Analiza genomu człowieka przy wykorzystaniu NGS w kontekście diagnostyki medycznej

dr inż. Tomasz Gambin [1,2]

[1]Instytut Informatyki, Politechnika Warszawska

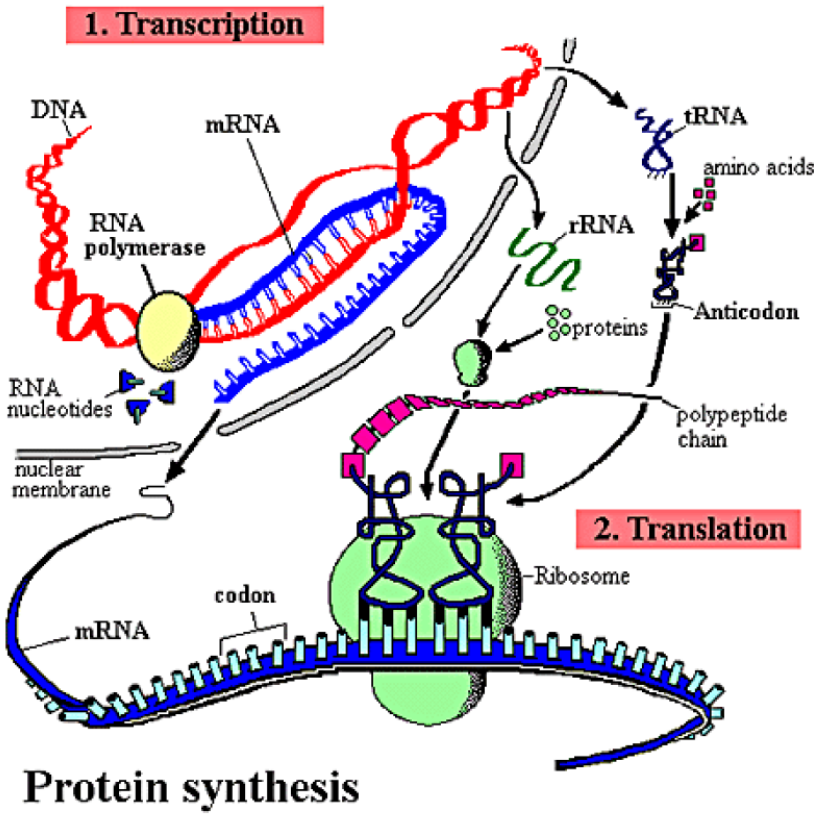
[2]Zakład Genetyki Medycznej, Instytut Matki i Dziecka w Warszawie

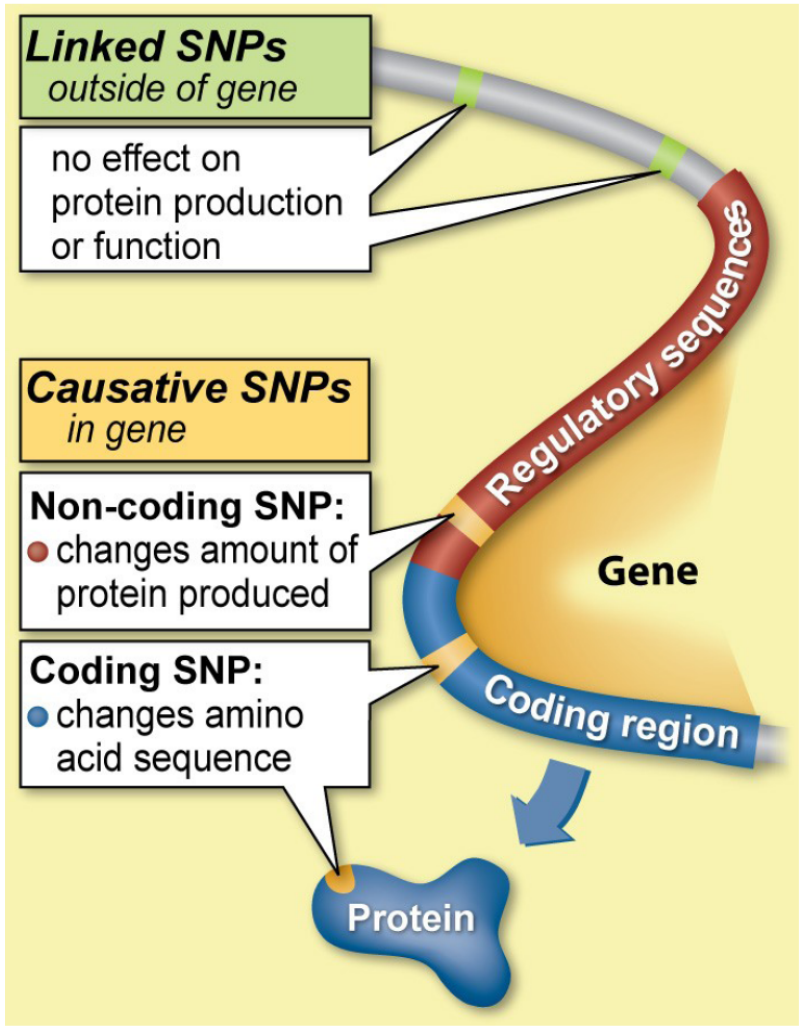
Spis treści

Genom ludzki

- 23 pary chromosomów: chr1,...,chr22, X, Y
- w sumie 2 x 3 Gbp
- Genom referencyjny – publicznie dostępna sekwencja, wygenerowana na podstawie sekwencji kilku zdrowych osób
- Wariant genetyczny – różnica pomiędzy genomem osoby badanej a genomem referenyjnym
- Różnice genetyczne pomiędzy dwoma osobami:
  - 0.1 - 0.4 % genomu (3-12 Mbp) - stanowią warianty pojedynczych nukleotydów (ang. Single Nucleotide Variants, SNVs)
  - 0.5 % genomu (15Mbp) - Zmiany strukturalne w tym zmiany liczby kopii

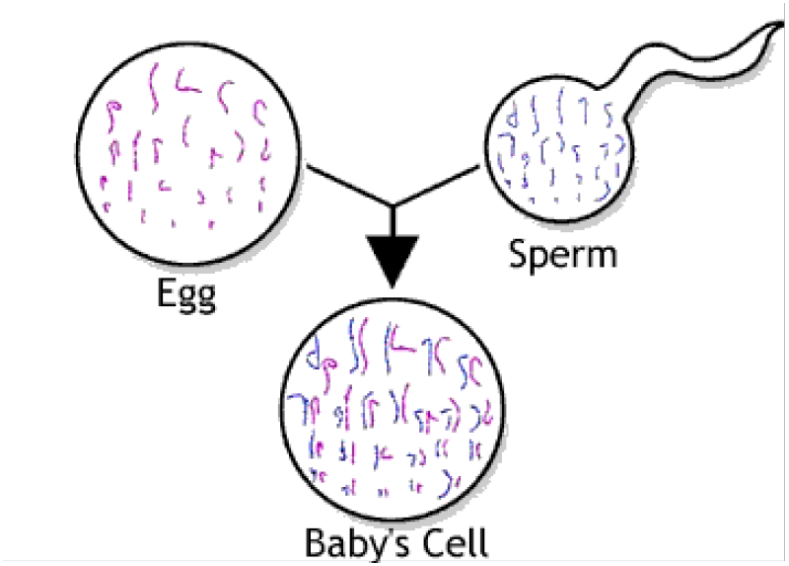
Czemu warianty są istotne?





Skąd się biorą warianty?

- Dzieci dziedziczą większość DNA od rodziców
- Jednak, ok. 60–100 zmian pojedynczych nukleotydów oraz kilka zmian strukturalnych pojawia się ***de novo***.



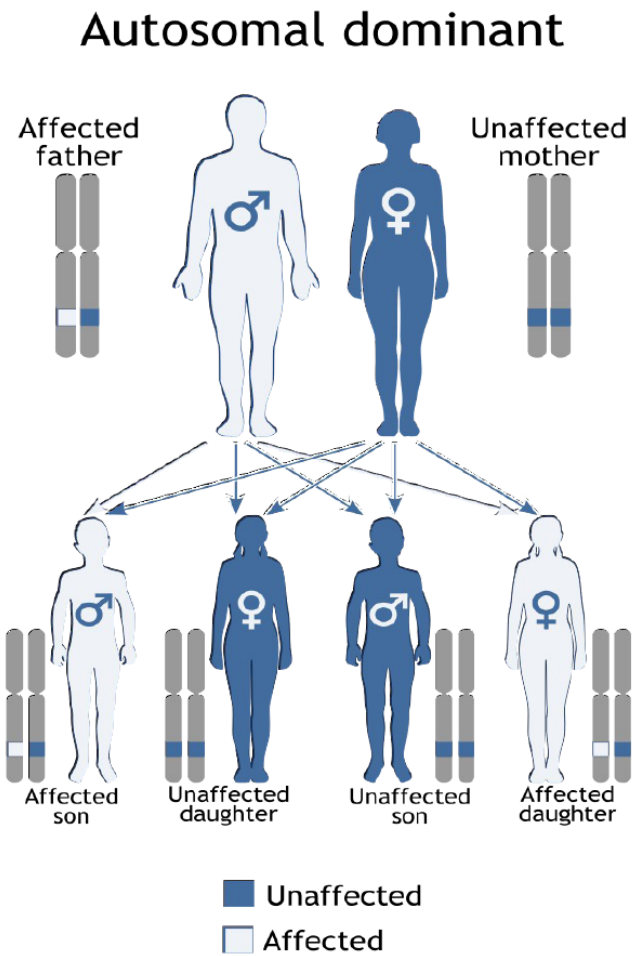
Choroby Mendlowskie, modele dziedziczenia

