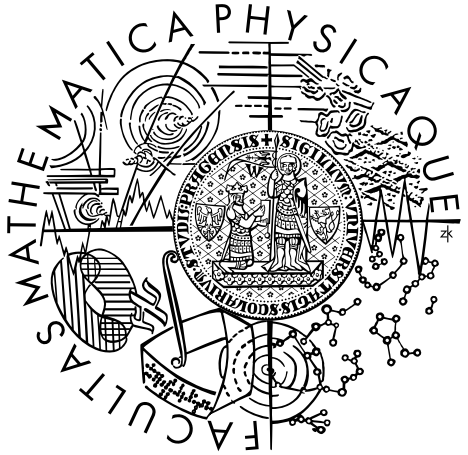
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Vojtěch Tuma

Sumarizace genových expresních čipů z volně žijících druhů

Katedra teoretické informatiky a matematické logiky

Vedoucí diplomové práce: Mgr. Libor Mořkovský
Studijní program: Informatika
Studijní obor: ITI (1801T010)

Praha 2015

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Sumarizace genových expresních čipů z volně žijících druhů

Autor: Vojtěch Tuma

Katedra: Katedra teoretické informatiky a matematické logiky

Vedoucí diplomové práce: Mgr. Libor Mořkovský, Katedra zoologie, Přírodovědecká fakulta UK

Abstrakt: Pro zkoumání exprese exonů a genů v organismech se používají genové expresní čipy. Genové expresní čipy jsou vytvořeny podle genomů laboratorních kmenů modelových organismů. Pro zpracování naměřených dat se používají sumarizační algoritmy, nejčastěji gcRMA, PLIER nebo IterPLIER. Při použití expresních čipů pro zkoumání volně žijících druhů jsou naměřené hodnoty ovlivněny rozdílností genomů zkoumaných a modelových organismů. Navrhujeme zlepšení výsledků vyřazením částí genomu ovlivněných známými rozdíly mezi druhy ze sumarizace. Odstranění ovlivněných částí může zlepšit sumarizaci, především na exonové úrovni.

Klíčová slova: Expresní čip gcRMA PLIER Sumarizace Affymetrix

Title: Summarization of gene expression arrays from free living species

Author: Vojtěch Tuma

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Mgr. Libor Mořkovský, Department of Zoology, Faculty of Science, Charles University

Abstract: Gene expression arrays are used to assess expression of exons and genes of organisms. The design of expression arrays is based on a genome of laboratory strains of model organisms. The most frequent summarization algorithms used to process data from measurements are gcRMA, PLER and IterPLIER. When using expression arrays to research free living species, the measured values are influenced by differences in genomes of free living and model organisms. We propose a method to improve the results by removing parts of genomes influenced by known differences between species from the summarization. Removing influenced parts can improve summarization, especially on exon level.

Keywords: Expression array gcRMA PLIER Summarization Affymetrix

Děkuji vedoucímu práce, Mgr. Liboru Mořkovskému, za veškerou podporu během práce a neocenitelné rady při řešení problémů. Děkuji konzultantům RNDr. Františku Mrázovi, CSc. a RNDr. Radce Reifové, Ph.D. za konzultace a řešení formální stránky práce.

Výpočetní zdroje byly poskytnuty uskupeními CESNET LM2015042 a CERIT Scientific Cloud LM2015085, pod záštitou programu „Projects of Large Research, Development, and Innovations Infrastructures“.

Děkuji svým rodičům za trpělivost.

Obsah

Úvod

Bioinformatika je poměrně mladá vědecká disciplína spojující tak vzdálené obory, jako jsou biologie a informatika. Bioinformatika se velmi rychle vyvíjí a přináší stále nové poznatky především z oblasti molekulární biologie a genetiky. S využitím nejmodernějších technologií umožňuje zkoumání podobností a odlišností různých organismů.

Na poli genetiky a genomiky jsou to právě bioinformatické postupy a stále rostoucí výpočetní síla počítačů, které umožňují sekvenování a anotaci genomu, t.j. čtení a popis genetického materiálu daného organismu, a zkoumání jeho změn. Znalost samotné sekvence DNA jedince ale ještě nestačí na popis fungování organismu. Z celého genomu jedince se v každé buňce využívají různé geny v různém množství v procesu označovaném jako *exprese*.

Jedním z nástrojů k určování, které geny jsou v buňkách daného organismu v danou dobu aktivní, jsou tzv. expresní čipy. Expresní čipy jsou navrhovány podle konkrétního modelového organismu, pro jehož zkoumání jsou určeny. Jedním z nejrozšířenějších modelových organismů, pro který jsou expresní čipy dostupné, je laboratorní kmen myši domácí (*Mus musculus*).

Motivace

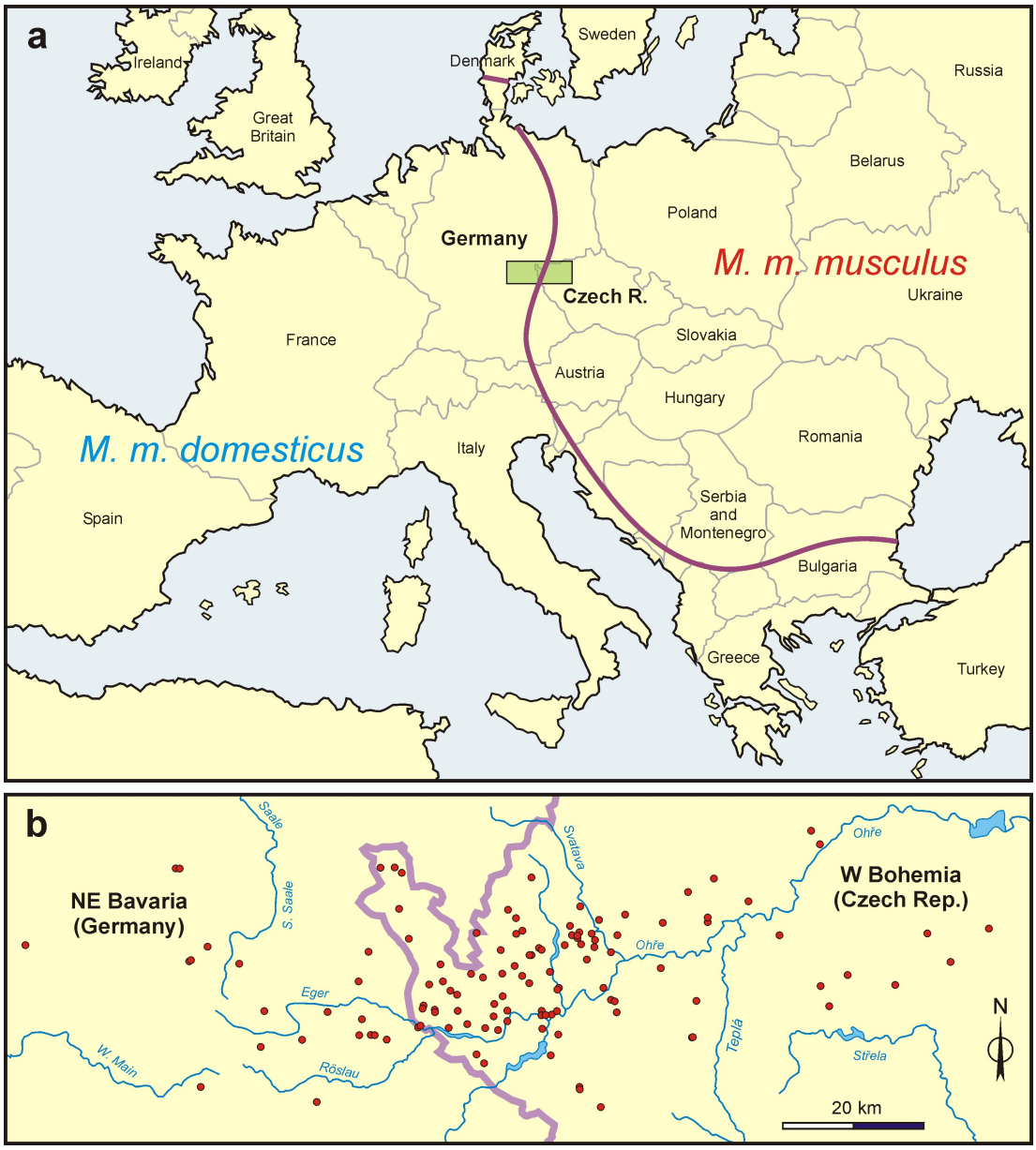
Ve střední Evropě žijí dva poddruhy myši domácí, *Mus musculus domesticus* a *Mus musculus musculus*, které se potkávají v tzv. hybridní zóně. Hybridní zóna je vhodným místem pro studium speciace, tedy evolučního procesu oddělování druhů. Jedním z projevů speciace může být rozdílná exprese genů. Pro poměření genové exprese lze použít expresní čipy.

V hybridní zóně byl prováděn sběr vzorků tkání divoce žijících myší [1]. Mapa oblasti, kde probíhal výzkum je na obrázku 1. Vzorky byly použity v experimentu na expresních čipech Affymetrix Mouse ST Exon Array, navržených pro zkoumání exprese u myší.

Avšak volně žijící poddruhy myší se mírně liší v sekvenci DNA od laboratorního kmene, podle kterého byly expresní čipy vytvořeny. To by mohlo mít vliv na měření na čipech a následné zpracování výsledků – tzv. *sumarizaci*.

Cíl práce

Odlišnost volně žijících kmenů a designového kmene může zanést do výsledků sumarizace zkreslení. Cílem této práce je ověřit, zda jsou nejčastěji používané metody sumarizace expresních čipů – (gc)RMA, PLIER a IterPLIER – odolné vůči odlišnosti kmenů, případně navrhnout postup, jak výsledky zlepšit. Biologicky zajímavým vedlejším produktem jsou konkrétní sumarizace z experimentu. Cílem této práce není přímé porovnání výsledků jednotlivých sumarizačních metod navzájem v rámci daného biologického experimentu.



Obrázek 1: *(a) Hybridní zóna M. m. musculus a M. m. domesticus se nachází v okolí hranice rozšíření těchto dvou druhů (označené fialovou barvou). Zvýrazněná je oblast výzkumu. (b) Detail oblasti s vyznačenými lokacemi sběru vzorků. Zdroj: [1]*

Struktura práce

Platforma expresních čipů je popsána v kapitole 1. Sumarizační metody použité ke zpracování experimentálně získaných dat jsou popsány v kapitole 2. Kapitola 3 obsahuje informace o poddruzích zkoumaných myší. Návrh vylepšení sumarizace pro poddruhy odlišné od designového kmene je v kapitole 4. Zpracování dat z expresních čipů je popsáno v kapitole 5. Analýza získaných statistik je v kapitole 6. V příloze A jsou stručně popsány skripty použité pro zpracování a analýzu dat. V příloze B jsou shrnuty formáty souborů spojených se sumarizací a analýzou.

1. Expresní čipy

Všechny buňky jednoho organismu mají stejný řetězec DNA. Asi nejznámější genetickou analýzou je sekvenování DNA. Slouží k přečtení řetězce, ze kterého lze určit, co může být na základě DNA vytvořeno buněčnými pochody. Expresní analýza slouží k určení částí DNA, které se v buňce aktuálně používají a v jakém množství.

Použití expresních čipů v dnešní době ustupuje, ale velké množství dat z expresních čipů je dostupné v různých databázích, pro nemodelové organismy mohou být jediným zdrojem informací [2]. V současnosti se na zachycení RNA profilu používá přímé sekvenování (RNA-Seq).

1.1 Exon, transkript, gen

Geny (úseky DNA kódující proteiny) jsou v buňce převáděny na proteiny procesem zvaným *genová exprese*. Geny jsou přepisovány do řetězců RNA (*transkripce*), poté proběhne *splicing* (*stříhání*). Vystřižené části genu se nazývají *introny*, části, které projdou splicingem, jsou označovány *exony*. Exprimované exony jednoho genu jsou spojeny do *transkriptu*, nazývaného *messenger RNA* (mRNA). Podle řetězce mRNA jsou na ribozomech vytvářeny proteiny procesem nazývaným *translace*.

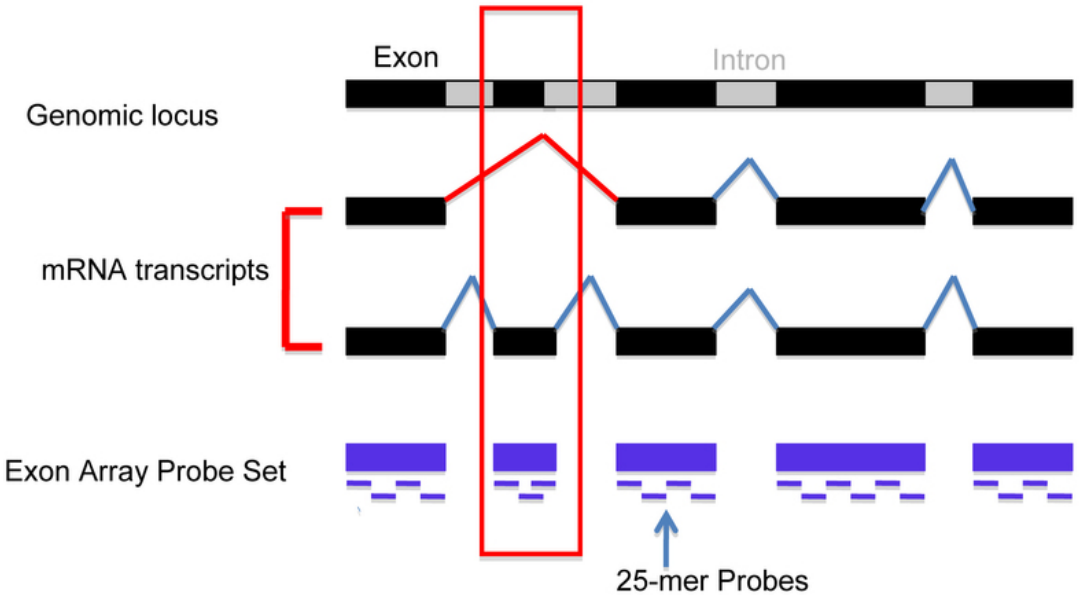
Procesy transkripce a stříhání sekvence mohou probíhat různě. Jeden úsek DNA může kódovat různé proteiny, zároveň mohou být v buňkách různých tkání exprimovány rozdílné části DNA. Rozdíl v expresi může být i v množství mRNA takto vytvořené, tedy v množství produkovaných proteinů. Znázornění alternativního sestřihu je na obrázku 1.1.

Účelem expresních čipů je zachycení aktuální „chemické situace“ v buňce – změření koncentrace proteinů, které jsou aktuálně v buňce vytvářeny. Expresní čipy k měření používají mRNA, mezistupeň expresního procesu. Rozdíly v expresi je možné zachytit na čipech a porovnat je s jinými tkáněmi či organismy.

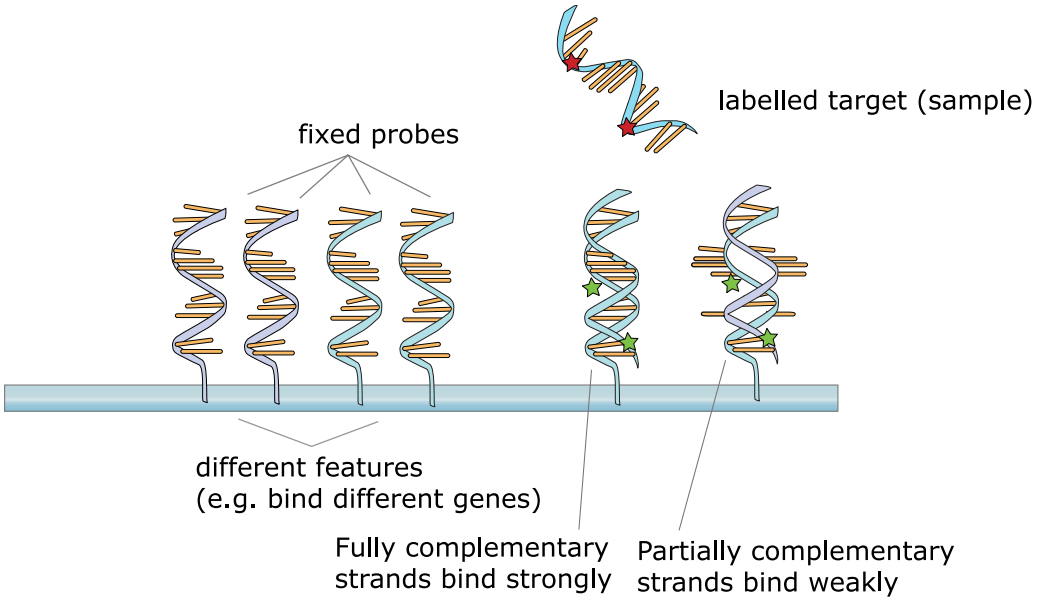
1.2 Design čipů

Expresní čip sestává z prostorově oddělených políček obsahujících *próby*. Próby jsou tvořeny jednovláknovým řetězcem DNA o délce 25 bází, řetězce jsou jedním koncem pevně přichyceny k podkladové destičce. Próby na čipu (ozn. *feature*) jsou komplementární k řetězcům *znaků* (*target*), jejichž obsah ve vzorku je záměrem pozorování. Schéma 1.2 přibližuje použití prób na čipu. Každá próba je v dokumentaci čipu popsána souřadnicemi umístění na čipu, sekvencí, pozicí v genomu a referencí na exon a transkript, který reprezentuje.

