

Bazy danych stosowane w analizie wyników badań molekularnych

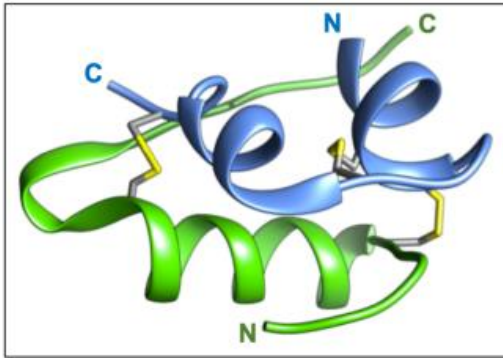
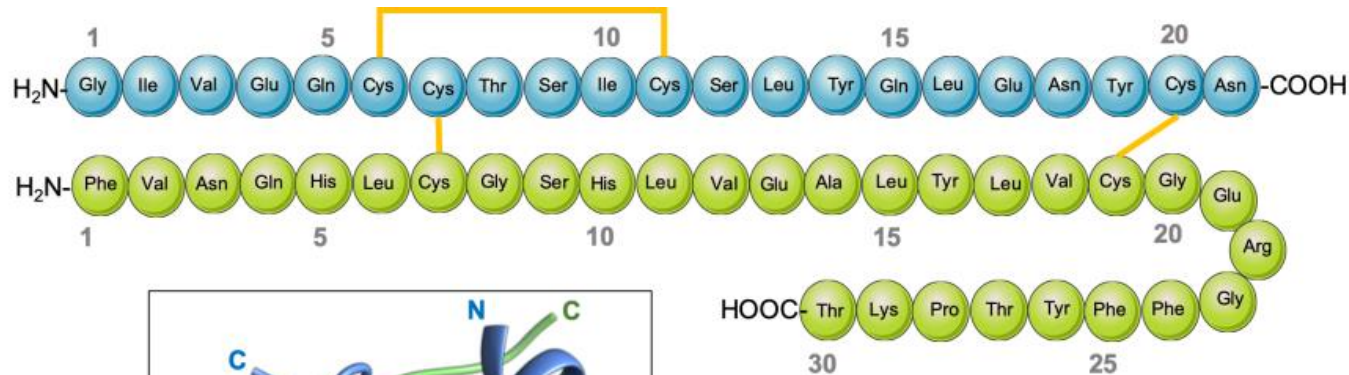
dr n. med. Mateusz Dawidziuk

Zakład Genetyki Medycznej

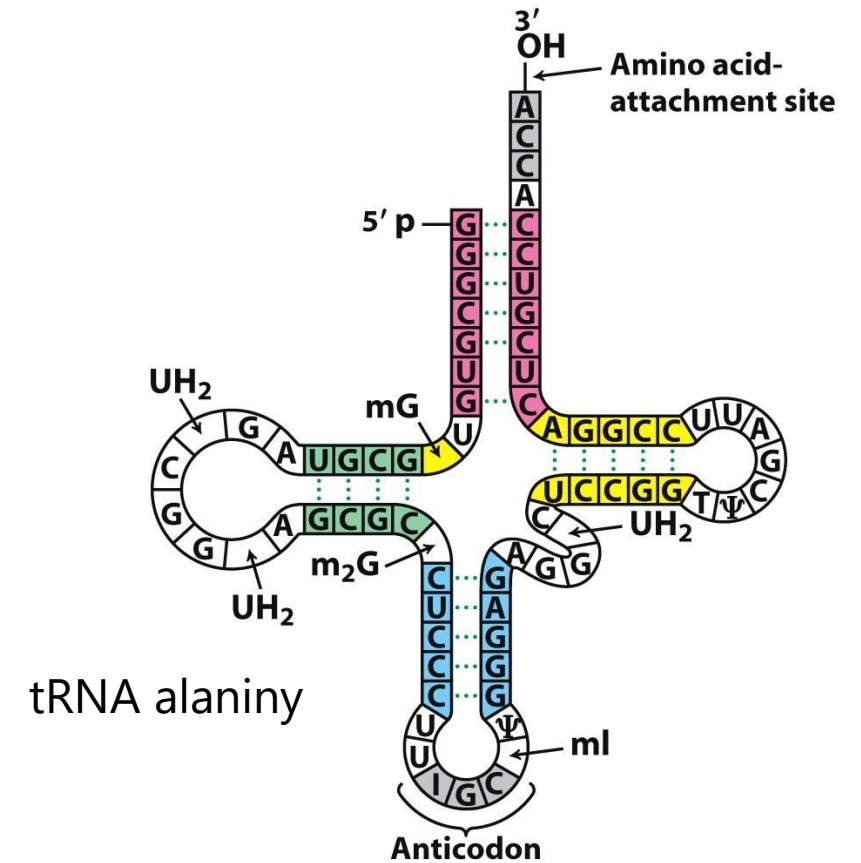
Instytut Matki i Dziecka w Warszawie

Sekwencje cząsteczek biologicznych

- Pierwsza sekwencja białkowa – insulina
- Pierwsza sekwencja nukleotydowa – tRNA alaniny



