Bioinformatická analýza mutací

Kamila Réblová

Rozvoj nástrojů pro data mining:

- rozvoj bioinformatických nástrojů (např. strukturní a predikční metody)
- vznik odvozených databází

Archivace sekvenčních dat (databáze), snadná manipulace, dostupnost

Primární (sekvenční) databáze nukleových kyselin:

NCBI (National Center for Biotechnology Information)
-> GenBank http://www.ncbi.nlm.nih.gov/Genbank/

EBI (European Bioinformatics Institute) -> European Molecular Biology Laboratory (EMBL)-Bank http://www.ebi.ac.uk/embl/

DNA Data Bank of Japan (DDBJ)
https://www.ddbj.nig.ac.jp/index-e.html

Formáty sekvencí:

Fasta formát

>sp|P00439|PH4H_HUMAN Phenylalanine-4-hydroxylase OS=Homo sapiens GN=PAH PE=1 SV=1
MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNGAISLIFSLKEEVGALAKVLRLFEENDV NLTHIESRPSRLKKD
EYEFFTHLDKRSLPALTNIIKILRHDIGATVHELSRDKKKDTVPW FPRTIQELDRFANQILSYGAELDADHPGFKD
PVYRARRKQFADIAYNYRHGQPIPRVEYM EEEKKTWGTVFKTLKSLYKTHACYEYNHIFPLLEKYCGFHEDNIP
QLEDVSQFLQTCTGF RLRPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDRSF
A QFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKQGDSIKAYGAGLLSSFGELQYCLSE KPKLLPLELEKTAIQ
NYTVTEFQPLYYVAESFNDAKEKVRNFAATIPRPFSVRYDPYTQR IEVLDNTQQLKILADSINSEIGILCSALQKIK

GenBank format

A sequence file in GenBank format can contain several sequences.
One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines.

An example sequence in GenBank format is:

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccgggccc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

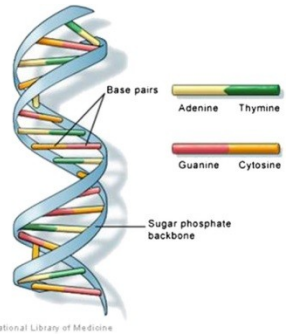
IUPAC nucleotide code

Nucleotide ambiguity code

(as defined in DNA Baser Sequence Assembler)

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

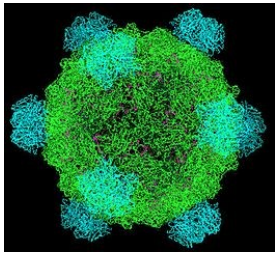
Large-scale DNA sequencing projects:



1953 - Odhalena struktura DNA… ???

1965 – přečtena sekvence tRNA kvasinky

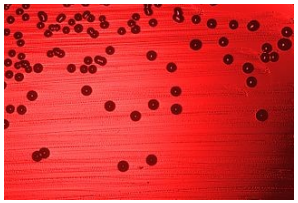
1974 – přečtena sekvence genomu bakteriofága ΦX174 **(**5,375 bp**)**



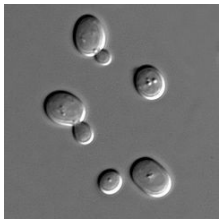
1988-1990 – založen Human Genome Project (HGP)
(probíhá 1990-2000/2003)

První organismy:

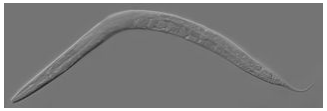
1995 – Haemophilus influenzae (1,830,140 bp)
(*epiglotitida, meningitida***)**



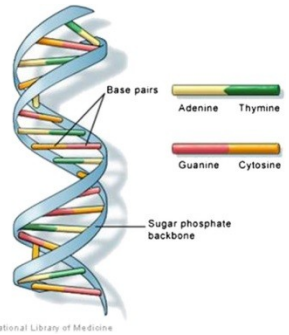
1996 – Saccharomyces cerevisiae (12,068,000 bp)



1998 – Caenorhabditis elegans (100,000,000 bp)



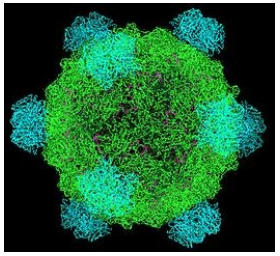
Large-scale DNA sequencing projects:



1953 - Odhalena struktura DNA, J. Watson, F. Crick, R. Franklin

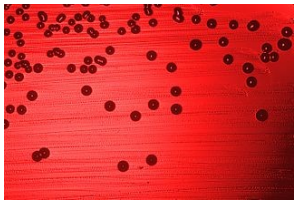
1965 – přečtena sekvence tRNA kvasinky

1974 – přečtena sekvence genomu bakteriofága ΦX174 **(**5,375 bp**)**



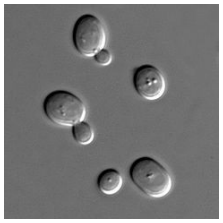
1988-1990 – založen Human Genome Project (HGP)
(probíhá 1990-2000/2003)

První organismy:

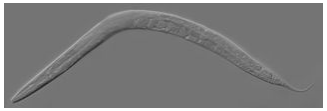


1995 – Haemophilus influenzae (1,830,140 bp)
(*epiglotitida, meningitida***)**

1996 – Saccharomyces cerevisiae (12,068,000 bp)



1998 – Caenorhabditis elegans (100,000,000 bp)

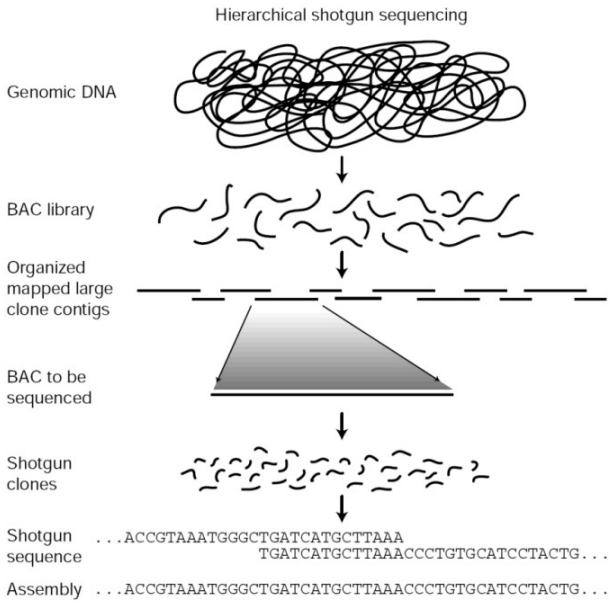


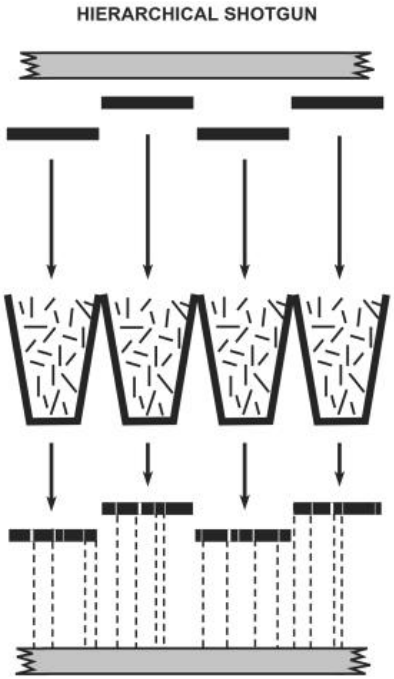
2001 – Lidský genom (3,200,000,000 bp)

Human Genome Project Consortium (2001) "A physical map of the human genome" Nature **409**:934–41

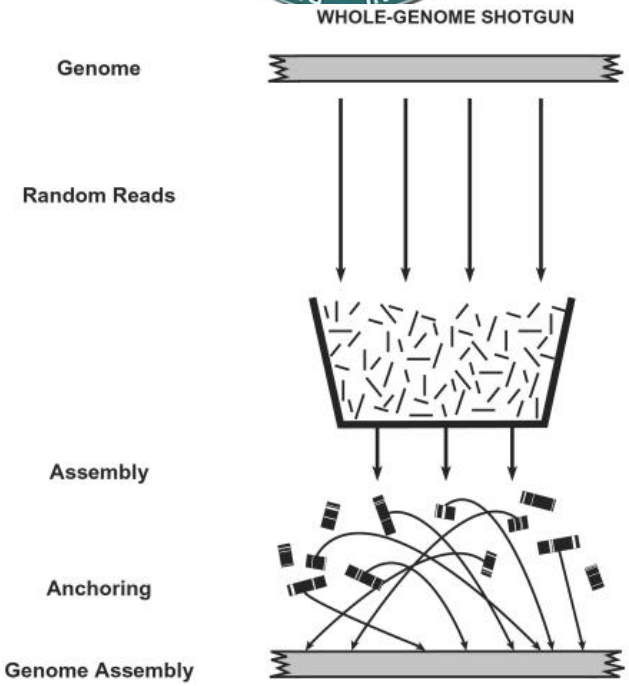
Dideoxy neboli ‘Sangerova metoda’

1977 – zavedena Sangerova metoda (Nobelova cena 1980)





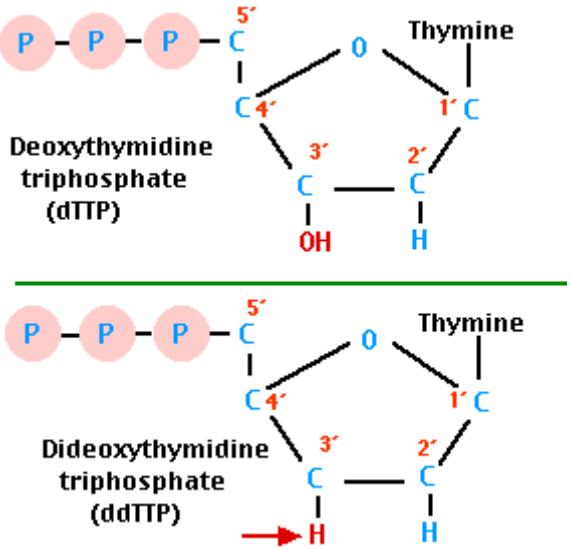
HGP



Celera Company

Sangerova metoda

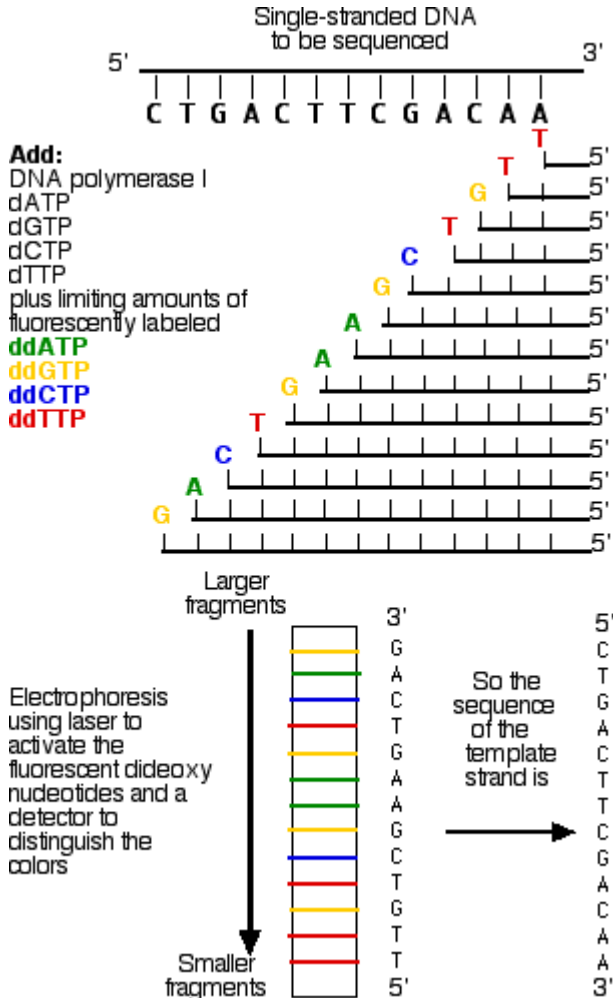
Původně gelová elektroforéza a radioaktivní značení
=> fluorescenčních značení
dideoxynukleotidtrifosfátů v kombinaci s kapilární elektroforézou.
Takto značené oligonukleotidy jsou detekovány laserem



dATP
dGTP
dCTP
dTTP

DNA polymerase I

ddATP
ddGTP
ddCTP
ddTTP



e.g. The concentration of ddATP should be 1% of the concentration of dATP.

Genom

3,2 mld párů bází

2% genomu → proteiny
zbytek – “junk” DNA -
Non-coding = tRNA, rRNA, Transpozony

Co ovlivňují? - 3D strukturu DNA - Chromosomalní uspořádání, regulace procesů v buňky (dělení)

Proteom

Proteom je soubor proteinů, které jsou produkované z genomu daného organismu
Je dynamický, zavisí na věku, ale také typu tkáně a stavu oragnismu.

Mikrobiom

1:10? 10:1 (bakterie ca. 1 kg)

3 200 000 000 párů bází
23 párů chromozómů
Délka DNA= 1 metr

	člověk	octomilka
známé protein-kódující geny	?	?

3 200 000 000 párů bází
23 párů chromozómů
Délka DNA= 1 metr

	člověk	octomilka
známé protein-kódující geny	20-25 tis	13 tis

3 200 000 000 párů bází
23 párů chromozómů
Délka DNA= 1 metr

	člověk	octomilka
známé protein-kódující geny	20-25 tis	13 tis
transkripty	178 191	23 017

95-100% lidských genů podléhá sestřihu, min. 2 transkripty.

