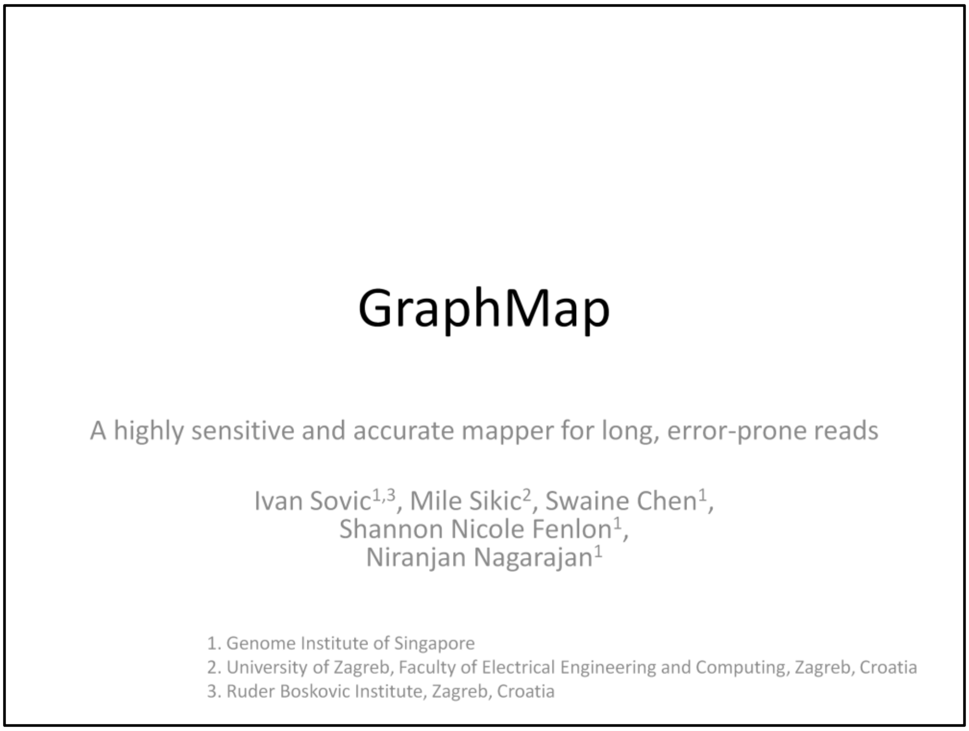


GraphMap

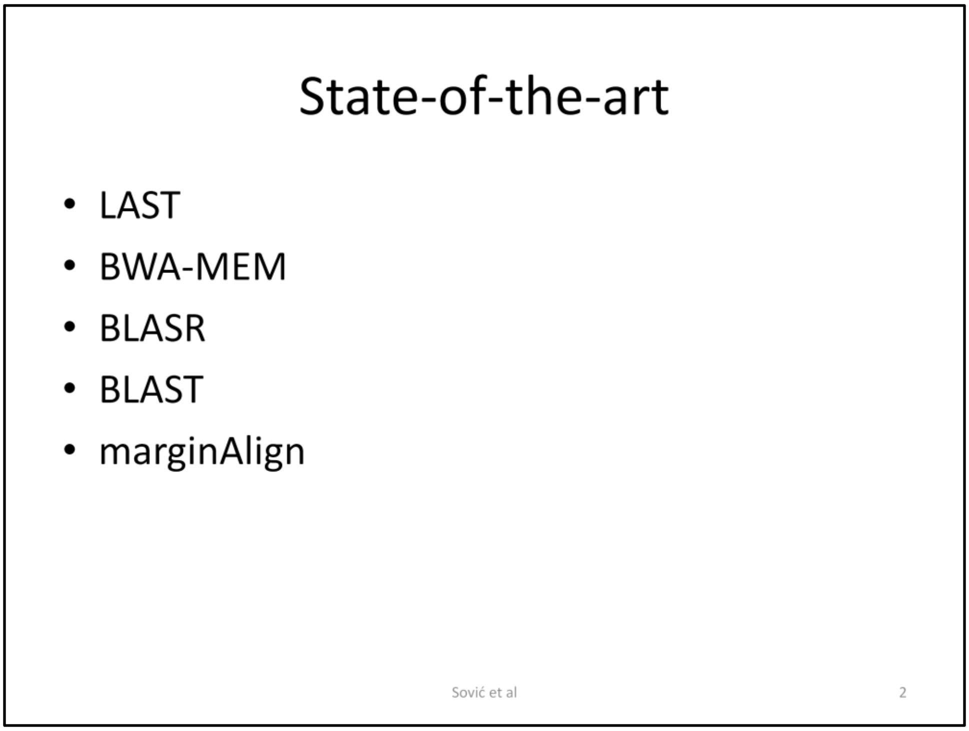
A highly sensitive and accurate mapper for long, error-prone reads

Ivan Sovic[1,3], Mile Sikic[2], Swaine Chen[1],
Shannon Nicole Fenlon[1],
Niranjan Nagarajan[1]

1. Genome Institute of Singapore
2. University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
3. Ruder Boskovic Institute, Zagreb, Croatia

GraphMap is a novel mapper targeted at aligning long, error-prone third-generation sequencing data.

It has been developed in collaboration between:
- Genome Institute of Singapore
- University of Zagreb, Faculty of Electrical Engineering and Computing, and
- Rudjer Boskovic Institute from Croatia