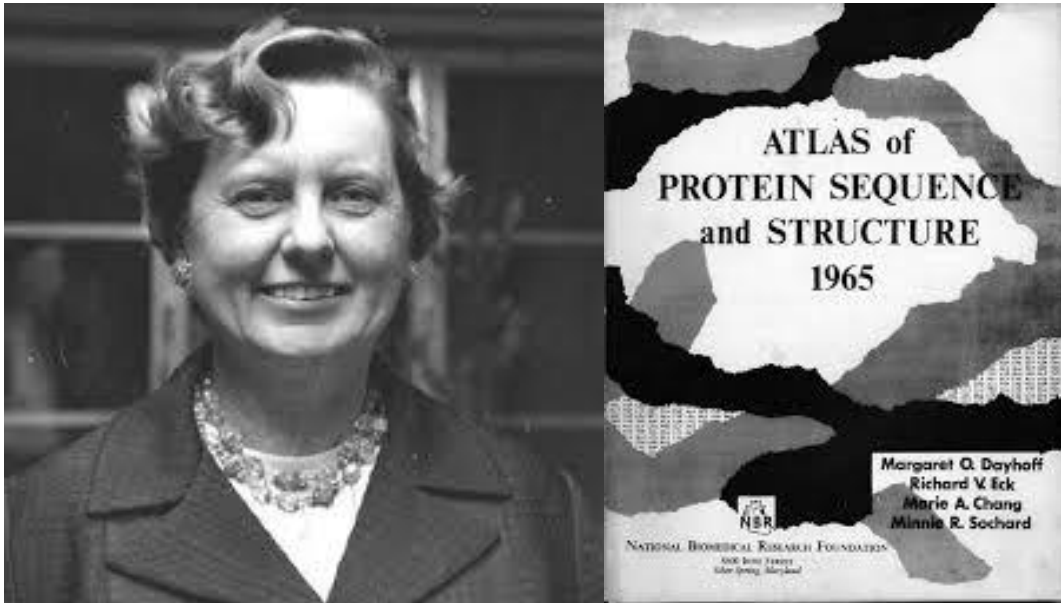
Insulina



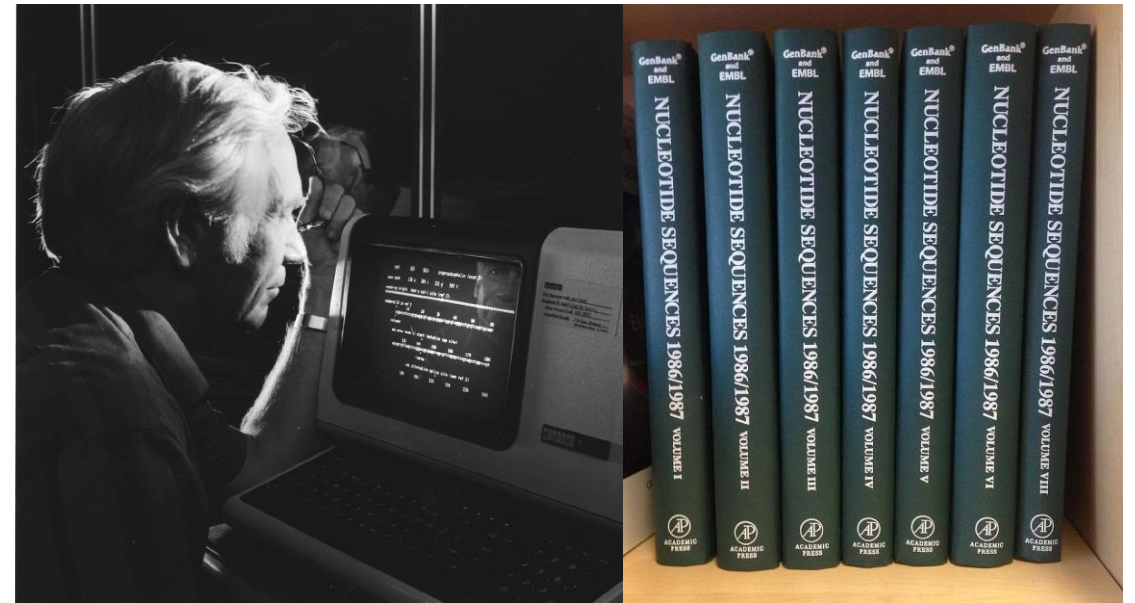
tRNA alaniny

Pierwsze biologiczne bazy danych

- Atlas of Protein Sequence – 65 sekwencji białkowych, podstawa do stworzenia bazy Protein Information Resource w roku 1984
- GenBank – 606 sekwencji nukleotydowych, rok 1982