Next (second) generation sequencing – rozvoj od 2005

'Next' = not Sanger Method

454 sequencing – pyrosequencing

Illumina – sequencing by synthesis/bridge amplification

SOLiD - Sequencing by Oligonucleotide Ligation and Detection

Third generation - Oxford nanopore technology (ONT)

Aviti

Nové platformy

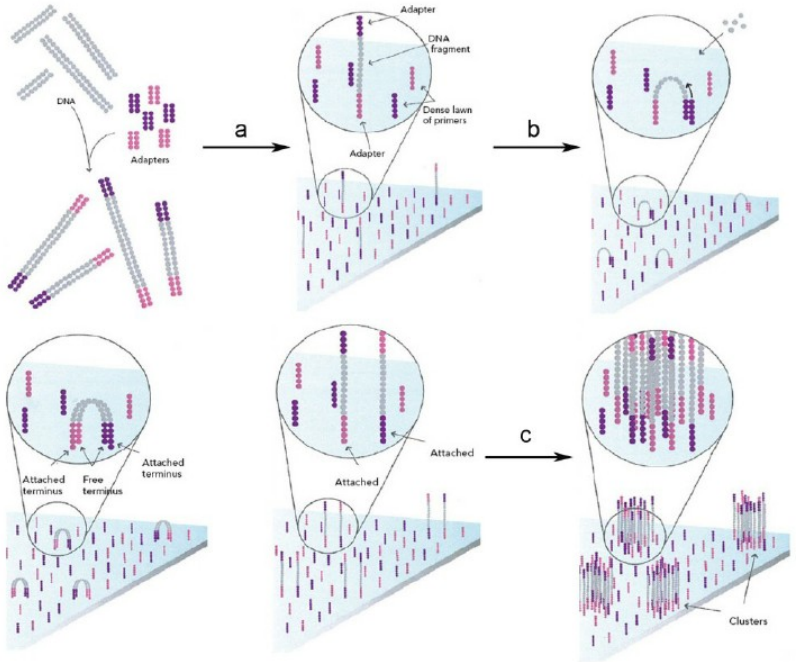
MGI

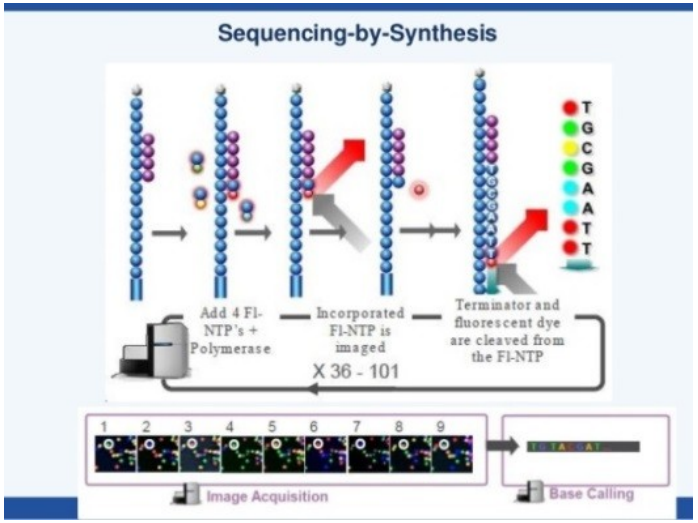
Capture seq.

(Použití sond, které vychytají cílené geny - sequence capture)

- Fragmentace (fyzikálně např. sonikace, nebo enzymaticky (DNAázy, Transpozáz) nebo chemicky (tepelné štěpení dvojmocnými ionty))

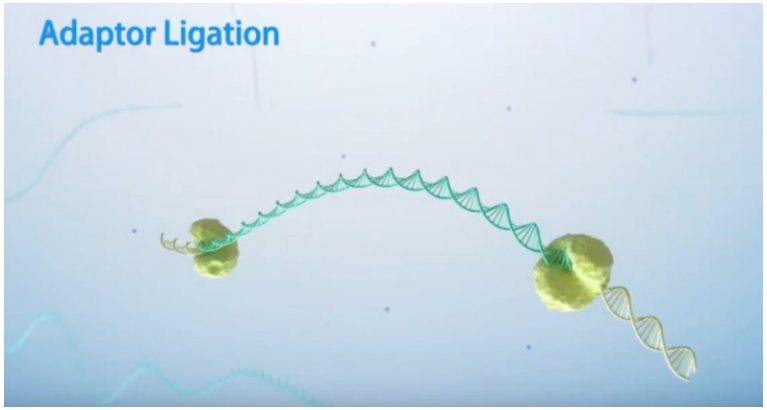
- Úprava konců – end repair (zarovnání), fosforylace, dATP
- Ligace adaptoru, ktere predstavuji „primery“
- Vychytání úseků pomocí hybridizačních song
- Amplifikace

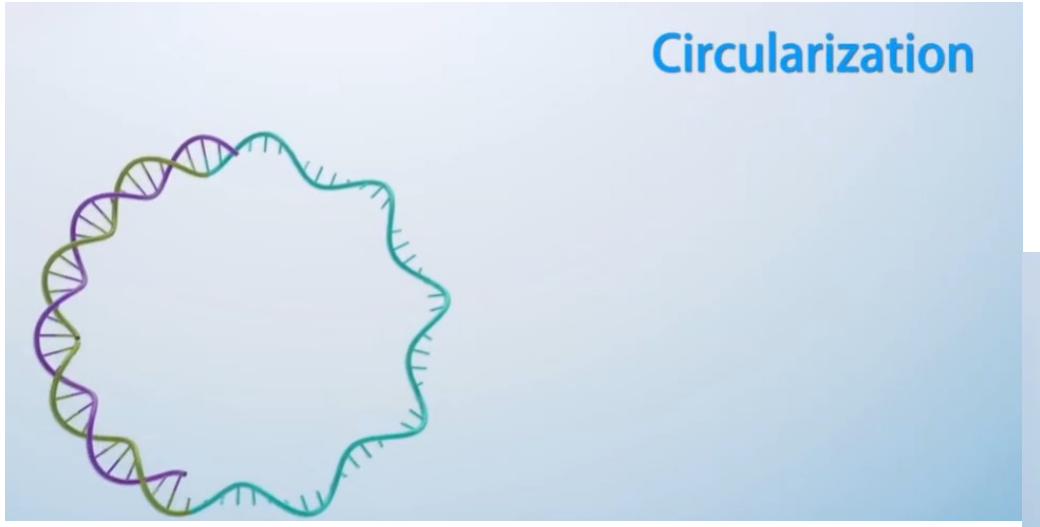


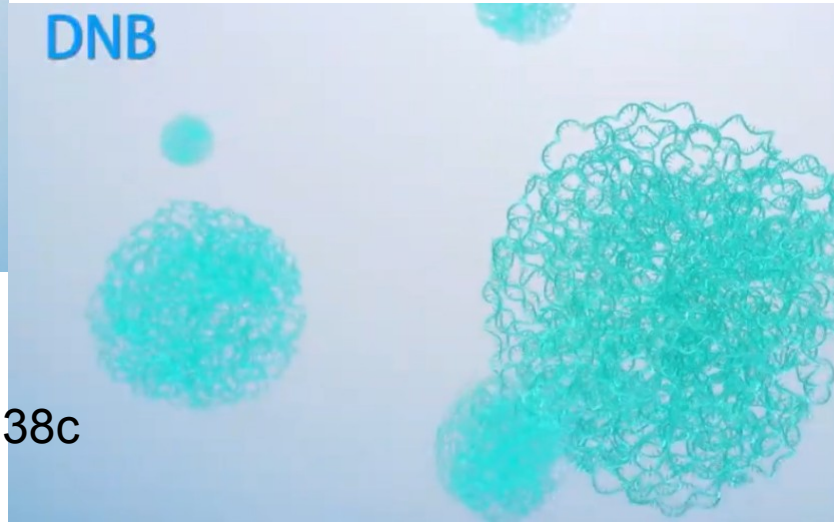


AVITI a MGI

Cirkularizace DNA

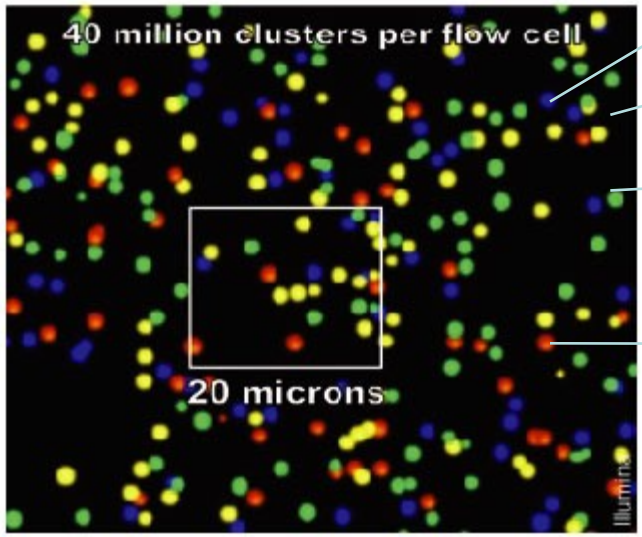






https://www.youtube.com/watch?v=xUVdJN0m38c

Tif fily -> bcl file (intensity)-> Fastq fily



A

G

C

T

Primární data

Fastq file

Mock_A_TATAGCGA-GACACCGT_L001_R1_001.fastq

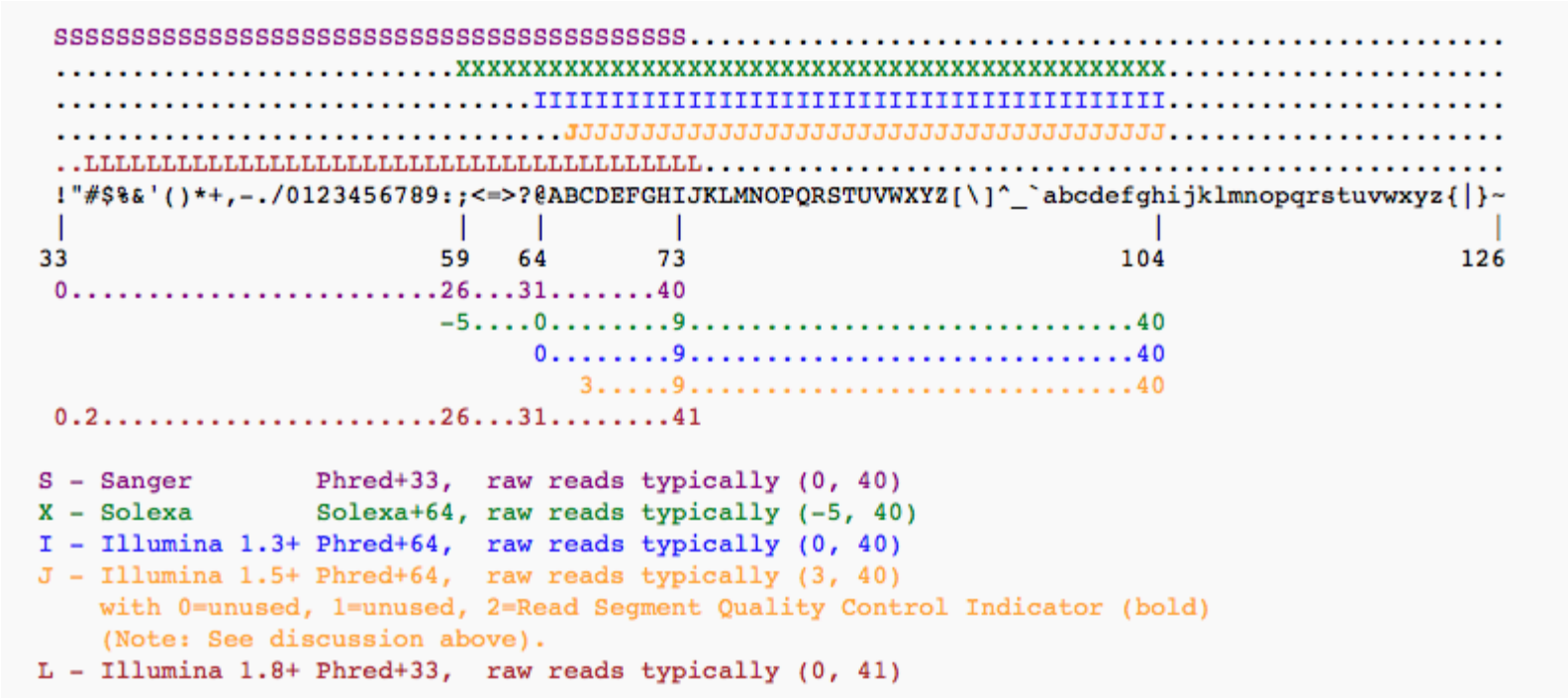
```
@HWI-M01141:63:A4NDL:1:1101:16668:1377 1:N:0:TATAGCGAGACACCGT
NACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGTTA
+
#>>AA>CAABBBGGGGGGGGFFGHFGEGGGFHHHGHFEGCEHHFEGGGGG@EEHHGGGHGHHC
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>>>A??AFAA1BGGEGGAAFGGCA0BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFAEFG1
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT
NACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTA
+
#>>AAABBBABBGGGGGGG?FGHGGGGGHHHHHHHHHGGGGHHHGGGGGGEGGGGGGEGG?FD
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCGGTTGTGCA
+
#>>AABFB@FBBGGGGGGGGGGHGGGGFHHHHHHHHHGGGGHHHGGGGGGGGGGGGGGEE?G4
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>AAAABBFABGGGGGGGECEGHGGEFFHHHHHHHGHGGGGHHHGGGGGGEFGHHGGGGEGHE
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCC
+
```

Info o běhu NGS

Sekvence readu

Kvalita readu

Primární sekvence



Mock_A_TATAGCGA-GACACCGT_L001_R1_001.fastq

```
@HWI-M01141:63:A4NDL:1:1101:16668:1377 1:N:0:TATAGCGAGACACCGT
NACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGTTA
+
#>>AA>CAABBBGGGGGGGFFGHFGEGGGFHHHGHFEGCEHHFEGGGGG@EEHHGGGHGHHC
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>>>A??AFAA1BGGEGGAAFGGCA0BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFAEFG1
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT
NACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTA
+
#>>AAABBBABBGGGGGGG?FGHGGGGGHHHHHHHHHGGGGHHHGGGGGGEGGGGGGEGG?FD
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCGGTTGTGCA
+
#>>AABFB@FBBGGGGGGGGGGHGGGGFHHHHHHHHHGGGGHHHGGGGGGGGGGGGGGEE?G4
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>AAAABBFABGGGGGGGECEGHGGEFFHHHHHHHGHGGGGHHHGGGGGGEFGHHGGGGEGHE
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCC
+
```