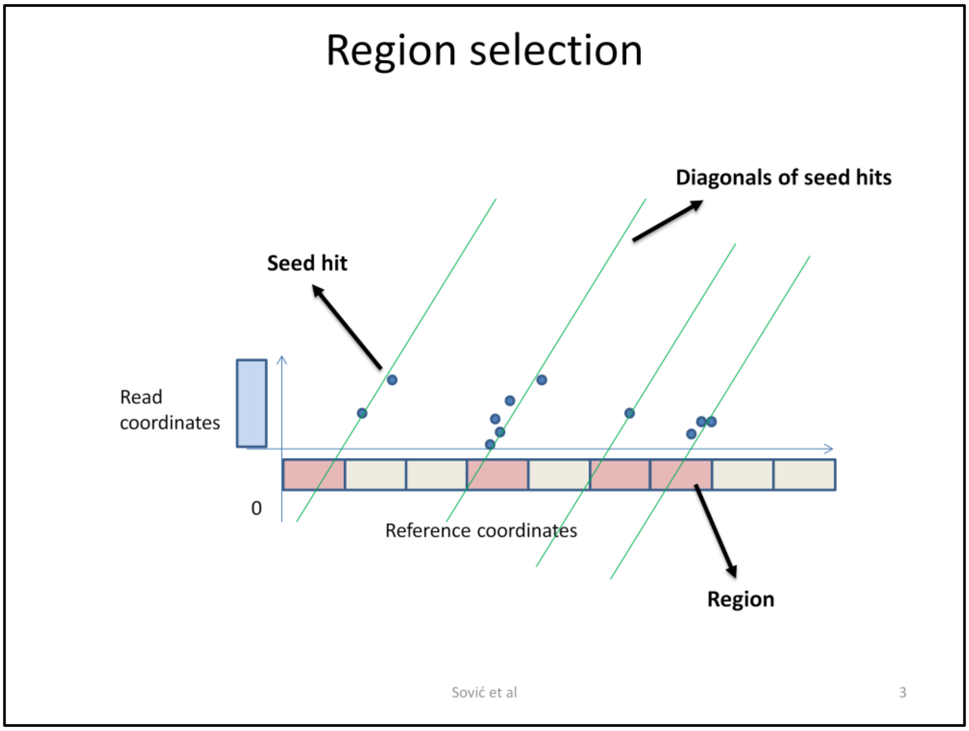


State-of-the-art

- LAST
- BWA-MEM
- BLASR
- BLAST
- marginAlign

Sović et al 2

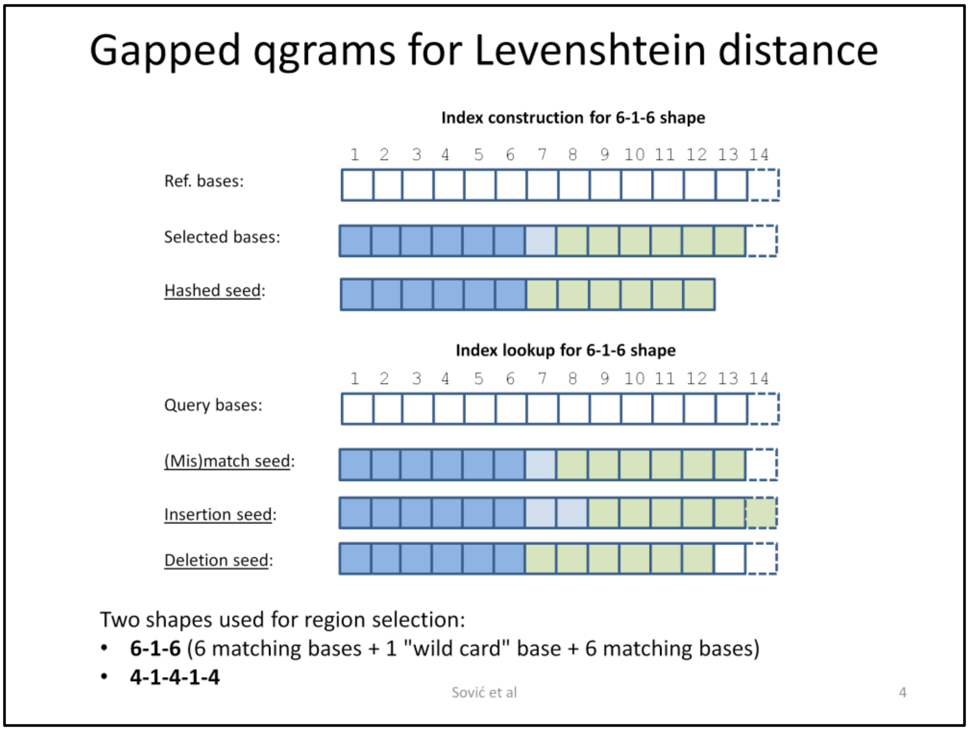
State-of-the-art mappers typically find a large number of short matching seeds and perform gapped or ungapped extensions around them.
Using gapped alignment to generate high scoring segment pairs makes mapping accuracy dependent of scoring parameters used for extension.
Instead, GraphMap takes a different approach and performs DP alignment only at the very last step, which makes it a robust and powerful tool to handle a wide range of error profiles with even the default parameters.



Region selection

Sović et al 3

GraphMap starts by roughly determining regions on the reference genome where a read could potentially be mapped in order to reduce the search space for the next step of the algorithm, while still providing very high sensitivity.
In short, region selection relies on finding and clustering seeds, and choosing only top scoring regions.
An important feature at this step is the type of seeds that GraphMap uses.



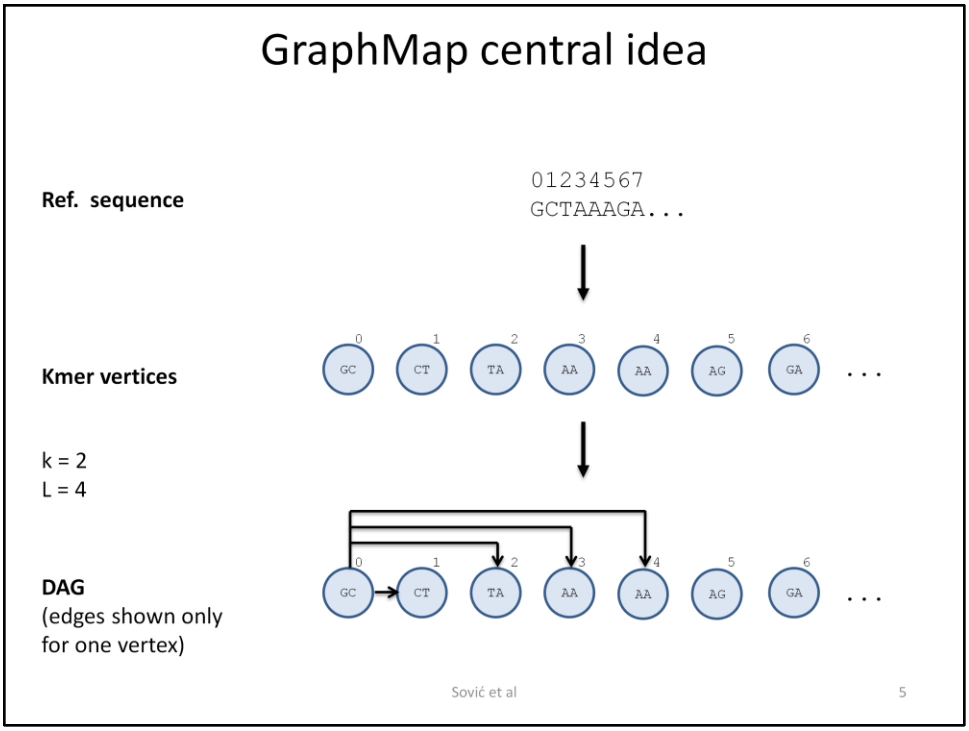
Gapped qgrams for Levenshtein distance

Two shapes used for region selection:

- **6-1-6** (6 matching bases + 1 "wild card" base + 6 matching bases)
- **4-1-4-1-4**

Sović et al 4

Mappers today usually find either a set of exact matches or allow only a small number of errors within a given Hamming distance from the indexed seed which allows only mismatches in comparison.
Instead, GraphMap implements gapped qgrams for Levenshtein distance matching which are perfectly suited for large number of indels such as in nanopore data. We specify qgrams with a shape that determines matching and “wild card” positions. Regions are then sorted by seed hit count for further processing one at a time.



