Big Data Genomics Pipelines

Processing and Analysing Big Data

mgr inż. Marek Wiewiórka [1] dr inż. Tomasz Gambin [1]
dr hab. inż. Michał Okoniewski [2]
prof. dr hab. inż. Henryk Rybiński

[1]Instytut Informatyki, Politechnika Warszawska

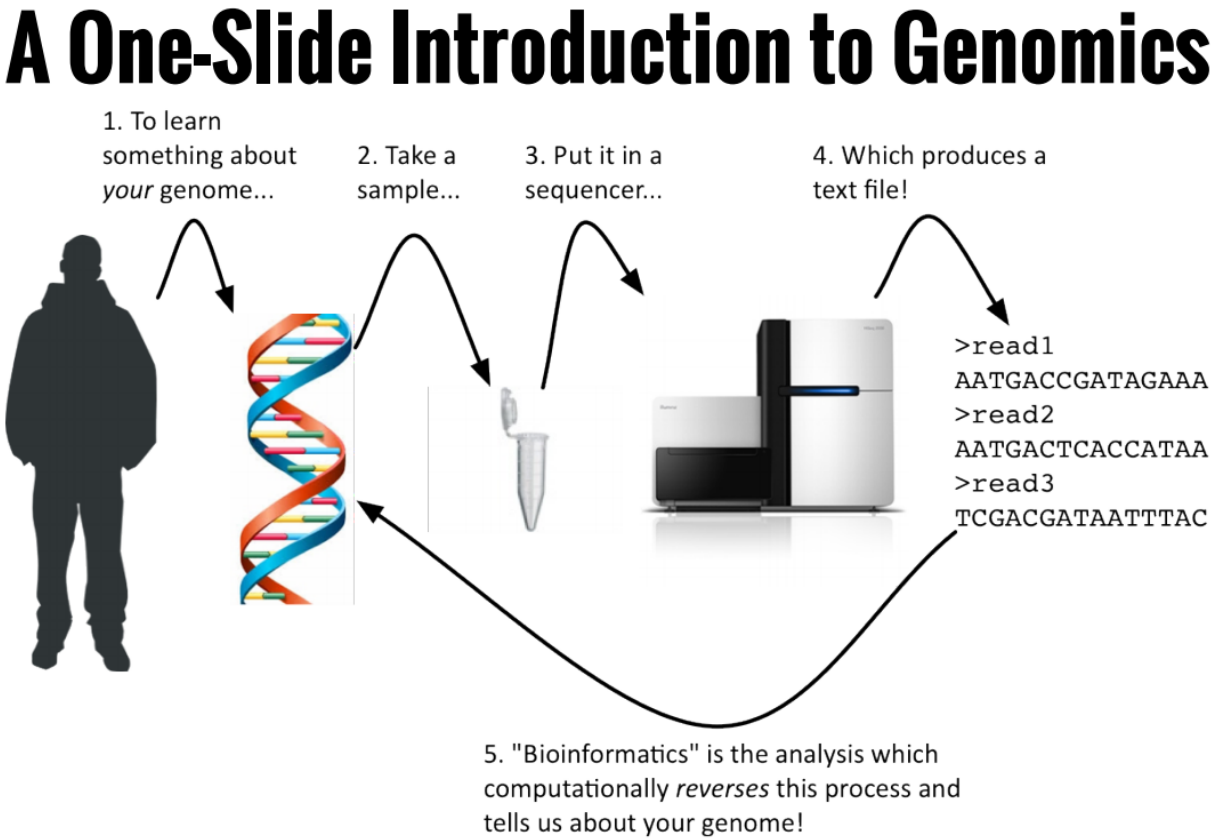
[2]Scientific IT Services, ETH Zurich

Plan prezentacji

Sekwencjonowanie nowej generacji

- różne technologie służące do wysokowydajnego, równoległego odczytywania kolejnych nukleotydów w cząsteczkach DNA;
- metody te umożliwiają wykonywanie sekwencjonowania i analiz:
  - całych genomów – DNA-Seq (*whole-genome sequencing*);
  - całych transkryptów (odwrotna transkrypcja na cDNA) – RNA-Seq;
  - eksomów *exome sequencing*;
  - immunoprecypitacji chromatyny – (*Chip-Seq*);
  - metylacji DNA (*Bisulfite-seq*);
  - wybranych fragmentów (*targeted sequencing panels*).
- pojedynczy sekwencer wykonuje odczyt krótkich fragmentów zazwyczaj w czasie od 1,2 dni do tygodnia i może generować ok. 0.7TB do 1.5TB surowych danych genomicznych, do tego należy doliczyć jeszcze przynajmniej drugie tyle danych kontrolnych.