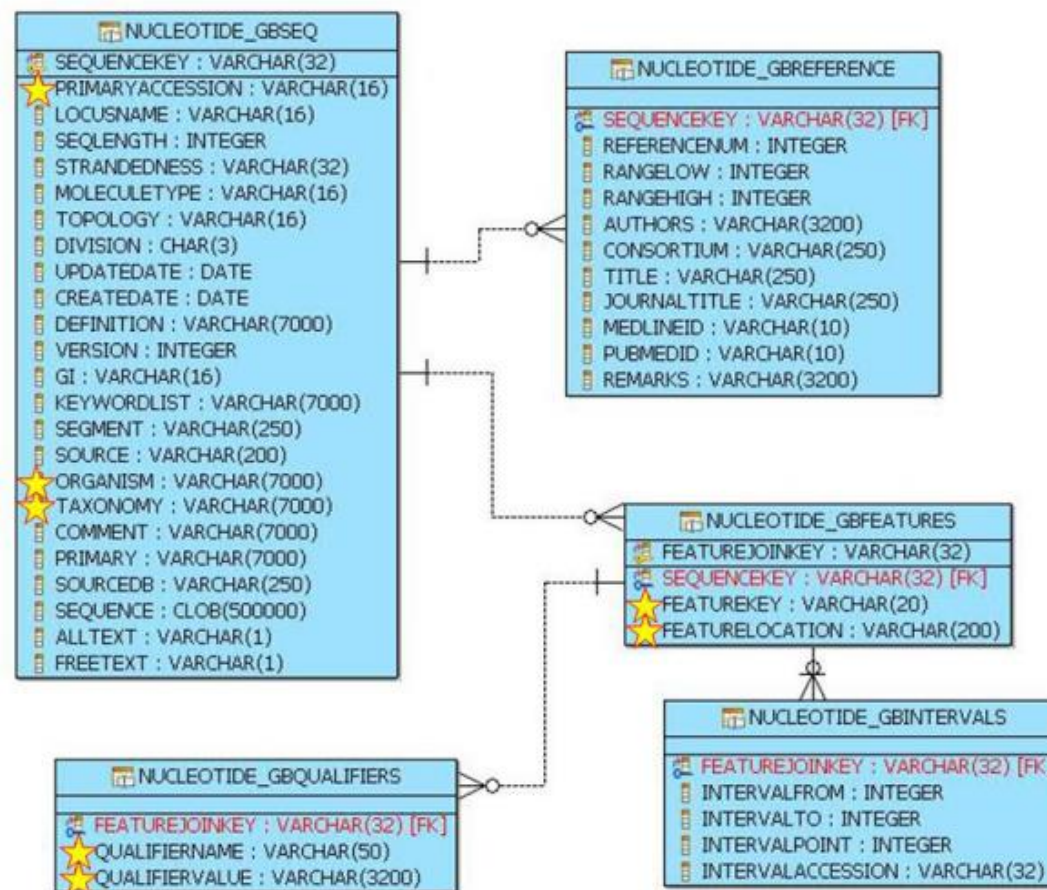
Margaret Dayhoff



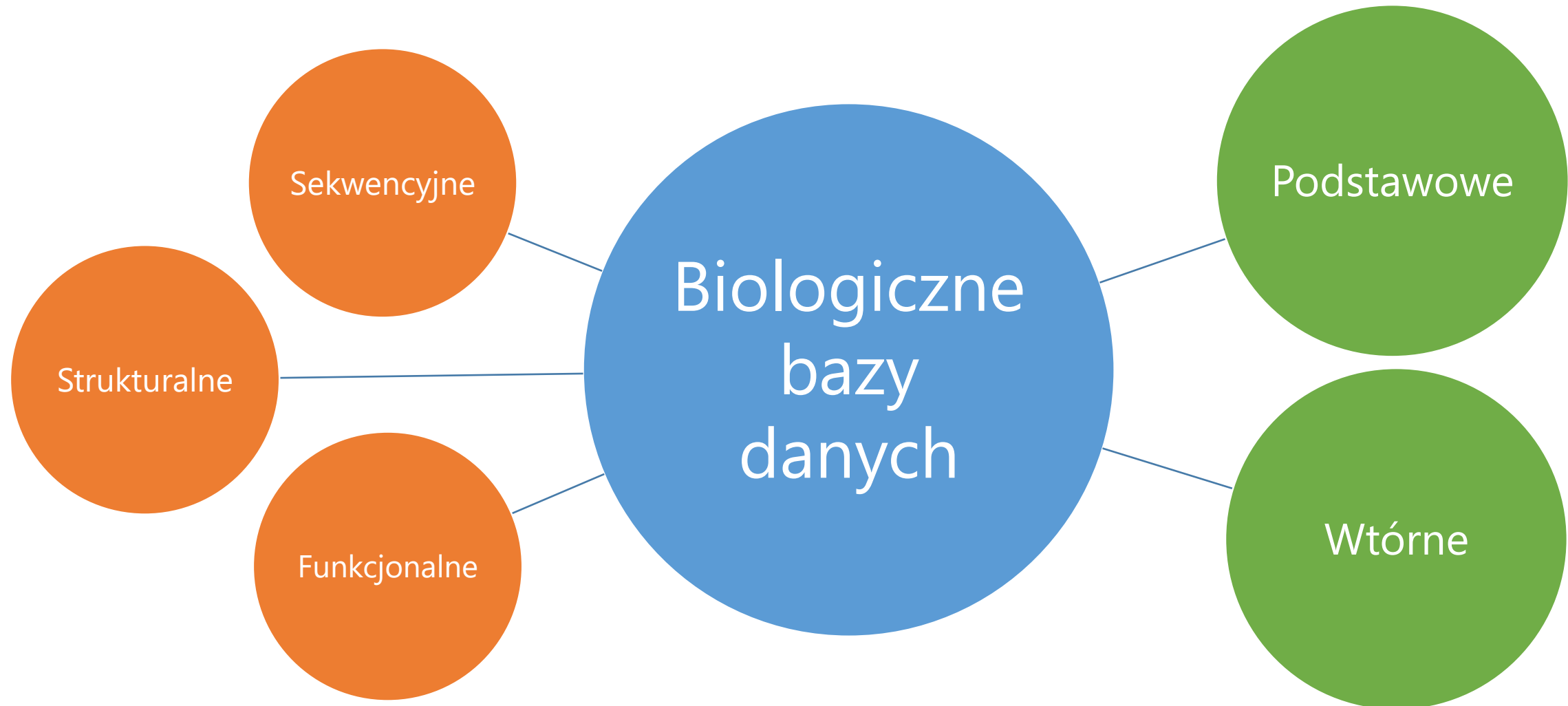
Walter Goad

Czym jest biologiczna baza danych

- Zbiór danych zaprojektowany w sposób umożliwiający pozyskiwanie informacji, ich wprowadzanie czy aktualizację
- Większość baz biologicznych to relacyjne bazy danych



