Fast and sensitive mapping of nanopore sequencing reads with GraphMap

Niranjan Nagarajan

Genome Institute of Singapore

Computing Revolution

1960's

1980's

2015

Genomics Revolution

2003

2011

2014

Research → Medical Genomics → Consumer Genomics

W
I
M
P

New Results

What's in my pot? Real-time species identification on the MinION

Sissel Juul, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett, Daniel J Turner

doi: http://dx.doi.org/10.1101/030742

Abstract | Info/History | Metrics | Preview PDF

Abstract

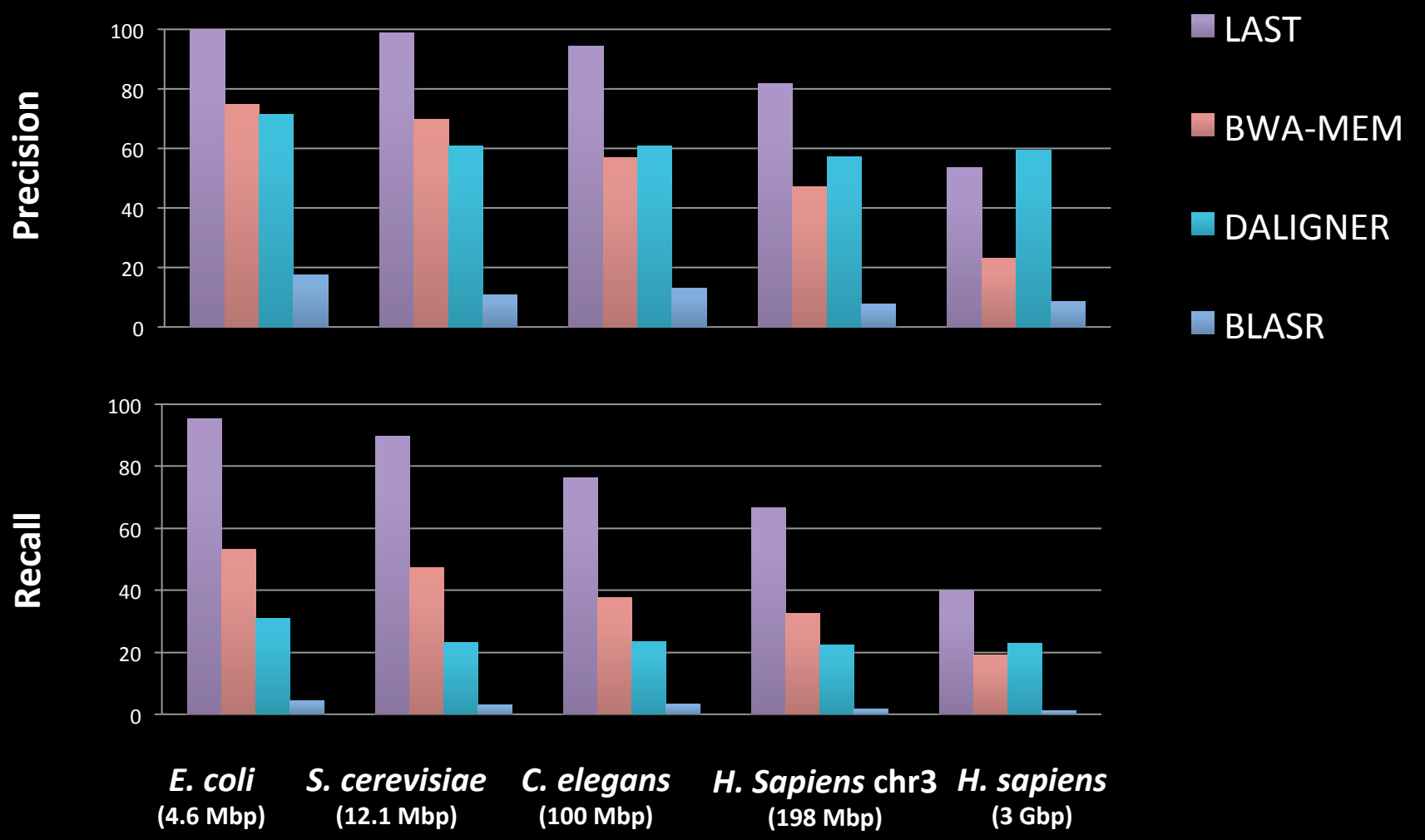
Whole genome sequencing on next-generation instruments provides an unbiased way to identify the organisms present in complex metagenomic samples. However, the time-to-result can be protracted because of fixed-time sequencing runs and cumbersome bioinformatics workflows. This limits the utility of the approach in settings where rapid species identification is crucial, such as in the quality control of food-chain components, or in during an outbreak of an infectious disease. Here we present What's in my Pot? (WIMP), a laboratory and analysis workflow in which, starting with an unprocessed sample, sequence data is generated and bacteria, viruses and fungi present in the sample are classified to subspecies and strain level in a quantitative manner, without prior knowledge of the sample composition, in approximately 3.5 hours. This workflow relies on the combination of Oxford Nanopore Technologies' MinION ™ sensing device with a real-time species identification bioinformatics application.

