Univerzita Karlova v Praze

Přírodovědecká fakulta

Molekulární a buněčná biologie, genetika a virologie



Expresní profilování jednotlivých buněk a jejich analýza

Single cell gene expression profiling and analysis

Dizertační práce

Ing. Vendula Novosadová
(rozená Rusňáková)

Vedoucí práce: Prof. Mikael Kubista, Ph.D.

Praha 2013

Prohlašuji, že tato práce byla vypracována samostatně s využitím řádně citovaných literárních zdrojů, pod vedením vedoucího práce a nebyla využita pro získání obdobného nebo jiného akademického titulu.

V Praze dne 5. listopadu 2013

……

Tato práce byla vypracována za podpory následujících grantů AVOZ50520701, GAČR P303/13/02154S a P303/10/1338; EMBO ASTF 110-2011 a ESF 2426.

Na tomto místě bych ráda poděkovala především mému školiteli Mikaelu Kubistovi za jeho plnou podporu v mém výzkumu a za inspirativní nápady. Bez něj by tato práce nemohla vzniknout.
Dále bych ráda poděkovala Mirce Anděrové za to, že mě dokázala držet při zemi a Andersu. Stahlbergovi za to, že mě naučil vše, co zná o qPCR.
Poděkování patří i mým kolegům z laboratoře za vytvoření přátelské a tvůrčí atmosféry. Jmenovitě děkuji Lukáši Valihrachovi za neskutečnou trpělivost.
Děkuji svým rodičům, kteří mě celý život podporovali v tom, co mě baví a přiměli mě dopsat tuto práci.
V neposlední řadě bych ráda poděkovala svému manželi Petru Novosadovi za veškerou pomoc, kterou mi poskytl při vzniku této práce a v životě vůbec.

Obsah

Obsah 1

Seznam zkratek 3

Souhrn 5

Abstract 6

1. Úvod 7

2. Cíle práce 8

3. Současný stav problematiky 9

3.1. Jednotlivé buňky a jejich význam pro výzkum 9

3.1.1. Stochastická exprese v buňkách 10

3.1.2. Možnosti analýzy jednotlivých buněk 13

3.1.3. RT-qPCR pro jednotlivé buňky 17

3.2. Analýza dat 21

3.2.1. Úprava dat a normalizace 22

3.2.2. Základní statistická analýza 25

3.2.3. Multivariantní analýza 26

3.2.3.1. Analýza hlavních komponent 26

3.2.3.2. Shluková analýza 28

3.2.3.3. Kohonenovy samoorganizující se mapy 29

3.3. Astrocyty 31

3.3.1. Základní rozdělení a morfologie astrocytů 31

3.3.2. Hlavní funkce za fyziologických podmínek 33

3.3.3. Patologie astrocytů 35

4. Materiál a metody 38

4.1. Použitý biologický materiál 38

4.2. Použité metody 38
5. Výsledky 39
5.1. Vypracování metody 39
5.2. Analýza dat 74
5.3. Aplikace získaných poznatků na astrocyty 118
6. Diskuze 148
6.1. Analýza jednotlivých buněk 148
6.1.1. Vliv chybějících dat na analýzu 148
6.1.2. Vizualizace výsledků 150
6.2. Heterogenita astrocytů 152
6.2.1. Odlišná odpověď astrocytů na stresové podmínky 152
6.2.2. Vývoj astrocytů v kůře mozkové 153
6.2.3. Vývoj a přeměna astrocytů po ischemii 155
7. Závěr 158
8. Seznam citované literatury 159
Další publikace autora 168

Seznam zkratek

Aqp	Aquaporin
AMPA	Kyselina(2-amino-3-(3-hydroxy-5-methyl-isoxazol-4-yl))propanová
ANOVA	Analýza rozptylu (Analysis of Variance)
CFP	Modrozeleně svítící protein (Cyan Fluorescent Protein)
Cq	Kvantifikační cyklus (Cycle of quantification)
CNS	Centrální nervová soustava
CSPG4	Chondroitin sulfát proteoglykanu
dsRed	Červeně svítící protein
E	Účinnost (Efficiency)
Eaat	Glutamátový transportér (Excitatory Amino Acid Transporter)
EGFP	Zeleně svítící protein (Enhanced Green Fluorescent Protein)
FACS	Fluorescencí aktivované buněčné třídění (Fluorescence-activated cell sorting)
FISH	Fluorescenční *in situ* hybridizace (Fluorescent *in situ* hybridization)
GFAP	Gliální fibrilární kyselý protein (Glial fibrialary acidic protein)
*Glu*l	Glutamin syntáza
Gria	Glutamátový ionotropní AMPA receptor
Grik	Glutamátový ionotropní AMPA kainátový receptor
Grin	Glutamátový ionotropní NMDA receptor
Hcn	Hyperpolarizací aktivované kationtové kanály (hyperpolarization-activated cation channels)
HR	Silně odpovídající astrocyty (High Response)
IPC	Mezidestičkový kalibrátor (Interplate Calibrator)
K2P	Draselný kanál se dvěmi doménami (two-pore potassium channel), protein
Kcnj	Slabě dovnitř draslík usměrňující kanál (potassium inwardly-rectifying channel)
Kcnk	Draselný kanál se dvěma doménami (two-pore potassium channel)
Kir	Kanál usměrňující dovnitř draselné ionty (inwardly rectifying K^+ channels), protein
KSOM	Kohonenova samoorganizující se mapa (Kohonen self-organizing map)
LR	Slabě odpovídající astrocyty (Low Response)
OGD	Nedostatek kyslíku a glukosy (Oxygen-glucose deprivation)

MIQE	Minimum informací pro publikování dat naměřených pomocí real-time kvantitativní polymerázové řetězové reakce (The Minimum Information for Publication of Quantitative Real-Time PCR Experiments)
miRNA	mikroRNA
mRNA	Mediátorová ribonukleová kyselina (messenger ribonucleic acid)
NMDA	N-metyl-δ-aspartát
oligo(dT)	oligo thymidin difosfát
PCA	Analýza hlavních komponent (Principal component analysis)
Pdgfrα	Receptor α pro růstový faktor odvozený z krevních destiček (Platelet-derived growth factor receptor α)
qPCR	Kvantitativní polymerázová řetězová reakce (Quantitative Polymarese Chain Reaction)
RNA	Ribonukleotidová kyselina (Ribonucleic acid)
RNA-FISH	Fluorescenční *in situ* hybridizace pro sledování RNA
RT-qPCR	Reverzní transkripce a kvantitativní polymerázová řetězová reakce (Reverse transcription quantitative real-time Polymerase Chain Reaction)
S100β	Vápník vázající protein
Snap25	Protein spojený se synaptozómem (synaptosomal-associated protein)
Trpv4	Osmosensitivní receptor (transient receptor potential cation channel)
YFP	Žlutě svítící protein (Yellow Fluorescent Protein)

Souhrn

Buňky jsou základním stavebním kamenem všech vyšších organismů. Pro pochopení komplexity tkání a celých orgánů je nutné porozumět buněčné heterogenitě a chování jednotlivých buněk. I buňky stejného typu mají v daném okamžiku odlišné chování a různé složení mRNA, které je způsobeno stochastickou expresí genů a proteinů. Kvantitativní polymerázová řetězová reakce (qPCR) je jednou z nejcitlivějších metod pro detekci mRNA, přesto se však při analýze jednotlivých buněk získává velké množství chybějících dat. Tento fakt je dán nepřítomností transkriptu nebo jeho malým množstvím, které je pod mezí detekce. Z tohoto důvodu je nutné vytvořit nové postupy pro statistickou analýzu. V této práci je diskutován potenciál genového expresního profilování jednotlivých buněk (single cells) pomocí vysokokapacitního přístroje Biomark se zaměřením na analýzu dat a jejich biologickou interpretaci.

Ukazuje se, že mezi dva klíčové kroky v analýze jednotlivých buněk patří normalizace a doplnění chybějících dat. Jednotlivé buňky nejsou normalizovány referenčním genem, ale jsou normalizovány na počet buněk, ze kterých je exprese měřena. Chybějící údaje jsou nahrazeny hodnotou rovné jedné čtvrtině nejmenšího měřitelného množství transkriptu v buňce. Práce se dále věnuje možnostem vizualizace dat získaných měřením exprese v jednotlivých buňkách, metodám charakterizace těchto buněk a jejich rozdělení do subpopulací. Vypracovaná metodika byla testována na několika biologických systémech. Jako první byl ukázán vývoj drápatky vodní. Dále byla metoda použita pro ohodnocení vlivu karcinogenů na kvalitu myších spermií a následně na vývoj embrya.

V této práci byly jako hlavní biologický systém použity kortikální astrocyty. Byly popsány změny genové exprese iontových kanálů a transportérů účastnících se změny objemu astrocytů a jejich aktivace během vývoje, v dospělosti a po poranění. qPCR data ukázala existenci dvou subpopulací astrocytů u dospělého jedince, které se od sebe liší v hladině exprese genů kódujících draselné a chloridové kanály. Pro objasnění heterogenity astrocytů během vývoje a po lokální cerebrální ischemii byla provedena qPCR a statistická analýza jednotlivých buněk z kortikálních GFAP/EGFP pozitivních buněk. Byly nalezeny tři subpopulace buněk, jak ve vývojovém stádiu, tak po poranění. Byla nalezena silná exprese dosud nepopsaných genů *Snap25, Hcn1, Hcn2 a Hcn3, Grin2a, Gria2, Gria3, Grm5 a Kcna3*, které mohou být použity jako nové markery pro reaktivní astrocyty.

Klíčova slova: qPCR, genové expresní profilování jednotlivých buně, vizualizace dat, astrocyty

Abstract

Cells are the basic units of life. Studying complex tissues and whole organs requires an understanding of cell heterogeneity and responses to stimuli at the single-cell level. Even the cells, which belong to the same cell type, behave differently at a specific moment and contain different amount of mRNA. Quantitative polymerase chain reaction (qPCR) is one the most sensitive methods for the detection of mRNA, however, gene expression profiling in single cells leads to a large amount of missing data due to the fact that the transcript is missing, or is below the level of detection. Therefore, it is necessary to establish a new statistical approach for analysis of single cells. In this thesis the potential of single-cell gene expression profiling using the high throughput instrument Biomark, focusing on data analysis and biological interpretation, is discussed.

Data normalization and handling of missing data are two important steps in data analysis that are performed differently at the single-cell level. Single cells are not normalized by reference genes but the number of cells as a normalizer is applied. Missing data are replaced by value, which is equaled one quarter of transcript amount in the cell. Furthermore it is shown how single-cell gene expression data can be viewed and how subpopulations of cells can be identified and characterized. The methodology has been tested on several biological systems. Firstly the development of *Xenopus laevis* was showed. Furthermore, high throughput qPCR was used for carcinogen effect determination on quality of mouse sperm and afterwards on embryo development.

Cortical astrocytes were used as the main biological systems. The changes in gene expression of ion channels and transporters participating in astrocyte swelling, volume regulation and activation of astrocytes in development, adultness and after injury were described. The qPCR data revealed the existence of two astrocytic subpopulations in adult animals markedly differing in their gene expression levels for K^+ channels, and Cl^- channels. To unravel astrocyte heterogeneity during postnatal development and after focal cerebral ischemia, single cell gene expression profiling in acutely isolated cortical GFAP/EGFP-positive cells from mouse brain was employed. Analyses revealed three subpopulations of astrocytes present within postnatal development and after focal cerebral ischemia. It found a high expression of these genes *Snap25*, *Hcn1*, *Hcn2* and *Hcn3*, *Grin2a*, *Gria2*, *Gria3*, *Grm5* and *Kcna3*, which have never been described before and they are possible new markers of astrogliosis.

Key words: qPCR, single cell gene expression profiling, data visualization, astrocytes

1. Úvod

Analýza jednotlivých buněk se stává důležitým nástrojem pro poznání biologických pochodů v buňkách a tkání. Umožňuje najít vzácné buňky s netypickou expresí, různé podskupiny buněk a vztahy mezi nimi. Na každou buňku se lze dívat mnoha způsoby: jak vypadá, z čeho se skládá, jaké biochemické reakce probíhají uvnitř atd. Jednou z metod, která pomáhá odhalit děje uvnitř buňky, je kvantitativní polymerázová řetězová reakce. Touto metodou je možné zjistit množství mRNA v buňce. Změny v hladině mRNA souvisí s vnějšími změnami a často jsou ukazateli změny stavu v organismu. Zatímco qPCR je již dlouhou dobu součástí diagnostických testů, jeho využití pro analýzu jednotlivých buněk má zatím význam pouze v základním výzkumu. S rozvojem qPCR jednotlivých buněk stoupají požadavky na množství analyzovaných genů a s tím roste i náročnost analýzy.

Tato práce je rozdělena na dvě části. První část se týká optimalizace metody sběru buněk, optimalizace reverzní transkripce, preamplifikace, qPCR a věnuje se zejména metodám statistické analýzy dat, která je klíčová pro pochopení a interpretaci výsledků. Vzhledem k omezenému množství materiálu je každý krok při analýze velmi důležitý a metody musí být optimalizovány s důrazem na co nejvyšší účinnost daného kroku. Data získaná měřením genové exprese v jednotlivých buňkách vyžadují specifický přístup především v normalizaci a zacházení s chybějícími údaji. V této práci jsou představeny různé možnosti normalizace a dále důležitost propojení jednotlivých univariantních a multivariantních analýz s důrazem na biologický význam výsledků.

Druhá část dizertační práce ukazuje uplatnění vypracované metody pro nekultivované astrocyty. Tyto podpůrné buňky neuronů mají význam za fyziologických i patologických podmínek. Práce se zabývá jejich charakterizací se zaměřením na gliové glutamátové transportéry a receptory iontových kanálů. Je ukázána heterogenita astrocytů a jejich různé subpopulace založené na rozdílné expresi především glutamátových receptorů/transportérů, draselných kanálů a markerů pro gliové buňky.

2. Cíle práce

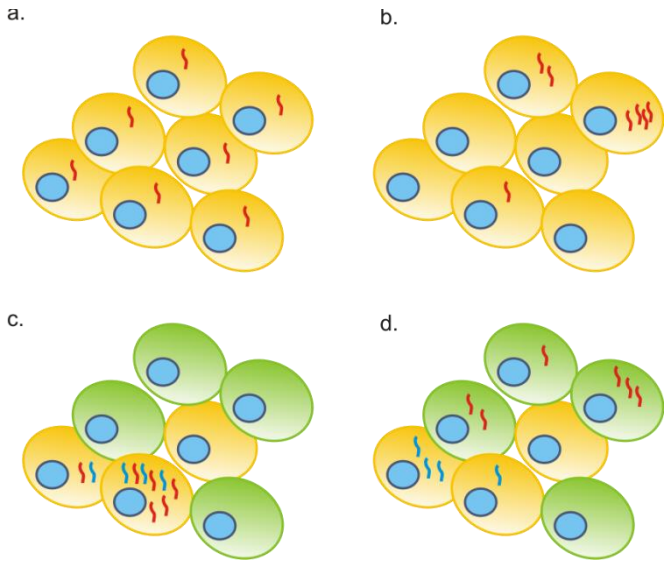
Hlavní cíle této práce lze definovat v následujících bodech:

- Vyvinout metodu pro sběr jednotlivých buněk a následnou analýzu pomocí vysokokapacitní qPCR.
- Vypracovat postup pro statistickou analýzu dat získaných pomocí uvedené metody.
- Aplikovat získané poznatky na různé biologické systémy.
- Popsat a charakterizovat genovou expresi v astrocytech, které byly vybrány jako modelový systém.

3. Současný stav problematiky

3.1. Jednotlivé buňky a jejich význam pro výzkum

Buňky jsou nejmenší jednotkou tkání, orgánů a celého organismu. Přestože jsou všechny buňky v rámci jednoho organismu geneticky téměř identické, liší se v morfologii, ve složení i v biochemických pochodech, které probíhají uvnitř buňky. Heterogenita jednotlivých buněk je známa již od padesátých let minulého století (Novick and Weiner, 1957). V té době ovšem nebylo technicky možné měřit změny ve složení mRNA a jednotlivých proteinů na úrovni jedné buňky. Každá individuální buňka je jedinečná i přesto, že pochází ze stejné buněčné kultury nebo ze stejné části tkáně. Liší se velikostí, množstvím exprimovaného proteinu a množstvím přepsané mRNA. Biologická data sesbíraná z tkáňových vzorků, které se skládají z tisíců buněk, jsou pouze jejich průměrem a nemohou tak vystihnout jedinečnost každé buňky. Není tak možné rozlišit jednotlivé buněčné subtypy, není znám původ jednotlivých transkriptů a jejich rozložení v jednotlivých buňkách (Obrázek 1: Heterogenita buněk), (Stahlberg and Bengtsson, 2010). Analýza jednotlivých buněk umožňuje najít korelace mezi jednotlivými geny a stanovit tak společný molekulární mechanismus jejich přepisu (Volfson et al., 2006).



Obrázek 1: Heterogenita buněk

a. Hypotetický případ, kdy každá buňka obsahuje právě jeden hledaný transkript. **b.** Pravděpodobnější situace, kdy většina buněk neexprimuje žádný transkript a pouze v jedné či dvou je exprese silná. **c.** Pozitivní korelace mezi dvěma geny, které jsou exprimovány současně ve stejné buňce. **d.** Negativní korelace mezi dvěma geny,kdy každý typ buňky (znázorněný různou barvou) exprimuje jiný transkript.

Analýza jednotlivých buněk má rovněž velké uplatnění pro medicínu, např. umožňuje charakterizovat kmenové buňky, které mohou mít velké uplatnění v regeneraci tkání. Je možné zaměřit se pouze na určitý typ buněk a objasnit metabolické dráhy, které vedou k jejich diferenciaci. Další oblastí, kde má analýza jednotlivých buněk velký význam, je charakterizace rakovinných buněk. Nádor je tvořen mnoha typy buněk, z nichž jen některé mají potenciál pro zhoubné dělení a mohou být cílem léčby. Tím, že nebude analyzována celá tkáň, ale jednotlivé buňky, je možno nalézt různá vývojová stádia bujení a lépe tak zacílit léčbu. Slibná budoucnost analýzy jednotlivých buněk je rovněž v diagnostice. Již nyní je možné například detekovat rakovinu prsou a plic pomocí analýzy pouhých pěti buněk (Pesta et al., 2013).

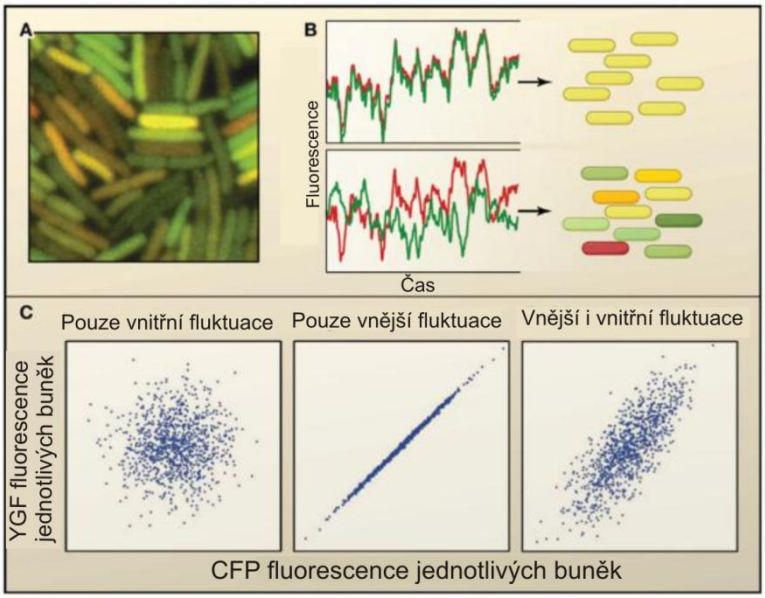
3.1.1. Stochastická exprese v buňkách

Genová exprese je silně stochastický, náhodný proces. Náhodnost v transkripci a translaci vede k rozdílům mezi buňkami na úrovni mRNA i proteinu. Tato různorodost se objevuje ve všech organismech bez ohledu na vývojový stupeň, od bakterií až po savce, a je dána jak biofyzikálními parametry genové exprese, tak i vlastnostmi následné genové sítě. Stochastická genová exprese ovlivňuje buňku v kladném i záporném smyslu. Může hrát roli při odpovědi na stres, v metabolismu, vývoji, buněčném cyklu, stárnutí atd. Je zodpovědná za rozdílnost dvou genově identických organismů (Raj and van Oudenaarden, 2008).

První experimenty zjišťující rozdílnost exprese v jednotlivých buňkách se datují do roku 1957, kdy bylo ukázáno, že produkce β-galaktosidázy je v buňkách různá a náhodná. Indukce laktózou zvýšila počet buněk produkujících β-galaktosidázu, ale nezvýšila expresi na stejnou úroveň ve všech buňkách (Novick and Weiner, 1957). V roce 1957 ještě nebylo možné měřit přímo genovou expresi, první práce na toto téma vznikla až v roce 1990, kdy byl zjišťován efekt různých dávek glukokortikoidu na expresi β-galaktosidázy a bylo potvrzeno předchozí pozorování (Ko et al., 1990).

S částečným vysvětlením stochasticity v expresi přišel koncept vnitřního a vnějšího šumu pozorovaný v bakteriích *E. coli* (Elowitz et al., 2002). V této práci byla kvantifikována variabilita v expresi dvou fluorescenčních proteinů, modrozeleného (CFP) a žlutého (YFP), pro které byly geny pod stejným promotorem vloženy do bakterie. Po následné kultivaci byly pozorovány bakterie s různou barvou a různou intenzitou zbarvení (Obrázek 2A). Tzn., že v každém organismu probíhala exprese fluorescenčních proteinů jinak, odlišná barva je dána kombinací CFP a YFP, zatímco intenzita jejich množstvím. Odlišnost v zabarvení se vysvětluje vlivem vnější a vnitřní fluktuace. Vnější fluktuace jsou ty, které ovlivní expresi obou proteinů stejně, např. množství RNA polymerázy nebo ribozomů. Vnitřní

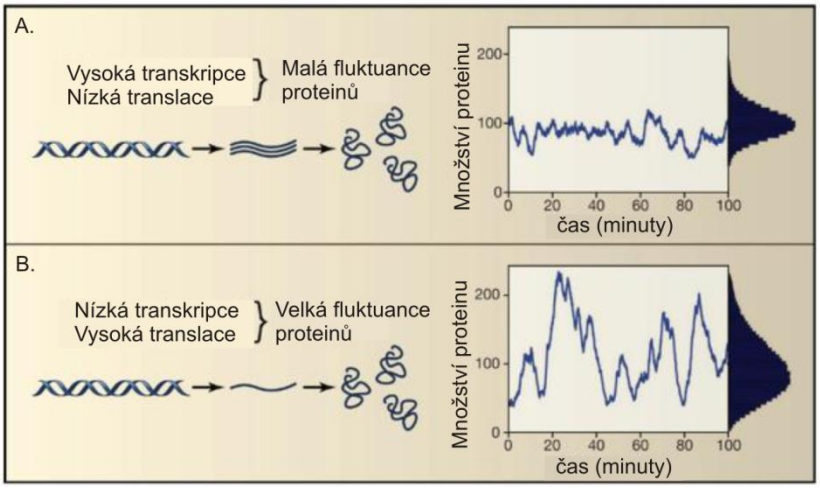
fluktuace jsou ty, které díky náhodnosti ovlivňují každý gen zvlášť, jsou nezávislé a nekorelované. Bylo ukázáno, že oba typy šumu závisí na promotoru. Pokud by byla stochasticita genové exprese závislá pouze na vnějším šumu, měly by bakterie stejnou barvu – zelenou. Ta by se lišila pouze v intenzitě (Obrázek 2B). V případě pouze vnitřního šumu by nebyla pozorována vůbec žádná souvislost mezi intenzitou modrozelené a žluté (Elowitz et al., 2002). Při zbarvení bakterií byla pozorována mírná korelace mezi CFP a YFP, poukazující na vliv jak vnější tak vnitřní fluktuace.



Obrázek 2: Ukázka vlivu vnitřního a vnějšího šumu

A. *E. coli* s vloženými plasmidy pro YGF a CFP. **B.** Pouze vnější fluktuace (horní obrázek), všechny bakterie mají v čase stejné zbarvení, které se liší intenzitou. Pouze vnitřní fluktuace (dolní obrázek), bakterie se liší intenzitou zbarvení i barvou zcela náhodně. **C.** Matematické znázornění pouze vnitřní fluktuace (pravý obrázek), pouze vnější fluktuace (prostřední obrázek) a vnitřní i vnější fluktuace zároveň (pravý obrázek). Převzato z Raj and van Oudenaarden (2008).

Variabilita v expresi proteinů záleží na změně v transkripci a translaci. Šum (měřený jako standardní odchylka dělená průměrem) může nepřímo záviset na rychlosti transkripce, ale nikoliv na translaci. Předpokládá se, že proteiny vznikají v translačních vzplanutích z jednotlivých transkriptů. V případě, že je intenzita transkripce vysoká, je variabilita v translaci malá, zatímco když je intenzita nízká, dochází ke zvýšení šumu v translaci, přestože průměrně zůstává translace stejná (Obrázek 3), (Ozbudak et al., 2002).



Obrázek 3: Vliv intenzity transkripce na translaci

A. V případě vysoké transkripce jsou všechny transkripty přepisovány a nedochází k proteinovým vzplanutím, množství proteinu je konstantní v čase. **B.** V případě nízké transkripce, je šum v translaci mnohonásobně vyšší, přestože zůstává průměrná translace stejná. Převzato z Ozbudak et al., (2002).

Při přímém pozorování transkripce v *E. coli*, kdy na vznikající mRNA byl navázán GFP, bylo zjištěno, že nejen proteiny, ale i mRNA je přepisována v transkripčních vzplanutích, jako kdyby byl gen náhodně přepínán mezi aktivním a neaktivním stavem (Golding et al., 2005). Stejný model byl navržen i pro eukaryotickou expresi. Při pozorování genové exprese ve vyšších eukaryotech se ukázalo, že fluktuace je zde ještě větší než v bakteriích a u jednobuněčných organismů. mRNA detekce v *Dictyostelium discoideum* potvrdila „teorii vzplanutí“, ale v porovnání s bakteriemi jsou transkripční vzplanutí méně častá a intenzivnější (Chubb et al., 2006).

Předpokládá se, že zde musí být nějaký regulátor transkripce zodpovědný za stochastické chování. Jako nejpravděpodobnější regulátor se jeví remodelace chromatinu. Když je chromatin v acetylovaném, tj. otevřeném stavu, gen je transkripčně aktivní, zatímco v kondenzovaném stavu je transkripce potlačena. Pro tuto teorii však zatím neexistují přímé důkazy (Raj et al., 2006; Xu et al., 2006; Levesque and Raj, 2013). Dalším regulátorem transkripce by mohla být existence preiniciačního komplexu na promotoru. Pokud by existoval pouze krátkou dobu, probíhala by transkripce „trhaně“. Transkripce neprobíhá po celém genomu rovnoměrně, ale pouze na určitých místech v tzv. transkripčních továrnách, ve kterých jsou aktivní geny přepisovány. Ukazuje se, že transkripční továrny, ve kterých se nachází velké množství genů, mají v jádře stálé místo a nepohybují se. Dío místa přepisu naopak cestují odhalené geny. (Osborne et al., 2004). Kompetice o místo v těchto „transkripčních továrnách“ může vést k náhodnému

přepisu genů a k velké fluktuaci v přepisu. Vzhledem k tomu, že transkripce a translace je v eukaryotech oddělený proces, není vliv transkripčních vzplanutí na množství proteinu tak velký jako u bakterií. Mimo odlišného místa vzniku může být dalším vysvětlením i delší stabilita proteinů. Platí, že proteiny s krátkou dobou životnostikopírují množství mRNA v buňce, ale jestliže degradují pomalu, fluktuace proteinů mizí a nezávisí na množství mRNA (Raj et al., 2006).

Náhodná exprese genů má velký význam pro heterogenitu bakterií, kvasinek a virů. Vzniklá variabilita jim umožnuje přizpůsobit se novým podmínkám, kdy náhodný vznik lépe fungujícího proteinu poskytuje konkurenční výhodu. Oproti tomu u vyšších eukaryot je diverzita organismů způsobena prostředím a změnou genomu. Stochastická exprese mRNA by neměla ovlivnit translaci, protože během vývoje je vznik určitých proteinů klíčový pro vytvoření tkání. Přesto se najde příklad, kde se náhodná exprese genů u vyšších eukaryot uplatní. Např. u myší během vývoje čichu vzniká tisíce odlišných čichových receptorů, každý je exprimován samostatně v čichových smyslových neuronech. Vlastnosti každého receptoru a jeho umístění na neuronu je náhodné, neexistuje žádné pravidlo a každé zvíře je odlišné od ostatních (Vassar et al., 1993). Na stejném náhodném principu funguje i vznik fotoreceptorů pro *Drosophilu* (Wernet et al., 2006).

Pro úplné objasnění transkripčních stochastických pochodů bude nutné vyvinout nové techniky analýzy pro jednotlivé buňky a propojit je se stávajícími.

3.1.2. Možnosti analýzy jednotlivých buněk

Analýza jednotlivých buněk může být rozdělena podle stavu buňky na *in vivo* a *in vitro* analýzu. Do *in vivo* analýzy řadíme především průtokovou cytometrii a mikroskopii. *In vitro* analýza v sobě nyní již zahrnuje téměř všechny metody používané pro analýzu tkání, jako jsou analýzy genomu (genomika), transkriptomu (transkriptomika), proteomu (proteomika), lipidomu (lipidomika) atd. Všechny „omiky“ usilují o identifikaci, kvantifikaci a charakterizaci všech součástí buňky včetně jejich umístění. Hlavní technologie „omik“ zahrnují DNA a RNA sekvenování, měření mikroarray, qPCR, hmotnostní spektrometrii a proteinové array (Wang and Bodovitz, 2010), (Tabulka I).

První možností pro analýzu jednotlivých buněk byl mikroskop. Nejprve byly pozorovány jednotlivé organely. S rozvojem možností vytvářet geneticky modifikované proteiny a vnášet fluorescenční značky přímo na vzniklé proteiny začalo být možné sledovat jejich přesné umístění a pohyb v buňce. Jako první byl sledován vznik mRNA *in vivo* v kvasinkách. Do sledovaného genu byla vložena vlásenková struktura z

nukleotidů, na kterou se váže fágový protein MS2 do příslušného genu a zároveň byl do kvasinek vložen plasmid pro expresi fúzního proteinu GFP-MS2. Po translaci se GFP-MS2 navázalo na vlásenkovou strukturu a tak bylo možno sledovat vznik a zánik mRNA v reálném čase (Bertrand et al., 1998). Pomocí fluorescenční *in situ* hybridizace (FISH) je možné sledovat chromozomy v dělících se buňkách, kdy je nukleotidová próba s fluorescenční barvou navržena tak, že obsahuje komplementární nukleotidy pouze pro jedno specifické místo na daném chromozomu. Po jejím navázání lze přesně určit místo a pohyb chromozomu v dělící se buňce. Stejným způsobem byla vyvinuta i technika RNA-FISH, kdy je nukleotidová próba komplementární k přepsané mRNA. Tyto metody se používají již delší dobu především při studiu nádorů (Tsuchiya, 2011). Teprve nedávno byla použita RNA-FISH na nefixovaných buňkách. Díky této metodě, mohly být přímo počítány mRNA molekuly v okamžiku svého vzniku, což umožnilo ukázat odlišnost mezi jednotlivými buňkami ze stejné buněčné linie a zachytit tak proces transkripčního vzplanutí. (Raj et al., 2006), (Kapitola 3.1.1).

Průtoková cytometrie a FACS patří k nejstarším typům analýz jednotlivých buněk. Během hodiny je možné analyzovat desítky tisíc buněk a měřit na nich až 18 povrchových proteinů (Chattopadhyay and Roederer, 2012). Výhodou je, že buňky zůstávají živé a mohou být použity pro další experimenty. Během měření je možné vytřídit pouze požadovanou skupinu buněk a tu analyzovat dalšími metodami. První experimenty v 80. letech umožnily např. rozlišit různé typy krvinek (Koziner et al., 1982) Později byly dokonce pomocí těchto metod charakterizovány jejich biochemické dráhy (Sachs et al., 2005). Další aplikací je analýza karyotypu. Použitím barev vázajících se na DNA je možné analyzovat jádra jednotlivých buněk a nalézt genetické abnormality jako jsou chromozomální delece a aneuploidie. (Trask, 2002). Dále je možné určit množství buněk v určitém buněčném cyklu pomocí značení syntetickým nukleosidem (BrdU). Na rozdíl od mikroskopických metod je možné analyzovat stovky tisíc buněk ve velmi krátkém čase.

Analýza genomu jednotlivých buněk patří mezi technicky nejnáročnější experimenty. Množství DNA v buňce se pohybuje v řádech femtogramů a musí být proto preamplifikováno. Poté následuje sekvenace. Vzhledem k vysokým finančním nákladům při analýze desítek buněk se obvykle nesekvenuje celý genom, ale jen jeho vybrané části. Nalezené rozdíly mezi buňkami, například v množství krátkých nekódujících tandemů, umožní určit vývoj tkání a nádorů (Frumkin et al., 2008). Další aplikací je sekvenace jednotlivých chromozomů z jedné buňky v mikrofluidním zařízení. To umožní studium jednotlivých alel nezávisle na sobě a získání přesné informace o haplotypech buňky (Fan et al., 2011).

Pro analýzu množství transkriptů je možné použít qPCR (Kapitola 3.1.3), měření na mikroarray a RNA sekvenování. Mikroarray umožní měření tisíce genů. Krátké specifické próby se váží na fluorescenčně značený vzorek a měří se síla signálu, která odráží množství transkriptů. Tato metoda je však pouze „semi-kvantitativní“, tzn. že její pomocí nelze určit přesné množství transkriptů ve vzorku. Její výhodou je velké množství genů, které je možné detekovat. Je možné měřit nejen jednotlivé transkripty, ale i nekódující RNA a miRNA. Byly již publikovány práce pro měření celého transkriptomu (Mock et al., 2008). Pro měření je typicky potřeba 1-2 μg mRNA, což odpovídá množství mRNA v 10^6-10^7 buňkách, proto je nutné mRNA pro měření mikroarrayí z jednotlivých buněk preamplifikovat (Kapitola 3.2.1), (Kurimoto et al., 2007).

RNA sekvenování celého transkriptomu (next-generation sequencing) umožní získat informace o nových genech, nových sestřihových variantách genů, mutacích, kódujících i nekódujících RNA. Stejně jako mikroarray je nutné pro RNA sekvenování celý transkriptom preamplifikovat, což může vnášet do reakce nepřesnosti. Nalezené nové transkripty je tedy nutné ověřit ještě další metodou. Obecně pro nalezení nového genu/transkriptu mezi již známými je nutné provést až 40 milionů reakcí. Oproti výše zmíněné technice mikroarray je RNA sekvenování kvantitativní metoda. První práce na sekvenaci RNA z jednotlivých buněk se začaly objevovat teprve nedávno (Tang et al., 2009) a stále se pracuje na vývoji plně optimalizovaného protokolu, který by mohl být považován za standard.

Poslední zde uvedená metoda je qPCR. Ta sice neumožní, ve srovnání s mikroarrayí a RNA sekvenováním, analyzovat tolik genů najednou, ale jedná se o plně optimalizovanou metodu s možností přesné kvantifikace jednotlivých transkriptů. Při použití vysokokapacitních přístrojů lze analyzovat až 96 genů a 96 vzorků najednou. Tato metoda je používána při hledání jemnějších rozdílů mezi buňkami a je jí věnována následující kapitola.

Technologie	Počet buněk	Počet genů	Způsob detekce	Detekovaný material / možnosti a omezení
Dynamická proteomika/ MS2 značení	100	1-2	Fluorescenčně značené proteiny	Živé buňky Geneticky upravené buněčné linie/ Rozlišení v čase
mRNA FISH	100	3-4	Fluorescenčně značené mRNA molekuly	Fixované tkáně nebo buňky
Průtoková cytometrie	Tisíce	Až 18	Obecná buněčná charakteristika fluorescenčně značené protilátky navázané na proteiny	Živé buňky Disociované tkáně Hlavně membránové proteiny
Analýza jednotlivých buněk pomocí qPCR	1-100	Až 10	Cq hodnoty udávají původní množství mRNA	Lyzované buňky/ Nutné velké množství reakcí
Vysokokapacitní qPCR	96 na čip	96 na čip	Cq hodnoty udávají původní množství mRNA	Lyzované buňky/ Nutná preamplifikace
Array pro jednotlivé buňky	1-10	20 000	Oligonukleotidové próby hybridizují na příslušný gen	Lyzované buňky/ Možnost detekce nových sestřihových variant Nutná preamplifikace
Sekvenování jednotlivých buněk	1-10	20 000	Převod genomu/transkriptomu do krátkých oligonukleotidů, jejich ligace a následné sekvenování	Lyzované buňky/ Možnost detekce nových sestřihových variant a nových genů Možnost multiplexování vzorků pomocí značení kódy Nutná preamplifikace

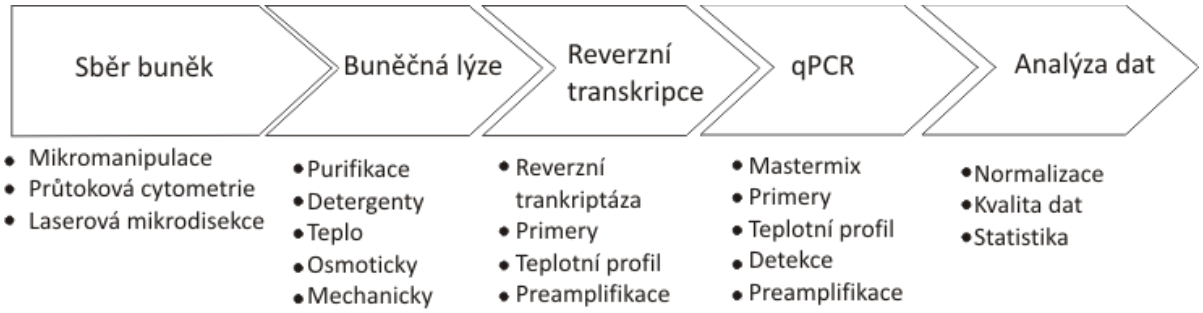
Tabulka I: Shrnutí technik pro analýzu jednotlivých buněk.

Každá technologie je charakterizována typem vstupního materiálu, způsobem detekce, možnostmi a omezením dané metody. Převzato z Kalisky et al., 2011.

3.1.3. RT-qPCR pro jednotlivé buňky

Jedna buňka obsahuje zhruba 1 pg mRNA, což odpovídá několika stovkám tisíců molekul přepsaných z asi deseti tisíců genů (Kuznetsov et al., 2002). Vysoká citlivost reverzní transkripce a qPCR teoreticky umožní detekovat i pouze jedinou molekulu.

Analýza jednotlivých buněk pomocí RT-qPCR se skládá z několika kroků (Obrázek 4).



Obrázek 4: Postup pro analýzu jednotlivých buněk pomocí RT-qPCR

Převzato ze Stahlberg a Bengtsson, 2010.

Prvním krokem RT-qPCR analýzy je sběr buněk. V současné době existují tři hlavní přístupy, a to pomocí průtokové cytometrie, mikromanipulátoru a laserové mikrodisekce. První dva postupy umožňují sběr živých buněk z rozdisociované tkáně. Buňky se sbírají přímo do lyzačního pufru v jamkách mikrotitrační destičky tak, aby bylo možné provést reverzní transkripci pouze přidáním potřebných reagencií. Výhodou průtokové cytometrie je možnost sesbírání několika stovek buněk ve velmi krátkém čase. Při použití fluorescencí aktivovaného buněčného třídění (FACS) je možné vybrat buňky určitého typu, které mohou být označeny buď přímo fluorescenční protilátkou, nebo se může jednat o transgenní organismus s vloženým genem pro fluorescenční protein (EGFP, YFP, dsRed). Tento protein má před svou sekvencí vložen stejný promotor jako má typický marker buňky (Brenner and Messing, 1996; Nolte et al., 2001; Rivers et al., 2008). Předpokládá se, že i přes náhodnost transkripce jsou oba geny přepisovány současně.

Při sběru pomocí mikromanipulátoru je možné sesbírat pouze několik desítek buněk, ale existuje možnost vybrat si konkrétní buňku, provést na ní další měření a teprve následně analyzovat její expresní profil. Toho se velmi často využívá při měření potenciálů na buňkách nervové soustavy (Liss et al., 2001). Nevýhodou je naopak delší doba sběru, během které může docházet k degradaci mRNA v buňce, případně může být exprese pozměněná, když se buňka snaží přizpůsobit novým podmínkám. Další

nevýhodou, která je společná s průtokovou cytometrií, je proces disociace buněk, jenž může rovněž pozměnit expresi některých genů (Stahlberg et al., 2011a). Navíc je takto ztracena informace o umístění analyzované buňky v tkáni.

Laserová mikrodisekce vyžaduje nejdříve fixaci tkáně, následuje výběr konkrétní buňky pod mikroskopem, vyříznutí požadované buňky a její přenesení do lyzačního pufru (Espina et al., 2006). Díky tomu je znám původ a morfologie buňky. Nevýhodu je však fixace tkáně, která je klíčová v celém procesu. Standardně používaná fixace formalínem způsobuje významnou degradaci RNA. Mnohem vhodnější je proto použití fixace etanolem nebo reverzibilním prokřižovacím činidlem, jako je dithiobis(sukcynimidyl)propionát (Xiang et al., 2004). Objevují se již práce, kdy je použita laserová mikrodisekce na živé, ničím nefixované tkáni (Podgorny, 2013). Tento postup by mohl v budoucnu odstranit nevýhodu chybějící znalosti o umístění buňky a zároveň pomoci získat vzorek s nedegradovanou RNA. Zatím se laserová mikrodisekce používá především pro analýzu nádorů, kdy je důležité oddělit nádorové buňky od zdravých (Joseph and Gnanapragasam, 2011).

Dalším krokem v RT-qPCR jednotlivých buněk je lýze. Buňky se nejčastěji sbírají přímo do co nejmenšího objemu lyzačního pufru. Tím je nejčastěji voda s přidanými chemikáliemi podporujícími rozpuštění buněčných obalů nebo přímo pufr pro reverzní transkriptázu (Bengtsson et al., 2005; Grundemann et al., 2008). Množství proteinů, solí a buněčných zbytků je v poměru k objemu, do kterého se buňky sbírají, natolik malé, že není nutné přečištění RNA. Pro účinnou reverzní transkripci je potřeba lyzovat buňku tak, aby se všechna mRNA stala dostupnou pro transkriptázu. Buněčná lýze může probíhat pomocí zahřátí, mechanicky, osmotickým tlakem nebo s použitím detergentů. Nejčastěji se jedná o kombinaci těchto přístupů. Bylo ukázáno, že pro některé typy savčích buněk je zvýšena účinnost lýze přidáním guanidin thiokyanátu do lyzačního pufru (Bengtsson et al., 2008). Množství detergentů je potřeba optimalizovat tak, aby následně neinhibovaly reverzní transkripci. Dále je potřeba zablokovat uvolněné RNázy buď přidáním komerčního inhibitoru RNáz a nebo přidáním chaotropních solí. Vzorky získané pomocí laserové mikrodisekce je možné přečistit pomocí komerčních kitů (Kamme et al., 2003).

Účinnost reverzní transkripce je nejvíce kritickým krokem celé metody RT-qPCR. Je ovlivněna typem reverzní transkriptázy a primovací metodou, tj. použitým typem oligonukleotidů, které se váží na mRNA a z nichž začíná přepis cDNA. Reverzní transkriptázy se dělí podle toho, zda mají RNázovou aktivitu a podle optimální teploty pro funkčnost. Pro primování je možno použít oligo(dT), náhodné hexamery či

specifické primery[1]. Účinnost transkripce je genově specifická a není možné říci, která strategie je nejlepší. Nejčastěji je doporučováno použití směsi hexamerů a oligo(dT) v poměru 1:1 (Stahlberg et al., 2004). Specifické primery se mohou vázat i nespecificky na mRNA (Stahlberg et al., 2004). Specifické primery pro RT by měly být navrženy na 3´ konec mRNA, který degraduje nejpozději. Dále by neměly být identické s primery použitými pro následné qPCR, jelikož účinnost qPCR v takovém případě klesá (Stahlberg and Bengtsson, 2010). Počet přepsaných transkriptů jednoho genu závisí nejen na jejich množství ve vzorku, ale i na celkovém množství RNA. Proto se doporučuje přidat do málo koncentrovaného vzorku nosič, nejčastěji tRNA (Stahlberg et al., 2004).

Množství genů analyzovaných z jedné buňky je limitováno množstvím transkriptů. Pro přesnou kvantifikaci je potřeba nejméně 20 molekul detekovaného genu (Bengtsson et al., 2008). Bez preamplifikace je možné naměřit z jedné buňky až 10 vysoce a středně exprimovaných genů. Pro zvýšení citlivosti qPCR je doporučeno analyzovat každou buňku pouze v jednom RT i qPCR replikátu, spíše než stejné množství cDNA použít na dvě měření. Při naředění cDNA, je vyšší pravděpodobnost, že jednotlivé transkripty nebudou rozděleny v reakci rovnoměrně. Dále bylo ukázáno, že variabilita mezi jednotlivými buňkami je větší než technická chyba měření (Tichopad et al., 2009).

Preamplifikace (namnožení vzorku) může zvýšit množství detekovaných genů až stokrát. Existují dva přístupy k preamplifikaci - buď namnožení RNA (používané hlavně pro mikroarray) nebo namnožení cDNA. Dále je možné preamplifikovat celý transkriptom nebo jen specifické geny, a to lineárně anebo exponenciálně. Při lineární preamplifikaci přibývá v každém kroku stejné množství cDNA/RNA, zatímco v pro exponenciální preamplifikaci platí, že se množství materiálu zdvojnásobí. Při analýze qPCR jednotlivých buněk se nejčastěji používá genově specifická exponenciální multiplexová preamplifikace, kdy jsou specifické primery smíchány dohromady a kdy je provedeno PCR s omezeným počtem cyklů. Při použití 20 cyklů je možné z jedné kopie cDNA získat při 100 procentní účinnosti preamplifikace více než půl milionu kopií. Účinnost preamplifikace závisí na použité polymeráze, teplotním režimu reakce, sekvenci specifických primerů a počátečním množství cDNA (Livak et al., 2013)

Pro samotné qPCR platí zásady uvedené v Bustin et al., 2009. Primery by měly mít vysokou efektivitu, nejlépe přes 95 %, tzn. že během každého cyklu by mělo dojít téměř k zdvojnásobení množeného materiálu. Dále musí mít vysokou senzitivitu, aby byly schopné zachytit jen jednu kopii cDNA a musí být specifické pro daný gen. Vzhledem k tomu, že každý gen je v buňce na genomové DNA pouze ve dvou

[1] Dva krátké oligonukleotidy (cca 20 bází), které jsou komplementární k cDNA a ohraničují na ní úsek většinou do 200 bází, který se amplifikuje

kopiích, není nezbytně nutné navrhovat primery tak, aby byl každý byl umístěn na jiném exonu a mezi nimi ležel intron jako při klasickém qPCR. I přes vysokou senzitivitu primerů platí, že při použití běžného qPCR instrumentu, je typická hodnota Cq (kvantifikační cyklus, z angl. Cycle of Quantification) mezi 30 a 40, tj. detekuje se maximálně 30 kopií cDNA. Proto je nezbytné, aby primery netvořily žádné dimery[2] do 45. cyklu. Dále se doporučuje použít k detekci spíše interkalační činidla, která se váží do velkého žlábku dvoušroubovice DNA, jako je SYBRGreen a EvaGreen, než hydrolyzační próby[3]. Interkalační činidla umožní zjistit teplotu tání produktu a detekovat dimery. Umožní také detekovat produkty, ve kterých je část amplifikována správně a část jako dimer. Takový výsledek může být hodnocen pouze kvalitativně (daný gen je přítomný), avšak ne kvantitativně. Pro analýzu jednotlivých buněk pomocí qPCR je typické, že velký počet reakcí neposkytuje žádný signál (Stahlberg and Bengtsson, 2010).

S rozvojem qPCR roste i množství analyzovaných genů. Vzhledem k tomu, že při analýze jednotlivých buněk je nutné jich analyzovat desítky, aby bylo ukázáno, že exprese nízko exprimovaných genů není jen náhodná, roste i finanční a časová náročnost experimentů. Proto je stále častější použití mikrofluidních čipů pro vysokokapacitní qPCR (high throughtput qPCR), ve kterých qPCR reakce probíhá pouze v několika nanolitrech. Na trhu existuje několik takových přístrojů. OpenArray®Real-Time PCR System (Applied Biosystems) pro měření 56x48 (56 genů ve 48 vzorcích) a dále 112x24, 168x16 a 224x12. Dalším systémem je Biomark™ HD System (Fluidigm) pro měření 48x48 a 96x96 a Smart Chip Real-Time PCR System (Wafergen) pro měření až 384 vzorků a celkem 5184 reakcí (Tabulka II).

[2] Krátké amplifikační produkty vznikající namnožením primerů, které se nenavázaly na templát.

[3] Oligonukleotid, který leží mezi primery. Při amplifikaci cDNA je rozštípán DNázovou aktivitou polymerázy a dochází k uvolnění fluorescenční značky a tím k zvýšení její koncentrace v roztoku.

	OpenArray (všechny typy)	Biomark 48x48	Biomark 96x96	Smart Chip Real-Time PCR System
Počet reakcí	2688	2304	9216	5184
Reakční objem	33 nl	10 nl	6,75 nl	100 nl
Doba pro nanesení vzorku na čip a průběh reakce	2,5 hodiny	2 hodiny	3,5 hodiny	3 hodiny

Tabulka II: Porovnání přístrojů OpenArray®Real-Time PCR System, Biomark HD System Smart Chip Real time PCR system

Pro první dva typy čipů je nutné vzorek preamplifikovat. Z Poissonova rozdělení plyne, že pro 99,3% šanci detekovat alespoň 1 molekulu je nutné, aby byla průměrná koncentrace 5 molekul v reakci. To znamená koncentraci 730 molekul/µl (Biomark™ HD System) respektive 151 molekul/µl (OpenArray®Real-Time PCR System), (Livak et al., 2013). Takto vysoké koncentrace nejsou pro „průměrnou buňku“ typické. Pro Smart Chip Real Time PCR Systém stačí koncentrace 50 molekul/µl, což je v některých případech již reálné množství.

Po proběhnutí qPCR experimentu, je nutná pečlivá úprava a analýza dat (Kapitoly 3.2 a 5.2).

3.2. Analýza dat

Výstupem z qPCR reakce jsou amplifikační křivky znázorňující exponenciální nárůst fluorescence, který odráží vzrůstající množství namnoženého produktu. Tyto amplifikační křivky jsou následně zarovnány pomocí softwaru na stejnou úroveň fluorescence tak, aby bylo možné je porovnat mezi sebou (anglicky baseline correction). Křivky jsou protnuty přímkou (anglicky threshold), vzniklý průsečík udává Cq hodnou. Naměřené Cq hodnoty jsou před samotným vyhodnocením experimentu upravovány a normalizovány tak, aby bylo možné porovnat vzorky mezi sebou. Existují dva základní přístupy jak hodnotit množství transkripty ve vzorku, a to relativní kvantifikace a absolutní kvantifikace. Relativní kvantifikace předpokládá, že ve všech vzorcích existují geny, které jsou exprimovány stejně a k nim se vztahují cílové geny (Kapitola 3.2.1). Absolutní kvantifikace vychází z co nejpřesnějšího zjištění o počtu molekul ve vzorku a není v této práci podrobněji diskutována.

3.2.1. Úprava dat a normalizace

Prvním krokem při úpravě dat, ať pro relativní nebo absolutní kvantifikaci, je mezidestičková kalibrace. Ta se používá pro sjednocení signálu, nastavení hladiny flourescence[4] (treshold) a baseline při měření na více destičkách. Pro kompenzaci měření se používá mezidestičkový kalibrátor (IPC), což je nejčastěji jeden vzorek měřený v triplikátu, pokaždé na stejné pozici destičky, za stejných podmínek. Přepočet Cq hodnot pak probíhá podle vzorce [1]:

$$\boldsymbol{Cq_{GOI} - \frac{1}{m} \sum_{j=1}^{m} (Cq_{IPC_j} - \frac{1}{n} \sum_{i=1}^{n} Cq_{IPC_i}) = Cq_{norm}} \qquad [1]$$

Kde Cq_{GOI} je naměřená hodnota pro danou destičku, m počet IPC na destičce, Cq_{IPC} Cq hodnota pro IPC, n celkový počet IPC a Cq_{norm} přepočítaná hodnota. Mezidestičkový kalibrátor není třeba používat v případě, kdy je experiment navržen tak, aby byly všechny vzorky pro jeden gen na jedné destičce.

Volitelným krokem v analýze dat je přepočet Cq hodnot na 100% účinnost (E) jednotlivých primerů, který se však používá pouze při relativní kvantifikaci [2].

$$\boldsymbol{Cq_{100\%} = Cq \frac{log(1+E)}{log(2)}} \qquad [2]$$

Nejčastěji se účinnost určí pomocí lineární regrese, kdy je měřena řada vzorků, které mají různou koncentraci. Naměřené Cq hodnoty jsou vyneseny proti dekadickému logaritmu koncentrace do grafu, kde jsou proloženy přímkou a směrnice této přímky určuje účinnost [3].

$$\boldsymbol{E = 10^{\frac{-1}{směrnice}} - 1} \qquad [3]$$

Další možností jak zjistit účinnost je přímo proložení amplifikační křivky. Pro tento postup bylo navrženo mnoho matematických modelů (Gentle et al., 2001; Tichopad et al., 2003; Boggy and Woolf, 2010). Přepočtením Cq hodnot se změní pouze jejich hodnota, která ovlivní relativní množství transkriptu, neovlivní ale výsledek statistických testů jako je t-test, ANOVA a neparametrické testy.

Před porovnáním vzorků mezi sebou při relativní kvantifikaci musí být vzorky normalizovány (srovnány) na stejný výchozí stav pomocí normalizátoru tak, aby byla vzata do úvahy rozdílná vstupní koncentrace vzorků, jejich kvalita a účinnost reverzní transkripce. Normalizátorem může být celkové množství RNA, množství ribozomální RNA nebo referenční gen. Množství mRNA se pohybuje mezi 5-7 % z celkového

[4] Flourescence se měří v exponenciální části křivky. Tyto části by měly být pro všechny reakce rovnoběžné a není tedy důležité přesné umístění tresholdu

množství RNA v buňce, proto je normalizace na celkové množství RNA a na množství ribozomální RNA téměř totožná. Tento typ normalizátorů předpokládá, že kvalita RNA je ve všech vzorcích stejná a není inhibována reverzní transkripce ani qPCR. Vzhledem k tomu, že toto nelze vždy zajistit, se jako normalizátory nejčastěji používají referenční geny. Za referenční gen je považován takový, který je stabilně exprimován na stejné hladině za všech testovaných podmínek u shodné populace vzorků. Nicméně neexistuje univerzální referenční gen, který by měl stejnou expresi ve všech tkáních (Bustin, 2000), proto by měl být ověřen set teoretických referenčních genů před každým novým experimentem. Nejčastěji jsou používany dvě metody pro nalezení nejvhodnějšího kandidáta – Genorm a Normfinder (Vandesompele et al., 2002; Andersen et al., 2004)

V první metodě se předpokládá, že nejstabilnější referenční geny mají stejný poměr exprese ve všech vzorcích bez ohledu na reakční podmínky a kvalitu vzorku. Pro každý potencionální referenční gen se vypočítá tzv. M-hodnota, která se rovná průměrné standardní odchylce počítané pro rozdíl Cq hodnot v každém vzorku z každého možného páru referenčních genů. Gen s nejvyšší M-hodnotou je nejméně stabilní a je vyloučen z dalších výpočtů. Pro ostatní geny se spočítá nová M-hodnota a vše se opakuje, dokud nezůstanou dva poslední geny, které jsou označeny jako nejvhodnější referenční geny. Pro tento výpočet musí být vybrány geny, které nejsou v jedné metabolické dráze, a nepředpokládá se, že jsou biologicky korelované. Tato metoda byla implementována do softwaru Genorm (Vandesompele et al., 2002).

Druhá metoda vychází z lineárního modelu variability pro daný gen [4],

$$\boldsymbol{Cq_{igj} = \alpha_{ig} + \beta_{gj} + \varepsilon_{igj}} \qquad [4]$$

kde Cq pro gen *ij*-tého vzorku ve skupině *g* je rovno celkové expresi genu *i* ve skupině *g* plus množství mRNA ve vzorku *j* a náhodnému biologickému a experimentálnímu faktoru *ε*. Umožňuje spočítat variabilitu uvnitř skupin, ale i mezi skupinami. Tím se odhalí geny, které jsou sice stabilní v rámci skupiny, ale mohou být ovlivněny faktorem určujícím typ skupiny (např. léčbou, věkem, pohlavím apod.). Aby referenční geny nebyly ovlivněny odlehlými hodnotami, je doporučeno hledat nejvhodnějšího kandidáta mezi 5-10 geny a v každé skupině mít nejméně 8 vzorků. V případě, že je jeden referenční gen nedostatečný, je možné použít průměr více genů. Tento model byl implementován do softwaru Normfinder (Andersen et al., 2004).

K zvýšení přesnosti měření se používají replikáty. Ty mohou být na několika úrovních – replikáty qPCR, reverzní transkripce nebo vzorků z jednoho subjektu. Pro celkovou analýzu je nutné z nich vypočítat průměr. Pokud není experiment navržen vyváženě (vždy stejné množství replikátů v každém kroku), je nutné dodržet pořadí výpočtů průměrů. Průměr z replikátu qPCR se počítá ještě před normalizací na referenční gen. Pouze v případě, že byl referenční a cílový gen měřen v jedné reakci pomocí dvou odlišně flourescenčně značených prób, počítá se průměr qPCR replikátů až po normalizaci.

Další úprava dat je již volitelná. Pro názornější grafické zpracování se přepočítávají Cq hodnoty na relativní množství, kdy nejvyšší hodnotě je přiřazena nulová exprese a naopak nejnižší hodonotě maximální exprese[5]:

$$N_{rel} = 2^{Cq_{min} - Cq} \qquad [5]$$

Některé typy statistických analýz (Kapitola 3.2.3) nelze provádět s chybějícími údaji. Existuje několik možností jak nahradit chybějící data. Ta se mohou nahradit průměrem biologických replikátů (v případě, že chybějící údaj je technickou chybou měření), interpolací na základě exprese naměřených u ostatních genů nebo nejvyšší naměřenou hodnotou Cq +2, příp. +4 (pokud se daný gen ve vzorku nevyskytuje nebo je pod hranicí kvantifikace). Podrobně se tímto tématem zabývá kapitola 6.1.1.

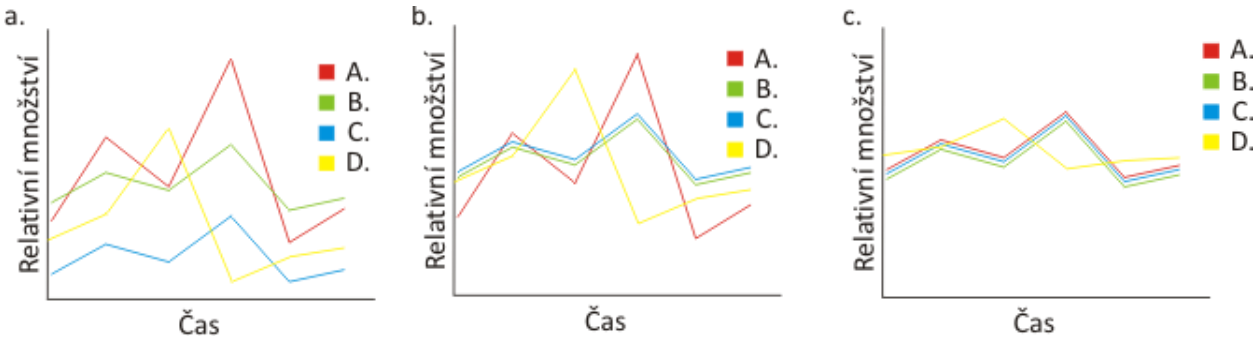
Pro multivariantní analýzy jsou data upravována centrováním (anglicky mean center) či škálováním (anglicky autoscale), (Obrázek 5). V případě centrování [6],

$$Cq_{centr} = Cq - Cq \qquad [6]$$

je od každé Cq hodnoty odečten průměr všech vzorků pro daný gen. Data tak mají průměrnou expresi 0. Tím jsou srovnána na stejnou hladinu a zdůrazní se velikost změn a ne velikost exprese. Další možností je škálování dat [7],

$$Cq_{scale} = \frac{Cq - Cq}{SD_{Cq}} \qquad [7]$$

kdy jsou normovaná data vydělena standardní odchylkou průměru. Data tak mají průměrnou expresi 0 a standardní odchylku 1. Tím se opět zdůrazní průběh exprese s potlačením vlivu její velikosti. Všechny geny v analýze tak mají stejnou váhu.



Obrázek 5: Vliv centrování a škálování

a. Naměřená genová exprese. Geny *B, C* mají identický profil, jen se liší v množství transkriptu. Gen *A* má podobný profil, ale větší rozdíly mezi maximální a minimální expresí. Gen *D* je zcela odlišný od ostatních. **b.** Centrovaná data, průměrná exprese všech genů je 0. Profil i relativní změny v profilu genů *B* a *C* jsou identické. **c.** Škálovaná data, průměrná exprese všech genů je 0 a standardní odchylka je 1. Profil i relativní změny v profilu genů *A, B* a *C* jsou identické.

3.2.2. Základní statistická analýza

Relativní exprese dvou genů v jednom vzorku je dána [8]

$$\frac{N_A}{N_B} = K_{RS}\frac{\eta_B\ 1+E_B\ ^{Cq_B-1}}{\eta_A\ 1+E_A\ ^{CqA-1}} \qquad [8]$$

Kdy K_{RS} je relativní senzitivita detekčních chemikálií pro geny *A* a *B*, η_A a η_B jsou výtěžky syntézy cDNA, Cq_A jsou hodnoty pro referenční gen, Cq_B hodnoty pro cílový gen, číslo *-1* v exponentu zahrnuje první cyklus, kdy nedochází ke zdvojení produktu, ale pouze k doplnění cDNA na dvouřetězcový templát (Stahlberg et al., 2005). Přestože se předpokládá, že výtěžky cDNA jsou ve všech vzorcích při dodržení stejných podmínek pro reverzní transkripci stejné, nemusí to být vždy pravda. Výtěžky syntézy cDNA však není snadné určit. To by vyžadovalo zkušební vzorek se známým množstvím kopií mRNA a se stejnou matricí jako je analyzovaný vzorek. Z tohoto důvodu se poměr dvou genů ve vzorku neurčuje a používá se komparativní kvantifikace, kdy se porovnávají dva vzorky mezi sebou (Pfaffl et al., 2002) [9]:

$$\frac{Vzorek\ 1}{Vzorek\ 2} = \frac{1+E_B\ ^{Cq_{B1}-Cq_{B2}}}{1+E_A\ ^{Cq_{A1}-Cq_{A2}}} \qquad [9]$$

Kdy Cq_A jsou hodnoty pro referenční gen, Cq_B hodnoty pro cílový gen. Při předpokladu 100% účinnosti pro oba geny se vzorec zjednoduší na [10]:

$$\frac{\boldsymbol{Vzorek\ 1}}{\boldsymbol{Vzorek\ 2}} = 2^{\ Cq_{B1} - Cq_{B2}\ -\ Cq_{A1} - Cq_{A2}\ } = 2^{\Delta\Delta Cq} \qquad [10]$$

Tento vzorec patří mezi nejpoužívanější zobrazení qPCR dat (Livak and Schmittgen, 2001).

Pro porovnání vzorků mezi sebou se používají nejčastěji data ve formátu ΔCq (cílový gen minus referenční gen). Ta bývají normálně rozdělena, což umožní použít pro srovnání parametrické testy typu t-test (pro dvě skupiny vzorků) či ANOVA (více skupin vzorků). Pokud data nejsou normálně rozdělena, je možné použít neparametrické testy.

Variabilita mezi studovanými vzorky se dělí na dva typy – hledanou, která je dána rozdílným ošetřením vzorku, a nechtěnou, která je zapříčiněna obecnou biologickou a technickou variabilitou. Pro odstranění nechtěné biologické variability se používá náhodný výběr vzorků a jejich velké množství. Pokud je to možné, používá se párové měření. Technická variabilita se minimalizuje použitím technických replikátů a to jak při qPCR, tak při reverzní transkripci a izolaci (Tichopad et al., 2009).

Další možností, jak se podívat na zkoumané geny, je hledání jejich vzájemné expresní podobnosti/různosti. Tu lze nejlépe vyjádřit korelačním koeficientem. Tento koeficient nabývá hodnot od mínus jedné pro geny, které se chovají opačně, až do plus jedné pro geny které se chovají z hlediska genové exprese stejně. Geny, které nemají mezi sebou žádnou závislost, mají hodnotu korelačního koeficientu blízkou nule. Naopak za silný korelační koeficient se považuje takový, jehož absolutní hodnota je vyšší než 0,6. Vztah mezi geny může být přímý nebo zprostředkovaný třetím genem. Početně to lze zjistit pomocí výpočtu tzv. parciální korelace (Stahlberg et al., 2011b).

3.2.3. Multivariantní analýza

S nárůstem množství dat stoupá i náročnost analýzy. Jednoduché statistické testy nejsou dostatečné a přehledné, proto se používají multivariantní analýzy zaměřené na vizualizaci dat. Pomocí nich se neanalyzuje gen po genu, ale na vzorek se pohlíží jako na celek s určitými vlastnostmi, které jsou dány expresí genů.

3.2.3.1. Analýza hlavních komponent

Analýza hlavních komponent je jednou z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Byla zavedena již v roce 1901 Pearsonem (Pearson, 1901) a nezávisle na něm Hotellingem v roce 1933 (Hotelling, 1933). Analýza hlavních komponent je typ analýzy, který umožňuje zobrazení vícerozměrného

prostoru do dvou- či třírozměrného. Je to způsob nalezení určitého vzoru v datech, který může zdůraznit podobnost i rozdíly v jednotlivých měřeních. Výhodou analýzy hlavních komponent je, že dochází k velkému snížení dimenze dat a relativně malému snížení informací o variabilitě.

Analýza hlavních komponent může být chápána jako transformace z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí směry maximálního rozptylu. Matematicky je hlavní komponenta definována jako lineární kombinace optimálně vážených pozorovaných proměnných.

Prvním krokem ve výpočtu je normalizace dat tak, aby jednotlivé proměnné měly průměr rovný nule (Kapitola 3.2.1). Dalším krokem je výpočet kovariantní matice pro data se stejnými jednotkami nebo korelační matice Pearsonových korelačních koeficientů pro data s různými jednotkami. Pro qPCR data se používají kovariantní matice. Pro každou dvojici proměnných se spočítá kovariance a data se seřadí do matice tak, aby na hlavní diagonále byly variance jednotlivých proměnných. Matice je pak souměrná podle hlavní diagonály. Předposledním krokem je výpočet vlastních vektorů a vlastních hodnot matice (anglicky eigenvector a eigenvalue). Vlastní vektor je takový vektor, který po vynásobení matice zprava dá opět násobek vlastního vektoru.

$$\mathbf{a}.\ \begin{matrix}2 & 3\\ 2 & 1\end{matrix} \times \begin{matrix}1\\ 3\end{matrix} = \begin{matrix}11\\ 5\end{matrix}$$

$$\mathbf{b}.\ \begin{matrix}2 & 3\\ 2 & 1\end{matrix} \times \begin{matrix}3\\ 2\end{matrix} = \begin{matrix}12\\ 8\end{matrix} = 4 \times \begin{matrix}3\\ 2\end{matrix}$$

Obrázek 6: Příklad nevlastního (a) a vlastního vektoru (b)

Mezi charakteristiky vlastních vektorů patří, že mohou být nalezeny pouze pro čtvercové matice, ale ne každá čtvercová matice má vlastní vektory. Pokud má matice o rozměrech *n x n* vlastní vektor, má jich právě *n*. Vlastní vektory jsou navzájem kolmé a zpravidla se píší tak, aby měly vždy velikost jedna. Ke každému vlastnímu vektoru je přiřazeno vlastní číslo, které určuje původní velikost vektoru. Po nalezení vlastních čísel a vlastních vektorů kovariantní matice jsou vlastní vektory seřazeny podle velikosti vlastních čísel od největšího po nejmenší a vytvoří se matice, kdy jednotlivé vlastní vektory tvoří sloupce, což jsou hledané hlavní komponenty. Pro vytvoření dvoudimenzionálního zobrazení se ponechají pouze první dva vlastní vektory, pro vytvoření třídimenzionálního první tři atd. Hlavní komponenty jsou seřazeny podle klesajícího rozptylu a většina informací o variabilitě dat je přitom soustředěna do první

komponenty a nejméně informací je obsaženo v poslední komponentě. Dále platí, že hlavní komponenty jsou nekorelované. Posledním krokem při zobrazování dat je transpozice srovnaných vlastních vektorů a normovaných dat a jejich vynásobení. Tím se získá matice, kterou již lze zobrazit (Smith, 2002; Shlens, 2005).

3.2.3.2. Shluková analýza

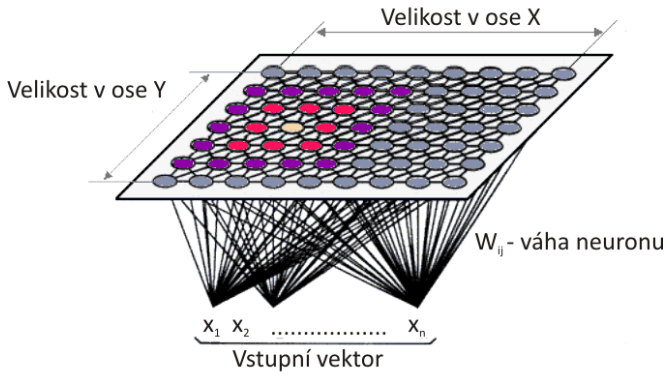
Shluková analýza (cluster analysis) rozděluje data do skupin na základě jejich podobnosti. Na rozdíl od analýzy hlavních komponent nedochází při zobrazení dat ke snížení dimenze a ke ztrátě informace. Rozdělení datového souboru na několik relativně homogenních podsouborů (shluků, clusterů) by mělo proběhnout tak, aby byly jednotky či objekty v jenom shluku co nejvíce si podobné, a naopak jednotky či objekty zařazené do jiného shluku si byly podobné co nejméně. Shlukovou analýzu dělíme na dva základní typy – parciální a hierarchické shlukování. Parciální shlukování dělí data do nepřekrývajících se skupin, takže každý objekt je v právě jednom subsetu. Naproti tomu hierarchické shlukování umožňuje tvorbu subclusterů.

Existuje velké množství možností jak data rozdělit do shluků. Mezi základní výpočty patří metoda nejbližšího souseda (single linkage), kdy je vzdálenost určována vzdáleností dvou nejbližších objektů z různých shluků. Metoda nejvzdálenějšího souseda (complete linkage) určuje vzdálenost shluků naopak, tedy vzdáleností dvou nejodlehlejších objektů z různých clusterů. Používá se především v případě, že objekty tvoří přirozeně oddělené shluky, ale nehodí se, pokud je tendence k řetězení objektů. Dalšími způsoby určení shluků jsou: podle vzdáleností jejich těžišť (centroid linkage), na základě průměrné vzdálenosti všech párů objektů z různých shluků (average linkage) nebo na základě minimální sumy čtverců (Wardova metoda). Stejně tak existuje velké množství možností jak počítat vzdálenosti mezi shluky. Nejčastěji se používá euklidovská vzdálenost, čtverec euklidovské vzdálenosti, manhattanská vzdálenost, maximální vzdálenost, Jaccardův koeficient, Pearsonův korelační koeficient a mnoho dalších (Lance and Williams, 1967).

Základními výchozími předpoklady shlukové analýzy jsou vzájemná nezávislost a stejná významnost vstupních proměnných a nezávislost čísel na jednotkách měření, tedy jejich bezrozměrnost. Výsledky analýzy závisí na volbě proměnných, zvolené míře vzdálenosti mezi objekty a shluky a na zvoleném algoritmu výpočtu. Aplikace shlukové analýzy je vhodná zejména tam, kde objekty projevují přirozenou tendenci se seskupovat. Grafickou podobou výsledků popsané metody je dendrogram (Kapitola 5.23.2.3).

3.2.3.3. Kohonenovy samoorganizující se mapy

Alternativou klasických metod shlukové analýzy jsou biologií inspirované algoritmy, zejména metody založené na umělých neuronových sítích. Kohonenova samoorganizující se mapa (Kohonen Self-Organizing Map –SOM), jejímž autorem je Teuvo Kohonen, patří do kategorie modelů, které se učí samoorganizací (Kohonen and Honkela, 2007). Cílem Kohonenových map je zobrazení vícerozměrných dat do dvourozměrného prostoru tak, aby byly podobné položky z mnoharozměrného prostoru umístěny v mapě blízko sebe. Stejně jako většina arteficiálních neuronových sítí, SOM může pracovat ve dvou fázích – učební (trénovací) na známém datovém setu a samotné mapovací, kdy se dělí neznámé vzorky do skupin. SOM se skládají z uzlů (neuronů), které jsou umístěny ve dvou vrstvách jako dvě hexagonální nebo pravoúhlé mřížky. Zobrazení zachovávající topologii, které vznikne při trénování SOM, má důležitou vlastnost: libovolné dva vzory, které jsou blízké ve vstupním prostoru, vytvářejí v síti odezvy v neuronech, které si jsou ve výstupním prostoru také blízké. Každý výstupní neuron je propojen s každým vstupním neuronem, vstupní neurony nejsou propojeny vůbec, kdežto výstupní ano (Obrázek 7).

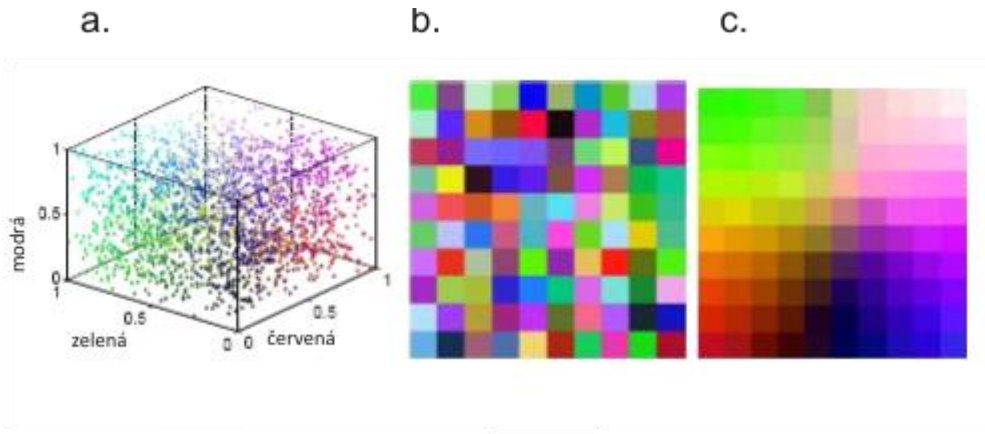


Obrázek 7: Jednoduché znázornění Kohonenovi samorganizující se mapy

Každý vstupní neuron je propojen s každým výstupním. Během učícího procesu dochází ke změně váhy neuronů tak, aby si byly sousední výstupní neurony co nejpodobnější (nejbližší okolní neurony v okolí vítězného neuronu jsou je znázorněny stejnou barvou) a zároveň odpovídaly co nejvíce vstupním.

Zjednodušený algoritmus výpočtu je následující. Při inicializaci jsou náhodně určeny váhy jednotlivých neuronů. Následují dva kroky, které se opakují až do dosažení požadovaného stupně konvergence. Prvním krokem je nalezení vítězného neuronu s nejvíce podobným modelovým vektorem (best matching unit) pro každou jednotku vstupního prostoru. Určení vítězného neuronu probíhá pomocí výpočtu euklidovské vzdálenosti mezi váhou neuronů a vstupní hodnotou vektoru. Ve druhém kroku se

modelové vektory nalezeného neuronu upraví tak, aby lépe odpovídaly datům ze vstupního prostoru, a zároveň se upraví i modelové vektory sousedních neuronů. Počet opakování dvou kroků by měl být alespoň pětsetkrát větší, než je počet neuronů v síti (Weinlichova and Fejfar, 2010).



Obrázek 8: Ukázka třídění dat pomocí SOM

a. Vstupní data v třírozměrném prostoru, pro jednoduchost je každý bod (vzorek) charakterizován svojí barvou. **b.** Počáteční váhy jednotlivých neuronů, před učícím procesem. Barvy neuronů jsou náhodné. **c.** Konečné váhy neuronů a zobrazení barev dvourozměrné prostoru

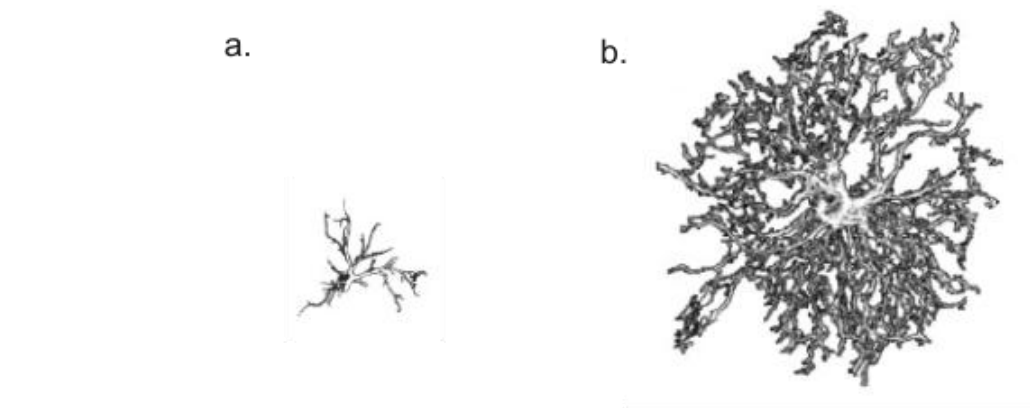
Při automatizovaném výpočtu Kohonenovy mapy se zadávají dva hlavní parametry, jedním je velikost mapy, druhým konečný rádius sousednosti. Větší mapy dávají v zobrazování lepší rozložení, ale nevýhodou je vyšší výpočetní složitost. Konečný rádius sousednosti udává hladkost zobrazení a měl by být nastaven vzhledem k šumu v datech.