Rodzaje biologicznych baz danych



Bazy podstawowe

- Dane uzyskane w sposób eksperymentalny
- Dane wprowadzane są do systemu przez badaczy
- Dane raz wprowadzone nie ulegają zmianie, mogą być aktualizowane ale informacje o wcześniejszej wersji nie są nadpisywane

- Przykłady:

- GenBank – sekwencje nukleotydowe



- UniProt – sekwencje aminokwasowe



- Protein Data Bank – struktury 3D białek



Bazy wtórne

- Informacje uzyskane na podstawie analizy danych pochodzących z baz podstawowych lub innych baz wtórnych
- Duży nacisk kładzie się na kurację danych
 - Algorytmy komputerowe
 - Manualna analiza i interpretacja uzyskanych danych

- Przykłady:

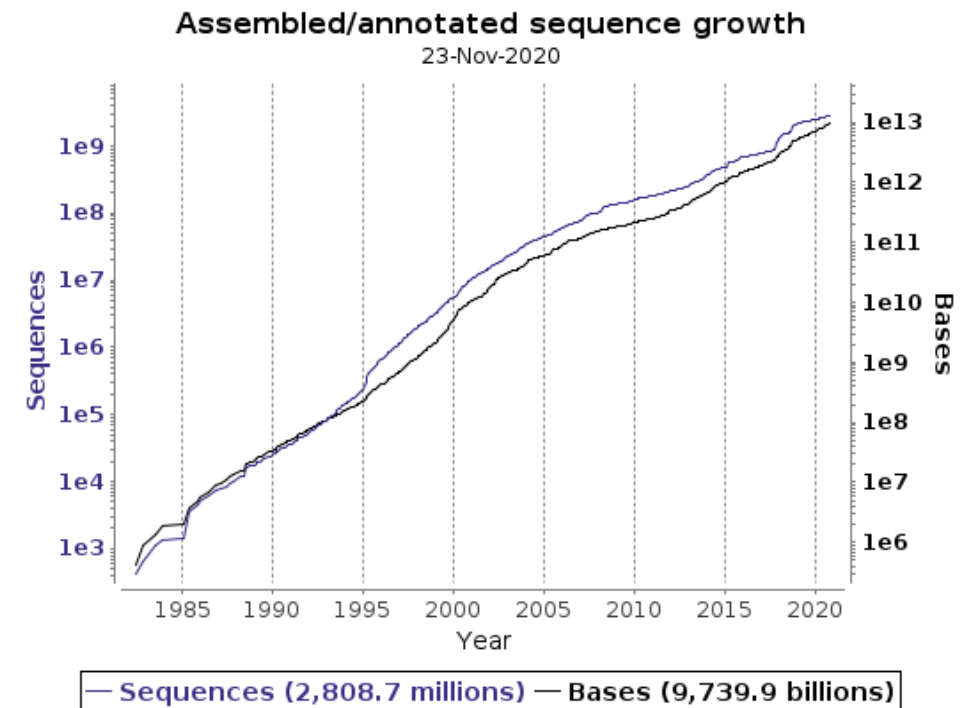
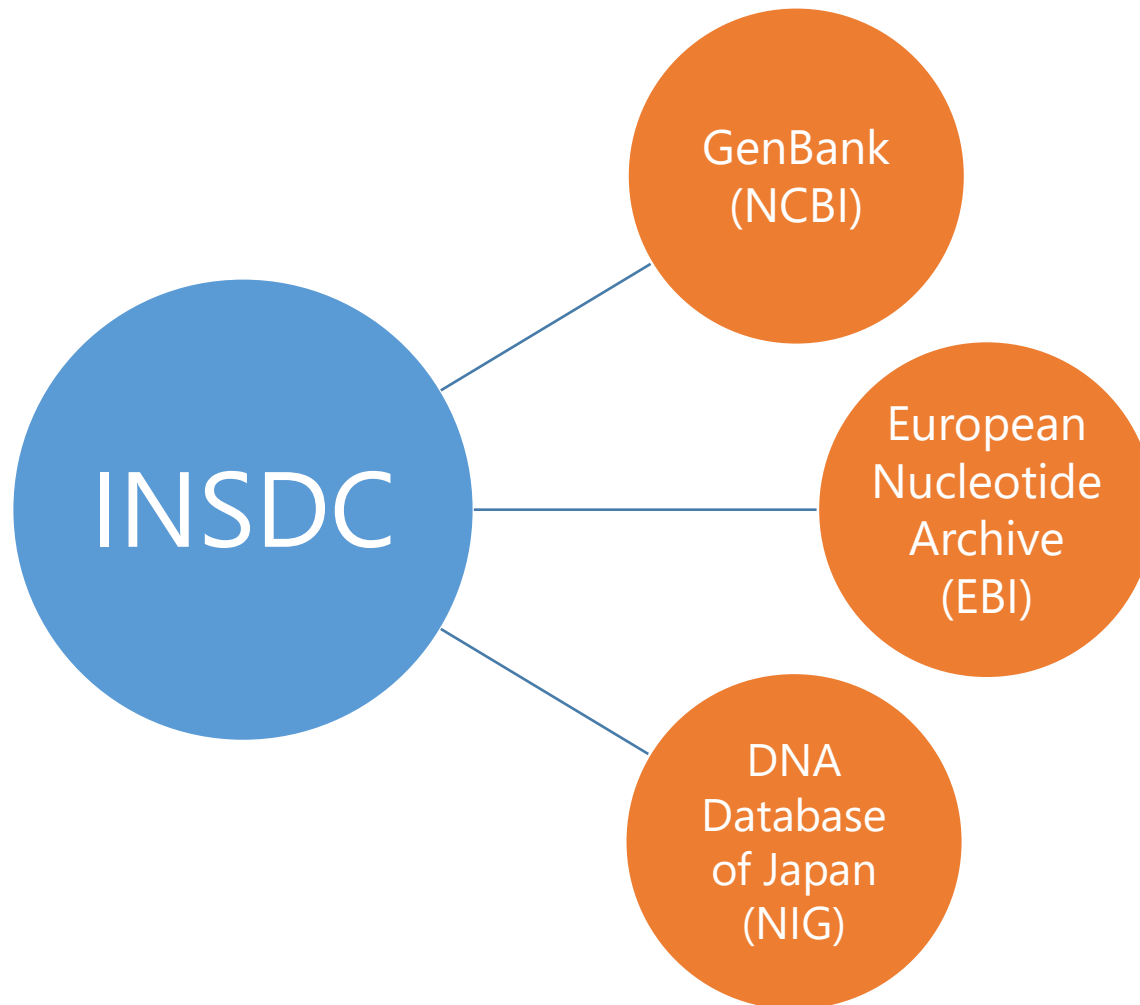
- Ensembl - sekwencje, warianty, funkcje, regulacja