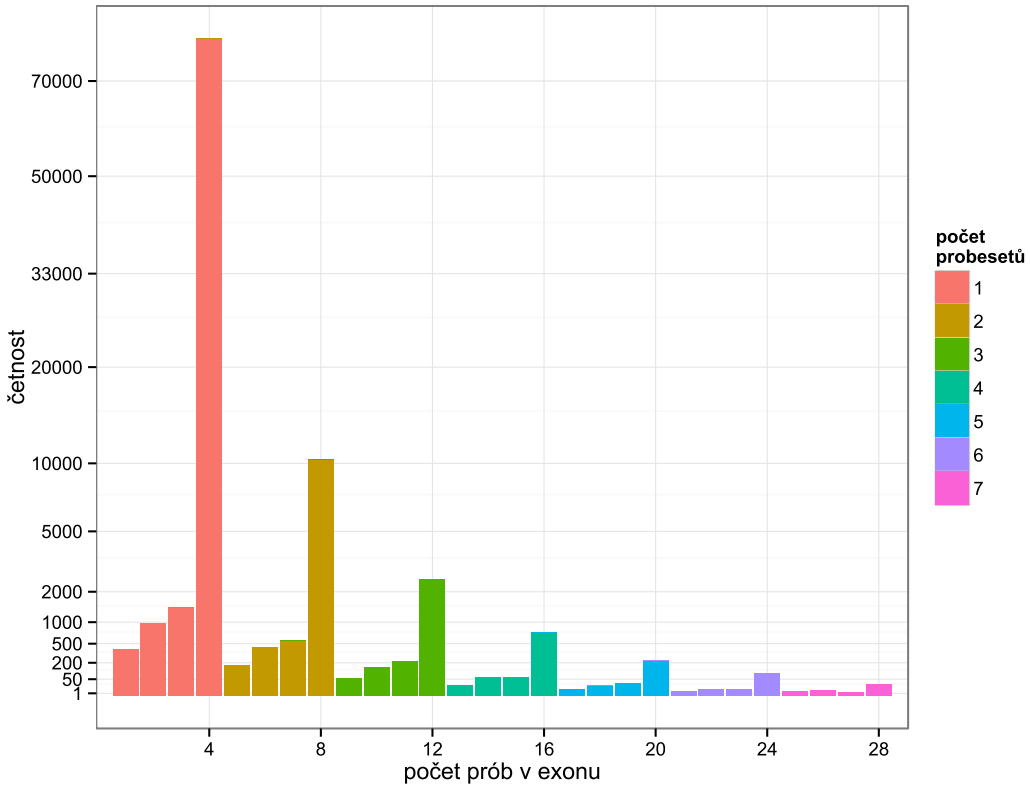
V terminologii expresních čipů je *transkriptová skupina* (transcript cluster) sdružení exonových oblastí odpovídajících známému nebo domělému genu. *Exonová skupina* (exon cluster) je skupina jedné nebo více sad prób pokrývajících souvislou část genové sekvence. Exonová skupina je rozdělena do více oblastí pro výběr prób (probe selection region, PSR), pokud transkriptové podklady naznačují hranici splicingu nebo polyadenylaci [5][6].



Obrázek 1.1: *Znázornění alternativního sestřihu genů. Při splicingu jsou vyřazeny introny, z exonů je vytvořen řetězec mRNA pro syntézu proteinů. Různý splicing daného genu (vyznačen rámečkem) může vytvářet různé transkripty, což se odrazí v naměřených signálech na expresních čipech. V posledním řádku jsou naznačeny sady prób reprezentující exony. Zdroj:[3]*



Obrázek 1.2: *Schéma prób na čipu. Řetězce přichycené na podkladu (feature) jsou komplementární k řetězcům ze vzorku (target), jejichž obsah ve vzorku zkoumáme. Próby se mohou navázat na částečně komplementární řetězce slabší vazbou. Zdroj:[4]*



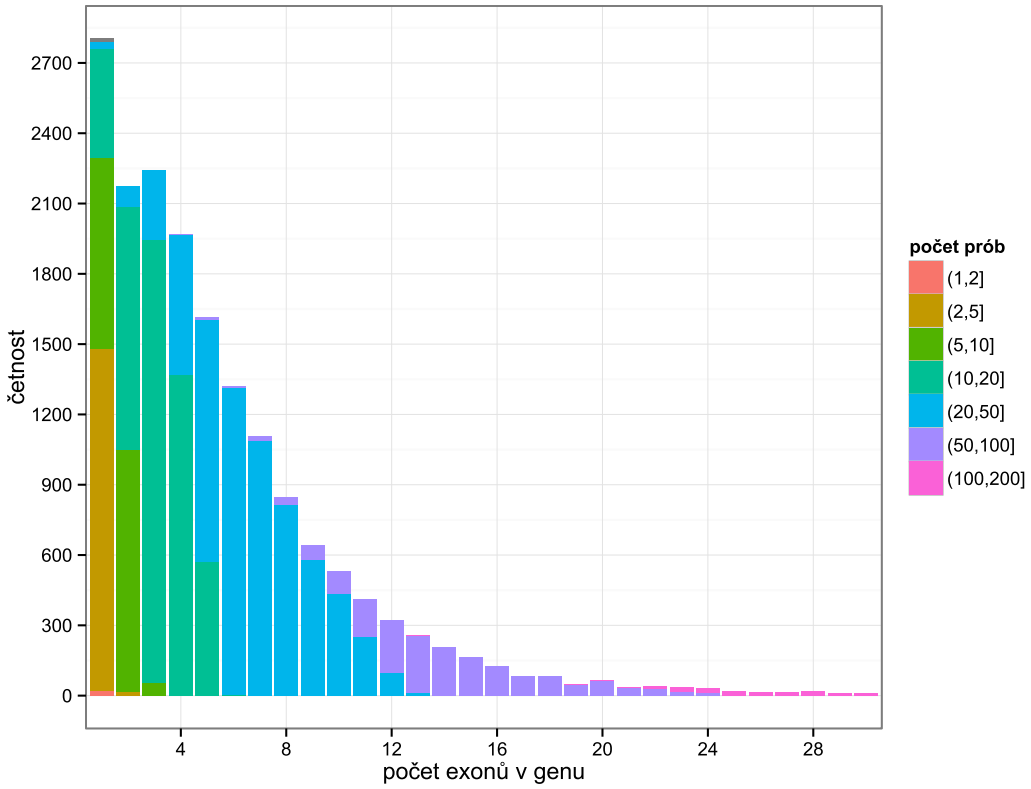
Obrázek 1.3: *Histogram pokrytí exonů próbami. Barva představuje počet probesetů reprezentujících daný exon. Pro přehlednost bylo vynecháno 24 exonů s 30 nebo více próbami.*

Při hybridizaci může docházet k *nespecifické vazbě* (NSB, non-specific binding) – hybridizace neúplně komplementárních řetězců. Pro korekci na NSB byly na dříve používaných *genových* expresních čipech dva typy prób v párech – PM (*perfect match*) próby pro detekci signálu a MM (*mismatch*) próby pro detekci NSB. MM próby mají téměř stejný řetězec jako jejich párová PM próba, liší se jedinou bází, obvykle uprostřed próby. Design čipu vycházel z předpokladu, že skutečnému signálu bude odpovídat rozdíl $PM - MM$, to je celkový signál zmenšený o NSB. Problémem je, že některé MM próby zachycují NSB i skutečný signál a často byly jejich hodnoty vyšší, než u odpovídajících PM prób [7].

Na novějších *exonových* expresních čipech už MM próby nejsou a pro korekci na NSB se používají speciální kontrolní próby a specifické vlastnosti řetězců. Koncept kontrolních prób vychází z poznatku, že NSB je značně ovlivněno obsahem G a C bází v řetězci. Kontrolní próby pokrývají rozpětí obsahu G a C bází a pro výpočet NSB pro PM próbu se používají kontrolní próby se stejným obsahem G-C bází.

Každý exon je reprezentován alespoň jednou sadou prób (*probesetem*). Sada prób obsahuje většinou čtyři próby, některé sady jich mají méně z důvodu omezené délky regionu pro výběr prób. Celkové počty prób v exonech jsou vidět na obrázku 1.3. Souhrn počtu exonů v genech je na obrázku 1.4.

V této práci jsou použita data získaná pomocí exonových expresních čipů GeneChip Mouse Exon 1.0 ST Array.



Obrázek 1.4: *Graf počtu exonů v genech. Barva přestavuje počet prób reprezentujících daný gen. Pro přehlednost bylo vynecháno 98 genů s více než 30 exony.*

Podklady pro exonové expresní čipy od firmy Affymetrix pochází ze dvou hlavních zdrojů [6]:

- databáze genů odvozených z cDNA[1] – RefSeq mRNA [8], GenBank [9], dbEST [10],
- Predikované geny – GENSCAN [11], Ensembl [12], Vega [13] a další.

Próby jsou klasifikovány do tříd podle úrovně průkaznosti. Nejvyšší úrovně (ozn. *core*) dosahují próby z nejspolehlivěji určených genů s úplnými záznamy v GenBank. Nižší úroveň (ozn. *extended*) obsahuje próby z genů s neúplnými záznamy v GenBank, případně prokázaných jinými zdroji. Do nejnižších úrovní (*Full, Free, Ambiguous*) jsou zařazeny próby z algoritmicky predikovaných genů a exonů.

1.3 Postup zpracování čipů

Z tkáně zkoumaného jedince je vytvořen vzorek, RNA z tohoto vzorku je převedena na cDNA a sekvence je označena fluorescentní látkou[2]. Vzorek je aplikován na čip – molekuly ve vzorku hybridizují s komplementárními próbami, podobnost

[1] cDNA – komplementární DNA syntetizovaná z mRNA pomocí enzymu reverzní transkriptázy

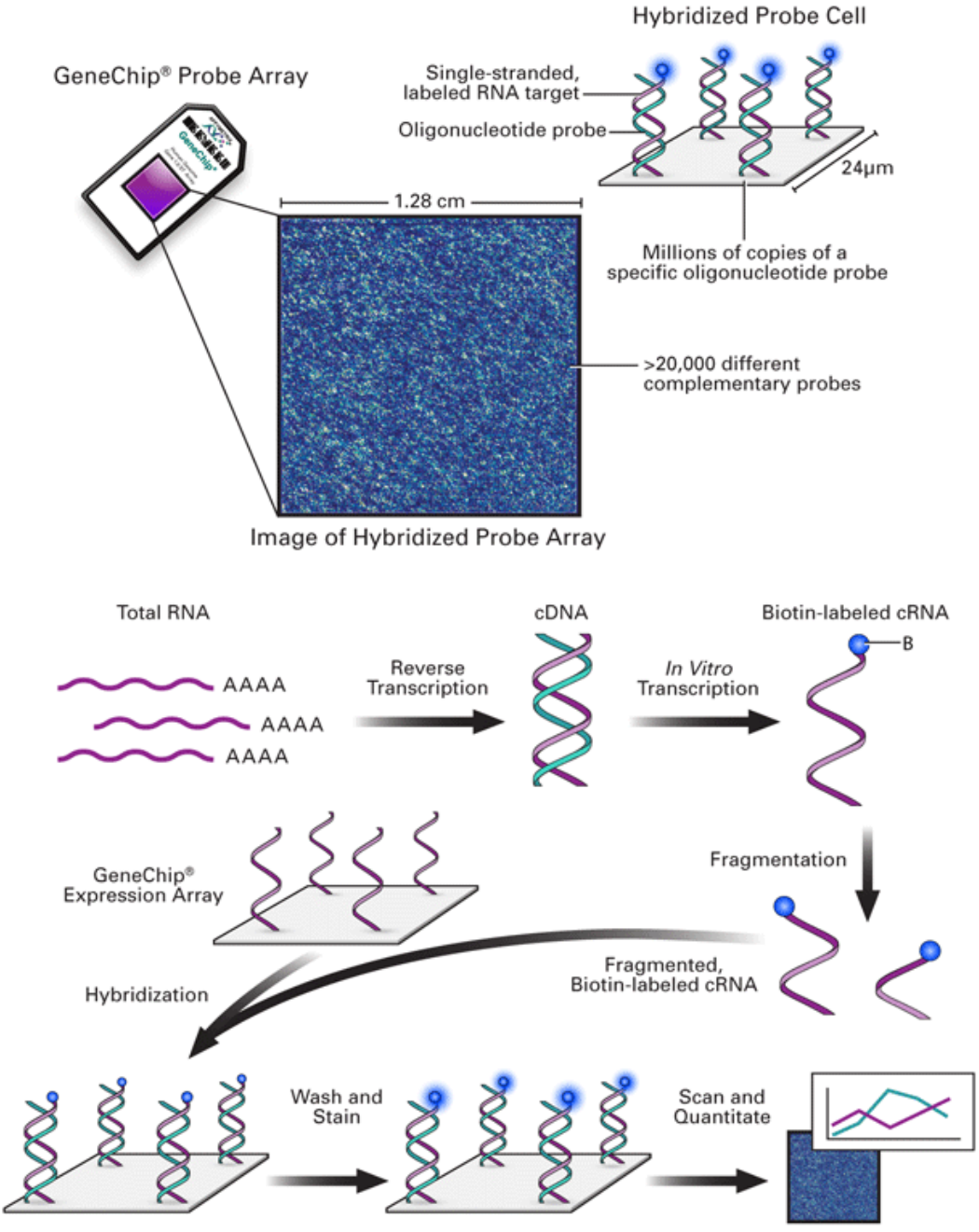
[2] Podrobnosti zpracování (jako např. druh indikační látky) se mohou lišit u konkrétních čipů.

navázaných řetězců určuje sílu vazby. Při následném omytí čipu jsou odplaveny molekuly vzorku s nedostatečně pevnou vazbou. Omytý čip je skenován pomocí konfokálního mikroskopu, tedy je zaznamenána intenzita vyzářeného světla při excitaci fluorescentního barviva laserem na vzorcích přichycených na próbách. Na obraz získaný skenováním je pomocí speciálních značek na čipu umístěna mřížka. Výsledné hodnoty intenzit jsou průměry hodnot získaných z pixelů v příslušných polích mřížky [14]. Shrnutí postupu je znázorněno na schématu 1.5.

1.4 Použití čipů pro příbuzné druhy

Způsob konstrukce expresních čipů vytváří omezení pro jejich použití. Sekvence próbumístěných na čipu jsou určeny při vytváření čipu – sekvence pochází z referenčních kmenů modelových laboratorních organismů. Při použití vzorku pocházejícího z jiného druhu, než podle kterého byl čip vytvořen, bude rozdílností DNA druhů (přítomností jednonukleotidových záměn – SNP) ovlivněna vazba próbu a tím i naměřený signál. Pro odstranění rozdílnosti je možné, ale velmi nákladné, vyrobit čip podle konkrétního nemodelového vzorku.

Přesnější výsledky ze vzorků z nemodelových organismů je možné získat pomocí moderní analýzy RNA-Seq. Tato metoda používá přímé sekvenování RNA ve vzorku, měření není zprostředkováno vazbou na předem určené sekvence jako u expresních čipů [15].



Obrázek 1.5: *Obraz získaný skenováním expresního čipu a schéma postupu zpracování čipu. Uvedené schéma znázorňuje jiný model čipu, než byl zkoumán v práci, proto nemusí souhlasit konkrétní rozměry a počet prób. Princip zpracování je ale stejný. Zdroj: Affymetrix*

2. Metody sumarizace expresních čipů

Výstupem zpracování expresního čipu jsou hodnoty intenzit světelné odezvy naměřené na jednotlivých políčkách čipu. Aby bylo možné tato data interpretovat jako míru exprese exonů, je třeba nejprve provést *sumarizaci*. Cílem sumarizace je sdružení hodnot pro probesety a normalizace rozdělení, protože chceme měřit efekt prób, ne efekt čipu. Sumarizační metody by měly být *robustní* a poradit si i s odlehlými hodnotami.

V době, kdy se expresní čipy staly populární metodou zkoumání genové exprese, byl pro jejich sumarizaci používán algoritmus MAS5 (2001) vyvinutý firmou Affymetrix. Algoritmus MAS5 používá jednoduchou korekci v podobě $PM - MM$ (odečtení hodnoty signálu „mismatch próby“ od hodnoty signálu příslušné párové „perfect match próby“). Tato korekce se ukázala být suboptimální a později byly vyvinuty algoritmy, které dosahují lepších výsledků [16]. Nakonec se v designu čipů upustilo od používání MM prób pro každou PM próbu a objevily se algoritmy využívající znalosti genetické sekvence prób [17].

Jako alternativa pro původní algoritmus MAS5 byl vyvinut algoritmus RMA (2003)[16]. S využitím poznatku, že vazebnost prób je ovlivněna obsahem G/C bází, byl algoritmus RMA vylepšen do podoby gcRMA (2004)[17]. Společnost Affymetrix uvedla nástupce MAS5, algoritmus PLIER (2005)[18]. Pro analýzu exprese na genové úrovni byl algoritmus PLIER rozšířen do podoby IterPLIER[19]. V následujících sekcích budou podrobně představeny algoritmy gcRMA, PLIER a IterPLIER.

Sumarizační algoritmy jsou implementovány v několika balících a knihovnách. V této práci bude použita implementace ze sady nástrojů *Affymetrix Power Tools*. Podrobnosti o nástrojích a postupu sumarizace jsou v kapitole 5.

2.1 gcRMA

Algoritmus gcRMA vychází z algoritmu RMA, ale používá background korekci založenou na znalosti zkoumaných řetězců [17].

2.1.1 RMA

Pro původní algoritmus MAS5 byla vyvinuta alternativa – algoritmus RMA (Robust Multiarray Average) [16]. Zkoumání expresních čipů vedlo k poznatku, že variabilita konkrétní próby napříč čipy je znatelně nižší, než variabilita mezi próbami na jednom čipu [20].

Algoritmus RMA se skládá z těchto základních kroků:

1. Background korekce,
2. Normalizace a logaritmizace,
3. Sumarizace (lineární model).

Původní background korekce použitá pro RMA předpokládá, že se signál zachycený na pozadí skládá z optického šumu a nespecifické vazby. Také zohledňuje poznatek, že MM próby na čipu mohou zachycovat navíc skutečný signál, což vede k záporným hodnotám při odečtu od intenzit PM prób. Proto RMA používá background korekci, jejímž výsledkem jsou pozitivní hodnoty.

Aby bylo možné intenzity naměřené na jednotlivých čipech porovnávat a sumarizovat, je nutné nejprve hodnoty normalizovat. V rámci RMA k tomu slouží *kvantilová normalizace*, která zajistí, že hodnoty na jednotlivých čipech budou mít stejné rozdělení. Následně jsou všechny hodnoty logaritmovány ($\log_2$). Tím se multiplikativní vztahy stávají aditivními.

Sumarizační model RMA může být zapsán jako

$$T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}, \;\; i = 1, \ldots, I, \;\; j = 1, \ldots, J, \tag{2.1}$$

kde T reprezentuje transformaci, která provádí background korekci, normalizaci a logaritmizaci PM intenzit, e_i reprezentuje expresní hodnotu na čipech $i = 1, \ldots, I$ v $\log_2$ škále, a_j reprezentuje efekt afinity pro próby $j = 1, \ldots, J$ v log škále, a ε_{ij} je chybový člen. Nyní lze nahlédnout, že jde o aditivní model pro log transformaci (upravených a normalizovaných) PM intenzit. Pro určení odhadu expresních hodnot e_i (log) je použita robustní regrese, tzv. *median polish* [21]. Výše uvedený postup je souhrnně označován jako RMA.

