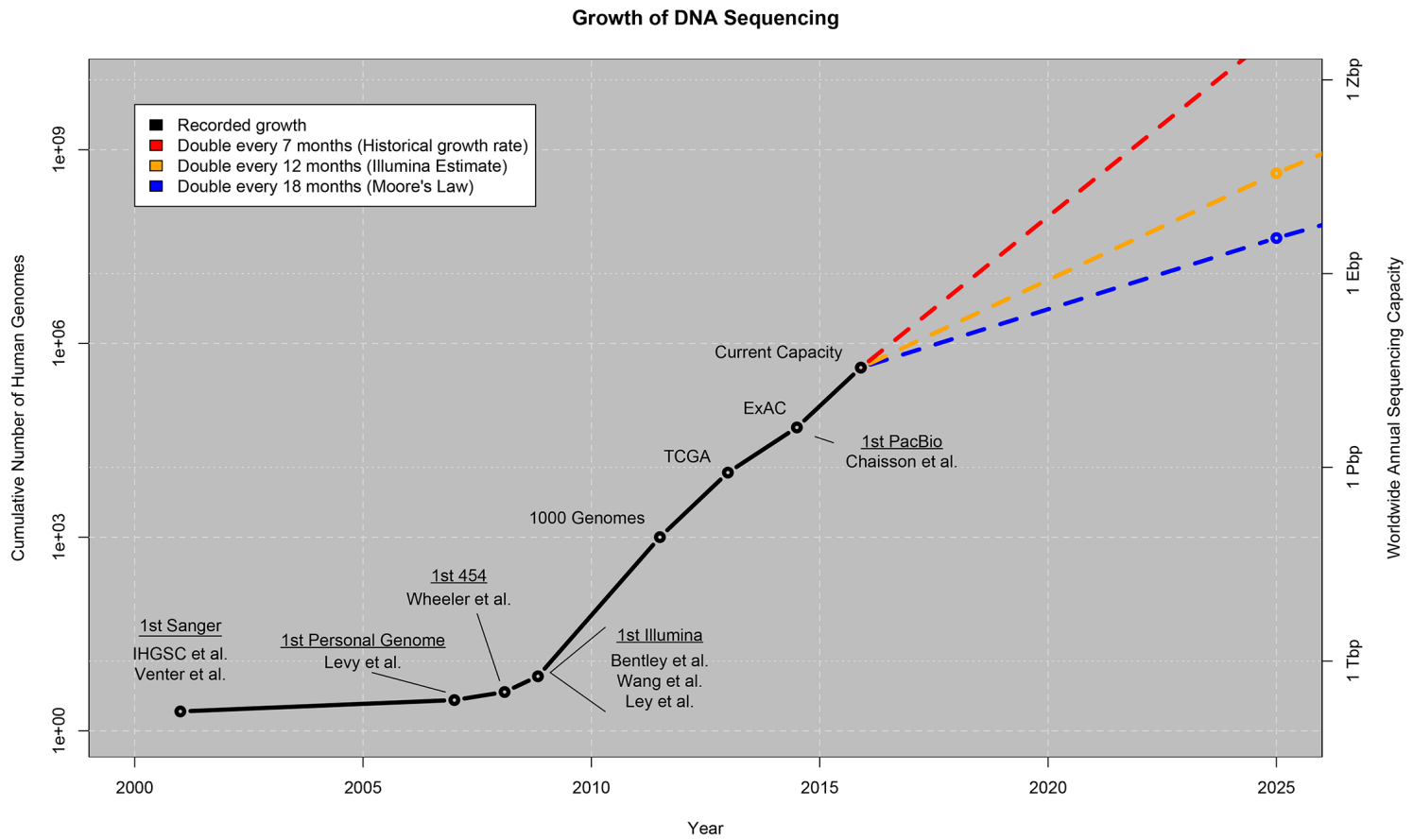
Sekwenatory trzeciej generacji

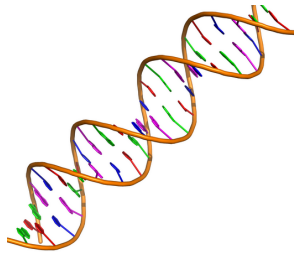
dr inż. Wiktor Kuśmirek
Zakład Sztucznej Inteligencji, Instytut Informatyki

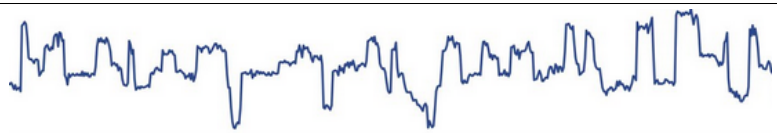
Znaczenie sekwencjonowania



Stephens, Zachary D., et al. "Big data: astronomical or genomical?."PLoS Biology 13.7 (2015): e1002195.