D
R
.

L
O
O

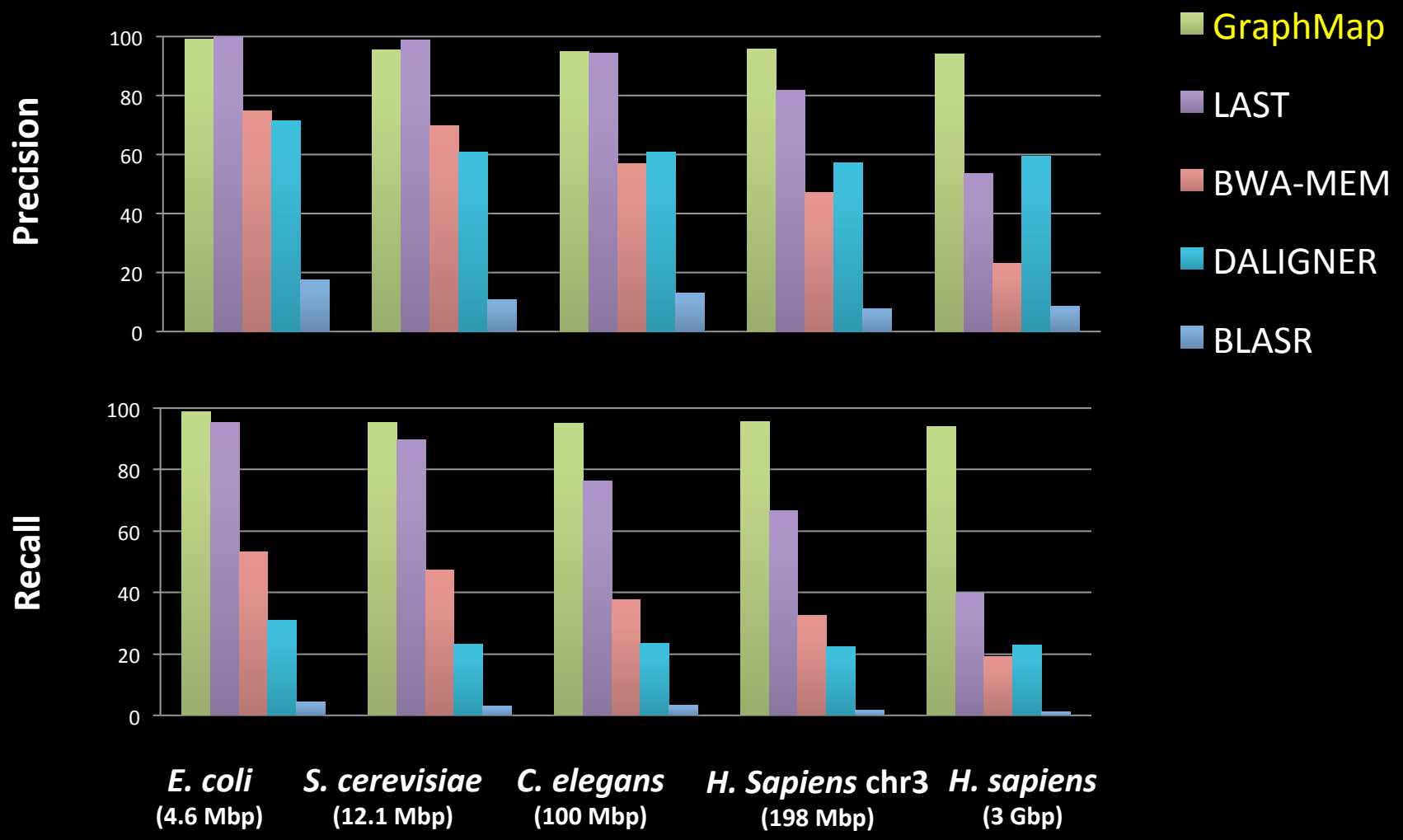
Operating System = Sequence Alignment

Mapping nanopore reads to human genome



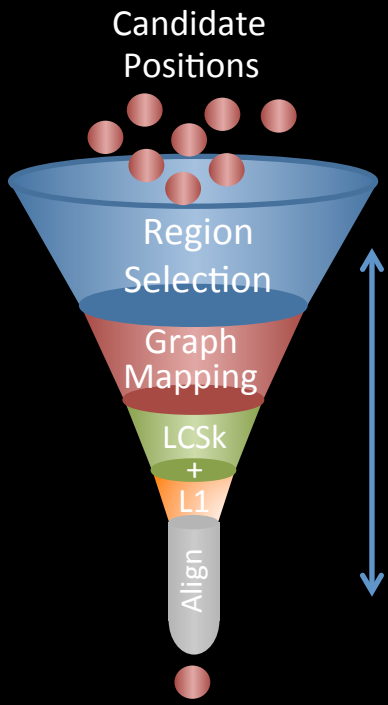
Simulated reads based on 1D error profiles for *E. coli* dataset from Quick *et al* 2014

Mapping nanopore reads to human genome



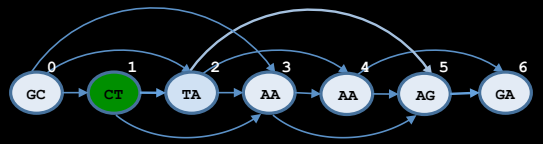
Simulated reads based on 1D error profiles for *E. coli* dataset from Quick *et al* 2014

