

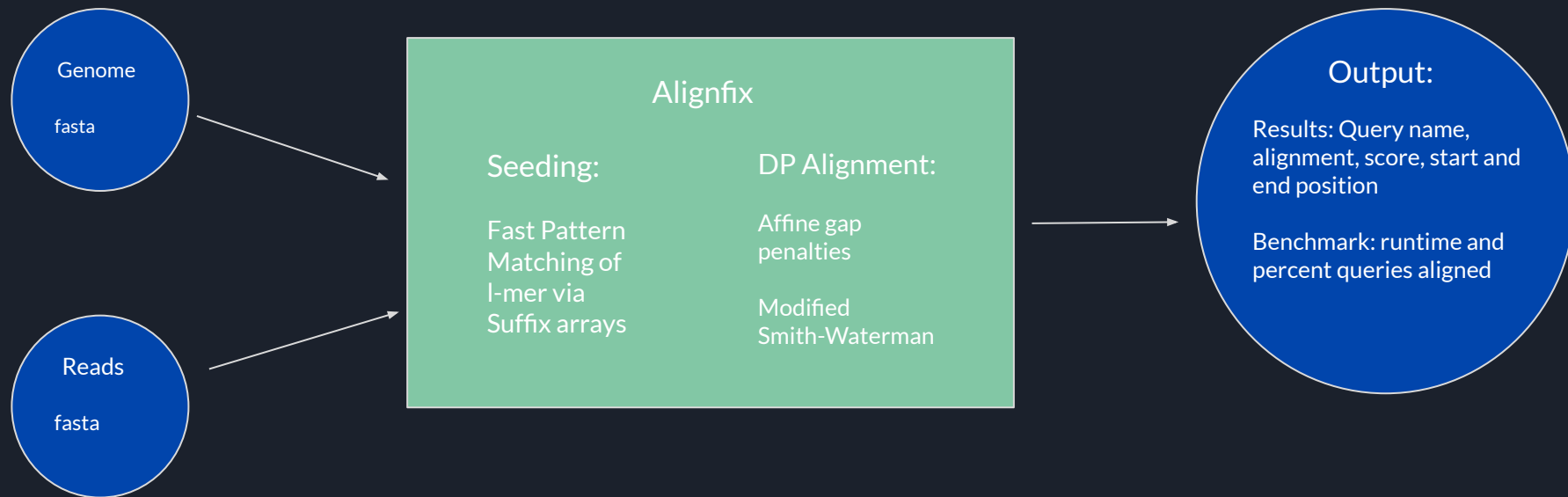


# AlignFix:

## A Seed and Extend Aligner

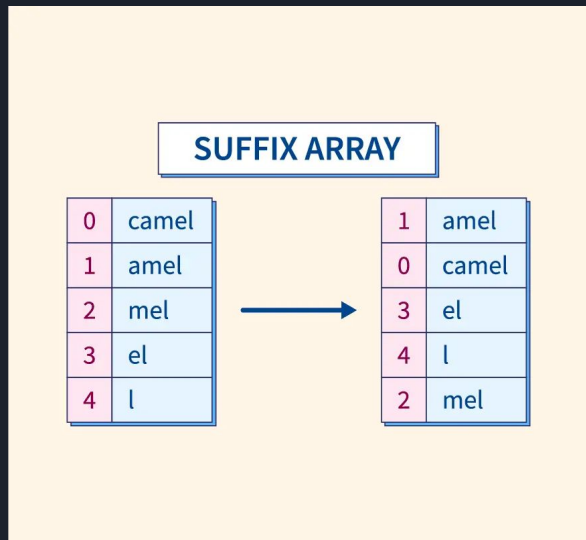
By: Siddharth Kaipa, Manish Sampath, Jason  
Chiu

