

Alignment-free Sequence Analysis Methods

Hector Espitia
hspitia@gatech.edu

PhD Student | Bioinformatics | Jordan Lab

Computational Genomics 2018 – Georgia Institute of Technology
Atlanta, 8th February, 2017

Outline

- Background
 - Sequence Similarity
 - Sequence Alignment (generalities, drawbacks)
- Alignment-free Methods
 - Classification
 - NGS Data Analysis
- STing: an Alignment-free Application
 - Sequence Typing
 - Multilocus Sequence Typing (MLST)
 - Performance (Typing, Gene Detection)
- Conclusions

Background

Sequence Similarity

- Knowledge derived from sequence similarity.
- Similar sequences tend to share features.
- Similarity: functional, structural and evolutionary inferences.

Sequence Alignment

- Sequence Alignment is a very useful “tool”: provides a similarity measure.
- 80’s-90’s: BLAST, FASTA, MAFFT, Muscle, ClustalW, PSI-BLAST, HMMER/Pfam, Mauve, BLASTZ, TBA.

Alignment-based Analysis Drawbacks

- Assumption of linearity and conservation in stretches of homologous sequences.
- Poor accuracy of alignment when sequence identity is below a critical point.
- Depends on multiple evolutionary assumptions about the sequences.

Alignment-based Analysis Drawbacks

- Computationally expensive (RAM and processing time).
- Not ideal for NGS-era (not scalable).

NGS-era requires rapid and accurate analysis at a high scale
(complete genomes, billions of sequences)

Alignment-free Methods

Alignment-free Sequence Analysis

“Any method that quantifies sequence similarity without producing/using alignment at any step of the algorithm application”

Zielezinski et al., 2017

Advantages:

- Less computationally expensive.
- Resistant to shuffling and recombination events.
- Evolutionary assumptions-free.

Alignment-free Sequence Analysis

“Any method that quantifies sequence similarity without producing/using alignment at any step of the algorithm application”

Zielezinski et al., 2017

Advantages:

- Less computationally expensive.
- Resistant to shuffling and recombination events.
- Evolutionary assumptions-free.

REVIEW

Open Access

Alignment-free sequence comparison:
benefits, applications, and tools



Andrzej Zielezinski¹, Susana Vinga², Jonas Almeida³ and Wojciech M. Karlowski^{1*}

Zielezinski et al. *Genome Biology* (2017) 18:186
DOI 10.1186/s13059-017-1319-7

Classification of Alignment-free Methods

- Word frequency-based, and
- Information-theory based.

- Other alignment-free methods:
 - Chaos game representation
 - Iterated maps
 - Graphical representation of DNA

Word Frequency-based Methods

Depend on the amount of shared words/ k -mers between sequences.

4-mer: ATTCGTC **CAAG**ATCTATG

Three steps:

- k -mer extraction and grouping.
- Frequencies quantification.
- Dissimilarity quantification.

Word Frequency-based Methods

Depend on the amount of shared words/ k -mers between sequences.

4-mer: ATTCGTC **CAAG** ATCTATG

Three steps:

- k -mer extraction and grouping.
- Frequencies quantification.
- Dissimilarity quantification.

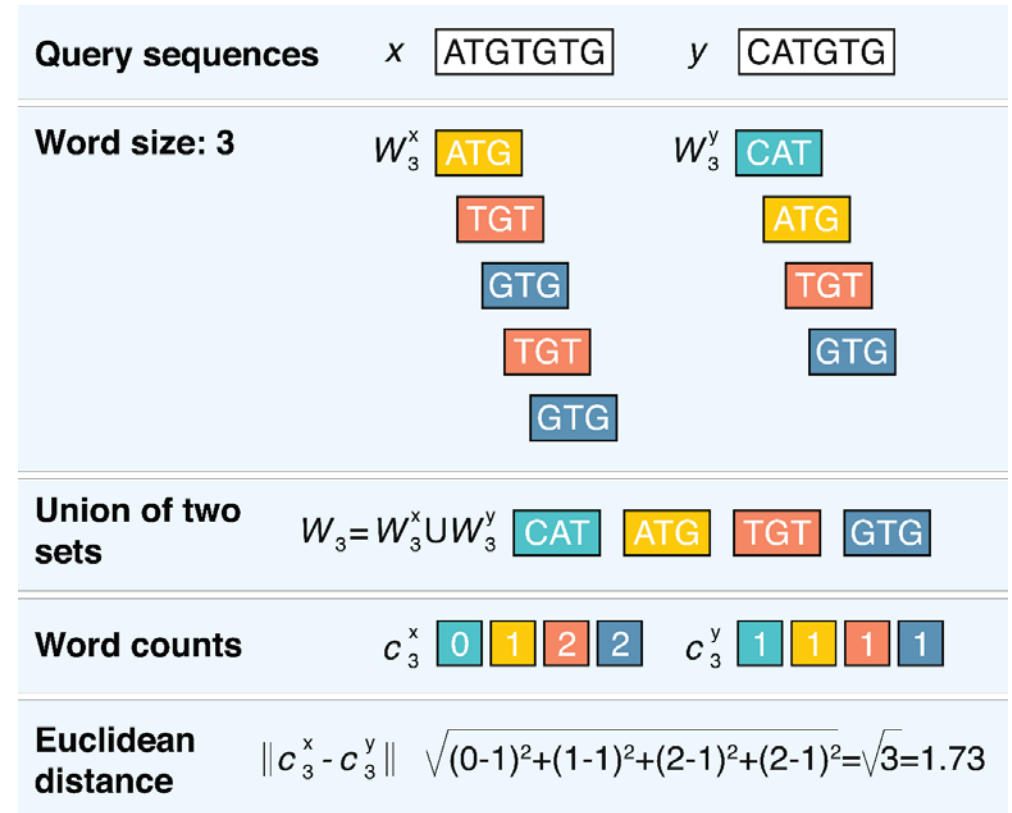


Fig. 1 Alignment-free calculation of the word-based distance between two sample DNA sequences ATGTGTG and CATGTG using the Euclidean distance.

Zielezinski et al., 2017

Information-theory Based Methods

Depend on the amount of shared information (complexity/entropy).

p

C	C	C	C	C	C	C	C
---	---	---	---	---	---	---	---

 less complex

q

C	G	A	T	G	T	G	A
---	---	---	---	---	---	---	---

 more complex

Two steps:

- Complexity calculation.
- Dissimilarity quantification.

Information-theory Based Methods

Depend on the amount of shared information (complexity/entropy).

p  less complex

q  more complex


Two steps:

- Complexity calculation.
- Dissimilarity quantification.

Query sequences

x  y  xy 

Lempel-Ziv complexity


 $c(x)=4$ $c(y)=5$ $c(xy)=7$

Normalized compression distance

$$\frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} = \frac{7-4}{5} = 0.6$$

Fig. 2 Alignment-free calculation of the normalized compression distance using the Lempel–Ziv complexity estimation algorithm. LempelZiv complexity counts the number of different words in sequence when scanned from left to right (e.g., for $s = \text{ATGTGTG}$, Lempel–Ziv complexity is 4: A|T|G|TG). Description of compression algorithms in alignment-free analysis has been reviewed extensively.

Zielezinski et al., 2017

Alignment-free Methods in NGS Data Analysis

- Transcript identification (Kallisto, Sailfish, Salmon).
- Genomic variability profiling (FastGT, LAVA).
- Assembly: error correction (Quorum, Lighter, Trowel), overlapping (MHAP algorithm, Miniasm), and scaffolding (LINKS).
- Metagenomics: species identification/taxonomic profiling (Kraken, CLARK, MASH, stringMLST, STing, Taxonomer).
- Phylogenetics (AAF, NGS-MC, kSNP).

Zielezinski *et al. Genome Biology* (2017) 18:186
DOI 10.1186/s13059-017-1319-7

Alignment-free for Research Purposes

Sequence similarity

- CAFE (desktop, GUI)
 - 28 distance measures.
 - Dissimilarity matrices.
 - Dendrograms, heatmaps, PCA and networks.
- Alfree (Web)
 - 38 distance measures.
 - Fully automated analysis.
 - Consensus phylogenetic tree.

STing

STing (Sequence Typing)

- A lightweight, alignment- and assembly-free application for the NGS era, that belongs to the group of word frequency-based methods.
- Two functionalities for NGS sample analysis

Sequence Typing

Prediction of the
Sequence Type (ST)