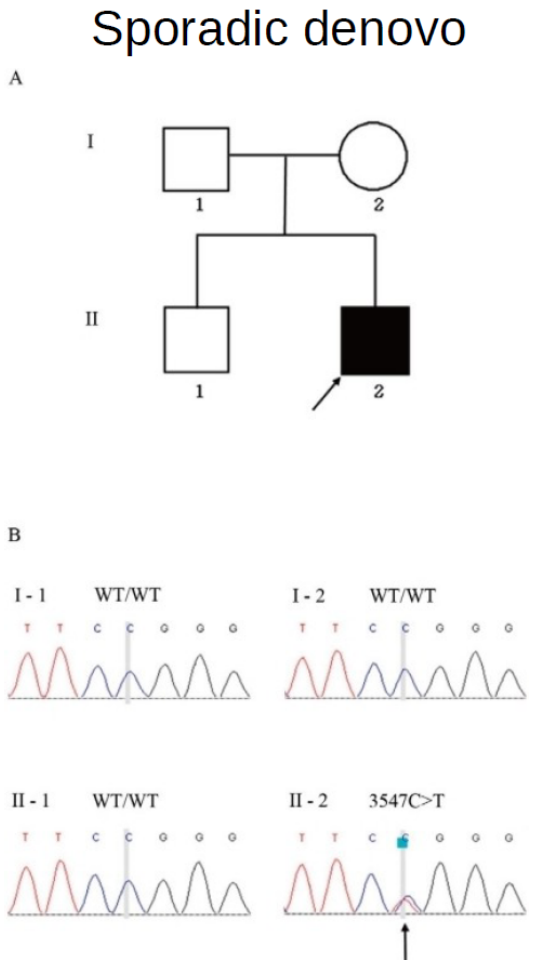
Choroby Mendlowskie

Choroby jednogenowe, czyli wywołane przez wariant(y) w pojedynczym genie.

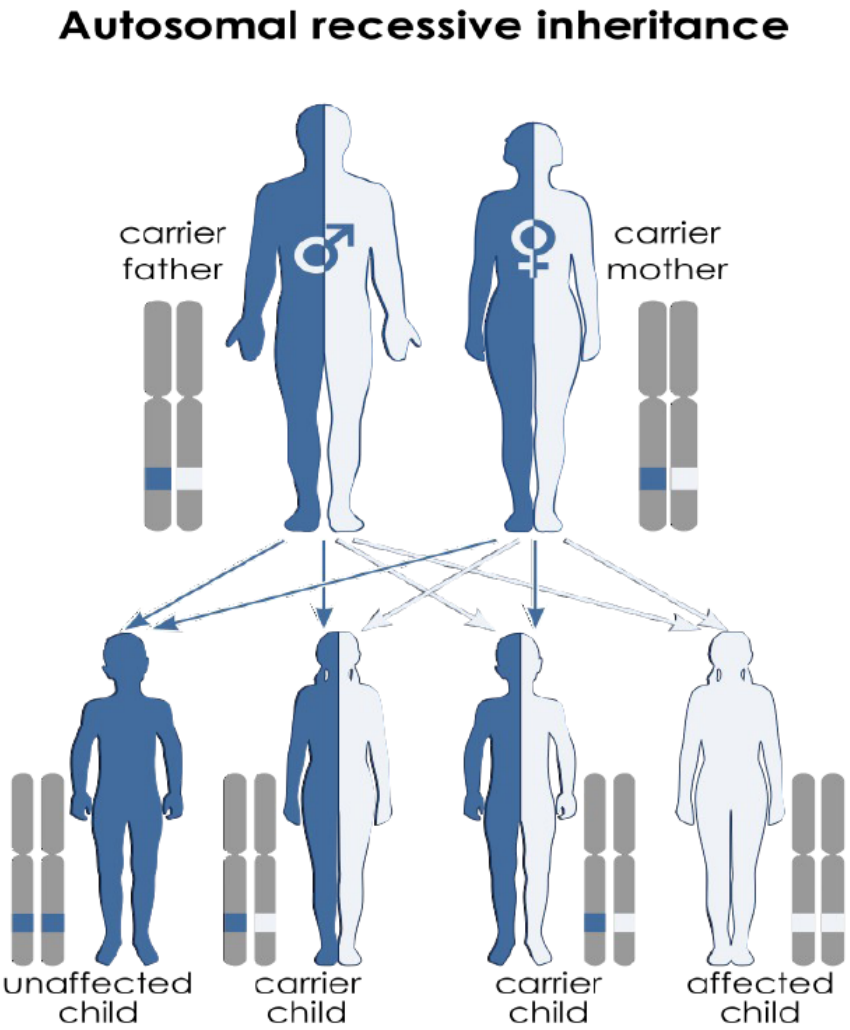
- Choroby autosomalnie dominujące
- Choroby autosomalnie recesywne
- Choroby sprzężone z chromosomem X (dominujące i recesywne)

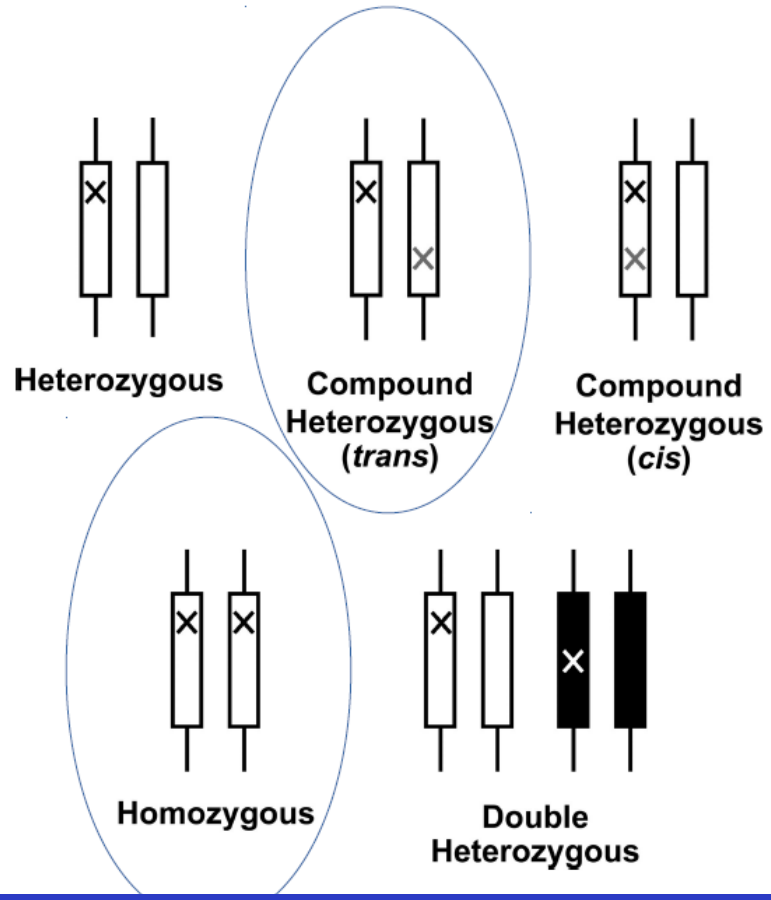
Choroby autosomalnie dominujące



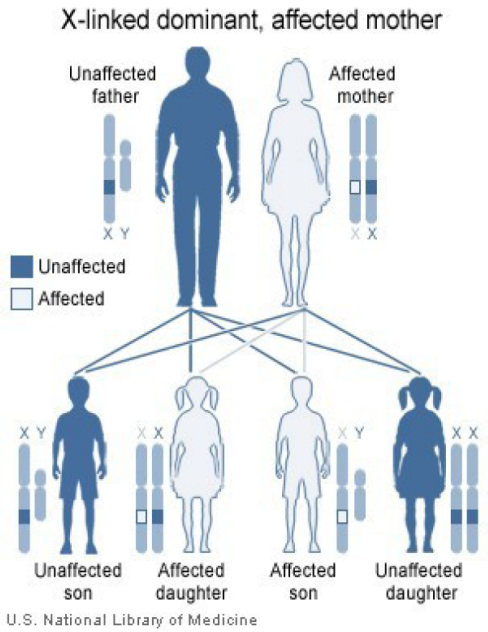


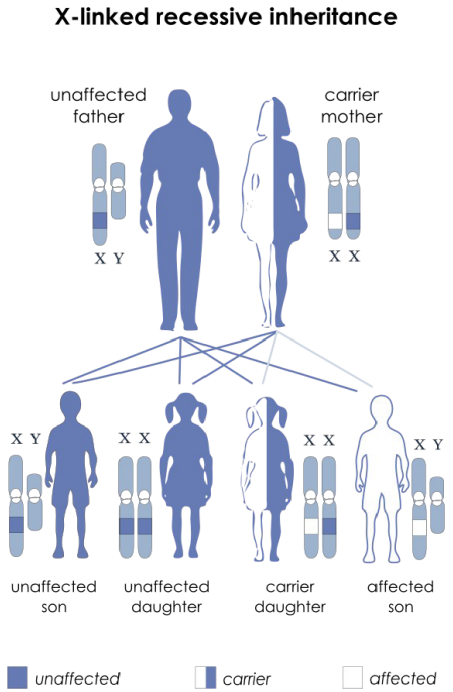
Choroby autosomalnie recesywne





Choroby sprzężone z X





Częstość chorób uwarunkowanych genetycznie

Choroby dominujące	Częstość występowania (na 1000 żywo urodzonych)
Pląsawica Huntingtona	0,5
Neurofibromatosis	0,4
Dystrofia miotoniczna	0,2
Torbielowatość nerek	0,8
Ślepota dominująca	0,1
Hipercholesterolemia	2,0
Sferocytoza wrodzona	0,2
Dentinogenesis imperfecta	0,1
Osteogenesis imperfecta	0,04
Zespół Marfana	0,05

Choroby recesywne	Częstość występowania (na 1000 żywo urodzonych)
Autosomalne	
Mukowiscydoza	0,4
Głuchota (różne postaci)	0,5
Ślepota (różne postaci)	0,2
Fenyloketonuria	0,08
Galaktozemia	0,025
Mukopolisacharydozy (różne postaci)	0,04
Glikogenozy	0,02
Sprzężone z chromosomem X	
Dystrofia mięśniowa Duchenne'a	0,14
Hemofilia A	0,01