# Overview of AlignFix



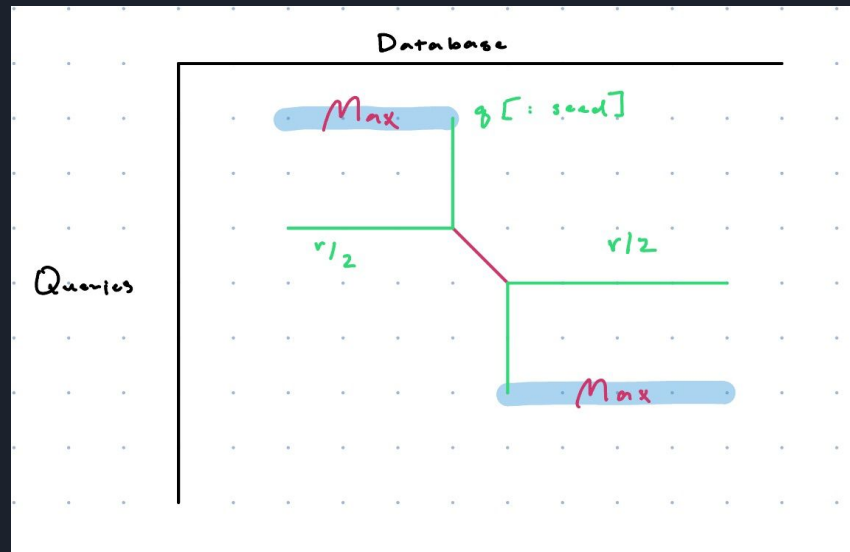
# Implementation Details and Algorithms

## Suffix Arrays



Given a keyword, you can apply binary search  
We stored indices in the database

## Alignment



Seed length -> 15  
r -> length of the query \* 2

For the top, we reverse the strings and compute the alignment

Finally, we start backtracking from the max of the bottom row

# Alignment

- Way of generating a score of the best alignment possible through a 2D matrix
- Backtracks from the best score to get the alignment
- Uses affine gap penalties: minimizes penalties for increasing the gap between nucleotides
- Together, an alignment is generated

		A	C	A	T	A	G
		0	0	0	0	0	0
A	0	1	0	1	0	1	0
A	0	1	0	1	0	1	0
T	0	0	0	0	2	1	0
G	0	0	1	0	1	1	2

Score = 288 bits (318), Expect = 2e-73  
 Identities = 262/325 (81%), Gaps = 8/325 (2%)  
 Strand=Plus/Plus

```

Query 1923 TCAGCCTACCATGAGAATAAGAGAAAAGA-AAATGAAGATCAAAAGCTTATTCATCTGTTT 1981
Sbjct 33774 TCAGACTACCTGAGAATAAGAGAAAAGAGAAATGAAGACCTAGA-CTTATCCATCTCTTT 33832

Query 1982 TTCTTTTCGTTGGTGTAAGCCAACACCCCTGTCTAAAAACATAAATTCCTTAAATCAT 2041
Sbjct 33833 TTCTTTTCGTTGGTTTAAACCAACCCCTGTCTAAAGTACACAAATTCCTTAAATAT 33892

Query 2042 TTTGCCCTCTTTTCTGCTGCTCAATTAA-AAAAAATGGAAGAATCTAATAGAGTGGT 2100
Sbjct 33893 TTTGCCCTCTTTTCTGCTGCTCAATTAA-AAAAAATGGAAGAATCTAATAGAGTGGT 33952
          Match=+2      Mismatch=-3

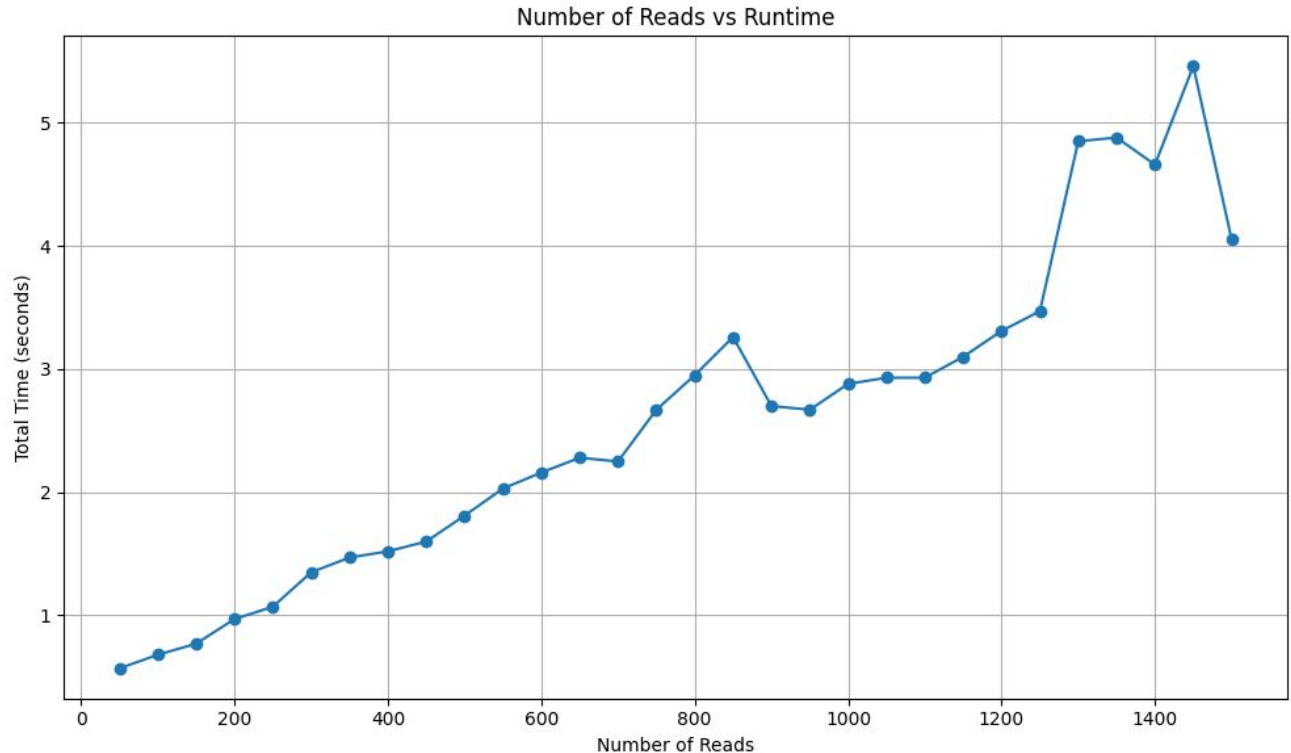
Query 2101 ACAGCACTGTTA-TTTTTCAAAGATGTGTGCTATCCTGAAAATTCGTAGGTTCTGTGG 2159
Sbjct 33953 CTATGACTGTTATTTTTTGAAGATGTGTGCTCAACCTGATAATTTGTAGGTTCTATGA 34012

Query 2160 AAGTTCACGTGT-          GGATTCTAGTTTCTTGTGGGCTA 2219
Sbjct 34013 AAATTCACATAT-          GGACTTCTAGTTCCTTCTGGATTA 34072
          Gap
          -(5 + 4(2)) = -13

Query 2220 AT-----TAAATAATCATTAACT 2240
Sbjct 34073 ATTGCATAAAAGAAACATTAATACT 34097
  