Gene Detection

Prediction of
presence/absence of a
gene of interest

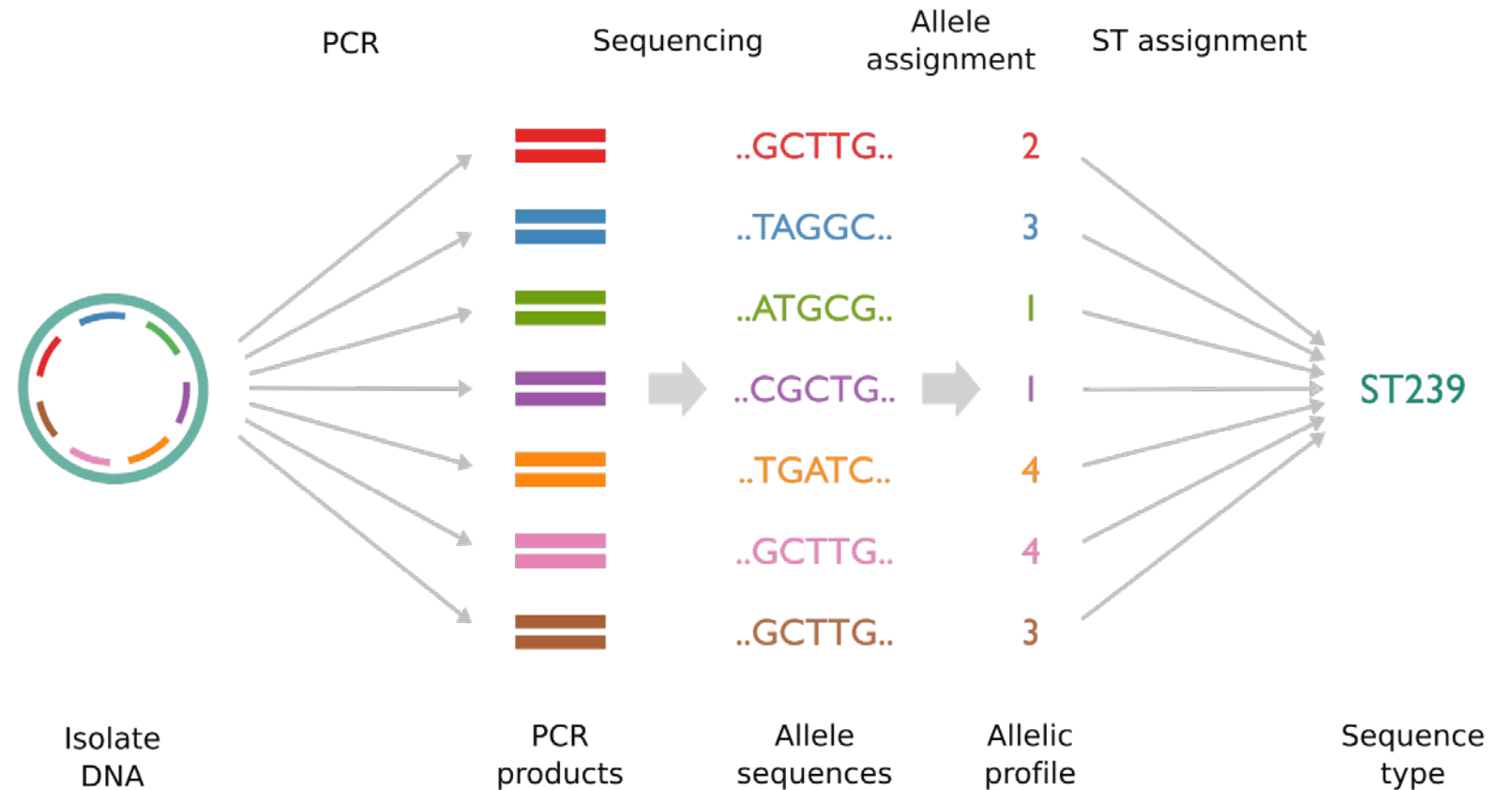
Sequence Typing

- Identifying organisms within a species.
- Human pathogens of one species can comprise very diverse set of organisms.
- Typing technique must have a good discriminatory power.

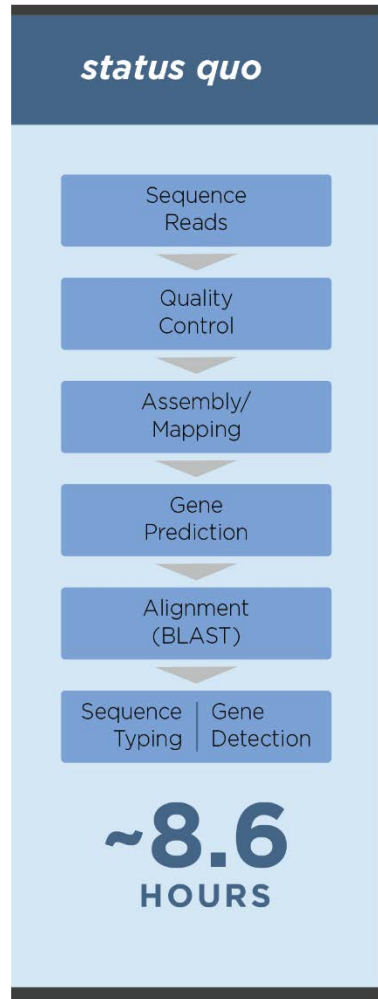


Multilocus Sequence Typing (MLST)

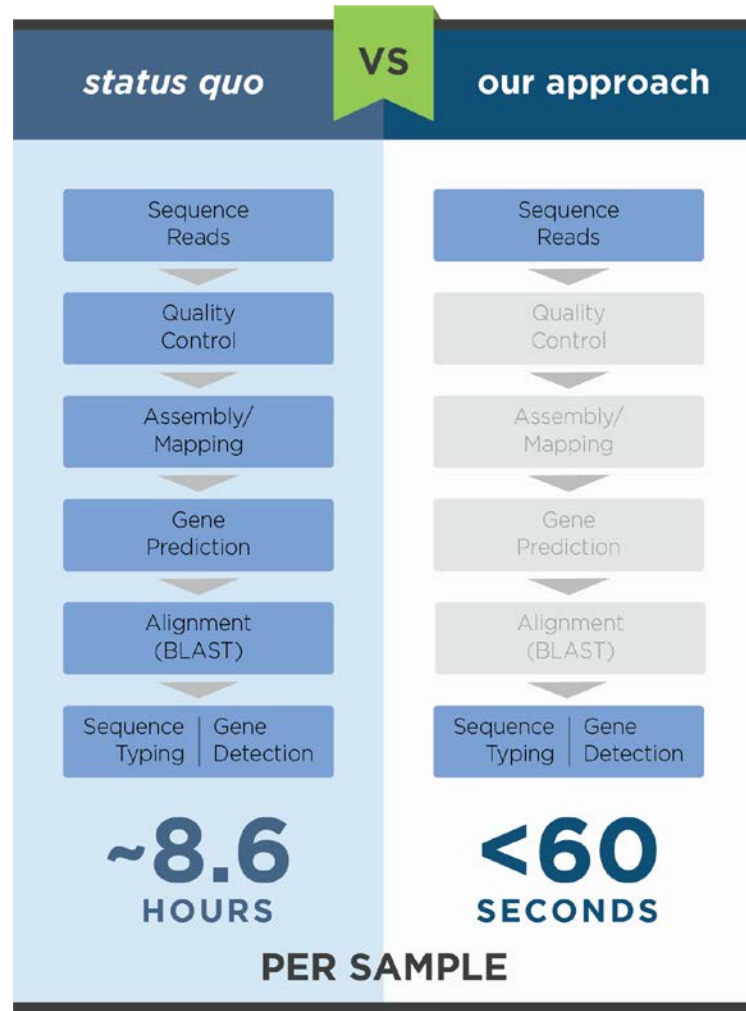
- Pre-NGSera
- Gene-based approach (7 housekeeping)
- Extensive information available (PubMLST, MLST.net)



MLST: Computational Methods with NGS Data

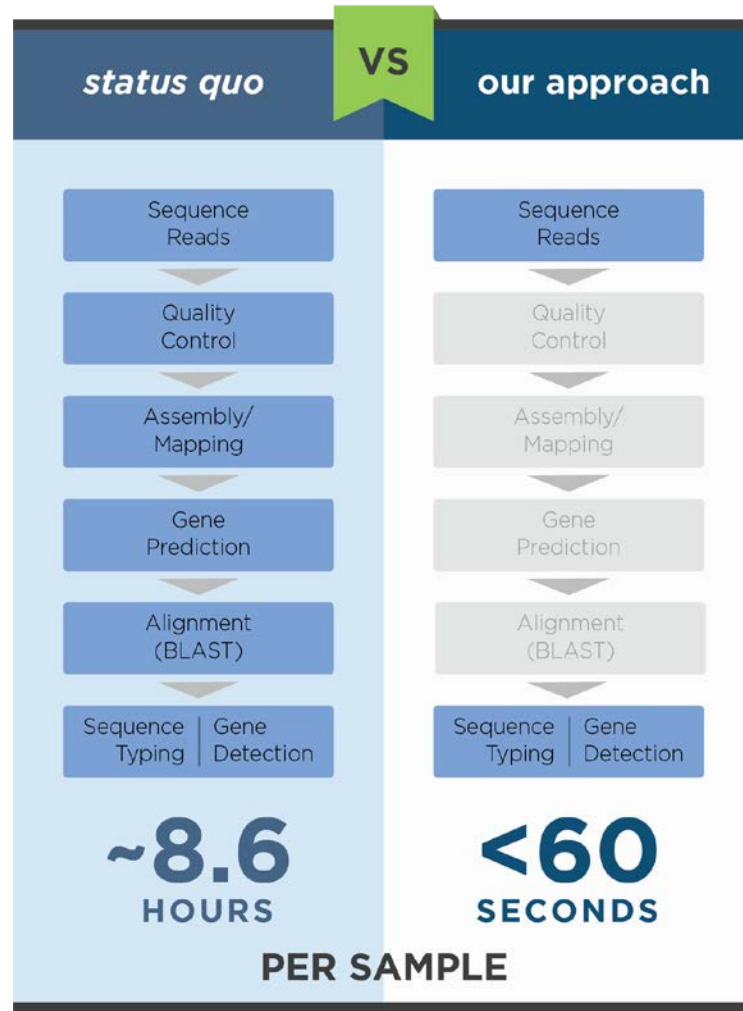


MLST: Computational Methods with NGS Data



Alignment- and assembly-free with minimum expertise and time required.

MLST: Computational Methods with NGS Data



Alignment- and assembly-free with minimum expertise and time required.