Phred 10 – 90% base call accuracy

Phred 20 – 99% base call accuracy

Phred 30 – 99.99% base call accuracy

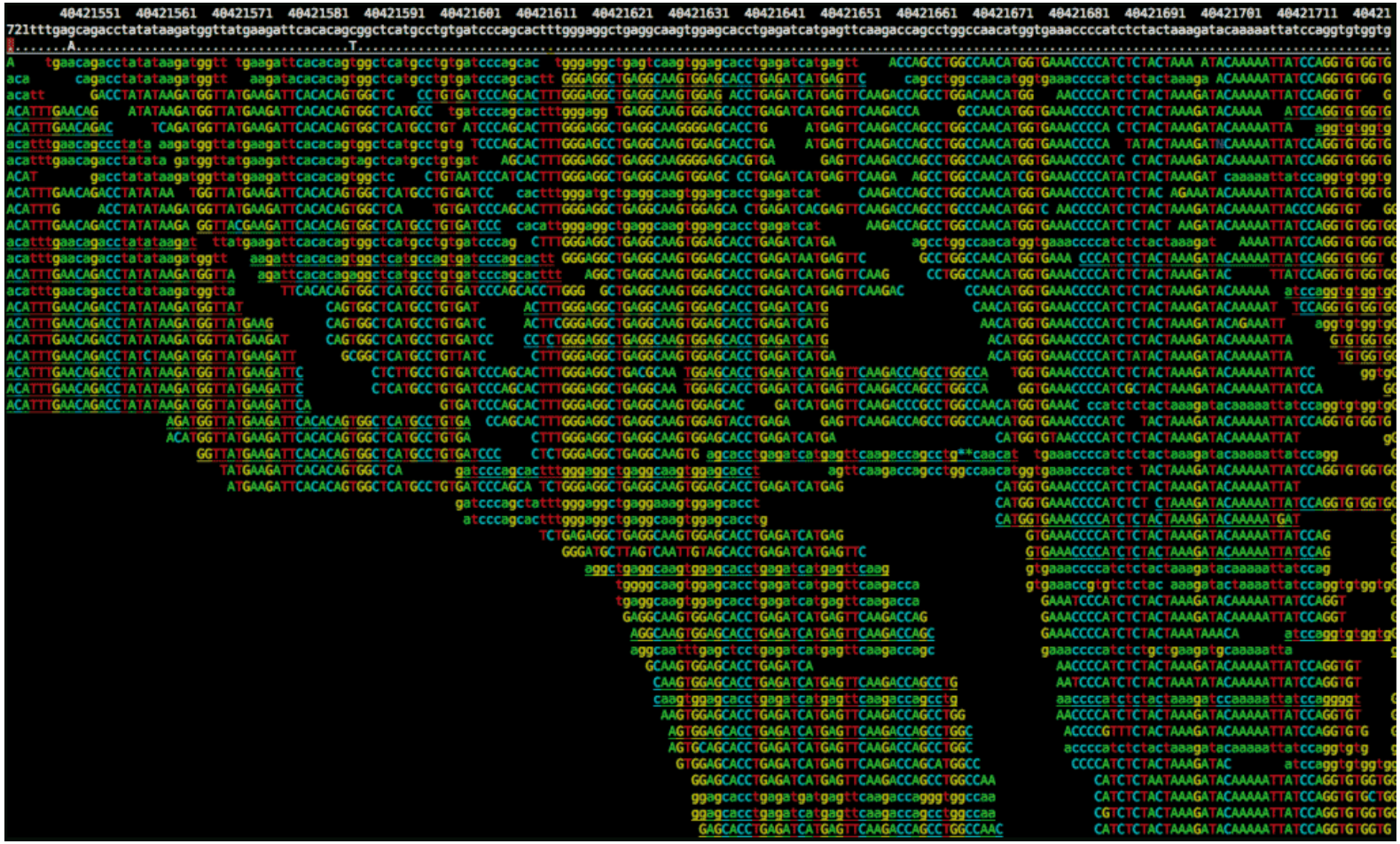
Phred 40 – 99.999% base call accuracy

Analýza (mutací) variant v sekvencích

1. Mapování readů na referenční genom (hg19, hg38, T2T)

sam file – textový soubor, bam file je jeho binární verze

Pro každý vzorek - R1.fastq a R2.fastq (tisíce - miliony readů)



2. Úprava bam souboru
Indexování, realignment indelů, odstranění duplikátů

3. Variant calling – statistické testy, porovnává nalezené rozdíly vůči referenci

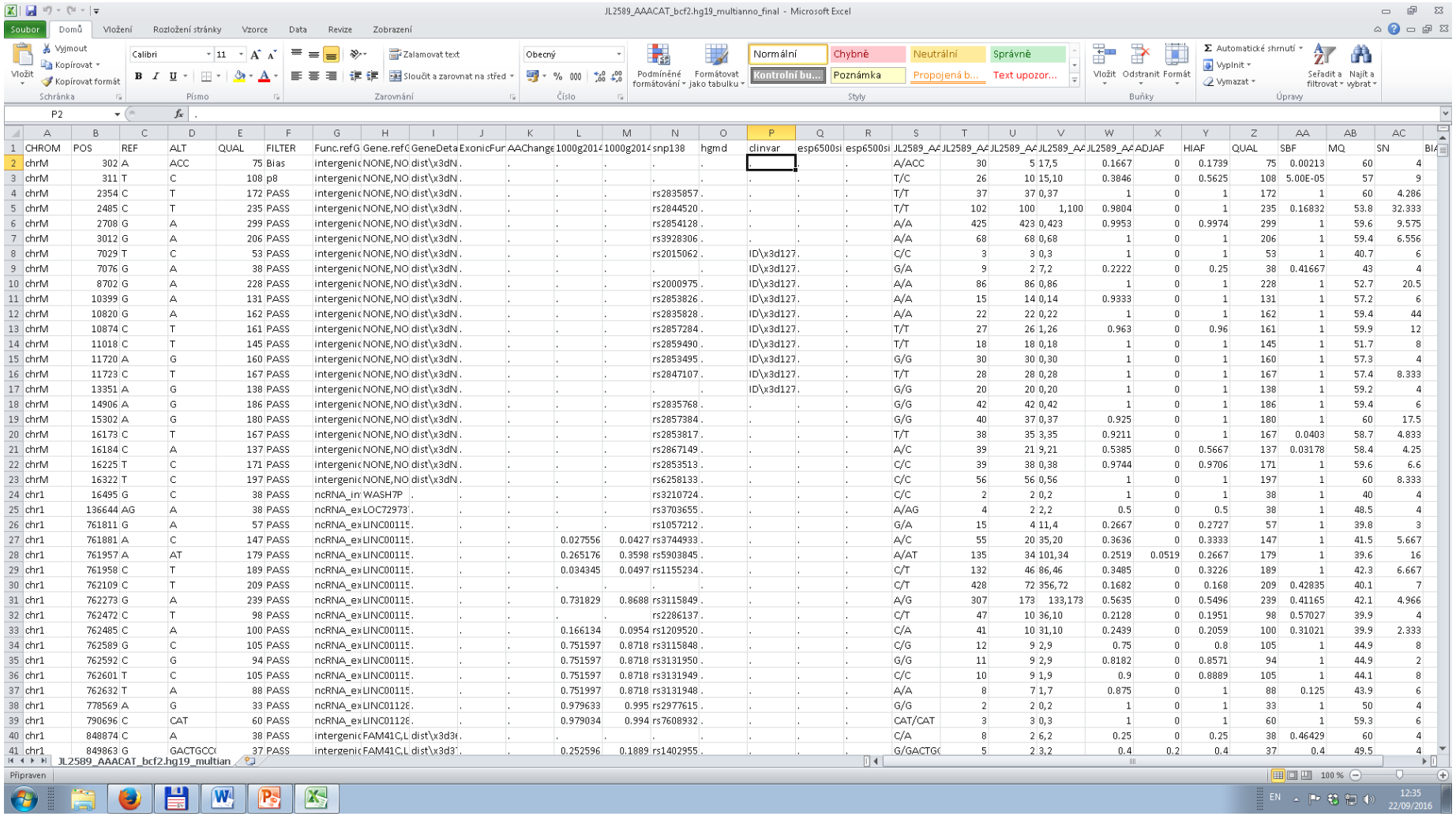
výsledek: vcf file – soubor nalezených genetických variant

Nastavení pro bioinformatickou proceduru např:

- Treshold for allele frequency – 0,2 = 20% (for germline)
- Mapping quality – 30
- Base quality – phred score
- Počet mismatchů v readu
- aj.

4. Anotace … u kodujících úseků přeloží do proteinové sekvence.
Zjistí četnost dané varianty proti databázím, např: 1000genomů, Clinvar, Gnomad…)
Přidá info z Omimu

Výsledný soubor -> xls file



JL2589_AAACAT_bcf2.hg19_multianno_final - Microsoft Excel

CHROM	POS	REF	ALT	QUAL	FILTER	Func.refG	Gene.refG	GeneDeta	ExonicFur	AAChange	1000g2014	1000g2014	snp138	hgmd	clinvar	esp6500si	esp6500si	JL2589_AA	JL2589_AA	JL2589_AA	JL2589_AA	JL2589_AA	ADJAF	HIAF	QUAL	SBF	MQ	SN	BIA
chrM	302	A	ACC	75	Bias	intergenic	NONE,NO	dist\x3dN										A/ACC	30	5	17,5		0	0.1739	75	0.00213	60	4	
chrM	311	T	C	108	p8	intergenic	NONE,NO	dist\x3dN										T/C	26	10	15,10		0	0.5625	108	5.00E-05	57	9	
chrM	2354	C	T	172	PASS	intergenic	NONE,NO	dist\x3dN					rs2835857					T/T	37	37	0,37		1	0	1	172	1	60	4.286	
chrM	2485	C	T	235	PASS	intergenic	NONE,NO	dist\x3dN					rs2844520					T/T	102	100	1,100		0.9804	0	1	235	0.16832	53.8	32.333	
chrM	2708	G	A	299	PASS	intergenic	NONE,NO	dist\x3dN					rs2854128					A/A	425	423	0,423		0.9953	0	0.9974	299	1	59.6	9.575	
chrM	3012	G	A	206	PASS	intergenic	NONE,NO	dist\x3dN					rs3928306					A/A	68	68	0,68		1	0	1	206	1	59.4	6.556	
chrM	7029	T	C	53	PASS	intergenic	NONE,NO	dist\x3dN					rs2015062		ID\x3d127			C/C	3	3	0,3		1	0	1	53	1	40.7	6	
chrM	7076	G	A	38	PASS	intergenic	NONE,NO	dist\x3dN							ID\x3d127			G/A	9	2	7,2		0.2222	0	0.25	38	0.41667	43	4	
chrM	8702	G	A	228	PASS	intergenic	NONE,NO	dist\x3dN					rs2000975		ID\x3d127			A/A	86	86	0,86		1	0	1	228	1	52.7	20.5	
chrM	10399	G	A	131	PASS	intergenic	NONE,NO	dist\x3dN					rs2853826		ID\x3d127			A/A	15	14	0,14		0.9333	0	1	131	1	57.2	6	
chrM	10820	G	A	162	PASS	intergenic	NONE,NO	dist\x3dN					rs2835828		ID\x3d127			A/A	22	22	0,22		1	0	1	162	1	59.4	44	
chrM	10874	C	T	161	PASS	intergenic	NONE,NO	dist\x3dN					rs2857284		ID\x3d127			T/T	27	26	1,26		0.963	0	0.96	161	1	59.9	12	
chrM	11018	C	T	145	PASS	intergenic	NONE,NO	dist\x3dN					rs2859490		ID\x3d127			T/T	18	18	0,18		1	0	1	145	1	51.7	8	
chrM	11720	A	G	160	PASS	intergenic	NONE,NO	dist\x3dN					rs2853495		ID\x3d127			G/G	30	30	0,30		1	0	1	160	1	57.3	4	
chrM	11723	C	T	167	PASS	intergenic	NONE,NO	dist\x3dN					rs2847107		ID\x3d127			T/T	28	28	0,28		1	0	1	167	1	57.4	8.333	
chrM	13351	A	G	138	PASS	intergenic	NONE,NO	dist\x3dN							ID\x3d127			G/G	20	20	0,20		1	0	1	138	1	59.2	4	
chrM	14906	A	G	186	PASS	intergenic	NONE,NO	dist\x3dN					rs2835768					G/G	42	42	0,42		1	0	1	186	1	59.4	6	
chrM	15302	A	G	180	PASS	intergenic	NONE,NO	dist\x3dN					rs2857384					G/G	40	37	0,37		0.925	0	1	180	1	60	17.5	
chrM	16173	C	T	167	PASS	intergenic	NONE,NO	dist\x3dN					rs2853817					T/T	38	35	3,35		0.9211	0	1	167	0.0403	58.7	4.833	
chrM	16184	C	A	137	PASS	intergenic	NONE,NO	dist\x3dN					rs2867149					A/C	39	21	9,21		0.5385	0	0.5667	137	0.03178	58.4	4.25	
chrM	16225	T	C	171	PASS	intergenic	NONE,NO	dist\x3dN					rs2853513					C/C	39	38	0,38		0.9744	0	0.9706	171	1	59.6	6.6	
chrM	16322	T	C	197	PASS	intergenic	NONE,NO	dist\x3dN					rs6258133					C/C	56	56	0,56		1	0	1	197	1	60	8.333	
chr1	16495	G	C	38	PASS	ncRNA_in	WASH7P						rs3210724					C/C	2	2	0,2		1	0	1	38	1	40	4	
chr1	136644	AG	A	38	PASS	ncRNA_ex	LOC72973						rs3703655					A/AG	4	2	2,2		0.5	0	0.5	38	1	48.5	4	
chr1	761811	G	A	57	PASS	ncRNA_ex	LINC0011						rs1057212					G/A	15	4	11,4		0.2667	0	0.2727	57	1	39.8	3	
chr1	761881	A	C	147	PASS	ncRNA_ex	LINC0011				0.027556	0.0427	rs3744933					A/C	55	20	35,20		0.3636	0	0.3333	147	1	41.5	5.667	
chr1	761957	A	AT	179	PASS	ncRNA_ex	LINC0011				0.265176	0.3598	rs5903845					A/AT	135	34	101,34		0.2519	0.0519	0.2667	179	1	39.6	16	
chr1	761958	C	T	189	PASS	ncRNA_ex	LINC0011				0.034345	0.0497	rs1155234					C/T	132	46	86,46		0.3485	0	0.3226	189	1	42.3	6.667	
chr1	762109	C	T	209	PASS	ncRNA_ex	LINC0011											C/T	428	72	356,72		0.1682	0	0.168	209	0.42835	40.1	7	
chr1	762273	G	A	239	PASS	ncRNA_ex	LINC0011				0.731829	0.8688	rs3115849					A/G	307	173	133,173		0.5635	0	0.5496	239	0.41165	42.1	4.966	
chr1	762472	C	T	98	PASS	ncRNA_ex	LINC0011						rs2286137					C/T	47	10	36,10		0.2128	0	0.1951	98	0.57027	39.9	4	
chr1	762485	C	A	100	PASS	ncRNA_ex	LINC0011				0.166134	0.0954	rs1209520					C/A	41	10	31,10		0.2439	0	0.2059	100	0.31021	39.9	2.333	
chr1	762589	G	C	105	PASS	ncRNA_ex	LINC0011				0.751597	0.8718	rs3115848					C/G	12	9	2,9		0.75	0	0.8	105	1	44.9	8	
chr1	762592	C	G	94	PASS	ncRNA_ex	LINC0011				0.751597	0.8718	rs3131950					G/G	11	9	2,9		0.8182	0	0.8571	94	1	44.9	2	
chr1	762601	T	C	105	PASS	ncRNA_ex	LINC0011				0.751597	0.8718	rs3131949					C/C	10	9	1,9		0.9	0	0.8889	105	1	44.1	8	
chr1	762632	T	A	88	PASS	ncRNA_ex	LINC0011				0.751997	0.8718	rs3131948					A/A	8	7	1,7		0.875	0	1	88	0.125	43.9	6	
chr1	778569	A	G	33	PASS	ncRNA_ex	LINC0112				0.979633	0.995	rs2977615					G/G	2	2	0,2		1	0	1	33	1	50	4	
chr1	790696	C	CAT	60	PASS	ncRNA_ex	LINC0112				0.979034	0.994	rs7608932					CAT/CAT	3	3	0,3		1	0	1	60	1	59.3	6	
chr1	848874	C	A	38	PASS	intergenic	FAM41C,L	dist\x3d3										C/A	8	2	6,2		0.25	0	0.25	38	0.46429	60	4	
chr1	849863	G	GACTGCC	37	PASS	intergenic	FAM41C,L	dist\x3d3			0.252596	0.1889	rs1402955					G/GACTG	5	2	3,2		0.4	0.2	0.4	37	0.4	49.5	4	

JL2589_AAACAT_bcf2.hg19_multian

Připraven

100 %

EN 12:35 22/09/2016

Proteinové sekvence????