Sekwencjonowanie DNA trzeciej generacji



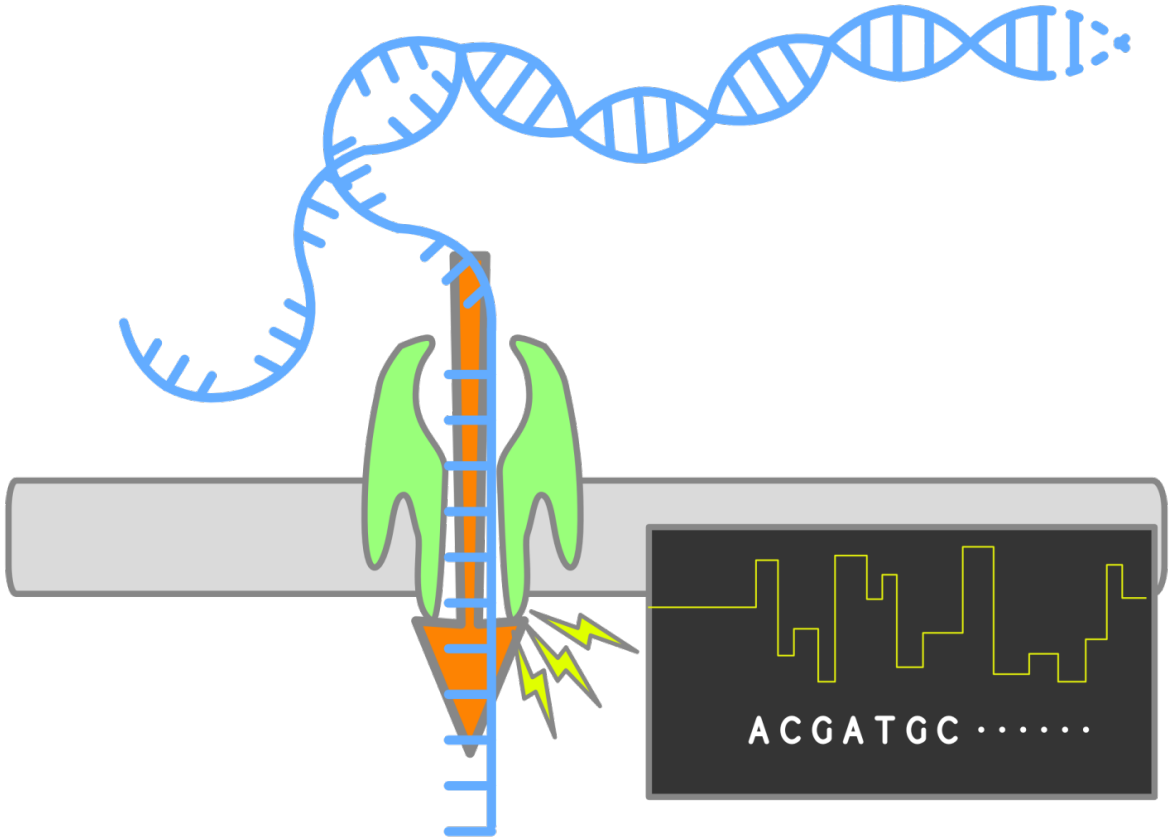


```
@read_id
TCAGTGTACTTCGTTCAGTTGCATATTGCTAAGGTTAAGACTACTTTCTGCCTTTGCGAGAACAGCACC
+
;:373:1,2,/+-;62,:28$(,$"$+'-.66(++.).31-05-/2-003,4,+,193;++-.9%&&%'
```

Sekwencjonowanie nanoporowe (I)

- sekwnecjonowanie nanoporowe polega na przejściu pojedyńczej nici DNA przez bardzo mały otwór nazywany nanoporem
- podczas przejścia nici DNA prąd na nanoporze zmienia swoją wartość w zależnośći od nukleotydów zajmujących nanopor
- zmiany wartości prądu są segmentowane w dyskretne zdarzenia tworząc serie dyskretnych wartośći prądu na nanoporze
- dyskretne wartośći przetwarzane są w procesie *basecallingu* w wynikowe sekwencje DNA stanowiące odczyty DNA

Sekwencjonowanie nanoporowe (II)



https://en.wikipedia.org/wiki/Nanopore_sequencing

Proces basecallingu

- Oparty na metodach sztucznej inteligencji
- Wejście: ciąg wartości oznaczającej prąd na porach sekwenatora
- Wyjście:
  - ciąg zrekonstruowanych nukleotydów
  - ciąg symboli oznaczający jakość odtworzenia pojedyńczego nukleotydu
- Dużo błędów (10% - 30%)