Genome analysis

stringMLST: a fast k-mer based tool for multilocus sequence typing

Anuj Gupta^{1,2}, I. King Jordan^{1,2,3} and Lavanya Rishishwar^{1,2,3,*}

Bioinformatics, 33(1), 2017, 119–121

doi: 10.1093/bioinformatics/btw586

Advance Access Publication Date: 7 September 2016

Application Note

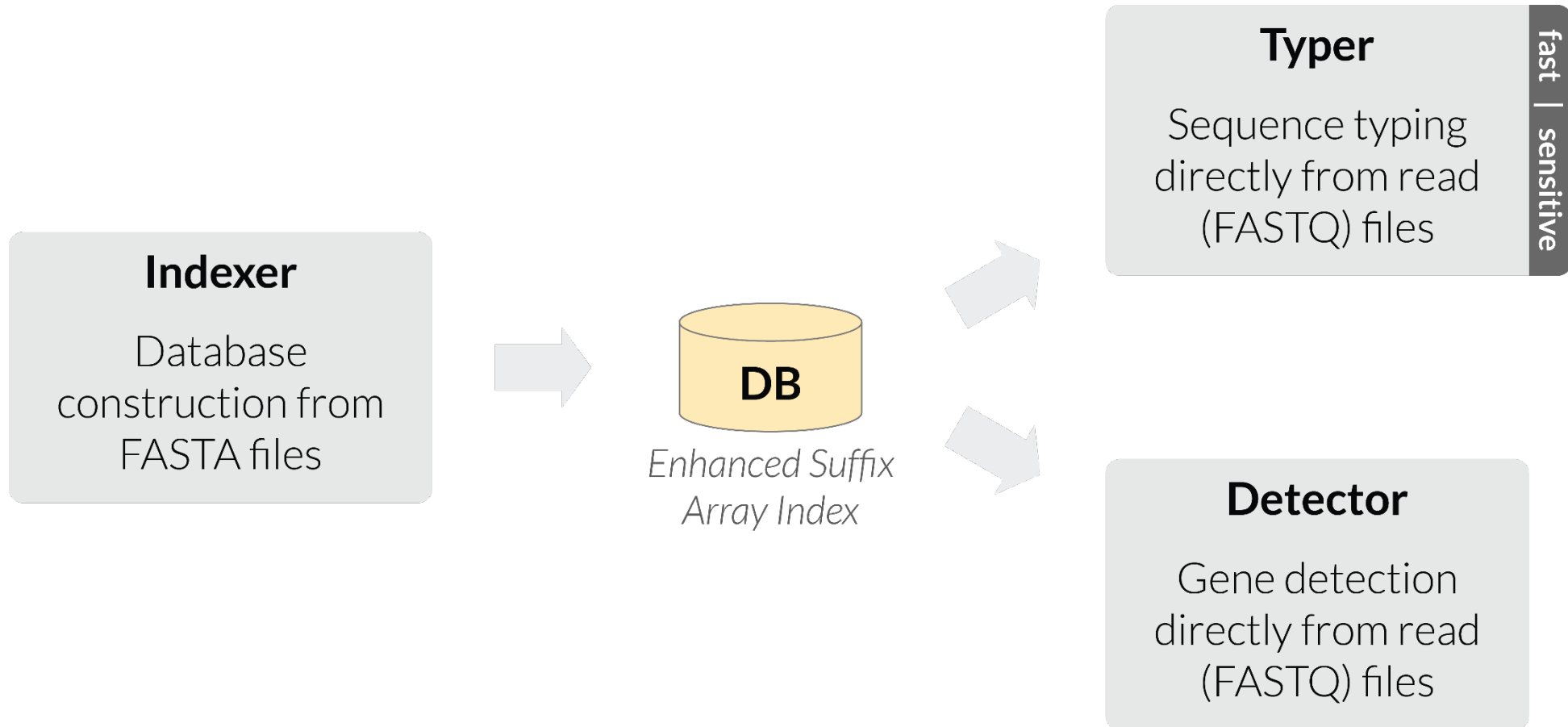
STing

- Addresses the shortcomings of its predecessor (stringMLST): speed and RAM consumption on larger typing schemes (rMLST, cgMLST).
- Uses Enhanced Suffix Arrays as core algorithm data structure.

Quick
determination of the
membership of an
input string

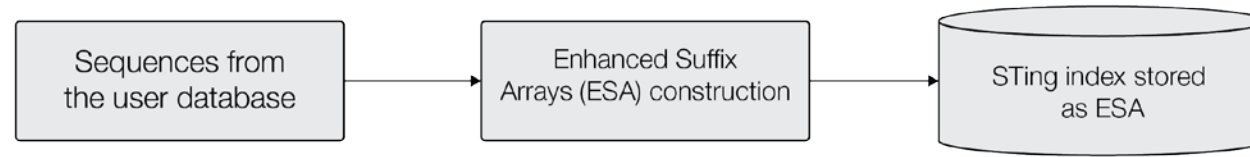
Search time
depends on
query length, not on
the DB size