2.1.2 Využití biologických vlastností sekvence

V tradičních hybridizačních postupech je pozorované nespecifické hybridizační pozadí často způsobeno částečnou homologií nukleových kyselin mezi dvěma jednostrannými vlákny s nedokonalou komplementaritou. Tento problém je spojen s chemismem bází v molekule nukleové kyseliny. Báze G/C v sekvenci vedou k silnější hybridizaci, protože každý G-C pár tvoří tři vodíkové vazby, zatímco každý A-T pár tvoří dvě vazby. Pro predikci specifické hybridizace na základě kompozice prób byl zaveden pojem *afinita próby* (též *vazebnost*), modelovaný jako suma efektů bází v závislosti na pozici:

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} \mathbb{I}_{b_k} \quad \text{pro} \quad \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l, \tag{2.2}$$

kde k = 1,...,25 značí pozici v próbě, j je písmeno báze, b_k reprezentuje bázi na pozici k. Funkce $\mathbb{I}_{b_k}$ je indikační funkce, která vrací 1, pokud k-tá báze je typu j, jinak 0. Hodnota $\mu_{j,k}$ je příspěvek k afinitě báze j na pozici k. Pro fixní j je efekt $\mu_{j,k}$ určen jako polynom stupně tři. Model je fitován na logaritmus intenzit z mnoha čipů metodou nejmenších čtverců.

2.1.3 Background model

Původní algoritmus nazývaný gcRMA [17] používá složitý background model přizpůsobený biologickým vlastnostem prób za použití veličiny α ze vztahu (2.2). Zároveň je ale výpočet odvozen pro použití na čipech s PM i MM próbami.

Při použití sady APT pro expresní čipy bez MM prób je model zjednodušen. Hodnota background korekce je vypočtena za pomoci mediánu intenzit prób se

stejným obsahem G/C bází, jako v dané PM próbě, který je od PM próby odečten [22]. Výpočet je následující:[1]

$$PM_{cor} = \frac{(I_{PM} - BKG) + \sqrt{(I_{PM} - BKG)^2 + 4 \cdot I_{PM} \cdot BKG \cdot L}}{2}, \qquad (2.3)$$

kde PM_{cor} je hodnota po background korekci, I_{PM} je naměřená intenzita próby (v lineární škále), BKG představuje medián intenzit prób se stejným obsahem G/C bází a L je ladící konstanta s výchozí hodnotou 0.005.

Sumarizace se v ostatních krocích shoduje s algoritmem RMA, liší se pouze background korekcí.

2.2 PLIER

Probe Logarithmic Intensity Error (PLIER) je algoritmus pro sumarizaci expresních čipů vyvinutý společností Affymetrix [18]. V porovnání s dříve používaným MAS5 algoritmus PLIER dává vylepšený signál (souhrnnou hodnotu pro probeset) započítáním experimentálně pozorovaných vlastností znaků. Náležitě řeší chybu v bodech s nízkým i vysokým signálem.

Konkrétně používá parametr *vazebná afinita próby*, který reprezentuje sílu signálu pozorovaného ve specifické koncentraci pro danou próbu. Afinita próby (také *feature response*) se počítá z dat napříč čipy. Chybový model, který je použit pro PLIER, předpokládá, že chyba je úměrná pozorované intenzitě, ne intenzitě po background korekci. Zároveň je v odvození metody použit předpoklad, že chyba MM prób je nepřímo úměrná chybě PM prób, což je na první pohled neintuitivní přístup. Bylo ale prokázáno, že tímto přístupem PLIER dosahuje dobrých výsledků, chybový model algoritmu PLIER má mnoho klíčových vlastností ideální chybové funkce [23].

2.2.1 Algoritmus

PLIER je založen na několika předpokladech o chování prób a signálů (targets). Prvním předpokladem je, že signál nesmí být negativní, ale může být nulový. Druhým předpokladem je lineární vztah celkového signálu (T, total response) a signálu koncentrace (t, target response), kde afinita próby (f, feature response) určuje sklon lineární závislosti:

$$T \sim f \cdot t. \qquad (2.4)$$

Na genovém expresním čipu jsou PM próby a MM próby umístěny fyzicky blízko sebe, proto PLIER předpokládá, že nespecifická vazba je velmi blízká pro PM a MM próby. Zároveň předpokládá, že nespecifická vazba se liší mezi různými místy na čipu a mezi experimenty (čipy).

Nejsignifikantnějším zdrojem variability je multiplikativní chyba intenzity, tedy chyba v opakovaných experimentech pro intenzitu signálu (I, feature intensity) je přibližně z log-normálního rozdělení (tedy $log(I)$ je přibližně normální). Předpokládá se, že nespecifická vazba B a celkový signál T dávají dohromady pozorovanou intenzitu ($I \sim T + B$), ale hodnota B se může lišit mezi vzorky.

[1] Vzorec byl odvozen ze zdrojového kódu balíku nástrojů APT.

PLIER vychází ze zjednodušeného modelu:

$$PM - MM = f \cdot t. \tag{2.5}$$

Za předpokladu multiplikativní chyby intenzit lze model vyjádřit jako:

$$e_{PM} PM - e_{MM} MM = f \cdot t, \tag{2.6}$$

kde e_{PM} a e_{MM} jsou náhodné veličiny pro PM a MM z log-normálního rozdělení ($log(I)$). Dobrý odhad afinity a signálu vede k dobrému odhadu e_{PM} a e_{MM} blízko 1 (tedy data odpovídají pozorováním s minimální chybou).

Pokud by byla chyba přesně log-normální, bylo by možné optimalizovat funkci $log(e_{PM})^2 + log(e_{MM})^2$ (analogicky metodou nejmenších čtverců). Je ale známo, že na čipech jsou odlehlé hodnoty a rozdělení chyby není přesně log-normální. Zároveň je hledání e_{PM} a e_{MM} s touto podmínkou výpočetně náročné. Je vhodné zdůraznit, že e_{PM} a e_{MM} nejsou skutečné chyby na čipu (ty jsou neznámé), ale hodnoty určující „správnost odhadu“. Proto je použit jednodušší model chybového členu.

Možné zjednodušení předpokládá, že $log(e_{PM})^2 = log(e_{MM})^2$. Jsou dvě možnosti, jak může tato rovnost nastat:

$$log(e_{PM}) = log(e_{MM}), \text{ nebo} \tag{2.7a}$$
$$log(e_{PM}) = -log(e_{MM}). \tag{2.7b}$$

Za použití předpokladu (2.7a), že chyba pro PM a MM v páru prób i je stejná, tedy $e_{PM} = e_{MM} = e$, lze odvodit vztah

$$PM_{ij} - MM_{ij} = f \cdot t \cdot e, \tag{2.8}$$

což je rovnice původního algoritmu MAS5. Nevýhody tohoto modelu jsou známy, obzvláště pro nízké intenzity [16].

PLIER nepředpokládá, že jsou si chyby pro PM a MM v jednom páru rovny, ale předpokládá vztah (2.7b), který lze upravit jako:

$$e_{PM} = \frac{1}{e_{MM}}. \tag{2.9}$$

Tento předpoklad je z biologického hlediska neintuitivní; próby jednoho páru na čipu sousedí a lokální změny by měly obě ovlivnit ve stejném směru. Bylo ověřeno [23], že PLIER s tímto předpokladem dosahuje dobrých výsledků a přibližuje se ideální chybové křivce.

Za použití předpokladu (2.9) má redukovaný model následující tvar:

$$e \cdot PM - \frac{MM}{e} = f \cdot t, \tag{2.10}$$

ze kterého lze vyjádřit:

$$e = \frac{f \cdot t + \sqrt{(f \cdot t)^2 + 4 \cdot PM \cdot MM}}{2 \cdot PM}. \tag{2.11}$$

Algoritmus PLIER hledá f a t taková, že průměrný „reziduál“ $r = log(e)$ je roven nule. Bez odlehlých hodnot by odhad signálů minimalizoval r^2. Na čipech jsou odlehlé hodnoty, proto je použita funkce podobná r^2 pro r blízké 0, ale

snižuje váhu konců rozdělení. Přesněji za pomoci aproximované Newtonovy metody minimalizuje robustní průměr hodnot v podobě Geman-McClureho funkce (z je ladící konstanta):

$$\frac{r \cdot r}{1 + r\frac{r}{z}}. \tag{2.12}$$

Výpočet začne návrhem odhadu signálu (koncentrace, *target response*) a afinity (*feature response*) pro každý čip a každý pár prób z dostupných dat, potom pomocí Newtonovy metody hledá lepší odhady, které přesněji odpovídají daným datům. Algoritmus končí, pokud nemůže najít lepší odhad.

2.2.2 PLIER bez MM prób

Výše uvedený algoritmus PLIER slouží pro sumarizaci genových čipů. PLIER je možné použít také pro sumarizaci modernějších exonových expresních čipů. Exonové čipy, na rozdíl od genových čipů, neobsahují MM próby, proto musí PLIER používat následující rovnici:

$$e = \frac{f \cdot t + BKG}{PM}, \tag{2.13}$$

kde BKG je odhad nespecifické vazby pozadí. Pro výpočet nespecifické vazby jsou na čipu speciální próby s různým obsahem G/C bází. Intenzity naměřené na těchto próbách se používají k určení nespecifické vazby na próbách se stejným obsahem G/C bází.

2.3 IterPLIER

Předchozí uvedené algoritmy lze použít pro sumarizaci na exonové i genové úrovni. Pro účel sumarizace na genové úrovni byl algoritmus PLIER vylepšen [19]. Toto vylepšení, nazvané IterPLIER, funguje dobře pouze s velkým počtem prób reprezentujících jeden vzorek. Pro genovou sumarizaci jsou probesety reprezentující jednotlivé exony sdruženy do tzv. *meta probesetu*, který reprezentuje celý gen.

Algoritmus IterPLIER opakovaně spouští sumarizaci pomocí PLIER a vyřazuje próby, které nejméně odpovídají celkovému signálu. Díky sloučení probesetů je dostupný dostatečný počet prób pro iterace. Tímto postupem IterPLIER identifikuje próby nejvhodnější pro odhad signálu a tedy dosahuje lepších výsledků, než samotný PLIER [19].

2.4 DABG

Algoritmus DABG není určen pro sumarizaci signálu, ale slouží jako kontrola kvality signálu a zároveň indikace exprese. Detekční proces DABG porovnává PM próby s rozdělením prób nespecifické vazby se stejným obsahem G/C bází. Výstupem porovnání jsou p-hodnoty, které jsou agregovány pro probeset. Výsledná hodnota DABG říká, s jakou pravděpodobností probeset zaznamenává šum na pozadí.

3. Zkoumaný genetický materiál

Nyní popíšeme rozdíly mezi genomy různých organismů. Nejprve obecné pojmy, poté konkrétní příklad z experimentu, který byl zdrojem dat pro tuto práci.

3.1 Mutace

Při porovnání řetězců DNA různých jedinců je možné narazit na rozdíly, souhrnně označované jako *mutace* nebo *genové variace*. Základní prvky genové variace jsou běžně klasifikovány jako:

- SNP (single nucleotid polymorphism) – záměna jedné báze v řetězci za jinou,
- inzerce / delece (souhrnně ozn. indel) – přidání / odebrání báze nebo části řetězce,
- inverze – převrácení části řetězce,
- duplikace – opakování části řetězce.

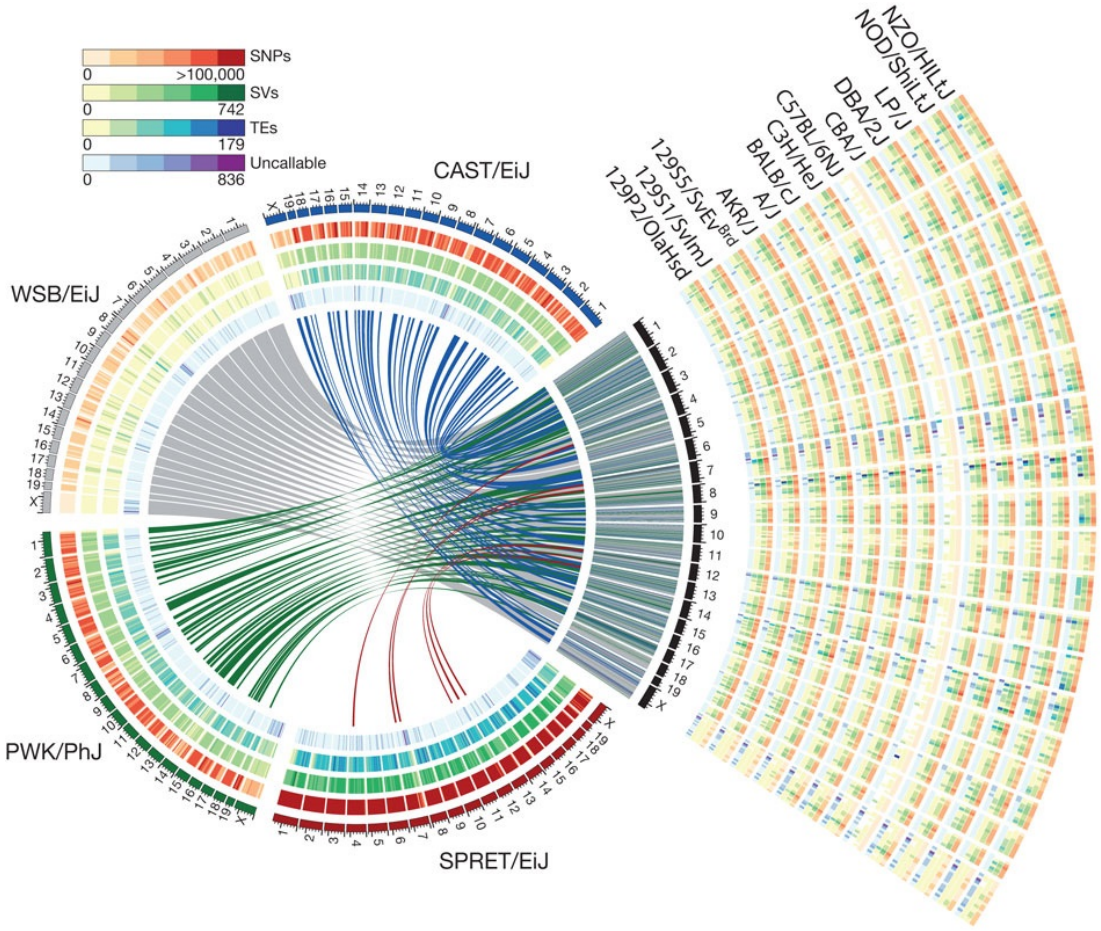
Rozdíly v DNA mohou být různého rozsahu a mohou mít na jedince různý vliv. Změny v řetězcích se nemusí nijak projevit, pokud jsou v intronech a nekódujících oblastech DNA. Změny v oblastech DNA kódujících protein mohou ovlivňovat produkt genu. Záměny v oblasti kódující protein, které vedou k syntéze stejného proteinu, jsou označovány jako *synonymní SNP*. Změny v regulačních oblastech genů ovlivňují míru jejich exprese.

3.2 Vztahy kmenů zkoumaných myší

Exonové čipy Mouse Exon 1.0 ST Array jsou vytvořeny podle referenčního genomu myši domácí (*Mus musculus*), který pochází z laboratorního kmene označovaného C57BL/6J. Tento kmen byl vyšlechtěn v první polovině 20. století z poddruhů *Mus musculus musculus*, *Mus musculus domesticus* a *Mus musculus spretus* a stal se nejpoužívanějším kmenem laboratorních myší [24].

Jedinci myší, jejichž tkáň byla sbírána a analyzována pomocí expresních čipů, patří k volně žijícím poddruhům *Mus musculus domesticus* a *Mus musculus musculus* [1]. Z těchto poddruhů byly vyšlechtěny laboratorní kmeny WSB/EiJ a PWK/PhJ. Jednotlivé kmeny se liší genetickou sekvencí, zároveň u nich mohou být různě exprimovány některé stejné nebo podobné geny. Předpokládáme, že rozdíly pozorované mezi laboratorními kmeny jsou srovnatelné i s rozdíly mezi původními volně žijícími poddruhy. Při zkoumání vzorků z volně žijících druhů na čipech za účelem měření exprese mohou být naměřená data ovlivněna jak skutečně odlišnou expresí, tak rozdíly v genetické sekvenci, které zapříčiňují sníženou vazebnost prób.

Odlišnostmi mezi genetickými sekvencemi jednotlivých laboratorních kmenů i volně žijících poddruhů se zabývá práce (Keane et al., 2011 [25]). Z jejích výsledků vyplývá, že rozdíl mezi referenčním kmenem C57BL/6J a kmenem



Obrázek 3.1: *Znázornění rozdílů kmenů laboratorních myší vzhledem k referenčnímu genomu. Čtyři kmeny odvozené od volně žijících druhů (CAST/EiJ, WSB/EiJ, PWK/PhJ a SPRET/EiJ) reprezentují po řadě M. m. castaneus, M. m. musculus, M. m. domesticus a M. spretus. Každému druhu odpovídá kruhová výseč s označením chromozomů (1, . . . , 19, X). Na pravé straně je zobrazeno 13 klasických laboratorních kmenů, referenční genom vychází z kmene C57BL/6. Červeně jsou označeny počty SNP, ostatní barvy reprezentují jiné změny genetické sekvence. SV – strukturální variace – změna v DNA, obvykle většího rozsahu než SNP. TE – transpozibilní element (transpozon), úsek DNA přesunutý na jiné místo; podmnožina SV. Uncallable – nerozhodnutelný genotyp; bez reference. Tmavší barva znamená více rozdílů. Spojnice uprostřed kruhu označují úseky, které jsou nejblíže referenci. Zdroj:[25]*

WSB/EiJ (*M. m. domesticus*) je menší (jsou si „bližší“), než rozdíl mezi referenčním kmenem a kmenem PWK/PhJ (*M. m. musculus*). Na obrázku 3.1 jsou znázorněny rozdíly kmenů laboratorních myší vzhledem k referenčnímu genomu.

Součástí výzkumu [25] jsou data popisující konkrétní genetické rozdíly v genomech jednotlivých kmenů, včetně *SNP*. Pomocí záznamů o SNP se pokusíme snížit negativní dopad genetických rozdílů mezi kmeny a zlepšit sumarizaci exonových čipů.

U každého jedince z obou poddruhů byly odebírány vzorky z ledvin (KID), sleziny (SPN) a varlat (TES). Ačkoliv buňky všech tkání jednoho jedince sdílí stejnou DNA, liší se genovou expresí, pro jejíž detekci slouží expresní čipy.

4. Využití expresních čipů pro nemodelové druhy

Jak bylo zmíněno v části 1.4, při použití vzorků z jedinců druhu, který se liší od designového druhu, podle kterého byl vytvořen čip, dochází ke zhoršení kvality měření. Próby na čipu a zkoumané řetězce ve vzorku nedosahují úplné komplementarity, pokud je v řetězci změna daná rozdílností druhů. V následující části navrhujeme možný způsob kompenzace.