3.3. Astrocyty

Lidský mozek obsahuje zhruba $1x10^{12}$ buněk. Podle nejnovějších výzkumů je poměr gliových a nervových buněk v celém mozku zhruba 1:1, záleží na místě původu (Azeved et al., 2009). Gliové buňky rozdělujeme na dvě skupiny – mikroglie, které mají mezodermální původ, a makroglie (ektodermální původ), mezi které patří oligodendrocyty, NG2 glie, ependymální buňky a astrocyty. Poslední jmenované jsou nejpočetnější buňky nervové tkáně, jejichž název byl odvozen podle jejich tvaru z latinského slova *astra* (hvězda). Astrocyty tvoří různorodou heterogenní populaci s mnoha funkcemi jak ve zdravé, tak i v poškozené tkáni. Pro jejich studium bylo vyvinuto mnoho transgenních modelů, kde jsou jednotlivé astrocyty značeny fluorescenčními proteiny (Nolte et al., 2001).

3.3.1. Základní rozdělení a morfologie astrocytů

Astrocyty můžeme rozdělit podle jejich morfologie a umístění v centrální nervové soustavě na dvě skupiny – protoplazmatické a vláknité (fibrilární) astrocyty. Fibrilární astrocyty se nachází hlavně v bílé hmotě, kde jsou zodpovědné za výživu axonů. (Nedergaard et al., 2003). Protoplazmatické astrocyty se nachází v šedé hmotě mozku a míchy, kde mají podobnou funkci jako fibrilární. Jejich výběžky jsou kratší, širší a více větvené v porovnání s fibrilárními astrocyty. Protoplazmatické astrocyty jsou organizovány v mikrodoménách, které jsou od sebe odděleny a fungují nezávisle. Každá mikrodoména může obsahovat více než 100 000 synapsí pokrytých astrocytálními výběžky (Bushong et al., 2002). Velikost a množství výběžků astrocytů závisí na jejich původu, kdy s vývojovým stupněm živočicha roste jejich počet a rozvětvenost. Protoplazmatické astrocyty u hlodavců mají 30-60 µm v průměru a pouze 3-4 výběžky, které obsahují gliální fibrilární kyselé proteiny (GFAP) a dotýkají se nejčastěji cév. Oproti tomu lidské protoplazmatické astrocyty jsou symetrické, větší a mnohem více komplexní (Obrázek 9). I přestože tělo buňky má průměr pouze 10 µm, výběžky jsou dlouhé až 200 µm a průměrně jich je na jedné buňce 40.



Obrázek 9: Astrocyt

Porovnání rozvětvenosti a velikosti astrocytů pro myš **(a)** a člověka **(b)**. Převzato z Nedergaard et al., 2003.

S komplexností organismu roste i množství astrocytů na jeden neuron. Na jeden astrocyt u pijavice připadá 25-30 neuronů (poměr 25-30:1), u háďátka je to poměr 6:1, pro hlodavce 3:1 a u člověka připadá na jeden neuron 1,4 astrocytu, tj. poměr 1:1,4 (Nedergaard et al., 2003).

Kromě obecného rozlišení na fibrilární a protoplazmatické astrocyty, existuje další rozdělení astrocytů, např. podle Emsleyho a Macklinse. Na základě tří komplementárních metod značení s využitím GFAP a S100β imunohistochemického barvení u transgenních myších, které exprimují zelený fluorescenční protein (GFP) pod lidským promotorem pro GFAP, rozdělili astrocyty do 9 skupin: tanycyty, radiální glie, Bergmannovy glie, protoplazmatické astrocyty, fibrilární astrocyty, velate glie, marginální glie, perivaskulární glie a ependymální glie (Emsley and Macklis, 2006).

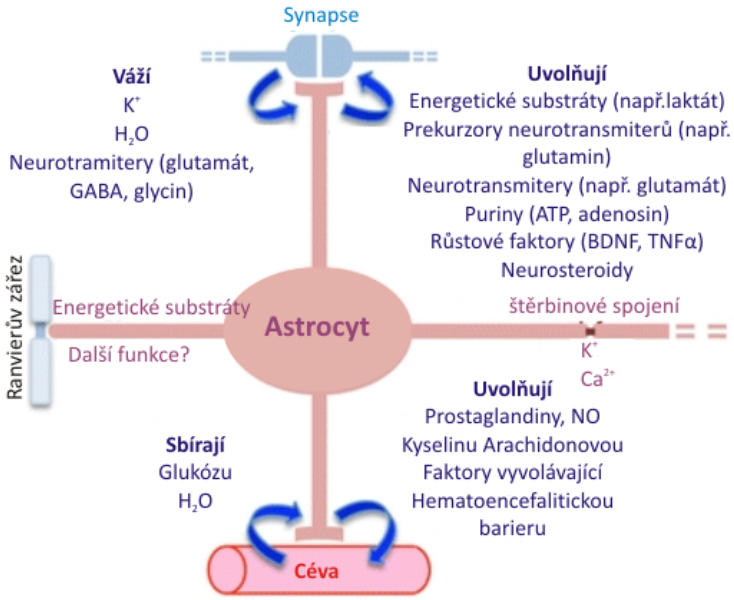
Za typický znak astrocytů je považována exprese GFAP, který se nejčastěji používá jako specifický marker a zároveň je to jeden z nejvíce syntetizovaných proteinů v mozku. Gliální filamenta, jejíž základní složka je kódována *gfap* genem, jsou tvořena ve vyšší míře ve fibrilárních astrocytech ve srovnání s protoplazmatickými (Panickar and Norenberg, 2005). Přestože je GFAP považován za univerzální marker astrocytů, ukázalo se, že imunochemickým barvením pro GFAP nelze detekovat všechny buňky, které jinak vykazují morfologické vlastnosti astrocytů (Walz, 2000). Proto se používají i další markery jako je S100β, vápník vázající protein, či glutamin syntáza (GLUL). Astrocyty dále exprimují ve velkém množství transportéry excitačních aminokyselin 1 a 2 (EAAT1, EAAT2). V roce 2008 byl nalezen další typický protein pro astrocyty, a to aldehyd dehydrogenáza 1 (ALDHL1), (Cahoy et al., 2008).

3.3.2. Hlavní funkce za fyziologických podmínek

Od doby svého objevení se neustále rozšiřují znalosti o významu a funkci astrocytů. Z počátku byly považovány za buňky, které spojují neurony s cévami a slouží jen jako „lešení“ pro neurony. Další objevenou funkcí astrocytů byla jejich schopnost vychytávat glutamát ze synaptické štěrbiny, a tím zabraňovat jeho dlouhodobému působen na postsynaptické receptory (Martinez-Hernandez et al., 1977). Glutamát je zachycen pomocí transportérů excitačních aminokyselin (EAAT1 a EAAT2). V astrocytech je glutamát přeměn glutamin syntázou na glutamin a je vrácen zpět do neuronů, kde je hydrolyzován glutaminázou

Gliové buňky jsou také zodpovědné za udržování homeostáze (Walz, 1989). Ukázalo se, že gliové buňky obsahují přenašeče a kanály pro draslík a vyrovnávají koncentraci draslíku v extracelulárním prostoru při depolarizaci neuronů. Draslík je přijat do buňky dovnitř usměrňujícími draselnými kanály (KIR), je přenášen podél astrogliálního syncytia pomocí mezerových spojů a uvolňuje se do extracelulárního prostoru v místech s nižší koncentrací draslíku (Kofuji and Newman, 2004). Draslík může být dále také distribuován spolu s chloridovými ionty za pomoci Na^+/K^+ATPáz. Celý pufrovací mechanismus je založen na polarizované expresi specifických iontových kanálů na membráně astrocytu. Nejvýznamnější draselný kanál KIR4.1 (kódovaný genem *Kcnj10*) je exprimován především na výběžcích astrocytů, které se dotýkají cév. KIR4.1 je slabě dovnitř draslík usměrňující kanál a dovolí tok iontů oběma směry, tím výborně udržuje množství draselných iontů v požadované koncentraci. Další z draselných kanálu KIR2.1 (kódovaný genem *Kcnj2*) je exprimován pouze na membráně astrocytů dotýkajících se neuronálních synapsí. Tento kanál je silně dovnitř usměrňující a dovolí pouze vtok draselných iontů, které jsou následně redistribuovány.

Další základní funkcí astrocytů je zásobovat energií (glukózou) neurony (Obrázek 10). Proteiny CX43 a CX30 zprostředkovávající štěrbinové spojení, které umožní intercelulární přesun glukózy a jejich metabolitů přes syncytium astrocytů z cév až do neuronů (Rouach et al., 2008). Astrocyty mají dva zdroje glukózy - mohou získat glukózu přímo z vlásečnic anebo přeměnit uskladněný glykogen glykogenlýzou na glukózu. Tento způsob se používá především při nedostatku glukózy. Pomocí glykolýzy je glukóza přeměněna na pyruvát, který je dále zpracováván v mitochondriích astrocytů nebo je pyruvát přeměněn na laktát a ten transportován dále do neuronů, které jsou schopny z něj získat energii.



Obrázek 10: Známé funkce astrocytů ve zdravé tkáni

Převzato z Sofroniew and Vinters, 2010..

Po dlouhou dobu byly astrocyty považovány za buňky, které nejsou excitovatelné. Teprve až v devadesátých letech byly nalezeny astrocyty, které byly schopny odpovědět na neuronální aktivitu pomocí oscilace koncentrace vápenatých iontů v intracelulárním prostoru buňky. Tyto oscilace slouží jako přenašeče signálů do sousedících buněk. Tímto způsobem se kalciové vlny mohou šířit na velkou vzdálenost v celé centrální nervové soustavě. Při změně intracelulárních koncentrací vápenatých iontů dochází k uvolnění „gliotransmiterů“, mezi které patří glutamát, ATP a D-serin (Hamilton and Attwell, 2010), které dále mohou aktivovat různé glutamátové receptory. Pokud jsou tyto látky uvolněny blízko synaptické štěrbiny, mohou modulovat synaptickou aktivitu blízkých neuronů. Tato spolupráce neuronů a astrocytů se nazývá tripartitní synapse (Araque et al., 1999).

Astrocyty jsou formovány u myší až v pozdním stádiu embryogeneze (Ge et al., 2012) a plně vyvinuté jsou mezi třetím a čtvrtým týdnem po narození (Freeman, 2010). Jejich původ, stejně jako původ reaktivních astrocytů vznikajících při astroglióze, není zatím zcela objasněn. Část astrocytů může mít původ v NG2 gliích (Alonso, 2005). Tyto buňky byly dlouhou dobu považovány pouze za prekurzory oligodendrocytů, později se však ukázalo, že mohou být i prekurzory pro část neuronů a astrocytů. NG2 glie jsou považovány za multipotentní dospělé neurální kmenové buňky. Jejich hlavním markerem je

exprese NG2 chondroitin sulfát proteoglykanu (CSPG4) a Pdgfrα (Receptor α pro růstový faktor odvozený z krevních destiček).

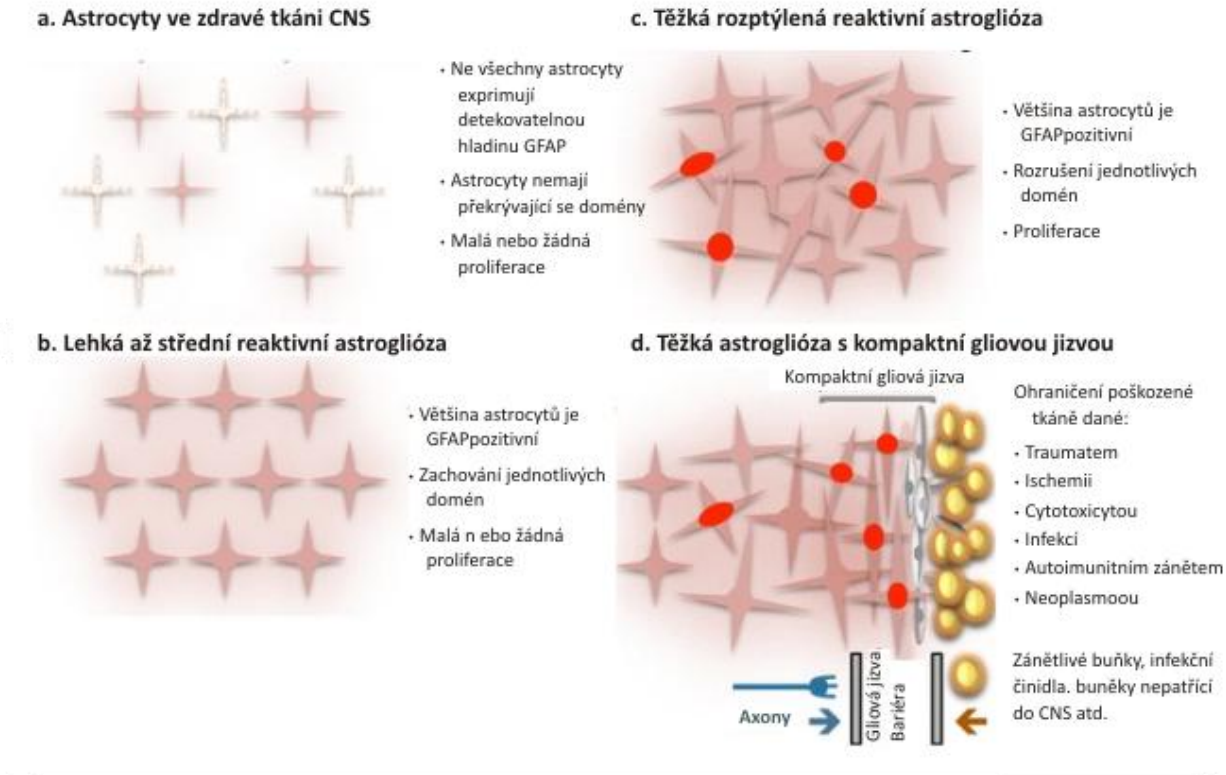
3.3.3. Patologie astrocytů

Mezi nejčastější poranění centrální nervové soustavy (CNS) patří ischemické poškození, které je následkem přechodného nebo trvalého snížení průtoku krve. Tím je sníženo zásobení kyslíkem a živinami, a dochází tak k narušení funkcí neuronů a gliových buněk. Obecně existují dva rozdílné typy mozkové ischemie: globální a fokální. Globální ischemie v mozku nastává po dočasném zastavení krevního oběhu v celém těle. Následkem toho se během sekund sníží mozkový průtok krve. Po pěti minutách pak dochází k nevratnému poškození tkáně. Fokální ischemie nastává po přechodném nebo trvalém zmenšení průtoku krve v oblasti některé z mozkových cév. V místě, kde je krevní průtok zcela zastaven, se vytvoří tzv. „infarktové jádro“, v němž dochází k odumření všech typů buněk (Nedergaard and Dirnagl, 2005).

Po mozkové ischemii jsou poškozeny nejen neurony, ale i astrocyty, které nemohou plnit svoji funkci. Astrocyty reagují na poškození změnou objemu, čímž dochází ke vzniku mozkového edému. Na zvětšování objemu astrocytů se podílí řada iontových kanálů a přenašečů. Může být způsobeno mnoha různými mechanismy, například inhibicí Na^+/K^+ATPázy, glutamátovou excitotoxicitou, acidózou a dalšími příčinami (Sofroniew and Vinters, 2010). Poškození astrocytů ovlivňuje neurony. Hlavní roli hraje exitotoxicita glutamátu, oxidativní stres a acidóza. Za fyziologických podmínek, kdy astrocyty zcela plní svojí funkci, obsahují vysoké koncentrace antioxidantů a substrátů pro antioxidanty (např. glutathion), které zvládají neutralizovat oxidační molekuly. Po poranění dochází ke zvýšení množství oxidantů a ke zvýšenému vylití glutamátu z astrocytů i z neuronů do extraceluárního prostoru (Swanson et al., 2004), což zhoršuje poškození astrocytů a neuronů. Astrocyty mohou přispívat k poškození mozku při jeho poranění, kdy přijímají ionty a aminokyseliny (Sofroniew and Vinters, 2010). Pro vyrovnání zvyšujícího se osmotického tlaku musí dále přijmout větší množství vody, což vede ke zvýšení objemu astrocytů a vytváření otoku, který způsobuje další ischemické poškození. Za toto zvětšování objemu zodpovídá mnoho iontových kanálů a transportérů (Mongin, 2007).

Astrocyty mají klíčovou roli po poranění CNS, kdy se mění jejich vzhled i funkce. Jsou zodpovědné za uvolňování cytokinů, hlavně TNFα, interleukinů, a produkují molekuly extracelulární hmoty. Při mírném poškození zůstává část astrocytů plně funkční a tkáň může být navrácena do téměř původního stavu (Thoren et al., 2005). V závislosti na typu poškození působí astrocyty na neurony pozitivně i negativně.

Po poranění dochází k reaktivní astroglióze, kdy se astrocyty přemění na tzv. reaktivní astrocyty. Vlastnosti reaktivních astrocytů a jejich funkce záleží na stupni poškození (Obrázek 11).



Obrázek 11: Různé stupně astrogliózy

a. Astrocyty ve zdravé tkáni CNS. **b.** Lehká až střední reaktivní astroglióza, při které se již mění exprese a funkční aktivita. Začínají se překrývat výběžky astrocytů a samotné astrocyty jsou připraveny buď na úplnou regeneraci tkáně nebo na další aktivaci. **c.** Těžká rozptýlená reaktivní astroglióza. Obsahuje změny v molekulární expresi, funkci a buněčnou hypertrofii. Objevují se nové astrocyty (na obrázku značené červenou značkou), domény původních astrocytů jsou zrušeny. **d.** Těžká reaktivní astroglióza s kompaktní gliovou jizvou ohraničující poškozenou tkáň. Nachází se zde nové astrocyty (na obrázku značené červenou značkou) a i další buňky (na obrázku šedivě), např. fibrilární buňky a další glie. V gliální jizvě se výběžky astrocytů překrývají. Zralá jizva zůstává po dlouhou dobu jako bariéra nejen pro axony, ale i pro zánětlivé buňky a buňky nepatřící do CNS. Převzato z (Sofroniew and Vinters, 2010).

Obecně jsou reaktivní astrocyty hypertrofní a mají zvýšenou expresi GFAP, vimentinu, nestinu a zvýšenou proliferaci. To je spojeno s tvorbou nových tenkých dlouhých cytoplazmatických výběžků. Gliální cytoplazmatické výběžky mohou v případě nekrózy nakonec vytvořit síťovinu kolem nekrotického místa a utvoří tak gliální jizvu (Pekny et al., 2007). Ta může tvořit bariéru mezi zdravou tkání a místem

poranění. Zabraňuje tak šíření zánětlivé odpovědi. Gliální jizva ale také brání průchodu neurotransmiterů, cytokinů, léků a dalších molekul potřebných pro normální funkci neuronů (Roitbak and Sykova, 1999).

4. Materiál a metody

4.1. Výčet použitých biologických materiálů

Drápatka vodní (*Xenopus laevis)*

GFAP/EGFP transgenní myší model [označení TgN(GFAP-EGFP)]l

Myš (*Mus musculus*), strain B6D2F1

4.2. Výčet použitých metod

Práce s RNA a DNA: Izolace RNA, stanovení koncentrace RNA, DNA, reverzní transkripce, navrhování primerů, preamplifikace, kvantitativní polymerázová řetězová reakce.

Statistické metody: Popisná statistika, statistické testy, výpočet korelací a parciálních korelací, analýza hlavních komponent, výpočty dendrogramu a arteficiálních neuronových sítí.

Ostatní metody: Fluorescencí aktivovaný sběr buněk (provedeno Z. Cimbůrkem, Servisní laboratoř mikroskopie a cytofluorometrie, Ústav molekulární genetiky), ischemické poškození kůry mozkové a disociace tkáně (provedeno P. Honsou a D. Džambou, Oddělení buněčné neurofyziologie, Ústav experimentální medicíny), imunochemická detekce GFAP, PDGFRα a CSPG4 (provedeno P. Honsou).

5. Výsledky

Hlavním cílem této práce bylo vyvinout a optimalizovat metodu pro sběr jednotlivých buněk a jejich analýzy pomocí qPCR, dále vypracovat postup pro analýzu jednotlivých buněk a aplikovat získané poznatky na biologický systém. Jako ukázkový biologický systém byly vybrány astrocyty, na kterých probíhala většina měření.

5.1. Vypracování metody

Prvním krokem při vypracovávání metody, která je vhodná pro jednotlivé buňky, byla optimalizace preamplifikace a měření pomocí vysokokapacitního qPCR přístroje. Naměřené výsledky byly prezentovány v těchto publikacích:

Dye Based High Throughput qPCR in Microfluidic Platform BioMarkTM. Svec D., **Rusnakova V.**, Korenkova V., Kubista M., Editors: Nolan T., Bustin S.A., CRC Press , PCR Technology*:* Current Innovations, Third Edition (2013), 448 Pages.

Preconceptional Paternal Glycidamide Exposure Affects Embryonic Gene Expression: Single Embryo Gene Expression Study Following *In Vitro* Fertilization. Brevik, A., **Rusnakova, V.,** Duale, N., Slagsvold, H. H., Olsen, A. K., Storeng, R., Kubista, M., Brunborg, G. and Lindeman, B., Reproductive Toxicology 32, (2011), p. 463-471, IF=3,141.

Paternal Benzo a Pyrene Exposure Affects Gene Expression in the Early Developing Mouse Embryo. Brevik, A., Lindeman, B., **Rusnakova, V.**, Olsen, A. K., Brunborg, G. and Duale, N., Toxicological Sciences **129** (2012), p. 157-165, IF=4,328.

První publikace – kapitola v knize *PCR Technology: Current Innovations* se zabývá vypracováním funkčního protokolu pro vysokokapacitní přístroj Biomark. Jako první laboratoř vůbec jsme začali používat k detekci PCR produktu interkalační barvy (Kapitola 3.2.1). Bylo nutné optimalizovat přístroj i protokol a vybrat vhodnou detekční chemii. Je zde ukázán vliv objemu reakce a detekčního systému na Cq hodnotu. Zatímco pro klasické qPCR se uvažuje Cq hodnota pro detekci jedné molekuly kolem 36 (v závislosti na nastavení tresholdu), pro vysokokapacitní přístroj Biomark platí hodnoty o cca 10 cyklů nižší. To je dáno vyšší citlivostí detekční kamery. Dále je zde ukázána nutnost použití specifické interkalační fluorescenční barvy jako je Chromofy a EvaGreen, které se váží do každého velkého žlábku

dvoušroubovice DNA na rozdíl od nejčastěji používané barvy SYBRGreen, který se váže nespecificky. Na této práci jsem se podílela optimalizací qPCR protokolu.

Po optimalizaci protokolu pro měření 2304 reakcí najednou a po otestování preamplifikace na tkáních byl optimalizovány tyto protokoly pro jednotlivé buňky, a to na myších oocytech ve spolupráci s A. Brevikem (Department of Chemical Toxicology, Division of Environmental Medicine, Norwegian Institute of Public Health). Jeho publikace (Brevik et al., 2011; Brevik et al., 2012) se zabývají vlivem karcinogenů na kvalitu spermií a následně na vývoj embrya. Pro přístroj Biomark se používá specifická exponenciální preamplifikace, založená na multiplexové PCR reakci, kdy je cDNA namnožena pomocí omezeného množství cyklů (14-21). Počet cyklů vhodný k namnožení daného transkriptu je nepřímo úměrný k jeho množství, vysoce exprimované geny, jako je 18S, se nedoporučuje preamplifikovat z důvodů rychlého vyčerpání chemikálií během reakce. Zvýšení množství dNTP, polymerázy, primerů a hořečnatých iontů vede následně k větší tvorbě dimerů. Nebyla prokázána korelace mezi účinností preamplifikace a samotnou účinností primerů. Pro zjištění kvality preamplifikace se používá výpočet rozdílu Cq hodnot mezi preamplifikovaným a nepreamplifikovaným genem. Hodnota rozdílu by měla být konstantní pro různé koncentrace transkriptu a pro různé vzorky. Na těchto pracech jsem se podílela na preamplifikaci vzorků, qPCR analýze a analýze dat.

23 Dye-Based High-Throughput qPCR in Microfluidic Platform BioMark™

David Svec, Vendula Rusnakova, Vlasta Korenkova, and Mikael Kubista

CONTENTS

23.1 Introduction 323
23.2 Materials and Methods 326
23.3 Results and Discussion 327
23.3.1 Calibration of Background Signal 327
23.3.2 Data Uniformity 331
23.3.3 Melting Curve Analysis 334
23.4 Conclusions 335
Acknowledgments 336
References 336

23.1 INTRODUCTION

In biological and medical studies, large-scale gene expression and transcriptome analyses are essential tools.[1] Based on the goals of the investigator, different methods such as RNA microarrays, massive parallel sequencing, or high-throughput real-time quantitative PCR (qPCR) can be used. There are currently four platforms available for high-throughput qPCR: the BioMark from Fluidigm, the OpenArray from Life Technologies, the LightCycler™ 1536 from Roche, and the Smartchip™ from Wafergen (Table 23.1). These instruments offer the broadest dynamic range, high versatility of experiment setup, low cost per sample, and the lowest demands on the amount of sample material, allowing analysis of minute sample amounts, including single cells.[2,3] One focus of our center is gene expression profiling of single cells and we have used the BioMark as well as the OpenArray system. When we acquired the BioMark, most of our assay libraries were based on nonspecific dyes and so we explored using these established dye-based assays in microfluidic qPCR and extensively optimized the experimental protocol. In this report, we focus on the use of the BioMark system, which uses microfluidic chips called "dynamic arrays" for gene expression profiling, "digital arrays" for digital PCR, and reusable genotyping arrays for single nucleotide polymorphism (SNP) applications.

When using the dynamic array, each sample is mixed with each assay combinatorially in the integrated fluidic circuits (IFC) resulting in $48 \times 48 = 2304$ or $96 \times 96 = 9216$ qPCR reactions. The complete workflow takes 3–4 h and requires approximately 1 h of hands-on time. When compared to conventional qPCR, the high-throughput platform requires 24 or 96 times less effort and time (considering 384- or 96-well blocks). Furthermore, when using the dynamic array, all the reactions are run in parallel under identical thermal and optical conditions; hence, no interplate calibration is needed for studies of up to 96 samples and less bias is introduced. The automatic mixing in the

K12140_C023.indd 323 12/27/2012 3:37:23 PM

K12140_C023.indd 324 12/27/2012 3:37:23 PM

TABLE 23.1
High-Throughput qPCR Platforms Available as of Autumn 2011[4–7]

	Fluidigm 48.48 Dynamic Array (+Access Array[a])	Fluidigm 96.96 Dynamic Array	Fluidigm FR 48.48 (Genotyping)	Fluidigm 192.24 (Genotyping)	Fluidigm 12.765 Digital Array	Fluidigm 48.770 Digital Array	Life Tech. Openarray	Wafergen Smartchip	Roche LightCycler 1536
Priming time	11 min	20 min	11 min	10 min	6 min	30 min	—	—	—
Loading time	60 min	95 min	60 min	30 min	40 min	40 min	30 min	>10 min	10–20 min
Number of samples	48	96	48 (reusable)	192	12	48	12–48 (3×)	1–384	1–384
Assays per sample	48	96	48 (reusable)	24	Single/multiplex	Single/multiplex	18–224 (3×)	1–1296	1–1536
qPCR reactions	2304[a]	9216	2304	4608	9180	36,960	3072 (3×)	5184	1536
Minimum input/ sample	5 µL	5 µL	5 µL	4 µL	8 µL	4 µL	3–5 µL	100–400 nL	500–2000 nL
Reaction volume	10 nL	6.75 nL	8 nL	8 nL	6 nL	0.85 nL	33 nL	100 nL	500–2000 nL
Detection	Probe/dye	Probe/dye	Probe	Probe	Probe/dye	Probe/dye	Probe/dye	Probe/dye	Probe/dye
Loader	MX	HX	WX	RX	MX	MX	AccuFill System	Nanodispenser	Innovadyne™
Launched	Spring 2007	Fall 2008	Fall 2010	May 2011	Fall 2006	Spring 2009	Spring 2009	Spring 2010	Summer 2009

[a] The Access array is for downstream sequencing.

loading station (IFC controllers with type names MX, HX, WX, and RX, specific for particular arrays) dramatically reduces manual liquid handling. Only 48 + 48 pipetting steps are needed to prepare 2304 qPCRs and 96 + 96 to prepare 9216 qPCRs. Pipetting error can be further reduced by using automatic dispensers and multichannel pipettes, as the arrays are compatible with standard footprints. The cost saving on reagents is substantial, since the BioMark consumes 200 times less reagents and sample material than conventional microtiter plate-based instruments running 20 µL reactions. The volume of the sample mix, composed of mastermix, including dsDNA binding dye + loading buffer + cDNA + ROX, pipetted in an array inlet is 5 µL, which is distributed into 48 or 96 reaction chambers. The reaction volume is 10 nL in the 48 × 48 dynamic array and 6.75 nL in the 96 × 96 dynamic array (Table 23.1).

Nanoliter-scale qPCR requires concentrated samples with a sufficient number of template molecules to allow homogeneous distribution into the reaction chambers. The reaction volumes vary somewhat among the microfluidic arrays, but are typically 2000 times smaller than conventional 20 µL qPCRs. In practice, this usually requires that the samples are preamplified.[8] Preamplification is required because, for example, if 2 µL of cDNA containing 30 copies or more of a target gene is used in a conventional 20 µL qPCR, the reproducibility of technical replicates will be satisfactory. But if the same sample is used in a 48 × 48 array, each chamber will contain an average of 0.015 copies. Only a few chambers will contain target, while most will be empty. Hence, the data would not be reproducible.[9] Poisson variability is higher in the BioMark than in conventional systems due to low template concentration yielding high Cq values (Figure 23.1).[10,11]

When the first BioMark dynamic array was launched in 2007,[12] hydrolysis probes were used exclusively for gene expression analysis. The cost of a probe assay is typically 5–10 times that of primers only, which can add a substantial amount to the cost of entire profiling experiments. Testing and validating assays requires substantial resources when using high-throughput profiling. The

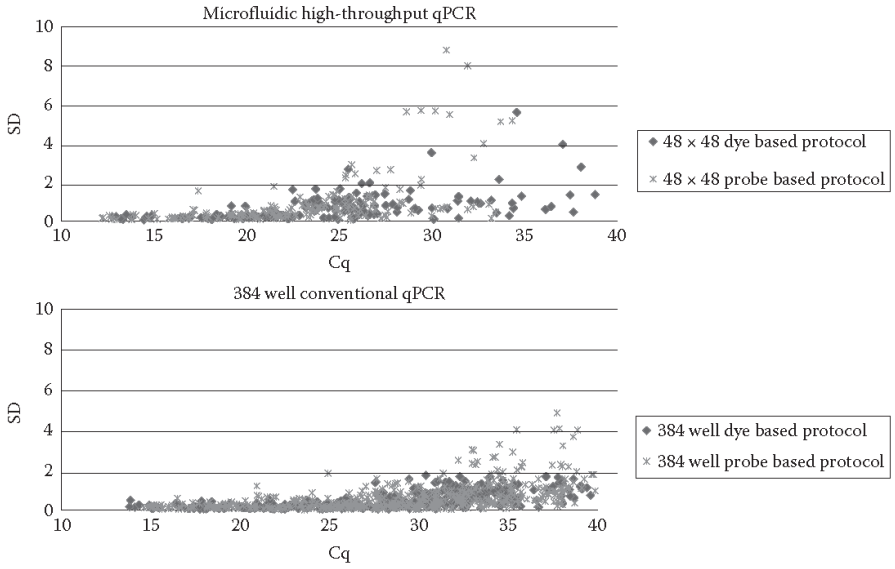


FIGURE 23.1 (**See color insert.**) Dependence of standard deviation on Cq in conventional and microfluidic qPCR systems. In the conventional system, 10 µL qPCR were run in a 384-well plate. The microfluidic system was a 48 × 48 array with 10 nL qPCRs. The same samples and assays were analyzed, although for the BioMark the material was preamplified through 14 cycles.

assay must be tested to determine efficiency, sensitivity, and specificity.[9] Assays in which primer dimers are formed should be excluded as they will result in unreliable data due to poor sensitive and compromised reproducibility. This factor is frequently ignored by those using probes for template detection, arguing that nonspecific products do not generate signal. However, parallel amplifications compete for reagents, which leads to an underestimation of the specific target.[13] When using dye-based assays, melt profiles of PCR products can be recorded after each run with no additional cost and these can be used to reveal the presence of competing reactions. Studies of SNPs and methylation patterns can be done at minimal cost by using a similar strategy but by applying post-amplification, high-resolution melt analysis (HRM). If sequencing of the PCR product is required, the 48 × 48 access array can harvest all PCR products in one pool, ready to be sequenced, instead of microinjecting the nanoliter-scale volumes one by one.

There are some important technical differences between conventional systems using microtiter plates and the microfluidic platforms. The BioMark optical system uses a xenon lamp, optical filters (by default: three out of four filter positions are mounted with filters that are compatible with detection of ROX (6-carboxy-X-rhodamine), FAM (5-carboxyfluorescein), VIC/JOE (6-carboxy-4′,5′-dichloro-2′,7′-dimethoxyfluorescein, a charged coupled device (CCD) 9M pixel camera, and a cooling system.[4] The passive reference dye, ROX, is always used in the BioMark platform to identify the chambers and to normalize the signal (Figure 23.3). Before the start of cycling, the CCD camera is focused onto the wells. Then, the autoexposure function sets an exposure time for each fluorescence channel based on the background signal. The objective is to set a baseline such that even a minute increase in fluorescence can be detected with high sensitivity by the CCD chip. The autoexposure calibration is performed by collecting several images with different exposure times. An optimum exposure time is calculated and is then used for the acquisition of qPCR signals. With increasing background, the exposure time to register fluorescence during qPCR is shortened. If the exposure time during acquisition is too short, images are underexposed, which leads to poor-quality data and, in the worst-case scenario, amplification, which is not detected above the background.

The calibration setting protocol dates back to the time when the BioMark was introduced in November 2006,[14] and only probe-based assays were used. This included using a calibration temperature of 20°C, and initially the calibration temperature was not adjustable. The calibration settings for Fluidigm's EvaGreen protocol based on the use of Taqman Gene Expression Master Mix® (TGEMM, Life Technologies) use a different approach to that presented here. This requires that a new detector profile is created in the data collection software artificially to reprogram the sensitivity of the CCD. This setting is only optimum for the combination of EvaGreen and TGEMM and can lead to under- or overexposure of the CCD if used for other combinations of dye and mastermix. Here, we present a new calibration approach of the CCD in the BioMark that does not depend on the dye and master mix used, and included the use of the novel dye Chromofy™.

23.2 MATERIALS AND METHODS

Intensity of fluorescence, loading uniformity, and the performance of reporter dyes in the microfluidic qPCR assays using the 48 × 48 dynamic arrays were tested. A GAPDH qPCR assay developed within the SPIDIA project (www.spidia.eu) with efficiency >90% and no primer dimer formation in conventional qPCR was used for tests on the BioMark platform with dye-based and probe-based detection and using several master mixes. The same primer pair was used for all studies: human GAPDH assay targeting NM_002046.3, 91 bp amplicon, Fwd primer: CCTCCACCTTTGACGCT, Rev primer: TTGCTGTAGCCAAATTCGTT, probe FAM-BHQ1: AGCTTGACAAAGTGGTCGTTGAGGGCAATG. Primers (MWG Operon) were desalted and the hydrolysis probe was HPLC purified (MWG Operon). Final concentrations in the reaction chamber were 400 nM of each primer, 200 nM hydrolysis probe, and recommended amount of 2× DA Assay Loading Reagent (Fluidigm). The selection of mastermixes used with the probe were ABI Taqman Gene Expression MasterMix (TGEMM) according to Fluidigm's protocol. Reporter dyes were used

in KAPA SYBR Fast Universal mix™ and Finnzymes SYBR Flash mix™. These mixes were supplied either with SYBR Green I or without SYBR Green I to which we added Chromofy or EvaGreen. Chromofy 5000× (TATAA Biocenter, 6.85 mM) was diluted with nuclease-free water to a 400× working stock. The final concentration in the reaction chamber was 6×. EvaGreen 20× (Biotium) was used in the final 1× concentration in qPCR. Tests were performed using the Universal human cDNA (BioChain Institute Inc.) in different concentrations or after preamplification: preamplfication of 20 genes (SPIDIA) was carried out in Biorad IQ Supermix™ and contained 5 µL of universal cDNA in 50 µL reaction volume, 50 nM each primer final concentration, and used a amplification protocol consisting of 3 min 95°C followed by 18 cycles of 20 s 95°C, 1 min 58°C, and 30 s 72°C. Test reactions consisting of 5 µL of sample mix contained 0.5 µL of 20× DA sample loading reagent (Fluidigm) and 2.5 µL of mastermix. Concentrations of ROX reference dye (50× Invitrogen, stock 25 µM) in test arrays varied from 25, 50, to 125 nM in qPCR. Prior to loading of samples and assay reagents into the inlets, the array was primed in the NanoFlex™ 4-IFC Controller (Fluidigm), 5 µL of sample mix was pipetted into sample inlets, and 5 µL of assay mix into the assay inlets. Technical replicates were performed on the assay as well as on the sample level to evaluate repeatability. All dynamic arrays were thermocycled using the BioMark Real-Time PCR System (Fluidigm) under the following conditions: calibration temperature was manually set to 20°C or 72°C made possible with a change in the Windows registry using a *.reg file provided by Fluidigm. The ROX/FAM-MGB detection channel was selected. The temperature protocol was 600 s 95°C activation, 40 cycles followed by either a two-step probe protocol: 95°C 20 s, 60°C 60 s (signal acquisition at 60°C), or a three-step dye protocol: 95°C 20 s, 58°C 20 s, and 72°C for 30 s (signal acquisition at 72°C). Cycling was followed by melt curve analysis with amplicon incubated from 65°C to 95°C at 0.5°C increments and 1 s wait between each increment. Data were collected using the BioMark Data Collection software with separate protocols for the qPCR and for the melting analysis. Data were analyzed using the Fluidigm Real-Time PCR Analysis v3.0.2. and the Fluidigm Melting Curve Analysis v.3.0.2. Linear baseline correction was used and threshold was manually adjusted to be as close as possible to the background. A set of validated assays that had been designed for the EU project SPIDIA (www.spidia.eu) were used for comparison of the BioMark data with measurements performed on a conventional qPCR platform (Roche LightCycler 384). Images of PCR tubes with GAPDH PCR product were taken using stereomicroscope with green fluorescent protein filter. The melting temperature of GAPDH product in KAPA SYBR mix is 81°C.

23.3 RESULTS AND DISCUSSION

When the BioMark system was installed in our lab, several high-throughput projects were planned based on panels of established dye-based assays. The first choice of dye was SYBR Green™, which is the most popular qPCR reporter dye. However, SYBR Green I turned out to perform very poorly on the Biomark (Figure 23.3). After extensive screening of other dyes, we identified Chromofy as a strong, alternative candidate. The performance of the assays (hydrolysis probes, Chromofy) was evaluated on standard human cDNA at different concentrations. The uniformity of the loading process was tested using 48 assay replicates (GAPDH) on cDNA at different concentrations analyzed in sample triplicates.

23.3.1 Calibration of Background Signal

While DNA binding dyes usually generate higher-plateau fluorescent signal than probes and a larger ΔRn value in conventional qPCR systems,[15–17] we found the opposite to occur in the initial runs with SYBR Green on the BioMark. The poor signal had a major impact on data quality in terms of low reproducibility and poor sensitivity. We found that this was caused by inappropriate calibration settings of the CCD camera, which was optimized for probe-based detection. The calibration temperature was set to 20°C by default. The background fluorescence of free probes

does not change appreciably with temperature. However, in a reaction mix containing free dye, there is negligible fluorescence, although the reaction mix contains primers and possibly other residual DNA to which the dye may bind. This bound dye fluoresces and its intensity decreases with temperature (Figure 23.2).[15,18]

The strong background signal of the dye at 20°C during calibration results in short exposure time during acquisition, which leads to underexposure during qPCR. To improve the quality of measured data in dye-based reads, we performed background calibration at elevated temperature. For the dyes, we found calibration at 72°C to perform well (Figure 23.3). We were not able to test higher calibration temperature because this is limited in the BioMark Data Collection software. This limit depends on the thermal protocol and for some reason that is unclear, this limit is set to

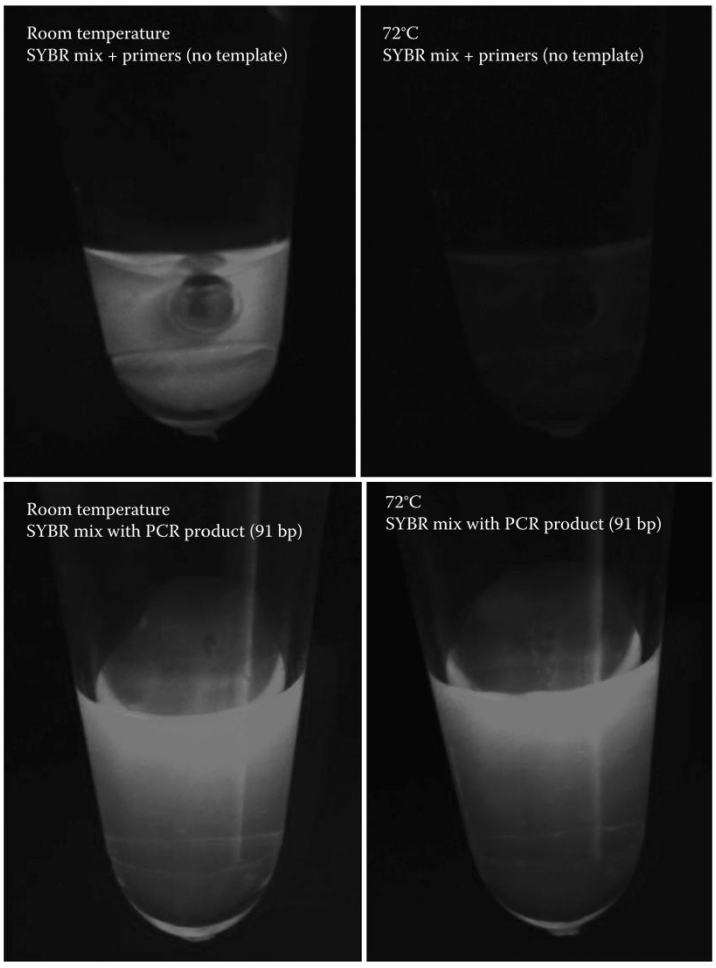


FIGURE 23.2 (**See color insert.**) Dye fluorescence in conventional qPCR at different temperatures. Left: two images show a tube with 10 µL SYBR Green master mix (KAPA Universal SYBR mix) in the presence of primers (400 nM per GAPDH primer) before cycling at room temperature and at 72°C. Right: two images show the tube after cycling in the presence of PCR product at room temperature and at 72°C.

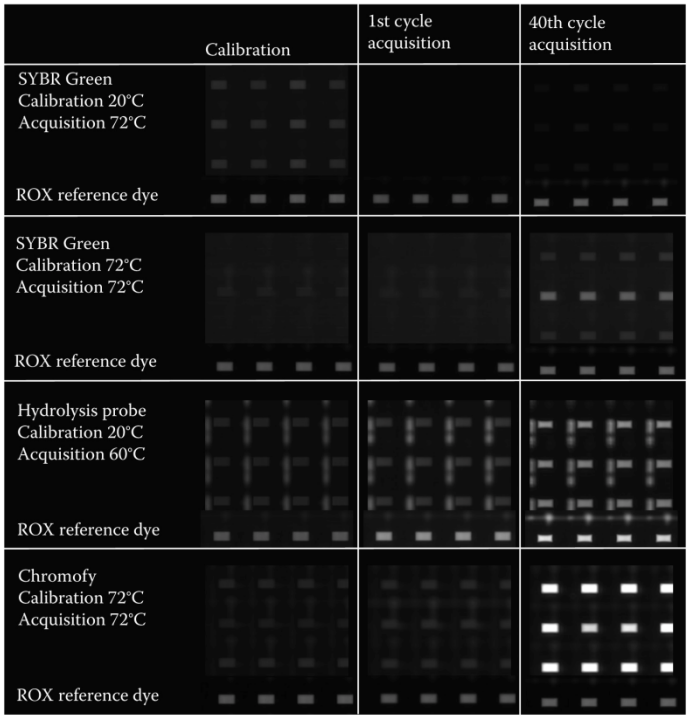


FIGURE 23.3 Camera images showing 12 qPCR chambers in a 48 × 48 dynamic array. Top row: 20°C calibration. Dye signal is strong during calibration and exposure time is set to 0.39 s. During acquisition at 72°C, no signal is detected after one cycle, and even after 40 cycles, the signal is very weak. Second row: 72°C calibration. Dye signal is weak during calibration and exposure time is set to 3.6 s. Already after one cycle, chambers are distinguishable, and after 40 cycles, strong signals develops. Third row: Using hydrolysis probes, fluorescence at calibration temperature is low and exposure time is set to 2.15 s. Bottom: Chromofy signal calibrated at 72°C, which results in 1.94 s exposure time. ROX reference dye signal shows that fluorescence is independent of calibration temperature. First to fourth row concentration and exposure time: 125 nM, 0.33 s; 125 nM, 0.23 s; 25 nM, 2.15 s; 125 nM, 0.119 s.

the temperature at which signal is acquired. In most three-step PCR protocols, this is at 72°C, while for two-step probe-based PCR, acquisition is typically at 60°C.

To demonstrate the effect of calibration temperature, three dynamic arrays with identical sample and assay setup were used, but signal was generated using either hydrolysis probes, SYBR Green I or Chromofy (Table 23.2). For each array, autoexposure at 20°C was performed. The run was then aborted and new autoexposure was measured at 72°C followed by running qPCR. We observed that the exposure time was almost 10 times longer after calibrating at 72°C compared to calibration at 20°C. There was a slight decrease of exposure times when using hydrolysis probes and including ROX, which suggests that there was only a minor increase of background fluorescence from these labels with decreased temperature. ΔFAM is the difference between initial raw fluorescence signal and the signal measured after the last cycle. Raw fluorescence data measured in the reporter channel are normalized by dividing by the raw data of the passive ROX reference channel and are denoted Rn. Rn values and raw fluorescence data were accessed using the real-time PCR analysis software module (Fluidigm). The maximum values in initial and terminal cycles within the array were read.

TABLE 23.2
Effect of Calibration Temperature on Detection Chemistry

	Hydrolysis Probe—Mix A				Eva Green—Mix A				SYBR Green—Mix B				Chromofy—Mix B			
	ROX (s)	FAM (s)	ΔFAM Raw Signal	Max ΔRn	ROX (s)	FAM (s)	ΔFAM Raw Signal	Max ΔRn	ROX (s)	FAM (s)	ΔFAM Raw Signal	Max ΔRn	ROX (s)	FAM (s)	ΔFAM Raw Signal	Max ΔRn
20°C	1.11	0.63	x	x	1.13	0.22	x	x	1.21	0.391	x	x	1.01	0.53	x	x
72°C	0.88[a]	0.37[a]	4000[a]	0.3[a]	0.84	3	16,000	0.6	0.9	3.15	1600	0.12	0.79	2.73	24,000	2.6

Note: Identical setup of assays and samples are compared. Mix A = TGEMM, Mix B = KAPA Fast Universal.
[a] 60°C.

K12140_C023.indd 330 12/27/2012 3:37:29 PM

(+)Rn represents fluorescence values for a reaction that contain all components, including template, while (–)Rn is the background reading for a negative sample. For positive reactions, the (–)Rn signal is obtained from the early cycles before fluorescence starts to increase. The normalized signal ΔRn = (–)Rn + (+)Rn is the fluorescence signal generated using particular ROX concentration and run settings (exposure times).

The qPCR was completed using calibration readings at 72°C and data are presented in Table 23.2. To demonstrate higher sensitivity with calibration at lower temperature, we performed a second study using a different set of dynamic arrays (Table 23.3).

One sample was analyzed using SYBR Green I as the reporter for a set of assays using different calibration temperatures in the qPCR (Table 23.4). Samples were analyzed in triplicates using 48 assays. This generated a total of 144 data points. The assays were loaded in all chips in the same order, which minimized the uniformity bias experienced with SYBR Green I (see below). We observed lower Cq values, reduced standard deviations of technical replicates, and higher call rates after calibration at 72°C than with calibration at 20°C temperature.

23.3.2 Data Uniformity

Sample mixes are conventionally loaded into the IFCs from one edge of the array to fill all the reaction chambers. The surface of the polydimethylsiloxane (PDMS)[19] that contacts the sample mix is much greater than the surface of the polypropylene in conventional microtiter plates. To validate the dye-based protocol, a uniformity test was performed, with the identical GAPDH assay loaded into all 48 assay inlets. Finnzymes Dynamo master mix was used to test the performance of SYBR Green I in first array (1 μL of 20× diluted universal cDNA per 5 μL sample mix) and Chromofy performance was tested on a second array (1 μL of 40× diluted universal cDNA per 5 μL sample mix). Both arrays were run using autoexposure at 72°C. Figure 23.4 shows how Cq varies from the

TABLE 23.3
Effect of Calibration Temperature for Different qPCR Chemistries

Calibration Temperature (°C)	ROX (s)	FAM (s)	Probe/Dye	Master Mix	Array	ΔFAM Raw Signal	Max ΔRn
20	2.159	1.893	Probe	A	48 × 48	25,000	0.4
60	1.516	1.635	Probe	A	48 × 48	19,000	0.9
20	2.21	0.391	SYBR Green	C	48 × 48	1100	0.14
60	1.87	3.15	SYBR Green	C	48 × 48	6000	0.8
72	1.75	3.35	SYBR Green	C	48 × 48	5500	0.7
72	0.34	2.33	Chromofy	C	48 × 48	36,000	2
72	0.328	1.386	Chromofy	C	96 × 96	46,000	2.4

Note: Mix A = TGEMM, Mix C = Finnzymes SYBR Flash Mix.

TABLE 23.4
Sensitivity and Reproducibility

	20°C Calibration	60°C Calibration	72°C Calibration
Mean SD of triplicates	1.14	0.65	0.26
Missing values (%)	15.28	34.30	4.17

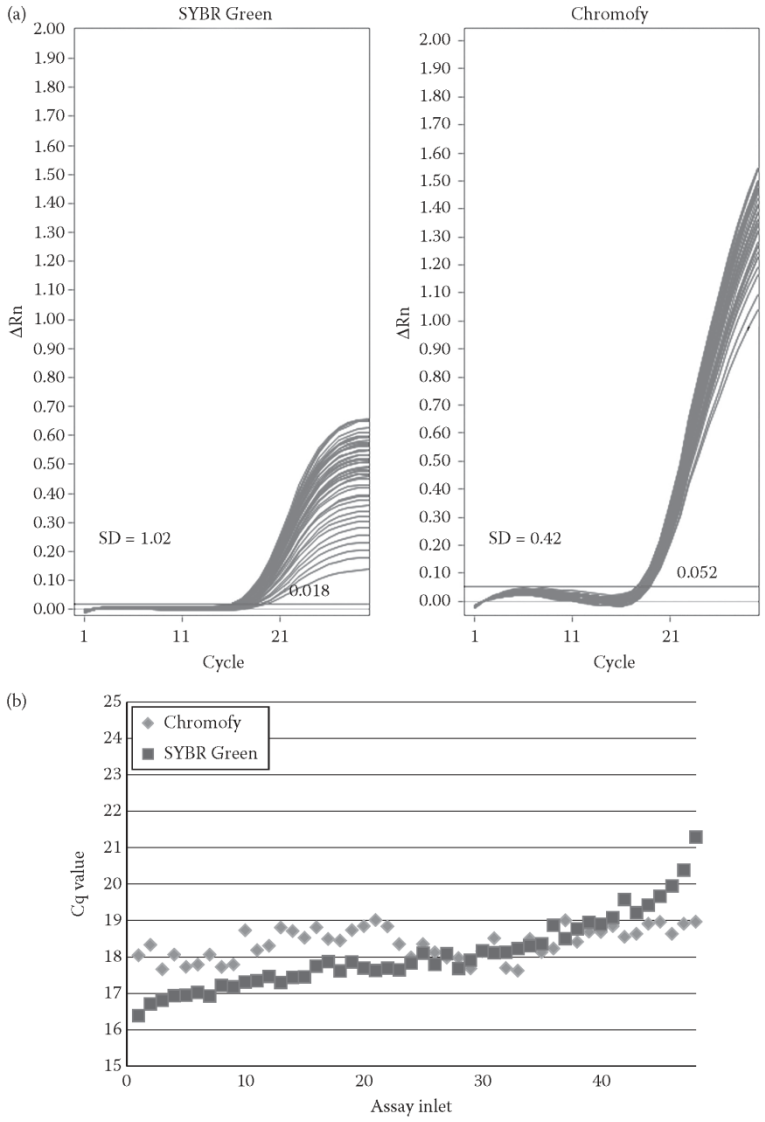


FIGURE 23.4 Amplification curves (a) and Cq values (b) showing uniformity of SYBR and Chromofy assays in 48 × 48 arrays. All 48 assay inlets were loaded with GAPDH primers.

loading influx. For the SYBR Green I, reporter Cq increases with five cycles from the inlet to the last-filled reaction chamber. For Chromofy, the increase is much more modest (Table 23.5). This drift in Cq probably reflects adsorption of the reporter dye to the PDMS of the IFCs (Figure 23.5).[19]

Q1

After optimizing the calibration of the CCD and reagent conditions of the Chromofy protocol, we compared the performance with the recommended probe protocol (Figure 23.6, Table 23.6).

TABLE 23.5
Uniformity of SYBR and Chromofy Assays Run in a BioMark 48 × 48 Channel

	Mean Cq (48 Replicates)	SD (48 Replicates)	ΔCq (Max – Min)	ΔCq (Mean of Five Left Chambers – Mean of Five Right Chambers)
SYBR Green	18	1.01	4.9	3.38
Chromofy	18.3	0.4	1.3	0.9

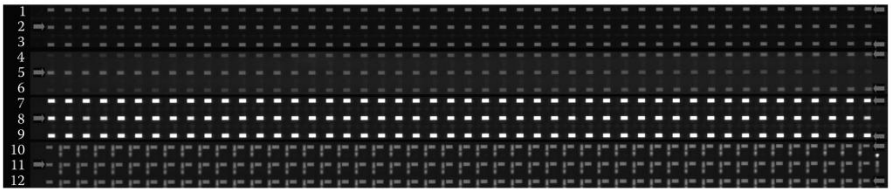


FIGURE 23.5 CCD images showing selected qPCR chambers in ROX and FAM channels. Lines 1–3 are ROX signals and lines 4–6 show SYBR Green signal. Lines 7–9 show Chromofy signal and lines 10–12 hydrolysis probe signal. Blue label indicates inlets. **Q2**

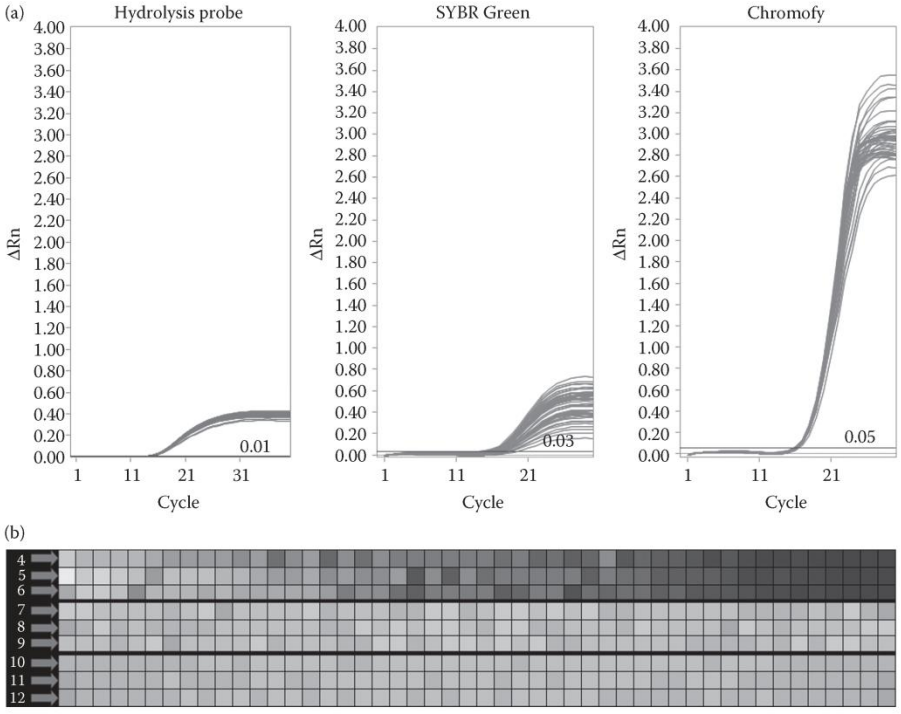


FIGURE 23.6 (**See color insert.**) Amplification curves (a) and heatmap produced with the BioMark software. Blue squares represent high Cqs, yellow low Cqs. Lines 4–6 show SYBR Green signal, lines 7–9 show Chromofy, and lines 10–12 show probe signal.

TABLE 23.6
Reproducibility and Loading Uniformity of Probe- and Dye-Based Assays

	Mean Cq (48 Replicates)	SD (48 Replicates)	ΔCq (Max – Min)	ΔCq (Mean of Five Left Chambers – Mean of Five Right Chambers)
Probe	14.64	0.18	0.80	0.09
SYBR Green	16.70	1.00	4.96	3.44
Chromofy	16.11	0.14	0.75	0.30

TABLE 23.7
Effect of Calibration Temperature on Melt Curve Analysis

Calibration Temperature (°C)	ROX (s)	FAM (s)	Probe/Dye	Master Mix	Array	Decrease of FAM Fluorescence	–dRn/dT	Average T_m (°C) GAPDH Product	SD (48 Replicates)
20	2.65	1.76	SYBR Green	C	48 × 48	9000	0.02	N/A	N/A
60	1.38	3.64	SYBR Green	C	48 × 48	20,000	0.03	N/A	N/A
65	1.58	3.13	SYBR Green	C	48 × 48	21,000	0.03	81.31	0.16
65	0.3	0.87	Chromofy	C	48 × 48	18,000	0.03	83.52	0.21
65	0.1	0.33	Chromofy	B	48 × 48	21,000	0.03	86.71	0.19
65	0.35	0.22	Chromofy	C	96 × 96	17,000	0.03	85.05[a]	0.14[a]

Note: Mix A = TGEMM, Mix B = KAPA Fast Universal, Mix C = Finnzymes SYBR Flash Mix.
[a] 96 replicates.

The Chromofy assay shows comparable reproducibility and loading uniformity to the probe-based assay. The optimized conditions were for the probe assay: 0.625 µL of universal cDNA per 5 µL of sample mix, for the SYBR Green I assay: 0.05 µL cDNA per 5 µL sample mix, and for the Chromofy assay: 0.2 µL cDNA per 5 µL sample mix.

23.3.3 Melting Curve Analysis

Optimized calibration temperature also improves the readings of fluorescence during melt curve analysis. The BioMark data collection software has a surprising limit for the maximum calibration temperature that can be used. This is the temperature at which acquisition of the HRM signal starts. For example, if we run a melt curve analysis from 65°C to 95°C, the highest possible calibration temperature is 65°C. The performance of HRM analysis under various conditions is summarized in Table 23.7.

The melting temperature (T_m) is fairly homogeneous across the array when using SYBR Green I or Chromofy reporter (Figure 23.7), demonstrating that it is not appreciably affected by the progressive dye adsorption revealed above. Poor reproducibility is only observed at the edges of the array and is most likely due to evaporation through the partially permeable MSDS caused by extensive heating during thermal cycling.

PCR products can be categorized based on the melting profiles when using version 2.6.× (and higher) of the Fluidigm Melting Curve Analysis. This is useful to identify qPCRs with nonspecific amplification products such as primer dimers. An example of this application on single-cell material is shown in Figure 23.8.

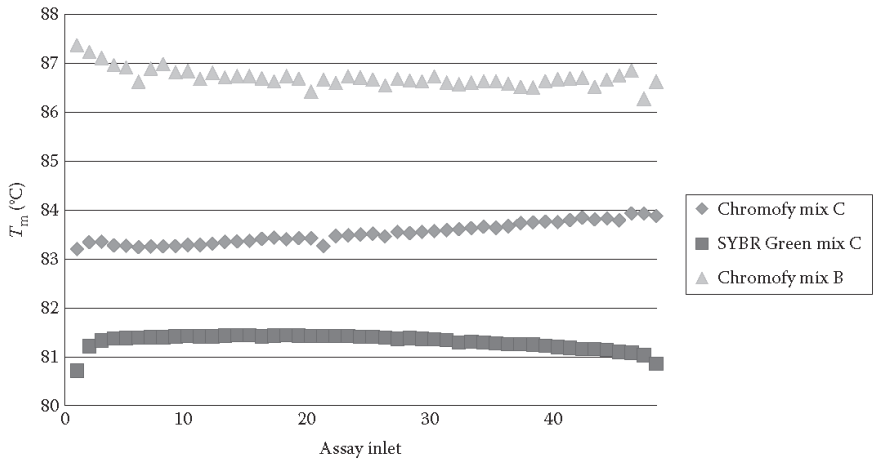


FIGURE 23.7 **(See color insert.)** T_m uniformity across the 48 × 48 array. Mix B = KAPA Fast Universal, mix C = Finnzymes SYBR Flash mix.

23.4 CONCLUSIONS

In the BioMark IFC, the fluorescence reporter signal collected during qPCR is normalized to the signal from the passive reference dye ROX based on the background signals collected before the onset of the experiment. In the standard BioMark protocol, the background signal is collected at ambient temperature. ROX is a negatively/uncharged dye at neutral pH that does not interact with DNA. It has a rigid structure and constant, very high fluorescence quantum yield that is essentially independent of temperature. Hydrolysis probes are most frequently based on fluorescein (FAM) reporter and rhodamine quencher (TAMRA) dyes, which have similar properties. Therefore, when using hydrolysis probes, there is essentially no dependence of fluorescence intensity on temperature. Since neither ROX nor hydrolysis probe fluorescence depends on temperature, it is possible to record background signal at ambient temperature even though fluorescence during PCR is recorded at substantially higher temperature.

Most free DNA binding dyes used as reporters in qPCR are cationic asymmetric cyanines that bind preferentially to double-stranded DNA but show residual binding to single-stranded DNA.[18] The strong fluorescence induced upon binding is induced by the dye being held in a tight conformation, usually in the groove of the double helix or squeezed in between bases in intercalative mode. The tight binding restricts internal motion within the dye that dissipates excitation energy through nonradiative processes, giving rise to intensive fluorescence. The tightness of the binding is temperature dependent. At elevated temperatures, binding is looser, thus increasing nonradiative decay and thereby reducing fluorescence emission. The background fluorescence of asymmetric cyanine dyes such as SYBR Green I and Chromofy at ambient temperature is therefore much lower than their fluorescence at the temperatures used to record fluorescence during qPCR. As demonstrated in this report, calibration of the CCD at ambient temperature when using dye reporters results in poor quality of the fluorescence signals measured during the qPCR, and the performance can be dramatically improved by calibrating the CCD at high temperature. We also show that the common reporter dye SYBR Green I adsorbs to the PDMS of the ICS resulting in a large drift of Cq values with the distance from the inlet. The adsorption is much less with dye Chromofy.

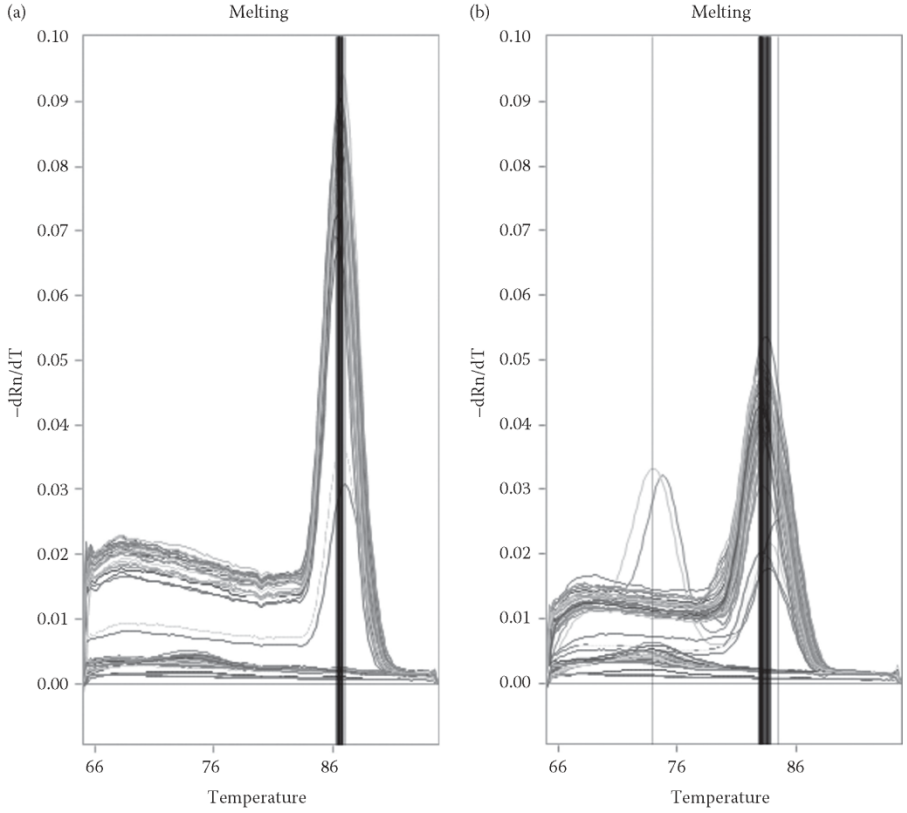


FIGURE 23.8 High-throughput melting analysis of PCR products. (a) Only specific PCR products formed. (b) Some reactions have formed aberrant PCR products characterized by lower T_m.

ACKNOWLEDGMENTS

This work was supported in part by the European Community Seventh Research Framework Programme project SPIDIA (http://www.spidia.eu) Grant Agreement no. 222916 and to the Biotechnology Institute of the Czech Academy of Sciences and the TATAA Biocenter and in part by AV0Z50520701 granted by Ministry of Youth, Education and Sports of the Czech Republic. David Svec is in part supported by the PhD fellowship from Charles University.

REFERENCES

1. Tang, F., Lao, K. and Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat Methods* **8**, S6–S11, 2011.
2. Stahlberg, A. et al. Defining cell populations with single-cell gene expression profiling: Correlations and identification of astrocyte subpopulations. *Nucleic Acids Res* **39**, e24, 2011.
3. Stahlberg, A. and Bengtsson, M. Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods* **50**, 282–288, 2010.
4. BioMark™ HD System.

5. OpenArray® Real-Time PCR Platform.
6. The SmartChip Cycler.
7. LightCycler® 1536 Real-Time PCR System.
8. Mengual, L., Burset, M., Marin-Aguilera, M., Ribal, M.J. and Alcaraz, A. Multiplex preamplification of specific cDNA targets prior to gene expression analysis by TaqMan arrays. *BMC Res Notes* **1**, 21, 2008.
9. Bustin, S.A. et al. The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611–622, 2009.
10. Rutledge, R.G. and Stewart, D. Assessing the performance capabilities of LRE-based assays for absolute quantitative real-time PCR. *PLoS One* **5**, e9731. **Q3**
11. Wang Z, S.J. Determination of target copy number of quantitative standards used in PCR-based diagnostic assays, in *Gene Quantification* (ed. F. F) pp. 31–43 (Birkhäuser, Boston, 1998). **Q4**
12. Heid, C.A., Stevens, J., Livak, K.J. and Williams, P.M. Real time quantitative PCR. *Genome Res* **6**, 986–994, 1996.
13. Lind, K., Stahlberg, A., Zoric, N. and Kubista, M. Combining sequence-specific probes and DNA binding dyes in real-time PCR for specific nucleic acid quantification and melting curve analysis. *Biotechniques* **40**, 315–319, 2006.
14. Fluidigm's BioMark™ System Sales Surpass the Century Mark.
15. Wittwer, C.T., Herrmann, M.G., Moss, A.A. and Rasmussen, R.P. Continuous fluorescence monitoring of rapid cycle DNA amplification. *Biotechniques* **22**, 130–131, 134–138, 1997.
16. Newby, D.T., Hadfield, T.L. and Roberto, F.F. Real-time PCR detection of Brucella abortus: A comparative study of SYBR green I, 5′-exonuclease, and hybridization probe assays. *Appl Environ Microbiol* **69**, 4753–4759, 2003.
17. Arikawa, E. et al. Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC Genomics* **9**, 328, 2008.
18. Jan Nygren, N.S. and Kubista, M. The interactions between the fluorescent dye thiazole orange and DNA. *Biopolymers* **46**, 39–51, 1998, 1997. **Q5**
19. Perkel, J.M. Microfluidics—Bringing new things to life science. in *Science Magazine* (AAAS OPMS, 2008).

Reproductive Toxicology 32 (2011) 463–471

Contents lists available at SciVerse ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

Preconceptional paternal glycidamide exposure affects embryonic gene expression: Single embryo gene expression study following *in vitro* fertilization

Asgeir Brevik[a,*], Vendula Rusnakova[b], Nur Duale[a], Hege Holte Slagsvold[a], Ann-Karin Olsen[a], Ritsa Storeng[c], Mikael Kubista[d], Gunnar Brunborg[a], Birgitte Lindeman[a]

[a] Norwegian Institute of Public Health, Division of Environmental Medicine, Department of Chemical Toxicology, P.O. Box 4404, Nydalen, N-0403 Oslo, Norway
[b] Dept. of Gene Expression, Institute of Biotechnology AS CR, v.v.i., Videnska 1083, Prague 4, 142 20, Czech Republic
[c] Oslo University Hospital - Rikshospitalet, 0027 Oslo, Norway
[d] Tataa Biocenter, Odinsgatan 28, 411 03 Göteborg, Sweden

ARTICLE INFO

Article history:
Received 2 March 2011
Received in revised form 10 August 2011
Accepted 17 September 2011
Available online 29 September 2011

Keywords:
Single-cell gene expression
Glycidamide
Acrylamide
Mouse embryo

ABSTRACT

Recognition of early determinants of disease onset has sparked an interest in paternally transmitted factors and their impact on the developing embryo. Acrylamide (AA), a widely distributed xenobiotic compound, is converted to its active metabolite glycidamide (GA) by the CYP2E1 enzyme. Based on its capacity to induce dominant lethal mutations, we hypothesized that paternal GA exposure would have a negative impact on embryonic genome activation, via GA-DNA and protamine adducts persisting in the fertilizing sperm. Using a combination of *in vitro* fertilization (IVF) techniques and RT-qPCR single embryo gene expression (SEGE), we studied the expression of key DNA repair genes and genes important for embryo development, at the 1-, 2-, 4- and 8-cell stage of the developing mouse embryo. Compared to controls paternal GA-exposure gave rise to an altered pattern of embryonic gene expression, with an initial reduced expression at early stages followed by increased expression at the 8-cell stage.
© 2011 Elsevier Inc. All rights reserved.

1. Introduction

A considerable part of infertility problems among couples is believed to be caused by male factors [1], and there is evidence of a decline in semen quality in many industrialized countries [2,3]. A parallel increase in testicular cancer observed over the last decades [4–6] raises concern over the impact of a modern lifestyle on the male reproductive system. Environmental factors are believed to be implicated in both the diminishing male reproductive health and the increased cancer incidence, but the part of the modern environment posing the greatest threat towards the male reproductive system has not been determined.

Germ line gene–environment interactions are currently being explored in an attempt to understand the dissonance between male reproductive health and the modern environment. Concern has been expressed on the possibility that paternal DNA damage is propagated across multiple generations. Post-meiotic male germ cells are sensitive to induction of heritable genomic damage particularly during the last few weeks of spermatogenesis [7,8]. All major DNA repair pathways seem to be less functional in late spermatids and sperm [9–12]. We have previously reported on excision repair activities in different testicular cell types [10,13–15] confirming that DNA lesions are not repaired in late spermatids and sperm. High levels of DNA lesions are thus present in the sperm [15–17]. Embryonic development may be compromised when oocytes are fertilized with sperm containing higher than normal levels of DNA- and protamine-adducts.

Acrylamide (AA) is one of many chemicals that have been shown to induce DNA damage in the male germ line [18–20], dominant lethal mutations representing one critical consequence [20,21]. In addition to being an industrial chemical and a component of coffee and cigarette smoke, AA is formed by high temperature processing of glucose and, it is generally present at low levels in many carbohydrate rich fried foods like French fries, potato crisps, crisp bread, bread, and biscuits [22–24]. AA is oxidized to the reactive electrophilic epoxide glycidamide (GA) [25] by the CYP2E1 enzyme [26–28]. A large part of the damages attributed to AA exposure is believed to be caused by this reactive metabolite [26,28–30]. DNA and protamine alkylation have been suggested as mechanisms by which AA induces germ cell mutagenic effects [19,31,32], the most commonly reported DNA adducts being N7(2-carbamoyl-2-hydroxyethyl)-guanine (N7-GA-Gua) and to a lesser extent N3(2-carbamyl-2-hydroxyethyl)-adenine (N3-GA-Ade) [26,33–35].

Global activation of the embryonic genome constitutes the most critical event at early stages of mammalian development. Maternal proteins and RNAs support development after fertilization, whereas a number of zygotic and embryonic genes are expressed

* Corresponding author. Tel.: +47 2107 6452.
E-mail address: asgeir.brevik@fhi.no (A. Brevik).

0890-6238/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.reprotox.2011.09.005

in a stage specific manner leading to embryonic genome activation [36]. The maternal to zygotic transition can be subdivided into two interrelated processes; first, a subset of the maternal mRNAs and proteins is eliminated; second, zygotic transcription is initiated [37]. The maternal to zygotic transition is a highly coordinated and extremely complex biochemical symphony and it has been estimated in the mouse that about 15,700 genes are expressed during preimplantation development [38]. Paternal exposure to the anticancer alkylating agent cyclophosphamide has been shown to alter the expression of DNA repair genes in the rat pre-implantation embryo [39,40]. A relevant question is whether less potent and more widely distributed environmental chemicals can affect the developing embryo via the same exposure route.

Based on our previous findings of very high levels of DNA lesions in sperm [15,16] and the ability of GA to form DNA and protamine adducts, we hypothesized that paternal preconceptional exposure to GA would have an impact on early embryonic transcription and on the activation of the embryonic genome. We studied the expression of key DNA repair genes and genes important to embryo development following paternal acute exposure to GA seven days prior to fertilization. To our knowledge this is the first time RT-qPCR single zygote/embryo gene expression techniques are used in studies on reproductive toxicity. We demonstrate that the early embryonic transcription of multiple genes is affected after paternal germ cells exposure to GA. The RT-qPCR single embryo gene expression technique has potential as a tool to study early developmental toxicity.

2. Methods

2.1. The exposure of male mice from which sperm was derived for IVF

Exposed males (strain B6D2F1 from Charles River Laboratories, 8–12 weeks of age) received one i.p. injection of GA (61 mg/kg body weight) dissolved in phosphate buffered saline eight days prior to the IVF experiment. Timing of the exposure to GA was based on pilot studies and knowledge about the most susceptible stage of spermatogenesis with respect to dominant lethal mutations. Similarly aged control males received an equivalent volume of phosphate buffered saline. At the day of the IVF experiment males were killed by cervical dislocation. Cauda were surgically removed and collected in an eppendorf tube containing M2 medium (500 μl, Sigma). Using a pair of micro scissors a few incisions were made in the cauda and the sperm was allowed to disperse for 10 min in a small drop (250 μl) HTF medium (EmbryoMax, Millipore) under liquid paraffin (MediCult) before transfer to the IVF dishes. Experiments are based on oocytes from 75 females and sperm from 10 males (5 exposed and 5 controls) altogether.

2.2. Super-ovulation and in vitro fertilization (IVF)

Females (strain B6D2F1 from Charles River Laboratories, 4–6 weeks of age) were injected i.p. with pregnant mare serum hormone gonadotropine (PMSG, Folligon from Intervet) (5 IU) three days prior to the IVF procedure. Two days later (that is, the day before the IVF) animals received an additional i.p. injection of human chorionic gonadotropine (HCG, Ovitrelle from Serono) (5 IU). Mice were killed by cervical dislocation and oviducts were collected in M2 medium (Sigma). Egg clutches (10–20 oocytes) embedded in cumulus cells were extracted from each oviduct. Oocytes were transferred to IVF-dishes and incubated in a droplet of HTF sperm containing medium under liquid paraffin for 4.5 h (37 °C). Oocytes from one side of the animal were combined with sperm from gycidamide exposed animals and oocytes from the other side were combined with sperm from control animals. Hence, oocytes from all animals were present in both the control group and the exposed group. After 4.5 h the fertilized oocytes (zygotes) were washed 5× in KSOM medium (EmbryoMax Millipore) before they were transferred to a drop of KSOM (200 μl) in a petri dish (35 mm) under liquid paraffin (MediCult). Samples from the 1-cell stage were collected immediately after fertilization. The rest of the zygotes were allowed to grow to harvest at the 2-cell, 4-cell or 8-cell stage. Upon harvest zygotes/embryos were collected in micro tubes filled with 5 μl lysis medium (CelluLyser, TATAA) and then frozen at −70 °C.

2.3. Reverse transcriptase and cDNA synthesis

Samples where thawed before a total of 14.5 μl mastermix and reverse transcriptase enzyme (Roche transcriptor first strand cDNA synthesis kit, Cat no. 04896866001) was added to all samples. The samples were then incubated in a thermal cycler unit (Eppendorf mastercycler) according to the following protocol: 10 min at 25 °C, 30 min at 50 °C, 5 min at 85 °C, and then held at 4 °C. After the RT reaction was completed samples were again frozen at −70 °C.

2.4. Preamplification of cDNA

Preamplification was used to increase the number of template molecules. This is a necessary step because the cDNA synthesis does not yield sufficient number of molecular copies of the template molecules which can be analyzed with confidence in parallel singleplex reactions. A high similarity between the gene expression measurements of preamplified and non-preamplified samples has been reported [41]. Preamplification PCR was run in 20 μl volume containing 4 μl of cDNA, 2 μl of a mixture of all forward and reverse primers (500 nM each), 10 μl of AmpliTaq Gold360 Master mix (Applied Biosystems) and water. PCR primers were designed using the online Universal Probe Library System from Roche. A CFX 96 cycler (Bio-Rad) was used for the preamplification with the cycling conditions: polymerase activation at 95 °C for 10 min, followed by 18 cycles (95 °C 15 s, 59 °C 1 min and 72 °C 1 min). The product of the preamplification reaction was diluted to 80 μl and stored at −20 °C. The robustness of the preamplification was validated by comparing qPCR expression levels of Chek (highly expressed) and Mlh3 (typically low expression) with and without preamplification in tissues from liver, ovaria and embryo. The relative expression of the two genes was similar when analyzing data with and without preamplification.