GraphMap central idea

Ref. sequence

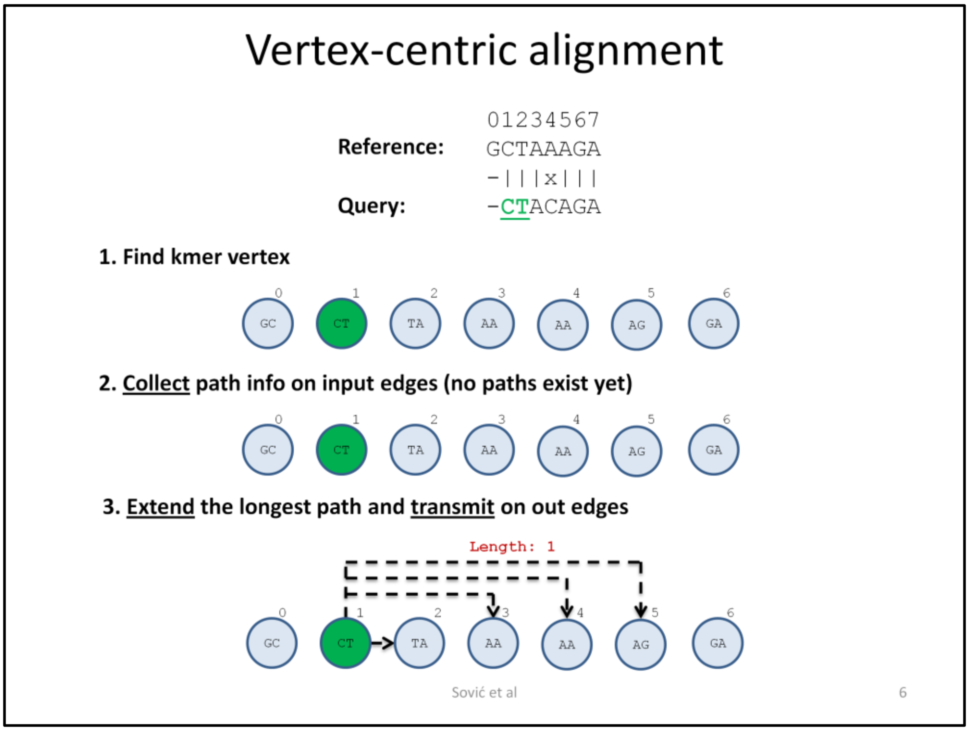
01234567
GCTAAAGA...

Kmer vertices

k = 2
L = 4

DAG
(edges shown only
for one vertex)

The central part of GraphMap is to develop a target sequence into a directed acyclic kmer graph,
where each vertex has L directed outbound edges, connecting the vertex to L following vertices.
In this example kmers are of size 2 and the number of edges is equal to 4.
In reality, GraphMap uses k = 6 and L = 9.



Vertex-centric alignment

```
            01234567
Reference:  GCTAAAGA
            -|||x|||
Query:      -CTACAGA
```

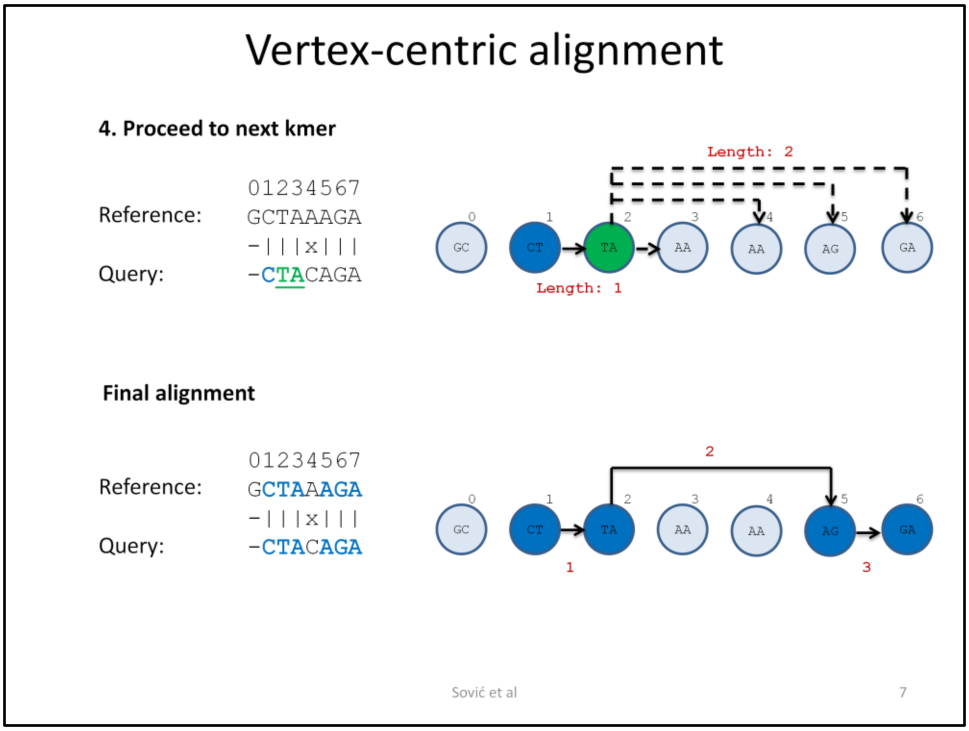
1. Find kmer vertex

2. Collect path info on input edges (no paths exist yet)

3. Extend the longest path and transmit on out edges

Sović et al

For each consecutive kmer in the query, all matching vertices are looked up and processed individually in a vertex centric manner.
For a vertex, data from its input edges is collected, and the edge belonging to longest previously traversed walk is chosen. The walk is extended, and transmitted to all outbound edges of the vertex simultaneously.



Vertex-centric alignment

4. Proceed to next kmer

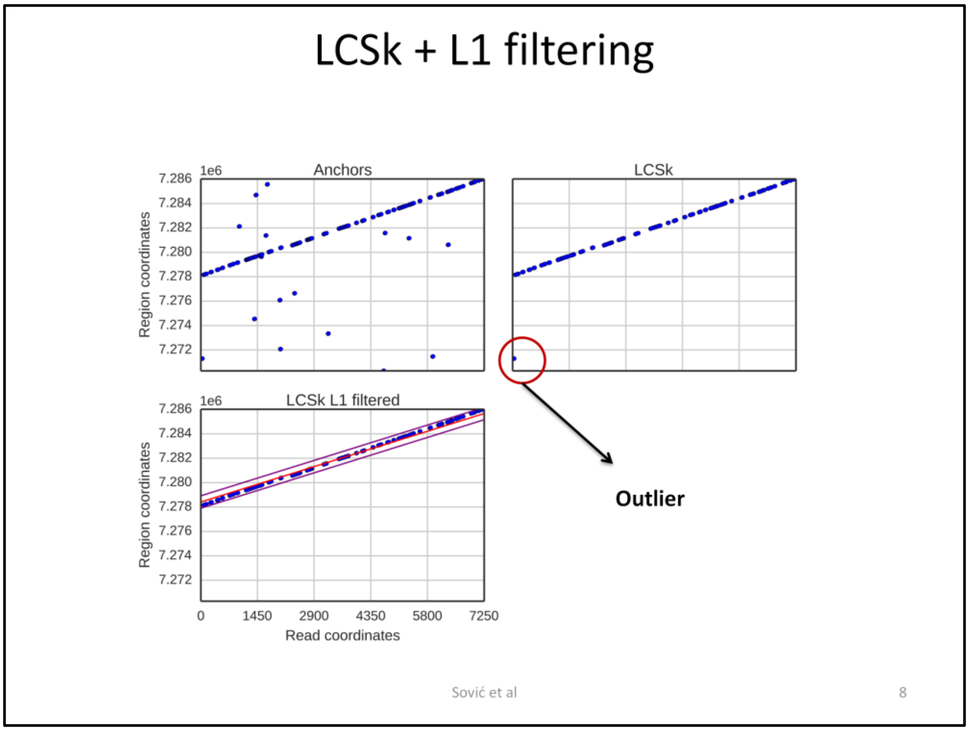
```
            01234567
Reference:  GCTAAAGA
            -|||x|||
Query:      -CTACAGA
```

Final alignment

```
            01234567
Reference:  GCTAAAGA
            -|||x|||
Query:      -CTACAGA
```

Sović et al 7

After the entire query is processed, edges contain information of graph traversals. Ideally, when error rates are low like in Illumina or even PacBio reads, a single path can cover almost the entire query, like in the example shown in figure.
However, when reads are very erroneous, the query is usually not covered by only one path.
Instead, we end up with a list of walks, or anchors which are long inexact matches between two sequences.