4.1 Odstranění prób zasažených SNP

Próby umístěné na čipu jsou vytvořené podle genomu referenčního organismu. Čím odlišnější je organismus, ze kterého pochází vzorek použitý na čipu, tím větší je pravděpodobnost, že část odpovídající nějaké próbě bude obsahovat změnu. Nejčastější a nejsnáze zpracovatelnou změnou je SNP – v dalším postupu budeme rozdíly a změnami v řetězcích myslet právě SNP.[1]

Při výpočtu sumarizace pro probeset může být hodnota snížena vlivem nižšího naměřeného signálu z próby se změnou. Z toho důvodu může být lepší zasaženou próbu vyřadit ze sumarizace. Cílem je odstranění předem známé variability v datech ještě před jejich zpracováním. Idea vyřazení prób zasažených SNP je znázorněna na schématu 4.1.

Se záměrem prozkoumat, zda odstranění prób zasažených SNP ovlivní sumarizaci expresních čipů, provedeme následující:

1. Z dostupných dat obsahujících informace o rozdílech laboratorních kmenů myší extrahujeme záznamy o SNP v genomech kmenů WSB/EiJ a PWK/-PhJ vzhledem k referenčnímu genomu.

2. Identifikujeme próby na expresních čipech, které reprezentují oblasti genů zasažené SNP.

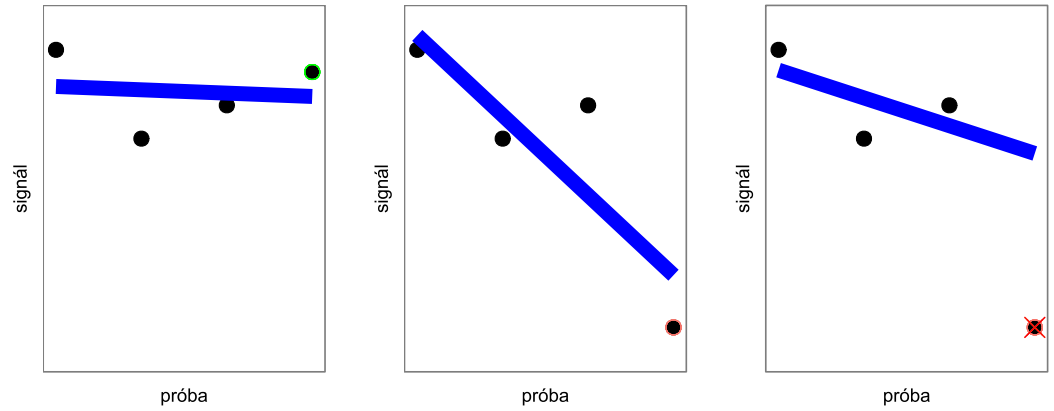
3. Provedeme sumarizaci čipů metodami gcRMA, PLIER a IterPLIER.

4. Pro otestování vlivu našeho zásahu porovnáme sumarizaci původních dat se sumarizacemi s vyřazením prób zasažených SNP a s vyřazením náhodných prób.

5. Provedeme analýzu výsledků sumarizace a porovnáme vliv vyřazení prób na sumarizaci dat z různých tkání a poddruhů myší.

6. Na závěr zhodnotíme robustnost sumarizačních metod vzhledem k systémové variabilitě dané rozdílností uvedených poddruhů.

[1] SNP jsou nejčastěji uchovanou změnou genomu, jelikož ostatní změny mají výrazně větší účinek na výsledný protein a jsou proto úspěšněji odstraňovány selekcí.



(a) (b) (c)

Obrázek 4.1: *Idea zlepšení sumarizované hodnoty probesetu vyřazením próby zasažené SNP (umístěna nejvíce vpravo). (a) Sumarizace probesetu s nezasaženou próbou. (b) Sumarizace probesetu s próbou zasaženou SNP. (c) Sumarizace probesetu s odstraněnou zasaženou próbou.*

4.2 Výběr spolehlivých dat

Při přípravě experimentu byla učiněna následující rozhodnutí:

- použití *core* probesetů s podklady z nejspolehlivějších zdrojů (RefSeq, úplné mRNA GenBank záznamy),
- použití záznamů o SNP s vysokou kvalitou – nejpravděpodobnější výskyty SNP; stejné jako v práci (Keane et al., 2011[25]).

Vybráním probesetů s nejvyšší spolehlivostí se snažíme dosáhnout snížení šumu v signálu.

4.3 Sumarizace

Pro každou tkáň (KID, SPN, TES) z obou poddruhů (*musculus*, *domesticus*) a pro každý algoritmus (PLIER, Iter-PLIER, gcRMA) jsou provedeny tři sumarizace:

1. sumarizace obsahující všechny próby z core probesetů,
2. sumarizace s vyřazením prób zasažených SNP,
3. sumarizace s vyřazením náhodných prób.

Výše uvedené sumarizace jsou provedeny na exonové úrovni a na genové úrovni.

5. Zpracování dat

Zde popíšeme základní informace týkající se výpočtu sumarizace a analýzy.

5.1 Mapování zásahů prób

Informace o próbách jsou uloženy v designovém souboru čipu formátu GFF (viz přílohu B.2.5). Soubory formátu VCF (viz přílohu B.3.1) obsahují informace o rozdílech kmenů myší vzhledem k referenčnímu genomu. Důležitou informací pro mapování je pozice na chromozomu, odkud pochází sekvence próby. Záznamy SNP obsahují souřadnice v genomu, kde byla změna zaznamenána a ve kterých kmenech ke změně došlo.

Zkombinováním informací o próbách ze souboru GFF a informací o SNP ze souboru VCF byly nalezeny próby, které byly pravděpodobně zasaženy SNP. Tyto próby tvoří tzv. *kill-list*, tedy seznam prób vyřazených ze sumarizace. Podle seznamu vyřazených prób je vytvořen stejně dlouhý seznam náhodně vybraných prób. Sumarizace s vyřazením náhodných prób bude sloužit jako kontrola míry efektu, který má vyřazení prób zasažených SNP na výsledek sumarizace.

Extrakci informací relevantních pro sumarizaci a analýzu provádí skripty pro bash (Unix shell) a python a jsou přiloženy k této práci, viz přílohu A.

5.2 Sumarizace

Pro sumarizaci exonových čipů byla použita sada nástrojů Affymetrix Power Tools (*APT*) [26], multiplatformní řešení pro analýzu a sumarizaci čipů vyvíjené firmou Affymetrix. Nástroj pro sumarizaci exonových čipů pomocí zvoleného algoritmu se nazývá `apt-probeset-summarize` [27].

Vstupem pro zpracování dat z exonových čipů jsou následující soubory:

- CEL – vlastní data; obsahuje naměřené intenzity políček na čipu (prób),
- CLF, PGF, BGP, QCC, PS a MPS – knihovní soubory pro sumarizaci obsahující identifikátory a další informace. Formáty souborů jsou popsány v příloze B.

Příklad volání:

```
apt-probeset-summarize \
   -a quant-norm.sketch=0.bioc=false,pm-gcbg,plier \
   -c MoEx-1_0-st-v1.r2.clf \
   -p MoEx-1_0-st-v1.r2.pgf \
   -b MoEx-1_0-st-v1.r2.antigenomic.bgp \
   -s core.ps \
   -o $out \
   --kill-list probes.kill \
   --qc-probesets MoEx-1_0-st-v1.r2.qcc \
   --temp-dir $SCRATCHDIR/data \
   $data_dir/D[1-9]*KID.CEL;
```

Parametr `-a` nastavuje vlastní průběh sumarizace. V tomto případě provede kvantilovou normalizaci, upraví hodnoty PM prób podle mediánu intenzit prób s podobným obsahem G/C bází a spustí kvantifikaci pomocí algoritmu PLIER. Doplňkové nastavení `sketch` umožňuje při kvantilové normalizaci použít pouze podmnožinu z dat pro ušetření paměti. Nastavení `bioc` může být použito pro zachování kompatibility s nástroji z frameworku Bioconductor [27]. Výsledkem jsou sumarizované hodnoty intenzit pro každý probeset ze zadaného seznamu.

5.3 Výpočet na MetaCentru

MetaCentrum VO[28] je virtuální organizace české Národní Gridové Iniciativy MetaCentrum NGI. Akademickým pracovníkům a studentům členů sdružení CESNET poskytuje bezplatně výpočetní a úložné kapacity. Pro účely práce byly tyto zdroje využity pro uložení všech potřebných souborů a pro jejich přípravu a zpracování. Výpočetní cluster umožňuje spuštění velkého množství paralelních výpočtů, díky čemuž bylo reálné provést potřebný počet běhů sumarizace a analýzy v rámci diplomové práce. Skripty pro zadávání úloh jsou přiloženy k práci, viz přílohu A

5.4 BEST – robustní bayesovský odhad

Bayesovský odhad je statistická metoda pro určování parametrů dat. Podstatou bayesovské analýzy je přesouvání a zpřesňování *věrohodnosti* odhadované hodnoty parametrů.

K porovnání výsledků sumarizace použijeme robustní bayesovský odhad zvaný BEST (Bayesian Estimation Supersedes the T-test). Zde použijeme původní implementaci pro R z publikace (Kruschke, 2013 [29]). Na rozdíl od *t-testu* dokáže BEST nejen určit, zda jsou dvě rozdělení rozdílná, ale poskytuje navíc míru rozdílu.

Studentovo rozdělení (t-rozdělení) je pravděpodobnostní rozdělení podobné normálnímu, navíc umožňuje parametrizovat pro „silné konce“, je tedy odolnější proti odlehlým hodnotám. Parametr t-rozdělení se nazývá *normalita* (ν), příklad t-rozdělení s různou normalitou je na obr. 5.1.

Analýza BEST hledá t-rozdělení, které nejlépe aproximuje data. V párovém testu hledá střední hodnotu (μ) a odchylku (σ) pro každý soubor dat a normalita t-rozdělení (ν) stejný pro oba soubory – celkem pět hodnot.

BEST pomocí MCMC (Markov Chain Monte Carlo) odhaduje hodnoty parametrů t-rozdělení, která nejlépe odpovídají datům. V průběhu MCMC jsou vytvořeny řádově tisíce odhadů hodnot parametrů. Ze statistik odhadů je určen interval nejvyšší hustoty (HDI, highest density interval), do kterého patří 95 % určených hodnot. V tomto intervalu se s 95% pravděpodobností vyskytuje zkoumaný parametr.

S pomocí analýzy BEST hledáme následující:

- vliv vyřazení prób zasažených SNP,
- porovnání vlivu vyřazení prób zasažených SNP s vlivem vyřazení náhodných prób,

- souhrn výsledků analýzy pro různé čipy,
- porovnání analýzy pro různé tkáně a pro zkoumané poddruhy.

V dalším budeme analýzu BEST používat jako párový test, kde prvním vstupem bude vždy sumarizace původních dat z daného čipu (bez odstranění prób), druhým vstupem budou data s odstraněnými próbami zasaženými SNP, případně s odstraněnými náhodnými próbami. V každém z párových testů jsou použity pouze probesety ovlivněné vyřazením prób. Do testu jsou použity pouze probesety s p-hodnotou $DABG < 0.05$.

Z hodnot, které lze určit ze statistiky BEST, nás bude zajímat rozdíl středních hodnot (difference of means):

$$\mu_1 - \mu_2, \tag{5.1}$$

rozdíl odchylek rozdělení:

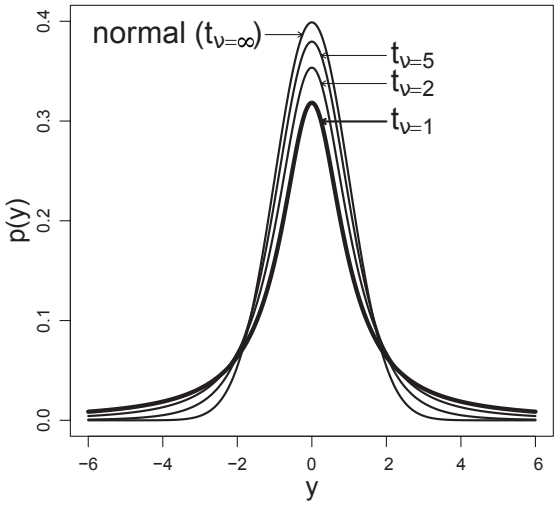
$$\sigma_1 - \sigma_2 \tag{5.2}$$

a velikost účinku (effect size):

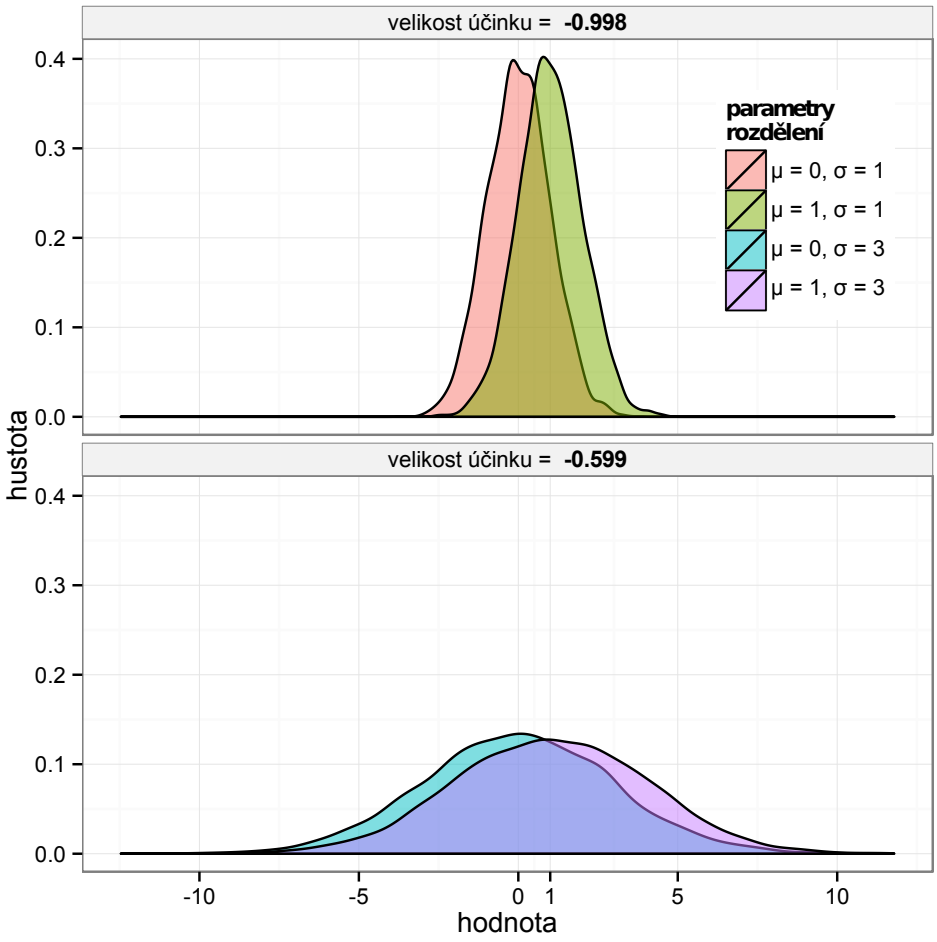
$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}. \tag{5.3}$$

Budeme sledovat, zda je možné rozdíly parametrů pokládat za nenulové a pokud ano, jakých hodnot nabývají. Negativní hodnoty rozdílů, jak je zřejmé z rovnic, znamenají větší hodnoty parametrů pro druhou sadu dat.

Velikost účinku představuje rozdíl středních hodnot normalizovaný na průměrnou odchylku. Tím je znázorněna míra vlivu změny v datech vzhledem k jejich variabilitě. Pro ilustraci uvedeme příklad na obrázku 5.2. Jde o dva vzorky z normálního rozdělení posunuté o 1 v kladném směru (zvýšení střední hodnoty). V prvním případě má rozdělení menší směrodatnou odchylku, než rozdělení ve druhém případě. Proto má posun v prvním případě absolutně větší velikost účinku, ačkoliv jde v obou případech o posun o stejnou hodnotu. Je vhodné zdůraznit, že při zvýšení střední hodnoty je velikost účinku záporná.



Obrázek 5.1: *Příklad t-rozdělení s různou parametrizací. Pro hodnotu $\nu = \infty$ je t-rozdělení normální. Ve všech případech jsou parametry $\mu = 0$ a $\sigma = 1$. Zdroj:[29]*



Obrázek 5.2: *Příklad velikosti účinku pro různě parametrizované vzorky normálního rozdělení. V obou případech jde o posunutí střední hodnoty o 1. Rozdělení na levé straně má menší odchylku než rozdělení na pravé straně. Proto je i při posunu o stejnou hodnotu rozdílná velikost účinku.*

6. Analýza

V experimentu bylo zpracováno 45 exonových čipů zkoumajících vzorky ze tří tkání celkem 8 jedinců poddruhu *M. m. musculus* a 7 jedinců poddruhu *M. m. domesticus*. Sumarizace na exonové a genové úrovni byly provedeny pro původní data, data s vyřazením prób zasažených SNP a s vyřazením náhodných prób.

Výsledky sumarizací byly zpracovány v prostředí pro statistické výpočty R [30]. Při zpracování jsou používána zkrácená označení tkání (ledviny – KID, slezina – SPN, varlata – TES) a poddruhů (*musculus* – M, *domesticus* – D).

6.1 Načtení dat

K načítání dat pro zpracování v prostředí R [30] slouží funkce a skripty přiložené k této práci, viz přílohu A. Výstupem načítacích funkcí je datová struktura obsahující tabulky s výsledky sumarizace pro každou zadanou sumarizační metodu, tkáň a poddruh. Intenzity získané sumarizačními metodami jsou agregovány přes exony, navíc jsou k intenzitám přiřazeny i hodnoty DABG (maximum přes probeset). Posledními agregovanými informacemi jsou počet SNP v exonu, počet prób zasažených SNP a celkový počet prób v exonu.

Pro snazší práci s daty je výhodnější tato přeformátovat do tzv. dlouhého formátu. V takovém je na jednom řádku vždy jeden záznam hodnot určený identifikátory čipu, genu a exonu, případně probesetu.

