Analiza genomu człowieka przy wykorzystaniu NGS w kontekście diagnostyki medycznej

dr inż. Tomasz Gambin [1,2]

[1]Instytut Informatyki, Politechnika Warszawska

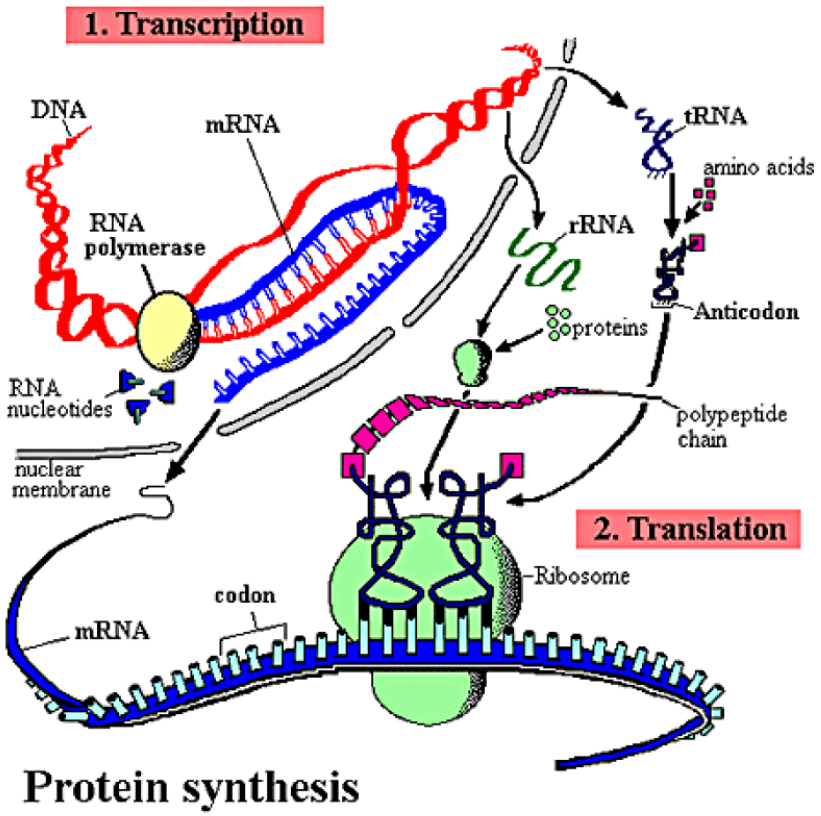
[2]Zakład Genetyki Medycznej, Instytut Matki i Dziecka w Warszawie

