Metode sekvenciranja u preciznoj medicini

izv. prof. Mile Šikić
Sveučilište u Zagrebu,
Fakultet elektrotehnike i računarstva
Laboratorij za bioinformatiku i računalnu biologiju
Bioinformatics Institute, A*STAR, Singapore

Nieprecizna medicina

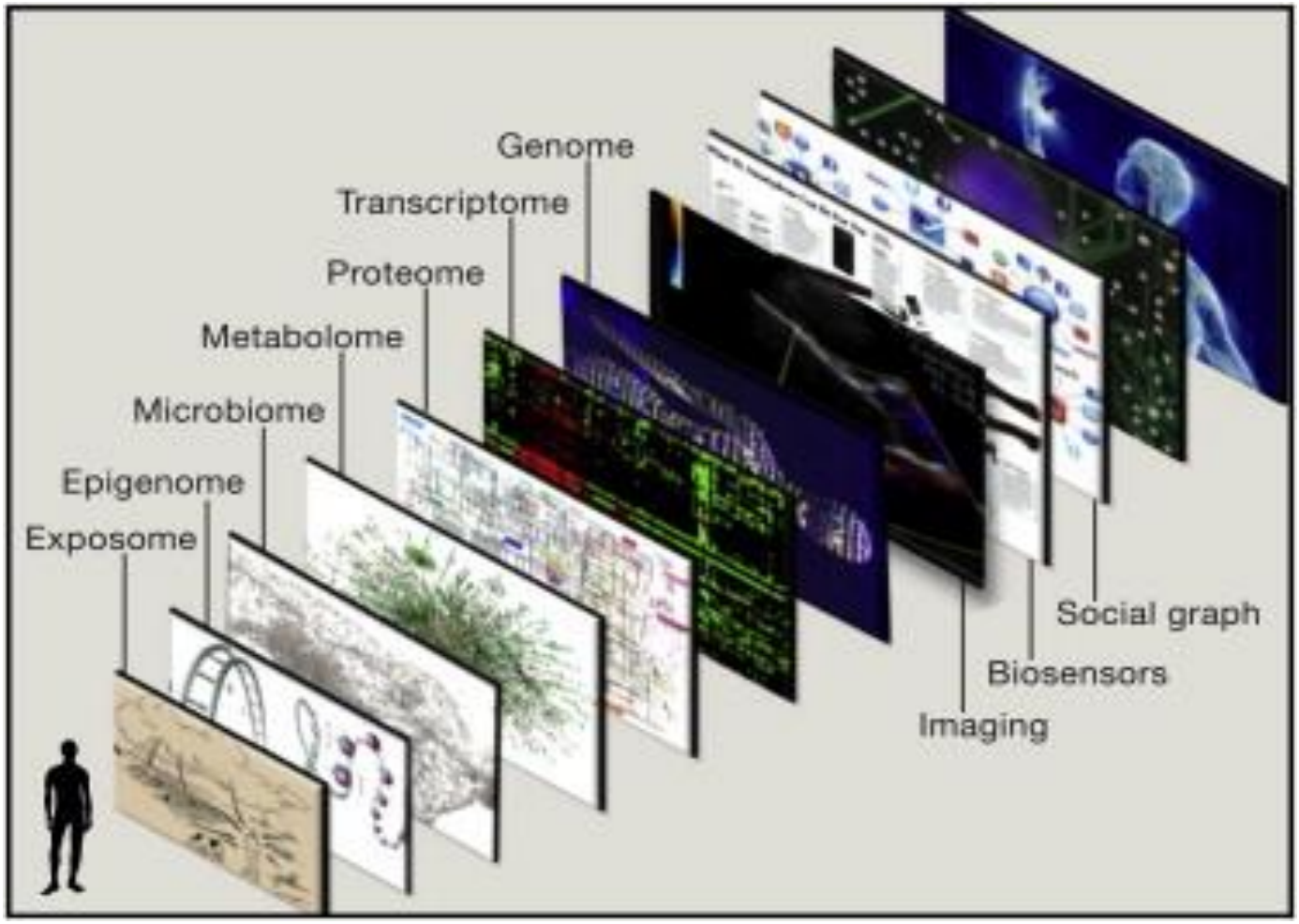
Adapted from Nicholas J. Schork Personalized medicine: Time for one-person trials, Nature, 2015

Medicina danas (US)

- >60% lažno pozitivnih nalaza u masovnim pregledima*
- 1 od 4 pacijenta su se dodatno oboljeli u bolnici
- >12 milijuna ozbiljnih dijagnostičkih pogrešaka
- Medicinske pogreške i nuspojave su 4-ti vodeći uzročnik smrti
- U ~80% pacijenata vodeći prepisani lijekove ne djeluju

*Godišnja mamografija u 10 godina

Precizna medicina



Topol, E. Individualized Medicine from Prewomb to Tomb, Cell, Volume 157, Issue 1, p241–253, 27 March 2014

Inicijativa za preciznu medicinu ($215 million)

Mission statement:

To enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized care.

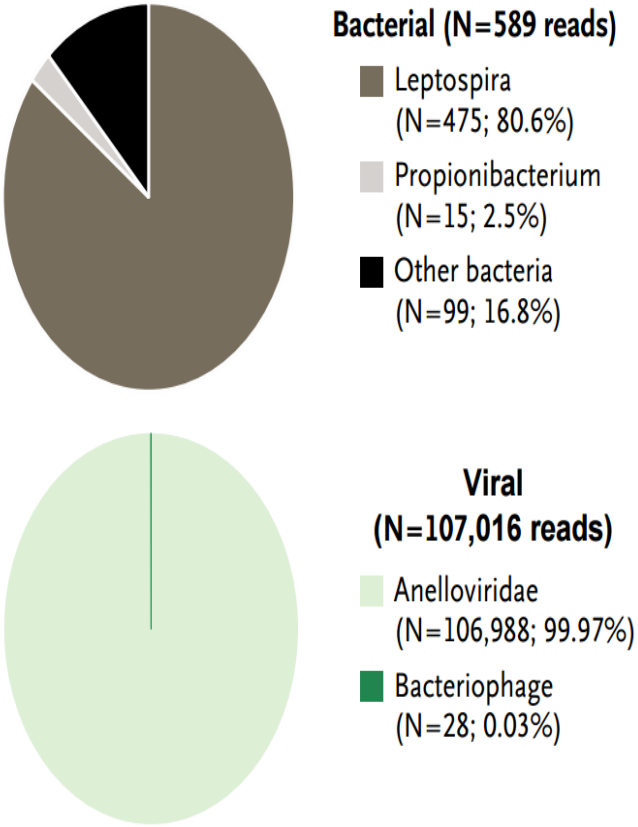
Stara nasuprot nove medicine

Old Medicine	New Medicine
Population-Based	Individualized
One-Off, Doctor's Office	Real-Time Streaming, Real World
Doctor Ordered Data	Patient Generated Data
Doctor's Notes, Unshared	Our Notes, Patient Edited
Information Owned by Doctors and Hospitals	Information Owned by Rightful Owner
Expensive, Big-Ticket Tech	Cheap Chips, Moore's Law
Data Limited	Panoromic