Proteinové sekvence

Proteinové sekvence jsou odvozeny přepisem kódujících DNA sekvencí (CDS)

Každý region DNA má 6 čtecích rámců

5' 3'

atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

Čtecí rámec začíná atg (Met) u většiny druhů a končí stop kodonem (taa, tag or tga).

Proteinové sekvence

Proteinové sekvence jsou odvozeny přepisem kódujících DNA sekvencí (CDS)

Každý region DNA má 6 čtecích rámců, 3 v každém směru

5' 3'

atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
M P K L N S V E G F S S F E D D V *

2 tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
C P S * I A * R G F H H L R T M Y

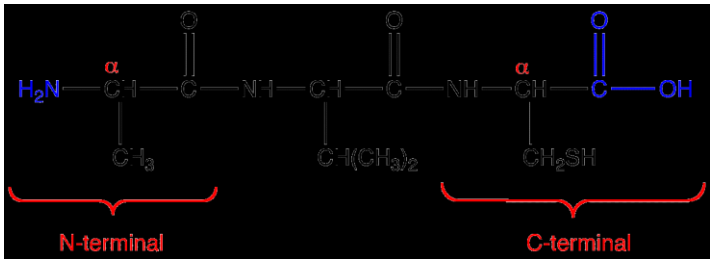
3 gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
A Q A E * R R G V F I I * G R C I

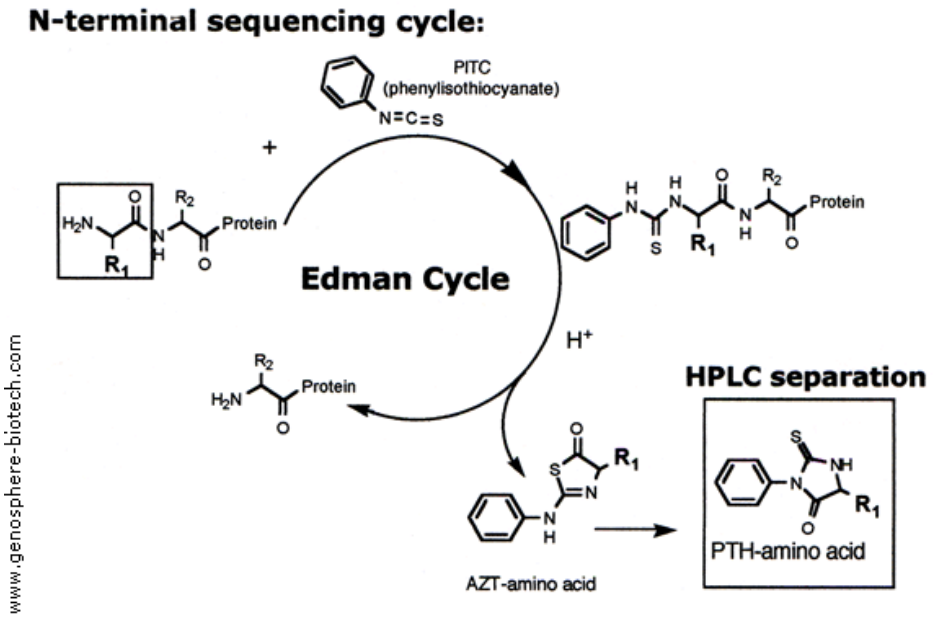
Čtecí rámec začíná atg (Met) u většiny druhů a konční stop kodonem (taa, tag or tga).

Proteinové sekvence

Další zdroje proteinových sekvencí

- přímé sekvenování Edmanova degradace
(odbourávání aminokyselin z N-konce a jejich identifikace)





-

+ MS/MS

„nepřímé“ sekvenování MS/MS experimenty
Získaná spektra se porovnávají vůči databázi

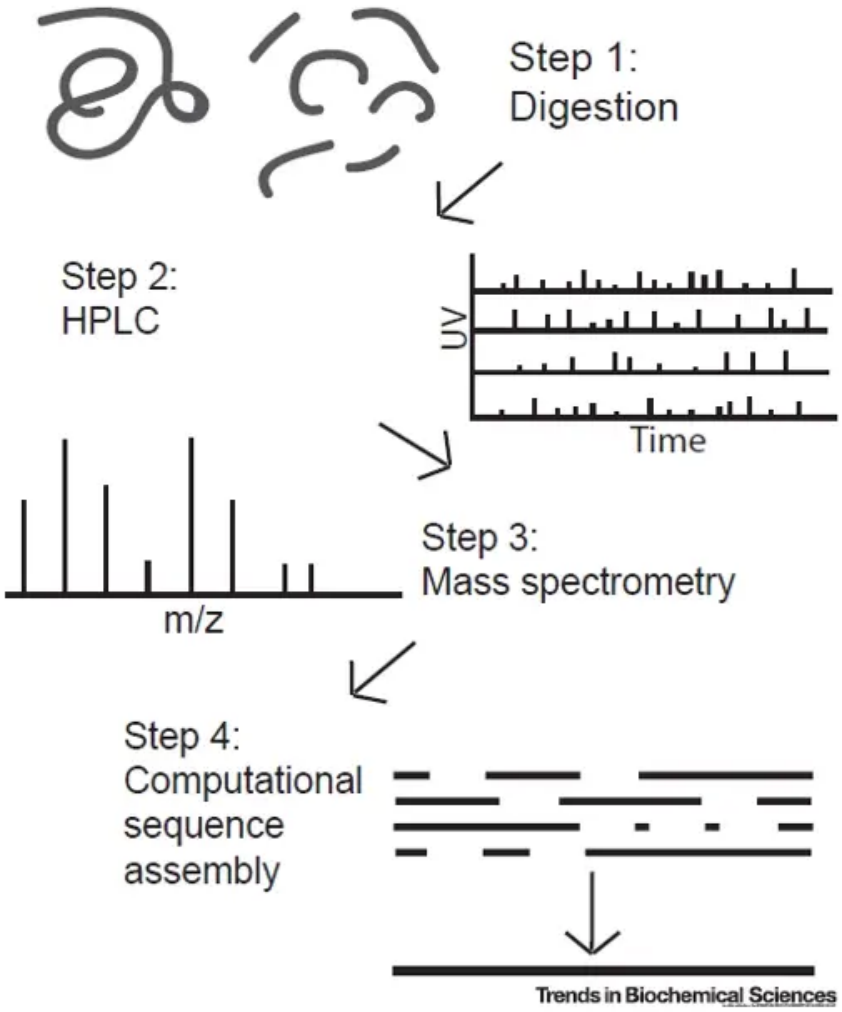
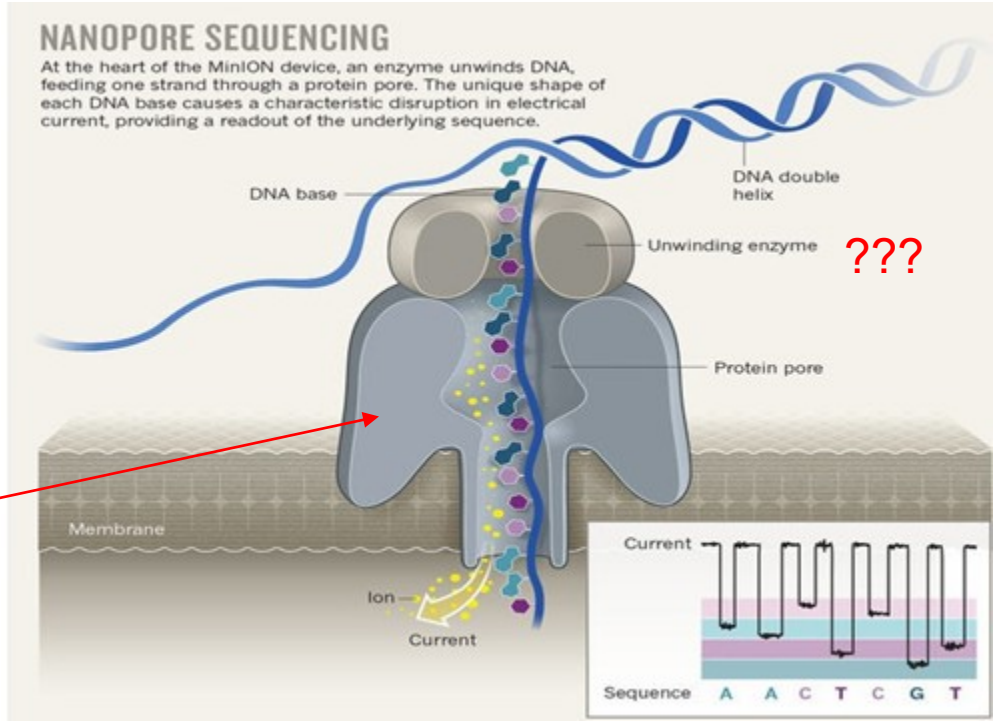
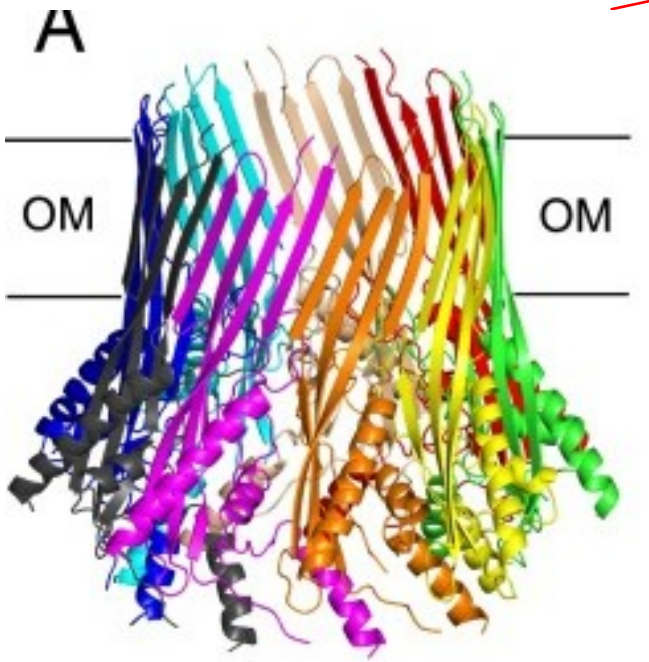


Figure 1. Typical Mass Spectrometry Workflow [1]

DNA
Nanopore sequencing

pore-forming proteins, such as an engineered version of CsgG





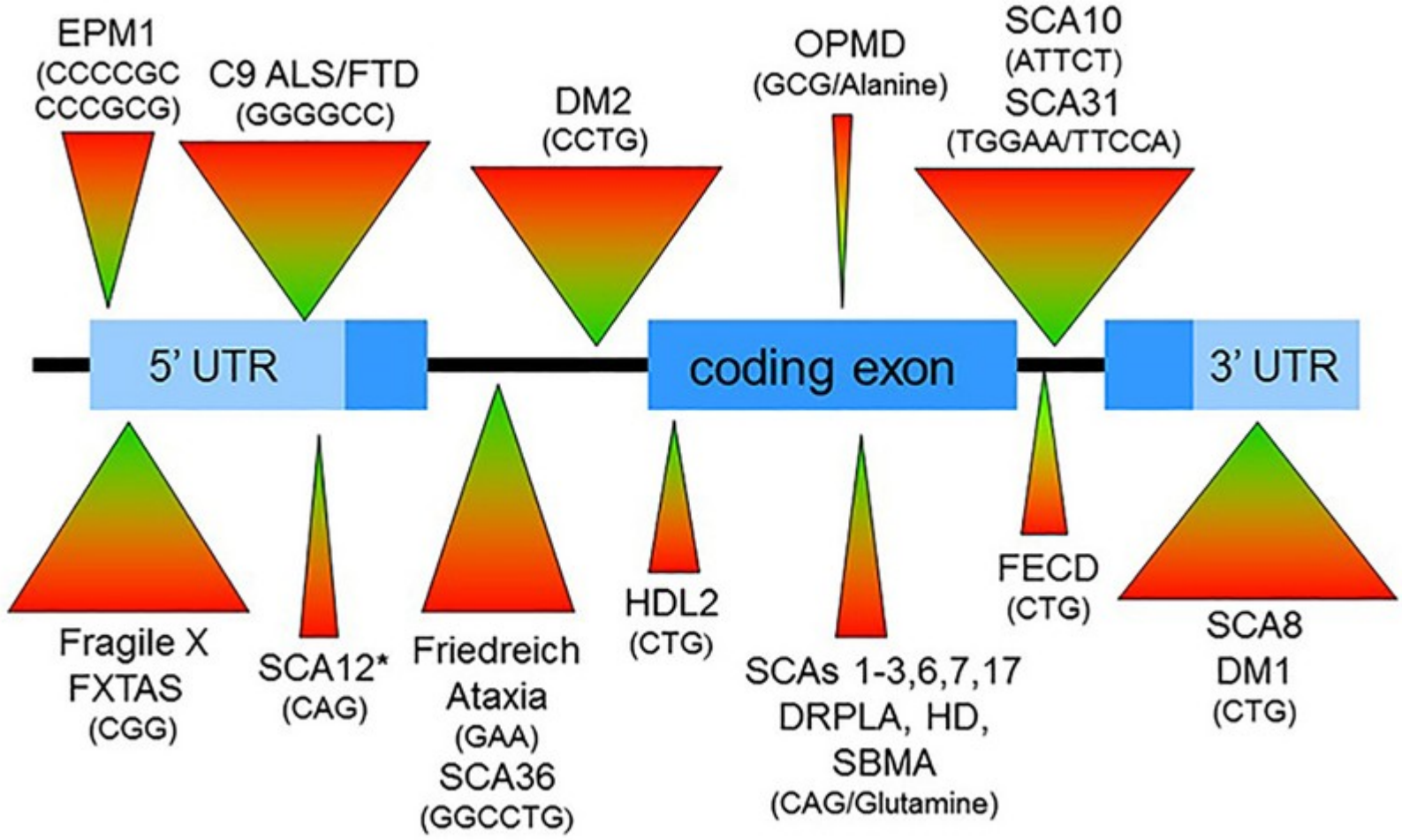
https://www.youtube.com/watch?v=RcP85JHLmnI&ab_channel=OxfordNanoporeTechnologies

