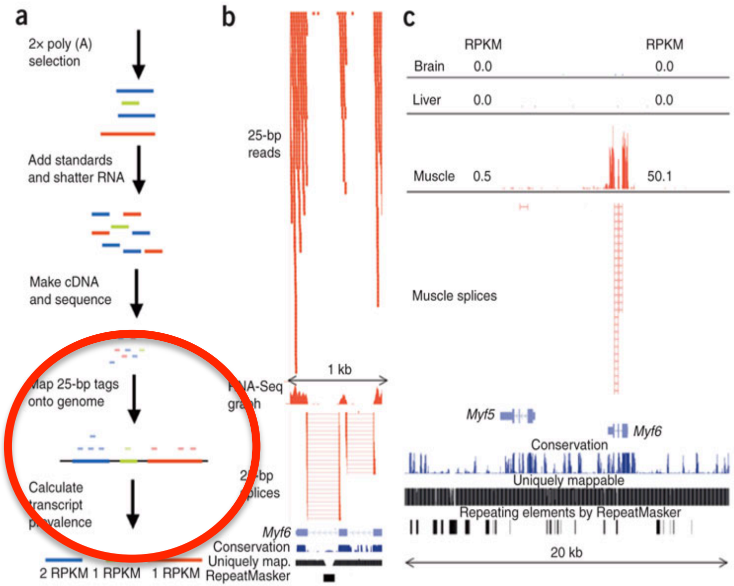
Short Read Alignment in Cloud Computing

Xuanyi Li
Xinyi Wang

Rationale

- Improvement in sequencing technology:
- Critical need to accelerate data analysis
- Short reads (30-50bp) alignment (RNA-seq)
    - Large input size
    - Parallel nature

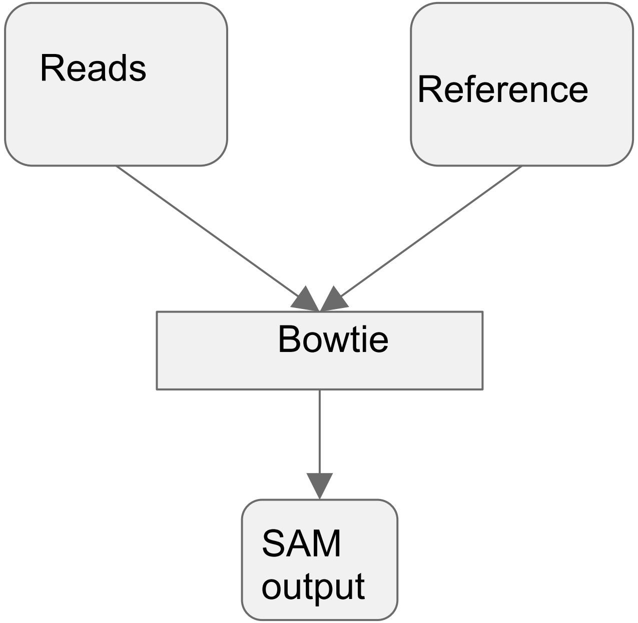
◆ **Goal: convert <u>bowtie</u>, an open-source short-read-aligner, to the Cloud.**



[1] Mortazavi, Ali, et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods* 5.7 (2008): 621-628.