Author: Eric Topol

Učinkovita dijagnoza neuroleptospiroze koristeći sekvenciranje

Joshua Osborne, 14 godina



The New York Times

The NEW ENGLAND JOURNAL of MEDICINE

Projekt određivanja ljudskog genoma

Određivanje ljudskog genoma u brojkama

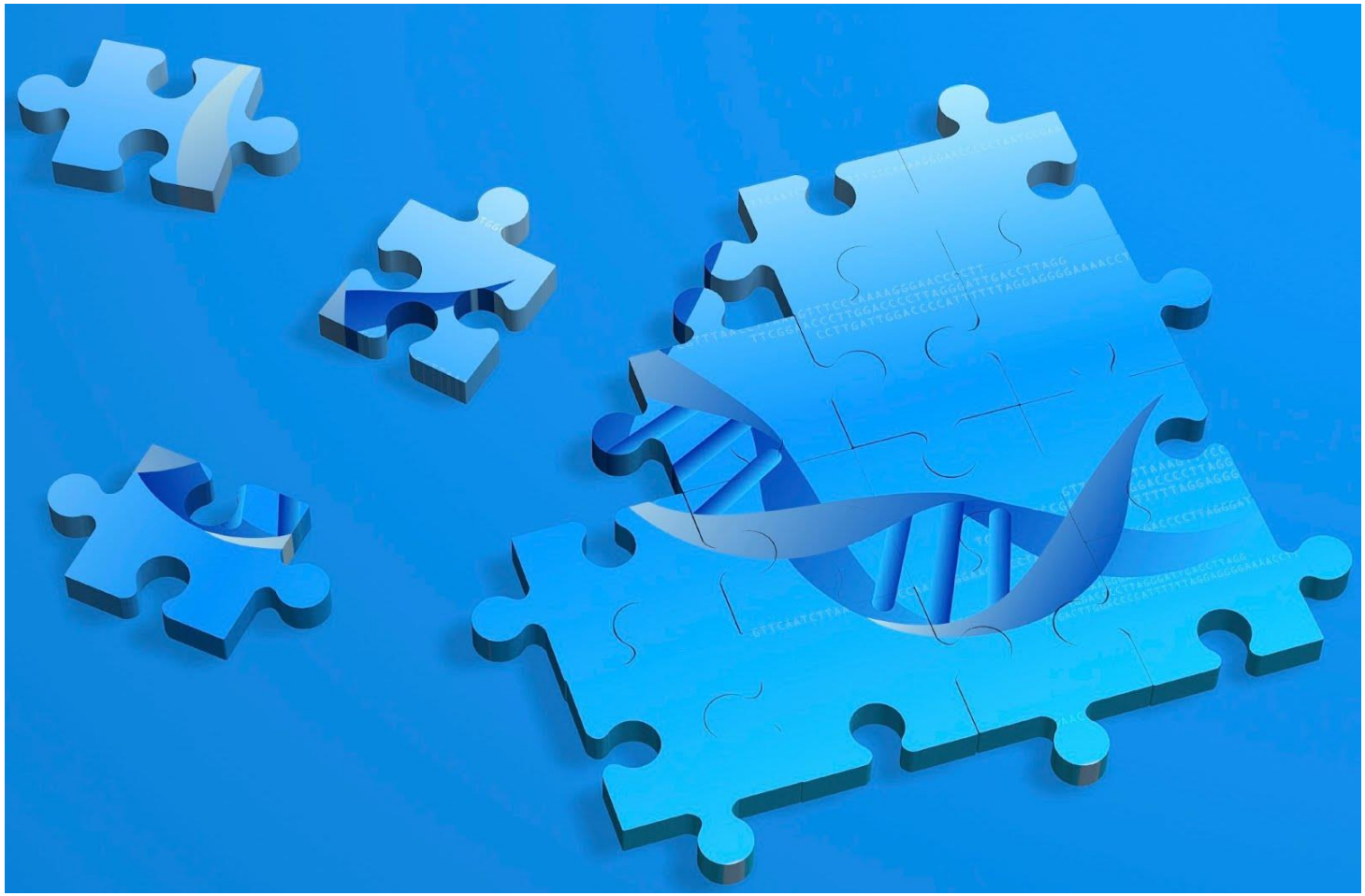
- 3 milijarde nukleotida
- 3 milijarde USD
- 13 godina
- 6 uključenih država (US, UK, Francuska, Njemačka, Japan, Kina)
- 20 ustanova

