

Analiza genomu człowieka przy wykorzystaniu NGS w kontekście diagnostyki medycznej

dr inż. Tomasz Gambin ^{1,2}

¹Instytut Informatyki, Politechnika Warszawska

²Zakład Genetyki Medycznej, Instytut Matki i Dziecka w Warszawie

ZSI-Bio research group

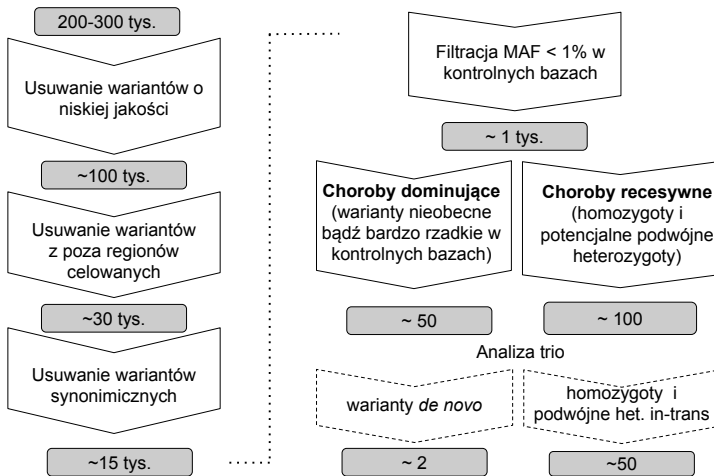


Spis treści



- 1 Filtrowanie i priorytetyzacja wariantów
- 2 Detekcja zmian strukturalnych
- 3 Poszukiwanie nowych genów chorobowych
- 4 Wyzwania w analizie danych z NGS

Filtrowanie wariantów



Identyfikacja wariantów potencjalnie patogennych



GTATGGGGCCAAGAGATATATCT
CGGCTGTCATCACTTAGACCTCAC
TGGGCATAAAGTCAGGGCAGAGC
TGCATCTGACTCCTGAGGAGAGT
TGGTATCAAGGTTACAAGACAGGT
CACTCTCTGCGCTATTGGTCTAT

ClinVar



Znane mutacje w znanych genach



OMIM®

Nowe mutacje w znanych genach

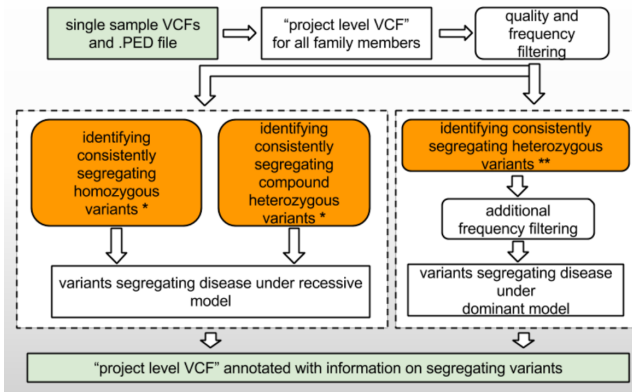


- Analiza danych w rodzinach
- Priorytetyzacja wariantów oraz korelacja genotyp-fenotyp
- Analiza zmian liczby kopii (CNV)

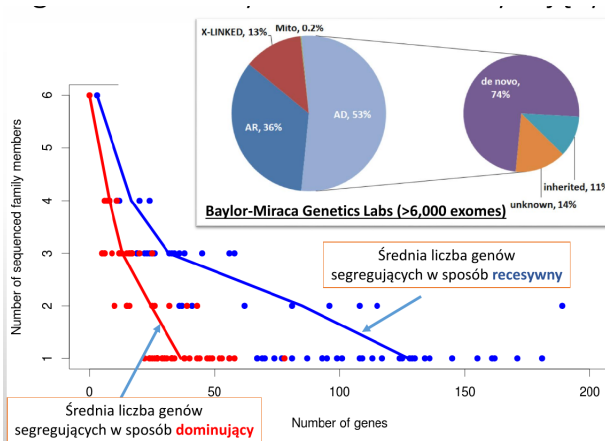
- Testy asocjacyjne
- Wymiana wiedzy z innymi grupami

