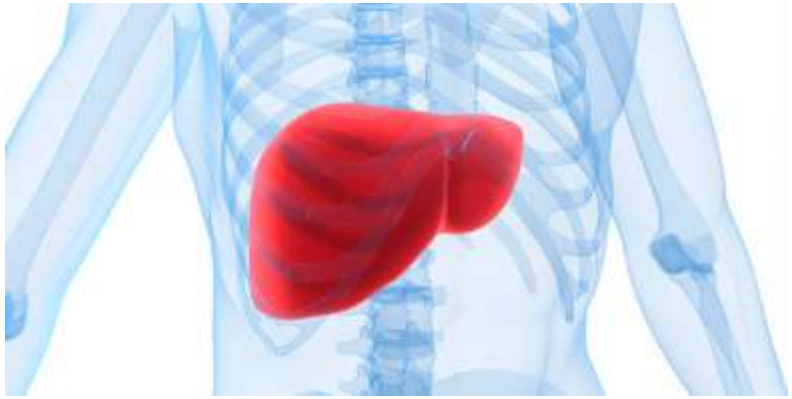
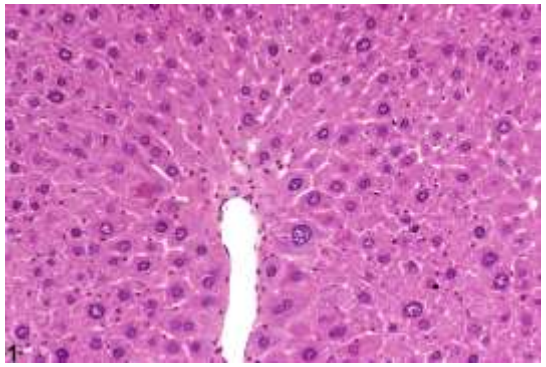
Zpracování dat a design experimentu single cell proteomika

Karel Harant

Kolik různých proteinů může být ve vzorku

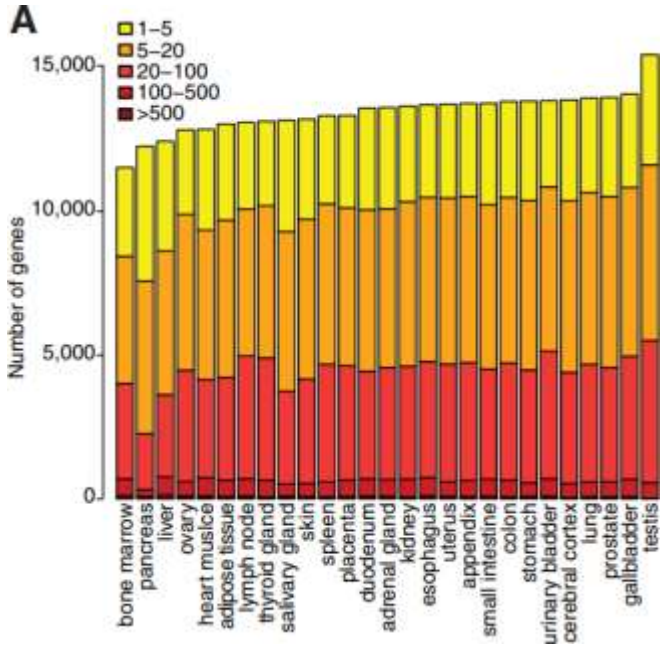
20 000
genů





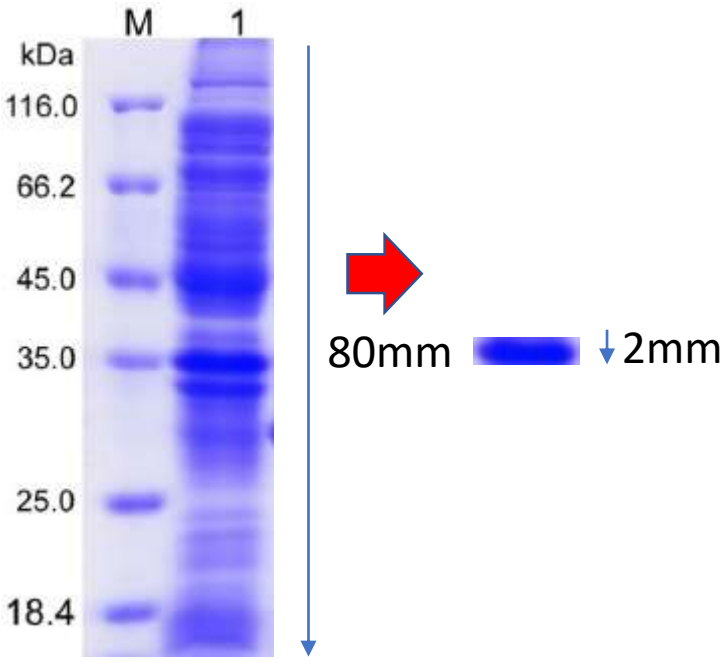
V jedné tkáni je přepisováno kolem 12 000 genů

Proteomicky lze identifikovat více než 9000 proteinů



Fagerberg,MCP,2014

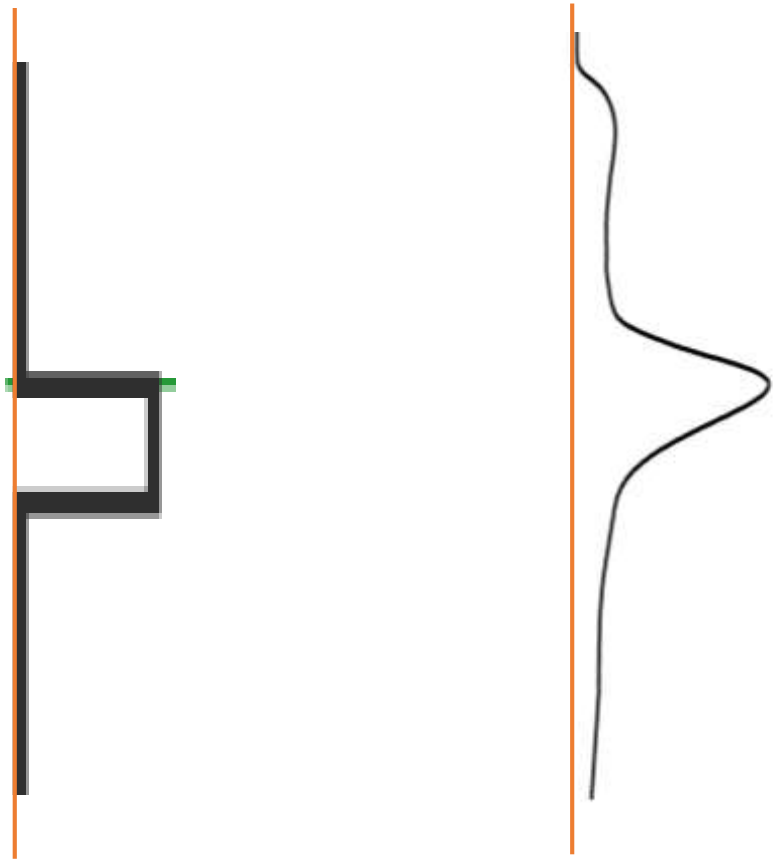
SDS gel



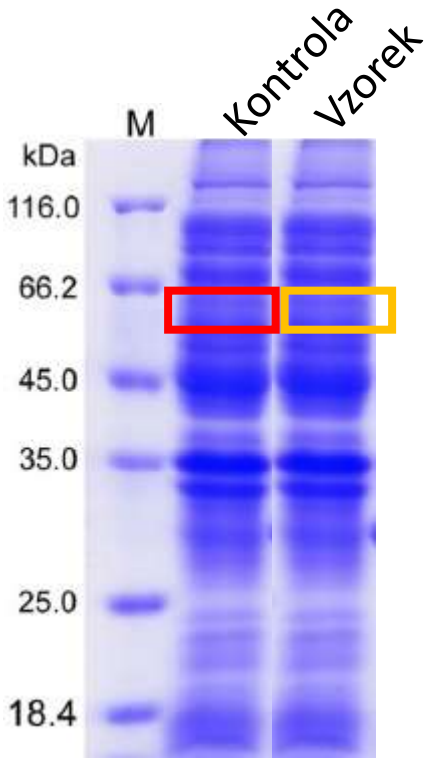
12000 proteinů

Cca 300 proteinů
Reálně však více

SDS separační profil



- Obdržíte identifikaci více než poloviny proteinů ve vzorku
- Libovolný protein bude identifikován s více než 50% pravděpodobností
- Měřit jeden vzorek, nebo jednu podmínku nemá smysl
- Vždy je potřeba porovnat stav který zkoumáte s kontrolou
- Relativní rozdíl je to co Vás zajímá



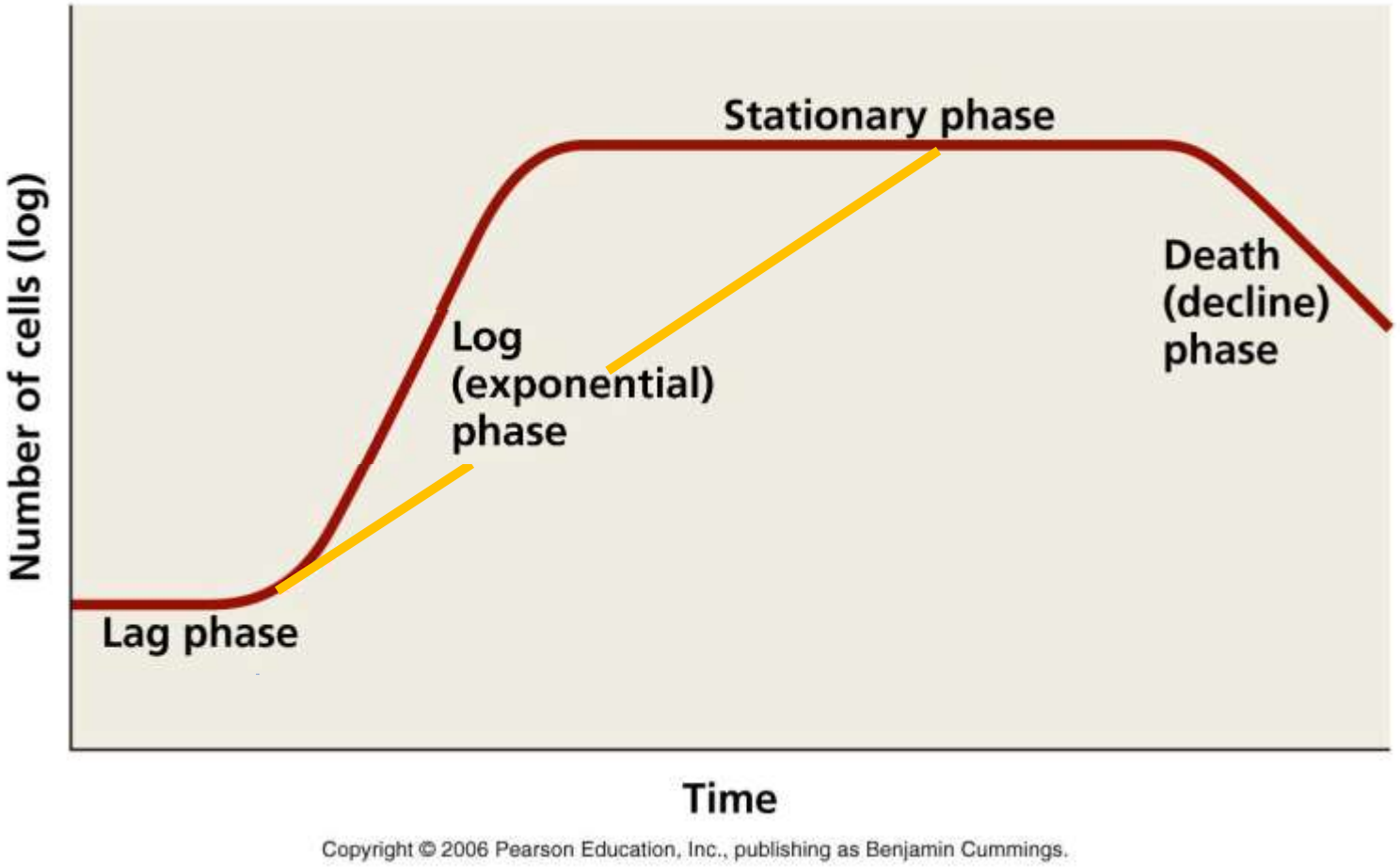
X

Návrh experimentu

- Definice otázky na níž hledáme odpověď
- Studium literatury pro výběr vhodné metody
- Výběr kontrol
  - Co nejvíce omezit variabilitu netýkající se položené otázky
    - Věk
    - Výživa
    - Prostředí
    - Pohlaví

Nevhodný výběr kontroly - příklad

- Bakteriální kultury
- Různá rychlost růstu kontroly a studovaného vzorku
- Odběr ve stejném čase
- Vzorek a kontrola mají různou růstovou fázi



Nevhodný výběr kontroly - příklad

- Zavedení sondy pro výživu vyžaduje chirurgický zákrok
- Kontrolní myš musí též absolvovat chirurgický zákrok s podobným poškozením
- Efekt z hojení poranění je větší než efekt z rozdílné výživy

Kontrola

Vzorek

Technický vs Biologický replikát

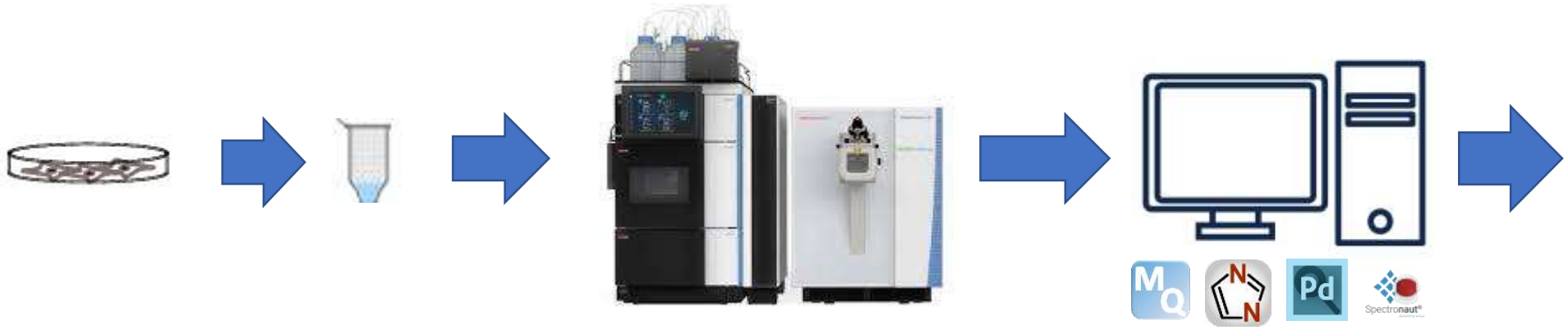
Biologický

- Člověk/živočich
- Kultivace očkované z různých konzerv
- V separátních kultivačních nádobach

Technický

- Podchytí experimentální variabilitu
- Opakovaný nástřik
- Opakovaná příprava vzorku
- Opakovaný odběr ze stejné kultivace

Zpracování dat



	LFQ intensity 1394-4...	LFQ intensity 1394-4...	LFQ intensity 1394-4...	LFQ intensity 1394-4...	LFQ intensity 1394-5...	LFQ intensity 1394-5...	LFQ intensity 1394-5...	LFQ intensity 1394-5...	LFQ intensity 1394-5...	C: Only identifi... by site	C: Reverse	C: Potential contam...	N: Peptides	N: Razor + unique peptides	N: Unique peptides	N: Seque... covera...	N: Unique + razor	N: Unique sequen...	N: Mol. weight [kDa]	N: Q-value	N: Score	N: Intensity	N: MS/M coun
Type	Main	Main	Main	Main	Main	Main	Main	Main	Main	Catego...	Catego...	Catego...	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Num
1	0	0	0	0	0	0	0	0	0	+		+	10	3	3	18.4	8.6	8.6	49.493	1	-2	775670...	9
2	923460...	120280...	118970...	188370...	136490...	0	121710...	144200...	127970...				2	2	2	26.4	26.4	26.4	7.9021	0	39.996	116780...	149
3	0	0	0	0	0	0	0	0	0	+		+	14	3	1	16.2	3.6	1.7	63.165	1	-2	645730...	5
4	202840...	160940...	149790...	103440...	132200...	253480...	122830...	137430...	978000...			+	3	3	3	16.5	16.5	16.5	24.409	0	13.695	179690...	431
5	539950...	528290...	172550...	384970...	211010...	519800...	207510...	303750...	342150...			+	36	32	7	71.6	66.3	21.6	51.621	0	323.31	218570...	1296
6	0	0	0	0	0	0	0	0	0	+		+	11	1	1	17.5	4	4	57.769	1	-2	5163100	1
7	8448300	0	0	0	0	0	0	0	0			+	44	3	2	59.9	6.6	3.5	60.044	0	25.478	608400...	65
8	0	0	0	0	0	0	0	0	0			+	4	4	4	24.1	24.1	24.1	22.975	0.0062...	1.8459	127900...	39
9	0	0	0	0	0	0	0	0	0			+	1	1	1	10.7	10.7	10.7	18.974	0.0011...	2.7073	6334400	1
10	304220	273070	0	315590	0	294830...	145670	0	295340			+	38	38	33	71.1	71.1	64.9	69.366	0	174.33	189110...	95
3625	0	0	0	0	0	0	0	0	0		+		3	3	3	0	0	0	126.14	0.0070...	1.7964	126440...	4
3626	5222300	3164200	0	4602800	1942300	1493100	4313500	3848000	0		+		3	3	3	0	0	0	100.23	0.0012...	2.9061	583360...	69
3627	0	0	0	0	0	204480...	0	0	0		+		4	4	4	0	0	0	102.63	0.0051...	1.8886	391180...	10
3628	0	0	0	0	0	0	0	0	0		+		6	6	6	0	0	0	93.068	0.0093...	1.6621	756120...	15
3629	2784700	0	3823500	0	0	0	0	0	0		+		6	6	6	0	0	0	95.379	0.0023...	2.5512	498340...	27
3630	0	0	1981300	0	0	0	0	0	0		+		2	2	2	0	0	0	24.84	0.0062...	1.8464	250890...	19
3631	164650...	168130...	165330...	174110...	141250...	142570...	164000...	153360...	166990...				54	54	53	75.3	75.3	75.3	75.338	0	323.31	135030...	3491
3632	154780...	170530...	148030...	178920...	120850...	123790...	161920...	146910...	136810...				8	6	6	35.2	32.4	32.4	28.261	0	63.933	288850...	285
3633	151450...	162000...	183940...	161950...	144860...	144560...	169760...	169360...	182910...				30	30	30	71.8	71.8	71.8	52.088	0	323.31	438780...	1575

Informace ve výsledcích

Identifikátory

- Identifikátor z FASTA databáze
- Jméno genu
- Jméno proteinu
- Popis proteinu

Kvalitativní ukazatele

- Score Identifikace
- Q value Identifikace
- Počet unikátních
- Počet peptidů

Data

- Intezity

Identifikátory

T: Protein IDs	T: Majority protein	T: Protein names	T: Gene names	T: id	T: Fasta headers
Text	Text	Text	Text	Text	Text
P00044;P62894;CON__P62894	P00044	Cytochrome c iso-1	CYC1	53	CYC1_YEAST_YJR048W_CYC1 Cytochrome c iso-1 OS=Saccharomyces cerevisiae (str...

Kvalitativní ukazatale

matrix13 | matrix14 | matrix15 | matrix16 | matrix17 | matrix18 | matrix19 | matrix20 | matrix24 | 1394_export | matrix.

N: Peptides	N: Razor + unique peptides	N: Unique peptides	N: Seque... covera...	N: Unique + razor	N: Unique sequen...	N: Mol. weight [kDa]	N: Q-value	N: Score
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
12	12	9	73.4	73.4	50.5	12.182	0	233

N: PG.Qvalue	N: PG.Cscore	T: PG.ProteinAccessions
Numeric	Numeric	Text
0.00012773	41.5598	Q86X27
0	42.3311	P62875

Obecný postup zpracování proteomických dat

Příprava

Transformace
Normalizace

Kontrola kvality

Kontrola rozložení dat a středových hodnot

Kontrola podobnosti pomocí PCA

Identifikace rozdílů

Rozdělení vzorků do skupin

Statistické testy a vizualizace

Příprava na interpretaci

Funkční anotace

Go term analýza

Normalizace




- Dobrá praxe – normalizace na celkový protein při přípravě vzorku
- Většina používaných vyhledávacích nástrojů produkuje normalizovaná data nejčastěji je implementovaná MaxQuantLFQ normalisation
  - Předpoklad velké podobnosti vzorků na vstupu
  - Předpoklad shodného nástřiku
  - Neznormalizuje příliš velké rozdíly (násobky)

- Pokud je i tak potřeba data normalizovat
  - Pravděpodobně nebyl experiment ideálně proveden
- Vícero metod – podobná efektivita
  - Globální medián
  - Loess
  - Quantile

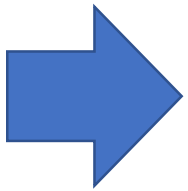
Normalizace na median

Suma intenzit 600 700 800 800 700 600

	LFQ intensity 1-1	LFQ intensity 1-2	LFQ intensity 1-3	LFQ intensity 3-1	LFQ intensity 3-2	LFQ intensity 3-3
Type	Main	Main	Main	Main	Main	Main
Group1	ctrl1	ctrl1	ctrl1	3	3	3
1	24.5548	24.2626	24.3077	24.3901	25.7452	24.2662
2	30.8355	30.5738	30.747	30.5588	30.4589	30.5707
3	23.8336	23.8302	23.4919	23.4799	22.6654	22.599
4	22.8469	21.7399	21.9448	22.7756	22.794	22.8207
5	22.5228	22.8234	23.874	23.565	24.1907	24.1253
6	29.5208	29.2778	29.2701	29.5693	29.5607	29.3691
7	23.9238	23.695	23.4772	23.428	22.9664	23.2279

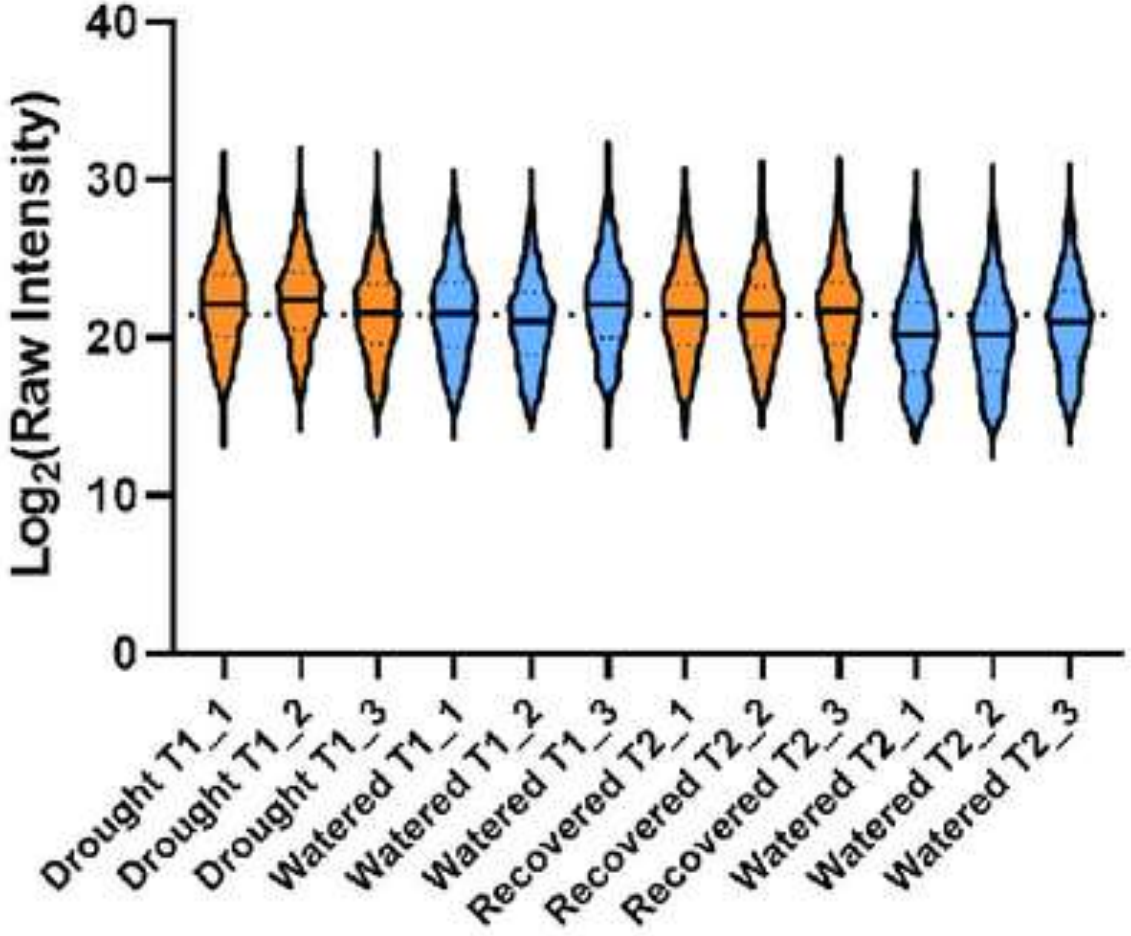
Normalizační Faktor 1,16 1 0,875 0,875 1 1,16