- InterPro - sekwencje, motywy, domeny i rodziny białek

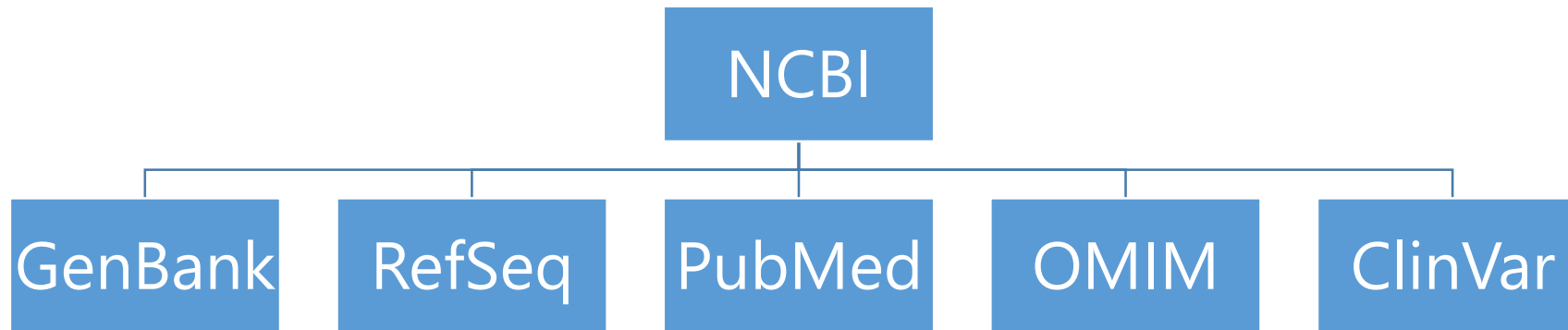


Bazy sekwencji nukleotydowych



National Center for Biotechnology Information

- Główne zadania
 - Przechowywanie informacji
 - Tworzenie narzędzi do analizy danych
 - Rozpowszechnianie wiedzy



www.ncbi.nlm.nih.gov

Wybór bazy danych

Kategorie baz danych

Wyszukiwanie w wybranej bazie

NCBI Resources How To matdawidziuk My NCBI Sign Out

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Remembering Mark Boguski

14 Apr 2021

It is with much sadness that we recently learned of the passing of Mark Boguski, MD PhD a former Senior Investigator in

April 7 Webinar: Recent and upcoming enhancements to NCBI BLAST and Primer-BLAST services!

29 Mar 2021

Join us on April 7, 2021 at 12PM eastern

Conserved Domain Database version 3.19 is available!

17 Mar 2021

The Conserved Domain Database (CDD) version 3.19 is now available. This

[More...](#)

You are here: NCBI > National Center for Biotechnology Information

Support Center

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	GenBank	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	Reference Sequences	NCBI News & Blog
Training & Tutorials	Domains & Structures	BLAST	Gene Expression Omnibus	NCBI FTP Site
Submit Data	Genes & Expression	Nucleotide	Genome Data Viewer	NCBI on Facebook
	Genetics & Medicine	Genome	Human Genome	NCBI on Twitter
	Genomes & Maps	SNP	Mouse Genome	NCBI on YouTube
	Homology	Gene	Influenza Virus	Privacy Policy
	Literature	Protein	Primer-BLAST	
	Proteins	PubChem	Sequence Read Archive	
	Sequence Analysis			
	Taxonomy			
	Variation			