GraphMap Design

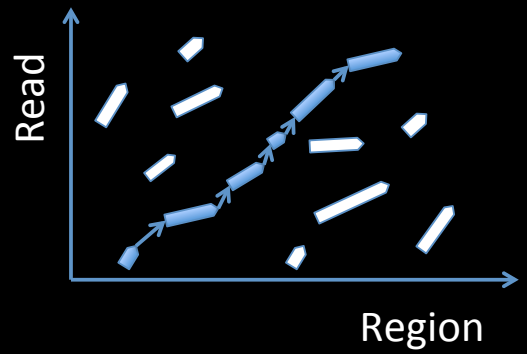


http://www.nature.com/ncomms/2016/160415/ncomms11307/full/ncomms11307.html

1. Gapped Spaced Seeds



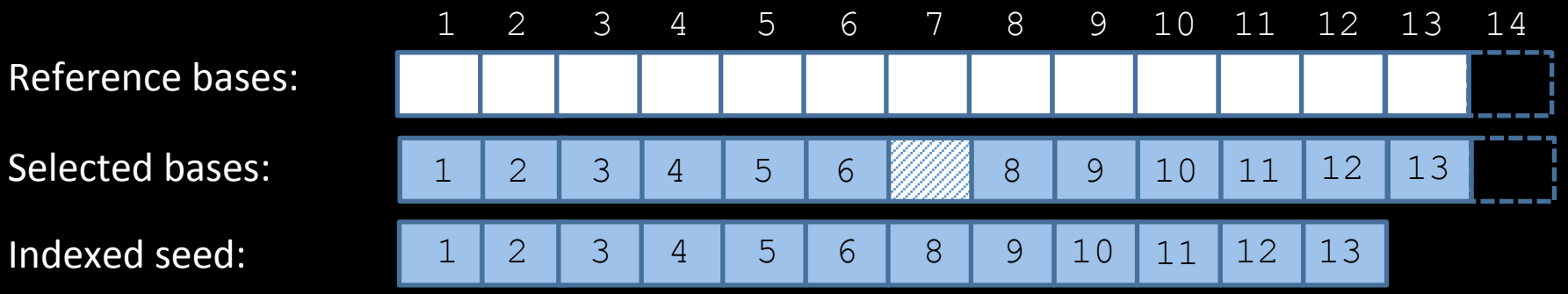
2. Graph Mapping



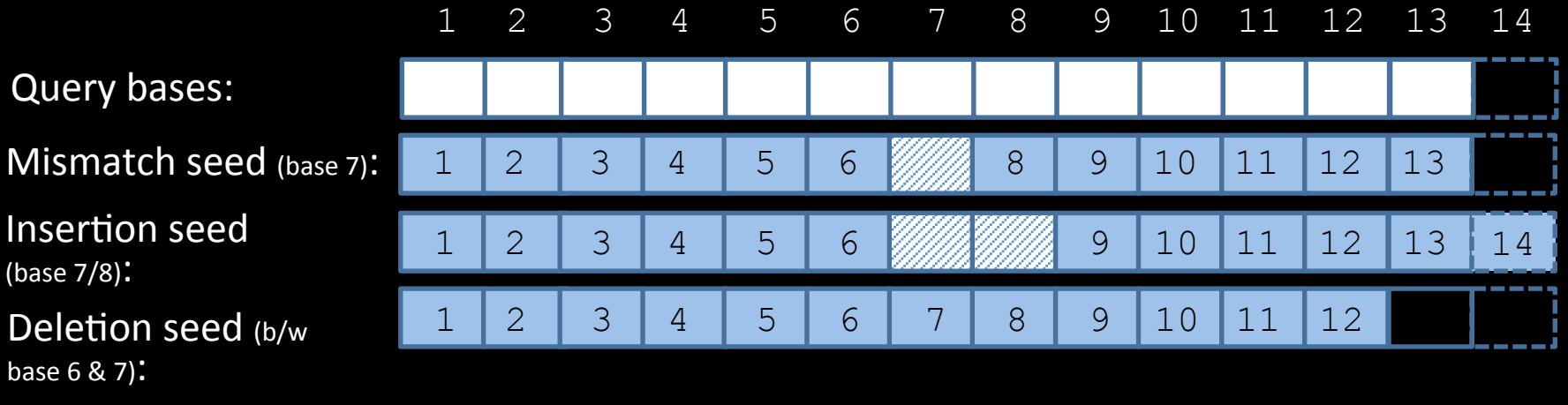
3. LCSk Chaining

1. Gapped Spaced Seeds

Index construction