STing - Structure



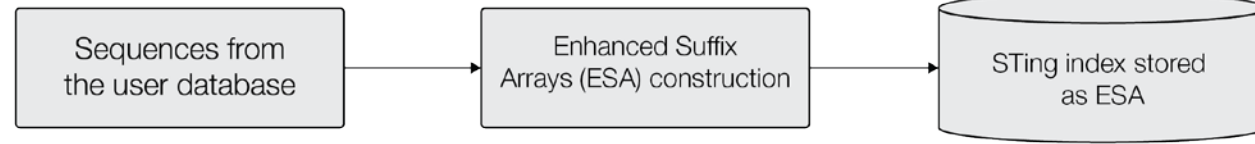
Algorithm Overview

(A) Database Indexing



Algorithm Overview

(A) Database Indexing

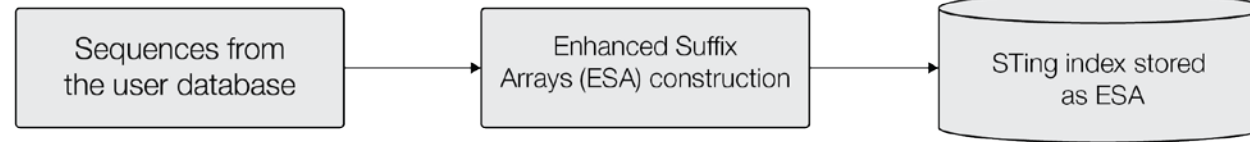


Typer



Algorithm Overview

(A) Database Indexing

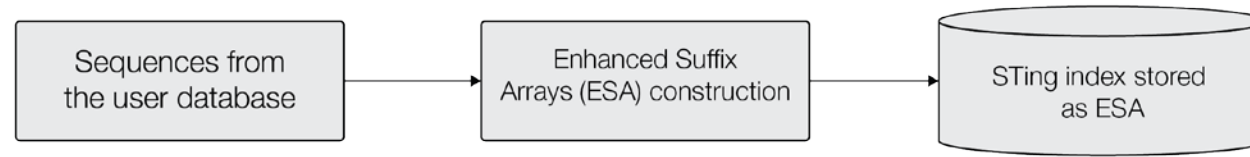


Detector

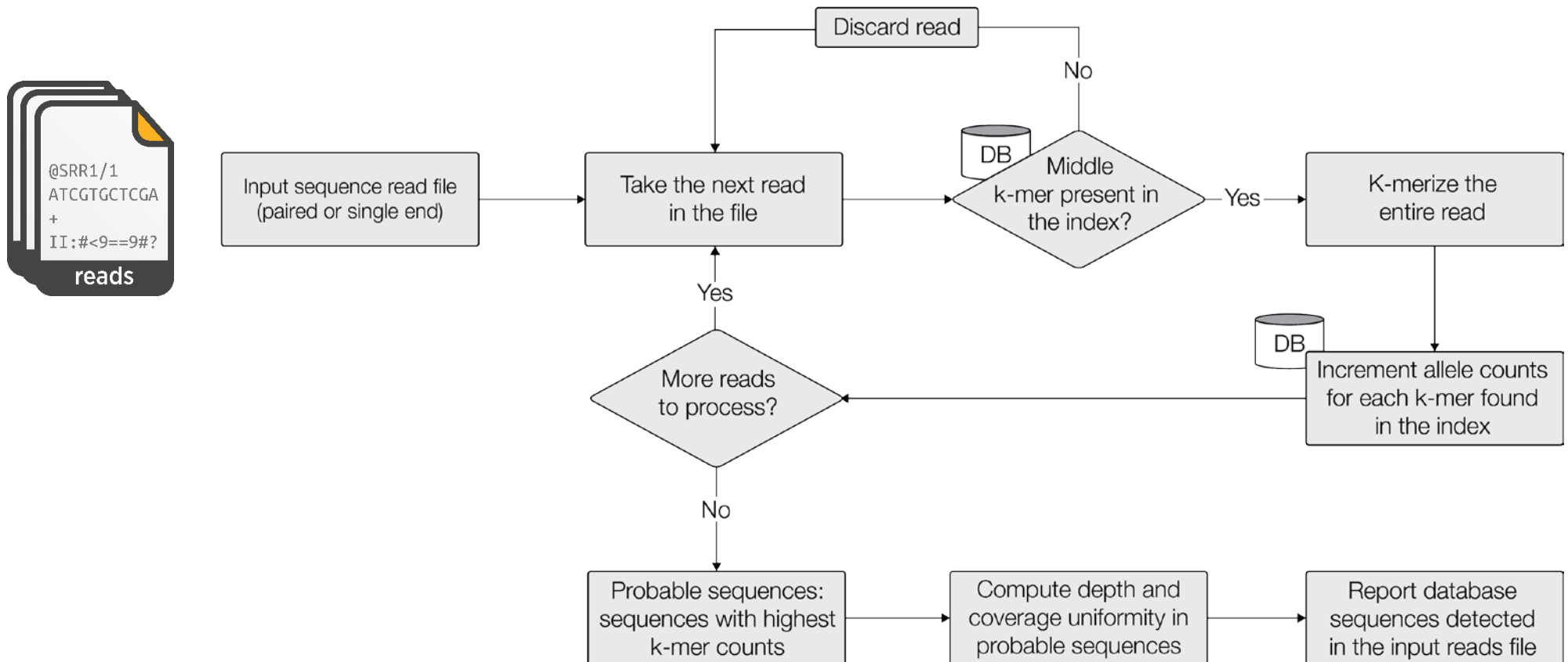


Algorithm Overview

(A) Database Indexing

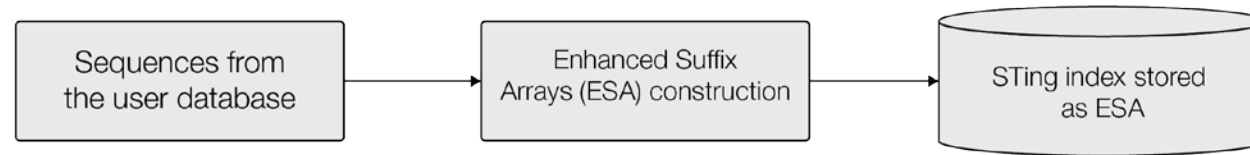


(B) Sequence Variant Detection

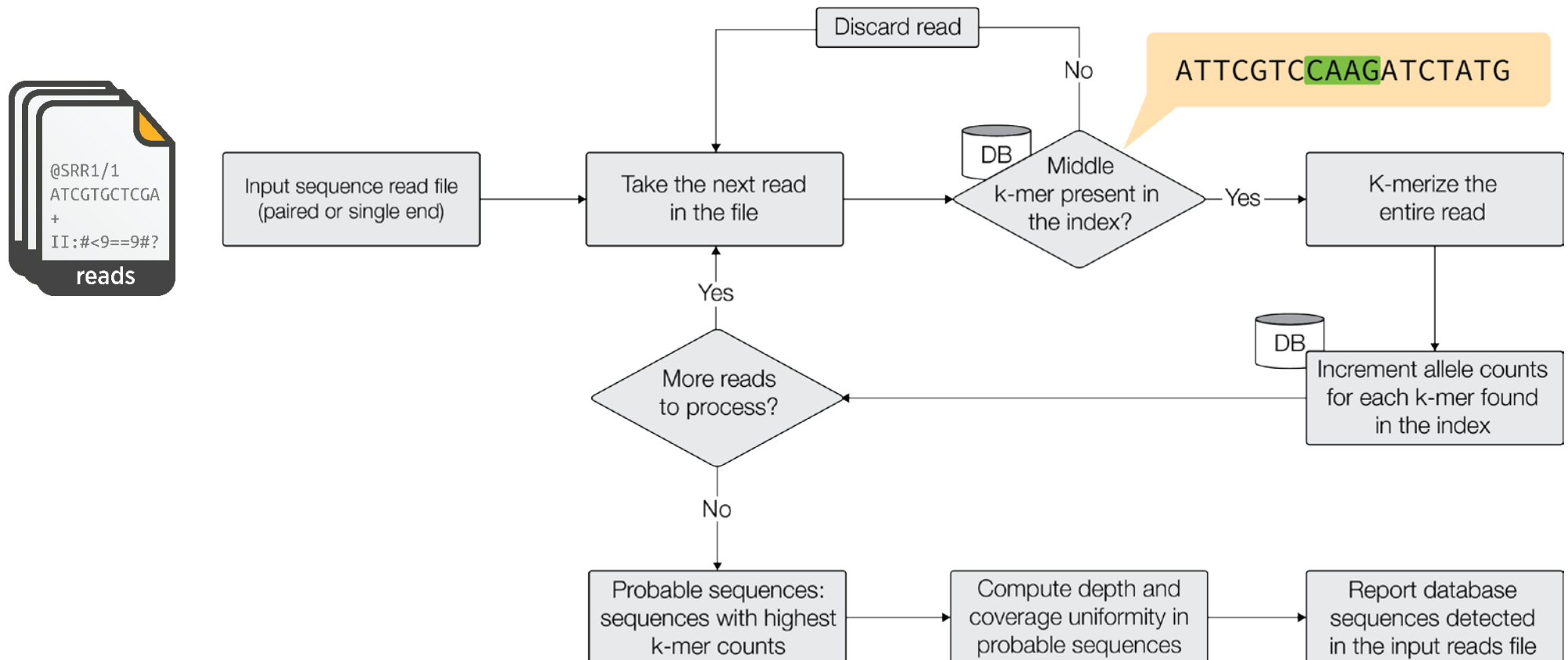


Algorithm Overview

(A) Database Indexing

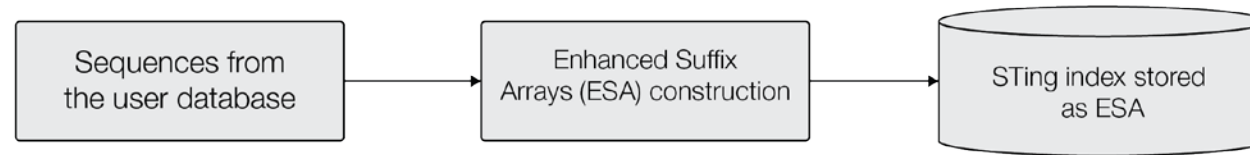


(B) Sequence Variant Detection

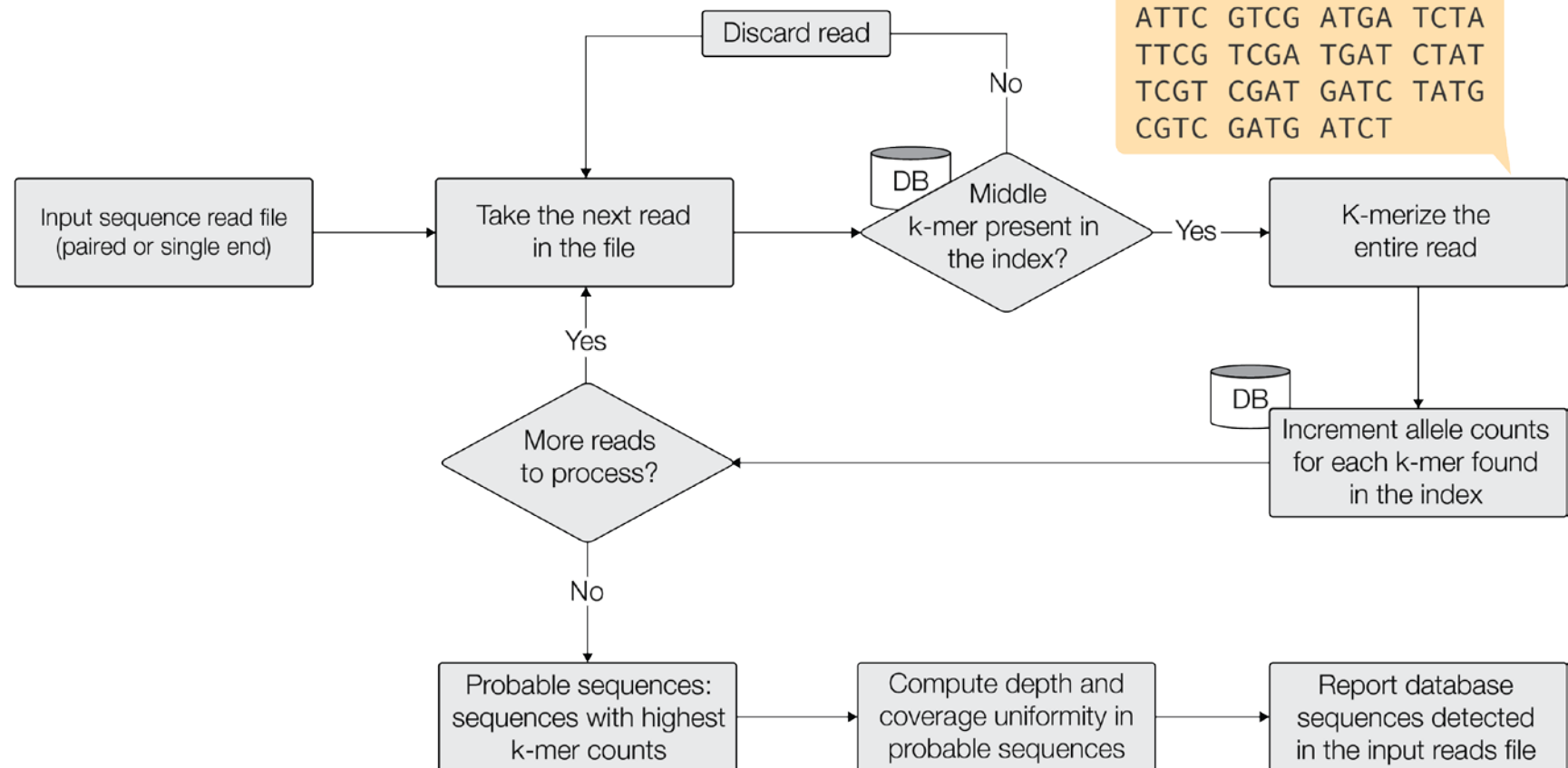
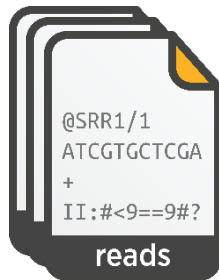