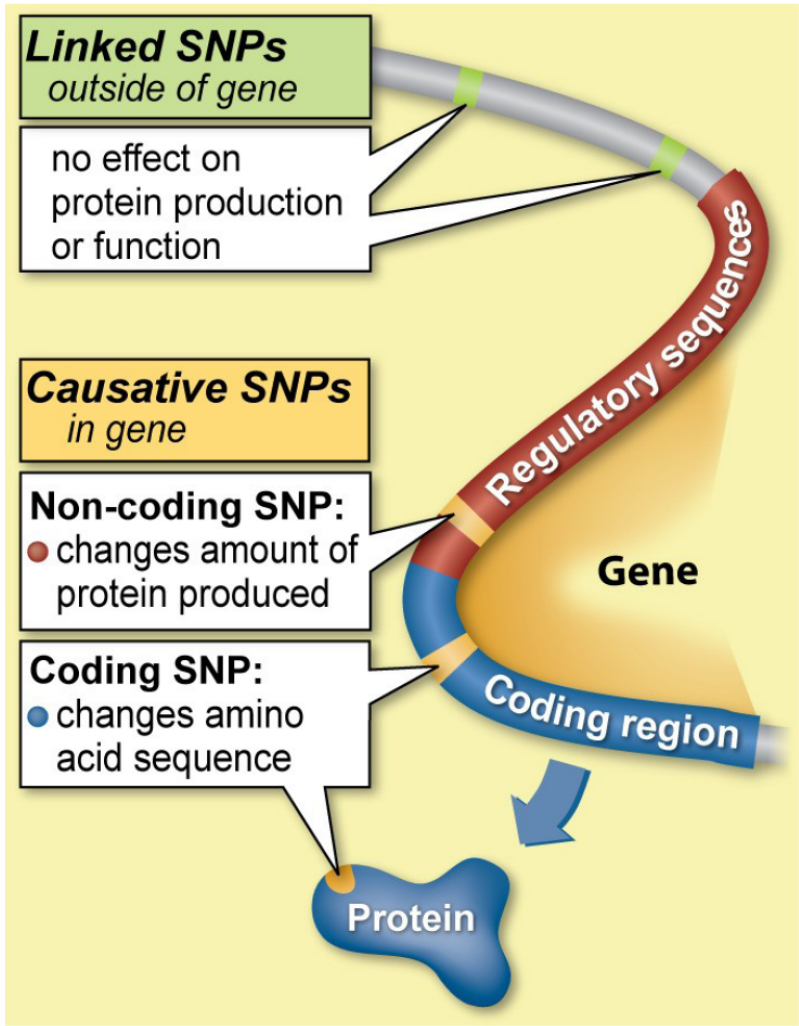
Spis treści

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

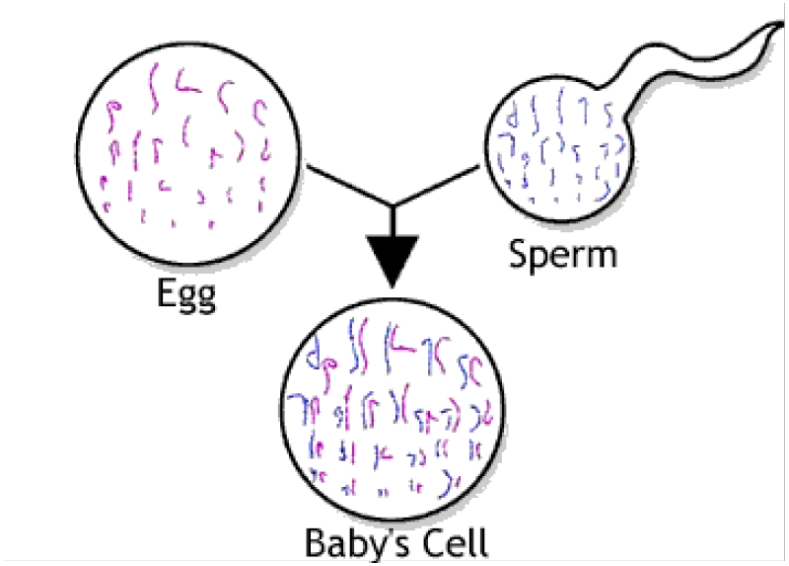
Czemu warianty są istotne?





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***.



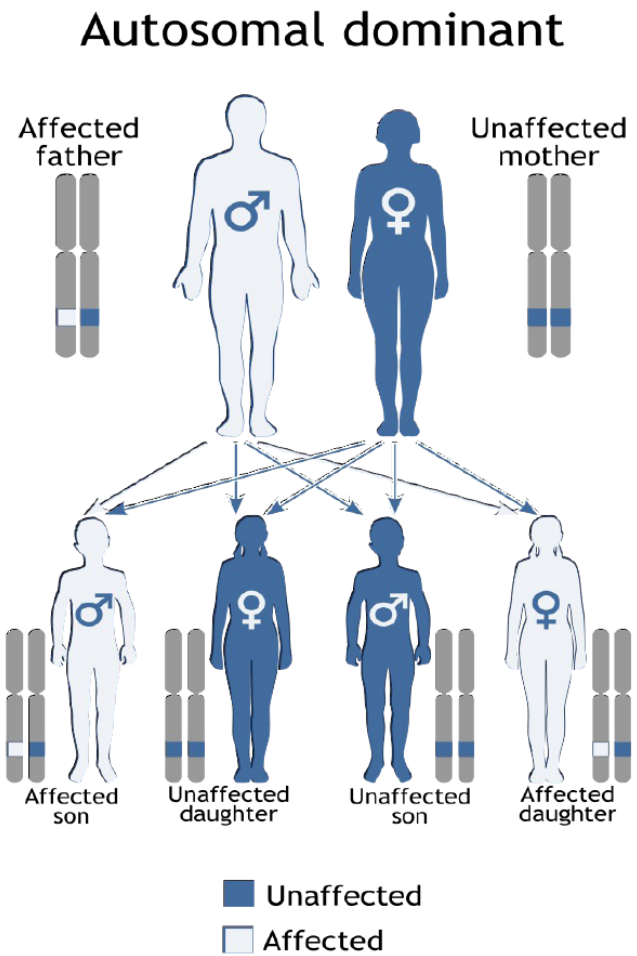
Choroby Mendlowskie, modele dziedziczenia