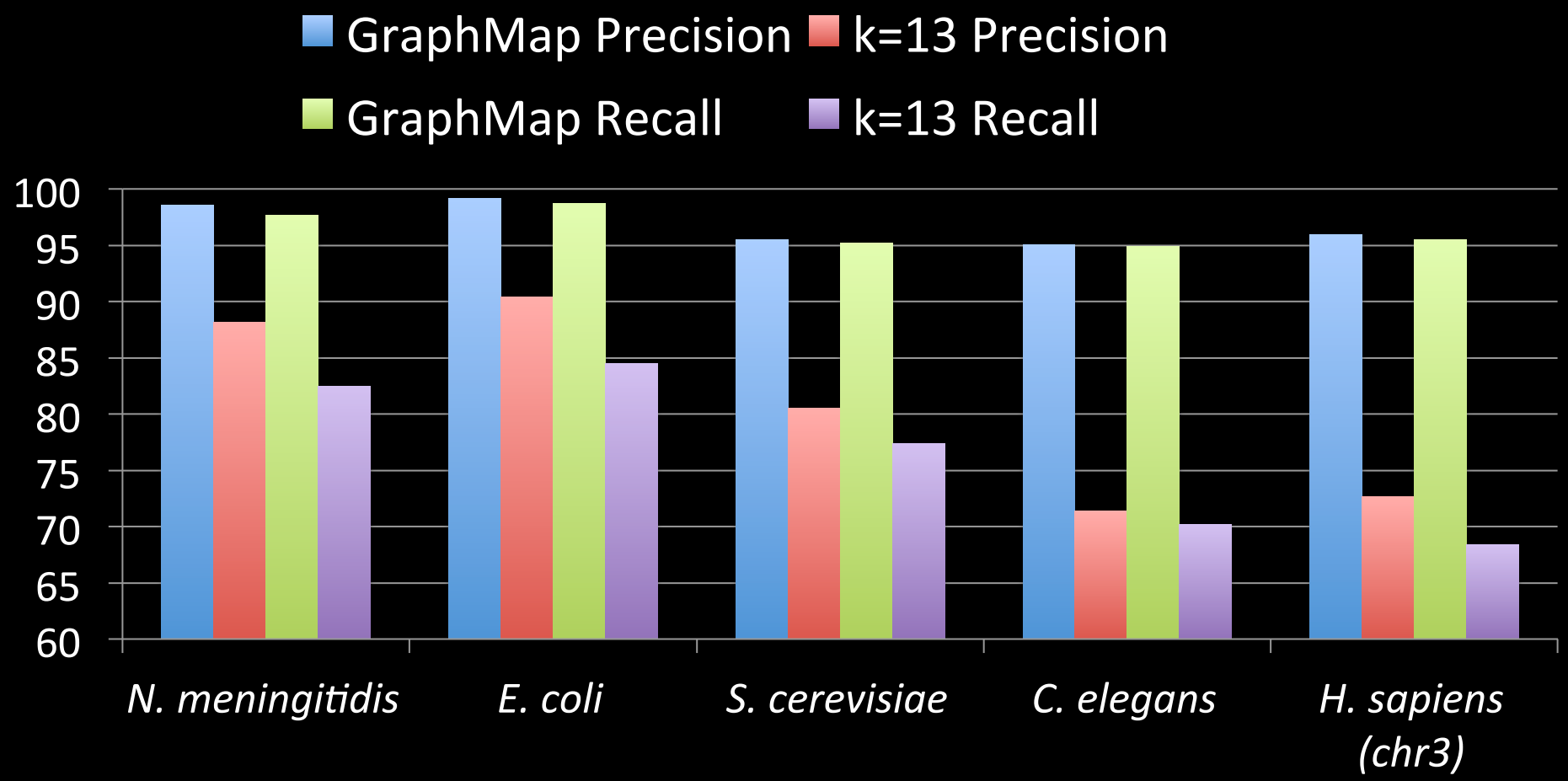


Index lookup



Shapes: 6-1-6, 4-1-4-1-4

Gapped spaced seeds retain sensitivity and recall

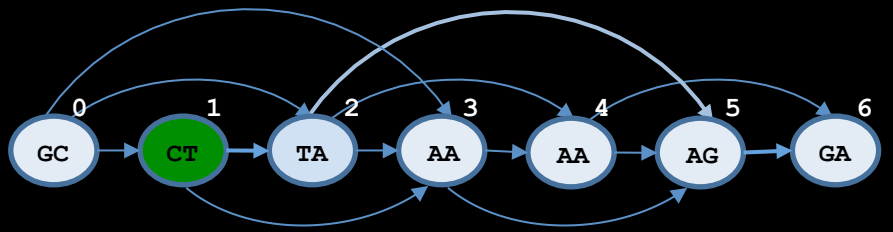


2. Graph Mapping to get Anchors

Graph based search for MEMs with indels

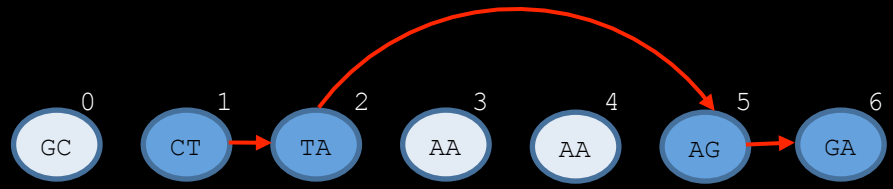
Initial graph

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



