Analiza genomu człowieka przy wykorzystaniu NGS w kontekście diagnostyki medycznej

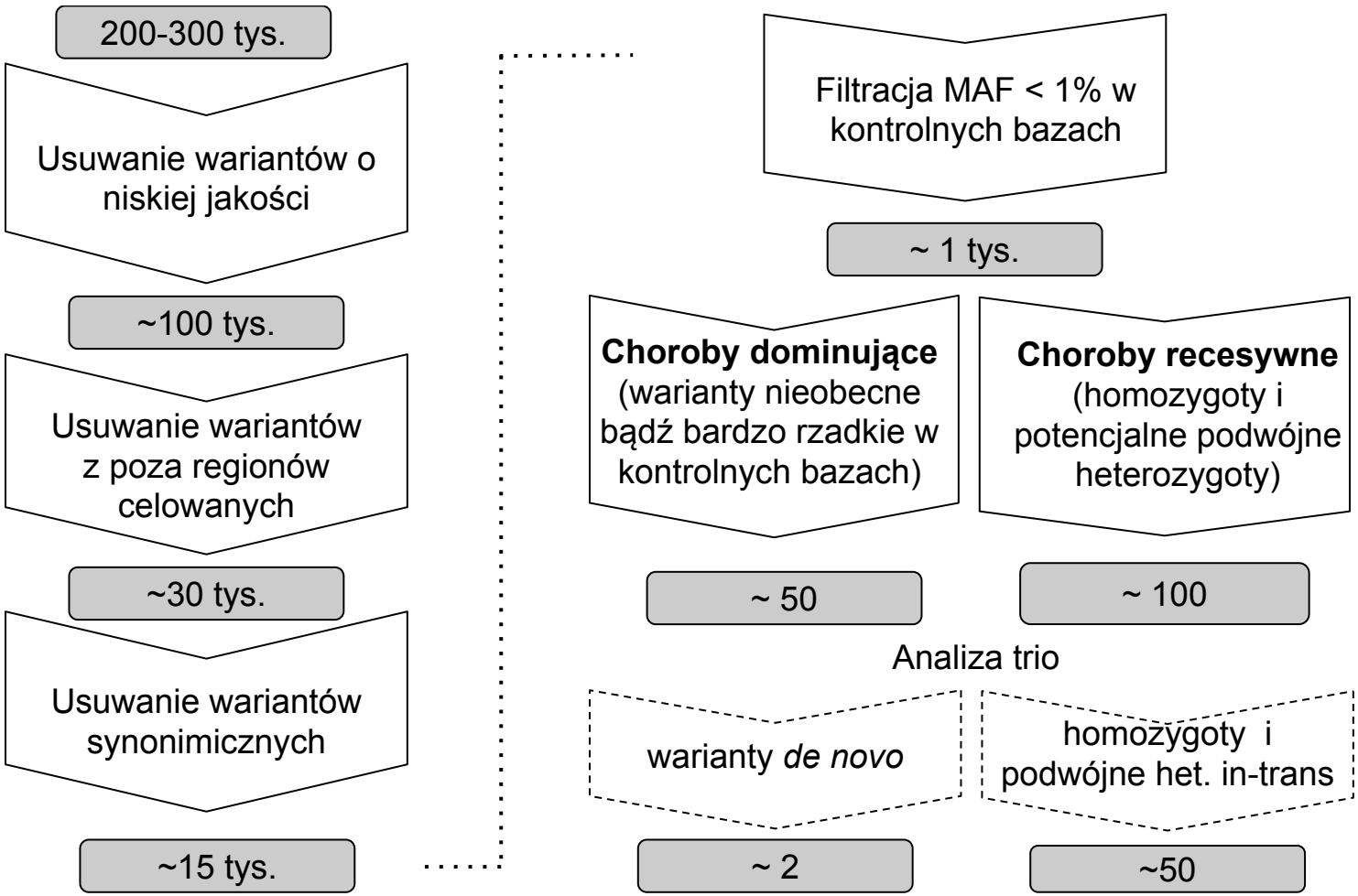
dr inż. Tomasz Gambin [1,2]

[1]Instytut Informatyki, Politechnika Warszawska

[2]Zakład Genetyki Medycznej, Instytut Matki i Dziecka w Warszawie

Spis treści

Filtrowanie wariantów



Identyfikacja wariantów potencjalnie patogennych

HGMD®

Znane mutacje w znanych genach

OMIM®

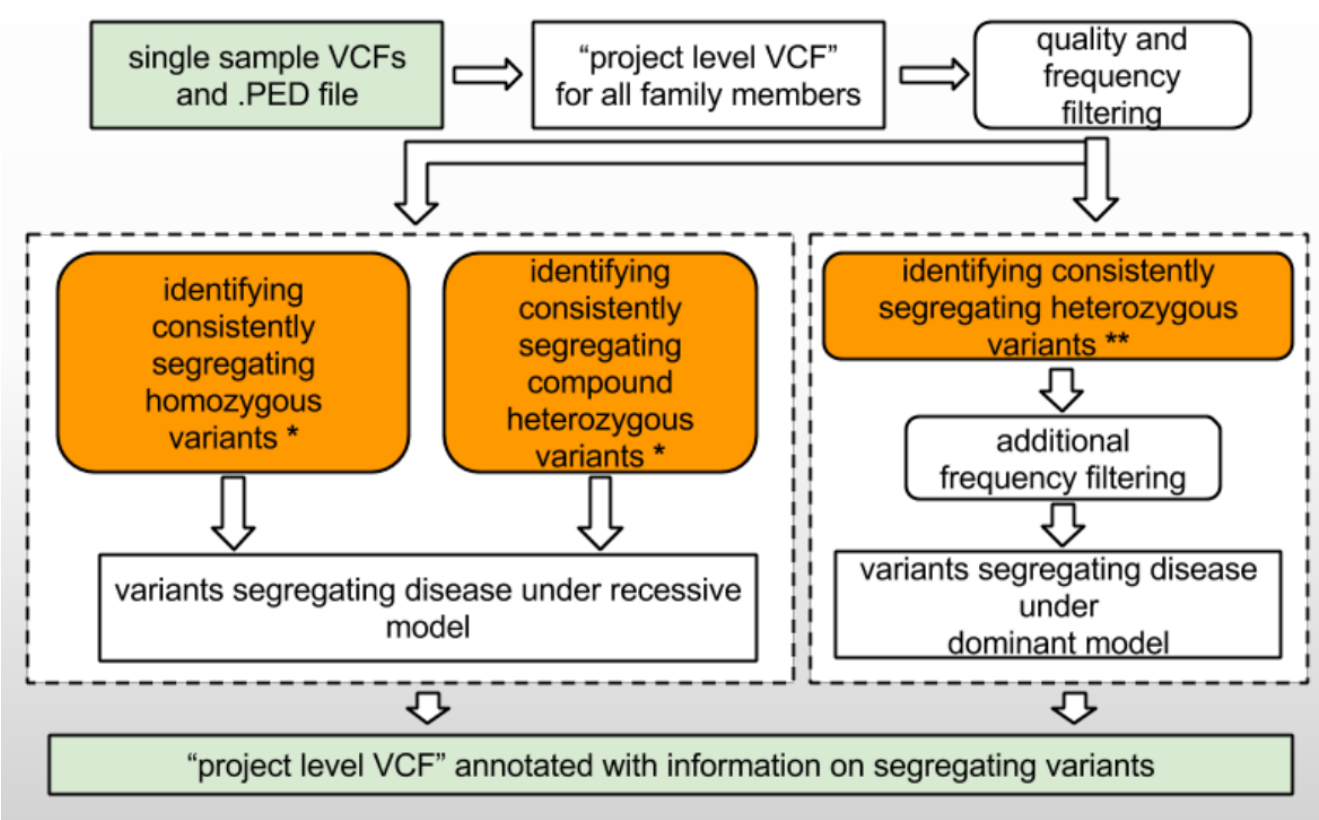
- Analiza danych w rodzinach
- Priorytetyzacja wariantów oraz korelacja genotyp-fenotyp
- Analiza zmian liczby kopii (CNV)