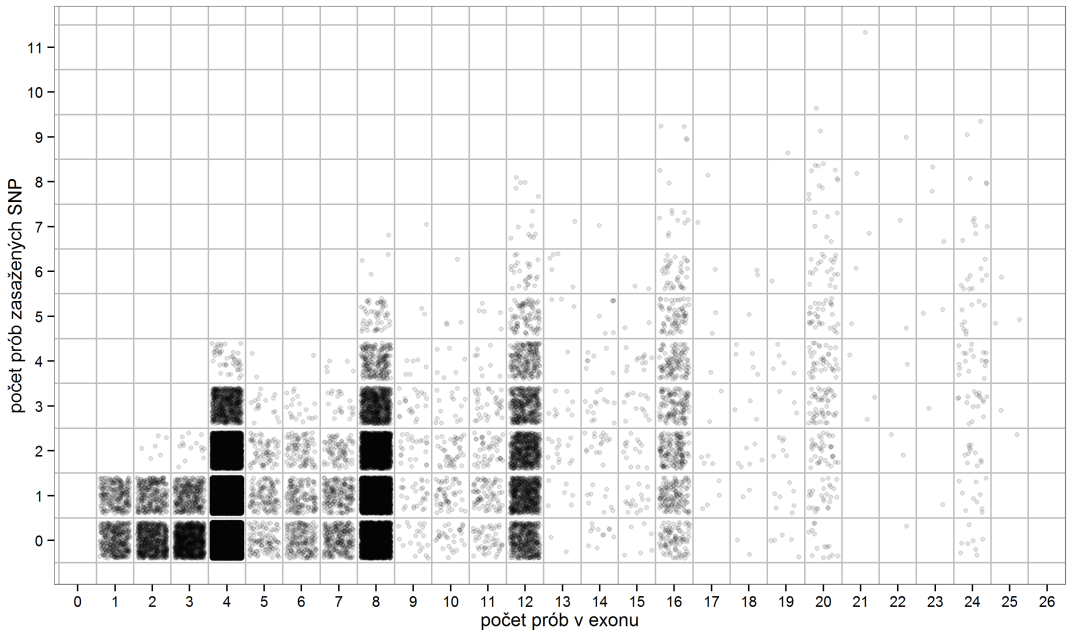
6.2 Explorace dat

Než přistoupíme k rozboru výsledků sumarizačních metod, prozkoumáme, kolik prób bylo zasaženo SNP, a jak odstranění prób zasažených SNP ovlivňuje detekci signálu.

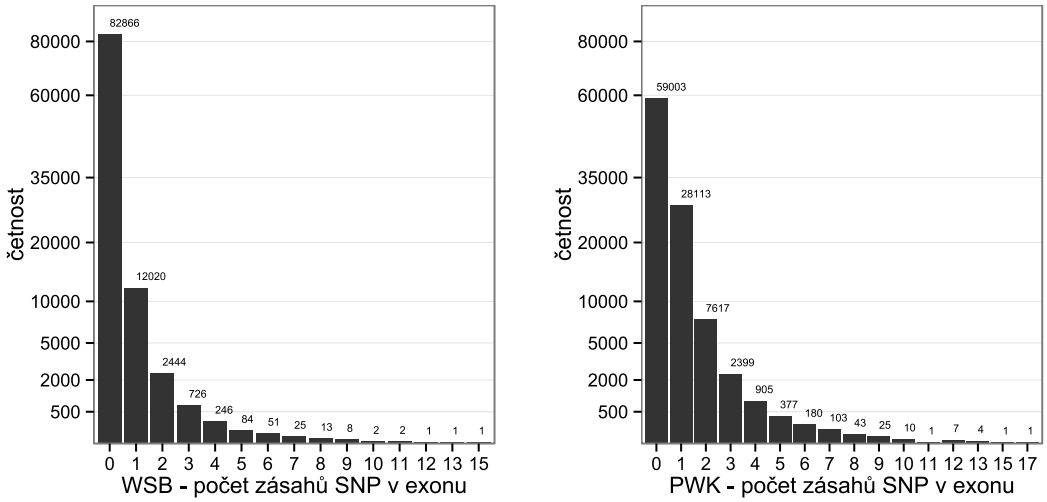
6.2.1 Próby zasažené SNP

Nyní se zaměříme na SNP a zasažené próby. Počet prób zasažených SNP vzhledem k počtu prób v exonu je znázorněn na obrázku 6.1. Výrazná pole se zvýšenou četností jsou patrná proto, že exony mají většinou počet prób dělitelný čtyřmi (viz obr. 1.3).

Z práce o rozdílech kmenů myší [25] je patrné, že se liší počet SNP zde zkoumaných poddruhů vzhledem k referenčnímu kmeni. U prób je tento rozdíl stále patrný. Na histogramech 6.2 jsou znázorněny celkové počty exonů, jejichž próby byly zasaženy SNP. V souladu s předpokladem podobnosti s referenčním genomem je v poddruhu *M. m. domesticus* méně prób zasažených SNP, než v poddruhu *M. m. musculus*. V následující analýze jsou zahrnuty pouze exony a geny ovlivněné vyřazením prób. Jak bylo uvedeno výše, za detekovaný signál jsou považována měření prób s hodnotou DABG menší než 0.05.



Obrázek 6.1: *Znázornění počtu prób v exonech a počtu prób zasažených SNP v exonech. Pro přehlednost bylo vynecháno 58 exonů s více než 25 próbami.*



(a) *Mus musculus domesticus* (b) *Mus musculus musculus*

Obrázek 6.2: *Histogram prób zasažených SNP v exonech jednotlivých kmenů.*

6.2.2 DABG

První hodnotou, která nás ve výstupu sumarizace zajímá, je DABG (*detection above background*). Hodnota DABG představuje p-hodnotu detekce šumu. Próby s nízkou p-hodnotou vykazují signál odlišný od pozadí a jejich naměřená hodnota je považována za detekovaný signál. Konstanta 0.05 je považována za optimální prahovou p-hodnotu DABG pro minimalizaci chyby I. a II. druhu [31].

Konkrétní otázka zní: změní se hodnoty DABG vyřazením prób zasažených SNP? Z porovnání hustoty rozdělení usuzujeme, že rozdělení DABG se ošetřením výrazně nemění. Při pohledu na graf hustoty rozdělení 6.3 předpokládáme, že rozdělení DABG se ošetřením výrazně nemění[1]. Podle analýzy variance (ANOVA) nelze zamítnout nulovou hypotézu shodného rozdělení (neuvedeno).

6.3 Sumarizace na exonové úrovni

V následujících sekcích se budeme postupně věnovat jednotlivým sumarizačním metodám na exonové úrovni. Za použití analýzy BEST nahlédneme vliv vyřazení prób na sumarizaci. Výstupem párové analýzy BEST je rozdíl ošetřených a původních dat. Srovnání rozdílů ukazuje směr a velikost vlivu ošetření.

Důležitým pozorováním, které je potřeba učinit, je srovnání odhadů parametrů pro jednotlivé čipy. Příkladem srovnání čipů je obrázek 6.4[2] – vzorky z ledvin jedinců poddruhu *M. m. domesticus* (D KID) sumarizované metodou gcRMA. Je zřejmé, že variabilita mezi čipy je vzhledem k ošetření zanedbatelná. Stejné výsledky vychází pro všechny čipy a všechny parametry analýzy BEST (rozdíl středních hodnot, rozdíl odchylek, síla efektu) na exonové úrovni. V dalších porovnáních proto budeme používat sjednocení odhadů parametrů přes všechny čipy odpovídající analýzy. Znázornění porovnání odhadů pro všechny čipy a sumarizace jsou přiložena k práci, viz přílohu A.2.3.

6.3.1 gcRMA

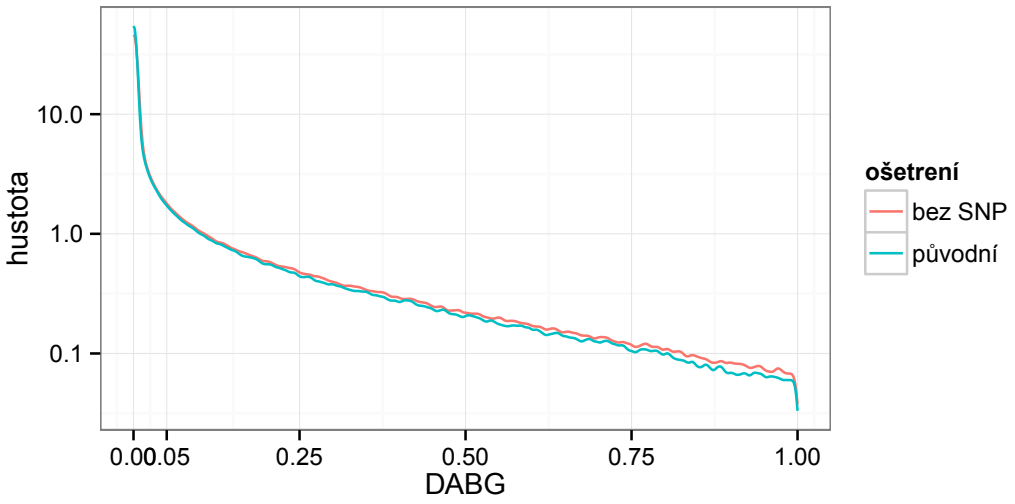
V této části se zaměříme na výsledky sumarizace pomocí gcRMA. Na obrázku 6.5 je grafické znázornění analýzy BEST pro sumarizaci čipu s označením *D1100TES* (vzorek z varlete jedince poddruhu *M. m. domesticus*). Porovnává původní data a data s odstraněnými próbami zasaženými SNP. Z analýzy je patrné, že rozdíl středních hodnot je nenulový, stejně jako velikost účinku (*effect size*). Z histogramů lze odvodit, že data mají rozdělení blízké t-rozdělení, jehož parametry jsou odhadovány analýzou BEST.

Porovnání velikosti účinku mezi poddruhy je znázorněno na obrázku 6.6. Velikost účinku ve vztahu (5.3) zahrnuje střední hodnoty i odchylky rozdělení. Samostatné porovnání těchto veličin je v příloze A.2.3.

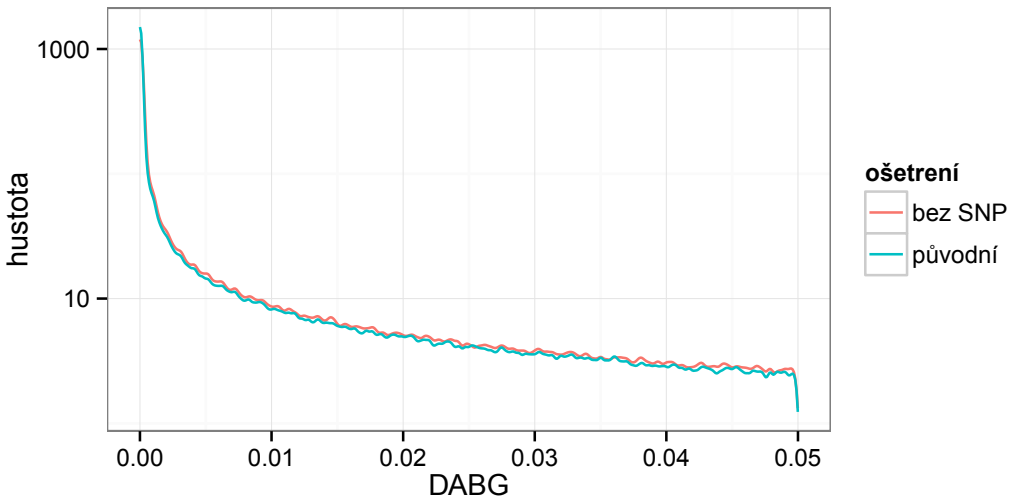
Při odstranění náhodných prób je rozdíl středních hodnot pro oba poddruhy podobný, zároveň dosahuje podobné velikosti účinku. Vyřazení prób zasažených SNP způsobuje zvýšení střední hodnoty a má (absolutně) větší velikost účinku než náhodné vyřazení pro oba kmeny. To znamená, že odstranění zasažených

[1] Konkrétně jde o sumarizaci čipu *D1100KID*

[2] Uvedené znázornění je kombinací tzv. *houslového grafu (violin plot)* a *krabicového grafu (box plot)*. Zobrazuje hustotu (histogram) a kvartily rozdělení dat.



(a) *[0.00 – 1.00]*



(b) *[0.00 – 0.05]*

Obrázek 6.3: *Srovnání hustoty rozdělení DABG (log transformace) pro původní data a s odstraněním SNP. Na celém rozsahu (a) a na rozsahu p-hodnot přijímaných jako detekovaný signál (b). Pozorujeme, že rozdělení se po vyřazení prób výrazně nemění.*

prób způsobuje celkové zvýšení průměru sumarizovaných signálů, což je v souladu s hypotézou v kapitole 4.1.

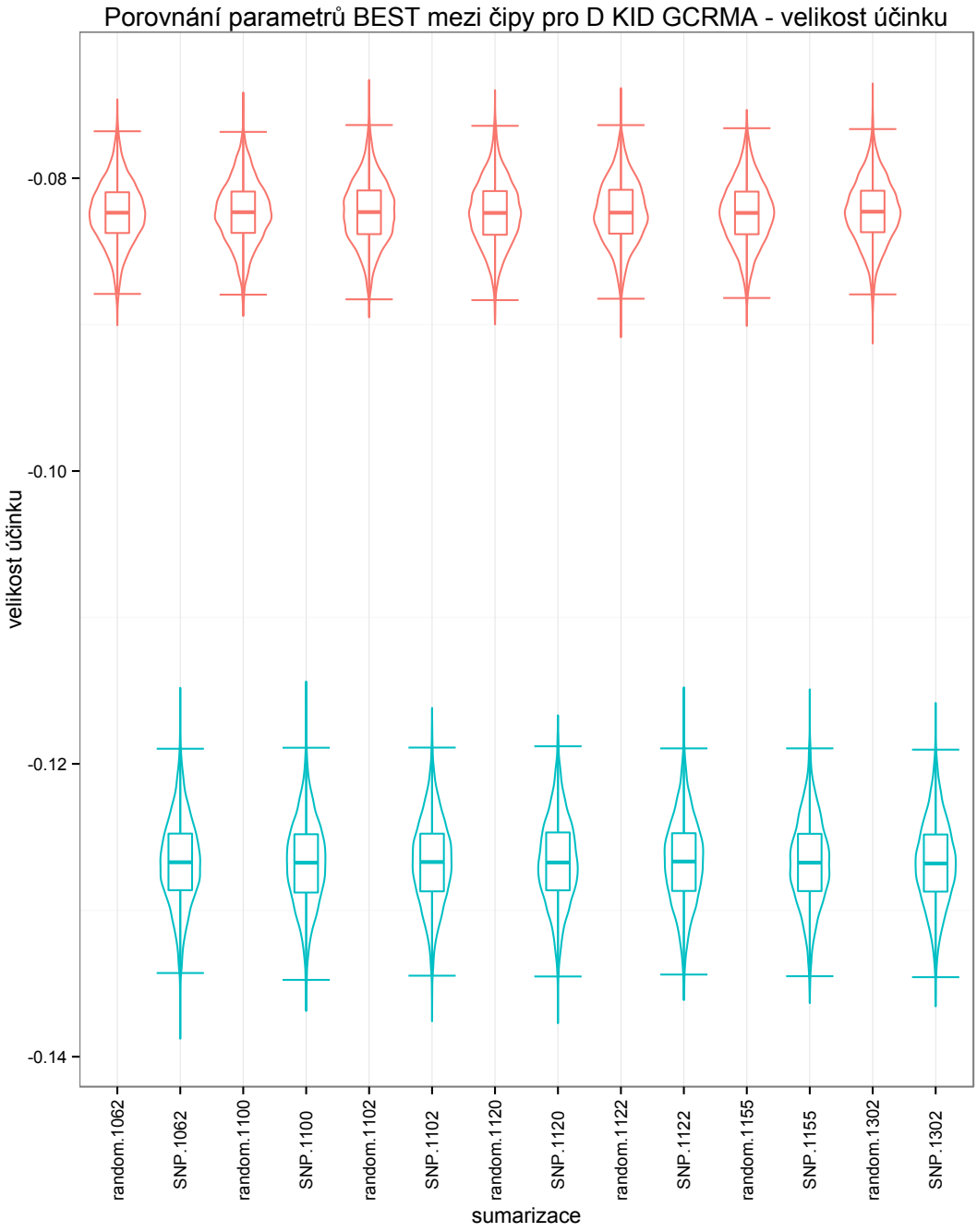
U poddruhu *musculus* je velikost účinku větší než u poddruhu *domesticus*, což je v souladu s poznatkem, že próby poddruhu *musculus* jsou vzhledem k referenčnímu kmeni častěji zasaženy SNP (viz kap. 3.2). Odstranění většího počtu zasažených prób (dle předpokladu s nižším signálem) způsobuje větší rozdíl středních hodnot před ošetřením a po něm u vzdálenějšího poddruhu.

6.3.2 PLIER

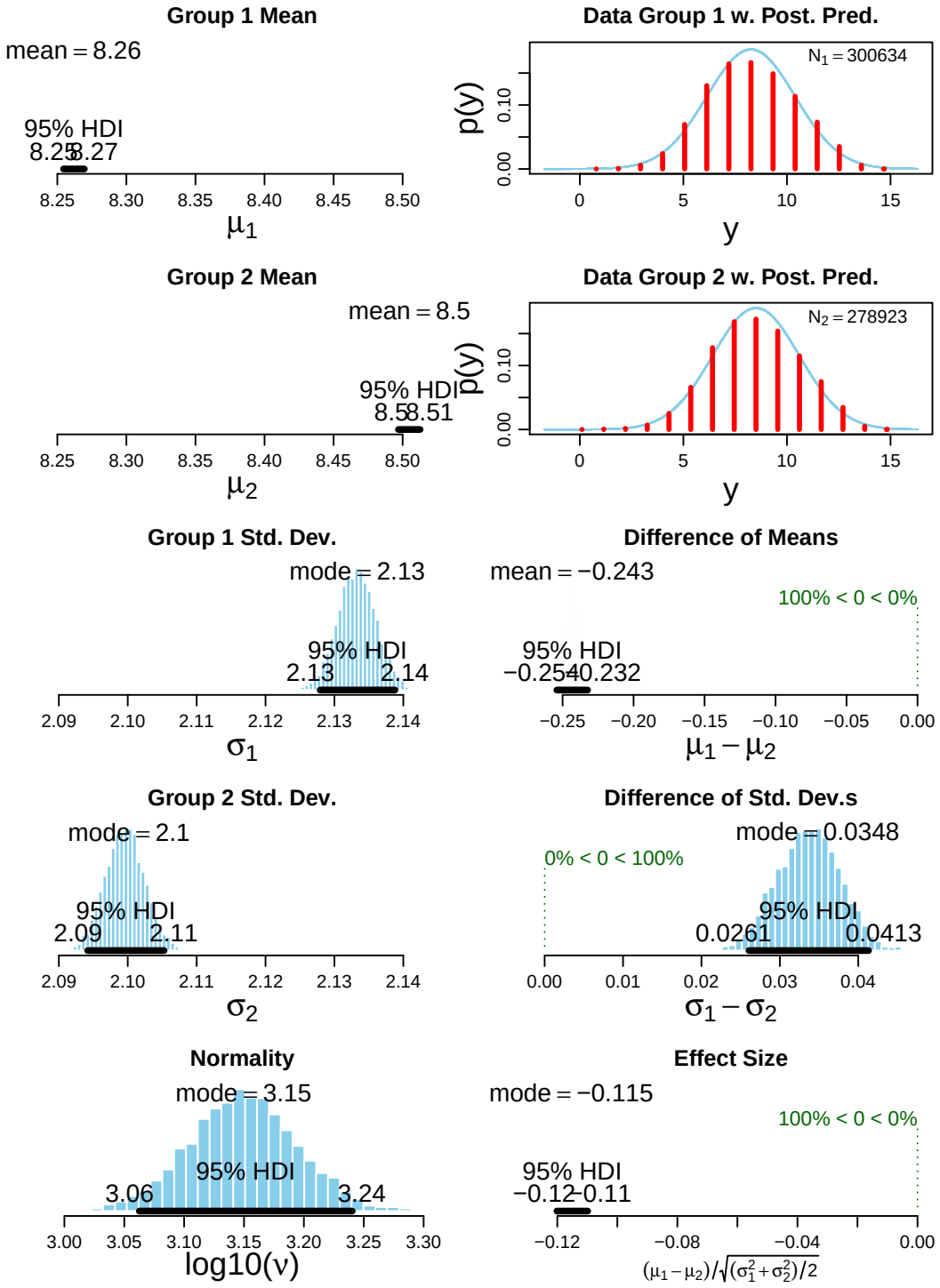
Pro sumarizaci pomocí metody PLIER na exonové úrovni jsou rozdíly mezi odhady parametrů velmi podobné, jako rozdíly v případě použití gcRMA. Na obrázku 6.7 je znázorněno porovnání velikostí účinku pro sumarizaci metodou PLIER na exonové úrovni.