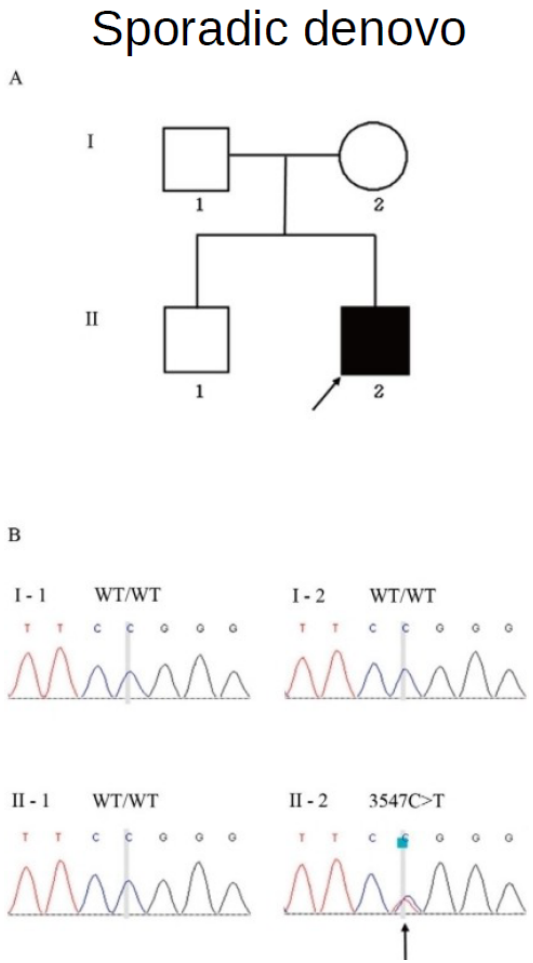
Choroby Mendlowskie

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

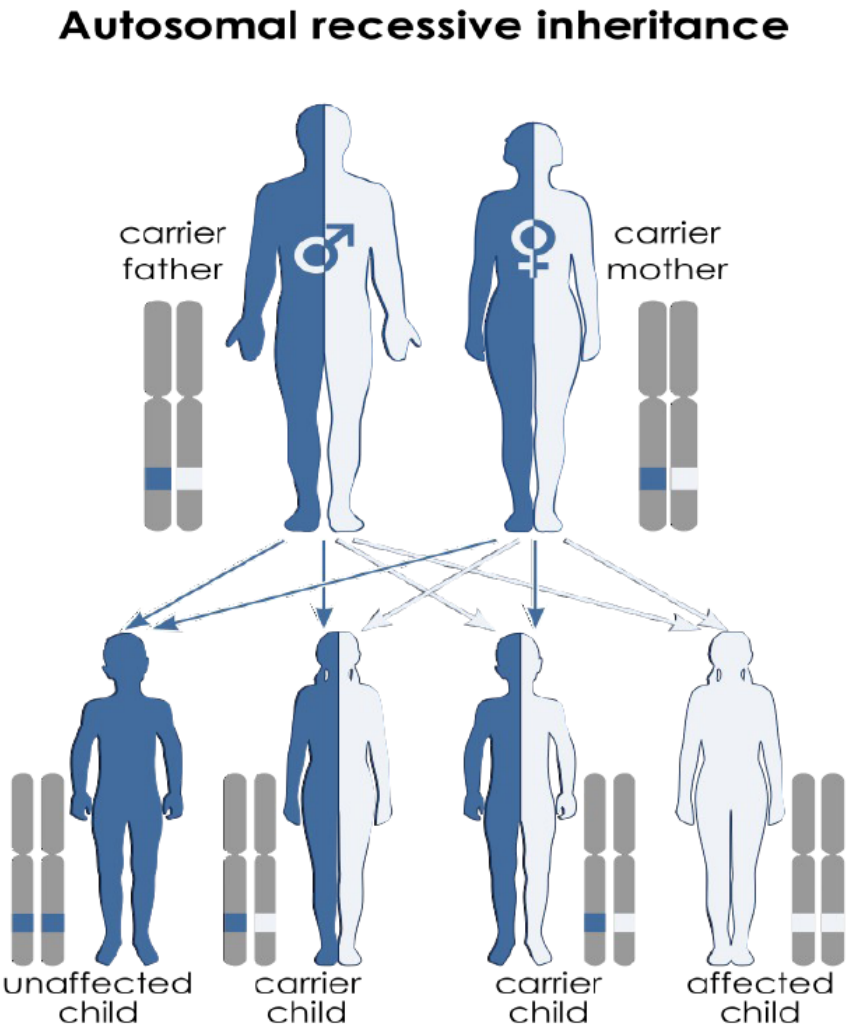
- Choroby autosomalnie dominujące
- Choroby autosomalnie recesywne
- Choroby sprzężone z chromosomem X (dominujące i recesywne)