DNA sekvenciranje

DNA sekvenciranje

DNA sekvenciranje

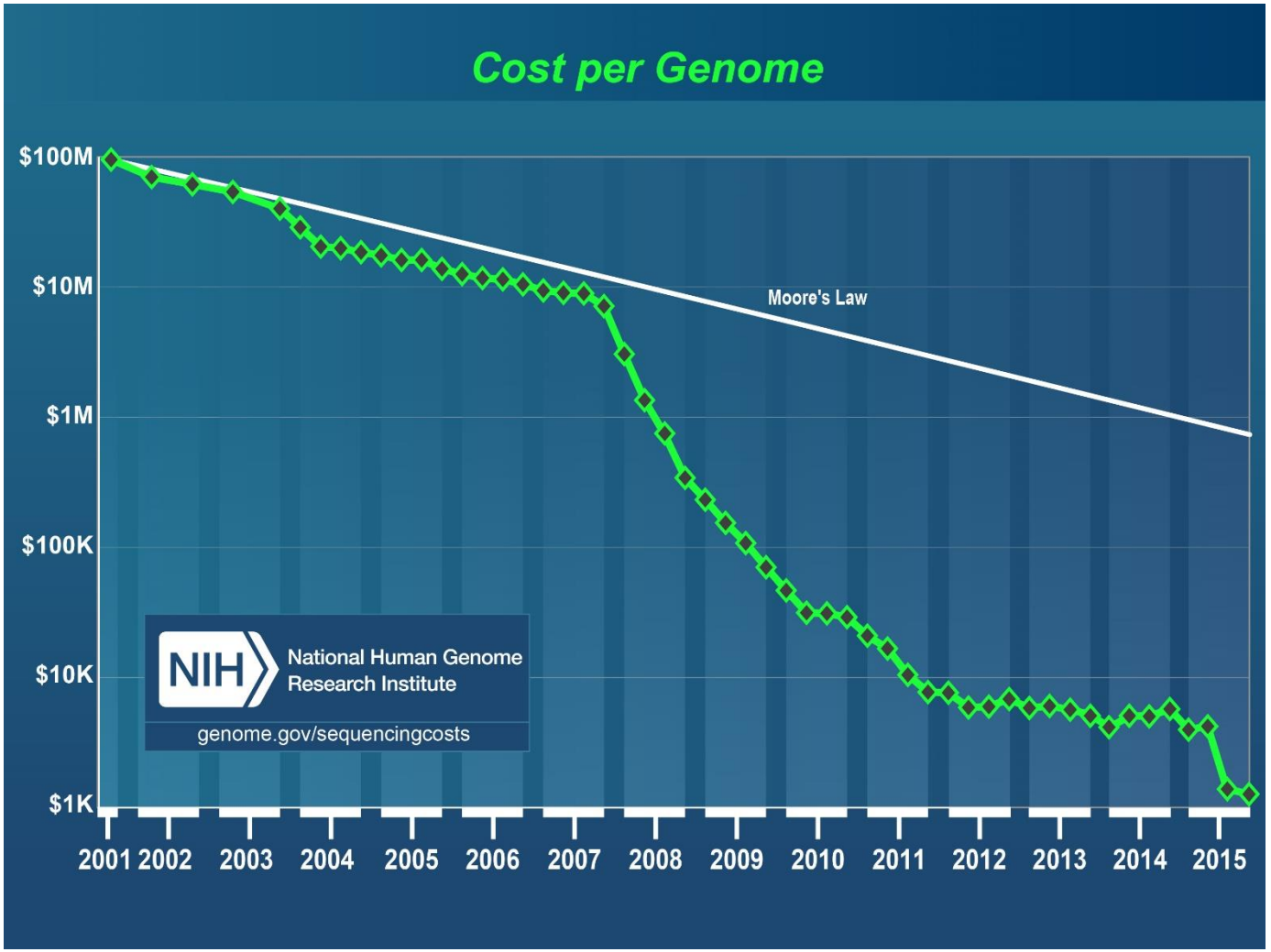
DNA sekvenciranje



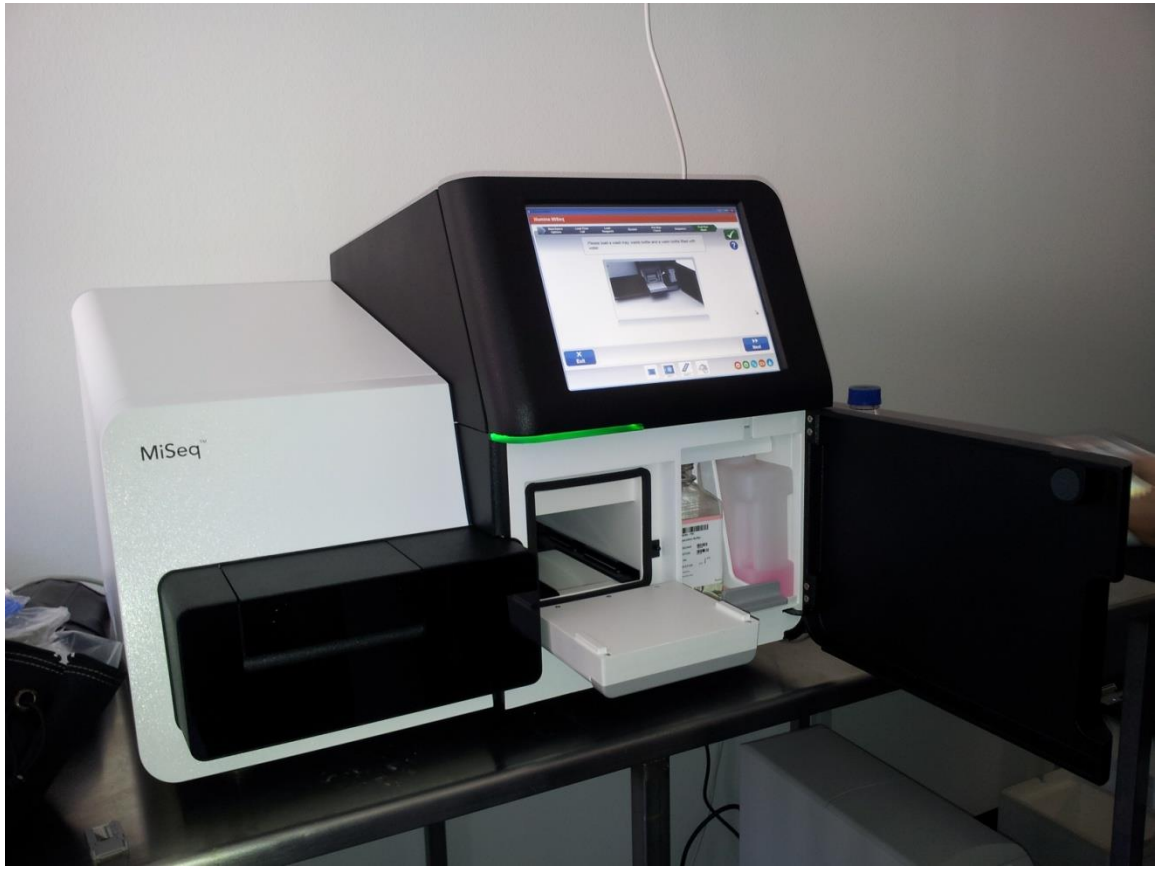
Projekt sekvenciranja ljudskog genoma

Izvor: Wikipedia

Troškovi sekvenciranja