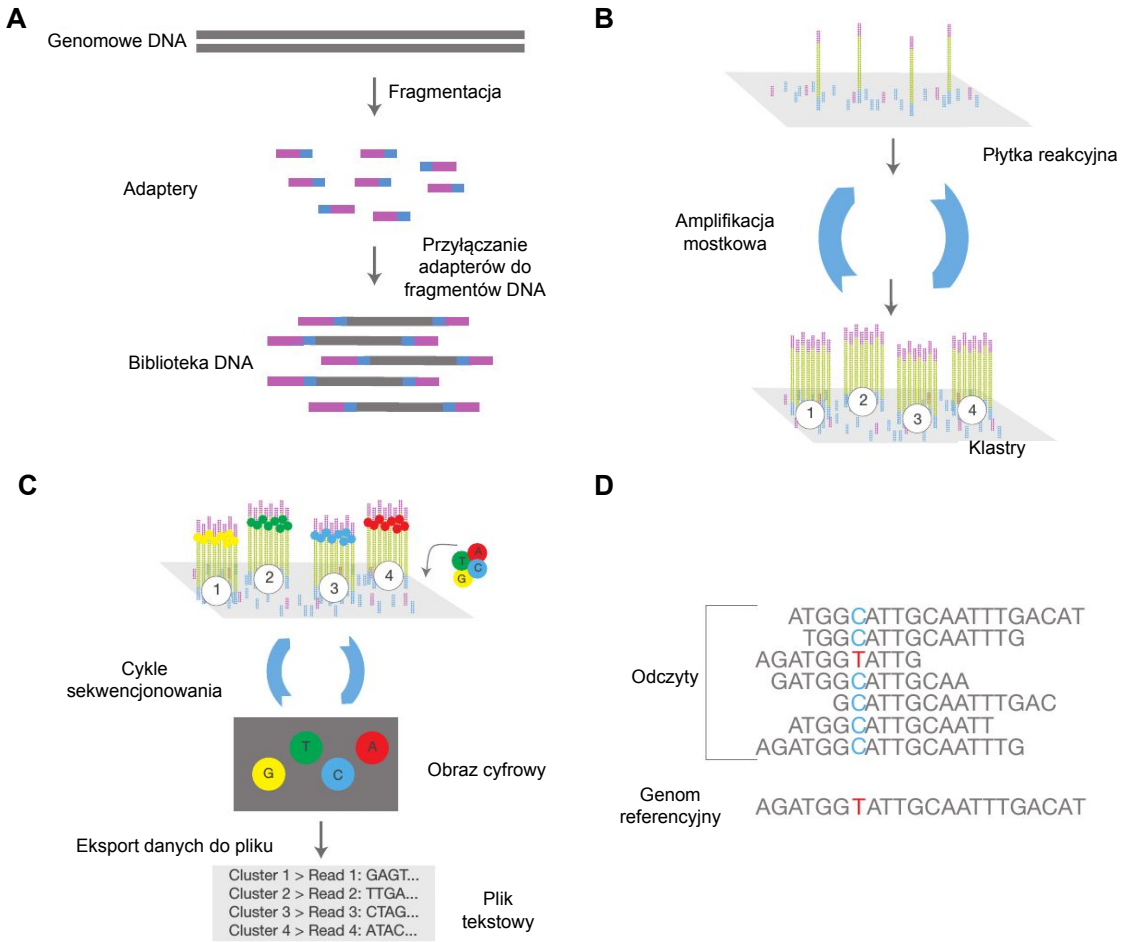
https://pl.wikipedia.org/wiki/Choroby_genetyczne_czlowieka

Spis treści

Przykładowe zastosowania kliniczne NGS

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

(Re)sekwencjonowanie w technologii SBS



https://www.illumina.com/content/dam/illumina-marketing/

Sekwencjonowanie całogenomowe vs. celowane

Całogenomowe

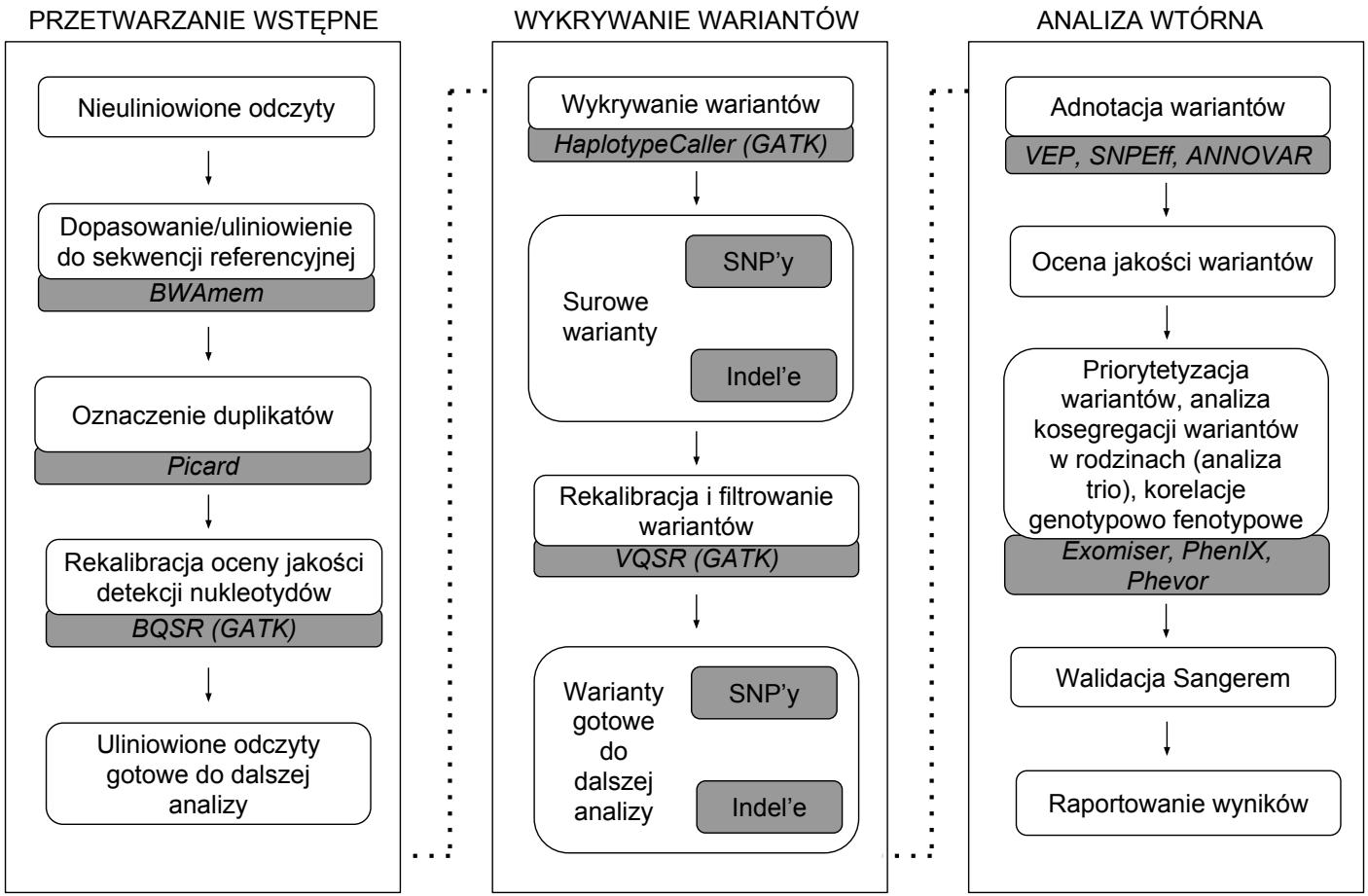
- Stosunkowo równomierne pokrycie odczytami całego genomu
- Łatwiejsza identyfikacja zmian strukturalnych
- Wysoki koszt
- Duże wolumeny danych (100Gb/ pacjent)
- Względnie mniejesze pokrycie niż w przypadku sekw. celowanego (mozaiki)

Celowane (np. całoeksomowe, panele genów)

- Nierównomierne pokrycie odczytami całego genomu
- Ograniczone możliwości wykrywania zmian strukturalnych
- Niższy koszt sekwencjonowania
- Mniejszy wolumen danych
- Większe pokrycie odczytami (do kilkuset x w zastosowaniach klinicznych)