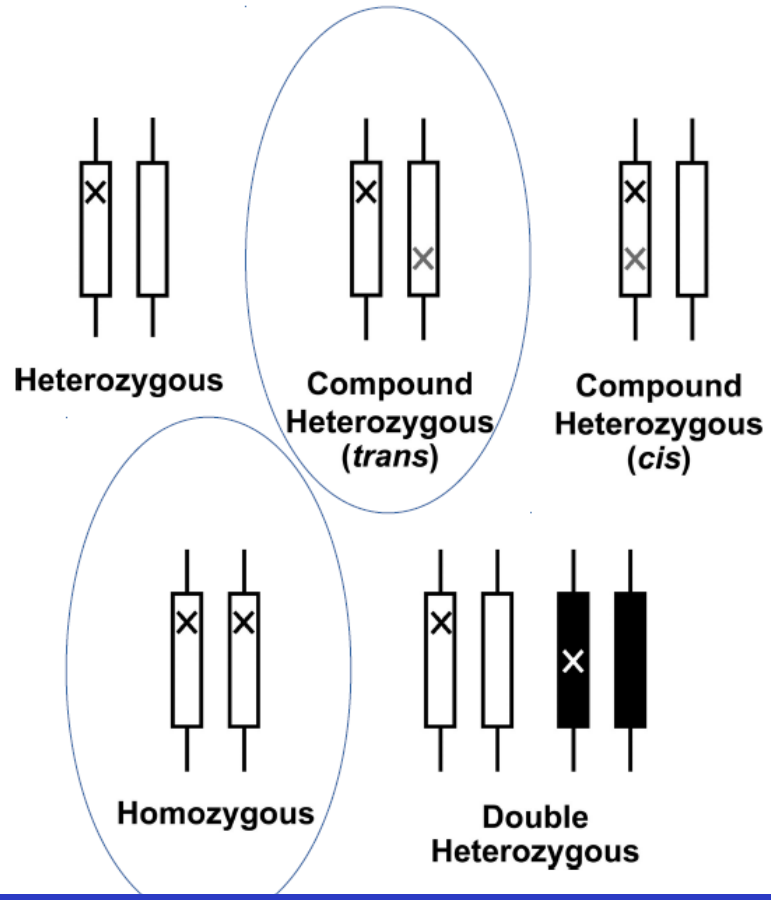
Choroby autosomalnie dominujące



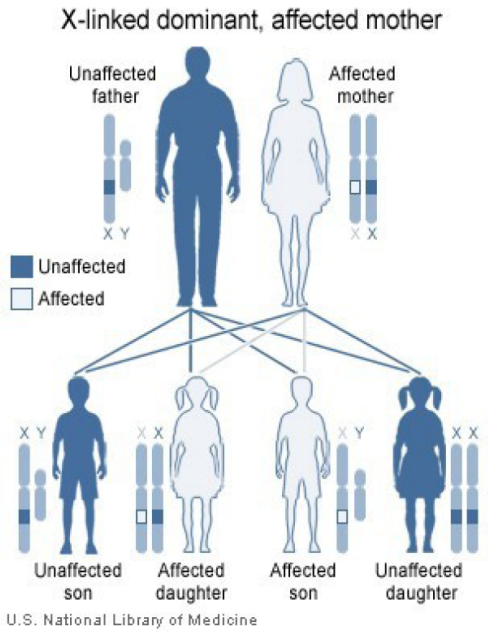


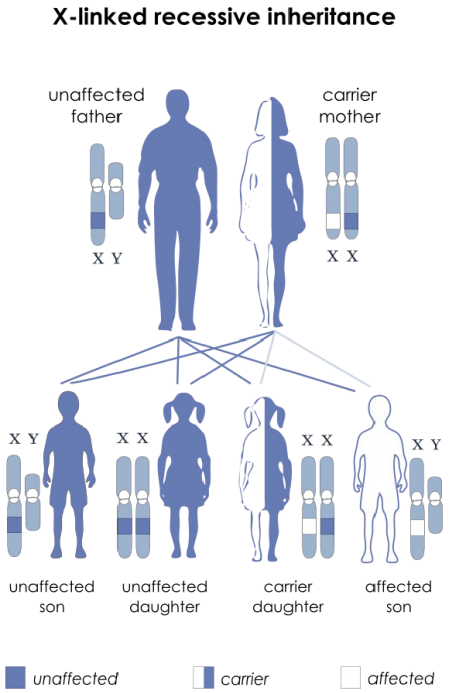
Choroby autosomalnie recesywne





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

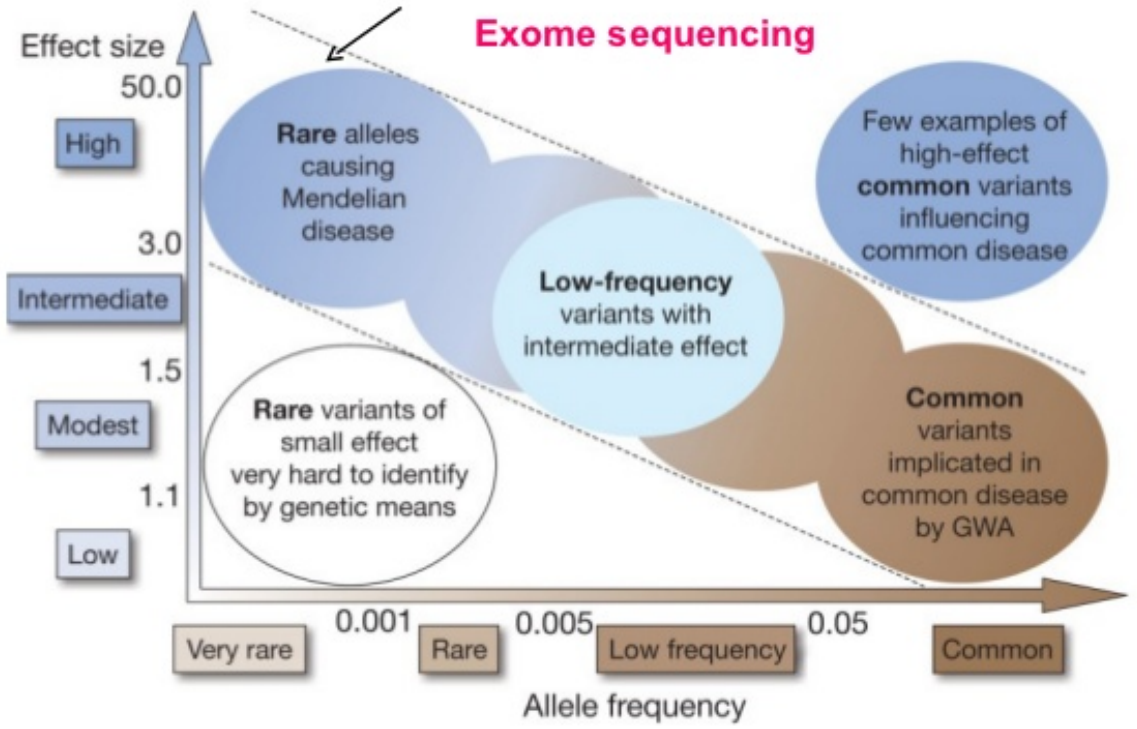
https://pl.wikipedia.org/wiki/Choroby_genetyczne_czlowieka