Przebieg eksperymentu



Rysunek: Źródło:`http://www.slideshare.net/TimothyDanford/tdanford-spark`

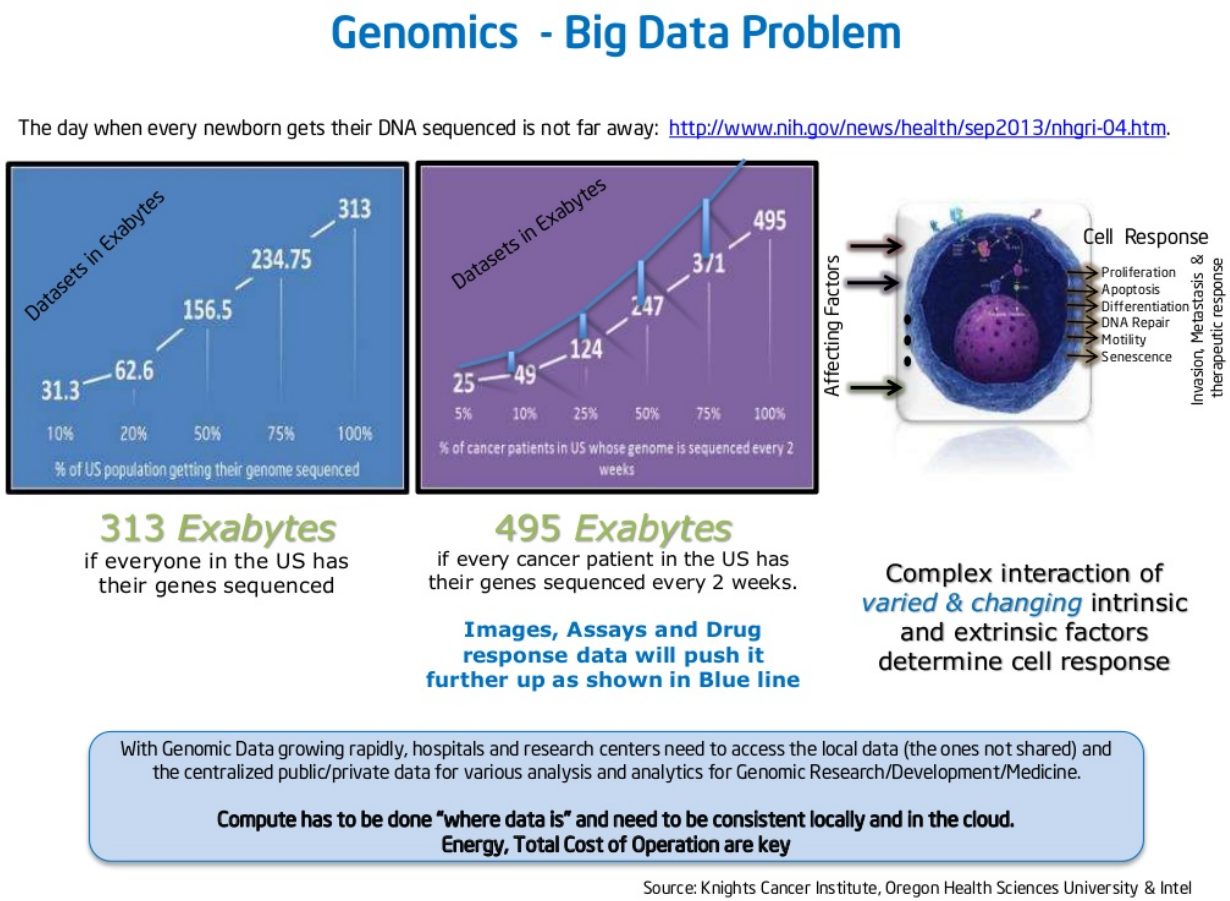
Plan prezentacji

Dlaczego big data 1/3 – typowe wolumeny danych

- typowy plik BAM (Binary Alignment Map) dla 1 próbki z sekwencjonowania cało genomowego to ok. 100-150GB ($3*10^9$ bp, pokrycie 30 z kompresją danych);
- rozmiar pliku zależy zarówno od liczby sekwencjonowanych regionów jak i średniej głębokości(pokrycia);
- przykładowo dla Illumina X Ten można oszacować, iż wolumen danych może sięgać 320*30*2*3GB $\approx$ 60TB/tydzień i ponad 3PB/rok!;
- powyższe dane obejmują tylko surowe dane-niezmapowane (FASTQ), dane zmapowane (BAM) to kolejne 30-45TB/tydzień.

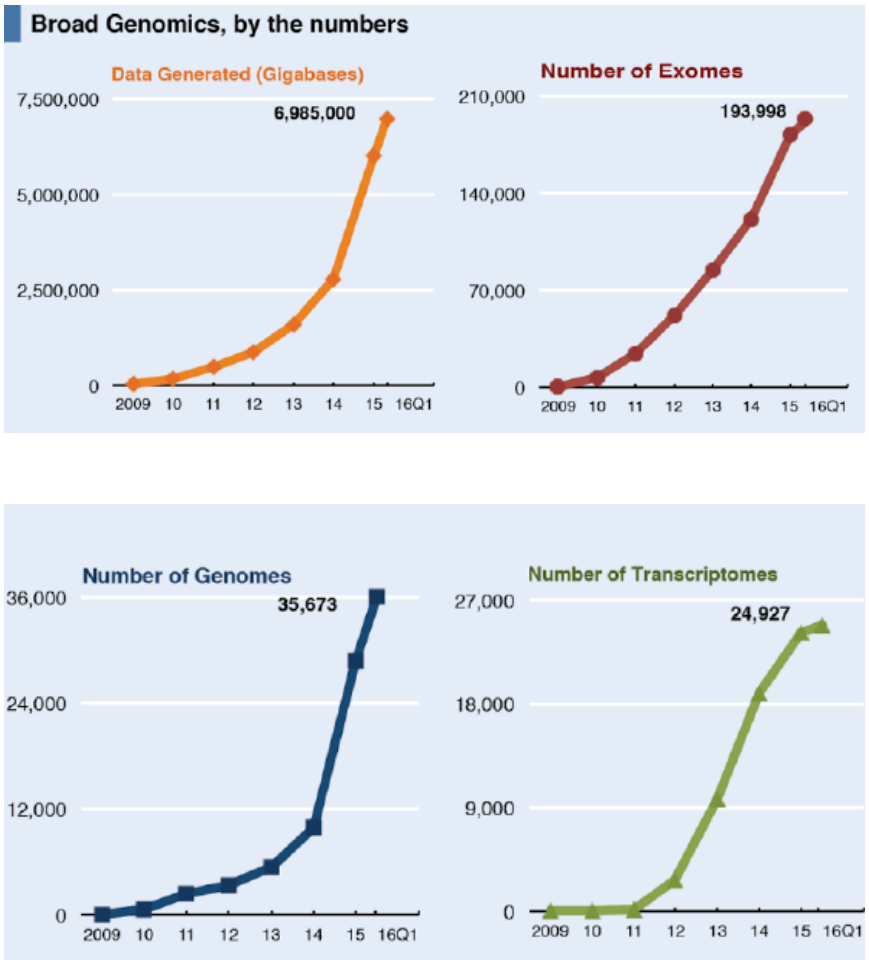
Rysunek: Porównanie przepustowości sekwencerów Illumina (liczba próbek),
Źródło:`http://res.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf`

