

Analiza wyników badań NGS w praktyce

dr hab. Agnieszka Charzewska

Zakład Genetyki Medycznej

Instytut Matki i Dziecka w Warszawie

*KURS: Warsztaty z analizy wyników cytogenetycznych, cytogenetyki molekularnej oraz molekularnych
z uwzględnieniem sekwencjonowania następnej generacji*

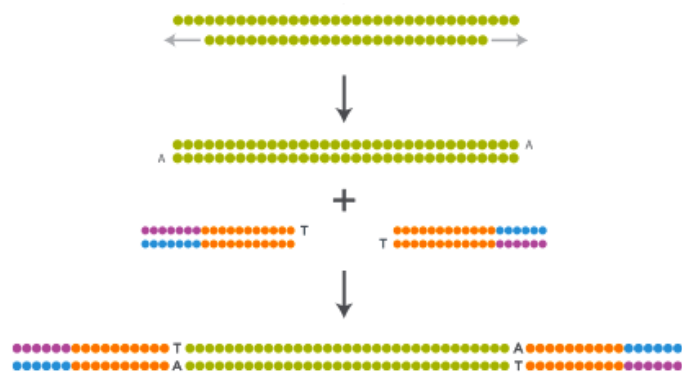
Warszawa, 16.04.2021

Plan prezentacji

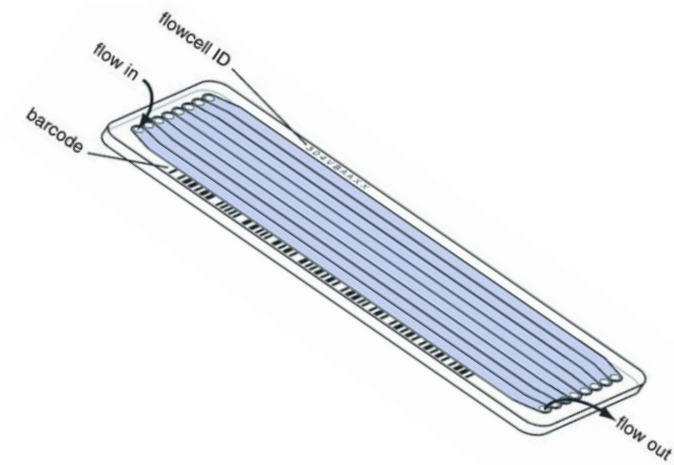
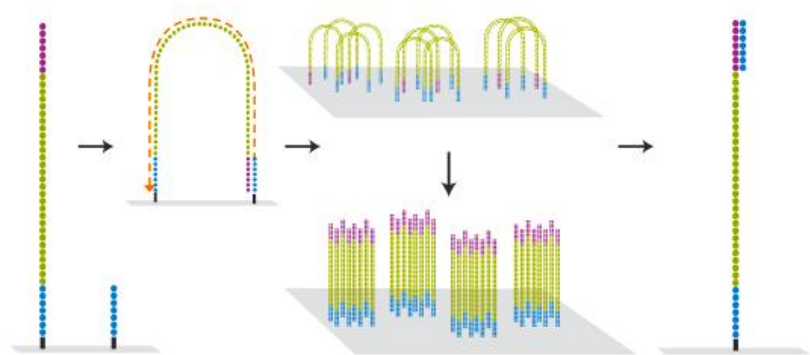
1. NGS (next generation sequencing)
2. Etapy analizy danych NGS – potok przetwarzania danych, identyfikacja wariantów
3. Przeszukiwanie baz danych wariantów
4. Predykcje bioinformatyczne
5. Modele dziedziczenia na przykładach
6. Praktyczna analiza danych NGS

Etapy NGS

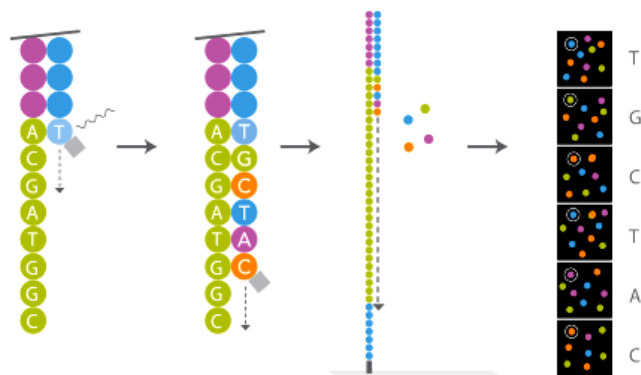
1. Przygotowanie biblioteki



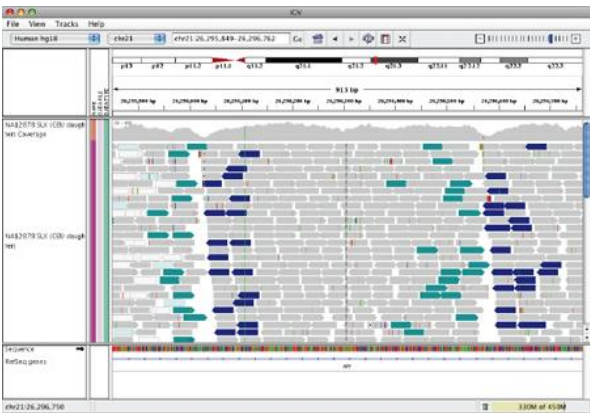
2. Tworzenie klastrów



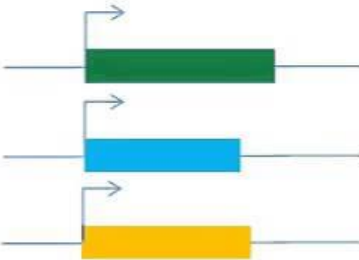
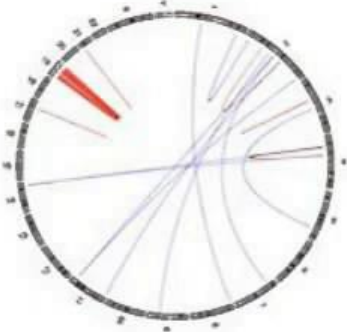
3. Sekwencjonowanie



4. Analiza danych



Rodzaje genomowych badań NGS

	PANEL	EKSOM	GENOM
	40 - 400 genów	25,000 genów	kodujące - niekodujące DNA
			
	Wysokie pokrycie	Średnie pokrycie	Niskie pokrycie
			
koszt	+	++	++++
Ilość danych	+	+++	++++
CNV	+/-	+/-	+/(strukturalne)
nowe geny?	-	+	+

Analiza bioinformatyczna – potok przetwarzania danych (pipeline)



Surowe sekwencje odczytów - FASTQ

- Wszystkie sekwencje odczytywane przez sekwenator zawarte są w pliku FASTQ wraz z parametrami jakości

```
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:ATCACG
TGC GGAAGGATCAT TGTGGAATTCTCGGGTGCCAAGGA ACTCCAGTCACATCACGA
TCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAAAAAATTA
B@CFFFFFFHFFHJIIGHIHIJJIIJJGDCHIIJJJJJJJJGJGIHHEH@)=F@
EIGHHEHFFFFFFDCBBD:@CC@C:<CDDDD50559<B#####
```