Spis treści

Przykładowe zastosowania kliniczne NGS

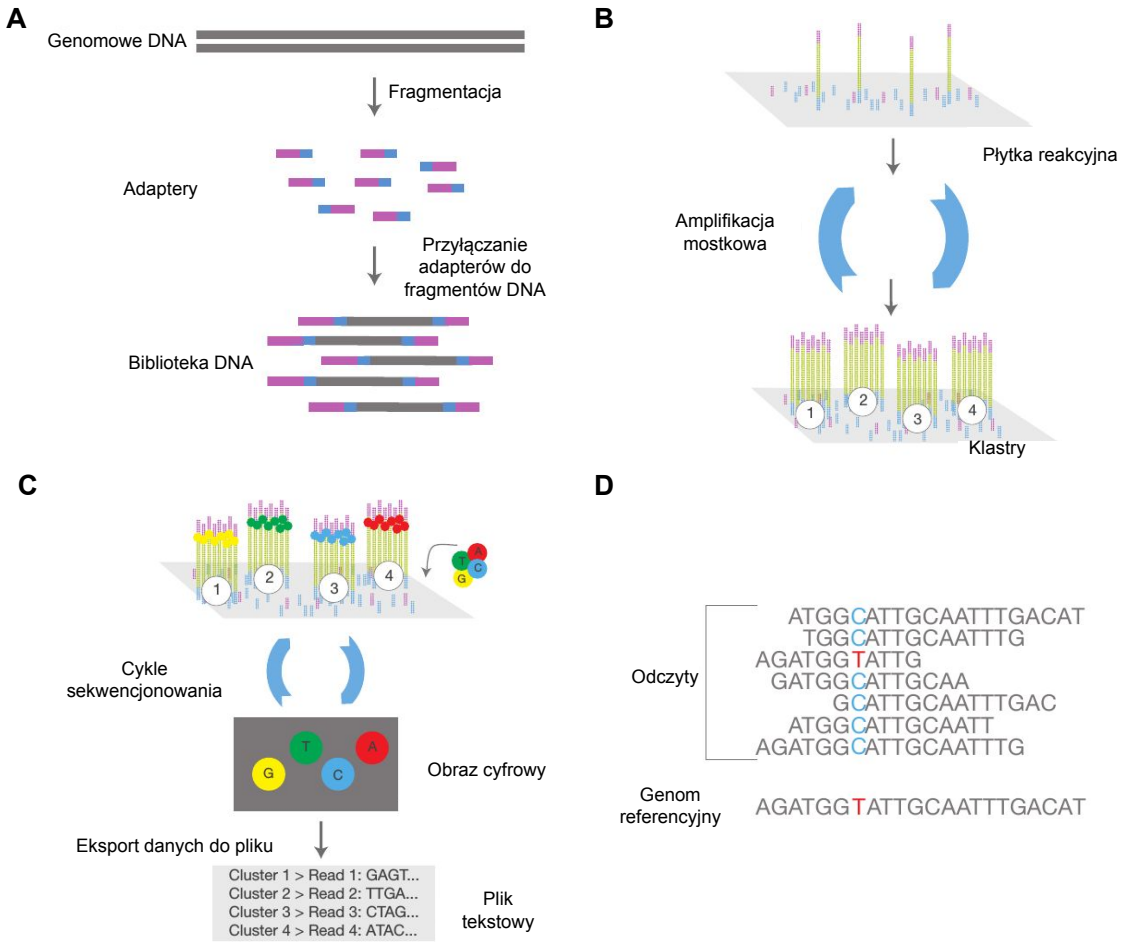
- **diagnostyka chorób rzadkich**;
- diagnostyka chorób uwarunkowanych genetycznie, które przebiegają w sposób nietypowy lub bezobjawowy;
- personalizowana terapia medyczna;
- badania przesiewowe (ang. *screening*) – w szczególności *new born screening*;
- identyfikacja markerów nowotworowych (*liquid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

Dlaczego potrzebujemy informacji o wszystkich wariantach?



Manolio TA, et al (2009). Finding the missing heritability of complex diseases. Nature., 461(7265), 747