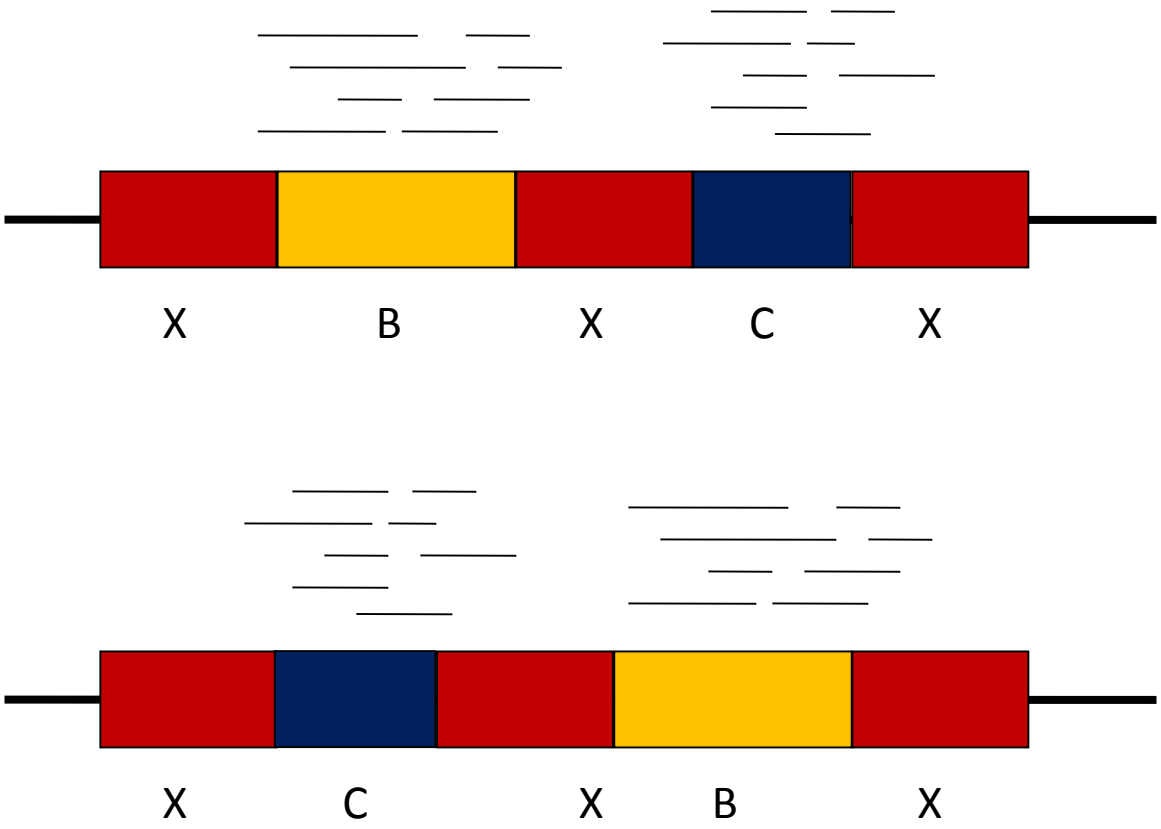


Druga generacija sekvenciranja

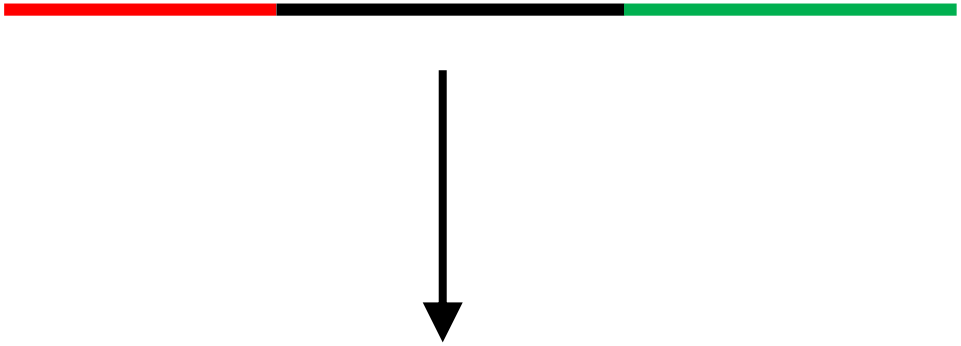


Izvor: Illumina

Problemi u sastavljanju genoma (ponavljanja)

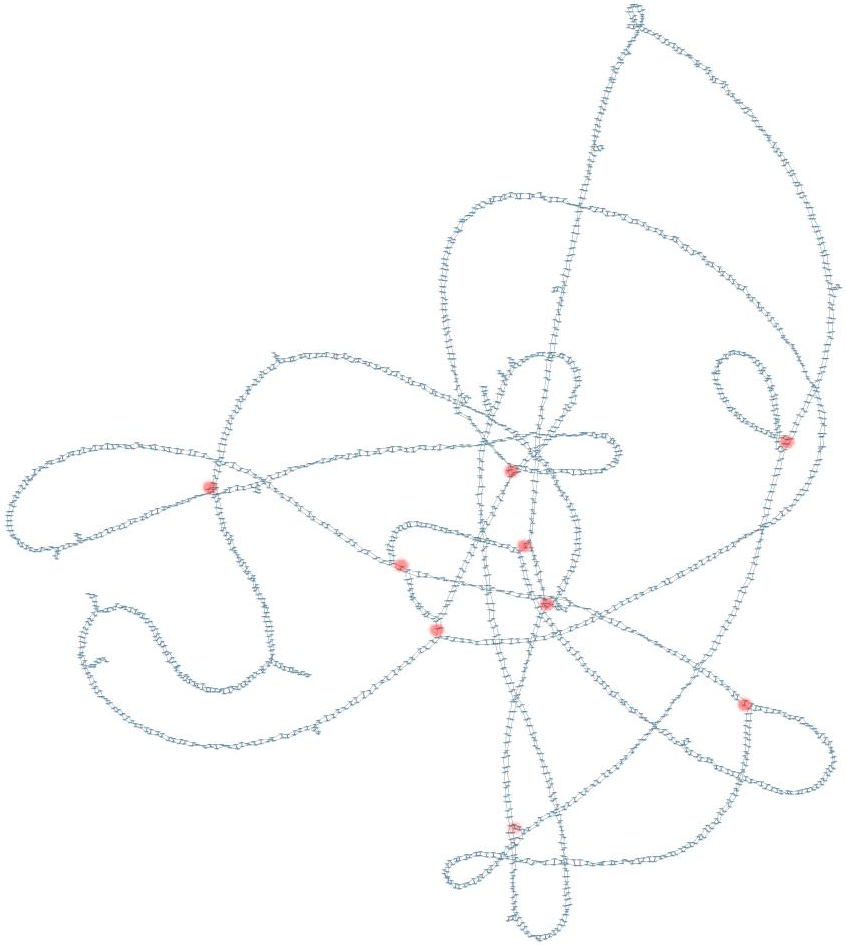


Problemi u sastavljanju (kimerna očitanja)