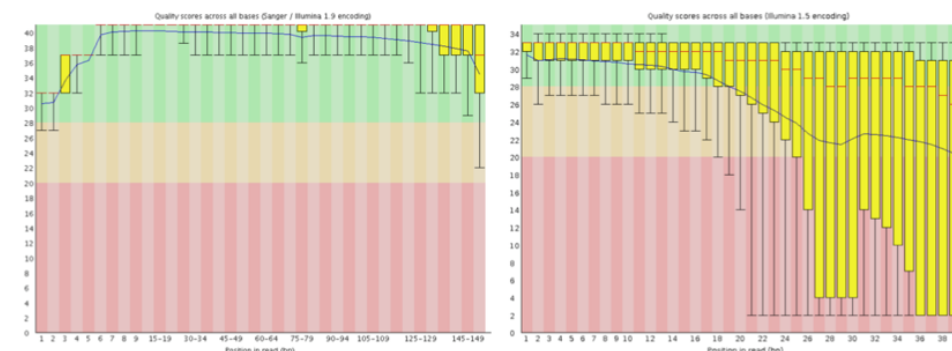
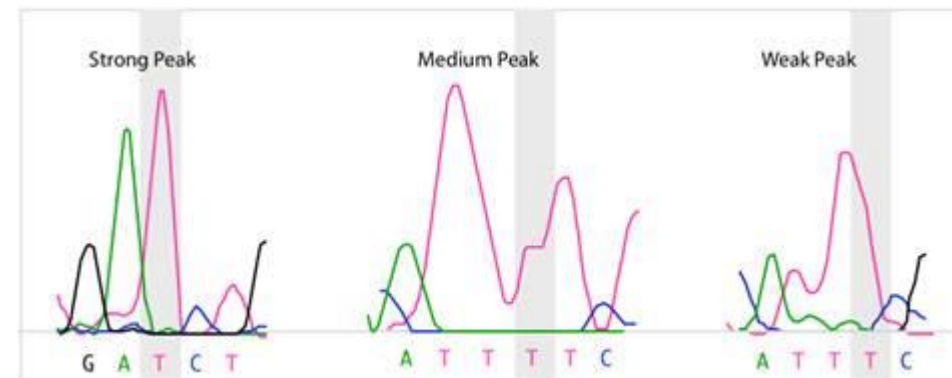
ID maszyny i numer runu, ID flowcell i ścieżka, numer obszaru w obrębie ścieżki, koordynat „x” w obrębie obszaru, koordynat „y” w obrębie obszaru, numer odczytu z pary, czy odczyt przeszedł filtr jakości, sekwencja nukleotydowa indeksu

Jakość odczytu w postaci „Phred Quality Score” (Q) zakodowana w standardzie ASCII

Phred Quality Score (Q)

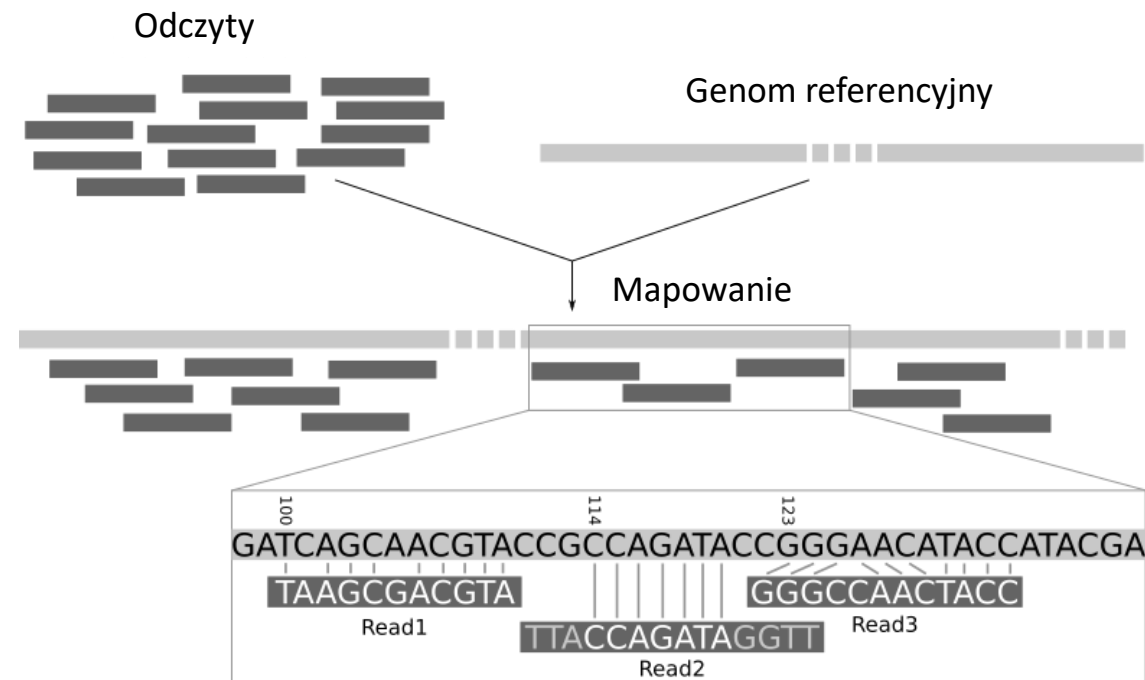
$$Q = -10 \times \log_{10} P$$

- P oznacza prawdopodobieństwo, że nukleotyd został zidentyfikowany niepoprawnie
- Dla
Q = 10, precyzja odczytu nukleotydu – 90%
Q = 20 – 99%
Q = 30 – 99.9%
- Sanger – oceniany na podstawie właściwości pików (m.in. wysokość, szerokość, lokalizacja)
- Illumina – oceniany na podstawie właściwości poszczególnych obszarów w klastrach (m.in. intensywność sygnału, stosunek sygnału do szumu)



Przyrównanie do sekwencji referencyjnej

- Polega na zmapowaniu odczytów z plików FASTQ do sekwencji referencyjnej genomu
- Najczęściej do mapowania wykorzystuje się metodę BWT (Burrows Wheeler Transform)
- Trudności sprawiają:
 - Krótkie odczyty (~100pz)
 - Odczyty mogą mapować się w kilku lokalizacjach genomowych
 - Obecność regionów repetytywnych (stanowią około połowę genomu ludzkiego)



Zmapowane odczyty - SAM/BAM

- Format plików przechowujący zmapowane odczyty (SAM, w skompresowanej formie binarnej – BAM)

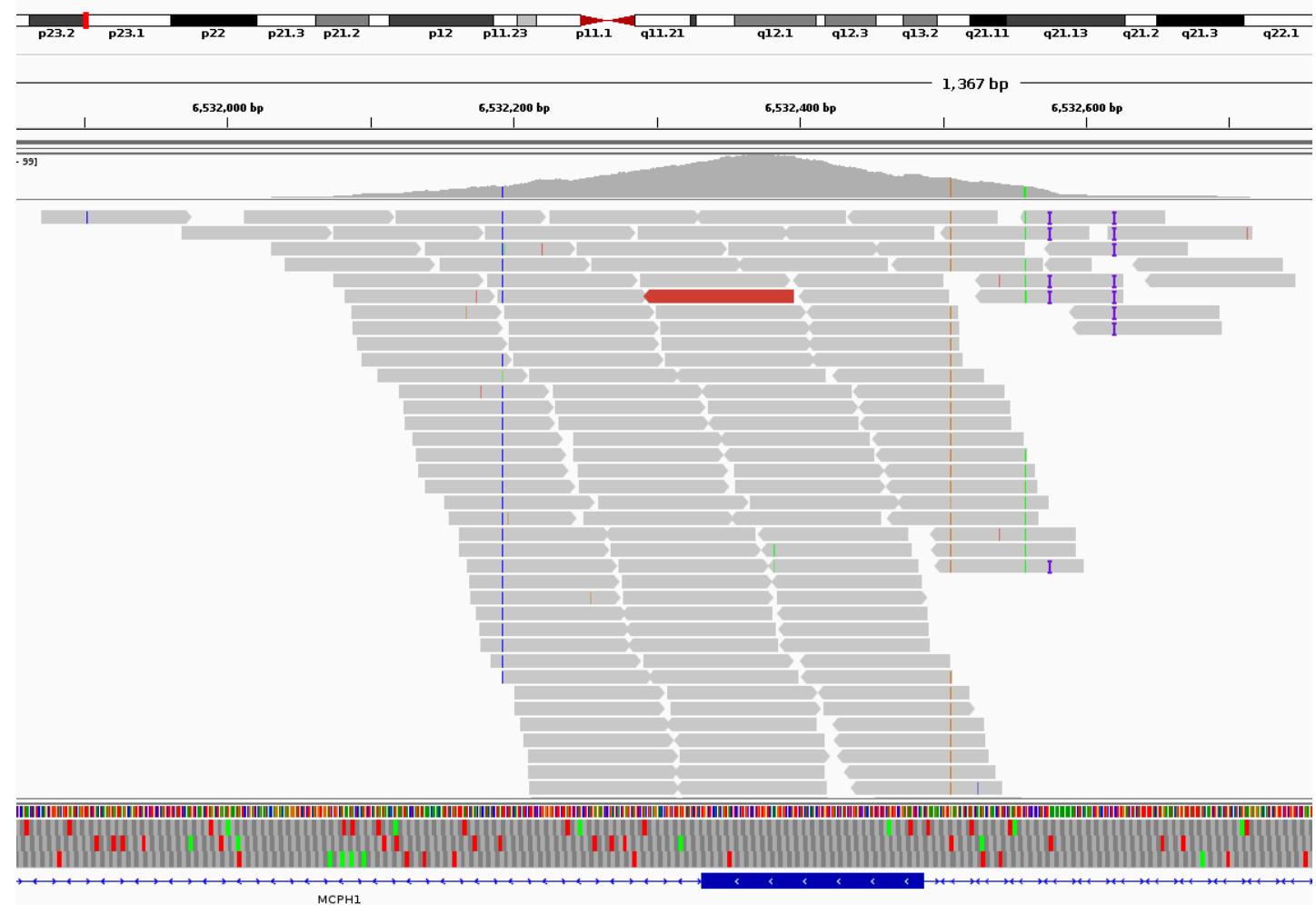
M01121:5:000000000-A2DTN:1:2111:20172:15571 163 chrM
480 60 148M2S 524 195

AATCTCATCAATACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTG
CTAACCCCATACCCCGAACCAACCAAAACCCCAAAGACACCCCCCACAGTTTA
TGTAGCTTACCTCCTCAAAGCAATAACCTGAAAATGTTTAGACGGG

BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEAGGAGFFHAEFDGG?E?EGG
GFGHFGHF?FFCHFHO0E@EGFGGEEE1FFEEHGBGEFFFGGGG@</01BG2
12222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-

Identyfikator, wartość FLAG, nazwa sekwencji referencyjnej, pozycja
pierwszego nukleotydu w sekwencji, jakość mapowania, ciąg znaków CIGAR,
pozycja pierwszego nukleotydu w sekwencji następnego odczytu, długość

Zmapowane odczyty - plik .bam w przeglądarce IGV Genomics Viewer



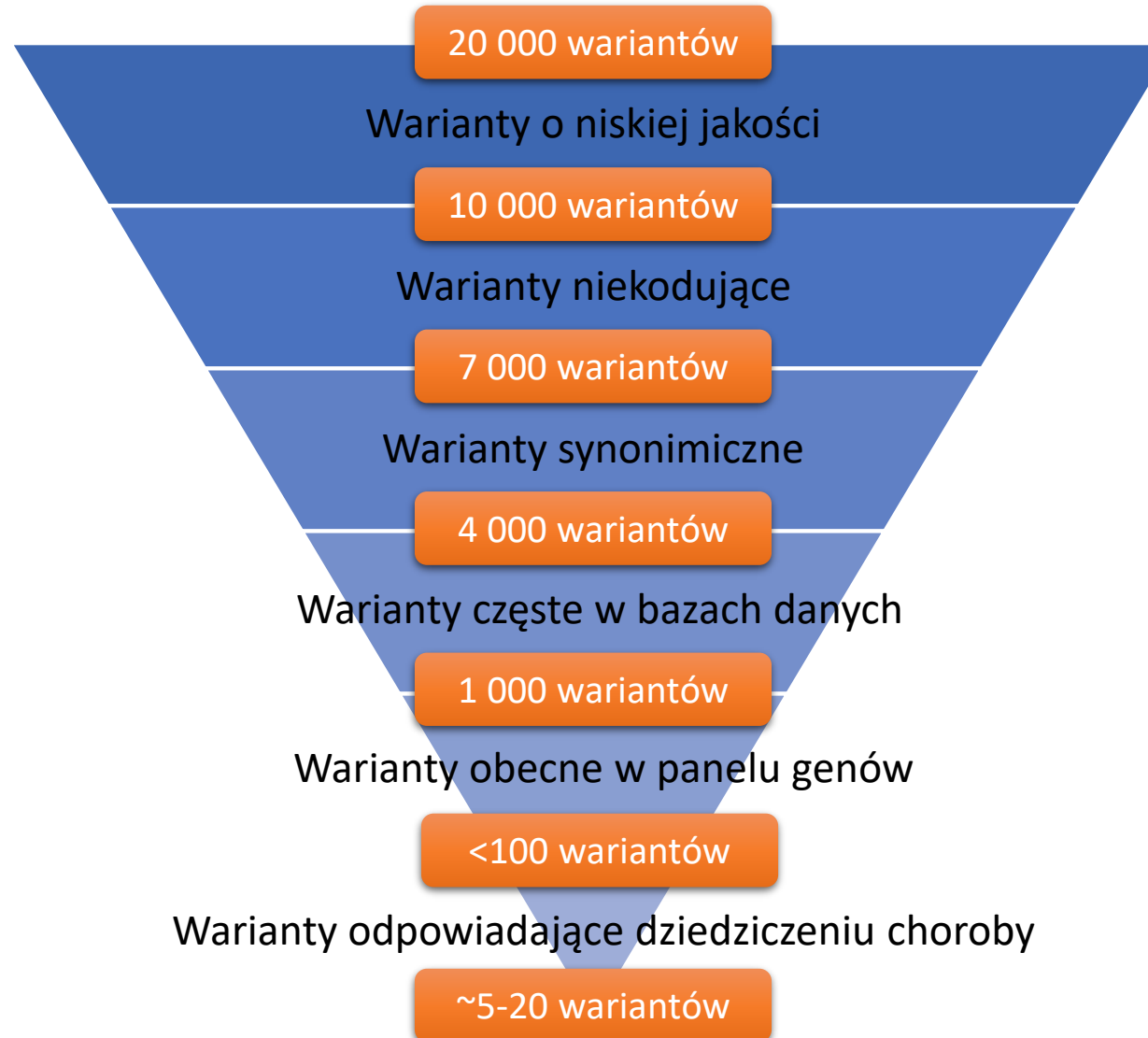
Wykrywanie wariantów - VCF (variant call format)

- Format do przechowywania wykrytych wariantów w sekwencji
- Nagłówek oznaczony znakami ## opisuje pola używane w dalszej części pliku

#CHROM	POS	ID	REF	ALT	QUAL		
20		14370	rs6054257	G	A	29	
20	17330	.	T	A	3		
20	1110696	rs6040355	A	G	67		
20		1230237	.	T	.	47	
20		1234567	.		GTC	GTCT	50

- Adnotacja może być związana z genem (nazwa, nr OMIM), z pozycją wariantu (ekson, intron, sekwencje regulatorowe, niekodujące) lub z samym wariantem (częstość w bazach populacyjnych, patogenność w przypadku znanej zmiany, wyniki algorytmów predykcyjnych)
- Najpopularniejsze programy do adnotacji wariantów: ANNOVAR, Ensembl VEP

Etapy analizy wariantów WES



Sprawdzamy obecność wariantu w bazach danych

(publiczne bazy danych vs. wewnętrzne bazy wariantów)

Bazy populacyjne

gnomAD
Exome Variant Server
dbSNP
dbVar

Częstość populacyjna
(osoby zdrowe) – czy zmiana jest
częsta w danej populacji ?

Bazy kliniczne

OMIM
Clinvar
HGMD
LOVD
Decipher

Czy wariant jest zarejestrowany
w bazach klinicznych (osoby chore)
i jaki ma status patogenności?

Wewnętrzne bazy danych wariantów

Czy wariant występuje
z dużą częstością
w bazie wewnętrznej
i wynika z błędu
maszynowego ?