Mutacje w nowych genach

Analiza wariantów w rodzinach



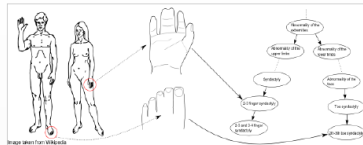
Analiza wariantów w rodzinach



Zbieranie danych fenotypowych



Human Phenotype Ontology (Kochler S, et al., *Nucl. Acid Res.* 2014)



PhenoTips

System bazodanowy do gromadzenia danych fenotypowych bazujący na ontologii HPO

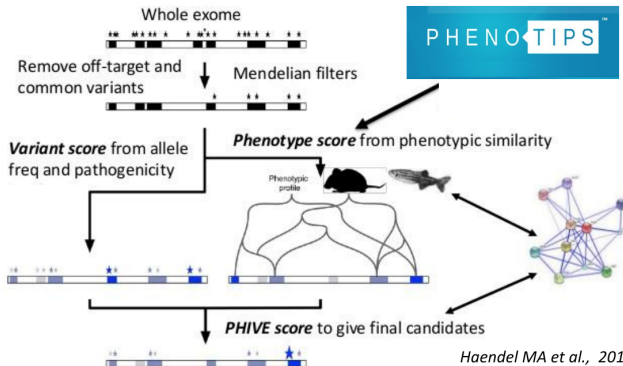
(Girdea M et al., *Hum Mut*, 2013)

Wykorzystanie informacji fenotypowej do priorytetyzacji wariantów



EXOMISER

<http://www.sanger.ac.uk/science/tools/exomiser>



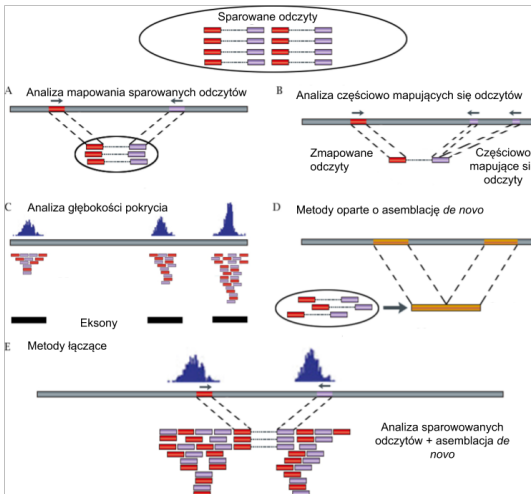
Haendel MA et al., 2015

Spis treści

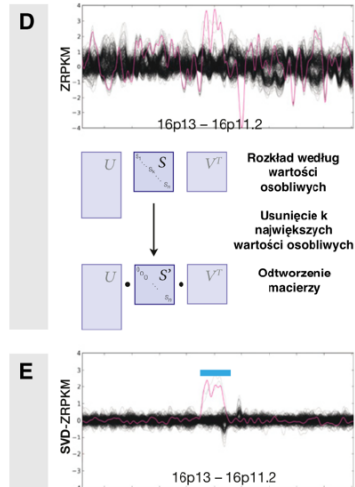
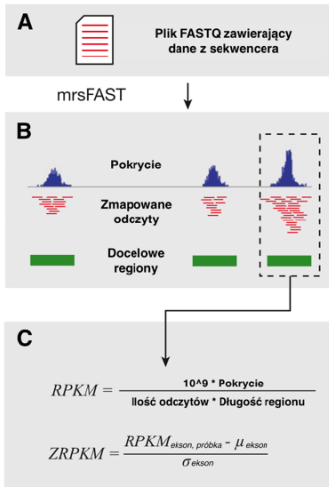


- 1 Filtrowanie i priorytetyzacja wariantów
- 2 Detekcja zmian strukturalnych
- 3 Poszukiwanie nowych genów chorobowych
- 4 Wyzwania w analizie danych z NGS

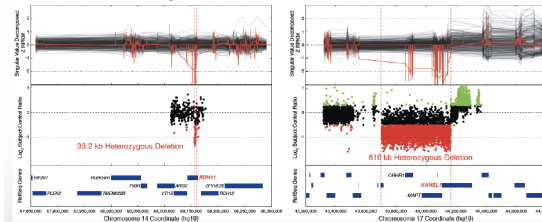
Rodzaje metod do wykrywania zmian liczby kopii, translokacji, inwersji



Przykład algorytmu wykorzystującego analizę głębokości pokrycia



Wykrywanie zmian liczby kopii z sekwencjonowania celowanego c.d.