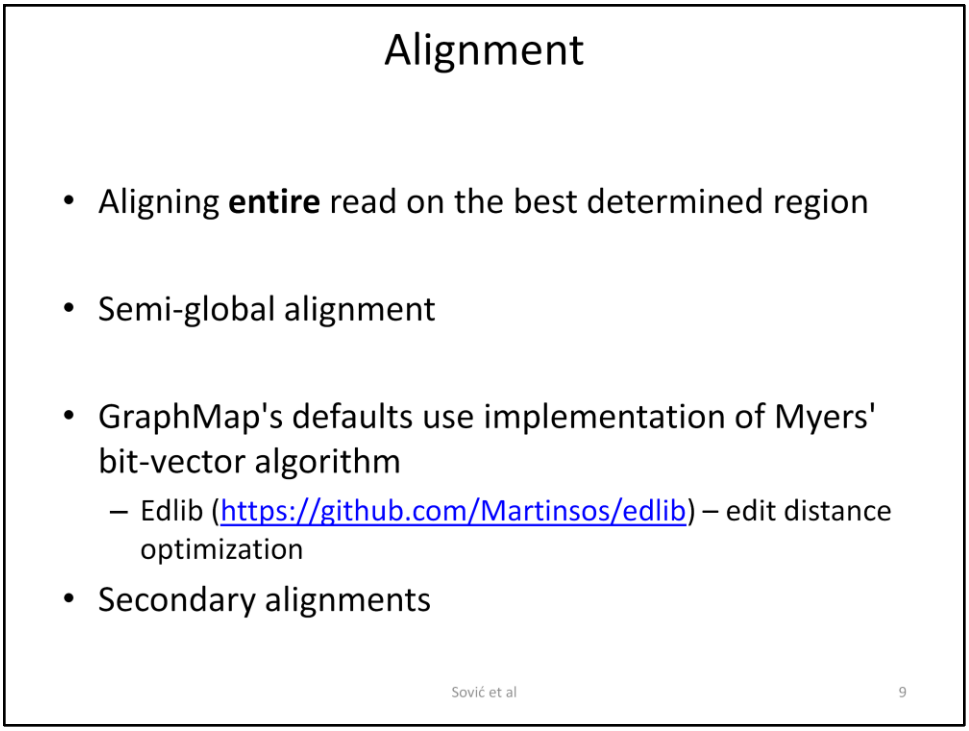
LCSk + L1 filtering

Outlier

Sović et al

8

Many outlier anchors might occur off the main diagonal mostly because of short repeats.
We apply the Longest Common Subsequence in k-length Substrings algorithm on all anchors in order to obtain one monotonically increasing list of anchors.
Leading and trailing outliers can still occur, because LCS is a global similarity measure.
We filter such outliers by fitting the L1 line and filtering anchors outside the confidence interval.

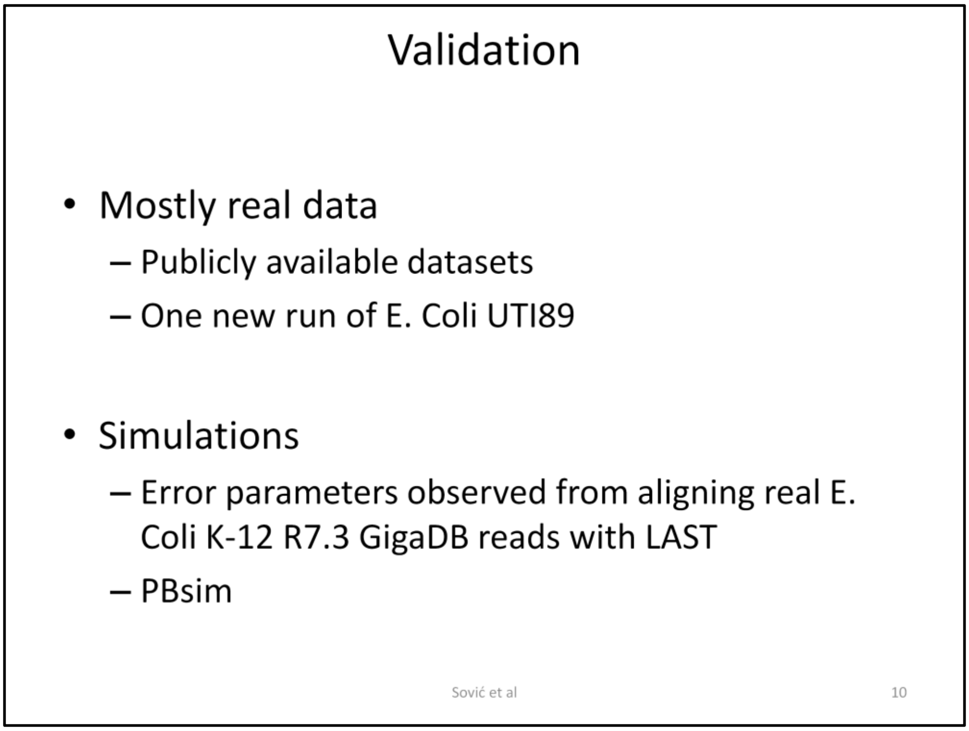


Alignment

- Aligning **entire** read on the best determined region
- Semi-global alignment
- GraphMap's defaults use implementation of Myers' bit-vector algorithm
  - – Edlib (https://github.com/Martinsos/edlib) – edit distance optimization
- Secondary alignments

Sović et al 9

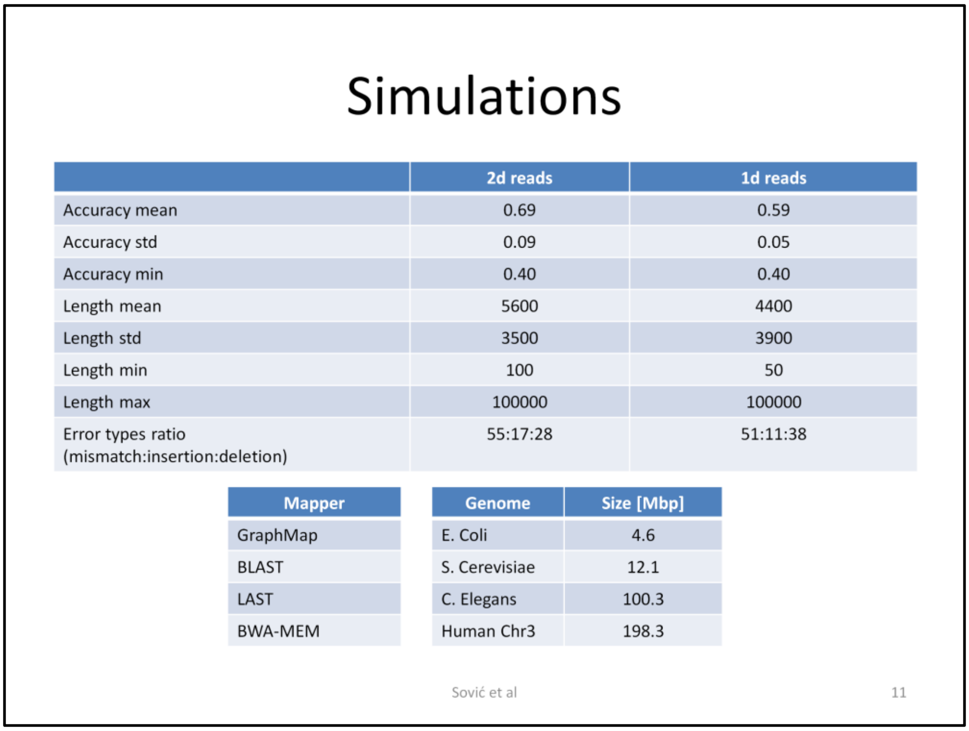
After all regions have finished processing, they are sorted and the best scoring region is chosen.
GraphMap aligns the entire read using semi-global alignment algorithm.
The default parameters of GraphMap use an implementation of Myers’ bit-vector approach to alignment,
but it's also possible to use custom alignment scoring scheme and a semi-global implementation of alignment with affine gaps.
GraphMap can also output all secondary alignments to within a specified margin of difference from the top score.



