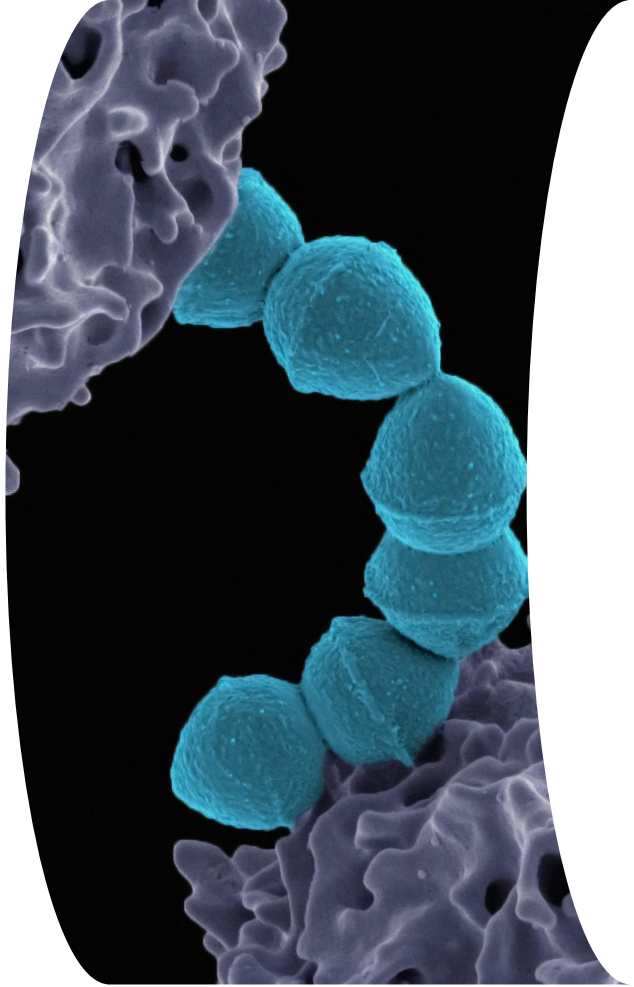
wellcome connecting science

ACORN

DNA READ ALIGNMENT



23 May 2024

[Virtual, Across Africa and Asia]

Arun Gonzales Decano
Senior Bioinformatician - University of Oxford

WCS ACORN - Bioinformatics for Antimicrobial Resistance - Virtual Course

Streptococcus Pyogenes
Photo by National Institute of Allergy and Infectious Diseases on Unsplash

Learning Objectives

1. Understand the fundamental principles of DNA read alignment and its significance in genomic analysis.
2. Gain familiarity with common algorithms and tools used for DNA read alignment.
3. Develop practical skills in performing DNA read alignments using bioinformatics software and interpreting alignment results for downstream analysis.

Read alignment is the process of mapping short DNA sequences, known as reads, to a reference genome or another set of sequences.

Importance of mapping reads

Fill gaps in incomplete assembled genomes.

Identify enrichment level of target sequence e.g. gene copy number.

Comparison between two or more genomes e.g. reference genome vs. sample.

Importance of mapping reads

Variant Calling: It helps identify genetic variations such as single nucleotide polymorphisms (SNPs) and insertions/deletions (indels).

Gene Expression Analysis: By aligning reads from RNA sequencing experiments, we can measure gene expression levels.

Structural Variation Detection: It enables the detection of large-scale genomic rearrangements like duplications, inversions, and translocations.

Workflow

1. **Preprocessing:** Quality filtering and adapter trimming to ensure high-quality reads.
2. **Alignment:** Mapping reads to the reference genome using advanced alignment algorithms.
3. **Post-processing:** Filtering and refining alignments to improve accuracy and reliability.
4. **Variant Calling:** Identifying genetic variants based on the aligned reads.

Applications

Genomic Medicine: Enabling personalized medicine by identifying disease-causing mutations.

Agricultural Genomics: Studying crop genomes for improved breeding and crop yield.

Microbiome Analysis: Characterizing microbial communities by aligning metagenomic reads.

Algorithms in Read Alignment: Requirements

Scalability: With the exponential growth of sequencing data, algorithms need to handle large datasets efficiently.

Accuracy: Alignment algorithms must produce accurate results to ensure reliable downstream analyses.

Flexibility: Different sequencing technologies and experimental setups require algorithms that can adapt to diverse data types.

Algorithms in Read Alignment: Seed-and-Extend

Seed Search: Identify short, exact matches (seeds) between the read and the reference genome.