Detekce expanzi pomocí nanoporového sekvenování

Repeat expansions cause many neurologic diseases

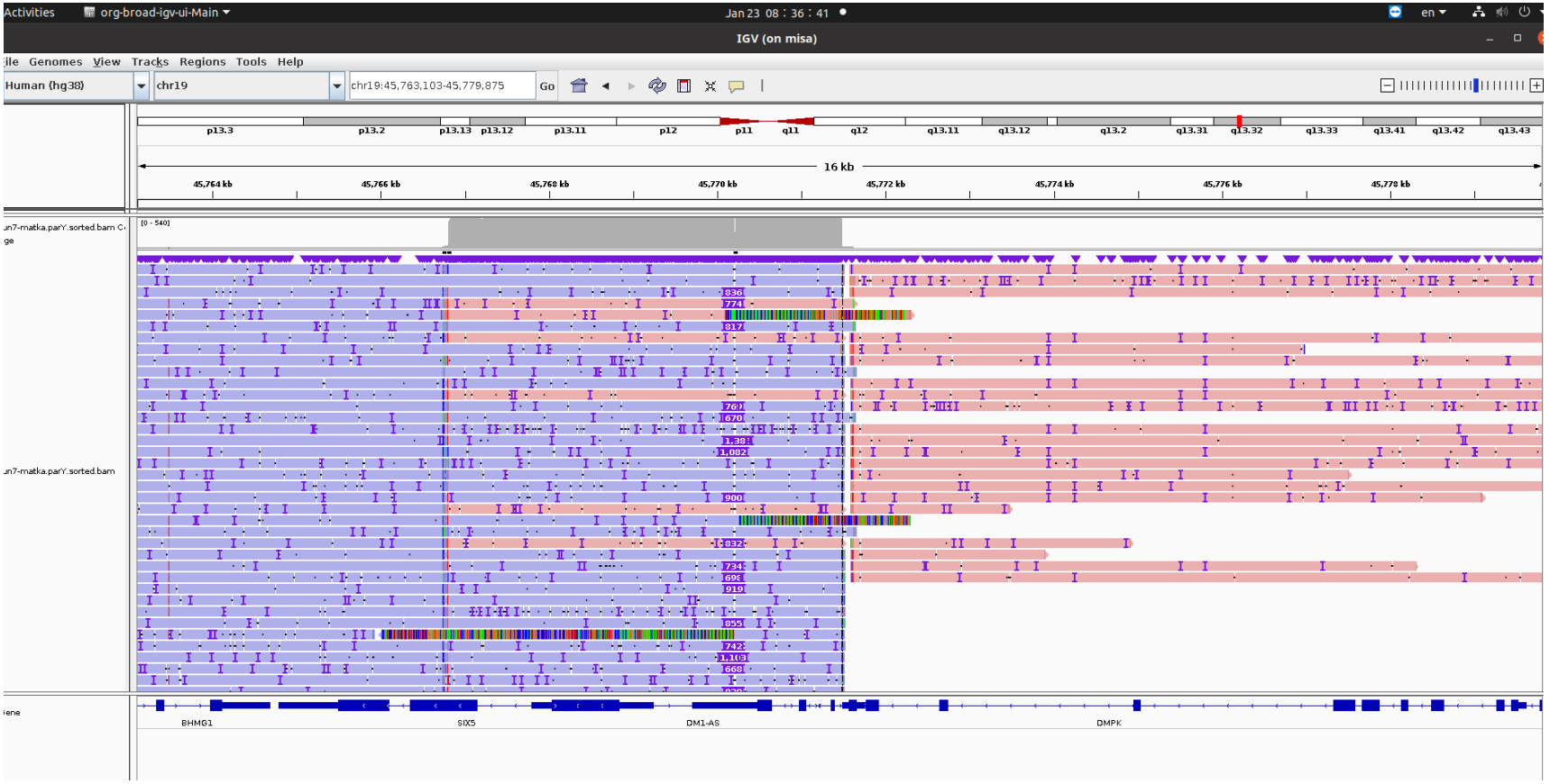
- CAG – at least 10 diseases (Huntington disease, spinal and bulbar muscular atrophy, dentatorubr pallidoluysian atrophy and seven SCAs)
- CGG – fragile X, fragile X tremor ataxia syndrome, other fragile sites (GCC, CCG)
- CTG – myotonic dystrophy type 1, Huntington disease-like 2, spinocerebellar ataxia type 8, Fuchs corneal dystrophy
- GAA – Friedreich ataxia
- GCC – FRAXE mental retardation
- GCG – oculopharyngeal muscular dystrophy
- CCTG – myotonic dystrophy type 1
- ATTCT – spinocerebellar ataxia type 10
- TGGAA – spinocerebellar ataxia type 31
- GGCCTG – spinocerebellar ataxia type 36
- GGGGCC – C9ORF72 frontotemporal dementia/amyotrophic lateral sclerosis
- CCCCGCCCCGCG – EPM1 (myoclonic epilepsy)

Lokalizace expanzí v genu

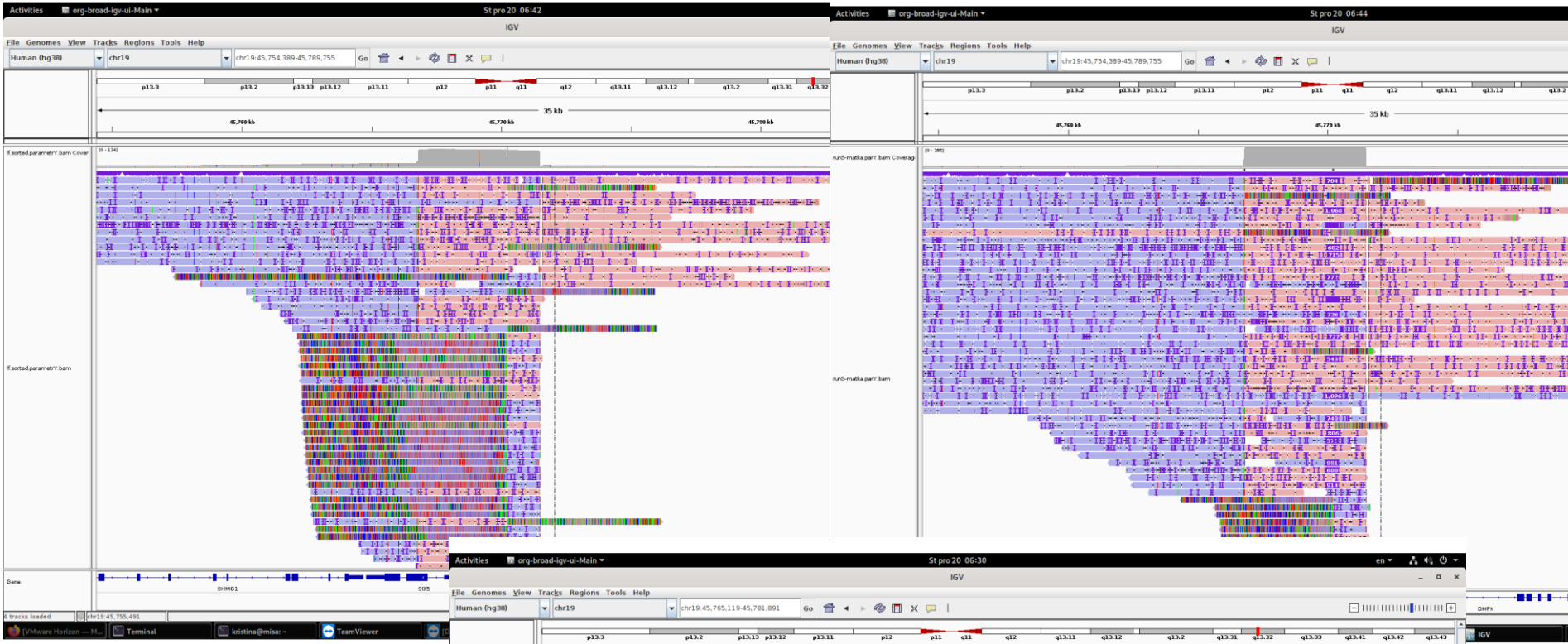


markedly different sizes of pathogenic expansions are suggested by the varying sized triangles

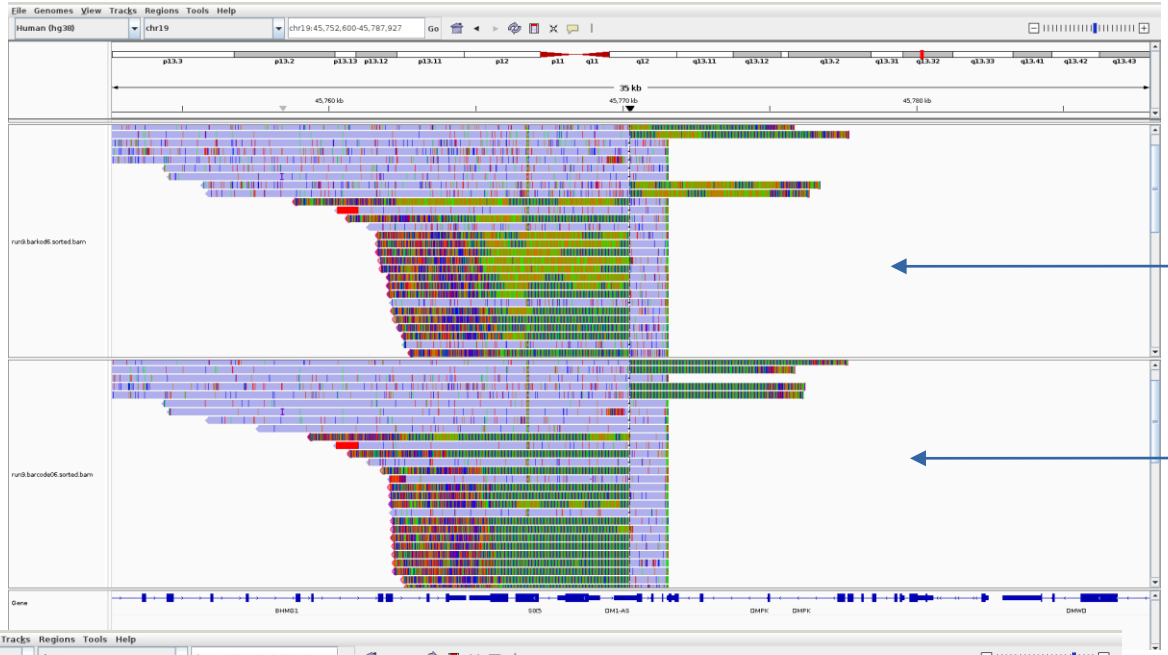
Expanze repetic CTG v genu DMPK – myotonicka dystrofie