AGACGACTTTAGATACTGGGTACTAGAACCCTCAGG

Problemi u sastavljanju

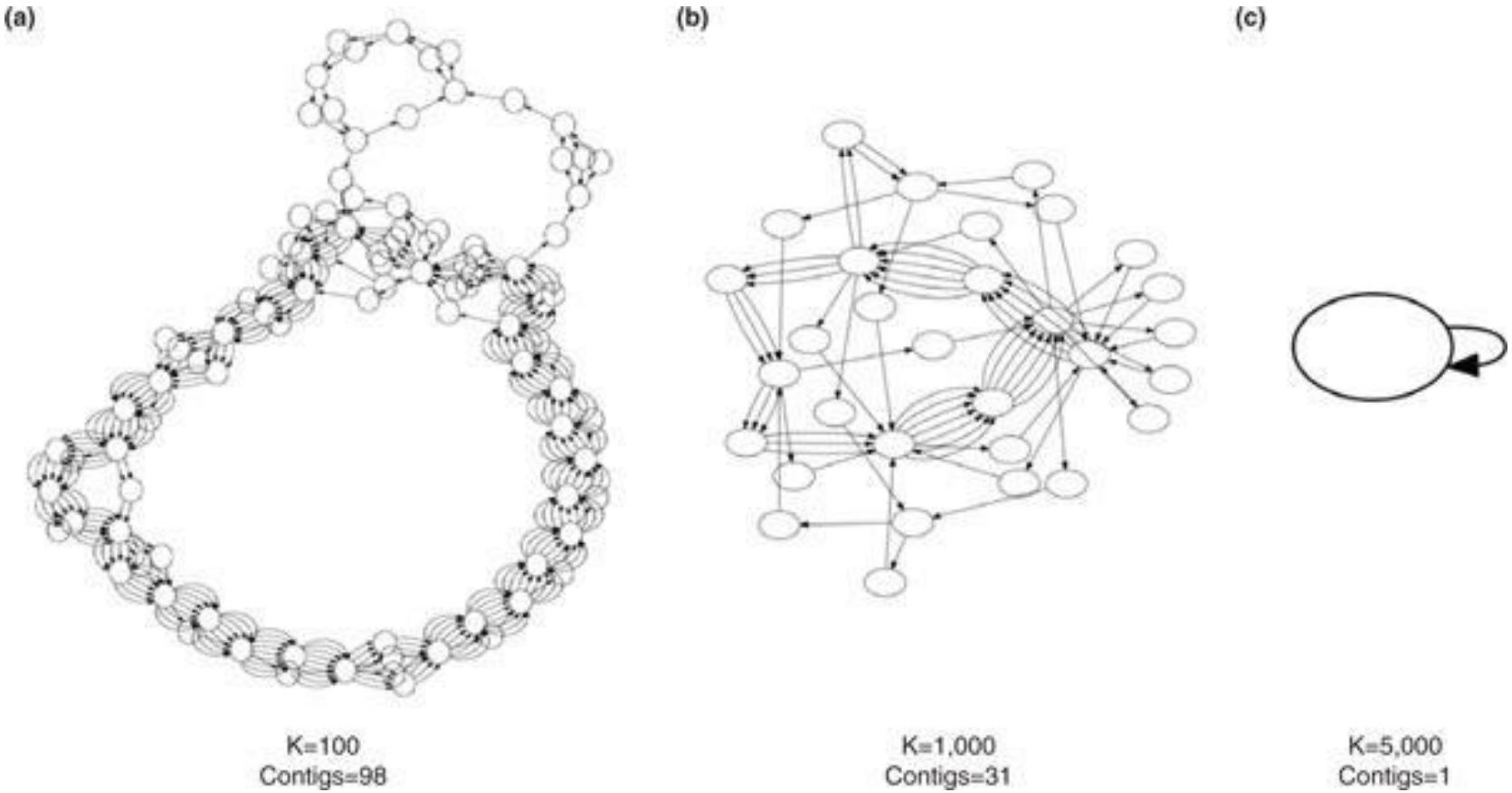


Dugačka očitanja

Pacific Biosciences Sequel

Oxford nanopore technologies MinION MK I

Dugačka očitanja otpetljavaju graf



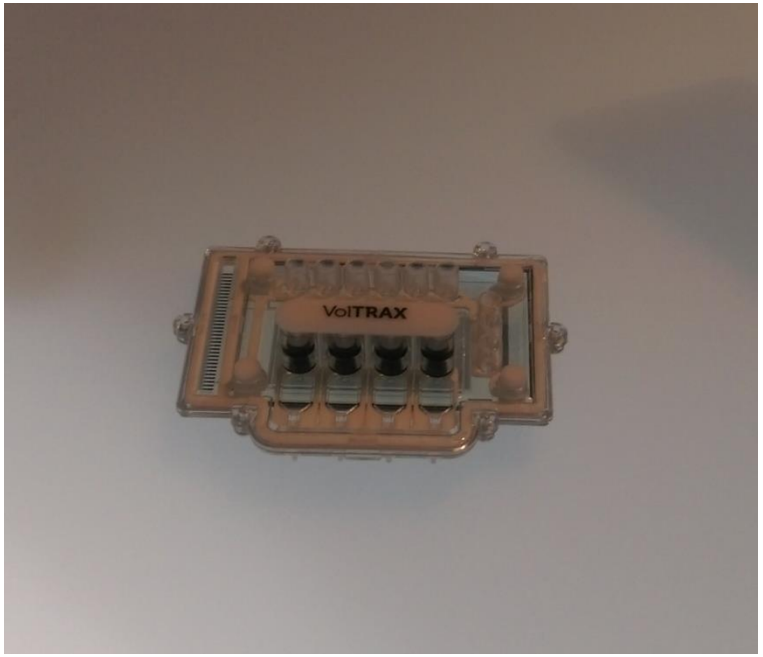
PromethION

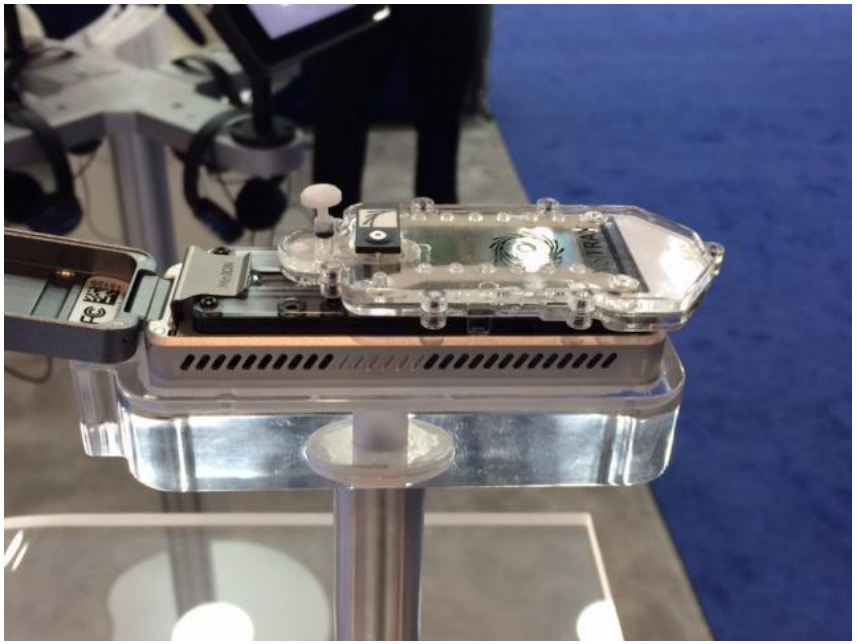
- Novija generacija uređaja za sekvenciranje
- Puno veća propusnost

Izvor: Oxford Nanopore Technologies

Pripremanje uzoraka

Pripremanje uzorka - Voltrax





Izvor: Oxford Nanopore Technologies

SmidgION

Detekcija ebole

Photograph: Tommy Trenchard/EMLabs

- Pomoć u identificiranju mutacija u realnom vremenu
- Bez potrebe za prebacivanje uzoraka krvi u visoko kontrolirane laboratorije
- Rezultati unutar 24 sata
- J. Quick et al. Real-time, portable genome sequencing for Ebola surveillance, *Nature*, February 2016

Temeljni računalni problemi

- Sastavljanje genoma – genom za vrstu unaprijed poznat (mapiranje)
- Sastavljanje genome – genom za vrstu nepoznat (*de novo* sastavljanje)
- Sastavljanje kada je u genomu prisutno nekoliko uzoraka
- Određivanje prisutnih RNA molekula