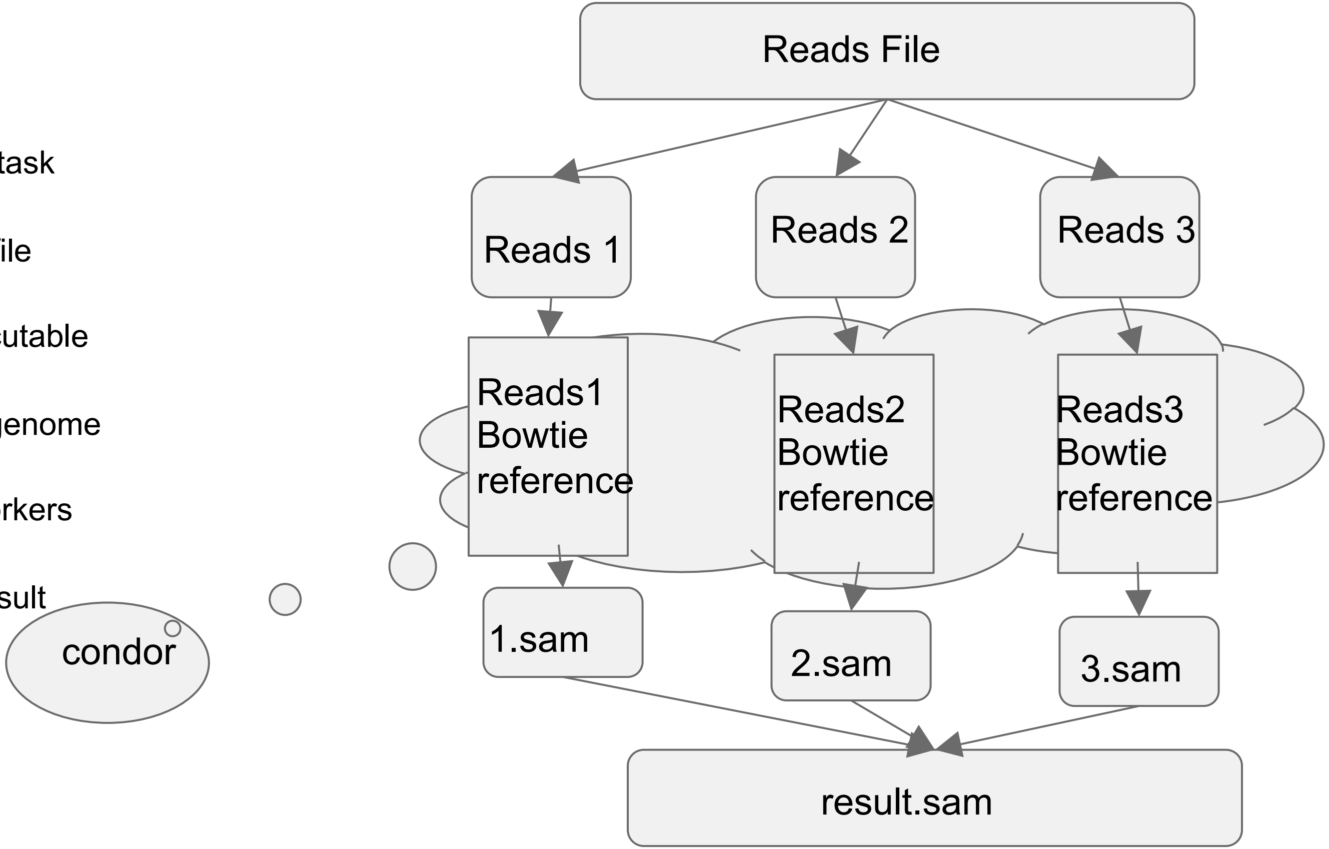
Bowtie

- Ultrafast, memory efficient, short aligner
- Works best with short reads and long reference genomes
- Output in SAM format
    - Read name, reference strand aligned to, string representing differences etc..
- Default parameters used:
    - (-k 1): output the first valid alignment encountered
    - Number of cores: 1



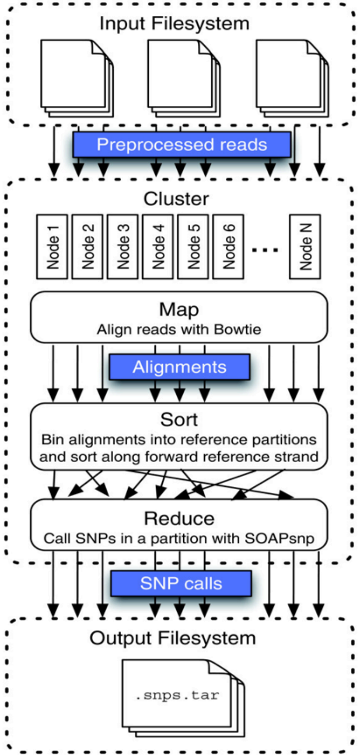
Approach one: Bowtie in workqueue

- Split reads files
- Each read file is a task
    - Small read file
    - Bowtie executable
    - Reference genome
- condor_submit_workers
- Combine output result



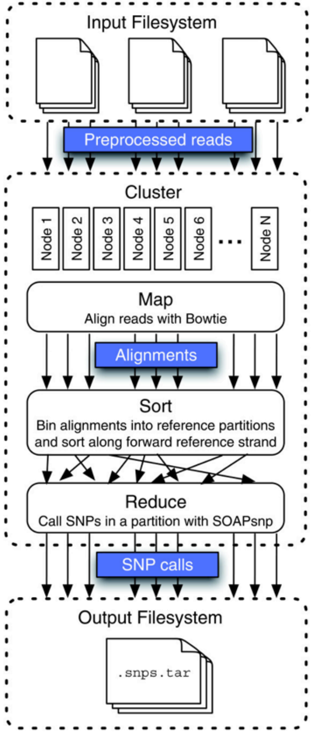
Approach two: Implementation with MapReduce