Validation

- Mostly real data
  - Publicly available datasets
  - One new run of E. Coli UTI89
- Simulations
  - Error parameters observed from aligning real E. Coli K-12 R7.3 GigaDB reads with LAST
  - PBsim

We validated GraphMap intensively, mostly on real data.
However, simulations were also performed in order to determine the behavior when different error profiles are applied to data.
In the absence of a dedicated MinION simulator, we used PBsim to which we provided parameters determined from LAST's alignments of real data.



Simulations

	2d reads	1d reads
Accuracy mean	0.69	0.59
Accuracy std	0.09	0.05
Accuracy min	0.40	0.40
Length mean	5600	4400
Length std	3500	3900
Length min	100	50
Length max	100000	100000
Error types ratio (mismatch:insertion:deletion)	55:17:28	51:11:38

Mapper
GraphMap
BLAST
LAST
BWA-MEM

Genome	Size [Mbp]
E. Coli	4.6
S. Cerevisiae	12.1
C. Elegans	100.3
Human Chr3	198.3

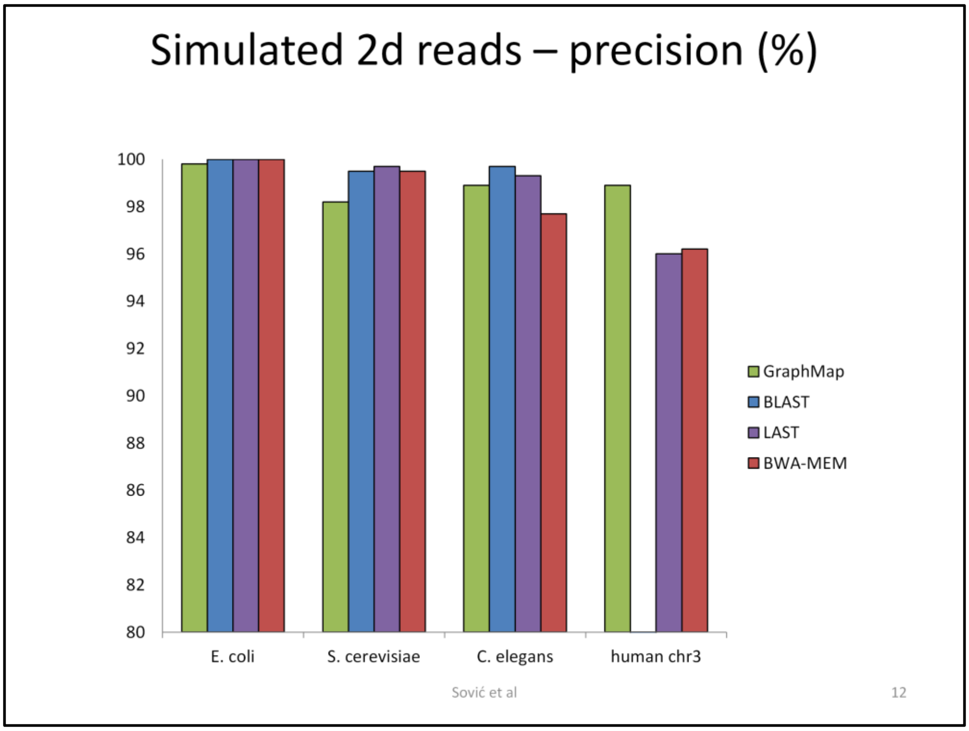
Sović et al

11

These parameters include the mean accuracy, mean standard deviation, error type ratio and read lengths.

We determined these parameters separately for 1d and 2d reads, and generated simulations of 4 genomes of different sizes.

The datasets were used to determine the precision and recall for mapping the reads into correct position on the reference genome.

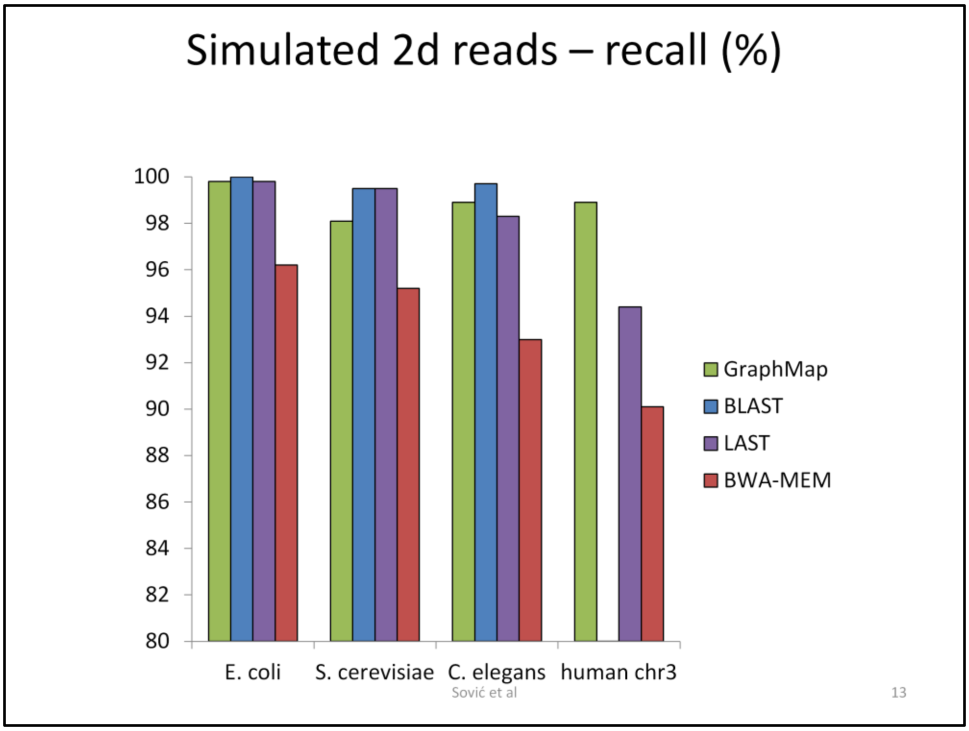


Simulated 2d reads – precision (%)

Sović et al

12

Due to the long running time of BLAST, it has not been evaluated on the human chromosome.
The obtained results show that all mappers have similar precision on 2d reads.
Since 2d reads are more accurate, this was expected.
However, it is important to note that, for longer genomes, while the precision of LAST and BWA-MEM drops, the precision of GraphMap is steady.