Nowe mutacje w znanych genach

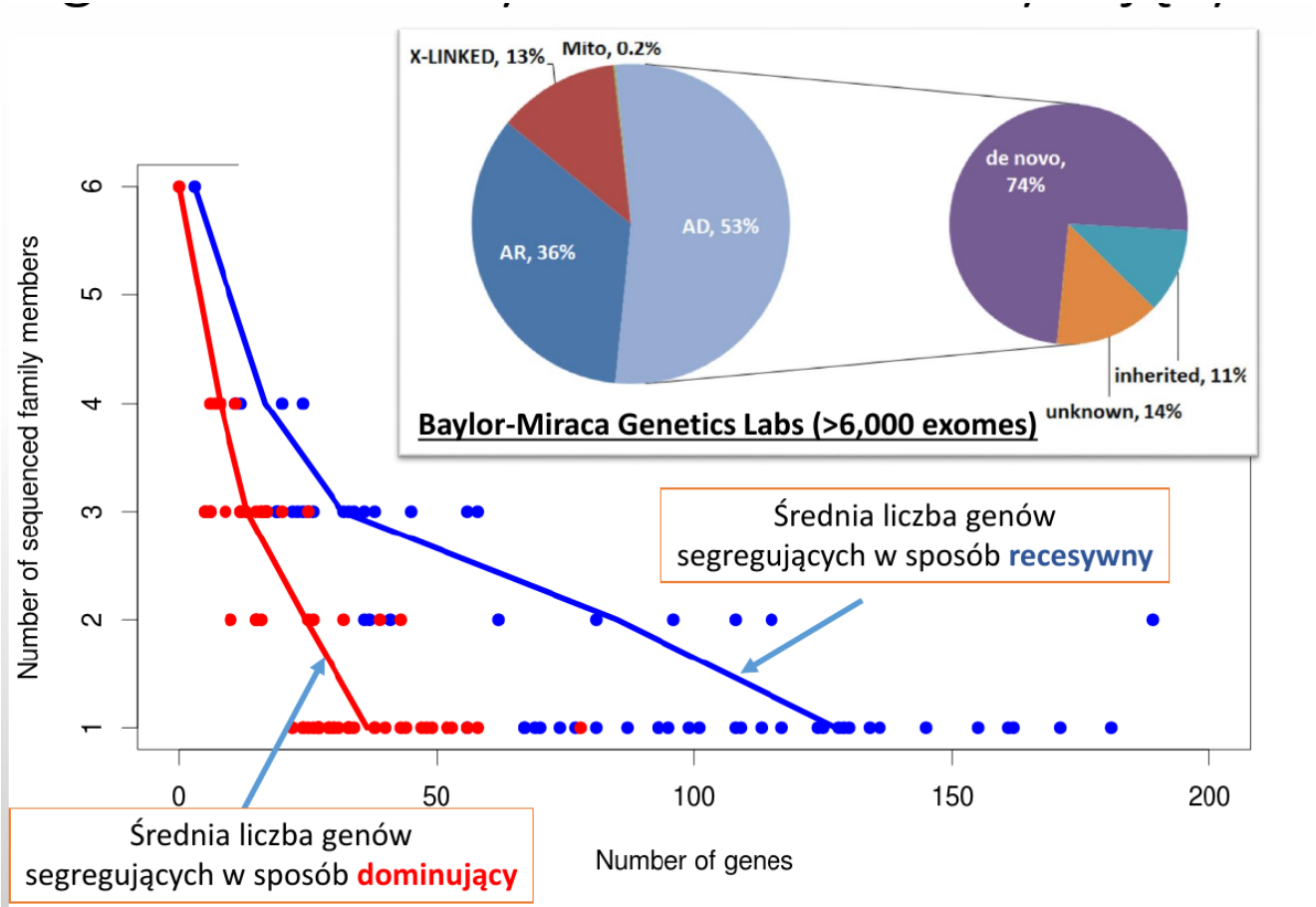
- Testy asocjacyjne
- Wymiana wiedzy z innymi grupami