Dlaczego big data 2/3



Rysunek: Genomika jako problem big data, Źródło: Knights Cancer Institute, Oregon Health Sciences University & Intel

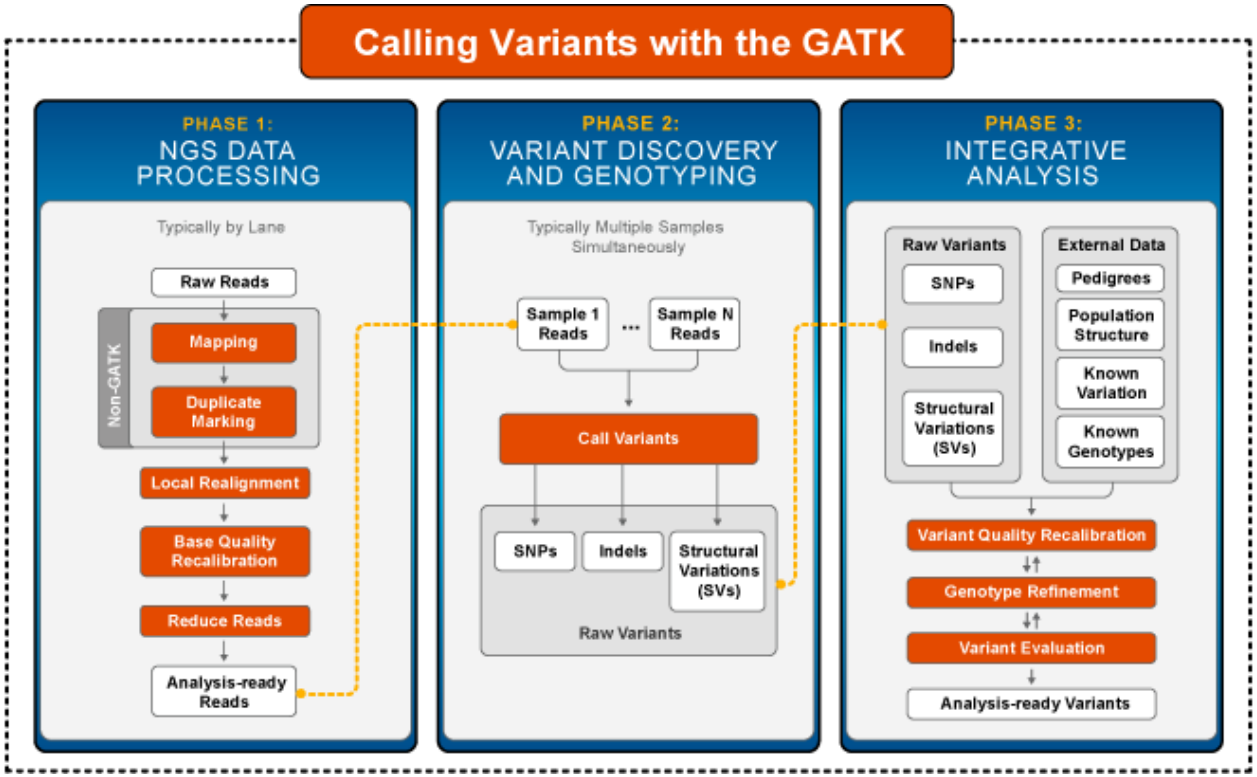
Dlaczego big data 3/3



Rysunek: Liczba próbek sekwencjonowanych z podziałem na typ w ostatnich latach, Źródło: Broad Genomics

Plan prezentacji

Przykładowy przebieg analizy danych z NGS i formaty



Rysunek: Analiza danych z sekwencjonowania DNA , Źródło:`http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0-retired`

Wybrane analizy – DNA/RNA/Bisulfite-seq

- znajdowanie nowych wariantów różnego typu:
  - SNV – *single nucleotide variant*;
  - małe zmiany strukturalne np. insercje, delecje, duplikacje;
  - duże zmiany strukturalne – zmiany liczby kopii.
- pomiar ekspresji różnicowej genów (eksonów);
- odkrywanie nowych transkryptów, isoform genów;
- odkrywanie biomarkerów stopnia metylacji skorelowanych z różnymi zmiennymi objaśniającymi;
- poznawanie genomów wybranych gatunków – sekwencjonowania *de-novo*.