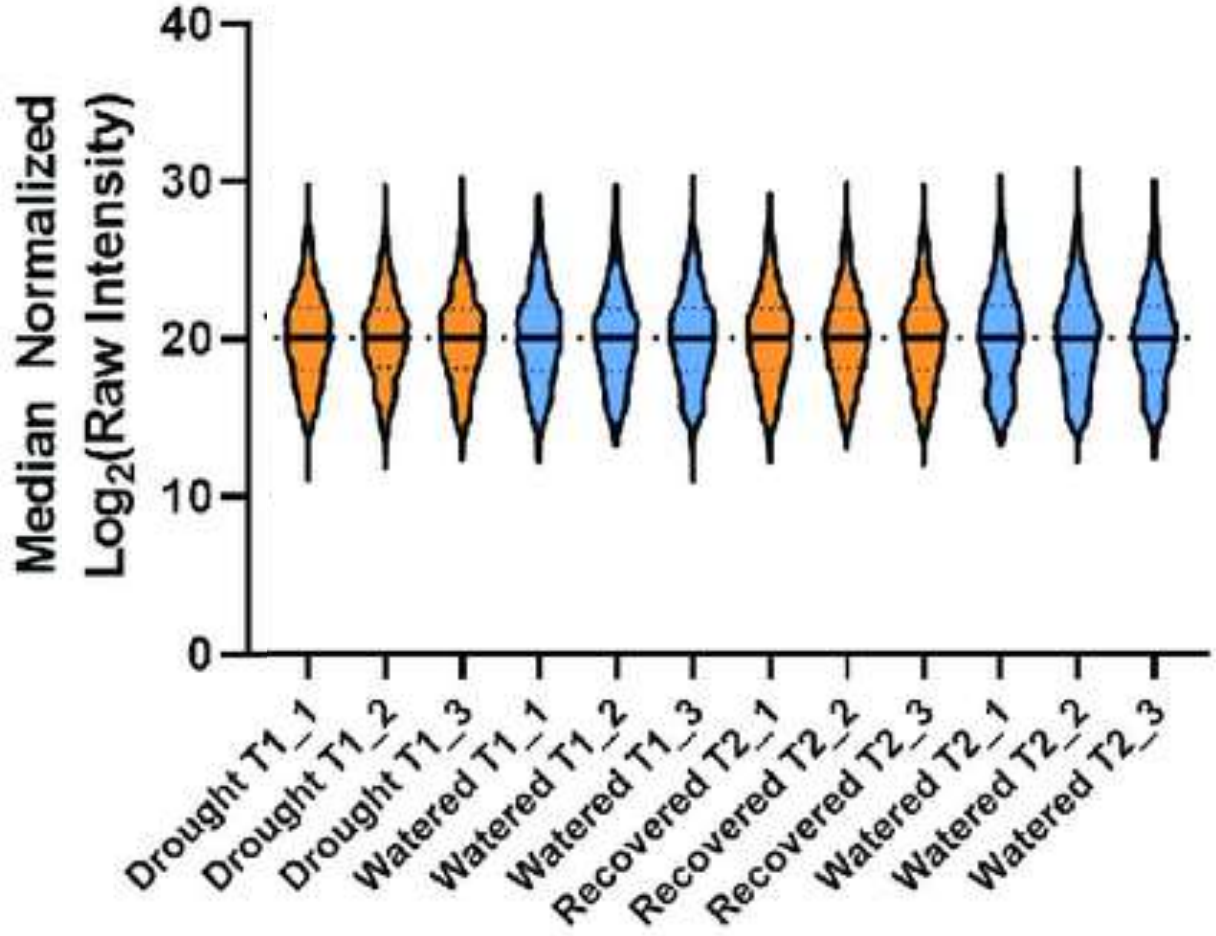
Medián 700



Suma intenzit 700 700 700 700 700 700

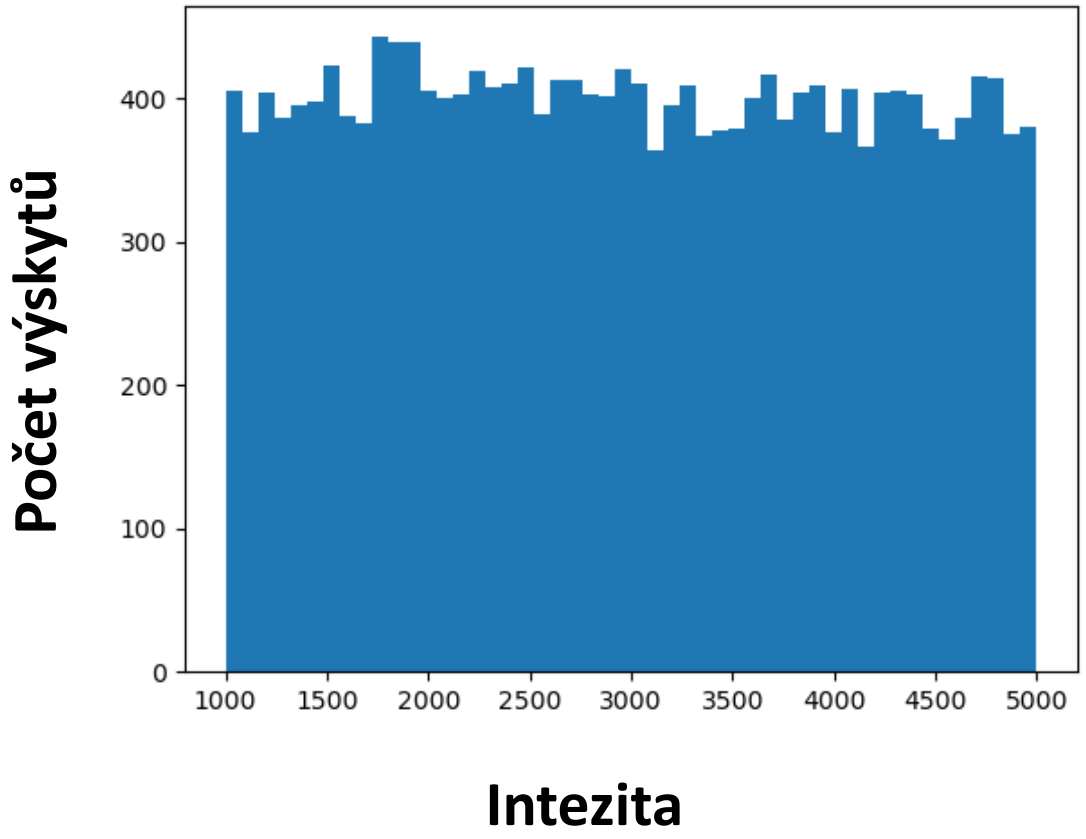
	LFQ intensity 1-1	LFQ intensity 1-2	LFQ intensity 1-3	LFQ intensity 3-1	LFQ intensity 3-2	LFQ intensity 3-3
Type	Main	Main	Main	Main	Main	Main
Group1	ctrl1	ctrl1	ctrl1	3	3	3
1	24.5548	24.2626	24.3077	24.3901	25.7452	24.2662
2	30.8355	30.5738	30.747	30.5588	30.4589	30.5707
3	23.8336	23.8302	23.4919	23.4799	22.6654	22.599
4	22.8469	21.7399	21.9448	22.7756	22.794	22.8207
5	22.5228	22.8234	23.874	23.565	24.1907	24.1253
6	29.5208	29.2778	29.2701	29.5693	29.5607	29.3691
7	23.9238	23.695	23.4772	23.428	22.9664	23.2279


Log2(Raw Intensity)
40
30
20
10
0
Drought T1_1
Drought T1_2
Drought T1_3
Watered T1_1
Watered T1_2
Watered T1_3
Recovered T2_1
Recovered T2_2
Recovered T2_3
Watered T2_1
Watered T2_2
Watered T2_3


Median Normalized
Log2(Raw Intensity)
40
30
20
10
0
Drought T1_1
Drought T1_2
Drought T1_3
Watered T1_1
Watered T1_2
Watered T1_3
Recovered T2_1
Recovered T2_2
Recovered T2_3
Watered T2_1
Watered T2_2
Watered T2_3

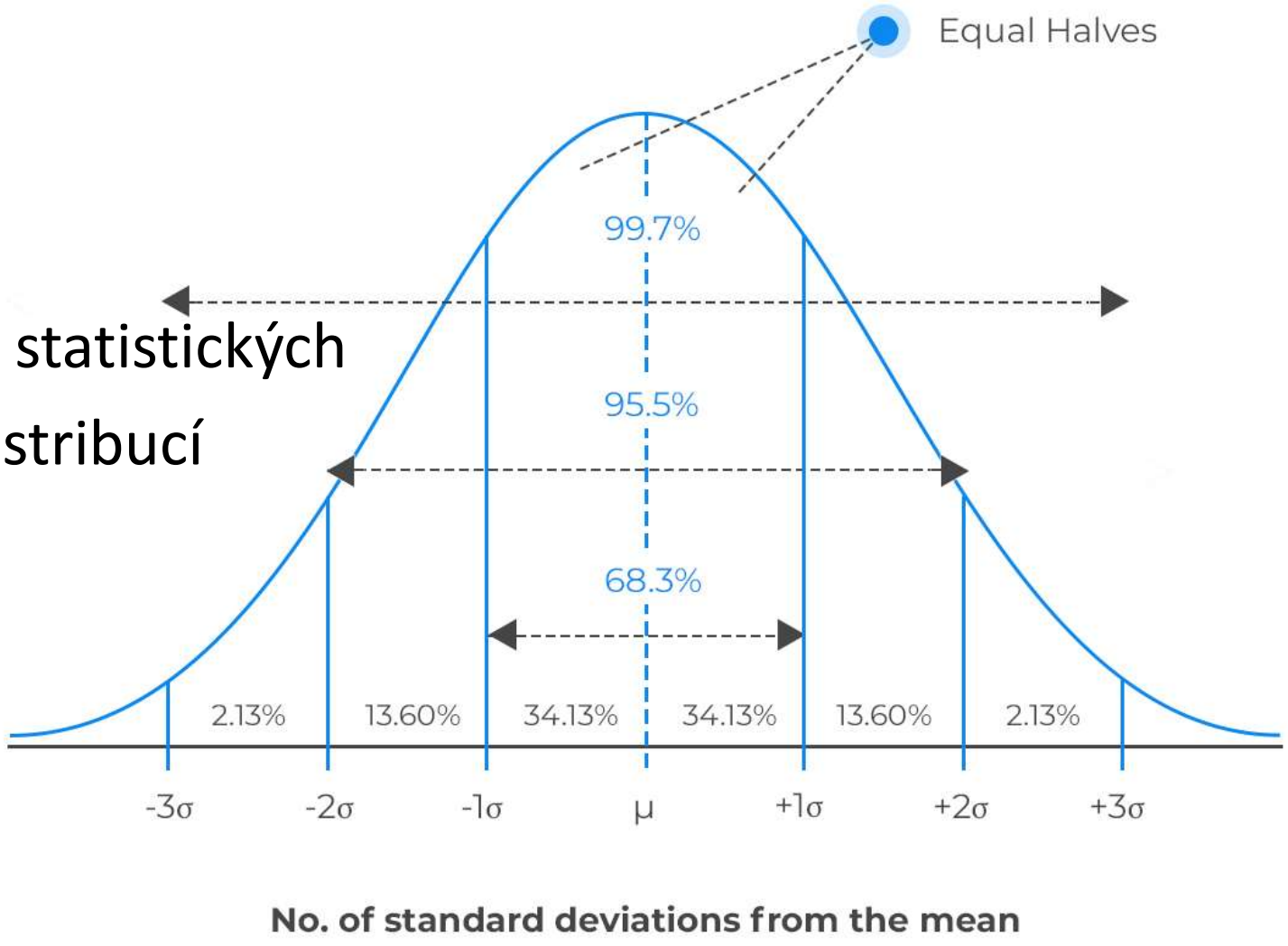
Histogram

- Typ sloupcového grafu
- Data rozdělíme na definovaný počet segmentů – binů
  - Interval mezi největší a nejmenší hodnotou rozdělíme na 50 shodných intervalů
- Vyneseme do sloupce počet hodnot spadajících do každého intervalu

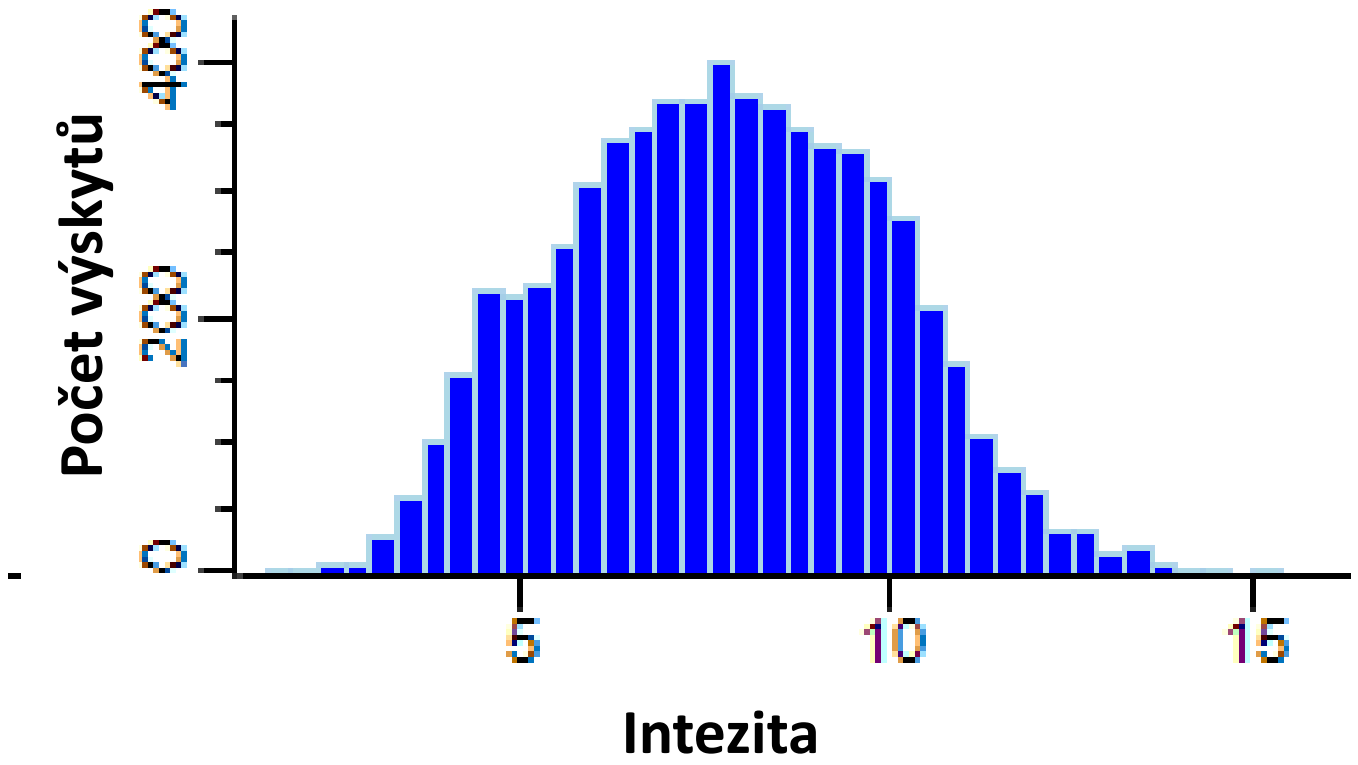


Gausovská distribuce (Normální distribuce)

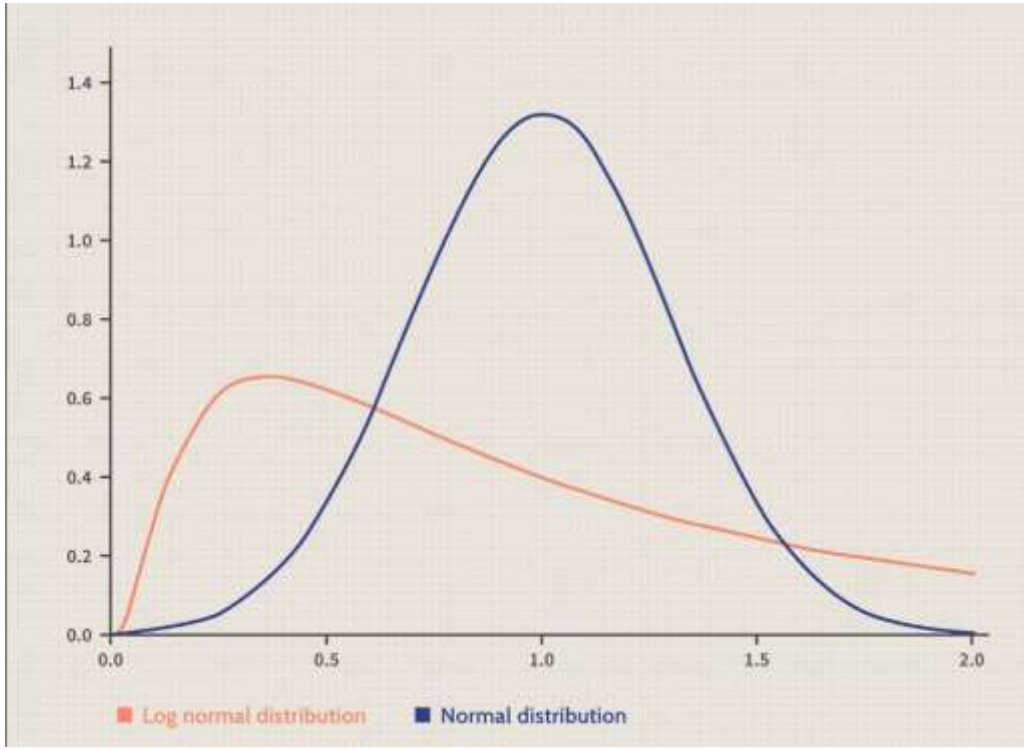
- Častá u přírodních jevů
- Biologická data
  - Výška v populaci
  - Proteomická data
- Správná funkce parametrických statistických metod je podmíněna normální distribucí

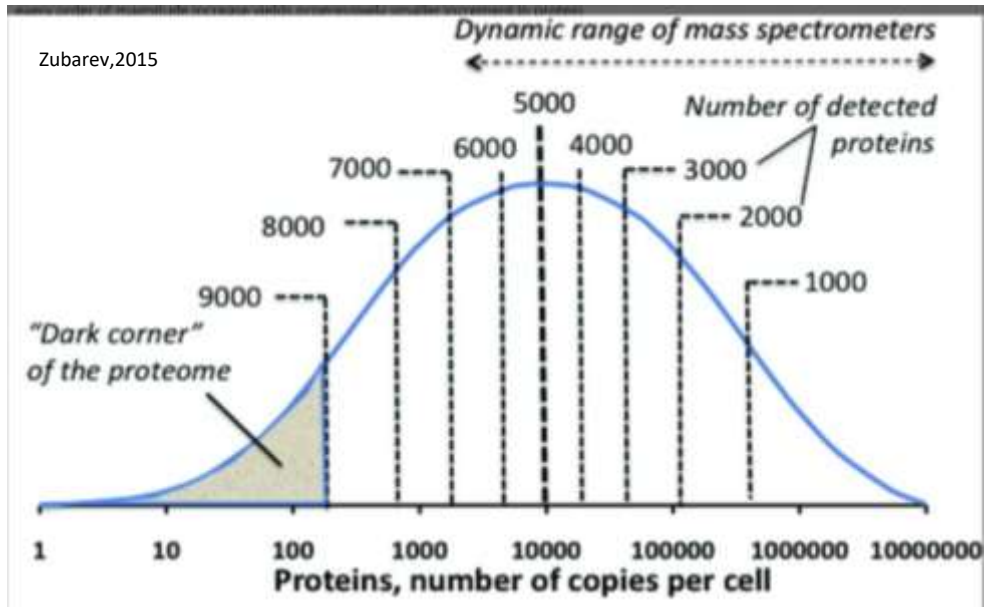


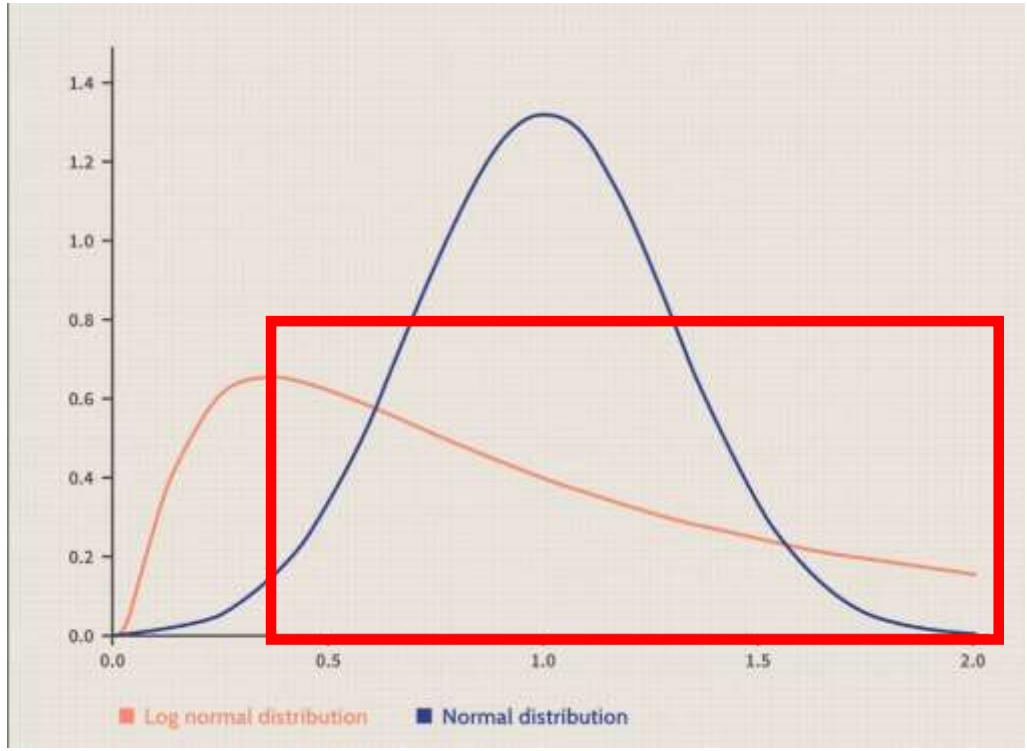
- Vizualizace rozdělení dat
  - Lze tak ověřit distribuci dat
  - Kontrola středové hodnoty dat

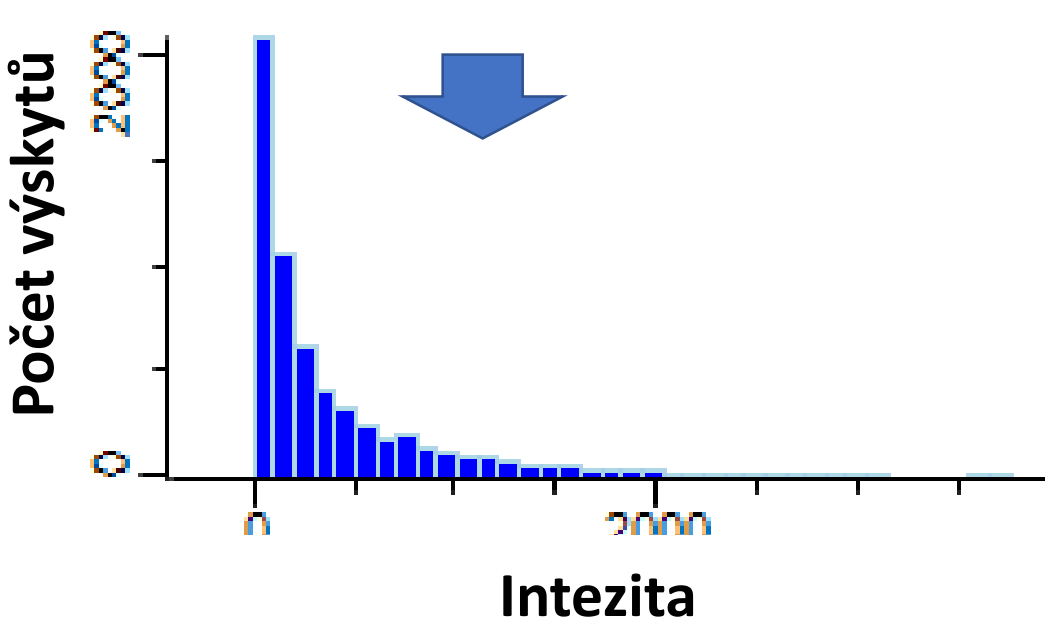


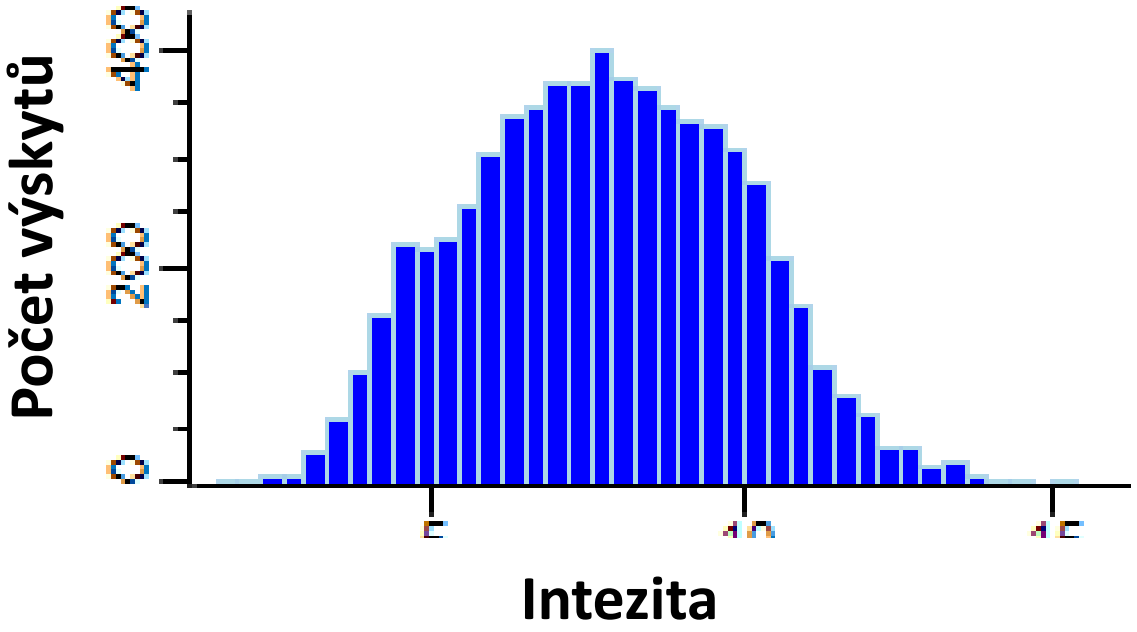
Lognormální distribuce






1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Log normal distribution
Normal distribution


Počet výskytů
2000
0
0
2000
Intezita


Počet výskytů
400
200
0
Intezita

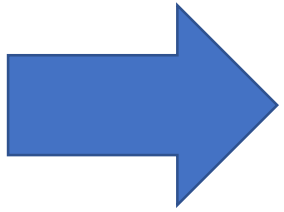
Metody redukující dimenzionalitu dat

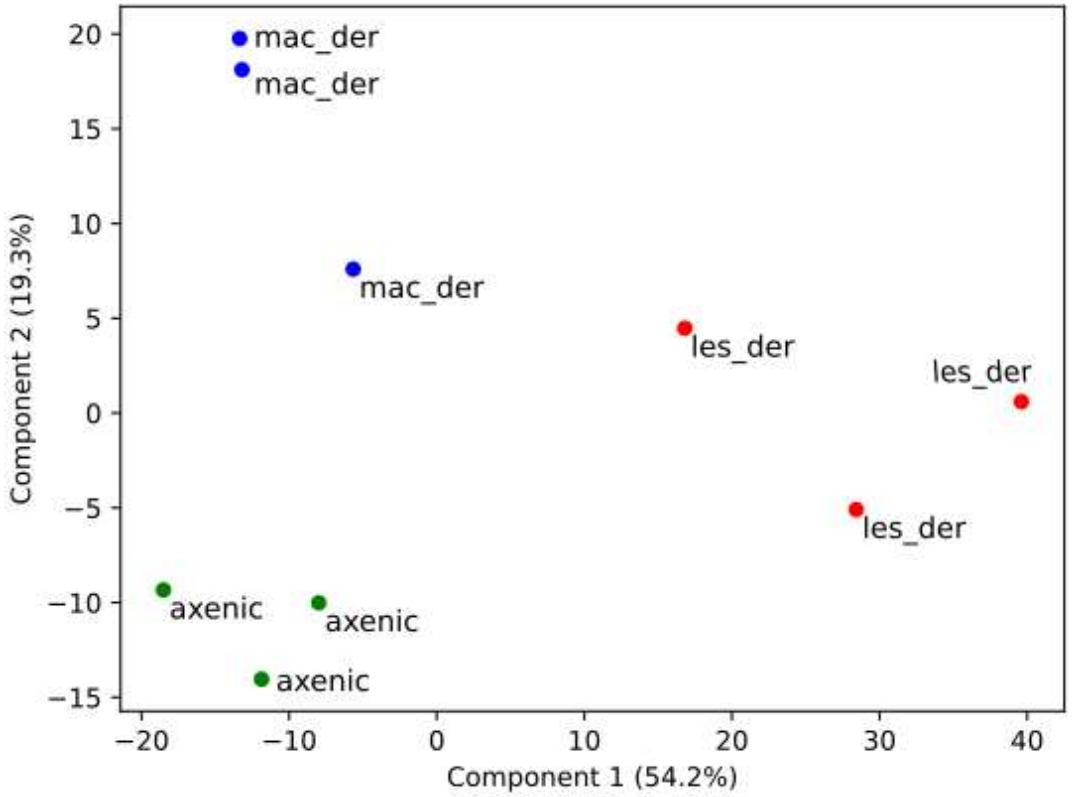
PCA - *Principal Component Analysis*

Analýza hlavních komponent

Principal Component Analysis, **PCA**

	LFQ intensity 1-1	LFQ intensity 1-2	LFQ intensity 1-3	LFQ intensity 3-1	LFQ intensity 3-2	LFQ intensity 3-3
Type	Main	Main	Main	Main	Main	Main
Group1	ctrl1	ctrl1	ctrl1	3	3	3
1	24.5548	24.2626	24.3077	24.3901	25.7452	24.2662
2	30.8355	30.5738	30.747	30.5588	30.4589	30.5707
3	23.8336	23.8302	23.4919	23.4799	22.6654	22.599
4	22.8469	21.7399	21.9448	22.7756	22.794	22.8207
5	22.5228	22.8234	23.874	23.565	24.1907	24.1253
6	29.5208	29.2778	29.2701	29.5693	29.5607	29.3691
7	23.9238	23.695	23.4772	23.428	22.9664	23.2279
4434	28.3192	28.4554	28.4873	28.3725	28.7116	28.4981
4435	22.62	22.7617	22.825	22.5495	22.5323	22.5352
4436	25.6781	26.0049	25.7859	25.5292	25.8811	25.7345
4437	23.0281	23.7413	23.3013	23.1111	23.1813	22.3912
4438	25.2451	25.5161	25.3303	25.624	25.3515	25.2891
4439	22.3205	22.1204	22.2094	24.0372	23.4178	23.0194





PCA využití

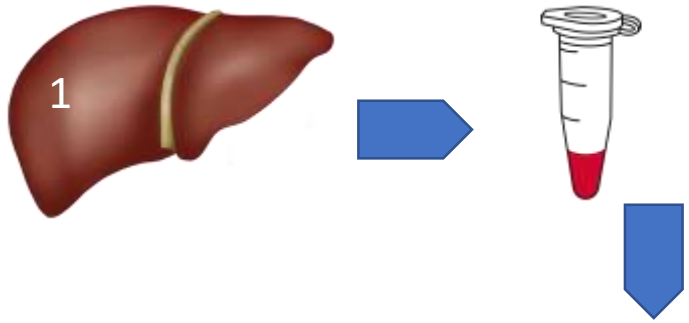
- Metoda převede tisíce hodnot do x a y koordinátů
- Hledá to v čem jsou vzorky podobné
- Pozice v dvourozměrném prostoru popisuje podobnost vzorků
- Využití pro posouzení separace skupin vzorků
- Identifikaci odlehlých vzorků