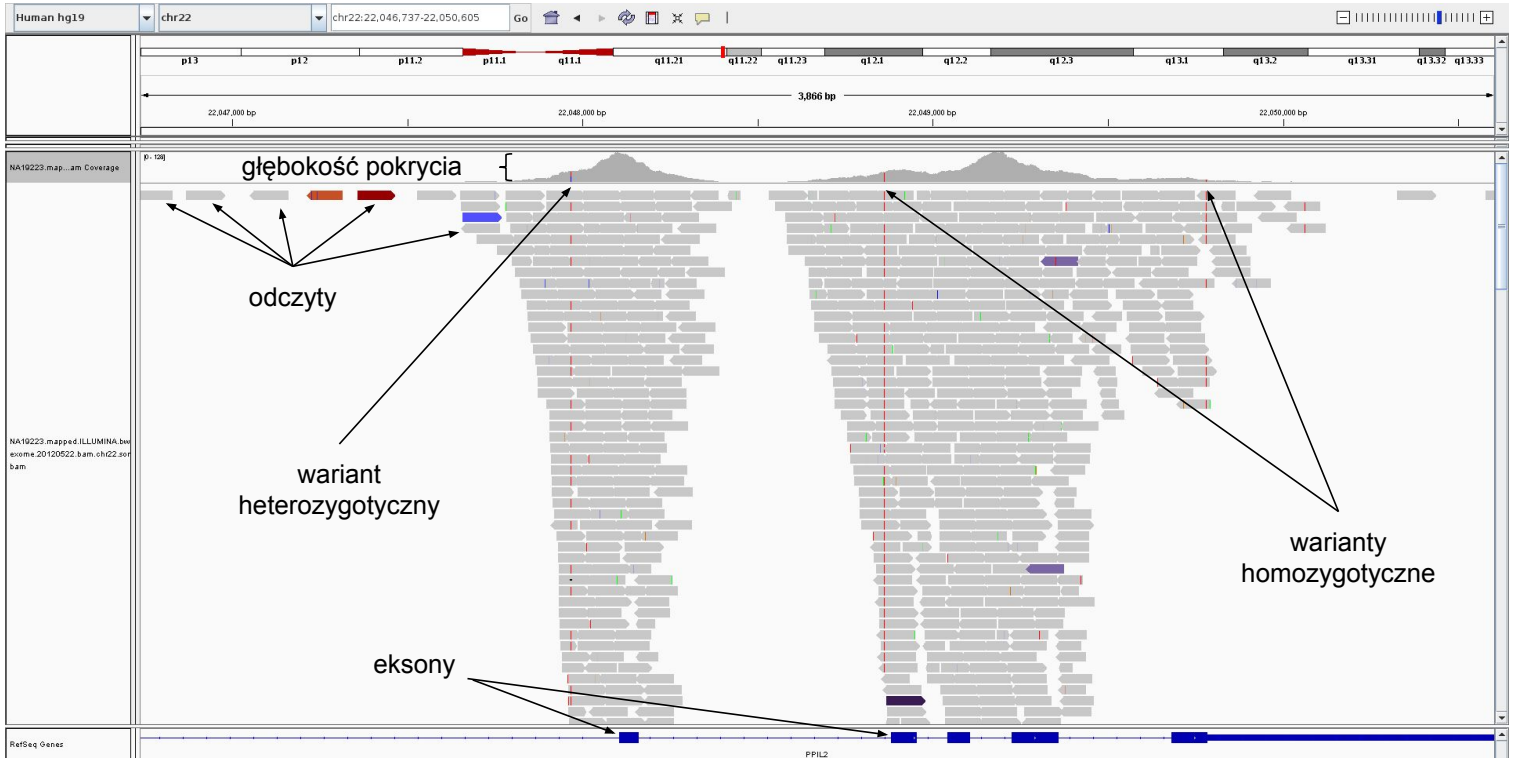
Spis treści

Tok przetwarzania



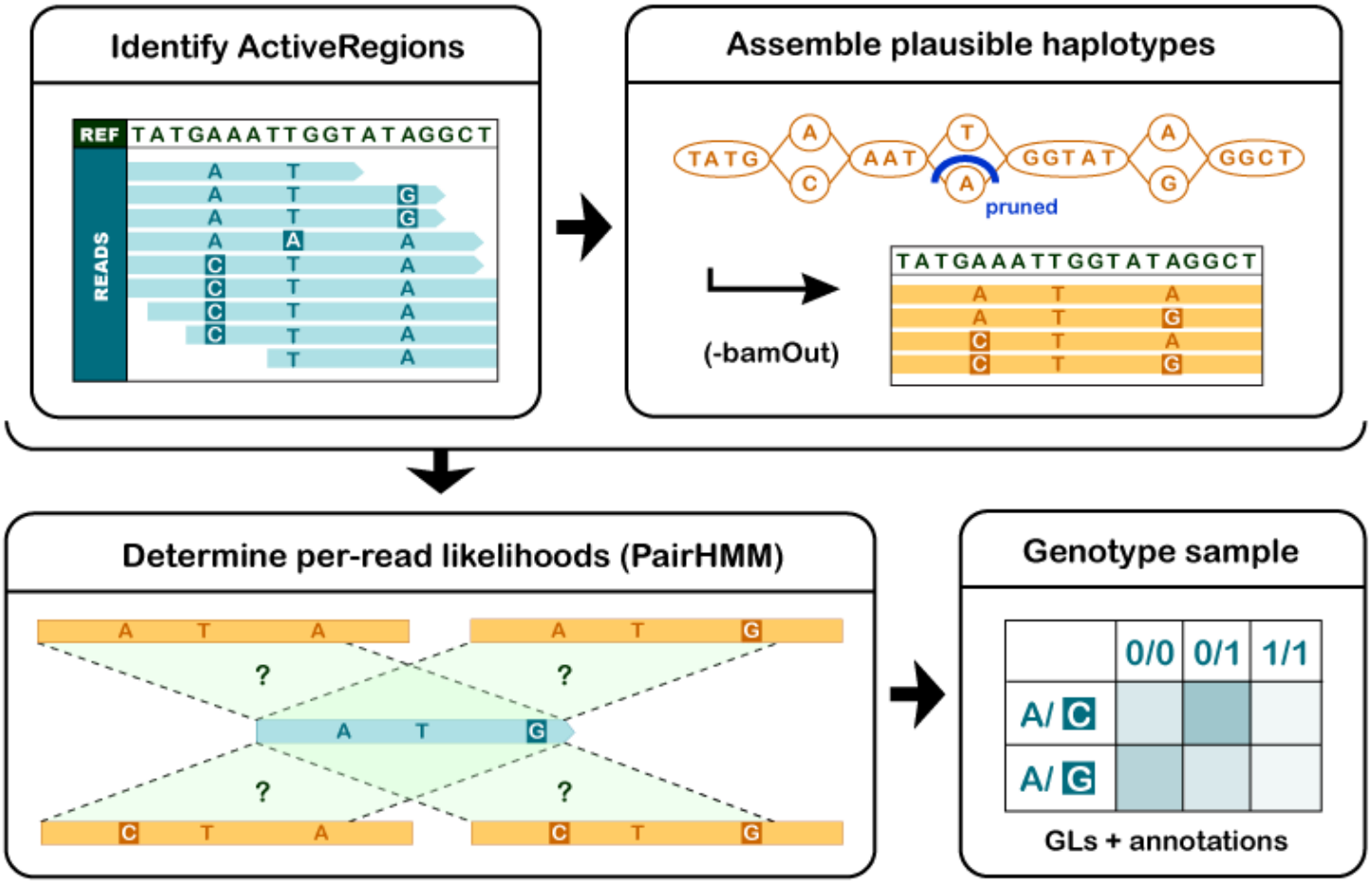
Główne kroki: mapowanie/uliniowienie –> wykrywanie wariantów –> adnotowanie –> priorytetyzacja i interpretacja

Zmapowane/uliniawione odczyty – pliki BAM



Do mapowania wykorzystuje się transformację BWT (Burrows Wheeler Transform): `http://slideplayer.com/slide/9095176/`