Błędy w odczytach nanoporowych

```
Query  16962    ACGGTTCCTAAATTATTGG-TTTAGATGATGACGAAAACCGTG-CAAAATTGTGATGAGC  17019
               ||||||| ||||||||||| ||||||||||||||||||||||| |||||||||||||||
Sbjct  1922838  ACGGTT-CTAAATTATTGGTTTTAGATGATGACGAAAACCGTGCCAAAATTGTGATGAGC  1922896

Query  17020    TATGACGCAAAAA-CAAAGCTTAGCGCGGAGAGCAATCAGTCTAACCAAGCGATGACC--  17076
               |||||||||||||| ||||||||| | ||  ||||||||||||||||||||||||||  
Sbjct  1922897  TATGACGCAAAAAACAAAGCTT-G-GC--AGAGCAATCAGTCTAACCAAGCGATGACCGT  1922952

Query  17077    -GAATTGACACATTATG--AAAATTGATGATCAGAAAG-TACGCTGAATCAATTGGTCG-  17131
                |||||| |||||||||  |||||||||||||||||||| |||||||||||||||||| 
Sbjct  1922953  TGAATTGGCACATTATGAAAAAATTGATGATCAGAAAATTACGCTGAATCAATTGGTCGG  1923012

Query  17132    CAATTTACGCGATTTATCGTTTTCCAGATGGAACAACGGTAAATGCGGGGATCAATG--C  17189
               |||||| |||  |||||||| |||||||||||||||||||||| | | ||||||   | 
Sbjct  1923013  CAATT-ACG--ATTTATCG-TTTCCAGATGGAACAACGGTAAATGCCGAGGTCAATGCAC  1923068

Query  17190    AAGGCAAAATTGTCTCAGAAGA-T-AAATTGTGTATTTACGGATTATTTATTACTGAAAG  17247
               |||||||||||||||  |||| | ||||||||||||||| | |  ||||||||||||| 
Sbjct  1923069  AAGGCAAAATTGTCTCGAAAGACTCAAATTGTGTATTTACAGGTG--TTATTACTGAAAG  1923126

Query  17248    TGCATTAGCG-ATACTGCAAACTATCAACTGACCGATAATAAGTGTGATGCTTTAAAAAA  17306
               |||||||||| |||||||||||||||||||||||||||||||||||||||  |||||||  
Sbjct  1923127  TGCATTAGCGAATACTGCAAACTATCAACTGACCGATAATAAGTGTGATGCGCTAAAAAA  1923186

Query  17307    TAATTCGA---GTTATGTTGTAGTGAATGAGGCAC-TAGAGCCTATGAGTTTTAGACTGG  17362
               |||||||    |||||||||| ||  |||| | || ||||||||||||||||||||||  
Sbjct  1923187  TAATTCGAAAGGTTATGTTGTGGTAGATGAAG-ACTTAGAGCCTATGAGTTTTAGACTGG  1923245

Query  17363    TATCTGACACGGTTGCCTCAAAAGATGTT-GGGCATTTG--CAATCTTAAAAATAAAA-G  17418
               |||||||||||||||||||||||||||| || |||||||  |||||||||||||||| | 
Sbjct  1923246  TATCTGACACGGTTGCCTCAAAAGATATTTGGGCATTTGCACAATCTTAAAAATAAAAAG  1923305

Query  17419    GCA---GATTCATTTCGTTGTCTTTTTACCCCAACCTATCTTGCGCCGTTGAAACAGCTA  17475
               |||   |||||||||||||||| ||||||||||||||||||||||||||||||||||||  
Sbjct  1923306  GCAACGGATTCATTTCGTTGTC-TTTTACCCCAACCTATCTTGCGCCGTTGAAACAGCTA  1923364

Query  17476    ACATGGTTTGAAGATAATAATCTACAC--GATTGATTATTTTCTGCCCATAA-AGTTATG  17532
               |||||||||||||||||||||||||||  |||||||||||||  |||  || |||||||  
Sbjct  1923365  ACATGGTTTGAAGATAATAATCTACACGTGATTGATTATTTTC--CTTATGTGAGTTATG  1923422

Query  17533    CGCGGCCAAGGCTGCATCATGAAGTTTTACGACTCGAGCAGTCAG-TTCAAACCACAACA  17591
               |||   ||||||||||||||||||||||||||||||||||||  || | ||||||||||||  
Sbjct  1923423  CGCATTAAAGGCTGCATCATGAAGTTTTACGACTCGAGCGATCGGATTCAAACCACAACG  1923482

Query  17592    GATTAAATATTCTTCGCGTGTTTCCGTTTCCCGTTTGAATTAATGCATCAACGGCATGCT  17651
                ||||||||||||||||||||||||||||||||||| | ||||||||||||||||||||  
Sbjct  1923483  AATTAAATATTCTTCGCGTGTTTCCGTTTCCCGTTT-AGTTAATGCATCAACGGCATCAC  1923541

Query  17652    GGACTTCTTTTTCGAATAAACTTTCACGATTTTGTCTAGCTTGTAAGAGGGTGACTATGT  17711
               |||||||||| |||||||||||  ||||||||||||| ||||| |||||  | || ||  
Sbjct  1923542  GGACTTCTTTTCCGAATAAACTTT--CGATTTTGTCTAGC--GTAAG-GGGTG--T--GT  1923592

Query  17712    CTTCAAC--TATCGTGGAGGAAGGGCGGTAATAATATAAAACTAAGAAAATTGGTGCTCA  17769
               |||||||  |||||||| |||||||||||||||||||||  | ||||||||||||||||  
Sbjct  1923593  CTTCAACGCTATCGTGGA-GAAGGGCGGTAATAATATAAACTTCTGAAAATTGGTGCTCA  1923651

Query  17770    ATTAATGGGTTGGCCACACCTTTAATATGGTAATCAAGGCAATGTCCTTGAGGACCATAC  17829
               ||||||| |||| |||||||||||||||||||||||| | |||  | ||||||||||||  
Sbjct  1923652  ATTAATGAGTTGACCACACCTTTAATATGGTAATCAAAGTAAT--CA-TGAGGACCATAC  1923708
```

Proces basecallingu - algorytmy

- ukryte modele Markov'a (HMM) z dekodowaniem przy pomocy algorytmu Viterbi'ego - Metrichor, Nanocall
- rekurencyjne sztuczne sieci neuronowe RNN - Deepnano, BasecRAWller, Nanonet
- dwukierunkowy WaveNets - Wavenano
- konwolucyjne sztuczne sieci neuronowe CNN - Causalcall, Chiron
- algorytm klasyfikacji czasowej CTC - Causalcall, Chiron

Wykorzystanie długich odczytów DNA

Asemblacja *de novo* długich odczytów

- algorytm OLC - Overlap-Layout-Consensus
- proces kosztowny czasowo i pamięciowo
- wiele możliwośći optymalizacji obliczeń, np. fazy Overlap
- wyniki asemblacji *de novo* długich odczytów typowo dłuższe niż sekwencje wynikowe asemblacji *de novo* krótkich odczytów
- sekwencje wynikowe asemblacji *de novo* długich odczytów typowo zawierają więcej błędów niż wyniki asemblacji *de novo* krótkich odczytów DNA

Hybrydowa asemblacja *de novo* (I)