Alignment graph

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

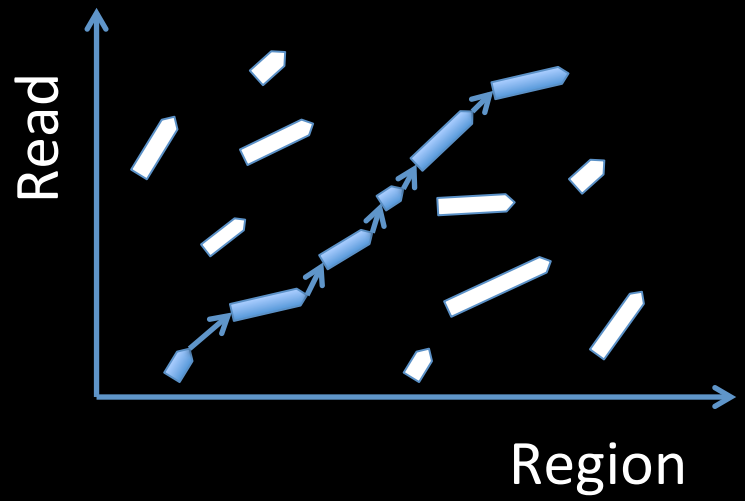


Final anchor graph

Vertex-centric parallelism

3. Chaining Anchors with LCSk

Motivation: Longest Common Subsequence (LCS) provides fast alignments but allows arbitrary insertions and deletions



Pavetic *et al* 2014

O(n•log(n))