6.3.3 IterPLIER

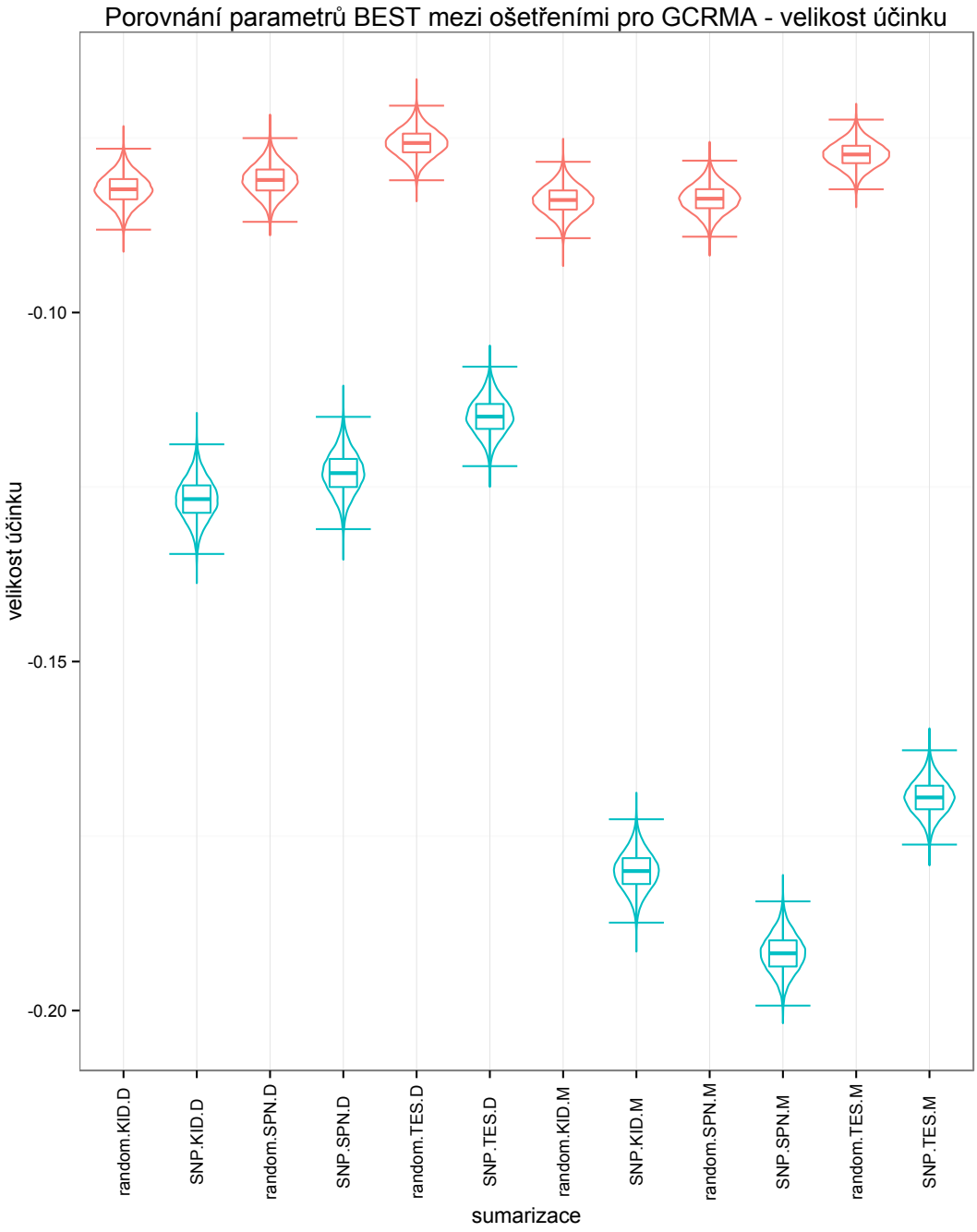
Pro sumarizaci na exonové úrovni jsou výsledky získané pomocí metod PLIER a IterPLIER stejné. Metoda IterPLIER nemá při sumarizaci na exonové úrovni k dispozici sdružení prób do skupin reprezentujících geny, ze kterých by bylo možné vyřazovat nevyhovující měření. Zanedbatelné rozdíly mezi výsledky metod PLIER a IterPLIER v analýze BEST jsou způsobeny náhodností procesu odhadování parametrů.



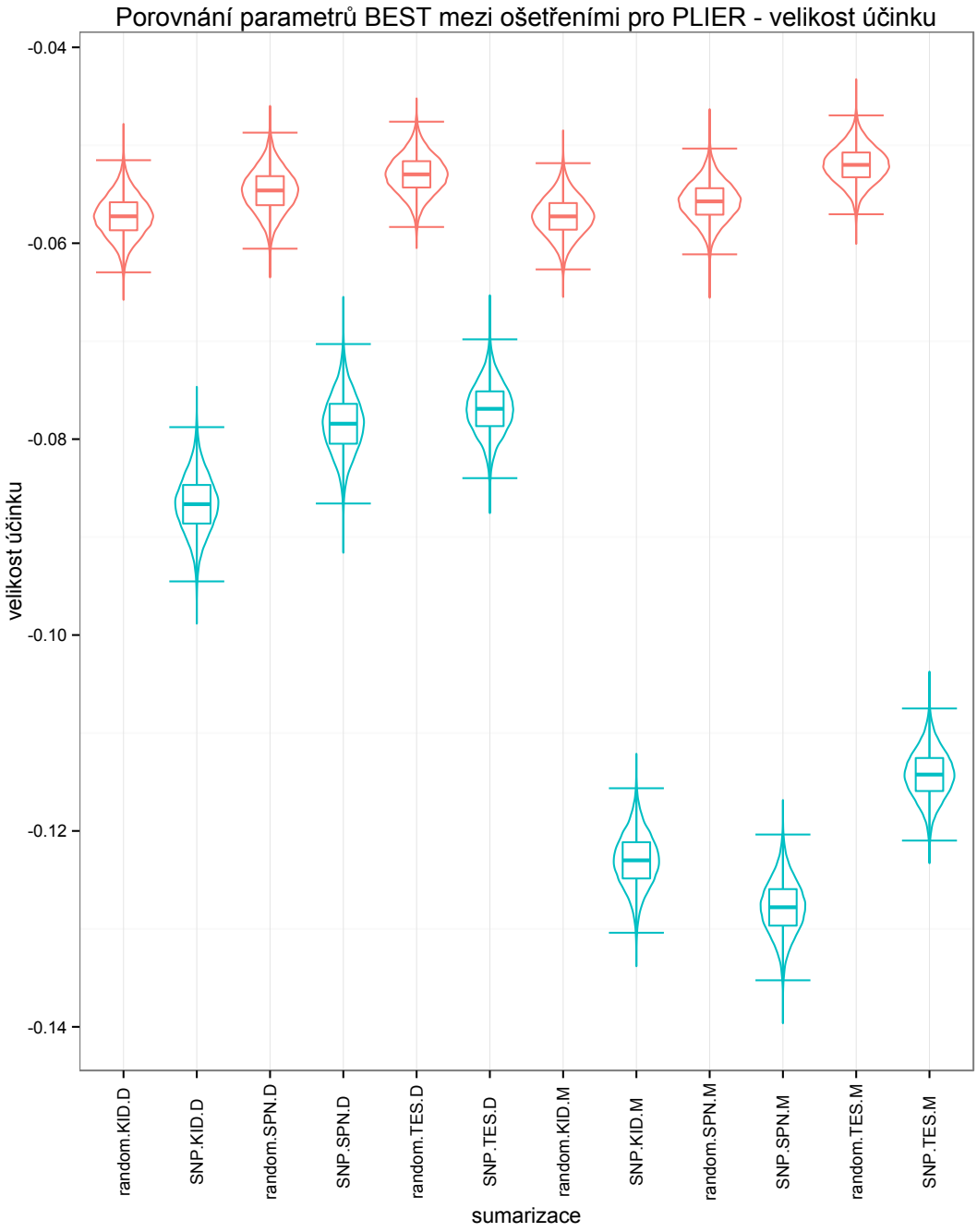
Obrázek 6.4: *Srovnání velikosti účinku (effect size) vyřazení prób pro jednotlivé čipy – příklad pro D KID. Vzhledem k vyřazení prób je variabilita mezi čipy zanedbatelná. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*



Obrázek 6.5: *D1100TES – Zobrazené grafy na levé straně jsou po řadě odhady středních hodnot rozdělení, odhady odchylek a normalita (koeficient t-rozdělení). Na pravé straně jsou histogramy vstupních dat a proložená t-rozdělení. Poslední tři histogramy jsou nejzajímavější pro porovnání vstupních dat – rozdíl středních hodnot, rozdíl odchylek a velikost účinku. Ve všech případech jsou hodnoty v HDI různé od nuly.*



Obrázek 6.6: *gcRMA: Porovnání velikosti účinku (effect size) vyřazení prób při sumarizaci obou poddruhů na exonové úrovni. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*



Obrázek 6.7: *PLIER: Porovnání velikosti účinku (effect size) vyřazení prób při sumarizaci obou poddruhů na exonové úrovni. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*

6.4 Sumarizace na genové úrovni

Rozdíly v odhadech parametrů rozdělení pomocí BEST mezi jednotlivými čipy jsou i v případě sumarizací na genové úrovni zanedbatelné (viz přílohu A.2.3). Další porovnání používají sjednocení odhadů přes odpovídající čipy.

6.4.1 gcRMA

V přehledu dat sumarizovaných pomocí gcRMA na genové úrovni (obr. 6.8) je patrné, že t-rozdělení určené parametry BEST neaproximuje data tak dobře, jako na exonové úrovni. Největšího odklonu od t-rozdělení dosahuje rozdělení dat ze sumarizací vzorků z ledvin poddruhu *M. m. domesticus* při odstranění náhodných prób. Přesto srovnáme odhady parametrů pro odhalení případných tendencí.

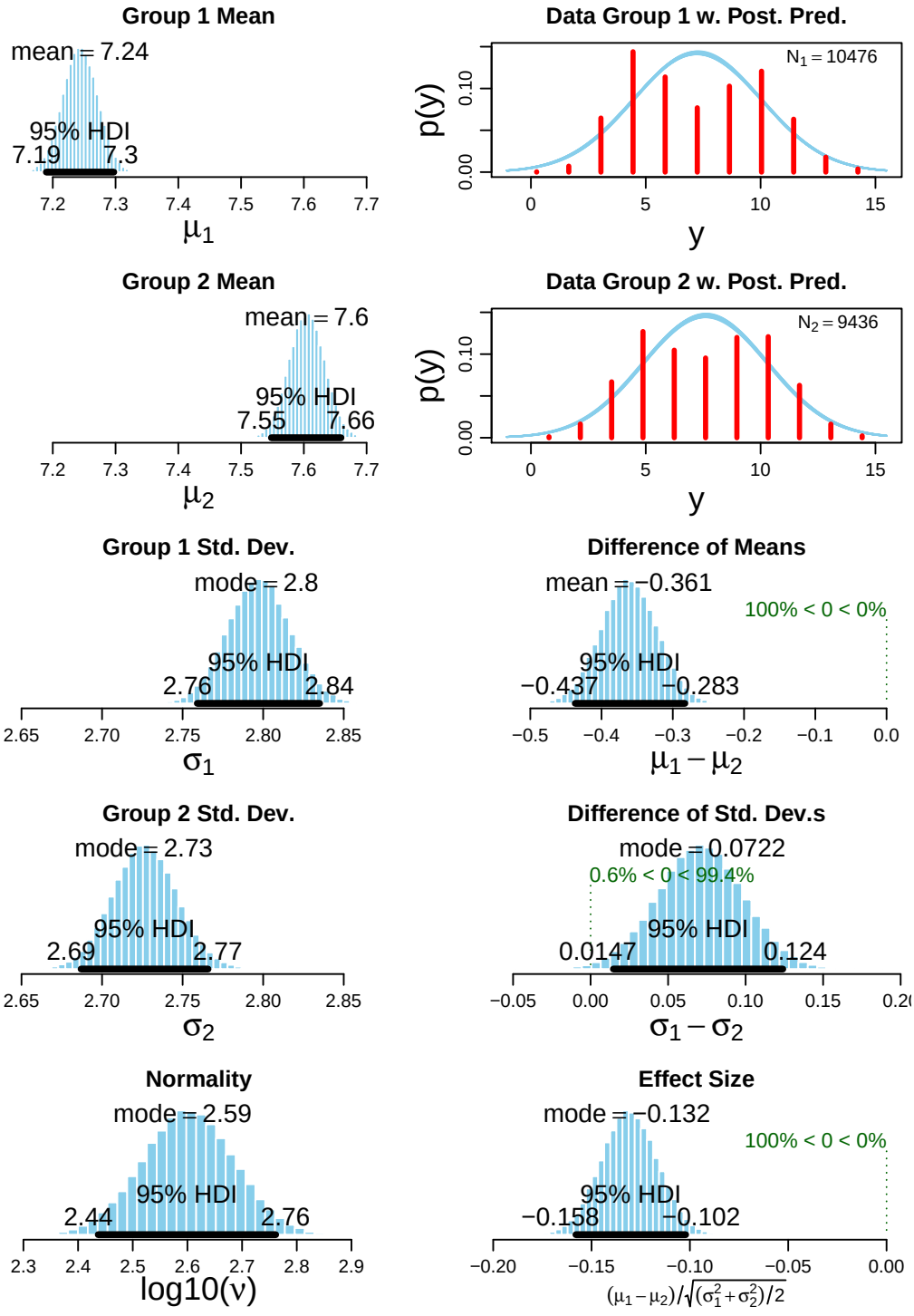
Srovnání velikostí účinku pro gcRMA na genové úrovni je na obrázku 6.9. Vztahy mezi ošetřeními mají stejný směr, jako na exonové úrovni, ale rozdíly jsou menší.

6.4.2 PLIER

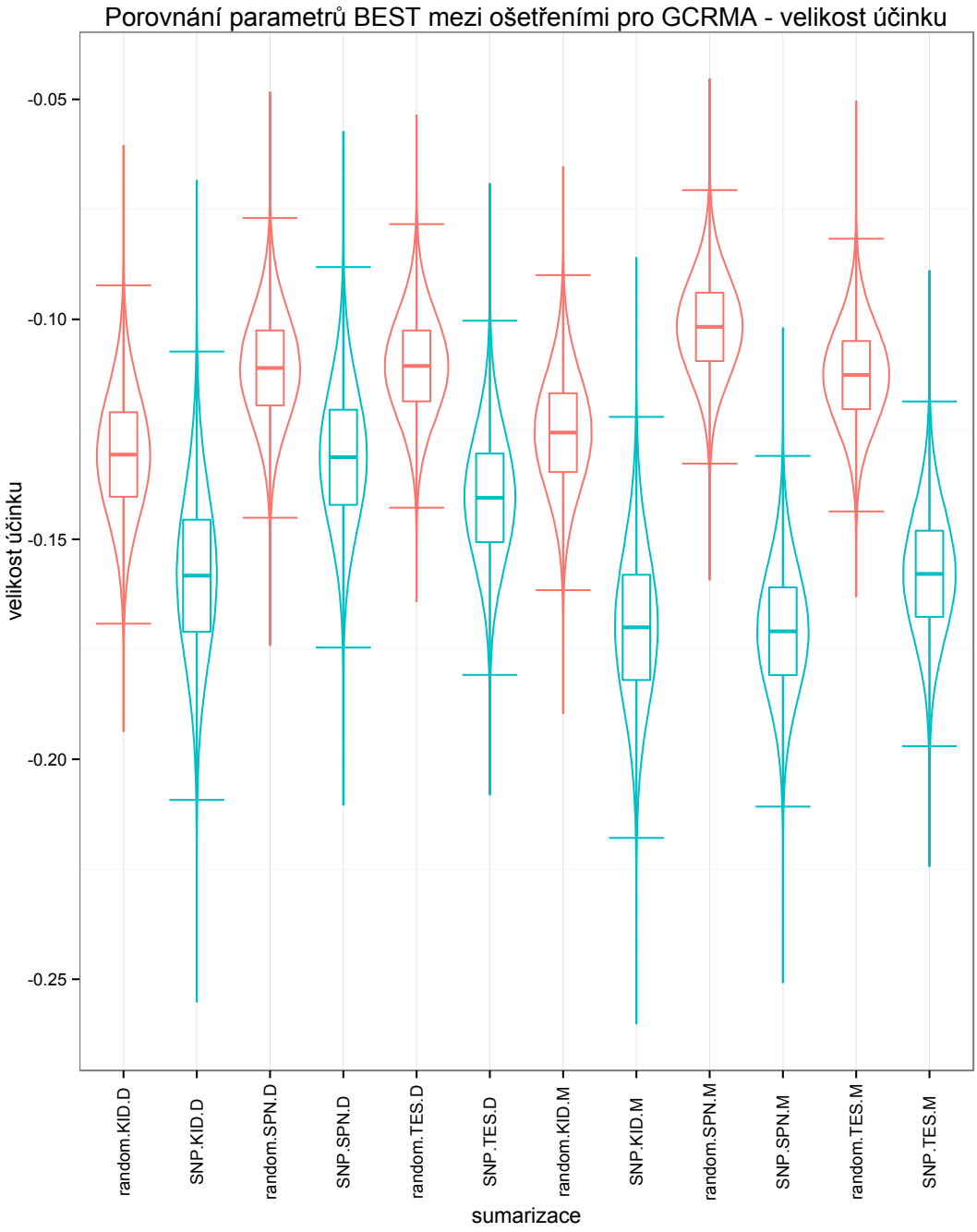
Srovnání velikostí účinku pro sumarizaci metodou PLIER na genové úrovni jsou na obrázku 6.10. Z analýzy je patrné, že velikost účinku obou ošetření je pro většinu tkání srovnatelná. Tento výsledek by mohl znamenat, že metoda PLIER je robustnější ve srovnání s gcRMA, kde jsou rozdíly mezi ošetřeními patrné i na genové úrovni.