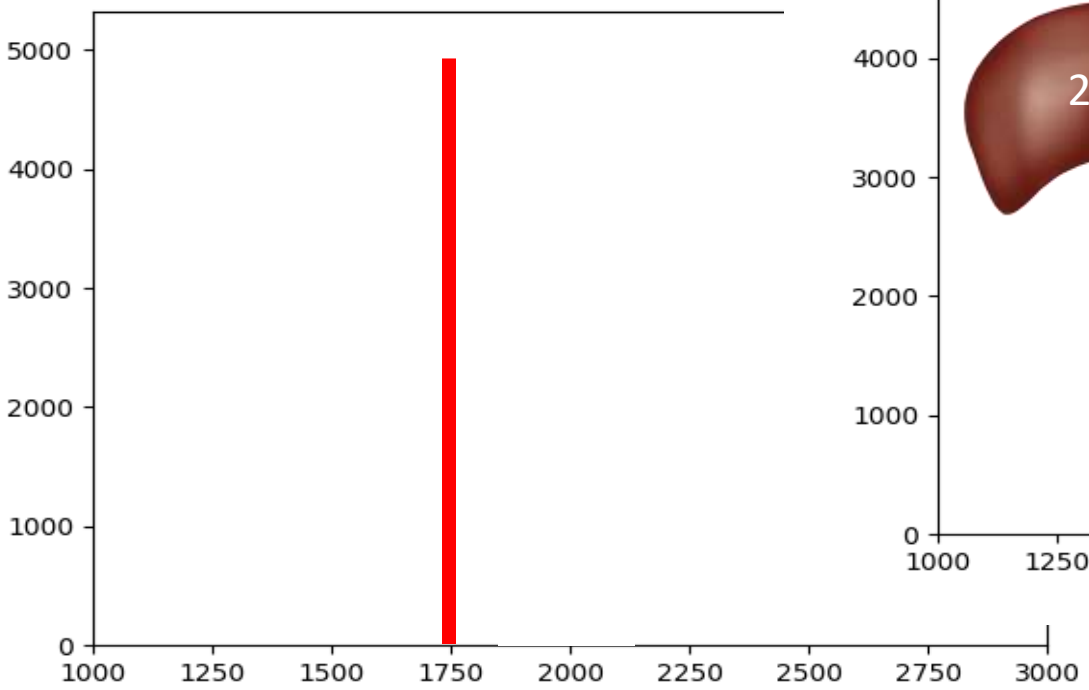
Obecný postup zpracování proteomických dat

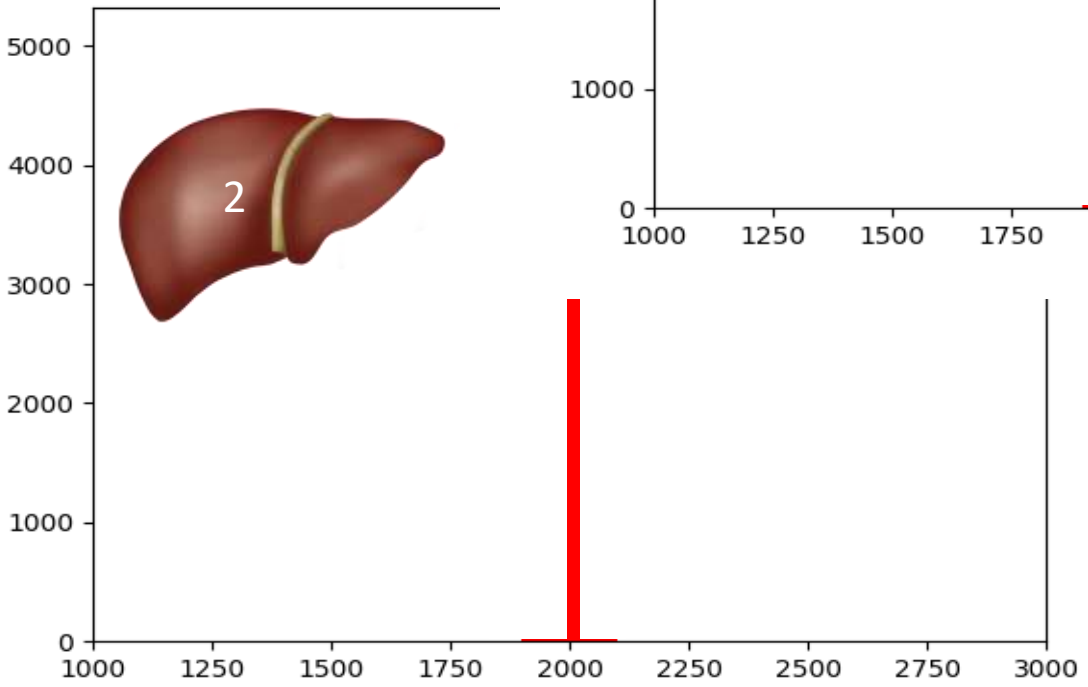
Příprava	Logaritmování Filtrování Normalizace
Kontrola kvality	Kontrola rozložení dat a středových hodnot Kontrola podobnosti pomocí PCA
Identifikace rozdílů	Rozdělení vzorků do skupin Statistické testy a vizualizace
Příprava na interpretaci	Funkční anotace Go term analýza

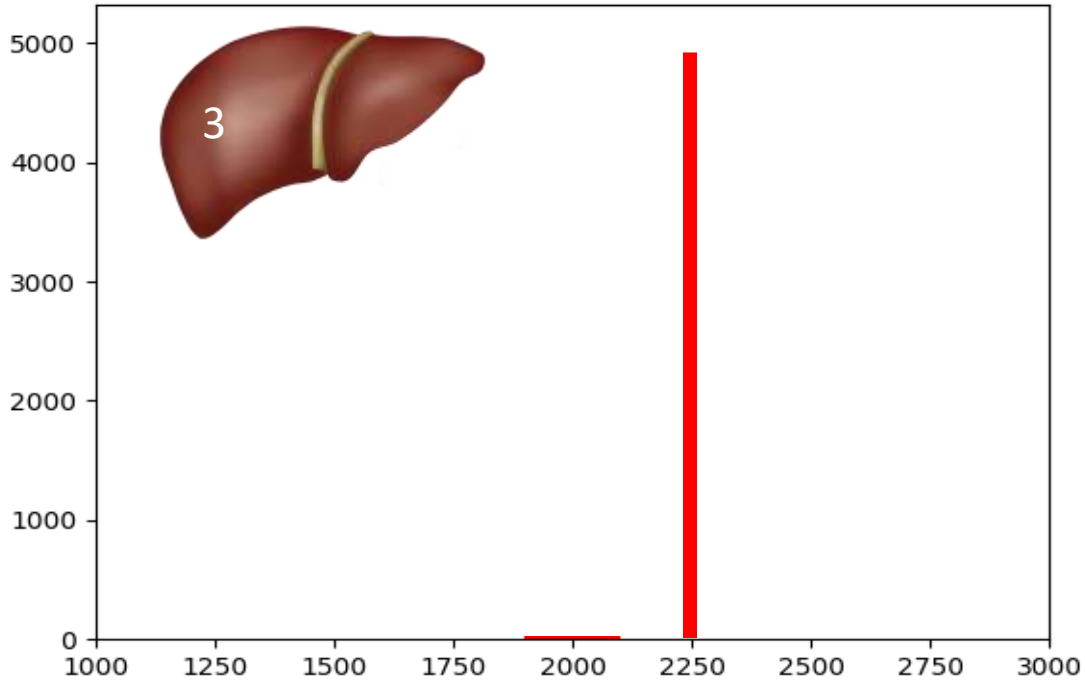
Statistické testování

Množství proteinu - co měří





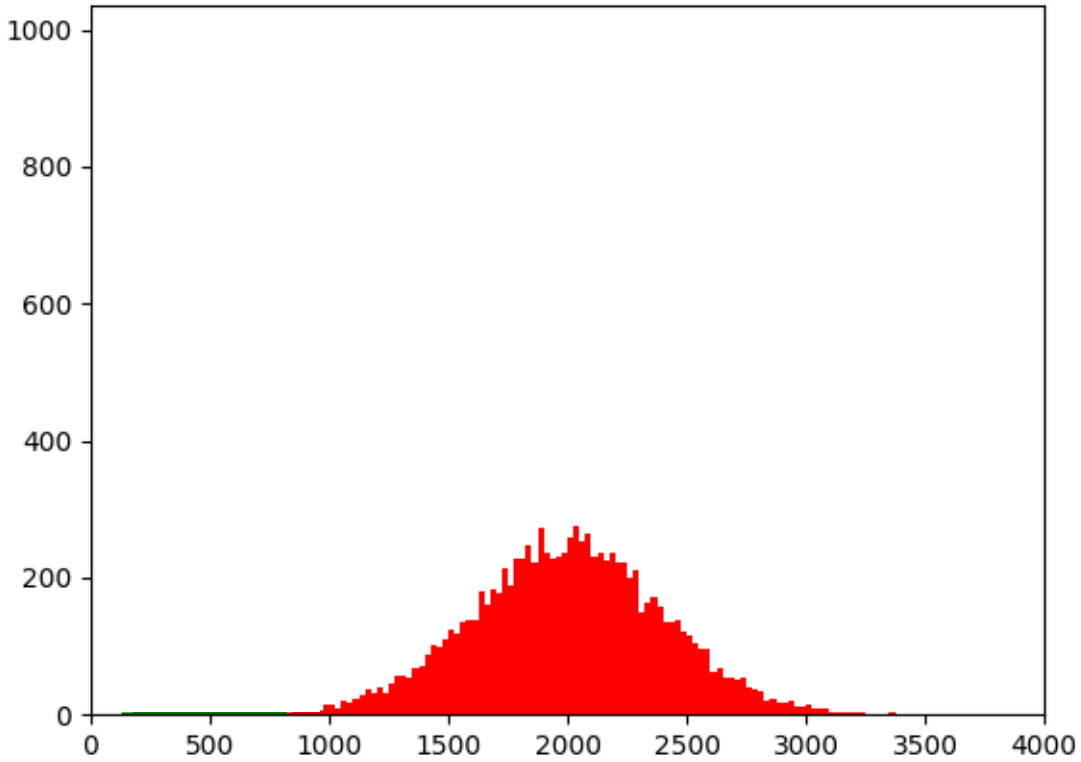


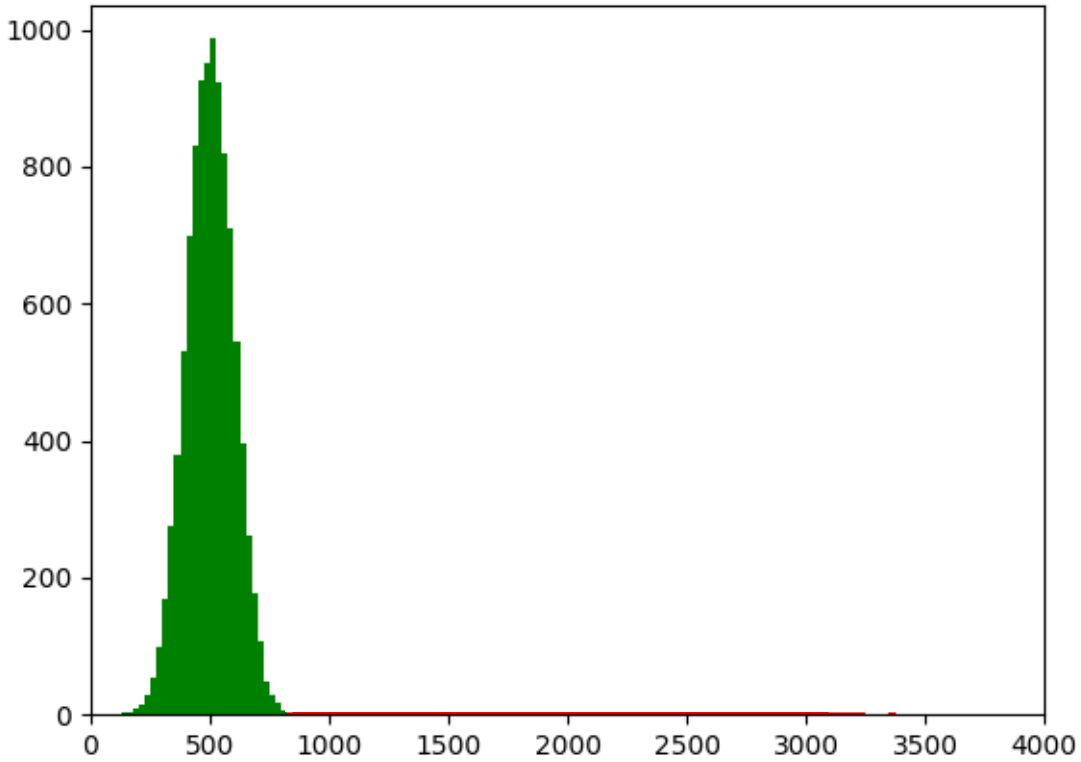


Distribuce proteinu

Popsána průměrem a směrodatnou odchylkou naměřených hodnot

Více naměřených hodnot – vyšší pravděpodobnost že distribuci zachytíme přesně





Technická a biologická variabilita

Homogenizace

Měření Koncentrace

Efektivita štěpení

Chromatografie

MS instrumentace

Degradace

Pipetování

Z naší zkušenosti by měla být do 20%

Coeficient of variation - Variační koeficient CV

- Udává jaký rozptyl dostaneme při opakovaném měření identického vzorku
- Kvalitativní metrika analytické metody

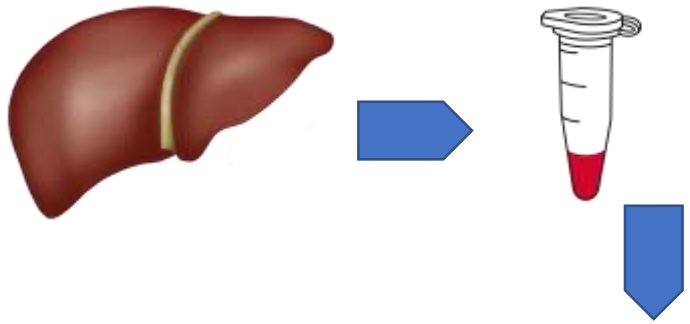
$$\frac{\text{Směrodatná odchylka}}{\text{průměr}} * 100 = CV\%$$

Příklad

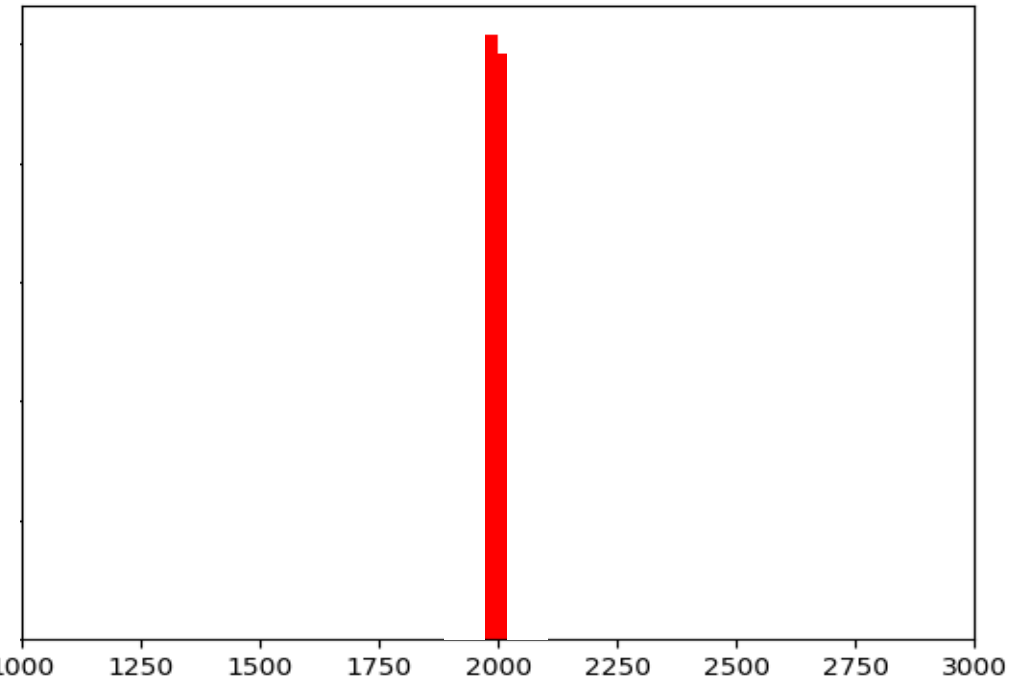
Získané hodnoty
90,100,110

Průměr 100, Směrodatná odchylka 8.1, CV=8.1%

Technická variabilita

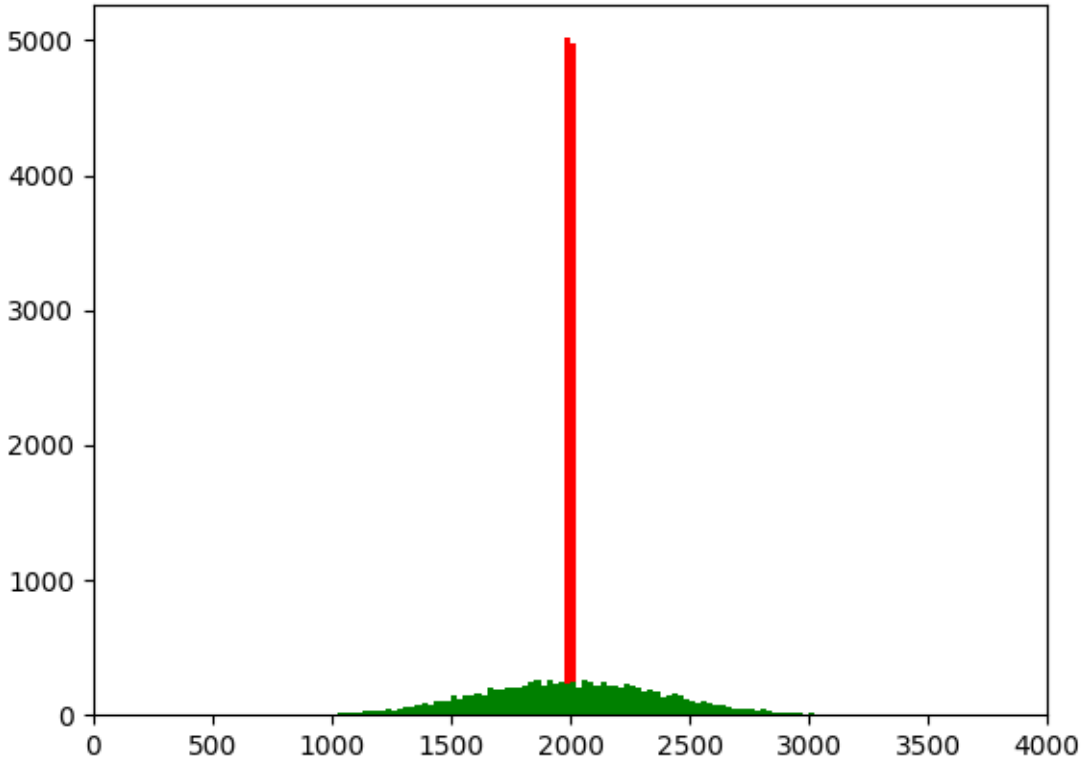


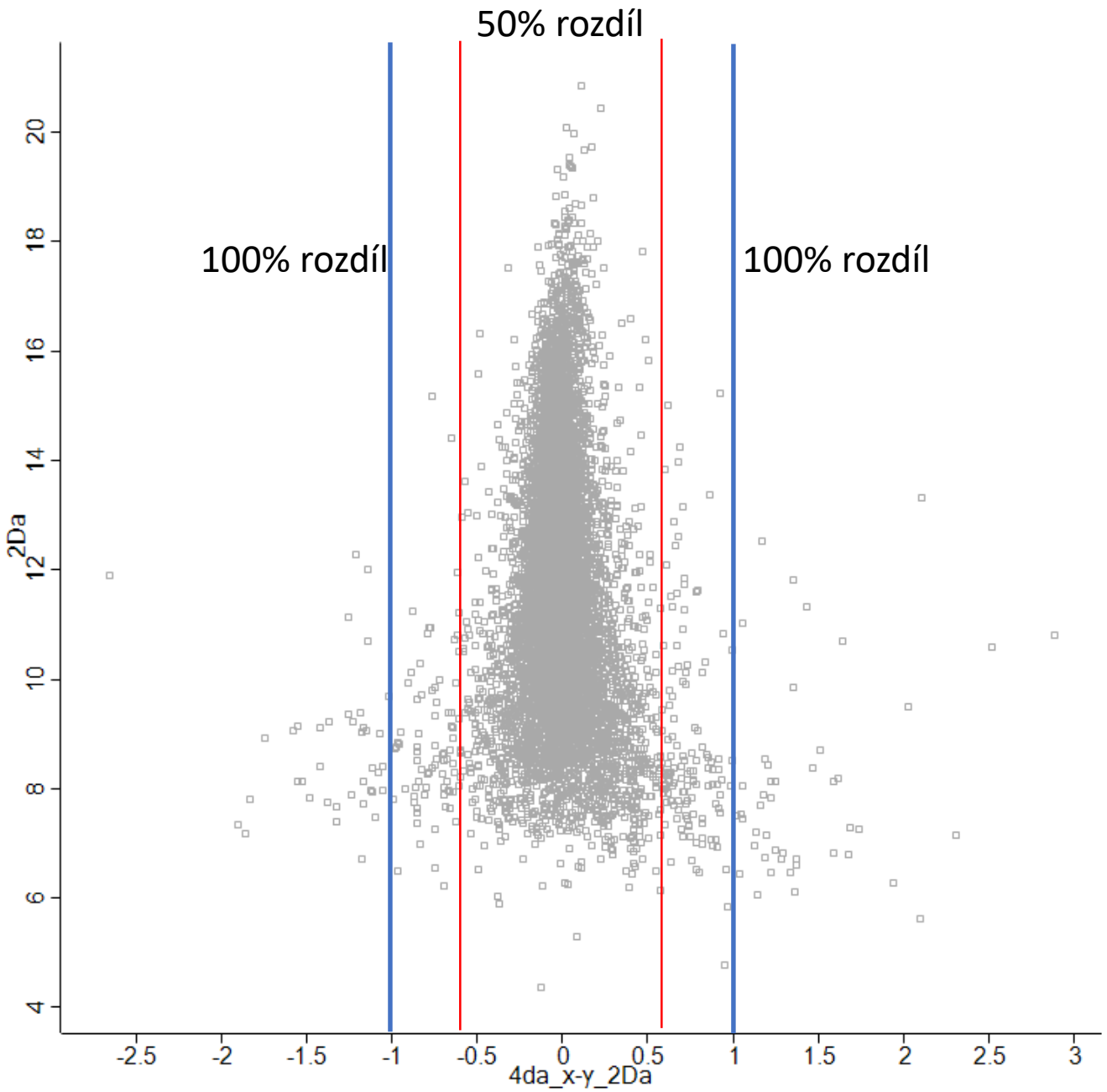
Skutečný stav



Zelená – stav zobrazený měřením se CV=5%

Zelená – stav zobrazený měřením se CV=20%





Rychlá screeningová metoda
Median CV 21%
6 replikátů, náhodně rozděleno na dvě skupiny.

Osa X – binární logaritmus poměrů průměrů těchto skupin

Osa Y – intezita v binárním logaritmu

Teoretická distribce s CV 1% - Modrá

Změřená kvantifikační metodou s CV 25% Červená

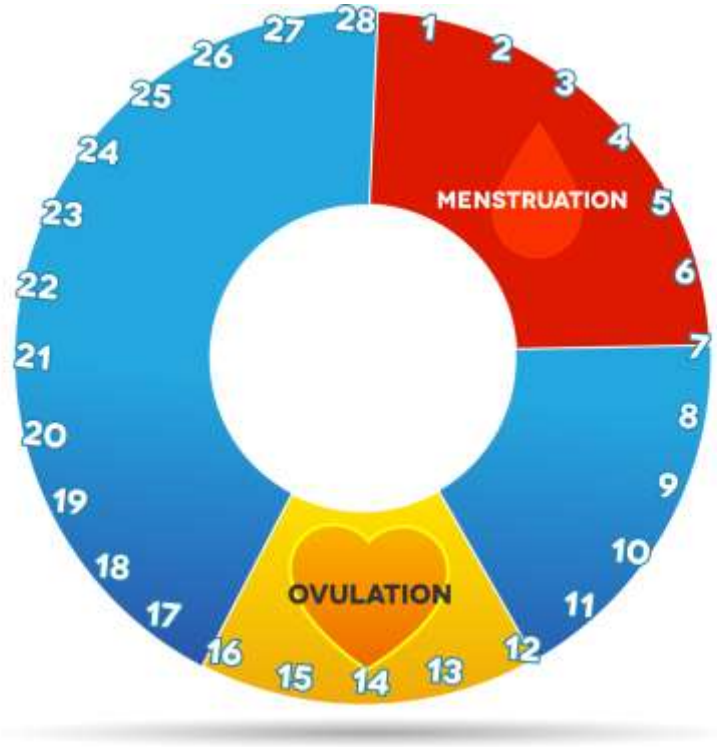
Statistické testy

- Experiment se snaží odpovědět na otázku jak se od sebe liší kontrola a vzorek
- Statistické testy umí vyvrátit nulovou hypotézu

Follicle-stimulating hormone

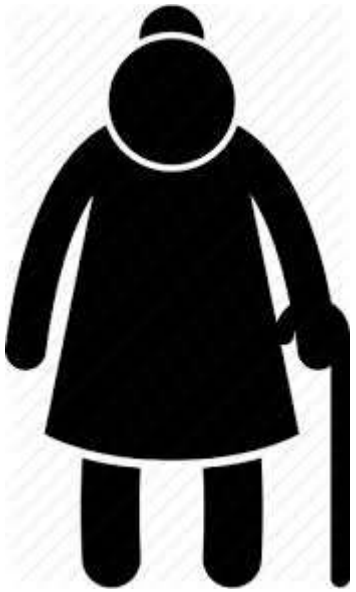
- https://en.wikipedia.org/wiki/Follicle-stimulating_hormone