(Re)sekwencjonowanie w technologii SBS



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

Sekwencjonowanie całogenomowe vs. celowane

Całogenomowe

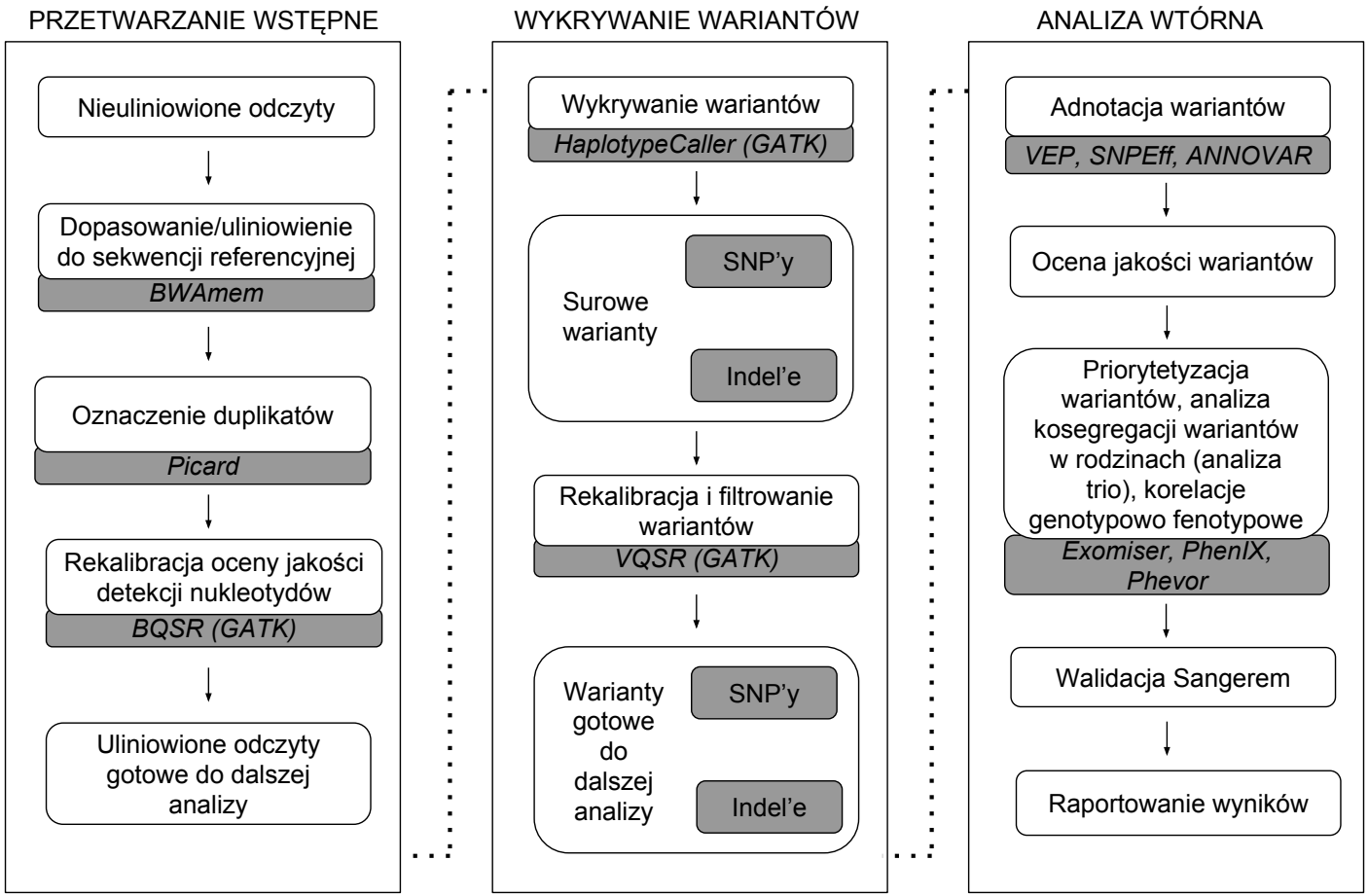
- Stosunkowo równomierne pokrycie odczytami całego genomu
- Łatwiejsza identyfikacja zmian strukturalnych
- Wysoki koszt
- Duże wolumeny danych (100Gb/ pacjent)
- Względnie mniejesze pokrycie niż w przypadku sekw. celowanego (mozaiki)

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)