Seed Extension: Extend the seeds to align the entire read, allowing for mismatches and gaps.

Algorithms in Read Alignment: Smith-Waterman

Dynamic Programming: It employs dynamic programming to find the optimal local alignment between the read and the reference genome.

Scoring System: Uses a scoring system to penalize mismatches, insertions, and deletions, producing accurate alignments.

Approximate string matching

Goal: all approximate matches between sequences where similarity score above threshold, or distance measure below a threshold

Similarity score: defined by optimal alignment which minimises number (or weight) of edits, or has maximum score

Distance threshold between two sequences

- Hamming distance: number of positions at which symbols are different (substitutions only)
- Levenshtein/edit distance: Number of single character deletions, insertions or substitutions required to transform one string into another
- Weighted edit distance: Different penalties for indels and substitutions

Source: Data Analysis Group at University of Dundee

Approximate string matching

AACTAGA-AC-TACTGA
AA-TACAGACTTAC-GA

- Matches: Black (12)
  Insertions/Deletions (INDELs): Blue (4)
  Mismatches: Red (1)
- Hamming distance = 1
- Edit distance = 5
- Different technologies need different scoring

Example from Canzar and Salzberg (2017)
http://doi.org/10.1109/JPROC.2015.2455551
Data Analysis Group @ University of Dundee

Algorithms in Read Alignment: Burrows-Wheeler Transform (BWT)

Compression: It rearranges the reference genome to exploit local similarities, resulting in a more compact representation.

FM Index*: The transformed data structure enables efficient pattern matching and alignment.

*FM (Full-text Minute-space) index

FM Index

BWT -> Suffix Array -> Wavelet Tree -> Backward Search

Advantages:

- Space efficient
- Versatile
- Versatility

Algorithms in Read Alignment: FM Index-based

Bowtie: A fast and memory-efficient algorithm for aligning short reads to large genomes.

BWA (Burrows-Wheeler Aligner): Combines the BWT with the FM index to achieve high-speed alignment while maintaining accuracy.

Algorithms in Read Alignment: Graph-based

De Bruijn Graphs: Represent the reference genome as a graph of overlapping k-mers, enabling efficient alignment of short reads.

Graph Searching: Algorithms like SOAPdenovo and Velvet use graph traversal techniques to align reads and assemble genomes.

Algorithms in Read Alignment: Hybrid Approaches

GASSST (Global Alignment Short Sequence Search Tool): Integrates seed-and-extend with FM index-based methods for improved speed and sensitivity.

Stampy: Utilizes both seed-and-extend and graph-based techniques to handle diverse sequencing data types.

Choosing a read mapper

- What is the best mapper?
  - There isn't one…
- Many options with different capabilities
  - Paired ends?
  - Read length?
  - Gapped alignment?
  - Use quality scores?
  - Splice aware?
- Choice should be guided by your experiment

Choosing a read mapper

Speed: With the ever-increasing volume of sequencing data, alignment algorithms need to be fast and efficient.

Accuracy: Alignments must be accurate to ensure reliable downstream analyses.

Handling Variability: Genomes exhibit variations, such as mutations and structural changes, which must be accounted for during alignment.

Pairwise Sequence Alignment (PSA)

Pairwise Sequence Alignment is identifying regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

https://www.ebi.ac.uk/Tools/psa/

Pairwise Sequence Alignment (PSA)

- Alignment of two sequences originally carried out…
  - To identify subsequences which are positionally similar
  - To uncover evolutionary relationships
- First optimal alignment of two sequences: Needleman and Wunsch (1970)
  - Dynamic programming
  - Global alignment:
    - optimal alignment of full length of sequences
    - Attempt to align every base between sequences
    - Conserved domains may not be aligned

```
--T--CC-C-AGT--TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || |       | | ||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT--C
```

Data Analysis Group @ University of Dundee

Local alignments

- Optimal alignment of substring within sequences
- Useful for identifying locally conserved regions

```
ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA
                    |||| |||||| ||||||||||||||
                    TACTCACGGATGAGGTACTTTAGAGGC
```

- Smith-Waterman (1981) modification of Needleman-Wunsch

Data Analysis Group @ University of Dundee

Alignment scoring

- How do you asses if one alignment is better than another?
- Scoring – reward similarity, penalise differences

```
GCATCATCTCCG
||| |||| ||||
GCTTCATGTCCG
```