117-331 ng/l



247-572 ng/l

292-1462 ng/l



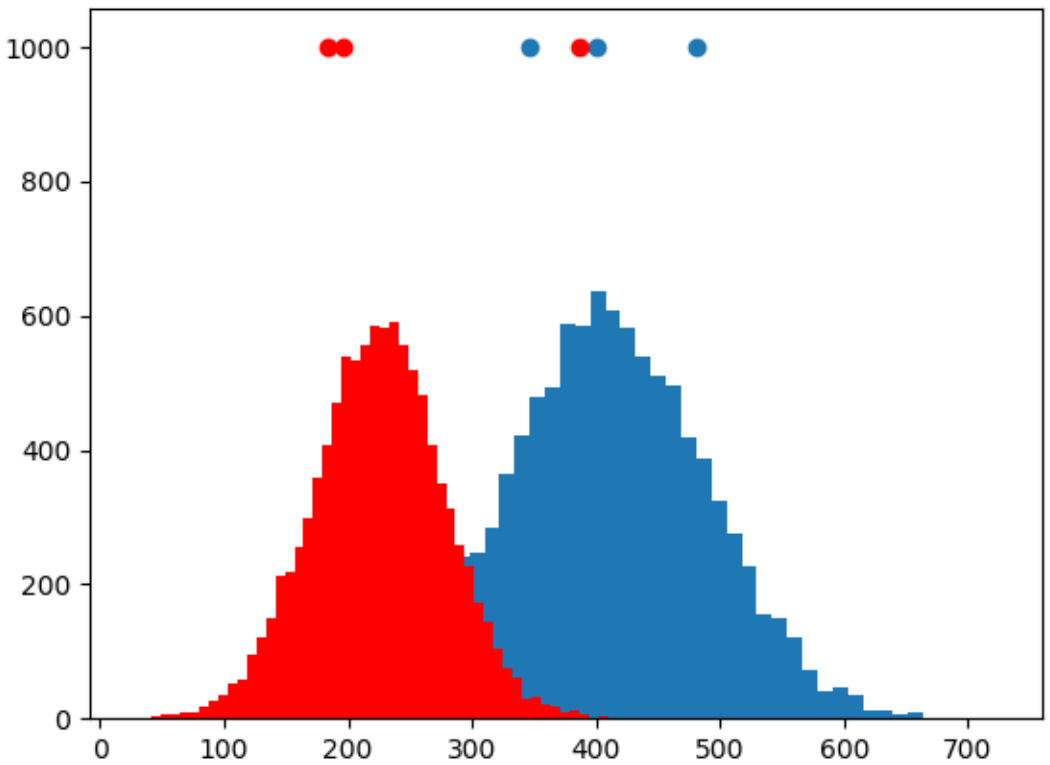
1088-7345 ng/l

1040-7540 ng/l

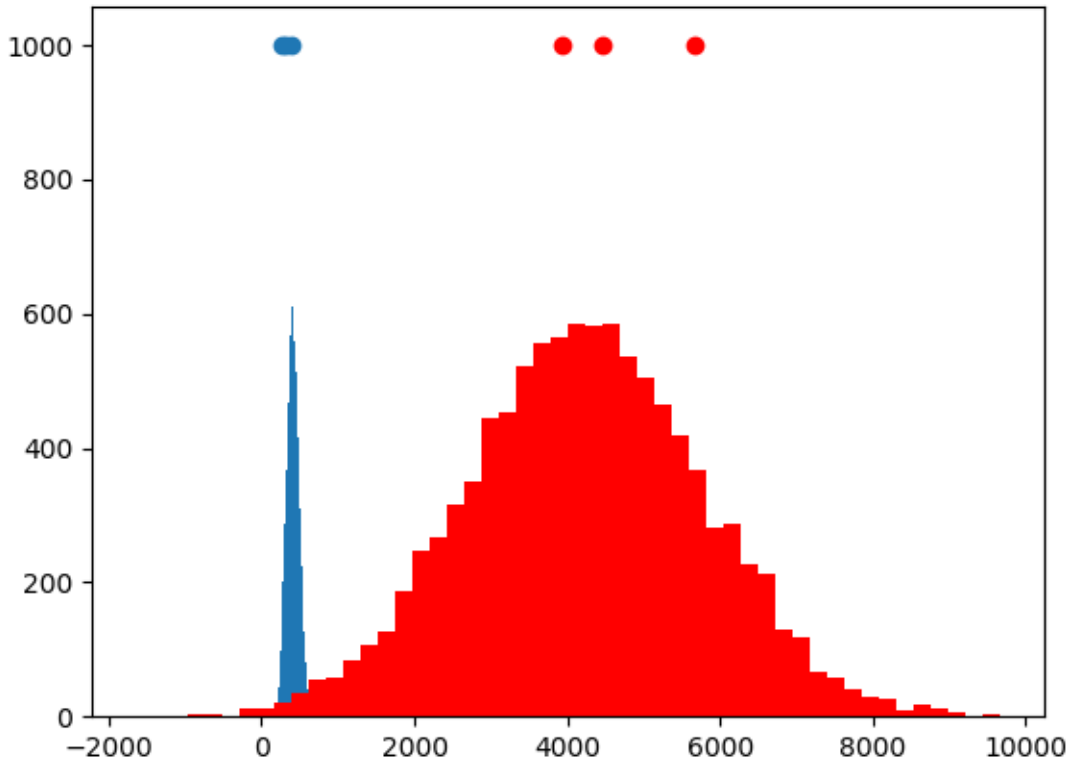
https://biologicalvariation.eu/meta_calculations

Všechna data pocházejí z wikipedie, grafy jsou generovány, nejedná se o skutečné měření

Po ovulaci / Před ovulací
224 ng/l 409 ng/l



Před ovulací/menopauza
409 ng/l 4216 ng/l

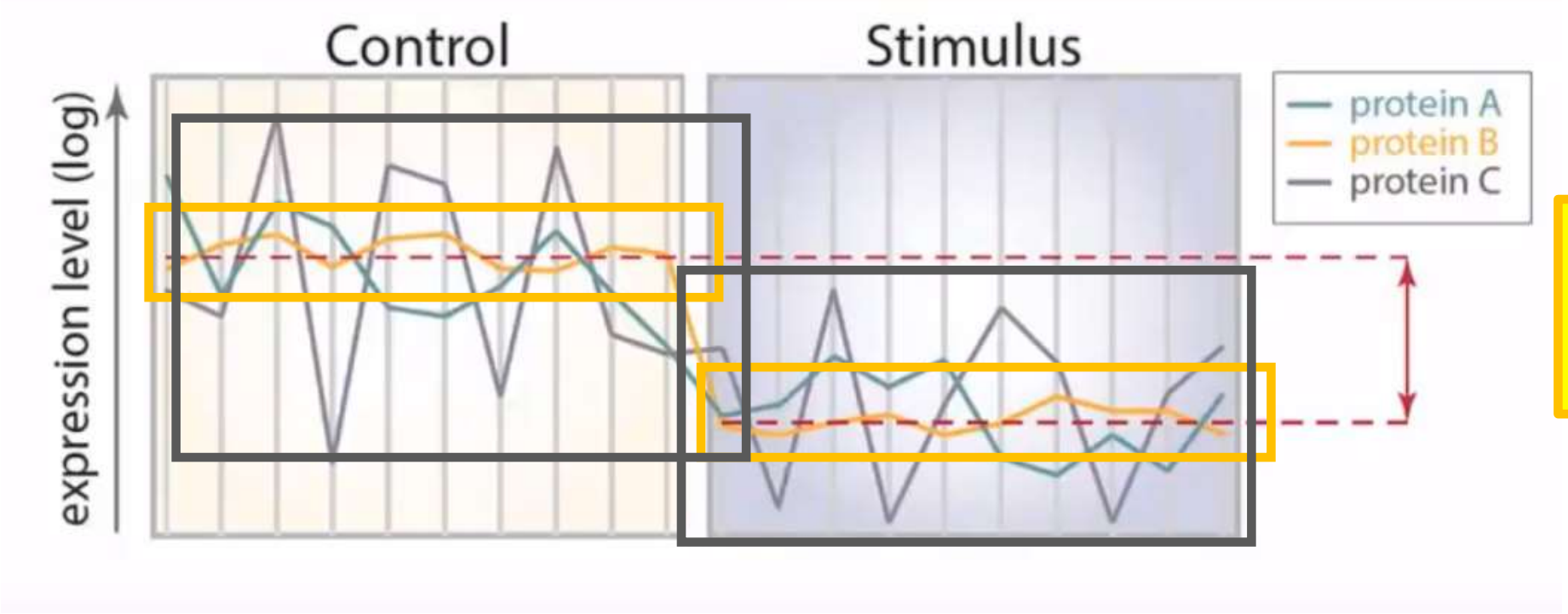


pValue

Hodnota pravděpodobnosti s jakou dostaneme stejnou nebo extrémnější distribuci dat náhodně

Hodnoty jsou v intervalu 0 až 1

1=100%



Relativní poměr proteinů je stejný

Statistická významnost je však jiná pValue

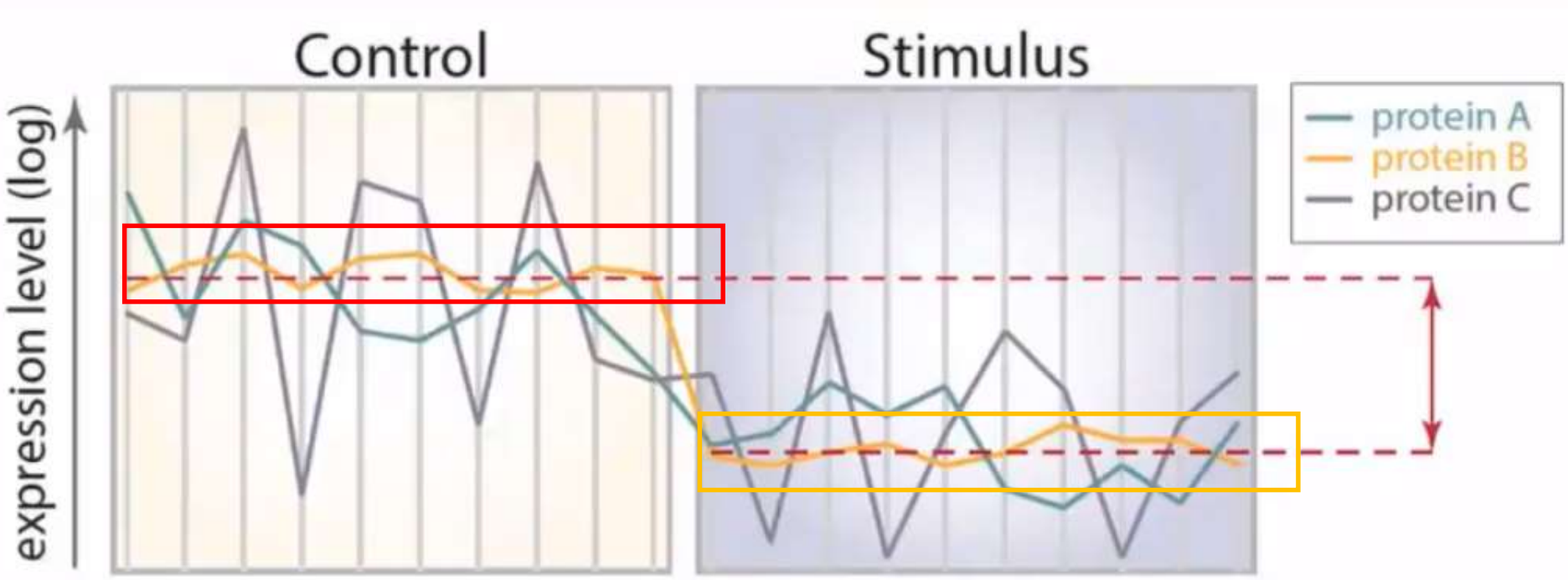
Univerzální hodnota pro více statistických testů

Není však jejich přímým výsledkem

Výsledek testu se přepočítává na pValue

0,05 -> 5%

Student T test pro proteomická data



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

T score – vyjadřuje podobnost dvou skupin

Bezrozměrná hodnota

Vysoká hodnota znamená že jsou rozdílné

Nízká hodnota znamená že jsou podobné

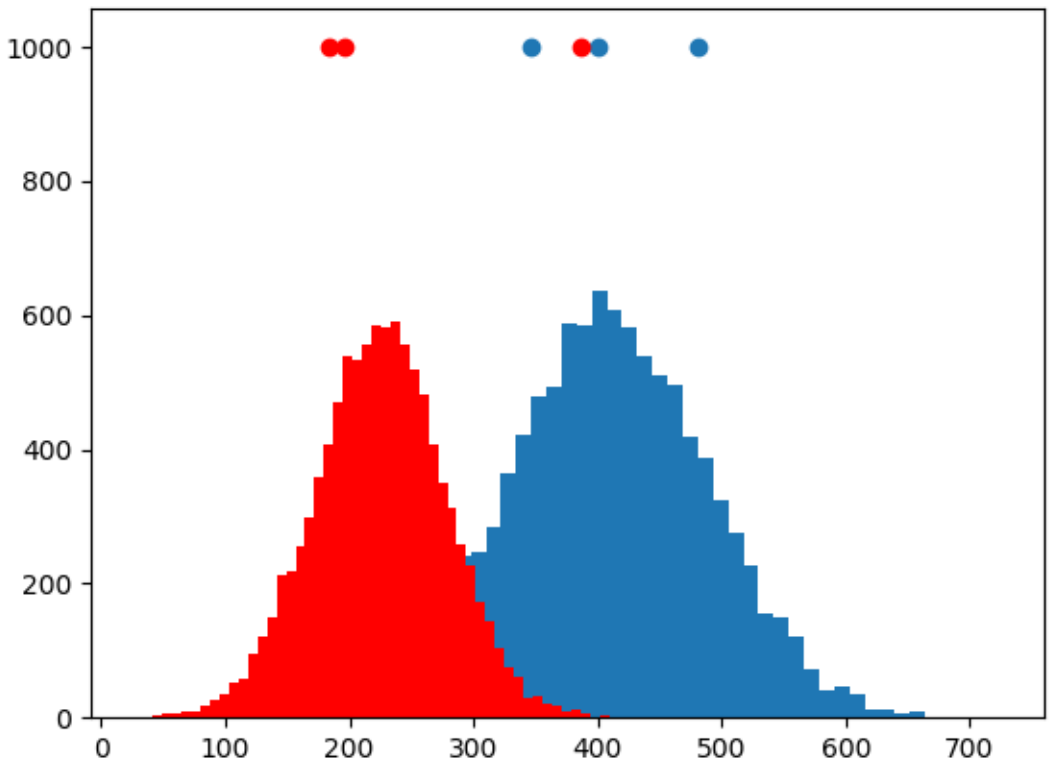
Student T test -> pValue

- Zadává se T score a počet stupňů volnosti (počet vzorků)
- Ručně je možné použít převodní tabulku (t distribution table)

Degrees of freedom *(df)*	.2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587

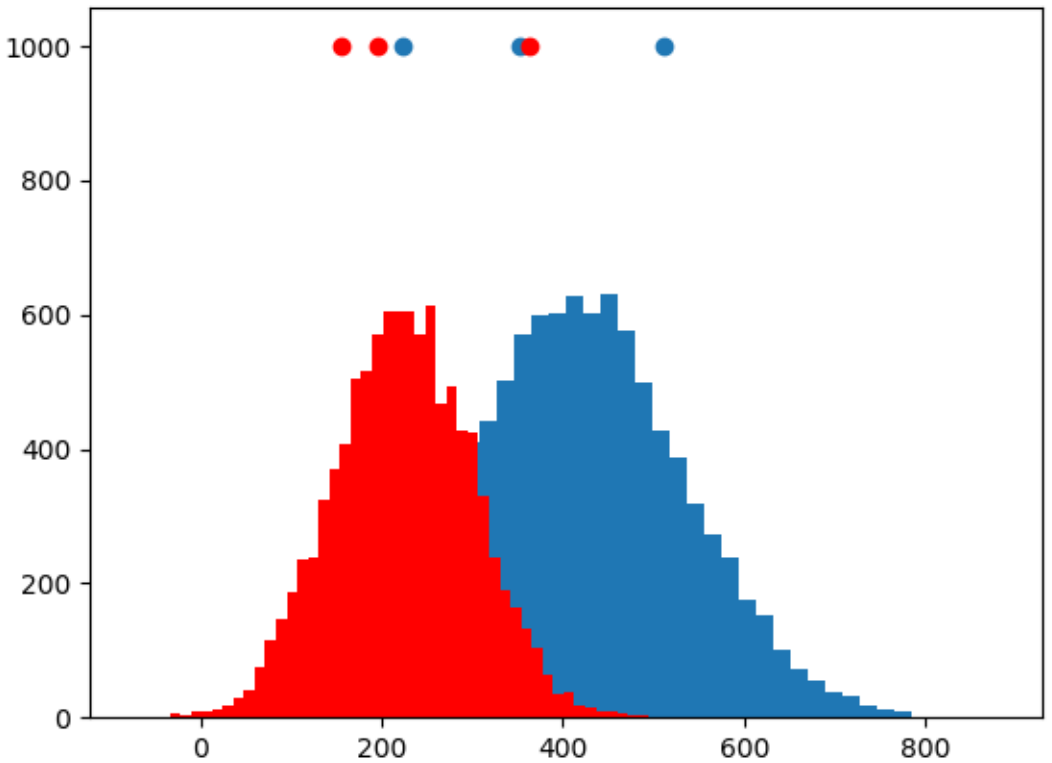
- Mnoho online kalkulátorů
- Excel funkce T.TEST vrací přímo pValue

Po ovulaci / Před ovulací
224 ng/l 409 ng/l



3 opakování, 68 ze sta testů úspěšných
4 opakování více než 80%

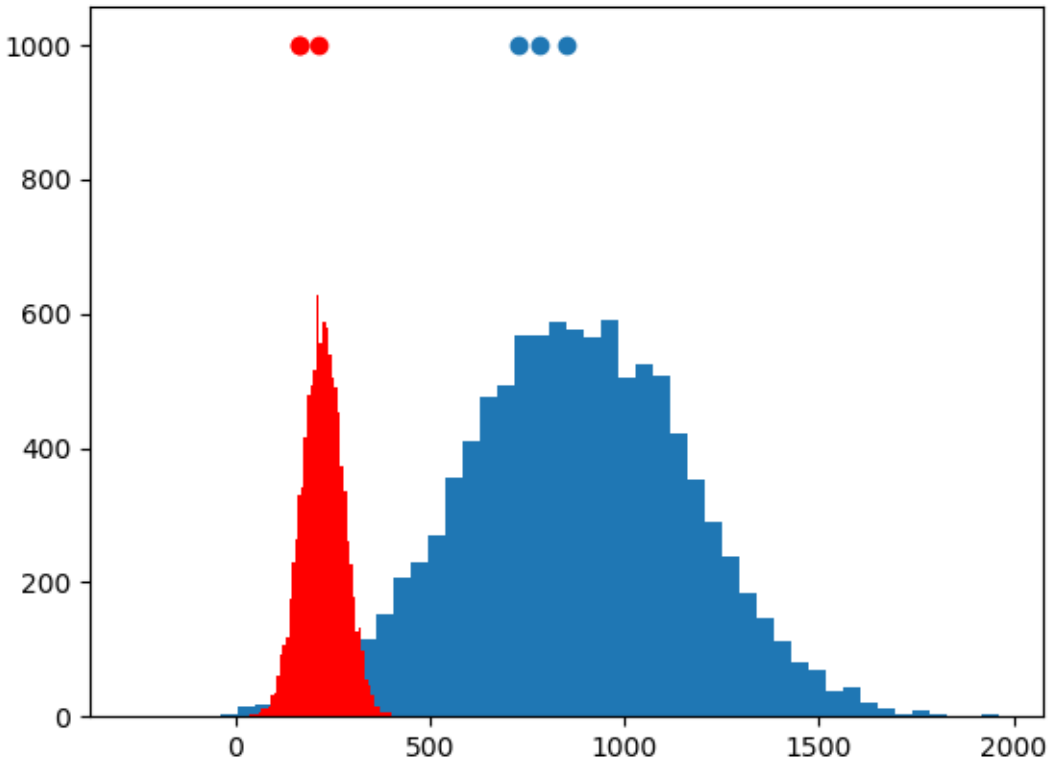
Po ovulaci / Před ovulací
224 ng/l 409 ng/
10% CV technická variabilita



3 opakování, 36 ze sta testů úspěšných
6 opakování více než 80%

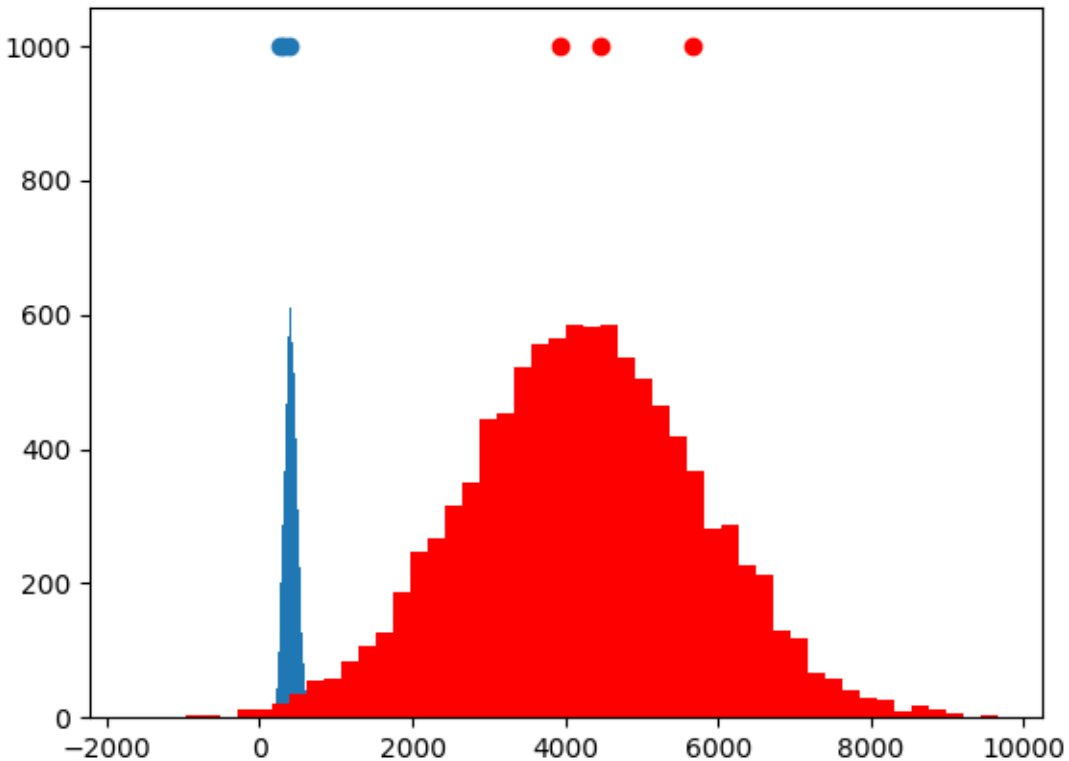
Všechna data pocházejí z wikipedie, grafy jsou generovány 10000 hodnot v daném intervalu definovaném průměrem a smodch, nejedná se o skutečné měření

Po ovulaci/Ovulace
224 ng/l 877ng/l



3 opakování, 87 ze sta testů úspěšných

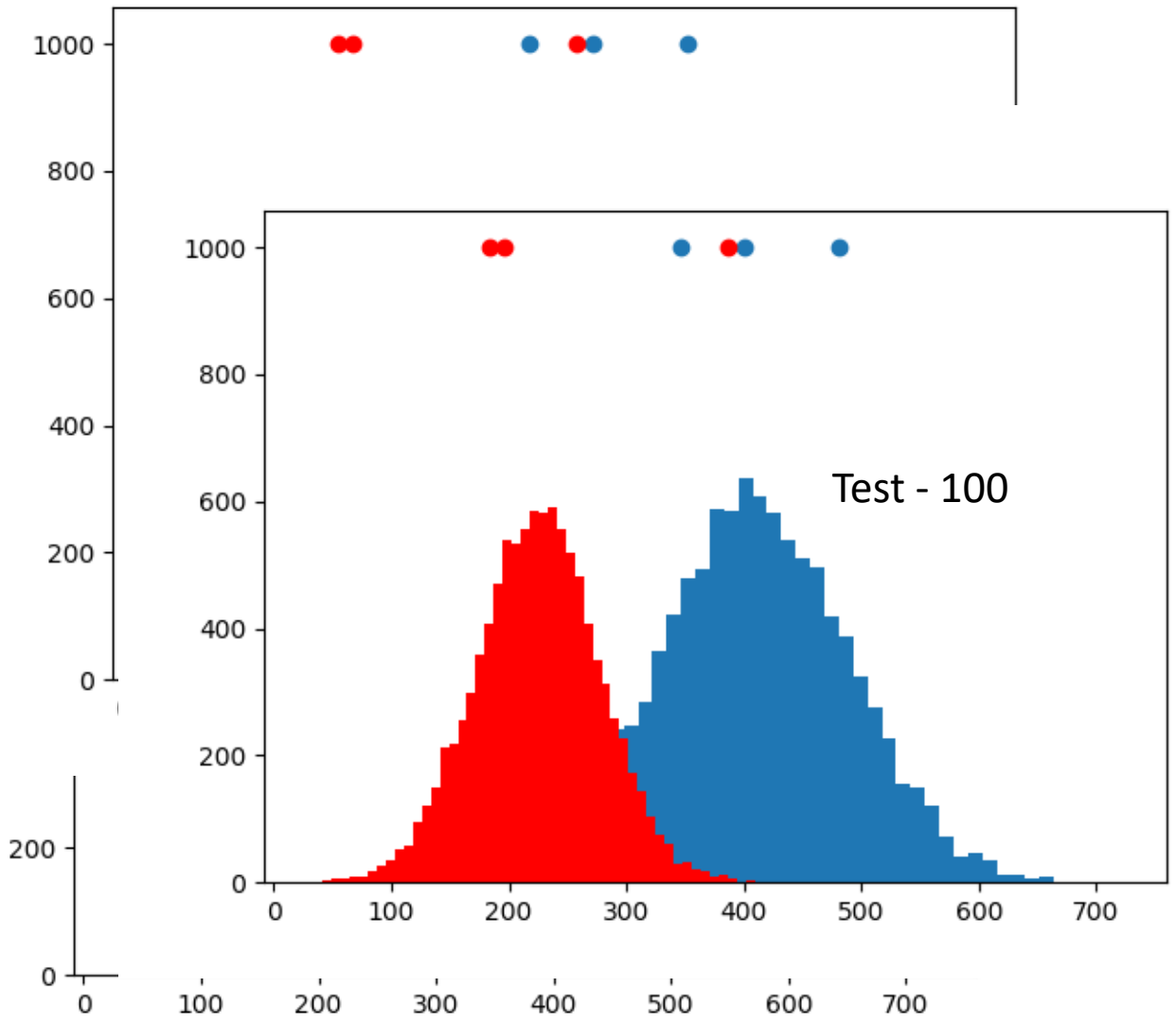
Před ovulací/menopauza
409 ng/l 4216 ng/l



3 opakování, 90 ze sta testů úspěšných

Všechna data pocházejí z wikipedie, grafy jsou generovány 10000 hodnot v daném intervalu definovaném průměrem a smodch, nejedná se o skutečné měření

Síla testu



Určení kolik vzorků musím změřit , abych odlišil dvě distribuce proteinu

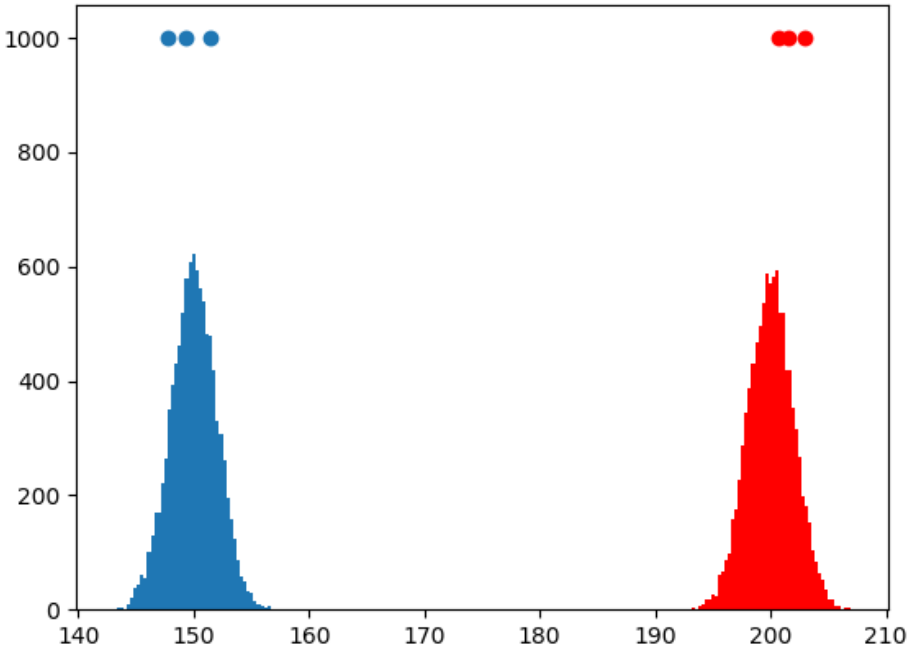
Vstupy :
1)průměr skupiny A a B
2)Směrodatná odchylka skupiny A a B
3)Požadovaná hladina signifikance

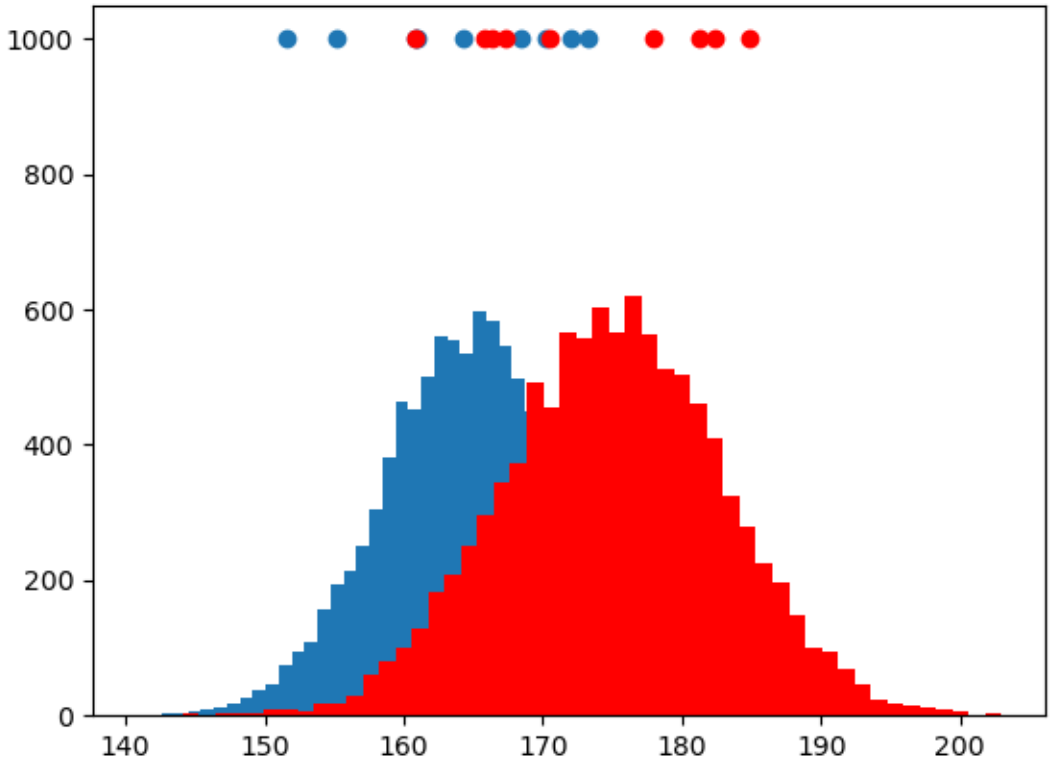
Výstup: počet vzorků na skupinu abych v 80% případů správně určil, že se distribuce liší

- Různé on line kalkulátory
- https://onlinestatbook.com/2/calculators/power_calc.html

https://www.youtube.com/watch?v=VX_M3tIyiYk

- Hodnoty do výpočtu získáme
  - Pilotního měření
  - Odvodit z volně dostupných dat
    - Proteomické experimenty v repozitáři
    - https://biologicalvariation.eu/meta_calculations
  - Literatury
  - Kvalifikovaného odhadu





Praxe

- Ve většině případů se měří tři hodnoty
  - Stačí pro výpočet pValue (t-Test)
  - Vzácnost vzorku
  - Časové
  - Finanční
    - Cena přípravy vzorku
    - Cena měření
  - Často to stačí
    - Kromě situací kdy to nestačí
    - Po změření už nelze napravit
- Při plánování velkých experimentů má smysl se nad možností výpočtu alespoň zamyslet

Počty vzorků - příklady

Orientační počty založené na zkušenostech v naší laboratoři – doporučujeme našim uživatelům*

- Imbrední myš
  - Shodný věk, pohlaví a podmínky chovu
  - Více než 3 vzorky (5-6) je dobré
- Divoká myš
  - 5 a více vzorků

- Imunoprecipitace
  - 3 jsou dostatečné

- Buněčná linie
  - 3 jsou dostatečné

- Fosfoproteomika z buněčné linie
  - 5 a více replikátů
- Lidská séra
  - Desítky budou málo

*jedná se o subjektivní názor

Hodnota pValue v experimentu

- Pro každý protein samostatný t Test
- Tisíce testů na experiment
- P Value 0,05 říká, že v 5% je platná nulová hypotéza
  - Obdržíme falešně pozitivní výsledek
- Proteomická data
  - Problém vícenásobného testování
  - **Při 5000 testech můžeme dostat 5000*0,05=250 falešně pozitivních výsledků**

Je použití fixní hodnoty správné?