2.5. High throughput qPCR

The sample reaction mixture had a volume of 5 μl and contained 1 μl of preamplified cDNA, 0.5 μl of SYBR Green Sample Loading reagent (Fluidigm), 2.77 μl AmpliTaq Gold360 Master mix (Applied Biosystems), 0.165 μl of Chromofy, diluted 1:25 (TATAA), 0.025 μl of ROX (Invitrogen). The primer reaction mixture had a final volume of 5 μl and contained 2.5 μl Assay Loading Reagent (Fluidigm) and 2.5 μl mixture of reverse and forward primers corresponding to a final concentration of 5 μM. The chip was first primed with an oil solution in the NanoFlex™ 4-IFC Controller (Fluidigm) to fill control wells of the dynamic array. Bubbles were carefully removed from the 5 μl reaction mixture and the mix was loaded into the sample wells, and 5 μl of the primer reaction mixtures was loaded into the assay wells of the dynamic array. The dynamic array was then placed in the NanoFlex™ 4-IFC Controller for automatic loading and mixing. After 55 min the dynamic array was transformed to the BioMark qPCR platform (Fludigm). The cycling program was 10 min at 95 °C for preactivation, followed by 30 cycles of denaturation at 95 °C for 15 s, annealing at 60 °C for 20 s, and elongation at 72 °C for 20 s. After completed qPCR cycling melting curves were collected between 60 and 95 °C with 0.5 °C increments.

2.6. qPCR basic data analysis

The BioMark qPCR platform is primarily designed for Taqman probes. Here we used next-generation Chromofy as an unspecific dye binding double stranded DNA with higher affinity and fluorescence compared to regular SYBR Green. Chromofy has excitation/emission spectra similar to FAM. Since there is no separate set up for Chromofy in the Biomark qPCR platform, we used FAM-MGB settings with slightly altered exposition temperature and a linear baseline correction derived from independent experiments (data not shown). ROX (emission at 645 nm) was always used for passive signal normalization.

2.7. Data preprocessing

Normalization of qPCR data is often done relative to the expression of reference genes, number of cells, weight of tissue (DNA/RNA spike) or total RNA concentration. Expression of mouse reference genes in temporal development of an embryo will not be constant and hence will likely introduce unwanted variability. As a consequence we chose to not to use endogenous reference genes in the present study, in line with recommendations from Sindelka et al. [42]. By not using endogenous reference genes emphasis was put on careful and highly standardized technical handling.

To ensure that measurements at low levels were well within the linear area of detection, all *Cq* values above 26 were coded as missing values. In addition, embryos with gene expression pattern radically different from the overall group mean were classified as outliers. The excluded outliers most likely represented deteriorating embryos that were about to die. Filtering criteria for missing values was set to 70%, which is the minimum percentage of existing values, and all the patterns with less than 70% existing values were removed. All estimates are mean values based on 10–15 single zygotes/embryos.

The information contained in the biological replicates was used to replace missing values in the remaining biological replicates when available. If all biological replicates gave missing data, they were all assigned the highest measured *Cq* of that particular gene +1. Since the highest measured *Cq* of a truly positive sample can be assumed to be the limit of detection (LOD) for that particular gene, assigning *Cq*(LOD) + 1 to the off-scale samples corresponds to a concentration that is half of the LOD. Relative expression (RQ) among the sample was calculated as [41]:

$$RQ = 2^{Cq_{min} - Cq}$$

Table 1
The genes included in the present analysis and their main function.

Symbol	Common name	Major function
Chek1[a]	Checkpoint homolog (*Schizosaccharomyces pombe*)	P53 signaling
Mlh3	mutL homolog 3	DNA repair
Xpa[b]	Xeroderma pigmentosum, complementation group A	DNA repair
Ercc2	Excision repair cross-complementing rodent repair deficiency 2	DNA repair
Zbed3[a]	Zinc finger, BED-type containing 3	DNA repair
Pms2[a]	Mismatch repair endonuclease (postmeiotic segregation increased 2	DNA repair
Prm2	Protamine 2	Sperm marker
Hdac1	Histone deacetylase 1	Epigenetic signaling
Ercc3	Excision repair cross-complementing rodent repair deficiency 3	DNA repair
Pcna[a]	Proliferating cell nuclear antigen	Embryo development
Spin1[b]	Spindlin 1	Embryo development
Pms1[a]	Postmeiotic segregation increased 1 (*Saccharomyces cerevisiae*)	DNA repair
Dnmt1[b]	DNA (cytosine-5-)-methyltransferase 1	P53 signaling
Polb[b]	Polymerase (DNA directed), beta	DNA repair
Kdm4a[b]	Lysine (K)-specific demethylase 4A	Epigenetic signaling
Xrcc1	X-ray repair cross complementing protein	DNA repair
Ercc1	Excision repair cross-complementing rodent repair deficiency 1	DNA repair
Tp53[a]	Tumor protein p53	P53 signaling
Nanog[a]	Nanog homeobox	Embryo development
Kdm4c[a]	Lysine (K)-specific demethylase 4C	Epigenetic signaling
Msh2	mutS homolog 2, colon cancer, nonpolyposis type 1 (*Escherichia coli*)	DNA repair
Ddb2	Damage-specific DNA binding protein 2	DNA repair
Mpg	N-methylpurine-DNA glycosylase	DNA repair
Pou5f1	POU class 5 homeobox 1	Embryo development
Cdh1	Adherin 1, type 1, E-cadherin (epithelial)	Embryo development
Ddb1	Damage-specific DNA binding protein 1	DNA repair
Fgfr2	Fibroblast growth factor receptor 2	Embryo development
Tdp1	Tyrosyl-DNA phosphodiesterase 1	DNA repair
Mre11a[a]	Meiotic recombination 11 homolog A (*S. cerevisiae*)	DNA repair
Tgfb1	Transforming growth factor, beta 1	Embryo development
Ogg1	8-Oxoguanine DNA glycosylase	DNA repair
Lig3	Ligase III, DNA, ATP-dependent	DNA repair
Dppa3	Developmental pluripotency associated 3	Embryo development
H2afz	H2A histone family, member Z	Embryo development
Mos	Oocyte maturation factor mos	Embryo development
Gtf2h3	General transcription factor IIH, polypeptide 3	DNA repair
Npm2	Nucleophosmin/nucleoplasmin 2	Sperm marker
Crabp1[b]	Cellular retinoic acid binding protein 1	Embryo development
Gapdh[b]	Glyceraldehyde-3-phosphate dehydrogenase	Metabolism/control gene

[a] Genes that responded to paternal GA exposure.
[b] Genes that were mainly off-scale (excluded from the analysis).

For each gene $\triangle Cq$ values were calculated by subtracting Cq from the lowest sample Cq, and then converting the difference to linear scale, as shown in the equation above. These relative quantities (RQ) indicate the level of expression, in each sample, of a particular gene relative to the sample in which the gene has the highest expression. Hence, the RQ of the sample with highest expression is set to one and all other samples for that gene have RQ < 1.

For further analysis with classic parametric tests, such as Student's t-test, linear regression and Anova, data must be normally distributed. Gene expression data are usually not normally distributed when expressed as relative quantities, but they will become normally distributed after logarithmic transformation to fold difference (FD). Traditionally log base 2 is used.

$$FD = \log_2(RQ)$$

Fold differences are equivalent to the corresponding delta Cq values for the purpose of statistical analysis. However fold differences provide useful control of the relative pivot point as defined by the RQ transformation step, the latter being preferred for visualization purposes.

2.8. Data scaling

Mean centering was performed to avoid highly expressed genes from being allowed to carry too much weight. The effect of the gene expression level was removed by subtracting the mean expression of every gene to the corresponding gene. For mean centered data a certain deviation from normal expression has the same weight independent of the expression level [41].

$$FD_{MC} = FD - \overline{FD}$$

Using the approach outlined above we compared differences in gene expression between developing mouse embryos originating from fertilization by spermatozoa from GA exposed fathers relative to control fathers right after fertilization, and at the 2-, 4- and 8-cell stages of embryo development.

2.9. Statistical analysis

Results are presented as the mean ± SEM. Student's t tests were used to test for differences between embryos of exposed fathers relative to control embryos. Values were considered to be significantly different when $p < 0.05$.

3. Results

Zygotes and embryos from four pre-implantation stages (1-, 2-, 4- and 8-cell) fertilized either with control sperm or with sperm from GA treated mice, were used to study the effect of paternal exposure to GA on the expression of a panel of 39 genes involved in DNA damage response and embryo development. These genes were selected based on prior knowledge and literature mining. The 39 genes are listed in Table 1.

3.1. Sample extraction and fertilization

A total of 105 zygotes/embryos (43 derived from GA exposed sperm and 62 from control sperm) at 4 different developmental stages were used for the final analysis. The number of embryos analyzed from each stage after pre-processing is shown in Table 2.

Known to vary between different strains of mice, the 1st (1–2) and 2nd (2–4) cell cycles of the embryo take between 16–20 h and 18–22 h, respectively [43,44]. In the current experiment we timed our sample extraction according to 24 h cycles; 2-, 4- and 8-cell time points equals 24, 48 and 72 h,

Table 2
The number of embryos in each treatment group after RT-qPCR pre-processing.

Stage	Group	Median	Range
1 cell	Control	17	11–17
	Exposed	14	7–14
2 cell	Control	13	9–13
	Exposed	12	9–12
4 cell	Control	16	9–16
	Exposed	11.5	5–13
8 cell	Control	11	8–11
	Exposed	10	6–10
Total	Control	56	46–57
	Exposed	46	36–49

respectively. Stages can be readily identified up to the 8-cell stage and careful visual inspection served as additional quality criteria upon sampling. Overall, cell cycle length is subject to considerable variation, not only among different embryos, but also among blastomeres within the same embryo [45–47].

For the 1-cell fertilized samples, both exposed and controls were collected immediately after the IVF procedure. Isolation of successful fertilizations was based on the appearance of polar bodies on the oocyte cell surface of healthy looking oocytes, when these were visible. Crude correlations based on RT-qPCR results indicate a clear difference between fertilized and unfertilized oocytes in terms of gene expression (supplementary figure 1), and this may tentatively be taken as a sign of successful fertilization. Samples from the 2-, 4- and 8-cell stage were exclusively chosen from the pool of structurally healthy looking embryos representative of the relevant stage.

3.2. Pre-processing and results from single cell RT-qPCR

Only those genes that passed the quality assurance criteria described in Section 2 were included in the down-stream analysis. From a total of 39 genes, a filtered list consisting of a matrix of dimension 30 genes × 43 GA exposed sperm derived embryos was used to search for transcripts showing fold differences (between embryos fertilized with sperm form GA-treated mice and those fertilized with control sperm from the same embryonal stage). The 43 GA exposed sperm derived embryos and the 62 control embryos were from four preimplantation stages (1-, 2-, 4- and 8-cell). Each stage consists of at least nine biological replicates. The fold difference in gene expression was derived by dividing each normalized relative quantity value by the average of stage-matched control values (RQ-value of treated embryo – RQ-value of average stage-matched control embryos). Gene expression levels are hence always relative to the control embryos at the same embryonal stage. The obtained gene expression ratios were used in the down-stream analysis.

An unsupervised hierarchical clustering analysis (average-linkage and Euclidean distance similarity measurement) was used to cluster genes and samples into groups based on their similarity, and the results were visualized in a dendrogram (Fig. 1). The dendrogram shows the similarity of the genes (rows), and similarity of the samples (columns) (Fig. 1). Although the row dendrogram and the column dendrogram are shown simultaneously on the heatmap, they are computed independent of each other. Fig. 1 shows the clustering analysis of gene expression alterations across embryo developmental stages. The clustering dendrogram of the 43 embryos from GA exposed sperm reveals two distinct branches: the first branch consists of embryos from 1- and 2-cell and some from the 4-cell stage, while all 8-cell stage and some 4-cell stage embryos were clustered close to each other in the second branch (Fig. 1). With a few exceptions, we observed an overall tendency towards reduced expression at the 1- and 2-cell stage together with an increased gene expression at the 8-cell stage in embryos of GA exposed fathers compared to controls (Fig. 1).

Genes with similar temporal expression patterns over a range of embryo development stages might share common regulators. Co-regulated and functionally related genes are probably going to be expressed (up and down) simultaneously, and will hence appear as clusters. One way to detect co-regulated genes is to investigate whether genes with similar expression patterns belong to the same cluster. By inspection of the clustering dendrogram, we observe the 30 genes studied cluster in three branches (Fig. 1). Group I consists of seven genes: Cdh1, Dppa3, H2afz, Kdm4c, Msh2, Pou5f1 and Xrcc1, Group II consists of 18 genes: Chek1, Ddb1, Ddb2, Ercc1, Ercc2, Ercc3, Hdac1, Lig3, Mlh3, Mos, Mpg, Mre11a, Pcna, Pms1, Pms2, Tdp1, Fgfr2 and Trp53, while Group III consists of five genes: Gtf2h3, Nanog, Npm2, Ogg1 and Zbed3. We then used the MouseNet server (http://www.functionalnet.org/mousenet/) to investigate the connectivity or co-occurrence network of these 30 genes. The interactions between these genes are constructed by Cytoscape® software [48], and shown in Fig. 2.

In Fig. 3 relative gene expression for the genes that were expressed significantly different between control embryos and embryos of GA exposed fathers is shown. A general trend of an initial down-regulation at 1- and 2-cell stage, followed by an up-regulation at the 8-cell stage seems to be present.

A separate single experiment was performed to study potential differences in development between control embryos and embryos of GA exposed fathers. The results suggest that there was a slightly lower formation of 2-celled embryos following fertilization with sperm from GA exposed fathers compared to controls (80% and 92%, respectively). Several of the embryos of GA exposed fathers also seemed to have slightly delayed development into the blastocyst stage.

4. Discussion

We have previously shown that DNA lesions induced in late spermatids or sperm persist in the sperm [15–17]. The consequences of paternally transferred DNA lesions are however still unclear. With respect to GA exposure heritable genomic changes and dominant lethal effects are most evident after exposure to these later stages of spermatogenesis and spermiogenesis [49,50], and it is likely that these effects are associated with persistent DNA- and protamine-adducts present in the fertilizing sperm. In this work we have investigated whether paternal exposure to GA at the spermatozoa stage has implications for gene expression during early embryo development. We have selected embryos with normal appearance at each stage in order to study responses different from those involved in preimplantation embryo lethality, which occurs at higher exposure doses. The most apparent effects observed in embryos of GA exposed sperm were that gene expression at the 1- and 2-cell embryonic stages were down regulated relative to controls followed by an up-regulation at the 8-cell stage.

The formation of GA-induced DNA and protamine adducts represents, the most likely mechanism by which parental GA-exposure impacts gene expression in the ensuing embryo, although other cellular mechanisms may also come into play. Protamine adducts are thought to introduce stresses in the sperm chromatin structure that are converted into DNA strand breaks upon fertilization. Such DNA lesions are repaired in the zygote, but may cause an increase in chromosomal aberrations if misrepaired as shown in zygotes from acrylamide exposed male mice [51]. In contrast, the major GA induced DNA adduct (N7-(2-carbamoyl-2-hydroxyethyl)guanine; N7-GA-Gua) has a slow repair kinetic [33,34] and is likely not fully removed before the first cleavage

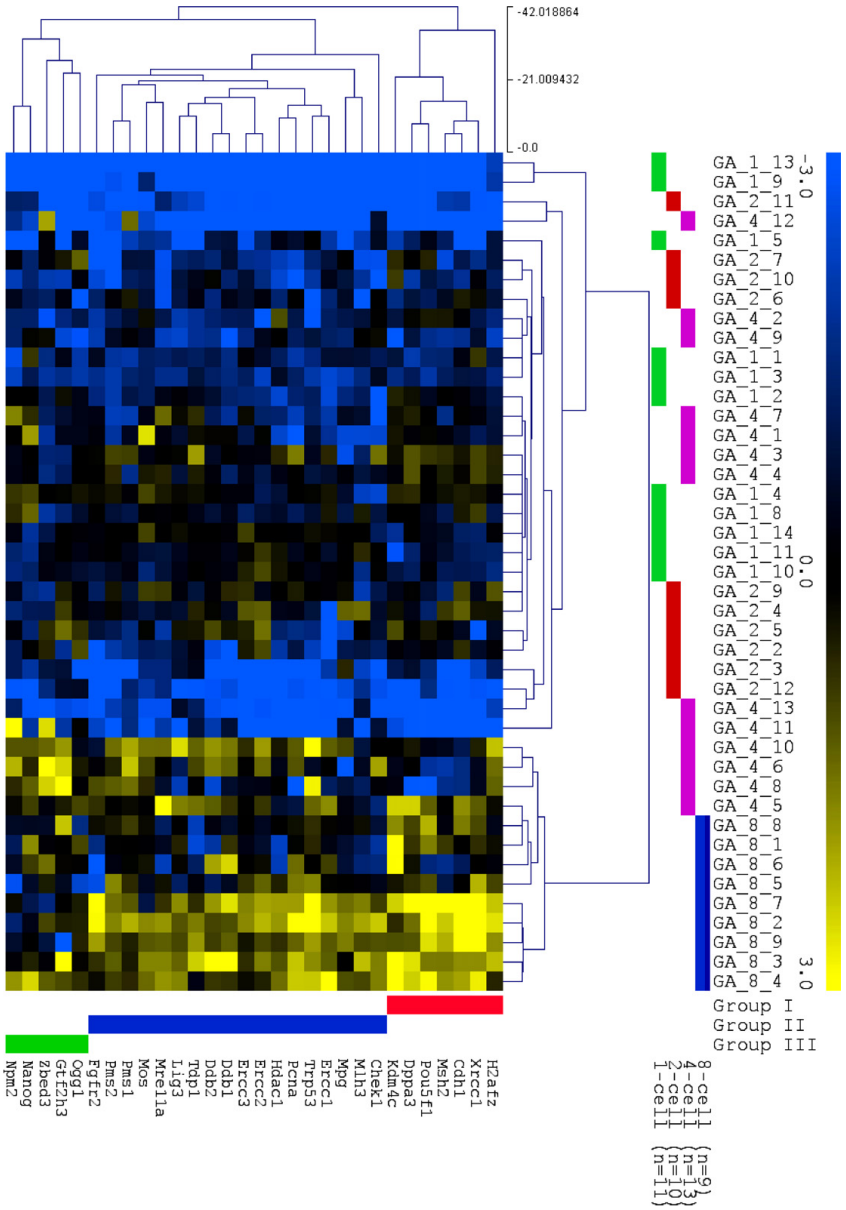


Fig. 1. The heatmap shows relative gene expression after normalization and mean-centering of the data. An unsupervised hierarchical clustering analysis is based on similarities in gene expression is included in the heatmap. Row and column dendrograms are computed independently of each other. Yellow samples represent up-regulated gene expression whereas bright blue samples represent down-regulated gene expression compared to average (black). Horizontal color bars indicate cell stage, vertical color bars indicate clustering after hierarchical clustering analysis. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

division. It is possible that other effects, beyond direct protamine and DNA adduct formation, could have contributed to the characteristic pattern of gene expression observed in the present study. The fact that the major reprogramming in the mouse embryo does not start until the 2-cell stage could indicate epigenetic effects as being responsible for the initial reduction in gene expression seen in embryos of GA exposed sperm compared to controls. There is convincing evidence that at least six components of the sperm

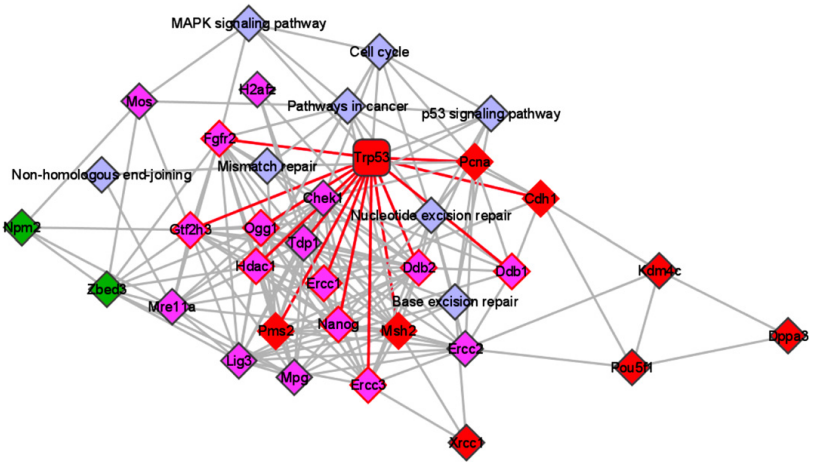


Fig. 2. A network analysis is shown indicating expression levels and connections at the 8 cell stage. Red symbols indicate genes that are up-regulated, whereas green symbols indicate genes that are down-regulated in embryos of exposed fathers relative to controls. Pathways are indicated by blue symbols. Red lines indicate connections to the p53 protein. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

nucleus are inherited from the paternal nucleus; DNA-methylation, sperm-specific histones, other chromatin-associated proteins (e.g. topoisomerase II), the perinuclear theca proteins, sperm-borne RNAs and the DNA-loop domain organization by the sperm nuclear matrix [52].

A separate IVF experiment was performed to test for potential differences in fertilization rates and developmental milestones among controls and embryos of GA exposed fathers. Based on these results, several embryos of GA exposed fathers seemed to have a delayed development into the blastocyst stage compared to controls. The suppressed transcription observed in GA exposed sperm derived embryos at the 1- and 2-cell stage could not simply be interpreted as an increased propensity of these cells to undergo apoptosis. It has been demonstrated by others that cell cycle checkpoints and apoptosis are largely absent in early stages of mouse embryo development [53–57].

The fact that paternal exposure can affect gene expression of the developing embryo justifies concern about potential human health effects manifested in the next generation. Trans-generational effects of DNA adducts have been reported previously. Smoking may serve as a classical example of xenobiotic exposure with potential transgenerational health effects. Sperm from smoking males typically exhibit higher levels of DNA fragmentation and base adduct formation [7]. Sperm from smoking males still retain their capacity to fertilize the oocyte, but it has been reported that the offspring of smoking males experience higher incidence of childhood cancer [8,58]. These studies exemplify the concept of paternal germ cell DNA damage creating undesired effects in the offspring.

Effects of paternal exposure manifested at the 2-cell stage of embryo development have been documented by other investigators. Marchetti et al. reported increased levels of chromosomal mosaicism in 2-cell mouse embryos after paternal

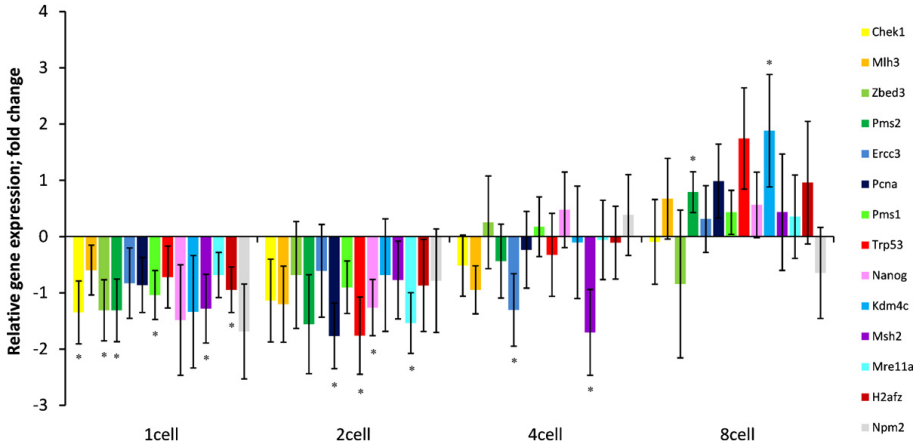


Fig. 3. Relative gene expression is shown for the genes that at some point during the 72 h study period responded significantly to paternal preconceptional exposure to GA. Results are mean values (±standard error).

exposure to acrylamide [59]. In another report cyclophosphamide, a widely used anti-cancer agent has been found to alter epigenetic histone H4 acetylation and methylation in preimplantation rat embryos at the 2-cell stage [60]. Paternal cyclophosphamide exposure has also been shown to induce DNA damage in the rat zygote [61] and to induce an altered expression of several DNA repair genes in cleavage stage embryos [39].

With respect to technical details in this experiment, the exposure level of GA is much higher than would be expected in a natural environment. Sperm protamine and DNA adducts are not repaired, and repeated acrylamide exposure leads to an accumulation of acrylamide induced chromatin adduction [32]. Using a single exposure enabled us to study the consequences of a defined exposure of the spermatozoa in a maturation stage strongly associated with induction of dominant lethality and heritable translocations. The dose chosen was lower than those used for studying dominant lethality [21,62,63], but slight to moderate increases in heritable translocations and chromosomal aberrations is reported at comparable levels in studies using acrylamide [51,64]. New studies are needed to examine potential gene expression perturbations in the developing embryo after chronic low-dose paternal exposure to relevant environmental exposures.

In contrast to the general down-regulation at the 1- and 2-cell stage, the transcriptional up-regulation observed in GA exposed sperm derived embryos at the 8-cell stage does not apply to all genes. The slightly more differentiated pattern at the latter stage could perhaps indicate that a more specific cellular response has been initiated in an attempt to deal with molecular damage inherited from GA exposed male germ-cells.

Among the DNA repair genes investigated Pms2 was significantly up-regulated at the 8-cell stage in the GA exposed sperm derived embryos compared to the controls. This gene, located on chromosome 7, codes for a DNA mismatch repair protein. It forms a heterodimer with MLH1 and this complex interacts with other complexes bound to mismatched bases. Mutations in these genes are associated with hereditary non-polyposis colorectal cancer (Lynch syndrome) [65]. Moreover, Pms2 is implicated in DNA damage signaling, a process which induces cell cycle arrest and can lead to apoptosis in case of major DNA damages. The up-regulation of this gene is consistent with the idea that the cell cycle may be arrested and with apoptosis being induced.

Embryos of GA exposed fathers were also found to have significantly increased expression of the histone demethylase Jmjd2C/Kdm4C at the 8-cell stage. In one study Jmjd2C/Kdm4C knockdown caused decreased proliferation of tumor cells, suggesting a contribution of Jmjd2C/Kdm4C to tumor growth [66]. Over-expression of Jmjd2C/Kdm4C has been found to enhance the expression of the Mdm2 oncogene [67].

We were not able to pinpoint clear-cut mechanistic reactions based on the horizontal branching of the dendrogram (Fig. 1). However, in a separate network analysis many of these genes were found to be p53 target genes (Fig. 2), and several of them even have p53 binding sites. The p53 association is interesting even though p53 signaling is not very well characterized at these early stages of development. In studies of preimplantation mouse embryos fertilized with sperm exposed to gamma irradiated paternal DNA damage did not severely reduce fertilization rates and implantation, but radiation reduced the number of fetuses [68–70]. Sperm irradiation resulted in a suppression of DNA synthesis in the zygote. The S-phase delay in the zygotes was p53-dependent and seemed to be dependent on p53 DNA-binding, but not on p53 transcriptional activity [68–70] Although a more differentiated response of embryos of GA exposed sperm is witnessed at the 8-cell stage compared to the 1- and 2-cell stage, we believe that the increased expression in some of the genes at this stage is best interpreted as being caused by an overall effect on the embryo rather than being clearly confined to particular pathways. A conservative evaluation of the genes that responded significantly is advisable. The overall tendency towards reduced gene expression at the 1-, and 2-cell stage followed by the increased expression at the 8-cell stage constitutes our most important observation. New dose-response studies should look at the consistency of the characteristic pattern of 1- and 2-cell down-regulation and 8-cell up-regulation across different doses. There is a need for more information on low-dose periconceptional toxicity, and gene expression analyses such as these may provide one tool for a much sought after quantitative evaluation of subtle gene or gene network perturbations that should be acknowledged in risk evaluations.

In summary early embryonic gene expression is affected by preconceptional paternal exposure to glycidamide. Persisting paternal GA DNA and protamine adducts or epigenetic changes are probable mechanisms to cause altered expression of genes related to embryonic stem cell pluripotency, DNA repair, and methylation dynamics in early mouse embryos. Single embryo gene expression represents a new technique with potential to advance our understanding of both reproductive toxicity and the extremely complex process of embryo development.

Conflict of interest

Stock Ownership: M. Kubista, TATAA Biocenter.

Acknowledgements

This work was supported by The Research Council of Norway, grant # 182048/V40, and The Grant Agency of Academy of Science Czech Republic (IAA500520809) and research goal AV0Z50520701 granted by Ministry of Youth, Education and Sports of the Czech republic.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.reprotox.2011.09.005.

References

[1] Hotaling JM, Walsh TJ. Male infertility: a risk factor for testicular cancer. Nat Rev Urol 2009;6(October (10)):550–6.
[2] Auger J, Kunstmann JM, Czyglik F, Jouannet P. Decline in semen quality among fertile men in Paris during the past 20 years. N Engl J Med 1995;332(February (5)):281–5.
[3] Carlsen E, Giwercman A, Keiding N, Skakkebaek NE. Evidence for decreasing quality of semen during past 50 years. BMJ 1992;305(September (6854)):609–13.
[4] Jacobsen R, Moller H, Thoresen SO, Pukkala E, Kjaer SK, Johansen C. Trends in testicular cancer incidence in the Nordic countries, focusing on the recent decrease in Denmark. Int J Androl 2006;29(February (1)):199–204.
[5] Richiardi L, Bellocco R, Adami HO, Torrang A, Barlow L, Hakulinen T, et al. Testicular cancer incidence in eight northern European countries: secular and recent trends. Cancer Epidemiol Biomarkers Prev 2004;13(December (12)):2157–66.
[6] Walsh TJ, Croughan MS, Schembri M, Chan JM, Turek PJ. Increased risk of testicular germ cell cancer among infertile men. Arch Intern Med 2009;169(February (4)):351–6.
[7] Fraga CG, Motchnik PA, Wyrobek AJ, Rempel DM, Ames BN. Smoking and low antioxidant levels increase oxidative damage to sperm DNA. Mutat Res 1996;351(April (2)):199–203.
[8] Ji BT, Shu XO, Linet MS, Zheng W, Wacholder S, Gao YT, et al. Paternal cigarette smoking and the risk of childhood cancer among offspring of nonsmoking mothers. J Natl Cancer Inst 1997;89(February (3)):238–44.
[9] Hamer G, Roepers-Gajadien HL, van Duyn-Goedhart A, Gademan IS, Kal HB, van Buul PP, et al. DNA double-strand breaks and gamma-H2AX signaling in the testis. Biol Reprod 2003;68(February (2)):628–34.
[10] Olsen AK, Lindeman B, Wiger R, Duale N, Brunborg G. How do male germ cells handle DNA damage? Toxicol Appl Pharmacol 2005;207(September (2 Suppl.)):521–31.
[11] van Loon AA, Sonneveld E, Hoogerbrugge J, van der Schans GP, Grootegoed JA, Lohman PH, et al. Induction and repair of DNA single-strand breaks and

DNA base damage at different cellular stages of spermatogenesis of the hamster upon in vitro exposure to ionizing radiation. Mutat Res 1993;294(August (2)):139–48.
[12] Xu G, Spivak G, Mitchell DL, Mori T, McCarrey JR, McMahan CA, et al. Nucleotide excision repair activity varies among murine spermatogenic cell types. Biol Reprod 2005;73(July (1)):123–30.
[13] Jansen J, Olsen AK, Wiger R, Naegeli H, de BP, van Der HF, et al. Nucleotide excision repair in rat male germ cells: low level of repair in intact cells contrasts with high dual incision activity in vitro. Nucleic Acids Res 2001;29(April (8)):1791–800.
[14] Olsen AK, Bjortuft H, Wiger R, Holme J, Seeberg E, Bjoras M, et al. Highly efficient base excision repair (BER) in human and rat male germ cells. Nucleic Acids Res 2001;29(April (8)):1781–90.
[15] Olsen AK, Andreassen A, Singh R, Wiger R, Duale N, Farmer PB, et al. Environmental exposure of the mouse germ line: DNA adducts in spermatozoa and formation of de novo mutations during spermatogenesis. PLoS One 2010;5(June (6)):e11349.
[16] Sipinen V, Laubenthal J, Baumgartner A, Cemeli E, Linschooten JO, Godschalk RW, et al. In vitro evaluation of baseline and induced DNA damage in human sperm exposed to benzo[a]pyrene or its metabolite benzo[a]pyrene-7,8-diol-9,10-epoxide, using the comet assay. Mutagenesis 2010;25(July (4)):417–25.
[17] Verhofstad N, van Oostrom CT, van BJ, Van Schooten FJ, van SH, Godschalk RW. DNA adduct kinetics in reproductive tissues of DNA repair proficient and deficient male mice after oral exposure to benzo(a)pyrene. Environ Mol Mutagen 2010;51(March (2)):123–9.
[18] Bjorge C, Brunborg G, Wiger R, Holme JA, Scholz T, Dybing E, et al. A comparative study of chemically induced DNA damage in isolated human and rat testicular cells. Reprod Toxicol 1996;10(November (6)):509–19.
[19] Sega GA, Generoso EE. Measurement of DNA breakage in specific germ-cell stages of male mice exposed to acrylamide, using an alkaline-elution procedure. Mutat Res 1990;242(September (1)):79–87.
[20] Shelby MD, Cain KT, Hughes LA, Braden PW, Generoso WM. Dominant lethal effects of acrylamide in male mice. Mutat Res 1986;173(January (1)):35–40.
[21] Marchetti F, Lowe X, Bishop J, Wyrobek AJ. Induction of chromosomal aberrations in mouse zygotes by acrylamide treatment of male germ cells and their correlation with dominant lethality and heritable translocations. Environ Mol Mutagen 1997;30(4):410–7.
[22] Mottram DS, Wedzicha BL, Dodson AT. Acrylamide is formed in the Maillard reaction. Nature 2002;419(October (6906)):448–9.
[23] Rosen J, Hellenas KE. Analysis of acrylamide in cooked foods by liquid chromatography tandem mass spectrometry. Analyst 2002;127(July (7)):880–2.
[24] Stadler RH, Scholz G. Acrylamide: an update on current knowledge in analysis, levels in food, mechanisms of formation, and potential strategies of control. Nutr Rev 2004;62(December (12)):449–67.
[25] Calleman CJ, Bergmark E, Costa LG. Acrylamide is metabolized to glycidamide in the rat: evidence from hemoglobin adduct formation. Chem Res Toxicol 1990;3(September (5)):406–12.
[26] Ghanayem BI, McDaniel LP, Churchwell MI, Twaddle NC, Snyder R, Fennell TR, et al. Role of CYP2E1 in the epoxidation of acrylamide to glycidamide and formation of DNA and hemoglobin adducts. Toxicol Sci 2005;88(December (2)):311–8.
[27] Settels E, Bernauer U, Palavinskas R, Klaffke HS, Gundert-Remy U, Appel KE. Human CYP2E1 mediates the formation of glycidamide from acrylamide. Arch Toxicol 2008;82(October (10)):717–27.
[28] Sumner SC, Fennell TR, Moore TA, Chanas B, Gonzalez F, Ghanayem BI. Role of cytochrome P450 2E1 in the metabolism of acrylamide and acrylonitrile in mice. Chem Res Toxicol 1999;12(November (11)):1110–6.
[29] Koyama N, Yasui M, Oda Y, Suzuki S, Satoh T, Suzuki T, et al. Genotoxicity of acrylamide in vitro: acrylamide is not metabolically activated in standard in vitro systems. Environ Mol Mutagen 2010;(March).
[30] Rice JM. The carcinogenicity of acrylamide. Mutat Res 2005;580(February (1–2)):3–20.
[31] Sega GA. Unscheduled DNA synthesis (DNA repair) in the germ cells of male mice—its role in the study of mammalian mutagenesis. Genetics 1979;92(May (1 Pt 1 Suppl.)):s49–58.
[32] Sega GA, Alcota RP, Tancongco CP, Brimer PA. Acrylamide binding to the DNA and protamine of spermiogenic stages in the mouse and its relationship to genetic damage. Mutat Res 1989;216(August (4)):221–30.
[33] Gamboa da CG, Churchwell MI, Hamilton LP, Von Tungeln LS, Beland FA, Marques MM, et al. DNA adduct formation from acrylamide via conversion to glycidamide in adult and neonatal mice. Chem Res Toxicol 2003;16(October (10)):1328–37.
[34] Maniere I, Godard T, Doerge DR, Churchwell MI, Guffroy M, Laurentie M, et al. DNA damage and DNA adduct formation in rat tissues following oral administration of acrylamide. Mutat Res 2005;580(February (1–2)):119–29.
[35] Segerback D, Calleman CJ, Schroeder JL, Costa LG, Faustman EM. Formation of N-7-(2-carbamoyl-2-hydroxyethyl)guanine in DNA of the mouse and the rat following intraperitoneal administration of [14C]acrylamide. Carcinogenesis 1995;16(May (5)):1161–5.
[36] Rodriguez-Zas SL, Schellander K, Lewin HA. Biological interpretations of transcriptomic profiles in mammalian oocytes and embryos. Reproduction 2008;135(February (2)):129–39.
[37] Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development 2009;136(September (18)):3033–42.
[38] Stanton JA, Macgregor AB, Green DP. Gene expression in the mouse preimplantation embryo. Reproduction 2003;125(April (4)):457–68.
[39] Harrouk W, Codrington A, Vinson R, Robaire B, Hales BF. Paternal exposure to cyclophosphamide induces DNA damage and alters the expression of DNA repair genes in the rat preimplantation embryo. Mutat Res 2000;461(November (3)):229–41.
[40] Harrouk W, Robaire B, Hales BF. Paternal exposure to cyclophosphamide alters cell-cell contacts and activation of embryonic transcription in the preimplantation rat embryo. Biol Reprod 2000;63(July (1)):74–81.
[41] Bergkvist A, Rusnakova V, Sindelka R, Garda JM, Sjogreen B, Lindh D, et al. Gene expression profiling—clusters of possibilities. Methods 2010;50(April (4)):323–35.
[42] Sindelka R, Ferjentsik Z, Jonak J. Developmental expression profiles of *Xenopus laevis* reference genes. Dev Dyn 2006;235(March (3)):754–8.
[43] Bowman P, McLaren A. Cleavage rate of mouse embryos in vivo and in vitro. J Embryol Exp Morphol 1970;24(August (1)):203–7.
[44] Harlow GM, Quinn P. Development of preimplantation mouse embryos in vivo and in vitro. Aust J Biol Sci 1982;35(2):187–93.
[45] Bischoff M, Parfitt DE, Zernicka-Goetz M. Formation of the embryonic–abembryonic axis of the mouse blastocyst: relationships between orientation of early cleavage divisions and pattern of symmetric/asymmetric divisions. Development 2008;135(March (5)):953–62.
[46] Fujimori T, Kurotaki Y, Komatsu K, Nabeshima Y. Morphological organization of the mouse preimplantation embryo. Reprod Sci 2009;16(February (2)):171–7.
[47] Kurotaki Y, Hatta K, Nakao K, Nabeshima Y, Fujimori T. Blastocyst axis is specified independently of early cell lineage but aligns with the ZP shape. Science 2007;316(May (5825)):719–23.
[48] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(November (11)):2498–504.
[49] Marchetti F, Wyrobek AJ. Mechanisms and consequences of paternally-transmitted chromosomal abnormalities. Birth Defects Res C: Embryo Today 2005;75(June (2)):112–29.
[50] Shelby MD. Selecting chemicals and assays for assessing mammalian germ cell mutagenicity. Mutat Res 1996;352(June (1–2)):159–67.
[51] Pacchierotti F, Tiveron C, D'Archivio M, Bassani B, Cordelli E, Leter G, et al. Acrylamide-induced chromosomal damage in male mouse germ cells detected by cytogenetic analysis of one-cell zygotes. Mutat Res 1994;309(September (2)):273–84.
[52] Yamauchi Y, Shaman JA, Ward WS. Non-genetic contributions of the sperm nucleus to embryonic development. Asian J Androl 2010;(October).
[53] Grinfeld S, Jacquet P. An unusual radiation-induced G2 arrest in the zygote of the BALB/c mouse strain. Int J Radiat Biol Relat Stud Phys Chem Med 1987;51(February (2)):353–63.
[54] Jurisicova A, Latham KE, Casper RF, Casper RF, Varmuza SL. Expression and regulation of genes associated with cell death during murine preimplantation embryo development. Mol Reprod Dev 1998;51(November (3)):243–53.
[55] Schmidt-Kastner PK, Jardine K, Cormier M, McBurney MW. Absence of p53-dependent cell cycle regulation in pluripotent mouse cell lines. Oncogene 1998;16(June (23)):3003–11.
[56] Mukherjee AB. Cell cycle analysis and X-chromosome inactivation in the developing mouse. Proc Natl Acad Sci USA 1976;73(May (5)):1608–11.
[57] Kamjoo M, Brison DR, Kimber SJ. Apoptosis in the preimplantation mouse embryo: effect of strain difference and in vitro culture. Mol Reprod Dev 2002;61(January (1)):67–77.
[58] Meseguer M, Martinez-Conejero JA, O'Connor JE, Pellicer A, Remohi J, Garrido N. The significance of sperm DNA oxidation in embryo development and reproductive outcome in an oocyte donation program: a new model to study a male infertility prognostic factor. Fertil Steril 2008;89(May (5)):1191–9.
[59] Marchetti F, Bishop J, Lowe X, Wyrobek AJ. Chromosomal mosaicism in mouse two-cell embryos after paternal exposure to acrylamide. Toxicol Sci 2009;107(January (1)):194–205.
[60] Barton TS, Robaire B, Hales BF. Epigenetic programming in the preimplantation rat embryo is disrupted by chronic paternal cyclophosphamide exposure. Proc Natl Acad Sci USA 2005;102(May (22)):7865–70.
[61] Barton TS, Robaire B, Hales BF. DNA damage recognition in the rat zygote following chronic paternal cyclophosphamide exposure. Toxicol Sci 2007;100(December (2)):495–503.
[62] Adler ID, Baumgartner A, Gonda H, Friedman MA, Skerhut M. 1-Aminobenzotriazole inhibits acrylamide-induced dominant lethal effects in spermatids of male mice. Mutagenesis 2000;15(March (2)):133–6.
[63] Generoso WM, Sega GA, Lockhart AM, Hughes LA, Cain KT, Cacheiro NL, et al. Dominant lethal mutations, heritable translocations, and unscheduled DNA synthesis induced in male mouse germ cells by glycidamide, a metabolite of acrylamide. Mutat Res 1996;371(December (3–4)):175–83.
[64] Adler ID, Reitmeir P, Schmoller R, Schriever-Schwemmer G. Dose response for heritable translocations induced by acrylamide in spermatids of mice. Mutat Res 1994;309(September (2)):285–91.
[65] de la Chapelle A. Genetic predisposition to colorectal cancer. Nat Rev Cancer 2004;4(October (10)):769–80.
[66] Cloos PA, Christensen J, Agger K, Maiolica A, Rappsilber J, Antal T, et al. The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. Nature 2006;442(July (7100)):307–11.

[67] Ishimura A, Terashima M, Kimura H, Akagi K, Suzuki Y, Sugano S, et al. Jmjd2c histone demethylase enhances the expression of Mdm2 oncogene. Biochem Biophys Res Commun 2009;389(November (2)):366–71.
[68] Shimura T, Toyoshima M, Taga M, Shiraishi K, Uematsu N, Inoue M, et al. The novel surveillance mechanism of the Trp53-dependent s-phase checkpoint ensures chromosome damage repair and preimplantation-stage development of mouse embryos fertilized with x-irradiated sperm. Radiat Res 2002;158(December (6)):735–42.
[69] Shimura T, Inoue M, Taga M, Shiraishi K, Uematsu N, Takei N, et al. p53-dependent S-phase damage checkpoint and pronuclear cross talk in mouse zygotes with X-irradiated sperm. Mol Cell Biol 2002;22(April (7)):2220–8.
[70] Toyoshima M, Shimura T, Adiga SK, Taga M, Shiraishi K, Inoue M, et al. Transcription-independent suppression of DNA synthesis by p53 in sperm-irradiated mouse zygotes. Oncogene 2005;24(May (20)):3229–35.

TOXICOLOGICAL SCIENCES **129(1)**, 157–165 (2012)
doi:10.1093/toxsci/kfs187
Advance Access publication May 28, 2012

Paternal Benzo[a]pyrene Exposure Affects Gene Expression in the Early Developing Mouse Embryo

Asgeir Brevik,* Birgitte Lindeman,* Vendula Rusnakova,† Ann-Karin Olsen,* Gunnar Brunborg,* and Nur Duale*,[1]

*Norwegian Institute of Public Health, Division of Environmental Medicine, Department of Chemicals and Radiation, PO Box 4404, Nydalen, N-0403 Oslo, Norway; and †Department of Gene Expression, Institute of Biotechnology AS CR, v.v.i., Videnska 1083, Prague 4, 142 02, Czech Republic

To whom correspondence should be addressed. E-mail: nur.duale@fhi.no.

Received March 14, 2012; accepted May 15, 2012

The health of the offspring depends on the genetic constitution of the parental germ cells. The paternal genome appears to be important; e.g., *de novo* mutations in some genes seem to arise mostly from the father, whereas epigenetic modifications of DNA and histones are frequent in the paternal gonads. Environmental contaminants which may affect the integrity of the germ cells comprise the polycyclic aromatic hydrocarbon, benzo[a]pyrene (B[a]P). B[a]P has received much attention due to its ubiquitous distribution, its carcinogenic and mutagenic potential, and also effects on reproduction. We conducted an *in vitro* fertilization (IVF) experiment using sperm cells from B[a]P-exposed male mice to study effects of paternal B[a]P exposure on early gene expression in the developing mouse embryo. Male mice were exposed to a single acute dose of B[a]P (150 mg/kg, ip) 4 days prior to isolation of cauda sperm, followed by IVF of oocytes from unexposed superovulated mice. Gene expression in fertilized zygotes/embryos was determined using reverse transcription-qPCR at the 1-, 2-, 4-, 8-, and blastocyst cell stages of embryo development. We found that paternal B[a]P exposure altered the expression of numerous genes in the developing embryo especially at the blastocyst stage. Some genes were also affected at earlier developmental stages. Embryonic gene expression studies seem useful to identify perturbations of signaling pathways resulting from exposure to contaminants, and can be used to address mechanisms of paternal effects on embryo development.

Key Words: **gene expression; benzo[a]pyrene; mouse embryo.**

Reduced male fertility and sperm quality are reported for many developed economies, suggesting that male reproductive health is challenged by environmental factors associated with a modern lifestyle (Auger *et al.*, 1995; Carlsen *et al.*, 1992). There are indications that the genetic constitution of the paternal genome may be challenged by environmental agents (Olsen *et al.*, 2010), potentially affecting the offspring (Marchetti and Wyrobek, 2005). Elevated exposure to such agents including polycyclic aromatic hydrocarbons (PAHs) has been reported in animal studies to affect reproduction and also the development of the fetus (Edwards *et al.*, 2006; Somers *et al.*, 2002). It is believed that transmittable effects may arise from accumulated DNA damage in mouse germ cells, which—beyond a critical stage of spermatogenesis—have a limited capacity for DNA repair (Marchetti and Wyrobek, 2008; Olsen *et al.*, 2005, 2010). There are concerns that undesired environmental exposure may lead to transgenerational effects in humans. Transmittable genetic damage in humans has been studied in populations exposed to radiation (Dubrova *et al.*, 1996), and there are reports indicating that children of smoking fathers have a higher risk of contracting childhood cancer (Ji *et al.*, 1997).

Benzo[a]pyrene (B[a]P) is a PAH, which has received much attention due to its carcinogenic and reprotoxic potential and its ubiquitous distribution. B[a]P was determined in 1933 as a component of coal tar and is recognized as the first compound being responsible for occupation-associated cancers, in the form of the sooty warts (cancers of the scrotum) suffered by chimney sweeps in 18th century England (Phillips, 1983). B[a]P is found in coal tar, automobile exhaust fumes (especially from diesel engines), and all smoke resulting from the combustion of organic material (including cigarette smoke). The dominant route of exposure in humans is through ingestion of contaminated food such as charbroiled meat (Hattemer-Frey and Travis, 1991). Cooked meat products have been shown to contain up to 4 ng/g of B[a]P (Kazerouni *et al.*, 2001), fried chicken contains up to 5.5 ng/g (Lee and Shim, 2007), and up to 62.6 ng/g have been measured in overcooked charcoal barbecued beef (Aygün and Kabadayi, 2005).

Phase I CYP450 enzymes convert B[a]P into reactive metabolites (Pelkonen and Nebert, 1982), e.g., benzo[a]-pyrene-7,8-diol-9,10-epoxide (BPDE), which bind covalently to DNA (Jeffrey *et al.*, 1977). B[a]P can also be metabolized via the aldo-keto reductase pathway, leading to increased levels of oxidative stress and formation of catechol DNA adducts (Penning, 2004). The metabolites react with a broad range of cellular molecules, and it follows that exposure to B[a]P may affect a number of cellular processes including gene transcription.

© The Author 2012. Published by Oxford University Press on behalf of the Society of Toxicology. For permissions, please email: journals.permissions@oup.com
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Our previous findings indicate that, although the sperm chromatin is highly compacted, B[a]P-derived DNA lesions may form in sperm after *in vivo* exposure of mice (Olsen *et al.*, 2010) and *in vitro* exposure of human sperm (Sipinen *et al.*, 2010). B[a]P exposure leads to the formation of DNA adducts in testis and sperm of exposed male mice (Verhofstad *et al.*, 2010) and humans (Zenzes *et al.*, 1999a). In a study of surplus human embryos from *in vitro* fertilization (IVF) treatment, the amount of B[a]P-derived DNA adducts in the embryo seemed to be more dependent on paternal than maternal smoking (Zenzes *et al.*, 1999b). This study also suggested that paternally transmitted BPDE-DNA adducts can persist past the first embryonic cleavage divisions. Furthermore, exposure of male mice to high doses of B[a]P in the week before fertilization has been shown to induce dominant lethal mutations (Generoso *et al.*, 1982; Shukla and Taneja, 2001).

Based on the ability of B[a]P to form DNA and protein adducts, and the reported multigenerational impact of chronic paternal B[a]P exposure (Mohamed *et al.*, 2010) on sperm production and sperm parameters, we hypothesized that a single acute paternal preconceptional exposure to B[a]P could affect early embryonic transcription and the activation of the embryonic genome. It is important if such exposure can affect the fetus even when administered to the epididymal (cauda) sperm. Following paternal acute exposure of male mice to B[a]P 4 days prior to fertilization, we studied the effect of B[a]P-exposed sperm cells on the expression of key DNA damage response genes and genes important to embryo development, in early embryonic stages. We demonstrate that early embryonic transcription of multiple genes is affected by paternal preconceptional exposure to B[a]P.

MATERIALS AND METHODS

Exposure of male mice. Male mice (B6D2F1 from Charles River Laboratories, 8–12 weeks of age) received one ip injection of B[a]P (150 mg/kg body weight) dissolved in corn oil, or corn oil only for controls, 4 days prior to isolation of their cauda sperm used for IVF treatment. Timing of the exposure to B[a]P was based on pilot studies and knowledge about the most susceptible stage of spermatogenesis with respect to dominant lethal mutations (Generoso, 1986). At the day of the IVF experiment males were killed by cervical dislocation. Cauda were surgically removed and transferred to an Eppendorf tube containing M2 medium (500 µl, Sigma). Using microscissors, a few incisions were made in the cauda and the sperm was allowed to disperse for 10 min into ~250 µl HTF medium (EmbryoMax, Millipore) under liquid paraffin (MediCult), before transfer of sperm to the IVF dishes. Experiments comprise sperm from 14 males (7 exposed and 7 controls) and oocytes from 84 females.

Superovulation and IVF. Female mice (B6D2F1 from Charles River Laboratories, 4–6 weeks of age) were injected ip with pregnant mare serum hormone gonadotropin (Folligon; Intervet) (5 IU) 3 days prior to the IVF procedure, to induce superovulation. Two days later (i.e., the day before IVF), animals received an additional ip injection of human chorionic gonadotropin (Ovitrelle; Serono) (5 IU). Mice were killed by cervical dislocation and oviducts were collected in M2 medium (Sigma). Egg clutches (10–20 oocytes) embedded in cumulus cells were extracted from each oviduct. Oocytes were transferred to IVF dishes (35-mm petri dish, Corning) and incubated in a droplet of HTF medium containing sperm under liquid paraffin for 4.5 h (37°C). Oocytes from one side of the animal were combined with sperm from B[a]P-exposed animals, and oocytes from the other side where combined with sperm from control animals. Hence, oocytes from all animals were present in both the control group and the exposed group. After 4.5 h the fertilized oocytes (zygotes) were washed 5 times in potassium simplex optimization medium (KSOM) (EmbryoMax, Millipore) before they were transferred to a drop of KSOM (200 µl) under liquid paraffin, in a petri dish (35 mm). Samples from the 1-cell stage were collected immediately after fertilization. Isolation of successful fertilizations was based on the appearance of polar bodies on the oocyte cell surface of healthy looking oocytes, when these were visible, and based on experience from previous experiments in which the majority of the oocytes were shown to be fertilized. The remaining zygotes were allowed to grow to be harvested at the 2-, 4-, 8-, and blastocyst cell stages. Upon harvest, zygotes/embryos were collected separately in micro tubes filled with 5 µl lysis medium (CelluLyser, TATAA Biocenter) and then frozen at –70°C.

Gene expression analysis—reverse transcription-qPCR. The reverse transcription (RT)-qPCR analysis of cell lysates was performed as previously described (Bergkvist *et al.*, 2010). In brief, cDNA synthesis was performed using a transcriptor first strand cDNA synthesis kit according to the manufacturer's recommendation (Roche, Cat. no. 04896866001). Hence, no RNA extraction was conducted. Samples were incubated in a thermal cycler unit (Eppendorf Mastercycler) according to the following protocol: 10 min at 25°C, 30 min at 50°C, 5 min at 85°C, and then held at 4°C. After RT reaction, samples were stored at –70°C until use. A preamplification step of the cDNA was included: preamplification PCR was run in 50 µl reactions containing 5 µl of cDNA, 5 µl of a mixture of all forward and reverse primers (250nM each), 25 µl of iQ Supermix (Bio-Rad), and water. PCR primers were designed using the online Universal Probe Library System from Roche, and all primers showed PCR efficiency higher than 90% (Supplementary table 1). A CFX 96 cycler (Bio-Rad) was used for the preamplification with the cycling conditions: polymerase activation at 95°C for 3 min, followed by 18 cycles (each of 95°C 15 s, 57°C 4 min, and 72°C 20 s). The product of the preamplification reaction was diluted 10 times and stored at –20°C. Then, a high throughput qPCR was performed as previously described in detail (Brevik *et al.*, 2011) using the BioMark qPCR platform (Fluidigm). The robustness of the preamplification was validated by comparing qPCR expression levels of *Chek1* (highly expressed) and *Mlh3* (typically low expression) with and without preamplification. The relative expression of the two genes was similar when analyzed with and without preamplification (Brevik *et al.*, 2011).

Data preprocessing. Prior to normalization, the raw data (*Cq* values) generated from qPCR experiments were preprocessed to ensure that measurements at low levels were well within the linear area of detection; all *Cq* values above 26 were coded as missing values. In addition, embryos with gene expression pattern radically different from the overall group mean were classified as outliers. The excluded outliers were considered to represent deteriorating embryos that were about to die. Filtering criteria for missing values was set to 70%, which is the minimum percentage of existing values, and all the patterns with less than 70% existing values were removed. All estimates are mean values based on 5–33 single zygotes/embryos. Missing values were replaced by using the information contained in the biological replicates when available. Because the highest measured *Cq* of a truly positive sample can be assumed to reflect the limit of detection (LOD) for that particular gene, assigning *Cq*(LOD) + 1 to the off-scale samples corresponds to a concentration that is half of the LOD. Following the preprocessing, we ended up with 51 genes out of 78 genes and 184 samples (embryos) passing the quality assurance criteria described above.

Identification of stably expressed normalizing genes. Following the raw data preprocessing steps, we chose an optimal normalization strategy. The gene expression stability of the 51 genes that passed our filtering criteria was analyzed using the geNorm algorithm (Vandesompele *et al.*, 2002), in order to identify the most stably expressed and putative normalizer genes. Raw *Cq* values were transformed to linear scale (relative quantities [RQs]), as shown in the equation below, before analysis.

$$\mathrm{RQ} = 2\left(Cq_{(\min)} - Cq_{(\text{sample})}\right),$$

TABLE 1
Differentially Expressed Genes (Based on Twofold Cutoff Value) Between Control Embryos and Embryos of Exposed Fathers at Various Stages of Embryo Development

Genes	1 cell	2 cell	4 cell	8 cell	Blastocyst
Adam2	−0.5	0.9	0.7	0.7	***−2.6***
Bcl2l2	−0.5	0.3	0.3	**1.0**	***−1.8***
Crem	−0.9	0.6	0.1	**1.0**	***−1.4***
Cyba	−0.5	0.1	−0.1	0.3	***−1.7***
H2afz	−0.4	0.6	0.4	0.0	***−1.3***
Hsp1b	−0.2	0.2	0.2	0.6	***−1.9***
Mos	0.4	0.7	0.0	**1.0**	***−1.9***
Pms1	0.5	−0.2	0.3	**1.4**	−0.4
PolB	−0.2	0.9	0.4	0.3	***−1.4***
Pdrg1	−0.6	−0.5	0.3	**1.7**	***−2.0***
Prdx2	0.2	0.4	0.0	0.2	***−3.4***
Prm2	0.5	0.5	0.9	0.7	***−4.6***
Rad50	0.3	0.1	−0.1	0.5	***−1.7***
Spin	0.1	0.1	−0.1	0.2	***−1.4***
Tdg	0.6	0.6	0.2	0.6	***−3.0***
Trp53	−0.1	0.2	0.5	0.2	***−1.8***
Usp16	0.3	0.0	0.2	0.3	***−1.1***
Zbed3	**1.7**	−0.6	−0.1	0.8	***−1.9***

Note. Bold values represent statistically significant up- and downregulated genes.

where Cq(min) is the sample with lowest Cq value. The geNorm calculates an expression stability M value (default limit, $M < 1.5$) and sample specific normalization factors (NFs). NFs were calculated as the geometric mean of the expression of the most stable normalizer. The selection of the optimal number of stably expressed genes was based on geNorm's pairwise variation analyses between subsequent NFs using a default cutoff value of 0.15 for the inclusion of additional normalizers (Vandesompele *et al.*, 2002). We also evaluated the suitability of the average gene expression value as a NF alongside of the 51 genes. For each individual sample, the average expression value was calculated based on the 51 genes that passed our filtering criteria. The three most stably expressed genes identified by geNorm were *Carm1*, *Kat2a*, and *Lig3* (Supplementary fig. 1). The geometric mean of the *Carm1*, *Kat2a*, and *Lig3* genes was used to calculate the NFs. The normalized relative expression quantity (NRQ) was generated by dividing the RQ of a sample by the sample specific NF, as follows:

$$\mathrm{NRQ} = \frac{\mathrm{RQ}}{\mathrm{NF}_{(\text{sample specific})}}$$

The fold difference (FD) between B[a]P-exposed and untreated control samples was then calculated by dividing the normalized RQ of B[a]P-exposed sample by the average of normalized RQ of cell stage–matched untreated control samples, as shown below:

$$\mathrm{FD} = \frac{\mathrm{NRQ(B[a]P\text{-}exposed\ samples)}}{\text{Average NRQ(untreated controls)}}$$

The FD was then log base 2 transformed; log2-transformed FDs are preferred for visualization. Using the approach outlined above, we compared gene expression in developing mouse embryos at the 1-, 2-, 4-, 8-, and blastocyst cell stages of embryo development, originating from fertilization with sperm from either B[a]P-exposed fathers or control fathers.

Statistical analysis. Gene expression data often deviate from normal distribution, but after log transformation they are normally distributed. Statistical significance between mean (log2-transformed) values was assayed by Student's t-test. Values were considered to be significantly different when $p < 0.05$. Statistical tests were performed using Stata, version 9.

RESULTS

Zygotes and early embryos from five preimplantation stages (1-, 2-, 4-, 8-, and blastocyst cell stages), fertilized either with sperm from B[a]P-treated male or from control mice, were used to study the effect of paternal exposure to B[a]P on the expression of a panel of 78 mouse mRNAs. It is well known that embryonic cell cycle duration is subject to considerable variation, not only among different embryos but also among blastomers within the same embryo (Bowman and McLaren, 1970). In the present experiment, samples were selected exclusively from the pool of healthy looking embryos representative of that particular stage.

A total of 184 *in vitro* fertilized embryos (derived from 92 B[a]P exposed and 92 control sperm) at five different developmental stages were used for the final analysis. A set of gene expression data was obtained from each embryo. Sample extraction was timed according to 24 h cycles, with the 1-, 2-, 4-, 8-cell, and blastocyst time points referring to 1, 24, 48, 72, and 120 h, respectively. The stages can be readily identified up to the 8-cell stage.

Gene Expression Analysis

We used only those genes that passed the quality assurance criteria described in the Materials and Methods section. After data normalization and filtering, we ended up with a matrix of 51 genes × 184 samples (embryos), and this matrix was used for downstream analysis. The normalized data were used to calculate FD between B[a]P exposed and control samples. A complete list of genes included in the downstream analysis is shown in Supplementary table 1.

We then conducted hierarchical clustering analysis (average linkage and Euclidean distance similarity measurement) in order to group samples from different embryonic stages based on their transcript level similarity using the MeV software version 4.7 (Saeed *et al.*, 2006). The hierarchical clustering analyses results were visualized in a dendrogram and are presented in Figure 1. By visual inspection of the clustering dendrogram, we observed that blastocyst stage samples clustered close to each other and most of the genes at this stage were downregulated following paternal B[a]P exposure; there were however a few samples from the earlier stages, which clustered together with the blastocysts (Fig. 1). The slight upregulation or lack of downregulation of three genes, *Mybl2*, *Pou5f1*, and *Xrcc1*, in five samples is responsible for the segregation of the blastocysts. Many of the 8-cell stage embryos (41%) cluster together in one group characterized by a marked upregulation of several genes; however the majority of the 8-stage embryos cluster together with samples from other developmental stages. In a subanalysis of the 13 samples that cluster together in the

heatmap at the 8-cell stage (Fig. 1), the following genes were significantly upregulated more than twofold in embryos of B[a]P-exposed fathers: *Adam2, Bcl2l2, Chek2, Crem, Ddb2, Ercc1, Ercc8, Fgfbp3, Hspa1b, Mos, Pms1, Pdrg1, Prm2, Rad50, Tdg, Ercc2, Zbed3* (data not shown).

To investigate if the effects of paternal B[a]P exposure were strongly manifested in particular pathways, we looked at relative gene expression based on functional classification of the genes. The results are shown in Figure 2. There is a tendency towards an upregulation of many genes at the 8-cell stage. Still, we found no sign of statistically significant pathway-based enrichment in the effects of B[a]P exposure. The most obvious pattern that emerges is the downregulation of gene expression in embryos of exposed fathers at the blastocyst stage. Table 1 shows the 16 significantly differentially expressed genes ($p < 0.05$ and ± 2-fold cutoff value) between B[a]P-exposed and control samples in one or several embryonic development stages. With the exception of *Pms1*, all of these genes were downregulated more than 2 times at the blastocyst stage following paternal B[a]P exposure.

Integrated Analysis of miRNA-mRNA Interaction

In a recent study, we investigated the effects of paternal exposure to B[a]P on global miRNA expression in the early developing mouse embryo and identified expression of 102 miRNAs (Brevik *et al.*, 2012). In order to identify potential genes whose mRNA might be targeted by these 102 miRNAs, we conducted an inverse correlation analysis of the integrated expression profile data of the 102 miRNAs and the 51 gene list from this study. We first searched for predicted mRNA targets of the 102 miRNA list using the miRWalk database (Dweep *et al.*, 2011). In order to determine how many of our B[a]P responsive genes could be considered predicted targets for B[a]P responsive miRNAs, we conducted similarity searches between the miRWalk database predicted target list, and our 51 membered gene list. We identified 37 genes from our list among the predicted gene target list targeted by 68 of the 102 B[a]P responsive miRNAs identified previously (Brevik *et al.*, 2012). Most of the genes were targeted by more than one miRNA. The expression data for the 37 genes and the expression data for the 68 miRNAs were combined to identify inversely correlated miRNA-mRNA pairs. In particular, we were interested in those miRNA-mRNA pairs showing upregulated miRNA expression level in one or several embryonic development stages, and downregulated mRNA level in the corresponding embryonic stage following paternal B[a]P exposure. We found 33 inversely correlated miRNA-mRNA pairs (19 miRNAs and 17 mRNAs). The resulting network of miRNA-mRNA pairs, constructed using CytoScape, version 2.6 (Cline *et al.*, 2007), is shown in Figure 3. A cluster consisting of miR-709, miR-455, miR-30b, miR-298 and miR-138, and *Ercc8*, *Rad1*, *Tdg*, *Trp53*, *H2afz*, *Mybl2*, *Pdrg1*, and *Pou5f1* was evident from the network analysis.

To ascribe biological relevance to the selected 51 genes, we conducted a functional enrichment analysis using the online FunNet Transcriptional Networks Analysis tool, www.funnet.info (Prifti *et al.*, 2008). The genes that passed our quality assurance criteria, and hence being the basis for Figure 1, were included in this enrichment analysis. The top seven biological processes represented by our gene list were DNA transcription, DNA damage response, cell cycle regulation, chromatin modification, oxidation-reduction processes, apoptosis, and embryo development. A complete table of 40 enriched GO biological processes is included as Supplementary table 2.

DISCUSSION

Changes in reproductive health has been reported for many Western countries and have been linked to exposure to environmental contaminants in association studies. Data supporting the importance of paternal exposure to testicular toxicants for fertility and health of the progeny have been reviewed by Hales and Robaire (2001). Recently, a study with mice documented paternally mediated transgenerational impact of B[a]P exposure on male mice fertility (Mohamed *et al.*, 2010). Using ^{32}P postlabeling, we recently showed very high levels of DNA adducts in human sperm exposed *in vitro* to the B[a]P metabolite BPDE (Sipinen *et al.*, 2010). We also detected B[a]PDE-N(2)-dG adducts in mouse epididymal spermatozoa, i.e., in *caput* sperm 4 days after 3 × 50 mg/kg ip B[a]P (Olsen *et al.*, 2005, 2010). In the present report, we demonstrate that paternal B[a]P exposure of epididymal sperm significantly affects subsequent gene expression in the developing IVF mouse embryo. Although the mechanisms are far from being fully understood, this altered gene expression may be induced by the presence of DNA adducts in sperm DNA. Alternatively, epigenetic modifications may (also) be involved.

Our main objective in this study has been to explore early embryonic transcriptional effects induced by paternal exposure to environmental chemicals. In a previous investigation (Brevik *et al.*, 2011), we reported upregulated gene expression at the 8-cell stage of embryonic development after paternal exposure to glycidamide (GA), the active metabolite of acrylamide. We also noted an initial downregulation at the 1- and 2-cell stage of embryo development following paternal GA exposure. In this study, we further analyzed gene expression patterns in samples from the blastocyst stage. The effects of paternal exposure to B[a]P did not give rise to an initial downregulation at the 1- and 2-cell stage, as opposed to in the GA study. However, a common effect of both these environmental chemicals was to upregulate gene expression in developing embryos at the 8-cell stage. Most of the studied genes were similarly upregulated after both GA and B[a]P exposure (~88% of the genes were upregulated in embryos derived from B[a]P-exposed fathers compared with controls;

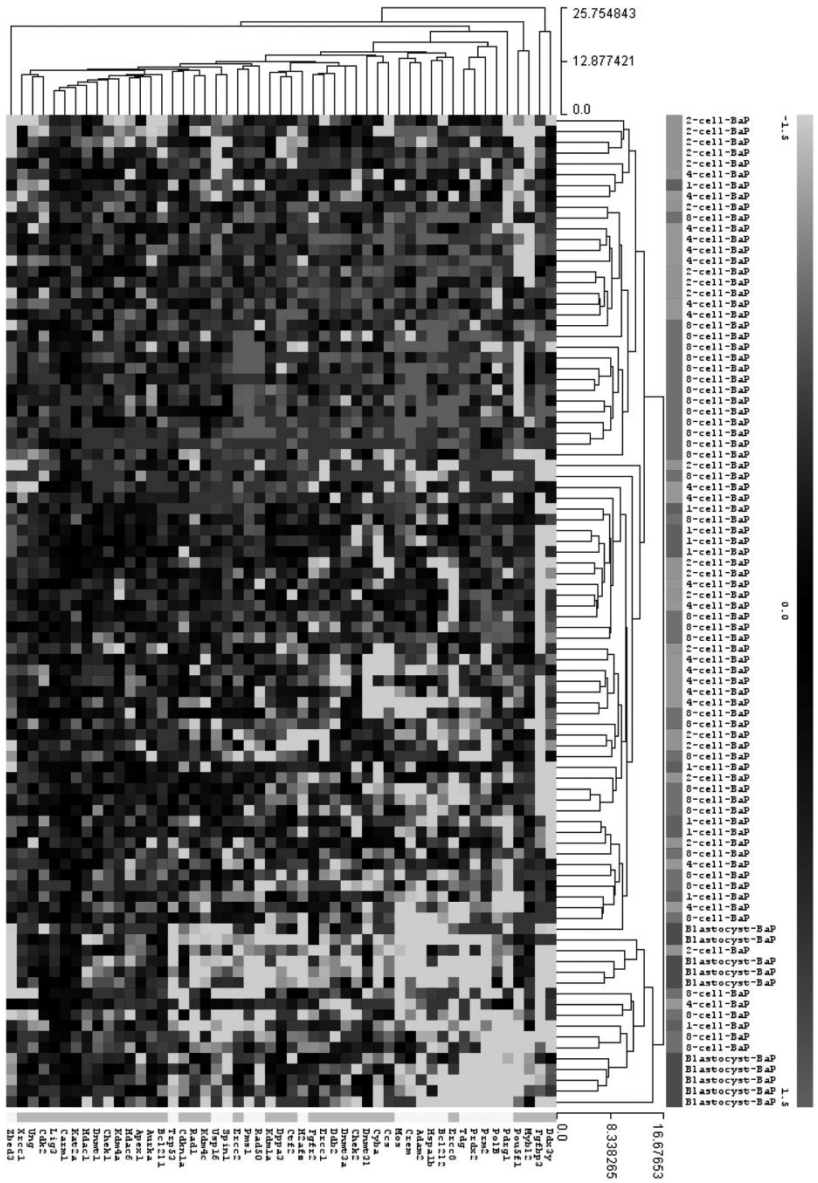


FIG. 1. Unsupervised hierarchical clustering analysis of the relative expression of 51 genes after filtering and normalization of the data. The hierarchical clustering analysis is based on similarities in gene expression. Genes are coded according to their expression patterns (the expression scale spans from -6.0 to 4.5). Samples are horizontally labeled based on the developmental stage they belong to. Vertically, labeled genes indicate more than twofold significantly downregulated expression levels at the blastocyst stage. The covariance value was used as distance metric in this complete hierarchical linkage clustering analysis.

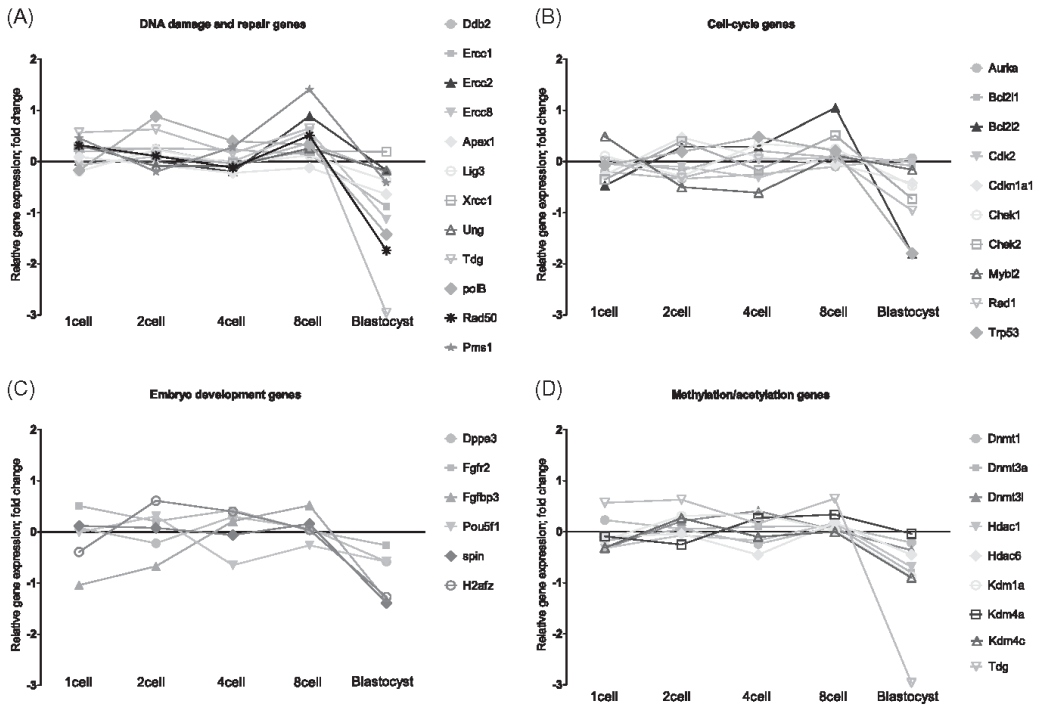


FIG. 2. Relative gene expression at various developmental stages for (A) genes involved in DNA damage repair; (B) genes involved in cell-cycle regulation; (C) genes involved in embryo development; and (D) genes involved in methylation and acetylation pathways. Expression in embryos of exposed fathers is shown relative to expression in control embryos.

for GA, the comparable gene ratio was 87%). Together these observations indicate that the 8-cell stage may represent a sensitive stage of mouse embryo development for chemically induced changes in gene expression. A striking observation in this experiment is the largely uniform downregulation of genes at the blastocyst stage (Figs. 1 and 2). Although no visible differences between the groups were evident upon careful morphological assessment in the microscope, the widespread downregulation of mRNAs in embryos of exposed fathers at the blastocyst stage could indicate that the number of cells was slightly different between the groups, or that the ratio between trophoblasts and embryoblasts was altered. For all practical purposes, counting cells in the microscope is virtually impossible at the blastocyst stage of embryo development without the use of nuclear staining, which was not employed in the present experiment. Previously it has been reported that paternal cyclophosphamide exposure causes decreased cell proliferation in developing embryos (Austin *et al.*, 1994). Furthermore, *Pou5f1* (Oct4) is required to develop a true inner cell mass at the blastocyst stage (Nichols *et al.*, 1998); the relatively low expression of *Pou5f1,* which we observed in embryos of exposed fathers at the 4-, 8-, and blastocyst cell stages (Fig. 2), could be consistent with a compromised development of the inner cell mass in embryos of exposed fathers. Further studies are needed to address whether important phenotypical changes are manifest due to transcriptional dysregulation. Nevertheless, we believe that our rigorous data normalization procedures should be sufficiently stringent to adjust for the effect of potential differences in cell numbers between the groups. The clustering of 41% of the embryos exhibiting significantly upregulated gene expression at the 8-cell stage represents another interesting observation (Fig. 1). It is uncertain why this characteristic pattern of broad upregulation is manifested only in some embryos, whereas others express a more balanced pattern. One explanation could be that some of the exposed embryos develop their full transcriptional competency at slightly different times show a developmental delay. The B[a]P exposed sperm is assumed to inflict variable levels of damage on the cellular machinery in the developing embryos, thus the clustered samples could reflect embryos with more severe lesions than their IVF siblings.

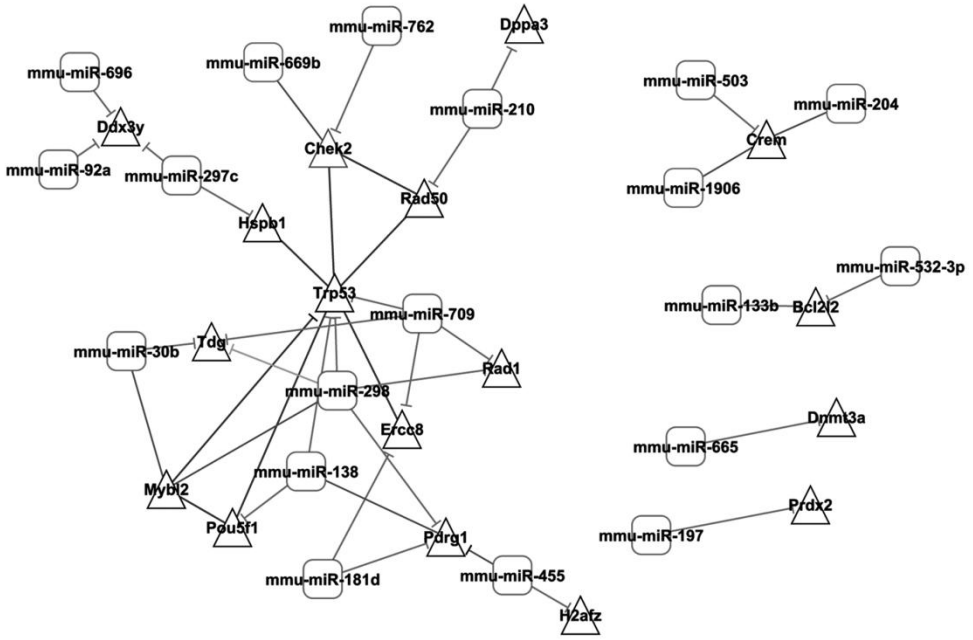


FIG. 3. miRNA-mRNA interaction network of 33 inversely correlated miRNA-mRNA pairs (19 miRNAs and 17 mRNAs). In this network, 16 out of 19 miRNAs (mmu-miR-133b, mmu-miR-138, mmu-miR-181d, mmu-miR-197, mmu-miR-210, mmu-miR-297c, mmu-miR-298, mmu-miR-30b, mmu-miR-455, mmu-miR-503, mmu-miR-532-3p, mmu-miR-665, mmu-miR-696, mmu-miR-709, mmu-miR-762 and mmu-miR-92a) were upregulated and the remaining three miRNAs (mmu-miR-1906, mmu-miR-204 and mmu-miR-669b) were downregulated, in previously determined of B[a]P dysregulated miRNAs (Brevik et al., 2012). These miRNAs are inversely correlated with the 17 mRNAs from this study. Connection lines represent miRNA-mRNA interaction; there are negative miRNA-mRNA pair correlation between the 16 upregulated miRNAs and downregulated mRNAs. There are also connection between genes, and inconsistently correlated miRNA-mRNA pairs.