Lista genów powiązanych z fenotypem

[PanelApp](#) [Panels](#) [Genes and Entities](#) [Activity](#) [Log in](#) [Register](#)

Genomics England PanelApp

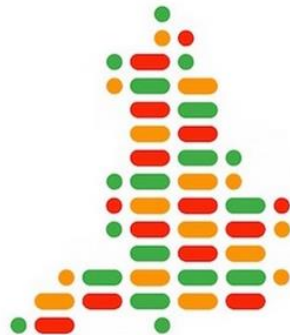
A crowdsourcing tool to allow gene panels to be shared, downloaded, viewed and evaluated by the Scientific Community

[Home](#)
[News](#)
[Navigate & Explore](#)
[Reviewers](#)
[Guidelines](#)
[API](#)
[FAQs](#)
[Contact, Content & Glossary](#)



Genomics
england

PanelApp



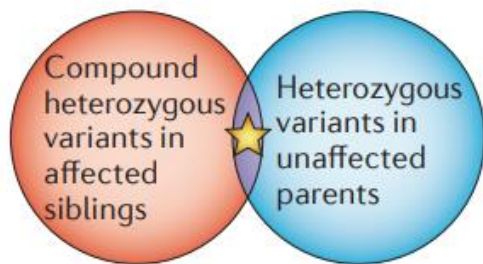
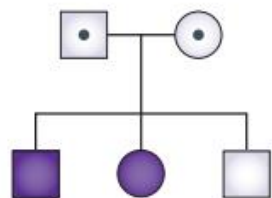
<https://panelapp.genomicsengland.co.uk/>

Algorytmy predykcyjne *in silico*

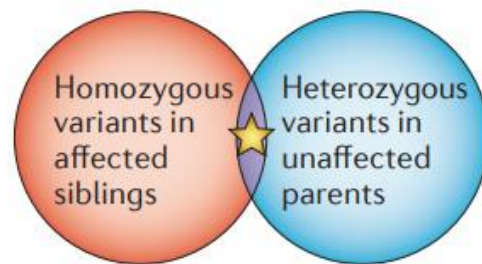
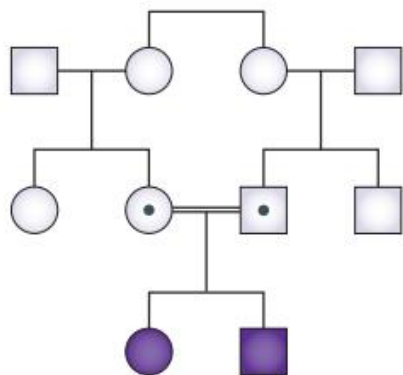
- Align-GVGD – różnice biochemiczne białka
- **SIFT** – konserwacja sekwencji białkowej
- **PolyPhen** – konserwacja sekwencji białkowej,
- **MutationTaster** – konserwacja sekwencji DNA i białka,
- MutationAssessor – konserwacja sekwencji białkowej
- FATHMM – homologia sekwencji
- Radial SVM – uczenie maszynowe i regresja logistyczna
- LR – regresja logistyczna
- VEST3 – uczenie maszynowe
- **CADD** – uczenie maszynowe
- GERP++ – konserwacja sekwencji DNA
- LRT – test ilorazu wiarygodności

Ustalamy model dziedziczenia

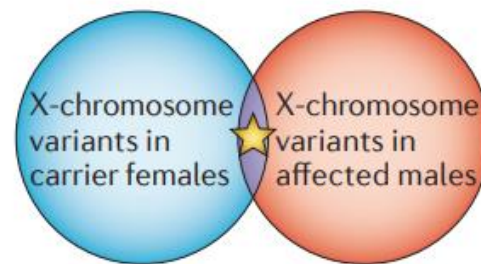
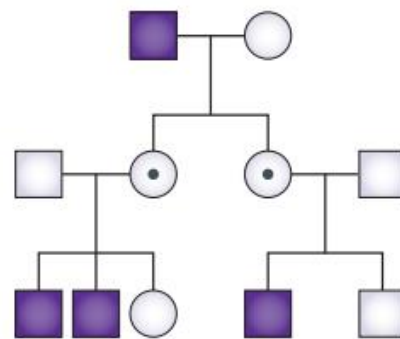
Autosomal recessive



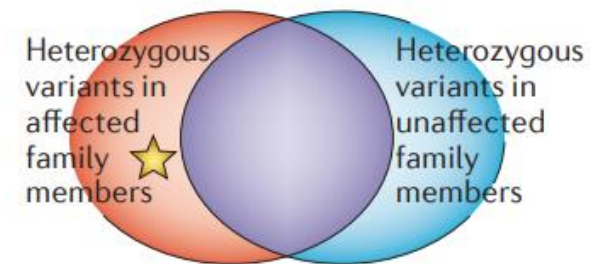
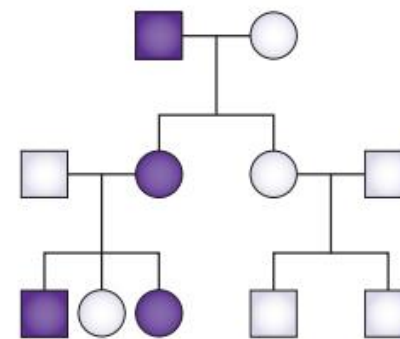
Consanguineous autosomal recessive



X-linked recessive

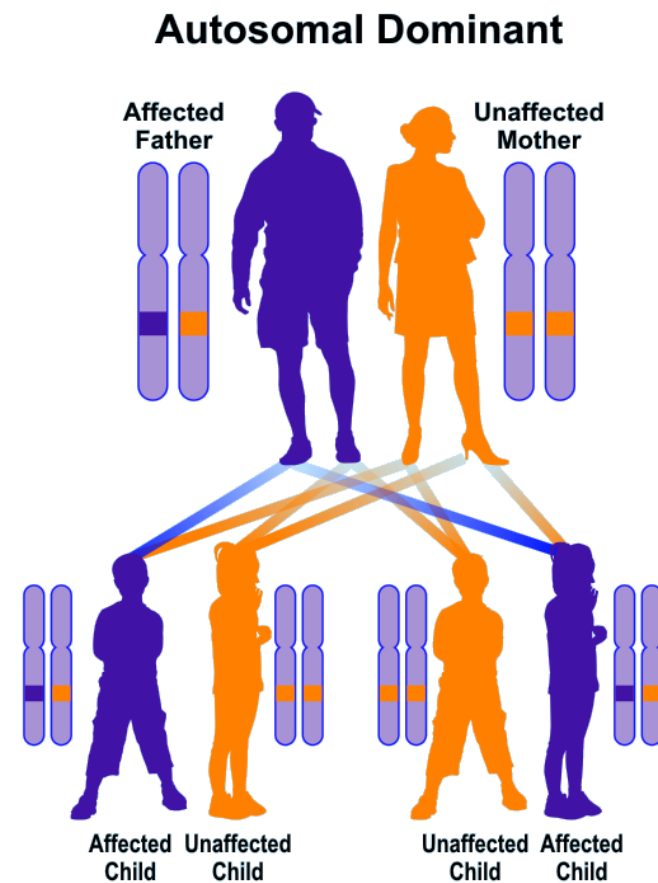


Autosomal dominant



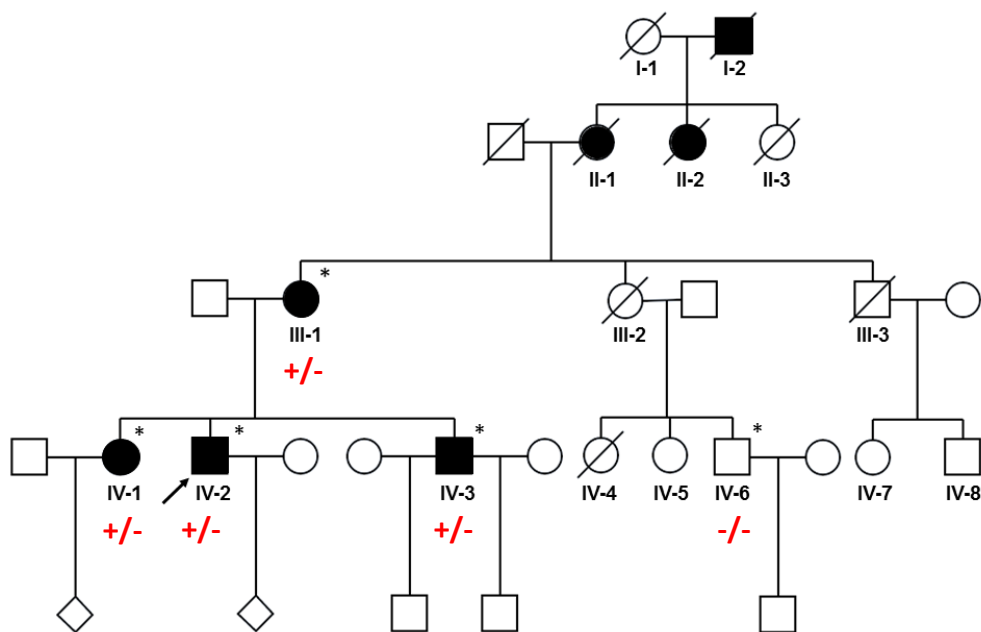
Choroby autosomalnie dominujące (AD)