Není

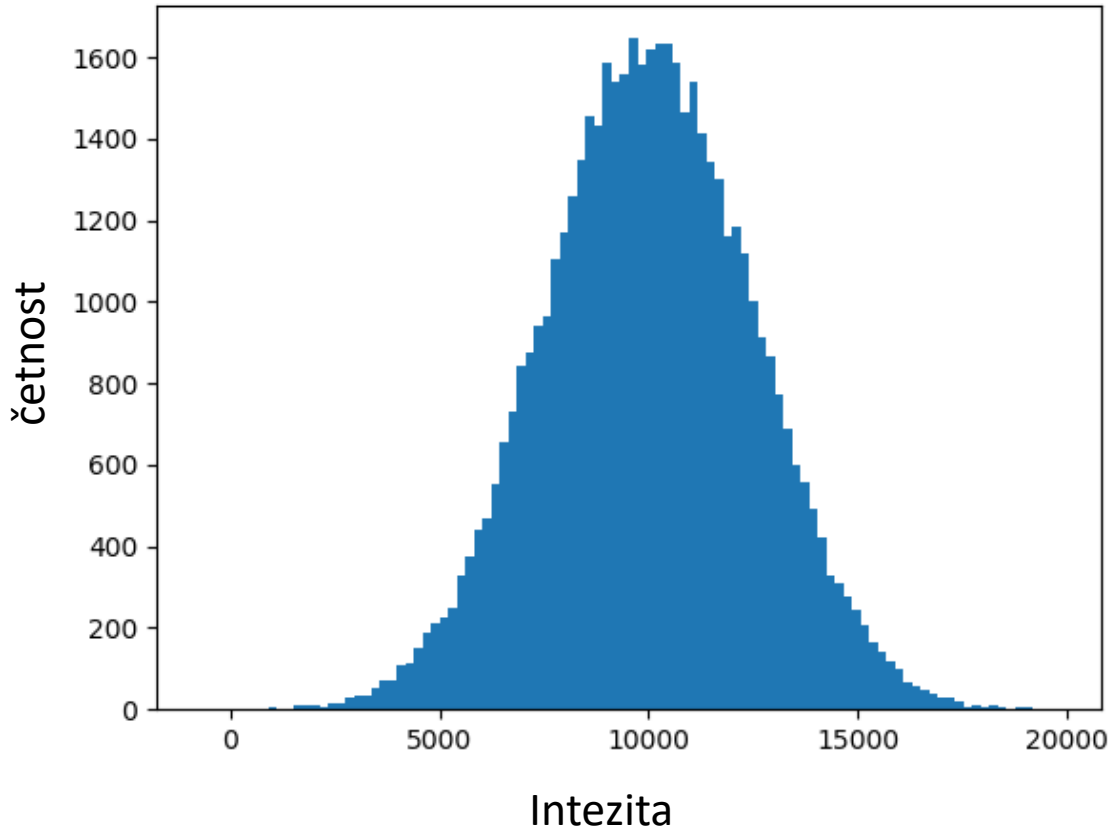
Při stovkách a tisících testů v rámci datasetu je třeba nástroj pro úpravu potřebné pValue

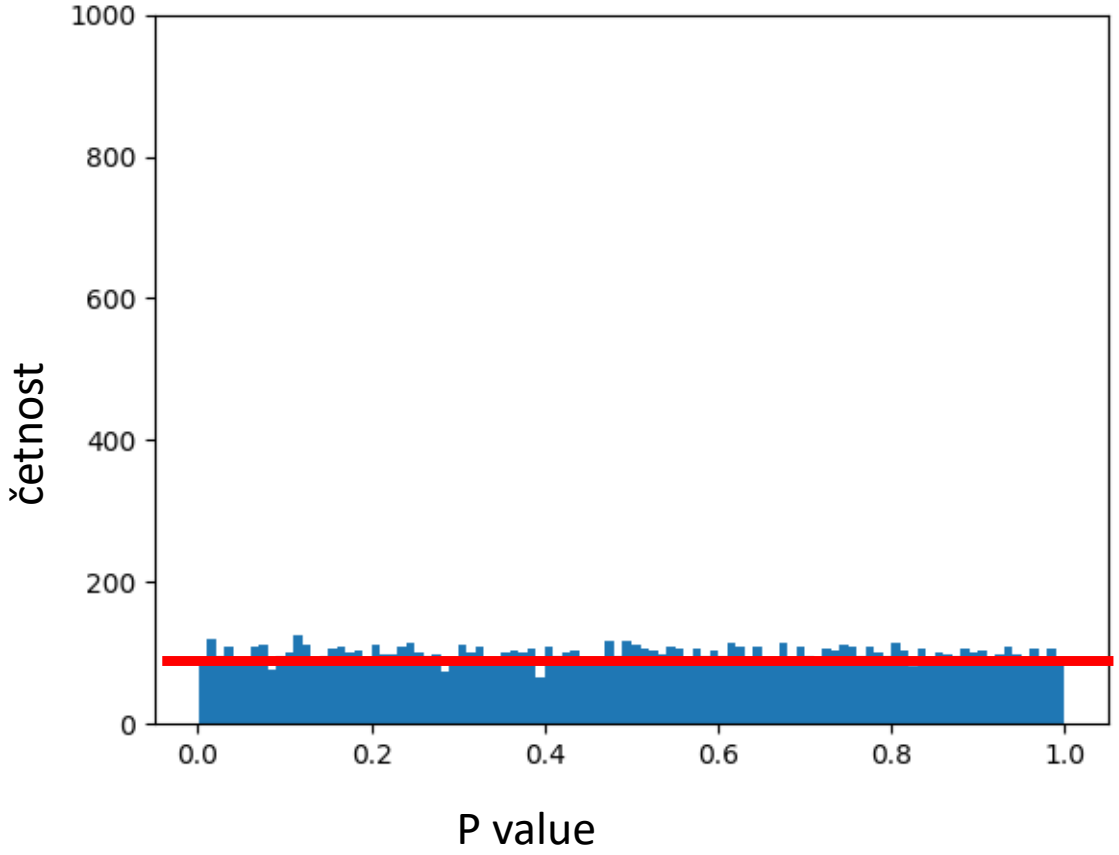
Bonferroni correction

- Nejstarší varianta
- dělení cílové pValue počtem testů -> α/m
- Příklad 20 testů v rámci analýzy a $\alpha=0,05$
  - 0,05/20=0,0025
- Velmi konzervativní, produkuje velké množství falešně negativních výsledků

Určení FDR výpočtem

10 000 výběrů tří a tří čísel -> výpočet pValue



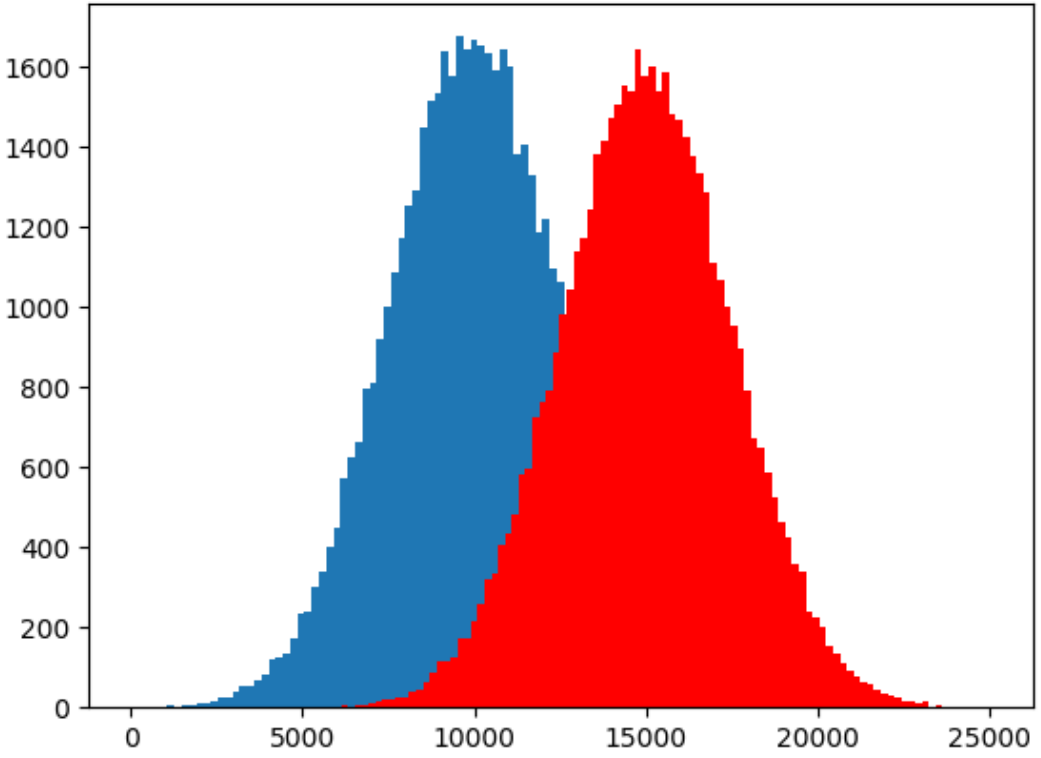


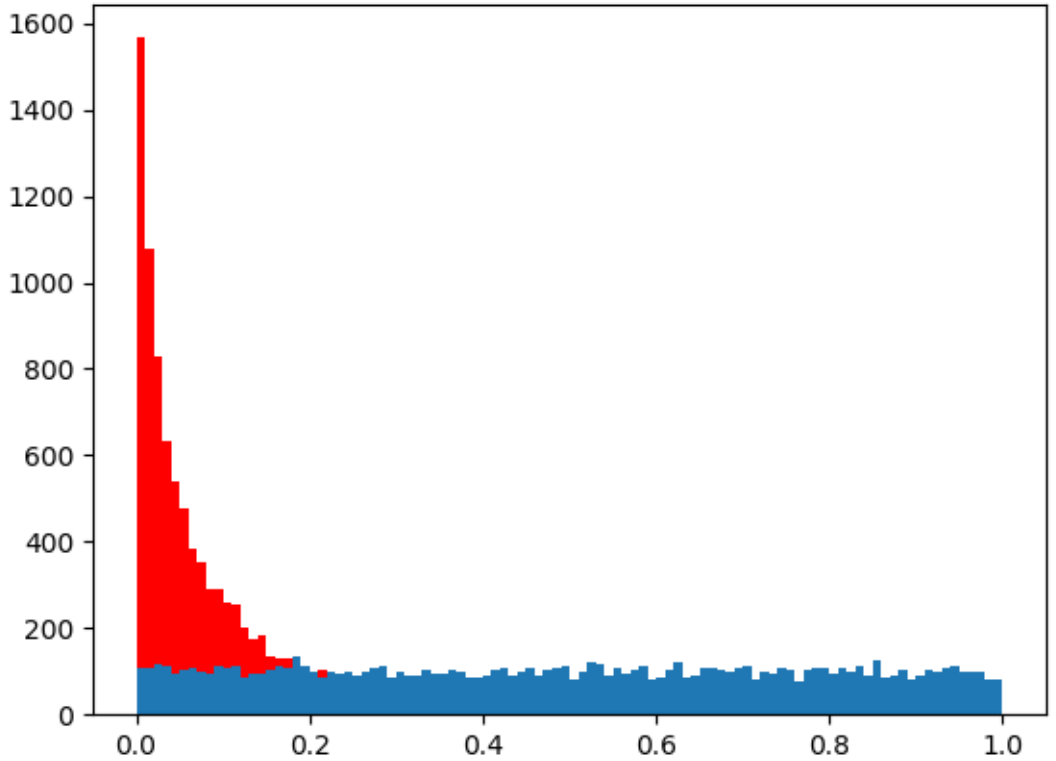
Modrá

10 000 výběrů tří a tří čísel z modré distribuce -> výpočet pValue

Červená

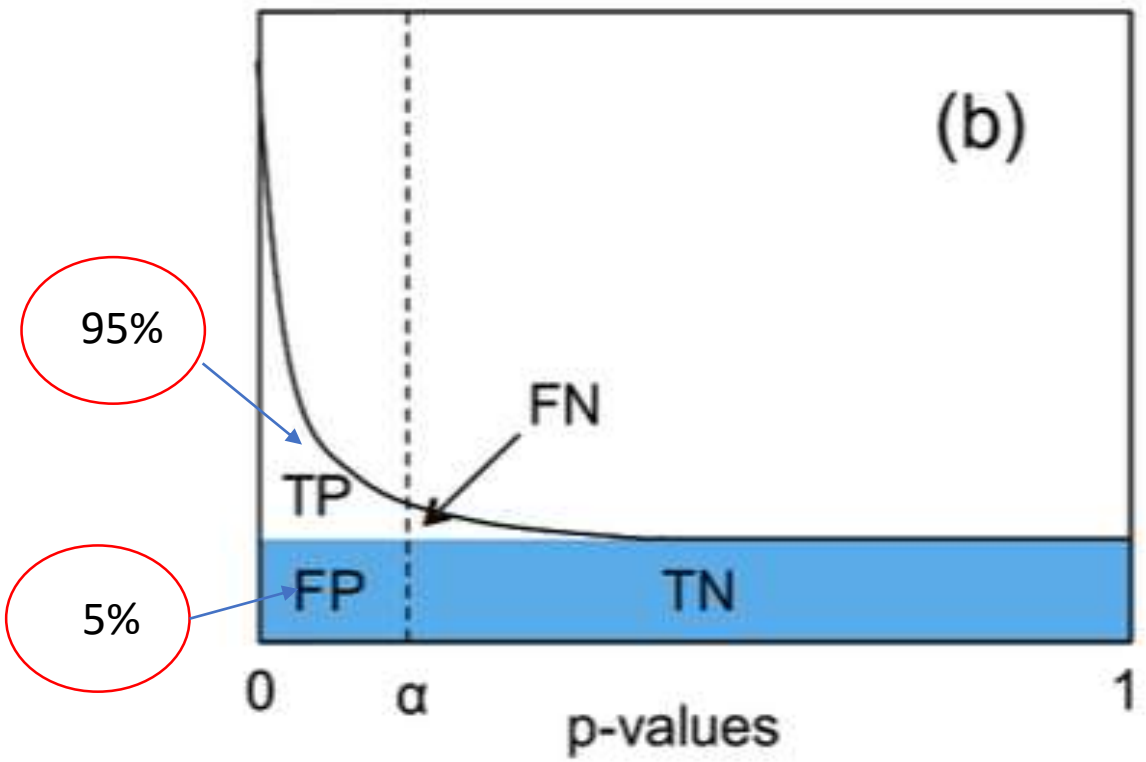
10 000 výběrů tří modrých a tří červených čísel -> výpočet pValue





FDR korigovaná pValue – permutation based

- Využití permutování dat
- Stovky až tisíce permutací
- Přímé určení Q value
- Citlivější než **Benjamini–Hochberg**



J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 479–498

A direct approach to false discovery rates

John D. Storey

Stanford University, USA

[Received June 2001. Revised December 2001]

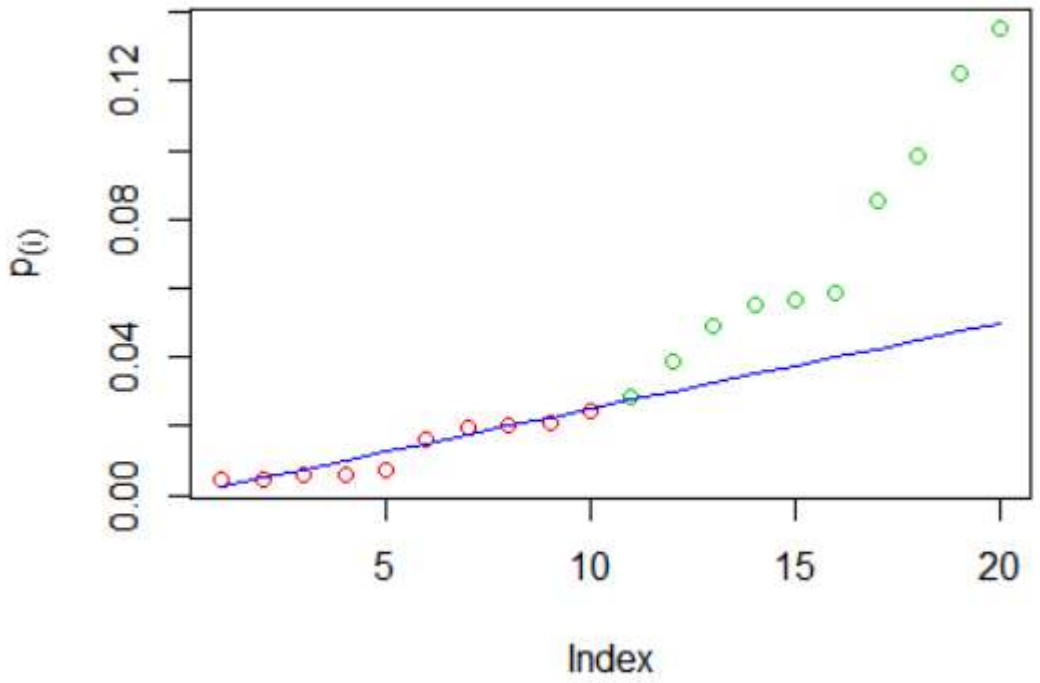
https://www.youtube.com/watch?v=T6J4b-WWebM

Adjusted pValue

Benjamini–Hochberg

FDR korigovaná pValue - **Benjamini–Hochberg**

- Seřadíme všechny pValues vzestupně
- Hledá se průsečík hodnot a přímky se sklonem **α/počet testů**
- Všechny pValues s pořadím nižším než pValue kde se obě vynesení protlnuly jsou považovány za signifikantní



JOURNAL ARTICLE

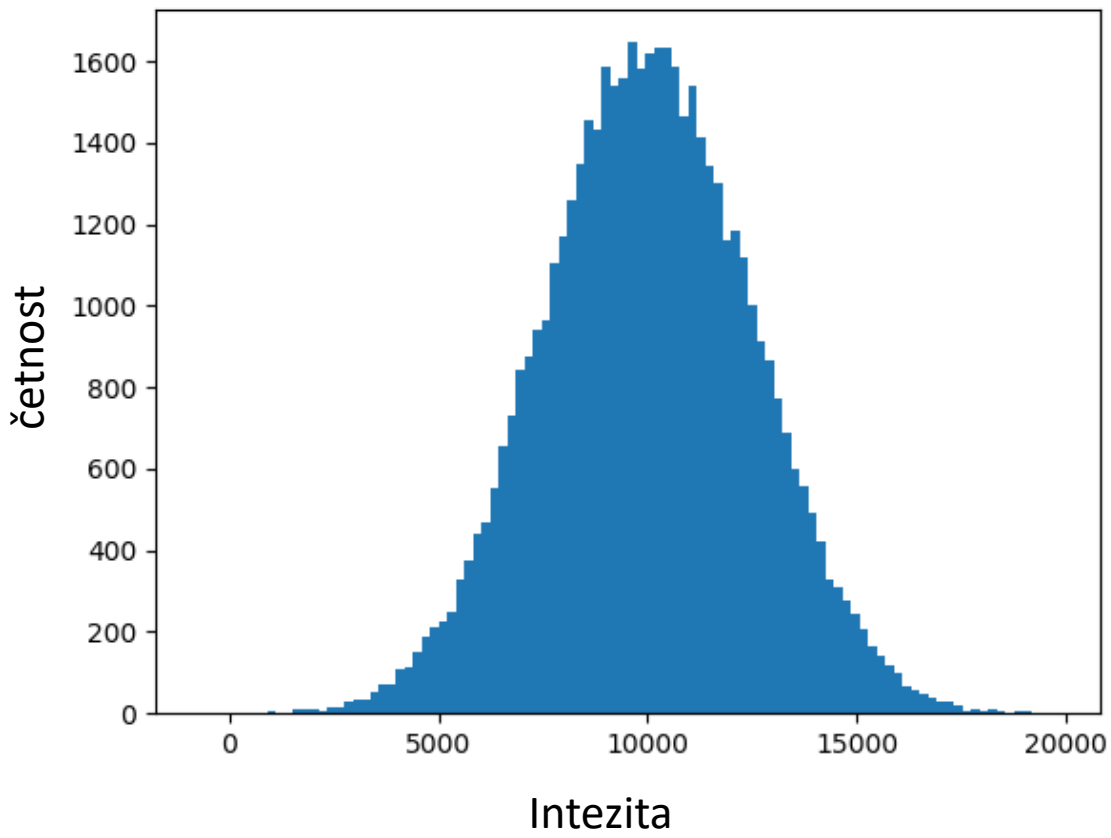
Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing

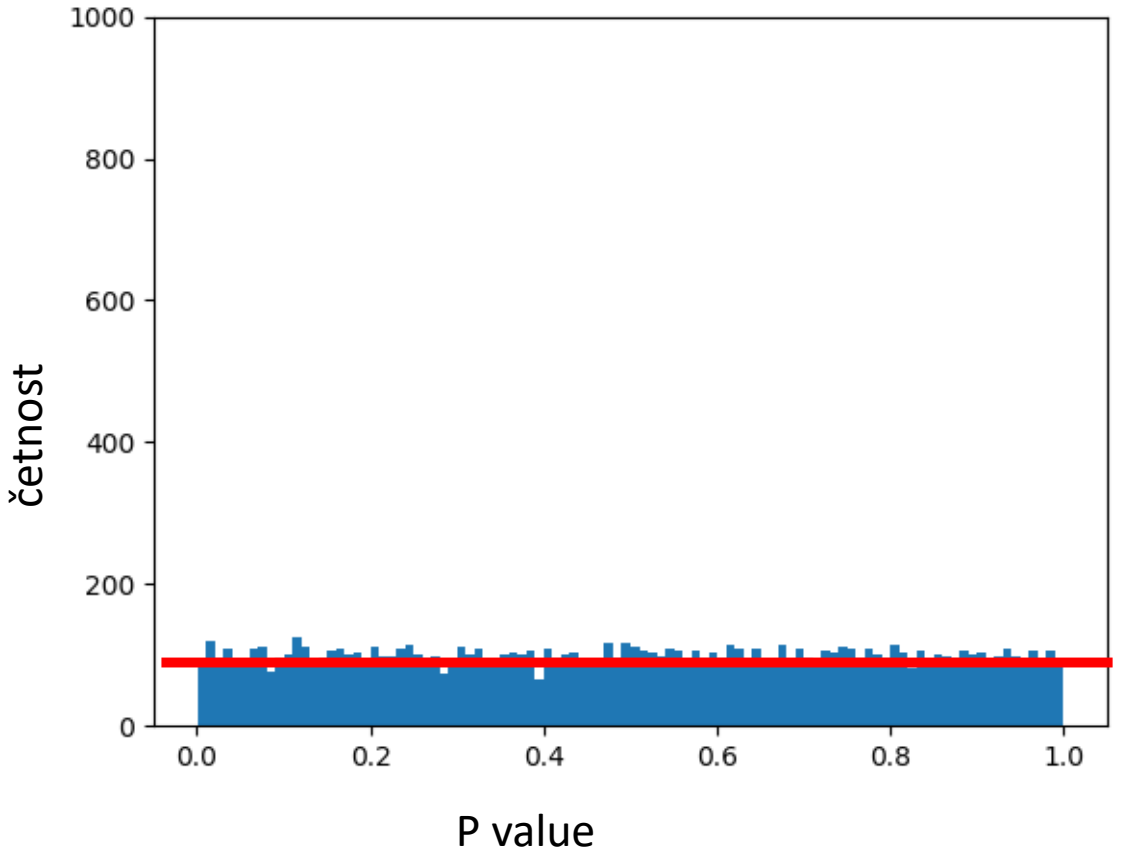
Yoav Benjamini and Yosef Hochberg

Journal of the Royal Statistical Society. Series B (Methodological)
Vol. 57, No. 1 (1995), pp. 289-300 (12 pages)
Published By: Oxford University Press

https://www.youtube.com/watch?v=K8LQSvtjcEo

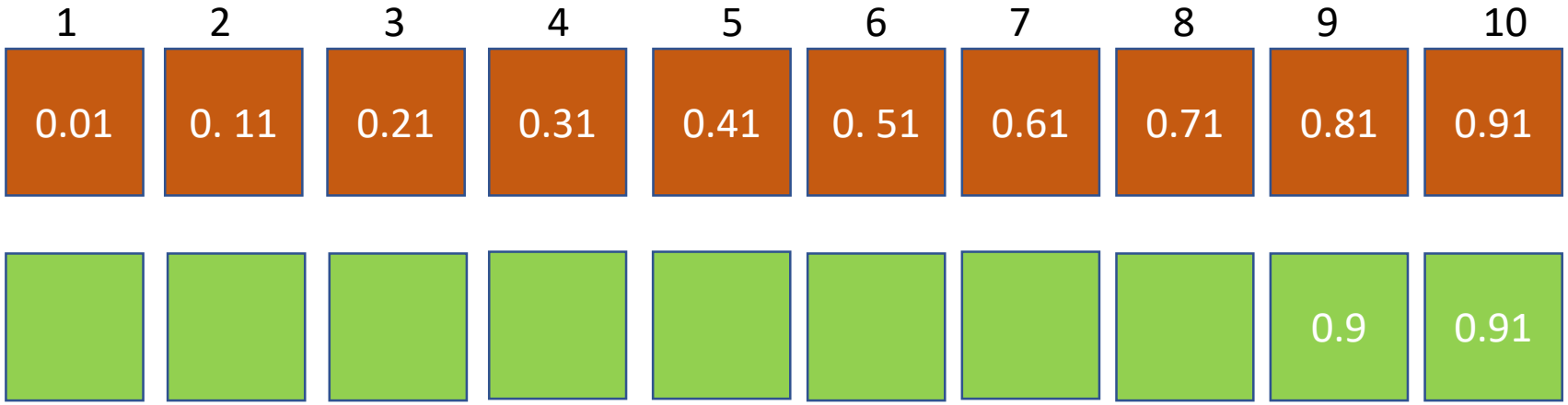
10 výběrů tří a tří čísel -> výpočet pValue