- korekcja długich odczytów DNA przez krótkie odczyty DNA
- asemblacja *de novo* długich, skorygowanych odczytów DNA
- proces korekcji długich odczytów:
  - czasochłonny
  - brak korekcji wszystkich błędów, w szczególności długich ciągów zawierających dużą liczbę błędów
- przykładowe aplikacje: Nanopolish, Canu

Hybrydowa asemblacja *de novo* (II)

- asemblacja *de novo* długich odczytów
- wyniki asemblacji *de novo* długich odczytów korygowane przez zbiór krótkich odczytów DNA
- proces czasochłonny: (I) asemblacje *de novo* długich odczytów, (II) korekcja błędów w sekwencjach wynikowych
- długość sekwencji wynikowych w głównej mierze zależna od liczby i długości długich odczytów a jakość sekwencji wynikowej - od jakośći i liczby krótkich odczytów
- przykładowe aplikacje: Canu, POLCA

Hybrydowa asemblacja *de novo* (III)

- zbudowanie grafu de Bruijn'a z krótkich odczytów
- rozwiązanie niejednoznaczności za pomocą długich odczytów zmapowanych na graf de Bruijn'a
- wygenerowanie wynikowych sekwencji DNA z grafu de Bruijn'a
- przykładowe aplikacje: ABySS, SPAdes, Velvet

Hybrydowa asemblacja *de novo* (IV)

- asemblacja *de novo* krótkich odczytów
- łączenie wyników asemblacji *de novo* krótkich odczytów za pomocą długich odczytów DNA
- metoda o wiele wydajniejsza i optymalniejsza czasowo od wcześniej omówionych
- przykładowe aplikacje: dnaasm, ABySS, LINKS, dnaasm-link