Przykładowe zastosowania kliniczne wyników analiz NGS

- personalizowana terapia medyczna;
- badania przesiewowe (ang. *screening*) – w szczególności *new born screening*;
- diagnostyka chorób rzadkich;
- diagnostyka chorób uwarunkowanych genetycznie, które przebiegają w sposób nietypowy lub bezobjawowy;
- szybka identyfikacja patogenów ożywionych (np. bakterii, wirusów, robaków pasożytniczych) – metagenomika;
- itd. . .

Plan prezentacji

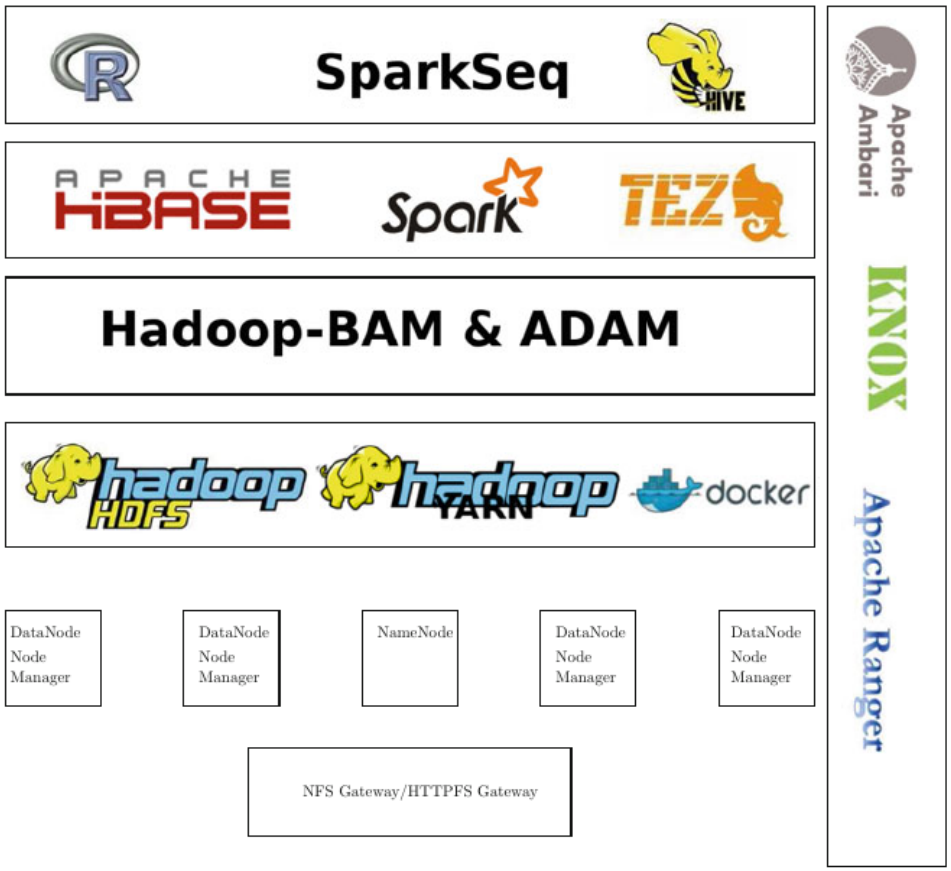
Wyzwania dla zastosowania big data w NGS 1/2

- tekstowe i binarne formaty danych specyficzne dla NGS, często zaprojektowane w sposób uniemożliwiający ich wydajne wykorzystanie w środowisku rozproszonym:
  - obecność scentralizowanych nagłówków;
  - wykorzystanie algorytmów/metod kompresji oraz formaty zapisu, które nie są łatwe do fragmentowania plików;
  - przechowywanie danych w sposób wierszowy pogarszające kompresję danych oraz wydajność zapytań na podzbiorach kolumn;
  - brak schematów danych bądź ich zdefiniowanie w sposób niekompletny/niespójny.
- heterogeniczność wykorzystanych technologii informatycznych:

Wyzwania dla zastosowania big data w NGS 2/2

- wykorzystanie bardzo wielu języków programowania (np. C, Python, Perl, R, shell);
- bardzo wiele istniejących narzędzi, które wykonują prawie te same lub podobne operacje;
- bardzo wiele metod zaimplementowanych w sposób sekwencyjny, ewentualnie umożliwiające ręczne wstępne popartycjonowanie danych w celu osiągnięcia równoległości – brak elastyczności obliczeń i alokacji zasobów;
- brak odpowiednich mechanizmów bezpieczeństwa do pracy środowisku rozproszonym i chmurowym.