Wykrywanie wariantów – GATK HaplotypeCaller

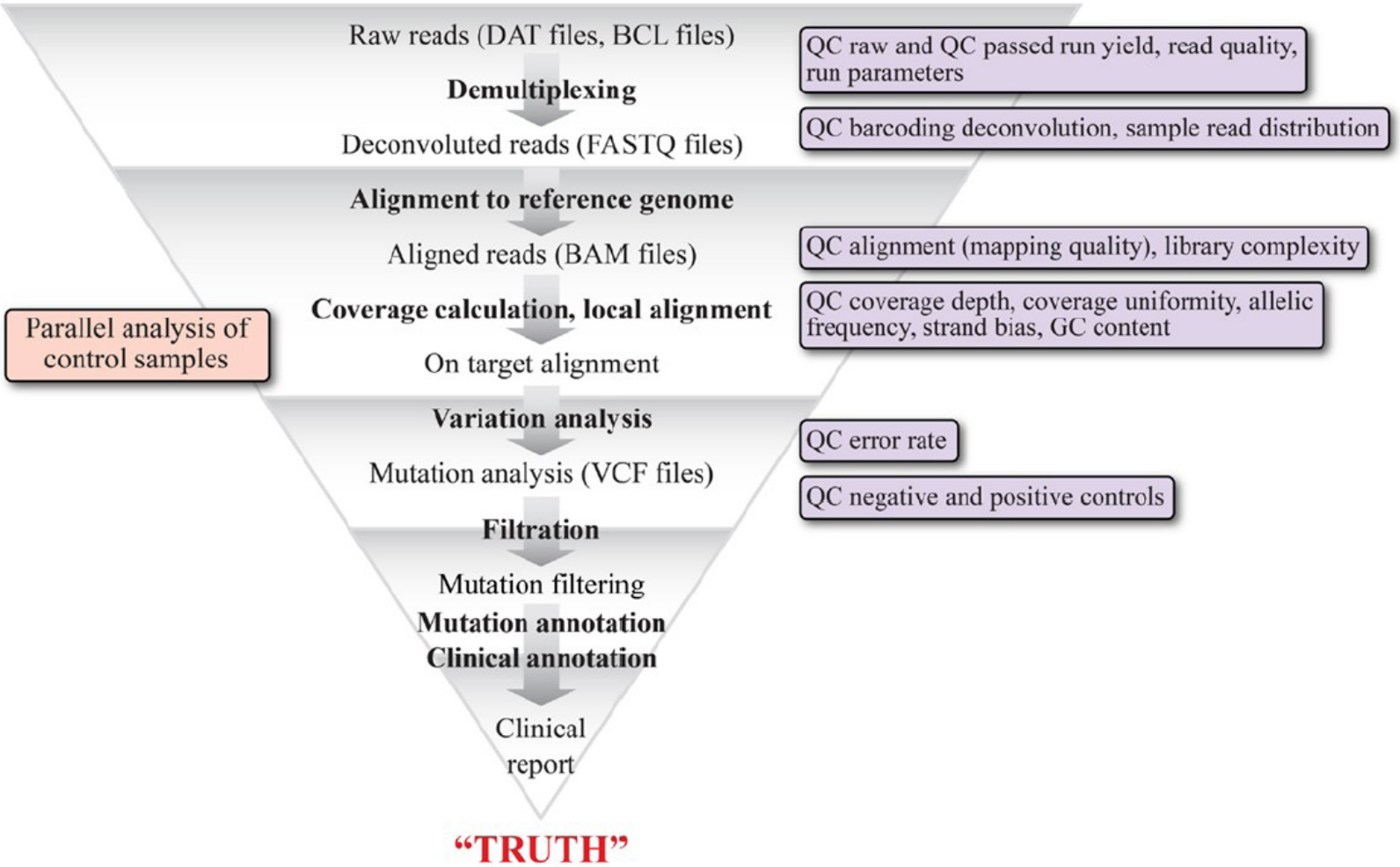


https://software.broadinstitute.org/gatk/

Adnotowanie wariantów

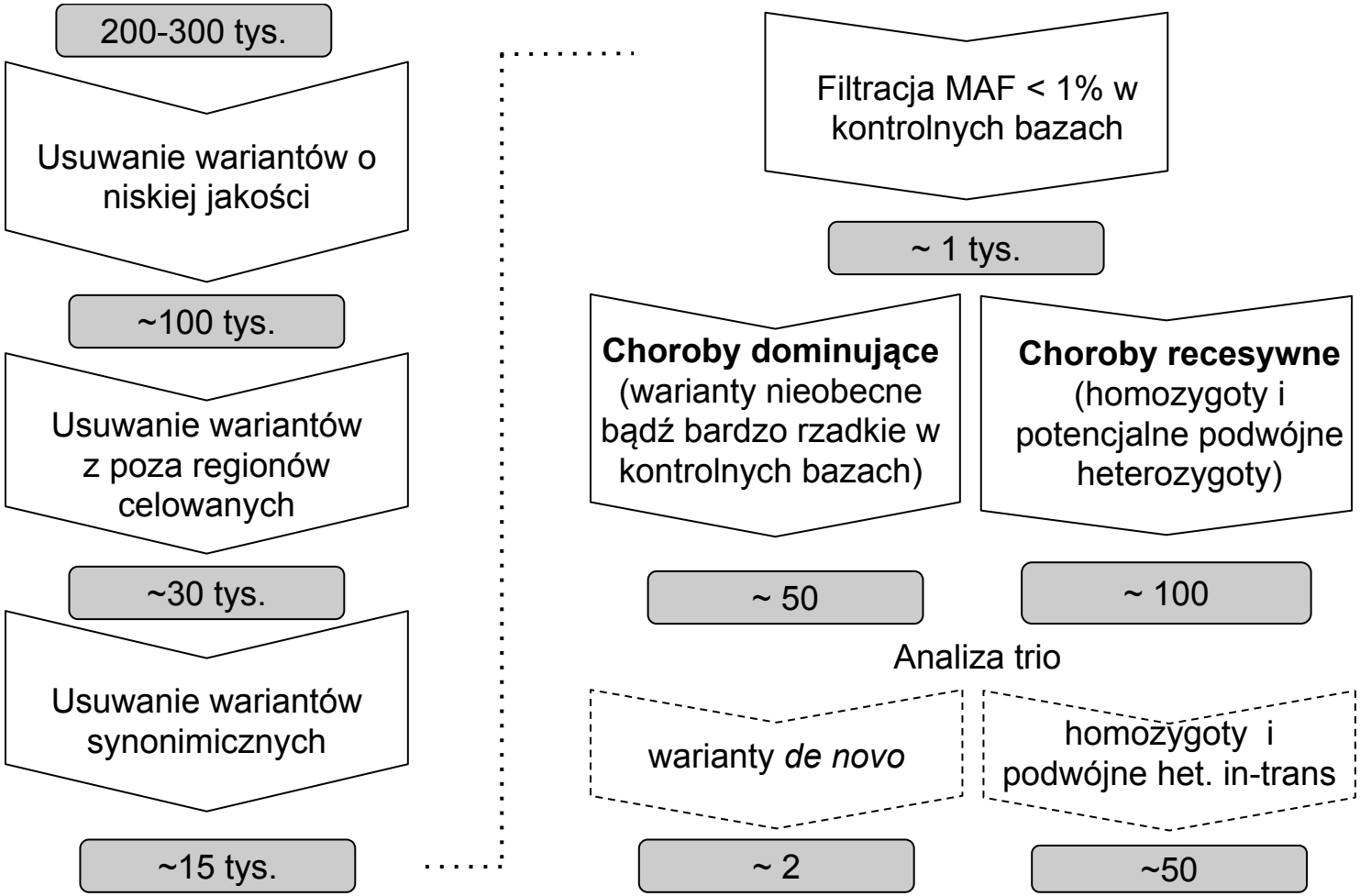
Kategoria	Komponenty
Efekt oddziaływania na białko	(VEP, SNPEff, ANNOVAR) x (Ensemble/Genecode, RefSeq)
Predykcje funkcjonalne (warianty missensowne oraz splicingowe)	dbscSNV, FATHMM, LRT, MetaLR, MetaSVM, MutationTaster, MutationAcessor, Polyphen2, SIFT
Predykcje funkcjonalne (warianty niekodujące)	CADD, FATHMM-MKL, Funseq, Funseq2, RegulomeDB
Częstości alleli	GnomAD, ExAC, UK10K, ESP, 1000Genomes
Znane warianty patogenne	ClinVar, COSMIC, GWAS catalog, GRASP
Konswerwacja	GERP++, phastCons, PhyloP, SiPhy
Epigenomika	Encode, FANTOM5, Roadmap
Opisy genów	Znane korelacje genotypowo-fenotypowe (OMIM, Medgen), ekspresja, interakcje gen-gen, ścieżki sygnałowe, fenotypy organizmów modelowych

Kontrola jakości



Spis treści

Filtrowanie wariantów



Identyfikacja wariantów potencjalnie patogennych

Znane mutacje w znanych genach

OMIM®

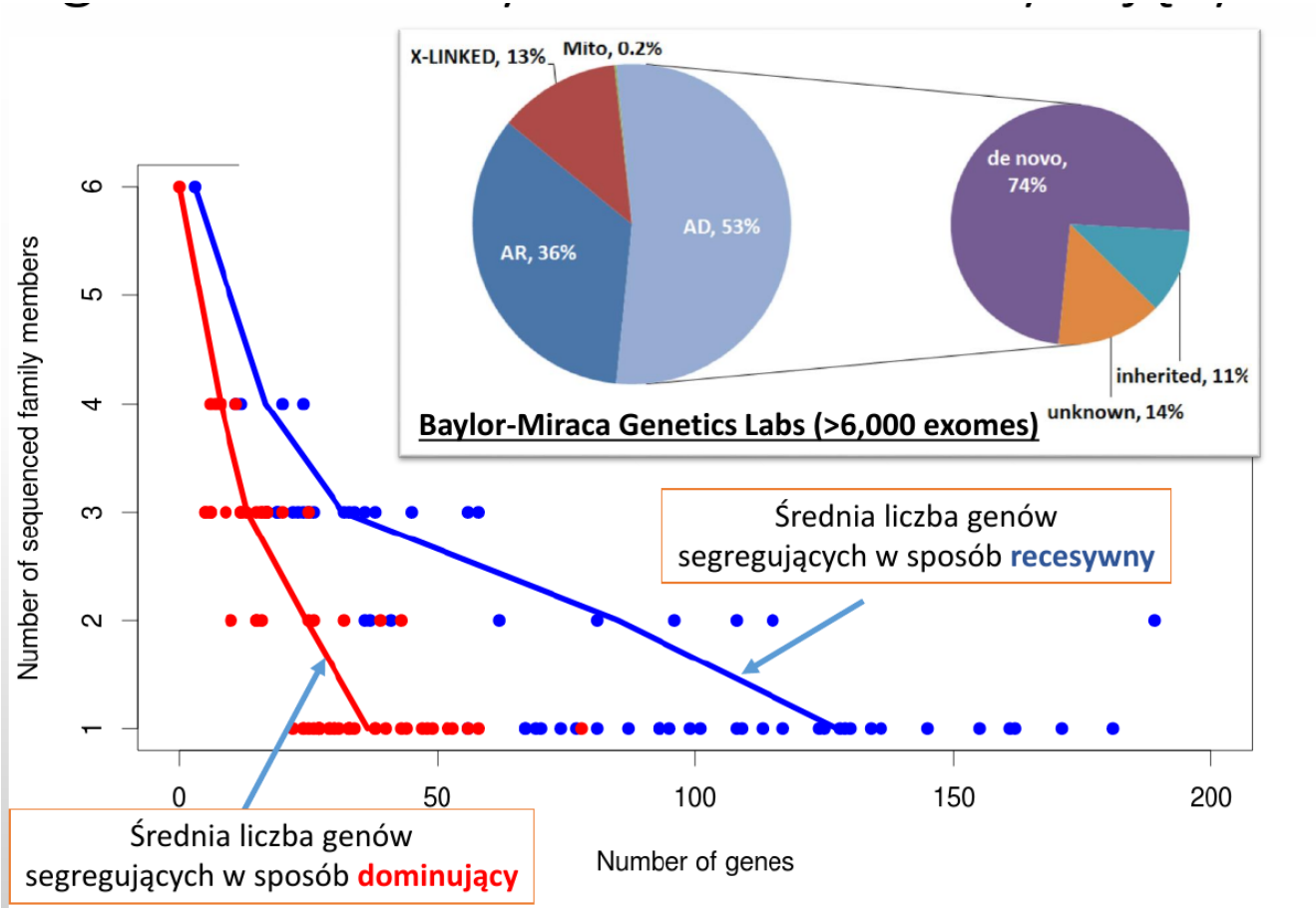
- Analiza danych w rodzinach
- Priorytetyzacja wariantów oraz korelacja genotyp-fenotyp
- Analiza zmian liczby kopii (CNV)

