Základy bioinformatiky
Tutorial 7: Studium exprese genů

V tomto tutorálu si ukážeme možnosti jak získat zajímavé výsledky z veřejně dostupných databasí experimentů zaměřených na studium genové exprese, konkrétně na úrovni koncentrace mRNA. Míra genové exprese vyjádřená jako koncentrace mRNA nás můžeme zajímat ze dvou důvodů. Zaprvé, pokud se liší koncentrace určité mRNA například ve zdravých a nádorových buňkách, pak je možné její měření použít pro diagnostické účely. Vzhledem k tomu, že rozdíl koncentrace mRNA jediného genu obvykle není signifikantní, je možné měřit koncentrace mnoha (klidně všech) genů a výsledky náležitě statisticky zpracovat. První motivací pro měření koncentrace mRNA je tedy odlišit od sebe různé typy buněk, tkání, onemocnění atd. Zadruhé, je možné například změřit exprese všech genů dobře zavlažovaného ječmene a ječmene vystaveného suchu, zjistit koncentrace mRNA a ty porovnat. Pokud je nějaký gen více nebo méně exprimován za sucha, pak je možné se na něj zaměřit při šlechtění či genových manipulacích s cílem získat odolnější ječmen. Cílem druhého typu experimentů je tedy pochopit molekulární podstatu nemoci, vývoje tkáně, adaptace na stres a podobně.

Koncentrace mRNA jednotlivých genů je možné měřit pomocí Northern blotingu nebo pomocí kvantitativní polymerasové řetězové reakci s reversní transkripcí (Q-PCR nebo RT-PCR s tím, že RT může znamenat jak reversní transkripci, tak i real-time). Pro měření a porovnávání koncentrací mRNA spousty, nejlépe všech genů najednou, byly vyvinuty metody založené na DNA-čipech (DNA-microarray). V dnešní době se navíc přidávají metody paralelního sekvenování.

DNA-čip je destička, ne které jsou navázány poly-/ nebo oligonukleotidové sondy. Tyto sondy jsou na čipu navázány v přesně určených místech. Sondy mohou být na čip nanášený pomocí mikromanipulátorů. Další možností je synthetisovat sondy přímo na čipu pomocí fotolithografie. Jinou používanou možností je imobilisace sond na kuličky, které jsou pak umístěny do jamek na sklíčku.

V běžném microarray experimentu je nejprve z různých vzorků isolována mRNA. Ta je pak přeložena do cDNA a amplifikována. Zároveň může být v tomto kroku fluorescenčně označena. Tento vzorek je pak nanesen na DNA-čip, provede se inkubace a promývání a nakonec se změří intensita fluorescence v různých bodech na čipu. Koncentrace mRNA je možné měřit buď přímo, nebo jako srovnání tak, že je každý vzorek značen jinou florescenční barvičkou, vzorky jsou naneseny na jeden čip a intensity fluorescence jsou měřeny při dvou vlnových délkách. Kromě komplementárních sond obsahují DNA-čipy často také „mismatch“ sondy pro korekci na obsah GC vs. AT a různé kontrolní sondy.

Zpracování dat z DNA-čipů je poměrně náročné a nepatří do základního kurzu bioinformatiky. Případné zájemce bych odkázal na statistický program *R* a jeho molekulárně-biologickou knihovnu *BioConductor*. V rámci prvotního zpracování je potřeba:

1. z naskenovaného obrázku chipu zjistit intensity v jednotlivých spotech,
2. pokud je jeden gen representován více sondami, je nutné z nich vypočítat koncentraci celé mRNA,
3. data je potřeba přeškálovat tak, aby bylo možné porovnávat více čipů mezi sebou
4. je potřeba statisticky vyhodnotit které sondy jsou více a které méně exprimované ve zdravých a nemocný, stresovaných a kontrolních buňkách atd.
5. výsledky je možné konfrontovat s metabolickými a signalizačními dráhami, vytvořit hypotesy atd.

Pokud by byla zjištěna exprese genů například jen pro jeden vzorek ječmene za sucha a jeden zavlažovaný, pak by nebylo možné provést statistické porovnání. Proto jsou obvykle pro každý druh vzorku prováděna tři nebo více biologických opakování.

Výsledky z DNA-čipů a jiných technik studia genové exprese naleznete v databasi *Gene Expression Omnibus* (GEO, http://www.ncbi.nlm.nih.gov/geo/). Na této stránce se pokusíme nalézt

informace o vlivu kokainu na expresi genů v mozku. Proto do řádku pro zadání dotazu zadejte název této přírodní látky. Jednou z mnoha studií, kterou vám GEO najde, je studie s označením **GDS1608**. Zkratka GDS představuje GEO dataset. Jedná se o výsledky jedné nebo více studií, které byly dány dohromady kurátory GEO. Kliknutím na odkaz se vám zobrazí stránka, kde se dozvíte, že se jedná o studii vlivu kokainu na části mozku myší, které jsou spojeny s pocitem odměny. Rovněž zde najdete odkaz na článek a informaci o čipu, který byl použit. Čipy a postupy, které byly použity, jsou katalogizovány jako GEO platforms (GPL). Můžete zde kliknout na vlastní sérii tak jak byla uložena do database autory článku, tedy GEO series (GSE). V této položce je možné se dostat až k hrubým datům (GEO samples, GSM).

Na tomto místě si ukážeme možnost jak zjistit míru exprese vybraného genu, například glyceraldehyd-3-fosfátdehydrogenasy, kterou najdete pod zkratkou **gapdh**. Klikněte na *find genes* a do políčka *Find gene name or symbol* zadejte tuto zkratky a zmáčkněte *Go*. Zajímavější je možnost nalézt geny, které jsou rozdílně exprimované v různých vzorcích. Klikněte na *Compare 2 sets of samples*. Stránka nám nabízí různé statistické testy. My vybereme oboustranný dvouvýběrový Studentův t-test. Jako hladinu pravděpodobnosti zvolíme 0,01. V kroku 2 máme vybrat dva výběry. Když kliknete na odkaz, objeví se vám nové okno, kde si můžete navolit vzorky. Pokud bychom chtěli porovnávat dvě mozkové tkáně neovlivněné kokainem, pak bychom mohli zadat tři vzorky „*control*“ pro jednu tkáň a tři vzorky „*control*“ pro druhou. Pokud bychom chtěli studovat vliv kokainu na například *nucleus accumbens*, pak bychom vybrali tři vzorky z této tkáně ovlivněné kokainem a tři kontrolní. Zmáčknutím *Query Group A vs. B* se vám objeví seznam diferenciálně exprimovaných genů na dané úrovni pravděpodobnosti.

Další možností jak analysovat genovou experesi je heatmap. Jedná se o graf, na kterém je každý gen zobrazen jako barevný čtvereček na obdélníkové mozaice. Každý řádek representuje jeden gen, každý sloupec pak representuje jeden vzorek. Fialová barva značí, že je gen nadprůměrně exprimován v daném vzorku v porovnání s ostatními vzorky. Zelená značí opak. Jak geny tak vzorky jsou analysovány hierarchickým shlukováním a jejich stromy jsou nad a vedle vlastní mozaiky. S hetamap je možné si různě hrát a zoomovat ji.

Poslední položkou je *Experiment design and value distribution*. Ta slouží k porovnání kvality dat a k získání informací o designu experimentu.