```

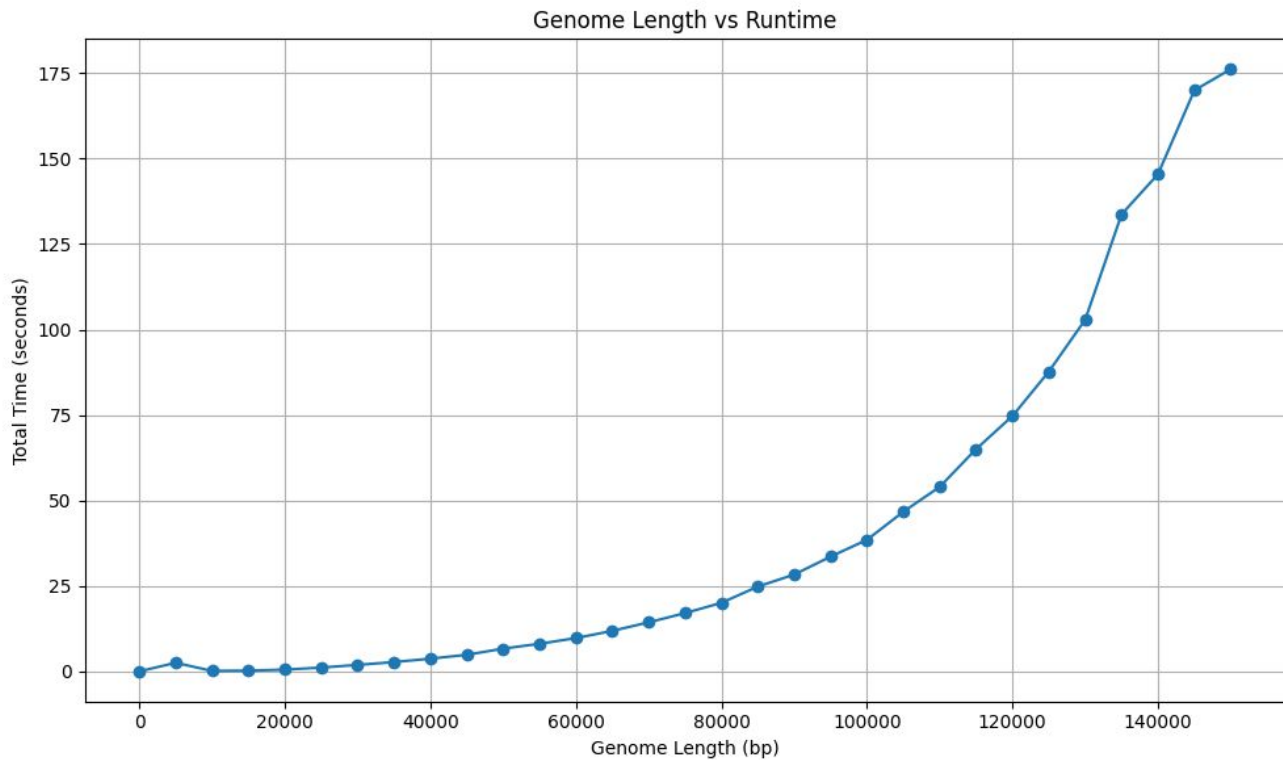
# Benchmarking

The graph shows approximately a linear relationship between the number of reads and runtime.