Expanze repetic CTG v genu DMPK – myotonicka dystrofie

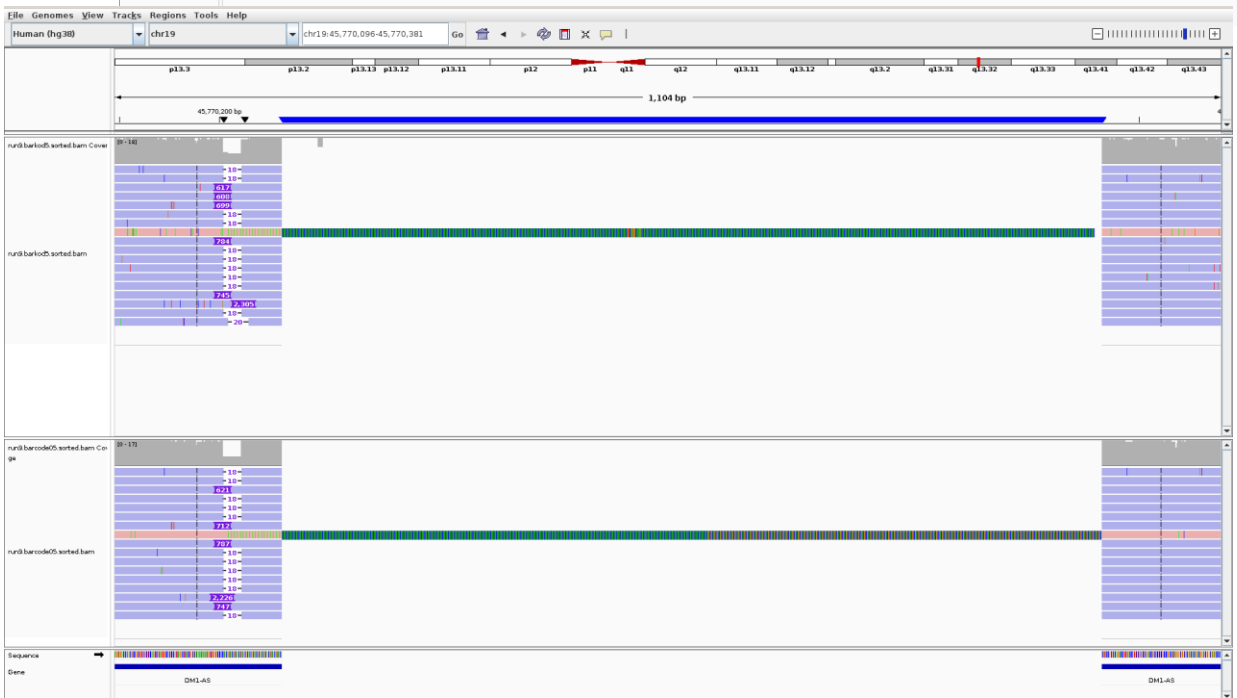


Problém při basecallingu repetitivních oblastí



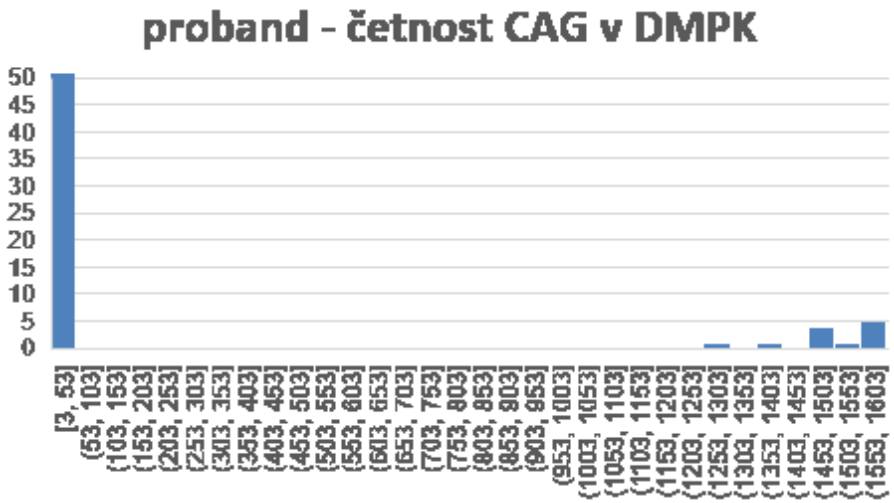
Základní basecalling

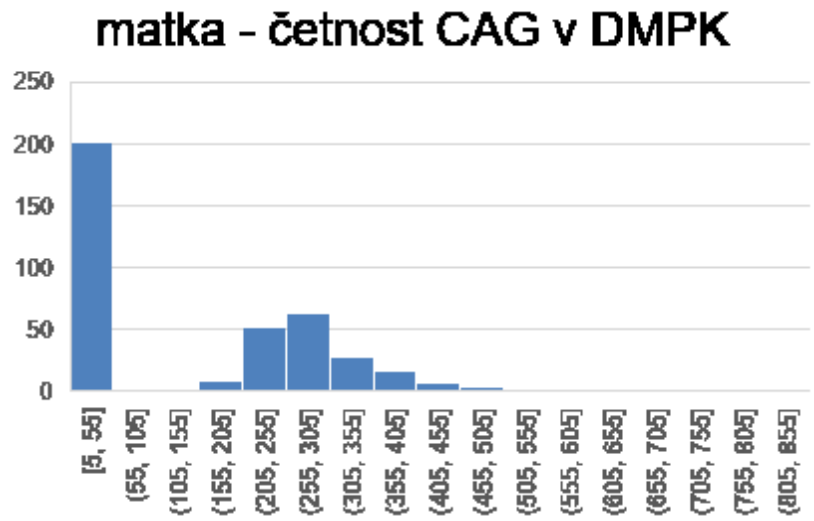
Zpřesněný basecalling

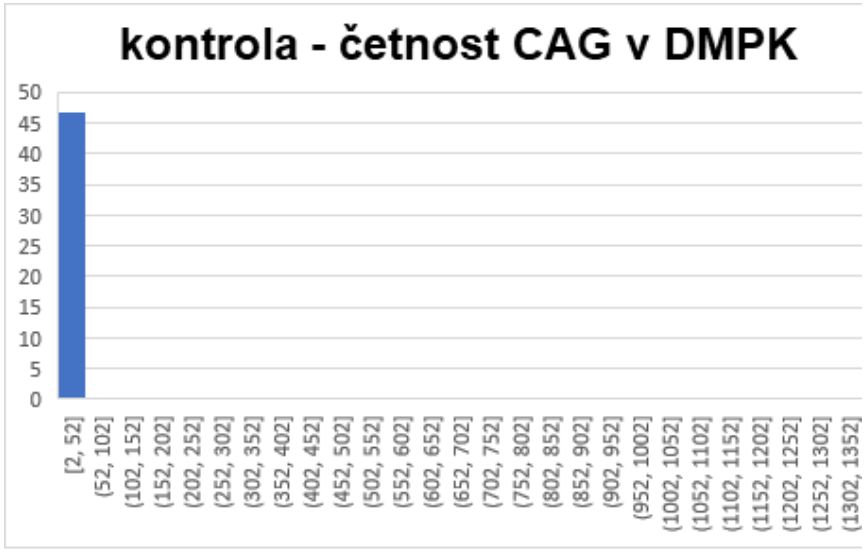


Základní basecalling

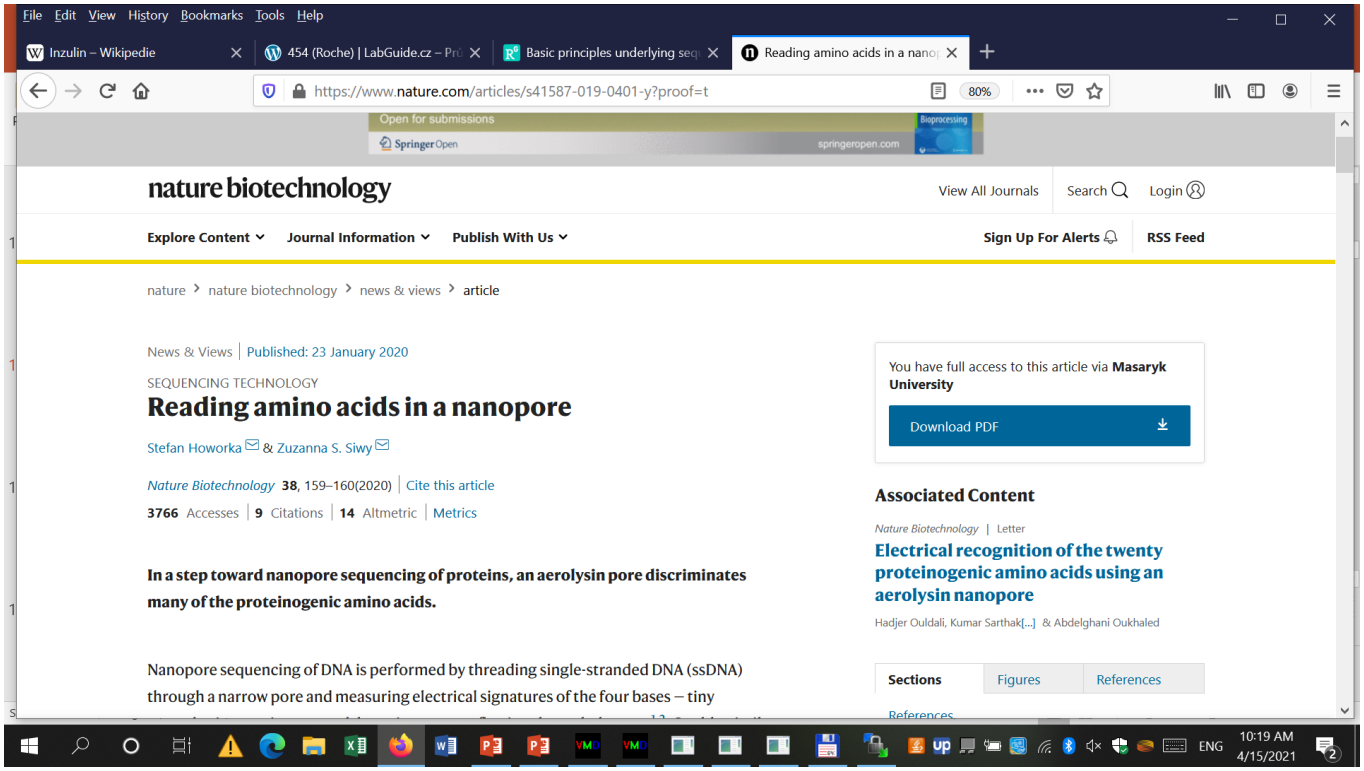
Zpřesněný basecalling


proband - četnost CAG v DMPK
50
45
40
35
30
25
20
15
10
5
0
[3, 53]
(53, 103]
(103, 153]
(153, 203]
(203, 253]
(253, 303]
(303, 353]
(353, 403]
(403, 453]
(453, 503]
(503, 553]
(553, 603]
(603, 653]
(653, 703]
(703, 753]
(753, 803]
(803, 853]
(853, 903]
(903, 953]
(953, 1003]
(1003, 1053]
(1053, 1103]
(1103, 1153]
(1153, 1203]
(1203, 1253]
(1253, 1303]
(1303, 1353]
(1353, 1403]
(1403, 1453]
(1453, 1503]
(1503, 1553]
(1553, 1603]


matka - četnost CAG v DMPK
250
200
150
100
50
0
[5, 55]
(55, 105]
(105, 155]
(155, 205]
(205, 255]
(255, 305]
(305, 355]
(355, 405]
(405, 455]
(455, 505]
(505, 555]
(555, 605]
(605, 655]
(655, 705]
(705, 755]
(755, 805]
(805, 855]


kontrola - četnost CAG v DMPK
50
45
40
35
30
25
20
15
10
5
0
[2, 52]
(52, 102]
(102, 152]
(152, 202]
(202, 252]
(252, 302]
(302, 352]
(352, 402]
(402, 452]
(452, 502]
(502, 552]
(552, 602]
(602, 652]
(652, 702]
(702, 752]
(752, 802]
(802, 852]
(852, 902]
(902, 952]
(952, 1002]
(1002, 1052]
(1052, 1102]
(1102, 1152]
(1152, 1202]
(1202, 1252]
(1252, 1302]
(1302, 1352]

Protein Nanopore sequencing???



Main issues ??

Unlike 4 nucleotides that make up DNA, proteins are composed of 20 amino acids, each with a different charge and virtually 100s of potential post-translational modifications [2]. Moreover, it has proved difficult to control the translocation of peptides across a nanopore as enzymes capable of ratcheting the peptide at a controlled rate have been difficult to identify and conjugate with existing nanopores.

- Mutace/varianty (odchylky od referenčnÍ sekvence)

NGS:

panel genů

Exom → WES – whole exom sequencing (2% genomu)

Genom → WGS – whole genome sequencing

Sekvenování identifikuje (známé) varianty nebo odhalí nové

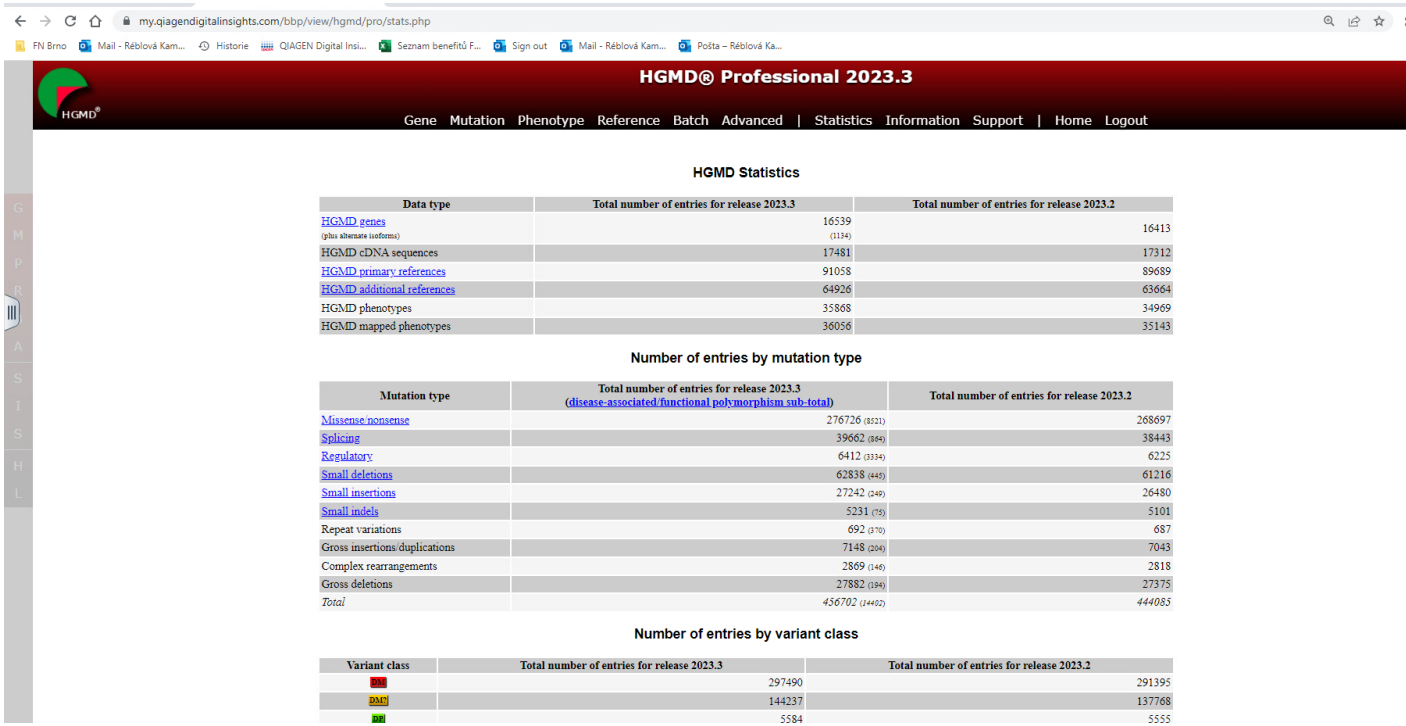
(jsou skutečně kauzální???)

Je nalezená mutace popsaná????

mutační databáze: germinální mutací (**HGMD**), somatické (COSMIC)

Specializované: (**PAHdb**, **Leiden**, **IARC TP53 Database – somatické/germinální**)

Human Gene Mutation Database (HGMD) – databáze mutací lidských genů v kódujících oblastech. *(vychází z primárních dat, manuální a Automatické prohledávání)*



Data type	Total number of entries for release 2023.3	Total number of entries for release 2023.2
HGMD genes (plus alternate isoforms)	16539 (1134)	16413
HGMD cDNA sequences	17481	17312
HGMD primary references	91058	89689
HGMD additional references	64926	63664
HGMD phenotypes	35868	34969
HGMD mapped phenotypes	36056	35143

Mutation type	Total number of entries for release 2023.3 (disease-associated/functional polymorphism sub-total)	Total number of entries for release 2023.2
Missense/nonsense	276726 (8521)	268697
Splicing	39662 (864)	38443
Regulatory	6412 (3334)	6225
Small deletions	62838 (445)	61216
Small insertions	27242 (249)	26480
Small indels	5231 (75)	5101
Repeat variations	692 (370)	687
Gross insertions/duplications	7148 (204)	7043
Complex rearrangements	2869 (146)	2818
Gross deletions	27882 (194)	27375
Total	*456702 (14402)*	*444085*

Variant class	Total number of entries for release 2023.3	Total number of entries for release 2023.2
DM	297490	291395
DM?	144237	137768
DP	5584	5555

Co když mutace není popsaná?

Jak rozlišit mezi:

Bežnou variantou [single nucleotide polymorphisms (SNPs)]

a

kauzální mutací?

Přístupy studia efektů missense mutací:

- Funkční analýzy (exprese proteinu, stabilita, specifické charakteristiky)
  časově a finančně náročné

- Automatické *in silico* programy (SIFT, PolyPhen, SNPs3D, FODLX..)
  rychlé a spolehlivost průměrně 70 %

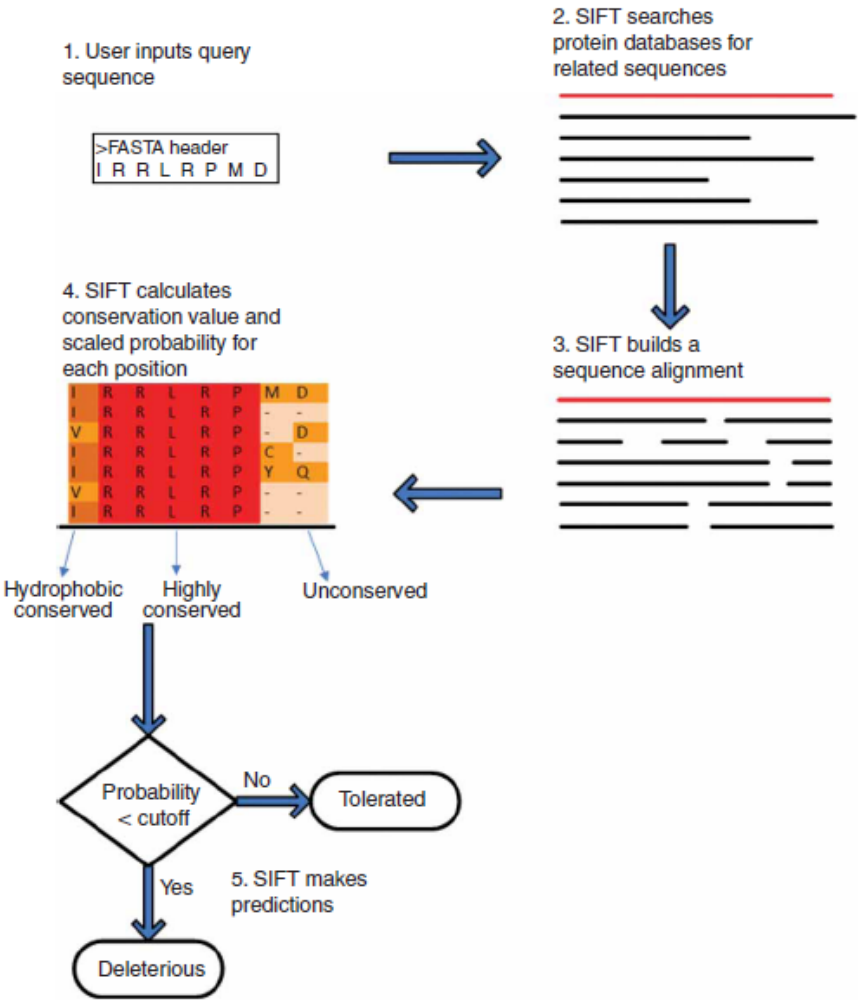
- Strukturní analýza pomocí molekulového modelování
  časově a finančně středně náročné, podhled do mechanismu účinky na atomární úrovni, je to ale stále predikce

Automatické analýzy missense mutací pomocí bioinformatických nástrojů

SIFT - Sorting Tolerant From Intolerant (http://sift.jcvi.org)

SIFT assumes that important positions in a protein sequence have been conserved throughout evolution and therefore substitutions at these positions may affect protein function.

