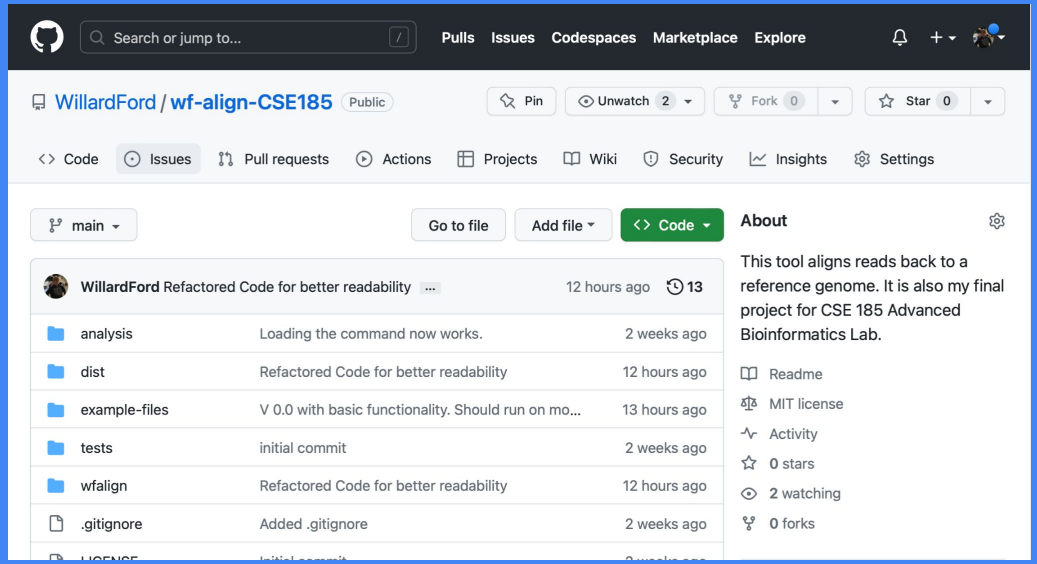
wf-align

A read aligner tool by Willard Ford



Current alignment tools are really good

Tool	Method	Run time	Space
BWA MEM	Maximal Exact Matches	Linear	Linear
Bowtie2	Dual Direction Backttracking	Almost Linear	Linear
STAR	Sequential Maximum Mappable Seed	Close to Linear	Linear

Article | Open Access | Published: 10 May 2023

A draft human pangenome reference

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

+ Show authors

Nature **617**, 312–324 (2023) | Cite this article

Reference: https://www.nature.com/articles/s41586-023-05896-x

wf-align aligns reads to reference genome

Inputs:

- Fasta Reference
- Fastq Reads

Outputs:

- SAM Alignment File

*Only aligns exact matches

output.sam

```
@HD VN:1.6 SO:unknown
ERR10000004.1840810 0 * 0 255 221M 0 0 221 CGTGCTAGCGCTATCATAGGTTGTAACCATACAGGTGTT
ERR10000004.1840811 0 * 0 255 217M 0 0 217 ATGTCTATGCAGATTCATTTGTAATTAGAGGTAATGAAG
ERR10000004.1840812 0 * 0 255 221M 0 0 221 TGTTGTTTGTAACAGTTTACTCACACCTTTTGCTCGTTG
ERR10000004.1840813 0 NC_045512.2 18245 255 221M 0 0 221 ACCCTAACATGTTTATCACCCGCGAAG
ERR10000004.1840814 0 * 0 255 221M 0 0 221 CAAACAAGCTAGTCTTAATGGAGTCACATTAATTGGAGA
ERR10000004.1840815 0 * 0 255 220M 0 0 220 CCGTGCTTTAACTGGAATAGCTGTTGAACAAGACAAAAA
ERR10000004.1840816 0 * 0 255 221M 0 0 221 TGGAAGAAACTAAGTTCCTCACAGAAAACTTGTTACTTT
ERR10000004.1840817 0 * 0 255 221M 0 0 221 ACAGATTTAATGTTGCTATTACCAGAGCAAAAGTAGGCA
ERR10000004.1840818 0 * 0 255 221M 0 0 221 GGTTACAGAGAAGGCTATTTGAACTCTACTAATGTCACT
ERR10000004.1840819 0 NC_045512.2 4924 255 221M 0 0 221 TGACAATCTTAAGACACTTCTTTCTTT
ERR10000004.1840820 0 * 0 255 221M 0 0 221 TTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTCAAC
ERR10000004.1840821 0 * 0 255 221M 0 0 221 GTGCGTTGTTCGTTCTATGAAGACTTTTTAGAGTATCAT
ERR10000004.1840822 0 * 0 255 220M 0 0 220 CAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCA
ERR10000004.1840823 0 * 0 255 200M 0 0 200 AGTGCAAATTTGTTATCAGCTAGAGGATGAAAAGGTGAA
ERR10000004.1840824 0 NC_045512.2 7382 255 220M 0 0 220 CAAATGGCCCCGATTTCAGCTATGGTT
ERR10000004.1840825 0 * 0 255 220M 0 0 220 ACATGATGAGTTAACAGGACACATGTTAGACATGTATTC
ERR10000004.1840826 0 * 0 255 221M 0 0 221 TGCTCAATACACTTCTGCACTGTTAGCGGGTACAATCAC
ERR10000004.1840827 0 * 0 255 221M 0 0 221 GTGCGTTGTTCGTTCTATGAAGACTTTTTAGAGTATCAT
ERR10000004.1840828 0 * 0 255 220M 0 0 220 CCAGGAACTAATCAGACAAGGAACTGATTACAAATATTG
ERR10000004.1840829 0 * 0 255 214M 0 0 214 GTGTATACTGCTGCCGTGAACATGAGCATGAAATTGCTT
ERR10000004.1840830 0 NC_045512.2 16113 255 221M 0 0 221 ACATGATGAGTTAACAGGACACATGTT
ERR10000004.1840831 0 NC_045512.2 16113 255 220M 0 0 220 ACATGATGAGTTAACAGGACACATGTT
```

Tool	Method	Run time	Space
wf-align	BWA Backtracking	O(Reads Length * Reference Length)	Linear