Architektura big data dla NGS

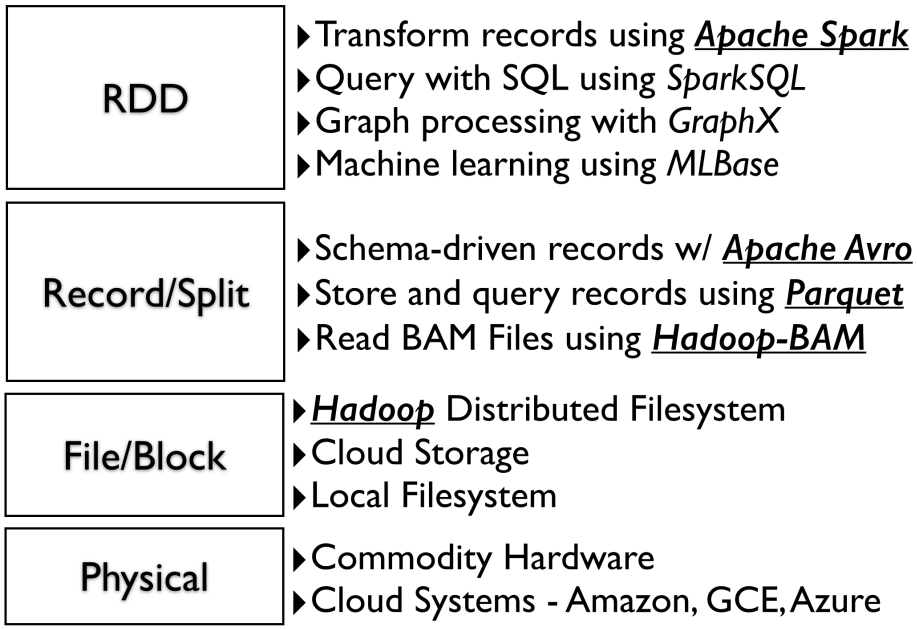


Rysunek: Stos komponentów wykorzystanych analizach big data dla NGS, Źródło: Opracowanie własne ZSI-Bio

Plan prezentacji

Rodzina formatów i zarys architektury – ADAM

- biblioteki do analiz genomicznych opartych na Apache Spark;
- schematy Avro dla danych genomicznych;
- dane przechowywane w formacie Apache Parquet;
- narzędza obsługiwane z linii poleceń (adam-shell, adam-cli).

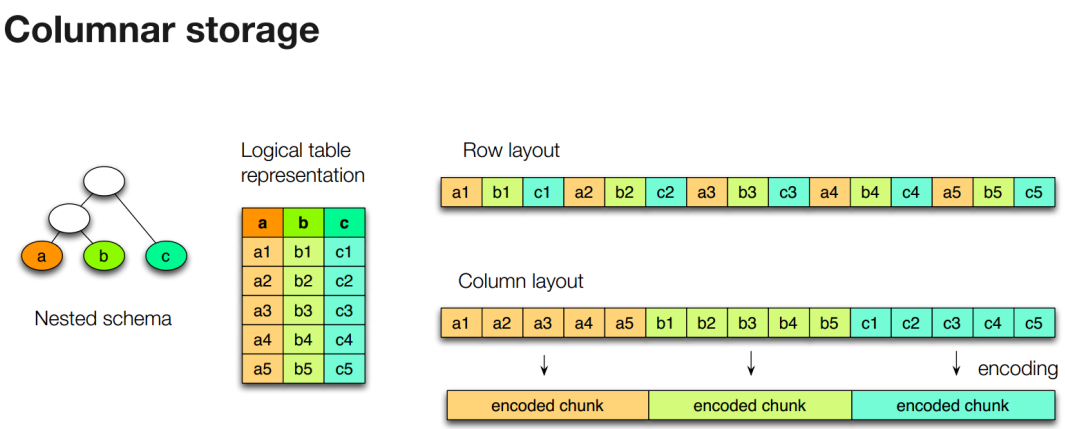