During the first cell divisions after fertilization, extensive genome-wide demethylation of 5-methylcytosine of both the paternal and the maternal genome occurs, and interference with this process may interfere with embryo development or offspring phenotype. In this study, we included a set of genes related to DNA methylation; *Dnmt1*, *Dnmt3a*, *Dnmt3l*, and *Tdg* and indirectly *Polβ*. *Tdg* and *Polβ* showed significantly altered gene expression levels in early embryos of the B[a]P-exposed fathers. *Tdg* is believed to participate in the active demethylation of the paternal genome (reviewed by Nabel and Kohli [2011]). The altered expression of *Tdg* and *Polβ* during the first cell divisions after fertilization with B[a]P sperm may interfere with the epigenetic status in the early embryo; this warrants further studies.

Bcl2l2, *Trp53, Pdrg1*, and *Pms1* were among the genes that were significantly upregulated at the 8-cell stage (Table 1). *Bcl2l2* has antiapoptotic properties and belongs to a family of prosurvival molecules. It is targeted by two miRNAs, miR-133b and miR-582-3p, and the expression of *Bcl2l2* was inversely correlated with miR-133b at the blastocyst stage (Fig. 3). Recently, a similar inverse correlation was reported in cancer cells, where overexpression of miR-133b was shown to induce apoptosis through downregulation of *Bcl2l2* (Crawford *et al.*, 2009). Another interesting response at the 8-cell stage was the upregulation of *Pdrg1*. *Pdrg1* emerges as a novel cancer biomarker because the expression of this gene has been found to be upregulated in multiple malignancies including cancers of the colon, rectum, ovary, lung, stomach, breast, and uterus (Jiang *et al.*, 2011). Jiang and coworkers also reported selective regulation of *Pdrg1* by agents inducing genotoxic stress. Four different B[a]P dysregulated miRNAs were found to target *Pdrg1* in the current experiment (Fig. 3). More research is however needed to fully characterize *Pdrg1* modulation by its multiple B[a]P responsive miRNA associations. The upregulation of *Pms1* at the 8-cell stage could indicate increased DNA mismatch repair activity. This protein belongs to the mutL/hexB family, and although its DNA repair function has not been proven, it is assumed to have a role in mismatch repair because it forms heterodimers with *Mlh1*, a commonly recognized mismatch repair protein. We recently observed a nonsignificant

upregulation of *Pms1* at the 8-cell stage in embryos of GA exposed fathers (Brevik *et al.*, 2011).

From the matching of miRNAs with their target mRNAs, it is clear that several B[a]P responsive genes, e.g., *Trp53*, *Pdrg1*, *Ddx3y*, *Crem*, *Check2*, and *Bcl2l2,* are targeted by several B[a]P responsive miRNAs (Fig. 3). Based on the present data, we are unable to thoroughly assess the strength of each miRNAs' contribution to the overall regulation of each gene. Rather than activation of a specific molecular mechanism, the matched miRNA and mRNA analyses suggest that the preconceptional paternal B[a]P exposure elicits broad genotoxic stress responses in the developing *in vitro* mouse embryo, with p53 protein as a central player. This study highlights some anticorrelated miRNA-mRNA pairs in the developing mouse embryo resulting from paternal exposure to B[a]P, but more studies are needed to thoroughly understand how key miRNA-mRNA regulatory modules in the developing embryo are affected by environmental contaminants. Considering the relatively narrow time window between exposure and sperm collection, the effects demonstrated in this experiment are mediated by sperm residing in the epididymis at the time of B[a]P treatment. The spermatogenesis is generally susceptible to xenobiotic exposure, but the most sensitive stages vary between different chemicals demonstrating different modes of action. We speculate that, in an evolutionary perspective, pathway-spanning effects that force the embryo to enter a recovery mode, is initiated by the B[a]P-exposed sperm; embryos of exposed fathers may require more time to cope with critical transitions. It is worth noting that the reported gene expression downregulation in embryos of exposed fathers is relative to controls. Hence, the observed downregulation may be interpreted as a delay or attenuation of a natural upregulation of several genes in control embryos.

Based on the present data, one cannot decide whether the B[a]P-induced effects are due to adducts in sperm DNA or may result from other sperm modifications. Our recent report showing BPDE adducts in highly compacted sperm DNA (Olsen *et al.*, 2005, 2010), and also high levels of BPDE adducts in human sperm exposed to BPDE *in vitro* (Sipinen *et al.*, 2010) suggest strongly that direct DNA adducts might be relevant.

In conclusion, paternal exposure to B[a]P leads to altered gene expression in preimplantation mouse embryos. In particular, a significant downregulation of several genes at the blastocyst stage was observed. At the 8-cell stage a clear differential response to paternal B[a]P exposure was found, with nearly 41% of the embryos forming a distinct cluster with significantly upregulated expression of 13 genes. Analysis of the embryonic gene expression suggests a dysregulation of genes related to DNA repair, cell cycle, *Trp53*, and MAPK signaling. It is well known that post-translational modifications may affect the functionality of transcribed gene products. It is at his point, however, not clear to what extent dysregulated gene expression at these early stages of embryo development may result in phenotypic changes, possibly predisposing the individual to disease later in life. This is an intriguing question that clearly deserves further attention. This study represents a first exploration of paternally mediated transcriptomal changes in early embryo development, induced by a common environmental chemical.

SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci.oxfordjournals.org.

FUNDING

The Research Council of Norway (182048/V40); The Czech Republic Ministry of Education, Youth and Sports (AVOZ50520701).

ACKNOWLEDGMENTS

The authors would like to thank Jill Andersen for her help in handling experimental animals.

Authors' Roles: A.B., B.L., G.B., N.D., and A.K.O. contributed to the study conception and design. B.L. and A.B. performed the animal experiments and sample extraction. V.R. performed the RT-qPCR. A.B. and N.D. carried out the data analysis and drafted the manuscript. All authors critically reviewed, contributed to and approved the final version of the manuscript.

REFERENCES

Auger, J., Kunstmann, J. M., Czyglik, F., and Jouannet, P. (1995). Decline in semen quality among fertile men in Paris during the past 20 years. *N. Engl. J. Med.* **332,** 281–285.

Austin, S. M., Robaire, B., Hales, B. F., and Kelly, S. M. (1994). Paternal cyclophosphamide exposure causes decreased cell proliferation in cleavage-stage embryos. *Biol. Reprod.* **50,** 55–64.

Aygün, S. F., and Kabadayi, F. (2005). Determination of benzo[a]pyrene in charcoal grilled meat samples by HPLC with fluorescence detection. *Int. J. Food Sci. Nutr.* **56,** 581–585.

Bergkvist, A., Rusnakova, V., Sindelka, R., Garda, J. M., Sjögreen, B., Lindh, D., Forootan, A., and Kubista, M. (2010). Gene expression profiling—Clusters of possibilities. *Methods* **50,** 323–335.

Bowman, P., and McLaren, A. (1970). Cleavage rate of mouse embryos in vivo and in vitro. *J. Embryol. Exp. Morphol.* **24,** 203–207.

Brevik, A., Lindeman, B., Brunborg, G., and Duale, N. (2012). Paternal benzo[a]pyrene exposure modulates microRNA expression patterns in the developing mouse embryo. *Intl. J. Cell Biol.* **2012,** 407431.

Brevik, A., Rusnakova, V., Duale, N., Slagsvold, H. H., Olsen, A. K., Storeng, R., Kubista, M., Brunborg, G., and Lindeman, B. (2011). Preconceptional paternal glycidamide exposure affects embryonic gene expression: Single embryo gene expression study following in vitro fertilization. *Reprod. Toxicol.* **32,** 463–471.

Carlsen, E., Giwercman, A., Keiding, N., and Skakkebaek, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. *BMJ* **305,** 609–613.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2,** 2366–2382.

Crawford, M., Batte, K., Yu, L., Wu, X., Nuovo, G. J., Marsh, C. B., Otterson, G. A., and Nana-Sinkam, S. P. (2009). MicroRNA 133B targets pro-survival molecules MCL-1 and BCL2L2 in lung cancer. *Biochem. Biophys. Res. Commun.* **388,** 483–489.

Dubrova, Y. E., Nesterov, V. N., Krouchinsky, N. G., Ostapenko, V. A., Neumann, R., Neil, D. L., and Jeffreys, A. J. (1996). Human minisatellite mutation rate after the Chernobyl accident. *Nature* **380,** 683–686.

Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk–database: Prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform.* **44,** 839–847.

Edwards, T. M., Moore, B. C., and Guillette, L. J. Jr. (2006). Reproductive dysgenesis in wildlife: A comparative view. *Int. J. Androl.* **29,** 109–121.

Generoso, W. M. (1986). Relationship between alkylation sites and induction of dominant lethals and heritable translocations in mice. *Prog. Clin. Biol. Res.* **209B,** 493–500.

Generoso, W. M., Cain, K. T., Hellwig, C. S., and Cacheiro, N. L. (1982). Lack of association between induction of dominant-lethal mutations and induction of heritable translocations with benzo[a]pyrene in postmeiotic germ cells of male mice. *Mutat. Res.* **94,** 155–163.

Hales, B. F., and Robaire, B. (2001). Paternal exposure to drugs and environmental chemicals: Effects on progeny outcome. *J. Androl.* **22,** 927–936.

Hattemer-Frey, H. A., and Travis, C. C. (1991). Benzo-a-pyrene: Environmental partitioning and human exposure. *Toxicol. Ind. Health* **7,** 141–157.

Jeffrey, A. M., Weinstein, I. B., Jennette, K. W., Grzeskowiak, K., Nakanishi, K., Harvey, R. G., Autrup, H., and Harris, C. (1977). Structures of benzo(a) pyrene-nucleic acid adducts formed in human and bovine bronchial explants. *Nature* **269,** 348–350.

Ji, B. T., Shu, X. O., Linet, M. S., Zheng, W., Wacholder, S., Gao, Y. T., Ying, D. M., and Jin, F. (1997). Paternal cigarette smoking and the risk of childhood cancer among offspring of nonsmoking mothers. *J. Natl. Cancer Inst.* **89,** 238–244.

Jiang, L., Luo, X., Shi, J., Sun, H., Sun, Q., Sheikh, M. S., and Huang, Y. (2011). PDRG1, a novel tumor marker for multiple malignancies that is selectively regulated by genotoxic stress. *Cancer Biol. Ther.* **11,** 567–573.

Kazerouni, N., Sinha, R., Hsu, C. H., Greenberg, A., and Rothman, N. (2001). Analysis of 200 food items for benzo[a]pyrene and estimation of its intake in an epidemiologic study. *Food Chem. Toxicol.* **39,** 423–436.

Lee, B. M., and Shim, G. A. (2007). Dietary exposure estimation of benzo[a]pyrene and cancer risk assessment. *J. Toxicol. Environ. Health Part A* **70,** 1391–1394.

Marchetti, F., and Wyrobek, A. J. (2005). Mechanisms and consequences of paternally-transmitted chromosomal abnormalities. *Birth Defects Res. C Embryo Today* **75,** 112–129.

Marchetti, F., and Wyrobek, A. J. (2008). DNA repair decline during mouse spermiogenesis results in the accumulation of heritable DNA damage. *DNA Repair (Amst.)* **7,** 572–581.

Mohamed, E., Song, W. H., Oh, S. A., Park, Y. J., You, Y. A., Lee, S., Choi, J. Y., Kim, Y. J., Jo, I., and Pang, M. G. (2010). The transgenerational impact of benzo(a)pyrene on murine male fertility. *Hum. Reprod.* **25,** 2427–2433.

Nabel, C. S., and Kohli, R. M. (2011). Molecular biology. Demystifying DNA demethylation. *Science* **333,** 1229–1230.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95,** 379–391.

Olsen, A. K., Andreassen, A., Singh, R., Wiger, R., Duale, N., Farmer, P. B., and Brunborg, G. (2010). Environmental exposure of the mouse germ line: DNA adducts in spermatozoa and formation of de novo mutations during spermatogenesis. *PLoS One* **5,** e11349.

Olsen, A. K., Lindeman, B., Wiger, R., Duale, N., and Brunborg, G. (2005). How do male germ cells handle DNA damage? *Toxicol. Appl. Pharmacol.* **207,** 521–531.

Pelkonen, O., and Nebert, D. W. (1982). Metabolism of polycyclic aromatic hydrocarbons: Etiologic role in carcinogenesis. *Pharmacol. Rev.* **34,** 189–222.

Penning, T. M. (2004). Aldo-keto reductases and formation of polycyclic aromatic hydrocarbon o-quinones. *Meth. Enzymol.* **378,** 31–67.

Phillips, D. H. (1983). Fifty years of benzo(a)pyrene. *Nature* **303,** 468–472.

Prifti, E., Zucker, J. D., Clement, K., and Henegar, C. (2008). FunNet: An integrative tool for exploring transcriptional interactions. *Bioinformatics* **24,** 2636–2638.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A., and Quackenbush, J. (2006). TM4 microarray software suite. *Meth. Enzymol.* **411,** 134–193.

Shukla, Y., and Taneja, P. (2001). Antimutagenic effect of black tea extract using 'rodent dominant lethal mutation assay'. *Toxicology* **168,** 269–274.

Sipinen, V., Laubenthal, J., Baumgartner, A., Cemeli, E., Linschooten, J. O., Godschalk, R. W., Van Schooten, F. J., Anderson, D., and Brunborg, G. (2010). In vitro evaluation of baseline and induced DNA damage in human sperm exposed to benzo[a]pyrene or its metabolite benzo[a]pyrene-7,8-diol-9,10-epoxide, using the comet assay. *Mutagenesis* **25,** 417–425.

Somers, C. M., Yauk, C. L., White, P. A., Parfett, C. L., and Quinn, J. S. (2002). Air pollution induces heritable DNA mutations. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 15904–15907.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3,** 34.

Verhofstad, N., van Oostrom, C. T., van Benthem, J., van Schooten, F. J., van Steeg, H., and Godschalk, R. W. (2010). DNA adduct kinetics in reproductive tissues of DNA repair proficient and deficient male mice after oral exposure to benzo(a)pyrene. *Environ. Mol. Mutagen.* **51,** 123–129.

Zenzes, M. T., Bielecki, R., and Reed, T. E. (1999a). Detection of benzo(a)pyrene diol epoxide-DNA adducts in sperm of men exposed to cigarette smoke. *Fertil. Steril.* **72,** 330–335.

Zenzes, M. T., Puy, L. A., Bielecki, R., and Reed, T. E. (1999b). Detection of benzo[a]pyrene diol epoxide-DNA adducts in embryos from smoking couples: Evidence for transmission by spermatozoa. *Mol. Hum. Reprod.* **5,** 125–131.

5.2. Analýza dat

Statistická analýza dat je klíčová pro jejich následnou interpretaci. Analýza jednotlivých buněk pomocí qPCR je teprve v počátku a neustále se hledají vhodné postupy, jak z dat vytěžit co nejvíce a ukázat je co nejsrozumitelněji. Přesto zatím nebyl pro jejich statistickou analýzu vypracován žádný doporučený postup, jako je uveden pro klasické qPCR experimenty (Bustin et al., 2009). Následující publikace se zabývají vypracováním metody pro analýzu a vizualizaci dat.

Gene expression profiling – Clusters of possibilities, Bergvist, A., **Rusňáková, V.**, Šindelka R., Garda J. M. A., Sjorgreen B., Lindh D., Forootan A. and Kubista M., Methods 50 (2010), p. 323-335, IF = 3,641.

GenEx: Data analysis software, Kubista M., **Rusnakova V.**, Svec D., Sjorgreen B., Tichopad A.**,** Quantitative Real-time PCR in Applied Microbiology (editor M. Filion, 2012), p. 63-84.

RT-qPCR work-flow for single-cell data analysis, Stahlberg A., **Rusňáková V.**, Forootan A., Anděrová M. and Kubista M., Methods 59 (2013), p. 80-88, IF = 3,641.

The added value of single-cell gene expression profiling, Stahlberg A., **Rusňáková V.** and Kubista M., Briefings in Functional Genomics 59 (2013), p. 80-88, IF = 4,210.

V publikaci *Gene expression profiling – Clusters of possibilities* (Bergkvist et al., 2010) je ukázáno, jak lze využít analýzu hlavních komponent, dendrogramy, heat-mapy a samoorganizující se mapy pro analýzu dat. Použití těchto uvedených typů analýz bylo provedeno na vývoji drápatky vodní (*Xenopus laevis)*, od vajíčka po pulce. Neoplodněná vajíčka obsahují mateřskou RNA pro maternální geny, jež jsou zodpovědné hlavně za oogenezi a počáteční vývojové procesy. V pozdějším vývoji dojde k degradaci této RNA a embryo si začne tvořit svojí vlastní zygotickou RNA. Tento okamžik se nazývá mid-blastulární transice (MTB). Vzorky byly měřeny na vysokokapacitním qPCR přístroji Biomark a analyzovány pomocí softwaru GenEx, ve kterém jsou implementovány výše uvedené statistické postupy pro vizualizaci vícerozměrných dat. Všechny čtyři typy analýz rozdělily vzorky do tří skupin (časný vývoj, MTB a pozdní vývoj) na základě exprese měřených genů. Zároveň rozdělily geny na maternální, mid-blastulární a zygotické. Mezi testovanými geny byly takové, které mají známý původ i zcela neznámé. Přidáním neznámých genů mezi známé umožní lépe zjistit jejich zařazení. Jedná se však pouze o explorativní analýzu, důležitou pro vytvoření hypotézy, která se následně potvrdí na novém vzorku (konfirmační analýza). Bez potvrzení výsledku konfirmační analýzou nejsou naměřená data statisticky významná. Na této práci jsem se podílela naměřením dat a jejich následnou analýzou.

Kapitola v knize *Quantitative Real-time PCR in Applied Microbiology* představuje software GenEx, který se používá pro analýzu qPCR dat. Tento statistický software se skládá se ze dvou modulů – modulu pro úpravu dat a modulu pro samotnou analýzu. Úprava dat je možná, tak jak bylo popsáno v kapitole 3.2.1. Modul pro analýzu obsahuje 9 různých typů analýz, a to výpočet hlavních komponent, shlukovou analýzu, nástroje pro tvorbu umělých neuronových sítí, výpočet regrese, nástroje pro hledání referenčních genů, statistické porovnání skupin vzorků a návrh experimentu. Principy těchto analýz jsou převážně diskutovány v kapitolách 3.2.2 a 3.2.3. Při vývoji tohoto programu jsem se podílela na jeho testování a vylepšování jeho funkčnosti pro uživatele.

Publikace *RT-qPCR work-flow for single-cell data analysis* a *The added value of single-cell gene expression profiling* (Stahlberg et al., 2013b; Stahlberg et al., 2013a) shrnují vypracovanou metodu analýzy jednotlivých buněk a ukazují, jak získat další informace o fungování buněk, pokud se analyzují samostatně. Je zde kladem důraz na způsob nahrazení chybějících dat a na normalizaci. Analýza jednotlivých buněk se oproti klasickému qPCR experimentu odlišuje v několika bodech. Největší rozdíl je v normalizaci. Klasicky se qPCR data normalizují pomocí referenčního genu (Vandesompele et al., 2002; Andersen et al., 2004), což však není v případě analýzy exprese u jednotlivých buněk vhodný postup z důvodu existence transkripčních vzplanutí. V nich se hladina exprese daného genu liší v čase a jednotlivé buňky tak nejsou synchronizovány (Kapitola 3.1.1). Proto se data normalizují pouze na množství buněk. Další odlišností oproti klasickému qPCR experimentu je množství chybějících dat. Až v 70 procentech případů údaje chybí nebo jsou hodnoty Cq natolik vysoké, že nejsou důvěryhodné. Takto nedůvěryhodným datům je přiřazena nejčastěji hodnota limitu detekce pro daný gen. Chybějící data, která nebyla naměřena vůbec, jsou nahrazeny hodnotou limitu detekce pro daný gen (vyjádřené jako Cq hodnota) plus 2 nebo plus 4, tím se exprese chybějících dat rovná jedné čtvrtině, respektive jedné šestnáctině nejmenšího množství transkriptu daného genu v buňce. Dalším rozdílem oproti klasickému qPCR je počet vzorků nutný k analýze. Vzhledem k tomu, že 20 % buněk exprimuje až 75 % transkriptů daného genu (Stahlberg et al., 2011b), je nutné měřit 100 a více buněk, aby byly výsledky statisticky důvěryhodné. Naměřené hodnoty se prezentují nejčastěji pomocí histogramů, které také slouží k prvnímu odhadu, zda se ve vzorku vyskytuje více typů buněk.

K zjištění heterogenity buněk se dále používají multivariantní analýzy (Kapitola 3.2.3), pomocí kterých se buňky rozdělí do skupin a na jednotlivé subpopulace. Ty musí být dále dány do biologického kontextu a statisticky popsány (počet pozitivních buněk ve skupině, průměr, medián, rozdělení, minimum, maximum atd.). Vzhledem k tomu, že data nejsou normálně rozdělena, se průměrný počet kopií v buňce

vyjadřuje geometrickým průměrem a ne aritmetickým (Bengtsson et al., 2008). Pro porovnání jednotlivých subjektů, ze kterých byly měřeny pouze jednotlivé buňky, je nutné najít referenční gen, který by vyrovnal rozdílnou expresi a množství buněk. Pro jeho určení se sečte celkové množství transkriptů pro každý gen a subjekt zvlášť a nalezne se nejstabilnější gen (geny) vhodný k normalizaci (Kapitola 3.2.1). Přestože je gen stabilní pro jednotlivé subjekty, není stabilní uvnitř populace buněk. Jeho exprese se mezi jednotlivými buňkami může lišit až 1000-násobně.

Na uvedených pracích jsem se podílela sběrem buněk, jejich qPCR analýzou a kompletním statistickým zpracováním. Data byla měřena na astrocytech a jejich biologická interpretace je diskutována v následující kapitole.

Methods 50 (2010) 323–335

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Gene expression profiling – Clusters of possibilities

Anders Bergkvist [a], Vendula Rusnakova [b], Radek Sindelka [c], Jose Manuel Andrade Garda [d,1], Björn Sjögreen [a], Daniel Lindh [a], Amin Forootan [a], Mikael Kubista [b,e,*]

[a] MultiD Analyses AB, Gothenburg, Sweden
[b] Laboratory of Gene Expression, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[c] Whitehead Institute, Cambridge, UK
[d] Department of Analytical Chemistry, University of Corunna, Campus da Zapateira, 15071, Corunna, Spain
[e] TATAA Biocenter AB, Odinsgatan 28, SE-411 03 Gothenburg, Sweden

ARTICLE INFO

Article history:
Accepted 12 January 2010
Available online 15 January 2010

Keywords:
qPCR
Statistical analysis
Xenopus laevis
Hierarchical clustering
Heatmap
PCA
SOM
Hypotheses

ABSTRACT

Advances in qPCR technology allow studies of increasingly large systems comprising many genes and samples. The increasing data sizes allow expression profiling both in the gene and the samples dimension while also putting higher demands on sound statistical analysis and expertise to handle and interpret its results. We distinguish between exploratory and confirmatory statistical studies. In this paper we demonstrate several techniques available for exploratory studies on a system of *Xenopus laevis* development from egg to tadpole. Techniques include hierarchical clustering, heatmap, principal component analysis and self-organizing maps. We stress that even though exploratory studies are excellent for generating hypotheses, results have not been proven statistically significant until an independent confirmatory study has been performed. An exploratory study may certainly be valuable in its own right, and there are often not enough resources to report both an exploratory and a confirmatory study at the same time. However, exploratory and confirmatory studies are intimately connected and we would like to raise that awareness among qPCR practitioners. We suggest that scientific reports should always have a hypothesis focus. Reports are either hypothesis generating, from an exploratory study, or hypothesis validating, from a confirmatory study, or both. In either case, we suggest the generated or validated hypotheses be specifically stated.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Modern scientific endeavours invariably involve reproducibility. To cite Werner von Siemens (1816–1892), 'good science needs good measurement'. As scientific questions continue to be resolved, the complexity of the remaining scientific questions increase. Complex scientific questions are likely to be obscured by unrelated processes that we as scientists often ascribe to random variables or noise. A distinction between the desired scientific knowledge and unrelated noise often requires statistical analysis. Although this is simple in principle, human intellect does not naturally seem to be optimally suited to parse such relationships and mistakes are often made because of subjective decisions. In this paper we will therefore attempt to describe some basic checkpoints that would facilitate correct handling of exploratory statistical analyses. We will also describe several alternative statistical methods relevant for pattern recognition by multivariate data analysis that are common for complex scientific questions.

A central checkpoint in statistical analysis is the hypothesis. We distinguish between an exploratory statistical study, if the aim of the study is to generate one or several hypothesis from a given data set, and confirmatory statistical study, if the aim of the study is to validate a given hypothesis or set of hypotheses by collection and analysing a new data set. We advocate that all statistical studies should provide a clear description of the hypothesis used, in case of a confirmatory study, or generated, in case of an exploratory study.

The statistical rigour required in a confirmatory study is very strict since we want to use the results for validation. Prior to collection and any testing of data, we need to define the hypothesis, what statistical test we are going to perform and what criteria we are going to use to judge whether we can or cannot reject the hypothesis, including whether the test will be one- or two-sided. It is important to make these definitions before data collection because a posterior selection of hypotheses, test and/or criteria may bias the conclusion and/or lead to an unintentional introduction of multiple testing which may compromise any statistical signifi-

* Corresponding author. Address: TATAA Biocenter AB, Odinsgatan 28, SE-411 03 Gothenburg, Sweden. Fax: +46 31 152890.
E-mail addresses: andrade@udc.es (J.M.A. Garda), mikael.kubista@tataa.com (M. Kubista).
[1] Fax: +34 981 167065.

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.01.009

cance found in this manner. In contrast to a confirmatory study, an exploratory study does not need to be defined prior to data collection, since the purpose of the exploratory study is to formulate a suitable hypothesis. Any approach that results in such a hypothesis may be a valid approach. However, it will not be clear whether the newly found hypothesis itself is statistically valid on the desired population and in the desired experimental setting until it has been validated in a confirmatory study.

Many mathematical and computational tools are available for exploratory studies. In this paper we will focus on visualization, clustering and partition methods for this purpose.

1.1. Embryonic development in Xenopus laevis

As an example to illustrate the power of visualization, clustering and partition methods for exploratory studies we will use measurements on transcripts coded by genes that are crucial for early embryonic development of African clawed frogs, *Xenopus laevis*.

RNA expression can be studied on two levels: as a spatial RNA distribution and a temporal RNA expression. The spatial expression profiles can be measured within different organs or different parts of the embryo [1], while temporal expression profiles are determined as function of development stage or time. Temporal expression profiling during early development of mammalian systems is limited by the small amounts of protein and RNA (few hundreds of pg per embryos) and time and money consumed preparing embryos. In contrast hundreds or thousands of embryos can be easily obtained by stimulating and in vitro fertilizing a single amphibian female. Further, amphibian eggs and embryos, such as those of *Xenopus laevis*, contain several μg of total RNA and proteins [2]. These features have made *Xenopus laevis* one of the most popular organisms for developmental studies. Two groups of mRNAs can be distinguished, based on temporal expression. One set, called maternal, is synthesized during oogenesis inside the mother's body. Proteins coded by maternal genes often have many roles in oogenesis and early developmental processes. Maternal mRNA molecules are translated into functional proteins in the oocyte during oogenesis and in the embryo after fertilization. At some point, the embryo reaches a stage when specialized gene products are needed and zygotic transcription is initiated. This point is called mid-blastula transition (MBT) and takes place at different developmental stages in different organisms. For example, mammalian zygotic transcription usually initiates after few cell divisions, while in *Xenopus* MBT occurs after 12 cell divisions in the gastrulation stage. The genes expressed after MBT are called zygotic genes.

Thirty-one genes that are crucial for early development of *Xenopus* were selected for high throughput qPCR expression profiling. Genes such as VegT, disheveled, p53, ubiquitin (ubc), Vg1, Xdazl, Xcad2, Oct-60, DEADSouth, alpha-tubulin, Stat3, U3-snoRNA, cytokeratin, Est1, Xmam1, An1, An2, 18S rRNA, mitochondrial cytochrome C (mt-cytC) and Wnt11 were previously found to be key components for early development and expressed in oocyte [3]. Similarly we selected genes, which were found to be important for developmental stages around and after MBT and therefore predicted to be expressed, such as siamois, chordin, HNF-3beta, Pax6, goosecoid, derriere, follistatin, cerberus and N-CAM. Thirteen developmental stages in biological triplicates from oocyte to tadpole were collected for high throughput qPCR analysis.

Primers and samples were loaded into a microfluidic chip, run and analyzed in the high-throughput BioMark qPCR platform [4]. Chips for dynamic qPCR analysis, which were used in our experiment, allowed us to run 48 cDNA samples times 48 primer pairs in parallel in a single run. Each run thus resulted in 2304 independent reactions with all cDNA sample:gene primer pair combinations. The analysis software GenEx developed by MultiD Analyses AB [5] was used for analysis and visualization of the data. Many other softwares are available to perform the studies presented herein; for instance, SPSS (SPSS Inc.), StatGraphics (StatPoint Technologies, Inc.) or open-code R and/or Matlab programs which can be downloaded freely, to mention but a few which are employed commonly.

1.2. Paper outline

The data analysis and visualization is performed using the analysis software GenEx, developed by MultiD Analyses AB [5].

Data measured with the BIOMARK platform are read and automatically annotated by the GenEx software. Analysis then starts with data preprocessing. This consists of normalization, imputation of missing data, removal of outliers, and scaling of data. Here normalization means scaling with endogenous reference or control samples to reduce systematic variations in the data. Scaling refers to rescaling needed to make the data analysis algorithms well conditioned. For example, mean centering of the data.

Some insight can be gained by visualizing the preprocessed data without further analysis. We will discuss visualization before proceeding to clustering. For clustering, we will consider four different methods to group the genes: three agglomerative hierarchical clustering methods, a divisive clustering approach, clustering by principal component analysis (PCA), and clustering by self organized maps (SOM).

2. Description of method

2.1. Sample collection and RNA isolation

Xenopus laevis females were stimulated by 500 U of hCG and kept overnight at 22 °C. Males were anesthetized in 0.1% tricaine solution for 20 min and testes were removed. A homogenized testes solution was used for in vitro fertilization by pouring to freshly squeezed oocytes. 0.1× Marc's Modified Ringers (MMR) medium was added after 5 min incubation at room temperature. After about 30 min after fertilization MMR medium was replaced by 2% cysteine solution to remove jelly coat. After short incubation, cysteine solution was removed and embryos were washed five times with 0.1× MMR solution. Thereafter the developing embryos were kept in 0.1 MMR at 25 °C.

Xenopus laevis embryos were staged according to [6]. Three sets of embryos (three embryos from the same female in each set) were collected from developmental stages 1, 2, 5, 6.5, 9, 10, 11, 13, 15, 21–22, 24–25, 38 and 44 and frozen at −70 °C. Total RNA from each sample was extracted using the RNeasy Mini kit (Qiagen) according to the manufacturer's instructions, including on column DNase treatment. Total RNA was eluted into 30 μl of elution buffer. Concentration of total RNA was measured with a Nanodrop instrument (Thermo Scientific). The RNA quality was analyzed on an Experion system (Bio-Rad).

2.2. Reverse transcription

cDNA was produced starting with 100 ng of total RNA, 1.5 μl of mixture 10 μM oligo-dT and 10 μM random hexamers (1:1) and water. The total volume was 6.7 μl. After incubation at 72 °C for 10 min, 100 U of MMLV reverse transcriptase (Promega), 12 U RNasin (Promega), 5 nmol dNTP and 2 μl buffer (5×) were added to a total volume of 10 μl, and incubated at 37 °C for another 70 min. The product was subsequently diluted to 100 μl and frozen.

2.3. Primer design and preamplification

Primers for the amplification of 31 selected genes were designed using Primer3 [7] and Beacon Designer (Premier Biosoft).

Primers' specificities and assay efficiencies were tested on control cDNA (mixture of cDNA from the three developmental stages). Criteria to accept a primer pair were: specific amplification of control cDNA with *Cq* lower than ~35, one peak in melting curve analysis, and no amplification of negative controls.

Preamplification PCR was run in 20 μl containing 2 μl of cDNA, 1 μl of all forward and reverse primers (500 μM each), 10 μl of Sigma A mastermix (kindly provided by Sigma, not yet a commercially available product) and water. CFX 96 cycler from Bio-Rad was used for preamplification with the cycling conditions: predenaturation at 95 °C for 2 min., followed by 14 cycles (95 °C 15 s., 59 °C 1 min. and 72 °C 1 min.). Product of the preamplification was diluted from 20 to 80 μl (4×) and stored at −20 °C. Preamplification efficiency was validated by comparing qPCR results of template that was and was not preamplified. Differences in *Cq* values between preamplified and not preamplified samples were similar for all genes [8], reflecting minimal bias and thus confirming the reliability of the preamplification.

2.4. High throughput qPCR performed on BioMark system

For qPCR analysis using the BioMark dynamic array (Fluidigm) a cDNA sample reaction mixture and a primer reaction mixture were prepared. The sample reaction mixture had a final volume of 5 μl and contained 1 μl of cDNA, 0.5 μl of SYBR Green Sample Loading reagent (Fluidigm), 2.77 μl Sigma A mastermix (Sigma, not provided yet), 0.165 μl of Chromophy, diluted 1:25 (TATAA), 0.025 μl of ROX (Invitrogen) and 0.1 μl of JumpStart DNA Taq polymerase (Sigma). The primer reaction mixture had a final volume of 5 μl and contained 2.5 μl of Assay Loading reagent (Fluidigm) and 2.5 μl mixture of reverse and forward primers corresponding to a final concentration of 10 μM. The chip was first primed with oil solution in the NanoFlex™ 4-IFC Controller (Fluidigm) to fill control valves. Bubbles were carefully removed from 5 μl of preamplified cDNA in sample reaction mixture and loaded into the sample wells, and 5 μl of the primer reaction mixtures was loaded into the assay wells of the dynamic array. The dynamic array was then placed on the NanoFlex™ 4-IFC Controller for automatic loading and mixing. After about 55 min the dynamic array was transferred to the BioMark qPCR platform (Fluidigm). The cycling program was 3 min at 95 °C for preactivation, followed by 30 cycles of denaturation at 95 °C for 15 s, annealing at 60 °C for 20 s, and elongation at 72 °C for 20 s. After completed qPCR melting curves were collected between 60 and 95 °C with 0.5 °C increments.

2.5. qPCR basic data analysis

An automatic exposure time with 72 °C calibration temperature was set up for measurement of fluorescence. Fluorescence signals were measured in the two channels: ROX and FAM-MGB. The raw FAM fluorescent data were normalized to the ROX signal. A linear baseline correction was used and the same threshold level was used for all assays. Quality threshold was set to 0.65.

2.6. Data preprocessing

qPCR data are frequently normalized by one of several options, including the expression of reference genes, number of cells, weight of tissue, DNA/RNA spike and total RNA concentration [9]. Expression of common *Xenopus laevis* reference genes in temporal developmental studies was found to be highly variable and unsuitable for normalization [10]. Endogenous reference genes are often powerful ways of reducing confounding variability introduced by technical handling. By not using endogenous reference genes, we emphasize careful technical handling. It may be more challenging to verify studied effects under these conditions, although it is certainly possible given that the studied biological variation is sufficiently larger than the confounding variation.

PCR products that gave unacceptable melting curves were classified as off-scale. Furthermore a cut-off of 28 cycles was used, and all *Cq*'s above 28 were treated as off-scale. The limit of 28 cycles was chosen due to our experience of this as a limit of reliable detection of true products for our BioMark instrument given the prior preamplification step. Measurements above 28 cycles were therefore judged to be extra sensitive to experimental handling errors. Off-scale *Cq* values were removed from the data set and subsequently treated as missing data. Patterns in remaining replicates indicated that missing data were due to experimental handling errors rather than concentration limiting errors. The information contained in the biological replicates was used to replacing missing data by the average of remaining biological replicates when available. If all biological replicates gave missing data, they were assigned the highest measured *Cq* for that particular gene +1. Since the highest measured *Cq* of a truly positive sample can be assumed to be the limit of detection (LOD) for that particular gene, assigning Cq(LOD)+1 to the off-scale samples corresponds to a concentration that is half of the LOD. Being below LOD is not equivalent to the sample being truly negative, because of sampling ambiguity. Hence, this is a rational correction of off-scale data for downstream processing of the results. The correction was further validated by reanalyzing the data assigning Cq(LOD)+2 to the off-scale measurement. This gave indistinguishable results evidencing that the correction does not affect the conclusions reached. How to deal with missing (outlying) data is a hot topic in statistics as any approach involves '*inventing the datum*' to some extent. The use of the average of the remaining values or, better, their mean assures that, at least, we do not disturb the major patterns on the dataset, which many times seems reasonable. Nevertheless, the reader should take into account that many other methods exist to cope with this difficult issue.

The relative expression among samples was calculated as [11]:

$$RQ = 2^{Cq_{\min} - Cq}$$

ΔCq values were calculated for each gene *Cq* by subtracting it from the lowest sample *Cq* and then converting the difference to linear scale as shown in the equation above. These relative quantities (RQ) indicate the level of expression, in each sample, of a particular gene relative to the sample in which the gene has highest expression. Hence, the RQ of the sample with highest expression for a particular gene is set to one and all other samples for that gene have $RQ < 1$.

Data must be normally distributed for analysis with parametric tests, such as the Student's *t*-test, linear regression, and ANOVA. If data transformed to RQ is normal distributed, further statistical analysis of data on this format may be preferred. However, gene expression data are usually not normal distributed when expressed as relative quantities, but usually become normal distributed by logarithmic transformation to fold differences (FD). Traditionally, log base 2 is used:

$$FD = \log_2(RQ)$$

Once the data has been transformed, it is, then, required to asses that outliers are not present into the series and, so, that normality is not clearly violated and that the parametric tests can be applied. The Grubbs' test [12–14] is generally recommended. For statistical analysis purposes fold differences are equivalent to the corresponding ΔCq values. However, for visualization purposes, the fold differences are preferred since they provide useful control over the relative pivot point as defined in the RQ transformation step.

2.7. Data pre-treatment

In general we analyze data with a hypothesis (or objective) in mind. This may have been determined *a priori* and our ambition is to validate it (confirmatory study) or our ambition is to browse through the data to propose a new suitable hypothesis (exploratory study). Regardless of whether the hypothesis is known *a priori* or not, we often know that some sort of comparison will be needed to study our data. To emphasize certain aspects of such a comparison it is therefore often appropriate to remove effects that may be present on the data, typically due to the different scales and ranges of variation of the genes. This requires scaling the data before further analysis. Here we describe two options: mean centering and autoscaling [11].

Genes are expressed at very different levels. Most genes have only a few transcripts per cell, while some few have tenths of thousands. In an analysis the few highly expressed genes will have much higher weight and may totally dominate the result. If this is not desired, the effect of the genes expression levels can be removed by subtracting the mean expression of every gene to the corresponding gene.

$$FD_{MC} = FD - \overline{FD}$$

Such transformed data are called mean centered. For mean centered data a certain deviation from normal (=mean) expression has the same weight independently of the expression level (or scale) of that gene.

Low expressed genes may show higher relative standard deviation (coefficient of variation) than high expressed genes, but hardly normal standard deviation. While a low abundant gene present in less than 10 copies can have standard deviations of perhaps 3–5 copies, a high expressed gene present in 100.000 copies will have a standard deviation of at least a few thousand copies. To remove also the effect of the magnitude of the change, data are further divided with the standard deviation:

$$FD_{AS} = \frac{FD - \overline{FD}}{SD} = \frac{FD_{MC}}{SD}$$

These new data are called autoscaled. An important drawback of autoscaled data is that subjects which vary randomly are given the same importance as subjects with systematic information because all variables are scaled to unit variance. Genes, whose expression is not sensitive to the studied parameters, should therefore be avoided, since they contribute only with noise to the analysis.

2.8. Direct visualization methods

As described in the introduction, our goal with the present report is to illustrate how an exploratory study can produce one or several testable hypotheses for future confirmatory statistical analyses. A hypothesis can be generated in many ways; however, it stands to reason that a good visualization improves a researcher's chances of proposing a good hypothesis. Here we distinguish between direct visualization methods that visualize the pre-treated or raw expression values of each (gene or sample) subject in the data set. Later we will see that visualizations of indirect descriptors, such as similarity measures or selected linear combinations of expression values, of the data set may be useful alternative ways to illustrate the data set.

One of the direct visualization tools is the *scatter plot*. In the scatter plot a subject (a sample or an mRNA target) is represented by a point in a coordinate system based on the expressions on each coordinate axis. However, due to human sensory limitations only two or three dimensions can conveniently be visualized in this way.

2D and 3D-line plots are ways to visualize one or several expression profiles. It is particularly useful to present trend studies where expression is measured as a function of time, drug load or other metric variable. However, when handling many and large expression profiles it may be difficult to distinguish features in such a plot. Once groups in the data set have been identified, typically by other means, the profiles in the trend plot can be coloured to emphasize different trend groups.

2.9. Hierarchical clustering methods

From a technical point of view, clustering can be performed either agglomeratively (i.e., by iteratively joining subjects (and small clusters) to form increasingly larger clusters, until all subjects have been accounted for in a comprehensive group), this is called *Hierarchical clustering*, or divisively (i.e., by iteratively breaking up clusters until only individual entities remain), this is called *Divisive clustering*. The latter option is not used frequently despite advanced algorithms exist, among them the '*k-means*' and the '*medoids*' ones. Both are partitioning methods in the sense that groups are found out from the initial overall set of data (which initially constitutes one group). The two are closely related and differ in how they minimize the distances, being the *medoids* method less sensitive to outliers. Medoids can be defined as those objects of a cluster, whose average (dis)similarity (mathematically, similarity and dissimilarity are closely related) to all the objects in the cluster is minimal; i.e., they are the most centrally located point in each cluster. The *medoids* or '*partitioning around medoids*' (PAM) algorithm clusters the objects (expression profiles, samples, etc.) into any of k clusters according to their (dis)similarity to each medoid. Its major difficulty is that k has to be decided in advance by some other means. PAM has been used successfully in the context of gene expression profiling [15,16]. Nevertheless, as agglomerative methods are employed much more frequently than the divisive ones, we will focus on them.

Clustering requires the selection of both a measure of similarity between samples (or genes) and a clustering algorithm.

2.10. Similarity measures

Clustering techniques are based on predefined criteria of similarity. In one of its most intuitive forms similarity is measured just as distances between points in the multidimensional space that is described by the expression vectors of assay samples or target genes. In fact, similarity has to be calculated as the inverse of the distance (the higher the distance, the lower the similarity). Note that it is equally possible to cluster samples (defined by the measured genes) and variables themselves (in this case the genes are grouped as a function of their expression throughout the different samples). To generalize the explanations, the term 'subject' will be used here to indistinctly refer to samples and variables.

There are many ways to mathematically define distance and, unfortunately, no '*golden rule*' can be given for a particular application. In general, the scientist has to try different options and select the most suited one (to what he/she is looking for). A note of caution is needed here as this does not mean that the scientist has to use the clustering methods to '*demonstrate*' its *preliminary* ideas; recall that clustering methods are intended to discover patterns among the datasets and, accordingly, something unexpected (or 'new' groups) should be studied carefully [17]. Here only some common distances will be discussed. The Euclidean distance is very intuitive as it generalizes the well-known Pythagoras' theorem.

$$E_{12} = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

E_{12} is the distance between samples 1 and 2 and x_{ji} represents the expression of each of the n genes ('i') measured on each sample ('j'). Small differences between each of the expression measurements x_{ji} result in a small distance E, which in turn can be interpreted as a measure of high similarity.

Another class of similarity measure is the correlation coefficient. In contrast to the distance measures correlation coefficients emphasize variational relationships between expression vectors. By variational relationship we mean that the differences between each of the expression measurements need not be small, but vary consistently throughout the sample vectors. For expression profiling purposes this is particularly useful to detect positive as well as negative correlations. The classical Pearson's correlation coefficient for two samples x and y, is [12,13]

$$r_{xy} = \frac{n\sum_{j=1}^{n} xy - \sum_{j=1}^{n} x \sum_{j=1}^{n} y}{\sqrt{\left[n\sum_{j=1}^{n} x^2 - \left(\sum_{j=1}^{n} x\right)^2\right]\left[n\sum_{j=1}^{n} y^2 - \left(\sum_{j=1}^{n} y\right)^2\right]}}$$

Here r_{xy} is the correlation between genes expression vectors of samples x and y where x_j and y_j indicate the expression values of the jth gene in the data set of n genes for the x and the y samples, respectively.

2.11. Clustering methods

As seen, measuring similarity between pairs of samples or variables (expression profiles) is relatively simple using a distance measure. Once samples (genes) with the highest similarity (lowest distance) were found, they are merged in a 'cluster' and the process is repeated. However, similarity measures between clusters need further definitions. To address this issue, a *clustering algorithm* has to be selected by the analyst. Four common options are reviewed here:

In the single linkage (or nearest neighbour) method groups are fused according to the distance between their nearest subjects, which are taken as representatives of their corresponding groups. The complete linkage (or furthest neighbour) method, behaves in the opposite way, as the distance between groups is now defined as the distance between their most remote subjects. The unweighted pairs or average linkage defines distance between groups as the average of the distances between all pairs of individuals in the two groups. It is sometimes also referred to as UPGMA (Unweighted Pair-Group Method using Arithmetic averages) [18,19]. It is a compromise between the single and complete linkages. Ward's method is more complex as it calculates the increase in the variance of the distances for the different possibilities of joining clusters. For each possibility, internal variance is computed as the sum of distances between each sample in the group and the group's centroid. The clustering that yields the lowest increment on the sum of the internal variances is then selected. Note that the Ward's method considers cluster analysis as an analysis of the variance problem, instead of using distance metrics. Therefore, a sample that was initially classified in a group might be removed from that group in a next step. Ward's method tends to produce compact clusters. In contrast, the single linkage tends to produce elongated groups, which sometimes are hard to interpret. The complete linkage tends to produce large numbers of groups.

2.12. Clustering visualization

The output of the clustering method is a figure which resembles a tree, which is called *dendrogram*, and it displays the distances among the individuals and the groups being formed. By careful selection of appropriate similarity measures (Table 1), and clustering algorithms (Table 2) the analyst may thus be able to highlight different aspects of the data set. Using the characteristics of different similarity measures and clustering algorithms (Tables 1 and 2), the significance of the groups has to be interpreted by the analyst and, so, explain the rationale of each group (i.e., what differentiates a group from the others). An advantage of the hierarchical clustering compared to the direct visualization methods is that a high dimensionality (a large number of genes and samples) of the data set is reduced to a convenient two-dimensional representation of subject similarities.

Hierarchical clustering can be performed either for the genes (comparing samples' expression profiles) or for the samples (comparing genes' expression profiles). The two classifications can be combined to produce a *heatmap* of the data set. The heatmap is a colour-coded two-dimensional mosaic that describes the whole expression matrix (samples vs. gene targets), each tile coloured with a different intensity according to the pre-processed data. In addition, the data set is reordered in each dimension of the mosaic according to the dendrograms calculated for samples and genes, respectively. The heatmap literally adds another dimension of information presented by the dendrogram, which may facilitate its interpretation.

2.13. Principal component analysis and its visualization

Principal components analysis, PCA, is a powerful approach to circumvent the dimensionality problem of scatter plots by projecting the high-dimensional data set onto two or three dimensions for easy visualization. Here, the original coordinate system of the data set (i.e., the measured expression profiles) is projected onto a new space with a lower number of new variables, so called principal components (PC's) or factors. Each PC is a linear combination of the subjects. Thus, when illustrating samples, each coordinate axis is a linear combination of genes' expression levels from the original measurements, and when illustrating genes each coordinate axis is a linear combination of sample expression levels from the original measurements.

By mathematical definition, the PC's are extracted in successive order of importance. This means that the first PC explains most of the information (variance) present in the data, the second less, and so forth. Therefore, we can use the first two or three PC coordinates (termed *scores*) not only to obtain a projection of the whole data set onto a conveniently small dimension, but also to obtain the projection that accounts for the most relevant variability in the data set. Variance from experimental design conditions is expected to be systematic, while confounding variance is expected to be random. Since the last PC's explain a very low amount of information, they can be considered to include noise or random information and can therefore be ignored. In this way, PCA can be a very efficient tool to separate systematic effects from noise.

2.14. Self-organizing maps

The *self organizing map, SOM,* or Kohonen artificial neural network, approach is an advanced iterative method based on comparisons, rather than an algorithm as used in hierarchical clustering and PCA. It is based on the use of artificial neural networks and their capability to 'learn', i.e., to change their internal values (called *weights*) to be as similar as possible to the measured samples. It can be used to visualize multivariate data sets in two dimensions (the

Table 1
Similarity measures.

Similarity measure	Effect
Euclidian distance	Favors direct one-to-one similarity. Results depend on scaling of expression profiles.
Pearson correlation	Emphasize variational relationships between expression vectors. Detects both positive and negative correlations.

Table 2
Clustering methods.

Clustering method	Effect
Single linkage	It has a chaining effect, useful for extended clusters with irregular shapes.
Complete linkage	Produces small compact clusters.
Ward's method	Produces very small and compact clusters.

graph obtained is called a *Kohonen map*). For the SOM approach, the analyst defines the size of a matrix of nodes, typically in two dimensions. The SOM algorithm then starts by associating the weights of each node with a random expression profile of the same dimensions (i.e., number of genes, if the object of study is gene expression profiles or number of samples, if the object of study is samples expression profiles) as the data set to be analyzed. Next, each expression profile in the real data set is sequentially assigned to the node that most resembles its own profile. The assigned expression profile is allowed to influence the expression profile in the node and its neighbours by a weighted factor that is large for the central node and progressively smaller for more distant neighbours. The procedure of assigning expression profiles and allowing them to influence the values of the weights of the related neuron is repeated iteratively until convergence is achieved. Finally, each node receives a code, according to the sample(s) that was(were) assigned there and this is plotted in a 2D plot (the map), where the distribution of the samples can be visualized and, hopefully, the groups interpreted.

An advantageous property of the SOMs relative to, for example, PCA, is that the distances between the expression profiles of neighbouring nodes in the map are non-linear. This property may allow detection of otherwise obscured similarity patterns. Another advantage of the SOM is that the size of the node map may be adjusted to seamlessly transition from clustering (no prior assumptions of number of clustering classes) by using a large node map to partitioning (pre-defining a number of clustering classes) by using a small node map. For example by using a node map of dimensions 31×31 on a data set with 31 expression profiles, plenty of space is available for the process to distribute the expression profiles and the resulting clustering structure is thus free from any restraints. In this situation individual relationships between expression profiles can be studied under favourable conditions, but boundaries between groups may be difficult to discern. On the other hand, if a node map of dimensions 10×1 is used on the same data set the process is restrained to a maximum of 10 partitions. In this case, individual relationships between expression profiles are less apparent, but boundaries between groups are highlighted.

2.15. Analysis of visualization results

As when applying any chemometrics or bioinformatics technique, validation is a key stone. One has to be aware that by their own nature pattern recognition techniques (like clustering) yield results that, sometimes, may be surprising for the scientist. Even though hierarchical dendrograms and heatmaps provide powerful visualizations of data sets, they hold drawbacks in their own nature. This problem is well-known in clustering and can only be detected by close examination and interpretation of the final dendrogram and, probably, by repeating the clustering using different distances and clustering algorithms.

Many studies have been devoted to evaluating clustering quality. Some common terminology includes compactness, connectedness and spatial separation [17].

Compactness is a consequence of various clustering algorithms tendencies to keep intra-cluster variation small. Algorithms and parameters that push in this direction have a tendency to give results that are appropriate for spherical clusters, but may fail if the clusters have more complex shapes. Clustering algorithms that implement parameters to favour *connectedness* tend to give results that are appropriate for arbitrary shaped clusters, but lack robustness when there is little separation between clusters (typically, this happens with the single linkage method). *Spatial separation* is rarely used by itself as an algorithm objective, but is usually combined with the other objectives above.

In practical terms it is highly difficult to evaluate these parameters with an unknown dataset and, therefore, in this paper our ambition was to perform an exploratory study. We are therefore not interested in defining any objective measure of cluster quality, but we will use the terminology described above to subjectively describe some patterns that may help us propose hypotheses for future confirmatory studies. Many times we are mining in our (almost unknown) dataset and a multitude of different methods can be applied to draw conclusions. A careful and sound biological explanation of the results is mandatory before accepting any model. Recall always that, then, your model will be accepted under the explicit condition that the resulting conclusions are nothing more than hypotheses that will need to be analyzed in an independent confirmatory study before they can be considered truly validated.

3. Results

Many scientific questions remain on early development of *Xenopus laevis*. Typical questions may be: What is the function of each or gene? At what phase of development are they active? How are they regulated? How do they influence each other? With the increased availability of sequence data and gene-specific characterizations, large scale investigations into these matters are becoming feasible. However, with the increased amount of data we also face new challenges. Specifically: the challenge of finding new, testable, and biologically relevant hypotheses. Here we used the data set of 31 selected genes, measured at 13 stages throughout *Xenopos laevis* development, to illustrate how hypotheses may be generated from a multidimensional data set in an exploratory study.

Data organization is arguably one of the most time-consuming steps in large scale multivariate data analysis. The separate analysis algorithms can typically be performed quickly at a click of a button once input data has been arranged in appropriate format.

In our case, we have collected expression data of 31 genes in 13 developmental stages. Biological triplicates were collected for each stage. The triplicates were organized in the data set under the samples dimension since each replicate can be said to be of the same gene, but not of the same sample. The full data set thus forms a matrix of dimension 39×31. It is also useful to assign classes, when available, to the data set. GenEx allows for classification columns and rows to classify samples and genes. Analysis methods applied to the data read the data in rows, columns or the whole matrix. Analysis methods that are applied to either rows or columns thus analyze the data from either the gene or the sample perspective. By transposing the data matrix, analyses can be shifted from gene to sample perspective or vice versa. In the case of the Xenopus data we have clear classification of the samples as the mid-blastula transition is a well-known transition state during development. The 13 samples can therefore readily be classified as early, midblastular or late stages. Fig. 1 shows a 3D line-plot of sample expression profiles for each of the Xenopus developmental stages. To reduce cluttering and improve visibility the biological replicates have been averaged for Fig. 1. Off-sets may be present in the data sets for technical reasons, unrelated to the biological questions of interest, in the expression profiles of each sample. To further improve visibility and remove off-sets we therefore mean centered the data for Fig. 1. The colours of the 3D lines correspond to the three classes of developmental stages: early (red), midblastular (blue) and late (green) stages. Features that are characteristic for each group are readily visible and the distinction between the groups is clear although best appreciation for this is achieved on a computer screen where the 3D line plot can be rotated for improved 3D visualization.

Heatmaps are an alternative representation of expression profiles. The mosaic representation of the heatmap in Fig. 2 illustrates expression amplitude by colour intensities. It is often useful to recognize patterns in the heatmap mosaic and compare them to the expression profiles in a 3D line-plot such as shown in Fig. 1. The birds-eye-view of the heatmap allows this representation to be less sensitive to information overload and cluttering. In Fig. 2 the biological replicates have been retained and mean centered in the samples dimension. Comparisons should therefore primarily be performed in the samples dimension of the heatmap, i.e., treating it as a set of columns. For a 3D line-plot this large number of samples would have obscured any interpretation, but heatmap produces a very comprehensive overview. In addition, the rows and columns of the heatmap are organized through hierarchical clus-

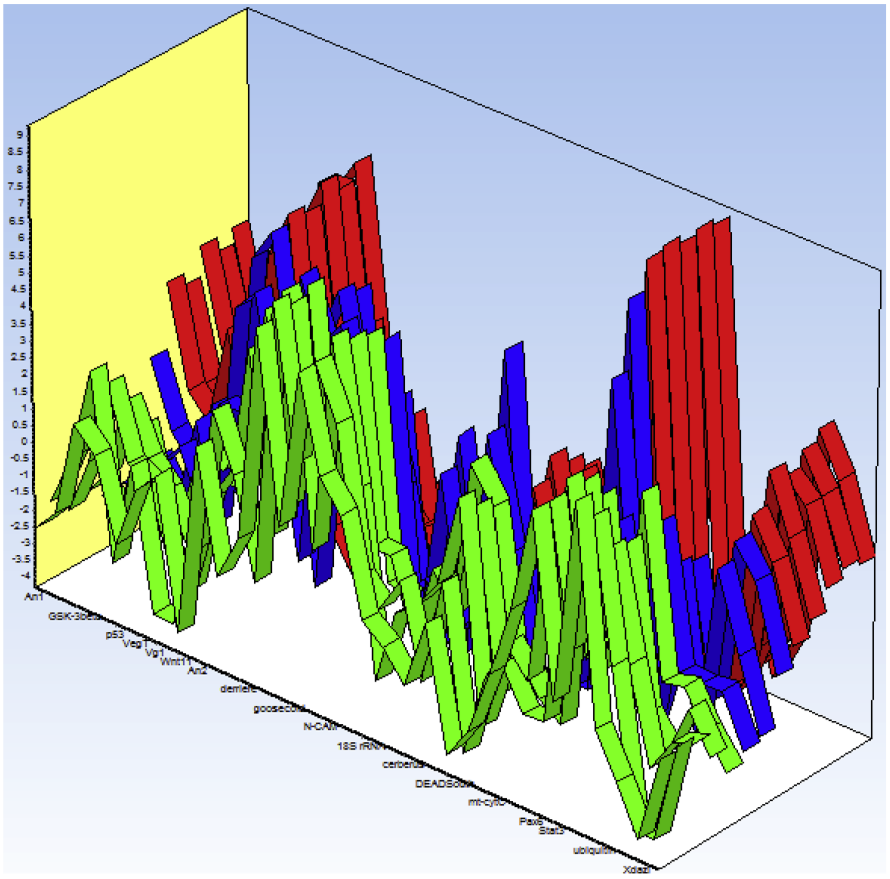


Fig. 1. A 3D line plot of mean centered sample expression profiles in *Xenopus laevis*. Red expression profiles are early, blue are intermediate and green are late developmental stages.

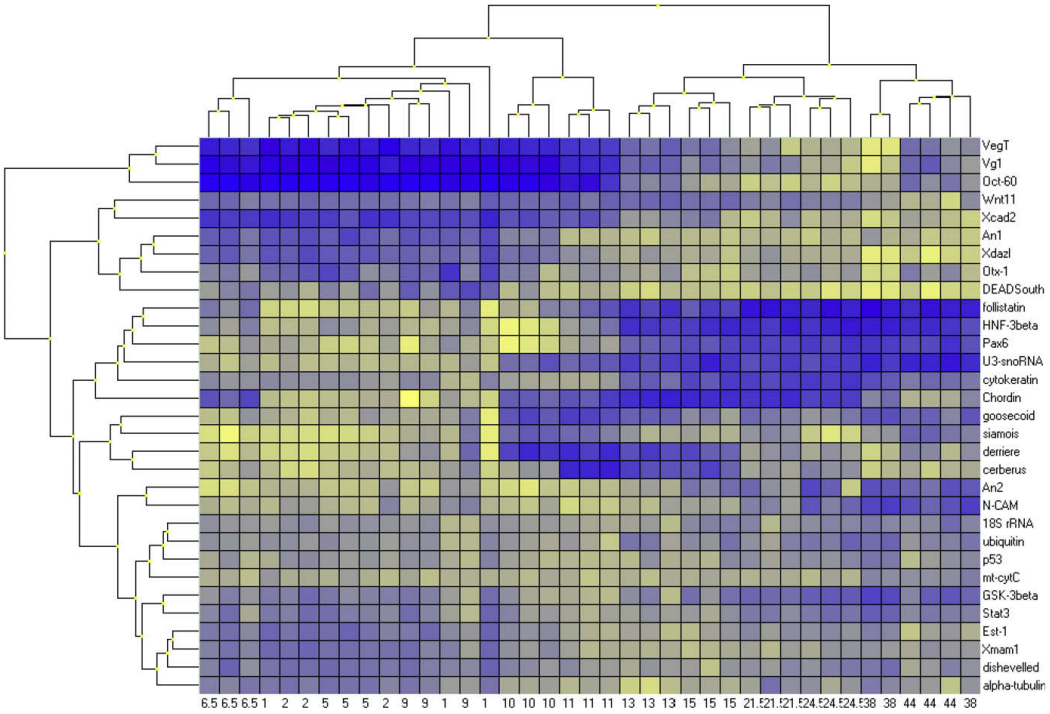


Fig. 2. Data with biological replicates, mean centered in the samples dimension. Light yellow corresponds to low expression, dark blue corresponds to high expression. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

tering. The dendrograms that indicate the relationships between the samples (columns) and genes (rows) thus gives additional information to the comprehensive overview. In our current data set we see that we have three distinct groups corresponding to the early (1, 2, 5, 6.5, and 9), midblastular (10 and 11) and late (13, 15, 21–22, 24–25, 38, and 44) developmental stages. We also see good reproducibility among the biological replicates, although some samples (notably sample 1) have variabilities large enough to make them overlap with several other neighboring developmental stages.

A heatmap is also very useful to characterize relationships between genes (Fig. 3). To emphasize the developmental characteristics of the gene expression profiles we average the biological replicates. This will give us a better estimate of the mean expression at each developmental stage and also reduce the information content in the heatmap for a more comprehensive overview. By autoscaling in the gene dimension we emphasize comparisons between gene expression profiles (rows in the heatmap), normalized so that the variability within each expression profile will be equal between different expression profiles. By inspecting the gene expression dendrogram on the left side of the heatmap in Fig. 3 we observe that, in the most simple, non-trivial clustering tier, they are organized into three distinct groups. The topmost genes (goosecoid, siamois, derriere, cerberus and chordin) have their highest expression levels primarily at the MBT (midblastular genes). The genes in the middle group (HNF-3beta to ubiquitin) have their highest expression levels primarily at the later developmental stages (zygotic genes). These group assignments are summarized in Table 3 and compared to assignments from Kubista et al. [20]. The bottom group of genes (An1–Wnt11) has high expression at the early developmental stages (maternal genes). However, there are also several interesting substructures in this data set. For instance, the maternal genes disheveled, Xmam1, Est-1 and alpha-tubulin seem to have an increased expression at the later stages of development compared to other maternal genes. Likewise, the midblastular genes goosecoid and siamois seem to have an increase expression at the developmental stages 38 and 44 compared to other MBT genes. The zygotic genes seem to divide into two characteristic groups one (genes HNF-3beta to U3-snoRNA) of which seems to have rather even expression throughout the later developmental stages (stages 13–44) and the other (genes GSK-3beta to ubiquitin) which seem to have a distinct increase in expression in the last few stages of development (stages 24, 25–44). These observations may inspire several interesting hypothesis to be validated in future confirmatory studies.

Now that we have used the heatmap and the associated hierarchical clustering to identify groups of genes, we can improve visibility in the 3D line-plot of the gene expression profiles by colouring each expression profile according to which group it has been assigned. Fig. 4 shows the autoscaled gene expression profiles of the Xenopus data set. Without the colour assignments it would be hard to discern patterns in the data set, but with the colour assignments the 3D line-plot literally gives another dimension of detailed information on the expression profiles. Now, we have a very good idea of what-is-going-on in our data set and, therefore, are ready to interpret the multivariate models that different multivariate techniques can offer us.

As explained above, hierarchical clustering is a way to visualize similarities in a multidimensional data set. An alternative is to produce a projection of the multidimensional data onto a two-dimen-

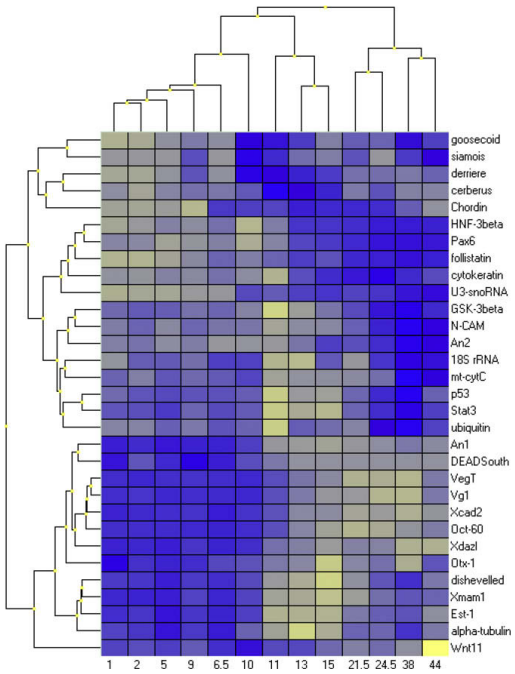


Fig. 3. Data autoscaled in the gene dimension. Samples 1–9 are from early stages in development, 10–13 are from the Mid-blastula transition, and 14–38 are from late stages in development. Light yellow corresponds to low expression, dark blue corresponds to high expression. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

sional scatterplot through PCA. The first two principal components of the mean centered samples expression profiles in the Xenopus data set is shown in Fig. 5. These two principal components account for 75.8% of the variation in the data, even though it is only a two-dimensional representation of a 31-dimensional data set (each sample vector contain expression measurements from 31 genes). As we saw in the heatmap dendrogram of the samples expression profiles (Fig. 2), the early, MBT, and late stages tend to be similar within each group than to members of other groups. An advantage of the PCA representation over the dendrogram is that the order of similarities is easier to discern. In Fig. 5 we see that there is a general trend from the early stages through the MBT stages to the late stages. However, there are some subpatterns that may be observed here too. For example, for the genes in this study it seems that stage 6.5 is more similar to later stages than stage 9. This we can interpret as a temporary regression in the development in stage 9 before development proceeds in stage 10 and beyond. Furthermore, stage 11 seems to constitute a significant jump in development compared to the other developmental stages in this study, as can be seen in the relatively large separation between this cluster of biological replicates and others. This fits with our expectations that the MBT is a dramatic transition during development. Finally, it seems that the Xenopus development goes through a local extreme around stage 38 before finally settling in (stage 44) to a state that is more similar to earlier developmental stages (stages 21–22 and 24–25).

The PCA can also be performed for the gene expression profiles. Fig. 6 shows that maternal, MBT and zygotic gene groups form distinct groups in the first two principal components of autoscaled gene expression profiles from the Xenopus data set. In this case, the first two PCs account for 80.5% of the total variation in the data. It is more than what we observed for the PCA in Fig. 5, although in this case, since we have averaged the biological replicates, the PCA is based on a 13-dimensional data set (of sample expression measurements). In either case it is a great focus on the systematic variation in the data set on only two easily visualized dimensions. In Fig. 6, the separation is clear between each group. Neither of the groups is perfectly connected. The maternal cluster may be interpreted to have three subgroups, and the MBT and zygotic groups may be interpreted to have two subgroups each. There seems to be a trend of sequential similarities of gene expression profiles from maternal to zygotic genes with the MBT genes being a separate group from this sequence of gene expression profiles. From a biological point-of-view it may make sense that different genes are activated and deactivated in sequence as they are required for each developmental stage. To further evaluate this as a tentative hypothesis we may take advantage of the inherent strengths and flexibilities of the SOM. Despite a two dimensional PC scores plot has been presented here, three dimensional plots can often be of much use as more information will be taken into account and, so, interesting features or patterns can appear.

Fig. 7 shows the SOM analogous to the PCA plot in Fig. 6. It is based on autoscaled gene expression profiles of the Xenopus data set. The default setting of the SOM in GenEx is a map of size equal to the number of expression profiles in each dimension. In our case we have 31 gene expression profile and thus a 31×31 matrix of SOM nodes. An obvious characteristic of the SOM is that the data points are distributed more evenly across the plot area. This is due to the fact that the plot area is non-linear and neighboring data points may be as similar or dissimilar as other data points further away in a different direction. This cause a greater separation among data points, and reduces the compactness of the clusters. With a large node matrix as the one in Fig. 7 it may be challenging to discover or confirm clustering tendencies. However, an interesting feature of the SOM is that the size of the node matrix can easily be reduced to funnel the data points into a specific number of available partitions. Keeping the notion of the tentative hypothesis of sequential activation of gene expression from the previous paragraph, we may suggest a specific organization of the SOM node matrix. In Fig. 8 we have three representations of 10×1 SOM matrices. By selecting the size of one side of the node matrix to

Table 3
Xenopus gene group assignments. Gene names in bold indicate genes assigned to different groups in Kubista et al. [20]. Chordin, derriere, goosecoid, and siamois were assigned to the late group in [20]. GSK-3beta and p53 were assigned to the early group in [20].

Developmental assignment from this study	Genes also analyzed in Kubista et al. 2006	Genes not analyzed in Kubista et al. 2006
Early	VegT, Vg1, dishevelled	Oct-60, Xcad2, Xdazl, An1, DEADSouth, Otx-1, Xmam1, Est-1, alpha-tubulin, Wnt11
MBT	Cerberus, **chordin**, **derriere**, **goosecoid**, **siamois**	
Late	Follistatin, HNF-3beta, N-CAM, **GSK-3beta**, **p53**	An2, Pax6, U3-snoRNA, 18S rRNA, mt-cytC, Stat3, ubiquitin, cytokeratin

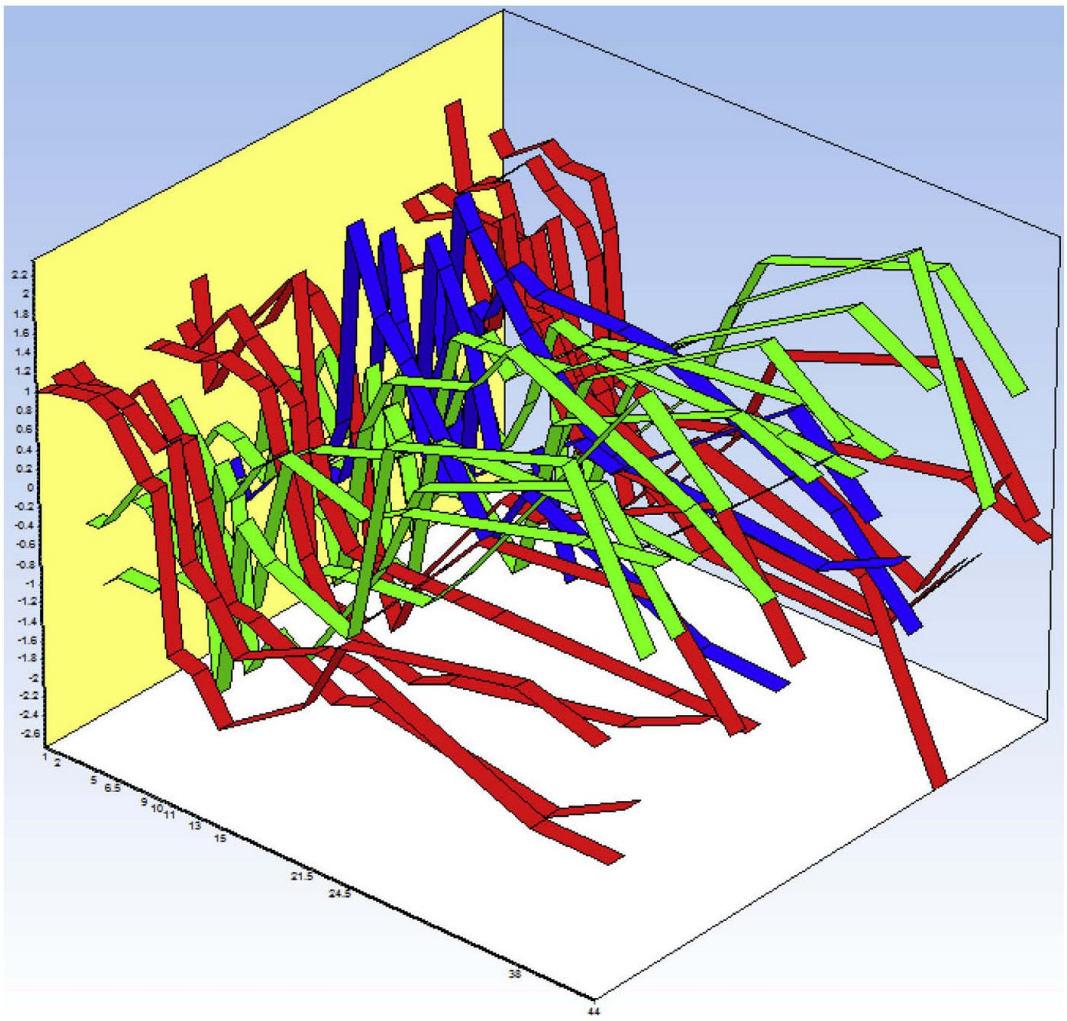


Fig. 4. A 3D line plot of autoscaled gene expression profiles in *Xenopus laevis*. Red expression profiles show maternal genes, green show zygotic genes and blue show midblastular or unknown genes.

one we naturally obtain a node structure conducive to investigate sequential patterns. Because there is an element of random starting points for each SOM generation, the SOMs are not identical. However, by comparing repeated SOM runs and identifying consistent patterns, the SOMs can be used to propose testable hypotheses. For example, here we find that expression profile of VegT seem to reach its higher levels before the expression profile of Otx-1 does. This hypothesis may be biologically sensible since VegT has previously been associated with mesoderm formation, and Otx-1 has been associated with brain development [2].

4. Concluding Remarks

From the observations we have made in the current study, we may propose several hypotheses for further study and potential validation in future confirmatory studies. Some examples may be:

- The average gene expression during the late stages (21–22, 24–25, 38 and 44) relative to the average gene expression during the early stages (1, 2, 5, 6.5, and 9) is higher for Otx-1 than for VegT.
- The expression of mt-cytC is more than 4 times larger during stage 38 than during stage 24-25.
- The average expression of cerberus is larger during the MBT stages (10 and 11) than during an average of other developmental stages (1, 2, 5, 6.5, 9, 21–22, 24–25, 38 and 44).

These examples illustrate how exploratory studies, and clustering and visualization techniques are powerful tools to generate hypotheses. It may be tempting to draw scientific conclusions already at this point of the analysis, in particular if the tendencies in the data seem to be very strong. However, without hypotheses defined before the study, we need to be mindful that we may, consciously or unconsciously, be testing a potentially unlimited

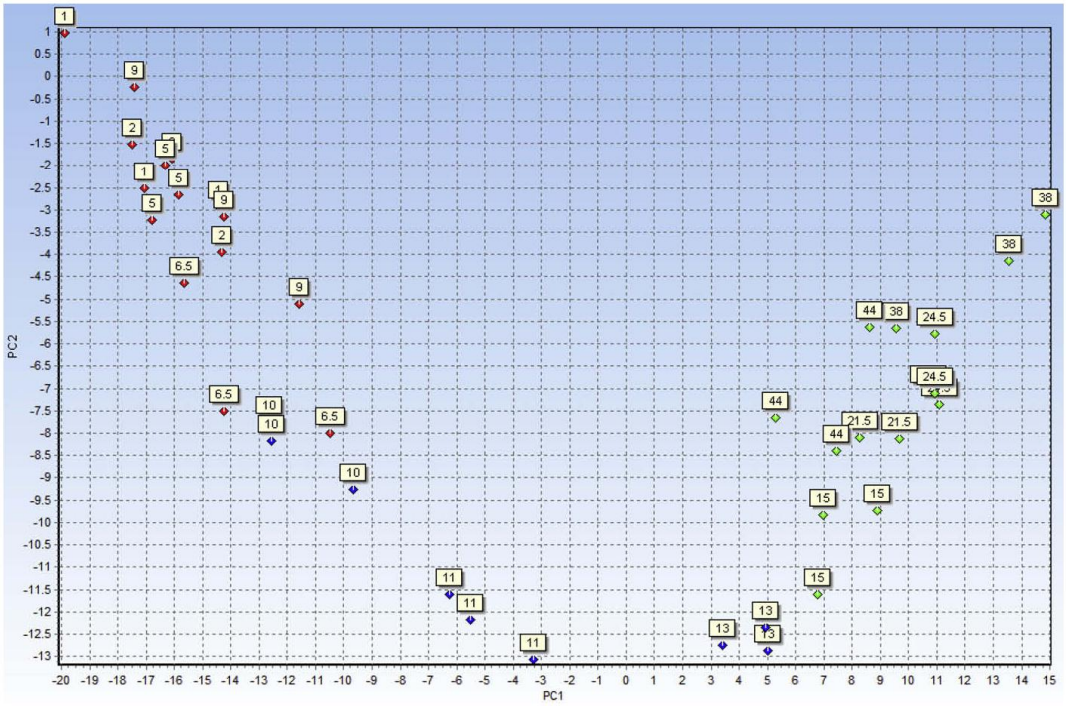


Fig. 5. PCA on Xenopus samples (developmental stages), mean centered, each developmental stage represented by triplicate measurements. Early stages in red, MBT transition stages in blue and late stages in green. For brevity, stages 21–22 and 24–25 are indicated by labels 21.5 and 24.5, respectively.

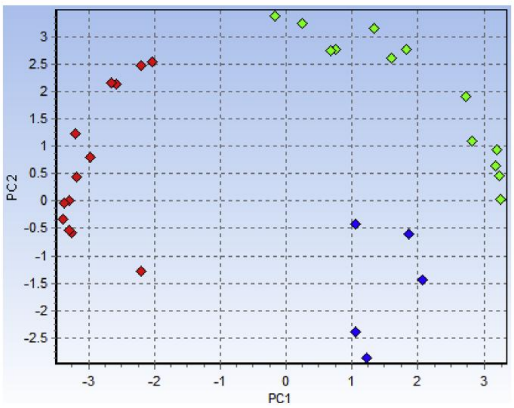


Fig. 6. PCA on Xenopus genes. Autoscaled in genes dimension, maternal genes in red, zygotic genes in green and other genes in blue.

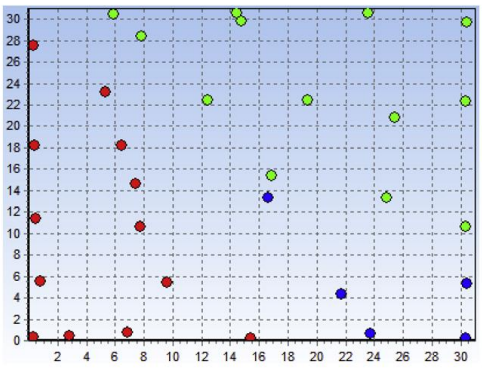


Fig. 7. SOM on Xenopus gene expression data with dimensions 31 × 31. Maternal genes in red, zygotic genes in green and other genes in blue.

number of hypotheses simultaneously. Testing many hypotheses simultaneously require corrections of the underlying statistics parameters in order to maintain overall significance level of the study [21]. Without these corrections there is a risk that observations of random events may be interpreted as biologically relevant processes. The generation of hypotheses is nevertheless an important part of the scientific process. We conclude that we have many useful tools for hypothesis generation and that statistical validation of new-found hypotheses is an integral part of the scientific process, even though it may be left to future studies.