- Niezbędna jest grupa kontrolna (kilkadziesiąt - kilkaset próbek)
- Typowe metody (np. Conifer, XHMM) składają się z czterech kroków: [1] Wyliczenie głębokości pokrycia (np. RPKM); [2] Normalizacja (np. z-RPKM); [3] Wygładzenie/redukcja szumu (np. SVD, PCA);
- Zazwyczaj wykrywają zmiany większe od 3 eksonów
- Wymagają weryfikacji ortogonalną metodą biologiczną (aCGH, FISH, PCR)

Wykrywanie delecji homozygotycznych

i) Calculation of RPKM values – **196,907 calls/genome**

ii) Removal of exons with median RPKM values less than 7 – **183,304 calls/genome**

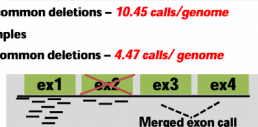
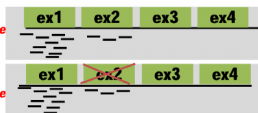
iii) Identification of exons with 0 or low number of reads (RPKM ≤ 0.65) – **2,521 calls/genome**

iv-a) Removal of low quality/common deletions – **10.45 calls/genome**

v) Removal of low quality samples

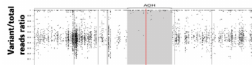
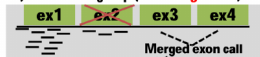
iv-b) Removal of low quality/common deletions – **4.47 calls/genome**

vi) Merging consecutive exon calls and removal of calls < 50bp – **3.36 calls/genome**

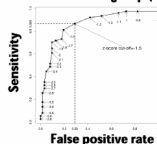


BAM FILES

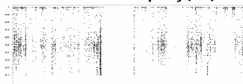
vii) AOH filtering step (**0.6 calls/genome**)



viii) Z-score based filtering step (**0.16 calls/genome**)



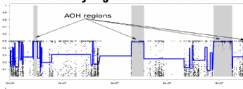
Detection of AOH regions
Calculation of B-allele frequency (BAF)



Transformation = $|BAF - 0.5|$

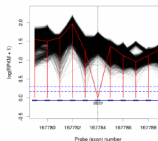


Circular binary segmentation and AOH calling



VCF FILES

BA06555 : chr7:33407378-33407475 (1 exons) size=3.4



Spis treści



- 1 Filtrowanie i priorytetyzacja wariantów
- 2 Detekcja zmian strukturalnych
- 3 Poszukiwanie nowych genów chorobowych
- 4 Wyzwania w analizie danych z NGS

Centers for Mendelian Genomics



Ponad 300 nowych genów
chorobowych
Chong et al. *AJHG* , 2015



James Lupski



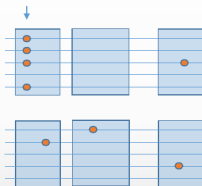
Eric Boerwinkle



Richard Gibbs

Testy asocjacyjne rzadkich wariantów

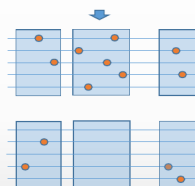
Testy dla pojedynczych wariantów



przypadki
testowe

kontrola

Testy typu Burden, SKAT



Filtracja wstępna

rzadkie warianty
(na podstawie częstości i alg. predykc.)

warianty LoF
(stopgain, frameshift, splicing)

choroby recesywne:
(homozygoty i
podwójne heterozygoty)

- Opracowanie wstępnej listy genów kandydujących pozwala zredukować efekt poprawki stosowanej przy wielokrotnym testowaniu hipotez