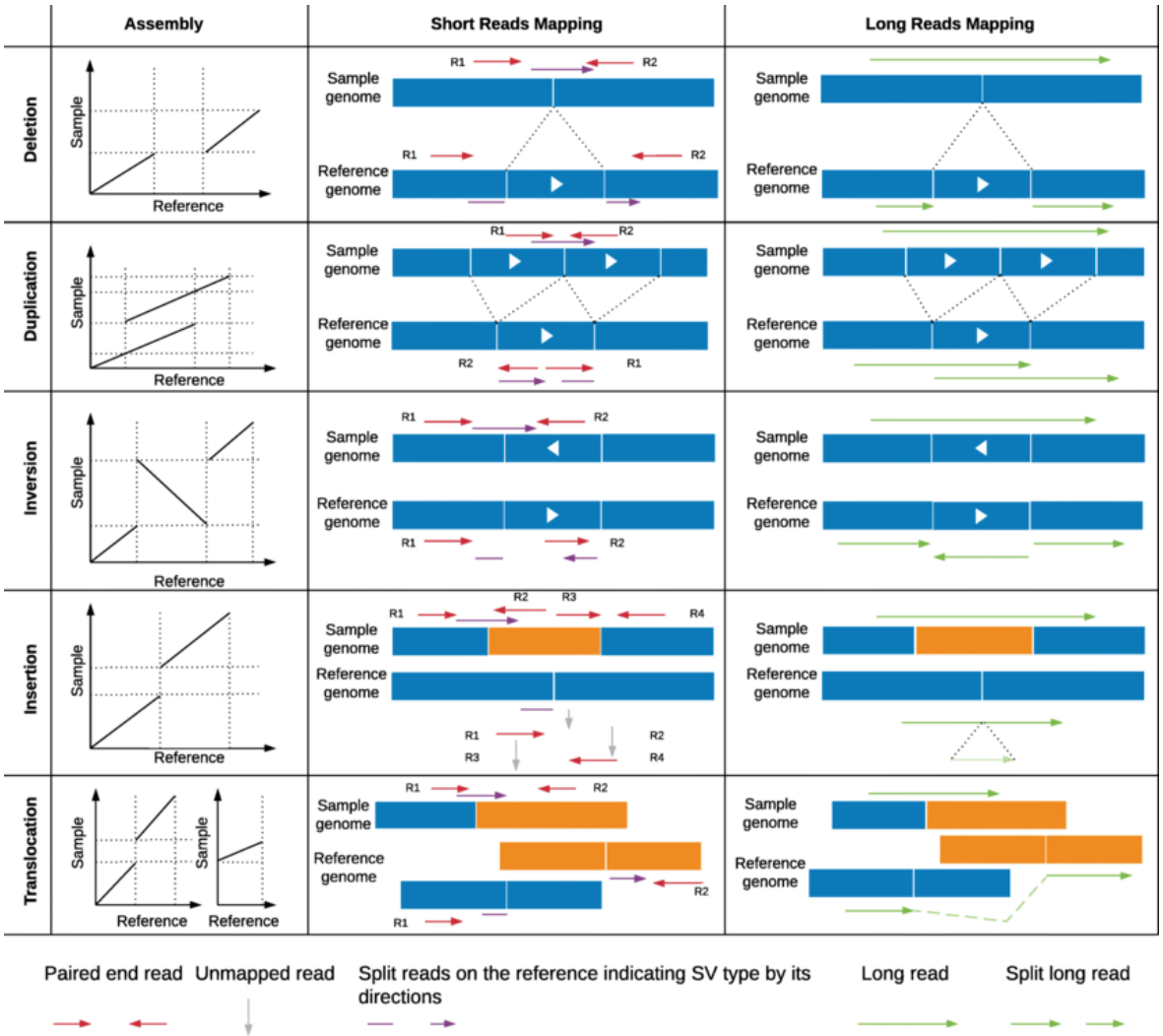
Hybrydowa asemblacja *de novo* - efekt

Assembly set:	PET1	PET1 + PET2	PET1, 2 + MP1	PET1, 2 + MP1, 2	PET1, 2 + MP1, 2 + ONT1	PET1, 2 + MP1, 2 + ONT1, 2
Number of scaffolds	4805	4688	2346	2342	902	719
Total scaffolds size [Mbp]	162.29	162.89	170.80	170.84	176.55	177.07
Longest scaffold [Mbp]	0.439	0.487	3.8	3.8	6.78	6.94
N50 scaffold [kbp]	69.7	84.2	842.2	844.2	1737	2331
Number of contigs	7424	6487	7049	7050	7127	7118
Total contigs size [Mbp]	162.12	162.78	167.66	167.66	167.93	167.95
Longest contig [kbp]	265.4	472.6	472.6	472.6	472.6	472.6
N50 contig [kbp]	46.3	56.1	73.5	73.5	75.0	75.1
Complete (BUSCOs)	630 (64.4%)	628 (64.2%)	646 (66.0%)	647 (66.2%)	649 (66.4%)	646 (66.0%)
Complete and single-copy	621 (63.5%)	620 (63.4%)	637 (65.1%)	636 (65.0%)	639 (65.3%)	638 (65.2%)
Complete and duplicated	9 (0.9%)	8 (0.8%)	9 (0.9%)	11 (1.1%)	10 (1.0%)	8 (0.8%)
Fragmented	107 (10.9%)	105 (10.7%)	92 (9.4%)	93 (9.5%)	90 (9.2%)	90 (9.2%)
Missing	241 (24.6%)	245 (25.0%)	240 (24.5%)	238 (24.3%)	239 (24.4%)	242 (24.7%)

Wykrywanie zmian strukturalnych

- mapowanie długich odczytów DNA na genom referencyjny
- porównanie miejsc mapowań interesujących fragmentów odczytu DNA na genom referencyjny
- im dłuższe odczyty tym lepiej
- im mniej błędów w odczytach DNA tym lepiej

Wykrywanie zmian strukturalnych



Mahmoud, Medhat, et al. Structural variant calling: the long and the short of it. Genome biology 20.1 (2019): 1-14.

Wielkość danych genomowych

Dane genomowe

- genom referencyjny

```
>ref
ACTACGACGCATCAAGCAGACTACACAGCAGCCCCAGCATACAGCAGACGACATCACAGCAGACGACGA
ACATACGACAGCAGCAGACTACATCAGCAGACAGCAGCAGCAGCGACGACGACAGCAGCAGCAGCAGAC
ACATCAGACGACATCATCATCATACGAGAGACGACGACGACGACAGCAGCAGCAGCAGACGACGACGAG
ACTACAGCAGCAGCATCATCATCATCATCATCATCATTTTTCATACTACATCATCATCATCATACTTTT
ACTACAGAGAGCATACTACATCATCATCAGACGCAGACGACGACATTTTACTACATCAGCAGCAGCAGG
ACTACAGCAGCAGACGACGACGACTACTCATCATCATCATACTACTCATACTACTACTACTACTACTCC
```

- odczyty DNA (FASTA)

```
>0
ACTACGAACGACTACGACGACATCATCAGCAGACGACTCATCAGCAGCATTTTCCAGACGCATACTACG
>1
ACTCGGGCGATCATCGACAGCAGCATCATCAGACGAGGAGGGCACATCATACGACAGCAGCGACGACGA
>2
ACTACATCAGAGGAGACTCATCTACTACGGGCAGCATCATCATCAGACGAGAGAGGACCTCATCACTGC
>3
ACTAGGGACATCATCAGACGACAGCATACTACTTTTCATCAGCGGGACCTACATCAGACGGGGCACACT
```

Dane genomowe - człowiek - studium przypadku

- sekwencjonowanie Whole Genome Sequencing
- pokrycie genomu odczytami: 30x
- genom referencyjny: 3 Gbp
- długość odczytu: 100 bp
- liczba odczytów: 3 Gbp * 30 / 100 bp = 900 000 000
- 900 000 000 odczytów po 100 bp
- (minimalna - bez identyfikatorów) wielkość pliku z odczytami (FASTA): 900 000 000 * 100 / 1024 / 1024 / 1024 = 83.8 GB
- (minimalna - bez identyfikatorów) wielkość pliku z odczytami (FASTQ): 83.8 GB * 2 = 167.6 GB

Zadanie

Zaproponuj ile należy przygotować przestrzeni dyskowej na nieskompresowany plik w formacie FASTQ. Paramerty sekwencjonowania:

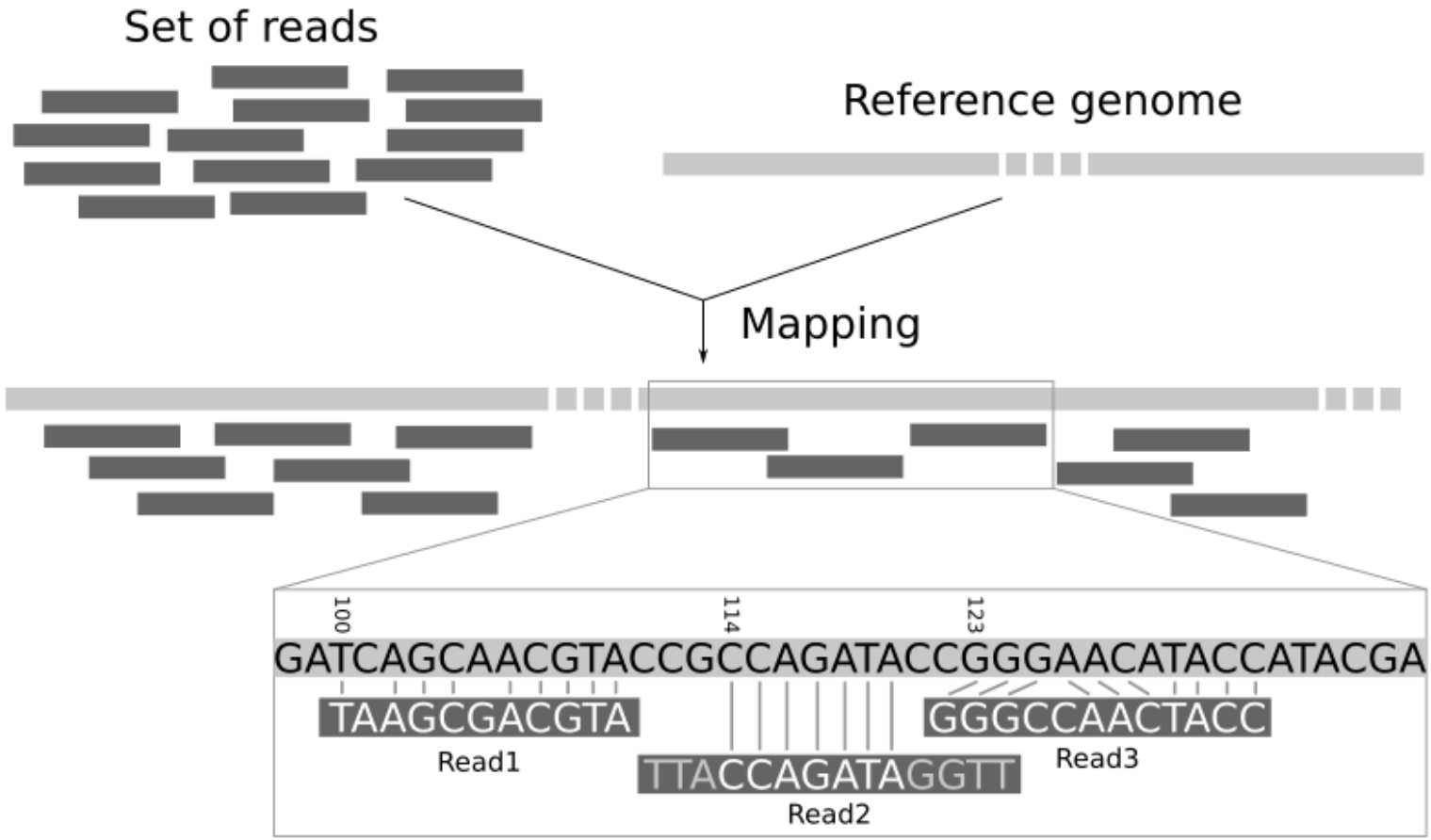
- sekwencjonowanie Whole Genome Sequencing
- pokrycie genomu odczytami: 50x
- genom referencyjny: 500 Mbp
- długość odczytu: 100 bp

W szacowaniu pomiń identyfikatory sekwencji oraz linie oddzielające sekwencje DNA od symboli jakości.

„Naiwna” kompresja sekwencji DNA

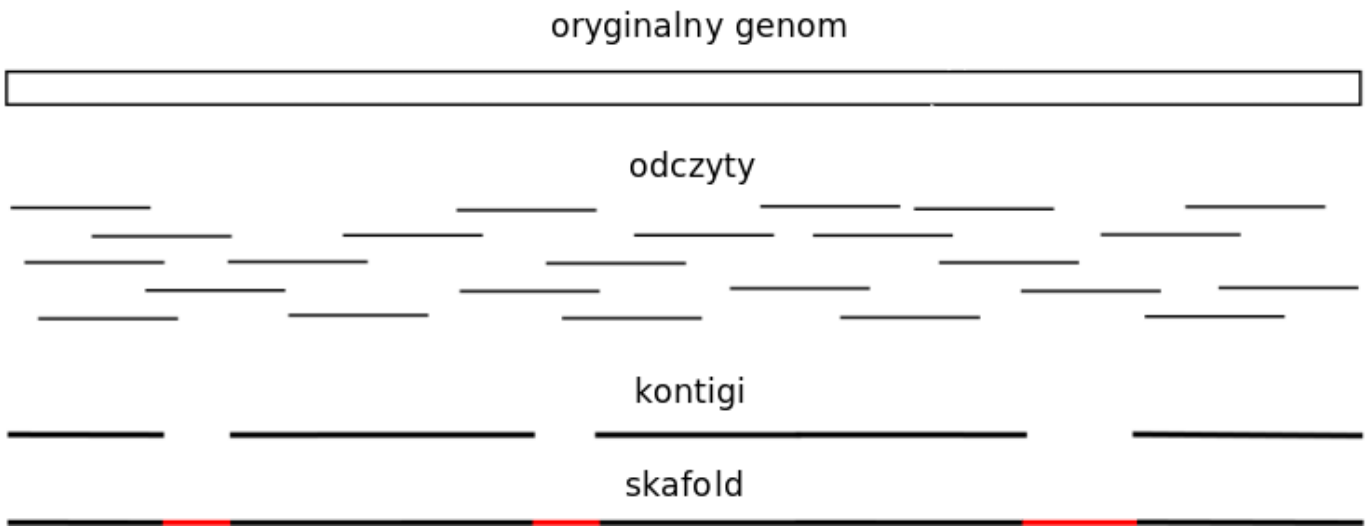
- sekwencja typowo złożona z czterech symboli - A, C, G, T
- do kompresji wystarczą dwa bity na jeden znak, przykład:
  - A - 00
  - C - 01
  - G - 10
  - T - 11
- przewidywany współczynnik kompresji: 4
  - pojedyńczy nukleotyd - 1 bajt, czyli 8 bitów
  - po kompresji - pojedyńczy nukleotyd - 2 bity