Algorithm Overview

(A) Database Indexing

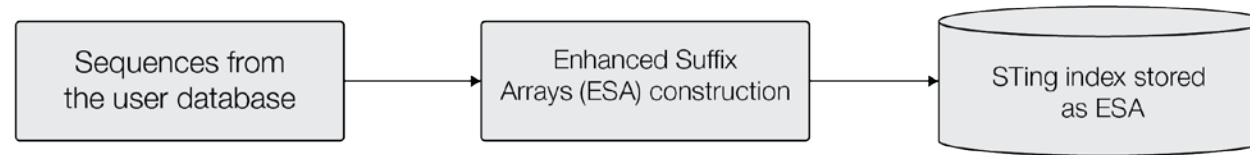


(B) Sequence Variant Detection

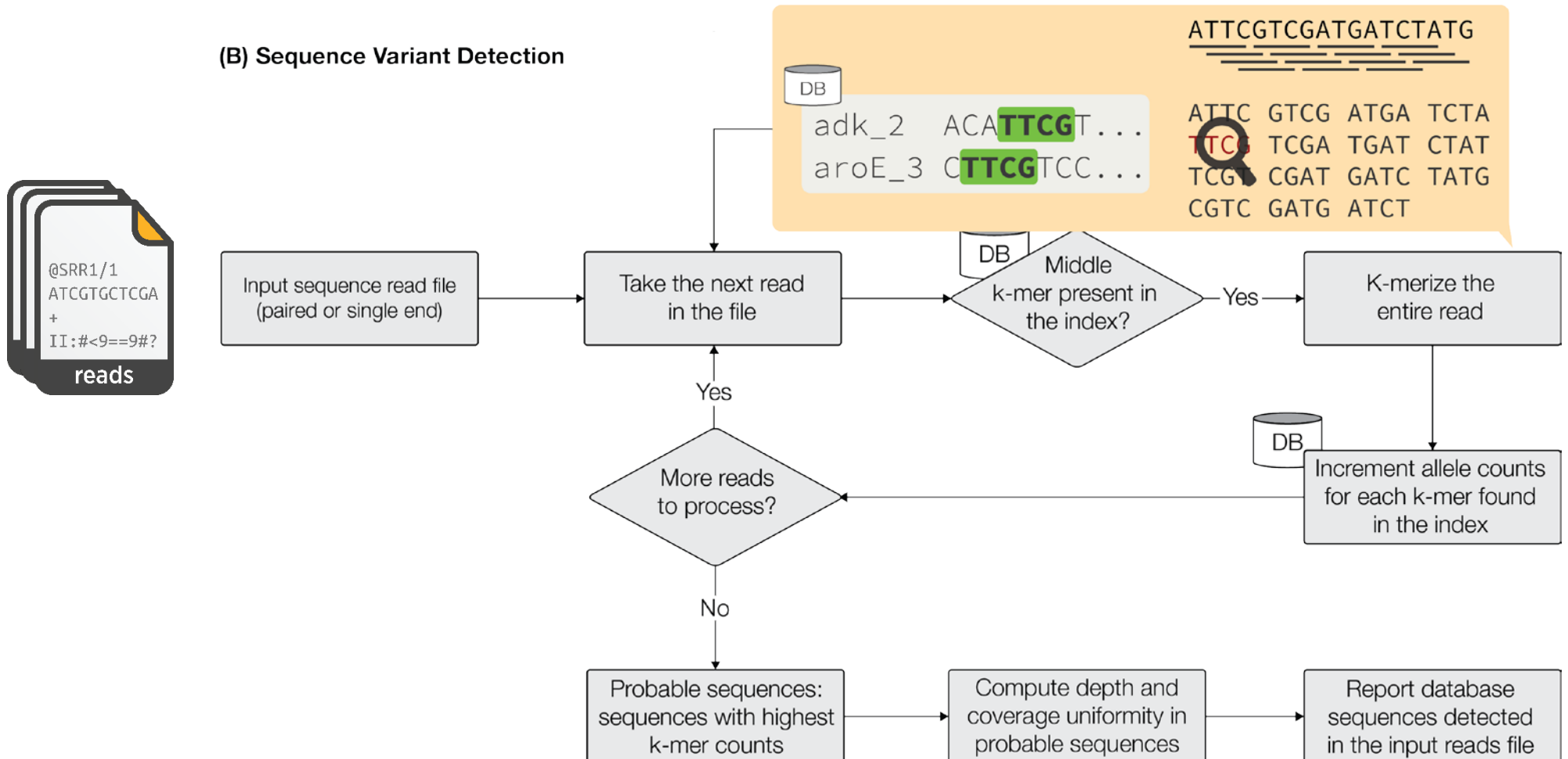


Algorithm Overview

(A) Database Indexing

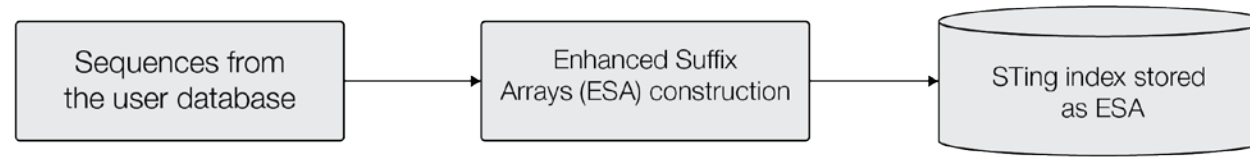


(B) Sequence Variant Detection

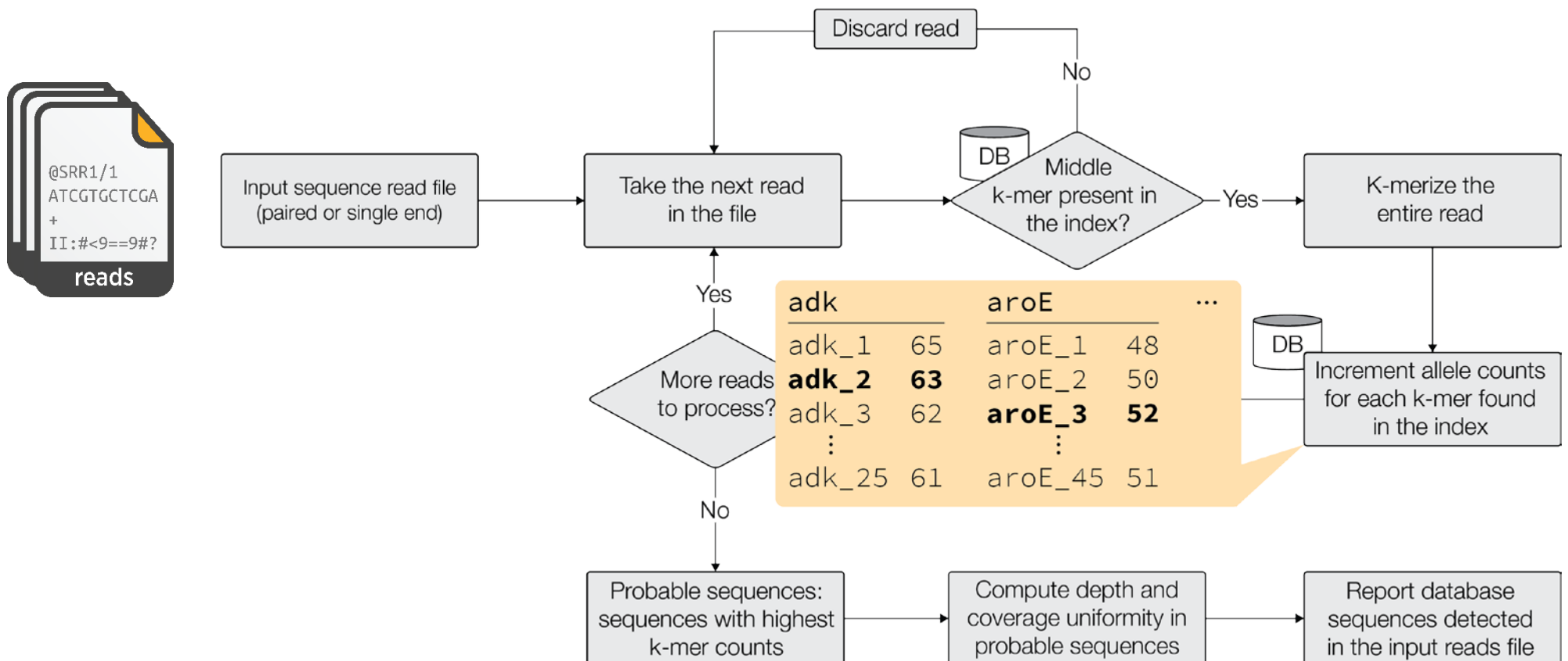


Algorithm Overview

(A) Database Indexing

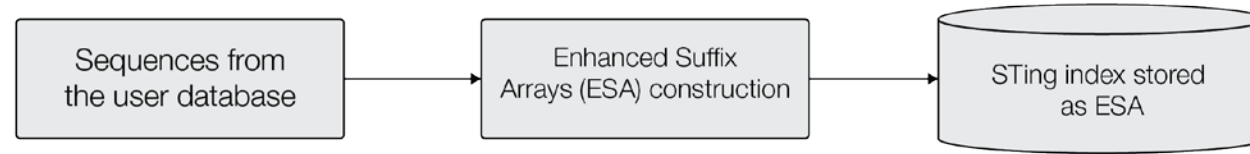


(B) Sequence Variant Detection

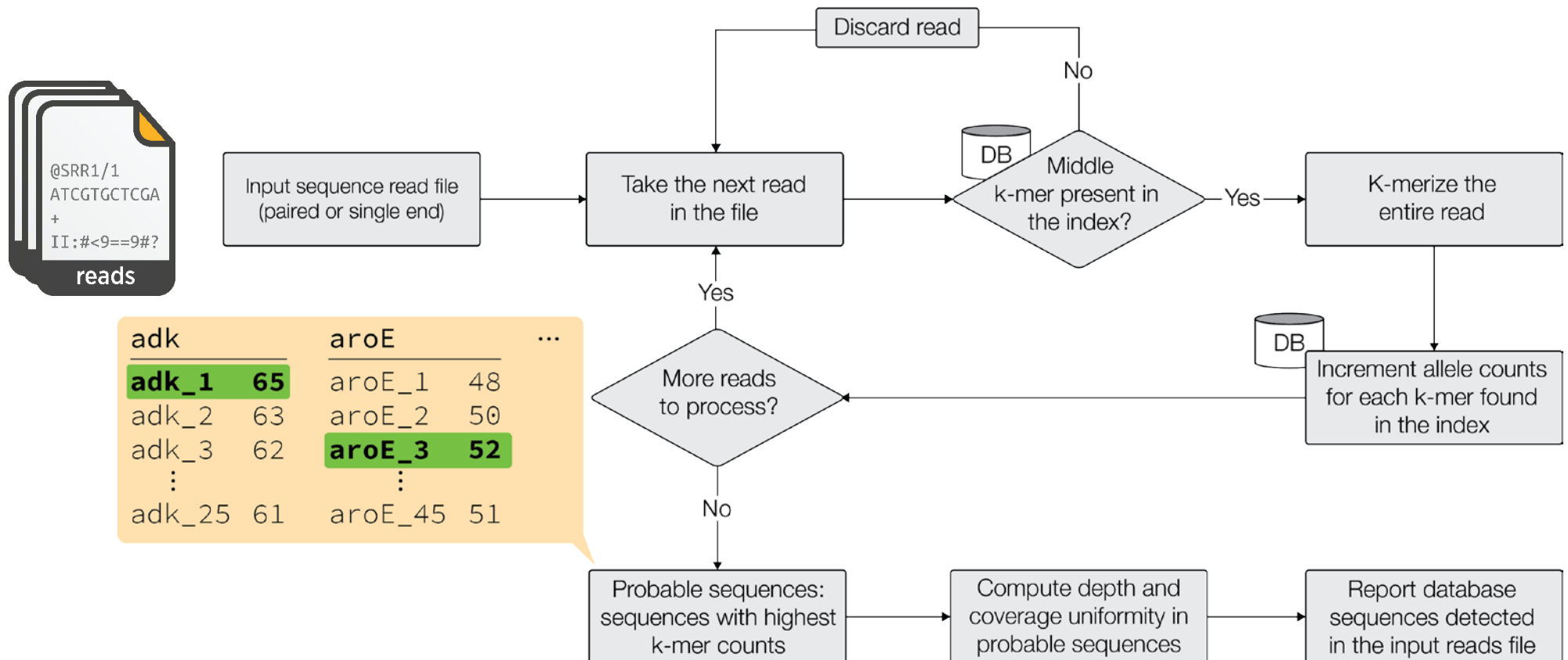


Algorithm Overview

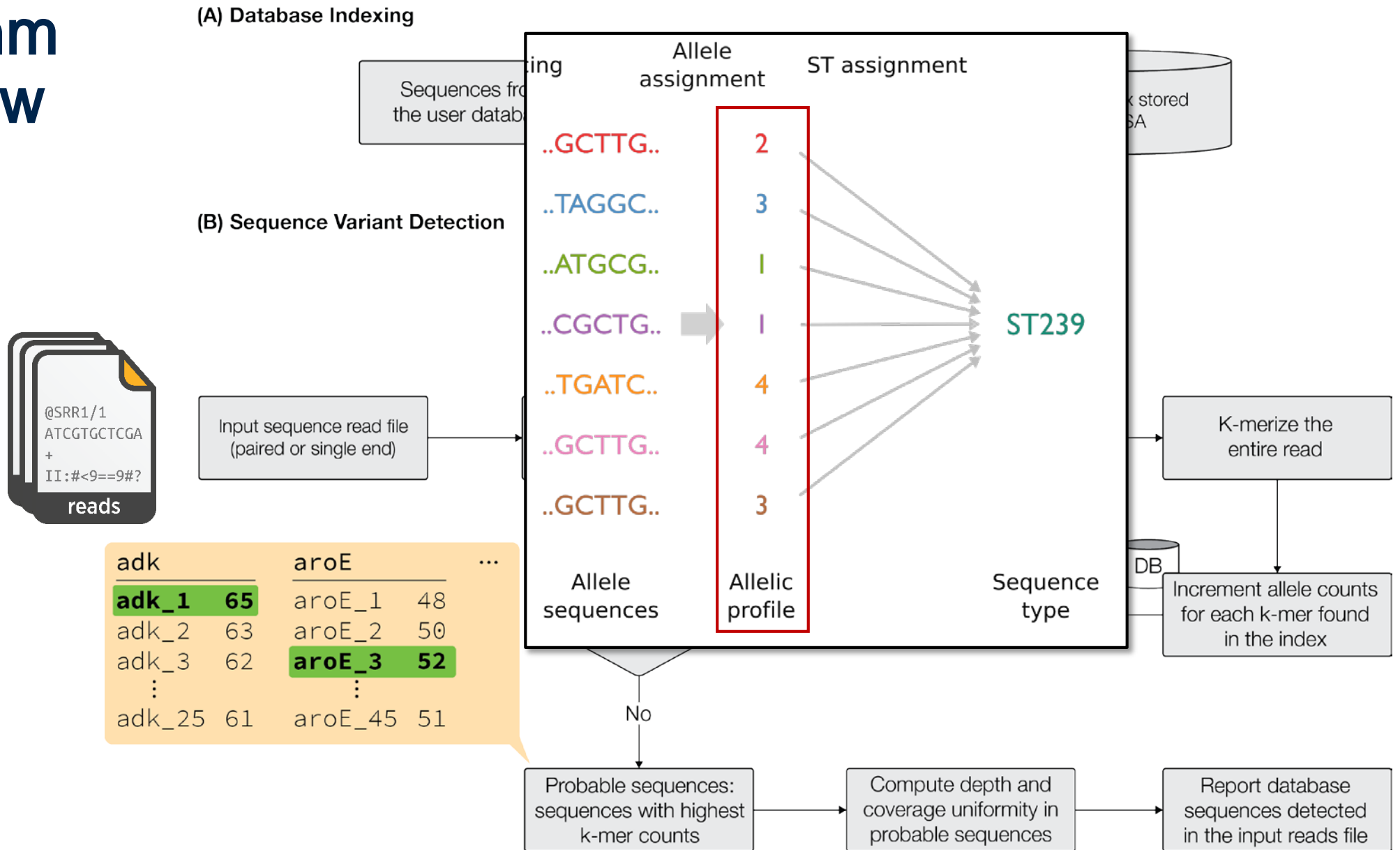
(A) Database Indexing



(B) Sequence Variant Detection

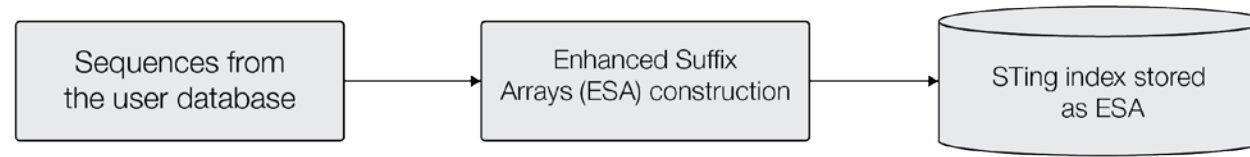


Algorithm Overview

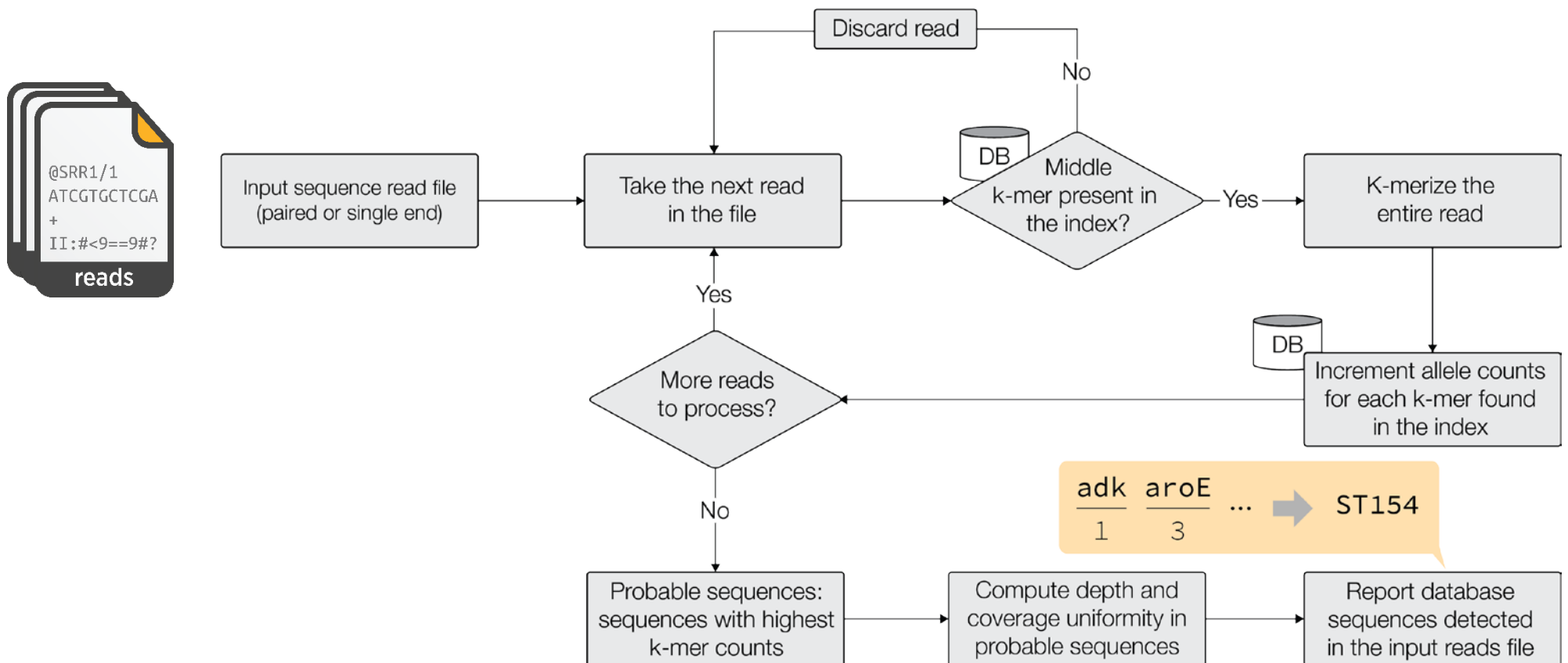


Algorithm Overview

(A) Database Indexing



(B) Sequence Variant Detection