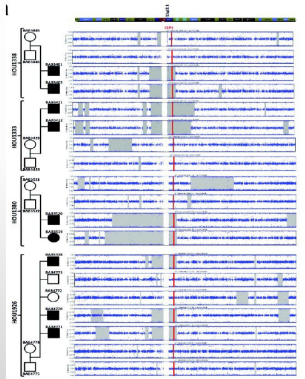
Odkrycia nowych genów chorobowych



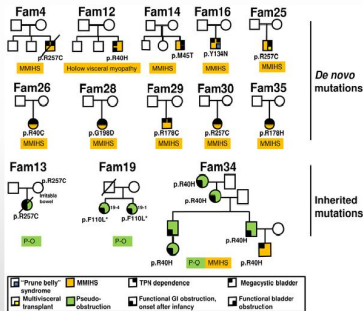
Wady rozwojowe mózgu *CLP1*

MMHS(Berdon syndrome) *ACTG2*

Consanguineous families from Turkey



Karaca et al., *Cell*, 2014



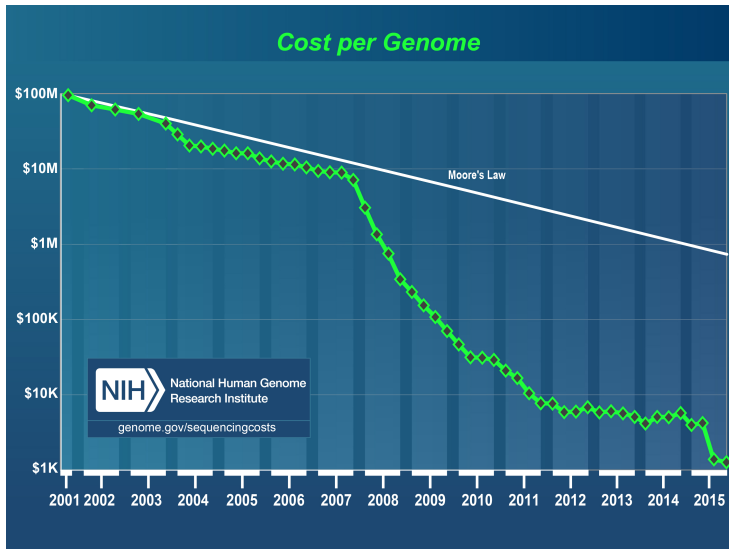
Wangler et al., *PloS Genetics*, 2014

Spis treści



- 1 Filtrowanie i priorytetyzacja wariantów
- 2 Detekcja zmian strukturalnych
- 3 Poszukiwanie nowych genów chorobowych
- 4 Wyzwania w analizie danych z NGS

Koszt sekwencjonowania

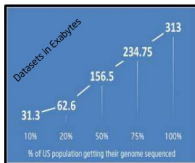


Genomika jako problem Big Data

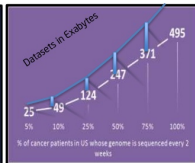


Genomics - Big Data Problem

The day when every newborn gets their DNA sequenced is not far away: <http://www.nih.gov/news/health/sep2013/nhgri-04.htm>.

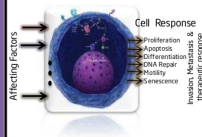


313 Exabytes
if everyone in the US has
their genes sequenced



495 Exabytes
if every cancer patient in the US has
their genes sequenced every 2 weeks.

**Images, Assays and Drug
response data will push it
further up as shown in Blue line**



**Complex interaction of
varied & changing intrinsic
and extrinsic factors
determine cell response**

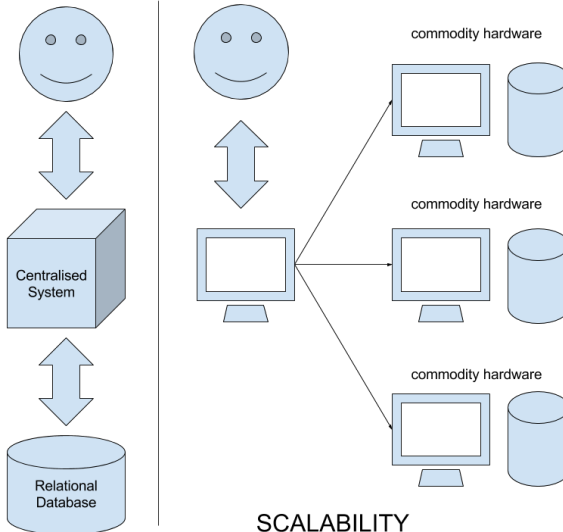
With Genomic Data growing rapidly, hospitals and research centers need to access the local data (the ones not shared) and the centralized public/private data for various analysis and analytics for Genomic Research/Development/Medicine.