Rysunek: [Źródło:https://github.com/bigdatagenomics/adam/tree/master/docs

ADAM – Avro + Parquet

Avro:

```
{"namespace": "example.avro",
"type": "record",
"name": "User",
"fields": [
    {"name": "name", "type": "string"},
    {"name": "favorite_number",  "type": ["int", "null"]},
    {"name": "favorite_color", "type": ["string", "null"]}
]
}
```



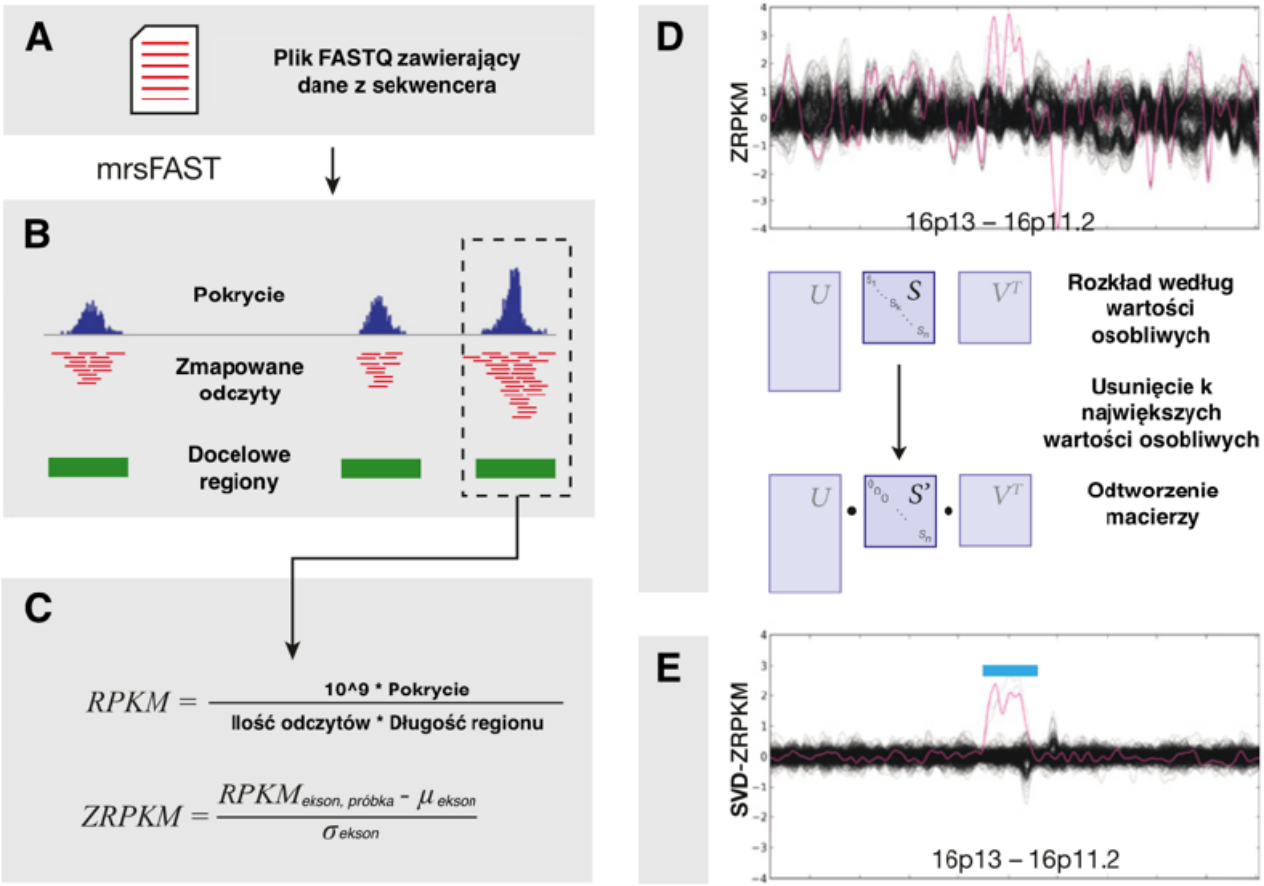
Rysunek: Źródło: http://www.slideshare.net/cloudera/hadoop-summit-36479635