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
[Policies and Guidelines](#) | [Contact](#)

Identyfikatory sekwencji w NCBI

- Numer GI (GenInfo Identifier)
 - Ciąg liczb kolejno nadawanych każdej sekwencji w bazie
 - Gdy sekwencja ulegnie zmianie, nadawany jest nowy numer
- Numer wersji sekwencji (Accession version)
 - Grupuje numery GI konkretnej sekwencji w uporządkowane serie
 - Składa się z przedrostka, numeru sekwencji oraz przyrostka oznaczającego numer wersji sekwencji

Accession version	GI
NC_000001.8	51511461
NC_000001.7	42406218
NC_000001.6	42405892
NC_000001.5	37623929
NC_000001.4	29824572
NC_000001.3	29824110
NC_000001.2	27777714
NC_000001.1	22539468

Identyfikatory sekwencji w NCBI

- N**G**_numer.wersja gen
 - N**M**_numer.wersja mRNA
 - N**P**_numer.wersja białko
 - N**R**_numer.wersja niekodujące RNA
 - N**C**_numer.wersja kontig, sekwencja genomowa
-
- N - sekwencje pozyskane eksperymentalnie, podlegają kuracji
 - X - sekwencje uzyskane najczęściej w sposób automatyczny, nie podlegają kuracji

Wyszukiwanie w bazach NCBI

(MCPH1[Preferred Symbol]) AND Human[Organism]

[Edit](#) [Clear](#)

Builder

	Preferred Symbol ▼	MCPH1	–	Show index list
AND ▼	Organism ▼	Human	–	Show index list
AND ▼	All Fields ▼		– +	Show index list

[Search](#) or [Add to history](#)

www.ncbi.nlm.nih.gov/books/NBK49540/

PubMed

- Wyszukiwarka artykułów naukowych
- Oparta na bazie MEDLINE zawierającej dane bibliograficzne
- Poszczególne rekordy indeksowane są za pomocą słownika MeSH

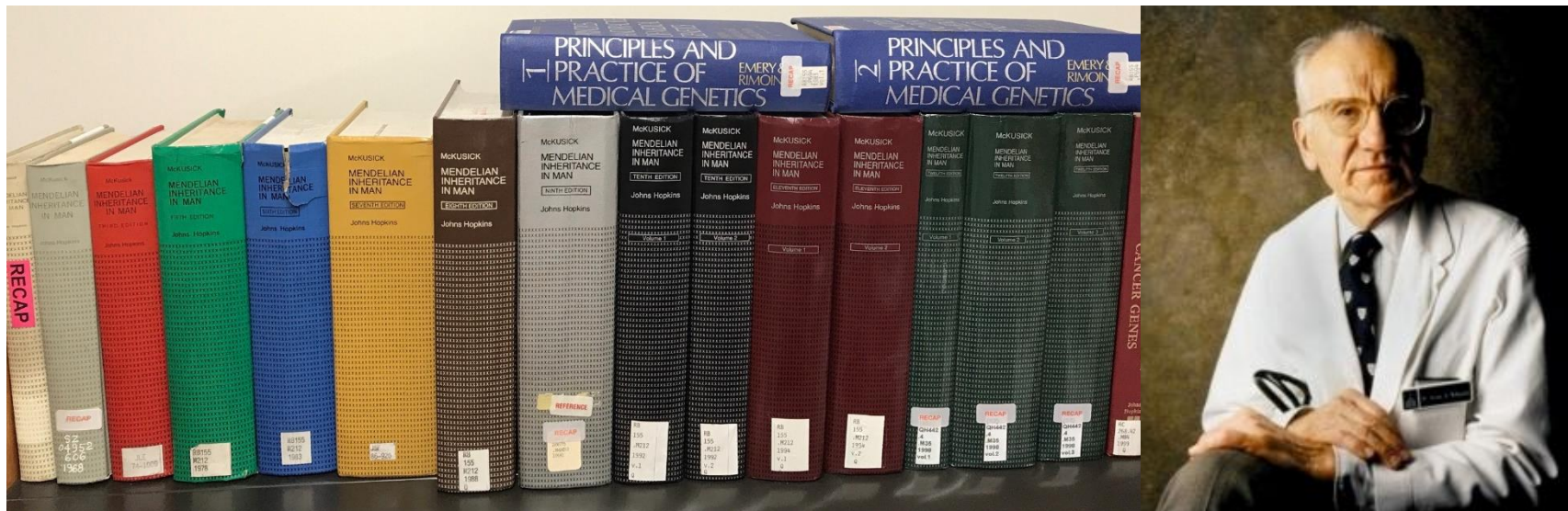
www.pubmed.ncbi.nlm.nih.gov

<https://meshb.nlm.nih.gov/search>

OMIM

Online Mendelian Inheritance in Man

- Kompendium chorób człowieka o podłożu genetycznym
- Zawiera opisy chorób genetycznych o dziedziczeniu mendlowskim
- Informację o ponad 16000 genów



www.omim.org

OMIM kategorie zespołów genetycznych

- 1 – Choroba została umiejscowiona na mapie genetycznej na podstawie asocjacji z genem, ale konkretny defekt nie został poznany
- 2 – Choroba została opisana na podstawie analizy sprzężeń, nie jest znana konkretna mutacja powodująca występowanie choroby
- 3 – Podstawa molekularna choroby jest znana, warianty patogenne powodujące chorobę zostały znalezione w genie
- 4 – Choroba spowodowana jest delecją lub duplikacją większego regionu genomu, może obejmować wiele genów