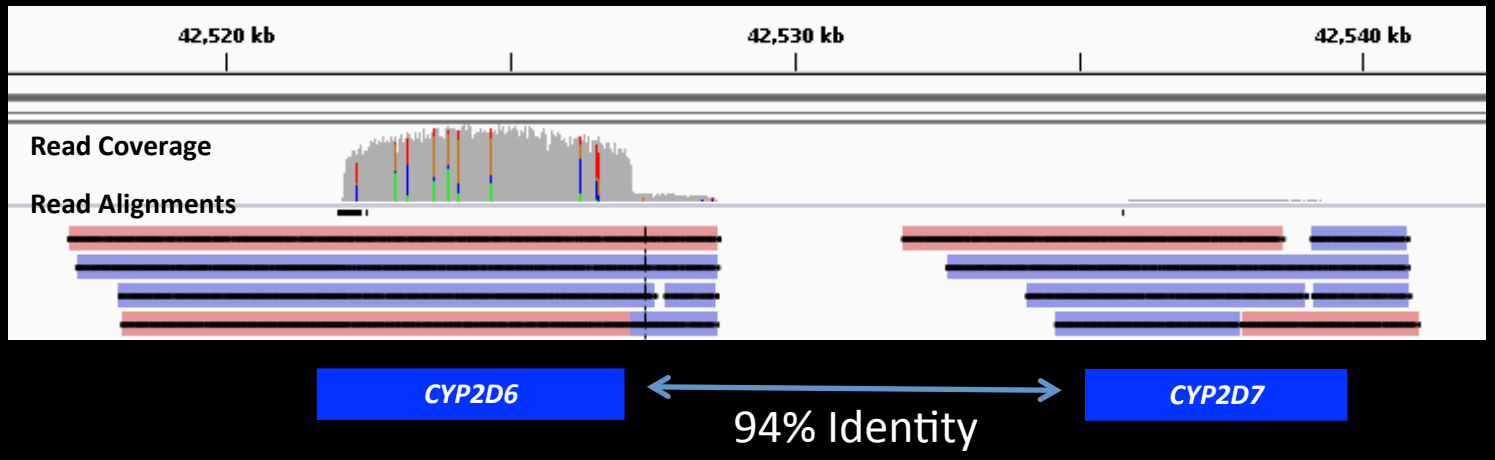
Approximate alignments evaluated to select best region

Other Features

1. Different Alignment Options (marginAlign)
2. Mapping Quality, BLAST-like E-value
3. Circular Genome Support
4. Technology Agnostic (w/o tuning parameters)
   - Illumina, PacBio, Ion Torrent

https://github.com/isovic/graphmap

Application: SNV Calling (Ammar et al. 2015)

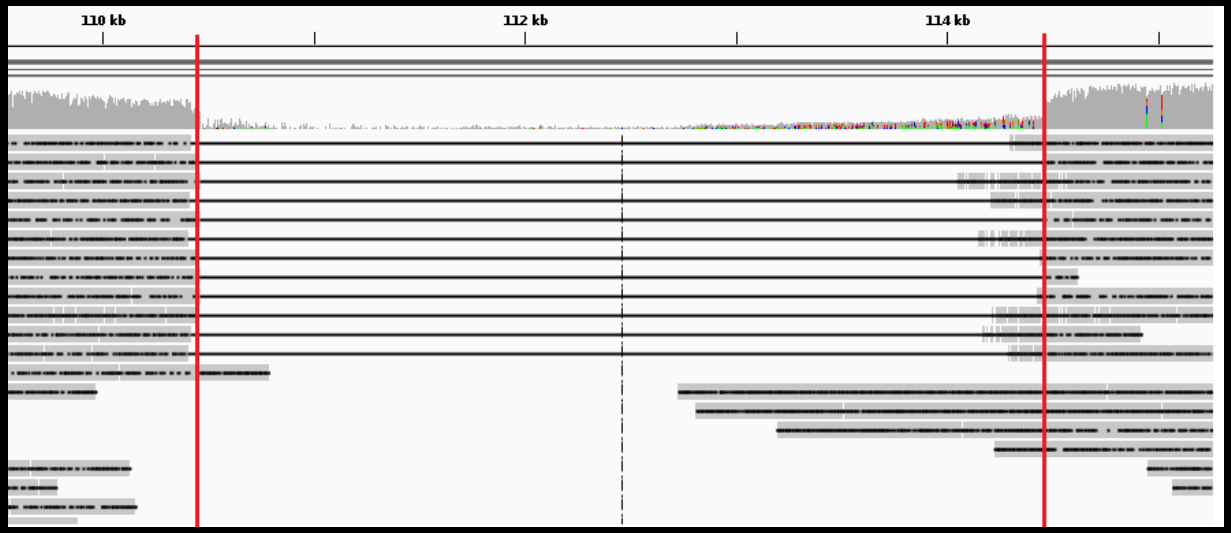


***CYP2D6* (chr22)**
# of on-target reads
GraphMap: 6879
BWA-MEM: 5900
LAST: 5032
BLASR: 2284
marginAlign: 4683
DALIGNER: 1451