Mutacje w nowych genach

Analiza wariantów w rodzinach



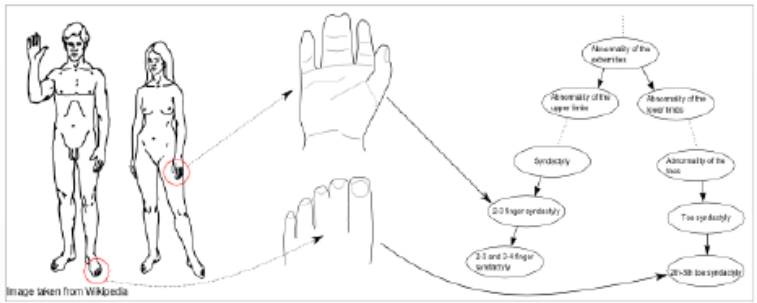
Analiza wariantów w rodzinach



Zbieranie danych fenotypowych

Human Phenotype Ontology

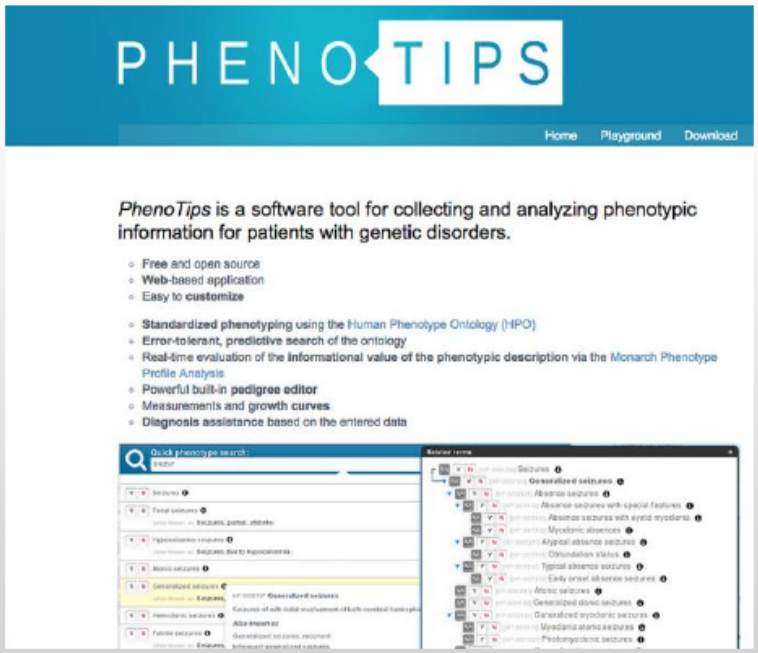
(Kochler S, et al., *Nucl. Acid Res.* 2014)



PhenoTips

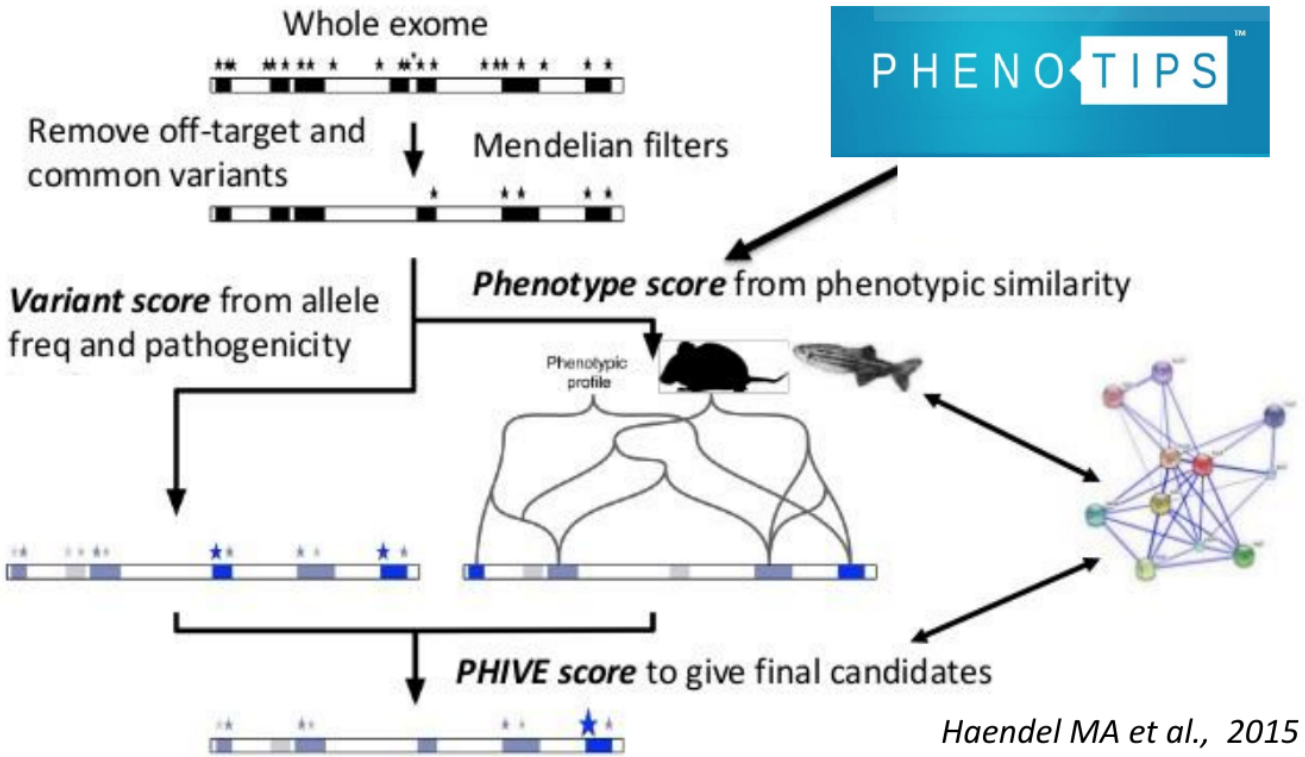
System bazodanowy do gromadzenia danych fenotypowych bazujący na ontologii HPO