dbSNP

- Baza wariantów genetycznych
- Podstawienia pojedynczych nukleotydów, małe warianty INDEL, powtórzenia mikrosatelitarne
- Informacje o pozycji genomowej, częstości w populacji, konsekwencjach zmiany na poziomie molekularnym oraz znaczeniu klinicznym

rs748011724

Current Build 154
Released April 21, 2020

Organism	<i>Homo sapiens</i>	Clinical Significance	Reported in ClinVar
Position	chr8:6445347 (GRCh38.p12) ?	Gene : Consequence	MCPH1 : Stop Gained
Alleles	T>C / T>G	Publications	0 citations
Variation Type	SNV Single Nucleotide Variation	Genomic View	See rs on genome
Frequency	G=0.000032 (8/249570, GnomAD_exome) G=0.000048 (6/125568, TOPMED) G=0.000017 (2/120628, ExAC) (+ 3 more)		

Variant Details

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

ALFA Allele Frequency (New)
The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project [page](#) including descriptions, data access, and terms of use.
Release Version: 20200227123210

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	11176	T=0.99991	C=0.00000, G=0.00009
European	Sub	8134	T=0.9999	C=0.0000, G=0.0001
African	Sub	676	T=1.000	C=0.000, G=0.000
African Others	Sub	14	T=1.00	C=0.00, G=0.00
African American	Sub	662	T=1.000	C=0.000, G=0.000

ClinVar

- Baza wariantów genetycznych
- Powiązania pomiędzy konkretnymi wariantami a zespołami genetycznymi
- Informacje o patogenności zmiany oraz wiarygodności źródła informacji

NM_024596.5(MCPH1):c.1625T>G (p.Leu542Ter)[Cite this record](#)

Interpretation: Pathogenic/Likely pathogenic

Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

Submissions: 2 (Most recent: Nov 14, 2018)

Last evaluated: Oct 31, 2018

Accession: VCV000586140.1

Variation ID: 586140

Description: single nucleotide variant

Variant details

Conditions

Gene(s)

Aggregate interpretations per condition

Interpreted condition	Interpretation	Number of submissions	Review status	Last evaluated	Variation/condition record
not provided	Likely pathogenic	1	criteria provided, single submitter	Jun 13, 2018	RCV000712273.1
Primary autosomal recessive microcephaly 1	Pathogenic	1	criteria provided, single submitter	Oct 31, 2018	RCV000763602.1

European Bioinformatics Institute

- Zadania zbliżone do NCBI
- Główne bazy danych:
 - EMBL-Bank - baza danych genetycznych
 - Ensembl - przeglądarka genomowa
 - UniProt - baza sekwencji białkowych

<https://www.ebi.ac.uk/services>

Identyfikatory sekwencji w Ensembl

Dla człowieka:

- ENS**G**numer.wersja Gen
- ENS**T**numer.wersja Transkrypt
- ENS**P**numer.wersja Białko
- ENS**E**numer Ekson

Dla innych gatunków w nazwie umieszczony jest skrót:

- ENS**DARG**numer.wersja Gen danio pręgowany
- ENS**MUST**numer.wersja Transkrypt mysz domowa
- ENS**RNOE**numer Ekson szczur

UniProt

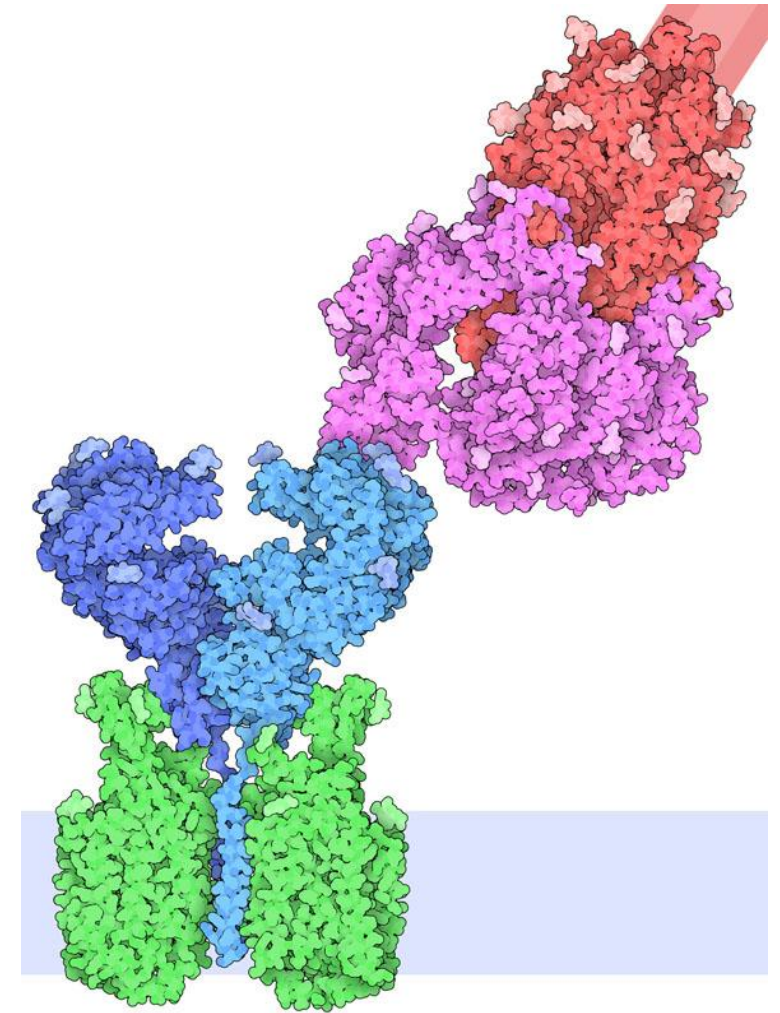
- Dwie części:
 - Swiss-Prot - dane kurowane
 - TrEMBL - dane uzyskane automatycznie
- Informacje o sekwencjach aminokwasowych, motywach, domenach czy rodzinach białek
- Narzędzia do analizy - BLAST, Clustal Omega