6.4.3 IterPLIER

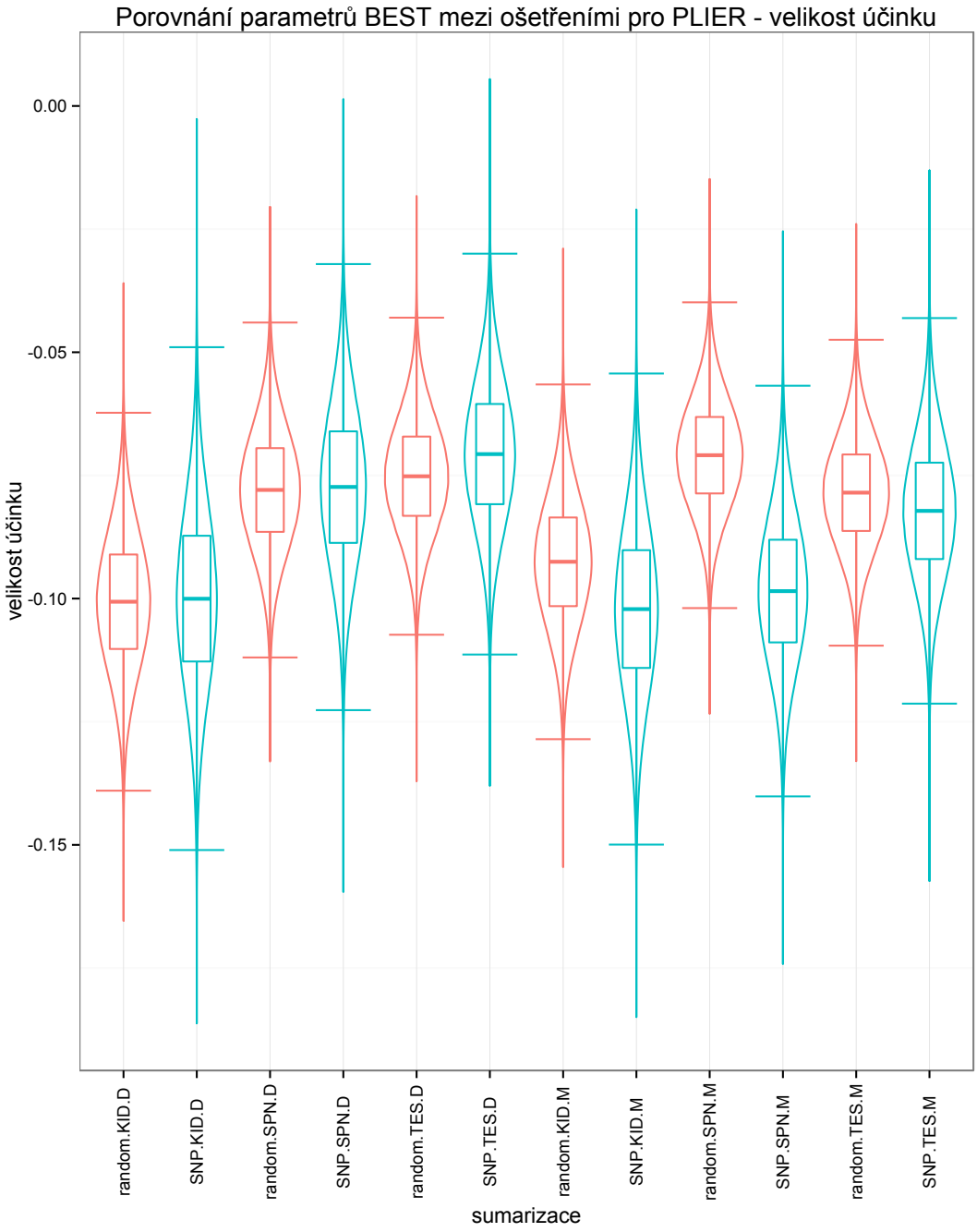
Sumarizační metoda IterPLIER je určena pro sumarizaci expresních čipů na genové úrovni. Celkové výsledky jsou v tomto případě odlišné od sumarizace pomocí metody PLIER, avšak odstranění prób má u obou metod stejný účinek. Srovnání velikostí účinku odstranění prób na sumarizaci IterPLIER je na obrázku 6.11.



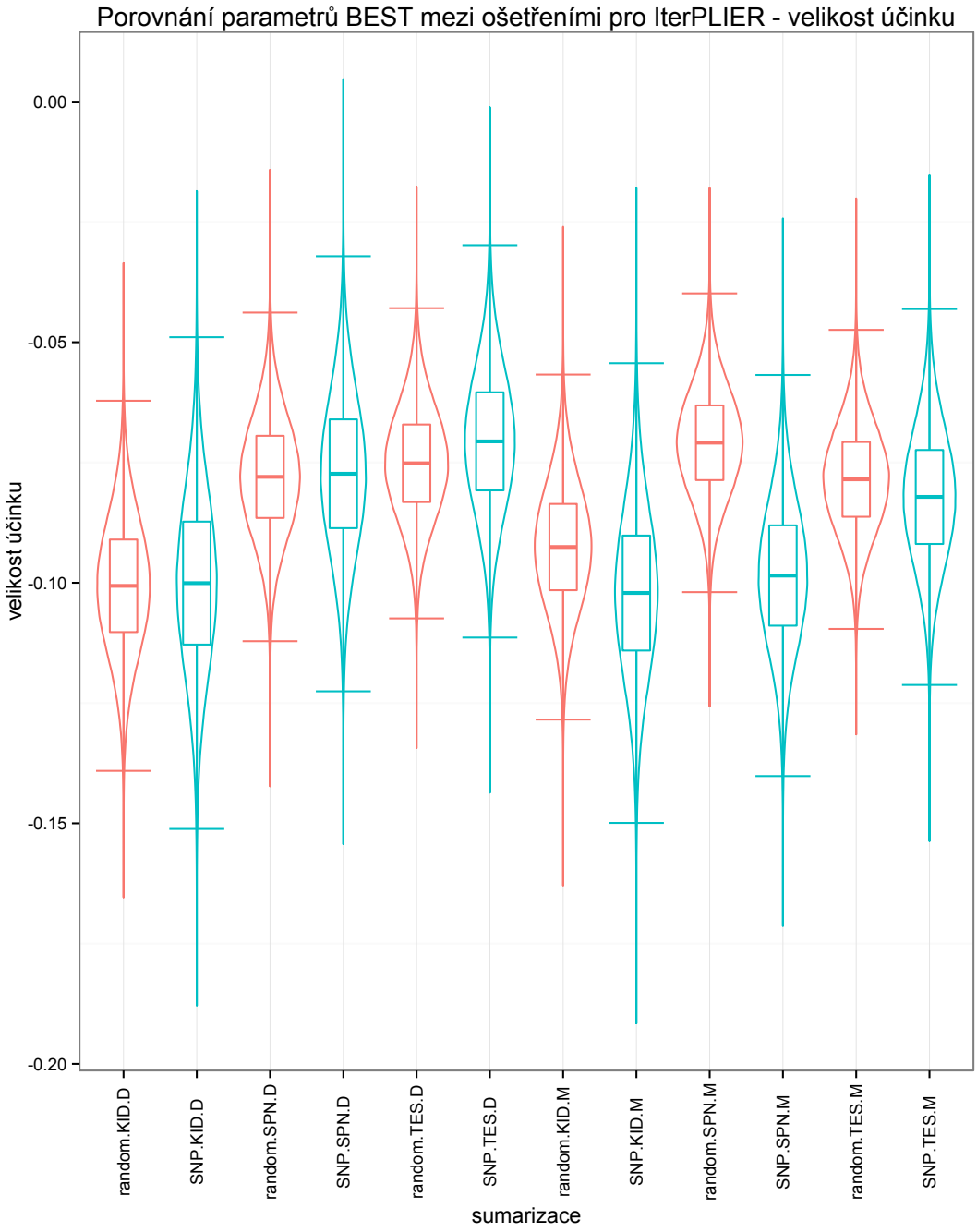
Obrázek 6.8: *D1062KID gcRMA Random – Z histogramů je patrné, že na genové úrovni je aproximace rozdělení dat t-rozdělením méně přesná, než v případě exonové úrovně (zobrazen nejhorší případ).*



Obrázek 6.9: *gcRMA: Porovnání velikosti účinku (effect size) vyřazení prób při sumarizaci obou poddruhů na genové úrovni. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*



Obrázek 6.10: *PLIER: Porovnání velikosti účinku (effect size) vyřazení prób při sumarizaci obou poddruhů na genové úrovni. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*



Obrázek 6.11: *IterPLIER: Porovnání velikosti účinku (effect size) vyřazení prób při sumarizaci obou poddruhů na genové úrovni. Modře je označeno ošetření odstraněním zasažených prób a červeně je označeno ošetření odstraněním náhodných prób.*

Závěr

V této práci byly zkoumány algoritmy gcRMA, PLIER a IterPLIER pro sumarizaci expresních čipů se vzorky z tkání volně žijících jedinců. Cílem bylo zjistit, zda jsou sumarizační metody robustní vůči rozdílnostem designového kmene a volně žijících poddruhů. Pro zvýšení přesnosti sumarizace byl navržen postup odstranění prób ze sumarizace s využitím známých rozdílů v genomu zkoumaných poddruhů.

Využití znalosti SNP

Ukázali jsme, že lze využít znalost rozdílů designových a zkoumaných druhů. Rozdíly ve výsledcích byly patrné především při sumarizaci na exonové úrovni. Pro vylepšení výsledků sumarizace vzorků nemodelových organismů na exonové úrovni má smysl použít metodu odstranění prób ovlivněných známou rozdílností v genomu.

Při sumarizaci na genové úrovni se ukázalo, že sumarizační metody jsou natolik robustní, že odstranění zasažených prób – v porovnání s náhodným vyřazením prób – výsledky výrazně nezlepšilo. V analýze byla zkoumána sumarizovaná data jako celek. Při zkoumání konkrétního genu může mít odstranění zasažených prób značný vliv i v případě, že celková sumarizace nebude výrazně zlepšena.

Na sumarizace PLIER a IterPLIER má v našem případě odstranění prób stejný efekt.

Použití

V této práci byla uvedena obecná metoda, která může vést ke zlepšení výsledků sumarizace dat z exonových čipů při použití vzorků z nemodelových organismů. K práci je přiložena implementace, která byla použita pro ověření metody na konkrétním experimentu. Více o skriptech a použití naleznete v příloze A.

Diskuse

Výsledky naznačují, že v mnoha případech vyřazení prób zasažených SNP zvyšuje sumarizovaný signál, což je ve shodě s hypotézou v kapitole 4.1. Odstranění prób zasažených SNP je možné použít pro zlepšení výsledků při sumarizaci expresních čipů, zejména na exonové úrovni. Zřejmým nedostatkem tohoto postupu je nutnost znát předem odlišnosti genomů referenčního a zkoumaného druhu. Navíc odstranění prób z nedostatečně reprezentovaných exonů může způsobit jejich úplné vyřazení ze sumarizace.

Seznam použité literatury

[1] Miloš Macholán, Stuart JE Baird, Pavel Munclinger, Petra Dufková, Barbora Bímová, and Jaroslav Piálek. Genetic conflict outweighs heterogametic incompatibility in the mouse hybrid zone? *BMC Evolutionary Biology*, 8(1):271, 2008. [cit. 2015-12-01]. Dostupné z: `http://www.biomedcentral.com/1471-2148/8/271`.

[2] Array express database. [cit. 2015-12-01]. Dostupné z: `https://www.ebi.ac.uk/arrayexpress/`.

[3] Xiaoling Zhang, Roby Joehanes, Brian H Chen, Tianxiao Huan, Saixia Ying, Peter J Munson, Andrew D Johnson, Daniel Levy, and Christopher J O'Donnell. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet*, 47(4):345–352, April 2015. Dostupné z: `http://dx.doi.org/10.1038/ng.3220`.

[4] Picture explaining hybridization. [cit. 2015-07-21]. Dostupné z: `https://commons.wikimedia.org/wiki/File:NA_hybrid.svg`.

[5] Affymetrix. Exon Glossary | Affymetrix. [cit. 2015-07-17]. Dostupné z: `http://www.affymetrix.com/support/help/exon_glossary/index.affx`.

[6] Affymetrix. Genechip exon array design. [cit. 2015-07-16]. Dostupné z: `http://media.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf`.

[7] Y. Wang, Z.-H. Miao, Y. Pommier, E. S. Kawasaki, and A. Player. Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. *Bioinformatics*, 23(16):2088–2095, August 2007. Dostupné z: `http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm306`.

[8] Bethesda (MD): National Library of Medicine (US). *The NCBI handbook*. National Center for Biotechnology Information, 2002. Dostupné z: `http://www.ncbi.nlm.nih.gov/books/NBK21101`.

[9] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. GenBank. *Nucleic Acids Research*, 36(Database issue):D25–D30, January 2008. Dostupné z: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238942/`.

[10] MS. Bokuski, TM. Lowe, and CM. Tolstoshev. dbest – database for "expressed sequence tags", 1993. Dostupné z: `http://www.ncbi.nlm.nih.gov/pubmed/8401577`.

[11] C Burge and S Karlin. Prediction of complete gene structures in human genomic dna. *Journal of molecular biology*, 268(1):78—94, April 1997. Dostupné z: `http://dx.doi.org/10.1006/jmbi.1997.0951`.

[12] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle. Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755, January 2014. Dostupné z: `http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1196`.

[13] Jennifer L. Harrow, Charles A. Steward, Adam Frankish, James G. Gilbert, Jose M. Gonzalez, Jane E. Loveland, Jonathan Mudge, Dan Sheppard, Mark Thomas, Stephen Trevanion, and Laurens G. Wilming. The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Research*, 42(D1):D771–D779, January 2014. Dostupné z: `http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1241`.

[14] Jennifer L. Hall, Suzanne Grindle, Xinqiang Han, David Fermin, Soon Park, Yingjie Chen, Robert J. Bache, Ami Mariash, Zhanjun Guan, Sofia Ormaza, and others. Genomic profiling of the human heart before and after mechanical support with a ventricular assist device reveals alterations in vascular signaling networks. *Physiological genomics*, 17(3):283–291, 2004. [cit. 2015-12-01]. Dostupné z: `http://physiolgenomics.physiology.org/content/17/3/283.short`.

[15] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, 9(1):e78644, January 2014. [cit. 2015-12-01]. Dostupné z: `http://dx.plos.org/10.1371/journal.pone.0078644`.

[16] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, February 2003. Dostupné z: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC150247/`.

[17] Zhijin Wu, Rafael A Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–917, December 2004. [cit. 2015-03-22]. Dostupné z: `http://www.tandfonline.com/doi/abs/10.1198/016214504000000683`.

[18] Affymetrix. Guide to probe logarithmic intensity error (plier) estimation. Technical report, Affymetrix, 2005. Dostupné z: `http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf`.

[19] Alan J. Williams, Simon Cawley, John E. Blume, Hui Wang, and Tyson A. Clark. *Gene Signal Estimates from Exon Arrays*. Google Patents, 2012. US Patent 8,170,808 [cit. 2015-03-10]. Dostupné z: http://www.google.com/patents/US8170808.

[20] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001. [cit. 2015-03-23]. Dostupné z: http://www.pnas.org/content/98/1/31.short.

[21] J.D. Emerson and D.C. Hoaglin. Analysis of two-way tables by medians. *Understanding Robust and Exploratory Data Analysis*, pages 165–210, 1983.

[22] Alan J. Williams, Simon Cawley, John E. Blume, Hui Wang, and Tyson A. Clark. *Exon Array Background Correction*. Google Patents, 2005. US Patent 8,170,808 [cit. 2015-03-10]. Dostupné z: https://www.google.com/patents/US8170808.

[23] Terry M. Therneau and Karla V. Ballman. What does PLIER really do? *Cancer informatics*, 6:423, 2008. [cit. 2015-03-09]. Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2623311/.

[24] Wm S. Murray and C. C. Little. The genetics of mammary tumor incidence in mice. *Genetics*, 20(5):466, 1935. Dostupné z: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208626/.

[25] Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellaker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jérôme Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, José Afonso Guerra-Assunçao, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint, and David J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, September 2011. [cit. 2015-06-10]. Dostupné z: http://www.nature.com/doifinder/10.1038/nature10413.

[26] Affymetrix power tools 1.17.0 [software], 2015. [cit. 2015-07-21]. Dostupné z: http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx#1_2.

[27] Affymetrix. *Affymetrix Power Tools: MANUAL: apt-probeset-summarize (1.17.0)*. [cit. 2015-07-19]. Dostupné z: http://media.affymetrix.com/support/developer/powertools/changelog/apt-probeset-summarize.html.

[28] Metacentrum vo. [cit. 2015-07-21]. Dostupné z: http://www.metacentrum.cz/cs/VO/metavo.

[29] John K. Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573–603, 2013. [cit. 2015-11-25]. Dostupné z: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0029146.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. Dostupné z: http://www.R-project.org/.

[31] Tyson A. Clark, Anthony C. Schweitzer, Tina X. Chen, Michelle K. Staples, Gang Lu, Hui Wang, Alan Williams, and John E. Blume. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biology*, 8(4):R64, April 2007. [cit. 2015-12-02]. Dostupné z: http://genomebiology.com/2007/8/4/R64/abstract.

[32] NetCDF: Overview. [cit. 2015-07-21]. Dostupné z: https://www.unidata.ucar.edu/software/netcdf/docs/.

[33] A. Calvin. Package affxparser for bioconductor. 2011. [cit. 2015-06-06]. Dostupné z: ftp://ftp.au.freebsd.org/pub/bioconductor/packages/2.8/bioc/manuals/affxparser/man/affxparser.pdf.

[34] File format: Tsv. [cit. 2015-07-21]. Dostupné z: http://media.affymetrix.com/support/developer/powertools/changelog/file-format-tsv.html.

[35] GFF (General Feature Format) Specifications Document. [cit. 2015-06-09]. Dostupné z: https://www.sanger.ac.uk/resources/software/gff/spec.html.