**Compute has to be done "where data is" and need to be consistent locally and in the cloud.
Energy, Total Cost of Operation are key**

Source: Knights Cancer Institute, Oregon Health Sciences University & Intel

Figure: Źródło: Knights Cancer Institute, Oregon Health Sciences University & Intel

Podjęcie tradycyjne vs. Big Data



The Hadoop Ecosystem



- Hadoop - środowisko do przetwarzania rozproszonego wielkich zbiorów danych, zapewniające:
 - Skalowalność – aplikacja uruchomiona dla 10GB danych wykona się tak samo dla 10PB
 - Automatyczne zrównoleglenie i odporność na błędy
- Do ekosystemu hadoop należy:
 - Rozproszony system plików – HDFS (Hadoop Distributed File System)
 - Rozproszone silniki obliczeń – MapReduce, Spark, Flink
 - Interfejsy SQL – Hive, Impala
 - Rozproszone bazy danych: HBase, Cassandra
 - Wiele innych ...

Wybrane projekty Big data w genomice



- Big Data Genomics (ADAM) <http://bdgenomics.org/>
- Hadoop-BAM <http://seqpig.sourceforge.net/>
- SeqPig <http://seqpig.sourceforge.net/>
- Seal <https://github.com/ilveroluca/seal>
- SparkSeq <https://bitbucket.org/mwiewiorka/sparkseq/>
- SparkSW
- VariantSpark <https://github.com/BauerLab/VariantSpark>
- SeqHBase <http://seqhbase.omicspace.org/>
- Halvade <https://github.com/ddcap/halvade>
- GenoMetric Query Language http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQL/

Problemy w wykorzystaniu narzędzi Big Data



- Formaty plików nie są przystosowane do przetwarzania rozproszonego
 - Centralne nagłówki
 - Kompresja
 - Podział oparty o wiersze
 - Niespójne definicje formatów
- Różnorodność wykorzystywanych narzędzi
 - C, Python, Perl, R, shell
 - Sekwencyjne schematy przetwarzania – ciężkie do podziału
 - Brak mechanizmów bezpieczeństwa – konieczne przy przetwarzaniu w chmurach obliczeniowych
- Aby móc w pełni korzystać z zasobów chmurowych narzędzia muszą być przeprojektowane i zaimplementowane od nowa.

Korzyści wynikające z wykorzystania narzędzi Big Data

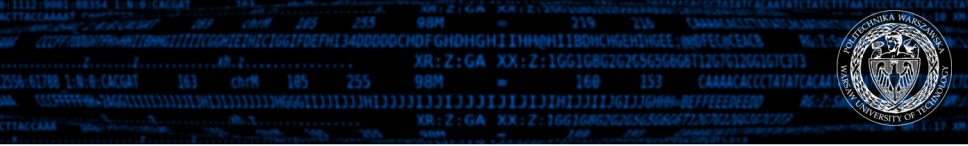


- **Szybsza analiza**
 - Szybkie eksomy/genomy (np. w przypadkach zagrożenia życia)
 - Zwiększone możliwości testowania, kalibracji - poprawienie skuteczności algorytmów
- **Implementacja w chmurach obliczeniowych**
 - Zniesienie barier utrudniających dzielenie danych
 - Łatwość skalowania rozwiązań
- **Nowe możliwości badawcze**
 - Integracja danych z wielkich projektów genomowych
 - Integracja danych z wielu platform NGS

Future...



Reaching the
~~\$1000~~ \$100 genome.



Dziękuję za uwagę
Tomasz Gambin

<http://zsibio.ii.pw.edu.pl>
tgambin@gmail.com