Inne projekty Big data w genomice

- Hadoop-BAM[4] (2012) `http://seqpig.sourceforge.net/`
- SeqPig[7] (2014) `http://seqpig.sourceforge.net/`
- Seal[6] (2011) `https://github.com/ilveroluca/seal`
- SparkSeq[8] (2014) `https://bitbucket.org/mwiewiorka/sparkseq/`
- SparkSW [9] (2015)
- VariantSpark [5] (2015)
- SeqHBase[2] (2015) `http://seqhbase.omicspace.org/`
- Halvade[1] (2015) `https://github.com/ddcap/halvade`
- GenoMetric Query Language[3] (2015) `http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQL/`

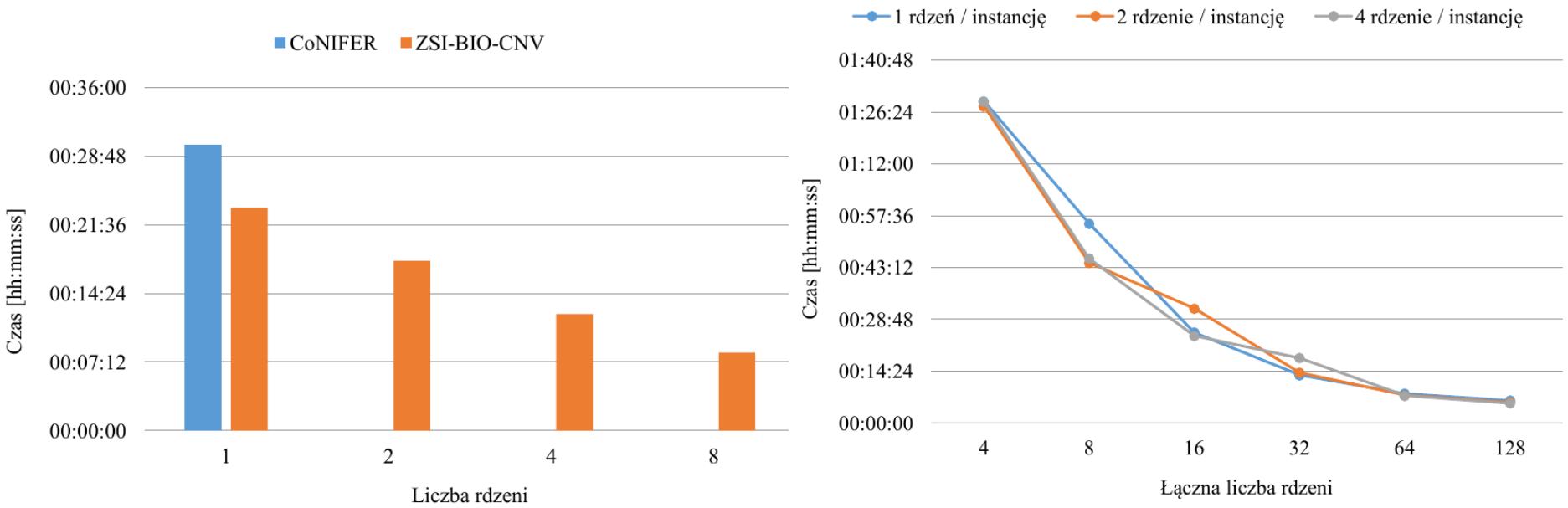
Plan prezentacji

Analizy liczby kopii – CNV (*Copy Number Variation*) 1/2



Rysunek: Zasada działania algorytmu CoNIFER, Źródło: Opracowanie własne ZSI-Bio

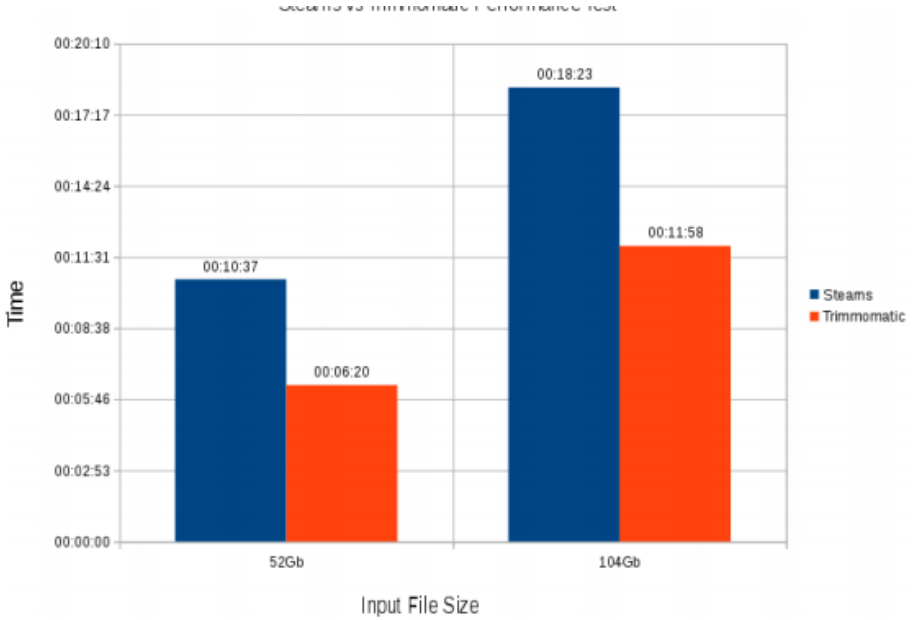
Analizy liczby kopii – CNV (*Copy Number Variation*) 2/2

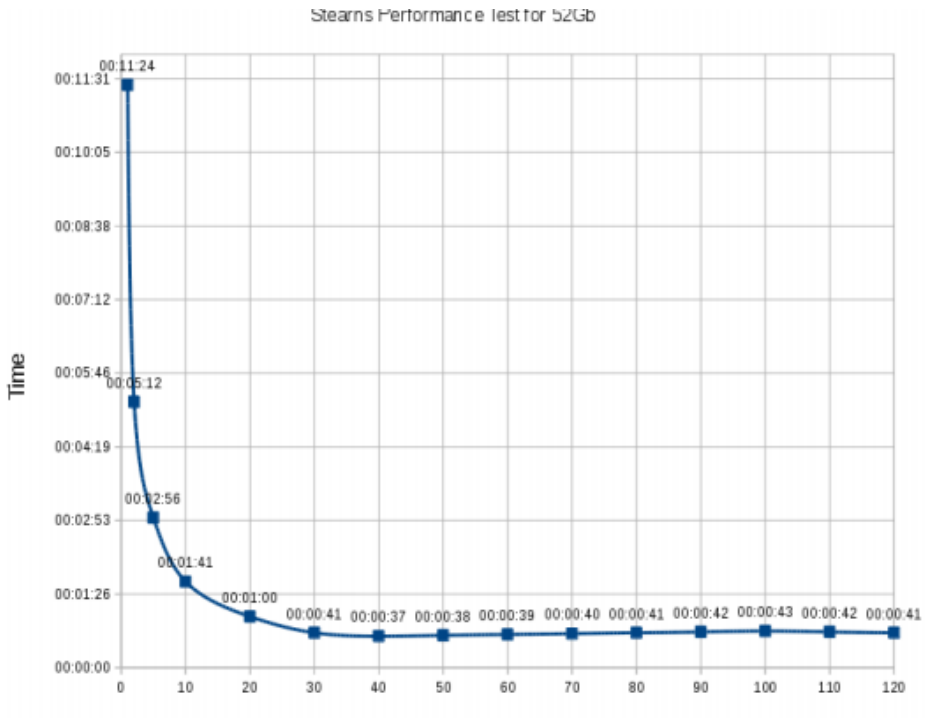


Rysunek: Czas obliczeń detekcji CNV w funkcji liczby rdzeni (a) CoNIFER i ZSI-BIO-CNV, b) Skalowalność ZSI-BIO-CNV), Źródło: Opracowanie własne ZSI-Bio

Trimming i kontrola jakości

- trimming krótkich odczytów wg różnych kryteriów jakości oraz usuwanie adapterów Illuminy