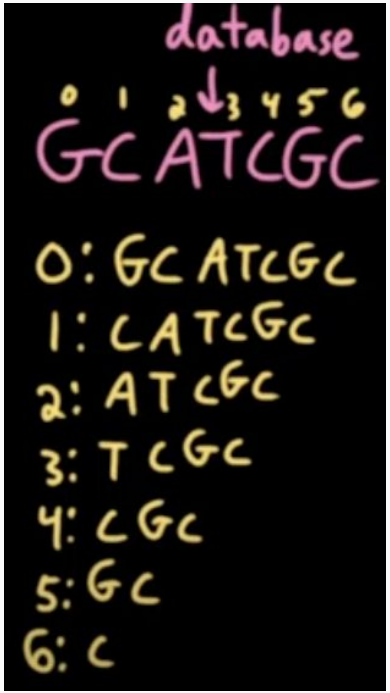
Building auxiliary data structures

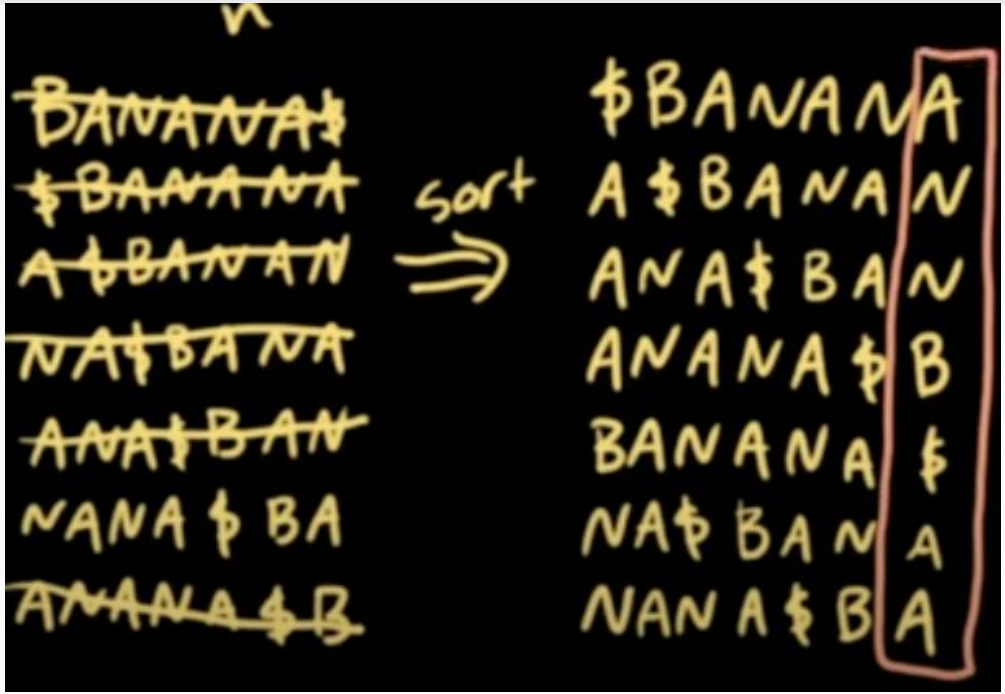
n = length of reference genome

Structure	Method	Run time	Space
Suffix Array	Prefix Doubling	O(nlogn)	O(n)
Burrows Wheeler Transformation	Directly Read from SA	O(n)	O(n)
Last to First Alignment	Radix Sort	O(n)	O(n)
Count Vector	Iterating through First Column of BWT Matrix	O(n)	O(1)

If time:

Suffix Array and Burrows Wheeler Transformation



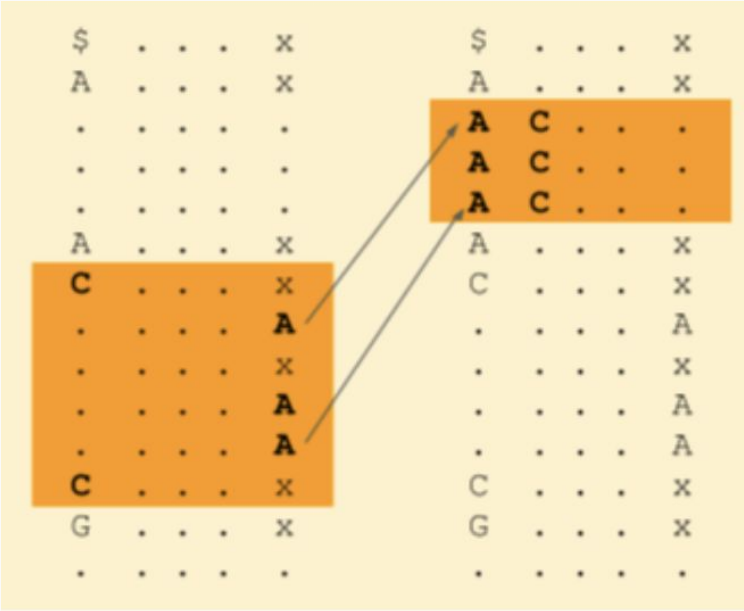


Searching for perfect matches

n = length of reference genome
m = length of reads

Method	Run time	Space
Exact Backtracking	O(m * n / (size of alphabet))	O(n / (size of alphabet))

By repeatedly grabbing valid matches in the BWT matrix we can narrow our search to the perfect alignments.