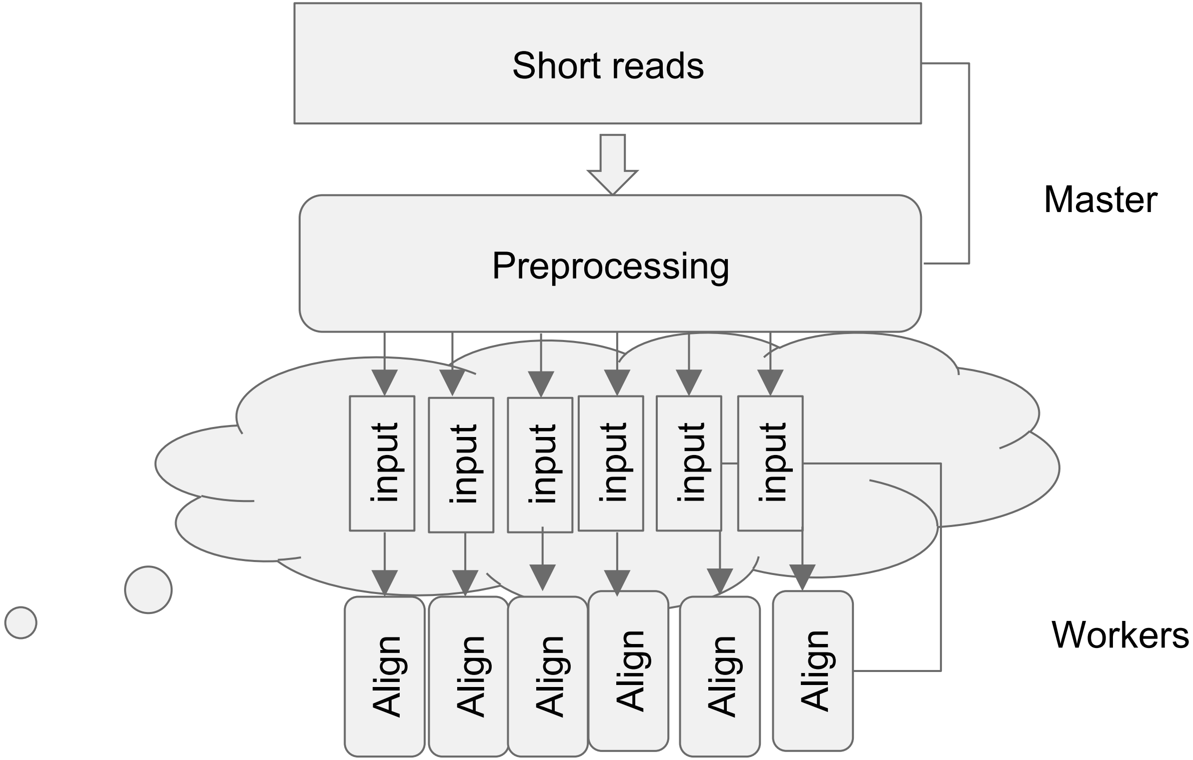
- Crossbow
  - Alignment & SNP calling
- Map & Reduce wrapper
  - MapWrap.pl
  - BinSort.pl
  - ReduceWrap.pl



1. Langmead, Ben, et al. "Searching for SNPs with cloud computing." *Genome Biol* 10.11 (2009): R134.

Implementation Crossbow





Testing Datasets

E_coli

Short Reads

Dataset	# of spots	# of bases	size
Small	8M	321.2M	**2.2Gb**
Full	20M	720M	**4.3Gb**

Reference Genomes

- Ecoli: **5.4 Mb**

Mouse

Short Reads

Dataset	# of spots	# of bases	size
Small	6M	485.9M	**3.2Gb**
Full	56M	4G	**26.6Gb**

Reference Genomes

- Mouse Chromosome 17: **29Mb**

Performance: scaling up input data

- 20-worker-runs

	Ecoli_small	Ecoli_full	Mouse_small	Mouse_full
Wq_bowtie	63.54(s)	162.54(s)	411.79(s)	1089.07(s)
Wq_crossbow	254.43(s)	498.63(s)	494.78(s)	1441.69(s)