Despite the power of the multivariate methods, we would like to stress that scientist must be aware that their sound knowledge and scientific skills are critical to validate/interpret/ascertain what they have found. The many clustering methods, distance metrics, and

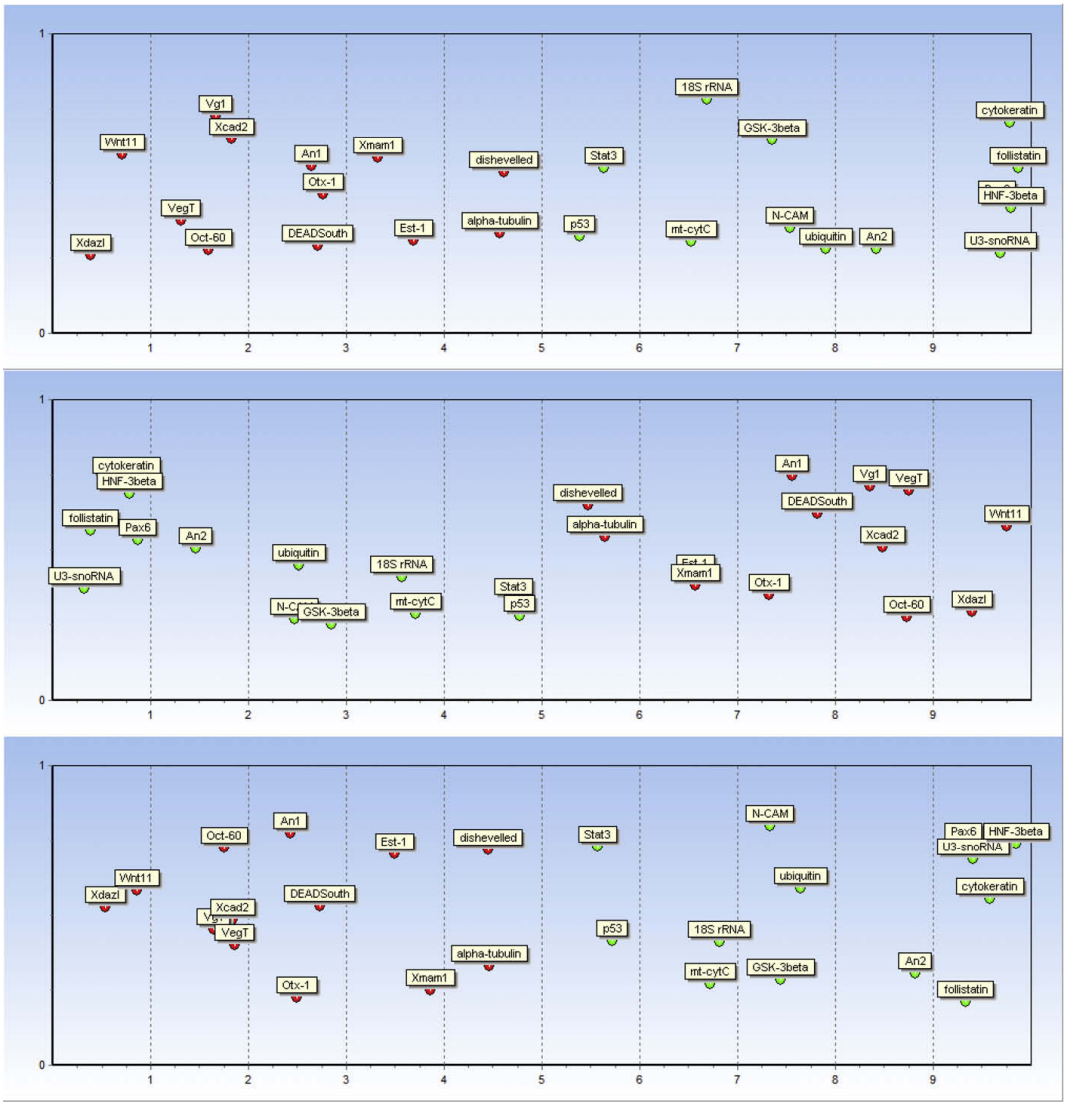


Fig. 8. Three SOMs on Xenopus genes with dimensions 10 × 1 show transition from maternal genes (red) to zygotic genes (green). Each SOM pattern is generated from a process that is influenced by random starting conditions and therefore the SOM patterns do not need to be identical. In the examples here we nevertheless see that there are consistent patterns in the data, such that the ordering of the gene expression profiles is preserved.

their combinations, make it too easy to bias the results to a subjective hypothesis on the data structure/patterns. Key point here is to be open to discover something unexpected but interesting.

Acknowledgments

Part of this work was supported by Grant agency of Academy of Science the Czech Republic (GAAV IAA500970904, IAA500520809) and research goal AV0Z50520701 Granted by Ministry of Youth, Education and Sports of the Czech Republic, together with support from Grant agency of the Czech Republic (GACR 301/09/1752).

References

[1] R. Sindelka, J. Jonák, R. Hands, S.A. Bustin, M. Kubista, Nucleic Acids Res. 36 (2) (2008) 387–392.
[2] J.B. Bowes, K.A. Snyder, E. Segerdell, R. Gibb, C. Jarabek, E. Noumen, N. Pollet, P.D. Vize, Nucleic Acids Res. 36 (2008) D761–D767.
[3] M.L. King, T.J. Messitt, K.L. Mowry, Biol. Cell 97 (1) (2005) 19–33.
[4] http://www.fluidigm.com

[5] http://www.multid.se
[6] P.D. Nieuwkoop, J. Faber, Normal Table of *Xenopus laevis*, Garland Publishing, Inc., New York & London, 1994.
[7] http://frodo.wi.mit.edu/primer3/
[8] R Sindelka, et al., to be published (2010).
[9] A. Bergkvist, A. Forootan, N. Zoric, L. Strömbom, R. Sjöback, M. Kubista, Genet. Eng. Biotechnol. News 28 (13) (2008) 26–28.
[10] R. Sindelka, Z. Ferjentsik, J. Jonák, Dev. Dyn. 235 (3) (2006) 754–758.
[11] M. Kubista, R. Sindelka, A. Tichopad, A. Bergkvist, D. Lindh, A. Forootan, G.I.T. Lab. J. 9–10 (2007) 33–35.
[12] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, fourth ed., Pearson, Harlow, 2000.
[13] L.L. Havilcek, R.D. Crain, Practical Statistics for the Physical Sciences, Edit ACS, Washington, US, 1988.
[14] V. Barnett, T. Lewis, Outliers in Statistical Data, third ed., Willey, Chichester, UK, 1994.
[15] M.J. van der Laan, K.S. Pollard, J.F. Bryan, A New Partitioning around Medoids Algorithm, The Berkeley Electronic Press, University of California, Berkeley, 2002. paper 105.
[16] M.J. van der Laan, J.F. Bryan, Biostatistics 2 (2001) 1–17.
[17] J. Handl, J. Knowles, D.B. Kell, Bioinformatics 21 (2005) 3201–3212.
[18] G.H. Lance, W.T. Williams, Comput. J. 9 (4) (1966) 373–380.
[19] J.H. Ward, J. Am. Stat. Assoc. 58 (1963) 236–244.
[20] M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, R. Sindelka, R. Sjöback, B. Sjögreen, L. Strömbom, A. Ståhlberg, N. Zoric, Mol. Aspects Med. 27 (2006) 95–125.
[21] H. Motulsky, Intuituve Biostatistics, Oxford University Press, New York, 1995. ISBN: 0-19-50-8607-4.

Briefings in Functional Genomics Advance Access published February 7, 2013
BRIEFINGS IN FUNCTIONAL GENOMICS. page 1 of 9 doi:10.1093/bfgp/elt001

The added value of single-cell gene expression profiling

Anders Ståhlberg, Vendula Rusnakova and Mikael Kubista

Abstract

Cells are the basic unit of life and they have remarkable abilities to respond individually as well as in concert to internal and external stimuli in a specific manner. Studying complex tissues and whole organs requires understanding of cell heterogeneity and responses to stimuli at the single-cell level. In this review, we discuss the potential of single-cell gene expression profiling, focusing on data analysis and biological interpretation. We exemplify several aspects of the added value of single-cell analysis by comparing the same experimental data at both single-cell and cell population level. Data normalization and handling of missing data are two important steps in data analysis that are performed differently at single-cell level compared with cell population level. Furthermore, we discuss how single-cell gene expression data can be viewed and how subpopulations of cells can be identified and characterized.

Keywords: *single-cell biology; single-cell analysis; RT–qPCR; single-cell transcriptomics; gene expression analysis*

INTRODUCTION

The great value in studying cell diversity and biological processes at single-cell level has been recognized for long time, but the lack of sensitive and specific techniques to measure few molecules has been a major obstacle. During the past decade, refinement of established methods and also development of new technologies have opened up new possibilities to analyze single cells and even single molecules in highly reproducible manners. Today, most analytes including DNA, RNA and protein, as well as other biological parameters, such as cell morphology and action potential, can be measured with high precision at the single-cell level. Microscopy, fluorescence-activated cell sorting, real-time quantitative polymerase chain reaction (qPCR), mass spectroscopy, microarrays and next-generation sequencing are some commonly applied techniques to study individual cells. In general, one class of analytes is analyzed at a time with one of the methods. Recently, we developed a strategy to analyze DNAs, RNAs and proteins in the same single-cell using qPCR, opening up for multi-analyte correlation analysis in single cells [1]. For overview and comparison of different techniques, see recent reviews [2–5].

Early single-cell work focused on understanding the seemingly stochastic variations in mRNA and protein levels observed among individual cells. The variability was explained by transcriptional and translational bursts [6–8] and mathematical models, including the master equation, were developed to describe underlying biological processes [4,6]. However, other types of kinetics have been reported [9]. Mathematical modeling of transcriptional and translational kinetic at single-cell level is discussed elsewhere [4,6,9,10].

Currently, characterization of cell heterogeneity in mixed as well as in seemingly homogenous cell

Corresponding author. Anders Ståhlberg, Department of Pathology, Sahlgrenska Cancer Center, Sahlgrenska Academy at University of Gothenburg, Box 425, 40530 Gothenburg, Sweden. Tel: +46 31 7866735; E-mail: anders.stahlberg@gu.se

Anders Ståhlberg, PhD, is a principle investigator at the Sahlgrenska Cancer Center, University of Gothenburg, Gothenburg, Sweden. His primary research interest is to understand molecular mechanisms in tumor development and stem cell differentiation.

Vendula Rusnakova, MSc, is a PhD student at Charles University in Prague, Prague, Czech Republic. She is developing new approaches of single-cell gene expression profiling, including astrocyte characterization.

Mikael Kubista, PhD, is a head of the department of gene expression at the Institute of Biotechnology of the Czech Academy of Sciences, Prague, Czech Republic and the founder of TATAA Biocenter and MultiD Analyses (both Gothenburg, Sweden).

© The Author 2013. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

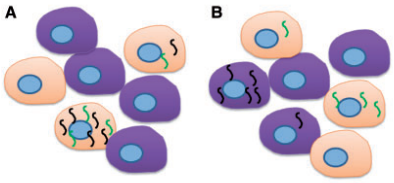


Figure 1: Biomarker analysis at single-cell level. Single-cell analysis can be used to define genes with correlated expressions (**A**) or uncorrelated expressions (**B**) in specific cell types.

populations is the focus area of single-cell studies. Established cell types as well as previously unknown subpopulations have been identified and characterized in detail at single-cell level using qPCR [11–19]. Analysis at single-cell level eliminates many of the factors that confound gene expression profiling at cell population level. Correlation of transcript levels within individual cells has proven to be a most powerful identifier of cell types and distinct subpopulations of cells (Figure 1A and B) [11–13]. Another application of single-cell analysis is to define and characterize rare cell types such as stem cells and circulating tumor cells [20–22]. One challenge when studying rare cells is that a large number of cells must be screened and tested to find those cells that are relevant. Flow cytometry can handle large numbers of cells, but most other single-cell techniques do not offer the required throughput. An alternative approach is to enrich for the targeted cells before analysis, taking advantages of distinctive physical or biological properties of the rare cells [21,22]. Further development of technologies such as high-throughput droplet PCR opens for new possibilities for high-throughput screening of rare cells [23].

The workflow of single-cell analysis requires delicate sample preparation, development of sensitive assays and often sophisticated data analysis. With time, as experience is accumulating, improved and simplified protocols for single-cell analysis will become available. In this review, we discuss important aspects of data pre-processing and how single-cell gene expression data can add value to our understanding of cell diversity and expression pathways controlling biological processes. To emphasize key aspects of single-cell gene expression profiling, we here analyze an experimental data set at both single-cell and cell population level by comparing the results.

SINGLE-CELL VERSUS CELL POPULATION ANALYSIS

To illustrate different aspects of single-cell gene expression profiling, we re-analyzed a previously published data set at both single-cell and cell population level [24]. The original data set consisted of the expression profiles for 41 genes measured in 303 individual astrocytes at four time points: 0, 3, 7 and 14 days after induced brain ischemia in mice, with day 0 being control [24]. Here, we focus our analysis on a smaller number of genes selected based on their expression features rather than biological relevance. The single-cell gene expression features we report for the selected genes are not new; they have been observed before [1,11–19].

Cell population data analysis follows the Minimum Information for Publication of Quantitative Real-Time PCR Experiments guidelines [25], while single-cell data analysis is performed as described [5,24]. The workflows for the two data pre-processing schemes are shown in Figure 2 [13,24]. Two steps differ fundamentally: missing data and normalization. It should be noted that the order of the data pre-processing steps in Figure 2 can be partly swapped, since some data processing is independent of the order. Furthermore, some steps will be canceled if the required information for that particular analysis is missing.

DATA PRE-PROCESSING: MISSING DATA

Missing data are common in scientific measurements and a classical problem in statistical analysis. The underlying cause of the missing data determines how they should be handled for downstream analysis. There are two fundamentally different reasons to miss data in qPCR: (i) the qPCR failed due to a technical error and (ii) the biological sample contained fewer target molecules than the workflow with all dilution steps, losses and imperfections could detect (limit of detection). Missing data due to technical failure should be replaced using information from (preferably) technical replicates or (less preferably) from similar samples using imputation [24]. However, missing values due to too few target molecules should be replaced by zeros in linear scale and by a cutoff value in logarithmic scale. Technical failures are handled early in the data pre-processing scheme, while transcript levels below limit of detection are

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

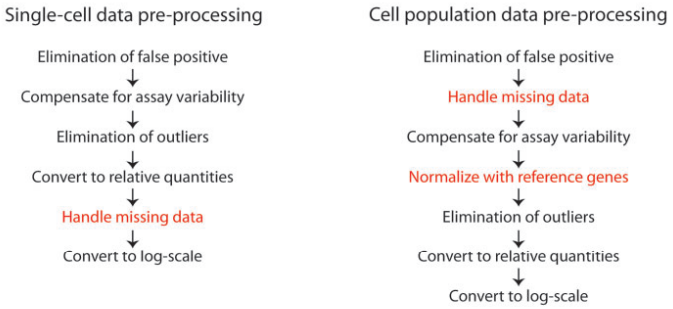


Figure 2: Simplified workflows for data pre-processing using qPCR. Handling of missing data and normalization are performed differently in single-cell and cell population analysis.

processed later in the workflow (Figure 2). How can one tell whether missing data are due to technical failures or expression levels below the limit of detection?

The transcriptome of typical mammalian cells consists of about 300 000 mRNA molecules represented by some 10 000 expressed genes [26]. Some genes are highly expressed (>100 transcripts per cell), but most genes have a mean expression of less than one molecule per cell. In the bacteria *Mycoplasma pneumoniae*, the mean mRNA count was determined to be as low as 0.04 molecules per cell [27]. The observed variability in expression level also scales to the average number of transcripts for respective gene [24]. If samples are analyzed in replicates, technical failures are usually readily recognized by inspecting the replicate measurements. However, single-cell profiling is usually performed without technical replicates to maximize the number of cells analyzed. In single-cell qPCR, extensively validated protocols and highly optimized assays are used, which reduces the risk for technical failures. A pragmatic approach to handle missing data in single-cell gene expression profiling is to assume that all missing data are due to too few molecules. In conclusion, high frequency of missing data is expected for most genes at single-cell level due to low mean expression levels and fluctuations over time (Figure 3).

Figure 4 shows the binary (transcript detected/not detected) expression profiles for all the genes and time points at single-cell level for the experimental astrocyte data set. The 10 selected genes had overall intermediate to high expression levels, still large amounts of missing data are observed.

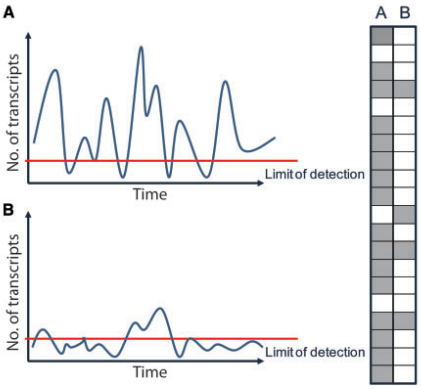


Figure 3: Missing data. Single-cell data sets have many missing data due to too few molecules in the measured sample. Total number of transcripts over time is shown for high (**A**) and low (**B**) abundant genes, respectively (left diagrams). The binary panel (right plot) shows the frequency of cells with detectable number (above limit of detection) of transcripts (gray). The limit of detection for high-performance assays is typically between one and five molecules.

DATA PRE-PROCESSING: NORMALIZATION

Comparison of classical samples containing many cells requires normalization to account for differences in the amount of material in the analyzed samples [25]. Several normalization strategies for qPCR data have been suggested. Most popular is to normalize with the expression of one or a few (two to five)

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

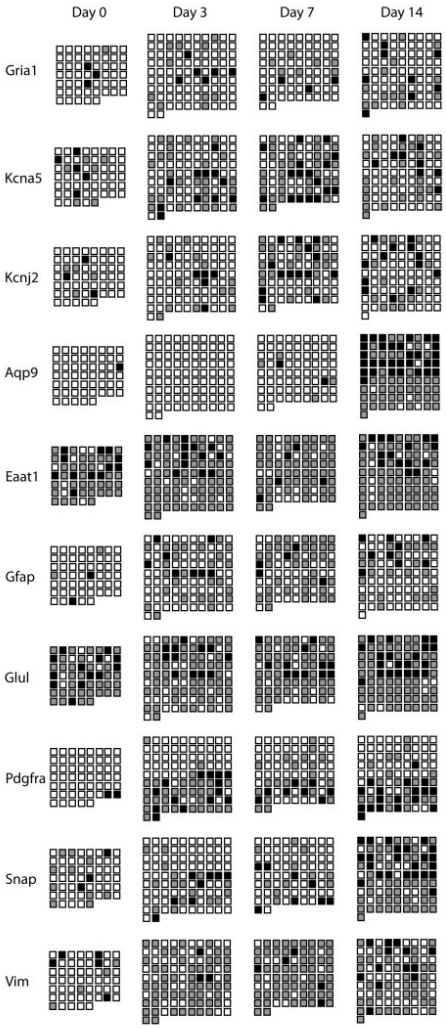


Figure 4: Binary single-cell analysis. Single cells with (gray and black) and without (white) measurable transcript levels are indicated. The cells with highest expression levels that represent 70% of all transcripts of any given gene at each time point are indicated in black. $n_{Day0} = 53$, $n_{Day3} = 92$, $n_{Day7} = 74$ and $n_{DayH} = 82$.

reference genes that are constantly expressed in the compared samples.

To analyze the single-cell data set at cell population level, we summed the numbers of each transcript in every cell and each time point. Then, the

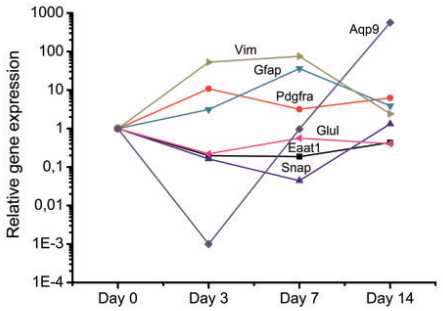


Figure 5: Cell population analysis. Relative gene expression profiles at four time points (days 0, 3, 7 and 14) are shown, with day 0 being the control (i.e. before brain ischemia). The relative gene expression level at day 0 was arbitrarily set to 1 for each gene. *Aqp9* expression was zero at day 3, but for presentation purposes it is set to 0.001.

genes with most constant expression were identified for normalization of the data using analysis of variance as described in NormFinder [28]. The three genes *Gria1*, *Kcna5* and *Kcnj2* were found to be the best choice for normalization. Relative gene expression was calculated by normalizing to the geometric average expression of the three validated reference genes [29]. Figure 5 shows the normalized gene expression over time for the selected target genes *Aqp9*, *Eaat1*, *Gfap*, *Glul*, *Pdgfra*, *Snap* and *Vim*. The gene expression was arbitrarily set to 1 at day 0 for all genes (Figure 5). This choice of scaling has no effect on the downstream statistics, but provides an intuitive baseline for sample comparison [30]. This cell population analysis reveals that expression of *Gfap*, *Pdgfra* and *Vim* is transiently upregulated, while *Eaat1*, *Glul* and *Snap* is transiently downregulated, and expression of *Aqp9* is first downregulated and then strongly upregulated (Figure 5). Clearly, the cell population analysis has caught the essence of the overall biological response induced by the brain injury.

As mentioned, statistical analysis identified *Gria1*, *Kcna5* and *Kcnj2* as optimal reference genes at cell population level. From single-cell analysis, we note that only very few cells host the majority of these transcripts at each time point (Figures 4 and 6). Furthermore, the *Gria1*, *Kcna5* and *Kcnj2* transcripts are found in different cells. Hence, if we would normalize expression of the target genes to that of the reference genes

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

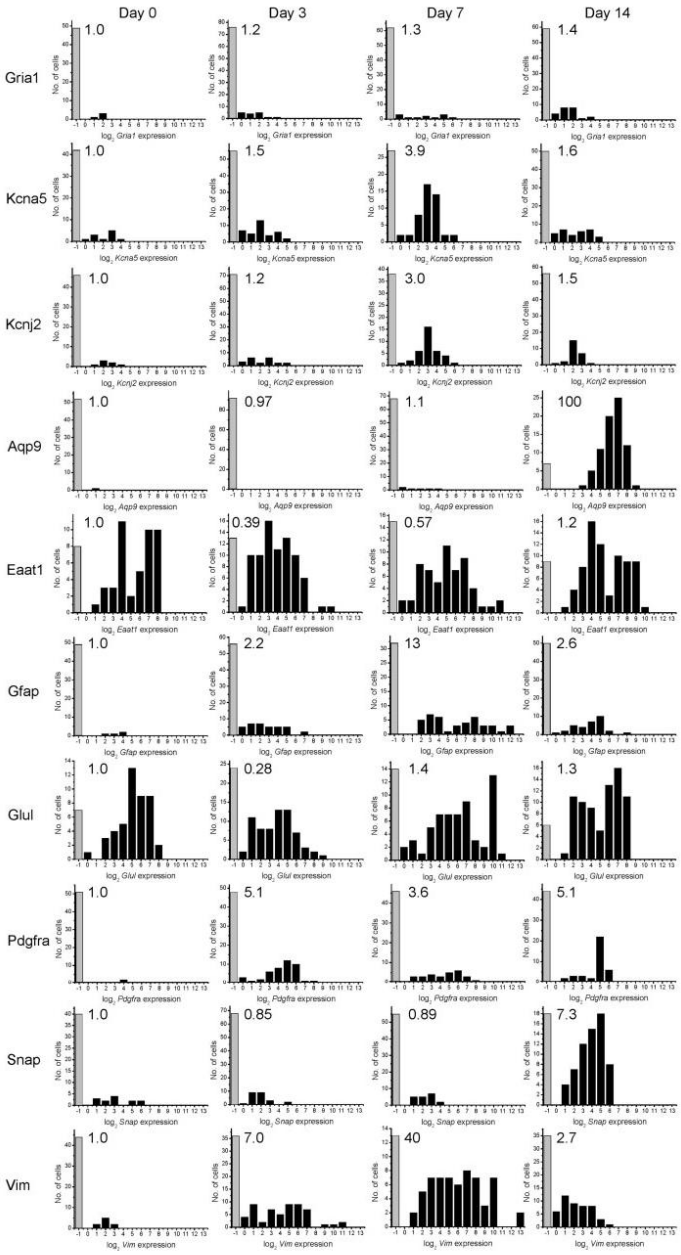


Figure 6: Distribution of transcripts among individual cells. Histograms showing relative gene expression levels (log_2-scale) for all genes at days 0, 3, 7 and 14. Day 0 is control (i.e. before brain ischemia). Cells without transcripts are indicated in gray. The typical gene expression of each gene and time point was calculated as geometric average by assigning missing data 0.5 transcripts, since calculations of geometric averages cannot include zeros. The relative mean expression was arbitrarily set to a value of 1 at day 0 for each gene.

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

separately in each individual cell, we would introduce very large unphysiological bias. The reason normalization to reference genes is applicable in cell population studies is due to the large number of cells analyzed, which averages out the large natural variation among individual cells. However, the averaging also eliminates much of the differences in expression contributed by subpopulations. There is also risk that the reference genes selected have different expression profiles in existing subpopulations; the Normfinder algorithm does not sense this.

Traditional cell population analysis is performed on thousands of cells, sometimes even many more. The interest to analyze samples containing fewer cells is, however, rapidly increasing. For example, fine needle aspirates collected frequently in minimal invasive diagnostics contain often only a few hundreds of cells for downstream analysis. In our data set, 52–92 cells were analyzed at each time point. One consequence of analyzing samples with so few cells is that estimated expression of low expressed genes is uncertain and results may become biased. For example, cell population analysis indicates that *Aqp9* is substantially downregulated at day 3, followed by dramatic upregulation at day 14 (Figure 5). Inspecting the single-cell data we find that only one cell expressed *Aqp9* at day 0 and none at day 3 (Figures 4 and 6). This vital biological information is lost in the cell population analysis, and the conclusion reached relied on expression in only one cell, corresponding to less than 1% of the cells (1 out of 145 cells expressed *Aqp9* during days 0–3). Single-cell data allow for more detailed and careful interpretation of the *Aqp9* expression data. One way to improve the accuracy of *Aqp9* expression analysis is to increase the amount of material analyzed. However, this is possible only to a certain extent. There is a maximum amount of material that can be reliably reversed transcribed and qPCR amplified in a single reaction. Furthermore, standard RNA extraction protocols purify total RNA or total mRNA and are not suitable to enrich rare transcripts. Consequently, quantification of low abundant transcripts will usually have poor precision when studying samples using common population analysis. In summary, single-cell analysis adds value by identifying the number of cells that actually express any given transcript.

The intuitive way to present single-cell gene expression data is per cell. Importantly, single-cell data should not be normalized using reference genes, since no gene has ever been reported to have constant and stable expression among individual cells. Instead, very high variability in specific mRNA levels among cells is seen due to transcriptional kinetics (Figure 3) [6,7]. One option could be to perform global normalization, i.e. to normalize to the mean expression of all genes. But this is complicated since we are not analyzing all the genes with qPCR and those that are studied are not randomly selected. The studied genes rather reflect specific biological processes that are induced or stimulated to change. Normalization to the mean average of the studied genes is then likely to introduce bias. Furthermore, data comparison between studies that use slightly different markers is also biased, since the 'global' mean is not truly global, but is the mean of those genes that actually are measured. Finally, single-cell profiling studies have many missing data, and there is no good way to handle those in global normalization. In brief, methods to normalize cell population data are not suitable for single-cell data. We have found that the most useful and natural way to present and report single-cell gene expression data is per cell.

TRANSCRIPTS ARE LOGNORMALLY DISTRIBUTED AMONG INDIVIDUAL CELLS

Cells that contributed to 70% of the total number of a particular transcript were in minority at population level in our experimental data set (indicated as black in Figure 4). For example, only 6 out of 92 cells contributed with 70% of all the *Gria1* transcripts at day 3. The reason few cells contribute with majority of the transcripts for any particular gene is the highly skewed distribution of transcripts normally observed among individual cells (Figure 6). Gene expression data from single cells can usually be adequately modeled with a lognormal distribution, although recent advances in theoretically biology suggest that the underlying distribution is more complex [4,6]. A consequence of lognormal distribution is that the geometric average reflects the median cell, rather than the normal arithmetic average that is obtained by dividing the total number of transcripts with the number of cells analyzed (Figure 6). The geometric average can only be determined from single-cell studies, not from studies of cell populations.

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

TRANSCRIPT DISTRIBUTIONS ARE INFORMATIVE AT SINGLE-CELL LEVEL

Figure 6 illustrates some gene expression responses to the induced brain trauma at single-cell level in our experimental data set. *Aqp9* and *Snap* are strongly activated in most cells at day 14 after having hardly been expressed in any cells at earlier time points. *Gfap* and *Vim* are transiently expressed with a peak at day 7, when *Gfap* and *Vim* expression is also observed in most cells. Deeper inspection of the data reveals that *Gfap* and *Vim* are expressed in the same cells, suggesting that their activations are synchronized (Figure 4 and Table 1). The *Gfap* and *Vim* expression profiles at single-cell level are in concordance with the cell population analysis. In contrast, *Aqp9* and *Snap* are expressed in few cells only during days 0–7 and while single-cell profiling reveals the correct picture, the expression analyzed at population level is stochastic, since the amount of transcripts in the only or very few expressing cells present vary.

Heavily skewed and even bimodal distributions of transcripts among single cells are observed for *Eaat1* and *Glul* and indicate presence of multiple cell types, or subpopulations of cells that perhaps are cells in different reversibly interchangeable states (Figure 6). The distribution of *Eaat1* transcripts is bimodal at day 0 and at day 14, while cells with high transcript levels are missing days 3–7. *Glul* shows similar expression pattern and its expression also correlates with *Eaat1* expression (Table 1). These data suggest that one cell subpopulation characterized by high *Eaat1* and *Glul* expression present initially disappears at day 3 to reappear later. From our data, we cannot tell whether expressions of *Eaat1* and *Glul* are uniformly downregulated in this group of cells or whether the subpopulation does not exist at days 3–7, since we are not measuring the same cells over time.

Pdgfra appears upregulated similarly to *Gfap* and *Vim*, when considering only expression levels (Figures 5 and 6). But, inspecting Figure 4 and Table 1, we see that *Pdgfra* is expressed in different cells than the other genes, in particular in cells that do not express *Glul*. This example shows that analyzing population data only gives part of the picture; analyzing correlations between gene expressions among the single cells also provides information about cell subpopulations. This is particularly interesting when cells are exposed to stimuli, and can reveal regulatory gene expression networks [13,15,17].

IDENTIFICATION AND CHARACTERIZATION OF DEFINED SUBPOPULATIONS

Principal component analysis, hierarchical clustering and self-organizing maps (SOMs) are algorithms that can be used to classify cells based on their expression profiles, dividing them into cell states or subpopulations [11–13,15]. However, the subpopulations will not become meaningful until they can be biologically interpreted.

We have shown that SOM can be used to divide genes and/or samples into subpopulations based on their expression profiles [13,29]. Here, we use SOM to classify the cells at day 14 into two subpopulations (Figure 7A). To interpret the biological relevance of the two subpopulations, we calculated the expression of all genes in the typical cell of the two

Table 1: Spearman correlation coefficients for all single cells

	Grial	*Kcna5*	*Kcnj2*	*Aqp9*	*Eaat1*	*Gfap*	*Glul*	*Pdgfra*	*Snap*	*Vim*
Grial	**1**									
Kcna5	0.10	**1**								
Kcnj2	**0.16**	**0.37**	**1**							
Aqp9	0.14	−0.02	−0.02	**1**						
Eaat1	0.12	0.04	0.01	**0.21**	**1**					
Gfap	0.13	**0.22**	0.11	0.05	**0.25**	**1**				
Glul	0.10	**0.15**	**0.20**	**0.21**	**0.53**	**0.23**	**1**			
Pdgfra	**0.17**	**0.19**	0.08	0.03	−0.08	0.06	**−0.34**	**1**		
Snap	**0.28**	0.10	0.10	**0.53**	0.10	0.08	0.09	**0.16**	**1**	
Vim	0.14	**0.28**	0.15	−0.10	0.12	**0.42**	0.06	**0.33**	−0.02	**1**

Notes: Bold indicates ≥99% significance. A correlation coefficient of 1 reflects perfect correlation, −1 reflects perfect anti-correlation and 0 is absence of correlation.

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

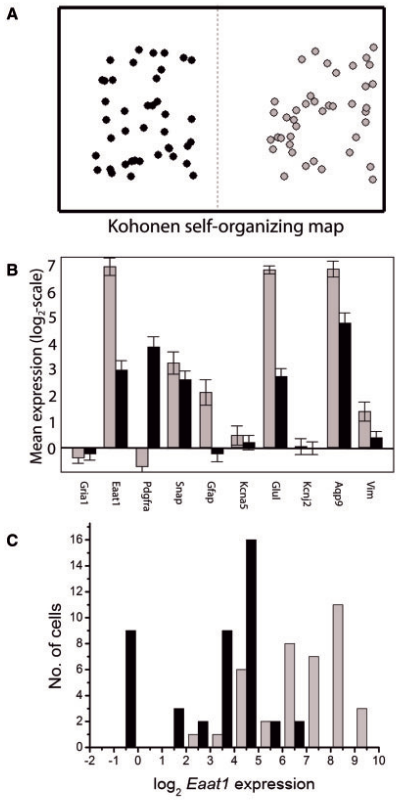


Figure 7: Classification and characterization of subpopulations. **(A)** Two subpopulations of cells (gray and black) were defined using Kohonen SOM at day 14. **(B)** Characterization of the typical cell in each subpopulation. **(C)** Distributions of *Eaat1* transcripts show that the originally observed bimodality in *Eaat1* expression levels (Figure 6) can be accounted for by the two subpopulations identified using SOM. Parameters for the SOM were 2×1 map, 0.1 learning rate, two neighbors and 10 000 iterations. The observed clustering was reducible and independent of parameter settings.

subpopulations (Figure 7B). *Aqp9*, *Eaat1*, *Gfap*, *Glul* and *Vim* are upregulated and *Pdgfr* is downregulated in one subpopulation (gray group) relative to the other subpopulation (black group). The gene expression profile we find for the astrocytes in the subpopulation indicated as gray in Figure 7B has expression features commonly attributed to activated astrocytes. In Figure 7C, we plot the distributions of *Eaat1* transcripts among cells for the two subpopulations separately. As expected, the bimodal pattern of *Eaat1* seen in Figure 6 is resolved by the classification into groups by means of SOM. Closer inspection of Figures 4 and 6 and Table 1 shows that the SOM clustering picks up the major trend at day 14. If we instead use SOM to generate three subpopulations, the third group reveals no new feature relative to the first two groups. The third group shows characteristics of an intermediate state between the two subpopulations (Figure 7). These cells may be in transition between cell states, perhaps astrocytes becoming activated [13]. In conclusion, we present a strategy to identify subpopulations of cells using SOM and show how these subpopulations can be characterized.

FUTURE PROSPECTS

Techniques are rapidly developing for single-cell applications, and the importance of single-cell analysis is realized in many fields of research and diagnostics. Most studies of individual cells are still explorative, but with the improvement of protocols, robust methods will soon be available that will introduce these methods also into diagnostic and prognostic applications.

Key Points

- Single-cell analysis is an emerging concept in molecular biology and diagnostics.
- Comparison between cell population and single-cell gene expression profiling highlights the advantages of single-cell analysis.
- Handling of missing data and sample normalization are two data pre-processing steps that are performed fundamentally different in single-cell analysis compared with cell population analysis.
- Biomarker analysis including correlation studies at single-cell level opens up new possibilities to understand biological processes with refined resolution.
- Single-cell analysis can be used to identify and characterize novel subpopulations of cells.

FUNDING

The authors' work on single cells was partly supported by grants from Assar Gabrielssons Research Foundation; Johan Jansson Foundation for Cancer Research; Socialstyrelsen; Swedish Society for Medical Research; The Swedish Research Council (to A.S. 521-2011-2367); EMBO Short Term Fellowship (to V.R.); ESF Functional Genomics Short Visit Grant (to A.S. and V.R.); BioCARE National Strategic Research Program at University of Gothenburg; Wilhelm and Martina Lundgren Foundation for Scientific Research; the grant of Czech Ministry of Education ME10052; the research project AV0Z50520701 and GACR GA P303/10/1338.

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

References

1. Ståhlberg A, Thomsen C, Ruff D, Aman P. Quantitative PCR analysis o f DNA, RNAs, and proteins in the same single cell. *Clin Chem* 2012;**58**:1682–91.
2. Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet* 2011;**45**:431–5.
3. Wu M, Singh AK. Single-cell protein analysis. *Curr Opin Biotechnol* 2012;**23**:83–8.
4. Larson DR, Singer RH, Zenklusen D. A single molecule view of gene expression. *Trends Cell Biol* 2009;**19**:630–7.
5. Wang D, Bodovitz S. Single-cell analysis: the new frontier in 'omics'. *Trends Biotechnol* 2010;**28**:281–90.
6. Raj A, Peskin CS, Tranchina D, *et al.* Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 2006;**4**:e309.
7. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. *Curr Biol* 2006;**16**:1018–25.
8. Yu J, Xiao J, Ren X, *et al.* Probing gene expression in live cells, one protein molecule at a time. *Science* 2006;**311**: 1600–3.
9. Zenklusen D, Larson DR, Singer RH. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* 2008;**15**:1263–71.
10. Paulsson J. Summing up the noise in gene networks. *Nature* 2004;**427**:415–9.
11. Guo G, Huss M, Tong GQ, *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyte. *Dev Cell* 2010;**18**:675–85.
12. Dalerba P, Kalisky T, Sahoo D, *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011;**29**:1120–7.
13. Ståhlberg A, Andersson D, Aurelius J, *et al.* Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res* 2011;**39**:e24.
14. Narsinh KH, Sun N, Sanchez-Freire V, *et al.* Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. *J Clin Invest* 2011;**121**: 1217–21.
15. Norrman K, Strömbeck A, Semb H, Ståhlberg A. Distinct gene expression signatures in human embryonic stem cells differentiated towards definitive endoderm at single-cell level. *Methods* 2013;**59**:59–70.
16. Liss B, Franz O, Sewing S, *et al.* Tuning pacemaker frequency of individual dopaminergic neurons by Kv4.3L and KChip3.1 transcription. *EMBOJ* 2001;**20**:5715–24.
17. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 2005;**15**:1388–92.
18. Warren L, Bryder D, Weissman IL, Quake SR. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci USA* 2006;**103**:17807–12.
19. Gibson JD, Jakuba CM, Boucher N, *et al.* Single-cell transcript analysis of human embryonic stem cells. *Integr Biol* 2009;**1**:540–51.
20. Powell AA, Talsaz AH, Zhang H, *et al.* Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. *PLoS One* 2012;**7**: e33788.
21. Pantel K, Brakenhoff RH, Brandt K. Detection, clinical relevance and specific biological properties of disseminating tumor cells. *Nat Rev Cancer* 2008;**8**:329–40.
22. Alix-Panabieras C, Schwarzenbach H, Pantel K. Circulating tumor cells and circulating tumor DNA. *Annu Rev Med* 2012;**63**:199–215.
23. Markey AL, Mohr S, Day PJ. High-throughput droplet PCR. *Methods* 2010;**50**:277–81.
24. Ståhlberg A, Rusnakova V, Forootan A, *et al.* RT-qPCR work-flow for single-cell data analysis. *Methods* 2013;**59**: 80–8.
25. Bustin SA, Benes V, Garson JA, *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 2009;**55**:611–22.
26. Velculescu VE, Madden SL, Zhang L, *et al.* Analysis of human transcriptomes. *Nat Genet* 1999;**23**:387–8.
27. Maier T, Schmidt A, Güell M, *et al.* Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol Syst Biol* 2011;**7**:511.
28. Andersen CL, Jensen JL, Orntoft TF, *et al.* Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004;**64**:5245–50.
29. Ståhlberg A, Elbing K, Andrade-Garda JM, *et al.* Multiway real-time PCR gene expression profiling in yeast *Saccharomyces cerevisiae* reveals altered transcriptional response of ADH-genes to glucose stimuli. *BMC Genomics* 2008;**6**: 170.
30. Kubista M, Andrade JM, Bengtsson M, *et al.* The real-time polymerase chain reaction. *Mol Aspects Med* 2006;**27**: 95–125.

Downloaded from http://bfg.oxfordjournals.org/ at Goteborgs Universitet on February 7, 2013

Methods 59 (2013) 80–88

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

RT-qPCR work-flow for single-cell data analysis

Anders Ståhlberg [a,b,*], Vendula Rusnakova [c], Amin Forootan [b,d], Miroslava Anderova [e], Mikael Kubista [a,c,*]

[a] TATAA Biocenter, Gothenburg, Sweden
[b] Sahlgrenska Cancer Center, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden
[c] Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
[d] MultiD Analyses AB, Gothenburg, Sweden
[e] Department of Cellular Neurophysiology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic

ARTICLE INFO

Article history:
Available online 25 September 2012

Communicated by Michael W. Pfaffl

Keywords:
RT-qPCR
Single-cell data analysis
Single-cell biology
Data pre-processing
Missing data
Gene expression profiling

ABSTRACT

Individual cells represent the basic unit in tissues and organisms and are in many aspects unique in their properties. The introduction of new and sensitive techniques to study single-cells opens up new avenues to understand fundamental biological processes. Well established statistical tools and recommendations exist for gene expression data based on traditional cell population measurements. However, these work-flows are not suitable, and some steps are even inappropriate, to apply on single-cell data. Here, we present a simple and practical workflow for preprocessing of single-cell data generated by reverse transcription quantitative real-time PCR. The approach is demonstrated on a data set based on profiling of 41 genes in 303 single-cells. For some pre-processing steps we present options and also recommendations. In particular, we demonstrate and discuss different strategies for handling missing data and scaling data for downstream multivariate analysis. The aim of this workflow is provide guide to the rapidly growing community studying single-cells by means of reverse transcription quantitative real-time PCR profiling.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

In translational molecular research tissue heterogeneity is major complication. Tissues consist of several cell types that respond differently to stimuli. When studying the effects of environmental changes or responses to drugs only some of the cells respond and they may respond differently. Non-responsive cells only confound the measured signal and obscure analysis. The introduction of single-cell analysis has opened up for new possibilities to study tissue heterogeneity by detecting differences even among seemingly identical cells. Several strategies for single-cell analysis have been described and used to study various experimental systems [1–3]. Today, single-cell gene expression profiling using reverse transcription quantitative real-time PCR (RT-qPCR) is the most commonly used method to study individual cells. Single-cell RT-qPCR has been applied to many cell types including neurons [4], astrocytes [5,6], embryonic stem cells [7–9] and beta-cells [10].

Cell collection can be handled in high throughput using fluorescence activated cell sorting (FACS). Specific cells from body fluids and in vitro cultures can be enriched on the basis of surface markers using FACS. Individual cells can also be generated from most tissues by careful dissociation and then collected by FACS, but the context from which the cell is taken is lost during preparation [5,6]. Other means to extract individual cells are microaspiration [4,10] and laser capture microdissection [11,12]. Single-cells are then lysed and if possible no further purification or washing is performed. Purification-free lysis minimizes RNA losses [13,14]. Lysis is followed by reverse transcription, pre-amplification and finally qPCR [9,15–17]. If fewer than ten genes are analyzed and they are reasonably high expressed pre-amplification may not be needed [4–7]. Single-cell RT-qPCR measurements should be performed to the highest possible extent according to the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines [18].

When designing experiments the studied effect should be maximized relative to the confounding variation. Confounding variation has two main contributions: (1) Intersubject variation, caused by the natural biological heterogeneity among the studied subjects that give rise to different gene expression levels; (2) Technical variation, introduced by imprecision in the processing of the samples, comprising the steps of sampling, transport, storage, extraction, reverse transcription, pre-amplification, and qPCR. The confounding variation is reduced by using appropriate controls and references and by performing biological and technical replicates. The measured cycle of quantification (Cq) values are then

* Corresponding authors. Addresses: Sahlgrenska Cancer Center, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Box 425, 40530 Gothenburg, Sweden (A. Ståhlberg); TATAA Biocenter AB, Odinsgatan 28, 411 03 Gothenburg, Sweden (M. Kubista).
E-mail addresses: anders.stahlberg@neuro.gu.se (A. Ståhlberg), mikael.kubista@tataa.com (M. Kubista).

1046-2023/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ymeth.2012.09.007

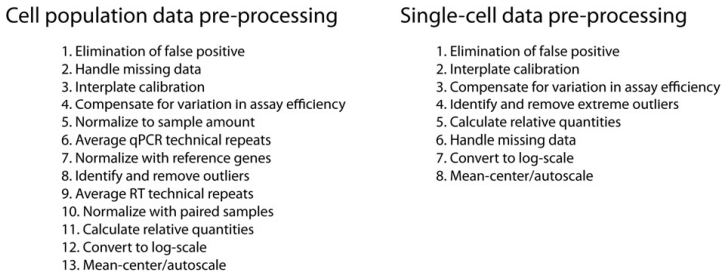


Fig. 1. Workflow for pre-processing of RT-qPCR data at cell population and single-cell level.

pre-processed, taking advantage of the controls and references used, to remove confounding variation ending up with as accurate Cq-values as possible for statistical analysis to extract biologically relevant information. Fig. 1 left shows the general qPCR data pre-processing workflow to reduce confounding variation and prepare data for statistical analysis for traditional samples based on large number of cells. The workflow lists the steps that may be relevant in appropriate order. In practice all pre-processing steps are not needed, since some steps cancel the effect of other steps. Some pre-processing steps considered routine when analyzing traditional samples are not only unsuitable but also inappropriate at single-cell level, since they would increase confounding variation. Other steps, hardly significant in cell population analysis are critical when analyzing individual cells. Here, we describe the steps relevant when pre-processing single-cell RT-qPCR expression data and present a convenient and robust workflow (Fig. 1, right).

2. Description of method

2.1. Data set

A previously unpublished single-cell data set is used to illustrate the single-cell data analysis workflow. Expression of 41 genes was measured in 303 single astrocytes collected by FACS from mouse brains before (day 0) and after (day 3, 7 and 14) induced ischemia, using the GFAP/EGFP mouse model described elsewhere [6]. The cells were lysed, reverse transcribed, pre-amplified and analyzed with qPCR using the BioMark qPCR platform (Fluidigm) as described [5,13]. The detailed experimental protocol will be published elsewhere. The applied data set is representative for single-cells analyzed by RT-qPCR with respect to distribution of transcripts among cells, number of positive cells, and gene expression levels [5–10]. The pre-processing steps and the workflow is illustrated using GenEx (ver. 5.3, MultiD), but in principle any statistical software can be used.

2.2. Data arrangement

In most experiments groups are compared. Most common is to arrange data with Cq-values (the explained variables) in columns headed with experimental group labels (the explanatory variables). This layout provides easy overview of data, and basic statistics such as means and standard deviations (SD) are easily calculated (Fig. 2A). However, this arrangement of data is not practical for more advanced studies involving more than one factor, multiple markers, replicate measurement, multiplate measurements, etc. A more flexible layout is to arrange samples in rows and all variables in columns (Fig. 2B). Today, this is standard arrangement of data in most statistical software. The format is readily generalized to any number of markers and additional columns and rows can be added that specify the experimental design. In GenEx, the explanatory variables are referred to as classification columns and classification rows and have labels starting with #. In the example shown in Fig. 2C, #Repeat indexes qPCR technical replicates (samples with the same index are technical replicates on the qPCR level). These are expected to be highly similar and shall be averaged during data pre-processing. #Group indexes treatment groups that eventually shall be compared using a statistical test. Finally, the study is paired, meaning that each subject received both treatments and a sample was collected after each treatment. Paired study designs are more powerful, because the pairing eliminates much of the systematic subject variation between the compared groups.

2.3. Elimination of false positives

Amplification curves with atypical shapes are usually not processed correctly. Aberrant amplification curves also indicate sample specific problem, such as enzymatic inhibition, and should be removed from further analysis or reanalyzed. The quality of amplification curves is usually performed by visual inspection. For high throughput qPCR analysis, manual inspection is tedious and some automatic approaches based on kinetic analysis that indicate or remove suspicious data are available, but their reliability under various conditions still remains to be validated (www.labonnet.com and www.azurepcr.com) [19].

If reporter dyes are used and melting curve analysis has been performed it should be used to validate that amplification is specific. If aberrant PCR products are formed in addition to the expected product, data should be interpreted qualitatively only, since the Cq-value cannot be trusted. When probes are used melting curve analysis cannot be performed and any non-specific PCR products, which can influence the measured Cq-values by competing for reagents, will go unnoticed. In principle, a non-specific dye can be added into the probe based reaction to monitor formation of aberrant PCR products by melting curve analysis in a separate detection channel [20]. Tables S1A and B shows the data set before and after validation based on melting curve analysis.

2.4. Interplate calibration

Single-cell gene expression profiling experiments tend to be large scale because of the underlying lognormal variation of transcripts among the individual cells, which requires large number of cells to be analyzed to reach statistical significance [5,10,21]. Typically, 50 cells need to be profiled at each condition, and with current workflow 25–100 different transcripts are readily measured per cell. This leads to many thousands of reactions to be analyzed. Even with high throughput platforms multiple plates have to be run. In the processing of the measured raw data the qPCR

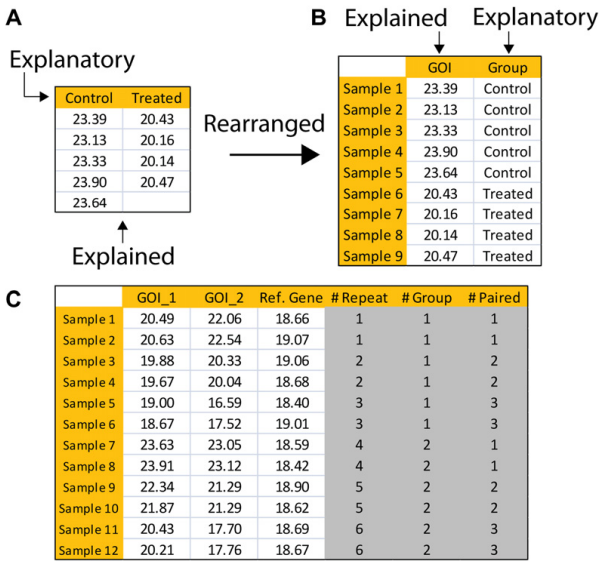


Fig. 2. Data arrangement. (A) Classical data arrangement with Cq-values (the explained variables) in columns headed with the experimental group labels (the explanatory variables). (B) Data arrangement with explained and explanatory variables in different columns is preferred for more advanced experimental designs. (C) Example of data arrangement with one column identifying samples, three columns with explanatory variables (# Repeat, # Group and # Paired) and three columns with explained variables (GOI 1, GOI 2 and Ref. Gene). GOI, Gene of interest; Ref. Gene, Reference gene.

instrument software subtracts baseline and reads out Cq-values as the crossing points of the amplification curves at threshold level or by calculating the maximum of the second derivative. The baseline corrections and the Cq estimations are performed independently for each run, which may introduce systematic variation. This bias is assay independent and can be compensated for running an interplate calibrator, which is a common sample assayed on all plates. Bias can also be introduced by variable assay performance over time. This can be differences in primer, probe, and mastermix batches or external factors such as the time to prepare the experiment. Table S1C shows the data set after interplate calibration. Indexes in the column labeled #plate indicates the independent qPCR runs and #IPC indicates the interplate calibrator. Global interplate calibration was applied to our data set. The average Cq-value of all genes for respective IPC ($Cq_{IPCAverage}$) was used to mean-center data between different runs using the equation:

$$Cq_{GOI}^{Calibtrated} = Cq_{GOI} - \frac{1}{m}\sum_{j=1}^{m}\left(Cq_{IPCAverage} - \frac{1}{n}\sum_{i=1}^{n}Cq_{IPCAverage}\right)$$

where $Cq_{GOI}^{Calibrated}$ and Cq_{GOI} are Cq-values after and before interplate calibration, respectively. m is the number of interplate calibrators in run m (qPCR replicates), and n is the total number of interplate calibrators (m = 1 and n = 8 in our data set, n = number of qPCR runs $*$ number of qPCR replicates). Calibrators can also be used to compensate for technical bias introduced during lysis, RT and pre-amplification [14,15,22,23].

2.5. Assay efficiency correction

PCR efficiency is assay dependent and influenced by factors such as amplicon length and sequence, mastermix composition, and temperature profile. PCR efficiencies can be estimated from standard curves [18,24] and used to correct the measured Cq-values using the equation:

$$Cq_{E=100\%} = Cq_E\frac{\log(1+E)}{\log(2)}$$

where Cq_E is the uncorrected Cq value and E ($0 \leqslant E \leqslant 1$) is the PCR efficiency. The importance of the correction depends on application. Most single-cell data are analyzed without additional normalization and autoscaled by genes (see below). Data scaled this way are independent of PCR efficiency and therefore not affected by the correction. In practice, there may be small effect of PCR efficiency introduced by the handling of off-scale and missing data (see below), since it influences which cells are above the off-scale threshold Cq. Figure S1 shows that the effect of PCR efficiency correction in a Principal Component Analysis (PCA) plot is negligible. Figure S2 shows how the calculated distribution of transcripts among cells is affected by the PCR efficiency correction. Comparing mean expression levels between groups, using e.g. *t*-test or non-parametric tests, the calculated p-values are not affected by the PCR efficiency correction, while the difference between the mean expressions is affected. Assuming 100% PCR efficiency when the true efficiency is lower also underestimates the relative quantities (RQ) and the variability in transcript levels between individual cells. Differences in pre-amplification efficiencies between genes affect the data in a similar way, i.e., negligible effect on analyses performed with autoscaled data and on calculated *p*-values. No correction for any variation in PCR efficiency among genes during pre-amplification is performed here.

2.6. Missing and off scale data

Analyzing small samples there will be missing data (reactions that do not give rise to any Cq-value) and off-scale data (reactions that give Cq-values too high to be trusted). How missing data appear in the data file depends on the qPCR instrument used. Some instruments leave them blank, other indicate them with "NAN" (= Not A Number), while some assign an extreme value to them,

A

	GOI_1	GOI_2	GOI_3	GOI_4	GOI_5	# Replicate	# Group
Sample 1	25.69	18.46	22.34	21.05	14.45	1	1
Sample 2	25.23	18.34	22.24	20.39	14.65	1	1
Sample 3	25.76	18.56	23.10	20.66	14.33	1	1
Sample 4	999	25.38	25.23	23.55	17.86	2	1
Sample 5	999	25.01	24.97	22.74	17.85	2	1
Sample 6	999	24.18	24.76	22.89	17.67	2	1
Sample 7	999	999	28.32	999	22.17	3	2
Sample 8	999	999	26.23	25.45	21.81	3	2
Sample 9	999	999	25.97	25.98	22.22	3	2
Sample 10	21.92	13.29	19.11	19.83	13.52	4	2
Sample 11	21.70	13.37	19.51	19.75	13.43	4	2
Sample 12	21.89	13.23	18.94	19.74	13.39	4	2

B

	GOI_1	GOI_2	GOI_3	GOI_4	GOI_5	# Replicate	# Group
Sample 1	25.69	18.46	22.34	21.05	14.45	1	1
Sample 2	25.23	18.34	22.24	20.39	14.65	1	1
Sample 3	25.76	18.56	23.10	20.66	14.33	1	1
Sample 4		25.38	25.23	23.55	17.86	2	1
Sample 5		25.01	24.97	22.74	17.85	2	1
Sample 6		24.18	24.76	22.89	17.67	2	1
Sample 7			28.32		22.17	3	2
Sample 8			26.23	25.45	21.81	3	2
Sample 9			25.97	25.98	22.22	3	2
Sample 10	21.92	13.29	19.11	19.83	13.52	4	2
Sample 11	21.70	13.37	19.51	19.75	13.43	4	2
Sample 12	21.89	13.23	18.94	19.74	13.39	4	2

C

GOI_1	# Replicate
	1
17.0	1
16.8	1

Averaging

GOI_1	# Replicate
16.9	1
17.0	1
16.8	1

D

GOI_1	GOI_2	# Replicate
22.5		1
22.0	17.0	1
21.8	16.8	1

Imputation

GOI_1	GOI_2	# Replicate
22.5	17.5	1
22.0	17.0	1
21.8	16.8	1

Fig. 3. Missing data. (A) Data can be missing due to technical failure or to too low gene expression level. (B) Any extreme Cq-values (e.g. 999) reported by the instrument software should be eliminated. (C and D) Missing data can be replaced using information from technical replicates, either by the average of the successful measurements or by imputation, which also considers systematic differences between replicate samples based on information from other genes.

such as 999. In particular, the latter has to be amended in the pre-processing of qPCR data, since an extreme value will have profound influence on any downstream parametric statistical analysis and the conclusions reached will be dominated by the extreme values rather than on those measurements that produce reliable data. In any workflow extreme Cq-values assigned by the instrument software should be removed or replaced with "NAN" (Fig. 3A).

Before handling missing data it helps realizing there are two distinctly different reasons a reaction does not produce a Cq-value: (1) the reaction chamber contained template molecules but the PCR failed (2) the reaction did not produce any product because there was no template in that particular reaction chamber. These two cases should be handled differently. If technical replicates are available (Fig. 3B), missing data due to reaction failure can be restored based on replicate information. The simplest approach is to restore them by the average of the successfully recorded replicates (Fig. 3C). A more advanced and accurate approach is imputation. Imputation accounts also for any systematic differences between replicates based on information from other assays (Fig. 3D). Imputation shall only be applied to sets that have common replicates. Typically, imputation is used for technical replicates; if a biological replicate is missing it is usually better to exclude it. If several levels of technical replicates are available, imputation can be repeated at each level. For single-cell expression data technical replicates are usually not available, since it is more cost efficient to analyze more cells than to collect replicate data. Should technical (qPCR) replicates be available they are averaged at this stage of pre-processing.

Very high Cq-values reported by the instrument software are often not reliable even when the correct PCR product is formed, since they correspond to very few molecules or even fractions of a molecule and have become large due to various problems such as failed amplification of single molecule targets in the initial cycles, delayed amplification due to competing reactions, or inhibition. They may also be false positives (primer-dimers and other aberrant products). These Cq-values are not useful and should be discarded. A pragmatic approach is to delete all Cq-values above a certain threshold. In principle, Cq-values higher than the Cq-value expected for a reaction starting with a single template molecule ($Cq_{N=1}$) should not be trusted. $Cq_{N=1}$ depends on instrument, PCR efficiency, applied detection chemistry, and instrument settings. On classical microtiter plate based instruments $Cq_{N=1}$ is usually around 35–37 cycles. On high throughput platforms using smaller reaction volumes and more sensitive

optics $Cq_{N=1}$ is some 10 cycles lower. $Cq_{N=1}$ is assay dependent and can in principle be determined as the intercept of a qPCR standard curve with known concentrations. However, this is not practical, since it requires preparing purified starting material (preferably cDNA) of known concentration for every target and should be run in a matrix similar to that of the single-cell experiments. Pragmatic is to delete all Cq-values above a certain threshold, Cq_{Cutoff}, which is the same for all the assays. A reasonable Cq_{Cutoff} can be chosen by inspecting control charts (Fig. 4). The first control chart shows the variation of Cq-values across samples for the different assays (Fig. 4A). In this chart cells that have overall very few transcripts (all Cq-values are high) are identified (Table S1D) and may be considered failed or at least anomalous. Furthermore, poor performing assays are identified by lines that never or rarely reach below Cq_{Cutoff}. A second control chart presents the measured Cq-values of each assay in a box and whiskers plot (Fig. 4B). Box and whiskers plots are expected to be symmetric because of the underlying lognormal distribution of transcripts among individual cells. Asymmetry may indicate that the number of copies of the particular transcript is too low to be reliably detected giving rise to truncation at high Cq. However, distributions of transcripts can have many features and genes should in general not be eliminated due to their expression characteristics. Tentative outliers are identified using standard statistical tests and are indicated with symbols. A third control chart shows the fractions of missing Cq-values, off scale Cq-values ($Cq > Cq_{Cutoff}$) and valid Cq-values ($Cq \leqslant Cq_{Cutoff}$) for each gene (Fig. 4C). This plot identifies assays that are not contributing with meaningful information. The control charts can be calculated and inspected for different Cq_{Cutoff}-values, guiding the selection of an appropriate Cq_{Cutoff} for further analysis.

Most single-cell expression data analyses can be performed directly on Cq-values, but results become more intuitive and suited for visualization if expression values are converted to relative copy numbers. RQ can be calculated assuming $Cq_{cutoff} = Cq_{N=1}$:

$$\text{Relative quantities of cDNA molecules (RQ)} = 2^{Cq_{cutoff} - Cq}$$

However, data should be expressed in $\log_2$-scale, since the underlying distribution of transcripts among cells is lognormal. A practical approach is to set all missing data to -1 in $\log_2$-scale, corresponding to 0.5 molecule in RQ. For the data in our example we applied $Cq_{Cutoff} = 27$ (Table S1E) and converted data to RQ using the equation above (Table S1F). Remaining missing data were assigned RQ = 0.5 (Cq Off scale = 27 + 1, Table 1F). The rational of assigning $Cq_{Cutoff} + 1$ to the missing data is that Cq_{Cutoff} is global representation of $Cq_{N=1}$, which is the Cq expected for a single template molecule. Hence, $Cq_{Cutoff} + 1$ represents a concentration lower than a single molecule, in fact, half that concentration. Of course, we cannot have 0.5 molecules in a test tube, but due to sampling ambiguity we cannot conclude a test sample is negative when we obtain a negative PCR analyzing an aliquot. We only know that the sample contains fewer molecules than we are able to detect with the current protocol. Even with a perfect PCR sampling ambiguity, which can be modeled by the Poisson distribution, results in a limit of detection (LOD) of 3 molecules at 95% confidence, i.e., the volume fraction analyzed shall in average contain 3 molecules for a reaction to give positive PCR in 95% of the cases [18]. For real samples LOD is typically higher because of losses in sample processing, RT and imperfect PCR. Analyzing experimental data, a more conservative approach is to set Cq_{Cutoff} to a lower value (25–26 cycles on a high throughput instrument) with a larger offset to the missing data. For example, assigning $Cq_{Cutoff} + 4$ to the missing data is equivalent to assigning a concentration to those samples that is 1/16 of LOD. The effect of the offset is to weight the importance of not detecting a transcript in a cell. We have found that a good strategy is to use a small offset when analyzing homogeneous cohort of cells that are expected to express the same genes, and

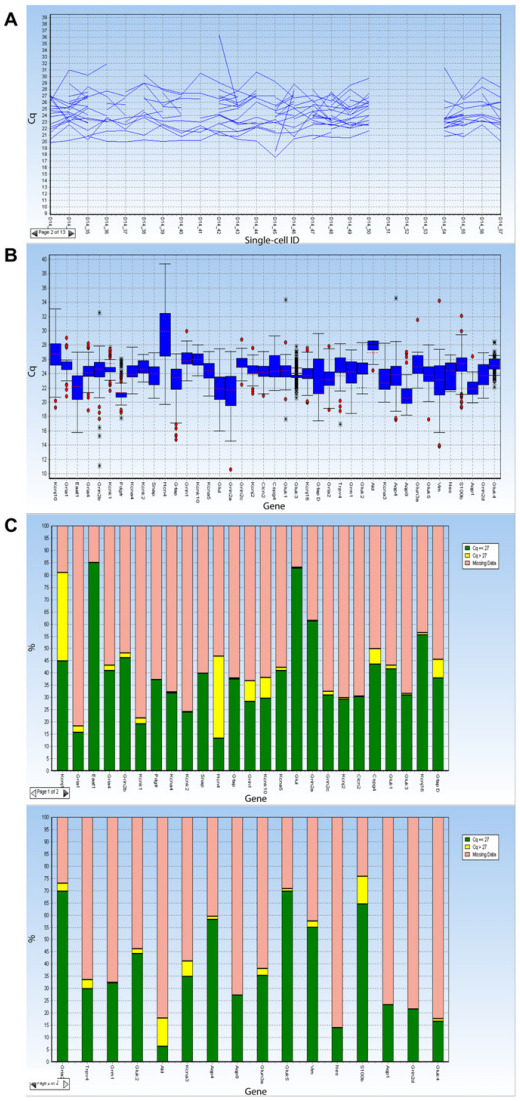


Fig. 4. Quality control charts. (A) Variation of Cq-values across samples for the different assays. In this chart anomalous single-cells that generated few reliable data (high rate of missing data and high Cq-values for all genes) are identified. Here, two cells (D14 51 and D14 53) showed no expression what so ever and should be eliminated from further analysis. (B) Box and whisker plot provides an overview of the spread of the genes' expressions. Potential outliers are indicated with circles and extreme outliers with stars. Potential and extreme outliers have expression levels outside 1.5*IQR and 3*IQR, respectively (IQR = Quartile 3–1). Quartile 1 and 3 represent the bottom and top of the box and are the 25th and 75th percentile of the gene expression values. Box and whisker plots with many potential outliers indicate genes (*Kcnk1*, *Gluk3* and *Gluk4*) with skewed distributions of transcripts among the individual cells. (C) Gene quality graph. The percentages of single-cells with Cq below and above Cq_{cutoff} and with missing data are shown. Genes with many Cq-values larger than Cq_{cutoff} (*Kcnj10* and *Hcn4*) have either very low expression and/or the assays are performing poorly.

a large offset when objective is to distinguish between cell types that exclusively express some markers. Figure S3 shows the effect of changing $Cq_{cutoff} = 27$ to $Cq_{cutoff} = 25$ on PCA, and Figure S4 shows the effect of offset when assigning different Cq-values to the missing data. For this particular data set the Cq_{Cutoff}-value and the offset have insignificant effect on the analysis result. This is usually the case. Still, we recommend users to test the effects of Cq_{Cutoff} and offset when analyzing new data to validate the robustness of their handling of missing data. One should also be cautious when key classification genes are expressed at low level with high Cq-values. It is then important to set Cq_{Cutoff} at a value that distinguishes between positive and negative cells with respect to this marker. In some cases it may be necessary to treat key classifiers separately, using an assay specific Cq_{cutoff}. Genes that may be biased due to the choice of Cq_{cutoff} are identified in the control charts (Figs. 4B and C). The rationale of handling missing data after, instead of before, the transformation to RQ is that RQ of the Cq_{cutoff}-value (and not of $Cq_{cutoff} + 1$) shall be assigned arbitrary expression of 1 (zero in $\log_2$-scale). The RQs after handling missing data and RQs in $\log_2$-scale are shown in Tables S1G and S1H, respectively. Conversion of relative quantities to the number of cDNA molecules or to the number of mRNA molecules requires calibration with standards [9,14,25]. Missing data due to few target molecules in the reaction vessel is in single-cell studies with good assays more common than reaction failure. Missing data can therefore be replaced with Cq_{cutoff} + offset instead of imputation or averaging even when qPCR replicates are available. In practice the two options to handle missing data will produce very similar results.

2.7. Basic statistics and distributions

To get an overview of the data we calculate some basic statics for all the genes studied, including the number (or fraction) of cells expressing each gene, and the mean and standard deviation (SD) of all the genes' expressions. Mean and SD are calculated on data in logarithmic scale, since the underlying distribution is lognormal, but can be converted to linear scale for presentation. The fraction of cells that express a particular gene correlates well with the gene's mean expression (Fig. 5A and B, Spearman correlation = 0.78, $P < 0.01$). The mean expression is calculated as the arithmetic average of the relative quantities expressed in logarithmic scale ($\log_2$). This is equivalent to calculating the geometric average of the data in linear scale, and reflects the medium number of transcripts per cell, which is the number of transcripts expected in the typical cell of the population. From Fig. 5C we calculated that the median number of molecules per reaction well was 3.1 ($RQ_{median} = 3.1$ assuming $Cq_{cutoff} = Cq_{N=1}$). To convert this to the median number of mRNA molecules per cell one must consider RT, pre-amplification, and qPCR efficiencies as well as all dilution steps. The SD of the measured Cq-values reflects variability, which is dominated by the heterogeneity of the cells. Usually SD scales with the average expression of the gene (Fig. 5D-E, spearman correlation = 0.98, $P < 0.01$). This correlation is quite linear, which is equivalent to the relative standard deviation (SD/mean, also known as the coefficient of variation, CV) being independent of concentration. For our data CV was roughly 100% (mean and SD are equal). A transcript that has deviant CV should be suspected to have differently regulated transcription or anomalous stability.

It is good idea to visualize the distributions of the different transcripts among the cells either as frequency histograms or violin plots (Fig. 6). From experience we know that under most conditions the distribution of a transcript among homogenous cells can be fitted with a lognormal distribution [5,8–10,14]. Recent theoretical considerations suggest that a Poisson-Beta distribution may be more appropriate to model single-cell expression data [26]. However, the two distributions have similar features and are both suitable for analysis of single-cell data with parametric statistics. Deviant distributions, in particular highly skewed and bimodal (two maxima) distributions, indicate that several cell types are present and/or that the cells respond differently to stimuli [5,10]. In our example the distribution of *Vim* transcripts among the cells changes over time. In the reference material (day 0) most cells had low/no *Vim* expression. Upon injury *Vim* was upregulated and reached maximum expression at day 7. At day 14 *Vim* was downregulated again, but did not return all the way to the basal expression level of the healthy brain.

2.8. Mean-centering and auto-scaling

Expression data can be analyzed one gene at a time using traditional univariate statistics, such as the *t*-test, Anova and regression. These tests assume data are normal distributed, which is typically satisfied for single-cell gene expression when expressed in logarithmic scale. Univariate statistical methods can be applied directly on the measured data or on calculated logarithmic expression quantities. They usually work well on single-cell data, but suffer from ambiguity if many genes are analyzed, since the false positive rate increases and many genes will appear differentially expressed due to chance only. An alternative, or rather a complementary approach, is to analyze the data using multivariate statistical methods. These methods take the collective response of all the genes into account. Popular multivariate statistical methods suitable for qPCR expression profiling include PCA, Hierarchical clustering, and the Self organizing map [27]. When using multivariate methods we must consider the expression levels of all the genes, since if no scaling is applied the more expressed genes will have higher weights in the analysis and will dominate the result. This is usually not desired. Rather, it is preferred to give all the genes same weights in the analysis, making them equally important in the data processing. This is accomplished by autoscaling the data, which is done by calculating *z*-score for each gene by subtracting its mean expression and divide by its SD [27,28]. Thus, a z-score of 2 indicates that a gene in a particular sample is overexpressed by two SDs relative to its mean expression in all the samples. Autoscaling is preferred when mining single-cell qPCR expression profiling data. There are few cases, though, when autoscaling is not suitable. When the panel includes genes that are not responsive to the studied conditions and their expression among the samples show only random variation, autoscaling will only increase the noise these genes contribute to the data set. If the number of non-responsive genes is large the data quality may be compromised by autoscaling. An option is then to only mean-center the data. Mean-centering is performed by subtracting the average expression of each gene, but not dividing with the SD. Mean-centered data have zero average expression for all the genes, but their levels still vary. In practice one often starts analyzing mean-centered data, identifying the responsive genes, and then inactivating those non-responsive for further analysis, which is performed on autoscaled data. Variable selection tools such as dynamic PCA are very useful for identification of the most responsive and relevant genes. Sometimes it is also of interest to analyze only a subset for samples. If data are mean-centered or autoscaled and samples are removed from the data set, the original scaling is not valid and has to be repeated. Software such as GenEx handles this automatically, but it has to be remembered if analyzing data manually or using spreadsheet program such as Microsoft Excel. If too many data are missing one should also use mean-centering and autoscaling with care, since the offset chosen may have pronounced effect on the scaling. This can be tested by reanalyzing data using a different offset as described above. One should also test the influence of genes with low expression levels and many missing data on the multivariate analysis result by reversibly inactivating them. Still, an option, if

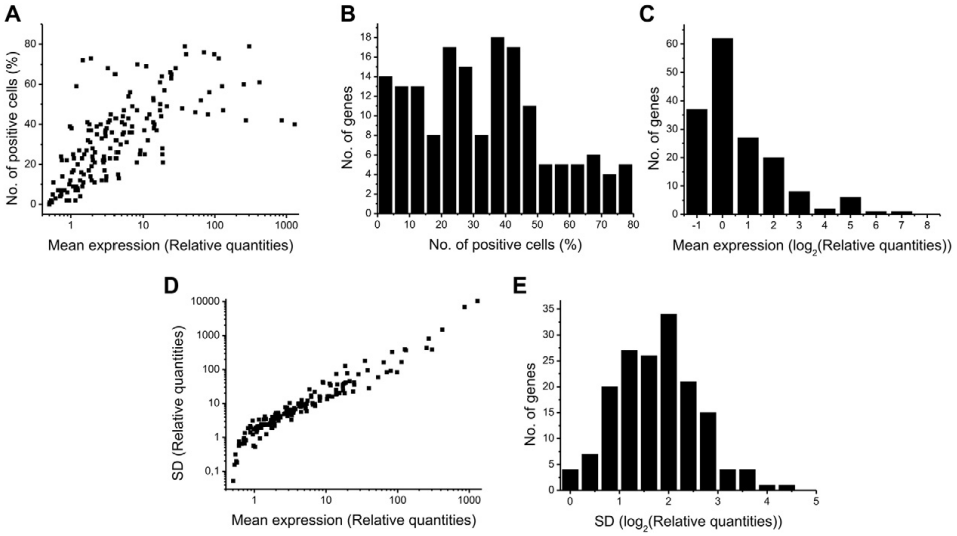


Fig. 5. Features of gene expression profiling at single-cell level. (A) Correlation between the number of cells containing a gene transcript and the mean expression of the same gene. (B) Number of genes versus percentages of positive expressing cells. (C) Diagram showing the frequency of expression levels among the genes. (D) Correlation between SD and mean expression (data in logarithmic scale). (E) Variation of expression levels among the cells presented as frequency distribution of the standard deviations of the genes' Cq-values. Genes' expressions at different time points after injury were treated as independent variables (N_{tot} = 164).

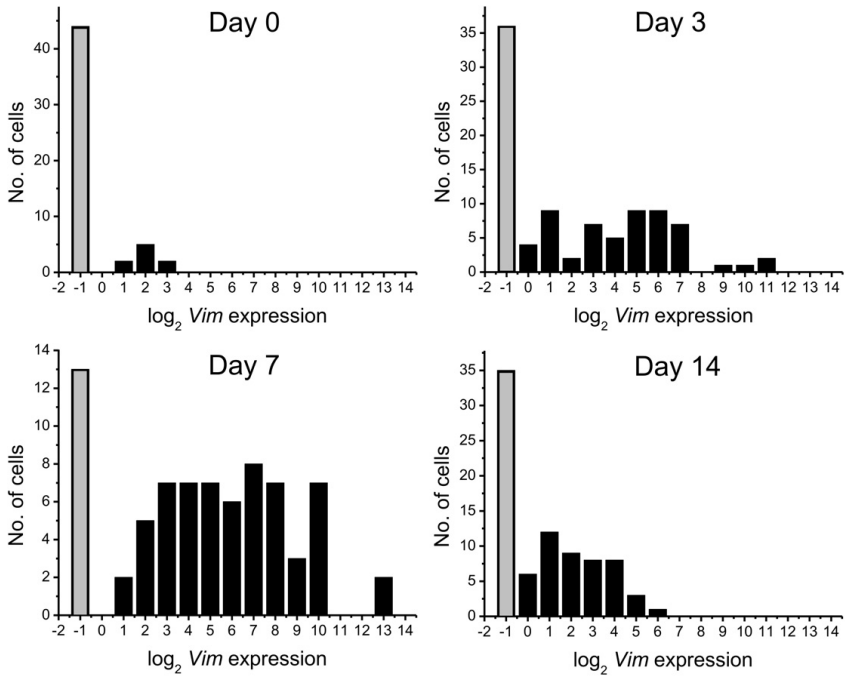


Fig. 6. Distribution of Vim transcripts among single-cells at different time points after injury. Frequency of cells with different expression levels of *Vim* before (day 0) and after injury (day 3, 7, and 14). The gray bars indicate the number of cells with no detectable expression of *Vim*.

there are many cell specific genes and the objective is to distinguish cell types, is to analyze the data in binary mode; i.e., considering only whether a gene is expressed or not in a cell. Mathematically this is equivalent to using a very large (infinite) offset when handling missing data. The offset used when handling missing data is actually setting the weight to the importance of a transcript not being present in a cell. Fig. 7 shows PCA with mean-centered (Fig. 7A) and autoscaled (Fig. 7B) data. Figure S5 shows the PCA loadings calculated with the two scaling options and the impacts of the genes. Only minor differences caused by the scaling are observed in the PCA scores and loadings for our data (Figs. 7 and S5).

Traditional autoscaling and mean-centering is applied to variables, which in qPCR are the genes, with the objective of giving them the same weight in the analyses [27,28]. For single-cell expression profiling one may also apply autoscaling or mean-centering to the samples. For each cell the average expression of all the genes is subtracted and optionally the data are divided by the standard deviation of the genes' expressions in that cell. Mean-centering the data per cell is particularly interesting, since it is equivalent to global normalization or at least truncated global normalization, since we only measure the expression of a fraction of all the genes. Such normalization will account for cell specific variations in extraction, RT yield, and pre-amplification/PCR efficiencies. Autoscaling data per cell standardizes the expression data to a common scale; for example, a cell type may be characterized by having certain genes expressed at least 2 SDs above the mean expression of all the other genes studied in that cell. Such scaling removes, for example, the influence of total expression level per cell and it allows comparison of single-cell with few cells data.

2.9. Preprocessing steps not suitable for single-cell analysis

As indicated by the workflow in Fig. 1 some steps of classical qPCR data pre-processing are not used on data from single cells. Traditional RT-qPCR data are almost always normalized to compensate for differences in the amount of sample material analyzed. Most common is to normalize with endogenous reference genes [18]. Their expression is not only expected to be proportional to the amount of material, but it also reflects any loss in the handling and processing of the samples. However, transcription in individual cells occurs in bursts and under normal conditions no gene has constant steady state level of transcripts [29]. Since the transcriptional bursts of most genes are uncorrelated one cannot use common reference genes for normalization. In principle global normalization based on the average expression of all the transcripts can be performed (or mean center data per sample, see above), but the robustness of this approach is still to be proven. Complications include the rather small number of genes typically analyzed per cell and the ambivalence in the handling of missing data. Another option may be to normalize to a highly abundant or degenerative transcript, such as the Alu repeat in human. This approach, however, is complicated by the genomic DNA background present.