Metrics

Runtime Metrics:

- Total Time
- IO Time
- Searching Time

Accuracy Metrics:

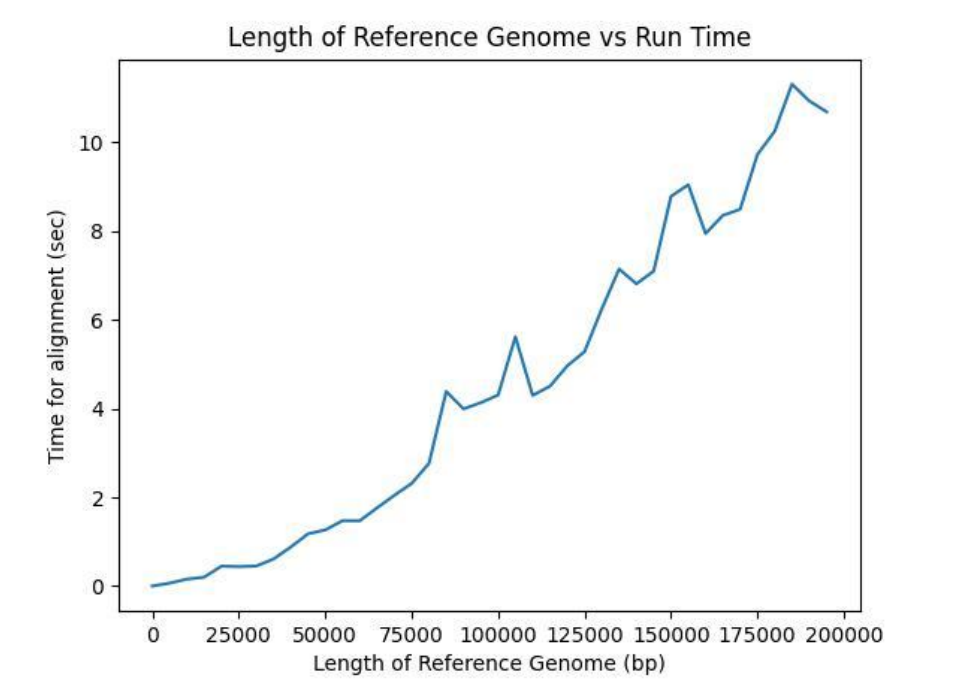
- Total Reads
- Percent Reads Mapped
- Percent Reads Uniquely Mapped

```
Metrics file wf-align run at 07:12:02

Runtime Metrics:
Total Time:	10.250350952148438
IO Reference Time:	 0.00072479248046875
Building Auxiliary Structures Time:	10.029198169708252
IO Reads Time:	 0.00031280517578125
Search Algorithm Time:	 0.22011518478393555

Accuracy Metrics:
Total Num Reads:	   50
Total Length of Reads:	 500
Total Length of Reference:	 180000
Percent reads aligned:	 0.14
Percent reads unaligned:	   0.86

```

Benchmarking

A number of analysis were performed that verify our results mentioned in previous slides.

Primarily we can see that the length of our reference genome linearly scales with the amount of time it takes to align.



SARS COV-2 Alignment

COVID reference genome and fastq files to generate metrics and sam file.

- 4.5 seconds
- ~20% of reads aligned

When the length of our reads are long exact matching filters out many reads. Found similar results with other SARS fastq files.

BWA MEM gets ~80% aligned across all SARS fastq files and near instantaneously

metrics.txt

```
Metrics file wf-align run at 07:09:17

Runtime Metrics:
Total Time: 4.658293962478638
IO Reference Time: 0.00040078163146972656
Building Auxiliary Structures Time: 0.3722350597381592
IO Reads Time: 0.013192176818847656
Search Algorithm Time: 4.272465944290161

Accuracy Metrics:
Total Num Reads:    5364
Total Length of Reads: 1157474
Total Length of Reference: 29903
Percent reads aligned: 0.18959731543624161
Percent reads unaligned:    0.8104026845637584
```

Reference Genome: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

References

SARS COV 2 Genome Analysis:

https://doi.org/10.1016/j.genrep.2020.100682

SARS COV 2 Reference Genome :

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

SARS COV 2 FASTQ Reads:

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR100/000/ERR10000000/ERR10000000.fastq.gz

Bowtie2 Information:

https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-aligner

BWA MEM:

https://bio-bwa.sourceforge.net/bwa.shtml

STAR Aligner:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/

Niema Moshiri CSE 100R Lectures:

https://www.youtube.com/watch?v=Lc-ACiJIrnM

https://www.youtube.com/watch?v=IzMxbboPcqQ