Seznam obrázků

Seznam použitých zkratek

DNA – deoxyribonukleová kyselina, úložiště genetické informace, nachází se v buněčném jádře (eukaryota)

cDNA – komplementární DNA syntetizovaná z mRNA pomocí enzymu reverzní transkriptázy

RNA, mRNA – ribonukleová kyselina, převod genetické informace z DNA do proteinů

PSR – probe selection region, oblast exonové skupiny určená pro výběr prób

NSB – non-specific binding, nespecifická vazba – hybridizace neúplně komplementárních řetězců

PM – perfect match probe, próba určena k detekci signálu

MM – mismatch probe, próba určena k detekci NSB

RefSeq, GenBank, dbEST – databáze genů

MAS, MAS5 – MicroArray Suite, sumarizační algoritmus firmy Affymetrix

RMA, gcRMA – Robust Multiarray Analysis – sumarizační algoritmus

PLIER – Probe Logarithmic Intensity Error – sumarizační algoritmus

IterPLIER – iterativní vylepšení sumarizačního algoritmu PLIER

DABG – Detection Above Background, algoritmus určující pravděpodobnost, že data zaznamenávají signál

SNP – Single Nucleotid Polymorphism, záměna jedné báze v řetězci DNA za jinou

C57BL/6J – referenční kmen myší

WSB/EiJ – kmen vyšlechtěný z *M. m. domesticus*

PWK/PhJ – kmen vyšlechtěný z *M. m. musculus*

KID – zkrácené označení tkání z ledvin

SPN – zkrácené označení tkání ze sleziny

TES – zkrácené označení tkání z varlat

APT – Affymetrix Power Tools, sada

GFF – soubor obsahující informace o čipu

VCF – soubor obsahující informace o rozdílech sekvencí

CEL – soubor obsahující naměřené intenzity na próbách

CLF, PGF, BGP, QCC, PS, MPS – knihovní soubory pro sumarizaci

BEST – Bayesian Estimation Supersedes the T-test, statistická analýza sumarizovaných dat

MCMC – Markov Chain Monte Carlo, stochastický algoritmus vytvářející velké množství odhadů

HDI – Highest Density Interval, interval do kterého spadá 95% hodnot množiny

A. Skripty

Zde jsou popsány skripty použité v práci, od přípravy dat přes jejich zpracování až po následné vyhodnocení. Skripty byly napsány v jazycích Python, bash a R. Skripty jsou přiloženy k této práci v elektronické podobě, navíc jsou dostupné v repozitáři GitHub[1].

A.1 Sumarizace

A.1.1 Předzpracování dat

Proces předzpracování dat prochází několika kroky, při kterých jsou vytvářeny pomocné soubory a výstupy. Celý proces předzpracování je proveden skriptem s názvem `prepare_data`. Komentáře ve skriptu popisují jednotlivé kroky a jejich vstupní a výstupní soubory. Předzpracování dat probíhá v následujících krocích:

1. zpracování GFF souborů,
2. rozbor atributů ze záznamů v GFF souborech,
3. zpracování VCF souborů,
4. mapování SNP a zasažených prób,
5. agregování ovlivněných probesetů,
6. vytvoření seznamů prób k vyřazení ze sumarizace (tzv. *kill-list*).

Některé kroky jsou implementovány pomocí unixových příkazů, složitější operace jsou prováděny pomocnými skripty v jazyce python. Skript `prepare_data` se nachází v adresáři `scripts` a není určen k přímému spouštění (viz následující část).

Po provedení předzpracování je vše připraveno pro spuštění vlastní sumarizace. Jedním z pomocných souborů vytvořených skriptem v průběhu předzpracování je tzv. `kill-list`, tedy seznam prób zasažených SNP, které mají být vyřazeny ze sumarizace.

A.1.2 Paralelní výpočet

Sumarizaci expresního čipu je možné spustit samostatně, přímo pomocí balíku nástrojů APT. Pro sumarizaci velkého počtu čipů je praktičtější provést paralelní zpracování. V této práci bylo využito služeb MetaCentra, které umožňuje velké množství paralelních výpočtů. Skript, který provede předzpracování dat a následně spustí sumarizace, se jmenuje `launch.sh`. Skript zadává úlohy do fronty pro výpočet v MetaCentru, a to ve všech kombinacích druhů, tkání a sumarizačních metod, které byly v práci použity.

Pro ještě větší paralelizaci je možné rozdělit seznamy probesetů určených k sumarizaci do více částí. Každý částečný seznam je zpracován samostatně, proto je nutné po dokončení sumarizace spojit mezivýsledky výpočtu pro každý čip dohromady. To provádí skript s názvem `join_results.sh`.

[1] Adresa repozitáře: `https://github.com/VojtechTuma/Sumarizace-expresnich-cipu`.

A.2 Analýza

Analýza výsledků sumarizace probíhá v prostředí R. Podobně jako v případě sumarizace je nejprve provedeno předzpracování.

A.2.1 Předzpracování

Předzpracování dat zahrnuje následující kroky:

1. načtení pomocných souborů vytvořených v předzpracování dat před sumarizací,
2. načtení výsledků sumarizace,
3. úprava výsledků získaných metodami PLIER a IterPLIER (dle [18]):

$$x \leftarrow \log_2(x + 16),$$

4. uložení do souborů pro následné zpracování.

A.2.2 Paralelní BEST

Pro výpočet dvoustranného testu BEST, který slouží k porovnání rozdělení dat, je třeba značná výpočetní síla. Z toho důvodu byl proveden paralelně v MetaCentru. Zadávání paralelních úloh pro všechny kombinace sumarizací obstarává skript `launch_mcmc.sh`, vlastní výpočet je prováděn skriptem `mcmc.R`. Výsledkem každého běhu je jeden soubor obsahující vstupní a výstupní data analýzy.

A.2.3 Obrázky

Z výsledků analýzy BEST jsou generovány dva typy obrázků. Příklady jsou uvedeny v textu práce, všechny vygenerované obrázky jsou přiloženy v samostatných souborech. Jde především o tři typy obrázků:

- přehled výsledků analýzy BEST pro jednotlivé čipy a ošetření, jako 6.8,
- violin plot se srovnáním odhadů parametrů (střední hodnoty, odchylky, velikosti účinku) mezi čipy jedné tkáně, jako 6.4,
- violin plot se srovnáním odhadů parametrů mezi tkáněmi a poddruhy, jako 6.6.

B. Souborové formáty

Zde jsou popsány formáty souborů použitých pro sumarizaci a analýzu dat. Některé položky mohou být specifické pro použité knihovní soubory platformy `MoEx-1_0-st-v1`.

B.1 Datové soubory

B.1.1 CEL

Soubory obsahující vlastní experimentálně získaná data. Jde o binární soubory v datovém formátu netCDF [32], které nelze otevřít v textovém režimu – pro přístup k datům slouží různé nástroje, např. *apt-CEL-extract* z balíku nástrojů Affymetrix Power Tools [26] nebo *affxparser* z knihovny Bioconductor pro R[33]. Soubor CEL vzniká zpracováním souboru DAT, ve kterém jsou uloženy intenzity pixelů v obraze skenovaného čipu.

Obsah souboru CEL (příklad):

- jméno souboru, verze (D1062KID.CEL, 1),
- počet sloupců a řádků matice intenzit a celkový počet políček (2560, 2560, celkem 6553600),
- algortimus získávání dat (Feature Extraction Cell Generation),
- parametry zpracování,
- typ čipu (`MoEx-1_0-st-v1`),
- hlavička souboru DAT,
- velikost okrajů buněk – počet pixelů na krajích buněk, které se mají při zpracování vynechat (4, počítá dohromady protilehlé okraje),
- záznamy buněk - hodnota intenzity, standardní odchylka a počet pixelů; hodnoty jsou uloženy po řádcích,
- seznam souřadnic uživatelem maskovaných záznamů,
- seznam souřadnic záznamů označených jako odlehlé hodnoty zpracovávajícím softwarem.

Informace obsažené v samotných CEL souborech nejsou dostatečné pro sumarizaci a agregaci výsledků na vyšší úrovni. Pro analýzu jsou využívány další soubory s doplňujícími informacemi.

B.2 Designové soubory

K expresním čipům jsou dostupné doprovodné soubory popisující sekvence prób, jejich umístění na čipu, unikátní identifikátory, rozdělení do skupin apod. Formát souborů PGF, CLF, BGP, PS a MPS vychází z formátu TSV (tab separated values) verze 1 a verze 2. Formát TSV nezaručuje pořadí sloupců a může obsahovat kromě povinných i další doplňující sloupce. V souboru formátu TSV se mohou vyskytovat komentáře a hlavičky. TSV v2 může mít navíc hierarchickou strukturu záznamů [34].

Soubory založené na formátu TSV, tedy PGF, CLF, BGP, PS a MPS, mají v hlavičce některé údaje společné, příklad je v tabulce B.1.

- typ čipu shodný s typem v CEL souboru (může obsahovat více záznamů, čímž indikuje kompatibilitu s více čipy),
- jméno a verze knihovny souborů pro daný čip (například soubory PGF a CLF určené pro společné použití by měly mít tyto hodnoty shodné),
- verze formátu, datum a další doplňkové informace.

B.2.1 PGF

Soubor formátu PGF definuje typy prób a probesetů a sdružuje PM a MM próby do párů u čipů, které mají MM próby. Informace o sekvencích prób je využívána při výpočtu GC pozadí. Příklad záznamu v souboru PGF je v tabulce B.3.

Informace v hlavičce:

- hlavičky hierarchické struktury.

Data jsou uložena v hierarchické struktuře. Uspořádání má tři úrovně – identifikátor probesetu (případně typ a jméno), *atom* a informace o próbách. Identifikátory probesetů musí být v rámci souboru PGF unikátní.

Atom seskupuje próby zaměřené na stejnou pozici v sekvenci. Pro expresní čipy jsou to obvykle páry PM a MM prób, nebo jen PM próby, jako v případě čipu Mouse Exon 1.0 ST, který neobsahuje párovou MM próbu pro každou PM próbu. Identifikátory atomů musí být unikátní v rámci probesetu i v rámci celého souboru PGF.

Nejnižší úroveň uspořádání dat obsahuje informace o próbách. Próba může být v jednom nebo více probesetech, ale v každém probesetu smí být nejvýše jednou. Identifikátor prób proto nemusí být v rámci souboru PGF unikátní. Doplňující informace bývají typicky obsah G a C bází, délka a zájmová pozice a sekvence próby. Obsah G a C bází je využíván při sumarizaci pro výpočet GC pozadí.

B.2.2 CLF

Soubor CLF (CEL layout file) obsahuje přiřazení identifikátorů prób k pozicím v CEL souboru. Příklad záznamu je v tabulce B.2.

Informace v hlavičce:

- počet řádků a sloupců v CEL souboru,

- hlavičky sloupců – identifikátor próby, souřadnice x a y, případně další sloupce,
- příznak *sequential* – pokud je nastaven na 1, vztah mezi identifikátorem a x-y souřadnicemi je deterministický. Pokud je tento příznak uveden, musí být uveden i parametr uspořádání,
- uspořádání identifikátorů po řádcích, nebo po sloupcích. Pokud je tento parametr uveden, musí být nastaven příznak *sequential*.

Při zpracování pomocí balíku nástrojů APT je nutné při použití PGF souboru přidat CLF soubor pro konverzi identifikátorů na souřadnice v CEL matici.

B.2.3 BGP

Soubor formátu BGP (background probe file) obsahuje seznam prób určených pro výpočet vazby nespecifického pozadí. Příklad záznamu je v tabulce B.4.

Informace v hlavičce:

- identifikátor próby (povinný sloupec),
- identifikátor a typ probesetu; typ, délka a sekvence próby, obsah GC odpovídá PGF souboru,
- x-y souřadnice próby v CEL souboru.

B.2.4 PS, MPS

Soubory PS (probeset) a MPS (meta-probeset) obsahují seznam probesetů, které mají být sumarizovány. V souboru PS je seznam tvořen identifikátory probesetů shodnými s identifikátory v ostatních souborech. Probesety určené souborem PS slouží pro sumarizaci na úrovni exonů. Soubor MPS definuje novou množinu probesetů na úrovni genů, každý nový probeset se skládá z probesetů na exonové úrovni.

Informace v hlavičce:

- identifikátor próby (povinný sloupec),
- druh organismu, jehož genom je reprezentován seznamem,
- verze genomu a verze v genomických databázích, datum sestavení genomu.

Soubor PS obsahuje sloupec s probesety. Soubor MPS obsahuje sloupec s definovaným identifikátorem probesetu na úrovni, genů následovaný identifikátorem transkriptové skupiny a seznamem probesetů na exonové úrovni a celkovým počtem prób v nich obsažených.

K dispozici je několik souborů PS a MPS, ve kterých jsou uloženy probesety dosahující určité úrovně spolehlivosti jejich určení (viz 1.2) [5]. Uživatel může zvolit, zda chce spolehlivější data, nebo větší množství dat. Probesety v souboru MPS jsou navíc vybírány tak, aby jejich próby byly vhodné pro zpracování na genové úrovni (např. jejich sekvence je jednoznačně spárovatelná s genem, který mají reprezentovat).

B.2.5 GFF

Soubor GFF obsahuje informace o genech a sekvencích. Formát GFF existuje ve více variantách, zde je použit standard GFF verze 2 [35]. Soubor GFF je rozdělen do sloupců oddělených tabulátorem, přičemž poslední sloupec může být rozdělen do dalších záznamů ve tvaru „`vlastnost hodnota;`“, k jedné vlastnosti může náležet jedna nebo více hodnot. Záznamy v posledním sloupci se mohou mezi řádky souboru lišit, a to obsahem i pořadím.

Povinné sloupce souboru GFF:

- název sekvence,
- zdroj znaku (program, který dělá predikci, anotace z databáze, experimentální ověření atd.),
- znak (feature),
- začátek a konec sekvence,
- skóre (pokud není hodnota dostupná, je použit znak „.“),
- vlákno (strand), nabývá hodnot „+“, „–“, nebo „.“ pokud není relevantní,
- fáze kodónu (frame / phase) značí posun začátku kodónu proti začátku sekvence (nabývá hodnot „0“, „1“, „2“, nebo „.“ pokud není relevantní).

B.3 Variace genomu myší

V této práci byla navržena metoda zlepšení sumarizace čipů se vzorky z nemodelových organismů vyřazením prób zasažených SNP. Zdrojem informací o SNP je soubor VCF s daty ze studie (Keane et al. 2001 [25]).

B.3.1 VCF

Soubor VCF obsahuje záznamy o rozdílech cílových sekvencí proti referenční sekvenci. První sloupce v souboru určují chromozom a pozici, na které se nalézá zaznamenaná změna. Sloupec ID obsahuje *rs* identifikátor změny v databázi dbSNP. Sloupce REF a ALT obsahují referenční a alternativní bázi, které jsou zaměněny. Alternativních bází může být více, hodnoty pro jednotlivé genomy určují, která z alternativních bází je pro daný genom platná.

Informace v hlavičce definují složky formátu dat; v příslušném formátovacím sloupci je uvedeno pořadí, v jakém jsou hodnoty uloženy v následujících sloupcích. Ukázka záznamu v souboru VCF je v tabulce B.5. Soubor obsahuje informace o velkém počtu kmenů, pro přehlednost jsou zobrazeny dva.

Tabulka B.1: Příklad společných údajů v hlavičkách knihovních souborů.

```
#%chip_type=MoEx-1_0-st-v1
#%chip_type=MoEx-1_0-st-ta1
#%lib_set_name=MoEx-1_0-st
#%lib_set_version=r2
#%create_date=Tue Sep 19 15:18:05 PDT 2006
#%guid=0000008635-1158704285-0293007264-1099755720-0690549889
```

Tabulka B.2: Příklad záznamu v souboru CLF

```
#%clf_format_version=1.0
#%rows=2560
#%cols=2560
#%sequential=1
#%order=col_major
#%header0=  probe_id  x  y
            1         0  0
            2         1  0
            3         2  0
            4         3  0
            5         4  0
            6         5  0
```

Tabulka B.3: Příklad hierarchického záznamu v souboru PGF

```
#%pgf_format_version=1.0
#%header0=  probeset_id  type
#%header1=                    atom_id
#%header2=                             probe_id  type   gc_count  probe_length  interrogation_position  probe_sequence
            4304920      main
                          1
                                       5994477   pm:st  10        25            13                      GTAACCTATTTCTACATGGACCTTG
                          2
                                       2030429   pm:st  10        25            13                      GACCTACCGTAACCTATTTCTACAT
                          3
                                       4309592   pm:st  11        25            13                      CCGTAACCTATTTCTACATGGACCT
                          4
                                       5663393   pm:st  10        25            13                      TACCGTAACCTATTTCTACATGGAC
```

Tabulka B.4: Příklad záznamu v souboru BGP

probeset_id	probeset_type	atom_id	probe_id	probe_type	gc_count	probe_length	probe_sequence	x	y
4305075	control->bgp->genomic	632	5588238	mm:st	17	25	CCCCTTCTCCGCCCAAGATCTGGCG	2317	2182
4305146	control->bgp->genomic	1025	5798994	mm:st	7	25	TTACTTTTCGTCTTGTCATATAGTT	593	2265
4305327	control->bgp->genomic	1751	3250028	mm:st	15	25	CGCCGGGTTGTCTACAACAGTCTGC	1387	1269

Tabulka B.5: Příklad záznamu v souboru VCF obsahujícího informace o SNP. Pro přehlednost jsou ze souboru zobrazeny informace jen o dvou kmenech myší.

```
##fileformat=VCFv3.3
##FORMAT=DP,1,Integer,"Read Depth"
##FORMAT=GQ,1,Integer,"Genotype Quality"
##FORMAT=GT,1,String,"Genotype"
##INFO=DP,1,Integer,"Total Depth"
##FORMAT=MQ,1,Integer,"Mapping Quality"
##FORMAT=HCG,1,Integer,"High Confidence Genotype. Can be 1(yes) or 0(no) for nonref alleles. It is always -1 for ref homs."
##FORMAT=ATG,1,Integer,"Above Threshold Genotype. Can be 1 (SNP) or 0 (Wildtype Reference) for genotypes above threshold *...*
##INFO=AC,-1,Integer,"Allele count in genotypes"
##INFO=AN,1,Integer,"Total number of alleles in called genotypes"
#CHROM  POS      ID  REF  ALT  QUAL  FILTER  INFO        FORMAT               *...*  PWK                WSB
1       3000054  .   C    T    -1    .       AC=2;AN=34  GT:ATG:MQ:HCG:GQ:DP         0/0:0:59:-1:79:14  0/0:0:60:-1:79:14
1       3000093  .   T    C    -1    .       AC=4;AN=34  GT:ATG:MQ:HCG:GQ:DP         1/1:1:52:1:65:21   0/0:0:60:-1:76:13
1       3000223  .   T    A    -1    .       AC=4;AN=34  GT:ATG:MQ:HCG:GQ:DP         1/1:1:59:1:67:10   0/0:0:60:-1:61:8
1       3000226  .   G    A    -1    .       AC=2;AN=34  GT:ATG:MQ:HCG:GQ:DP         0/0:0:58:-1:58:7   0/0:0:60:-1:58:7
```