- i.e. (from BLAST):
  - Match: score +1
  - Mismatch: score -2

- 10 matches and 2 mismatches = $(10 \times 1) + (2 \times -2) = 6$

Data Analysis Group @ University of Dundee

Tools for PSA

Basic Local Alignment Search Tool (BLAST)
Needle (EMBOSS)
BWA-MEM

Alignment file formats

Three main formats used for storing alignments:

SAM (Sequence Alignment Map)

- Plain text
- Uncompressed

BAM (Binary Alignment Map)

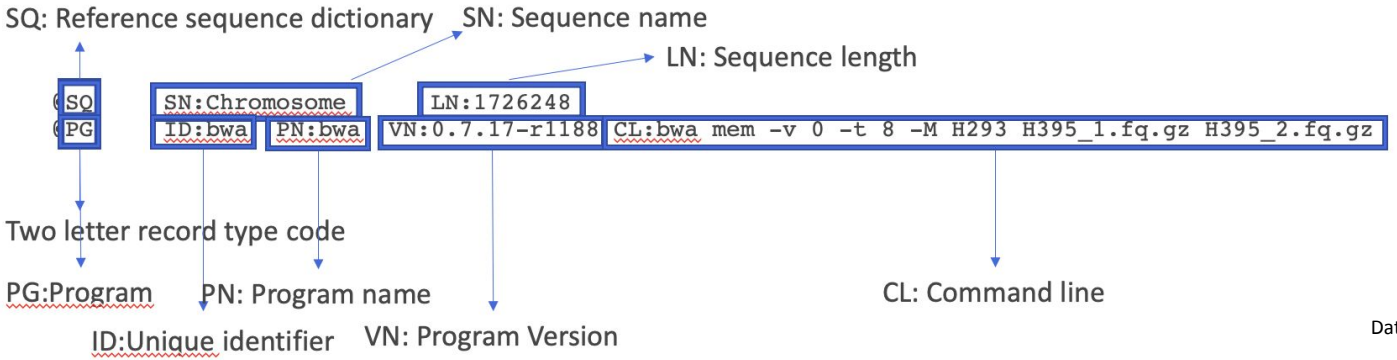
- Binary equivalent of SAM
- Compressed
- Can be indexed to allow random access

CRAM

- Uses Reference-based compression
- 45-50% space saving over BAM
- Not currently as well supported by tools

SAM Format

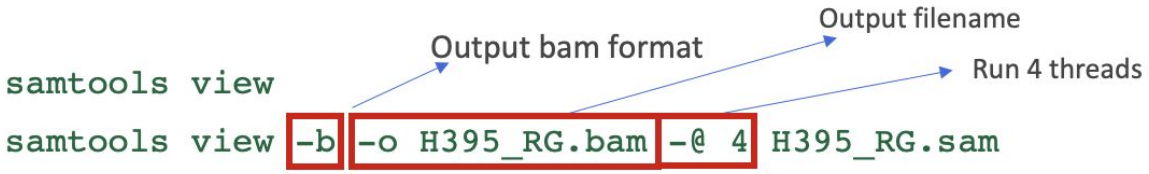
- Official definition document: https://samtools.github.io/hts-specs/SAMv1.pdf
- File contains two sections: header (optional) and alignments
- Human readable(ish) text format
- Header lines start with an '@'
- H395.sam contains two header lines:



Data Analysis Group @ University of Dundee

BAM Format

- Binary equivalent of SAM
- Much more compact – should always be used in preference to SAM
- Not human readable



Data Analysis Group @ University of Dundee

Multiple Sequence Alignment (MSA)

MSA employs progressive alignment method whereby pairwise alignment algorithm is used iteratively:

- First align the most closely related pair of sequences, then the most similar one to that pair, and so on.

https://www.majordifferences.com/

Tools for MSA

- MUSCLE: www.ebi.ac.uk/Tools/msa/muscle/
- T-Coffee: www.tcoffee.crg.cat/
- MAFFT: www.mafft.cbrc.jp/alignment/server/
- CLUSTALW: www.ebi.ac.uk/Tools/msa/clustalo/

Choosing a Reference Genome

What to align your reads to…

- Genome? Transcriptome?...it depends…
- Think about your experiment
- Include full range of sequences likely to be in samples
  - EBV? Host genome?

Your sample will almost never be identical to the reference sequence

- What is the closest related sequence?
- Is this the most useful? How complete is it?

Starting point is a fasta formatted representation of your reference

Beware of different genome versions

- Version and source should be reported in publications

Data Analysis Group @ University of Dundee

References

1. Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.
2. Burrows, M., & Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. Digital Equipment Corporation.
3. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.
4. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754-1760.

wellcome connecting science

ACORN

questions?