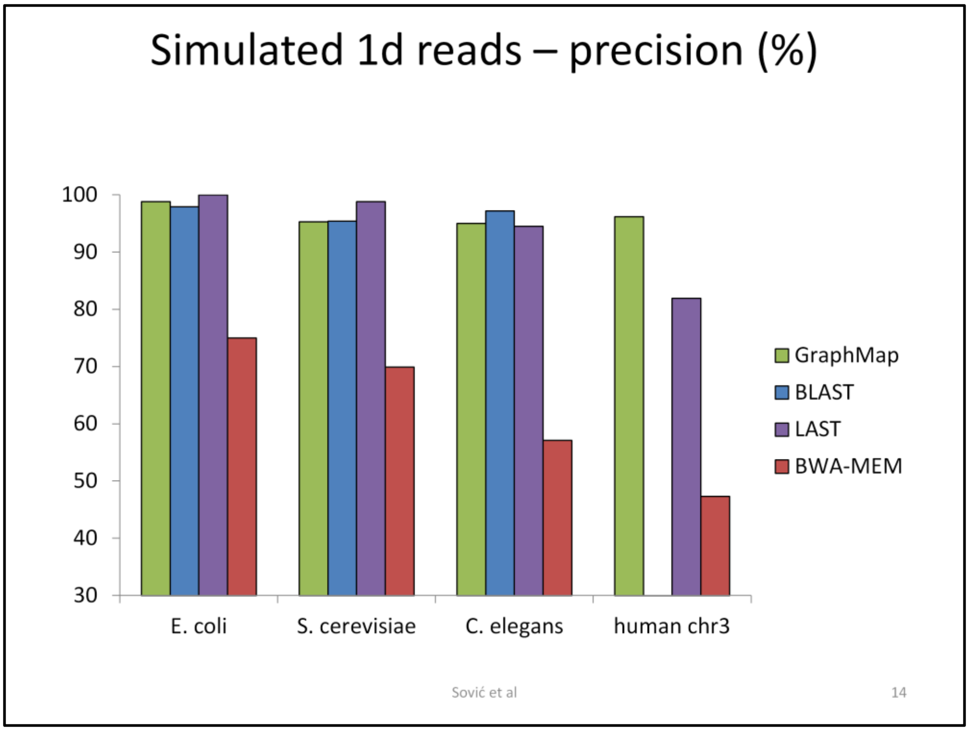


Simulated 2d reads – recall (%)

Sović et al

13

The results for recall show a similar pattern.
Generally, all mappers perform well. The recalls are above 90%.
However, comparing to the precision, the recall for longer genomes drops more significantly for LAST and especially for BWA-MEM.
On the contrary, the recalls obtained from GraphMap are stable for different genomes.

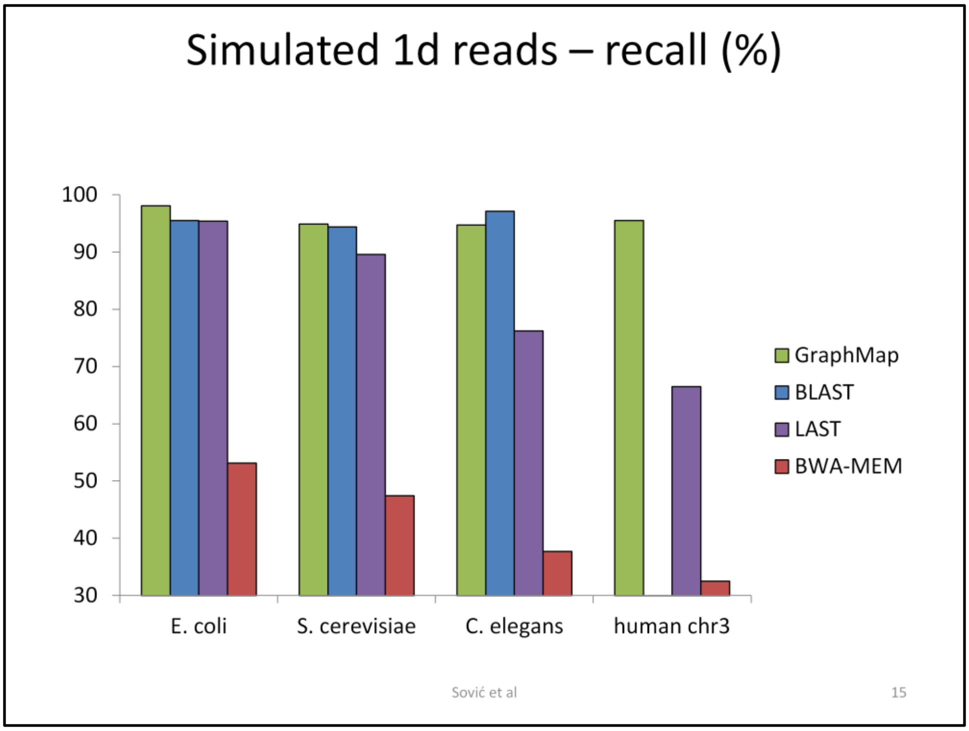


Simulated 1d reads – precision (%)

Sović et al

14

For 1d reads GraphMap still maintains very high precision on all genomes.
Much greater differences can be observed for LAST and BWA-MEM.
BWA-MEM’s performance never exceeds 80%, and on the human chromosome even drops below 50%.
LAST performs good on smaller genomes, but again, on the human chromosome its precision drops below 90%.



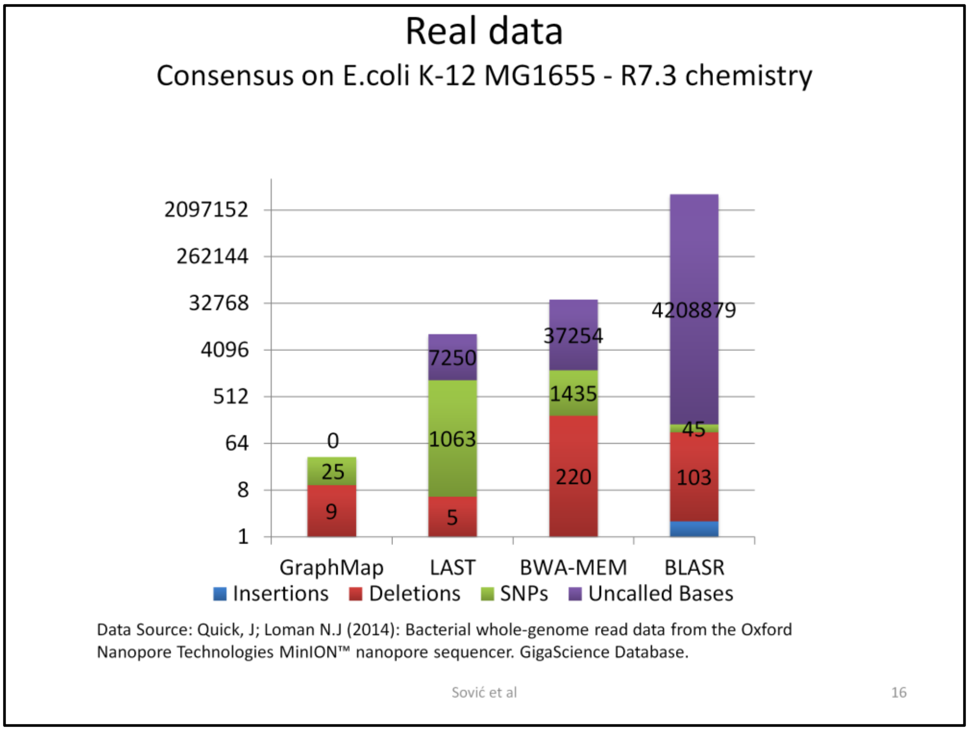
Simulated 1d reads – recall (%)

Sović et al

15

Recall on 1d reads shows even more dramatic changes in performance across different genomes.