0. 01	0.21	0.91	0.41	0.01	0.31	0.81	0.61	0.71	0. 51

https://www.youtube.com/watch?v=K8LQSvtjcEo

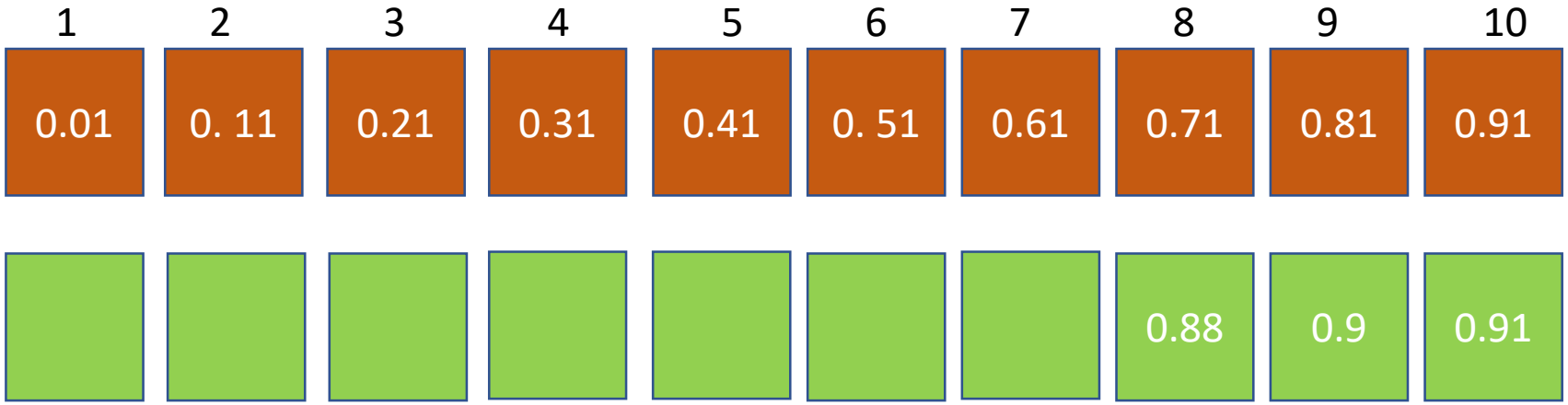


Menší z těchto dvou:

1) Předchozí upravená pValue — 0.91

2) Aktuální pValue*(celkem pValue/pořadí aktuální pValue) — 0.81*(10/9)=0.9

https://www.youtube.com/watch?v=K8LQSvtjcEo

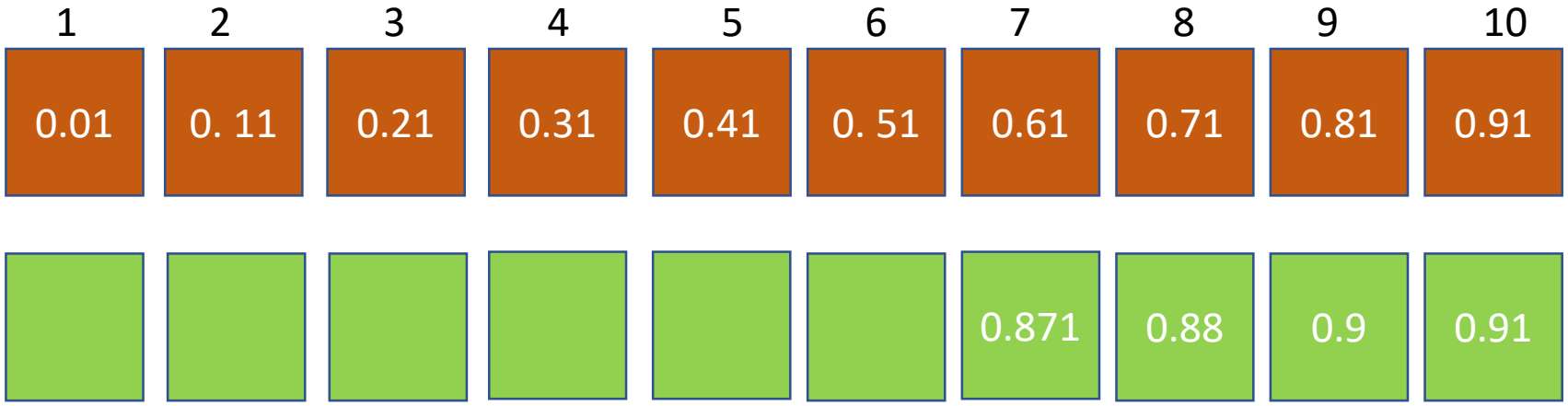


Menší z těchto dvou:

1) Předchozí upravená pValue — 0.9

2) Aktuální pValue*(celkem pValue/pořadí aktuální pValue) — 0.71*(10/8)=0.8875

https://www.youtube.com/watch?v=K8LQSvtjcEo

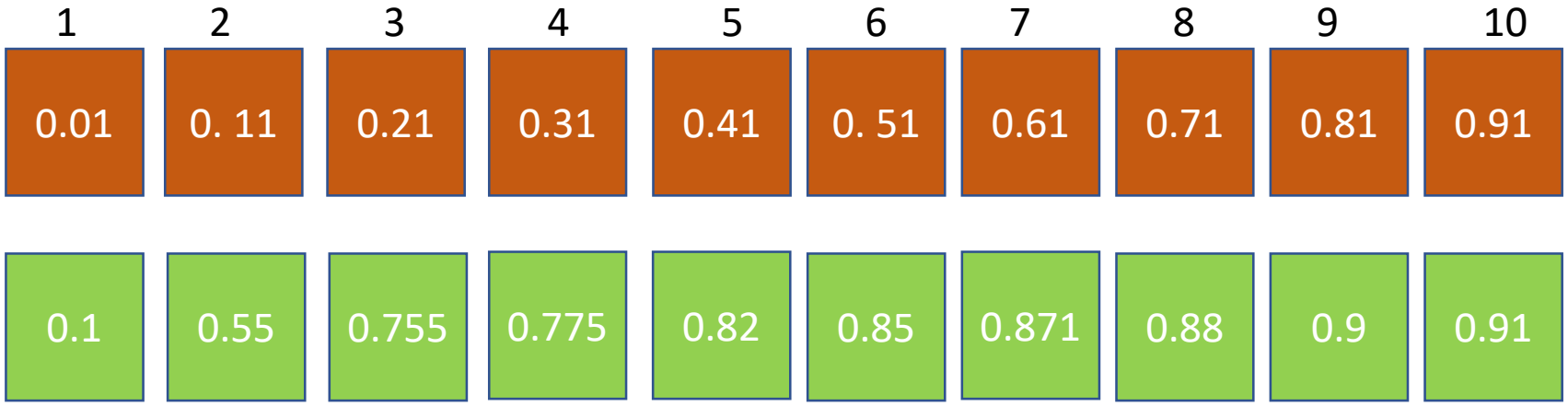


Menší z těchto dvou:

1) Předchozí upravená pValue — 0.88

2) Aktuální pValue*(celkem pValue/pořadí aktuální pValue) — 0.61*(10/7)=0.871

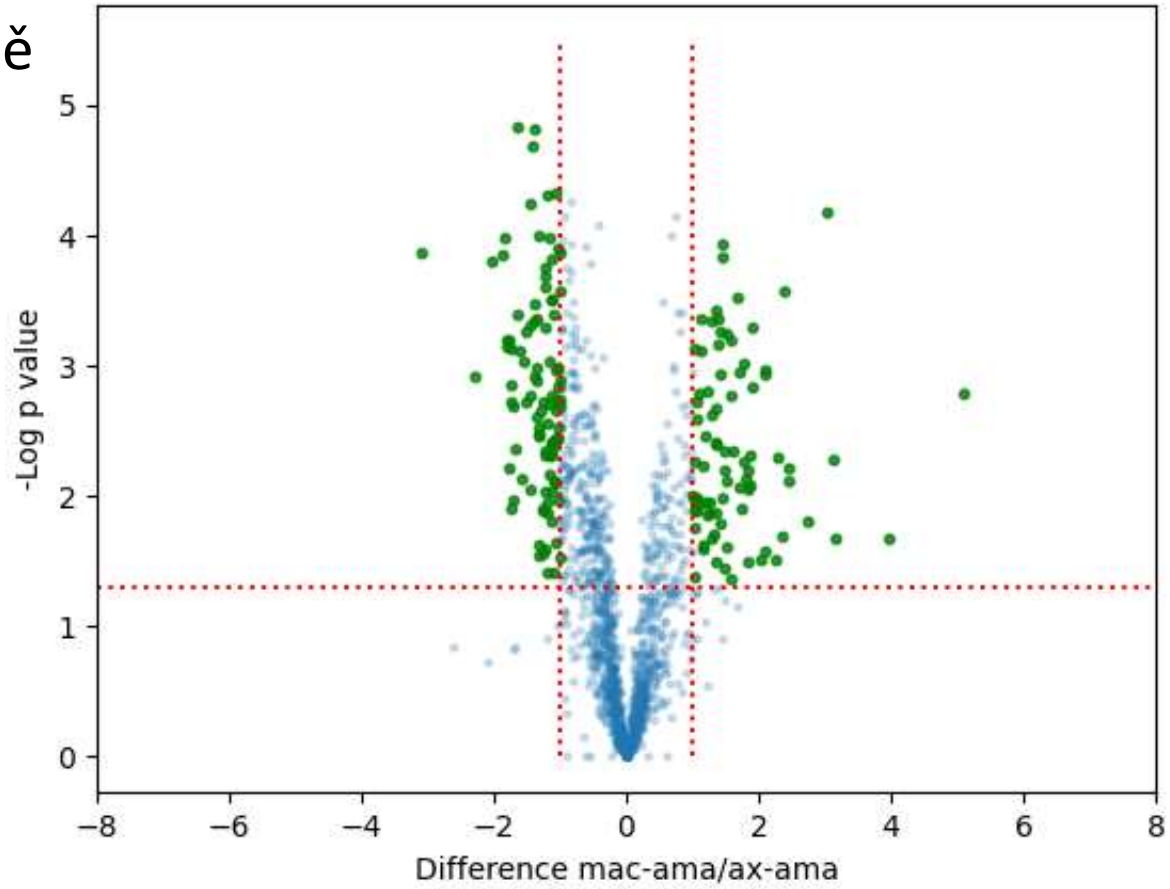
https://www.youtube.com/watch?v=K8LQSvtjcEo


1
2
3
4
5
6
7
8
9
10
0.01
0. 11
0.21
0.31
0.41
0. 51
0.61
0.71
0.81
0.91
0.1
0.55
0.755
0.775
0.82
0.85
0.871
0.88
0.9
0.91

Volcano plot

Volcano plot

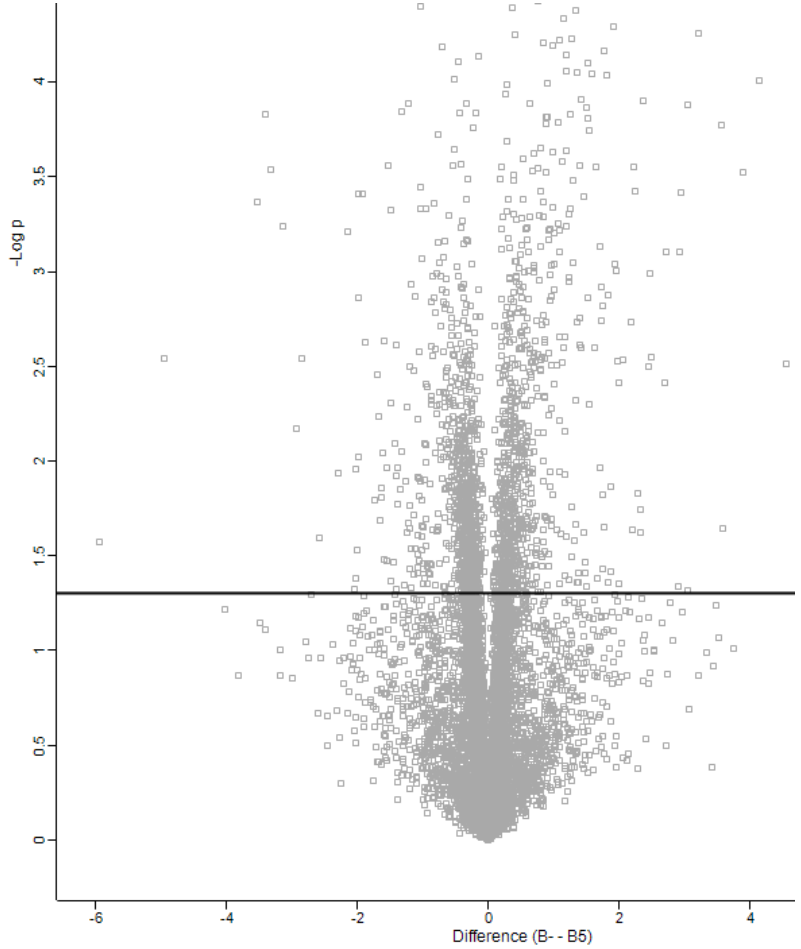
- Hladina significance – často používáno 0,05 a méně
- Pro účely vizualizace neprakticky nízké číslo
- Převedení na záporný logaritmus o základu 10
- 0,05 -> **1,3**
- Vynesení dat -> Volcano plot
  - Osa Y – pValue (–Log10)
  - Osa X – relativní změna v logaritmu o základu 2



- Vynesení pValue 0,05 a rozdílu 2x

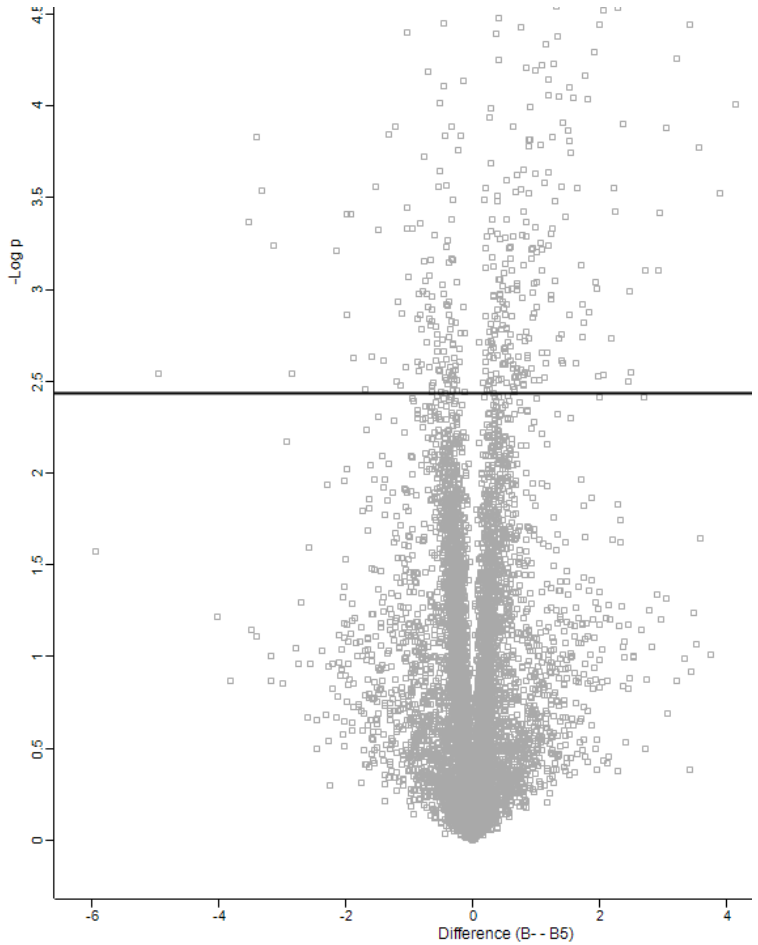
- Necílená proteomika není určena k detekci změn v nízkých desítkách procent
  - +/- **50%** je minimální realisticky zachytitelná změna

8148 kvantifikovaných proteinů



α=0,05 **bez korekce**

1672 signifikantních

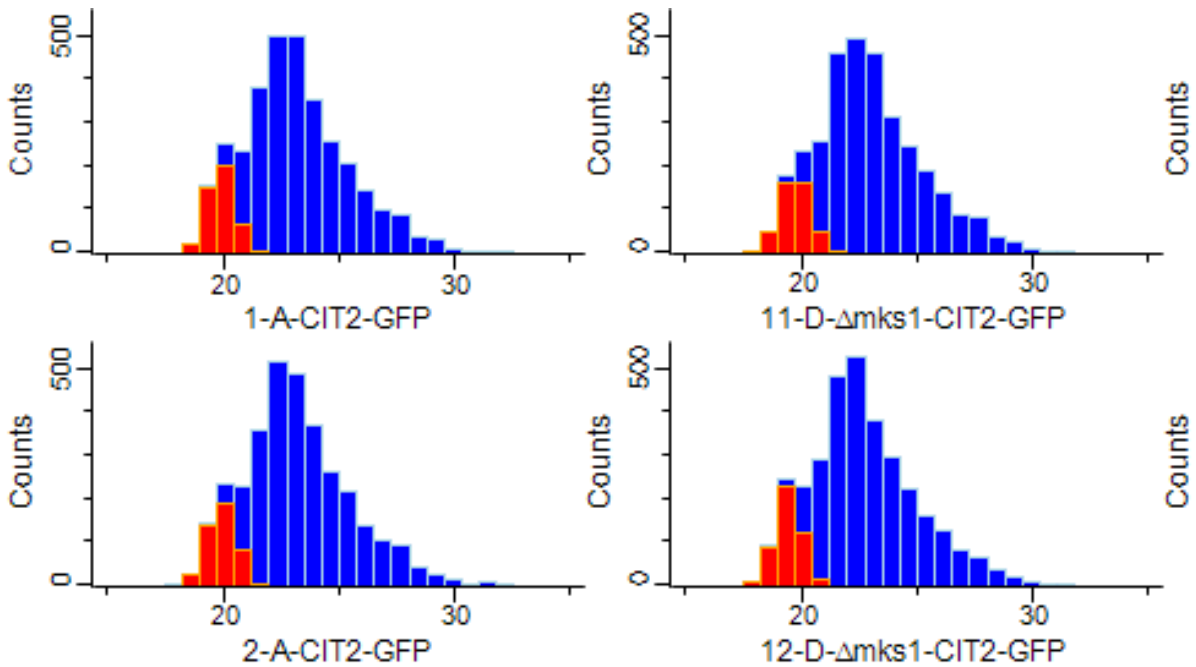


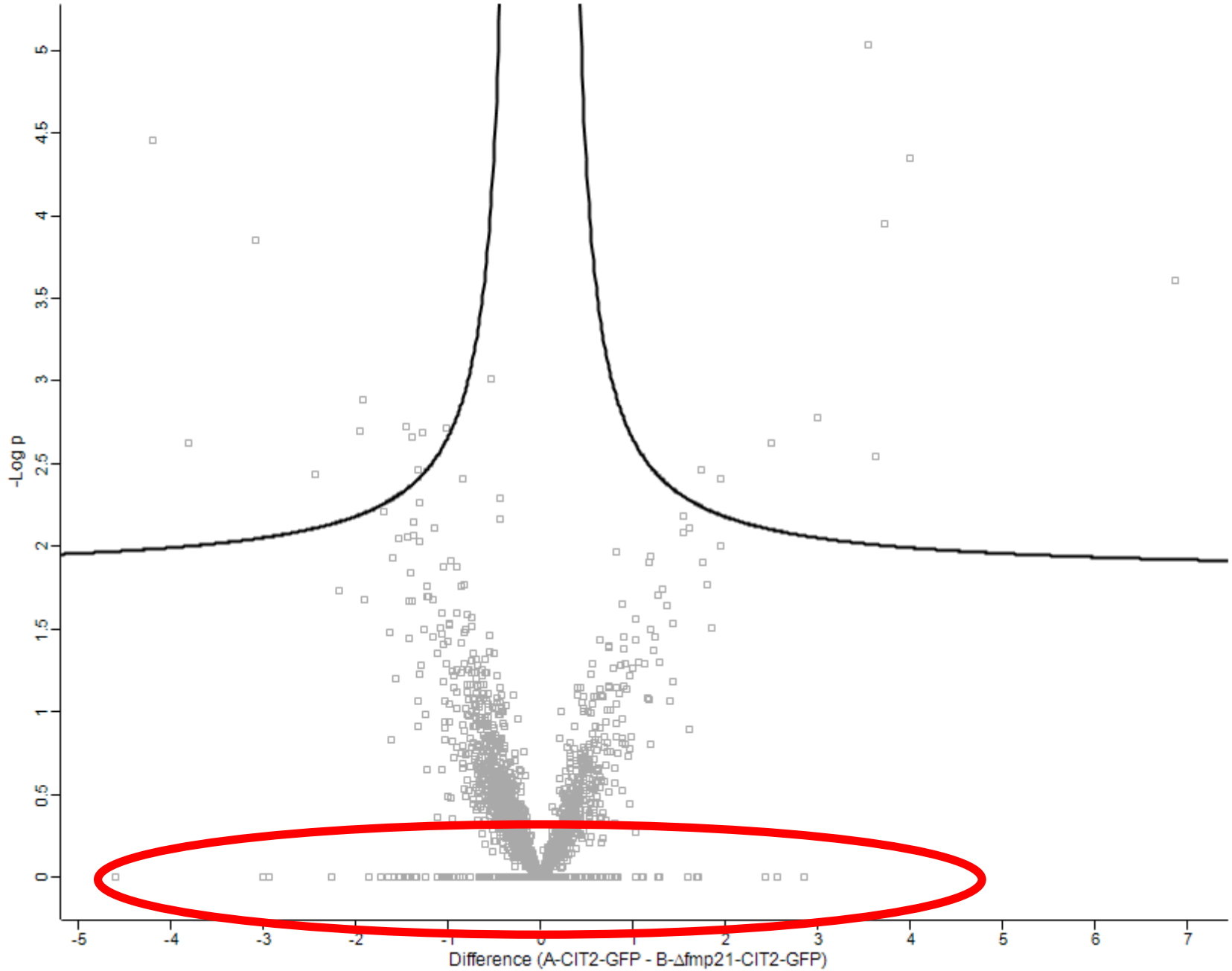
α=0,05 **5% FDR korekce**

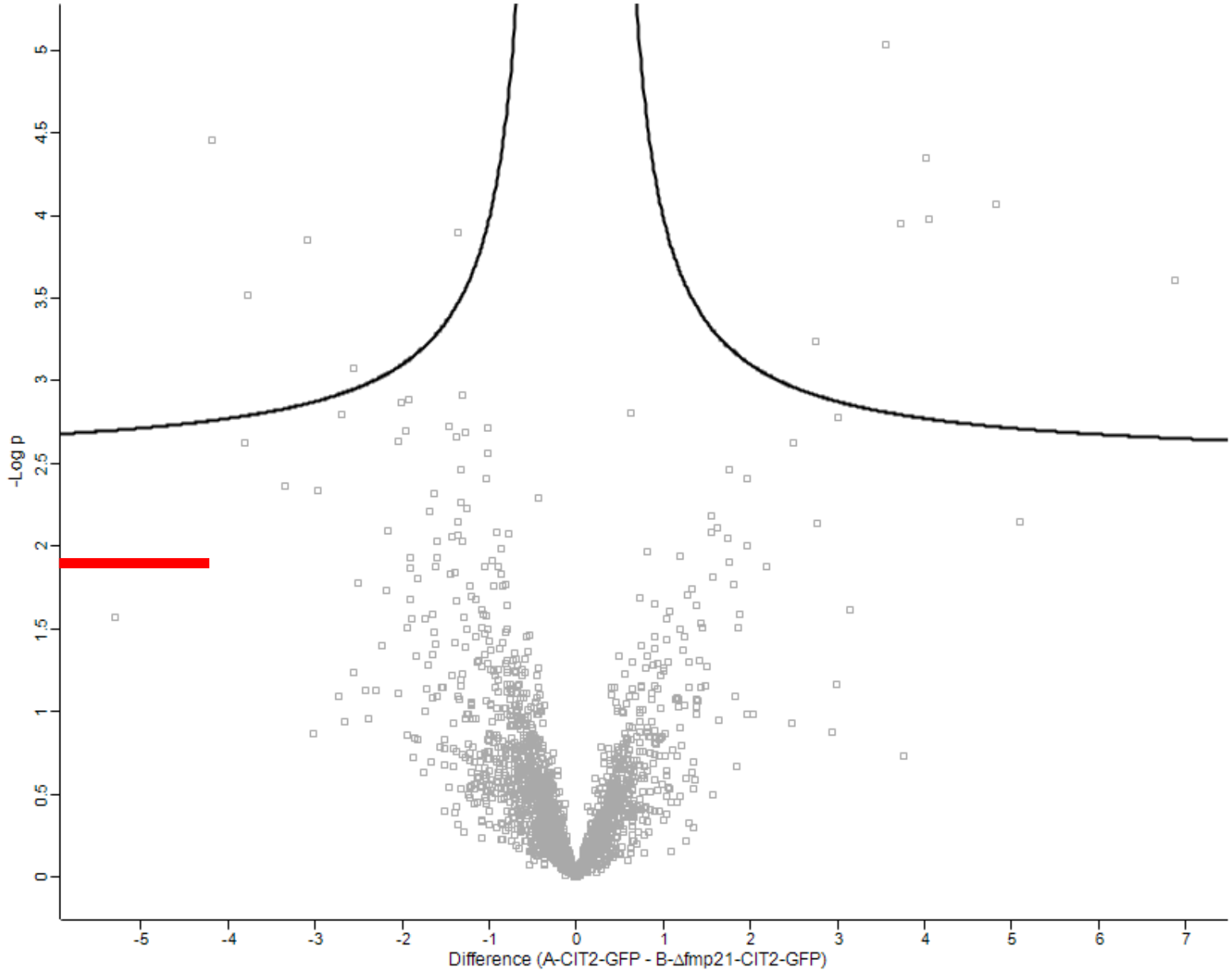
504 signifikantních

Chybějící hodnoty

- Část proteinů nemá kvantifikační hodnotu ve všech vzorcích
- Množství závisí na použité metodě měření
- Neexistuje ideální řešení
- Pro statistické zpracování je třeba hodnoty doplnit
- Doplnění vnese do dat šum
- Nejvhodnější je doplnit hodnoty na hranici citlivosti měření




-Log p
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
Difference (A-CIT2-GFP - B-Δfmp21-CIT2-GFP)


-Log p
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
Difference (A-CIT2-GFP - B-Δfmp21-CIT2-GFP)

Metody redukující dimenzionalitu dat

Lineární

unsupervised

PCA - Principal Component Analysis
T-SNE - t-distributed stochastic neighbourhood embedding (t-SNE)

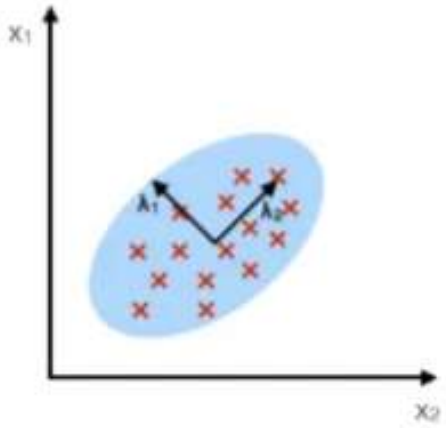
supervised

LDA - **Linear Discriminant analysis**
PLS -da - **Partial Least Squares discriminant analysis**
sPLS-da - **Sparse Partial Least Squares discriminant analysis**

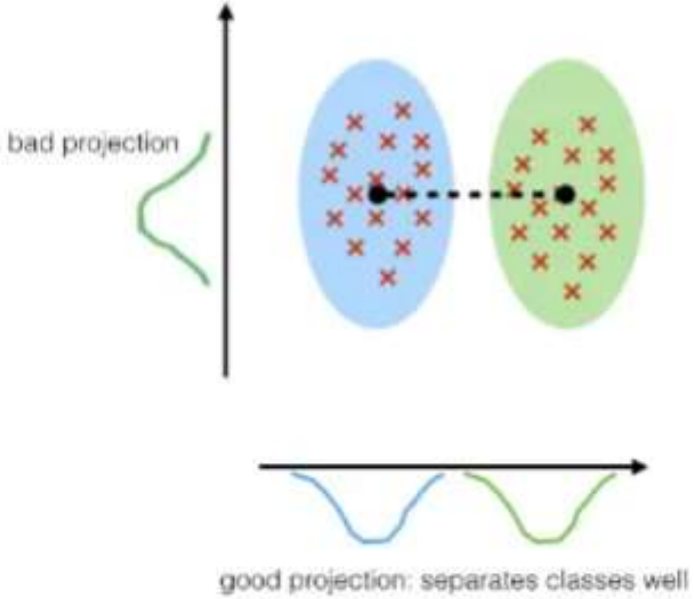
https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0