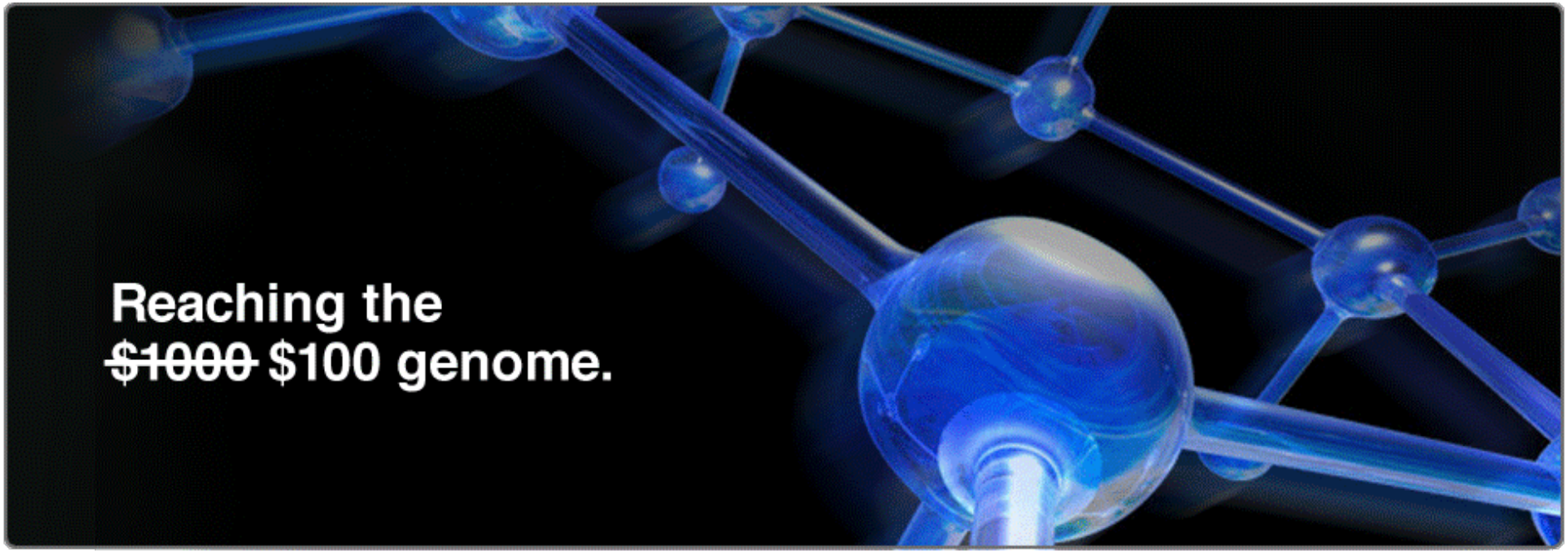
Rysunek: Czas obliczeń metody crop z w funkcji liczby rdzeni (a) Trimmomatic i ZSI-BIO-Stearns, b) Skalowalność ZSI-BIO-Stearns), Źródło: Opracowanie własne ZSI-Bio

Plan prezentacji

Dalsze kierunki badań przy użyciu technologii big data

- większa rozdzielczość analiz – w przypadku analiz RNA-Seq może to oznaczać przejście z poziomu genów/eksonów na poziom pojedynczych nukleotydów;
- większa skala analiz – genomika populacyjna, genomiczne hurtownie danych;
- integracja danych pochodzących z różnych analiz sekwencjonowania, np. (DNA-Seq, RNA-Seq i Bisulfite-Seq) z innymi danymi fenotypowymi;
- zwiększenie szybkości analiz – wymuszone przez wzrost przepustowości sekwencerów – projektowanie skalowalnych procesów przetwarzania i analizy danych;
- automatyzacja przetwarzania danych;
- przejście do strumieniowania danych;
- itd. . .

Przyszłość...



Dziękuję za uwagę.
Marek Wiewiórka

http://zsibio.ii.pw.edu.pl
marek.wiewiorka@gmail.com

Bibliografia I

Dries Decap, Joke Reumers, Charlotte Herzeel, Pascal Costanza, and Jan Fostier.
Halvade: scalable sequence analysis with mapreduce.
Bioinformatics, page btv179, 2015.

Min He, Thomas N Person, Scott J Hebbring, Ethan Heinzen, Zhan Ye, Steven J Schrodi, Elizabeth W McPherson, Simon M Lin, Peggy L Peissig, Murray H Brilliant, et al.
Seqhbase: a big data toolset for family based sequencing data analysis.
Journal of medical genetics, pages jmedgenet–2014, 2015.

Marco Masseroli, Pietro Pinoli, Francesco Venco, Abdulrahman Kaitoua, Vahid Jalili, Fernando Palluzzi, Heiko Muller, and Stefano Ceri.
Genometric query language: A novel approach to large-scale genomic data management.
Bioinformatics, page btv048, 2015.

Matti Niemenmaa, Aleksi Kallio, André Schumacher, Petri Klemelä, Eija Korpelainen, and Keijo Heljanko.
Hadoop-bam: directly manipulating next generation sequencing data in the cloud.
Bioinformatics, 28(6):876–877, 2012.

Bibliografia II


Aidan R O'Brien, Neil FW Saunders, Yi Guo, Fabian A Buske, Rodney J Scott, and Denis C Bauer.
Variantspark: population scale clustering of genotype information.
BMC genomics, 16(1):1, 2015.


Luca Pireddu, Simone Leo, and Gianluigi Zanetti.
Seal: a distributed short read mapping and duplicate removal tool.
Bioinformatics, 27(15):2159–2160, 2011.


André Schumacher, Luca Pireddu, Matti Niemenmaa, Aleksi Kallio, Eija Korpelainen, Gianluigi Zanetti, and Keijo Heljanko.
Seqpig: simple and scalable scripting for large sequencing data sets in hadoop.
Bioinformatics, 30(1):119–120, 2014.


Marek S Wiewiórka, Antonio Messina, Alicja Pacholewska, Sergio Maffioletti, Piotr Gawrysiak, and Michał J Okoniewski.
Sparkseq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision.
Bioinformatics, page btu343, 2014.

Bibliografia III

Guoguang Zhao, Cheng Ling, and Donghong Sun.
Sparksw: scalable distributed computing system for large-scale biological sequence alignment.
In *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on*, pages 845–852. IEEE, 2015.