STing: Sequence Typing

STing – Typing Dataset

Species	Scheme	# Locus	DB Size (sequences)	# Samples
<i>C. jejuni</i>	MLST	7	4,117	10
<i>C. trachomatis</i>	MLST	7	218	10
<i>S. pneumoniae</i>	MLST	7	3,319	10
<i>N. meningitidis</i>	MLST	7	5,325	1,009
<i>N. meningitidis</i>	rMLST	53	461,054	20
<i>N. meningitidis</i>	cgMLST	1,605	639,542	20

rMLST: Ribosomal MLST; cgMLST: Core Genome MLST

STing – Typing Dataset

Species	Scheme	# Locus	DB Size (sequences)	# Samples
<i>C. jejuni</i>	MLST	7	4,117	10
<i>C. trachomatis</i>	MLST	7	218	10
<i>S. pneumoniae</i>	MLST	7	3,319	10
<i>N. meningitidis</i>	MLST	7	5,325	1,009
<i>N. meningitidis</i>	rMLST	53	461,054	20
<i>N. meningitidis</i>	cgMLST	1,605	639,542	20

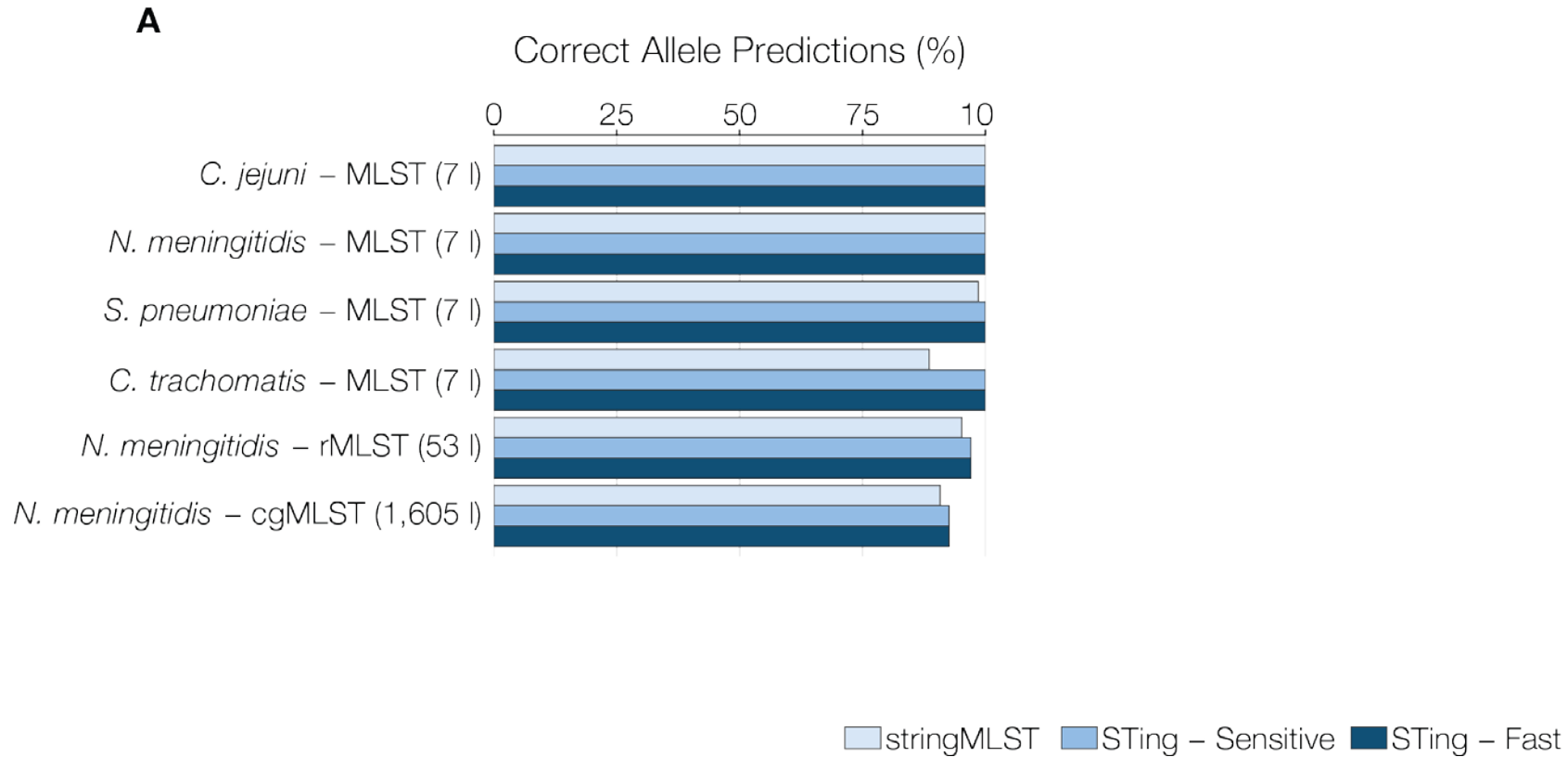
rMLST: Ribosomal MLST; cgMLST: Core Genome MLST

