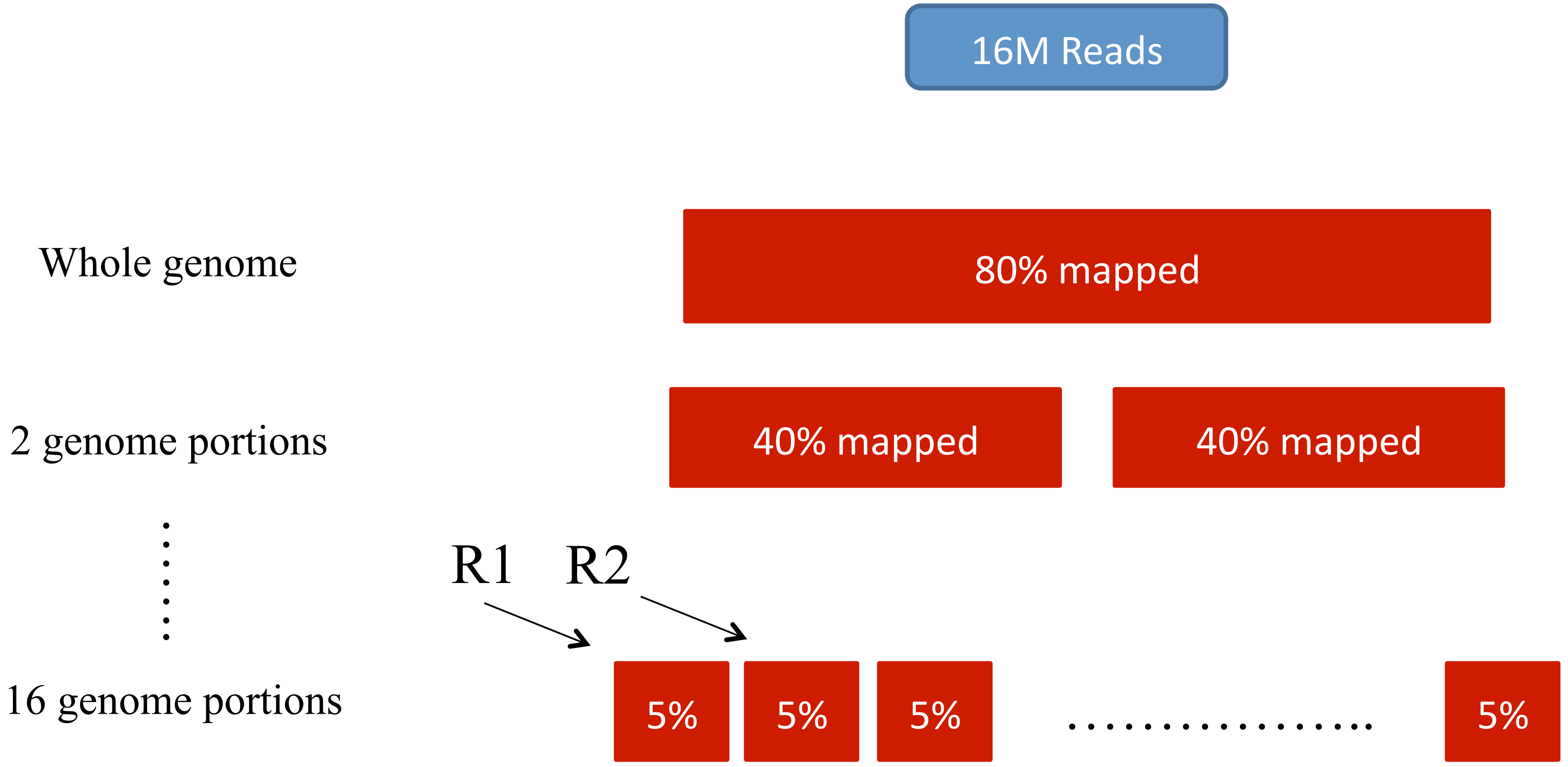


| Zebrafish genome | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4M reads | | | | | 16M reads | | | | | 64M reads | | | | |
| | (Index=7197, Match=951, Total=8148) | | | | | (Index=7197, Match=3878, Total=11075) | | | | | (Index=7197, Match=16438, Total=23635) | | | | |
| | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 |
| Data dist. | 41 | 29 | 28 | 30 | 47 | 122 | 123 | 104 | 86 | 83 | 661 | 590 | 531 | 487 | 475 |
| Indexing | 259 | 580 | 1263 | 2624 | 5718 | 259 | 580 | 1263 | 2624 | 5718 | 259 | 580 | 1263 | 2624 | 5718 |
| Matching | 1051 | 627 | 356 | 176 | 50 | 4222 | 2519 | 1430 | 689 | 187 | 20404 | 11422 | 6478 | 3252 | 1010 |
| Total | 1351 | **1236** | 1648 | 2830 | 5815 | 4603 | 3222 | **2797** | 3398 | 5988 | 21324 | 12592 | 8272 | **6362** | 7204 |