size

Performance: scaling up number of workers

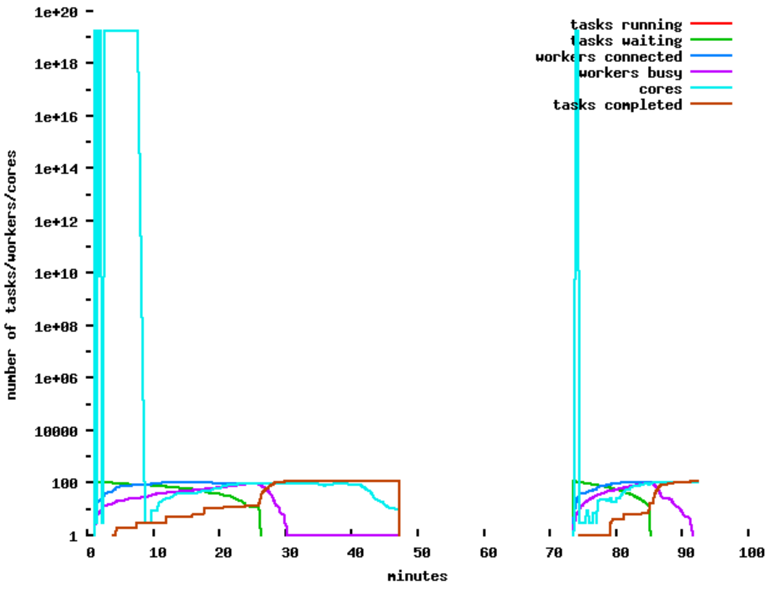
- Mouse-full-runs

	20 workers	50 workers	100 workers	150 workers
Wq_bowtie	1089.07(s)	663.71(s)	772.51(s)	739.64(s)
Wq_crossbow	1441.69(s)	2115.98(s)	1171.09(s)	2085.38(s)

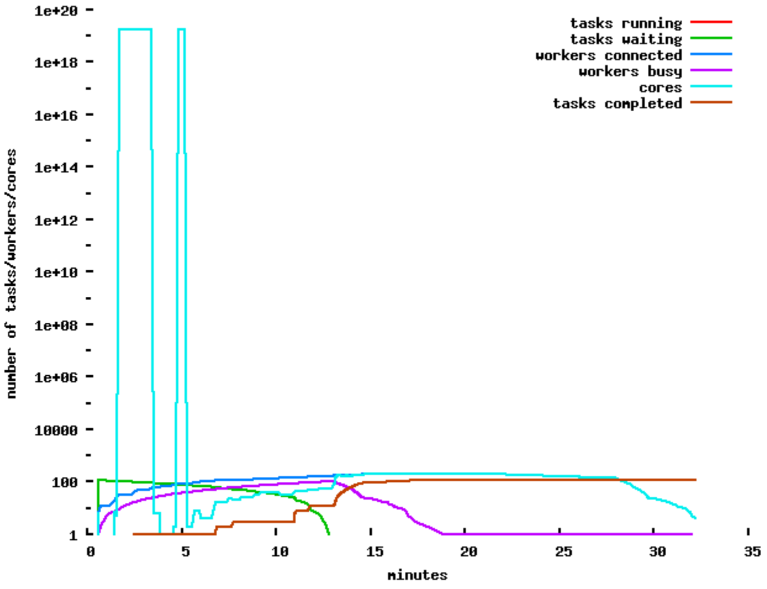
- Wq_Crossbow

	20 workers	50 workers	100 workers	150 workers
Speedup	21.81	14.85	26.85	15.07
Efficiency	1.09	0.30	0.27	0.10

Performance:stragglers



100-worker-run



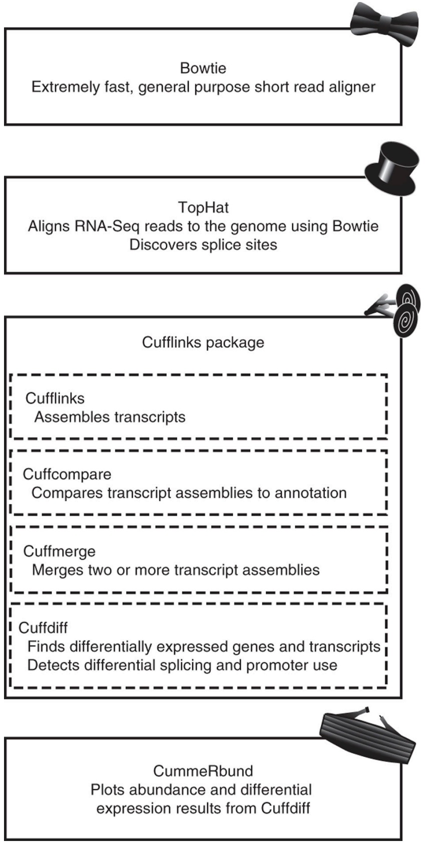
150-worker-run

Conclusion

- Short read alignment is suitable for the Cloud
- Problem with stragglers
- Wq_bowtie is faster; But wq_crossbow is part of an analysis pipeline in MapReduce model

Future direction

- Crossbow on hadoop
- Build wq_bowtie into a sequence analysis pipeline in cloud



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential
expression results from Cuffdiff

1. Trapnell, Cole, et al. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature protocols 7.3 (2012): 562-578.