Nowe mutacje w znanych genach

- Testy asocjacyjne
- Wymiana wiedzy z innymi grupami

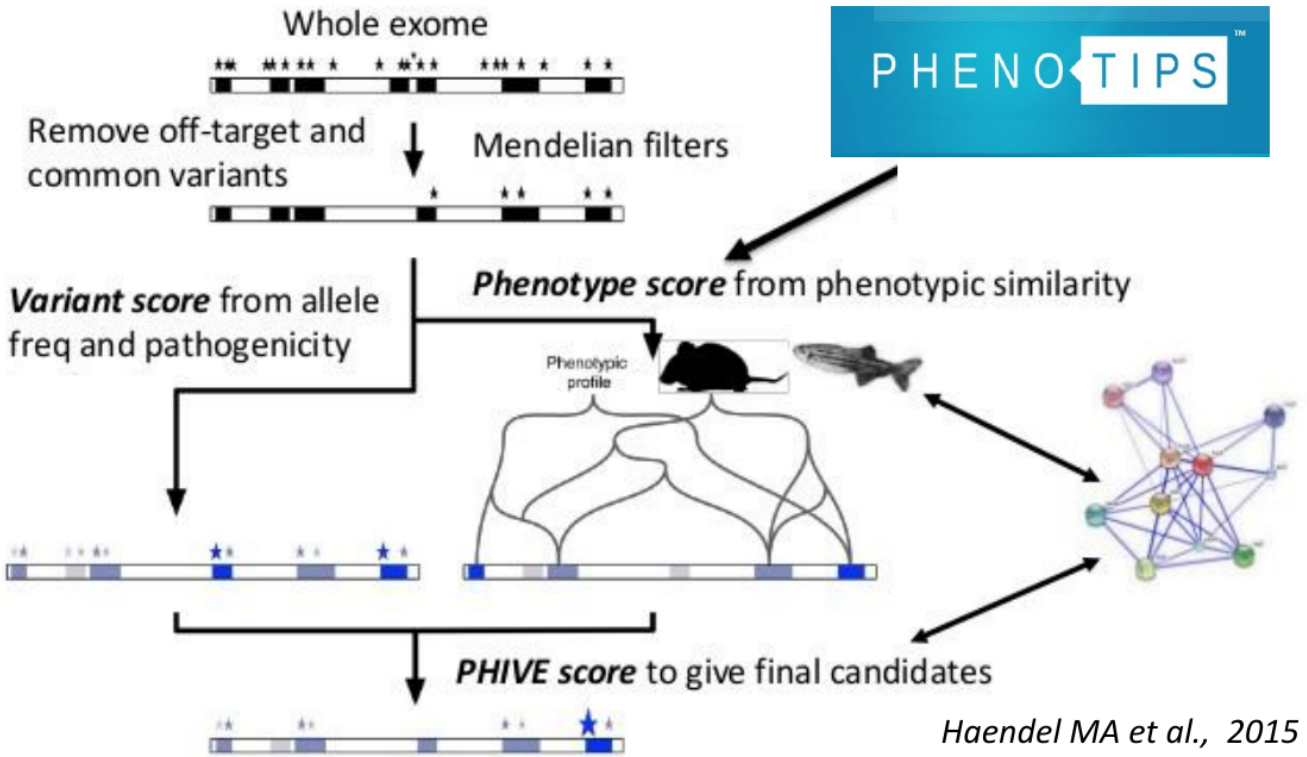
Mutacje w nowych genach

Analiza wariantów w rodzinach



Wykorzystanie informacji fenotyowej do priorytetyzacji wariantów

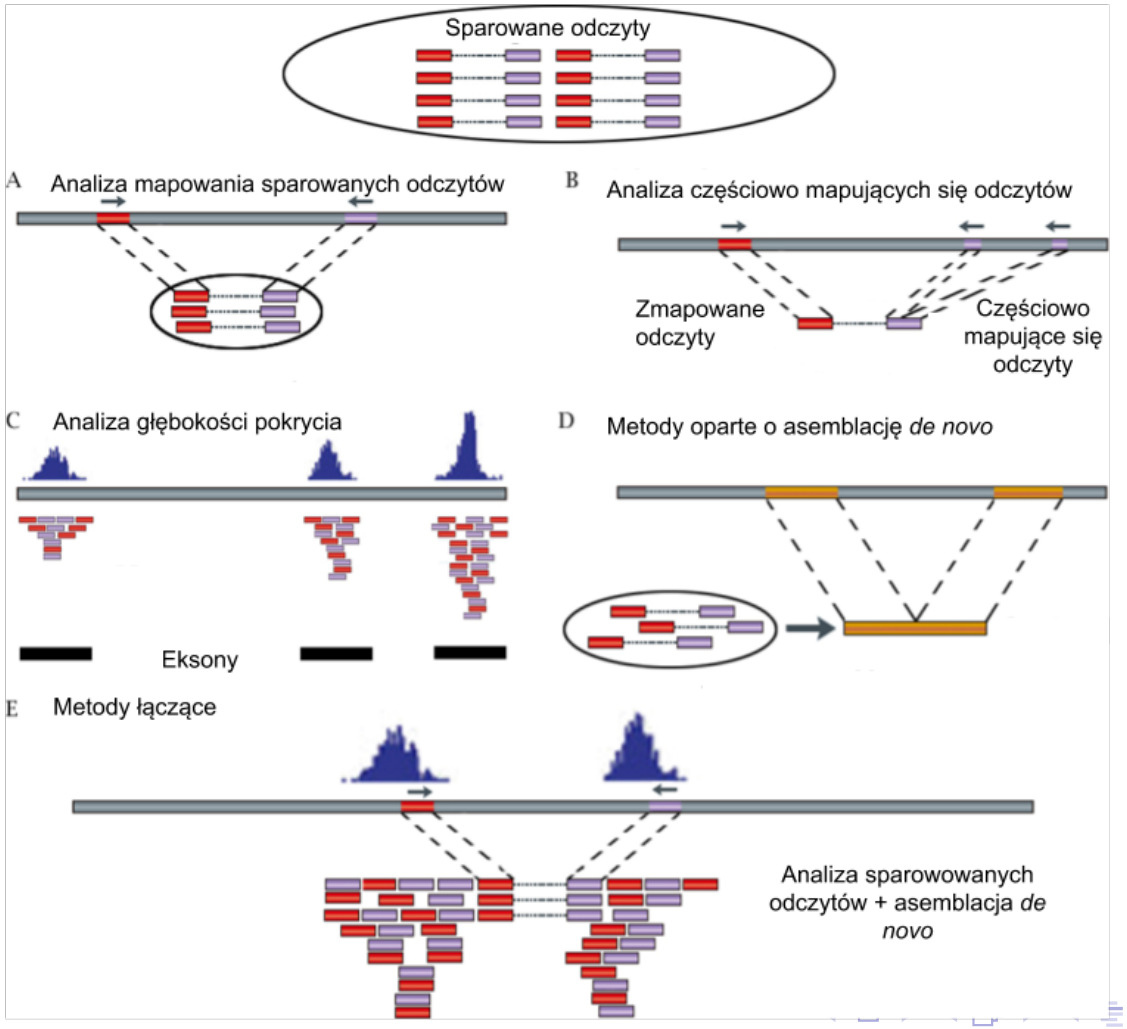
EXOMISER http://www.sanger.ac.uk/science/tools/exomiser