- Osoby chore są heterozygotami
- W rodowodach obserwujemy pionowy rozkład chorych
- Ryzyko przekazania choroby dziecku wynosi 50%



Dziedziczenie autosomalne dominujące AD – mutacja odziedziczona

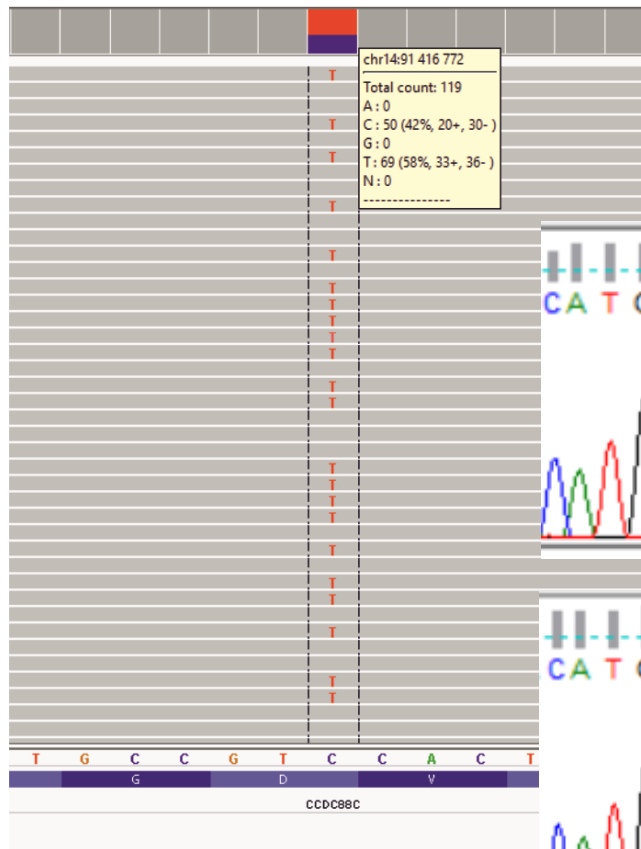
Ataksja rdzeniowo-mózdkowa, typ 40 (SCA40)



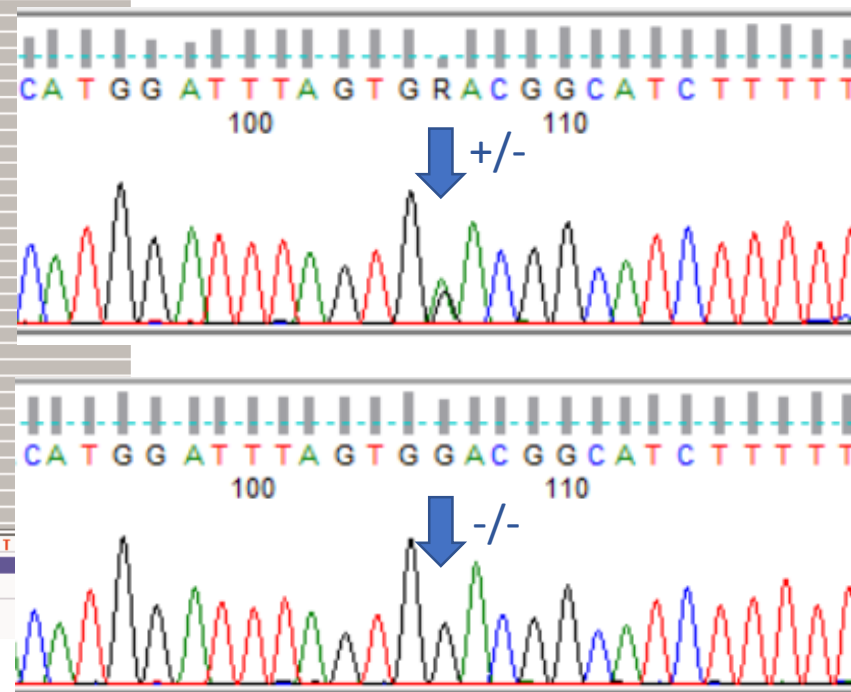
CCDC88C

c.127G>A

p.(Asp432Asn)



NGS

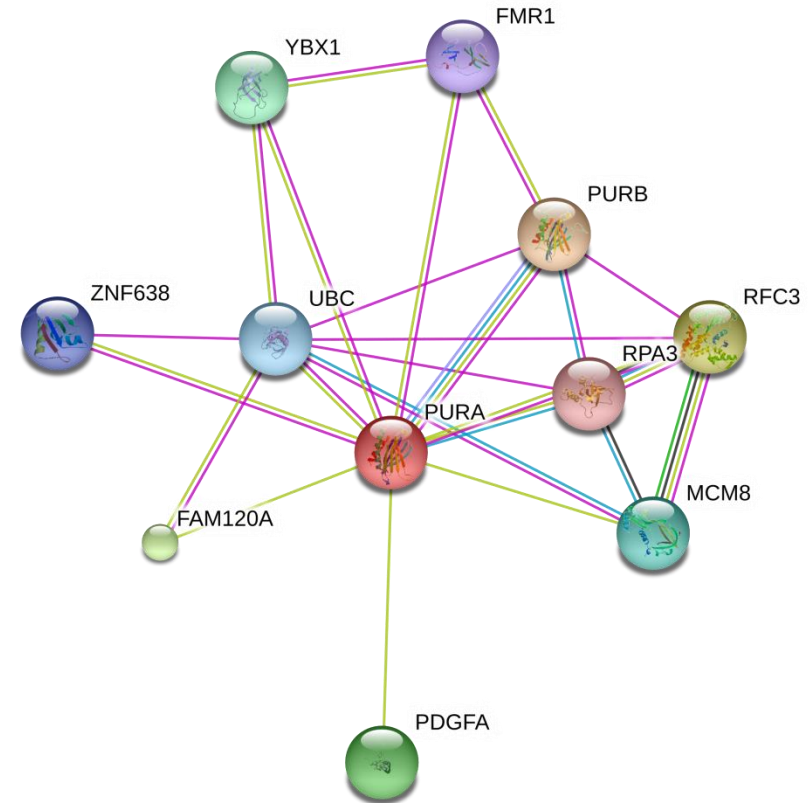


Sanger

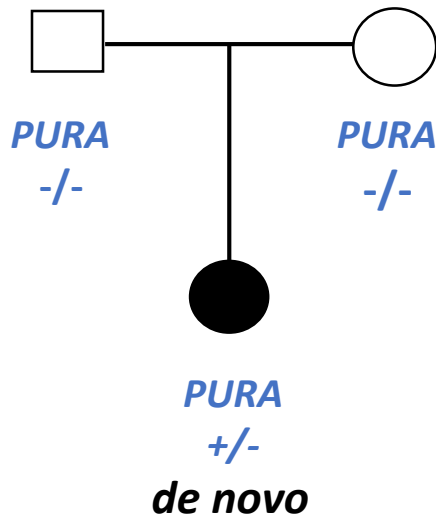
Dziedziczenie autosomalne dominujące AD – mutacja *de novo*

zespół PURA – Autosomalna dominująca NI, typ 31, fenotyp zbliżony do zespołu niedoboru transportera glukozy (SLC2A1)