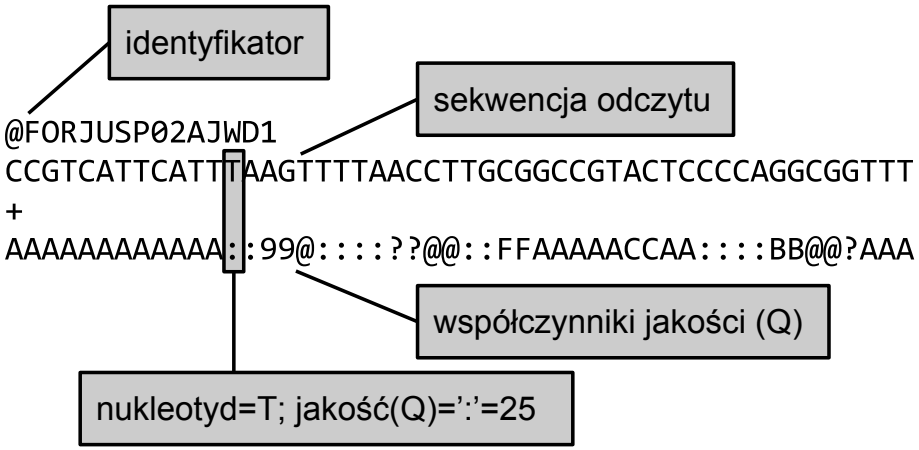
Spis treści

Tok przetwarzania



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

Surowe sekwencje odczytów – pliki FASTQ

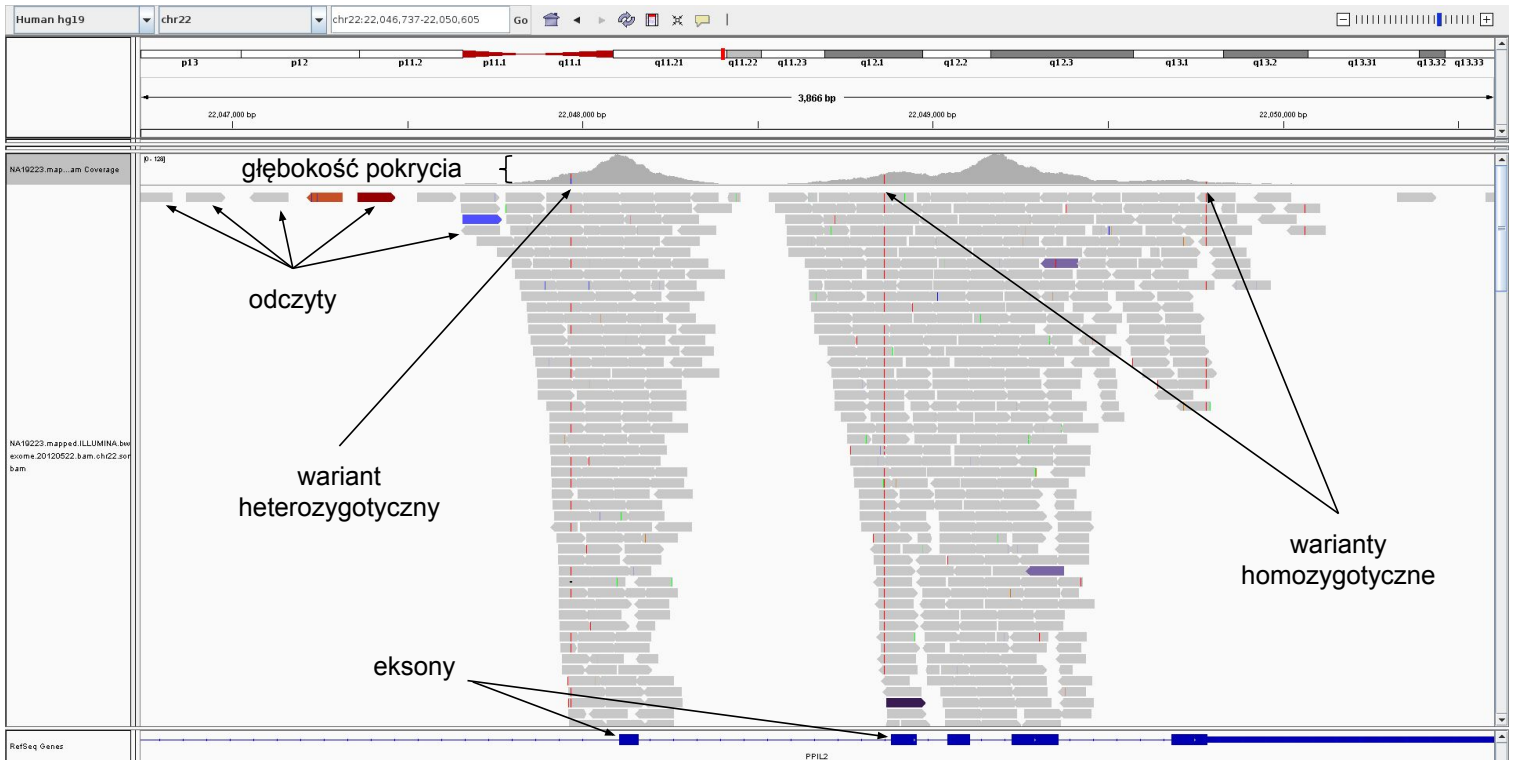


Współczynnik Q (Phred Quality Score)

$Q = -10 * log_{10}(P)$, gdzie P oznacza prawdopodobieństwo, że nukleotyd został zidentyfikowany niepoprawnie; np.:

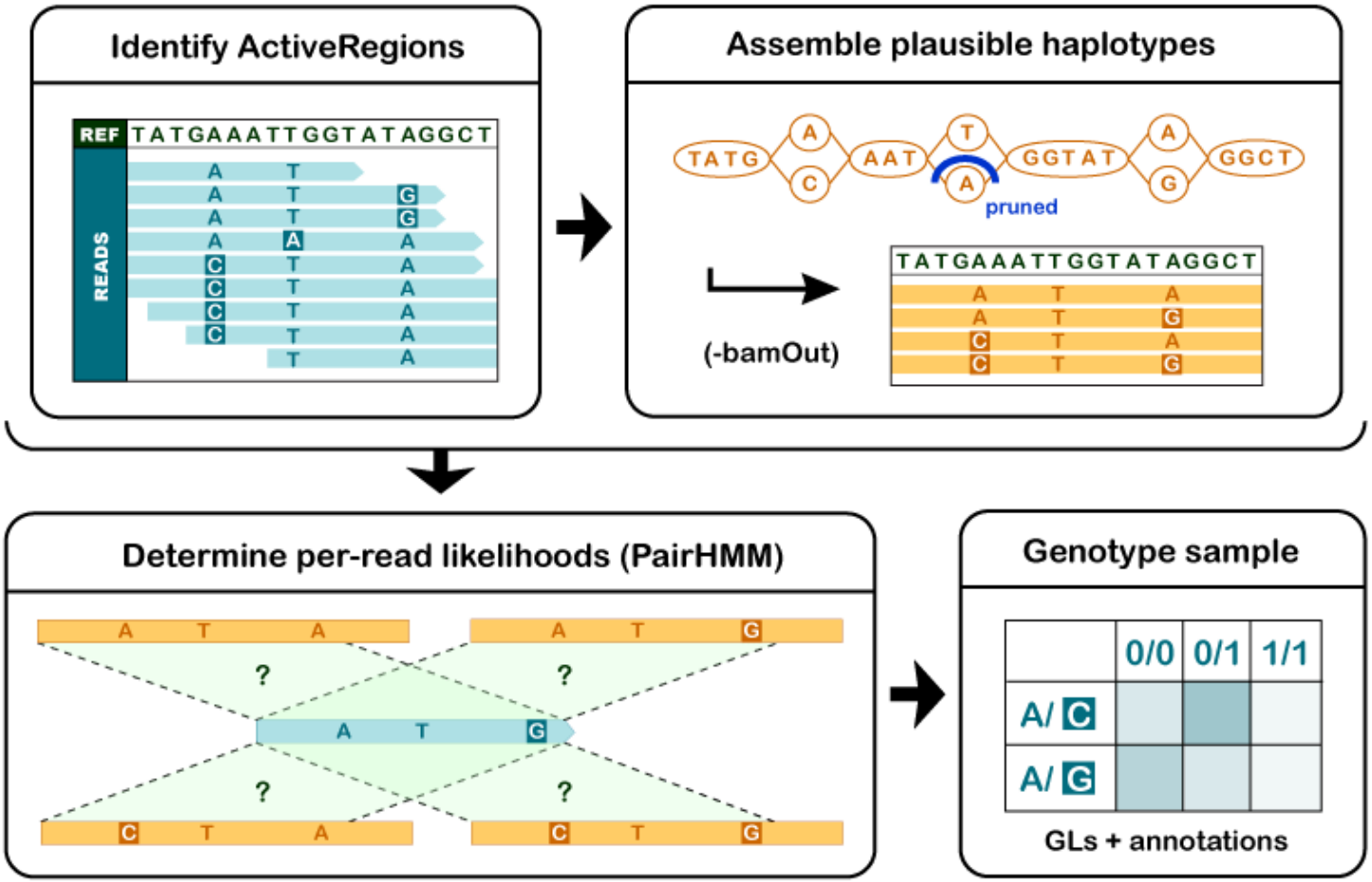
- $Q = 10$ – dokładność 90%,
- $Q = 20$ – dokładność 99%,
- $Q = 30$ – dokładność 99,9%, itd.