Dane genomowe - kompresja z mapowaniem odczytów

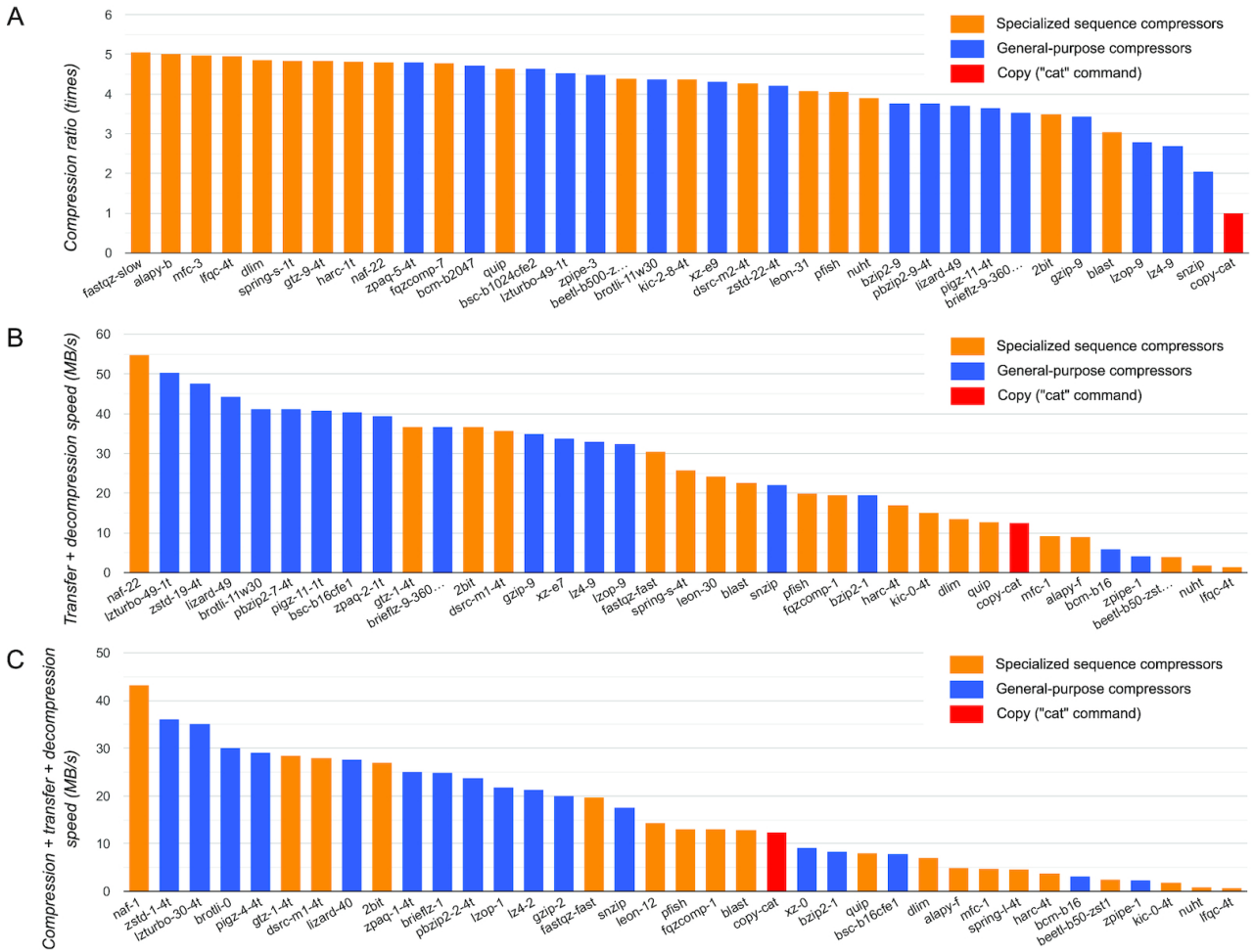


https://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Dane genomowe - kompresja z assemblingiem *de novo* i mapowaniem