PCA x LDA

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



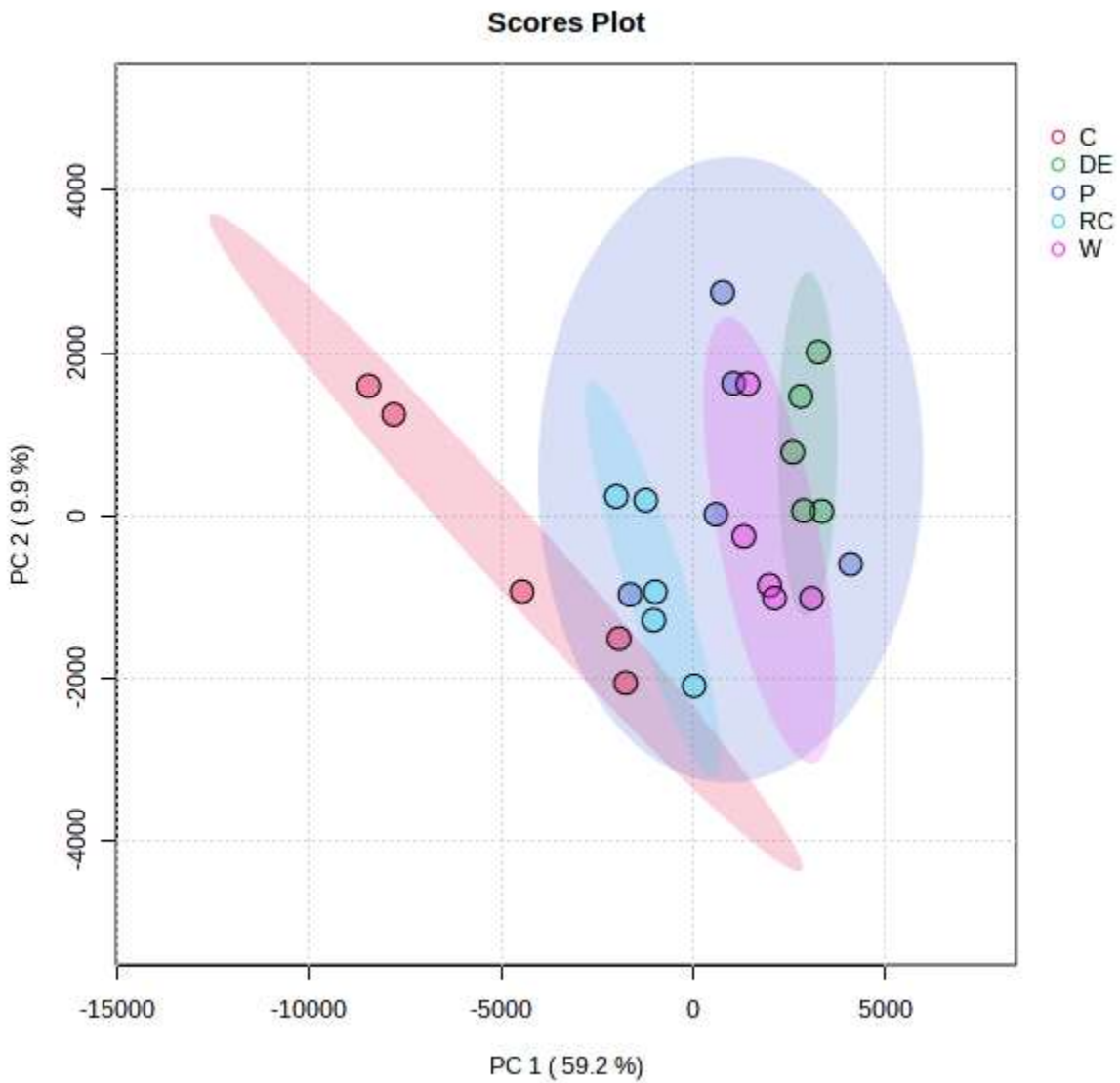
-Supervised metody - zadáváme které vzorky tvoří dohromady skupiny

-LDA zvýrazní to v čem se vzorky liší

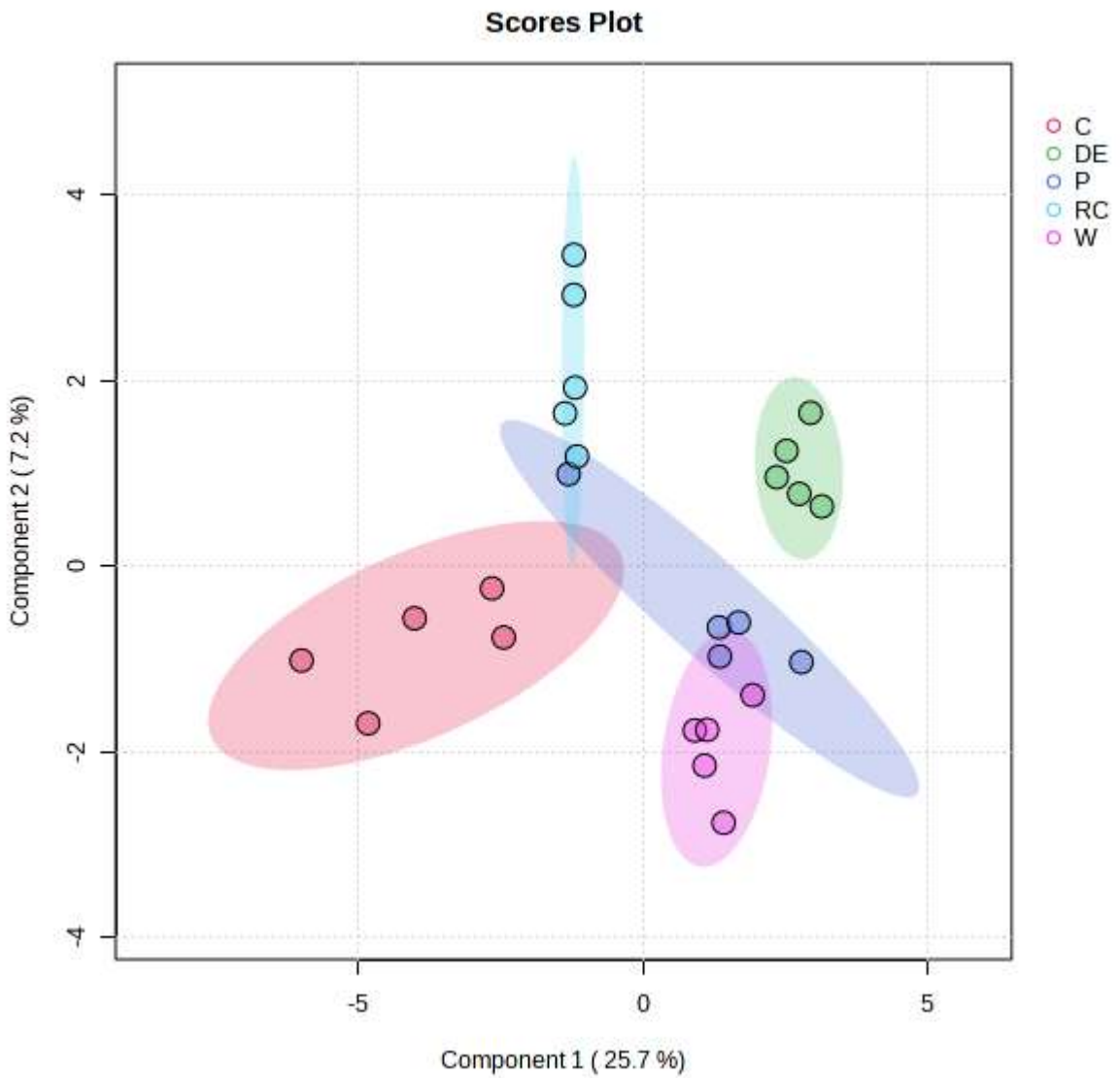
- Analýza identifikuje proteiny zodpovědné za separaci vzorků,
  - Vizualizace dat
  - Identifikace proteinů zodpovědných za separaci vzorků

Shodná data

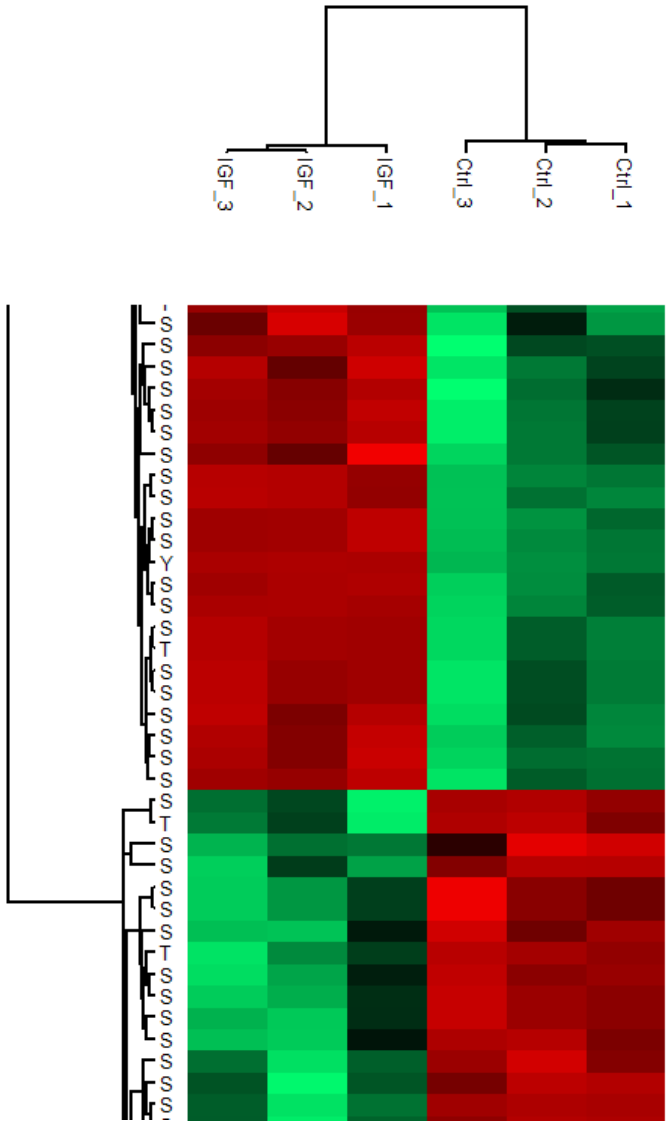
PCA



sPLS-DA

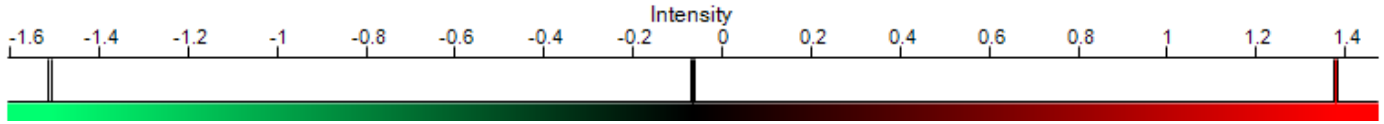


Heat Map



Metoda vizualizace

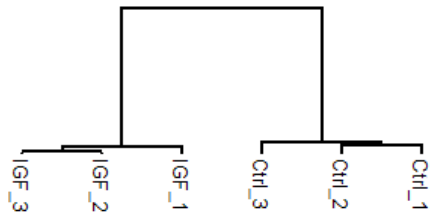
Hodnota má přiřazenou barvu dle škály

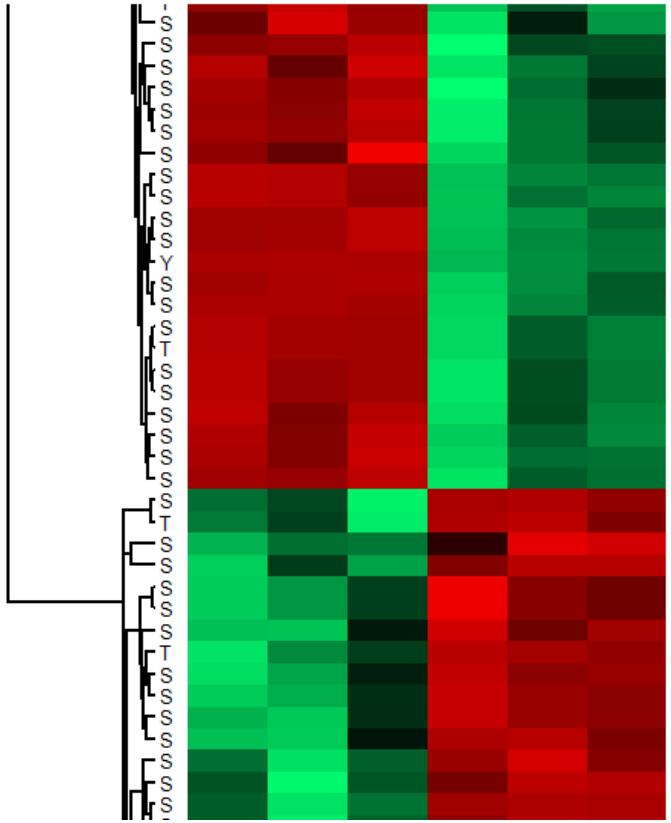


Řazení vzorků dle podobnosti

Často v kombinaci s **hierarchical clustering**

Hierarchical clustering





- Seskupování řádků a sloupců dle podobnosti
- Podobnost je vyjádřena jako vzdálenost
- Často Eukleidovská nebo Manhattan

Eukleidovská $\sqrt{(-1.8- -1.5)^2+(2-3)^2+(-1 - - 1.7)^2+(2.5-2)^2}$

Protein 1	-1.8	2	-1	2.5
Protein 2	-1.5	3	-1.7	2

Manhattan $(-1.8- -1.5)+(2-3)+(-1 - - 1.7)+(2.5-2)$

Funkční Anotace proteinů

Molecular function

Cellular component

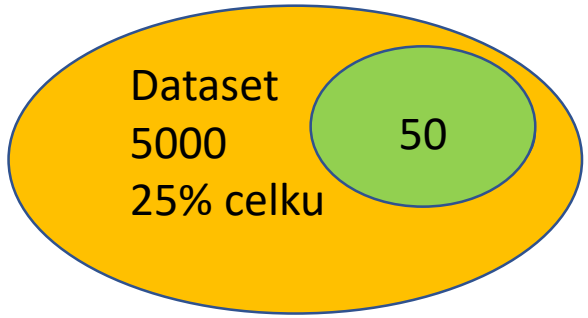
Biological process

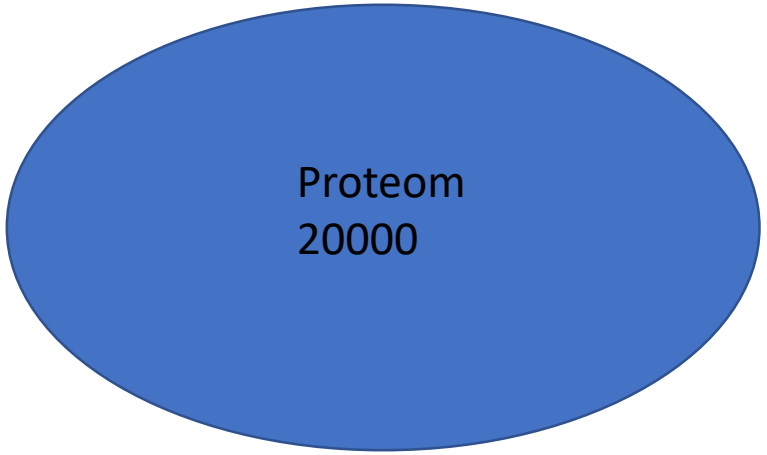
C: GOBP name	C: GOMF name	C: GOCC name	C: KEGG name
Category	Category	Category	Category
cellular macromolecule metabolic process;cellular...	adenyl nucleotide binding;adenyl ribo...	cell part;cytoplasmic part;intracellular membrane-bounded organelle;...	Zeatin biosynthesis
ameboidal cell migration;anatomical structure deve...	apolipoprotein binding;binding;cargo r...	cell body;cell part;cell surface;cytoplasmic part;cytoplasmic vesicle;e...	Wnt signaling pathway
anatomical structure formation involved in morpho...	binding;cation binding;ion binding;me...	cell part;cytoplasmic part;cytosol;intracellular membrane-bounded or...	Wnt signaling pathway
anatomical structure development;ATP-dependen...	adenyl nucleotide binding;adenyl ribo...	cell part;histone methyltransferase complex;intracellular membrane-b...	Wnt signaling pathway
biological regulation;biosynthetic process;canonica...	beta-catenin binding;binding;DNA bin...	cell part;cytoskeletal part;intracellular membrane-bounded organelle;...	Wnt signaling pathway
biological regulation;biosynthetic process;canonica...	beta-catenin binding;binding;DNA bin...	cell part;cytoskeletal part;intracellular membrane-bounded organelle;...	Wnt signaling pathway
cell projection organization;cellular component org...	binding;cation binding;ion binding;me...	apicolateral plasma membrane;cell part;cytoplasm;intracellular organ...	Wnt signaling pathway
aging;biological regulation;cardiac cell differentiati...	binding;identical protein binding;prote...	beta-catenin destruction complex;cell body;cell part;cell projection;cy...	Wnt signaling pathway
actin cytoskeleton organization;actin filament-base...	binding;identical protein binding;prote...	actin filament bundle;actomyosin;cell part;cytoplasm;cytoplasmic par...	Wnt signaling pathway

Zdroje anotací

- Rozsáhlé zdroje informací nad rámec FASTA databáze
- Modelové organismy dobře pokryté
- Mnoho různých databází
  - KEGG - **KEGG: Kyoto Encyclopedia of Genes and Genomes**
    - https://www.genome.jp/kegg/
  - https://geneontology.org/
- Někdy nejasná kvalita anotací

Gene Ontology (GO) term enrichment

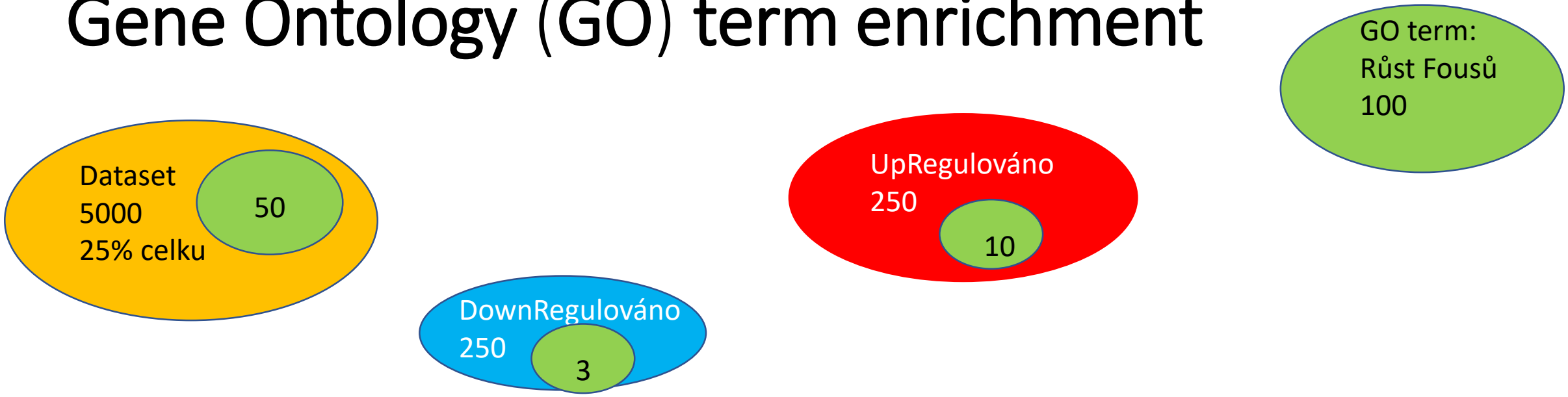




- Fischer exact test – statistický nástroj pro GO term **enrichment**

- GSEA – Gene Set Enrichment Analysis
  - https://www.gsea-msigdb.org/gsea/index.jsp
  - Software
  - Integrované anotační zdroje
  - Sofistikovanější metoda určení nabohacení

Gene Ontology (GO) term enrichment

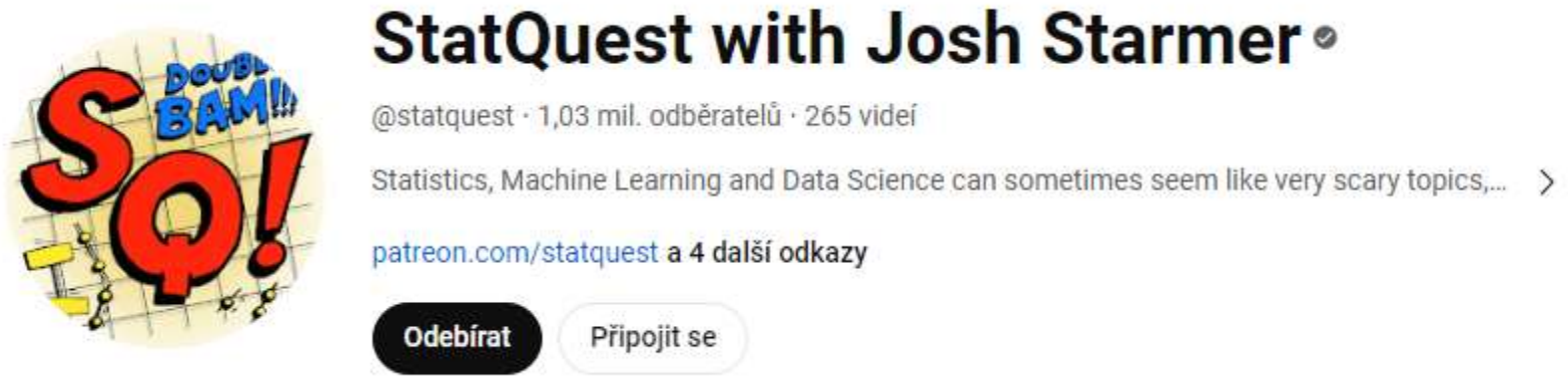


- Fischer exact test – statistický nástroj pro GO term **enrichment**

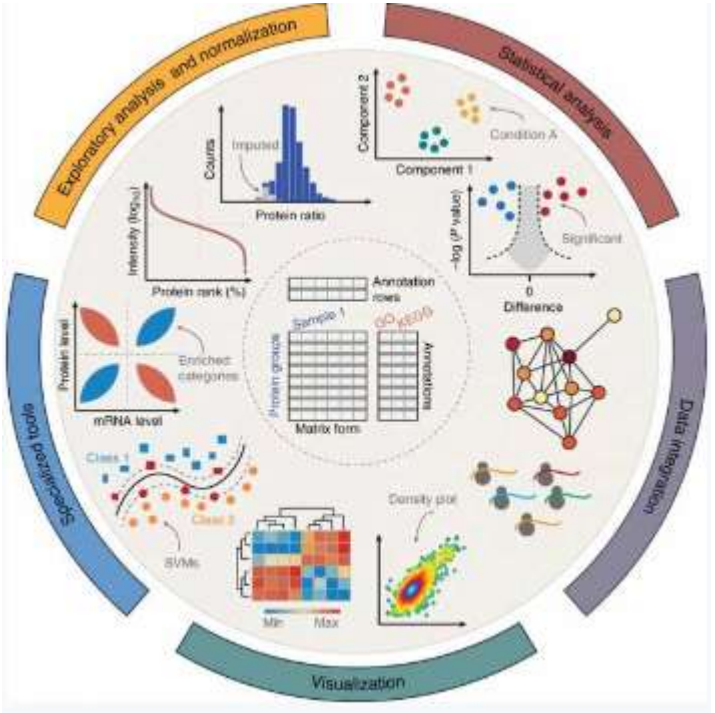
- GSEA – Gene Set Enrichment Analysis
  - https://www.gsea-msigdb.org/gsea/index.jsp
  - Software
  - Integrované anotační zdroje
  - Sofistikovanější metoda určení nabohacení

StatQuest with Josh Starmer

https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

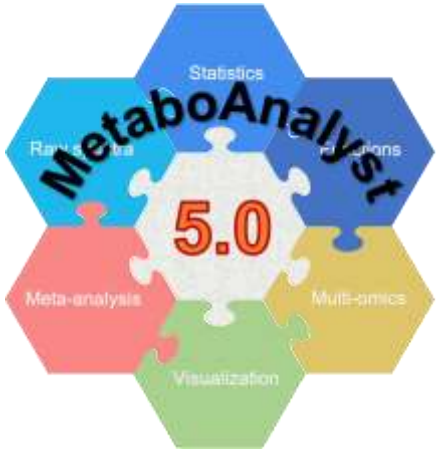


Nástroje



- Omezeně Excel
- Python
- R

- Perseus
  - zpřístupňuje v grafickém rozhraní vše co je potřeba pro analýzu
  - Není nutné skriptovat
  - https://maxquant.net/perseus/
  - Mnoho výukových videomateriálů
    - https://www.youtube.com/c/MaxQuantChannel
- MetaboAnalyst
  - webová aplikace primárně určená k analýze metabolomických dat.