<https://www.uniprot.org>

Protein Databank

- Zawiera struktury 3D białek
- Uzyskane metodą krystalografii, spektroskopii NMR czy mikroskopii krioelektronowej
- Pliki z koordynatami dla konkretnych cząsteczek
 - Lista atomów w cząsteczce
 - Lokalizacja 3D w przestrzeni
- Narzędzia do wizualizacji białek oraz narzędzia do ich analizy

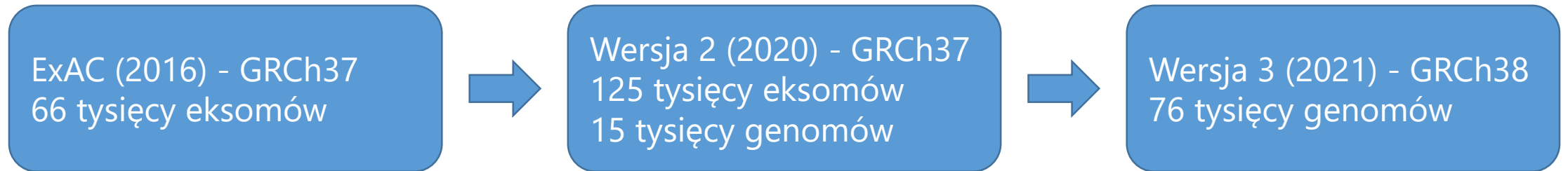
<https://www.rcsb.org>



gnomAD

Genome Aggregation Database

- Baza agregacyjna
- Dane z różnych wielkoskalowych projektów sekwencjonowania eksomowego i genomowego człowieka
- Informacje o częstości występowania danych wariantów w populacji, jakości sekwencjonowania czy też przeglądarkę genomową.

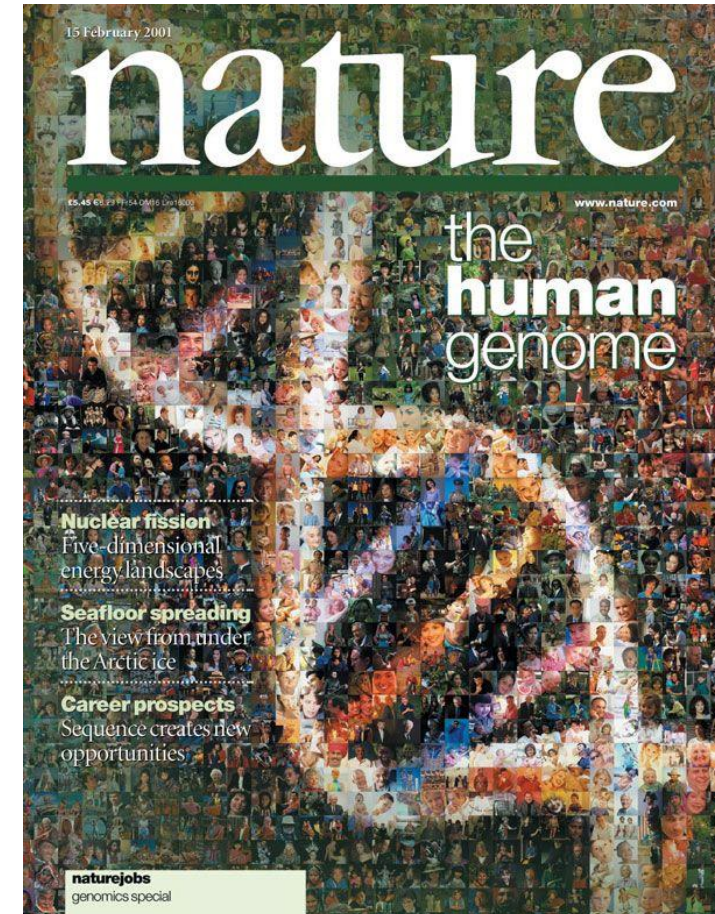


<https://gnomad.broadinstitute.org>

UCSC

- Baza sekwencji genomowych
- Przydatne narzędzia do analizy danych
 - Przeglądarka genomowa
 - Blat
 - Table browser
 - PCR *in-silico*
 - LiftOver

<https://genome-euro.ucsc.edu>



HGMD

Human Gene Mutation Database

- Informację o mutacjach związanych z chorobami genetycznymi
- Dane te z artykułów naukowych
- Główna baza informacji o konsekwencjach klinicznych mutacji
- Płatny dostęp
- Wersja darmowa - dane publikowane z 3 letnim opóźnieniem, ograniczona funkcjonalność

Dziękuję za uwagę