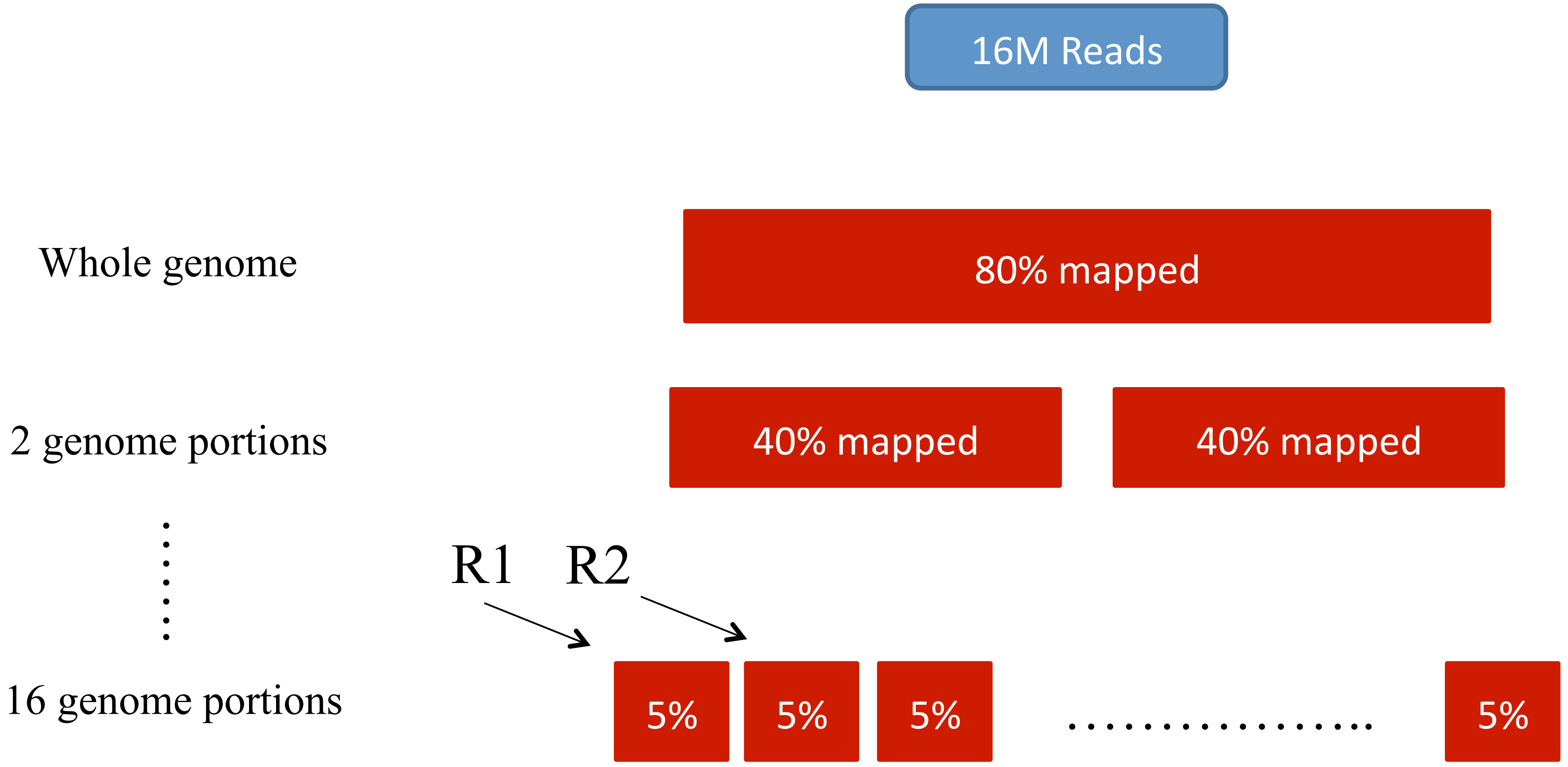
| 16M reads | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Lancelet genome | | | | | Zebrafish genome | | | | | Human genome | | | | |
| | (Index=3297, Match=2337, Total=5634) | | | | | (Index=7197, Match=3878, Total=11075) | | | | | (Index=20775, Match=3961, Total=24736) | | | | |
| | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 | 1x16 | 2x8 | 4x4 | 8x2 | 16x1 |
| Data dist. | 46 | 55 | 34 | 29 | 30 | 122 | 123 | 104 | 86 | 83 | 198 | 202 | 156 | 150 | 196 |
| Indexing | 155 | 318 | 716 | 1588 | 3230 | 259 | 580 | 1263 | 2624 | 5718 | 580 | 1253 | 2856 | 6338 | 14653 |
| Matching | 2031 | 1168 | 654 | 315 | 98 | 4222 | 2519 | 1430 | 689 | 187 | 5314 | 3244 | 1988 | 1316 | 322 |
| Total | 2233 | 1541 | **1405** | 1932 | 3358 | 4603 | 3222 | **2797** | 3398 | 5988 | 6092 | **4698** | 5000 | 7805 | 15170 |