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



Wykorzystanie informacji fenotyowej do priorytetyzacji wariantów

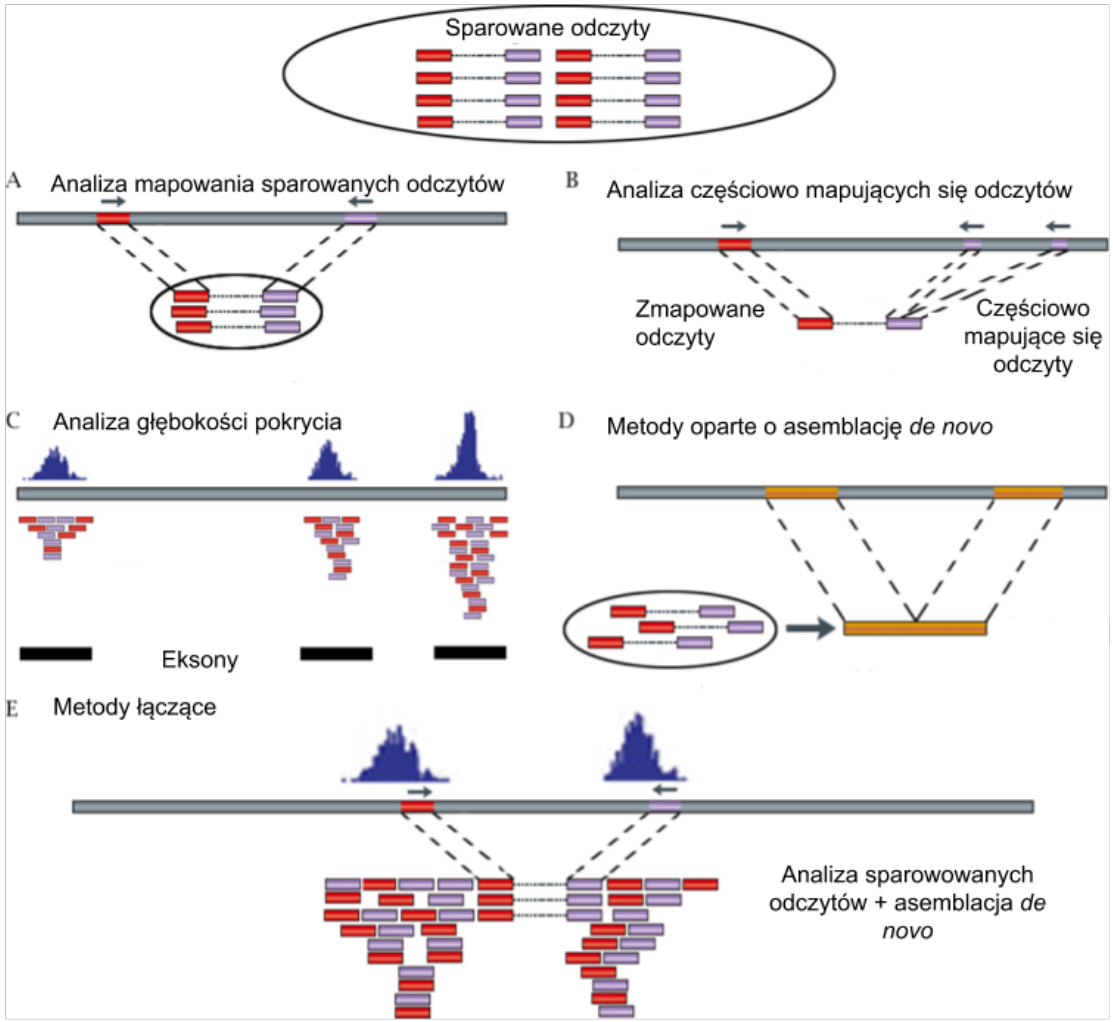
EXOMISER **http://www.sanger.ac.uk/science/tools/exomiser**