The accuracy of a RT-qPCR measurement is proportional to the number of molecules analyzed and if the number becomes very small accuracy is compromised due to sampling (Poisson) ambiguity. Instead of dividing a sample in duplicates that are analyzed separately and the results averaged, it is better to measure the single whole sample with twice the transcript amount [14]. Generally all samples should be processed keeping all dilution steps, including lysis, RT, pre-amplification and qPCR, to a minimum. As a rule of thumb the sample volume analyzed in a reaction vessel by qPCR should contain a minimum of 10–25 target molecules, which will contribute with sampling noise (SD) of 0.5–0.3 cycles. Notably, instead of aliquoting a cDNA sample with few molecules for separate singleplex qPCRs, it is better to first pre-amplify the cDNA. Even though preamplification introduces some variation it is less than the sampling error introduced when aliquoting a highly diluted

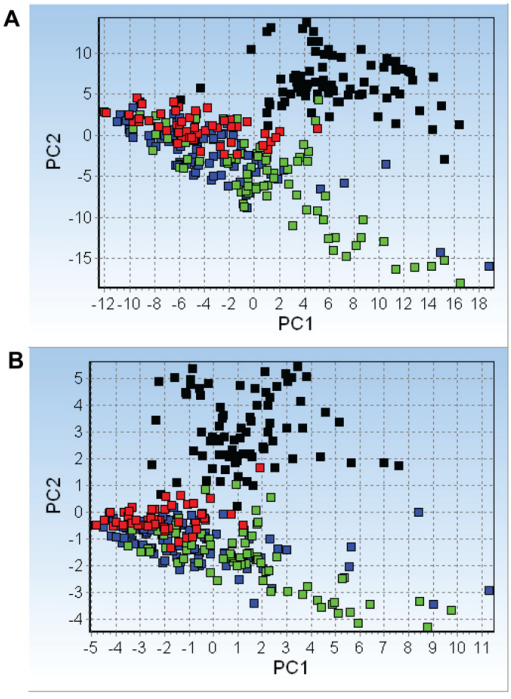


Fig. 7. Principal component analysis of single-cells based on their expression profiles. PCA based on mean-centered (A) and autoscaled (B) data. Loadings for principal component analysis are shown in Figure S5. Data points represent individual astrocytes collected from mouse brains before (day 0, red) and after (day 3, green, day 7, blue, and day 14, black) ischemia.

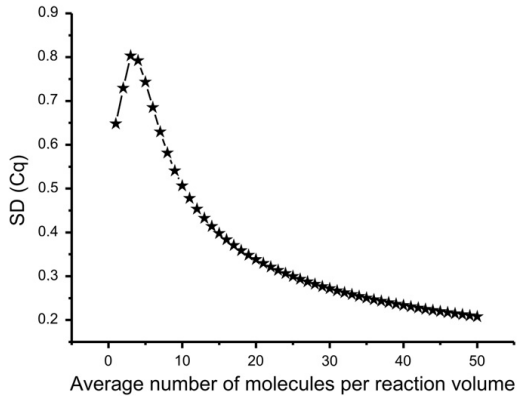


Fig. 8. Variation due to sampling ambiguity. Standard deviation of Cq-values of replicate samples caused by sampling ambiguity calculated by the Poisson distribution. The SD increases rapidly when the number of molecules decreases. The reason maximum SD is obtained at about 3 molecules is that only positive samples are considered.

sample (Fig. 8) and [15,22,23]. After pre-amplification the amount of transcripts is usually high enough to allow technical qPCR replicates to be measured, as a reassurance if a qPCR fails. From a cost-performance and statistical perspective, though, it is usually better to analyze larger number of single-cells than performing technical replicates. Of course, in the validation of assays, including controls, standard curves and LOD, technical replicates should be measured [18].

2.10. Classification algorithms

Analysis of data and data mining is usually hypothesis driven and therefore application dependent and the different methods available have been discussed elsewhere [5,27,28]. Here we only use PCA to visualize some of the important features of the pre-processing using an example data set collected on astrocytes harvested at different time points after brain injury in mice. PCA clearly reveals changes in gene expression profiles at the single-cell level occurring over time, reflecting heterogeneity and cell transformation induced by the injury. Other powerful methods to analyze and classify individual cells include hierarchical clustering, and the self organizing map [5,27,28].

3. Concluding remarks

Gene expression profiling using RT-qPCR took a major leap forward by the publication of the MIQE guidelines, which help users analyzing classical samples [18]. For single-cell profiling some aspects (e.g., reference gene normalization) of the standard workflow are not valid or applicable, while other options (e.g., mean-centering and autoscaling along samples) are almost compulsory. Here we present a complete workflow for the pre-processing of single-cell RT-qPCR data that is robust and prepares the data in a meaningful way for further analysis and mining. Using an example data set we also present several characteristics of single-cell expression data. For some pre-processing steps we present options and also recommendations based on the experience gathered so far. The aim of this workflow is to provide guidelines to the rapidly growing community studying single-cells by means of RT-qPCR profiling. To further stimulate exchange and experiences in the single-cell expression field we have made our example data available in supplement.

4. Disclosure

A.S., V.R., A.F. and M.K. declare stock ownership in TATAA Biocenter AB. A.F. and M.K. also declare stock ownership in MultiD.

Acknowledgements

This work was partly supported by grants from Assar Gabrielssons Research Foundation, Johan Jansson Foundation for Cancer Research, Socialstyrelsen, Swedish Society for Medical Research, The Swedish Research Council (A.S. 521-2011-2367), EMBO Short Term Fellowship (V.R.), ESF Functional Genomics Short Visit Grant (A.S. and V.R.), BioCARE National Strategic Research Program at University of Gothenburg, Wilhelm and Martina Lundgren Foundation for Scientific Research, the grant of Czech Ministry of Education ME10052, the research project AV0Z50520701, and GACR GA P303/10/1338.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2012.09.007.

References

[1] T. Kalisky, P. Blainey, S.R. Quake, Annu. Rev. Genet. 45 (2011) 431–435.
[2] M. Wu, A.K. Singh, Curr. Opin. Biotechnol. 23 (2012) 83–88.
[3] D.R. Larson, R.H. Singer, D. Zenklusen, Trends Cell Biol. 19 (2009) 630–637.
[4] B. Liss, O. Franz, S. Sewing, R. Bruns, H. Neuhoff, J. Roeper, EMBO J. 20 (2001) 5715–5724.
[5] A. Ståhlberg, D. Andersson, J. Aurelius, M. Faiz, M. Pekna, M. Kubista, M. Pekny, Nucleic Acids Res. 39 (2011) e24.
[6] J. Benesova, V. Rusnakova, P. Honsa, H. Pivonkova, D. Dzamba, M. Kubista, M. Anderova, PLoS One 7 (2012) e29725.
[7] A. Ståhlberg, M. Bengtsson, M. Hemberg, H. Semb, Clin. Chem. 55 (2009) 2162–2170.
[8] K.H. Narsinh, N. Sun, V. Sanchez-Freire, A.S. Lee, P. Almeida, S. Hu, T. Jan, K.D. Wilson, D. Leong, J. Rosenberg, et al., J. Clin. Invest. 121 (2011) 1217–1221.
[9] K. Norrman, A. Strömbeck, H. Semb, A. Ståhlberg, Distinct gene expression signatures in human embryonic stem cells differentiated towards definitive endoderm at single-cell level, Methods 59 (2012) 59–70.
[10] M. Bengtsson, A. Ståhlberg, P. Rorsman, M. Kubista, Genome Res. 15 (2005) 1388–1392.
[11] F. Kamme, R. Salunga, J. Yu, D.T. Tran, J. Zhu, L. Luo, A. Bittner, H.Q. Guo, N. Miller, J. Wan, et al., J. Neurosci. 23 (2003) 3607–3615.
[12] J. Gründemann, F. Schlaudraff, O. Haeckel, B. Liss, Nucleic Acids Res. 36 (2008) e38.
[13] A. Ståhlberg, M. Bengtsson, Methods 50 (2010) 282–288.
[14] M. Bengtsson, M. Hemberg, P. Rorsman, A. Ståhlberg, BMC Mol. Biol. 9 (2008) 63.
[15] B.C. Fox, A.S. Devonshire, M.O. Baradez, D. Marshall, C.A. Foy, Anal. Biochem. 472 (2012) 178–186.
[16] P. Dalerba, T. Kalisky, D. Sahoo, P.S. Rajendran, M.E. Rothenberg, A.A. Leyrat, S. Sim, J. Okamoto, D.M. Johnston, D. Qian, et al., Nat. Biotechnol. 29 (2011) 1120–1127.
[17] G. Guo, M. Huss, G.Q. Tong, C. Wang, L. Li Sun, N.D. Clarke, P. Robson, Dev. Cell 18 (2010) 675–685.
[18] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, Clin. Chem. 55 (2009) 611–622.
[19] T. Bar, M. Kubista, A. Tichopad, Nucleic Acids Res. 40 (2012) 1395–1406.
[20] K. Lind, A. Ståhlberg, N. Zoric, M. Kubista, Biotechniques 40 (2006) 315–319.
[21] A. Ståhlberg, M. Kubista, P. Åman, Expert Rev. Mol. Diagn. 11 (2011) 735–740.
[22] A.S. Devonshire, R. Elaswarapu, C.A. Foy, BMC Geomics 12 (2011) 118.
[23] M. Reiter, B. Kirchner, H. Müller, C. Holzhauer, W. Mann, M.W. Pfaffl, Nucleic Acids Res. 39 (2011) e124.
[24] A. Ståhlberg, P. Åman, B. Ridell, P. Mostad, M. Kubista, Clin. Chem. 49 (2003) 51–59.
[25] A. Ståhlberg, M. Kubista, M. Pfaffl, Clin. Chem. 50 (2004) 1678–1680.
[26] A. Raj, C.S. Peskin, D. Tranchina, D.Y. Vargas, S. Tyagi, PLoS Biol. 4 (2006) e309.
[27] A. Bergkvist, V. Rusnakova, R. Sindelka, J.M. Garda, B. Sjögreen, D. Lindh, A. Forootan, M. Kubista, Methods 50 (2010) 323–335.
[28] A. Ståhlberg, K. Elbing, J. Andrade-Garda, B. Sjögreen, A. Forootan, M. Kubista, BMC Genomics 9 (2008) 170.
[29] A. Raj, A. van Oudenaarden, Cell 135 (2008) 216–226.

GenEx - Data Analysis Software

Mikael Kubista,[1,2] Vendula Rusnakova,[2] David Svec,[1,2] Björn Sjögreen[3]and Ales Tichopad,[1,2]
[1]TATAA Biocenter, Gothenburg, Sweden, Department of Biotechnology, Czech Academy of Science, Prague, Czech Republic, [3]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, United States.

Abstract
As the qPCR field advances the design of experiments and the analysis of data is becoming more important and more challenging. Calculations of relative expression of a reporter gene to a reference gene in pairs of samples using the ΔΔCq method are no longer sufficient. Rather studies are based on multiple markers, designed with nested level, exploring or confirming the effect of multiple factors, occasionally in paired designs etc. Proper handling of such data requires software that supports the planning and design of experiments, and to analyze the data. Several software with these capacities are emerging. This chapter describes some of the the features of the most powerful of those: GenEx from MultiD Analysis.

Introduction
In the early days when quantitative real-time PCR (qPCR) was new the experiments performed were rather uncomplicated involving only small number of samples and analysis could be performed using a standard spreadsheet tools such as Microsoft Excel. However, as the technology develops advanced experimental designs are becoming common, and large data sets, often involving multiple plates or run on high throughput instruments are collected. These analyses require more information on how the experiment was set up to handle references, standards and controls appropriately, and correctly account for the variance and covariance in the measured data. On some high throughput instrument even the transfer of data from the instrument to the analysis software is challenging, requiring user intervention with obvious risk to accidently change or lose some of data. In parallel with the instrument becoming more advanced and powerful qPCR data analysis software is being developed. Today this development is heavily supported by the instrument companies, who have realized their customers' needs. The qPCR software on the market today is GenEx from MultiD,[1] which is supporting all leading qPCR instruments, Statminer from Integromics,[2] which is supporting qPCR instruments from Life Technologies, and qBaseplus from Biogazelle,[3] which is supporting Bio-Rad instruments. Since GenEx is currently the only generic platform and is also far most developed we will focus this chapter on it. Much of what is said, though, is general for qPCR analysis and applies also to the other software on the market.

Data arrangement
When groups are compared data are classically arranged with the measured Cq (explained variables) in columns headed with the experimental group label (explanatory variables). This arrangement provides easy overview of the data (Figure 1). However, it is not practical for advanced studies that may include more than one nominal factor or covariate (variable of metric character, e.g. time, age, dose etc.), multiple markers, replicate measurement, repeated measurements (same subject sampled repeatedly), multiplate measurements etc. More flexible is to arrange the data with samples as rows and all variables in columns (Figure 1). This is standard arrangement in statistical software today. The format is readily generalized to any number of markers and additional columns and rows can be added that specify the experimental design, indexes samples, markers, plates etc. In GenEx these are referred to as classification columns and classification rows and have labels starting with #. In the example shown in Figure 2 #Repeat indexes qPCR technical replicates (samples with the same index are technical replicates on the qPCR level). These are expected to be highly similar and shall be averaged during data pre-processing. #Treatment indexes treatment groups that eventually

shall be compared using a statistical test. Finally, the study is paired, meaning that each subject received both treatments and a sample was collected after each treatment. Paired study designs are more powerful, because the pairing eliminates much of the random as well as systematic subject variation between the compared groups. This elevates the power of the test is, and hence, the experiment requires fewer individuals. It can be, for example, samples collected from all subjects before treatment and a second set collected after treatment. It can also be positive and negative samples collected from the same subject or genetically similar individuals such as siblings, identical twins or clones used pair-wise. A special type of paired study design is repeated samplings, used in more than two subsequent measurements. In the general case the word "paired" is replaced by repeated. Specialized statistical procedures are available to analyze multiple repeated samples.

Data import

The experimental design is defined in part by deciding on the experimental factors and covariates involved in the experiment and in part when the samples and assay are mixed while dispensed into the qPCR containers. This information is critical for proper analysis and mining of the measured data. This has been realized by several of the leading qPCR instrument and assay providers. Roche LC480 software for RealTime Ready custom and focus panels, for example, names all genes and samples, indicates reference genes, and specifies technical and biological replicates at various levels. This information is transferred to GenEx and arranged appropriately for downstream analysis using a wizard. Similar high level user friendly solution is provided by Exiqon, who offer a customized version of GenEx with a powerful wizard to read their miRCURY LNA™ Universal RT microRNA PCR platform.[4] On the BIOMARK microfluidic platform from Fluidigm, technical and biological replicates are indicated by the naming of assays and samples, and appropriate classification columns are automatically created. Other qPCR instruments that generate data that can be read by GenEx include the Stratagene MX300X from Agilent, Realplex from Eppendorf, CFX96/384 from Bio-Rad, Eco from Illumina, and the many different qPCR platforms from Life Technologies. These efforts from the instrument manufacturers to transfer experimental design information automatically or at least semi automatically into GenEx substantially simplify the pre-processing needed to prepare qPCR data for statistical analysis

Data pre-processing

The days when the ΔΔCq method was sufficient to analyze qPCR data are past. Not that we are calculating differently, but the studies performed have become larger and the experiments more complex. In fact, for most studies performed today it is not even possible to write a closed form expression to calculate the resulting expression response, rather the measured data must be processed sequentially to account for the various aspects of the experiment. In particular it is essential to correctly define the statistical unit (often refer to as a subject when organisms are used). Each unit should be associated with a single value for each variable for use of common statistical methods. This value must, however, frequently be assembled from various measurements, i.e. responses of target and reference genes, estimated amplification efficiencies, etc. In order to integrate the process into a logical workflow GenEx provides an intuitive wizard with the following sequential operations:

1. **Interplate calibration.** Many studies cannot be fitted in a single experimental run or for practical reasons have to be extended over time. qPCR instruments perform base-line correction and set threshold separately for each run, which introduces a bias between the Cq values measured in different runs. This bias can be compensated for by performing a common amplification in all plates, where one and the same sample is analyzed for a

given assay. This sample is called Inter-Plate Calibrator, IPC. Any variation in the measured Cq's of the IPC among runs reflects systematic variations due to instrument factors and should be compensated for. It is sufficient to run a single IPC for each channel used in the instrument if a common threshold and base-line correction is used. It is not recommended to perform separate inter-plate calibrations for each target, since every correction adds confounding variation to the data unimportant corrections shall be avoided, since they may impair data quality rather than improving it. For the same reason it is good strategy to use a robust sample for IPC and analyze it in replicates. In multiplate experiments the runs and the interplate calibrators shall be indexed in classification columns.

2. **Efficiency correction.** If PCR efficiency has been estimated the measured Cq values can be corrected to account for suboptimal amplification. Typically, PCR efficiencies are estimated from serial dilutions run separately and only once. The PCR efficiencies may then be listed in a classification row for automatic correction in GenEx.

3. **Normalize with spike.** PCR efficiency depends on the sample matrix. Usually it is assumed that the sample matrix and thus the PCR efficiency is constant. But occasionally there are variations, which can be tested for using an exogenous spike added to the samples. Differential expression of the spike between the test and a standard sample reflects the sample's specific inhibition and can to some degree be accounted for.

4. **Normalize to sample amount.** Measured Cq values depend on the amount of sample. This can be the sample volume processed, amount of RNA used for reverse transcription, or cell count. If sample amount vary data may have to be normalized. The sample amounts shall be indicated in a classification column.

5. **Average qPCR replicates.** If qPCR replicates are available, they shall be indexed in a classification column, and their Cq values shall be averaged.

6. **Correct for genomic DNA background.** When quantifying RNA levels using RT-qPCR the assays may also amplify genomic copies of the target. The amount of genomic background can be assessed by measuring either RT^- controls or by the ValidPrime approach, developed by Henrik Laurell at Inserm. The contribution to Cq from the genomic background can be calculated and the Cq values corrected.

7. **Normalization with reference genes.** In expression studies normalization to endogenous controls such as stably expressed reference genes is popular. Using GenEx we can normalize to any number of reference genes; we can even normalize sets of reporter genes to sets of reference genes to match the genes' properties such as expression levels, stabilities etc. It is possible to normalize to the mean expression of all the genes (global normalization). Optionally, reference genes can be indexed in a classification row for automatic processing. Normalization to the expression of reference genes corresponds to calculating ΔCq in the classical approach.

8. **Average technical replicates.** If additional technical replicates are available, such as RT, extraction and sampling replicates, they shall be indexed in classification columns and averaged.

9. **Normalize with Reference Sample(-s)**. In some paired designs systematic variation can be reduced by normalizing to a paired sample during pre-processing

10. **Normalize with Reference Sample(-s)**. In some paired designs systematic variation can be reduced by normalizing to a paired sample during pre-processing

11. **Relative quantities**. An arbitrary reference level is selected (which corresponds to $\Delta\Delta$Cq in the classical approach) and data are converted to linear scale ($2^{-\Delta\Delta Cq}$ in the classical approach). The reference level can be the most expressed sample, the least expressed sample, mean expression of all the samples, mean expression of a group of samples, or percentage (sum of the expression in all samples is set to 1). It is also possible to covert the ΔCq values to linear scale ($2^{-\Delta Cq}$).

12. **Convert to log scale**. For statistical analysis with parametric methods the data shall be converted to logarithmic scale. Options are $\log_2$, $\log_{10}$, ln, and log(X+1).

All the steps in the workflow are never needed, since some cancel the effect of other. The appropriate steps depend on the particular experiment, the controls and references that are available and the analysis that will be performed.

In addition to the pre-processing work flow, GenEx has correction for missing data. GenEx recognizes two types of missing data, random missing (failed experiments) and non-random missing (off-scale data). There is built in handling of random missing data among technical replicates, which are replaced based on available information in the course of the pre-processing. This is most useful, since the missing information can be ignored and is automatically accounted for. There are also several means to handle non-random missing or off scale data that are due to too low target amounts, which may bias the biological effect and invalidate the statistical inference in majority of statistical tests employed. Outliers in the data can be tested for based on standard deviation and Grubb's outlier test.

The pre-processing of data is logged and stored in a log file.

Screening by correlation
Several companies including Roche, Exiqon, Life Technologies, Lonza, Qiagen and TATAA Biocenter offer pre-plated assays for smooth expression profiling and screening purposes. Data from those plates are readily read into GenEx. Rarely are all assays relevant for every study and a strategy is to analyze a few representative samples of each kind in a pilot study to identify differentially expressed genes to be used is a larger downstream study. This is readily done using the GenEx scatter plot (Figure 3). Replicated measurements can be compared to test the reproducibility between plates (top left), or screen for differentially expressed genes under two conditions (top right, bottom).

Correlations between genes' expressions can be quantified by calculating the Pearson or Spearman correlation coefficients. This is typically applied to larger number of samples and has, for example, been used to reveal correlations between genes expressed in individual cells.[5]

Preparing the data for analysis
Groups for comparison are created using the GenEx Data Manager. Treatment groups or treatment factors in multifactorial studies, such as studies of the effect of gender or covariates

such as age, time or drug load, can be indexed in classification columns and used to assign subjects into groups automatically. The groups are assigned colors and symbols for plotting. A neat feature is that colors and symbols can be set independently, which makes it possible to assign subjects to multiple groups and identify these in plots by the shape, size and color of the symbol. Even shades of colors can be used creatively to indicate various levels of covariates at ordinary scale (e.g. darker shade indicates higher drug load).

In the data manager subjects and genes can be removed temporarily from analysis to compare results based on analyses of subsets of the data. Data can also be mean centered (subtraction of the mean value) or autoscaled (subtraction of mean followed by division with the standard deviation) to change the weights of the genes/samples in analyses. This is particularly useful in expression profiling analysis, where genes having different expression levels can be assigned equal weights. For analyses that apply models based on measured data, such as the standard curve, reverse calibration, limit of detection, neural networks etc., samples (and genes) can be assigned either training or test. Training data are used to create the model, which then is applied to classify the test samples. The various analyses available in GenEx are listed in Table 1.

Standard curve and reverse calibration
Amounts of pathogens in field samples can be quantified using qPCR by comparing the measured Cq's of the field samples with those of standard samples by means of a standard curve. Representative data arrangement is shown in Figure 4. In addition to the measured data the concentrations of the standards is given in a classification column. Additional classification columns can be used to index replicates and to identify the standard and test samples. Any technical replicates shall be averaged during pre-processing and the average Cq shall be considered a single data point. Independently prepared standards are treated separately, while replicates of field samples are averaged and used as a single more precise estimate. Groups are created in the GenEx Data Manager and assigned either test or training status. Samples can also be reversibly removed from analysis (Figure 5). A confidence level is also set for the analysis.

The standard curve is the best straight line fit of the Cq values measured for the standard samples to their concentration in logarithmic scale (Figure 6). It is calculated using linear regression and defines the intercept, which is the Cq expected for a sample containing a single template molecule, and the slope. From the slope the PCR efficiency is estimated. GenEx also calculates the uncertainties in the estimates of the slope and the intercept, which are reflected by the dashed lines in the plot as the Working-Hotelling area, and the confidence interval for the PCR efficiency (Figure 6). It is essential to calculate the confidence information, since it reflects the precision of the estimated efficiency. In this example the precision of the estimated efficiency is quite high, because a large number (21) of standards was used and a wide concentration range was covered (6 logs). In the literature frequently standard curves are based on substantially lower number of standards also covering less broad concentration ranges. The PCR efficiencies estimated from such standard curves are highly uncertain and any corrections they lead to become unreliable. There are even cases when separate standard curves are measured for the same assay in different plates and the estimated PCR efficiencies are used to adjust the two sets of data. Such adjustment is nonsense, since any difference between the two estimates of efficiency reflects is likely to be caused by random variation in the measurements rather than any systematic difference between the runs.

The residual plot shows the deviations of the standard samples' measured Cq's and their predicted Cq's by the standard curve (Figure 6). If the straight line standard curve is adequate to model the data residuals should fluctuate randomly along the fitted estimate, while if the model is inadequate runs of positive and negative residuals will be seen. GenEx performs a statistical test for the number of runs and if they are too few warns the linear standard curve may be impropriate model. Outliers are readily identified visually in the residual plot and GenEx further uses the Grubb's test to support the removal of outliers. In general no more than one outlier should be removed from a standard curve.[6] If multiple outliers are indicated the approach used is likely to be unstable and should be overseen. When replicates are available, the residual plot also reveals if noise increases at low concentrations.

A reliable standard curve is critical for accurate estimation of the concentrations of field samples, which in GenEx are referred to as test samples. The estimates improve if the field samples are available in replicates that can be averaged to reduce confounding variation. As usual replicates are indicated in a classification column. Concentrations of the unknown samples are estimated by entering the standard curve at the measured Cq and reading out the log concentration on the x-axis (Figure 7). The Working-Hotelling area, which here reflects the prediction uncertainty, is wider than before because of the additional error contribution from the measured Cq. GenEx calculates confidence intervals for the estimated concentrations from the Working-Hotelling area. The confidence intervals are symmetric around the mean in logarithmic scale, while they are asymmetric around the mean in linear scale (Figure 7). The uncertainty in the estimates is often much poorer than people think. Even though the standard curve in the example was based on 7 concentration standards covering 6 logs, each measured in triplicates for a total of 21 readings, and the assay had 96% efficiency, the uncertainty in the estimated concentrations of the unknowns is substantial. For example, for "Test 1", estimated concentration is 46700 copies, with the 95% confidence interval: 31000 – 61000 copies! With less accurate standard curve the precision in the estimated concentration would be even lower.

Limit of detection

The limit of detection (LOD) is "the lowest amount of analyte in a sample that can be detected with (stated) probability, although perhaps not quantified as an exact value".[7,8] For classical tests when a signal is measured against a background LOD is estimated from the standard deviation of the blank readout at the standard curve intercept. This approach is, however, not applicable to qPCR, which due to its real-time readout, gives no reading for a negative sample (Cq for a blank sample is formally infinity). Instead, for an analytical process that involves qPCR LOD can be estimated from multiple standard curves.[9] A minimum of six is recommended and concentrations around the expected LOD should be assessed. The measured data are transferred to binary format indicating positive and negative PCR's and the fraction of positive calls at each concentration is calculated. LOD is the concentration at which replicates are positive at the stated rates (eg., 95 %). GenEx fits the measured positive rates at different concentrations to estimates the LOD (Figure 8).

Selecting reference genes

With qPCR we measure the amount of target molecules in a sample rather than their concentration. A large sample is expected to contain more target molecules than a smaller and to compensate for the effect of size we must normalize. There are several options how to normalize. Popular option in gene expression analysis is to normalize with reference genes, since this should not only compensate for variation in sample amount, but also for variations in extraction yield and RNA quality. Genes needed for basic housekeeping functions were in

the early days of PCR thought to have stable expression and could serve as references. Although, they often do, experience has shown this is not always true, and before a gene is used as reference this assumption should be validated. The criteria for a good reference gene are that it has stable expression among samples and that its expression is invariant of treatment. Stability of expression is reflected by the standard deviation (SD) of biological replicates. However, we cannot just take a set of samples, measure the Cq's and calculate the SD's, because we do not know how to normalize the samples for this exercise. Of course, we can use the same amount of RNA in the analyses, but then, how do we know that all samples were extracted and reverse transcribed with the same yields, that they have the same mRNA/totalRNA ratio and the same RNA quality? Furthermore, if we would assume we can evaluate genes' expression stabilities based on samples normalized to the same amount of RNA, then we have already decided total RNA is the best norm. The gene selected based on SD measured on samples having the same amount of RNA will be the gene that shows a variation that correlates the most with that of total amount of RNA, and we may then as well normalize to the amount of RNA directly. If we suspect total amount of RNA is not the best norm, we have to identify optimum reference genes based on other criteria.

Appropriate approach to select reference genes is a special form of analysis of variance, which in qPCR literature is best known as the tool Normfinder.[10] Normfinder is applied to a panel of candidate reference genes that is analyzed in a set of representative samples. In essence, Normfinder calculates a global average expression of all the genes in all the samples, to which the individual genes are compared. Based on this comparison SD for each candidate reference gene is estimated. Furthermore, if the samples are from different treatment groups, Normfinder separates the variation into an intragroup and an intergroup contribution. Figure 9 shows an example where reference genes were sought for an obesity study in mice, where wild-type mice and an obese strain should be compared. The genes were selected from the TATAA reference gene panel,[11] which was measured on seven representative mice from each strain. The intragroup variation estimated is the SD of the genes' in the different treatment groups, while the intergroup variation is differential expression and sums to zero for every gene over all the groups. Good reference genes shall have low intergroup variation in all groups and negligible intergroup variation. Good strategy using Normfinder is to inspect the calculated intra and intergroup variations to identify any genes that appear regulated or exceedingly unstable and remove them from the data set using the GenEx Data Manager. Normfinder analysis is then repeated without considering the groups, since the remaining genes are not regulated. This produces more robust result with a single SD estimated for each genes, based on which the genes can be compared (Figure 10). The gene with lowest SD is the optimum reference gene. GenEx also calculates the accumulated SD expected if multiple reference genes are used for normalization. If we use larger number of reference genes SD will decrease because random variations among the genes' expressions will partially cancel, but it will increase as less good reference genes are being included. Altogether this leads to a minimum in the accumulated SD plot indicating the number of reference genes that give lowest SD (Figure 10). However, analyzing additional genes cost money, and one should consider consider the degree of improvement and the overall noise contribution from the reference genes when making decision. In the example largest improvement is observed when including a 2nd reference gene; adding additional reference genes the improvement is small. Furthermore, the noise contribution from the best reference gene is only 0.05 cycles and as little as 0.04 cycles when using the two optimum reference gene candidates. Considering that the repeatability of qPCR instruments rarely is below 0.1 cycle (estimated as SD of technical replicates) using more than one and definately more than two reference genes in this study will hardly improve the quality of the data appreciably.

Using Normfinder normalization with reference genes can be compared to normalization with total RNA by adding an extra column in the data sheet with the RNA concentrations per analyzed sample in logarithmic scale (Figure 11). The algorithm is ignorant to the nature of the variables, and compares their variations. For the expression data in the example measured in flash frozen biopsies from mouse brains, normalization with total RNA is essentially as stable as normalization with PPIA, which for this study is the optimum reference gene. Our experience is that for samples with high quality RNA, normalization with total amount of RNA is often robust normalizer many times as good as any reference gene; in samples of poor RNA integrity or when expression may have been induced, normalization with reference genes is preferred.

An older method to identify optimum reference genes that still is popular is geNorm. It uses the same input data as Normfinder, but it does not consider groups; all samples are treated as being from a single population. Then geNorm sequentially eliminates the gene that shows highest variation relative to all the other genes based on paired expression values in all the studied samples. The variability is reflected by a so called M-value (Figure 12). Because of the elimination process, geNorm cannot identify an optimum reference gene, but ends up by suggesting a pair of genes that shows high correlation and should be suitable for normalization. The M-value is related to the SD, but as calculated the M-values for the genes are based on different sample sizes they are not comparable. Further, as the comparison of any individually judged candidate gene is performed toward a plurality of genes, assumed to resemble most closely the anticipated stable behavior, it is prone to systematic failure where group of co-regulated instable genes may be involved in the analysis. Any such co-regulated complex of instable genes may dominate over the stable genes and hence point at deviant genes as candidates. Usually, the gene rankings by geNorm and Normfinder are similar, which is reassuring. Should the rankings differ there is reason to suspect the selection includes one or more regulated genes, and the result should be interpreted with caution.

Relative quantification

Treatment groups are readily compared visually in bar graphs using descriptive statistics (Figure 13) and statistical comparison is made using ANOVA (one factor, two or more levels) or 2-way ANOVA (two factors, two or more levels each) or, in the case of two groups, with either t-test (paired/unpaired, 1-tail/2-tail) or non parametric tests (Mann-Whitney, Wilcoxon). The difference between the groups is shown in either linear or logarithmic scale and the confidence interval is indicated (Figure 14). Note that the confidence intervals of the mean differential expression are asymmetric when data are shown in linear scale. When expression of many genes is compared GenEx offers several means to control for the false discovery rate due to multiple testing including Bonferroni, Benjamini-Hochberg, Westfall & Young, and Benjamini-Yekutieli (Figure 15).

Expression profiling

T-test and ANOVA are univariate methods that analyze the expression of every gene separately, effectively assuming that the genes are expressed independently of each other. This is rarely the case; genes' expressions tend to be correlated. This correlation can be exploited in the analyses using multivariate statistical methods. GenEx offers several unsupervised as well as a selection of supervised methods to classify samples and categorize genes based on expression profiles. Unsupervised methods classify samples and genes based on the measured profiles only. They include classical hierarchical clustering combined with heat map, which can be based on various clustering schemes including the Ward's algorithm

and several distance measures including the Euclidian distance and the magnitude of the Pearson correlation. While the Euclidian distance clusters genes based on similarities and consider upregualtion and downregulation being opposite, hence, anticorrelated, distance based on the magnitude of the Pearson correlation considers upregulation and downregulation to be correlated. The latter is useful to classify, for example, genes that show the same temporal response to treatment independently of the genes being up or downregulated. The clustering is visualized in a dendrogram, which in GenEx can be mirrored in every node. Mirroring in a node changes the visual appearance of the dendrogram, producing an equivalent mathematical solution. A small Self Organized Map (SOM) can be used to force classification into a defined small number of groups based expression similarities. SOM can also be used to validate a classification model, based on the distribution of samples/genes in a large MAP. Principal Component Analysis (PCA) groups samples/genes based on correlated expression in reduce space. Figure 15 shows example of hierarchical clustering, SOM and PCA of genes expressed during early development of the African claw frog Xenopus Laevis.[12] Hierarchical clustering of sample and genes can be combined showing also the measured intensities in a heat map (Figure 16). Also the appearance of the heat map can be changed in GenEx within equivalent mathematical solutions by mirroring the dendrograms' in nodes.

Supervised methods require a training set of samples with known classification; for example, negative and positive samples, or short, medium and long term survivals. A model is developed based on the training set that can be used to classify new data (in GenEx called test data). The procedure is equivalent to regression based on standard curve above, but is based on multiple genes and the model does not have to be linear. Supervised method available in GenEx include Partial Least Squares (PLA), which is used to calculate a single standard curve based on the expression of multiple genes that can be used to predict concentrations or other measures of test samples. Potential Curves is a variant for prediction of new data based on PCA, and Artificial Neural Network (ANN) and Support Vector Machines (SVM) are multivariate non-linear methods to classify samples. Logistic regression, Probit, Receiver Operating Characteristics (ROC), and Survival Analysis are soon also available.

Experimental design

Designing experiments is more difficult than analyzing results. In a well designed experiment confounding variation is minimized and the number of subject is sufficient to obtain conclusive results. Good strategy is to perform a fully nested pilot study before specifying the test protocol for a larger study (Figure 16).[13] Figure 17 shows the result of a nested pilot study wherein three heifers were studied by collecting three blood samples of each that were reverse transcribed in triplicates, and each cDNA was analyzed in triplicate using qPCR. Using nested Anova the variation arising from the different experimental steps can be estimated and expressed either as standard deviations or as variance contributions.[14] While β-actine and Caspase 3 show generally low standard deviations in all steps, Interleukin 1-β and Interferon-γ levels varied substantially among the heifers. For liver samples, the picture was very different, with large variation in the sampling step of most genes.[13] Knowing the costs associated with the different experimental steps the follow up study can be cost optimized. For example, for genes exhibiting SD = 0.1 cycle for the qPCR and RT steps, 0.2 cycles for the sampling/extraction step and SD = 1 cycle for the variation among the animals, assuming a cost of 1 unit for the qPCR, 3 units for the RT, 10 units for sampling/extraction and 100 unit for each animal, having a total budget of 1000 units, best we can do is to analyze 8 animals, sampling each animal once, performing RT in triplicates and qPCR in duplicates. The total standard error (SE) for this study is expected to be about 0.36 cycles (Figure 18). Using the

same tool we also estimate the SD among animals analyzed with samples collected in singleplex, RT analyzed in triplicate and qPCR in duplicate to 1.02 cycles. This can then be fed into a Power analysis to estimate the number of animals needed to be able to measure a particular difference with certain confidence and power. If we accept 5 % false positive rate (95 % confidence) and 5 % false negative rate (95% power) we construct a graph from which we read out, for example, that to measure 2-fold difference (Delta Cq of 1) we need some 15 animals (Figure 18).

Conclusions

GenEx is the first professional software that offers the full palette of methods for qPCR data analysis and qPCR experimental design. Of course, there are still methods on the wish list but a most can be done with those already present. The user interface is attractive and output graphics is of high quality. Help file is OK and the audiovisual tutorials based on example data are excellent.[15] A big plus is the free support forum, where hundreds of GenEx users share experiences and advise.[16] The important support from virtually all leading qPCR instrument manufacturers is also most valuable.

Future trends

qPCR is rapidly becoming broadly accepted platform, which leads to needs for standardization and quality control. This will require future software to be GLP, ISO, andTitle 21 CFR Part 11 compliant. We also expect followers to Roche and Exiqon that already today offer excellent wizards for transfer of data and experimental design information to qPCR analysis software. Integration with Laboratory Information Management Systems (LIMS) is also expected.

References

[1]www.multid.se
[2]www.integromics.com
[3]www.biogazelle.com
[4]www.exiqon.com/qpcr-software
[5]Anders Stahlberg, Daniel Andersson, Johan Aurelius, Maryam Faiz, Marcela Pekna, Mikael Kubista, Milos Pekny. Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. Nucl. Acids Res. 1–12, doi:10.1093/nar/gkq1182 (2010).
[6]EP6-A, Vol 23, no 16 (2003). Approved guideline NCCLS
[7] WHO. Glossary of terms for Biological Substances Used for Texts of the Requirements. Expert Committee on Biological Standardization. Document WHO-BS/95.173. Geneva, World Health Organization; 1995.
[8]EP17-A, Vol 24, no 10 (2004). Approved guideline NCCLS.
[9]M. Burns and H. Valdivia. Modelling the limit of detection in real-time quantitative PCR. European Food Research and Technology Volume 226, Number 6, 1513-1524, DOI: 10.1007/s00217-007-0683-z
[10]Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 2004 Aug 1;64(15):5245-50.
[11] http://www.tataa.com/products/Gene-Expression-Assays-and-Panels/View-all-products.html
[12]Anders Bergkvist, Vendula Rusnakova, Radek Sindelka, Jose Manuel Andrade Garda, Björn Sjögreen, Daniel Lindh, Amin Forootan, Mikael Kubista. Gene expression profiling – Clusters of possibilities. Methods 50:323–335(2010).
[13]Tichopad A, Kitchen R, Riedmaier I, Becker C, Stahlberg A, Kubista M. Design and Optimization of Reverse-Transcription Quantitative PCR Experiments. Clinical Chemistry 55:101816–1823 (2009).

5.3. Aplikace získaných poznatků na astrocyty

Optimalizovaná metoda sběru buněk, preamplifikace, qPCR analýzy na vysokokapacitním qPCR přístroji, analýzy dat a jejich biologická interpretace byla publikována v těchto pracích:

Distinct Expression/Function of Potassium and Chloride Channnels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice, Benešová J., **Rusňáková V.,** Honsa P., Pivoňková H., Dzamba D., Kubista M. and Anděrová M., PlosOne 7(1) (2012),e29725, IF=3,730.

Heterogeneity of astrocytes: From Development to Injury – Single Cell Gene Expression, Rusnakova V., Honsa P., Dzamba D., Stahlberg A., Kubista M. and Anderova M., PLoS One 8(8) (2013),e69734, IF=3,730.

Publikace Distinct Expression/Function of Potassium and Chloride Channnels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice (Benesova et al., 2012) byla první, která ukázala heterogenitu astrocytů v kůře mozkové pomocí RT-qPCR analýzy jednotlivých buněk izolovaných přímo bez následné kultivace.

V mozkové kůře GFAP/EGFP transgenních myší byly nalezeny pomocí konfokální trojdimenzionální morfometrie *in situ* dvě populace astrocytů, které mění odlišně svůj objem při nedostatku glukózy a kyslíku (Oxygen-glucose deprivation, OGD), (Benesova et al., 2009). Tyto populace byly rozděleny na silně odpovídající (high reponse, HR), které zvýší svůj objem během 40 minut až o 40 % a na slabě odpovídající (low reponse, LR), které změní svůj objem pouze o necelých 5 %. Pomocí qPCR jednotlivých buněk byl charakterizován expresní profil EGFP pozitivních astrocytů se zaměřením na iontové kanály a transportéry, u kterých se předpokládalo, že se účastní změny objemu astrocytů. Bylo testováno 15 genů a 157 buněk. Data byla analyzována tak, jak bylo uvedeno v kapitole 5.2. Stejně jako konfokální morfometrie i qPCR odlišila dvě subpopulace astrocytů lišící se v expresi draselného kanálu umožňujícího vstup draslíku *Kcnj10*, K2P draselných kanálů *Kcnk1* a *Kcnk2* a chloridového kanálu *Clcn2*. Jejich selektivní inhibice ukázala rozdílné chování HR a LR astrocytů. U LR astrocytů ve srovnání s HR astrocyty byla nalezena snížená exprese *Kcnj10* a *Kcnk1* a naopak zvýšená exprese *Kcnk2*. Žádný rozdíl nebyl nalezen v expresi glutamátových receptorů *Eaat1* a *Eaat2* a dalších draselných kanálů - *Kcnj16*, *Kcnj2*, *Kcnk10*. Na této publikaci jsem se podílela sběrem vzorků a jejich měřením a následnou statistickou analýzou.

Publikace *Heterogeneity of -Astrocytes: From Development to Injury – Single Cell Gene Expression* ukazuje použití výše uvedených metod při analýze vývoje astrocytů. Zabývá se jejich původem, vývojem

při dospívání zvířete i jejich chováním po poranění. Je to vůbec první publikace, která ukazuje využití analýzy jednotlivých buněk v takovém rozsahu.

Pro částečné objasnění otázky ohledně původu astrocytů byly analyzovány buňky mozkové kůry GFAP/EGFP myší o různém stáří - 10 (P10), 20 (P20), 30 (P30) a 50 (P50) dní starých a dále myší 3 (D3), 7 (D7) a 14 (D14) dní po ischemii. Pomocí vysokokapacitního qPCR byla změřena exprese 47 genů. Analýza dat bez biologického kontextu byla ukázána v kapitole 5.2. Samoorganizující se mapy a analýza hlavních komponent identifikovala během vývoje myší existenci tři hlavních subpopulací astrocytů. První subpopulace tvořená hlavně buňkami z P10 byla charakterizována velkým množstvím nezralých buněk a celkově vysokou genovou expresí všech měřených genů. Většina buněk exprimovala markery NG2 glií - *Pdgfrα* a *Cspg4*. U druhé a třetí skupiny byl postupně pozorován úbytek nezralých buněk a celkové snižování genové exprese. Buňky P50 byly již převážně neaktivní a mezi 50 buňkami byly nalezeny pouze tři, které exprimovaly markery NG2 glií. Stejným způsobem byly analyzovány buňky po zranění. Jednotlivé astrocyty se postupně aktivovaly a měnily svojí genovou expresi. Mimo známé markery astrogliózy jako je zvýšená exprese *Gfap*, Vimentinu, *Aqp1*, *Aqp9* a glutamátového metabotropního receptoru 5(*Grm5*) byly identifikovány i další geny se silně zvýšenou expresí v reaktivních astrocytech - hyperpolarizací aktivované kationtové kanály *Hcn1*, *Hcn2*, *Hcn3* a glutamátové receptory *Grin2a* a *Gria2*, *Gria3*. Analýza korelačních koeficientů ukázala význam genů *Hcn2* a *Gria3*. Téměř polovina reaktivních astrocytů exprimovala markery NG2 glií na úrovni mRNA, přesto se však nepodařilo najít buňku, která by exprimovala GFAP i Pdgfrα zároveň na úrovni proteinu.

Na této práci jsem se podílela sběrem buněk, jejich qPCR i statistickou analýzou, grafickým zpracováním dat a psaním publikace.

OPEN ACCESS Freely available online

PLoS one

Distinct Expression/Function of Potassium and Chloride Channels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice

Jana Benesova[1,3], Vendula Rusnakova[2], Pavel Honsa[1,3], Helena Pivonkova[1,3], David Dzamba[1,3], Mikael Kubista[2,4], Miroslava Anderova[1]*

1 Department of Cellular Neurophysiology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic, **2** Laboratory of Gene Expression, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic, **3** Second Medical Faculty, Charles University, Prague, Czech Republic, **4** TATAA Biocenter, Gothenburg, Sweden

Abstract

Recently, we have identified two astrocytic subpopulations in the cortex of GFAP-EGFP mice, in which the astrocytes are visualized by the enhanced green–fluorescent protein (EGFP) under the control of the human glial fibrillary acidic protein (GFAP) promotor. These astrocytic subpopulations, termed high response- (HR-) and low response- (LR-) astrocytes, differed in the extent of their swelling during oxygen-glucose deprivation (OGD). In the present study we focused on identifying the ion channels or transporters that might underlie the different capabilities of these two astrocytic subpopulations to regulate their volume during OGD. Using three-dimensional confocal morphometry, which enables quantification of the total astrocytic volume, the effects of selected inhibitors of K^+ and Cl^- channels/transporters or glutamate transporters on astrocyte volume changes were determined during 20 minute-OGD *in situ*. The inhibition of volume regulated anion channels (VRACs) and two-pore domain potassium channels (K_{2P}) highlighted their distinct contributions to volume regulation in HR-/LR-astrocytes. While the inhibition of VRACs or K_{2P} channels revealed their contribution to the swelling of HR-astrocytes, in LR-astrocytes they were both involved in anion/K^+ effluxes. Additionally, the inhibition of Na^+-K^+-Cl^- co-transporters in HR-astrocytes led to a reduction of cell swelling, but it had no effect on LR-astrocyte volume. Moreover, employing real-time single-cell quantitative polymerase chain reaction (PCR), we characterized the expression profiles of EGFP-positive astrocytes with a focus on those ion channels and transporters participating in astrocyte swelling and volume regulation. The PCR data revealed the existence of two astrocytic subpopulations markedly differing in their gene expression levels for inwardly rectifying K^+ channels (Kir4.1), K_{2P} channels (TREK-1 and TWIK-1) and Cl^- channels (ClC2). Thus, we propose that the diverse volume changes displayed by cortical astrocytes during OGD mainly result from their distinct expression patterns of ClC2 and K_{2P} channels.

Citation: Benesova J, Rusnakova V, Honsa P, Pivonkova H, Dzamba D, et al. (2012) Distinct Expression/Function of Potassium and Chloride Channels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice. PLoS ONE 7(1): e29725. doi:10.1371/journal.pone.0029725

Editor: Michal Hetman, University of Louisville, United States of America

Received October 10, 2011; **Accepted** December 2, 2011; **Published** January 11, 2012

Copyright: © 2012 Benesova et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was supported by the grants: CZ:GA CR:GA P303/10/1338 and CZ:GA CR:GA 309/08/H079 from the Grant Agency of the Czech Republic and UEM AV CR – AV0Z50390703 from the Academy of Sciences of the Czech Republic. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have read the journal's policy and have the following conflicts: Mikael Kubista is employed by TATAA Biocenter (TATAA Biocenter AB, Göteborg, Odinsgatan 28, 411 03 Göteborg, Sweden). This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: anderova@biomed.cas.cz

Introduction

The generation of brain edema, ultimately leading to an intracranial pressure increase and brain herniation, markedly alters the course and treatment of cerebral ischemia. Among brain cell types, astrocytes have been reported to contribute predominantly to edema formation. They support neuronal function and provide for the maintenance of ionic and neurotransmitter homeostasis under physiological conditions; however, they may also contribute to the worsening of brain damage under pathological conditions. The uptake of ions and amino acids may turn into an excessive osmolyte influx leading to the generation of edema, which in turn contributes to further ischemic damage. Several transporters and ion channels have been described as participating in astrocytic swelling (for a review, see [1]). Astrocytic Na^+-dependent glutamate uptake, which is responsible for the termination of synaptic transmission and helps to protect neurons from the excitotoxic activity of glutamate, may reverse as a consequence of the ionic dis-balance accompanying energy depletion during severe ischemia and thus contribute to an increase of extracellular glutamate concentrations [2]. Additionally, K^+-Cl^- (KCC) and Na^+-K^+-Cl^- (NKCC) co-transporters, the major regulators of neuronal and astrocytic Cl^- and Na^+ gradients, have been found to participate in cell volume changes under pathological conditions. Increased NKCC activity has been shown to result in Na^+, K^+ and Cl^- accumulation, thus markedly contributing to astrocytic swelling [3], while KCCs, which are responsible for ion efflux, are rather involved in regulatory volume decrease (RVD) [4]. Nonetheless, a high extracellular concentration of K^+ ($[K^+]_o$) may induce the reverse operation of KCCs and thus contribute to cell swelling, as described in retinal Müller cells [5]. Furthermore, the depolarization of the astrocytic membrane

due to Na^+/K^+ ATP-ase inhibition results in excessive K^+ and Cl^- influx via ion channels [6]. Despite the fact that a prominent role in K^+ buffering has been suggested for the astrocytic Kir4.1 channel [7,8], two-pore domain K^+ (K_{2P}) channels, especially TWIKs and TREKs, have been shown to be responsible for glial K^+ uptake [9], and their contribution to glial volume homeostasis was described in Müller cells [10]. The activation of K_{2P} channels in response to ischemia has been demonstrated *in vitro* by Buckler and Honore [11] and the up-regulation of TREK-2 channels in conjunction with extracellular glutamate clearance was demonstrated in cultured rat astrocytes during anoxia/hypoglycemia by Kucheryavykh and co-authors [12]. Concurrently, astrocytic swelling is accompanied by RVD, which is also carried out by Cl^- and K^+ channels [13]. The main anion channels responsible for Cl^- and organic osmolyte efflux are volume-regulated anion channels (VRACs), which play a predominant role in RVD. However, an anion efflux via VRACs has also been shown to contribute to further ischemic brain damage [14,15] and moreover, VRACs have been reported to provide the anion influx that results in neuronal swelling [16].

In our recent *in situ* studies we have demonstrated that cortical astrocytes do not respond uniformly to hypoosmotic stress [17,18] and that two differently responding groups of astrocytes high response- (HR-) astrocytes and low response- (LR-) astrocytes are present in the cortex of GFAP/EGFP mice [19] based on the volume changes evoked by oxygen-glucose deprivation (OGD) [20]. In the present study, we aimed to elucidate the basic mechanisms potentially underlying the different ability of the two astrocytic populations to regulate their volume in response to OGD. We have focused on the role of chloride and potassium channels/transporters and excitatory amino acid transporters in astrocytic swelling during OGD. For the quantification of cell volume, three-dimensional confocal morphometry has been employed. Moreover, we have used real-time, single-cell quantitative PCR (qPCR) to explore the gene expression profiles of cortical GFAP/EGFP astrocytes with a focus on those ion channels and transporters participating in astrocyte swelling and volume regulation. We have used a similar approach to that of Stahlberg and co-authors [21]; they have identified astrocytic subpopulations in the mouse brain using single-cell qPCR profiling.

Materials and Methods

Ethics Statement

All procedures involving the use of laboratory animals were performed in accordance with the European Communities Council Directive 24 November 1986 (86/609/EEC) and animal care guidelines approved by the Institute of Experimental Medicine ASCR Animal Care Committee on April 17, 2009; approval number 85/2009.

Confocal 3D morphometry

Solutions. Artificial cerebrospinal fluid (ACSF) solution contained (in mM): NaCl 122.0, KCl 3.0, $CaCl_2$ 1.5, $MgCl_2$ 1.3, Na_2HPO_4 1.25, $NaHCO_3$ 28.0, D-glucose 10.0 (pH 7.4, in 95% O_2/5% CO_2). Osmolality was confirmed to be 300±5 mOsmol/kg with a vapor pressure osmometer (Vapro 5520, Wescor Inc., Logan, USA). Oxygen glucose deprivation (OGD) was achieved by saturating glucose-free ACSF with 5% O_2, 5% CO_2 and 90% N_2 ($ACSF_{OGD}$). All inhibitors and their concentrations are listed in Table 1. DMSO, ethanol and methanol, which were used for the preparation of inhibitor stock solutions, did not affect astrocyte volume changes at their final

Table 1. List of used inhibitors.

	c (μM)	Targeted proteins in astrocytes
Transporters		
DIOA	100**	KCC [4]
[(dihydroindenyl)oxy]alkanoic acid		
Bumetanide	100**	NKCC [48]
3-(Aminosulfonyl)-5-(butylamino)-4-phenoxybenzoic acid		
DL-TBOA	100*	EAAT1, 2 [53]
DL-threo-b-benzyloxyaspartic acid		
Cl channels		
NPPB	100*	Cl channels, VRACs [24]
5-Nitro-2-(3-phenylpropylamino)benzoic acid		
Tamoxifen	30***	VRACs [24]
(Z)-1-(p-Dimethylaminoethoxyphenyl)-1,2-diphenyl-1-butene		
DCPIB	30*	VRACs [24]
4-(2-butyl-6,7-dichloro-2-cyclopentylindan-1-on-5-yl)oxybutyric acid		
K^+ channels		
$BaCl_2$	100	Kir channels [26]
$BaCl_2$	1000	Kir, K_{2P} channels [27]
Quinine	200**	K2P channels [27]

KCC – K^+-Cl co-transporter, NKCC – Na^+-K^+-Cl co-transporter, EAAT – excitatory amino acid transporter, VRAC – volume-regulated anion channel, K_{2P} – two-pore domain potassium channel, Kir - inwardly rectifying potassium channel.
Stock solutions were dissolved at 1,000× final concentration in DMSO (*); ethanol (**) or methanol (***). $BaCl_2$ was dissolved at 1 mM concentration in double-distilled H_2O (ddH_2O).
doi:10.1371/journal.pone.0029725.t001

concentrations (data not shown). The majority of chemicals, including inhibitors, were purchased from Sigma-Aldrich (St. Louis, MO, USA). DL-TBOA was purchased from Tocris Bioscience (Bristol, UK).

Brain slice preparation. Quantification of astrocyte volume changes was carried out in acute brain slices of GFAP/EGFP mice [19]. Thirty to forty-day-old male/female mice were anesthetized with isoflurane followed by decapitation. Brains were dissected out and placed into ice-cold oxygenated ACSF. Transverse slices (400 μm) were cut using an HM650V vibratome (MICROM International GmbH, Waldorf, Germany). The slices were kept in continuously oxygenated ACSF, at room temperature (23 25°C), for at least 1 hour before starting experiments. To estimate the effect of each inhibitor on astrocytic swelling evoked by OGD, a single astrocyte was recorded in one slice and slices from 2 3 mice were used for each inhibitor.

Cell volume measurement and quantification. Changes in total astrocytic volume were measured and quantified using previously described method [17,18,20]. The slices were perfused with ACSF or $ACSF_{OGD}$ (perfusion rate ~5 ml/min). To examine the effect of inhibitors on astrocyte volume changes, they were applied during 40-minute OGD. All experiments were performed at room temperature (23 25°C). For cell volume quantification (Fig. S1) we have selected only brightly fluorescent cells at an approximate distance of ~30 μm from the slice surface. The astrocytes were recorded as a set of two-dimensional (2D) sectional images with a resolution of 1024×1024 pixels using a Leica TCS SP system confocal microscope (Leica Germany) with a Leica 40× water immersion objective (0,8) HCX APO (Leica, Germany). EGFP was excited by an Ar laser set at 488 nm, and the emitted signal was recorded over the range of 510 552 nm using a TD488/543/633 filter. Each three-dimensional (3D) image of the cell was sectioned into 70 80 consecutive 2D images at a uniform spacing of 1 μm. Data were collected at 10 minute intervals. Image processing and morphometric measurements were performed using the program CellAnalyst developed at the Department of Cellular Neurophysiology, Institute of Experimental Medicine, Prague, Czech Republic [18].

Statistics. Changes in total astrocyte volume are presented as the mean ± SEM. Statistical significance was determined by a two-tailed unpaired Student's *t*-test. Differences between the groups were considered statistically significant when $p<0.05$, very significant $p<0.01$, and extremely significant $p<0.001$.

Single-cell gene expression profiling

Preparation of cell suspensions from the cortex of GFAP/EGFP mice. GFAP/EGFP transgenic mice (30 days old) were deeply anesthetized with sodium pentobarbital (PTB, 100 mg/kg, i.p.), and perfused transcardially with cold (4 8°C) isolation buffer containing (in mM): NaCl 136.0, KCl 5.4, Hepes 10.0, glucose 5.5, osmolarity 290±3 mOsmol/kg. The forebrain was isolated by the removal of the olfactory lobes, cerebellum, and midbrain/hindbrain structures by dissection. To isolate the cerebral cortex, the brain was sliced in 1 mm coronal sections using a vibrating microtome HM650V (MICROM International GmbH, Walldorf, Germany), and the cerebral cortex was carefully dissected away from the ventral white matter tracks. The tissue was incubated with continuous shaking at 37°C for 90 minutes in 5 ml of a papain solution (20 U/ml) and 0.2 ml DNAase (both Worthington, Lakewood, NJ) prepared in isolation buffer. After papain treatment the tissue was mechanically dissociated by gentle trituration using a 1 ml pipette. Dissociated cells were layered on top of 5 ml of Ovomucoid inhibitor solution (Worthington) and harvested by centrifugation (140× g for 6 minutes). This method routinely yielded ~1.5 2 million cells per mouse forebrain.

Collection of single EGFP-positive cells. Dissociated cells were sorted using flow cytometry (BD FACSCalibur); they were kept on ice throughout the sorting procedure. Cell aggregates were removed by filtering with a 30 nm cell strainer (Becton Dickinson, USA). The flow cytometry instrument was manually calibrated to deposit single cells in the centre of each collection tube. Propidium iodide was added to the suspension of cells for checking viability. Single cells were collected into 96-well plates (Life Technologies, Czech Rep.) containing 5 μl nuclease-free water with bovine serum albumin (1 mg/μl, Fermentas, Biogen, Czech Rep.) and 20 U RNAseOut (Life Technologies, Czech Rep.) per well. The plates were then placed on a pre-cooled rack. The astrocytes were collected based upon their positivity for EGFP and their viability. Plates with collected cells were immediately frozen at −80°C.

cDNA synthesis. A modified protocol of A. Stahlberg [21] was used for cDNA synthesis. SuperScript III RT (Life Technologies, Czech Rep.) was used for reverse transcription. Lysed single cells in 5 μl water containing 0.5 mM dNTP (Promega, Germany), 1.0 mM oligo(dT15) (Invitrogen) and 1.0 mM random hexamers (Invitrogen) were incubated at 70°C for 5 min. 50 mM Tris HCl, 75 mM KCl, 3 mM $MgCl_2$, 5 mM dithiothreitol, 20 U RNaseOut and 100 U SuperScript III (all Invitrogen; final concentrations) were added to a final volume of 10 μl. Reverse transcription was performed at 25°C for 5 min, 50°C for 60 min, 55°C for 10 min and terminated by heating to 70°C for 15 min. All samples were diluted to 40 μl with water before qPCR.

Primer design and optimization of assays. Primers were designed using BeaconDesigner software (version 7.91, Premier Biosoft International). All primers except those for *Cspg4* and *Nkcc1* were designed to span an intron to avoid amplification of genomic DNA. BLAST (Basic Local Alignment Search Tool) searches revealed no pseudogenes. The primer sequences are shown in Table S1. All assays of single cells were optimized so as to not generate primer dimers before cycle 45, to have a PCR efficiency of at least 90%, and to amplify all known splice forms documented by the National Center for Biotechnology Information (NCBI). Calibration curves with purified PCR products (QIAquick PCR Purification Kit; Qiagen, Germany) were used to establish the linearity of the assays. The formation of the correct PCR products was confirmed by electrophoresis on 20 g/L agarose gels for all assays and by melting-curve analysis of all samples. Five individual cells per assay were tested, and no genomic DNA amplification was observed.

qPCR. A Biorad CFX384 (Biorad, Czech Rep.) was used for all qPCR measurements. To each reaction (10 μl) containing iQ SYBR Green Supermix (BioRad) and 300 nM of each primer (EastPort, Czech Rep.), we added 3 μl diluted cDNA. The temperature profile was 95°C for 3 min followed by 50 cycles of amplification (95°C for 15 s, 60°C for 15 s and 72°C for 20 s). All samples were analyzed by melting curve analysis.

Data processing and statistics. For all assays the limit of detection (LOD) was assumed to be the highest Cq (quantification cycle) value measured reflecting the expected product based on melt curve analysis. All Cq values that were above the limit of detection and hence were reflecting the formation of aberrant products, and any missing data, were replaced by the Cq at the limit of detection +1. The correction was performed separately for each gene and effectively corresponds to assigning a concentration to the off-scale measurements that is 50% of the concentration we reliably detect. Cq values were converted to relative quantities and transformed to a log scale [22].

Dendrograms, principal component analysis (PCA) and Kohonen self-organizing maps were calculated using GenEx software (ver. 5.3, MultiD, Sweden). The expression of each gene was mean-centred for profiling analysis with multivariate methods. Dendrograms were calculated using Ward's algorithm and Euclidean distance measure. Spearman correlation coefficients were calculated for all gene pairs. Because of the significant number of off-scale data that were corrected, the distribution of data was not necessarily Gaussian, and non-parametric statistical tests were used to assess significance. The expression levels measured in groups of cells were compared using the non-parametric Kruskal-Wallis test with Dunn's post-test. For Kohonen self-organizing maps (SOMs) and potential curve analysis, autoscaled expression values were used to give all genes equal weight in the clustering algorithms. The parameters used for the Kohonen SOMs were: 3x1 map, 0.10 learning rate, 2 3 neighbors and 10000 iterations. The calculated clusters did not depend on the parameters' settings.

Results

Astrocyte heterogeneity has been described previously as morphological differences, distinct functional properties and variations in the expression of receptors, ion channels and transporters [23]. In our previous study [20] we described two astrocytic subpopulations in the cortex of GFAP/EGFP mice differing in their volume changes induced by OGD. One subpopulation, referred to as high response-astrocytes (HR-astrocytes), markedly increased their total cell volume when exposed to OGD, while only small volume changes were evoked in the second subpopulation, referred to as low response-astrocytes (LR-astrocytes). Both subpopulations were morphologically identical, displayed passive currents and expressed EAAT1 and EAAT2 transporters. Here, we focus on elucidating the differences between these two subpopulations that might originate from the diverse functioning/expression of ion channels and transporters that are involved in astrocytic swelling and cell volume regulation. We have employed a wide range of transporter- and ion channel-inhibitors (Table 1) to study their effects on astrocyte volume changes *in situ* evoked by OGD.

Astrocyte swelling evoked by oxygen-glucose deprivation

To examine the effect of inhibitors on acute astrocytic swelling, brain slices were exposed to 40-minute OGD. The values of total astrocyte volume obtained during 40-minute OGD in the absence of any inhibitor were used as control values and compared to those obtained in the presence of inhibitors. The astrocyte volume at $t = 0$ was set to 100%, and the volume changes were expressed relative to this baseline as an increase in percentage. HR- and LR-astrocytes were classified using the same criteria as described previously [20]. Briefly, if the total cell volume increase was less than 10% (LR-astrocytes), the relative changes in total volume, cell soma volume and the volume of the processes were similar. If the total cell volume increase was greater than 10% (HR-astrocytes), the swelling of the cell processes exceeded the swelling of the soma. The total cell volume (V) of HR-astrocytes ($n = 13$) increased to $122.1 \pm 1.4\%$ after 20 minutes and to $135.9 \pm 2.6\%$ after 40 minutes of OGD. LR-astrocytes ($n = 8$) increased their volume to $108.4 \pm 0.8\%$ after 20 minutes and to $117.6 \pm 2.1\%$ after 40 minutes of OGD (Fig. 1, 2, 3 top). The effects of inhibitors were studied as follows: brain slices were first exposed to OGD alone for 20 minutes, followed by a 20 minute application of $ACSF_{OGD}$ together with an inhibitor (Fig. S2). Using this approach, an individual cell could be classified as either an HR- or LR- astrocyte based on its volume changes during the first 20 minutes. In order to evaluate the effect of inhibitors on astrocytic swelling, volume changes were evaluated in each cell individually and expressed as a percent cell volume increase/decrease related to the volume obtained after the first 20 minutes of OGD, which was set as 0% (Fig. 1, 2, 3 bottom).

Inhibition of glutamate and K^+-Cl transporters reduces the astrocytic swelling induced by OGD in both subpopulations, while inhibition of Na^+-K^+-Cl co-transporters affects only HR-astrocytes

The disruption of ionic and excitatory amino acid (EAA) homeostasis that occurs during cerebral ischemia results in enhanced glutamate/K^+ uptake by astrocytic transporters, leading to marked volume changes in astrocytes during ischemia [1]. In order to determine whether these transporters underlie the differences in volume changes between HR-/LR-astrocytes, we tested the effect of DL-TBOA, an inhibitor of the excitatory amino acid transporters EAAT1 and EAAT2, bumetanide, an inhibitor of Na^+-K^+-Cl^- co-transporters (NKCCs), and DIOA, an inhibitor of K^+-Cl^- co-transporters (KCCs), on the astrocyte swelling evoked by OGD (for concentrations see Table 1). The application of DL-TBOA led to a significant volume decrease in both astrocytic subpopulations (Fig. 1A); however, the course of inhibition was slightly different. While HR-astrocyte swelling remained unaffected during the first 10 minutes, it decreased by 6.9% after 20 minutes of $ACSF_{OGD}$ and DL-TBOA co-application ($n = 6$). In LR-astrocytes swelling was reduced already after 10 minutes, and after 20 minutes the swelling was reduced by 10.4% ($n = 6$). Similarly, the inhibitor of KCCs - DIOA - inhibited cell swelling in both astrocyte groups (Fig. 1B). A 20 minute application of $ACSF_{OGD}$ and DIOA reduced cell swelling by 7.1% in HR-astrocytes ($n = 9$) and by 10.2% in LR-astrocytes ($n = 9$). On the other hand, the inhibition of NKCCs by bumetanide led to the reduction of swelling in HR-astrocytes (by 7.0% after 20 minute $ACSF_{OGD}$ and bumetanide co-application; $n = 10$), while in LR-astrocytes the application of bumetanide did not significantly alter the extent of cell swelling ($n = 10$; Fig. 1C). In summary, the ion influx via NKCCs significantly contributed to HR-astrocyte swelling during OGD, while it was not involved in the volume changes observed in LR-astrocytes.

Inhibition of volume-regulated anion channels or two-pore domain K^+ channels during OGD reveals differences between HR- and LR-astrocytes

Since in many cell types, including astrocytes, VRACs are the predominant channels activated by swelling [13], we have examined the effects of three VRAC inhibitors NPPB, tamoxifen and DCPIB (for concentrations see Table 1). A non-selective inhibitor of anion channels, NPPB had an opposite effect on the two astrocytic subpopulations (Fig. 2A). The swelling of HR-astrocytes was decreased by 11.6% after 20 minutes of $ACSF_{OGD}$ and NPPB co-application ($n = 6$), while in LR-astrocytes the swelling was first reduced by 4.2% after 10 minutes and then it rather increased ($n = 9$). Similarly, tamoxifen and DCPIB also had a diverse effect on HR-/LR-astrocytes. Tamoxifen reduced the HR-astrocyte swelling by 6.2% after 20 minutes of its application ($n = 10$) and the DCPIB application led to the swelling reduction of 9.0% after 20 minutes ($n = 8$). On the other hand, the LR-astrocyte swelling was not appreciably affected by Tamoxifen (Fig. 2B), whereas DCPIB application led to their additional swelling (Fig. 2C). Moreover, the block of HR-astrocyte swelling by NPPB was more effective than that induced by the other

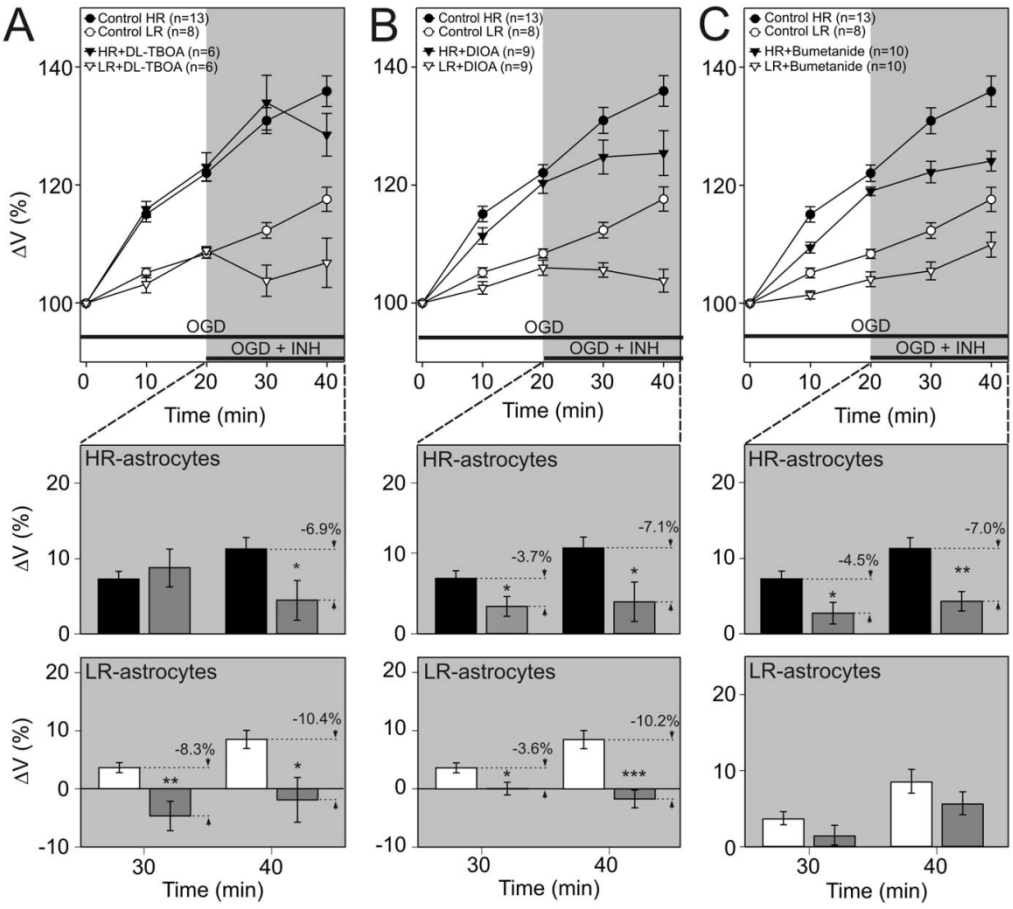


Figure 1. Inhibitors of glutamate transporters and K^+-Cl^- co-transporters reduce the OGD-induced swelling in both astrocytic subpopulations. The effect of 100 μM DL-TBOA, an inhibitor of excitatory amino acid transporters (**A**), 100 μM DIOA, an inhibitor of K^+-Cl^- co-transporter (**B**), and 100 μM bumetanide, an inhibitor of Na^+-K^+-Cl^- co-transporter (**C**). **A–C top:** Time course of volume changes in HR-astrocytes (filled circles) and LR-astrocytes (empty circles) during 40-minute OGD (control) and during 20-minute OGD followed by 20-minute OGD with the application of an inhibitor in HR-/LR-astrocytes (filled/empty triangles). **A–C bottom:** The effect of inhibitors was evaluated in each individual cell and expressed as the percent cell volume increase/decrease related to the maximal volume after 20-minute OGD, which was set as 0%. Note that the application of DL-TBOA and DIOA led to a swelling reduction in both HR- and LR- astrocytes, while bumetanide only affected the swelling in HR-astrocytes. Asterisks indicate significant differences from controls ($p<0.05$ (*, significant), $p<0.01$ (**, very significant), $p<0.001$ (***, extremely significant)).
doi:10.1371/journal.pone.0029725.g001

inhibitors, which is in accordance with the lower specificity of the inhibitor as reported previously by Abdullaev and co-authors [24], suggesting that more types of chloride channels are involved in HR-astrocyte swelling. In summary, DCPIB, which is currently the most selective inhibitor of VRACs known, reduced the swelling induced by OGD in HR-astrocytes, while in LR-astrocytes this inhibitor caused their additional swelling.

In view of previous findings that the main K^+ channels expressed in mature astrocytes are Kir4.1 and two-pore domain channels [25,26], we have examined the effect of $BaCl_2$ and quinine (for concentrations see Table 1) on the volume changes of HR-/LR-astrocytes during OGD. The application of 100 μM $BaCl_2$, which has been shown to preferentially inhibit Kir channels [26], led to an additional swelling after 20 minutes of OGD in both subpopulations of astrocytes. The swelling increased by 6.6% in HR-astrocytes (n = 8) and by 7.7% in LR-astrocytes (n = 10; Fig. 3A). Conversely, 1 mM $BaCl_2$, used as a blocker of both Kir and K_{2P} channels [27], completely inhibited the swelling of HR-astrocytes (9.9% reduction after 20 minutes; n = 7), while in LR-astrocytes it rather led to an additional swelling (not significant, n = 13; Fig. 3B). Equally, the application of quinine slightly reduced the swelling of HR-astrocytes (not significant; n = 5), while markedly increased the swelling of LR-

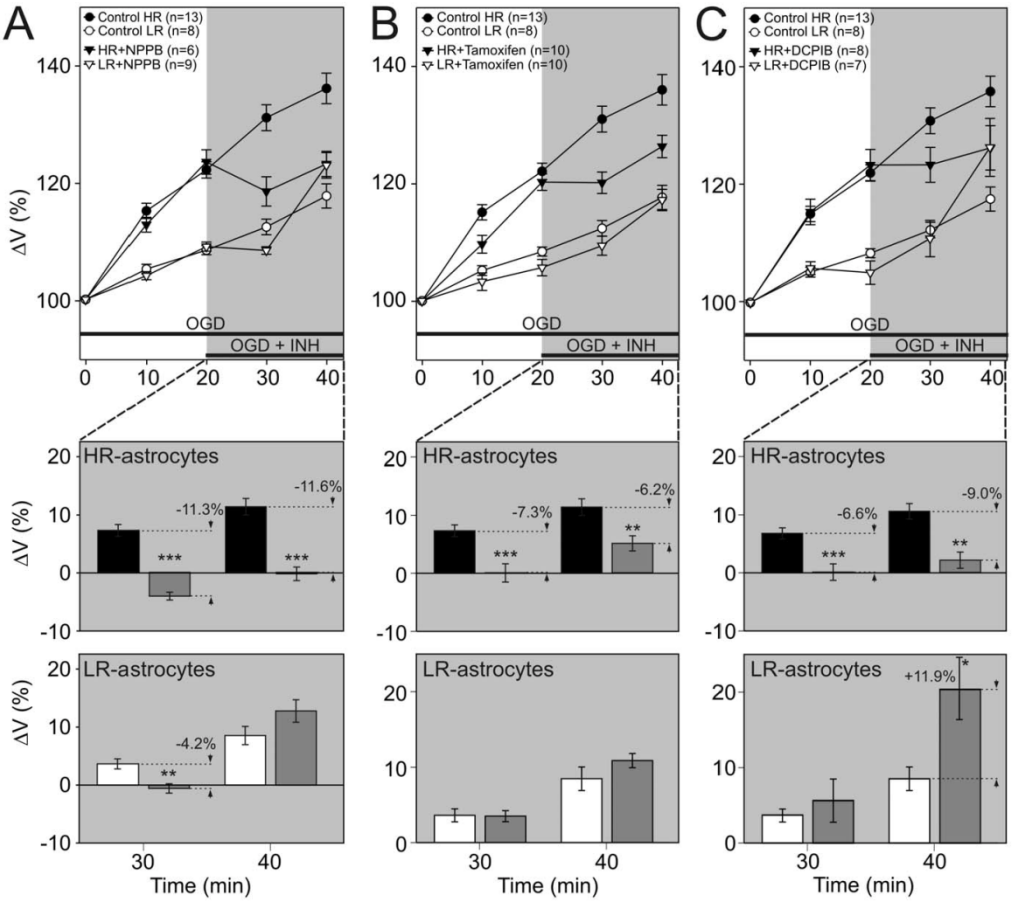


Figure 2. Inhibitors of Cl^- channels differently affect OGD-induced swelling in HR-and LR-astrocytes. The effect of 100 μM NPPB, a non-specific inhibitor of chloride channels (**A**), 30 μM Tamoxifen (**B**) and 30 μM DCPIB (**C**), inhibitors of volume regulated anion channels. **A–C top:** Time course of volume changes in HR-astrocytes (filled circles) and LR-astrocytes (empty circles) during 40-minute OGD (control) and during 20-minute OGD followed by 20-minute co-application of $ACSF_{OGD}$ plus an inhibitor in HR-/LR-astrocytes (filled/empty triangles). **A–C bottom:** The effect of the inhibitors was evaluated in each individual cell and expressed as the percent cell volume increase/decrease related to the maximal volume after 20-minute OGD, which was set as 0%. Note that in HR-astrocytes the application of VRAC inhibitors (Tamoxifen and DCPIB) reduced the swelling induced by OGD, while in LR-astrocytes the application of DCPIB resulted in an additional swelling. Asterisks indicate significant differences from controls ($p<0.05$ (*, significant), $p<0.01$ (**, very significant), $p<0.001$ (***, extremely significant)).
doi:10.1371/journal.pone.0029725.g002

astrocytes by 11.5% after 20 minutes (n = 7; Fig. 3C). In conclusion, the inhibition of Kir channels revealed that they contribute predominantly to the efflux of K^+ in both astrocytic subpopulations. On the other hand, the inhibition of K_{2P} channels disclosed their participation in K^+ uptake in HR-astrocytes, while in LR-astrocytes they rather contribute to K^+ efflux.