Bowtie sequential and parallel running time

## Experiments on Synthetic Data



## Experiments on Synthetic Data

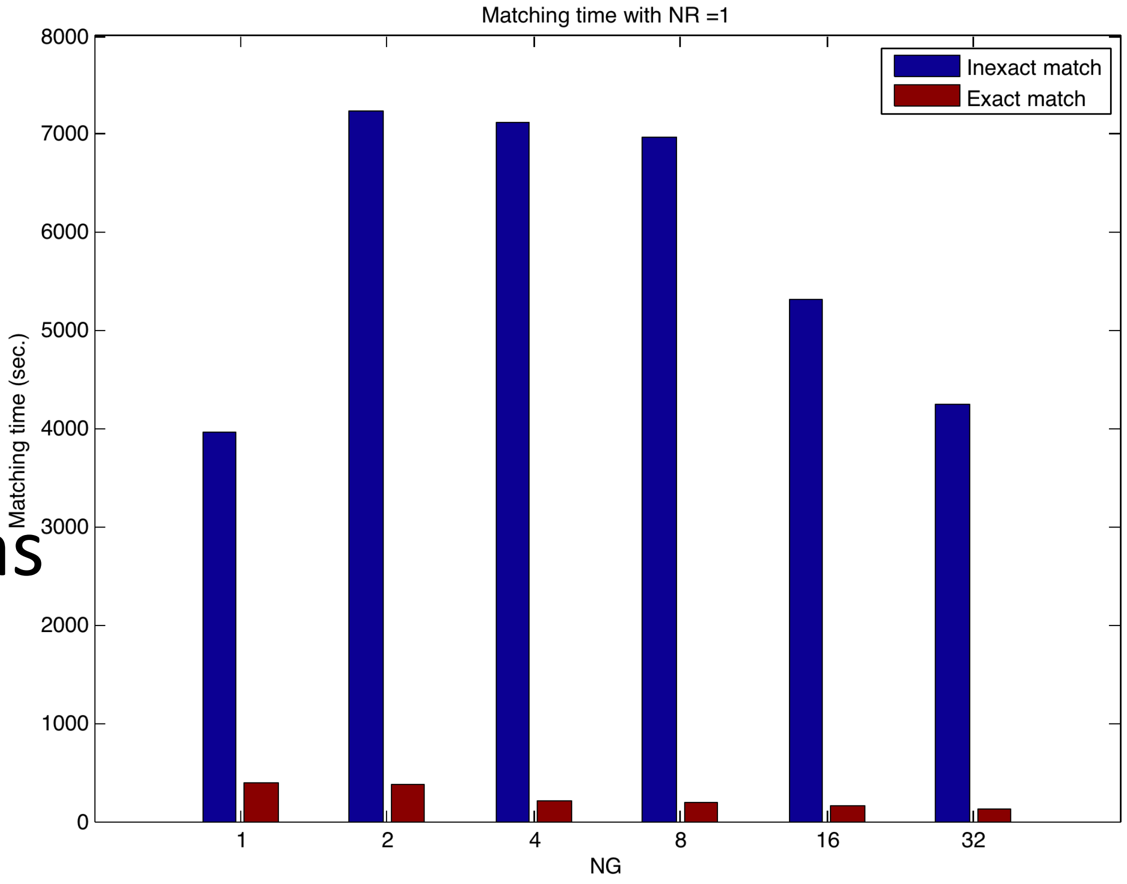


| | Mapped reads | Bowtie Matching Time |
|---|---|---|
| R1 | 90.04% | 121 |
| R2 | 4.2% | 482 |