STing – Typing Dataset

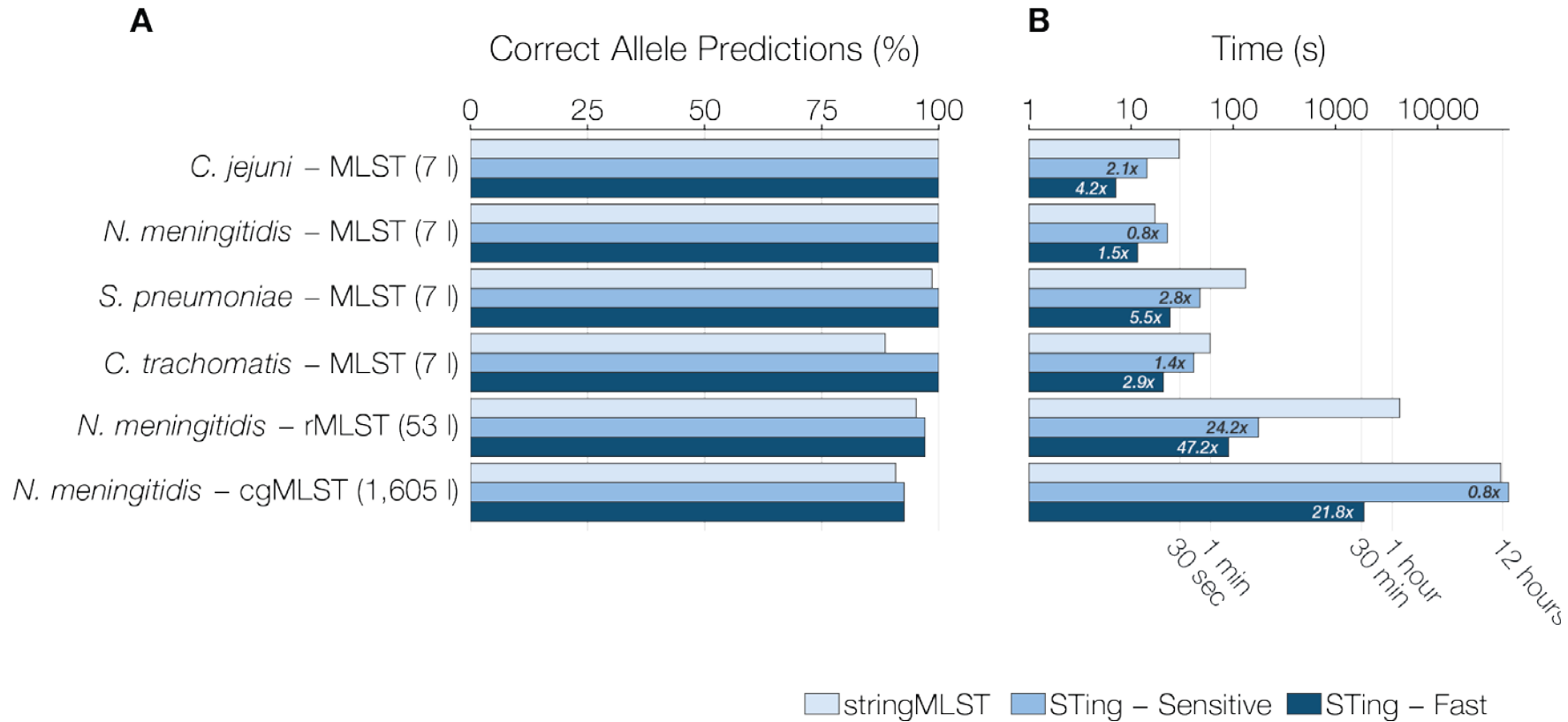
Species	Scheme	# Locus	DB Size (sequences)	# Samples
<i>C. jejuni</i>	MLST	7	4,117	10
<i>C. trachomatis</i>	MLST	7	218	10
<i>S. pneumoniae</i>	MLST	7	3,319	10
<i>N. meningitidis</i>	MLST	7	5,325	1,009
<i>N. meningitidis</i>	rMLST	53	461,054	20
<i>N. meningitidis</i>	cgMLST	1,605	639,542	20