Using sequence homology, SIFT predicts the effects of all possible substitutions at each position in the protein sequence.

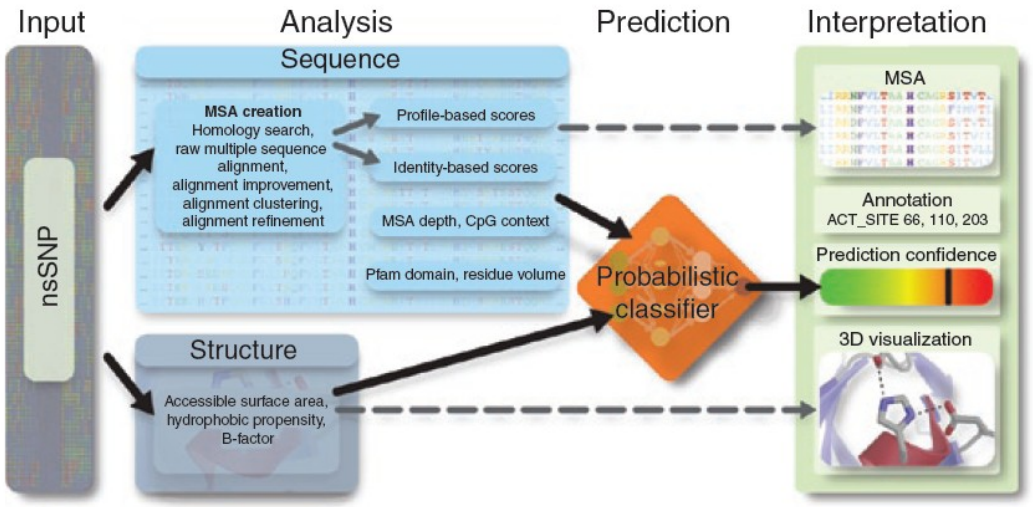


Úspěšnost predikce pro skupinu kauzálních mutací 69%.
Falešně pozitivní 19 %.

Automatické analýzy missense mutací pomocí bioinformatických nástrojů

Polyphen -2

PolyPhen-2 uses eight sequence-based and three structure-based predictive features



Structural features. Three additional features were selected for proteins with known 3D structures: 1) the accessible surface area of the wild-type amino acid residue, 2) the change in the hydrophobic propensity in the form of “knowledge-based potential”, and 3) crystallographic B-factor reflecting conformational mobility of the wild-type amino acid residue15.

Úspěšnost predikce 73%

Falešně pozitivní ~ 20 %.

Automatické analýzy missense mutací pomocí bioinformatických nástrojů

FOLDX (http://foldx.embl.de/)

Změna celkové volné energie
$\Delta\Delta G = \Delta Gmut - \Delta Gwt$

$\Delta Gmut$ - free energy difference between the folded and unfolded states

Empirical force field that was developed for the rapid evaluation of the effect of mutations on the stability, folding and dynamics of proteins and nucleic acids

$$\begin{aligned}\Delta G = {} & a \cdot \Delta G_{\mathrm{vdw}} + b \cdot \Delta G_{\mathrm{solvH}} + c \cdot \Delta G_{\mathrm{solvP}} + d \cdot \Delta G_{\mathrm{wb}} \\ & + e \cdot \Delta G_{\mathrm{hbond}} + f \cdot \Delta G_{\mathrm{el}} + g \cdot \Delta G_{\mathrm{kon}} + h \cdot \mathrm{T}\Delta \mathrm{S}_{\mathrm{mc}} \\ & + k \cdot \mathrm{T}\Delta \mathrm{S}_{\mathrm{sc}} + l \cdot \Delta G_{\mathrm{clash}}.\end{aligned}$$

(a . . . l) are relative weights of the different energy terms used for the free energy calculation

Výpočty na lokálním PC ne přes web rozhraní nyní.

Lze měnit počáteční nastavení: T(K), c(M)…

Automatické analýzy missense mutací pomocí bioinformatických nástrojů

FOLDX (http://foldx.embl.de/)

FoldX component	*Energy (kcal.mol-1)*
Backbone Hbond	-48.30
Sidechain Hbond	-18.40
Van der Waals	-62.33
Electrostatics	-2.79
Solvation Polar	84.34
Solvation Hydrophobic	-76.71
Van de Waals clashes	0.00
entropy side chain	34.93
entropy main chain	81.94
sloop_entropy	0.00
mloop_entropy	0.00
cis_bond	0.00
torsional clash	0.40
backbone clash	1.20
helix dipole	0.00
water bridge	0.00
disulfide 0.00	
electrostatic kon	0.00
partial covalent bonds	0.00
Total	-5.72

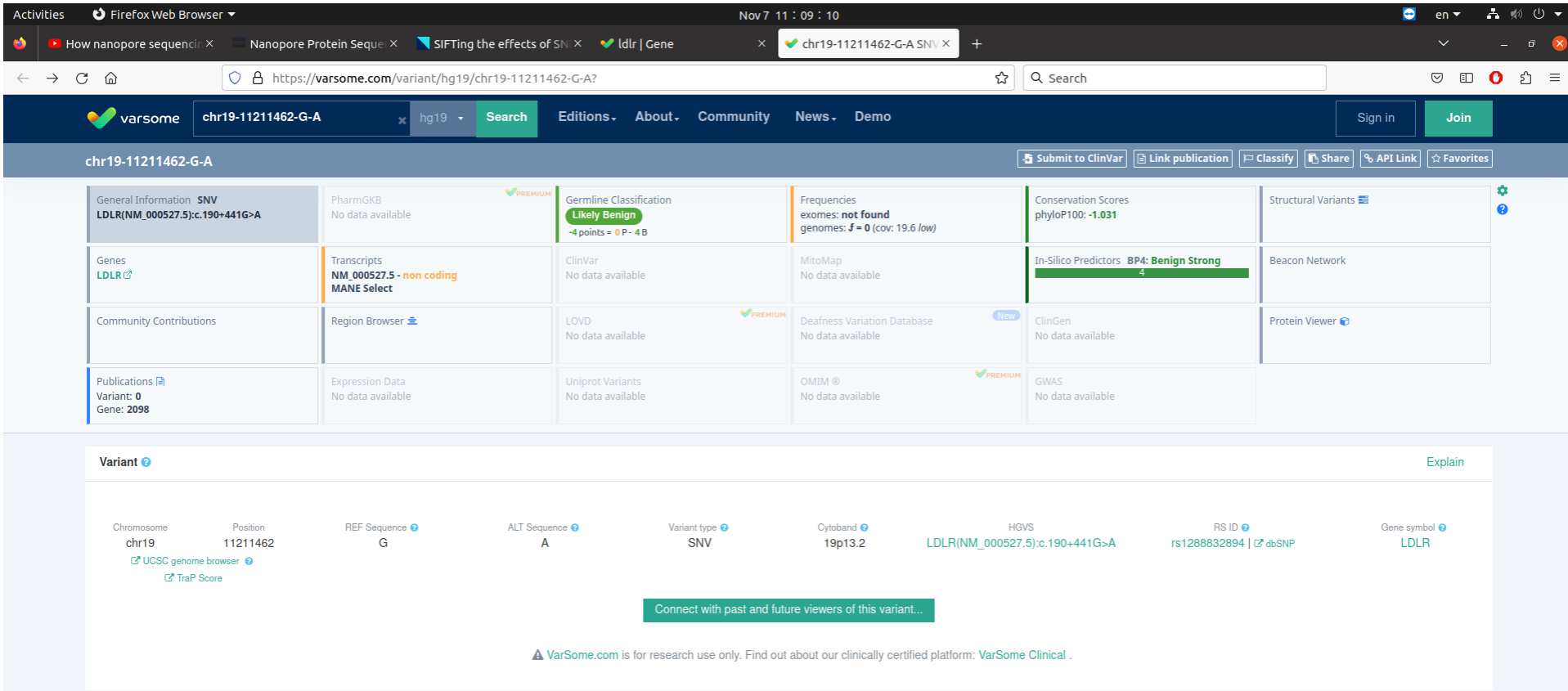
ΔΔG(change) = ΔG(MT) - ΔG(WT)

ΔΔG(change) > 1kcal/mol: the mutation is destabilizing

ΔΔG(change) < -1 kcal/mol: the mutation is stabilizing

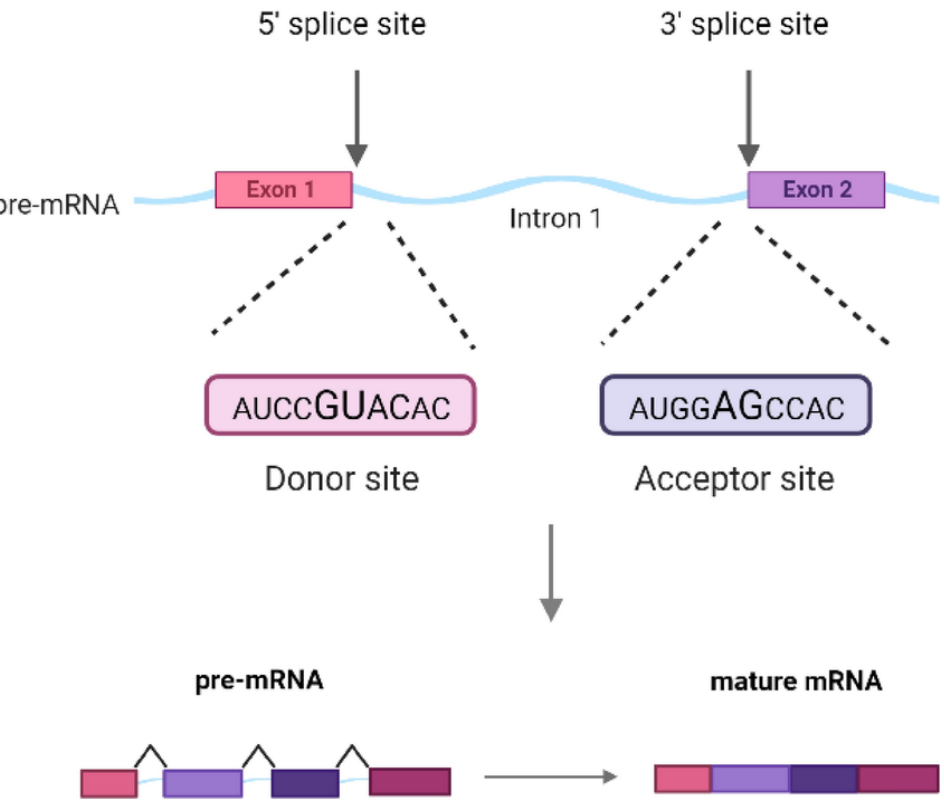
Úspěšnost predikce 60%

Konsensus mutačních programů….. např. Varsome



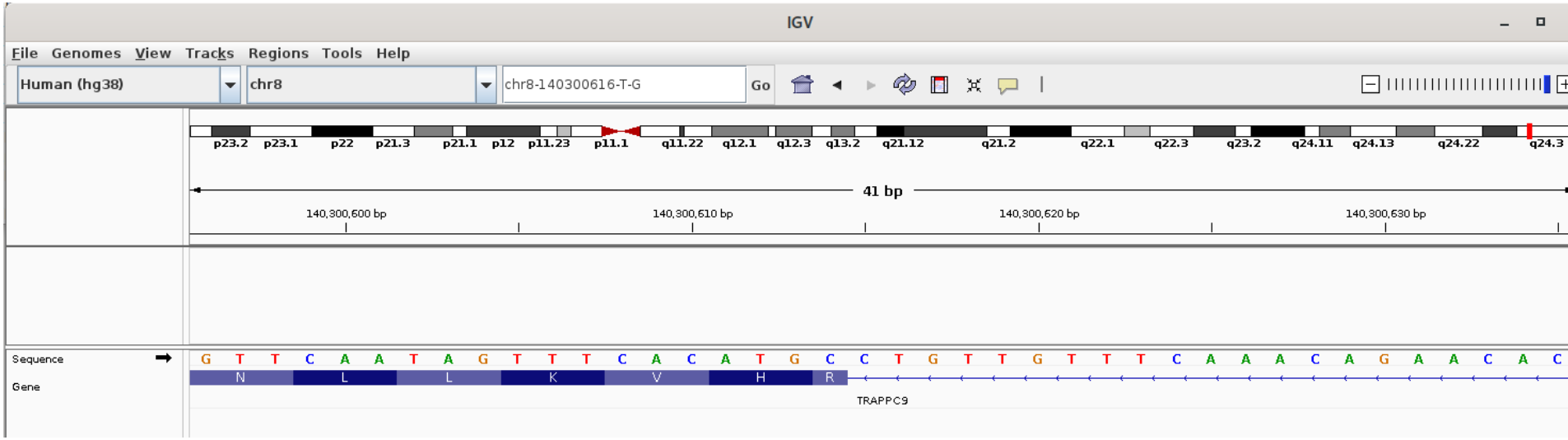
MetaRNN and MetaRNN-indel are pathogenicity prediction scores for human nonsynonymous SNVs (nsSNVs) and non-frameshift (NF) indels. They integrated information from 28 high-level annotation scores (16 functional prediction scores including SIFT, Polyphen2_HDIV, Polyphen2_HVAR, MutationAssessor, PROVEAN, VEST4, M-CAP, REVEL, MutPred, MVP, PrimateAI, DEOGEN2, CADD, fathmm-XF, Eigen and GenoCanyon, 8 conservation scores including GERP, phyloP100way_vertebrate, phyloP30way_mammalian, phyloP17way_primate, phastCons100way_vertebrate, phastCons30way_mammalian, phastCons17way_primate and SiPhy, and 4 allele frequency information from the 1000 Genomes Project, ExAC, gnomAD exome, and gnomAD genome) and produce an ensemble prediction model with a deep recurrent neural network (RNN). The final prediction is the likelihood of a nsSNV or NF indel being pathogenic.

Predikční nástroje na sestřihové varianty - nejčastěji rozhraní exon/intron



Small nuclear RNAs (snRNAs) are critical components of the spliceosome that catalyze the splicing of pre-mRNA.