The presence of two distinct populations is not linked to gender or location in the cortical layers

Since the sensitivity of astrocytes to OGD might be influenced by their location within the CNS, we quantified the incidence of HR-/LR- astrocytes based on their position in the cortical layers. The distribution of astrocytes between the two groups in the different cortical layers was as follows: I layer 46% HR, 54% LR (n = 46), II III layers 41% HR, 59% LR (n = 113), IV VI layers 48% HR, 52% LR (n = 133). Although females have been shown to be more resistant to ischemic injury than males in clinical studies as well as in animal models [28], gender does not underlie the existence of two astrocytic groups with a different sensitivity to ischemic conditions. In females, 45% of measured cells were classified as HR- and 55% as LR-astrocytes (total number of cells n = 193); similarly in males, 49% of cells belonged to HR- and 51% to LR- astrocytes (n = 168). In conclusion, neither the location of astrocytes in the cortical layers, nor the gender of the

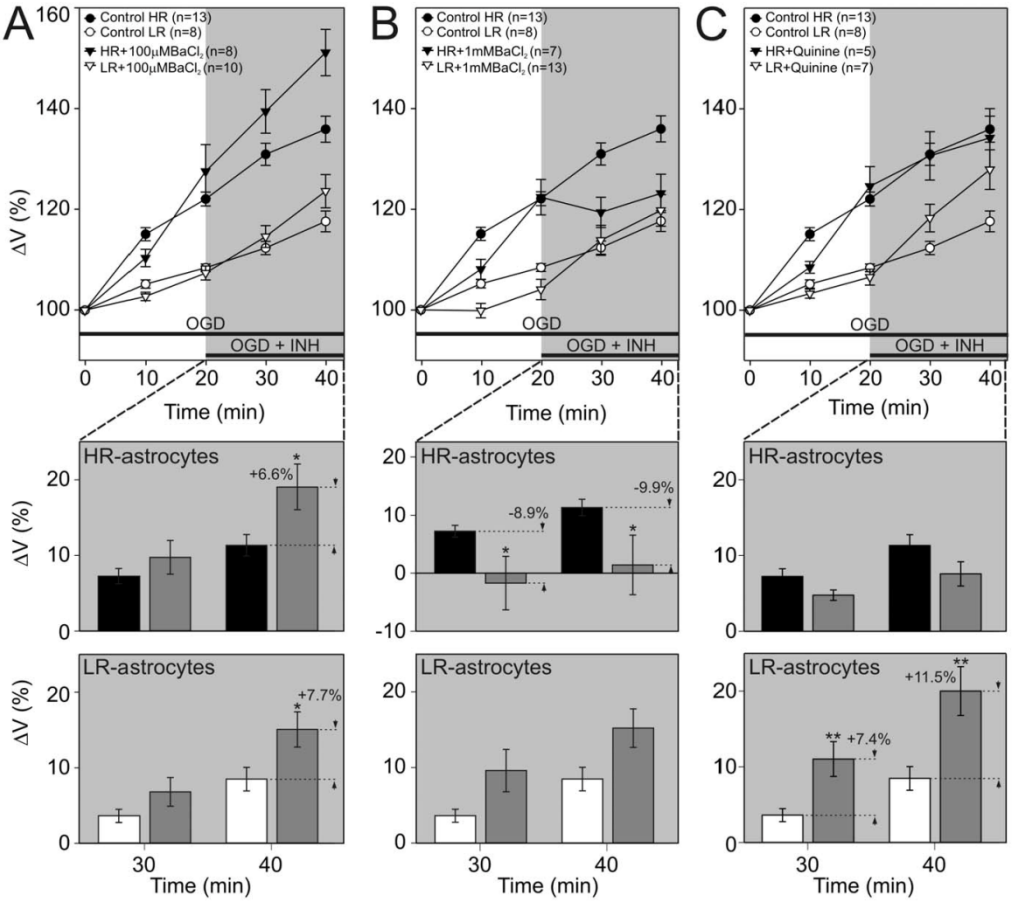


Figure 3. Inhibitors of K^+ channels have an opposite effect on HR-/LR-astrocytes when applied during OGD. The effect of 100 μM $BaCl_2$ (**A**), an inhibitor of inwardly rectifying potassium channels, 1 mM $BaCl_2$ (**B**) and 200 μM Quinine (**C**), inhibitors of two-pore domain potassium channels. **A–C top:** Time course of volume changes in HR-astrocytes (filled circles) and LR-astrocytes (empty circles) during 40-minute OGD (control) and during 20-minute OGD followed by 20-minute co-application of $ACSF_{OGD}$ plus an inhibitor in HR-/LR-astrocytes (filled/empty triangles). **A–C bottom:** The effect of inhibitors was evaluated in each individual cell and expressed as the percent cell volume increase/decrease related to the maximal volume after 20-minute OGD, which was set as 0%. Note that the application of 1 mM $BaCl_2$ or Quinine inhibited swelling in HR-astrocytes, while in LR-astrocytes it had the opposite effect. In contrast, the application of 100 μM $BaCl_2$ had the same effect in both groups of astrocytes. Asterisks indicate significant differences from controls ($p<0.05$ (*, significant), $p<0.01$ (**, very significant), $p<0.001$ (***, extremely significant)).
doi:10.1371/journal.pone.0029725.g003

mice underlies the presence of two astrocytic populations differing in their response to OGD (Fig. S3).

Single-cell gene expression profiling reveals astrocyte heterogeneity in the cortex of GFAP/EGFP mice

We have also considered that the different abilities of astrocytes to regulate their volume might originate from distinct cellular phenotypes characterized by ion channel/transporter expression profiles. Based on pharmacological analyses, we have selected genes encoding astrocytic channels and transporters that could be responsible for the divergent volume regulation in cortical astrocyte subpopulations and performed quantitative single-cell PCR analyses. EGFP-positive cells were collected using fluorescence-activated cell sorting (FACS). Collected brightly fluorescent astrocytes expressed high mRNA levels of *Eaat1*, *Eaat2* and *Aqp4*, which code the typical astrocytic markers: EAAT transporters 1 and 2 and AQP4 water channels [29]. EGFP-positive cells expressing genes typical of NG2 glia, namely *Pdgfr* and *Cspg4* (coding PDGFα receptors and chondroitin sulfate proteoglycan 4), were excluded from the analyses. The first round of screening, in which all genes listed in Table S1 were tested, revealed significant differences in gene expression among cortical astrocytes. For further analyses, we have chosen 16 genes (Table S1) that showed marked differences in their expression levels during the first round

of screening. The expression profiling was performed in two independent experiments (I and II). In experiment I, the expression profiles of *Nkcc1*, *Eaat1-2*, *Vdac1-3*, *Aqp4*, *Clcn2* and *Kcc1* were determined in 103 astrocytes. The numbers of astrocytes positive for individual genes are listed in Table S2. At first, we looked for correlations between the genes using Spearman's correlation. The correlation coefficient is a value between −1 and 1, where 1 reflects a perfect positive correlation, −1 reflects a perfect negative correlation and 0 indicates no correlation. Significant positive correlations were found between *Eaat1* *Eaat2*, *Eaat1* *Aqp4*, *Eaat1* *Vdac1*, *2*, *3*, *Eaat2* *Aqp4* and *Eaat2* *Vdac2*. The correlation coefficients for genes tested in experiment I are listed in Table S4. We have employed three independent analyses: principal component analysis (PCA) that enables the visualization of multidimensional data in a two- or three-dimensional plot [30], hierarchical clustering and Kohonen SOMs (Fig. 4 and 5). Based on these three independent mathematical algorithms, we have identified two subpopulations of astrocytes subpopulation 1 and subpopulation 2 that differ in their gene expression profiles. In experiment I (Fig. 4), both subpopulations expressed comparable levels of *Eaat1-2*, *Aqp4*, *Nkcc1*, *Kcc1* and *Vdac1-3*; however, in subpopulation 2 a markedly reduced expression of Clcn2 was detected. In experiment II, the expression of *Eaat1-2*, *Kcnj2*, *10*, *16*, *Kcnk1*, *2*, *10* and *Clcn2* was measured in 52 individual astrocytes. The numbers of astrocytes positive for individual genes after dividing the cells into two astrocytic

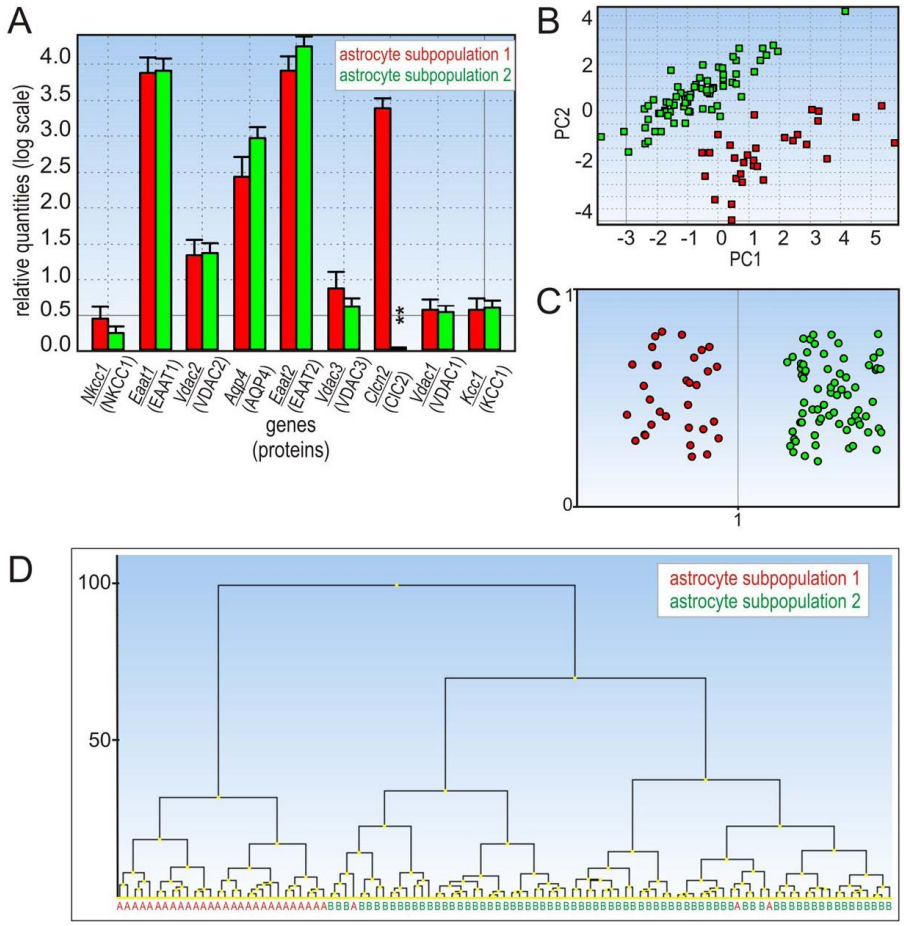


Figure 4. Experiment I: gene expression profiling of distinct astrocytic subpopulations. A: Bar plot with SEM for all the expressed genes; significant differences are indicated with asterisks ($p<0.05$ (*), $p<0.01$ (**), $p<0.001$ (***). **B:** Principal component analysis. The identification of 2 astrocytic subpopulations is along the first principal component, which accounts for most of the variation in the measured data. **C:** Clustering of astrocytes using Kohonen SOMs. The expression levels of all genes were mean-centered. Each dot represents one cell. **D:** Dendrogram based on all astrocytic genes. The y-axis shows the distance between groups.
doi:10.1371/journal.pone.0029725.g004

subpopulations are listed in Table S3. In order to compare mRNA levels from independent experiments and to follow the same astrocytic subpopulations that were identified in experiment I, we have analyzed *Eaat1*, *Eaat2* and *Clcn2* in both experiments. Significant positive correlations were found between *Eaat1* *Eaat2* and *Eaat2* *Kcnj10*. Interestingly, significant negative correlations were found between *Kcnk2* *Kcnk1* and *Clcn2* *Kcnk1*, which means that the majority of cells expressing *Kcnk1* (the gene encoding TWIK-1) do not express *Kcnk2* and *Clcn2* (encoding TREK-1 and ClC2 channels). The Spearman correlation coefficients for the genes tested in experiment II are listed in Table S5. In experiment II (Fig. 5), in addition to *Clcn2*, *Kcnj10*, *Kcnk1* and *Kcnk2* were also identified as genes separating the two astrocytic subpopulations. While *Kcnj10* and *Kcnk1* were highly expressed in subpopulation 2, their expression in subpopulations 1 was significantly lower. On the other hand, *Kcnk2* was highly expressed in subpopulation 1, while it was almost absent in subpopulation 2.

In summary, we found two astrocytic subpopulations in the cortex of GFAP/EGFP mice related to distinct expression profiles of *Clcn2*, *Kcnj10* and *Kcnk1*, 2 (genes encoding ClC2, Kir4.1, TWIK-1 and TREK-1).

Discussion

In the present study we have demonstrated that the two previously described cortical astrocytic populations HR- and

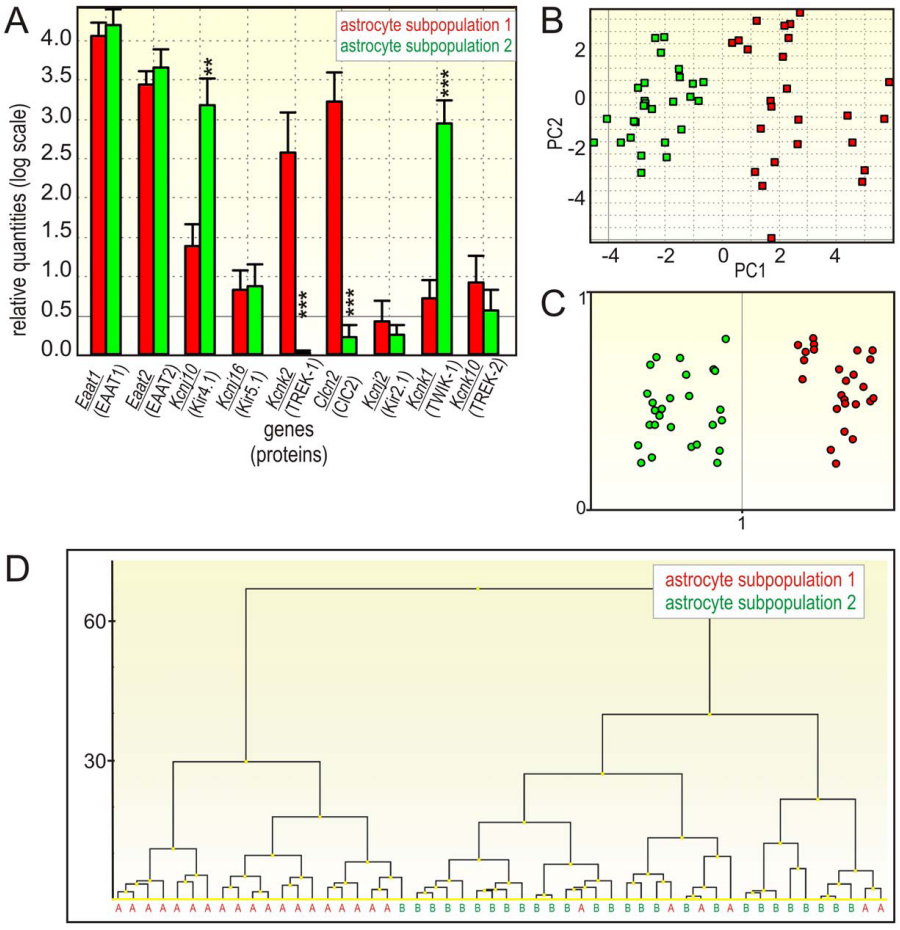


Figure 5. Experiment II: gene expression profiling of distinct astrocytic subpopulations. A: Bar plot with SEM for all the expressed genes; significant differences are indicated with asterisks ($p<0.05$ (*), $p<0.01$ (**), $p<0.001$ (***). **B:** Principal component analysis. The identification of 2 astrocytic subpopulations is along the first principal component, which accounts for most of the variation in the measured data. **C:** Clustering of astrocytes using Kohonen SOMs. The expression levels of all genes were mean-centered. Each dot represents one cell. **D:** Dendrogram based on all astrocytic genes. The y-axis shows the distance between groups.
doi:10.1371/journal.pone.0029725.g005

LR-astrocytes respond differently to Cl^- and K^+ channel inhibitors when applied during OGD. In accordance with the pharmacological experiments, quantitative single-cell PCR profiling revealed that cortical astrocytes are heterogeneous with respect to the gene expression profiles of the ion channels/transporters that participate in their volume regulation. We distinguished subpopulation 1, which is characterized by high gene expression levels of Cl^- channels and K^+ channels, namely ClC2 and TREK-1, and subpopulation 2, displaying high gene expression levels of K^+ channels Kir4.1, TWIK-1. As we hypothesize that the diverse HR- and LR-astrocyte swelling might be a consequence of different gene expression levels for ClC2 and TREK-1, both contributing strongly to astrocyte volume regulation, we propose a correlation of subpopulation 1 with LR-astrocytes and subpopulation 2 with HR-astrocytes. Their proposed correlation with the HR-/LR- classification scheme is summarized in Figure 6.

The most striking dissimilarity between HR- and LR- astrocytes was found during the inhibition of K_{2P} channels by 1 mM $BaCl_2$ and 200 μM quinine. Barium chloride (1 mM), which was shown to block the inwardly rectifying K^+ channel TWIK-1 [27], markedly reduced the swelling of HR-astrocytes, thus suggesting that TWIK-1 channels contribute to K^+ influx into HR-astrocytes. In contrast, quinine, which inhibits the outwardly rectifying K_{2P} channel TREK-1, increased the swelling of LR-astrocytes, thus implying that K_{2P} channels provide K^+ extrusion in LR-astrocytes. Since polymodal TREK channels are activated by mechanical stress/cell swelling, their role in RVD is highly possible [31]. These data are also in agreement with previous studies of Skatchkov and co-authors [10] in retinal Müller glia. Moreover, our quantitative single-cell PCR analysis revealed distinct astrocytic subpopulations: subpopulation 1 expressing extremely high levels of *Kcnk2* (a gene for TREK-1) and subpopulation 2 expressing high levels of *Kcnk1* (a gene for TWIK-1). Combining our pharmacological and PCR data, we propose that TREK-1 channels are uniquely expressed in LR-astrocytes and thus contribute to their efficient cell volume regulation during OGD by providing K^+ efflux. On the other hand, HR-astrocytes strongly express TWIK-1, which is responsible for enhanced K^+ uptake and their swelling. A similar heterogeneity in K_{2P} channel expression was demonstrated in hippocampal astrocytes [26]. Unequal TWIK-1/TREK-1 expression levels might originate from the local environment of each astrocyte in the cortex, as proposed in Fig. 7. We hypothesize that HR-astrocytes comprise cells locally exposed to a high extracellular K^+ concentration ($[K^+]_o$), thus predominantly expressing inwardly rectifying channels for effective K^+ uptake. LR-astrocytes, in contrast, might receive K^+ as a consequence of spatial buffering and release it into the ECS. Apparently, K_{2P} channels play an important role during ischemia as demonstrated by Kucheryavykh and co-authors [12], who showed the involvement of TREK-2 channels in response to anoxia/hypoglycemia in cultured rat astrocytes. Our present findings demonstrate the involvement of TREK-1 rather than TREK-2 in astrocytic swelling during OGD; however, such discrepancies might originate from the different properties of cultured astrocytes isolated from 1-day-old rats [12] and those in brain slices of 30-day-old mice, and moreover, from employing different models of brain injury. Nonetheless, in our recent work, Pivonkova and co-authors [25] described the increased expression of TREK-1 in hippocampal astrocytes in the CA1 region following global cerebral ischemia *in vivo*.

Furthermore, we have demonstrated that in HR- and LR-astrocytes the inhibition of Kir channels led to additional cell swelling, which suggests that Kir channels contribute rather to K^+ efflux than K^+ uptake after 40 minute OGD. This is in line with

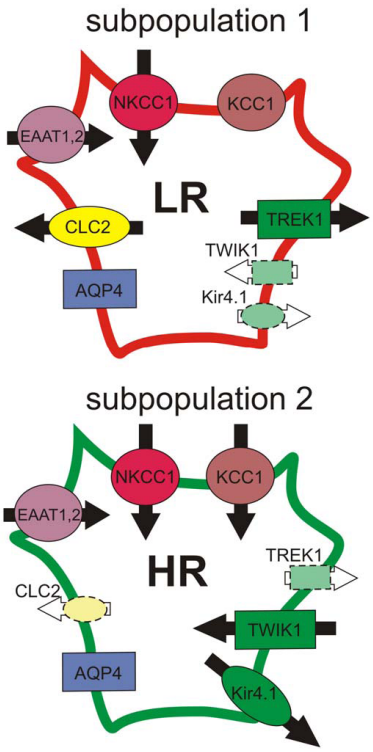


Figure 6. Two astrocytic subpopulations differing in the gene expression levels of K^+/Cl^- channels. We propose a correlation of subpopulation 1 with LR-astrocytes and subpopulation 2 with HR-astrocytes due to the diverse gene expression levels for ClC2 and TREK-1, which are responsible for K^+ and Cl^- efflux and thus contribute to cell volume regulation. Ion channels outlined with dashed line indicate their low gene expression levels. Arrows indicate the proposed direction of ion/EAA fluxes through the channels/transporters during OGD based on the effect of the relevant inhibitors. AQP4 – aquaporin channel (subtype AQP4), ClC2 - chloride channel (subtype ClC2), EAAT – excitatory amino acid transporter (subtypes EAAT1 and EAAT2), HR – high response astrocyte, TWIK1 and TREK1 – two-pore domain potassium channels (subtypes TWIK-1, TREK-1), KCC1 – K^+-Cl^- co-transporter (subtype KCC1), Kir4.1 – inwardly rectifying potassium channel, LR – low response astrocyte, NKCC1 – Na^+-K^+-Cl^- co-transporter (subtype NKCC1).
doi:10.1371/journal.pone.0029725.g006

data demonstrating the participation of Kir4.1 in RVD [32,33,34,35]. In contrast to the comparable effect of a Kir channel inhibitor (100 μM $BaCl_2$) on the extent of cell swelling in both astrocytic subpopulations, we found a significantly increased expression of *Kcnj10* (encoding Kir4.1) in subpopulation 2 that comprises HR-astrocytes. In general, an astrocyte is considered as a multifunctional unit in which the role of Kir channels can differ in the various plasma membrane regions of a given cell, providing K^+ uptake at sites facing the synapses and K^+ release from astrocyte endfeet that contact blood vessels [36]. Based on this assumption, changes in total astrocytic volume thus reflect both K^+ uptake as well as K^+ release. As blocking Kir4.1 channels by 100 μM barium resulted in a similar effect on LR- and HR-

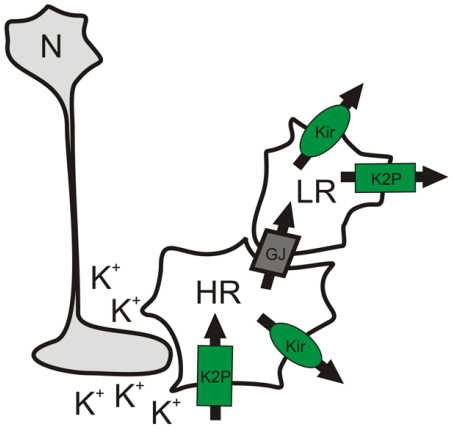


Figure 7. Proposed K^+ movement from HR- to LR-astrocytes during OGD. HR-astrocytes exposed to higher $[K^+]_o$ take up K^+ (by K_{2P} channels or co-transporters), which is either released via Kir channels or redistributed through the astrocytic syncytium and then extruded by Kir and K_{2P} channels in LR-astrocytes. HR – high response astrocyte, LR – low response astrocyte, N – neuron, K2P – two-pore domain potassium channel, Kir – inwardly rectifying potassium channel, GJ – gap junction.
doi:10.1371/journal.pone.0029725.g007

astrocyte volume, we suggest that in both types of astrocytes the proportion of Kir4.1 channels that provide K^+ uptake (causing astrocytic swelling) and those providing K^+ release (causing astrocytic shrinkage) are comparable. This might ultimately lead to the similar impact of Kir4.1 inhibition on astrocyte volume despite the observed differences in *Kcnj10* expression between the two subpopulations. We hypothesize that the high expression of *Kcnj10* in HR-astrocytes may mirror the increased requirements for Kir4.1 channel densities on HR-astrocyte membranes as these cells might be exposed to higher local extracellular concentration of K^+ than are LR-astrocytes, thus providing significantly greater K^+ uptake/K^+ release. This accords well with our previous findings, in which we described the presence of cortical astrocytes differing in Kir4.1 levels based on immunohistochemistry [20]. Nonetheless, a detailed analysis of different cellular compartments, such as astrocytic processes in the vicinity of neuronal synapses or those neighboring blood vessels, might elucidate the relative contribution of K^+ influx and efflux to astrocyte volume with respect to the local uptake/release provided by Kir4.1 channels. Additionally, both types of astrocytes express *Kcnj16* (encoding Kir5.1), and thus the ratio between the Kir4.1 homomer and the Kir4.1/Kir5.1 heteromer might play an important role in K^+ uptake as well [36]. Finally, we also cannot exclude the possibility that 100 μM barium may not have been sufficient to fully block the K^+ fluxes carried by Kir4.1 in brain slices that are 400 μm thick.

Similarly to K_{2P} channel inhibition, we found significant diversity between HR- and LR-astrocytes when chloride channels were inhibited using three different inhibitors. The inhibition of VRACs by tamoxifen and DCPIB led to a marked reduction of swelling in HR-astrocytes, which is in contrast to the proposed role of VRACs in RVD [37,38]. Nevertheless, these channels have been shown to be responsible for chloride influx leading to persistent neuronal swelling during a prolonged excitotoxic insult [16]. On the other hand, in LR-astrocytes the application of Tamoxifen during OGD caused no volume changes. Nevertheless, when DCPIB, so far the most specific known inhibitor of VRACs, was applied, a significant increase in astrocyte volume was detected. We hypothesize that LR-astrocytes efficiently regulate their volume by VRACs, while HR-astrocytes might be more ischemia-affected cells, in which volume regulation is suppressed. As expected, the application of NPPB, a non-selective anion channel blocker, led to the inhibition of swelling in both HR- and LR-astrocytes during OGD, demonstrating that different chloride channel types might be responsible for Cl^- movement leading to astrocytic swelling during OGD. Since the molecular structure of VRACs has not yet been characterized [39], we performed gene expression profiling of anion/chloride channels possibly involved in astrocytic volume changes, such as plasmalemmal voltage-dependent anion channels (VDAC1-3) and chloride channels ClC2-10. Interestingly, we found marked differences in the expression levels of *Clcn2* (a gene for the ClC2 channel), which has been reported to be activated by cell swelling [39]. *Clcn2* is highly expressed only in astrocytic subpopulation 1 (comprising LR-astrocytes), thus possibly contributing to their increased ability to regulate their cell volume.

Since it has been demonstrated that some ion channels are co-expressed in certain regions of the astrocytic plasma membrane, such as the endfeet [8,26,27,40], the importance of membrane protein interactions that might occur between different membrane units devoted to extracellular K^+ buffering and water homeostasis should also be considered. Kir channels and AQP4 water channels are among the first candidates for such interactions (specifically Kir4.1), as demonstrated previously [41,42]; however, regulatory volume decrease and volume homeostasis in astrocytes can also involve the co-localization/interaction of transient receptor potential vanilloid-4 (TRPV4) channels and AQP4 water channels occurring in a Ca^{2+}-dependent manner, as demonstrated by Benfenati and co-authors [43]. Despite the fact that we found comparable *Aqp4* expression in both astrocytic subpopulations, our preliminary data also revealed marked differences in *Trpv4* levels in cortical astrocytes of EGFP/GFAP mice (unpublished data), a finding that accords well with a previously published study on primary cultures of cortical astrocytes in which only a certain subpopulation of astrocytes (~60%) responded to a TRPV4 agonist or hypoosmotic stress [44]. However, further studies are necessary to clarify the role of TRPV4 channels in the diverse volume regulation of the two astrocytic subpopulations.

Finally, the inhibition of Na^+-K^+-Cl^- co-transporter also revealed differences between the two astrocytic populations. In HR-astrocytes we observed a marked reduction of swelling following bumetanide application during OGD, while the volume of LR-astrocytes remained unaffected. Our observations in HR-astrocytes are in agreement with those describing the involvement of NKCCs in astrocyte swelling and their contribution to cytotoxic edema [45 49]. The absence of a bumetanide effect on volume changes in LR-astrocytes might be explained by low NKCC protein levels, which were described by Yan and co-authors [50]. They found only a weak expression of NKCC1 in adult rat cortical astrocytes, while stronger expression was detected in perivascular astrocytes. Nevertheless, our PCR analysis did not reveal significant differences in the gene expression levels of *Nkcc1* among cortical astrocytes, and the overall expression levels were quite low in both astrocytic subpopulations.

No differences between LR- and HR-astrocytes were found during the inhibition of K^+-Cl^- co-transporters or the Na^+-dependent glutamate transporters EAAT1 (GLAST) and EAAT2 (GLT-1). KCC co-transporters contribute to the OGD-induced cell swelling in both subpopulations and accordingly no differences in the gene expression levels were found between subpopulation 1

and 2. The participation of KCCs in K^+/Cl^- uptake under pathological conditions has been suggested previously in studies performed in the chicken retina [5]. Similarly, a block of glutamate co-transporters did not reveal differences between HR- and LR-astrocytes. In both HR- and LR-astrocytes, glutamate transporters contribute to the cell volume increase, which is in accordance with previous studies [51,52].

We have proposed that differences in the extent of astrocytic swelling during OGD may be due to distinct ion channel and transporter expression. However, the possibility that HR- and LR-astrocytes face high/low $[K^+]_o$ and excitatory amino acid concentrations that could trigger distinct mechanisms of cell swelling and cell volume regulation has to be also considered. Although we have shown that their location in the cortical layers does not affect the extent of volume changes during OGD, we cannot rule out that the different volume regulation of astrocytes might partially correlate with their contact with different types of synapses or neurons with different sensitivity to ischemia. Since the application of inhibitors on acute brain slices unquestionably affects all cell types, we cannot overlook their inhibitory effect on neurons. As the inhibition of EAATs, KCCs, VRACs and K^+ channels led to the reduction of HR-astrocyte swelling, we suggest that the effect of the inhibitors might be partially achieved by a block of K^+ and EAA efflux from neurons, resulting in decreased astrocytic uptake.

Taken together, our results describe astrocytic subpopulations differing in their gene expression levels of K^+ and Cl^- channels, which participate in astrocyte swelling as well as their volume regulation. To the best of our knowledge, this is the first study showing astrocyte heterogeneity in the cortex with respect to distinct expression profiles of K^+ and Cl^- channels. Additionally, these findings support our earlier hypothesis that HR- and LR-astrocytes differentially regulate their volume during OGD based on their different expression levels of ion channels/transporters.

Supporting Information

Figure S1 3D-confocal morphometry of GFAP/EGFP astrocytes. An EGFP-labeled astrocyte sectioned into a uniformly spaced (1 μm) set of 2D parallel images [29]. The cell surface was found in each image using an edge-detecting algorithm, and the area of the image surrounded by the edge was calculated for each image (bottom). For the cell soma volume (highlighted by full line) and the total cell volume (highlighted by dotted line) calculations, the area of interest was chosen for each individual cell.
(TIF)

Figure S2 Scheme of the experimental sequence for measuring astrocyte volume changes during OGD. Superimposed confocal images of an EGFP-labeled cortical astrocyte in ACSF, during 20-minute OGD and during a 20-minute co-application of $ACSF_{OGD}$ and inhibitor. The volume changes were quantified every 10 minutes.
(TIF)

Figure S3 Distribution of two astrocyte populations in the cortex of GFAP/EGFP mice related to gender or their location in the cortical layers. A: Percent ratio of HR-(black)/LR- (white) astrocytes in the cerebral cortex of female (n = 193; left) and male (n = 168; right) mice. **B:** Percent ratio of HR-(black)/LR- (white) astrocytes in cortical layers I (n = 46), II III (n = 113) and IV VI (n = 133).
(TIF)

Table S1 Sequences of primers used for quantitative single-cell PCR. Genes used for analyses in experiments I and II are in bold.
(DOC)

Table S2 Number of cells positive for individual genes – experiment I.
(DOC)

Table S3 Number of cells positive for individual genes – experiment II.
(DOC)

Table S4 Experiment I - Spearman correlation coefficients. Significant correlations are in bold.
(DOC)

Table S5 Experiment II - Spearman correlation coefficients. Significant correlations are in bold.
(DOC)

Acknowledgments

The authors thank Helena Pavlikova for her technical assistance and James Dutt for helpful comments and suggestions.

Author Contributions

Conceived and designed the experiments: MK MA. Performed the experiments: JB VR PH DD HP MK MA. Analyzed the data: JB VR MK MA. Contributed reagents/materials/analysis tools: JB VR MK MA. Wrote the paper: JB MA.

References

1. Mongin AA (2007) Disruption of ionic and cell volume homeostasis in cerebral ischemia: The perfect storm. Pathophysiology 14: 183 193.
2. Phillis JW, Ren J, O'Regan MH (2000) Transporter reversal as a mechanism of glutamate release from the ischemic rat cerebral cortex: studies with DL-threo-beta-benzyloxyaspartate. Brain Res 880: 224.
3. Kahle KT, Simard JM, Staley KJ, Nahed BV, Jones PS, et al. (2009) Molecular mechanisms of ischemic cerebral edema: role of electroneutral ion transport. Physiology (Bethesda) 24: 257 265.
4. Ringel F, Plesnila N (2008) Expression and functional role of potassium-chloride cotransporters (KCC) in astrocytes and C6 glioma cells. Neurosci Lett 442: 219 223.
5. Ochoa-de la Paz LD, Lezama R, Toscano B, Pasantes-Morales H (2005) Mechanisms of chloride influx during KCl-induced swelling in the chicken retina. Pflugers Arch 449: 526 536.
6. Kimelberg HK (2005) Astrocytic swelling in cerebral ischemia as a possible cause of injury and target for therapy. Glia 50: 389 397.
7. Butt AM, Kalsi A (2006) Inwardly rectifying potassium channels (Kir) in central nervous system glia: a special role for Kir4.1 in glial functions. J Cell Mol Med 10: 33 44.
8. Nagelhus EA, Mathiisen TM, Ottersen OP (2004) Aquaporin-4 in the central nervous system: cellular and subcellular distribution and coexpression with KIR4.1. Neuroscience 129: 905 913.
9. Pasler D, Gabriel S, Heinemann U (2007) Two-pore-domain potassium channels contribute to neuronal potassium release and glial potassium buffering in the rat hippocampus. Brain Res 1173: 14 26.
10. Skatchkov SN, Eaton MJ, Shuba YM, Kucheryavykh YV, Derst C, et al. (2006) Tandem-pore domain potassium channels are functionally expressed in retinal (Muller) glial cells. Glia 53: 266 276.
11. Buckler KJ, Honore E (2005) The lipid-activated two-pore domain K+ channel TREK-1 is resistant to hypoxia: implication for ischaemic neuroprotection. J Physiol 562: 213 222.
12. Kucheryavykh LY, Kucheryavykh YV, Inyushin M, Shuba YM, Sanabria P, et al. (2009) Ischemia Increases TREK-2 Channel Expression in Astrocytes: Relevance to Glutamate Clearance. Open Neurosci J 3: 40 47.
13. Okada Y, Sato K, Numata T (2009) Pathophysiology and puzzles of the volume-sensitive outwardly rectifying anion channel. J Physiol 587: 2141 2149.
14. Kimelberg HK, Jin Y, Charniga C, Feustel PJ (2003) Neuroprotective activity of tamoxifen in permanent focal ischemia. J Neurosurg 99: 138 142.

15. Zhang Y, Zhang H, Feustel PJ, Kimelberg HK (2008) DCPIB, a specific inhibitor of volume regulated anion channels (VRACs), reduces infarct size in MCAo and the release of glutamate in the ischemic cortical penumbra. Exp Neurol 210: 514 520.
16. Inoue H, Okada Y (2007) Roles of volume-sensitive chloride channel in excitotoxic neuronal injury. J Neurosci 27: 1445 1455.
17. Chvatal A, Anderova M, Hock M, Prajerova I, Neprasova H, et al. (2007) Three-dimensional confocal morphometry reveals structural changes in astrocyte morphology in situ. J Neurosci Res 85: 260 271.
18. Chvatal A, Anderova M, Kirchhoff F (2007) Three-dimensional confocal morphometry - a new approach for studying dynamic changes in cell morphology in brain slices. J Anat 210: 671 683.
19. Nolte C, Matyash M, Pivneva T, Schipke CG, Ohlemeyer C, et al. (2001) GFAP promoter-controlled EGFP-expressing transgenic mice: a tool to visualize astrocytes and astrogliosis in living brain tissue. Glia 33: 72 86.
20. Benesova J, Hock M, Butenko O, Prajerova I, Anderova M, et al. (2009) Quantification of astrocyte volume changes during ischemia in situ reveals two populations of astrocytes in the cortex of GFAP/EGFP mice. J Neurosci Res 87: 96 111.
21. Stahlberg A, Andersson D, Aurelius J, Faiz M, Pekna M, et al. (2011) Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. Nucleic Acids Res 39: e24.
22. Bergkvist A, Rusnakova V, Sindelka R, Garda JM, Sjogreen B, et al. (2010) Gene expression profiling Clusters of possibilities. Methods 50: 323 335.
23. Matyash V, Kettenmann H (2010) Heterogeneity in astrocyte morphology and physiology. Brain Res Rev 63: 2 10.
24. Abdullaev IF, Rudkouskaya A, Schools GP, Kimelberg HK, Mongin AA (2006) Pharmacological comparison of swelling-activated excitatory amino acid release and Cl- currents in cultured rat astrocytes. J Physiol 572: 677 689.
25. Pivonkova H, Benesova J, Butenko O, Chvatal A, Anderova M (2010) Impact of global cerebral ischemia on K+ channel expression and membrane properties of glial cells in the rat hippocampus. Neurochem Int 57: 783 794.
26. Seifert G, Huttmann K, Binder DK, Hartmann C, Wyczynski A, et al. (2009) Analysis of astroglial K+ channel expression in the developing hippocampus reveals a predominant role of the Kir4.1 subunit. J Neurosci 29: 7474 7488.
27. Zhou M, Xu G, Xie M, Zhang X, Schools GP, et al. (2009) TWIK-1 and TREK-1 are potassium channels contributing significantly to astrocyte passive conductance in rat hippocampal slices. J Neurosci 29: 8551 8564.
28. Liu M, Dziennis S, Hurn PD, Alkayed NJ (2009) Mechanisms of gender-linked ischemic brain injury. Restor Neurol Neurosci 27: 163 179.
29. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci 28: 264 278.
30. Stahlberg A, Elbing K, Andrade-Garda JM, Sjogreen B, Forootan A, et al. (2008) Multiway real-time PCR gene expression profiling in yeast Saccharomyces cerevisiae reveals altered transcriptional response of ADH-genes to glucose stimuli. BMC Genomics 9: 170.
31. Enyedi P, Czirjak G (2010) Molecular background of leak K+ currents: two-pore domain potassium channels. Physiol Rev 90: 559 605.
32. Dibaj P, Kaiser M, Hirrlinger J, Kirchhoff F, Neusch C (2007) Kir4.1 channels regulate swelling of astroglial processes in experimental spinal cord edema. J Neurochem.
33. Hirrlinger PG, Wurm A, Hirrlinger J, Bringmann A, Reichenbach A (2008) Osmotic swelling characteristics of glial cells in the murine hippocampus, cerebellum, and retina in situ. J Neurochem 105: 1405 1417.
34. Obara-Michlewska M, Jiang H, Aschner M, Albrecht J (2010) Gain of function of Kir4.1 channel increases cell resistance to changes of potassium fluxes and cell volume evoked by ammonia and hypoosmotic stress. Pharmacol Rep 62: 1237 1242.
35. Obara-Michlewska M, Pannicke T, Karl A, Bringmann A, Reichenbach A, et al. (2011) Down-regulation of Kir4.1 in the cerebral cortex of rats with liver failure and in cultured astrocytes treated with glutamine: Implications for astrocytic dysfunction in hepatic encephalopathy. J Neurosci Res 89: 2018 2027.
36. Hibino H, Inanobe A, Furutani K, Murakami S, Findlay I, et al. (2010) Inwardly rectifying potassium channels: their structure, function, and physiological roles. Physiol Rev 90: 291 366.
37. Ernest NJ, Weaver AK, Van Duyn LB, Sontheimer HW (2005) Relative contribution of chloride channels and transporters to regulatory volume decrease in human glioma cells. Am J Physiol Cell Physiol 288: C1451 1460.
38. Parkerson KA, Sontheimer H (2003) Contribution of chloride channels to volume regulation of cortical astrocytes. Am J Physiol Cell Physiol 284: C1460 1467.
39. Nilius B, Droogmans G (2003) Amazing chloride channels: an overview. Acta Physiol Scand 177: 119 147.
40. Price DL, Ludwig JW, Mi H, Schwarz TL, Ellisman MH (2002) Distribution of rSlo Ca2+-activated K+ channels in rat astrocyte perivascular endfeet. Brain Res 956: 183 193.
41. Amiry-Moghaddam M, Frydenlund DS, Ottersen OP (2004) Anchoring of aquaporin-4 in brain: molecular mechanisms and implications for the physiology and pathophysiology of water transport. Neuroscience 129: 999 1010.
42. Hibino H, Higashi-Shingai K, Fujita A, Iwai K, Ishii M, et al. (2004) Expression of an inwardly rectifying K+ channel, Kir5.1, in specific types of fibrocytes in the cochlear lateral wall suggests its functional importance in the establishment of endocochlear potential. Eur J Neurosci 19: 76 84.
43. Benfenati V, Caprini M, Dovizio M, Mylonakou MN, Ferroni S, et al. (2011) An aquaporin-4/transient receptor potential vanilloid 4 (AQP4/TRPV4) complex is essential for cell-volume control in astrocytes. Proc Natl Acad Sci U S A 108: 2563 2568.
44. Benfenati V, Amiry-Moghaddam M, Caprini M, Mylonakou MN, Rapisarda C, et al. (2007) Expression and functional characterization of transient receptor potential vanilloid-related channel 4 (TRPV4) in rat cortical astrocytes. Neuroscience 148: 876 892.
45. Beck J, Lenart B, Kintner DB, Sun D (2003) Na-K-Cl cotransporter contributes to glutamate-mediated excitotoxicity. J Neurosci 23: 5061 5068.
46. Chen H, Luo J, Kintner DB, Shull GE, Sun D (2005) Na(+)-dependent chloride transporter (NKCC1)-null mice exhibit less gray and white matter damage after focal cerebral ischemia. J Cereb Blood Flow Metab 25: 54 66.
47. Su G, Kintner DB, Flagella M, Shull GE, Sun D (2002) Astrocytes from Na(+)-K(+)-Cl(−) cotransporter-null mice exhibit absence of swelling and decrease in EAA release. Am J Physiol Cell Physiol 282: C1147 1160.
48. Su G, Kintner DB, Sun D (2002) Contribution of Na(+)-K(+)-Cl(−) cotransporter to high-[K(+)](o)- induced swelling and EAA release in astrocytes. Am J Physiol Cell Physiol 282: C1136 1146.
49. Yan Y, Dempsey RJ, Flemmer A, Forbush B, Sun D (2003) Inhibition of Na(+)-K(+)-Cl(−) cotransporter during focal cerebral ischemia decreases edema and neuronal damage. Brain Res 961: 22 31.
50. Yan Y, Dempsey RJ, Sun D (2001) Expression of Na(+)-K(+)-Cl(−) cotransporter in rat brain during development and its localization in mature astrocytes. Brain Res 911: 43 55.
51. Koyama Y, Baba A, Iwata H (1991) L-glutamate-induced swelling of cultured astrocytes is dependent on extracellular Ca2+. Neurosci Lett 122: 210 212.
52. Schneider GH, Baethmann A, Kempski O (1992) Mechanisms of glial swelling induced by glutamate. Can J Physiol Pharmacol 70 Suppl: S334 343.
53. Shimamoto K, Lebrun B, Yasuda-Kamatani Y, Sakaitani M, Shigeri Y, et al. (1998) DL-threo-beta-benzyloxyaspartate, a potent blocker of excitatory amino acid transporters. Mol Pharmacol 53: 195 201.

OPEN ACCESS Freely available online

PLOS ONE

Heterogeneity of Astrocytes: From Development to Injury – Single Cell Gene Expression

Vendula Rusnakova[1], Pavel Honsa[2], David Dzamba[2], Anders Ståhlberg[3], Mikael Kubista[1], Miroslava Anderova[2]*

1 Laboratory of Gene Expression, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Prague, Czech Republic, **2** Department of Cellular Neurophysiology, Institute of Experimental Medicine, Academy of Sciences of the Czech Republic, Prague, Czech Republic, **3** Sahlgrenska Cancer Center, Department of Pathology, Sahlgrenska Academy at the University of Gothenburg, Gothenburg, Sweden

Abstract

Astrocytes perform control and regulatory functions in the central nervous system; heterogeneity among them is still a matter of debate due to limited knowledge of their gene expression profiles and functional diversity. To unravel astrocyte heterogeneity during postnatal development and after focal cerebral ischemia, we employed single-cell gene expression profiling in acutely isolated cortical GFAP/EGFP-positive cells. Using a microfluidic qPCR platform, we profiled 47 genes encoding glial markers and ion channels/transporters/receptors participating in maintaining K^+ and glutamate homeostasis per cell. Self-organizing maps and principal component analyses revealed three subpopulations within 10–50 days of postnatal development (P10–P50). The first subpopulation, mainly immature glia from P10, was characterized by high transcriptional activity of all studied genes, including polydendrocytic markers. The second subpopulation (mostly from P20) was characterized by low gene transcript levels, while the third subpopulation encompassed mature astrocytes (mainly from P30, P50). Within 14 days after ischemia (D3, D7, D14), additional astrocytic subpopulations were identified: resting glia (mostly from P50 and D3), transcriptionally active early reactive glia (mainly from D7) and permanent reactive glia (solely from D14). Following focal cerebral ischemia, reactive astrocytes underwent pronounced changes in the expression of aquaporins, nonspecific cationic and potassium channels, glutamate receptors and reactive astrocyte markers.

Citation: Rusnakova V, Honsa P, Dzamba D, Ståhlberg A, Kubista M, et al. (2013) Heterogeneity of Astrocytes: From Development to Injury – Single Cell Gene Expression. PLoS ONE 8(8): e69734. doi:10.1371/journal.pone.0069734

Editor: Michal Hetman, University of Louisville, United States of America

Received April 26, 2013; **Accepted** June 12, 2013; **Published** August 5, 2013

Copyright: © 2013 Rusnakova et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Academy of Sciences of the Czech Republic IBT AS CR [AVOZ50520701] and UEM AS CR [AV0Z50390703]; by Grant Agency of the Czech Republic [P303/13/02154S]; by an EMBO Short Term Fellowship [ASTF 110-2011] and an ESF Functional Genomics Short Visit Grant [2426]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: anderova@biomed.cas.cz

Introduction

Astrocytes comprise a heterogeneous cell type with several subgroups; for review see [1]. Even within the same brain region multiple astrocytic subgroups have been observed [2]. In addition to morphological differences, astrocytes show diversity in Ca^{2+} signalling, gap junction coupling, and the expression of membrane proteins such as K^+ channels, glutamate receptors and transporters; for review see [3]. It was recently shown that astrocyte heterogeneity also rests on the expression of inwardly rectifying (KIR) and two-pore-domain K^+ (K2P) channels [4]. In our recent work we demonstrated the presence of two distinct subpopulations of astrocytes that respond differently to oxygen-glucose deprivation, probably due to their different expression of chloride channels (ClC2), inwardly rectifying K^+ channels (KIR4.1) and K2P channels, such as TREK-1 and TWIK-1 [5]. Astrocytes also change their properties during development. In contrast to neurons, astrocytes are formed at late stages of embryogenesis (from E16 and onward) and during the first postnatal weeks [6], and the mouse cortex is not fully developed until between the 3rd and 4th week after birth [7]. The origin of astrocytes is not fully known; possibly they arise from distinct groups of progenitors [8], and some subpopulations may be generated from NG2 glia cells [9]. NG2 glia are characterized by their expression of NG2 chondroitin sulphate proteoglycan (CSPG4) and platelet-derived growth factor α receptor (PDGFαR). After their discovery, NG2 glia were found to be oligodendrocyte progenitor cells [10]. Later, NG2 glia were also shown to be progenitors of some groups of reactive astrocytes [11,12], which appear after CNS injury. This process, termed astrogliosis, is characterized by the increased expression of glial fibrillary acidic protein (GFAP) and Nestin, followed by the elevated expression of vimentin [13]. Another characteristic of astrogliosis is the high expression of aquaporins, especially of AQP1 and AQP9, which are proteins rarely expressed in healthy tissue [14,15]. Several publications have reported reactive astrocytes expressing proteoglycans [16,17] and glutamate metabotropic receptor 5 (*Grm5*) [18,19]. Traditionally, astrocytes are identified by the expression of GFAP [20]. However GFAP is not always expressed in non-reactive astrocytes, and GFAP is frequently not immunohistochemically detected in astrocytes in healthy tissue [21]. Other complementary markers are therefore used to identify astrocytes, including S100β [22], excitatory amino acid transporter 1 and 2 (EAAT1 and 2) [23], glutamine synthetase (GS) [24] and aldehyde dehydrogenase 1 family member L1 (ALDH1L1) [25]. Since non-glial cells also express some of these markers, detecting a combination of markers is needed for the reliable identification of astrocytes [26].

Altogether, there is evidence of astrocyte heterogeneity between, as well as within, CNS regions. To characterize the different types of astrocytes that appear under various conditions, single-cell studies are necessary. Single cell gene expression profiling is a novel and powerful method to study cell heterogeneity [27]. It has already been used to establish the presence of distinct subpopulations of astrocytes in a seemingly homogenous macroscopic population [5,28]. Single cell expression profiling is challenging because of the limited amount of RNA present in a single cell. Very efficient reverse transcription is needed, and for the analysis of more than some 10 genes pre-amplification is required. Cells express mRNAs in bursts [29,30], resulting in large natural variation in the number of any transcript among cells, which can be modelled with a lognormal distribution [31]. To obtain biologically relevant and statistically significant conclusions, rather large numbers of cells must be analysed [32]. A particularly powerful approach to distinguish different cells and subtypes of cells is to study the correlation of transcript levels within individual cells [33]. This kind of information is not available when studying classical samples composed of many cells.

Our goal was to characterize astrocytes based on changes in their gene expression profiles throughout development and after focal cerebral ischemia. We used single cell RT-qPCR profiling to analyse EGFP-positive ($EGFP^+$) cells collected by fluorescence-activated cell sorting (FACS) from the cortex of GFAP/EGFP transgenic mice with the objective of identifying and characterizing subpopulations of astrocytes. We focus our profiling effort on glial glutamate transporters and receptors as well as on ion channels, which are critical for the maintenance of the ionic/neurotransmitter homeostasis of neurons and glial cells under physiological conditions as well as in pathological states.

Materials and Methods

Induction of an Ischemic Lesion in Adult GFAP/EGFP Mice

All experiments were performed on cells from acutely isolated brains of GFAP/EGFP transgenic mice [line designation TgN(GFAPEGFP)], in which the expression of EGFP is controlled by the human GFAP promoter [34]. All procedures involving the use of laboratory animals were performed in accordance with the European Communities Council Directive 24 November 1986 (86/609/EEC) and animal care guidelines approved by the Institute of Experimental Medicine, Academy of Sciences of the Czech Republic (Animal Care Committee on April 17, 2009; approval number 85/2009).

Prior to the induction of focal ischemia, 50-day-old (P50) mice were anaesthetized with 1.5% isoflurane (Abbot) and maintained in 1% isoflurane using a vaporizer (Tec-3, Cyprane Ltd., Keighley). A skin incision between the orbit and the external auditory meatus was made, and a 1–2 mm hole was drilled through the frontal bone 1 mm rostral to the fusion of the zygoma and the squamosal bone and about 3.5 mm ventral to the dorsal surface of the brain. The middle cerebral artery (MCA) was exposed after the dura was opened and removed. The MCA was occluded by short coagulation with bipolar tweezers (SMT, Czech Republic) at a proximal location, followed by transection of the vessel to ensure permanent occlusion. The mice received 0.5 ml saline solution subcutaneously, and during the surgery the body temperature was maintained at 37±1°C using a heating pad.

Immunohistochemistry

The mice were deeply anaesthetized with pentobarbital (PTB, 100 mg/kg, i.p.) and perfused transcardially with 20 ml of saline followed by 20 ml of cooled 4% paraformaldehyde in 0.1 M phosphate buffer. Brains were dissected out, fixed with paraformaldehyde overnight, cryoprotected with sucrose and sliced into 30 μm coronal slices using a microtome (HM550, Microm International). The slices were incubated with 5% Chemiblocker (Millipore) and 0.2% Triton in PBS then with the primary antibodies at 4°C overnight. As primary antibodies, we used anti-GFAP (1:800, Sigma-Aldrich, mouse, conjugated with Cy3), anti-GFAP (1:800, Sigma-Aldrich, mouse), anti-NG2 (1:400, Chemicon, rabbit) and anti-PDGFαR (1:200, Santa Cruz, rabbit). Finally, the secondary antibody was applied for 2 hours, either goat anti-mouse IgG or goat anti-rabbit IgG conjugated with Alexa Fluor 660 (Molecular Probes). A Zeiss 510DUO LSM confocal microscope equipped with Arg/HeNe lasers was used for immunohistochemical analysis. Colocalization images and their maximum z projections were made using a Zeiss LSM Image Browser.

Preparation of Cell Suspensions from the Cortex of GFAP/EGFP Mice

GFAP/EGFP transgenic mice 10, 20, 30 and 50 days old (P10, P20, P30, P50) or 3, 7, and 14 days after middle cerebral artery occlusion (MCAO) (D3, D7, D14) were deeply anesthetized with PTB (100 mg/kg, i.p.) and perfused transcardially with cold (4–8°C) isolation buffer containing (in mM): NaCl 136.0, KCl 5.4, Hepes 10.0, glucose 5.5, osmolarity 290±3 mOsmol/kg. The forebrain was isolated by the removal of the olfactory lobes, cerebellum, and midbrain/hindbrain structures by dissection. To isolate the cerebral cortex, the brain (+2 mm to −2 mm from bregma) was sliced into 400 μm coronal sections using a vibrating microtome HM650V (MICROM International GmbH), and the dorsal cerebral cortex was carefully dissected out from the ventral white matter tracks. The tissue was incubated with continuous shaking at 37°C for 90 minutes in 5 ml of a papain solution (20 U/ml) and 0.2 ml DNase (both from Worthington) prepared in isolation buffer. After papain treatment, the tissue was mechanically dissociated by gentle trituration using a 1 ml pipette. Dissociated cells were layered on the top of 5 ml of Ovomucoid inhibitor solution (Worthington) and harvested by centrifugation (140×g for 6 minutes). This method routinely yielded ~2×10^6 cells per mouse. Cell aggregates were removed by filtering with a 30 μm cell strainer (Becton Dickinson), and the cells were kept on ice until sorting.

Collection of Single $EGFP^+$ Cells

Single cells were sorted using flow cytometry (BD Influx). The flow cytometer was manually calibrated to deposit a single cell in the centre of each collection tube. Hoechst 33258 (Life Technologies) was added to the suspension of cells to check viability. Initially, cell sorting was performed with a negative control (no fluorescent cells) in order to set the fluorescence threshold for collecting $EGFP^+$ cells and to avoid auto-fluorescent cells. After setting the threshold, all $EGFP^+$ cells that crossed this level of fluorescence were collected. Single cells were collected into 96-well plates (Life Technologies) containing 5 μl nuclease-free water with bovine serum albumin (1 mg/μl, Fermentas) and RNaseOut 20 U (Life Technologies). Plates were placed on a pre-cooled rack. The glial cells collected were positive for EGFP and viable. Collected cells in 96-well plates were immediately placed at −80°C. For analysing individual developmental or post-ischemic stages, the cells were isolated from 1 mouse (P10), 2 mice (P20), 1 mouse (P30), 1 mouse (P50), 2 mice (D3), 2 mice (D7) and 1 mouse (D14).

cDNA Synthesis

cDNA synthesis was performed with SuperScript III RT (Life Technologies). Lysed single cells in 5 µl water containing 0.5 µM dNTP (Promega), 1.0 µM oligo (dT15) (Invitrogen) and 1.0 µM random hexamers (Eastport) were incubated at 70°C for 5 min. 50 mM Tris–HCl, 75 mM KCl, 3 mM $MgCl_2$, 5 mM dithiothreitol, 20 U RNaseOut and 100 U SuperScript III (all Life Technologies) were added to a final volume of 10 µl. Reverse transcription was performed at 25°C for 5 min, then at 50°C for 60 min, followed by 55°C for 15 min and finally terminated by heating at 70°C for 15 min. Four µl from each sample was diluted 4 times and used for pre-testing. Five µl of cDNA was used for preamplification.

qPCR

Primers were designed using BeaconDesigner (version 7.91, Premier Biosoft International) as previously described [5]. Primer sequences are provided in Table S1. All single cells were pre-tested for the expression of glutamate transporter (*Eaat2*), glutamine synthetase (*Glul*), which are markers of astrocytes, and chondroitin sulphate proteoglycan (*Cspg4*) and platelet-derived growth factor α receptor (*Pdgfra*), which are markers of NG2 glia, to select samples for further gene expression profiling using assays for 47 genes. A CFX384 (Biorad) was used for all qPCR measurements. To each reaction (10 µl) containing iQ SYBR Green Supermix (BioRad) and 400 nM of each primer (EastPort), we added 3 µl of diluted cDNA. The temperature profile was 95°C for 3 min followed by 50 cycles of amplification (95°C for 15 s, 60°C for 15 s and 72°C for 20 s). All samples were subjected to melting curve analysis. The same experimental setup was used to test preamplification.

Preamplification and qPCR

The applied preamplification protocol was verified on samples from three animals. The RNA was extracted and cDNA was prepared to test the preamplification protocol. Each reaction contained 25 µl of iQ Supermix (BioRad), 5 µl of a mix of all primers (final concentration 25 nM each), 5 µl of non-diluted cDNA, and water added to a final volume of 50 µl. The temperature profile was 95°C for 3 min followed by 18 cycles of amplification (95°C for 20 s, 57°C for 4 min and 72°C for 20 s on a Biorad CFX96). The samples were diluted 5 times in water. The expression of all genes was measured in preamplified and non-preamplified samples. The average difference between preamplified and non-preamplified samples and the standard deviations of the differences were calculated. The same setup was then used for the preamplification of single cells.

High Throughput qPCR

The sample reaction mixture had a volume of 5 µl and contained 2.4 µl of diluted preamplified cDNA, 0.25 µl of DNA Binding Dye Sample Loading Reagent (Fluidigm), 2.5 µl SsoFast EvaGreen Supermix (Biorad), and 0.01 µl ROX (Invitrogen). The primer reaction mixture had a final volume of 5 µl and contained 2.5 µl Assay Loading Reagent (Fluidigm) and 2.5 µl of a mix of reverse and forward primers, corresponding to a final concentration of 4 µM. The chip was first primed with an oil solution in the NanoFlex™ 4-IFC Controller (Fluidigm) to fill the control wells of the dynamic array. The reaction mixture (5 µl) was loaded into each sample well, and 5 µl of the primer reaction mixtures was loaded into each assay well of the 48×48 dynamic array. The dynamic array was then placed in the NanoFlex™ 4-IFC Controller for automatic loading and mixing. After 55 min the dynamic array was transformed to the BioMark qPCR platform (Fluidigm). The cycling program was 3 min at 95°C for activation, followed by 40 cycles of denaturation at 95°C for 5 s, annealing at 60°C for 15 s, and elongation at 72°C for 20 s. After PCR, melting curves were collected between 60°C and 95°C with 0.5°C increments.

Data Pre-processing

qPCR data were pre-processed for expression analysis using GenEx (MultiD, version 5.3) as follows: Cq values registered from amplifications that generated melting curves with aberrant Tm were removed, Cq values larger than 26 were replaced with 26, and Cq values with products giving rise to a double peak in melting curves (corresponding to a mixture of expected and aberrant PCR products) were replaced with 26. All missing data, for each gene separately, were then replaced with the highest Cq +2, effectively assigning a concentration of 25% of the lowest measured concentration to the off-scale values. The Cq data were, for each gene separately, converted into relative quantities expressed relative to the sample with the lowest expression (maximum Cq) and finally converted to a logarithmic scale with base 2. The data were not normalized to any reference genes because of the large variation of all transcript levels among individual cells [33]; with this pre-processing, the levels are expressed per cell.

Analysis

Basic statistics were calculated. The non-parametric Mann-Whitney test was used to compare the expression of individual genes between groups of cells. Kohonen self-organizing maps (SOM) of size 3×1, dividing the cells into three groups, were trained using GenEx with the following parameters: 0.10 learning rate, 3 neighbors and 5000 iterations. The SOM analysis was repeated eight times with identical classification of the cells in six of the repeats. This classification of cells into three groups was substantiated with principal component analysis (PCA). The division of cells into three groups by SOM does not imply that there are three distinct subtypes of cells; the cells may also be changing their expression pattern from one state to another, and the SOM then divides the cells into different phases: those just starting to change, those that have undergone substantial change, and finally cells approaching completion. Distinct subtypes, when such are present, appear as separated clusters in PCA. For Kohonen SOM and PCA the data were mean-centered to reduce the effect of variation in the overall expression levels of the different genes. The Spearman correlation coefficients between genes were calculated for each group separately (SAS vers. 9.2) using all data, including the managed off-scale data, and then also for the truly positive expression values only. Only correlation coefficients for genes expressed in at least 50% of cells and having a p-value <0.05 were considered for further analysis. Indirect correlations between genes were established by calculating partial Spearman correlation coefficients (only genes expressed in at least 80% of the cells were considered). An interaction was considered indirect when the correlation coefficient lost significance (at 95% confidence) or was reduced to below 0.4 [28].

Results

$EGFP^+$ cells from GFAP/EGFP transgenic mice are today routinely used by several research groups [4,35] and characterizing their gene expression profile is highly desirable. Because of the possible heterogeneity of astrocytes, single cells were collected by flow cytometry, lysed, and analyzed individually by RT-qPCR. For preamplification and further analysis we selected $EGFP^+$ cells

that were also positive for the astrocytic markers *Glul* and/or *Eaat2* and EGFP$^+$ cells that were also positive for *Cspg4* and/or *Pdgfra*, which are markers of NG2 glia. In these cells the gene expression of the following markers was measured: glial cell markers (*Eaat1, Gfap, Gfapδ, Glul, S100b, Pdgfra, Cspg4*), developmental markers nestin and vimentin (*nes and vim*), aquaporins (*Aqp1, Aqp4, Aqp9*), subunits of glutamate ionotropic AMPA receptors (*Gria1, Gria2, Gria3, Gria4*), glutamate ionotropic NMDA receptors (*Grin1, Grin2a, Grin2b, Grin2c, Grin2d, Grin3a*), glutamate ionotropic kainate receptors (*Grik1, Grik2, Grik3, Grik4, Grik5*), glutamate metabotropic receptors (*Grm1, Grm3, Grm5*), outwardly rectifying K^+ channels (*Kcna3, Kcna4, Kcna5*), inwardly rectifying K^+ channels (*Kcnj2, Kcnj10, Kcnj16*), two pore domain K^+ channels (*Kcnk1, Kcnk2, Kcnk10*) and hyperpolarization-activated cyclic nucleotide-gated channels (*Hcn1, Hcn2, Hcn3, Hcn4*), voltage-gated chloride channels (*Clcn2*), transient receptor potential channels (*Trpv4*), and *Snap25* encoding synaptosomal-associated protein SNAP-25. The markers were selected based on our previous work [5], which revealed significant differences between astrocytic subpopulations in the expression of genes coding for membrane proteins.

For the profiling of a large number of genes on the Biomark platform, it is necessary to preamplify the cDNA; in this regard, we have extensively optimized and validated the preamplification protocol. Our validation scheme is based on three samples that are profiled with and without preamplification and the measured Cq values are compared. Unbiased preamplification should result in the same ΔCq for all of the genes. The reproducibility of the preamplification is reflected by the standard deviation (SD) of the three ΔCq replicates measured for each gene. For our panel and protocol the SD was below 0.5 cycles for almost all genes. Only one gene (*Grin2d*) showed a large standard deviation (SD = 3.7 cycles) (Figure S1) and was omitted in the expression analysis. Inspecting the validation results, we did not find correlation between assay efficiency and ΔCq.

All cells collected from the cortex were analysed for the expression of *Eaat2*, *Glul*, *Cspg4*, and *Pdgfra* to estimate the ratio between cells that express only astrocytic markers (*Eaat2* and *Glul*) and cells that also express at least one of the NG2 glia markers (Figure 1). The number of cells that expressed only astrocytic markers increased, while the number of cells that also expressed NG2 glia markers decreased during postnatal development. After ischemic injury there was a distinct increase in the number of cells expressing NG2 glia markers 3 days after MCAO, but their number decreased again at later post-ischemic stages. In neither system did the expression profile depend on the EGFP intensity of the cells or on their size. Notably, cells expressing NG2 glia markers were distributed across the entire sorter gate (Figure 2, green dots). For single cell qPCR expression profiling 64 cells were collected from P10, 60 cells from P20, 89 cells from P30, 50 cells from P50, 90 cells from D3, 74 cells from D7, and 78 cells from D14. The cDNA produced from these cells was preamplified and measured on the Biomark platform.

Characteristics of EGFP$^+$ Cells Collected from Different Stages of Postnatal Development

The majority of EGFP$^+$ cells isolated from P10 mice expressed *Cspg4* and/or *Pdgfra* (87%) and *S100b* (80%). More than 70% of these cells also expressed *Eaat1, Grik1-3, 5, Gria2-4* and *Kcnj16* (Figure 3A). Generally, cells at P10 were transcriptionally very active (Table 1). The expression profile of cells collected at P20 was different. *Cspg4* and/or *Pdgfra* were expressed in only 32% of the cells, and the expressions of *Kcnj16, Eaat1, Grik1, Grik3, Grik5, Gria2*, and *Gria4* were also reduced. *Gria3* expression remained high and the expression of *Hcn2, Grm5* and *Grin2a* increased. Cells at P30 and P50 displayed very similar expression patterns (Figure 3B). The expression of most genes was lower than at the earlier stages; only *Eaat1*, *Glul*, *Grin2a*, *Gria2* and *Aqp4* showed high expression. The fraction of cells positive for *Pdgfra* and/or *Cspg4* decreased to only 22% and 6% at P30 and P50, respectively.

Cells collected between P10 and P50 were divided by SOM into 3 groups based on their expression profiles (Figure 3C). This SOM classification was fully reproducible - repeated independent classifications predicted the same groups. This shows that the collection of cells is not homogeneous. SOM itself cannot tell if there are distinctly different groups or gradual changes, but the cells can be divided reproducibly into three subpopulations. The spatial separation between the clusters in the PCA suggests that the subpopulations are distinct (Figure 3B). Dividing the cells into more than three groups with SOM did not reveal any additional differences and was not supported by the PCA. We compared the average expression of the individual genes in the three subpopulations, which we henceforth refer to as A1, A2 and A3 (Table S2). Subpopulation A1 was characterized by the high expression of *Pdgfra, Cspg4, Grik1-3, Grik5, Gria2-4, Grin2a, Grin3a, Kcnj16, Eaat1, Glul, S100b, Snap25* and *Kcnj10* (Figure 3C) and was comprised mainly of P10 cells (composition: 64%, 19%, 13% and 2% of P10, P20, P30 and P50, respectively; Figure 3D). Subpopulation A2 was characterized by the low expression of most genes, including *Eaat1*, *Cspg4*, *Pdgfra*, *Grik5*, *Gria2* and *Grin3a*; only *Glul*, *Hcn2*, *Gria3* and *Grin2a* were highly expressed. Subpopulation A2 was comprised mainly of P20 cells (57%); cells from the other stages (P10, P30 and P50) were present at similarly low levels (Figure 3D). In addition to the classical astrocyte markers *Eaat1*, *Glul*, Aqp4, subpopulation A3 also expressed *Grik5*, *Gria2-3*, *Grin2a*, *Grm3* and *Kcnj16* at high levels and was comprised mainly of P30 and P50 cells (46% of each).

The relation between the expressions of different genes among the individual cells can be evaluated by calculating correlation coefficients. The calculation can include all cells, including those that did not express the genes of interest, or only those cells that had measurable expression. When all cells are considered, the dominant effect is whether a cell expresses both genes or not, while when cells with missing data are filtered out, the correlation coefficient is dominated by the expression levels of genes (Table S3, S4 and S5, Figure S2A–C). Although 95% of all cells in subpopulation A1 expressed *Pdgfra* as well as *Cspg4*, the correlation between their expressions was weak (r = 0.44), which means that they are expressed independently of each other in individual cells. We hypothesize that their expression might be regulated by different transcriptional factors and/or that such a weak correlation between *Pdgfra* and *Cspg4* expression might reflect the different stages of astrocyte maturation even within the A1 subpopulation, which comprises EGFP$^+$ cells from P10 and P20. The correlation between the expression of *Gria2* and *Cspg4* was negligible, while *Gria2* expression correlated with *Pdgfra* (r = 0.60) (Table S3). In subpopulation A3, correlated expression was found between *Glul* and *Eaat1* (r = 0.59) when considering only cells with valid expression data (74 cells) and between *Eaat1* and *Gria2* (r = 0.73). Notably, the correlation of *Eaat1* with *Glul* and of *Eaat1* with *Gria2* in subpopulations A1 and A2 was much lower, thus the increasing correlation between *Glul* and *Eaat1* clearly mirrors the maturation process of astrocytes.

Interestingly, gene expression profiling revealed that a large number of EGFP$^+$ cells express *Cspg4* as well as *Pdgfra* at P10, while at P50 we detected only a negligible number of EGFP$^+$ cells that co-expressed these markers (Figure 1). Furthermore, we also detected the co-expression of *Gfap* and *Cspg4/Pdgfra* in most of the cells at P10; this co-expression gradually decreased at P20 and

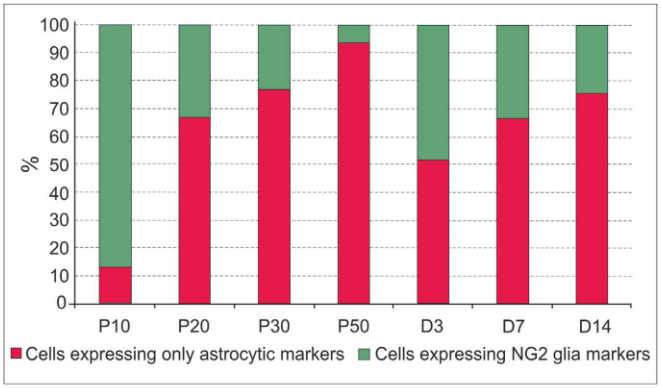


Figure 1. Characteristics of EGFP$^+$ cells in the cortex of GFAP/EGFP mice. (**A**) Percentage of EGFP$^+$ cells expressing either predominantly astrocytic or NG2 glia markers during postnatal development (P10, P20, P30, and P50) and at 3, 7 and 14 days (D3, D7 and D14) following focal cerebral ischemia. Cells expressing *Cspg4* and/or *Pdgfra* receptor are shown in green, while cells expressing only *Eaat2* and/or *Glul* are shown in red. The total numbers of the investigated cells were 137 for P10, 94 for P20, 157 for P30, 50 for P50, 148 for D3, 96 for D7 and 136 for D14.
doi:10.1371/journal.pone.0069734.g001

P30, and at P50 there was no co-expression detected. In order to confirm this observation on the protein level, we carried out immunohistochemical analyses, which revealed the co-expression of GFAP and PDGFαR or of GFAP and NG2 in EGFP$^+$ cells only at P10 and P20, but not at P50 (Figure 4, Figure S3).

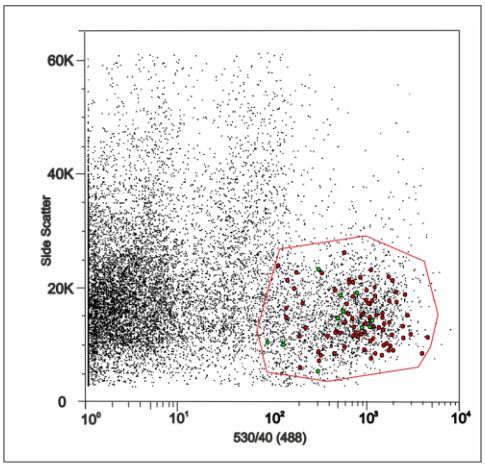


Figure 2. FACS analysis of EGFP$^+$ cells isolated from the cortex of 20-day-old mice. EGFP$^+$ cells sorted based on EGFP intensity (x axis) and side scatter intensity reflecting cell size (y axis). During the selection of cells by FACS, their position is automatically saved in the plot, and thus it is possible to determine the expression profile of each individual cell in the plot after qPCR analysis. The red line indicates the sorter gate and each dot represents one cell. Cells positive for *Cspg4* and/or *Pdgfra* are indicated in green while those positive only for *Eaat1* and/or *Glul* are shown in red. The latter are also drawn larger for better resolution. The small black dots represent all of the cells from which we selected our samples.
doi:10.1371/journal.pone.0069734.g002

Expression Profiling of Reactive Glia Following Focal Cerebral Ischemia

Three days after ischemia the expression profiles of the cells had changed. Approximately 50% of all EGFP$^+$ cells began to express *Pdgfra*, *Vim*, and *Grik5*, while *Aqp4*, *Glul* and *Eaat1* expression was decreased. Seven days after ischemia the cells were transcriptionally more active with particularly high expression of *Pdgfra*, *Vim*, *S100b*, *Eaat1*, *Glul*, *Gria2-4*, *Grin2a*, *Grin2b*, *Grik2*, *Grik5*, potassium channels *Kcna4*, *Kcna5*, *Kcnj10*, *Kcnj16*, *Aqp4* and *Hcn2*. Furthermore, the cells also began to express *Gfap* and *Gfapδ*. At this time, the expression of *Glul* returned to the basal level (P50, Figure 5A). Fourteen days after ischemia, the difference in expression compared to control (P50) was the largest. The expression of *Vim*, *Gfap*, *Gfapδ*, *Aqp4*, *S100b* and *Grm3* had dramatically decreased, while the expression of *Hcn1*, *Hcn2*, *Hcn3*, *Aqp1*, *Aqp9*, *Grm5*, *Gria3*, *Grin2a*, *Grin2d*, *Snap25* and *Kcna3* had markedly increased. At all post-ischemic stages, we found some cells expressing *Cspg4* and *Pdgfra*.

All cells (P50, D3, D7 and D14) were divided by SOM into 3 subpopulations and confirmed by PCA to have different expression profiles (Figure 5B). We will refer to these subpopulations as B1, B2 and B3. To characterize these subpopulations, we compared the expressions of the genes individually in histograms and by conventional univariate statistics. Almost all cells from P50 were assigned to subpopulation B1 (48 out of 50) (Figure 5C), and they comprised 35% of the B1 cells. 47% of B1 cells were from D3, 15% from D7, and only 3 B1 cells were from D14 (Figure 5D). The B1 subpopulation displayed the lowest overall transcriptional activity; high expression was observed only for *Eaat1*, *Glul*, *Aqp4* and *Gria2*. Subpopulation B2 was more transcriptionally active with the high expression of genes encoding glutamate ionotropic receptors, K^+ channels, *Grm3*, *Aqp4*, *Hcn2*, *Gfap*, *Gfapδ*, *Glul*, *Eaat1*, *S100b*, *Vim*, *Cspg4*, and *Pdgfra*. Most B2 cells were from D7 (66%), 31% were from D3 and only 2 cells were from P50. Notably, the two P50 cells also expressed *Pdgfra* and *Cspg4*. Subpopulation B3 was composed of only D14 cells. The very high expression of many genes makes this group unique, and there is no overlap in the PCA between B3 and the other subpopulations. Clearly, the B3 subpopulation is distinct. We also noted that the EGFP

Table 1. Statistics describing gene expression in different postnatal developmental stages or post-ischemic stages.