rMLST: Ribosomal MLST; cgMLST: Core Genome MLST

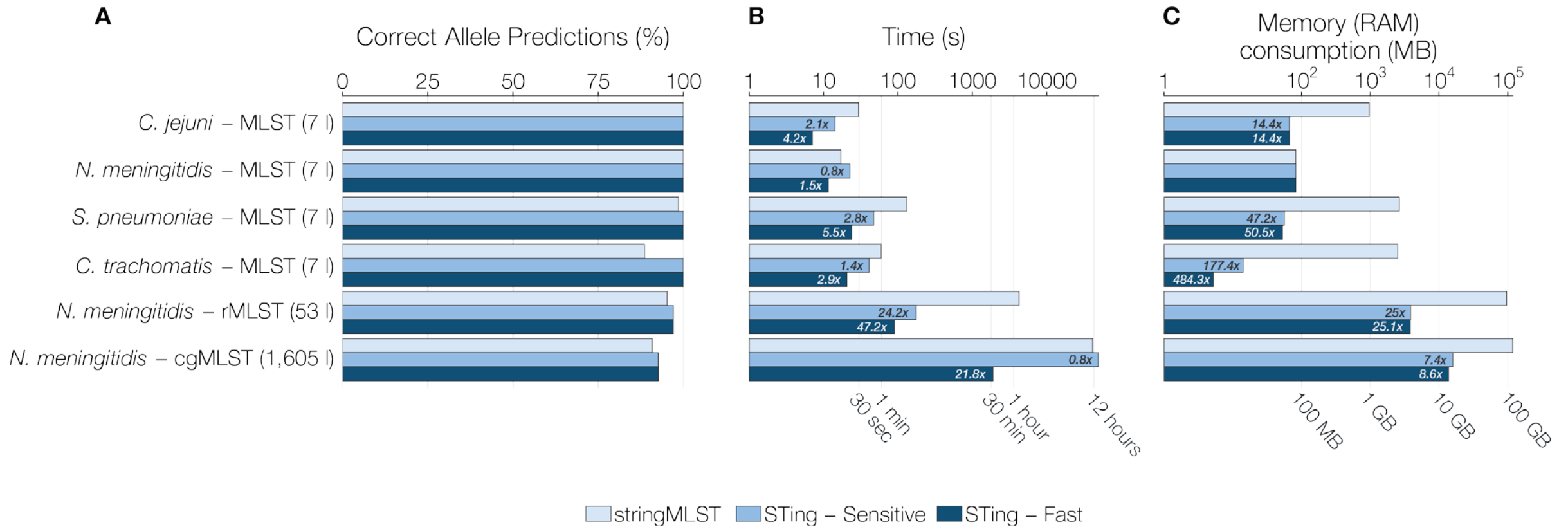
STing – Typing Performance



STing – Typing Performance

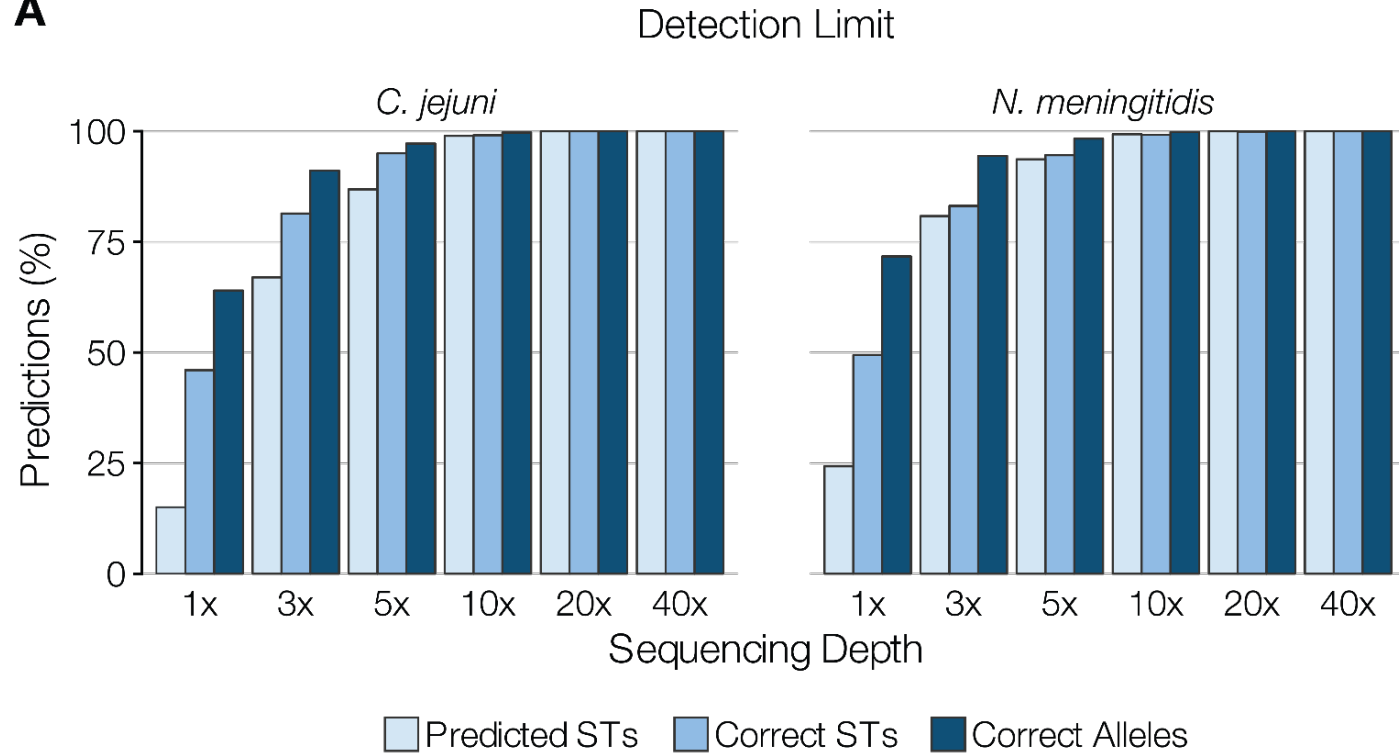


STing – Typing Performance



STing – Typing Performance

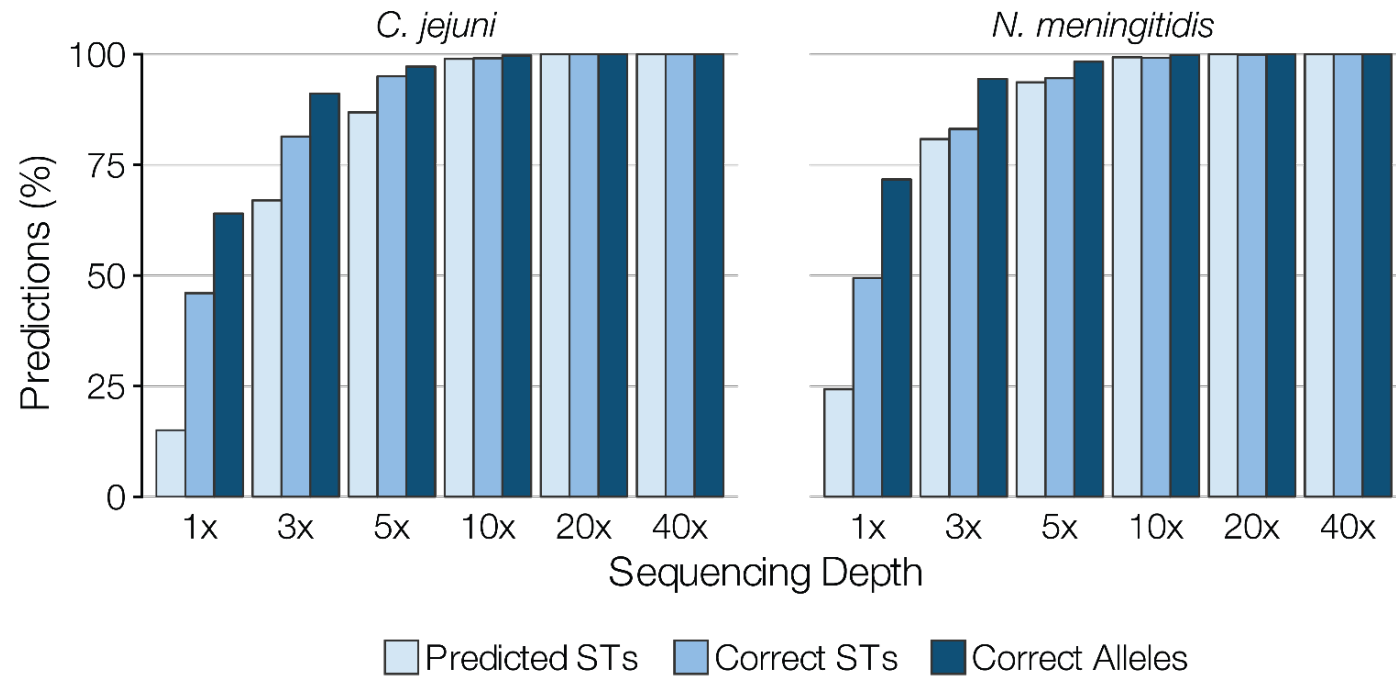
A



STing – Typing Performance

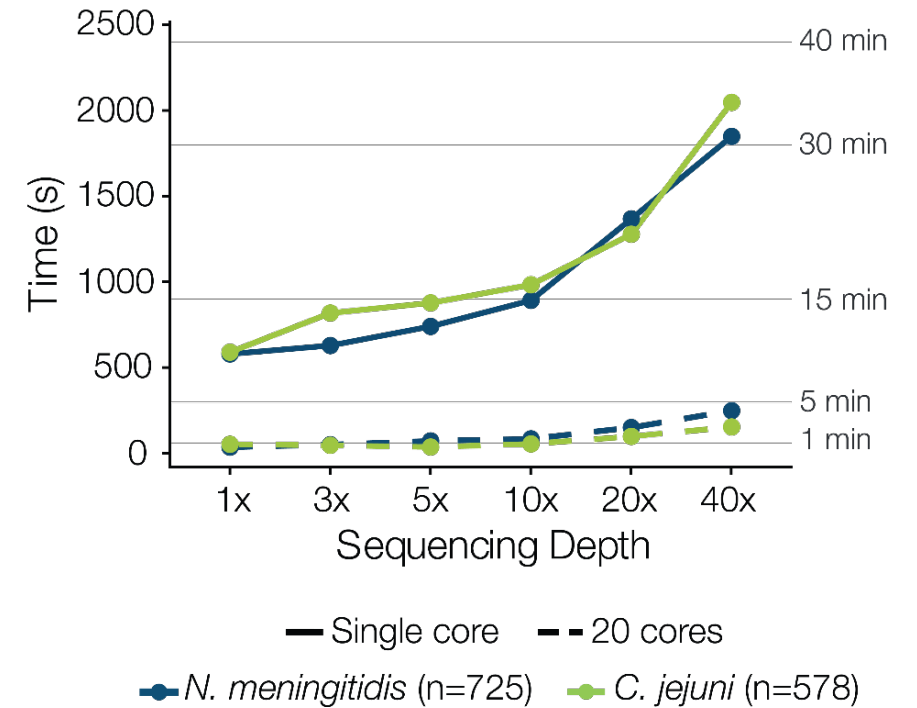
A

Detection Limit



B

Performance



STing: Gene Detection

STing – Gene Detection Dataset

- We evaluated whether we can detect AMR genes (**n=16**) from the sequence reads of **12 genomes** of nine species (positive samples)
- We artificially excised the AMR genes from each of the genomes to generate negative samples
- We simulated reads at **20x** and **40x** coverage from both positive and negative samples

STing – Gene Detection Performance

100% accuracy

Species/Strain	AMR Genes Present	AMR Genes											
		<i>aac2ic</i>	<i>aph6id</i>	<i>bl2a</i>	<i>ermB</i>	<i>pbp2x</i>	<i>pbp1a</i>	<i>pbp2b</i>	<i>ksgA</i>	<i>qacE</i>	<i>ermC</i>	<i>mepA</i>	<i>tetpA</i>
<i>M. tuberculosis</i>	<i>aac2ic</i>	■											
<i>K. pneumoniae</i>	<i>aph6id</i>		■										
<i>B. anthracis</i>	<i>bl2a</i>			■									
<i>C. difficile</i>	<i>ermB</i>				■								
<i>S. pneumoniae</i>	<i>pbp2x</i>					■							
<i>N. gonorrhoeae</i>	<i>ksgA</i> , <i>pbp1a</i> , <i>pbp2b</i>						■	■	■				
<i>N. meningitidis</i> 053442	<i>ksgA</i> , <i>pbp1a</i> , <i>pbp2b</i>						■	■	■				
<i>N. meningitidis</i> C FAM18	<i>ksgA</i> , <i>pbp1a</i> , <i>pbp2b</i>						■	■	■				
<i>N. meningitidis</i> Z2491	<i>ksgA</i> , <i>pbp1a</i> , <i>pbp2b</i> , <i>qacE1</i>						■	■	■	■			
<i>N. meningitidis</i> MC58	<i>ksgA</i> , <i>pbp1a</i> , <i>pbp2b</i> , <i>qacE1</i> , <i>ermC</i>						■	■	■	■	■		
<i>S. aureus</i>	<i>mepA</i>											■	
<i>C. acetobutylicum</i>	<i>tetpA</i>												■

■ Correct Predictions

STing – Other Applications

- Virulence factor (VF) gene detection (e.g. Shiga toxin and hemolysin loci).
- Antimicrobial (AMR) gene detection in fungal isolates.
- Gene detection in metagenome samples.

STing – Other Applications



Sung Im
PhD Student (Binf)

- Virulence factor (VF) gene detection (e.g. Shiga toxin and hemolysin loci).
- Antimicrobial (AMR) gene detection in fungal isolates.
- Gene detection in metagenome samples.

Conclusions

- Faster alternatives of analysis are necessary to face the challenges from the NGS-era.
- Although alignment-based analysis are slow, not scalable, they are irreplaceable! (e.g. annotation, ancestral DNA reconstruction, sequence evolution rate calculations).
- We applied the alignment-free paradigm for sequence typing and gene detection (accurately and efficiently).

Conclusions

- STing algorithm scales efficiently to genome-scale typing schemes (cgMLST).
- STing performs orders of magnitude better than existing tools.
- Possible applications of STing include culture-free diagnostics as well as virulence factor and antimicrobial resistance profiling directly from NGS reads.



STing Team!



Hector
Espitia



Aroon
Chande



Heather
Smith



Lavanya
Rishishwar



King
Jordan



Jordan Lab @ Georgia Tech

References

1. Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, *109*(5–6), 419–431. <https://doi.org/10.1016/j.ygeno.2017.06.007>
2. Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, *18*(1), 186. <https://doi.org/10.1186/s13059-017-1319-7>
3. Gupta, A., Jordan, I. K., & Rishishwar, L. (2017). stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics*, *33*(1), 119–121. <https://doi.org/10.1093/bioinformatics/btw586>
4. Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., & Sun, F. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, *15*(3), 343–353. <https://doi.org/10.1093/bib/bbt067>