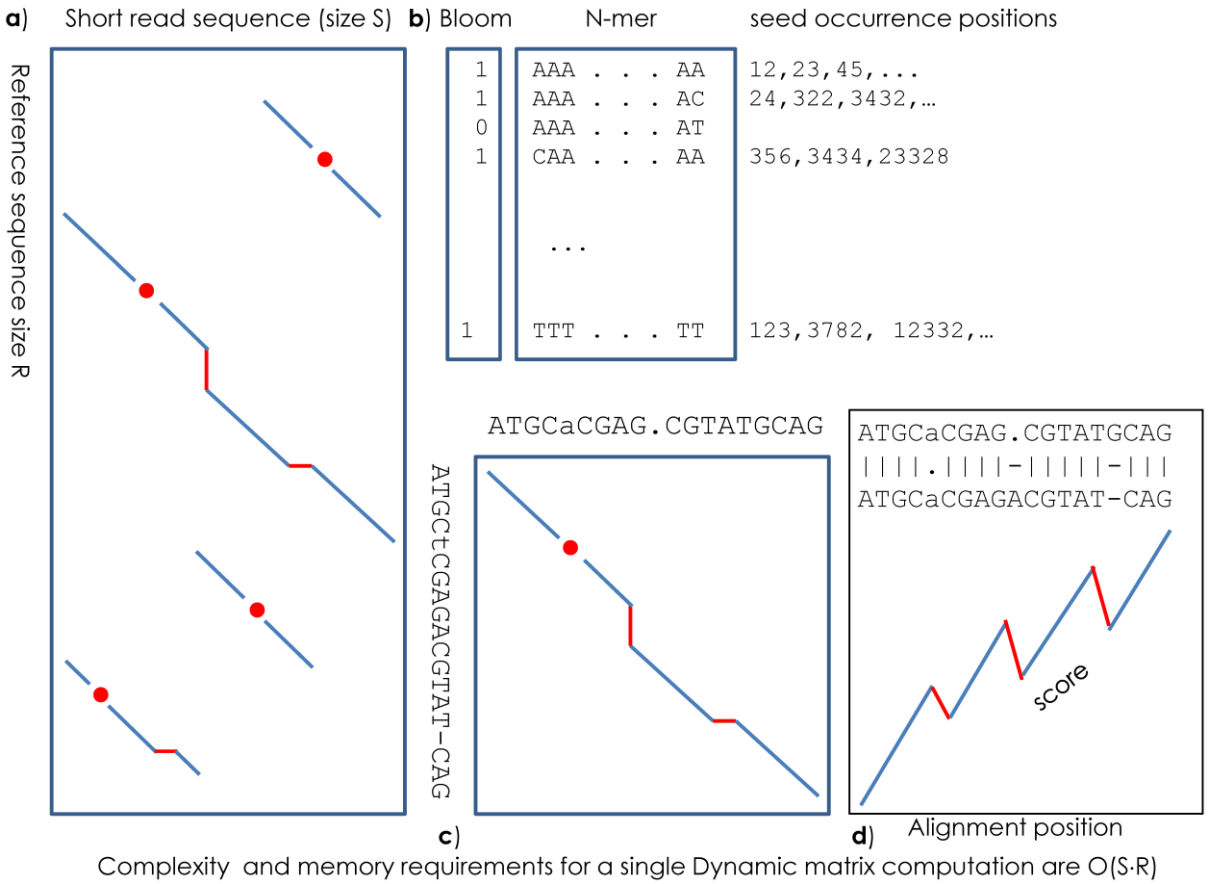
Mapiranje


>1:866511 in NA12878/NA12878.bam
Ref: TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGCTGTCCGCTGTCTCAGCGTGAGC
60> TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
60< TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----C
60> TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CC
60< TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCC
60< TCcGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CC
60> TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----cCC
60< TCCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----C
60< CCGAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CC
60> GAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCt----cCC
70< GAGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT
70< gagaGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT
60> AGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCctCCCT----CCCt----ccccc
70< AGAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----C
70> GAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CC
70< GAGGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CC
60< GGCGTCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
60> TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCt----CCC
60> TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCt----CCC
60< TCCTGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----CCC
60< ccTGCAGGTAGGaGCCGTGcTGTGCGTGCATAAGAGGGGGCCGTGACTcCCCTCCCTCCCT----CCCT----CCCACCCCT
60> TGCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCtga
60> GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTC----CCCTCCCT----CCCT----cCC
60< GCAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGAC
60< GCAGgtAGGAGCCGTGCTGtGCGTGCATAAGAGGGGGCCGtGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGAC
60< CAGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACC
60< AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCG
60< AGGTAGGAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCG
60< GAGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
60< gaGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
60< gaGCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCT
60> GCCGTGCTGTGCGtGCATAAGAGGGGGCCGTGACTCCCCtCCCTCCCT----CCCt----CCCACCCCtGACCgtgccctgc
60< GCCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGC
60< CCGTGCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCT
60< tgCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTC
60< tgctgtGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTC
60< GCTGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCaCCCCTGACCGTGCCCTGCTGTCT
60< TGTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGC
60< GTGCGTGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGCT
60< tgcgtgcATAAGAGGGGGcCgtgaCTCCCCTCCCTCCCT----cCCT----CCCAcCCCTGACCGTGCCCTGCTGTCTGCTG
60< TGCATAAGAGGGGGCCGTGACTCCCCTCCCTCCCT----CCCT----CCCACCCCTGACCGTGCCCTGCTGTCTGCTGTCCG
60> AGAGGGGGCCGTGACTC----CCCTCCCT----CCCTCCCccccccccctgaccctgccctgctgtctggtgtcccctggct
60< gaGGGGGCcGTGACTCCCCtCCCTCCCt----CCCt----CCCACCCCTGACCGTGCCCTGCTGTCTGCTGTCCGCTGTCTC

Mapiranje – optimalno (ljudski genom 3 milijarde nukleotida)

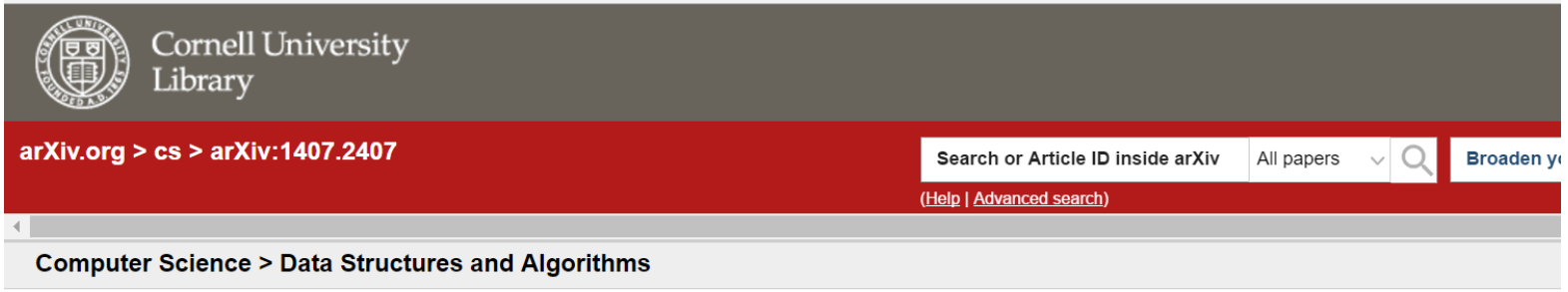
Algoritam	**Procesor**	**Vrijeme**
SW or NW	CPU i7	2000 godina
SW or NW	1000 * CPU i7	2 godine
SW or NW	GPU	60 godina
SW or NW	100 * GPU	7 mjeseci

Mapiranje postupak (heuristika)



Complexity and memory requirements for a single Dynamic matrix computation are O(S·R)

LCSk



$LCSk$++: Practical similarity metric for long strings

Filip Pavetić, Goran Žužić, Mile Šikić

(Submitted on 9 Jul 2014)

In this paper we present $LCSk$++: a new metric for measuring the similarity of long strings, and provide an algorithm for its efficient computation. With ever increasing size of strings occuring in practice, e.g. large genomes of plants and animals, classic algorithms such as Longest Common Subsequence (LCS) fail due to demanding computational complexity. Recently, Benson et al. defined a similarity metric named $LCSk$. By relaxing the requirement that the k-length substrings should not overlap, we extend their definition into a new metric. An efficient algorithm is presented which computes $LCSk$++ with complexity of $O((|X| + |Y|)\log(|X| + |Y|))$ for strings X and Y under a realistic random model. The algorithm has been designed with implementation simplicity in mind. Additionally, we describe how it can be adjusted to compute $LCSk$ as well, which gives an improvement of the $O(|X|\dot{|}Y|)$ algorithm presented in the original $LCSk$ paper.