	LAST	marginAlign	BWA-MEM	BLASR	DALIGNER	GraphMap
Precision (%)	94	100 (36)	96	100	93	**96**
True Positives	49	1 (107)	47	43	75	**86**

LoFreq: https://github.com/csb5/lofreq; results in parentheses for marginCaller

Application: Structural Variants (large Insertions and Deletions)



4kbp deletion spanned by GraphMap

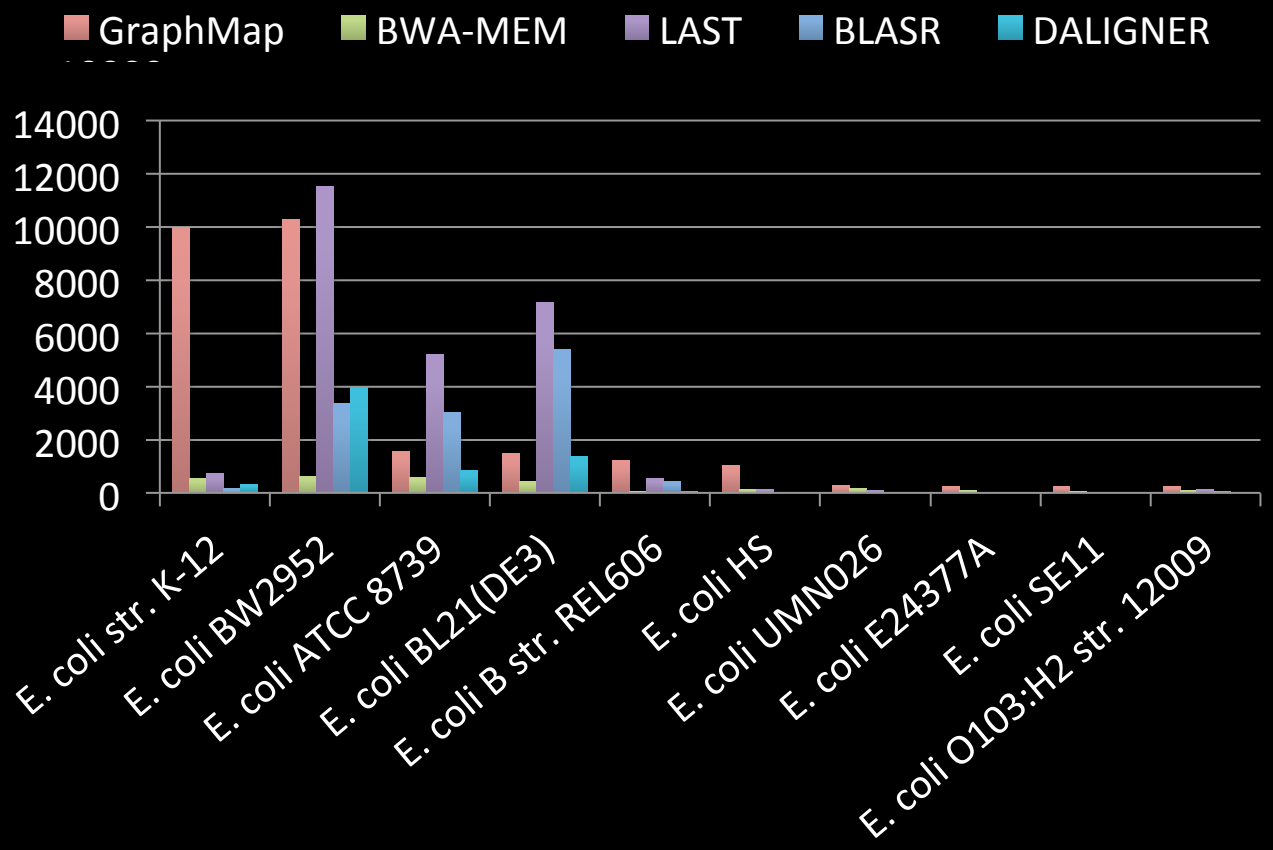
Data from Quick *et al* 2014 mapped to mutated reference

	LAST	marginAlign	BWA-MEM	BLASR	DALIGNER	GraphMap
Precision (%)	0	50	67 (90)	94	0	**100**
Recall (%)	0	5	10 (45)	75	0	**100**
F_1 Score (%)	0	9	17 (60)	83	0	**100**