NI
hipotonia
padaczka

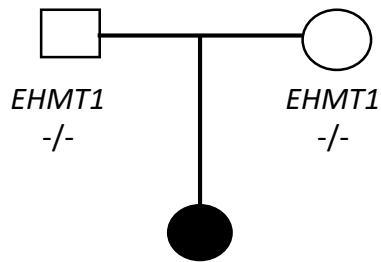


PURA
c.470T>G
p.Met157Arg

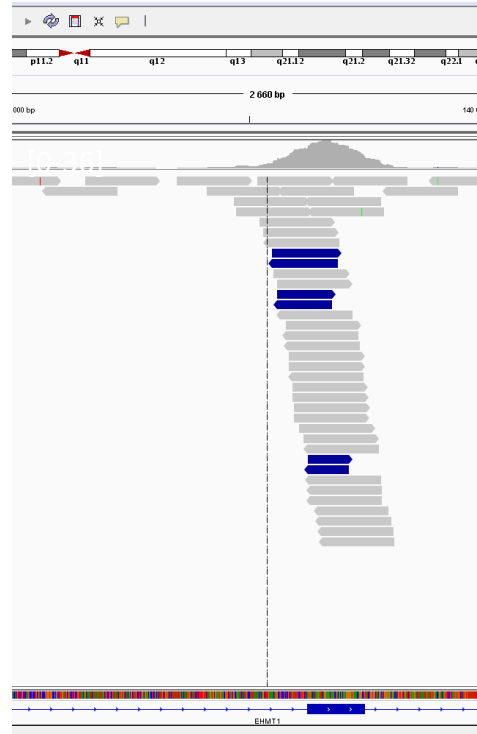


Dziedziczenie autosomalne dominujące AD – mutacja *de novo*

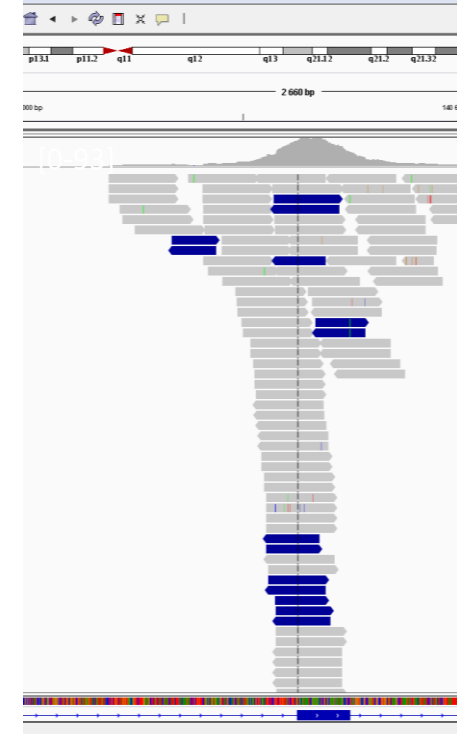
zespół Kleefstra delecja genu *EHMT1*



EHMT1:c.(?-1)_(3897-?)del
+/-
de novo



delecja genu *EHMT1* (*de novo*)

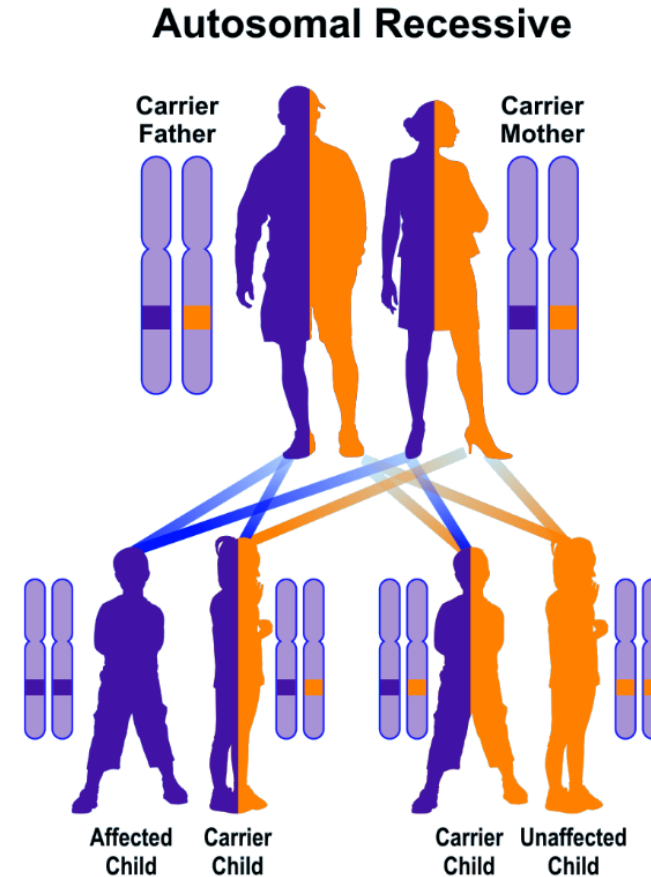


wt

Objawy kliniczne:
głęboka NI, brak mowy, hypotonia, autyzm,
zrośnięte brwi (*Synophrys*), wywinięta dolna warga,
hiperteloryzm, hipoplazja środkowej części twarzy

Choroby autosomalnie recesywne (AR)

- Osoby chore są najczęściej heterozygotyczne względem dwóch różnych patogennych mutacji w obrębie danego genu (rodzice niespokrewnieni)
- Osoby chore są homozygotyczne (najczęściej spokrewnieni rodzice)
- W rodowodach obserwujemy poziomy rozkład chorych
- Ryzyko przekazania choroby dziecku wynosi 25%

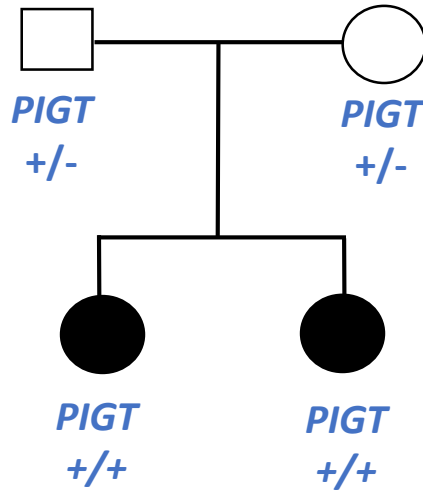


Dziedziczenie autosomalne recesywne AR

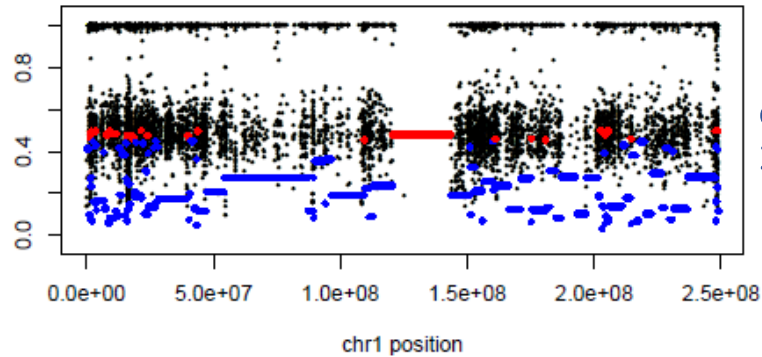
PIGT — Multiple congenital anomalies-hypotonia-seizures syndrome 3