Edlib

OXFORD ACADEMIC

Bioinformatics

Issues Advance Articles Publish ▾ Purchase Alerts About ▾

Article Contents

Comments (0)

CORRECTED PROOF

Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance

Martin Šošić; Mile Šikić ✉

Bioinformatics btw753. **DOI:** https://doi.org/10.1093/bioinformatics/btw753
Published: 31 January 2017 **Article history** ▾

Views ▾ PDF Cite Share ▾ Tools ▾

Summary: We present Edlib, an open-source C/C ++ library for exact pairwise sequence alignment using edit distance. We compare Edlib to other libraries and show that it is the fastest while not lacking in functionality and can also easily handle very large sequences. Being easy to use, flexible, fast and low on memory usage, we expect it to be easily adopted as a building block for future bioinformatics tools.
Availability and Implementation: Source code, installation instructions and test data are freely available for download at https://github.com/Martinsos/edlib, under the MIT licence. Edlib is implemented in C/C ++ and supported on Linux, MS Windows, and Mac OS.
Contact:mile.sikic@fer.hr
Supplementary information:
Supplementary data
are available at *Bioinformatics* online.

Issue Section: APPLICATIONS NOTE

GraphMap

Article | OPEN

Fast and sensitive mapping of nanopore sequencing reads with GraphMap

Ivan Sović, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen & Niranjan Nagarajan

Nature Communications **7**,
Article number: 11307 (2016)
doi:10.1038/ncomms11307
Download Citation

Bioinformatics DNA sequencing
Molecular biology

Received: 30 December 2015
Accepted: 11 March 2016
Published online: 15 April 2016

De novo sastavljanje (1000 CPU sati)

nature methods
Techniques for life scientists and chemists

ARTICLE PREVIEW

view full access options ▸

NATURE METHODS | BRIEF COMMUNICATION

A complete bacterial genome assembled *de novo* using only nanopore sequencing data

Nicholas J Loman, Joshua Quick & Jared T Simpson

Affiliations | **Contributions** | **Corresponding author**

Nature Methods **12**, 733–735 (2015) | doi:10.1038/nmeth.3444
Received 11 March 2015 | Accepted 22 May 2015 | Published online 15 June 2015

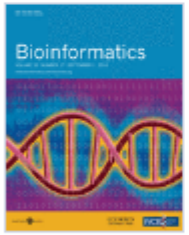
We have assembled *de novo* the *Escherichia coli* K-12 MG1655 chromosome in a single 4.6-Mb contig using only nanopore data. Our method has three stages: (i) overlaps are detected between reads and then corrected by a multiple-alignment process; (ii) corrected reads are assembled using the Celera Assembler; and (iii) the assembly is polished using a probabilistic model of the signal-level data. The assembly reconstructs gene order and has 99.5% nucleotide identity.

Može brže

OXFORD ACADEMIC

Bioinformatics

Issues | Advance Articles | Publish ▾ | Purchase | Alerts | About ▾



Volume 32, Issue 17
1 September 2016

Article Contents

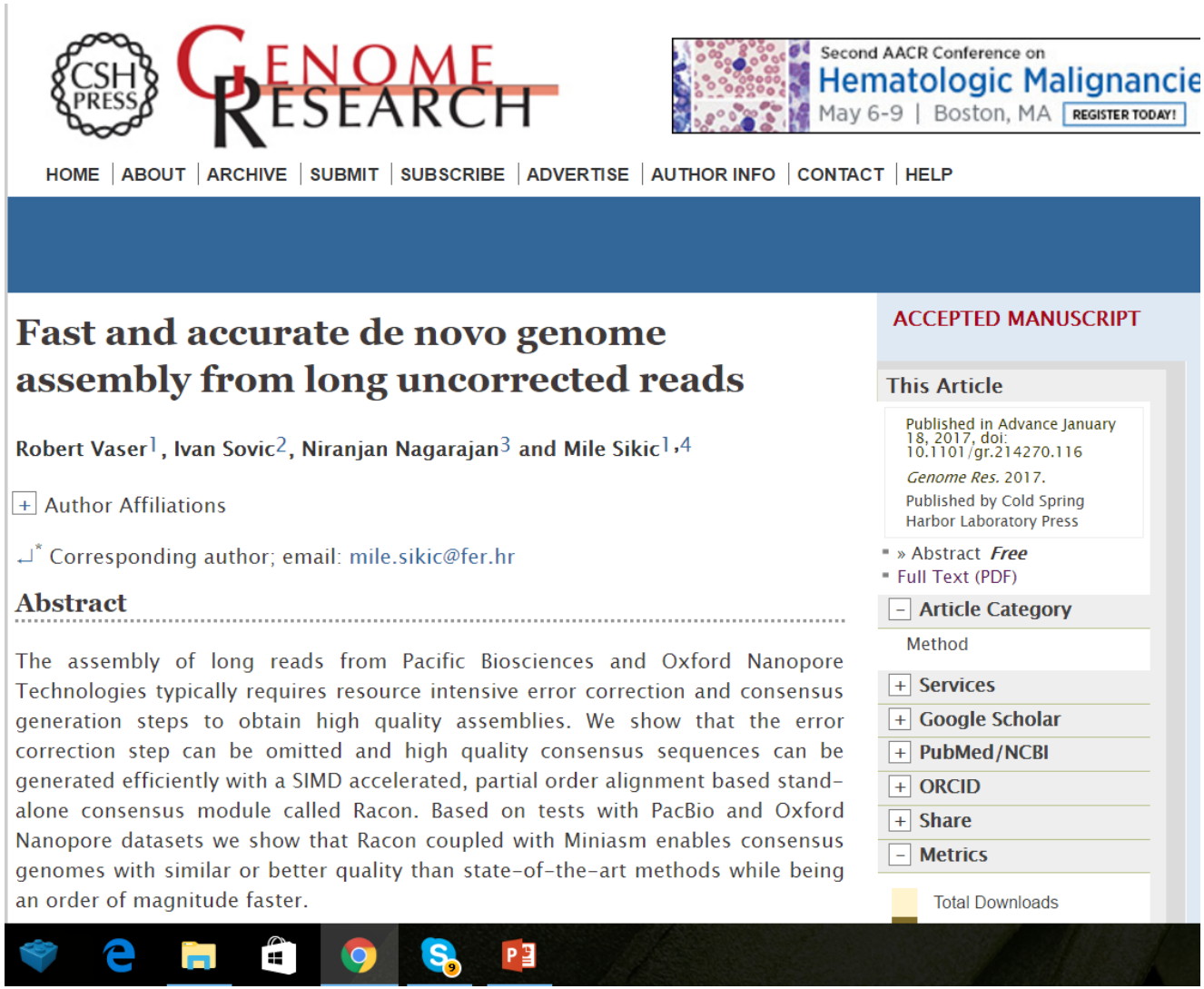
Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads

Ivan Sović; Krešimir Križanović; Karolj Skala; Mile Šikić ✉

Bioinformatics (2016) 32 (17): 2582-2589.
DOI: https://doi.org/10.1093/bioinformatics/btw237
Published: 09 May 2016 **Article history** ▾

❝ Cite | Share ▾ | Tools ▾

Racon



CSH PRESS GENOME RESEARCH

Fast and accurate de novo genome assembly from long uncorrected reads

Robert Vaser[1], Ivan Sovic[2], Niranjan Nagarajan[3] and Mile Sikic[1,4]

[+] Author Affiliations

[*] Corresponding author; email: mile.sikic@fer.hr

Abstract

The assembly of long reads from Pacific Biosciences and Oxford Nanopore Technologies typically requires resource intensive error correction and consensus generation steps to obtain high quality assemblies. We show that the error correction step can be omitted and high quality consensus sequences can be generated efficiently with a SIMD accelerated, partial order alignment based stand-alone consensus module called Racon. Based on tests with PacBio and Oxford Nanopore datasets we show that Racon coupled with Miniasm enables consensus genomes with similar or better quality than state-of-the-art methods while being an order of magnitude faster.

ACCEPTED MANUSCRIPT

This Article

Published in Advance January 18, 2017, doi: 10.1101/gr.214270.116

Genome Res. 2017.

Published by Cold Spring Harbor Laboratory Press

» Abstract *Free*

Full Text (PDF)

[−] Article Category

Method

[+] Services

[+] Google Scholar

[+] PubMed/NCBI

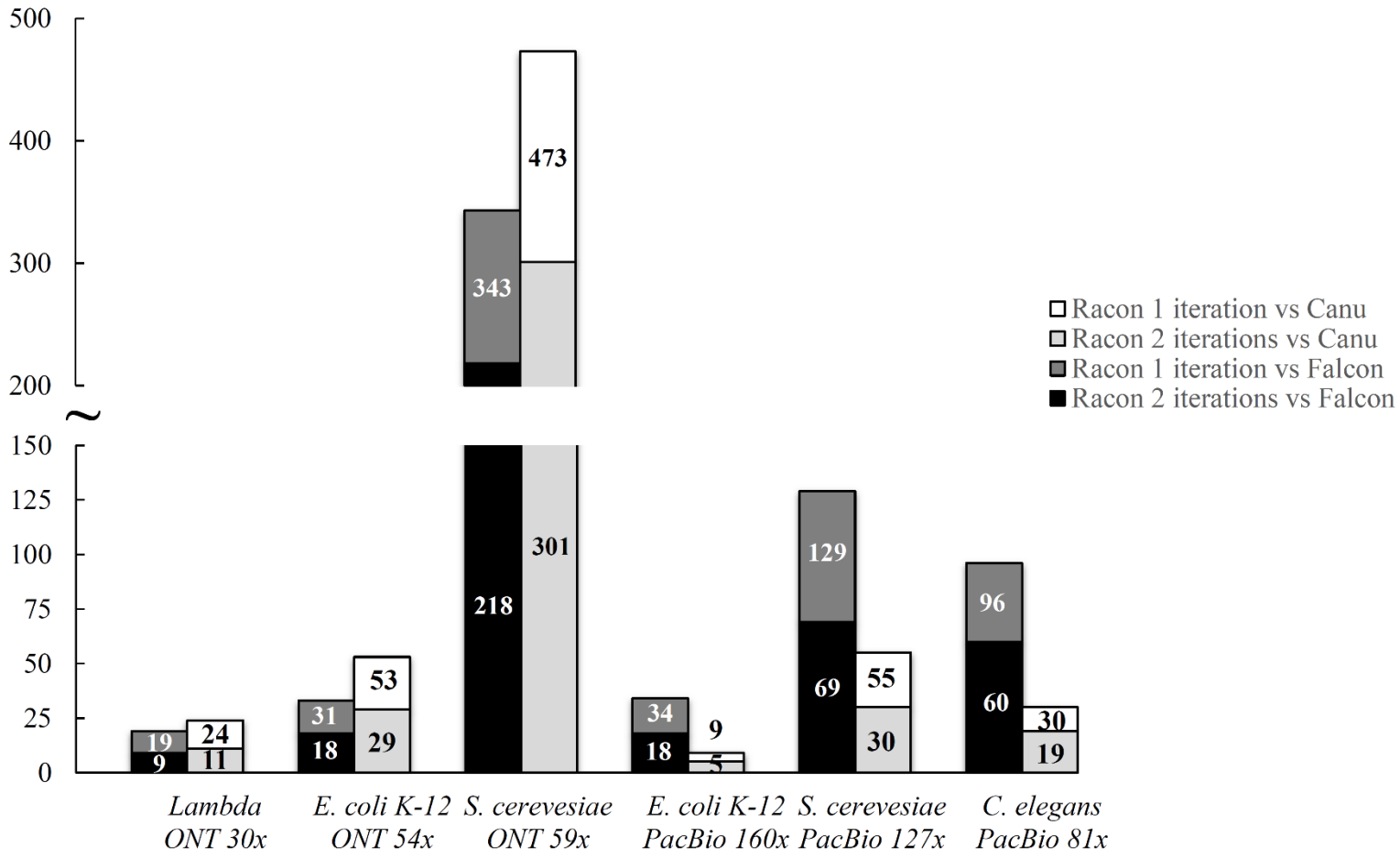
[+] ORCID

[+] Share

[−] Metrics

Total Downloads

Racon


500
400
300
200
150
125
100
75
50
25
0
19
9
24
11
31
18
53
29
343
473
218
301
34
18
9
5
129
69
55
30
96
60
30
19
Racon 1 iteration vs Canu
Racon 2 iterations vs Canu
Racon 1 iteration vs Falcon
Racon 2 iterations vs Falcon
Lambda ONT 30x
E. coli K-12 ONT 54x
S. cerevesiae ONT 59x
E. coli K-12 PacBio 160x
S. cerevesiae PacBio 127x
C. elegans PacBio 81x

Primjena - dijagnoza

“Naši” mikrobi

Genomi biljaka

Gdje smo sada?

- Genomi veličine bakterija – uspješno
- Eukarioti – fragmentirani genomi
- Ploidnost
- Metagenomi (de novo)

Laboratorij za bioinformatiku i računalnu bilogiju

https://www.facebook.com/ferlbcb/

Laboratorij za bioinformatiku i računalnu biologiju

- 1 poslijedoktorand
- 5 doktoranda
- 15 diplomskih studenata u području računarske znanosti
- Alumni
  - Carnegie Mellon, EPFL, ETH
  - Google, Facebook, Microsoft, Amazon,…

Projekti

- Algorithms for genome sequence analysis
- Utvrđivanje patogena iz niza sekvenciranih RNA podataka
- Utvrđivanje patogena iz niza sekvenciranih DNA podataka

Hvala vam!

Mile.Sikic@FER.hr