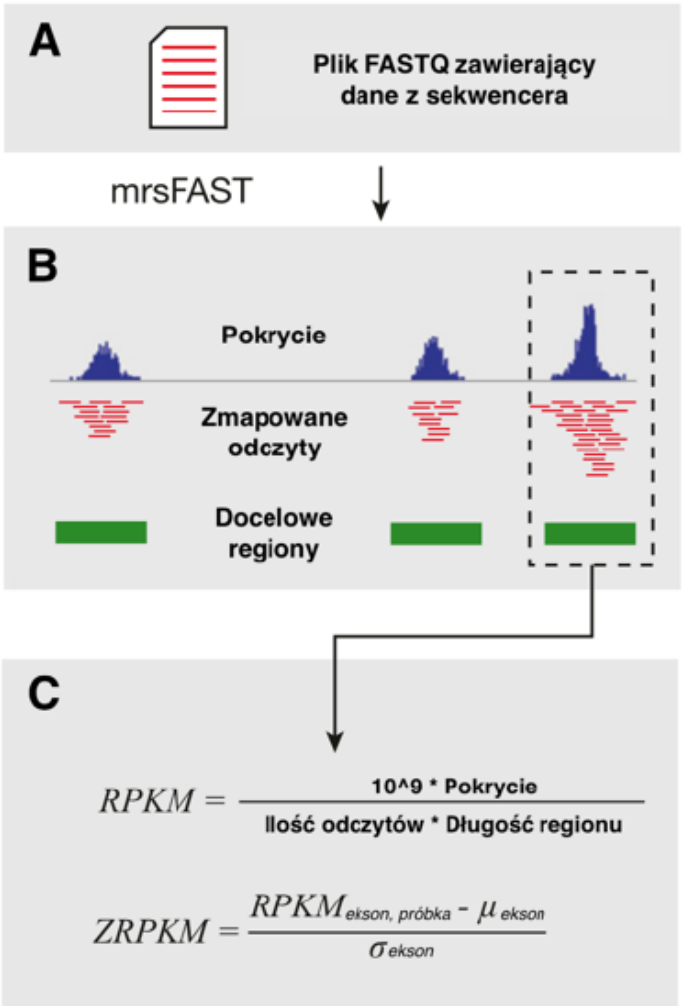
Haendel MA et al., 2015

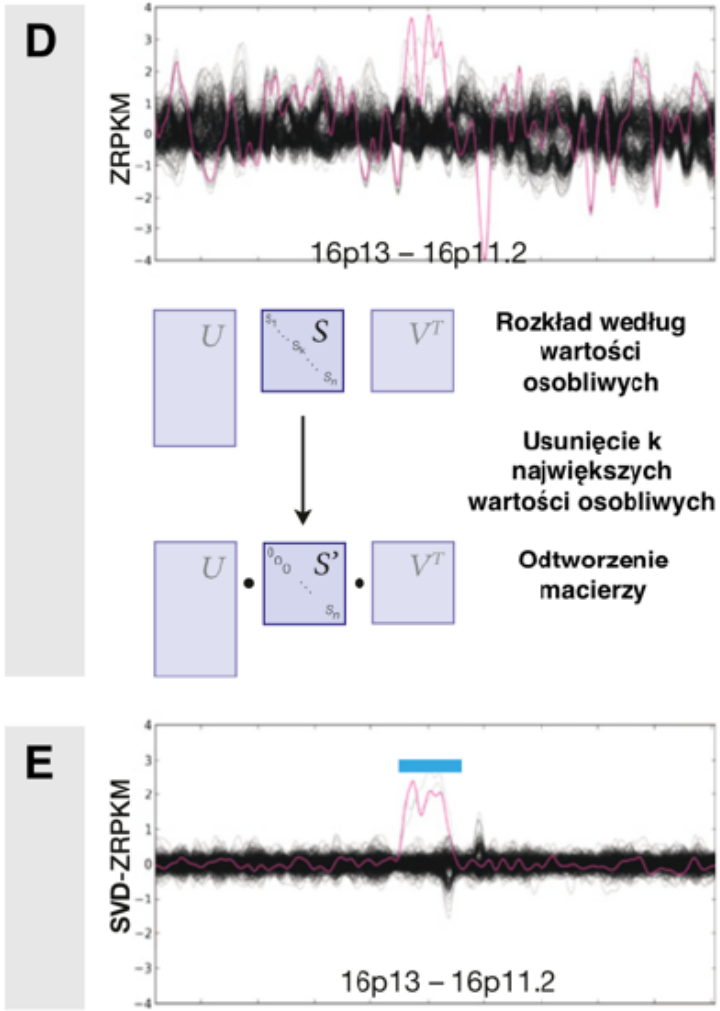
Spis treści

Rodzaje metod do wykrywania zmian liczby kopii, translokacji, inwersji



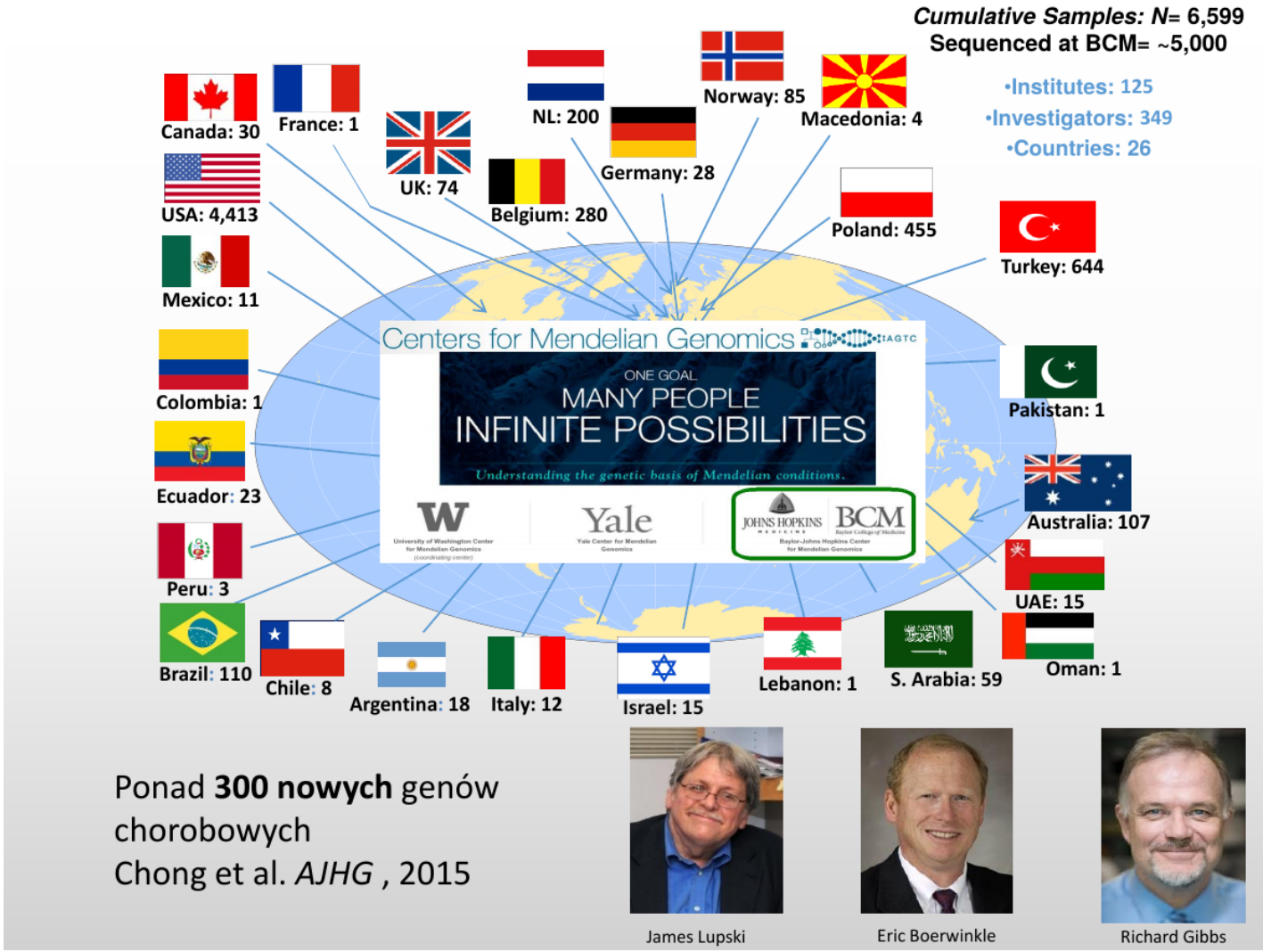
Przykład algorytmu wykorzystującego analizę głębokości pokrycia





Spis treści

Centers for Mendelian Genomics



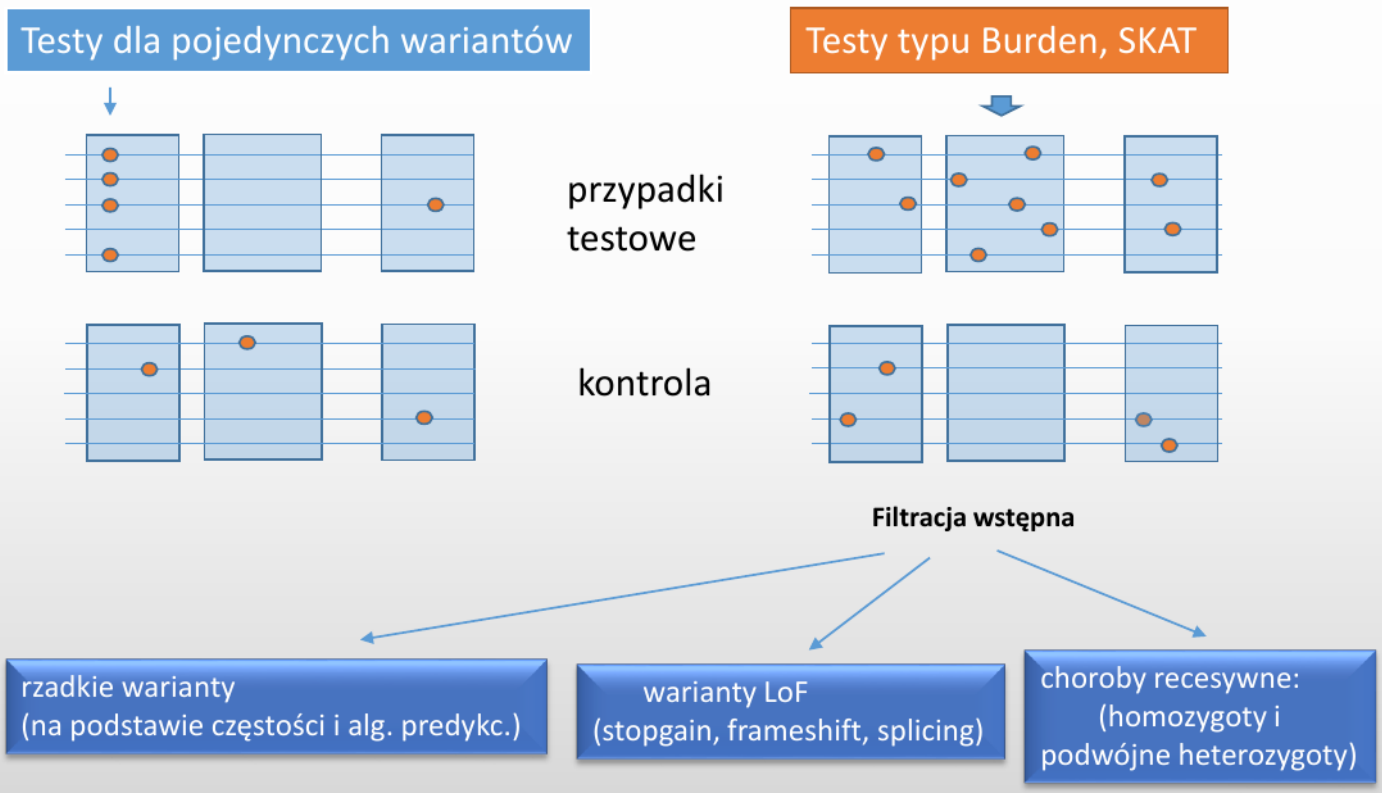
Ponad **300 nowych** genów chorobowych
Chong et al. *AJHG* , 2015