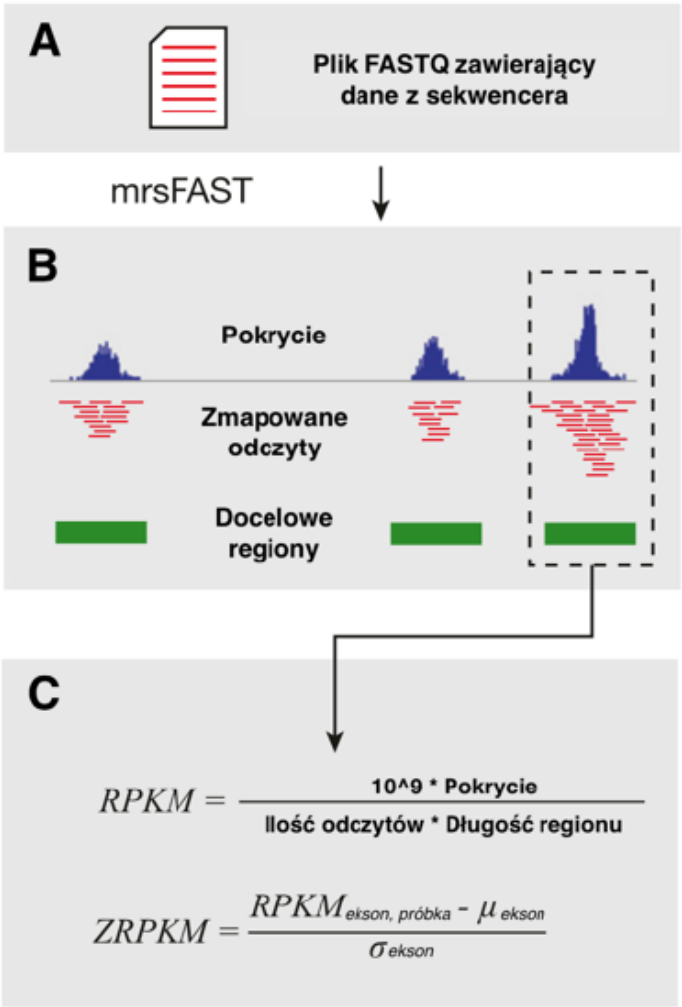
Haendel MA et al., 2015

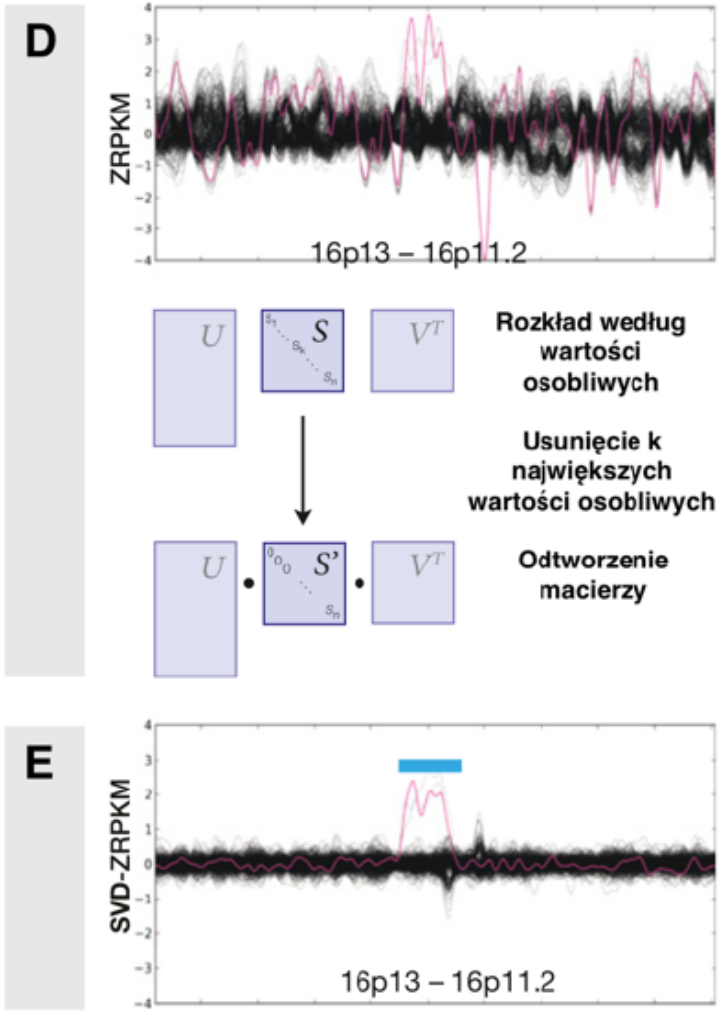
Spis treści

Rodzaje metod do wykrywania zmian liczby kopii, translokacji, inwersji

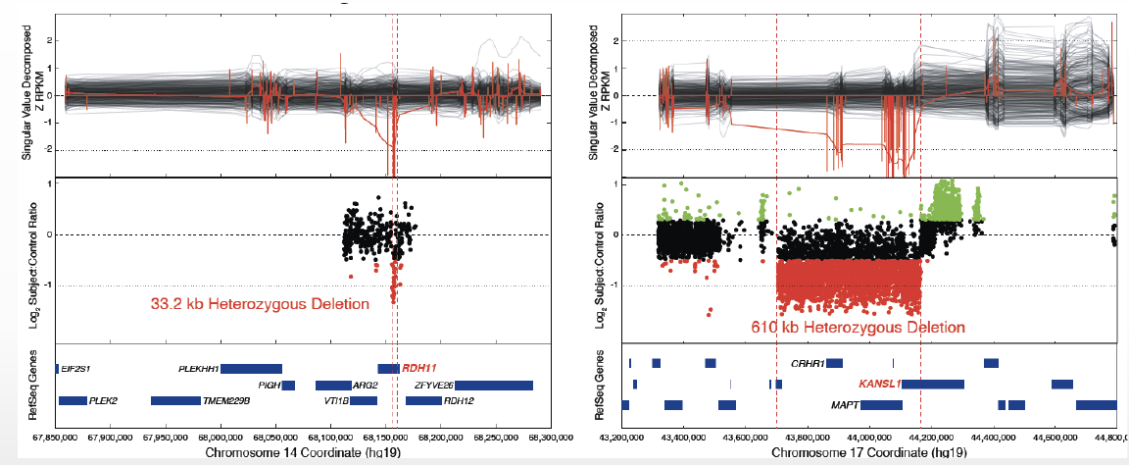


Przykład algorytmu wykorzystującego analizę głębokości pokrycia



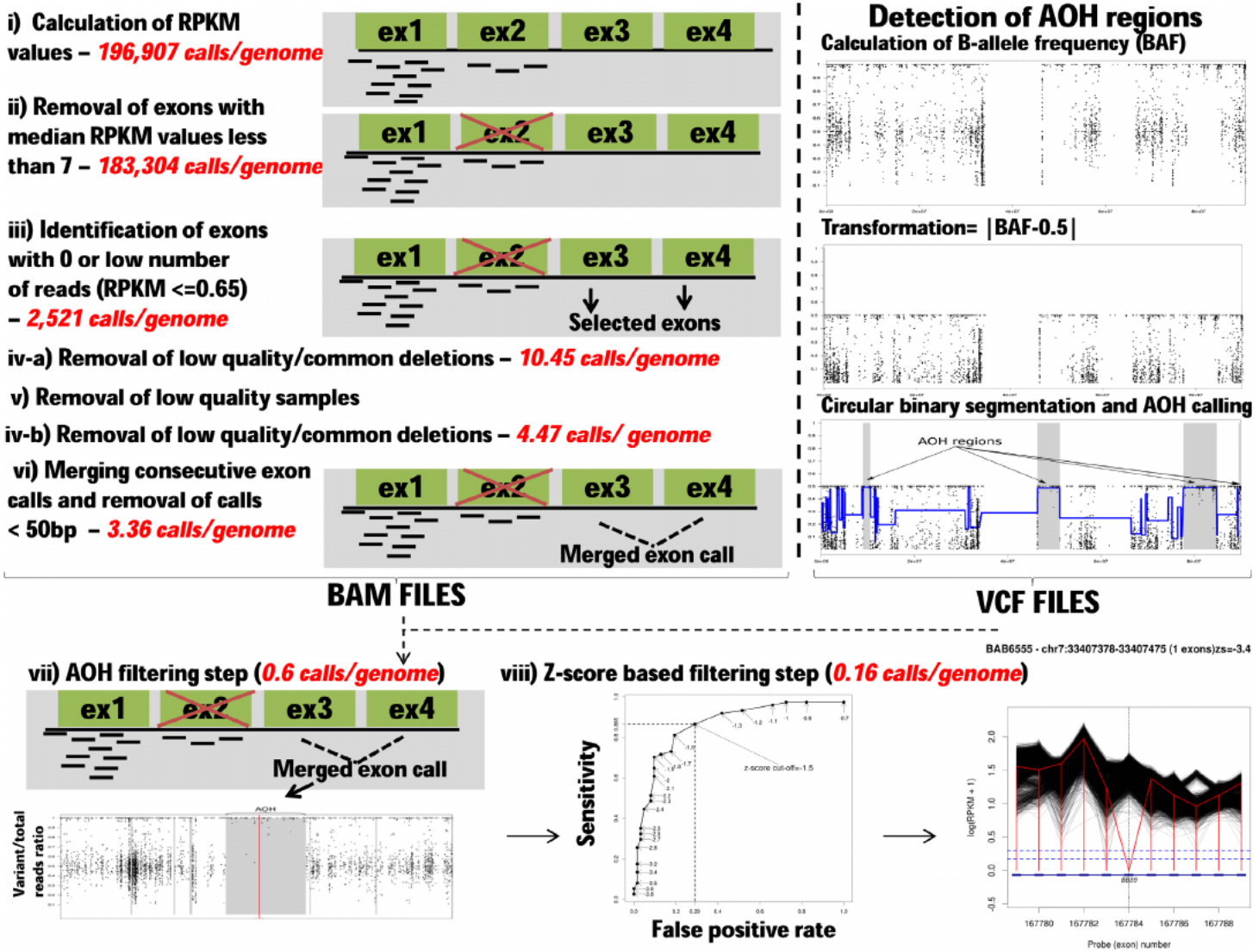


Wykrywanie zmian liczby kopii z sekwencjonowania celowanego c.d.:



- 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



Spis treści

Centers for Mendelian Genomics

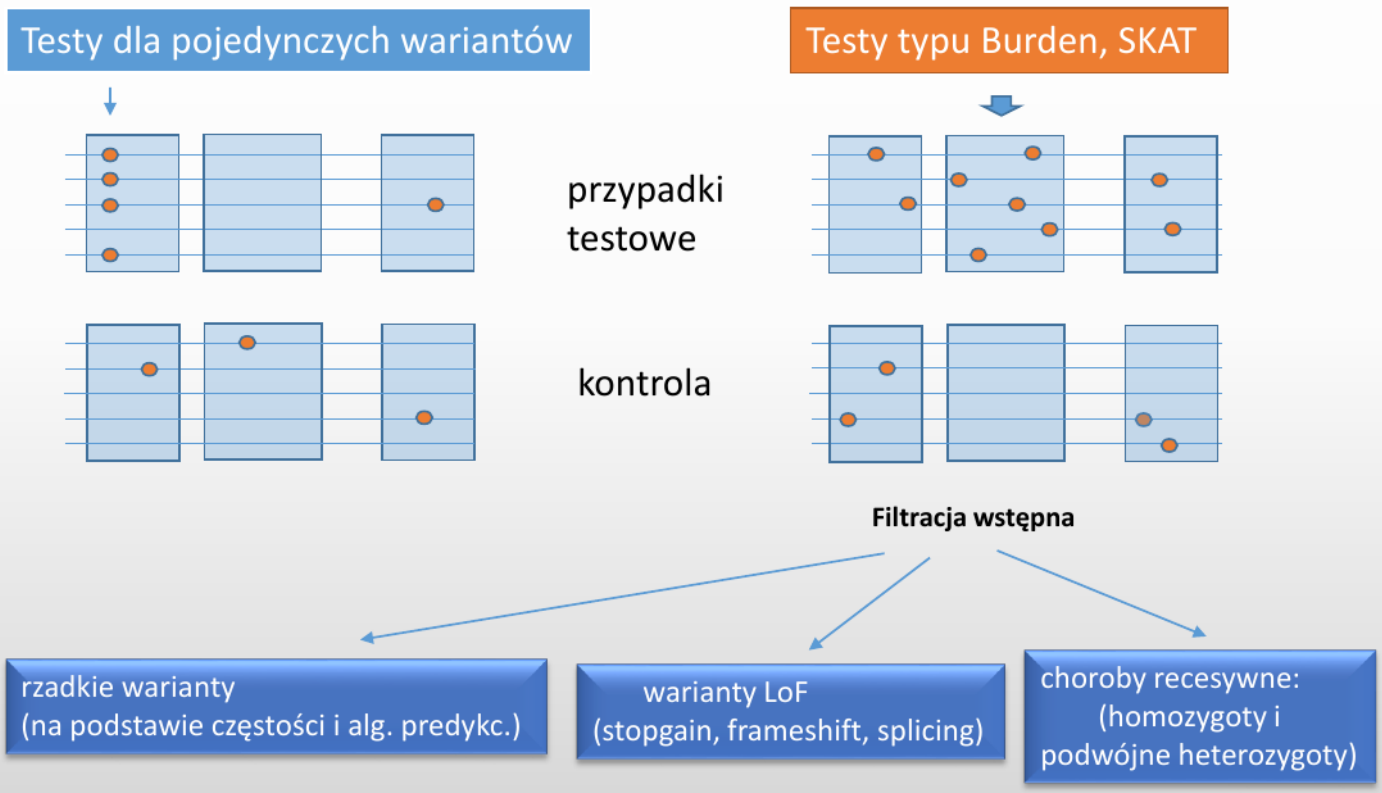
Ponad **300 nowych** genów chorobowych
Chong et al. *AJHG* , 2015

James Lupski

Eric Boerwinkle

Richard Gibbs

Testy asocjacyjne rzadkich wariantów

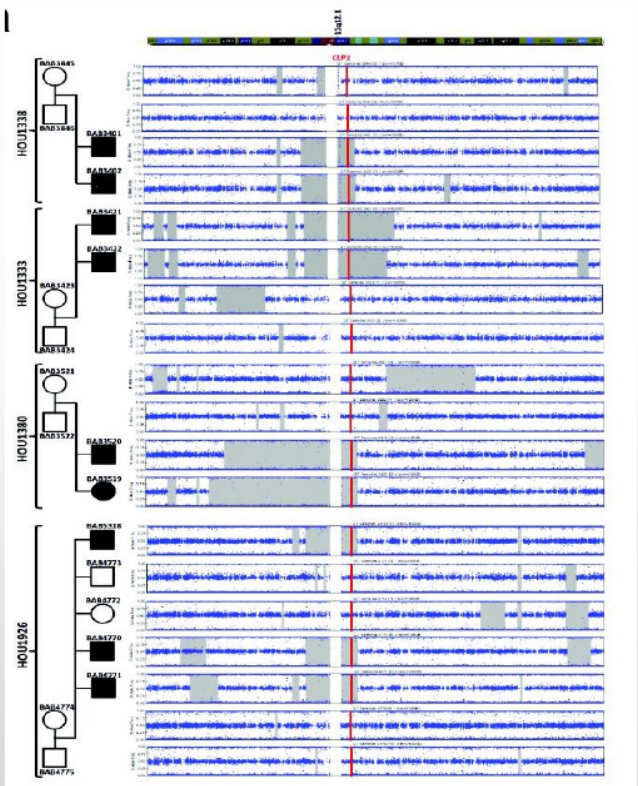


- **Opracowanie wstępnej listy genów kandydujących pozwala zredukować efekt poprawki stosowanej przy wielokrotnym testowaniu hipotez**