Predikční nástroje na rozhraní exon/intron – sestřihové varianty



Examples (on hg38):

chr8-140300616-T-G

6 31740453 G T

NM_001089.3(ABCA3):c.875A>T (p.Glu292Val)

[show more examples]

SPLICE AI

chr8-140300616-T-G

Genome version: hg19 hg38 **Gencode:** basic comprehensive

Max distance: 500 masked scores REF & ALT scores Submit

November 3, 2024
- added choice of Gencode basic or comprehensive transcripts
- updated to Gencode v47

June 20, 2024
- fixed SpliceAI visualizations to clarify positions of predicted splicing changes. See issue #70 for details.

[show older updates]

Related web tools:
liftover: for variants/positions/intervals (hg19 <=> hg38 <=> T2T)
TGG Viewer: igv.js-based web viewer for public reference tracks and private data in Google Storage buckets. Has custom track types for RNA-seq splice junctions and gCNV variants.

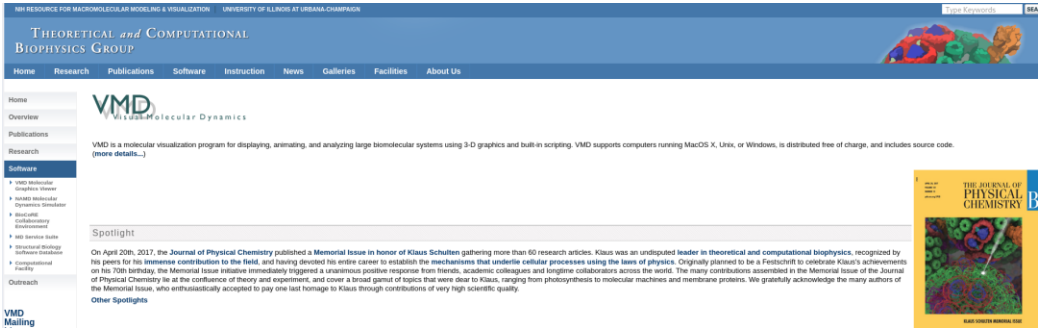
SpliceAI scores:

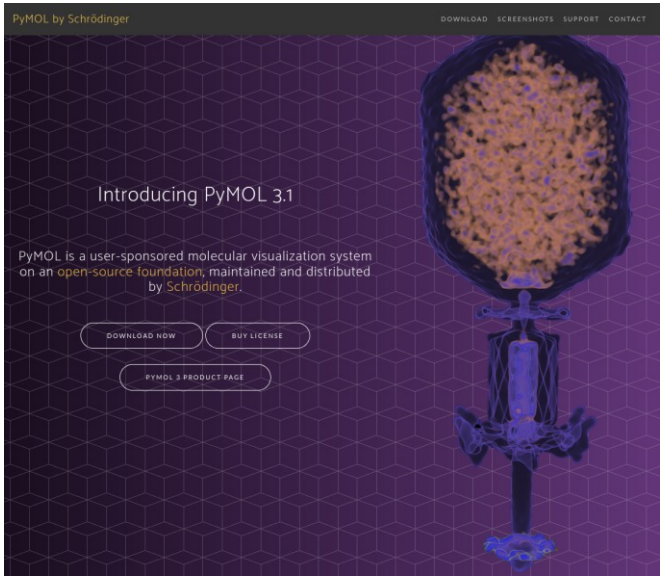
Variant	Gene	= MANE Select transcript	= non-coding transcript	Δ type	Δ score	position
chr8-140300616-T-G splice acceptor variant UCSC, gnomAD	TRAPPC9 (ENSG00000167632.18 / ENST00000438773.4 / NM_001160372.4) protein coding MANE Select transcript (minus strand) OMIM, GTEx, gnomAD, ClinGen, Ensembl, Decipher, GeneCards			Acceptor Loss	0.83	-2 bp
				Donor Loss	0.62	-147 bp
				Acceptor Gain	0.04	-32 bp
				Donor Gain	0.00	

Strukturní analýza pomocí molekulového modelování

Nutná podmínka – xray nebo nmr struktura!

Visualizační program: VMD, PYMOL





3D struktury proteinů – databáze PDB

File Edit View History Bookmarks Tools Help

RCSB PDB: Homepage

https://www.rcsb.org

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK

212,599 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

3D Structures | Enter search term(s), Entry ID(s), or sequence | Include CSM

Advanced Search | Browse Annotations Help

PDB-101 wwPDB EMDataResource NAKB wwPDB Foundation PDB-Dev

New: More Computed Structure Models (CSM) available Learn more

Welcome

Deposit

Search

Visualize

Analyze

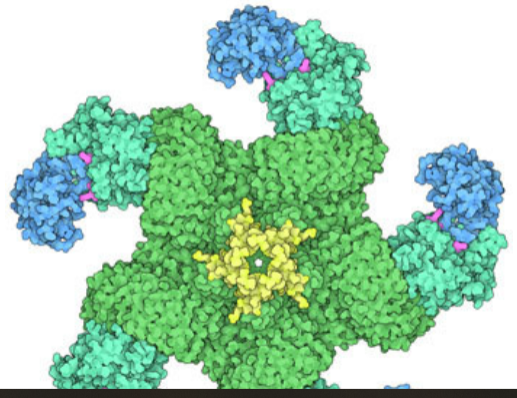
RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive

Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

November **Molecule of the Month**



Type here to search 6°C ENG 10:26 AM 11/24/2023

Databáze nukleových kyselin

NAKB
Nucleic Acid Knowledgebase

Full Search | Quick Search: enter PDB/NDB id, author or keyword

Home | Tools | Education | Standards | Download | About

Welcome to NAKB

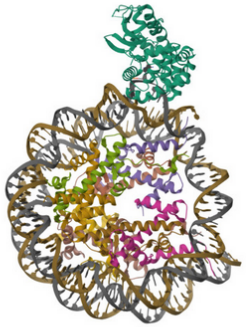
The Nucleic Acid Knowledgebase (NAKB), portal for 3D structural information about Nucleic Acids, is successor to the Nucleic Acid Database (NDB), as described in:

- Lawson CL, Berman HM, Vallat B, Chen L, Zirbel C (2024) The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. *Nucleic Acids Research* 52, D245. DOI
- Berman HM, Lawson CL, Schneider B (2022) Developing Community Resources for Nucleic Acid Structures. *Life* 12, 540. DOI
- **November 2023 online lecture (YouTube)** describing NAKB to the **CASP RNA special interest group**.

NAKB provides **Search**, **Report**, **Statistics**, **Atlas** and **Visualization** for all nucleic-acid containing experimentally determined 3D structures held by NDB and by the **Protein Data Bank (PDB)**, including all major methods: **X-ray**, **Electron Microscopy**, and **NMR**. For each structure, links are provided to **external resources** that annotate and analyze nucleic acid structures and their complexes.

Recently Released

All recent entries (29)



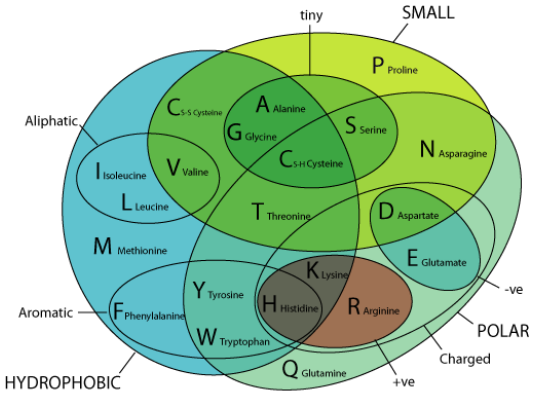
9EOZ
Protein/DNA
B-form double helix
hydrolase, nucleosome
Human OGG1 bound to a nucleosome core particle with 8-oxodGuo lesion at SHL6.0
Ren M, Gut F, Fan Y, Hopfner K-P, Zhou C
ELECTRON MICROSCOPY 3.1 Å *238.23 kDa*
Released 2024-11-06
Deposited 2024-03-16
PDB EMDB
PubMed DOI

Mutace narušuje strukturu proteinu

Vytvářejí sít kontaktů s okolními aminokyselinami
H-bonding, stacking, salt bridges

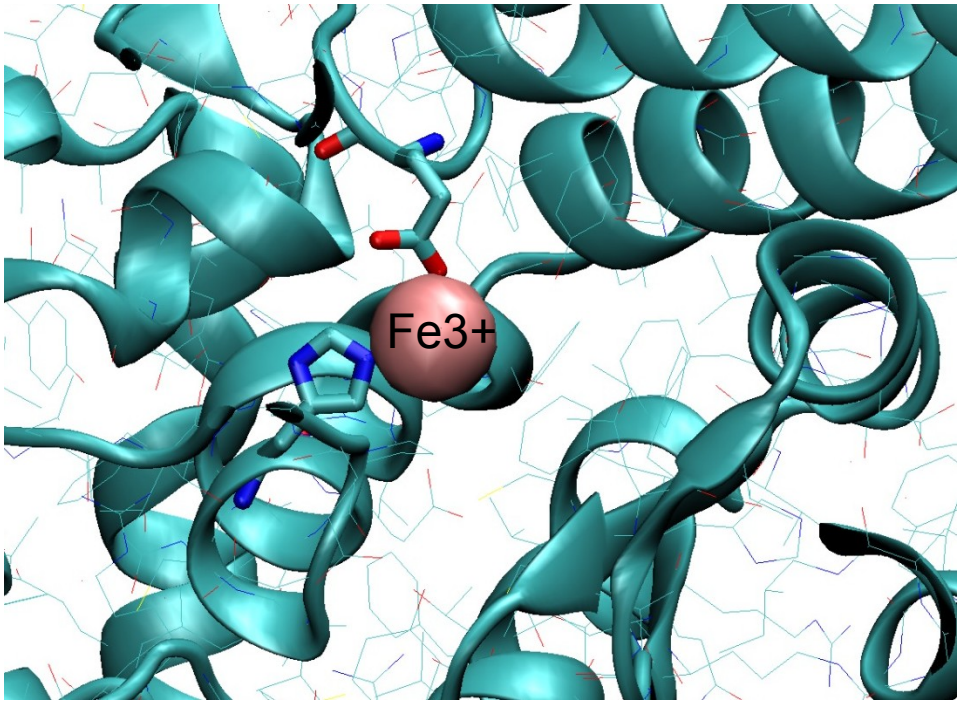
narušení struktury proteinu vede k destabilizaci proteinu, unfoldingu příp. agregaci

změna náboje,
polarity
velikost aminokyseliny



Mutace ve funkčních místech proteinu

(např. vazbu ligandu/kofaktoru, příp. vazba specifických iontů součást konformace aktivního místa)



Obr. PAH

Ukazatele kauzality

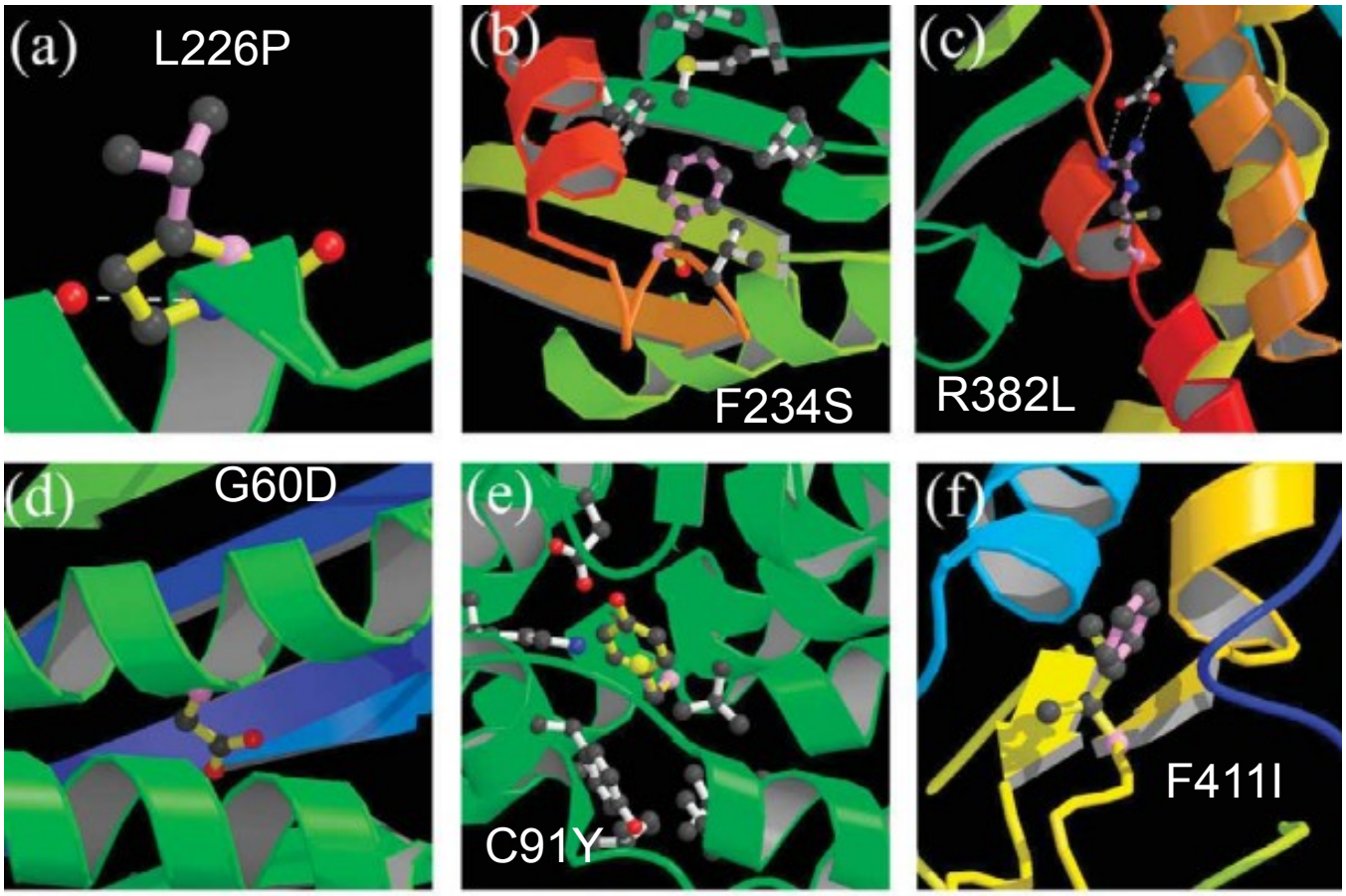
Hodnocené znaky:

- Specifické kontakty postraními řetězci AA
- Výskyt AA v aktivním místě
- Zanořenost AA v proteinu
- změna objemu AA
- změna náboje
- změna polarity
- Konzervovanost AA
- Přítomnost helix/turn breakers

1. Kauzální mutace uvnitř proteinu, polymorfismy spíš na povrchu

2. Kauzální mutace: wt mutace častěji vytváří vodíkové vazby a jiné kontakty

Examples of disease caused by structure destabilizing factors:
In silico analysis of mutations



Causal mutations:

bonds of wild-type side-chains are shown purple,
and bonds of the mutant side-chains are yellow.