https://www.metaboanalyst.ca

Single cell proteomics

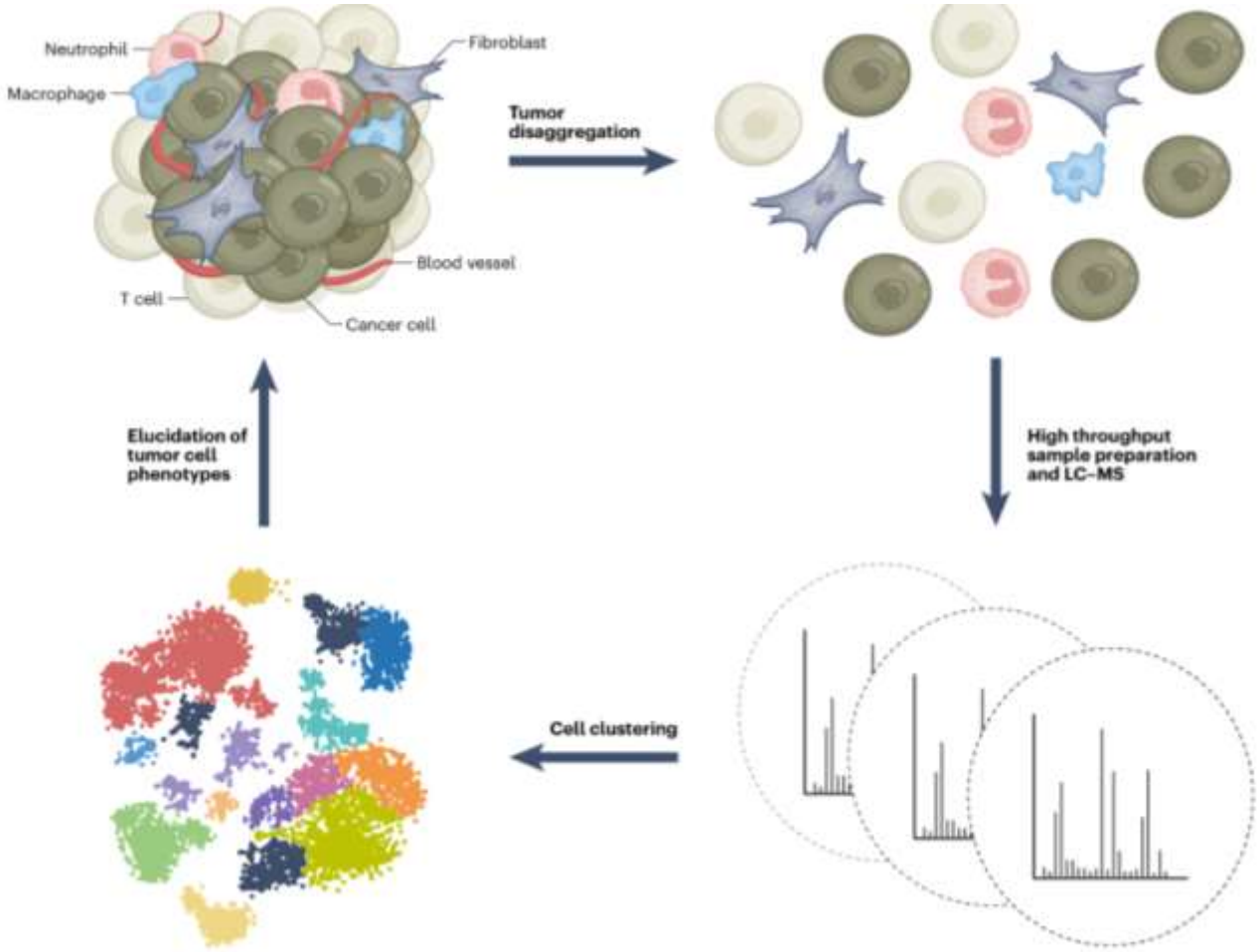
Proč single cell

Tkáň je tvořena mnoha druhy buněk

V buněčné kultuře existují subpopulace buněk

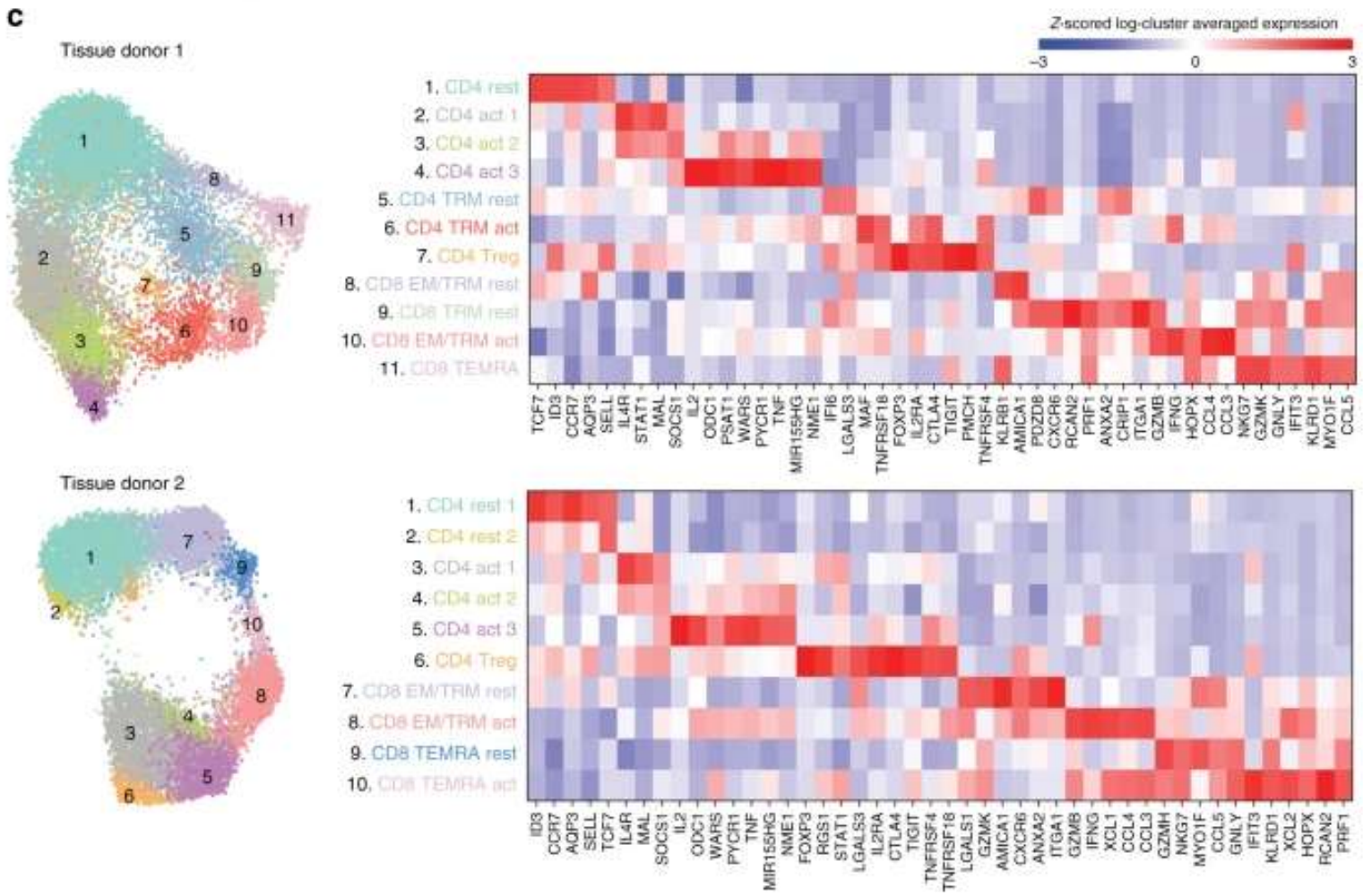
Dobře dokumentováno single cell RNA sekvenováním

Použití pro charakterizaci subpopulací buněk



2023,Geiger

Single cell RNA – příklad využití



c Identification of T cell subpopulations. UMAP embeddings colored by expression cluster along with heatmaps showing z-scored average expression of curated T cell subset marker genes that had a fold change >2 and $p < 0.05$ by the binomial test for at least one cluster.

Szabo, 2019

Limity a možnosti MS single cell

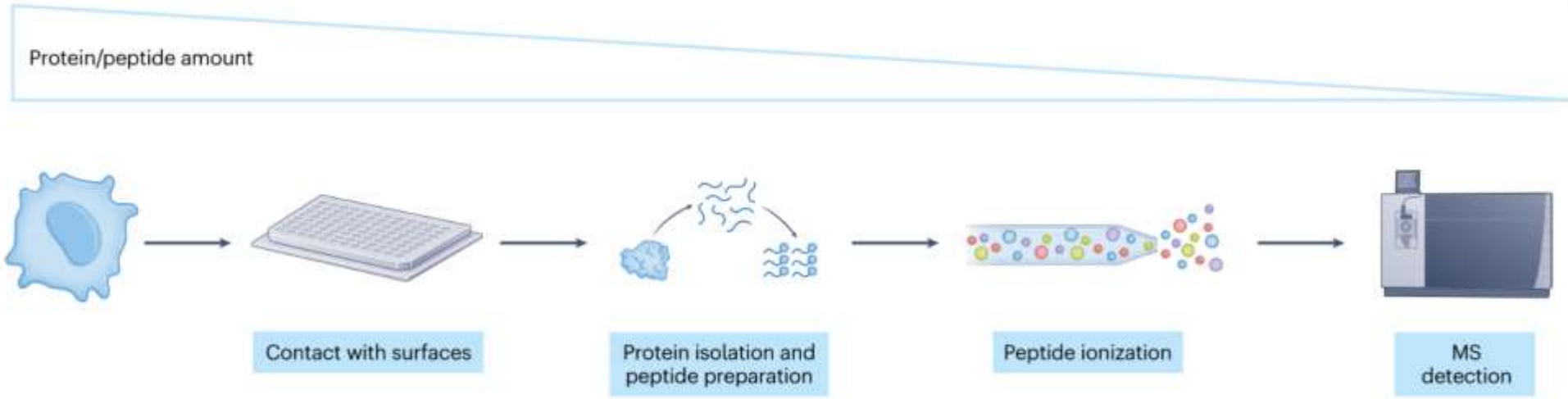
RNA sekvenování

- Amplifikační krok v průběhu analýzy
- Vysoký multiplexing
  - Desetitisíce buněk v rámci jedné analýzy
- Zavedená metoda, dostatek informací
- Komerčně dostupné systémy pro celý proces

MS based proteomika

- Absence amplifikačního kroku
  - Nízká Citlivost
- Žádný a nebo nízký multiplexing
  - Jedna buňka – jedna analýza
  - Maximálně desítky buněk denně

Darmanis 2023



500pg proteinu v jedné buňce

Většina je ztracena
při přípravě vzorku

X

Optimální nástřik 100ng-1ug

200 x více

Slavov 2018

Slavov Laboratory | Quantitative Biology

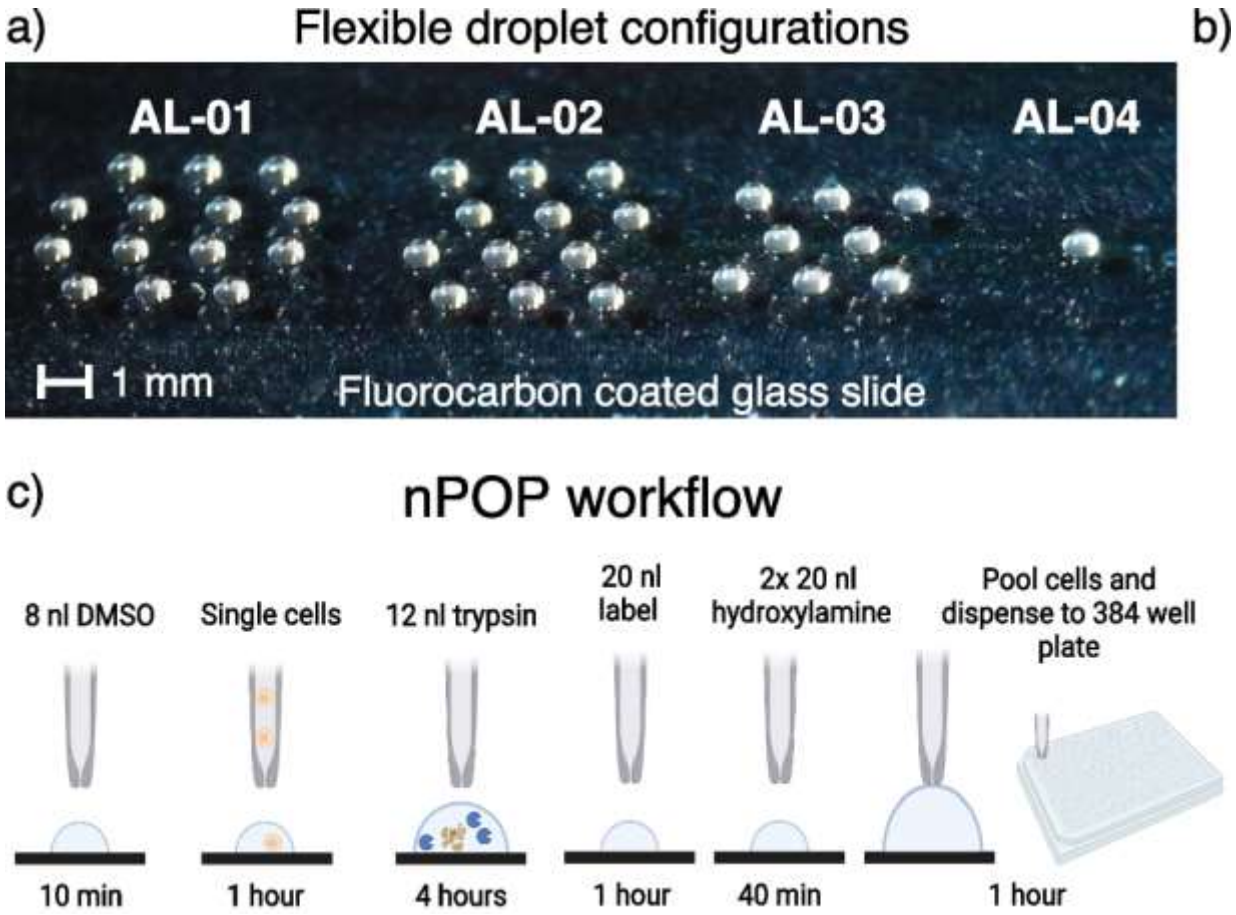
Seeking principles

in the coordination among protein synthesis, metabolism, cell growth and differentiation

https://slavovlab.net/

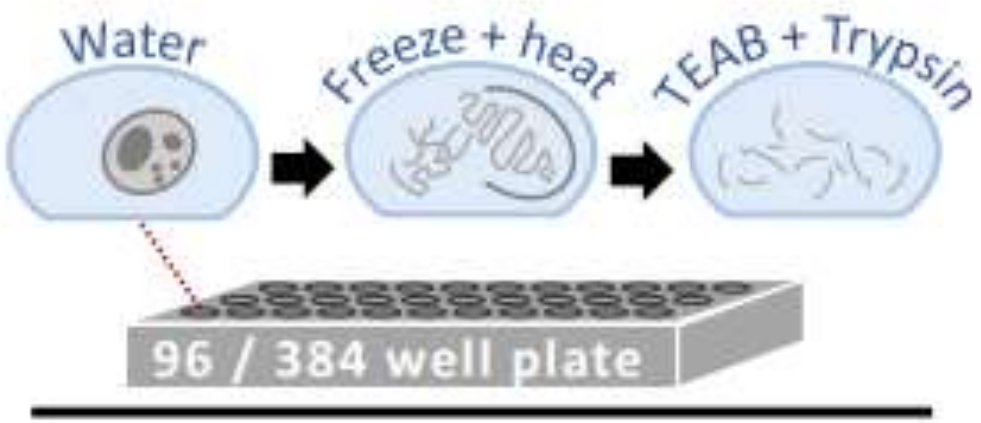
https://scp.slavovlab.net/ Protokoly a metody

Příprava vzorku - na sklíčku



Droplet sample preparation for single-cell proteomics applied to the cell cycle, Slavov , 2021

Příprava vzorku - v jamce



Automated sample preparation for high-throughput single-cell proteomics, Slavov 2018

Vynechat detergenty

Omezit objem

Omezit pipetování a přenášení vzorku

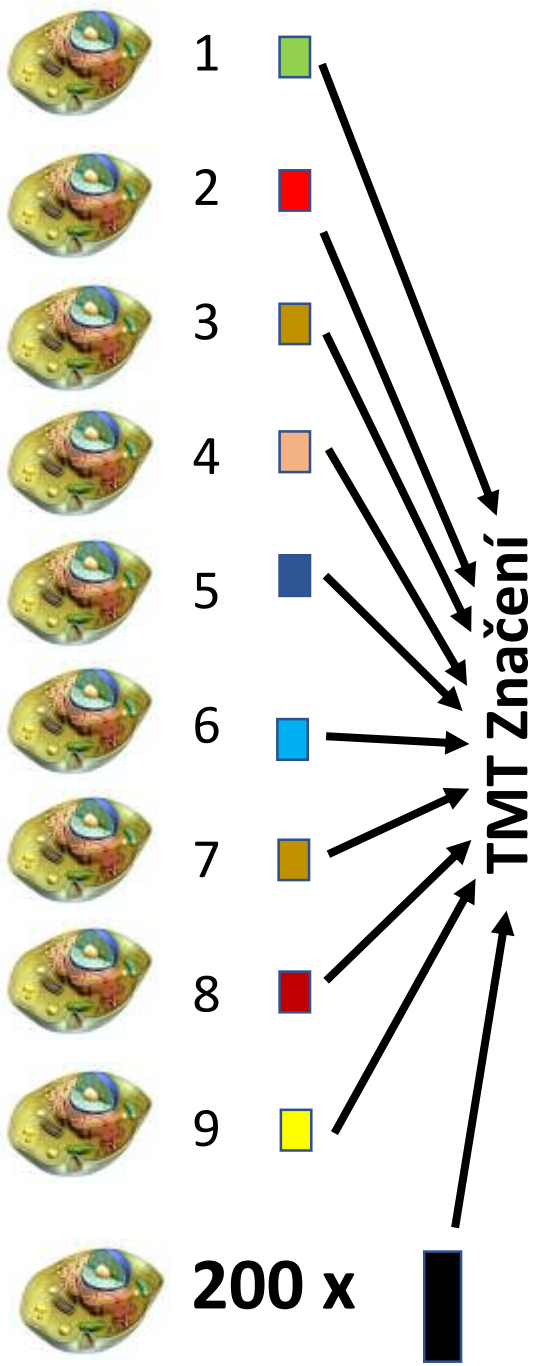
Jen přidávat ke vzorku chemikálie a pak vzorek rovnou nastříknout

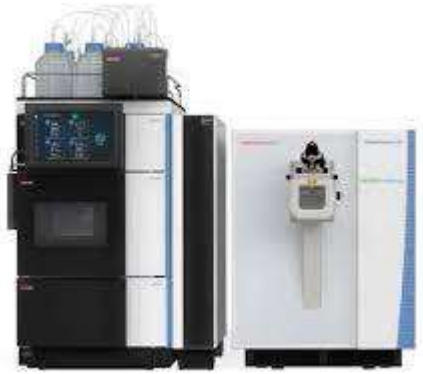
Citlivost a množství vzorků

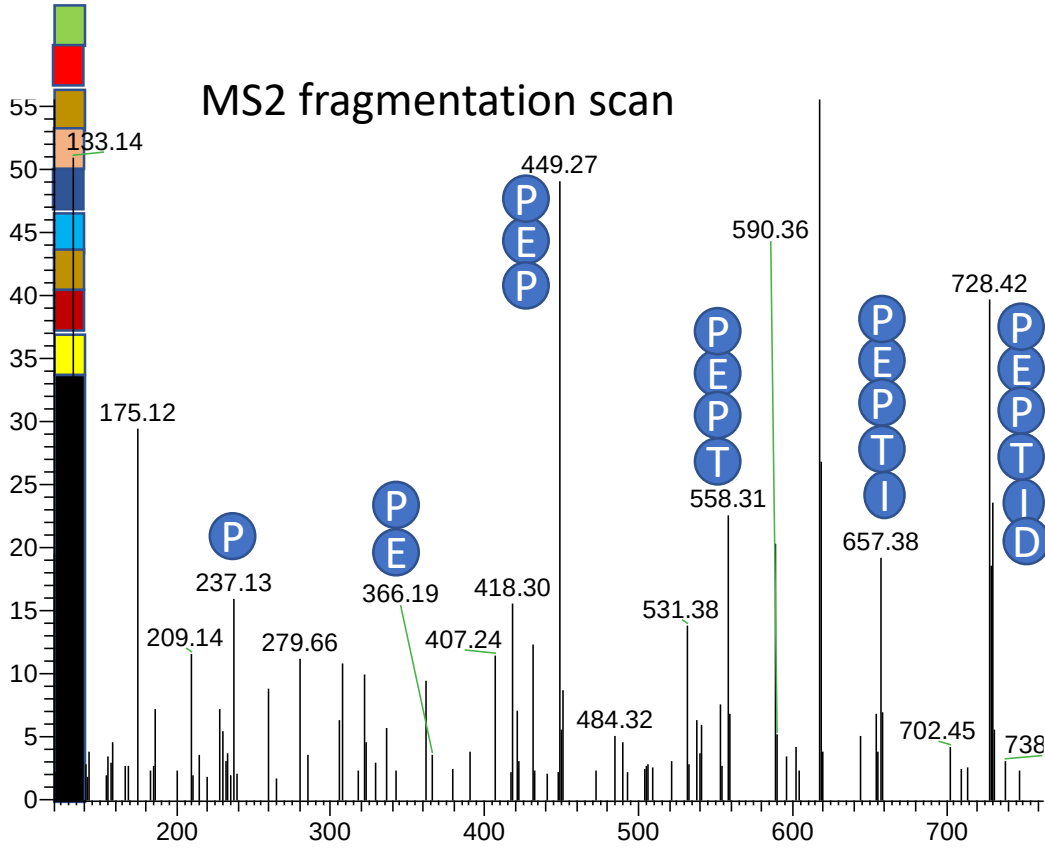
Technologické výzvy

řešení multiplexingem

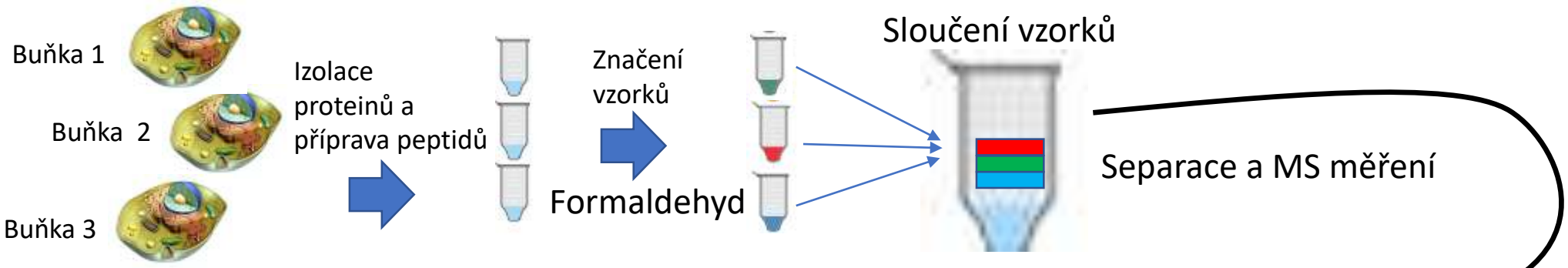
Isobarické značení , Carrier channel



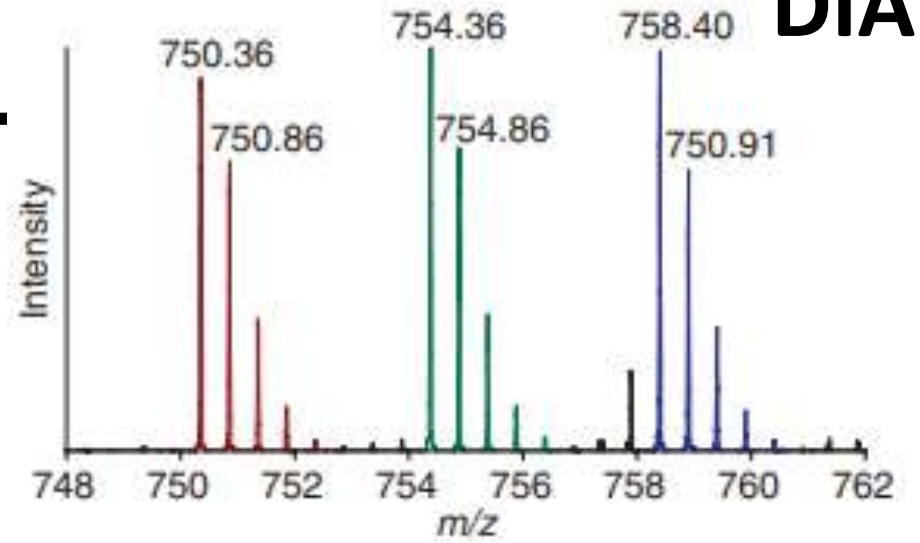




18 plex – 500 až 1000 buněk/den



DIA akvizice



Di-methylace

3 buňky na analýzu, 200 buněk/den

Metody redukující dimenzionalitu dat

Lineární

unsupervised

PCA - Principal Component Analysis

supervised

LDA - **Linear Discriminant analysis**
PLS -da - **Partial Least Squares discriminant analysis**
sPLS-da - **Sparse Partial Least Squares discriminant analysis**

Ne Lineární

unsupervised

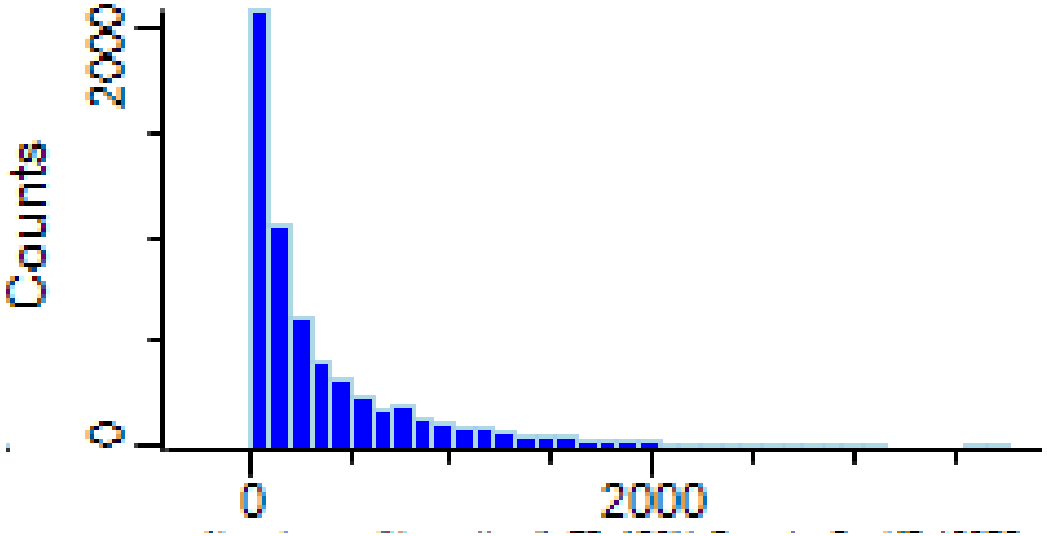
tSNE - **t-distributed stochastic neighbourhood embedding**
UMAP - **Uniform Manifold Approximation and Projection for Dimension Reduction**

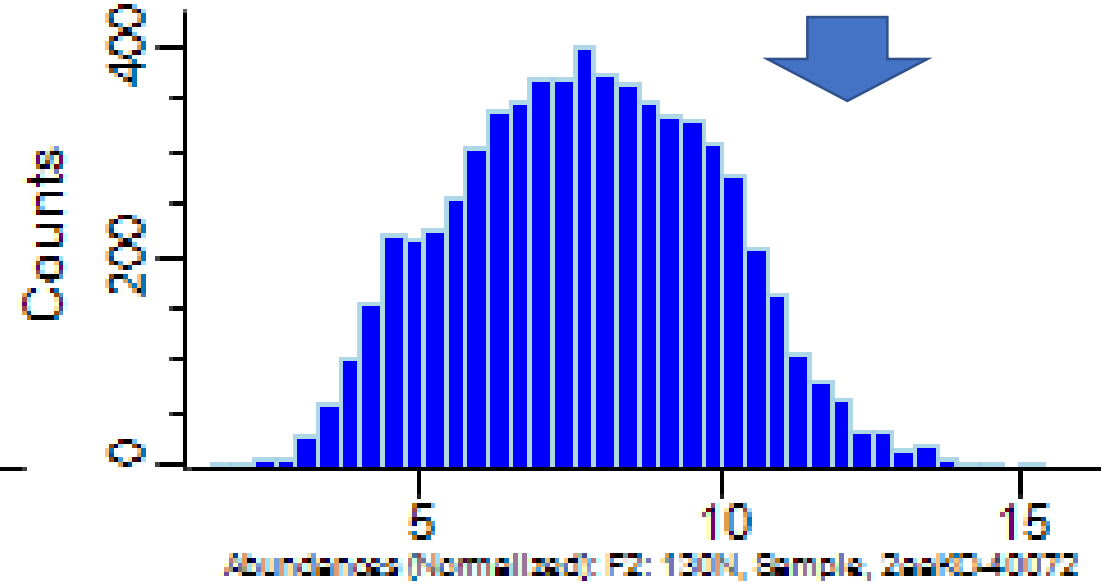
https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0

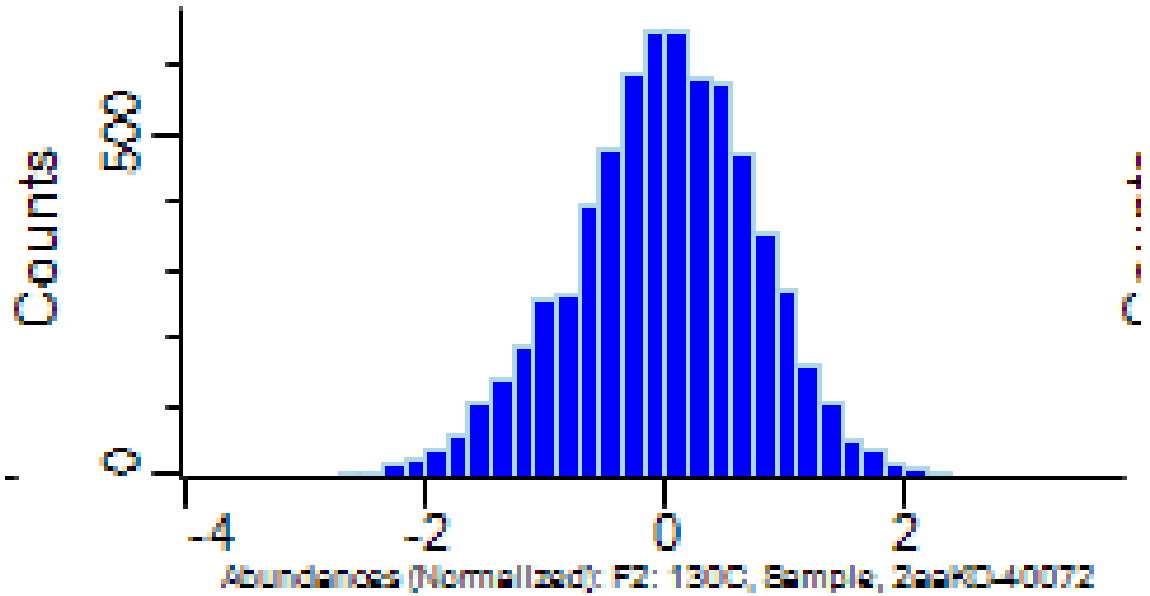
Škálování dat - Scaling

Několik možností
-Odečtení mediánu/průměru

-Převod na Z score
-odečtu průměr a zbytek vydělím směrodatnou odchylkou







Box Plot

String