Odkrycia nowych genów chorobowych

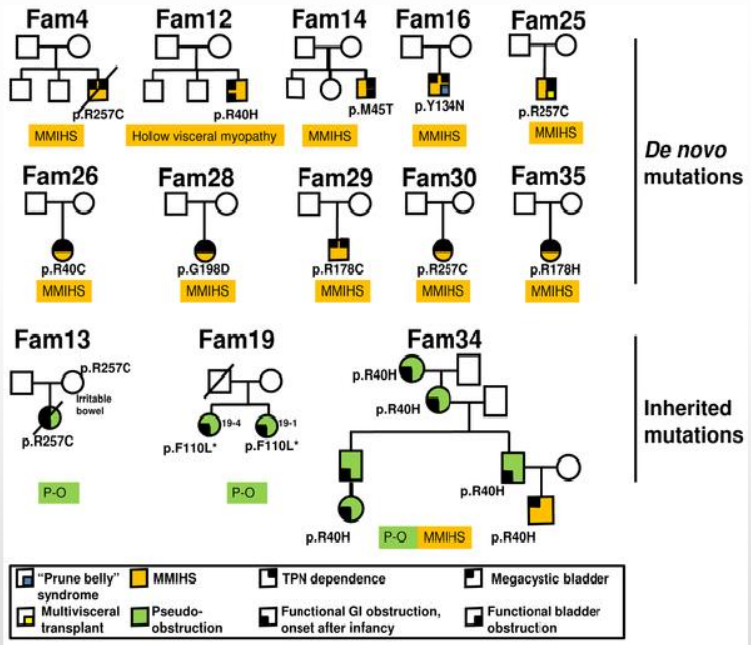
Wady rozwojowe mózgu *CLP1*

Consanguineous families from Turkey



Karaca et al., *Cell*, 2014

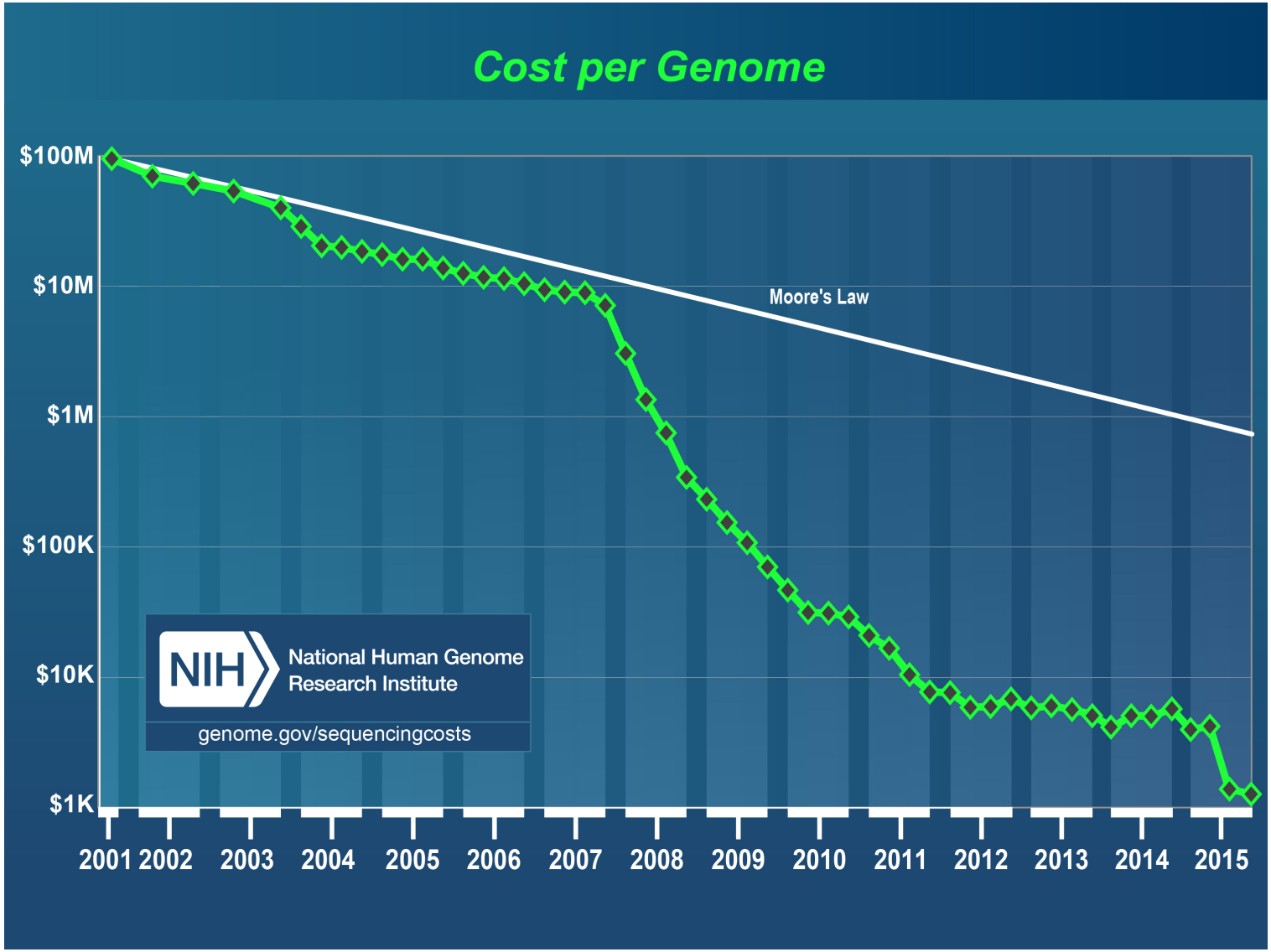
MMHIS(Berdon syndrome) *ACTG2*



Wangler et al., *PloS Genetics*, 2014

Spis treści

Koszt sekwencjonowania



Genomika jako problem Big Data

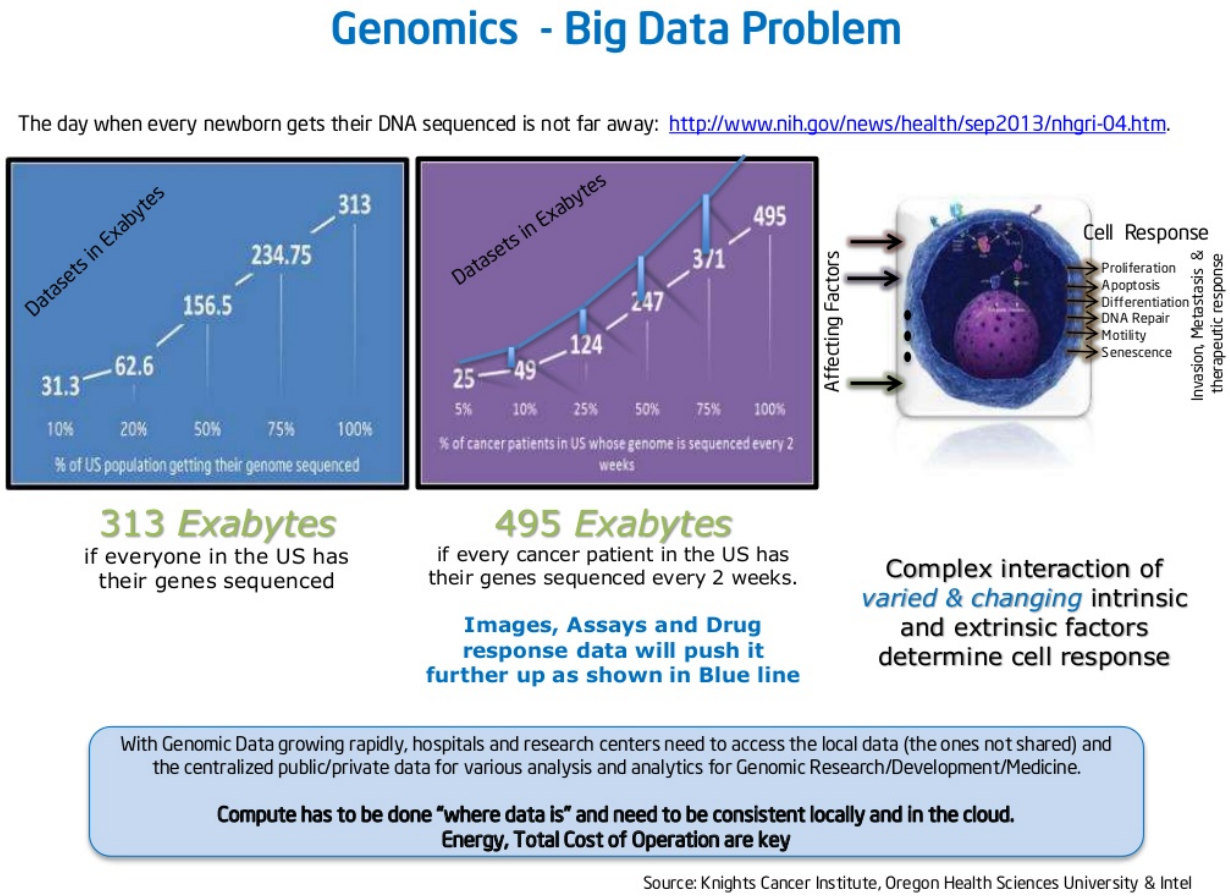