Window size > 20bp; AF > 15%; Results from LUMPY are in parentheses

Application: Pathogen Identification

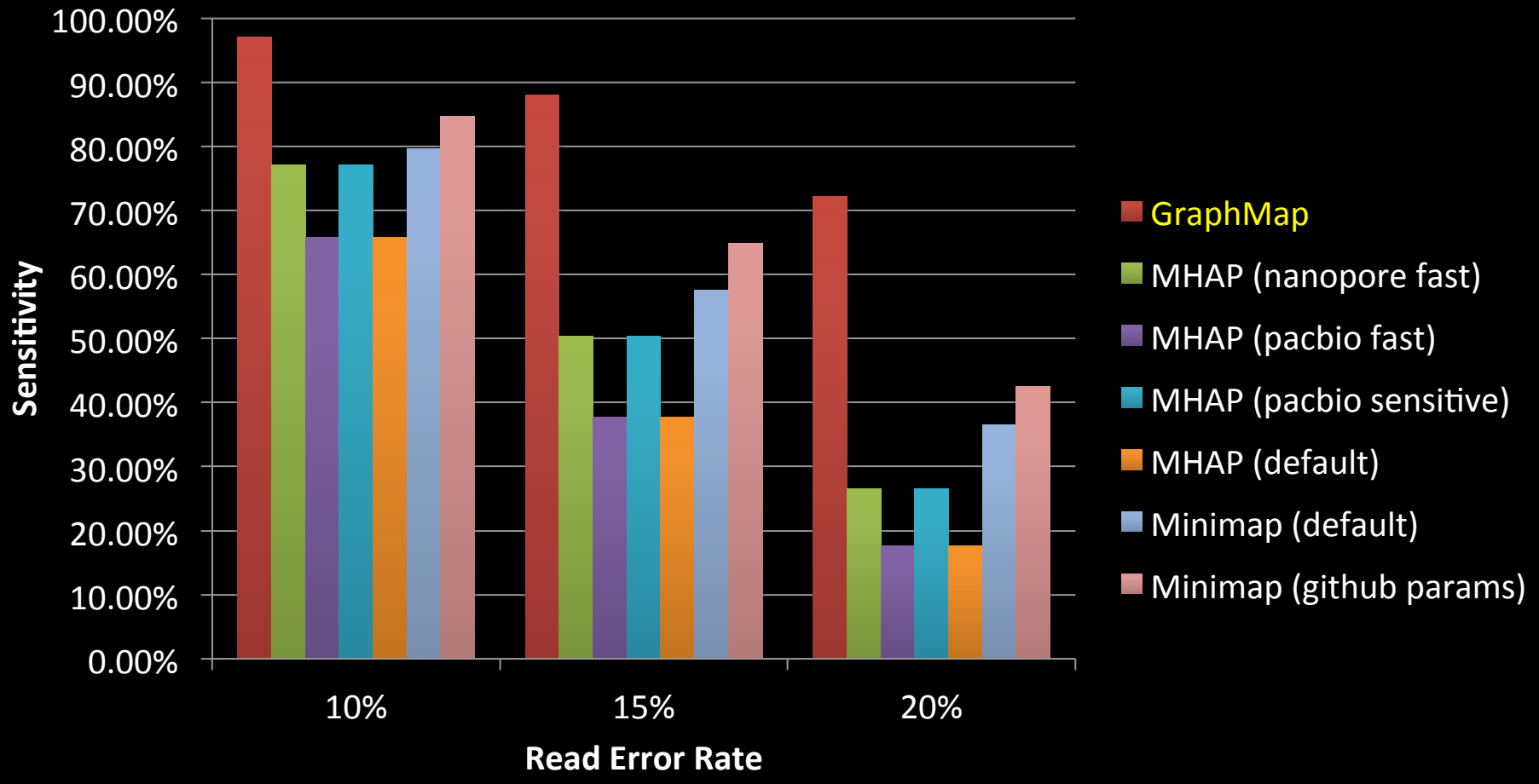
Database of 258 reference sequences



Data from Quick *et al* 2014

K-12 and BW2952 have >99% identity

GraphMap for Assembly



30X coverage simulated reads; Nanopore 2D error profile; *E. coli* genome

Ivan Sovic

Mile Sikic

Swaine Chen

Croatian Science Fundation project UIP-11-2013-7353

Genome Institute of Singapore
A*STAR

Oxford NANOPORE Technologies

https://github.com/isovic/graphmap
http://www.nature.com/ncomms/2016/160415/ncomms11307/full/ncomms11307.html


Genome