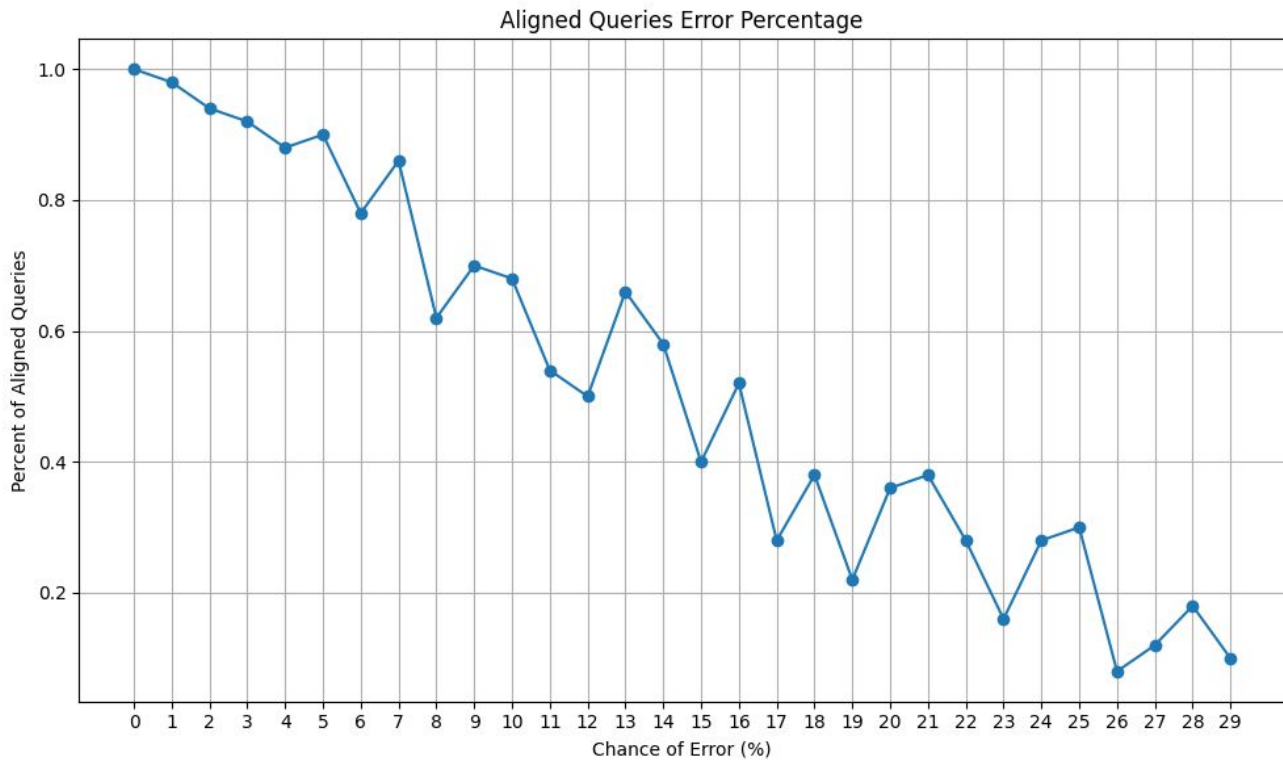
# Benchmarking

The graph shows approximately a quadratic relationship between genome length and runtime.



# Benchmarking

The graph shows approximately a linear relationship between the chance of error and percent of aligned queries.





# BWA MEM vs AlignFix

These results were taken based on the real world data.

Categories	AlignFix	BWA MEM
Time To Align	152.93 seconds	0.289 seconds
Percent of Queries Aligned	83%	99.93%





# Challenges

- Figuring out affine gap penalties
  - needs three matrices to keep track of
- Figuring out how pip install could be used to download AlignFix
  - Creating a package that can be used by the user



# Sources

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) “Basic local alignment search tool.” J. Mol. Biol. 215:403-410
- Heng Li, Richard Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics, Volume 25, Issue 14, July 2009, Pages 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324>