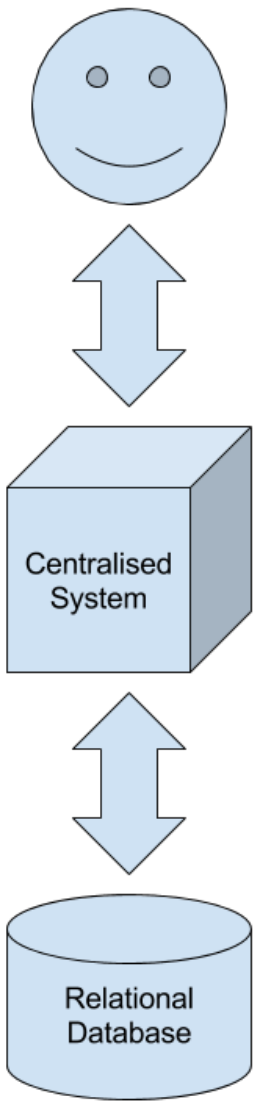
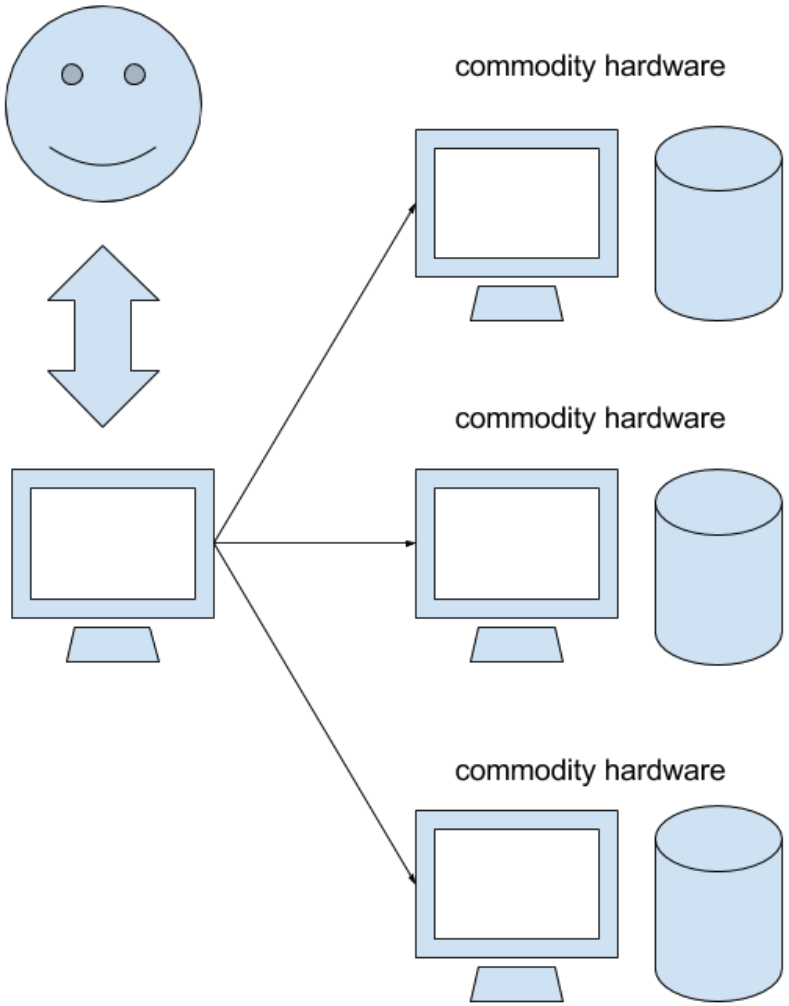


Figure: Źródło: Knights Cancer Institute, Oregon Health Sciences University & Intel

Podejście tradycyjne vs. Big Data





SCALABILITY

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 odporoność 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...

Dziękuje za uwage
Tomasz Gambin

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