James Lupski

Eric Boerwinkle

Richard Gibbs

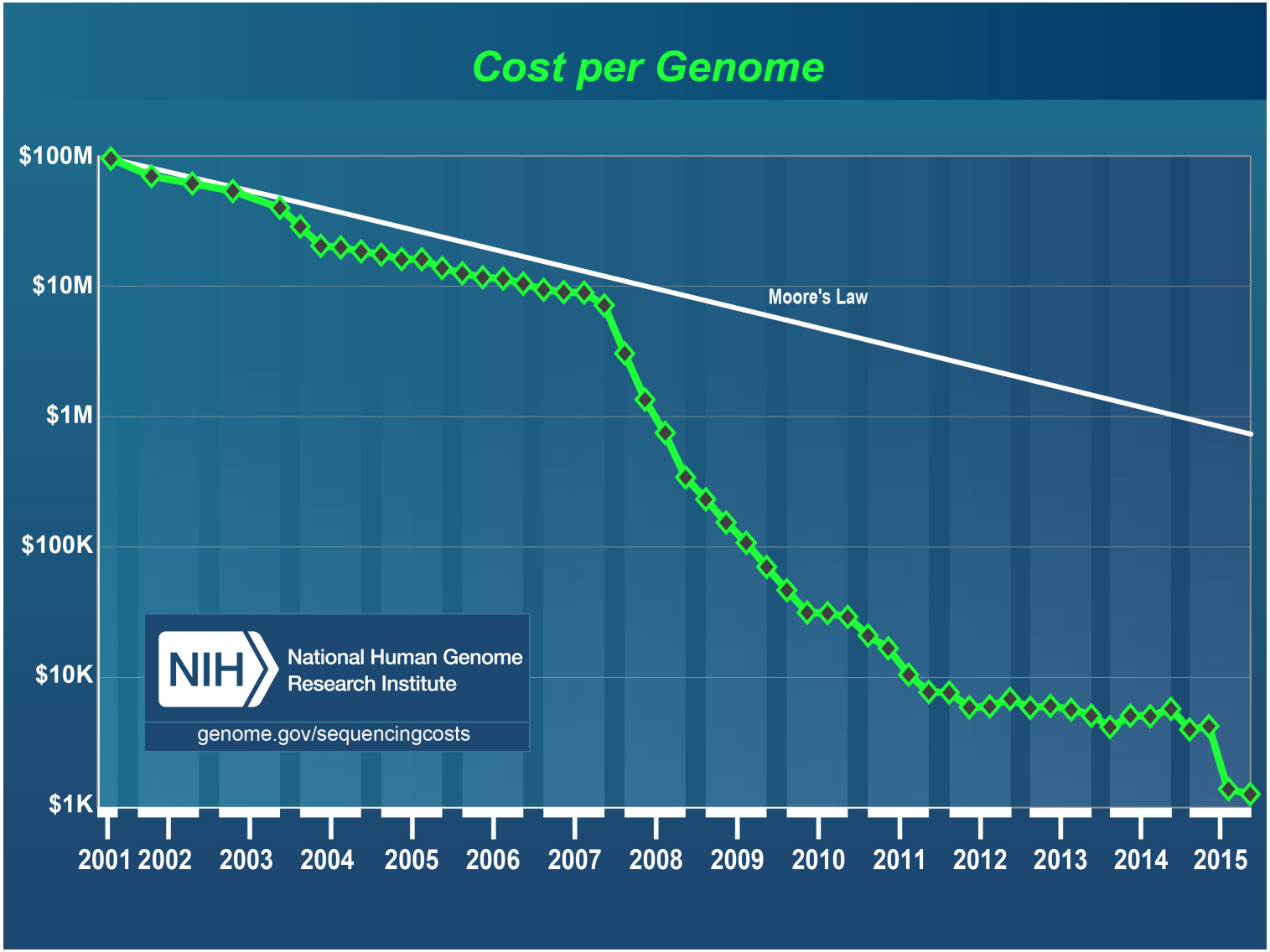
Testy asocjacyjne rzadkich wariantów



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

Spis treści

Koszt sekwencjonowania



Genomika jako problem Big Data

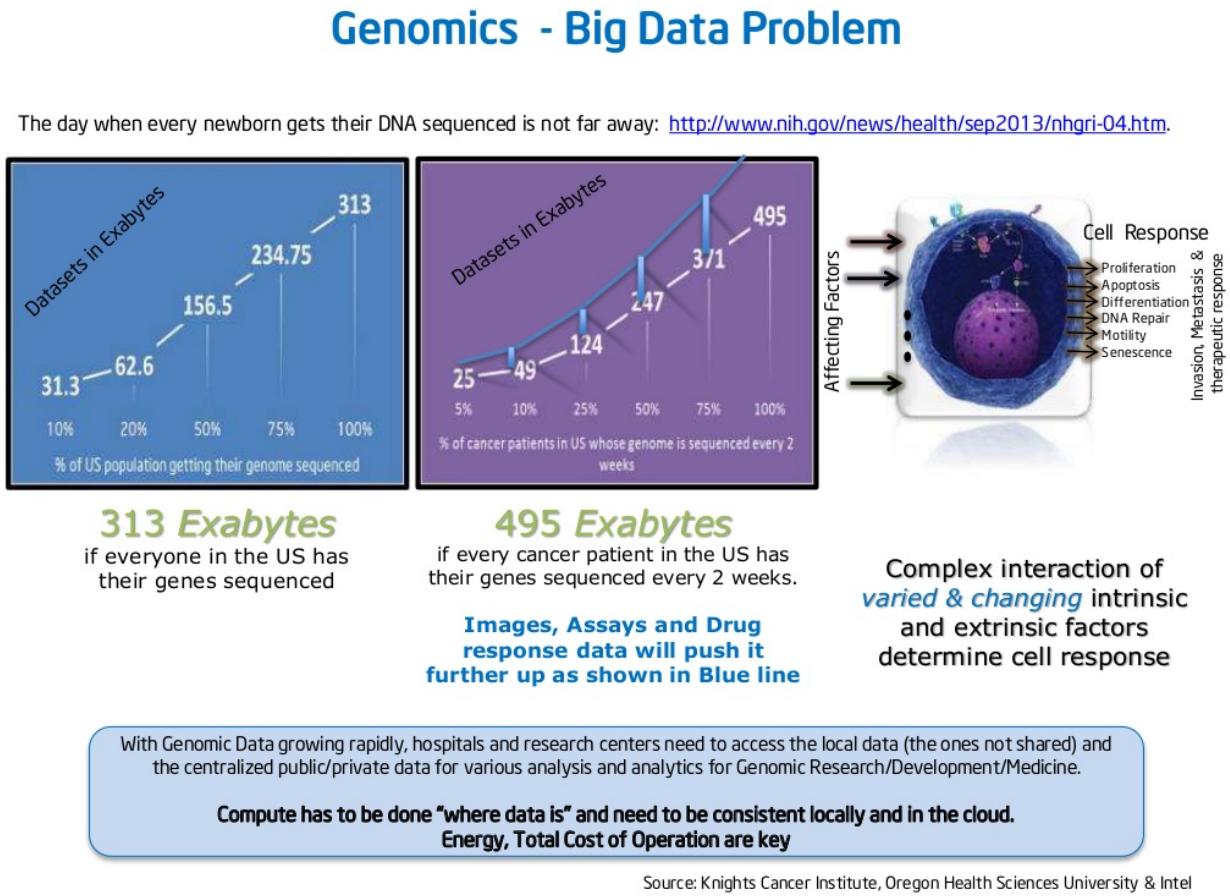
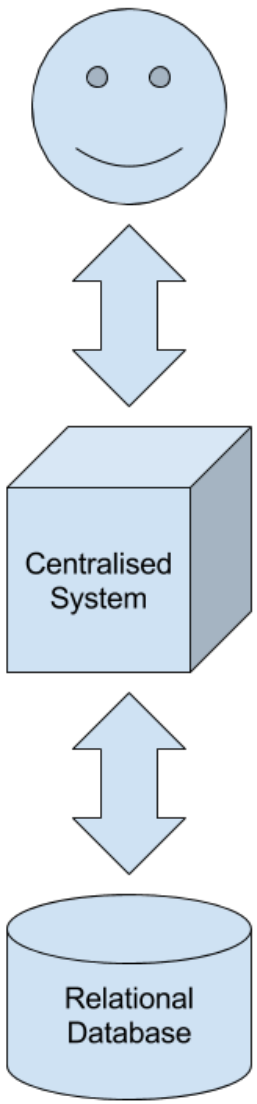
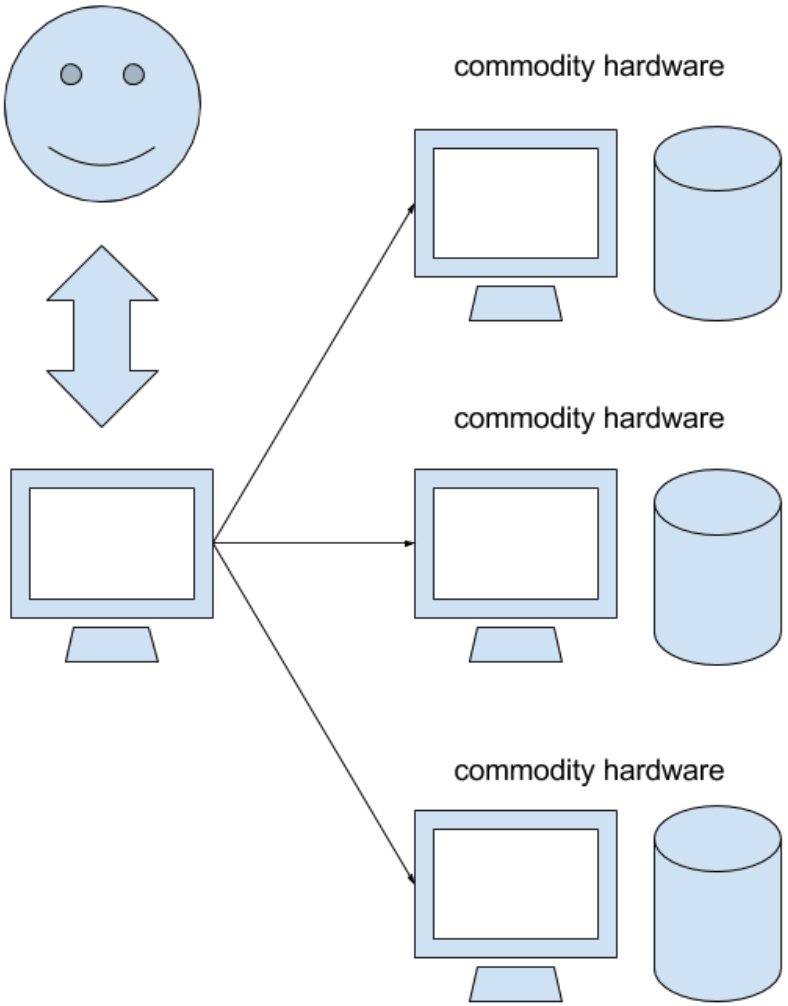


Figure: Źródło: Knights Cancer Institute, Oregon Health Sciences University & Intel

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 ...

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