**Inexact matching phase leads to increasing the running time**

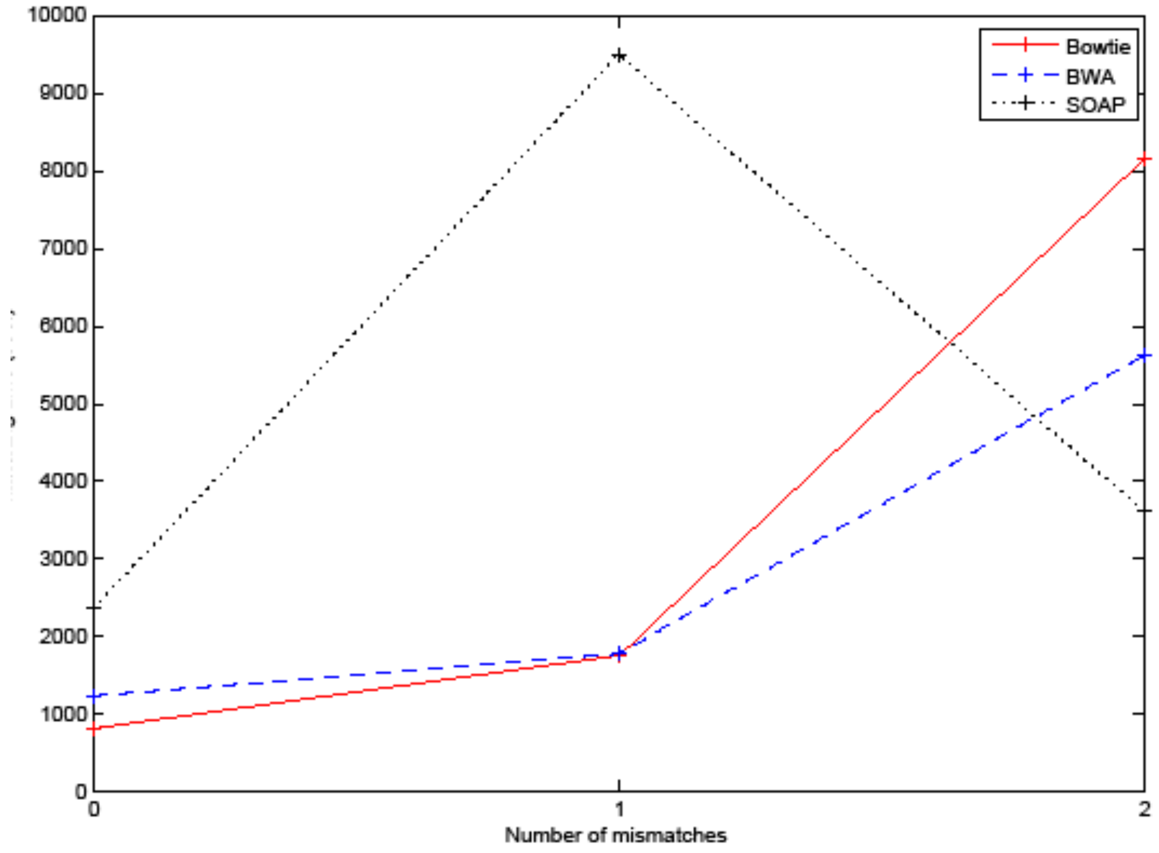
# Inexact vs Exact Matching in Bowtie

- R: 16M
- G: Human
- NG: 1, 2, 4, 8, 16, 32
- Inexact matching overhead is offset by decreased index size as NG increases but not enough!



# Varying Number of Mismatches

- G: Human, R: 21M
- Number of mismatches: 0, 1, 2
- SOAP is not designed for finding "at most" 1 mismatch: requires 2 executions



## Conclusions

- Experimented different data distribution strategies
- Tested on the state of the art mapping tools: Bowtie, BWA, and SOAP
- For Index+Matching Scenario  observed 2.2 to 10.7 speedup on 16 nodes
- Best data distribution strategy depends on:
  - Input scenario: genome size, number of reads, and number of nodes
  - Relative efficiency of the indexing and matching steps
    - algorithm used for inexact matching
- In case of building the index once, it is better to use read partitioning only
- Although our intention was not to compare the tools, yet ☺
  - Bowtie indexing is relatively slow
  - On human genome, Bowtie is faster for exact matching but when increasing the number of mismatches SOAP becomes the fastest

## Future Work: Our Goal

- Develop generic parallelization framework
  - Identify limitations due to the application scenarios and tools
- Find the “right” tool for the given problem
  - further analysis of tools and methods are needed
    - Quality vs Runtime tradeoffs

## Thanks

- For more information
  - umit@bmi.osu.edu
  - http://bmi.osu.edu/~umit or http://bmi.osu.edu/hpc
- Research at the HPC Lab is funded by