GraphMap is consistent in its very high recall, while for BWA-MEM and LAST only a fraction of reads are mapped on larger genomes.



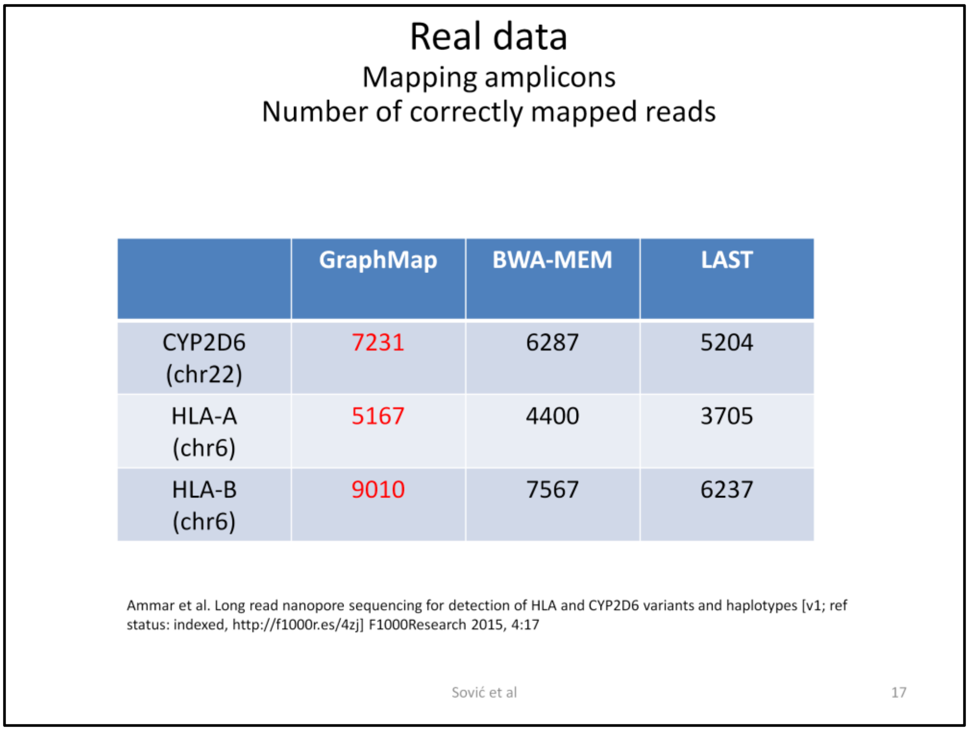
Real data

Consensus on E.coli K-12 MG1655 - R7.3 chemistry

Data Source: Quick, J; Loman N.J (2014): Bacterial whole-genome read data from the Oxford Nanopore Technologies MinION™ nanopore sequencer. GigaScience Database.

Sović et al 16

Next, we evaluated the mapping and alignment correctness on real data by counting the number of consensus variants and bases with insufficient coverage.
For this test, we used the publicly available E. Coli dataset with R7.3 chemistry which we mapped to the K-12 reference genome.
Consensus was called using a simple majority vote. If a certain position had coverage below 20x, the base was deemed uncalled.
GraphMap shows even more than an order of magnitude improvement over LAST, which is the next best performing method.



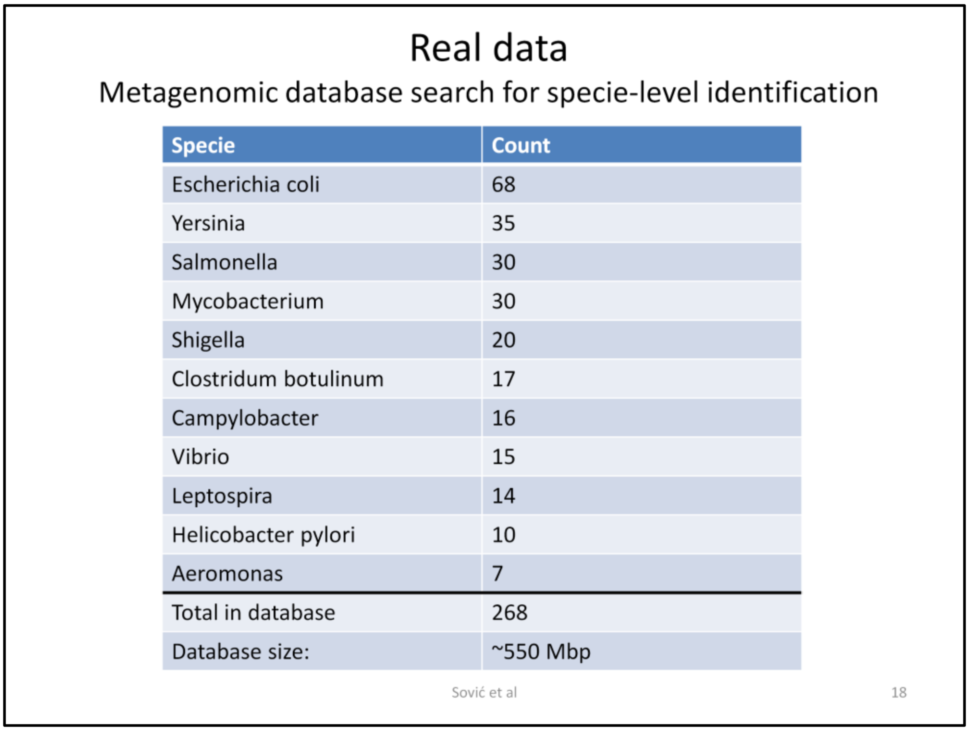
Real data

Mapping amplicons
Number of correctly mapped reads

	GraphMap	BWA-MEM	LAST
CYP2D6 (chr22)	7231	6287	5204
HLA-A (chr6)	5167	4400	3705
HLA-B (chr6)	9010	7567	6237

Ammar et al. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes [v1; ref status: indexed, http://f1000r.es/4zj] F1000Research 2015, 4:17

To further demonstrate the mapping position accuracy we used the publicly available amplicon sequencing dataset which is composed of a mixture of three amplicons coming from chromosome 6 of the human genome and chromosome 22.
The table shows counts of reads, both 1d and 2d, mapped to within the correct genomic region defined by the three sequenced genes.
GraphMap maps a significantly higher amount of reads to the correct genomic positions compared to other methods.