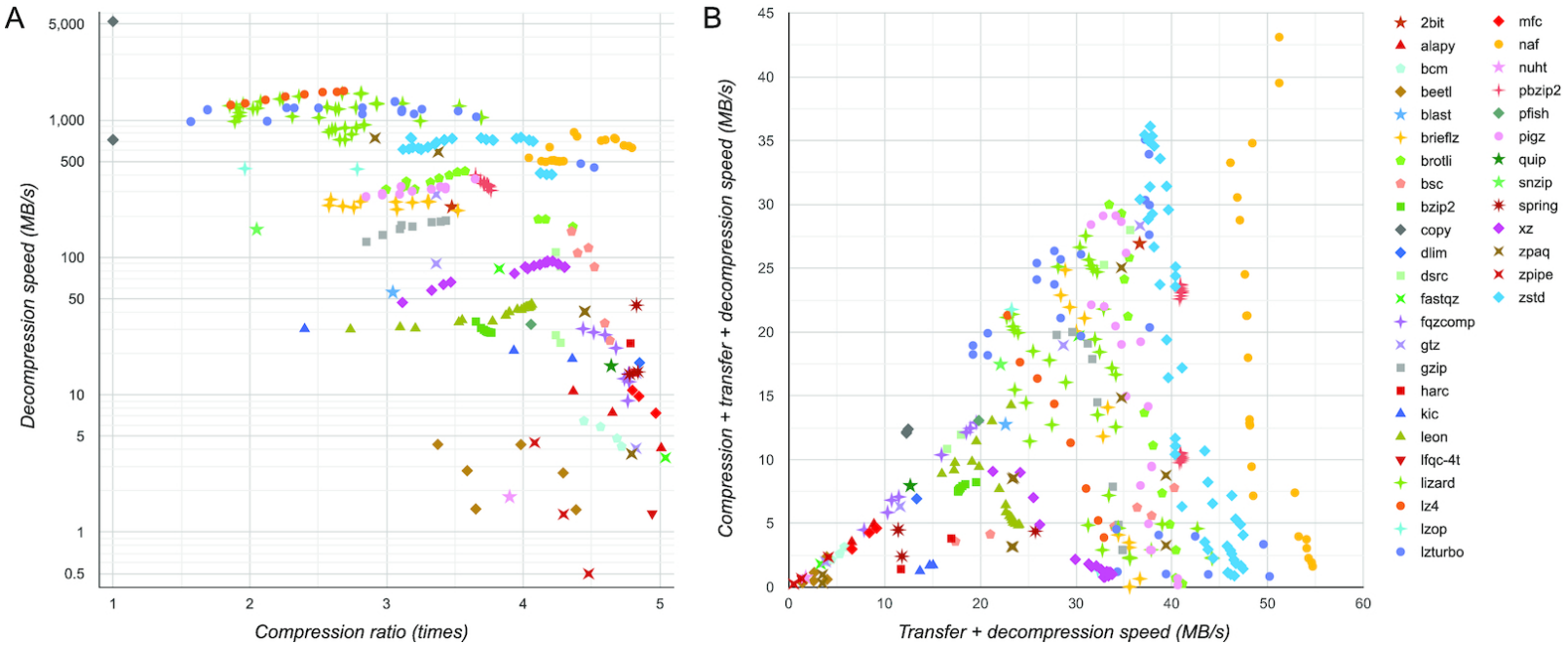


Porównanie algorytmów kompresji (I)



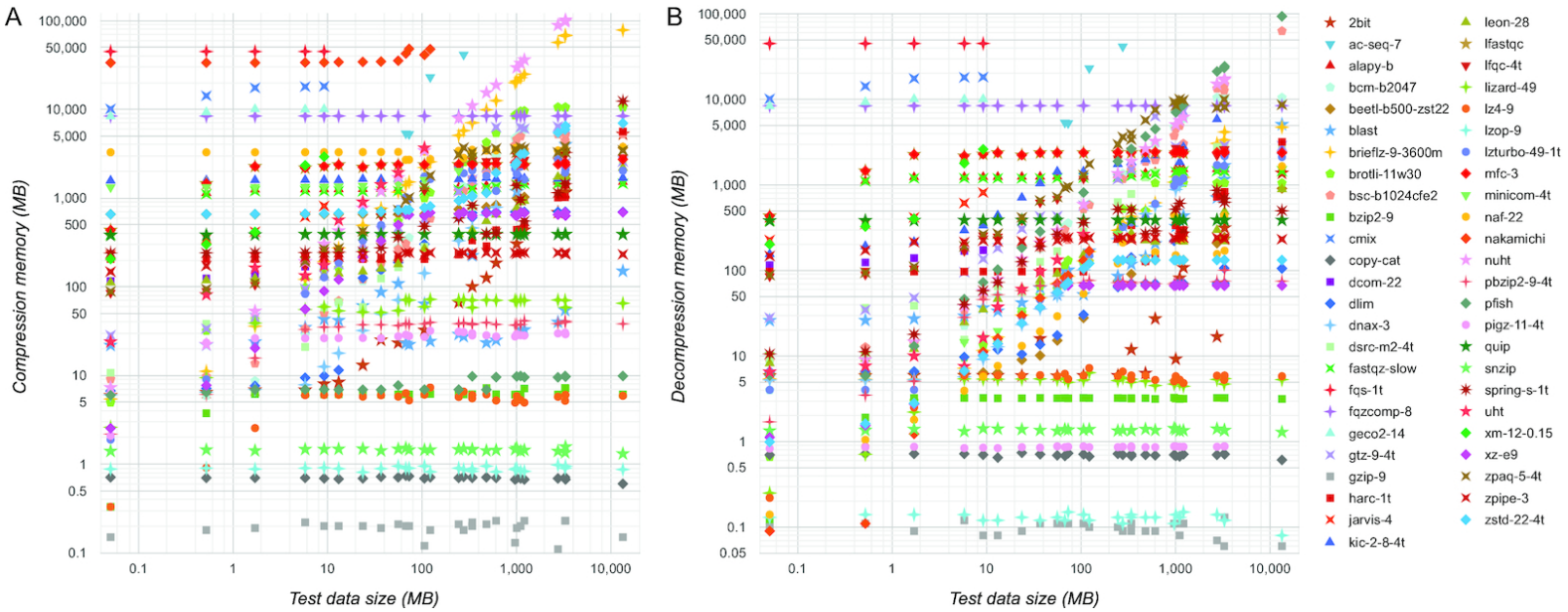
Kryukov, Kirill, et al. Śequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences."GigaScience 9.7 (2020): giaa072.

Porównanie algorytmów kompresji (II)



Kryukov, Kirill, et al. Śequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences."GigaScience 9.7 (2020): giaa072.

Porównanie algorytmów kompresji (III)



Kryukov, Kirill, et al. Śequence Compression Benchmark (SCB) database—A comprehensive evaluation of reference-free compressors for FASTA-formatted sequences."GigaScience 9.7 (2020): giaa072.

Dziękuję