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/

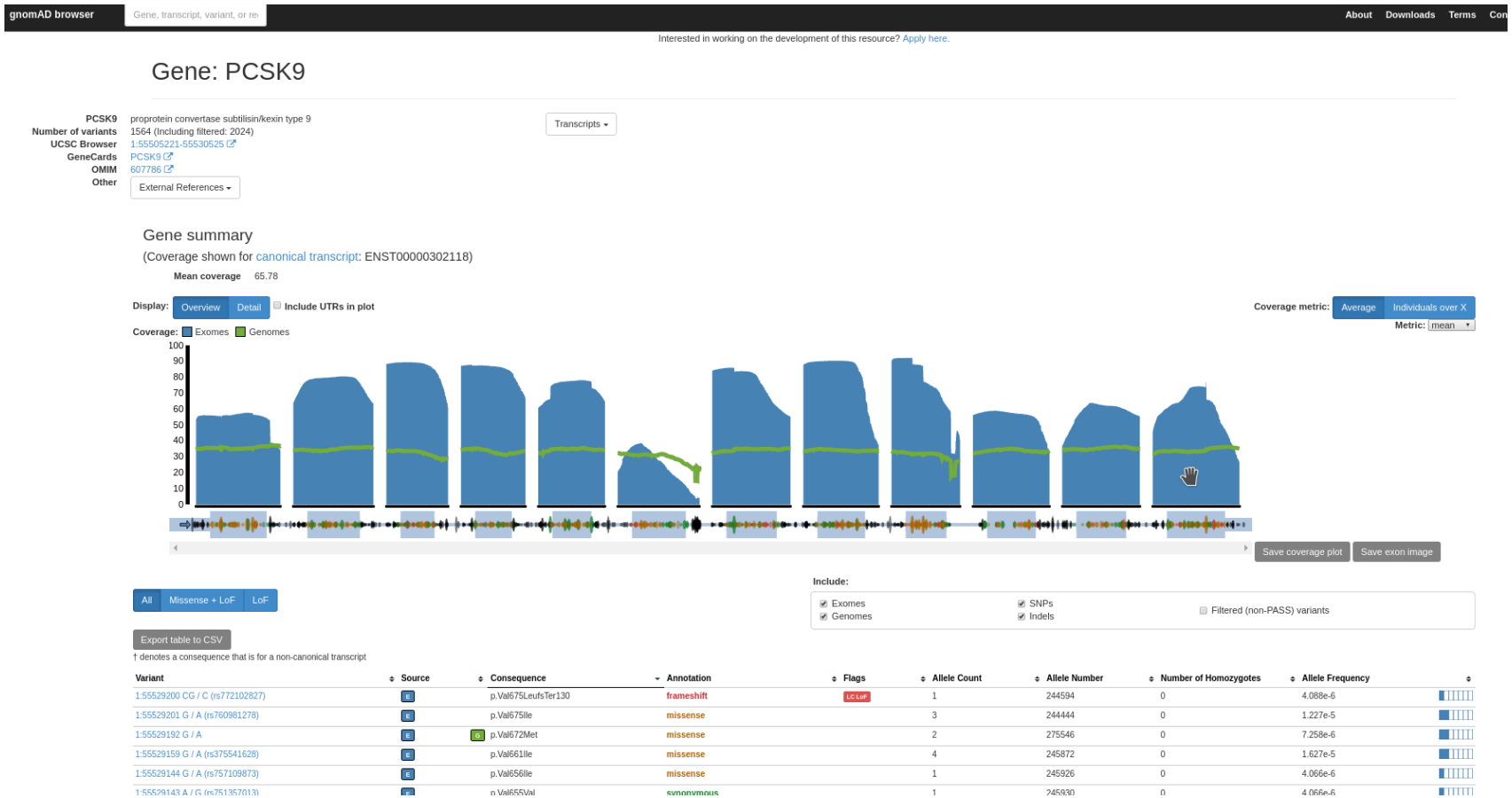
Lista wariantów - pliki VCF

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

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

Genome Aggregation Database (gnomAD)



- Zawiera informację o częstości wariantów z 123,136 sekwencji całoeksomowych oraz 15,496 sekwencji całogenomowych.

Kontrola jakości