PIGT

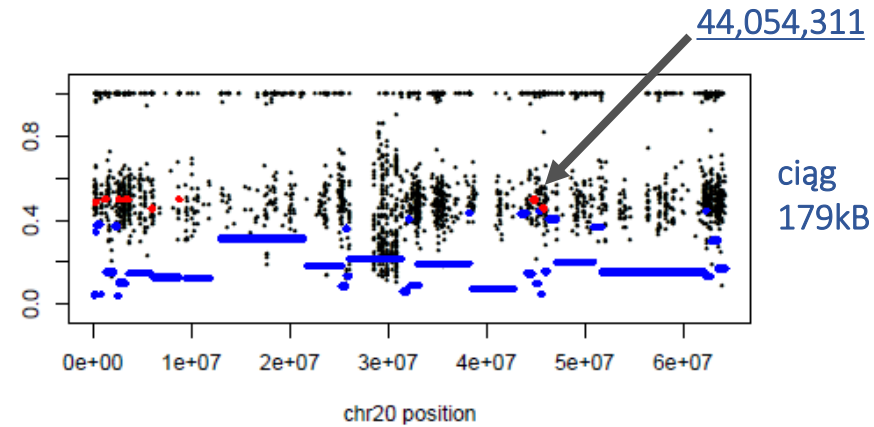
c.1582G>A
p.Val528Met



NI
makrocefalia
oczopląs
hipotonia
padaczka
hipoplazja mózdzku



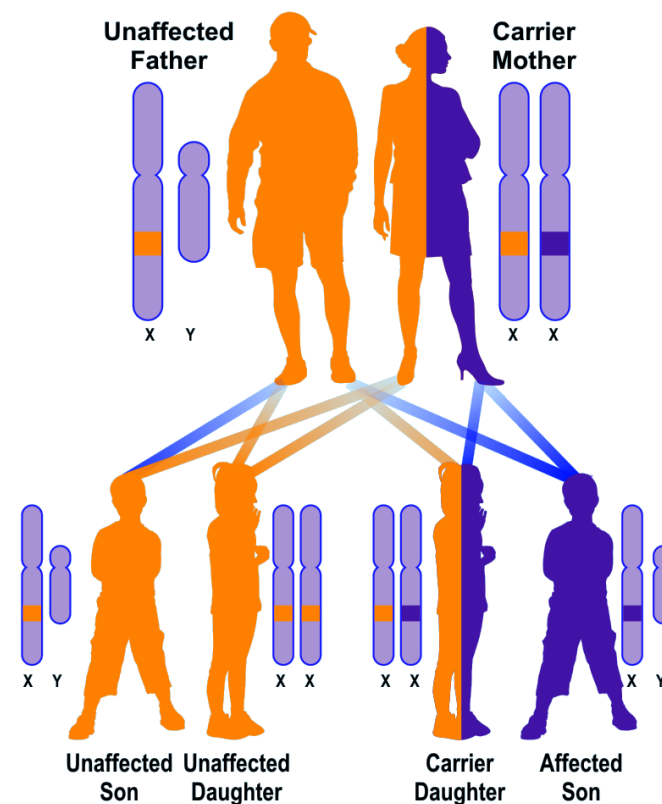
ciąg
23Mb



Choroby sprzężone z chromosomem X recesywne (XLR)

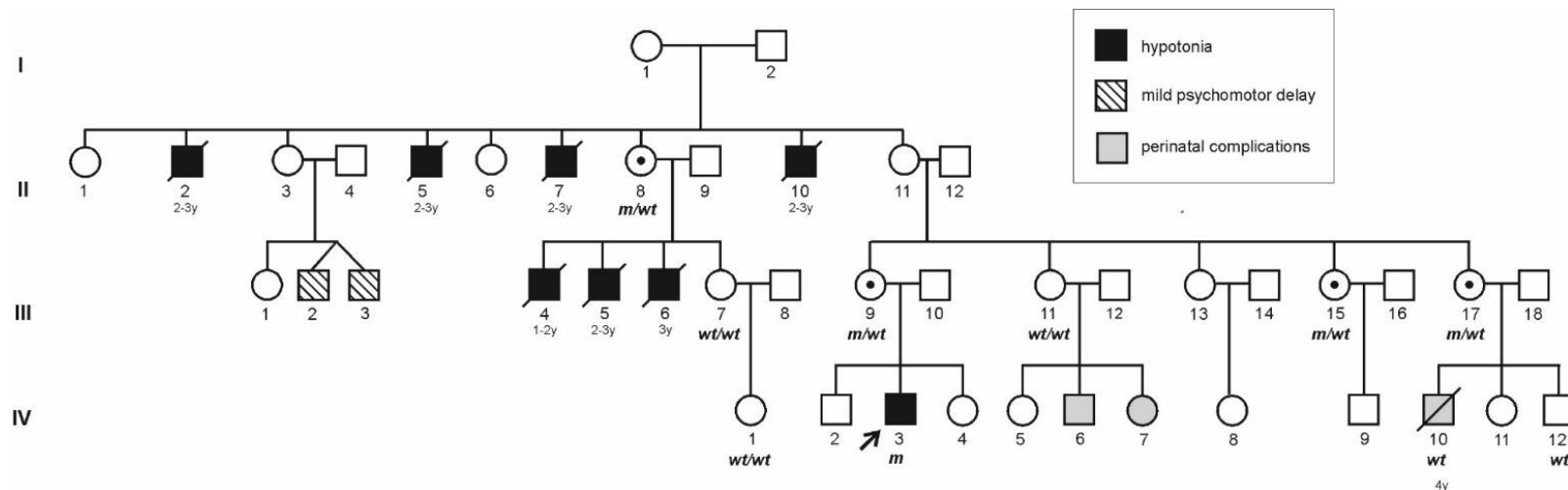
- Osoby chore to hemizygotyczni mężczyźni
- Przypadki chorych kobiet są bardzo rzadkie
 - Np. daltonizm – prawdopodobieństwo wystąpienia u mężczyzn 10-20 krotnie większe niż u kobiet

X-linked Recessive, Carrier Mother



Dziedziczenie sprzężone z chromosomem X, recesywne XLR

Zespół Allan-Herndon-Dudley – AHDS



7 zgonów małych chłopców w wyniku nieznanej choroby

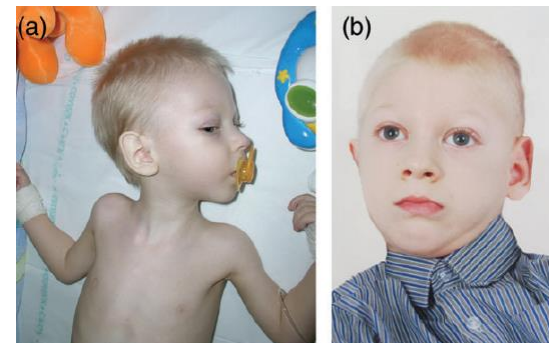
SLC16A2: c.940C<T, MCT8 p.Arg314*

T3/T4 – 2x norma

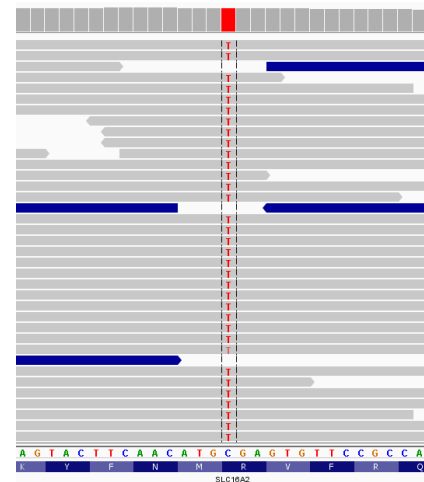
Mutacja w przenośniku hormonów tarczycowych = brak hormonów w mózgu

Możliwość leczenia !!! –

Pochodna hormonu tarczycowego - 3,5-diiodothyropropionic acid



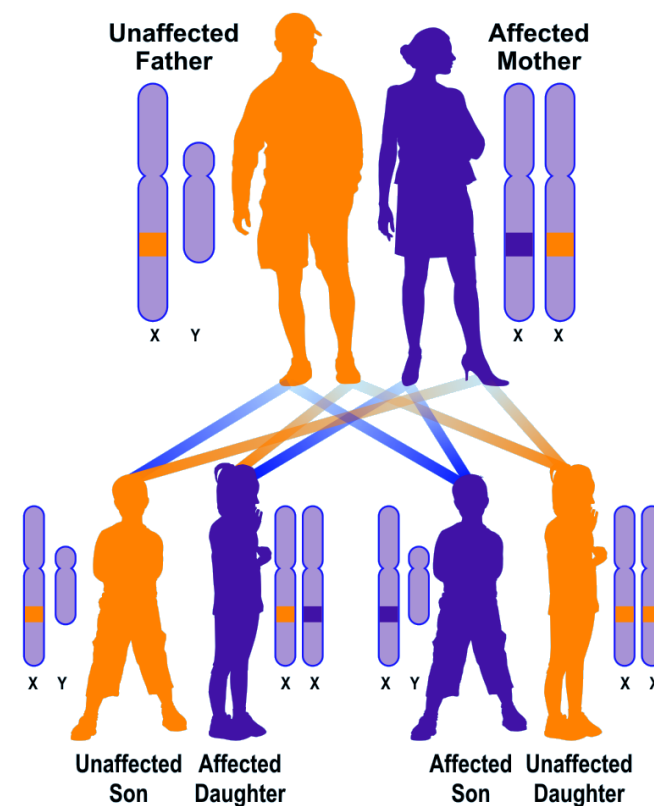
- hipotonia
- upośledzenie rozwoju psychoruchowego znacznego stopnia
- spastyczność kończyn
- brak mowy
- częste infekcje