Real data

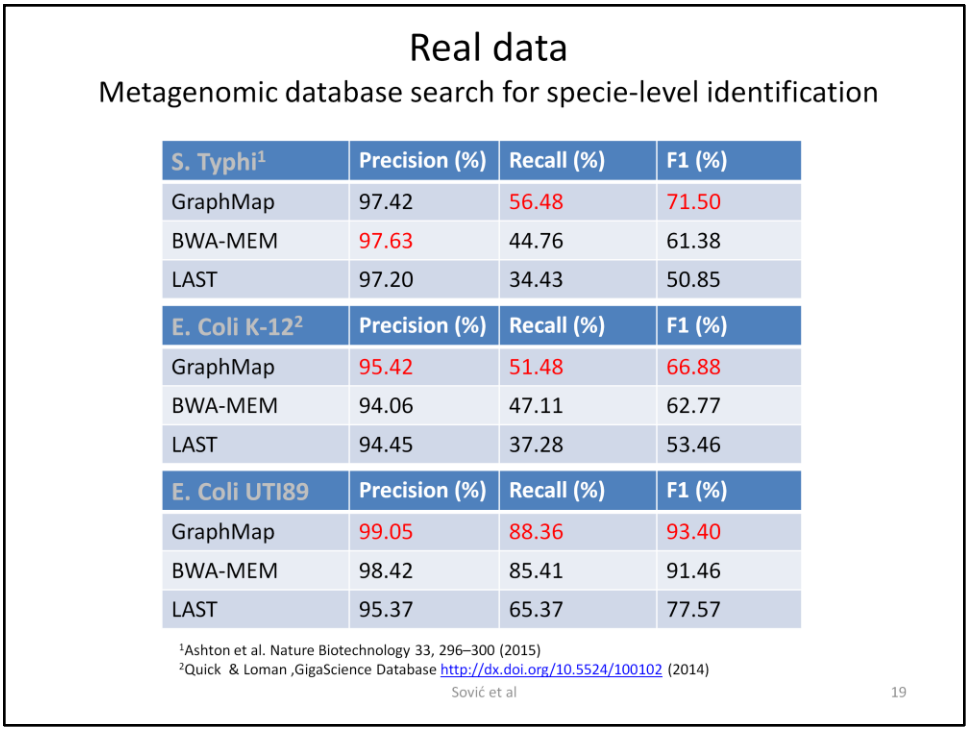
Metagenomic database search for specie-level identification

Specie	Count
Escherichia coli	68
Yersinia	35
Salmonella	30
Mycobacterium	30
Shigella	20
Clostridum botulinum	17
Campylobacter	16
Vibrio	15
Leptospira	14
Helicobacter pylori	10
Aeromonas	7
Total in database	268
Database size:	~550 Mbp

Sović et al

18

Next test on real data involved metagenomic database searching.
We compiled a ~550Mbp database of bacterial genomes which can commonly be found in drinking water.
Bacterial reference genomes were extracted from the NCBI's bacterial database, and include 268 genomes and plasmids of 12 species.
We then used three real MinION sequencing datasets to perform a search on this database in order to obtain the accuracy of mapping to the correct species.



Real data

Metagenomic database search for specie-level identification

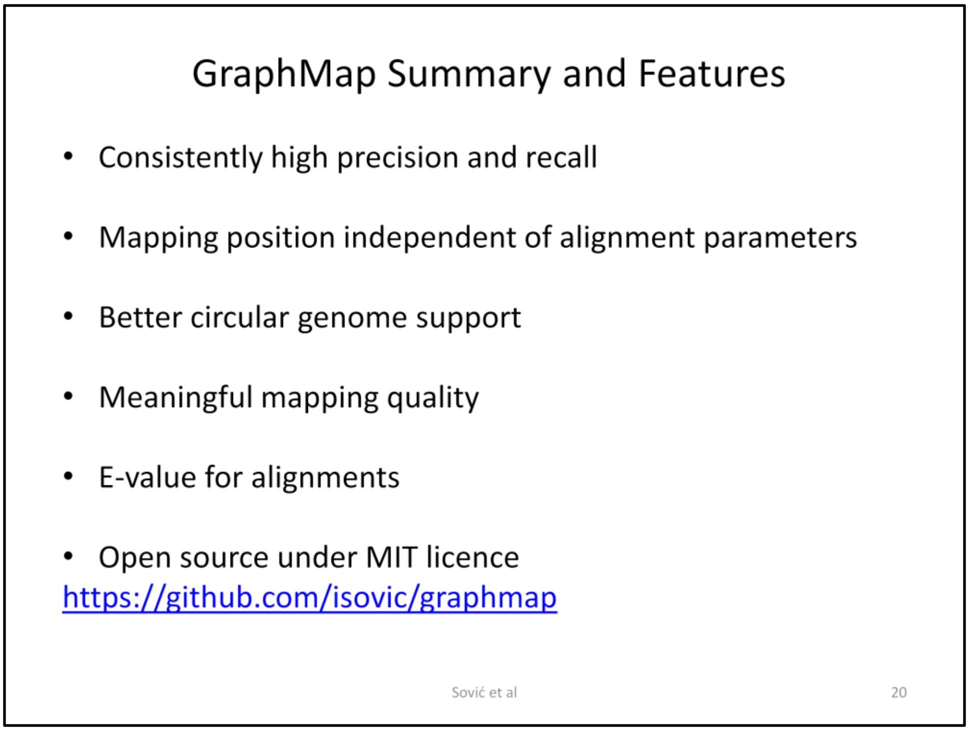
S. Typhi[1]	Precision (%)	Recall (%)	F1 (%)
GraphMap	97.42	56.48	71.50
BWA-MEM	97.63	44.76	61.38
LAST	97.20	34.43	50.85

E. Coli K-12[2]	Precision (%)	Recall (%)	F1 (%)
GraphMap	95.42	51.48	66.88
BWA-MEM	94.06	47.11	62.77
LAST	94.45	37.28	53.46

E. Coli UTI89	Precision (%)	Recall (%)	F1 (%)
GraphMap	99.05	88.36	93.40
BWA-MEM	98.42	85.41	91.46
LAST	95.37	65.37	77.57

[1] Ashton et al. Nature Biotechnology 33, 296–300 (2015)
[2] Quick & Loman ,GigaScience Database http://dx.doi.org/10.5524/100102 (2014)

Results obtained on all three datasets show that GraphMap performs better in all cases, which is reflected by the F1 measure.
Although precision is high for all mappers, the highest variation amongst methods can be seen in recall, especially for data generated on older versions of MinION sequencing workflows.



GraphMap Summary and Features

- Consistently high precision and recall
- Mapping position independent of alignment parameters
- Better circular genome support
- Meaningful mapping quality
- E-value for alignments
- Open source under MIT licence

https://github.com/isovic/graphmap

In summary,
GraphMap has consistently shown very high precision and recall across all datasets and performed tests.
An important feature of GraphMap is that the mapping position is not dependant of alignment parameters.

Other features include: option for mapping to circular genomes, meaningful mapping quality and E-value for alignments.
Finally,
GraphMap is open source and freely available on GitHub.
Give it a try! We are looking forward to your feedback!

GraphMap also provides a useful option for mapping on circular genomes which prevents the drop in coverage near chromosomal ends, often seen with other methods.
Because all equally good alignments are captured, GraphMap provides a meaningful mapping quality.
E-value is also calculated for each reported alignment and outputted to the SAM file in a custom field.