Proteomická data – export excel

Výstupem proteomické analýzy je seznam identifikovaných proteinů s kvantitativní hodnotou pro každý vzorek datasetu. Data byla procesována v programu Perseus(1). Vzorky jsou následně setříděny do skupin, které musí v případě požadavku na statistické vyhodnocení obsahovat alespoň triplikát. Po procesování jsou hodnoty intenzit převedeny do binárního logaritmu. Zachovány byly pouze ty identifikace, které mají v případě triplikátu alespoň dvě platné hodnoty intenzity v alespoň jedné skupině. Následně byly spočítány T-testy pro jednotlivá z požadovaných porovnání. Již od začátku pracujeme s LFQ intenzitami - výsledek hledání programem MaxQuant(2) nebo DIA-NN v případě dat měřených pomocí DIA. Jedná se tedy o data, která byla normalizována už v průběhu vyhodnocení a není nutná další normalizace.

V tomto bodě byla vygenerována tabulka, která se Vám dostává do rukou. Excelový soubor obsahuje tolik listů, jaký byl počet porovnání a jeden navíc. První list obsahuje souhrnně všechna data.

# Hlavním orientačním pomocníkem by měly být barvy jednotlivých sloupců a buněk. S výhodou lze pak použít funkci řazení dat podle barev.

# Tabulky mají čtyři barevné pod části:

# Šedá se týká identifikačních údajů proteinu – Každý protein v tabulce má své unikátní ID. K němu se pojí jméno genu (gene name) a jméno proteinu. U některých hůře anotovaných taxonů mohou tyto informace chybět nebo být nekompletní. Tato část obsahuje rovněž skrytý sloupec obsahující všechny proteinové skupiny pojící se s danou identifikací ne pouze tu označenou jako Majority.

# Světle modrá část se týká kvantifikace a statistického vyhodnocení. Ta obsahuje další tři barevné podčásti:

# světle oranžová obsahuje hodnotu fold change, tedy reálnou změnu poměru proteinu. Jedná se o odlogaritmovaný rozdíl průměrů binárních logaritmů intenzit. Směr změny ukazuje znaménko a barva pozadí. Pokud je poměr větší než 2, bude hodnota tučná a se žlutým pozadím u kladných čísel a tučná s modrým pozadím u záporných čísel. Pokud byla hodnota *fold change* označena T-testem jako signifikantní, bude tučná a červená. V případě, že byl protein detekován pouze v jedné skupině, není stanovena hodnota fold change. Tyto identifikace představují potenciálně zajímavé on/off stavy. Světle zelený sloupec obsahuje záporný dekadický log p-hodnot z T-testu. Pokud je hodnota červená a tučná, znamená to, že byla označen jako signifikantní. Pro výpočet byl použitý *two-sample test* a hladina signifikance byla stanovena pomocí *permutation based FDR*. (<http://coxdocs.org/doku.php?id=perseus:user:activities:MatrixProcessing:Tests:TwoSampleTestProcessing> ).

# Další světle modré sloupce, následující po p-value, obsahují průměrnou intenzitu v dané skupině. Zde lze dohledat hodnoty intenzit pro proteiny bez *fold change*. Doporučujeme data seřadit pomocí funkce vlastní řazení se třemi úrovněmi: fold change, median 1 a median 2. Průměr je udán pouze pro proteiny s více než jednou platnou hodnotou ve skupině. Dále zde najdete počet platných hodnot v procentech (ve skrytém sloupci pak v absolutní hodnotě). Zeleným pozadím s tučnou červenou hodnotou jsou označeny ty proteiny, které měly intenzitu v binárním logaritmu větší než cca 23 (hodnota hraniční pro odlišení signálu od šumu – počítaná pro každý dataset, takže se může lehce lišit) a byly ve všech vzorcích dané skupiny. Modré pole s tučným cihlově červeným textem pak značí hodnotu větší než cca 23 a zastoupenou alespoň v 50 % vzorků dané skupiny. Jelikož je stanovení hranice intenzity na 23 pouze orientační, berte prosím toto barevné označení pouze jako pomůcku.

# Oddíl označený šedým pozadím obsahuje kvalitativní parametry dané identifikace. Kvalita identifikace je popsána dvěma spolu souvisejícími hodnotami. *Score* je pravděpodobnost toho, že shoda peptidové, a potažmo proteinové sekvence s experimentálně naměřeným spektrem, je nenáhodná. Toto číslo bude vždy stejné, bez ohledu na velikost databáze nebo počet prohledávaných spekter. *Q-value* popisuje lokální hodnotu FDR dané identifikace. Hodnota *Q-value* závisí na velikosti databáze a počtu prohledávaných spekter. Čím více spekter prohledáváte anebo čím větší je databáze, tím větší *score* je třeba pro projití 1% hranicí. Proteiny s nenulovou *Q-value* byly identifikovány s menší spolehlivostí.

# Posledním oddílem bez barvy jsou intenzity naměřené v jednotlivých procesovaných vzorcích. Tyto intenzity byly použity pro výpočty hodnot v oddíle kvantifikace a statistika. Pro jednotlivé skupiny jsou vždy rozděleny barevně (jedna varianta je podbarvena světle zeleně).

1. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M. and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*, **13**, 731-740.

2. Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N. and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics*, **13**, 2513-2526.