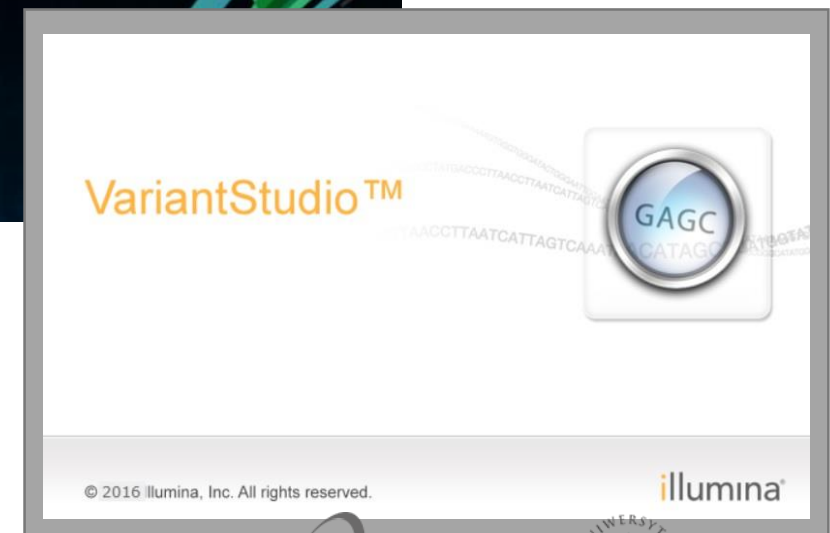
Choroby sprzężone z chromosomem X dominujące (XLD)

- Osoby chore to hemizygotyczni mężczyźni i heterozygotyczne kobiety
- Wpływ inaktywacji chromosomu X na wystąpienie fenotypu u kobiet
- Niektóre choroby są letalne dla chłopców

X-linked Dominant, Affected Mother



Aplikacje komercyjne do analizy danych NGS



Home

Annotation & Classification

Reports

Help

New

Open

Close

Save

Save As

Import VCF

Add Variants to Sample

Import Folder

View All Samples

Current Sample: NB-58112_S37

Remove Sample

Current:

Manage Favorites

Save

Select All

Copy

Column Order

Smaller

Larger

Save As Default

Apply Default

Restore Default

Project

Samples

Filter Favorites

Table Options

Layout

Filters

General

Variant

Gene

Consequence

Population Frequency

Cross Sample Subtraction

Family Based

Custom

Classification

Gene View

Gene	Variant	Chr	Coordinate	...	Type	Genotype	Exonic	Filters	Quality	GQX	Inherited From	Alt Variant Freq	Read Depth	Alt Read Depth	Allelic Depths	...	...	...	...	Transcript	
ESPNP	C>C/G	1	17026004	...	snv	het	yes	LowG...	3	10		50	2	1	1,1				1	NR_026567.1	non
ESPNP	G>G/A	1	17026043	...	snv	het	yes	LowG...	1	10		33.33	3	1	2,1				1	NR_026567.1	non
ESPNP	T>T/A	1	17026049	...	snv	het	yes	LowG...	1	10		33.33	3	1	2,1				1	NR_026567.1	non
N/A	C>C/G	1	17056378	...	snv	het	no	LowG...	8	2		100	1	1	0,1				0		
PADI2	G>G/T	1	17443113	...	snv	het	no	LowG...	8	5		100	1	1	0,1				1	NM_007365.2	intr
PADI1	C>T/T	1	17570955	...	snv	hom	yes	LowG...	39	9		100	2	2	0,2				2	NM_013358.2	3_p
N/A	C>C/G	1	18781186	...	snv	het	no	LowG...	9	5		100	1	1	0,1				0		
PAX7	C>C/T	1	19057480	...	snv	het	no	LowG...	9	5		100	1	1	0,1				1	NM_002584.2	intr
ALDH4A1	C>T/T	1	19199221	...	snv	hom	yes	OffT...	165	23		100	13	13	0,13				1	NM_003748.3	3_p
ALDH4A1	T>A/A	1	19199473	...	snv	hom	no	OffT...	859	40		100	67	67	0,67				1	NM_003748.3	intr
ALDH4A1	G>A/A	1	19199723	...	snv	hom	no	LowG...	52	11		100	3	3	0,3				1	NM_003748.3	intr
ALDH4A1	T>T/G	1	19200802	...	snv	het	no	LowG...	0	11		25	4	1	3,1				1	NM_003748.3	intr
ALDH4A1	G>C/C	1	19200908	...	snv	hom	no	OffT...	275	31		100	22	22	0,22				1	NM_003748.3	intr
ALDH4A1	G>C/C	1	19201184	...	snv	hom	no	LowG...	121	19		100	9	9	0,9				1	NM_003748.3	intr
ALDH4A1	T>TG/TG	1	19201211	...	insertion	hom	no	LowG...	203	9		100	6	5	0,5				1	NM_003748.3	intr
ALDH4A1	A>G/G	1	19201623	...	snv	hom	no	LowG...	58	12		100	4	4	0,4				1	NM_003748.3	intr
ALDH4A1	A>G/G	1	19201665	...	snv	hom	no	LowG...	110	17		100	8	8	0,8				1	NM_003748.3	intr
ALDH4A1	A>G/G	1	19201956	...	snv	hom	yes	LowG...	1853	7		100	151	151	0,151				1	NM_003748.3	syn
ALDH4A1	T>T/G	1	19202292	...	snv	het	no	LowG...	2	10		33.33	3	1	2,1				1	NM_003748.3	intr
ALDH4A1	T>C/C	1	19202507	...	snv	hom	no	LowG...	34	10		100	2	2	0,2				1	NM_003748.3	intr
ALDH4A1	T>C/C	1	19202770	...	snv	hom	no	OffT...	1058	39		100	84	84	0,84				1	NM_003748.3	intr
ALDH4A1	G>A/A	1	19202896	...	snv	hom	yes	PASS	1422	35		100	116	116	0,116				1	NM_003748.3	syn
ALDH4A1	T>C/C	1	19202917	...	snv	hom	yes	PASS	1346	35		100	111	111	0,111				1	NM_003748.3	syn

Variant 1 of 7365

Analiza danych NGS

Potok przetwarzania danych IMiD w oparciu o algorytmy Ensembl VEP

The screenshot shows the Ensembl VEP website. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar for species is also present. The left sidebar lists various resources under 'In this section', including VEP web interface, VEP command line, Data formats, and VEP FAQ. The main content area is titled 'Ensembl Variant Effect Predictor (VEP)' and explains that VEP determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence. It provides a list of information that can be determined using VEP, such as affected genes and transcripts, variant location, consequence on protein sequence, known variants, and SIFT and PolyPhen-2 scores.

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search all species...

Using this website | Annotation and prediction | Data access | **API & software** | About us

In this section

- VEP web interface
 - Input form
 - Results
- VEP command line
 - Tutorial
 - Download and install
 - Running VEP
 - Annotation sources
 - Filtering results
 - Custom annotations
 - Plugins
 - Examples and use cases
 - Other information
- Data formats
- Variant Recoder
- Haplosaurus
- VEP FAQ

Ensembl Variant Effect Predictor (VEP)

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **Genes** and **Transcripts** affected by the variants
- **Location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **Consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **Known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen-2** scores for changes to protein sequence
- ... And more! See [data types](#), [versions](#).

Analiza NGS na żywo

1. Analiza danych NGS w programie Variant Studio – encefalopatia padaczkowa, dziedziczenie AD, mutacja *stop gain*
2. Analiza danych NGS z zastosowaniem algorytmów Ensembl VEP i potoku przetwarzania danych IMiD
 - a. encefalopatia padaczkowa, dziedziczenie AD, mutacja *stop gain*
 - b. migrena hemiplegiczna, dziedziczenie AD, mutacja *missense*
 - c. niepełnosprawność intelektualna, dziedziczenie AD, mutacja *frameshift*
 - d. stwardnienie guzowe, dziedziczenie AD, *delecja*
 - e. niepełnosprawność intelektualna, dziedziczenie AR, *złożona heterozygota*
 - f. encefalopatia padaczkowa, dziedziczenie sprzężone z chromosomem X, nietypowe, mutacja *missense*
 - g. niepełnosprawność intelektualna, dziedziczenie sprzężone z chromosomem X, dominujące, mutacja *stop gain*
3. Samodzielna analiza danych NGS – zespół Lafory, dziedziczenie AR, mutacja *missens*

Dziękuję za uwagę
agnieszka.charzewska@imid.med.pl