	Percentage of positive cells							Average relative expression log_2						
Gene	**P10**	**P20**	**P30**	**P50**	**D3**	**D7**	**D14**	**P10**	**P20**	**P30**	**P50**	**D3**	**D7**	**D14**
All cells (count)	64	60	79	50	90	74	78							
Cspg4	77%	32%	46%	4%	49%	35%	38%	3.63	1.45	2.22	0.10	1.99	1.45	1.63
Pdgfra	78%	40%	51%	4%	49%	35%	42%	6.32	2.38	3.37	0.25	3.03	2.36	2.71
Eaat1	83%	65%	92%	90%	88%	80%	87%	4.21	2.35	5.41	6.00	4.48	4.84	5.81
Glul	63%	58%	92%	92%	76%	81%	92%	2.16	2.80	4.50	5.79	3.61	5.91	5.84
Gfap	8%	13%	27%	8%	40%	57%	40%	0.29	0.47	1.15	0.35	1.55	4.14	1.95
Gfap D	19%	20%	44%	32%	47%	64%	40%	0.40	0.55	1.27	0.75	1.66	3.75	1.40
S100b	80%	48%	62%	64%	72%	85%	82%	2.05	1.37	1.87	1.54	2.09	3.81	2.28
nestin	11%	3%	11%	0%	23%	16%	10%	0.24	0.11	0.46	0.00	0.92	0.68	0.45
Vim	31%	23%	30%	18%	60%	82%	55%	0.79	0.69	0.91	0.45	3.33	5.91	1.89
Snap25	61%	25%	35%	26%	27%	26%	77%	1.93	0.82	1.53	1.15	0.82	0.86	4.02
Aqp1	2%	50%	0%	0%	1%	4%	79%	0.03	1.13	0.00	0.00	0.02	0.13	4.89
Aqp4	33%	25%	65%	72%	48%	70%	59%	1.03	0.70	2.96	2.81	1.38	4.15	2.78
Aqp9	9%	40%	5%	2%	0%	8%	91%	0.24	1.26	0.22	0.04	0.00	0.23	6.77
Clcn2	27%	27%	22%	28%	33%	28%	31%	0.76	0.86	0.88	1.27	0.96	1.50	1.30
Hcn1	19%	43%	29%	12%	0%	22%	96%	0.46	2.21	1.24	0.48	0.00	0.94	8.81
Hcn2	59%	95%	0%	24%	43%	55%	95%	1.52	5.05	0.00	0.61	1.38	3.30	8.78
Hcn3	31%	62%	28%	10%	22%	34%	88%	0.65	2.08	0.74	0.26	0.51	0.84	6.55
Hcn4	67%	43%	84%	72%	24%	80%	27%	1.40	1.23	1.88	1.46	0.52	1.69	1.85
Trpv4	38%	17%	34%	28%	39%	34%	35%	1.17	0.40	1.16	0.81	1.13	1.27	1.25
Grik1	67%	30%	44%	14%	40%	47%	63%	2.37	0.91	1.70	0.43	1.28	1.89	2.26
Grik2	75%	58%	38%	30%	42%	61%	47%	2.67	1.65	1.39	0.90	1.31	2.87	1.52
Grik3	70%	32%	48%	8%	44%	31%	33%	2.54	1.11	2.17	0.21	1.62	1.28	1.25
Grik4	45%	15%	22%	4%	19%	30%	14%	1.14	0.42	0.76	0.08	0.53	1.12	0.37
Grik5	86%	47%	54%	72%	77%	68%	69%	3.45	1.42	2.14	2.38	3.26	3.39	3.07
Grin3a	67%	32%	61%	10%	43%	35%	51%	2.27	0.90	2.46	0.25	1.25	1.36	1.58
Gria1	41%	12%	16%	8%	18%	16%	28%	0.87	0.28	0.37	0.22	0.44	0.63	0.72
Gria2	91%	62%	52%	74%	76%	66%	77%	5.22	2.45	2.60	3.54	3.19	3.25	4.08
Gria4	75%	53%	44%	24%	44%	54%	42%	2.67	1.45	1.71	0.83	1.59	2.54	1.54
Grin1	13%	32%	39%	18%	28%	50%	50%	0.27	0.69	1.03	0.47	0.63	1.26	1.30
Grin2a	47%	77%	49%	48%	46%	54%	97%	1.67	3.20	2.45	2.58	1.71	3.11	8.56
Grin2b	61%	33%	33%	30%	44%	57%	59%	1.82	0.90	1.30	1.24	1.37	2.60	2.19
Grin2c	13%	8%	23%	48%	28%	30%	35%	0.29	0.17	0.57	1.55	0.79	0.96	0.94
Grin2d	2%	15%	14%	4%	8%	4%	64%	0.03	0.42	0.55	0.17	0.20	0.11	3.05
Gria3	84%	87%	41%	12%	53%	49%	100%	4.23	4.55	2.21	0.58	2.29	2.46	9.49
Grm1	30%	18%	27%	32%	37%	31%	31%	0.89	0.64	1.01	1.18	1.20	1.24	1.34
Grm3	20%	27%	33%	62%	44%	61%	41%	0.60	1.01	1.35	2.39	1.43	2.80	1.89
Grm5	38%	82%	29%	0%	30%	27%	92%	0.77	2.45	0.77	0.00	0.71	0.64	7.18
Kcna3	25%	67%	18%	14%	23%	27%	91%	0.52	1.47	0.33	0.31	0.52	0.68	6.00
Kcna4	33%	45%	38%	22%	19%	53%	37%	0.99	1.69	1.58	0.73	0.60	2.36	1.48
Kcna5	25%	30%	42%	22%	41%	64%	38%	0.79	0.89	1.58	0.64	1.29	2.83	1.33
Kcnj10	95%	45%	72%	74%	79%	82%	88%	2.05	1.12	1.33	1.48	1.92	3.04	1.96
Kcnj16	66%	30%	54%	52%	60%	55%	58%	2.37	0.96	2.11	2.15	2.25	2.64	2.38
Kcnj2	17%	17%	19%	14%	23%	49%	29%	0.51	0.56	0.73	0.44	0.81	2.12	0.96
Kcnk1	22%	15%	30%	24%	26%	11%	28%	0.53	0.44	1.10	0.65	0.62	0.43	0.94
Kcnk10	67%	25%	33%	20%	44%	50%	32%	1.54	0.55	0.78	0.48	1.02	1.23	0.79
Kcnk2	34%	17%	23%	4%	31%	12%	41%	1.09	0.41	0.80	0.16	0.87	0.42	1.35

doi:10.1371/journal.pone.0069734.t001

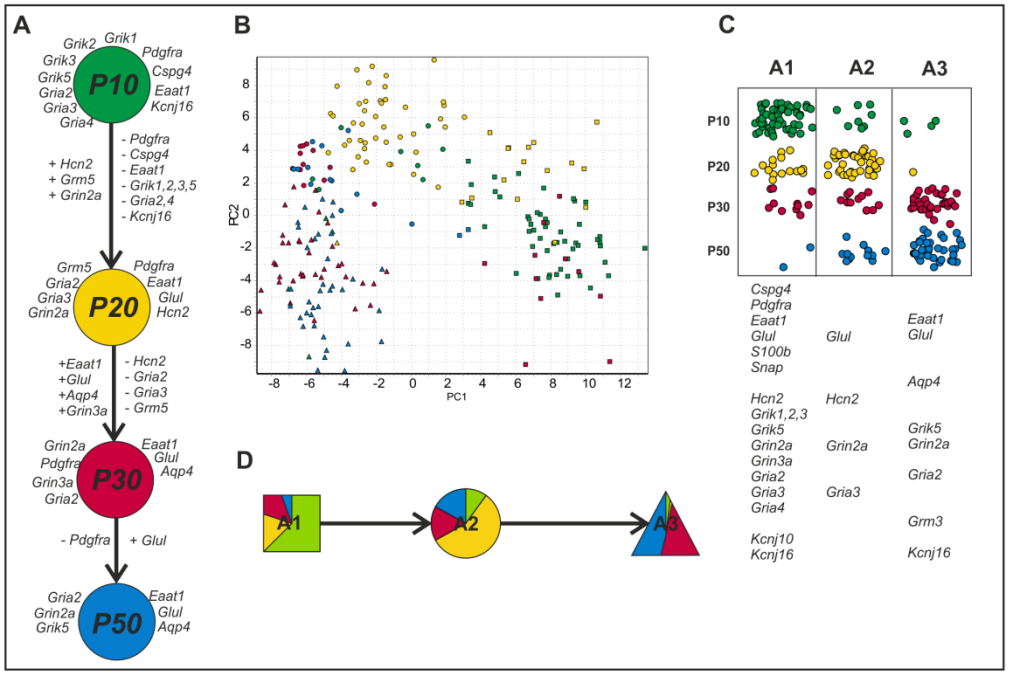


Figure 3. Gene expression profiling of cortical EGFP$^+$ cells during postnatal development. (**A**) Changes in the gene expression of highly expressed (log_2 relative average expression above 2.3 from all data) astrocytic/NG2 glia markers and membrane proteins at P10 (green), P20 (yellow), P30 (red), and P50 (blue). The listed genes show at least 2-fold up (+) or 2-fold down-regulation (−) between sequential developmental stages. (**B**) PCA clustering of cells from all development stages (data mean-centered along genes). The stages are indicated in color (P10 green; P20 yellow; P30 in red, and P50 in blue), and the three groups identified by SOM are indicated by symbols (squares for A1; circles for A2, and triangles for A3). (**C**) The incidence of cells from the three subpopulations at different postnatal developmental stages. The genes under the table indicate highly expressed genes (log2 relative average expression above 1.8 and at least 40% positive cells) in particular subpopulations. (**D**) The distribution of developmental stages in the subpopulations.
doi:10.1371/journal.pone.0069734.g003

fluorescence of these cells was much higher compared to that of cells classified as B1 and B2.

Generally, astrocytes isolated from post-ischemic stages (D3, D7 and D14) showed much higher gene expression than those collected from the postnatal developmental stages. Within each subpopulation B1-B3 we found many significant correlations. The strongest are listed in Tables S6, S7 and S8 and visualized in Figure S2D–F. Interestingly, the expression of *Pdgfra* and of *Cspg4* were strongly correlated when cells with non-measurable expression (r = 0.89, 0.71 and 0.87 in B1, B2 and B3) were included, but the correlation was lost when only cells with measurable expression were considered. The expressions of *Eaat1* and *Glul* in B1 and B3 were correlated (r = 0.66 and 0.79), and a correlation of *Eaat1* and *Aqp4* was found in B1, B2 and B3 (r = 0.43, 0.69, 0.81). A strong correlation was also found between the expressions of *Gfap* and *Gfapδ* in B1 and B2 (r = 0.76 and 0.87) and between the expressions of *Aqp1* and *Aqp9* in B3 (r = 0.78). The expressions of the latter two genes also correlated with those of *Hcn1, Hcn2, Hcn3, Hcn4, Grin2a, Grin2d, Gria3, Grm5, Kcna3, and Kcnk2*. In fact, the expressions of almost all of these genes were strongly correlated on the single cell level in B3 (r close to or larger than 0.7, Figure S4A). Such a strong correlation between the above-listed genes suggests significant alterations in the membrane properties of reactive astrocytes at D14, namely the appearance of *Hcn1–4*, which so far have been described only in neurons [36]. On the other hand, a significant negative correlation was found between the expressions of *Cspg4* or *Pdgfra* and *Glul, Aqp9, Hcn1, Grin2a* and *Gria3* (r<-0.55). This clearly demonstrates that astrocytes from D14 (B3 subpopulation) lose the expression of *Cspg4* and *Pdgfra* while they change into reactive astrocytes comprising a permanent glial scar.

The calculated partial Spearman correlation coefficients surprisingly indicated that there was no direct interaction between *Aqp9* and *Aqp1* (Figure S4B), but the observed correlation can be accounted for as being mediated via *Hcn2* or *Kcna3*. Further, *Aqp9* does not mediate any interaction between *Hcn1, Hcn2, Hcn3, Gria3, Grm5* and *Aqp1*. Out of the total 21 possible direct interactions, *Gria3* and *Hcn2* were involved in 10 (Figure S4C and S2D).

Further classification of the reactive glial cells in subpopulation B3 using SOM revealed that it is not a homogenous group; rather there are two subgroups, which we refer to as B3A and B3B. Their existence was clearly confirmed by PCA (Figure 6A). Viewing the data in a 3-dimensional PC1 vs. PC2 vs. PC3 space shows that the B3A/B3B subpopulations separate along PC3 (Figure 6B). Hence, the genes distinguishing B3A from B3B are those with high

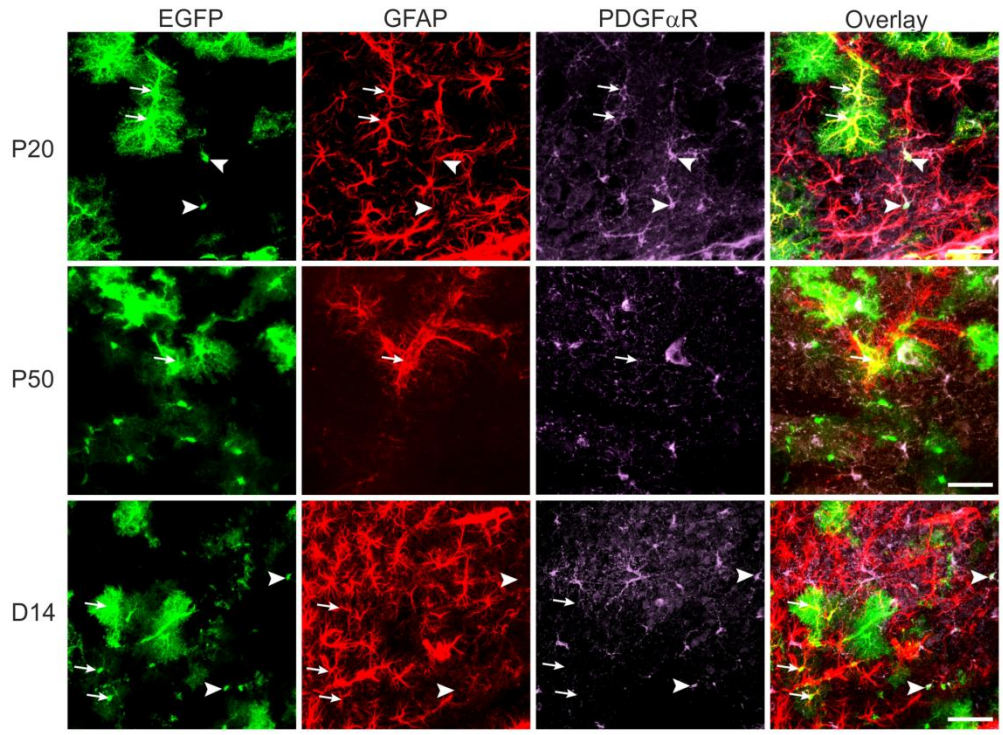


Figure 4. Immunohistochemical analysis of EGFP$^+$ cells. P20, P50, and D14 cells expressing EGFP (green), GFAP (red), and PDGFαR (violet). Arrowheads indicate cells co-expressing EGFP and PDGFαR, and arrows indicate cells positive for EGFP and GFAP. Note, at P20 there are cells positive for all three markers. Scale bars, 50 µm.
doi:10.1371/journal.pone.0069734.g004

weights in PC3, thus being different from the genes distinguishing B3 from B1/B2, which are those with high weights in PC1 and PC2. The expression of genes in the B3A and B3B subgroups were compared using the non-parametric Mann-Whitney test, and those with significant p-values after Bonferroni correction for multiple testing ($p<0.00109$) are shown in Figure 6C. The most important differences between the B3A and B3B subgroups are the down-regulation of *Pdgfra*, *Cspg4* and *Grik2* and the up-regulation of *Eaat1*, *Glul*, *Aqp1*, *Aqp4*, *Aqp9*, *Hcn1*, *Hcn2*, *Hcn3*, *Grin2a*, *Gria3*, *Grm3*, *Grm5* and *Kcna3* (Figure 6C).

Similarly as in the early stages of development, post-ischemic astrocytes also co-expressed *Gfap* and *Cspg4*/*Pdgfra* (approximately 40%); however, on the protein level we did not detect such co-expression (Figure 4).

Discussion

We show that cortical EGFP$^+$ cells collected during postnatal development as well as cells collected after focal cerebral ischemia are highly heterogeneous with respect to their gene expression profiles. Our findings demonstrate that the issue of astrocyte heterogeneity is complicated, but can be approached with single cell expression profiling, which reveals and distinguishes cell subpopulations. With single cell resolution we can identify rare cells, an approach impossible in classical bulk measurements [37]. Furthermore, single cell profiling avoids the ambiguity of normalization [38].

Astrocyte Heterogeneity during Postnatal Development

We have shown that during postnatal development, the number of EGFP$^+$ cells expressing the astrocytic markers *Glul* and/or *Eaat1* increases towards adulthood, while the number of cells expressing *Cspg4* and/or *Pdgfra* is reduced. Similarly to previous findings, we show that in addition to *Glul*, the expression of *Eaat1*, *Aqp4* and *Grin2a* is induced during development [39]. This is an opposite tendency to that reported by Stanimirovic and colleagues [40], but their experiments were carried out *in vitro*, in a purely astrocytic culture, where glutamate uptake and glutamine synthesis might be compromised. We also found a substantial decrease in *Grik2*, *Grik3*, *Gria2*, *Gria3* and *Gria4* expression during development, which is in agreement with previous data demonstrating the expression of kainate receptors in glial precursor cells, while mature astrocytes show an NMDA-specific responses [41,42]. Similar decline in *Grik1-5* expression during development was described recently by Cahoy and colleagues (Cahoy et al., 2008). In agreement with the findings of Matthias and colleagues [43], we found that EGFP$^+$ cells that express *Cspg4* and/or *Pdgfra* also express *Gria2-4* at high levels, while the low expression of *Gria2-4* is observed in EGFP$^+$

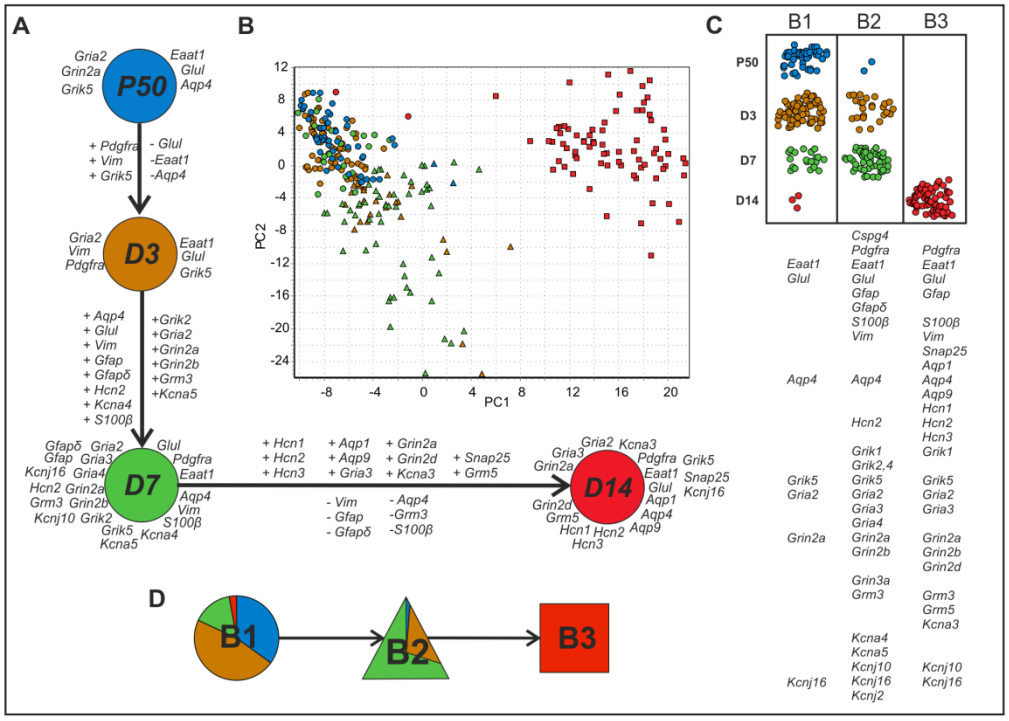


Figure 5. Change in the gene expression of EGFP$^+$ cells after MCAO. (**A**) Changes in the gene expression of highly expressed (log_2 relative average expression above 2.3 from all data) astrocytic/NG2 glia markers and membrane proteins at P50 (blue), D3 (orange), D7 (green), and D14 (red). The listed genes show at least 2-fold up (+) or 2-fold downregulation (−) between sequential stages. (**B**) PCA clustering of cells from all post-ischemic stages and P50 (data mean-centered along genes). The stages are indicated in color (P50 blue, D3 orange, D7 green, and D14 red), and the three groups identified by SOM are indicated by symbols (squares for B3; circles for B1 and triangles for B2). (**C**) The incidence of cells from the three subpopulations at post-ischemic stages and P50. The genes under the table indicate highly expressed genes (log2 relative average expression above 1.8 and at least 40% positive cells) in particular subpopulations. (**D**) The distribution of control and post-ischemic stages in individual subpopulations.
doi:10.1371/journal.pone.0069734.g005

astrocytes that do not express markers of polydendrocytes. These authors also claimed that EGFP$^+$ cells co-express glutamate receptors and transporters, which is supported by our single cell data. As shown in Table 1, we found the co-expression of glutamate receptors and glutamate transporters in individual EGFP$^+$ cells. Among cells collected at the postnatal developmental stages P10-P50, we distinguish 3 subpopulations here referred to as A1, A2 and A3 (Figure 3). A3 cells exhibit features previously described in mature astrocytes with very high expression of *Eaat1*, *Glul* and *Aqp4* [23]. Cells in A1 also express *Eaat1*, but they also have high levels of *Cspg4* and *Pdgfra*, which are traditionally associated with NG2 glia [44]. Nevertheless, *Cspg4* expression in P1-P17 astrocytes was already described elsewhere (Cahoy et al., 2008). Subpopulation A2 is the most unique with high expression of *Glul*, *Hcn2*, *Gria3*, and *Grin2a*, while the transcriptional activity of the other studied genes is rather low. Based primarily on the elevated expression of *Glul* and the low *Gria3* levels, we hypothesize that A2 is an intermediate stage between A1 and A3 (Table S2). This hypothesis is also supported by the prevalence of A2 cells at P20. It is still a matter of debate whether astrocytes can be derived from CSPG4- and/or PDGFαR-positive cells; some recent evidence suggests this possibility following CNS injuries [9,12]. Interestingly, the A1 subpopulation, mostly comprising cells from P10–20, is characterized by the increased expression of subunits of ionotropic AMPA/kainate receptors, such as *Grik1-3*, *Grik5*, *or Gria2-4*, which makes these cells sensitive to AMPA/kainate receptor–mediated glutamate injury. The presence of *Gria2* encoding GluA2 subunits suggests that AMPA receptors in the A1 subpopulation might be impermeable for Ca^{2+}. The subpopulations A3 and B1, which mainly represent mature astrocytes at P30-P50, show besides the expression of classical markers such as *Eaat1*, *Glul* and *Aqp4*, also the dominant expression of *Gria2*, suggesting AMPA receptor impermeability to Ca^{2+}. Moreover, *Kcnj10* and *Kcnj16*, encoding inwardly rectifying Kir4.1 and Kir5.1 channels, are both highly expressed in immature cortical glia, while in mature astrocytes *Kcnj16* prevails. Since heteromeric Kir4.1/5.1 and homomeric Kir4.1 have distinct ion channel properties, such differences in the expression of K^+ channels that participate in K^+ uptake might result in the diverse ability of astrocytes to take up/distribute high extracellular K^+ during increased neuronal activity or in pathological states, such as ischemia or trauma. Notably, even though mature/immature

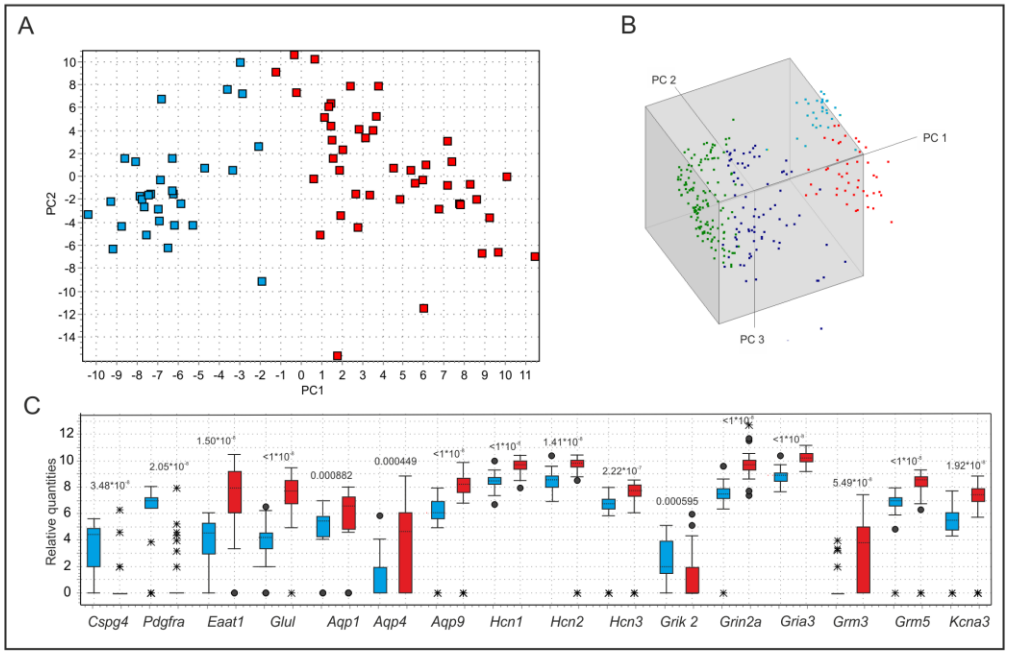


Figure 6. Classification of reactive glia (B3) into two subgroups. (A) PCA scatter plot of the cells in subpopulation B3 (reactive glia). Colors indicate the two groups (B3A red and B3B blue) obtained by decomposing the data by SOM analysis (data are mean-centered). (B) Three-dimensional PCA scatter plot of all post-ischemic cells and P50 reflecting the normal state for comparison - B1 (green), B2 (dark blue), B3A (red) and B3B (light blue). PC3 separates B3A from B3B. The first PC accounts for 32%, the second 14% and the third 8%, of the total variance in the data. (C) Box plots showing genes differentially expressed in B3A (red) and B3B (blue). The box indicates the median value and the 25th and 75th percentiles, the whiskers are 1.5× the distance of the 25th and 75th percentiles, the circles indicate outliers and the crosses extreme outliers. p-values (Mann-Whitney) are indicated for each gene.
doi:10.1371/journal.pone.0069734.g006

astrocytes have been shown to display different K^+ current profiles during development in the rat hippocampal CA1 region [45], such as outward rectifying, variably rectifying and passive current patterns, we found no significant differences in the transcription activities of inwardly and outwardly rectifying K^+ channels or K2P channels between the A1-A3 subpopulations. We found the expression of *Kcnj10* (encoding the Kir4.1 channel) in the $EGFP^+$ cells to be rather constant, while higher levels of *Kcnj16* (encoding Kir5.1) transcripts were found in cells from subpopulations A1 and A3. We did not see any correlation between *Aqp4* and *Kcnj10* expression on the cellular level in any of the subpopulations in agreement with previous data showing the lack of any direct interaction between Kir4.1 and AQP4 [46,47] and their different regulations of expression [48]. In agreement with recent findings [25,49] we showed that $EGFP^+$ mature cortical astrocytes from P30 and P50 (comprising the B3 subpopulation) express low levels of *Grm5,* as a result of a significant decrease in *Grm5* expression towards astrocytic maturation, and a high level of *Grm3*. Nedergaard and colleagues suggested that mGluR5 activity underlies glutamate–induced Ca^{2+} signalling only in immature postnatal astrocytes, while in mature astrocytes other transmitters, such as endocannabinoids, purines, norepinephrine, and acetylcholine, as well as changes in extracellular Ca^{2+}, may trigger astrocytic Ca^{2+} signalling. The activation of mGluR3, a Gi/Go receptor negatively coupled to adenylate cyclase, probably does not participate in glutamate-induced Ca^{2+} signalling as its activation did not trigger a Ca^{2+} increase in adult astrocytes [49]. Since the activation of mGluR3 inhibits cyclic adenosine monophosphate (cAMP) formation, it might participate in the regulation of processes such as astrocyte proliferation, the expression of glutamate receptor GLAST or GFAP expression [50,51].

Gene Expression in Post-ischemic Astrocytes

After CNS injury, such as trauma or ischemia, or in CNS neurodegenerative disorders, the properties of astrocytes change; the cells become reactive [52,53]. This is in line with our finding that the gene expression profile of astrocytes alters following focal ischemia. Particularly affected is the expression of *Gfap, Gfapδ, S100b* and *Vim*, which was also observed previously [13,53]. Fourteen days after ischemia, *Aqp1* and *Aqp9* were up-regulated as described earlier [14,15], and similarly to previous reports, we found upregulated *Grm5* [18,19]. Generally, *Aqp1* and/or *Aqp9* are elevated in hypertrophied, GFAP overexpressing astrocytes in a number of pathological states, such as ischemic or traumatic brain injury, Alzheimer's or Creutzfeldt-Jakob diseases [54–56]. In the present study, we detected their increased expression only in the B3 subpopulation, which comprises mostly $EGFP^+$ cells 14 days

after MCAO. Such a delayed increase in their expression suggests that they are probably not implicated in water movements during oedema formation/resolution, but they might play a role as metabolite channels facilitating the diffusion of glycerol and lactate, as suggested previously [57,58]. Moreover, the ischemic brain undergoes marked remodelling, which involves the migration of neighbouring cells to the injury site, thus the increased *Aqp1* expression may reflect a migratory phenotype of reactive astrocytes as AQP1 was shown to promote cell migration in epithelial cells [59] and glial tumors [60,61]. The B2 and B3 subpopulations also highly express AMPA receptor subunits, especially *Gria3*, which suggests that AMPA receptors may significantly contribute to Na^+/Ca^{2+} influx into the reactive astrocytes, and together with the strong expression of metabotropic glutamate receptors (*Grm1, Grm5*) these receptors possibly mediate Ca^{2+} oscillations and glutamate release, as mentioned previously [62]. Furthermore, an increase in *Grm5* expression in reactive astrocytes might lead to the increased activity of mGluR5 and thus contribute to astrocytic apoptosis, as described recently [63]. On the other hand, an anti-apoptotic effect mediated by mGluR3 activation that leads to reduced intracellular levels of cAMP was described in cultured astrocytes after their exposure to nitric oxide [64]. In addition, Aguado and colleagues [65] demonstrated that the activation of ionotropic glutamate receptors is required for the generation of correlated astrocytic network activity, essential for neuronal development and synaptic plasticity. In mature astrocytes under physiological conditions (P50), the Ca^{2+} permeability of AMPA receptors is diminished compared to immature stages (P10) due to the predominant expression of *Gria2*. Interestingly, mRNAs encoding kainate receptor subunits (*Grik1,2,3,5*) are also significantly increased in the B2 and B3 subpopulations; however, their function in reactive astrocytes and those from immature (P10) animals has not yet been elucidated.

Marked changes were also observed in the expression of different types of ion channels, such as outwardly and inwardly rectifying K^+ channels and chloride channels (ClC2), and moreover, Trpv4 expression almost doubled after ischemia, when comparing the B2 and B3 subpopulations to B1. Despite the fact that the expression of *Kcna3-5*, encoding the outwardly rectifying K^+ channels Kv1.3, Kv1.4 and Kv1.5, was comparable in the A1–3 subpopulations during development, it markedly increased in the post-ischemic B2 and B3 subpopulations. Such an increase might point to the fact that astrocytes enter the cell cycle and proliferate following ischemic injury, as shown previously [66]. In addition, an increase in *Clc2* expression could suggest the participation of astrocytes in delivering Cl^- to the site of intensive GABAergic transmission [67], or Cl^- efflux can play a role in astrocyte proliferation/migration as suggested for gliomas [68]. Furthermore, enhanced *Trpv4* expression following MCAO was detected in both the B2 and B3 subpopulations, which is in agreement with our recent data demonstrating an increased expression/function of TRPV4 following hypoxic/ischemic injury of the rat hippocampus [69]. TRPV4 upregulation might originate from its role as an osmosensor as shown in astrocytes [70] or from a possible role of this channel in Ca^{2+} signalling in post-ischemic astrocytes, which may result in astrocytic cell death [71]. A transient increase in *Nes* observed in cells 3 and 7 days after MCAO might support the hypothesis that these subpopulations comprise proliferative/newly derived cells. A striking finding is the high expression of *Hcn1-4*, which encodes hyperpolarization–activated non-specific cationic channels permeable for K^+/Na^+, so far described only in neurons. Despite the fact that their role in astrocytes is unknown, we hypothesize that they might have a role in determining the astrocyte resting membrane potential or they might interact with the glutamate release machinery as described recently for presynaptic terminals [72]. A question that still remains incompletely elucidated is the origin of reactive glia. Certain subpopulations of reactive astrocytes have been shown to arise from NG2 glia [73] or from proliferating astrocytes [74]. Based on our findings, we propose that the B2 subpopulation, which is characterized by an increase in the expression of several astrocytic markers including *Gfap, S100b, Gfapδ* and *Vim*, but also by a marked increase in *Nes, Pdgfra* and *Cspg4*, comprises astrocytes in an "intermediate stage" that increase in number as astrogliosis develops (Figure 5). We hypothesize that either NG2 glia give rise to the B2 subpopulation, which eventually develops into fully reactive astrocytes with a distinct expression profile as in subpopulation B3, or they might represent dedifferentiated mature astrocytes with increased proliferative activity. The latter possibility is supported by the fact that they display a relatively similar gene expression profile as $EGFP^+$ cells of the immature cortex (Figure 3). Interestingly, B1 and B2 show some similarity (e.g., the expression of *Pdgfra, Cspg4, Grik1-3, Grik5, Gria2-4, Grin2a* and *Kcnj16*) to the A1 subpopulation observed during early postnatal development (Table S2). Using partial Spearman correlation calculations, we did not identify any master gene in B3 that would account for all of the changes observed upon astrocyte reactivity that follows ischemic injury; however, it is clear that *Gria3* and *Hcn2* are of high importance. In subpopulation B3 we identified several, previously not recognized highly expressed genes: *Hcn1-3, Grin2a, Gria2, Gria3, Grm5* and *Kcna3*. We also found the high expression of *Snap25*, which is expressed only minimally in the other subpopulations and has been suggested to be a marker of cells with exocytotic capabilities under physiological conditions [25]. Similar observations were made in rats [75], where SNAP25 was found upregulated 3 and 12 days after injury. Despite the fact that synaptosomal-associated protein SNAP-25 has not been detected in astrocytes but SNAP-23 [76], our data revealed mRNA levels coding SNAP-25 were significantly increased in cortical astrocytes 14 days post-ischemia (B3 subpopulation). As astrocytes are known to release glutamate and aspartate in response to elevated intracellular calcium levels via a vesicular release mechanism in which SNARE proteins are implicated, high levels of *Snap25* might be a sign of a significant glutamate release from reactive astrocytes. We found a positive correlation between the expressions of *Snap25, Eaat1* and *Glul* and the absence of any correlation with *Cspg4* and *Pdgfra* in B3. We also found a strong correlation between *Eaat1* and *Aqp4* expression, which is in accordance with a recent report suggesting the close relation of excitatory amino acid transporters (*Eaat*) and *Aqp4* under pathological conditions with the massive release of glutamate [14]. Although an association between AQP4 and mGluR5 has been reported [77], we found no correlation on the gene expression level. On the other hand, we did find an extremely strong correlation between *Aqp1/Aqp9* and *Grm5* levels in B3.

Heterogeneity of reactive astrocytes in the vicinity of an ischemic lesion has been described previously [22,78,79]. Even within the B3 subpopulation of reactive astroglia, we found evidence for two subgroups using SOM and PCA: B3A and B3B. B3A is positive for traditional astrocytic markers, while B3B has a substantially higher expression of traditional NG2 glia markers (Figure 6). *Cspg4* has been proposed as a hallmark of reactive astrocytes [16]. Even though the entire B3 population has high overall transcriptional activity, much higher than B1 and B2, there is a difference in the expression levels of many genes between the B3A and B3B subgroups. Since we isolated cells from the vicinity of an ischemic lesion, it is possible that B3 comprises cells permanently converted to reactive glia, which are in different

states or at different distances from the site of the ischemic lesion. Employing Affymetrix GeneChip arrays, the gene expression profiling of reactive astrocytes by Zamanian and colleagues [53] provided transcriptome databases for two subtypes of reactive astrocytes that will be highly useful in generating new and testable hypotheses of their function, as well as for providing new markers to detect different types of reactive astrocytes.

Co-expression of PDGFαR and GFAP in EGFP$^+$ Cells

Our finding that at P10 almost all EGFP$^+$ cells express *Cspg4* and *Pdgfra* and that EGFP$^+$ cells co-express *Gfap* and *Cspg4*/*Pdgfra* was confirmed by immunohistochemistry. (Figure 4, Figure S3); [35]. Similarly, Cahoy and co-authors [25] also detected the high expression of *Cspg4* in immature astrocytes (P7, P17), which declined with astrocyte maturation; however, they never detected *Pdgfra* expression in astrocytes, but rather *Pdgfrb*, which is known to be expressed in pericytes [80]. Although GFAP is a traditional marker of astrocytes and we collected cells where the expression of EGFP is driven by a GFAP promoter, its mRNA level was very low, not at all comparable with the expression of *Cspg4* and *Pdgfra*. The GFAP protein must therefore be stable with slow turnover in the cortex. Nevertheless, *Gfap* and *Gfapδ* increased with astrocyte maturation similarly to the highly expressed *Glul, Eaat1* and *Aqp4*. Interestingly, in subpopulations A1, B2 and B3, *Cspg4* and *Pdgfra* are co-expressed on the mRNA level in 95% of cells, but the correlation between their expressions on the single cell level is not particularly strong ($r = 0.44$, $p = 0.001$ for A1, $r = 0.23$, $p = 0.151$ for B2, $r = 0.27$, $p = 0.159$ for B3). *Cspg4* and *Pdgfra* also show quite different correlations with other genes. This suggests *Cspg4* and *Pdgfra* are expressed independently of each other in individual cells, even in cells from the same subpopulation. The absence of correlation between highly expressed genes in individual cells of the same kind has been observed previously [33] and has been interpreted as the cells having independent transcription regulatory mechanisms, which leads to independent transcriptional bursts.

We do not think that the expression of *Cspg4* or *Pdgfra* in EGFP/GFAP cells isolated from the cortex of P10 animals was detected simply due to sensitivity in which low transcript levels are amplified. PDGFαR and NG2 are expressed in these cells also on the protein level as demonstrated by immunohistochemistry in Figure 4 for PDGFαR and in Figure S3 for NG2. These images show without any doubt that EGFP-positive cells (P10–20 mice) co-express GFAP and PDGFαR or NG2, the expression of which disappears at P50. More interestingly, in post-ischemic tissue the expression of both *Cspg4* and *Pdgfra* reappears; however, their expression on the protein level does not coincide with EGFP/GFAP-positive reactive astrocytes, which might suggest that newly generated reactive astrocytes lose both NG2 or PDGFαR when they reach the stage of fully developed reactive astroglia 14 days post ischemia, while those still expressing NG2 and/or PDGFαR might partially comprise cells in transition into reactive astrocytes or dedifferentiated astrocytes that display the gene expression profile of immature P10-P20 astrocytes (A1 subpopulation). This is also obvious from the division of the B3 subpopulation into two groups: B3A and B3B (Figure 6). Cells in the B3A group express *Cspg4* and *Pdgfra*, while the expression of *Eaat1, Glul* and *Aqp4* is significantly lower than in the B3B group. Therefore, NG2 or PDGFαR-positive cells may reflect astrocytes in a transitional stage similar to that described by Zhu and colleagues [81]. They showed that a subpopulation of GS$^+$ astrocytes is the progeny of NG2 cells, thus confirming the previous finding that NG2 cells give rise to protoplasmic astrocytes. They also showed the co-expression of aldehyde dehydrogenase (*Aldh1L1*, another marker of astrocytes) immunoreactivity in their somata and faint PDGFαR immunoreactivity in their processes, indicative of transitional cells that are losing PDGFαR and acquiring Aldh1L1. Similar data showing the co-expression of GFAP and NG2 in ischemic tissue were described in our recent publication [12].

The Basis of Astrocyte Heterogeneity

The heterogeneity of astrocytes even within an individual brain region, such as the cortex, might already arise during their development. Some astrocytes originate from cells in the ventricular zone via radial glia intermediates, while other astrocytes arise from immature cells that migrate in the perinatal period from the dorso-ventral subventricular zone into the brain [82]. Radial glia, which give rise to distinct subsets of forebrain neurons and later on to glial cells, must be highly heterogeneous themselves with respect to their progenitor function, which strongly depends on the expression of different sets of transcription factors such as Emx1 and Dach1. Progenitors expressing these transcription factors give rise, besides cortical and hippocampal CA1 neurons, to a subset of cortical and hippocampal astrocytes [83–85] and also to a subpopulation of astroglia in the dorsal roof of the subventricular zone [86]. How the developmental heterogeneity of astrocytes relates to their functional heterogeneity still remains unresolved. In addition, astrocytes might also display a spatial heterogeneity as they surround blood vessels and/or different synapses. Nonetheless, a diversity of astrocytic responses to pathological stimuli, such as oxygen-glucose deprivation or hypotonic stress, has been demonstrated in the cortex [5,87]. As reported earlier [88], there also exists a subpopulation of astrocytes that do not express GFAP and that were not considered in this study.

Conclusions

We have used single cell expression profiling and multivariate expression analysis to identify and characterize subpopulations of astrocytes present during postnatal development and after focal cerebral ischemia. We found that transcriptional activity decreases during development and that fully mature astrocytes express mainly *Eaat1, Glul, Aqp4* and *Kcnj10* and *Kcnj16*. After a brain injury the situation rapidly changes - the astrocytes become reactive and express *Eaat1, Glul, Aqp1, Aqp9, Snap25,* hyperpolarization-activated cation channels *Hcn1, Hcn2, Hcn3,* glutamate receptors *Grin2a, Gria2, Gria3, Grm5* and potassium channels *Kcna3* and *Vim*.

Supporting Information

Figure S1 Validation of pre-amplification. Three separate RNA samples were isolated from the cortex of three different mice, transcribed and preamplified. cDNA levels were measured with preamplification (P) and without preamplification (NP). The average difference (ΔCq) and the standard deviation of the difference between P and NP for each gene were calculated for each of the three RNA samples.
(TIF)

Figure S2 Schemas of all correlations higher than 0.6 for subpopulation A1 (A), subpopulation A2 (B), subpopulation A3 (C), subpopulation B1 (D), subpopulation B2 (E), and subpopulation B3 (F). Positive correlations are indicated by green lines, while red lines indicate negative correlations.
(TIF)

Figure S3 Immunohistochemical analysis of EGFP$^+$ cells. The expression of GFAP and NG2 was analyzed in the cortex of 10 days-old EGFP/GFAP mice. Note that at P10 there are EGFP-positive cells co-expressing GFAP and NG2.
(TIF)

Figure S4 Interactions between highly expressed "key" genes in subpopulation B3. (**A**) All significant correlations ($p<0.05$) between the expressions of gene pairs are indicated. (**B**) Partial Spearman correlation coefficients were calculated to separate direct and indirect interactions between genes. The lines indicate a direct correlation that remains significant after the removal of indirect (via a third gene) correlations. (**C**) Correlation that remains significant after removing the effect of Hcn2. (**D**) Correlation that remains significant after removing the effect of Gria3.
(TIF)

Table S1 PCR assay information and primer sequences.
(DOCX)

Table S2 Statistics describing the gene expression for each postnatal developmental subpopulation or post-ischemic subpopulations.
(DOCX)

Table S3 Correlation within development subpopulation A1.
(XLSX)

Table S4 Correlation within development subpopulation A2.
(XLSX)

Table S5 Correlation within development subpopulation A3.
(XLSX)

Table S6 Correlation within development subpopulation B1.
(XLSX)

Table S7 Correlation within development subpopulation B2.
(XLSX)

Table S8 Correlation within development subpopulation B3.
(XLSX)

Author Contributions

Conceived and designed the experiments: VR MK MA. Performed the experiments: VR PH DD. Analyzed the data: VR AS MK MA. Contributed reagents/materials/analysis tools: MK MA. Wrote the paper: VR MA.

References

1. Reichenbach WH (2005) Astrocytes and ependymal cells. In: Kettenmann H, Ransom BR, editor. Neuroglia. 2 ed. New York: Oxford University Press. 19 36.
2. Oberheim NA, Takano T, Han X, He W, Lin JH, et al. (2009) Uniquely hominid features of adult human astrocytes. The Journal of neuroscience : the official journal of the Society for Neuroscience 29: 3276 3287.
3. Matyash V, Kettenmann H (2010) Heterogeneity in astrocyte morphology and physiology. Brain research reviews 63: 2 10.
4. Seifert G, Huttmann K, Binder DK, Hartmann C, Wyczynski A, et al. (2009) Analysis of astroglial K+ channel expression in the developing hippocampus reveals a predominant role of the Kir4.1 subunit. The Journal of neuroscience : the official journal of the Society for Neuroscience 29: 7474 7488.
5. Benesova J, Rusnakova V, Honsa P, Pivonkova H, Dzamba D, et al. (2012) Distinct expression/function of potassium and chloride channels contributes to the diverse volume regulation in cortical astrocytes of GFAP/EGFP mice. PLoS One 7: e29725.
6. Ge WP, Miyawaki A, Gage FH, Jan YN, Jan LY (2012) Local generation of glia is a major astrocyte source in postnatal cortex. Nature 484: 376 380.
7. Freeman MR (2010) Specification and morphogenesis of astrocytes. Science 330: 774 778.
8. Zhang Y, Barres BA (2010) Astrocyte heterogeneity: an underappreciated topic in neurobiology. Current opinion in neurobiology 20: 588 594.
9. Trotter J, Karram K, Nishiyama A (2010) NG2 cells: Properties, progeny and origin. Brain research reviews 63: 72 82.
10. Stallcup WB, Beasley L (1987) Bipotential glial precursor cells of the optic nerve express the NG2 proteoglycan. The Journal of neuroscience : the official journal of the Society for Neuroscience 7: 2737 2744.
11. Alonso G (2005) NG2 proteoglycan-expressing cells of the adult rat brain: possible involvement in the formation of glial scar astrocytes following stab wound. Glia 49: 318 338.
12. Honsa P, Pivonkova H, Dzamba D, Filipova M, Anderova M (2012) Polydendrocytes display large lineage plasticity following focal cerebral ischemia. PLoS One 7: e36816.
13. Robel S, Berninger B, Gotz M (2011) The stem cell potential of glia: lessons from reactive gliosis. Nature reviews Neuroscience 12: 88 104.
14. Zelenina M (2010) Regulation of brain aquaporins. Neurochemistry international 57: 468 488.
15. Hwang IK, Yoo KY, Li H, Lee BH, Suh HW, et al. (2007) Aquaporin 9 changes in pyramidal cells before and is expressed in astrocytes after delayed neuronal death in the ischemic hippocampal CA1 region of the gerbil. Journal of neuroscience research 85: 2470 2479.
16. McKeon RJ, Jurynec MJ, Buck CR (1999) The chondroitin sulfate proteoglycans neurocan and phosphacan are expressed by reactive astrocytes in the chronic CNS glial scar. The Journal of neuroscience : the official journal of the Society for Neuroscience 19: 10778 10788.
17. Komitova M, Perfilieva E, Mattsson B, Eriksson PS, Johansson BB (2002) Effects of cortical ischemia and postischemic environmental enrichment on hippocampal cell genesis and differentiation in the adult rat. Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism 22: 852 860.
18. Ferraguti F, Corti C, Valerio E, Mion S, Xuereb J (2001) Activated astrocytes in areas of kainate-induced neuronal injury upregulate the expression of the metabotropic glutamate receptors 2/3 and 5. Experimental brain research Experimentelle Hirnforschung Experimentation cerebrale 137: 1 11.
19. Ulas J, Satou T, Ivins KJ, Kesslak JP, Cotman CW, et al. (2000) Expression of metabotropic glutamate receptor 5 is increased in astrocytes after kainate-induced epileptic seizures. Glia 30: 352 361.
20. Jacque CM, Vinner C, Kujas M, Raoul M, Racadot J, et al. (1978) Determination of glial fibrillary acidic protein (GFAP) in human brain tumors. Journal of the neurological sciences 35: 147 155.
21. Sofroniew MV, Vinters HV (2010) Astrocytes: biology and pathology. Acta neuropathologica 119: 7 35.
22. Walz W (2000) Controversy surrounding the existence of discrete functional classes of astrocytes in adult gray matter. Glia 31: 95 103.
23. Regan MR, Huang YH, Kim YS, Dykes-Hoberg MI, Jin L, et al. (2007) Variations in promoter activity reveal a differential expression and physiology of glutamate transporters by glia in the developing and mature CNS. The Journal of neuroscience : the official journal of the Society for Neuroscience 27: 6607 6619.
24. Mearow KM, Mill JF, Vitkovic L (1989) The ontogeny and localization of glutamine synthetase gene expression in rat brain. Brain research Molecular brain research 6: 223 232.
25. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. The Journal of neuroscience : the official journal of the Society for Neuroscience 28: 264 278.
26. Kimelberg HK (2004) The problem of astrocyte identity. Neurochemistry international 45: 191 202.
27. Klein CA, Zohlnhofer D, Petat-Dutter K, Wendler N (2005) Gene expression analysis of a single or few cells. Current protocols in human genetics/editorial board, Jonathan L Haines [et al] Chapter 11: Unit 11 18.
28. Ståhlberg A, Andersson D, Aurelius J, Faiz M, Pekna M, et al. (2011) Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. Nucleic acids research 39: e24.
29. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS biology 4: e309.
30. Raj A, van Oudenaarden A (2009) Single-molecule approaches to stochastic gene expression. Annual review of biophysics 38: 255 270.

31. Bengtsson M, Hemberg M, Rorsman P, Stahlberg A (2008) Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. BMC molecular biology 9: 63.
32. Ståhlberg A, Rusnakova V, Forootan A, Anderova M, Kubista M (2013) RT-qPCR work-flow for single-cell data analysis. Methods 59: 80 88.
33. Ståhlberg A, Bengtsson M (2010) Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. Methods 50: 282 288.
34. Nolte C, Matyash M, Pivneva T, Schipke CG, Ohlemeyer C, et al. (2001) GFAP promoter-controlled EGFP-expressing transgenic mice: a tool to visualize astrocytes and astrogliosis in living brain tissue. Glia 33: 72 86.
35. Benesova J, Hock M, Butenko O, Prajerova I, Anderova M, et al. (2009) Quantification of astrocyte volume changes during ischemia in situ reveals two populations of astrocytes in the cortex of GFAP/EGFP mice. Journal of neuroscience research 87: 96 111.
36. Wahl-Schott C, Biel M (2009) HCN channels: structure, cellular regulation and physiological function. Cell Mol Life Sci 66: 470 494.
37. Ståhlberg A, Rusnakova V, Kubista M (2013) The added value of single-cell gene expression profiling. Brief Funct Genomics.
38. Sindelka R, Ferjentsik Z, Jonak J (2006) Developmental expression profiles of Xenopus laevis reference genes. Dev Dyn 235: 754 758.
39. da Cunha A, Aloyo VJ, Vitkovic L (1991) Developmental regulation of GAP-43, glutamine synthetase and beta-actin mRNA in rat cortical astrocytes. Brain research Developmental brain research 64: 212 215.
40. Stanimirovic DB, Ball R, Small DL, Muruganandam A (1999) Developmental regulation of glutamate transporters and glutamine synthetase activity in astrocyte cultures differentiated in vitro. International journal of developmental neuroscience : the official journal of the International Society for Developmental Neuroscience 17: 173 184.
41. Palygin O, Lalo U, Pankratov Y (2011) Distinct pharmacological and functional properties of NMDA receptors in mouse cortical astrocytes. Br J Pharmacol 163: 1755 1766.
42. Ziak D, Chvatal A, Sykova E (1998) Glutamate-, kainate- and NMDA-evoked membrane currents in identified glial cells in rat spinal cord slice. Physiological research/Academia Scientiarum Bohemoslovaca 47: 365 375.
43. Matthias K, Kirchhoff F, Seifert G, Huttmann K, Matyash M, et al. (2003) Segregated expression of AMPA-type glutamate receptors and glutamate transporters defines distinct astrocyte populations in the mouse hippocampus. The Journal of neuroscience : the official journal of the Society for Neuroscience 23: 1750 1758.
44. Butt AM, Hamilton N, Hubbard P, Pugh M, Ibrahim M (2005) Synantocytes: the fifth element. Journal of anatomy 207: 695 706.
45. Zhou M, Schools GP, Kimelberg HK (2006) Development of GLAST(+) astrocytes and NG2(+) glia in rat hippocampus CA1: mature astrocytes are electrophysiologically passive. Journal of neurophysiology 95: 134 143.
46. Nagelhus EA, Horio Y, Inanobe A, Fujita A, Haug FM, et al. (1999) Immunogold evidence suggests that coupling of K+ siphoning and water transport in rat retinal Muller cells is mediated by a coenrichment of Kir4.1 and AQP4 in specific membrane domains. Glia 26: 47 54.
47. Connors NC, Adams ME, Froehner SC, Kofuji P (2004) The potassium channel Kir4.1 associates with the dystrophin-glycoprotein complex via alpha-syntrophin in glia. The Journal of biological chemistry 279: 28387 28392.
48. Liu H, Yang M, Qiu GP, Zhuo F, Yu WH, et al. (2012) Aquaporin 9 in rat brain after severe traumatic brain injury. Arquivos de neuro-psiquiatria 70: 214 220.
49. Sun W, McConnell E, Pare JF, Xu Q, Chen M, et al. (2013) Glutamate-dependent neuroglial calcium signaling differs between young and adult brain. Science 339: 197 200.
50. Lyon L, Kew JN, Corti C, Harrison PJ, Burnet PW (2008) Altered hippocampal expression of glutamate receptors and transporters in GRM2 and GRM3 knockout mice. Synapse 62: 842 850.
51. Masood K, Besnard F, Su Y, Brenner M (1993) Analysis of a segment of the human glial fibrillary acidic protein gene that directs astrocyte-specific transcription. J Neurochem 61: 160 166.
52. Pekny M, Nilsson M (2005) Astrocyte activation and reactive gliosis. Glia 50: 427 434.
53. Zamanian JL, Xu L, Foo LC, Nouri N, Zhou L, et al. (2012) Genomic analysis of reactive astrogliosis. The Journal of neuroscience : the official journal of the Society for Neuroscience 32: 6391 6410.
54. Hoshi A, Yamamoto T, Shimizu K, Ugawa Y, Nishizawa M, et al. (2012) Characteristics of aquaporin expression surrounding senile plaques and cerebral amyloid angiopathy in Alzheimer disease. J Neuropathol Exp Neurol 71: 750 759.
55. McCoy E, Sontheimer H (2010) MAPK induces AQP1 expression in astrocytes following injury. Glia 58: 209 217.
56. Rodriguez A, Perez-Gracia E, Espinosa JC, Pumarola M, Torres JM, et al. (2006) Increased expression of water channel aquaporin 1 and aquaporin 4 in Creutzfeldt-Jakob disease and in bovine spongiform encephalopathy-infected bovine-PrP transgenic mice. Acta Neuropathol 112: 573 585.
57. Badaut J, Petit JM, Brunet JF, Magistretti PJ, Charriaut-Marlangue C, et al. (2004) Distribution of Aquaporin 9 in the adult rat brain: preferential expression in catecholaminergic neurons and in glial cells. Neuroscience 128: 27 38.
58. Badaut J, Regli L (2004) Distribution and possible roles of aquaporin 9 in the brain. Neuroscience 129: 971 981.
59. Hara-Chikuma M, Verkman AS (2006) Aquaporin-1 facilitates epithelial cell migration in kidney proximal tubule. J Am Soc Nephrol 17: 39 45.
60. McCoy E, Sontheimer H (2007) Expression and function of water channels (aquaporins) in migrating malignant astrocytes. Glia 55: 1034 1043.
61. Papadopoulos MC, Verkman AS (2008) Potential utility of aquaporin modulators for therapy of brain disorders. Prog Brain Res 170: 589 601.
62. Bezzi P, Carmignoto G, Pasti L, Vesce S, Rossi D, et al. (1998) Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. Nature 391: 281 285.
63. Paquet M, Ribeiro FM, Guadagno J, Esseltine JL, Ferguson SS, et al. (2013) Role of metabotropic glutamate receptor 5 signaling and homer in oxygen glucose deprivation-mediated astrocyte apoptosis. Mol Brain 6: 9.
64. Durand D, Carniglia L, Caruso C, Lasaga M (2011) Reduced cAMP, Akt activation and p65-c-Rel dimerization: mechanisms involved in the protective effects of mGluR3 agonists in cultured astrocytes. PLoS One 6: e22235.
65. Aguado F, Espinosa-Parrilla JF, Carmona MA, Soriano E (2002) Neuronal activity regulates correlated network properties of spontaneous calcium transients in astrocytes in situ. J Neurosci 22: 9430 9444.
66. Anderova M, Vorisek I, Pivonkova H, Benesova J, Vargova L, et al. (2011) Cell death/proliferation and alterations in glial morphology contribute to changes in diffusivity in the rat hippocampus after hypoxia-ischemia. J Cereb Blood Flow Metab 31: 894 907.
67. Sik A, Smith RL, Freund TF (2000) Distribution of chloride channel-2-immunoreactive neuronal and astrocytic processes in the hippocampus. Neuroscience 101: 51 65.
68. Ransom CB, O'Neal JT, Sontheimer H (2001) Volume-activated chloride currents contribute to the resting conductance and invasive migration of human glioma cells. J Neurosci 21: 7674 7683.
69. Butenko O, Dzamba D, Benesova J, Honsa P, Benfenati V, et al. (2012) The increased activity of TRPV4 channel in the astrocytes of the adult rat hippocampus after cerebral hypoxia/ischemia. PLoS One 7: e39959.
70. Benfenati V, Amiry-Moghaddam M, Caprini M, Mylonakou MN, Rapisarda C, et al. (2007) Expression and functional characterization of transient receptor potential vanilloid-related channel 4 (TRPV4) in rat cortical astrocytes. Neuroscience 148: 876 892.
71. Pamenter ME, Perkins GA, McGinness AK, Gu XQ, Ellisman MH, et al. (2012) Autophagy and apoptosis are differentially induced in neurons and astrocytes treated with an in vitro mimic of the ischemic penumbra. PLoS One 7: e51469.
72. Neitz A, Mergia E, Eysel UT, Koesling D, Mittmann T (2011) Presynaptic nitric oxide/cGMP facilitates glutamate release via hyperpolarization-activated cyclic nucleotide-gated channels in the hippocampus. Eur J Neurosci 33: 1611 1621.
73. Zhao JW, Raha-Chowdhury R, Fawcett JW, Watts C (2009) Astrocytes and oligodendrocytes can be generated from NG2+ progenitors after acute brain injury: intracellular localization of oligodendrocyte transcription factor 2 is associated with their fate choice. The European journal of neuroscience 29: 1853 1869.
74. Komitova M, Serwanski DR, Lu QR, Nishiyama A (2011) NG2 cells are not a major source of reactive astrocytes after neocortical stab wound injury. Glia 59: 800 809.
75. Patanow CM, Day JR, Billingsley ML (1997) Alterations in hippocampal expression of SNAP-25, GAP-43, stannin and glial fibrillary acidic protein following mechanical and trimethyltin-induced injury in the rat. Neuroscience 76: 187 202.
76. Hepp R, Perraut M, Chasserot-Golaz S, Galli T, Aunis D, et al. (1999) Cultured glial cells express the SNAP-25 analogue SNAP-23. Glia 27: 181 187.
77. Illarionova NB, Gunnarson E, Li Y, Brismar H, Bondar A, et al. (2010) Functional and molecular interactions between aquaporins and Na,K-ATPase. Neuroscience 168: 915 925.
78. Anderova M, Antonova T, Petrik D, Neprasova H, Chvatal A, et al. (2004) Voltage-dependent potassium currents in hypertrophied rat astrocytes after a cortical stab wound. Glia 48: 311 326.
79. Bordey A, Lyons SA, Hablitz JJ, Sontheimer H (2001) Electrophysiological characteristics of reactive astrocytes in experimental cortical dysplasia. Journal of neurophysiology 85: 1719 1731.
80. Arimura K, Ago T, Kamouchi M, Nakamura K, Ishitsuka K, et al. (2012) PDGF receptor beta signaling in pericytes following ischemic brain injury. Curr Neurovasc Res 9: 1 9.
81. Zhu X, Zuo H, Maher BJ, Serwanski DR, LoTurco JJ, et al. (2012) Olig2-dependent developmental fate switch of NG2 cells. Development 139: 2299 2307.
82. Marshall CA, Suzuki SO, Goldman JE (2003) Gliogenic and neurogenic progenitors of the subventricular zone: who are they, where did they come from, and where are they going? Glia 43: 52 61.
83. Gorski JA, Talley T, Qiu M, Puelles L, Rubenstein JL, et al. (2002) Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. J Neurosci 22: 6309 6314.
84. Machon O, van den Bout CJ, Backman M, Rosok O, Caubit X, et al. (2002) Forebrain-specific promoter/enhancer D6 derived from the mouse Dach1 gene controls expression in neural stem cells. Neuroscience 112: 951 966.
85. Prajerova I, Honsa P, Chvatal A, Anderova M (2010) Neural stem/progenitor cells derived from the embryonic dorsal telencephalon of D6/GFP mice differentiate primarily into neurons after transplantation into a cortical lesion. Cell Mol Neurobiol 30: 199 218.
86. Honsa P, Pivonkova H, Anderova M (2013) Focal cerebral ischemia induces the neurogenic potential of mouse Dach1-expressing cells in the dorsal part of the lateral ventricles. Neuroscience.

87. Benesova J, Hock M, Butenko O, Prajerova I, Anderova M, et al. (2009) Quantification of astrocyte volume changes during ischemia in situ reveals two populations of astrocytes in the cortex of GFAP/EGFP mice. J Neurosci Res 87: 96–111.

88. Walz W, Lang MK (1998) Immunocytochemical evidence for a distinct GFAP-negative subpopulation of astrocytes in the adult rat hippocampus. Neurosci Lett. 257(3): 127–30.

6. Diskuze

6.1. Analýza jednotlivých buněk

Statistická analýza dat je klíčová pro pochopení a interpretaci výsledků. Povaha dat získaných z měření genové exprese u jednotlivých buněk vyžaduje specifický přístup. Jeden z největších problémů představuje existence transkripčních vzplanutí a s tím spojené chybějící hodnoty měření. Je důležité rozlišit mezi technickou chybou a nedetekovatelnou/nulovou expresí. Technické chyby jsou nejčastěji způsobeny přílišným naředěním vzorků nebo nedostatečnou účinností primerů. Při navrhování experimentu pro analýzu jednotlivých buněk je důležité vybrat geny tak, aby byly analyzovány jak geny se známým expresním profilem, tak i geny zcela neznámé. Známé geny pomohou ověřit technickou stránku celého experimentu a slouží tak jako kontrola. Dalším předpokladem pro kvalitní data je měření vysoce a středně exprimovaných genů.

6.1.1. Vliv chybějících dat na analýzu

Chybějící data je možné nahradit v případě technické chyby buď průměrem pro technické replikáty a nebo interpolovanou hodnotou na základě ostatních měřených vzorků a genů. Ani jedna z těchto variant není příliš vhodná pro data získaná měřením jednotlivých buněk. Při rozdělení buňky na technické replikáty dochází k naředění mRNA a nerovnoměrnému rozdělení transkriptů do jednotlivých reakcí, proto je mnohem výhodnější analyzovat co největší část buněčného obsahu a nedělit. Při měření exprese stejného genu na třech buňkách, kdy každá byla rozdělena na tři části, byly provedeny tři nezávislé reverzní transkripce a gen byl měřen v triplikátu, bylo ukázáno, že průměrná exprese mezi buňkami má standardní odchylku 1,99, zatímco standardní odchylka replikátu reverzní transkripce je pouze 0,3 a qPCR 0,51. Použitím technických replikátů tedy nelze zpřesnit měření a snížit standartní odchylku (Tichopad et al., 2009). Použití interpolace podle exprese ostatních vzorků a genů pro doplnění chybějících dat vnáší do analýzy umělou korelaci s ostatními geny, která nemusí mít biologický základ.

Vzhledem k těmto skutečnostem a k tomu, že při analýze jednotlivých buněk by měl být protokol velice dobře optimalizován, se s chybějícími daty zachází jako s nulovou hladinou exprese nebo hladinou exprese pod limitem detekce. Chybějící údaje se nahrazují nulovou hodnotou pro data v lineárním měřítku přepočtená na jednotlivé kopie, nebo hodnotou limitu detekce plus 2 (nebo plus 4) pro data v logaritmickém měřítku (publikace *RT-qPCR work-flow for single-cell data analysis* (Stahlberg et al., 2013b)). V případě, že je takto nahrazeno více než 10 % dat, je tím silně ovlivněno jejich rozdělení a

následná popisná statistika - výpočty průměru, standartní odchylek atd. (Graham, 2009). Pro analýzu je někdy vhodné ukázat data pouze v binární podobně: je exprese – není exprese, a až v dalším kroku se zabývat její velikostí (publikace *The added value of single-cell gene expression profiling* (Stahlberg et al., 2013a)).

Chybějící data ovlivňují možnost normalizace. Doporučený postup pro klasický qPCR experiment je normalizovat pomocí referenčních genů (Kapitola 3.2.1 a 3.2.1). To vychází z předpokladu, že kvalitní referenční geny jsou exprimovány ve všech zkoumaných subjektech stejně a umožní tak srovnat úroveň exprese na stejnou hladinu a vyrovnat technickou variabilitu. Matematicky to znamená snížit standardní odchylku tak, aby byla variabilita menší pro průměrné ΔCq ve srovnání s průměrným Cq hledaného genu, což se vždy nestane (Tichopad et al., 2009). Pro jednotlivé buňky je normalizace pomocí referenčních genů problematická. Vzhledem k transkripčním vzplanutím není vždy referenční gen v buňce detekován. Bylo ukázáno, že hladina exprese souvisí s velikostí buňky (Silbert et al., 2003) a normalizací na referenční gen lze dosáhnout snížení technického šumu (Veys et al., 2012), zatím ale nebyla publikována žádná data, která by presentovala stabilní referenční gen pro jednotlivé buňky nalezený některým z tradičních postupů (Genorm, Normfinder). Další variantou normalizace dat je použití globální normalizace, tj. normalizovat pomocí průměru celkové exprese (centrovat přes vzorek). Tato metoda se úspěšně používá pro normalizaci při analýze mikroarrayí měřených pro miRNA (Welle, 2013). Pro qPCR však není zcela vhodná, vzhledem k tomu, že se analyzují jen vybrané geny, které spolu velmi často souvisí. Tento postup dále přináší problém při porovnání jednotlivých studií, které používají odlišný set genů. Proto se většina vědecké obce kloní k normalizaci na „buňku“ zobrazení dat pro každou buňku zvlášť (Stahlberg and Bengtsson, 2010). Tento typ normalizace byl použit v diskutovaných publikacích o jednotlivých buňkách (Brevik et al., 2011; Benesova et al., 2012; Brevik et al., 2012; Rusnakova et al., 2013; Stahlberg et al., 2013b; Stahlberg et al., 2013a)

Také korelační analýza je silně ovlivněna množstvím chybějících dat. Pro málo exprimované geny, tj. chybí 70 % a více hodnot, korelační koeficient vyjadřuje pouze vztah je exprimovaný/není exprimovaný a hladina exprese je zohledněna pouze slabě. Pro silně exprimované geny (tj. chybí 20 % a méně hodnot) korelační koeficient vyjadřuje vztah velikosti exprese mezi jednotlivými geny a je zde možné posoudit jejich skutečný vztah v každé buňce. Korelační koeficient, jehož absolutní hodnota je nad 0,6, ukazuje na silnou závislost mezi dvěma geny, které jsou pravděpodobně přepisovány ve stejných transkripčních továrnách (Osborne et al., 2004). Dále lze spočítat parciální korelační koeficient, kdy se vztah mezi dvěma geny spočítá přímo bez vlivu ostatních. Pokud zůstane hodnota korelačního koeficientu stejná,

nemají ostatní geny na něj vliv a vztah mezi dvěma geny se označuje jako přímá vazba. Pokles korelačního koeficientu o 0,2 při jeho zachování statistické významnosti naznačuje, že je vztah všech zkoumaných genů rovnocenný. Pokud korelační koeficient poklesne pod 0,2 je vazba mezi dvěma geny zprostředkována dalšími. Touto cestou je možné nalézt „centrální gen“, který řídí ostatní (Stahlberg et al., 2011b). Výpočet parciální korelace, i s jeho následnou vizualizací, byl ukázán v publikaci *Heterogeneity of Astrocytes: From Development to Injury – Single Cell Gene Expression* (Rusnakova et al., 2013). Korelační koeficienty lze znázornit zobrazením do Euklidovského prostoru, kdy jsou jednotlivé geny navzájem propojeny podle velikosti korelace mezi nimi (Rusnakova et al., 2013, Fig. S2). Tento typ zobrazení ukazuje přehledně význam jednotlivých genů bez nutnosti číst velké tabulky. Na podobném principu je založeno i zobrazení pomocí paralelních koordinát, kdy jsou vzorky propojeny spolu podle svého významu a funkce (Meyer and Cook, 2000).

6.1.2. Vizualizace výsledků

Vizualizace dat získaných analýzou jednotlivých buněk má mnoho aspektů. Prvním je vybrání správného typu z matematického pohledu a nalezení biologické interpretace výsledku. Dalším úhlem pohledu je otázka, jak ukázat co nejvíce informací na jednom místě, aby byly výsledky srozumitelné. qPCR data se nejčastěji zobrazují ve sloupcových grafech. Pro jednotlivé buňky tento způsob není vhodný vzhledem k tomu, že data nejsou normálně rozložena a průměr nemá vypovídající hodnotu o expresi (Bengtsson et al., 2008). Ze stejného důvodu je nevhodné použít zobrazení pomocí krabicových grafů. Velké množství chybějících hodnot tyto grafy „roztahuje“ a posunuje dolů. Mnohem vhodnější metodu je použití histogramů, ve kterých je vidět rozdělení dat. Často jsou takto odhaleny i skupiny vzorků, které mohou pocházet z různých subpopulací na první pohled homogenní skupiny buněk. Nevýhodou je nutnost velkého množství dat, tak aby mohlo být rozdělení patrné. Další možností jak zobrazit data je binární přístup, kdy jsou pozitivní vzorky označeny jednou barvou a negativní vzorky druhou. Ani jeden z těchto způsobů vizualizace dat nepropojuje více genů či vzorků dohromady a nedokáže tak najít atypickou buňku, která se liší v expresi ve více genech.

Zobrazovací metoda, která zohledňuje příbuznost vzorků na základě exprese více genů je shluková analýza a její grafický výstup – dendrogram (Kapitola 3.2.3.2). Výhodou je převedení vícerozměrných dat do dvourozměrných. Pro výpočet je důležité správně zvolit metodu shlukování, která by měla být vybrána ještě před samotnou analýzou dat. Pro jednotlivé buňky se nejčastěji používá Wardova metoda, díky které vzorky tvoří malé kompaktní shluky. Shluková analýza může být počítána jak pro geny

(porovnání expresních profilů vzorků), tak pro vzorky (porovnání expresních profilů genů). Na příkladu vývoje drápatky (publikace *Gene expression profiling - Clusters of possibilities* (Bergkvist et al., 2010, Figure. 3)) bylo ukázáno použití obou dvou variant. Vzorky byly rozděleny podle stádia vývoje od vajíčka až po pulce a geny byly rozčleněny na tři skupiny - maternální, midblastulární a zygotické. Kombinace shlukové analýzy pro geny a pro vzorky je tzv. heat-mapa, ve které je velikost exprese znázorněna barevným přechodem. To plně nahrazuje sloupcové grafy a navíc ukazuje na vztah mezi jednotlivými vzorky a geny. Nevýhodou shlukových analýz je jejich velká citlivost k „abnormalitám" v expresi a odlišnost shluků podle použité shlukovací metody. Bylo prokázáno, že změna shlukovací metody vede ke tvorbě jiných shluků a tím i jiné interpretaci výsledků (Andrew, 2000). Další velkou nevýhodou shlukových analýz je, že jsou to hierarchické metody, tudíž se předpokládá, že všechna data spolu souvisí a mají společný „počátek". Z tohoto důvodu je tento typ analýzy vhodný pro vývojová data, ale poněkud méně pro hledání jednotlivých subpopulací (Gilbert et al., 2000).

Další metodou pro převod dat multidimenzionálního prostoru do dvoudimenzionálního je analýza hlavních komponent (PCA). Její nevýhodou je však ztráta části informace při transformaci dat. Čím více komponent zůstane zachováno, tím je informace celistvější. Pro nalezení podskupin stačí ve většině případů ukázat zobrazené první tři komponenty. S tím je polemizováno v publikaci F. Buettnera (Buettner and Theis, 2012), kde je navrženo použít Gaussovské procesní modely latentní proměnné (Gaussian process latent variable models), které pomohou odhalit menší podskupiny buněk. Otázkou zůstává, kolik podskupin dává ještě biologický význam. To bylo diskutováno v publikaci A. Stahlberga (Stahlberg et al., 2011b) a v publikaci V. Rusňákové (Rusnakova et al., 2013*)*, kde se autoři shodují v názoru, že počet podskupin závisí na množství vzorků, na reprodukovatelnosti a shodnosti jednotlivých analýz. Nevýhodou PCA je, že nerozděluje data přímo do skupin jako shluková analýza nebo SOM, ale zobrazí je v grafu, statistik si musí sám následně rozdělit jednotlivé vzorky do skupin. Tento přístup může být zavádějící, a proto je dobré ověřit navržené rozdělení další metodou, např. SOM.

Výhodou SOM oproti PCA je možnost zvolit si předem počet skupin. Při navolení velkého počtu (stejný počet jako je počet analyzovaných vzorků) se ukáže vztah mezi jednotlivými vzorky, které nejsou nuceny „si vybrat" mezi omezeným počtem skupin a mají dostatek prostoru na vytvoření přirozeného shluku (Fernandez and Balzarini, 2007). Naproti tomu menší počet skupin lze lépe charakterizovat a jsou zde jasné hranice mezi jednotlivými populacemi. Výpočet SOM představuje učební proces (na rozdíl od výpočetních algoritmů pro shlukovou analýzu a PCA), díky kterému je každé vytvořené rozdělení

jedinečné. SOM se používají jak pro navržení, tak i pro ověření hypotéz o složení a charakterizaci jednotlivých skupin (Stahlberg et al., 2011b).

Při hledání vhodné analýzy pro jednotlivé buňky je doporučeno nejdříve zobrazit data pomocí histogramů, aby bylo možné vytipovat geny, které budou mít rozhodující vliv na multivariantní zobrazení. Nalezené skupiny vzorků by se měly překrývat pro PCA, SOM i shlukovou analýzu (Stahlberg et al., 2008).

Možnosti vizualizace se s rozvojem počítačů a příslušných programů ubírají směrem k interaktivním mapám, ve kterých si uživatel může volit sám barvy pro různé skupiny genů, podívat se na analyzovaná data jako na celek, i na vlastnosti jednotlivých buněk (Gilbert et al., 2000). Software GenEx (Kapitola 5.2) jde touto cestou. Možnosti vizualizace dat jsou zde zaměřeny hlavně na SOM, PCA a dendrogramy.

Další část diskuze je věnována aplikaci analýzy jednotlivých buněk na astrocyty.

6.2. Heterogenita astrocytů

Heterogenita astrocytů i v jediné oblasti jako je kůra mozková vzniká už během vývoje. Některé astrocyty zde vznikly přeměnou intermediátorů radiálních glií z ventrikulární zóny, jiné se vyvinuly z nezralých buněk, které migrovaly v perinatální periodě do kůry z subventrikulární zóny (místo vzniku nových buněk CNS). Vzhledem k tomu, že z radiálních glií se mohou vyvinout jak neurony předního mozku, tak i gliové buňky, musí být už ve své podstatě vysoce heterogenní, což se musí odrazit i na jejich různorodosti. Jak heterogenita během vývoje souvisí s funkční heterogenitou, zůstává zatím neobjasněno. V této dizertační práci je kladen důraz na charakterizaci astrocytů z hlediska jejich genového profilu během jejich vývoje, za stresových podmínek a po fokální cerebrální ischemii. Pomocí RT-qPCR byly analyzovány EGFP/GFAP pozitivní buňky, které byly sbírány pomocí průtokové cytometrie. Zaměřili jsme se hlavně na analýzu gliových glutamátových transportérů a receptorů a iontových kanálů, které jsou nezbytné pro udržení homeostáze gliových buněk a neuronů jak za fyziologických podmínek, tak v patologickém stavu.

6.2.1. Odlišná odpověď astrocytů na stresové podmínky

Analýza jednotlivých buněk je metoda, která je vhodná k nalezení různých subpopulací buněk. Publikace *Distinct Expression/Function of Potassium and Chloride Channnels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice* (Benesova et al., 2012) ukazuje odlišné chování

dvou typů astrocytů na aplikaci inhibitorů pro chloridové a draselné kanály při nedostatku glukózy a kyslíku (OGD). Tyto typy astrocytů jsou označeny, jako silně odpovídají (HR) a slabě odpovídají (LR). HR astrocyty zvyšují svůj objem za stresových podmínek až o 40 %, zatím co LR astrocyty mění svůj objem pouze o 5 %. Jak farmakologické experimenty, tak qPCR analýzy ukázaly, že astrocyty se liší v expresi iontových kanálů a transportérů, které se účastní objemových změn astrocytů. Změna objemu astrocytů byla spojena s přijímáním iontů pro udržení osmotické rovnováhy.

Pro subpopulaci LR byla typická zvýšená exprese chloridového kanálu (*Clcn2*) a draselného kanálu *Kcnk2* (kóduje protein Trek1). Subpopulace HR vykazovala zvýšenou expresi *Kcnj10* a *Kcnk1* (kódují proteiny KIR4.1 a TWIK1), (Kapitola 3.3.2). *Kcnk1* se účastní přijímání K^+ iontů do buňky, zatímco *Kcnk2*, který může být aktivovaný stresem nebo při zvyšováním objemu buňky, je zodpovědný za vylučování K^+ iontů (Skatchkov et al., 2006; Enyedi and Czirjak, 2010). Rozdílná exprese *Kcnk1* a *Kcnk2,* které patří do skupiny draselných kanálů se dvěma doménami, může mít původ v rozdílném umístění astrocytů (Seifert et al., 2009). Předpokládá se, že HR astrocyty jsou v blízkosti vysokých extracelulárních koncentrací K^+, a proto hlavně exprimují kanály umožňující vstup draslíku do buňky. Zatímco LR astrocyty mohou získat K^+ jako následek pufrační činnosti a uvolnit ho do mezibuněčného prostoru. *Knck1* a *Kcnk2* mají také význam při poškození CNS, jak bylo již dříve ukázáno (Kucheryavykh et al., 2009; Pivonkova et al., 2010). Zvýšená exprese *Kcnj10* u HR astrocytů je v souladu s pozorovanou zvýšenou hustotou proteinu KIR4.1 na buněčné membráně astrocytů (Benesova et al., 2009) a je pravděpodobně dána polohou astrocytů v místě zvýšené koncentrace K^+ iontů, kde také dochází k častějším přenosům K^+ iontů z/do astrocytů. Zvýšená exprese *Clcn2*, který může být aktivován změnou buněčného objemu (Nilius and Droogmans, 2003) v LR astrocytech, může zapříčinit jejich zvýšenou schopnost regulovat svůj buněčný objem. Jak LR, tak HR astrocyty vykazovaly stejnou expresi glutamátových transportérů *Eaat1* a *Eaat2*, nepředpokládá se, že by se lišilyve velikosti nebo ve zralosti.

Odlišná změna objemů v astrocytech při nedostatku glukózy a kyslíku je pravděpodobně dána rozdílnou expresí iontových kanálů a transportérů. Ta může být vyvolána rozdílnou lokální koncentrací K^+ a aminokyselin. Schopnost různou měrou regulovat svůj buněčný objem může být spojena s umístěním astrocytu v kůře a s jeho sousedy.

6.2.2. Vývoj astrocytů v kůře mozkové

Během postnatálního vývoje dochází v kůře mozkové k velkým změnám, které jsou podle publikované práce (Rusnakova et al., 2013) dokončeny po 50 dnech života. Bylo ukázáno, že s věkem roste množství

EGFP/GFAP pozitivních (EGFP$^+$) buněk, které exprimují markery astrocytů jako je *Glul* a *Eaat1*, zatímco množství buněk, které exprimují markery NG2 glií, klesá (Rusnakova et al., 2013, Figure 1). Mimo těchto markerů astrocytů dále s věkem rostlo i množství mRNA pro *Aqp4* a *Grin2a*, ve shodě s publikovanými daty dříve (da Cunha et al., 1991), které je známkou zralosti a plné funkčnosti astrocytů. Na úrovni jednotlivých buněk bylo zjištěno, že EGFP$^+$ buňky, které exprimují *Cspg4* nebo *Pdgfrα* (Kapitola 3.3.2), exprimují také na vysoké úrovni podjednotky glutamátových ionotropních kainátových receptorů (*Grik*), zatímco buňky bez markerů NG2 glií *Grik* exprimují velmi slabě, což podporuje teorii, že buňky se znaky NG2 glií jsou vývojovým stádiem pro astrocyty (Matthias et al., 2003). Nalezený vzor exprese byl již publikován dříve (Matthias et al., 2003; Cahoy et al., 2008; Palygin et al., 2011) a ověřil tak technickou správnost experimentu (Kapitoly 5.2, 6.1.2 a 6.1.1).

Ve stádiu P10 většina EGFP pozitivních buněk exprimovala *Cspg4* a *Pdgfrα*, zatímco ve stádiu P50 byly nalezeny pouze 3 buňky s touto expresí. Ačkoliv celkově 95 % buněk exprimovalo jak *Cspg4*, tak *Pdgfrα* nejsou tyto geny navzájem korelované a naopak každý z nich je úzce korelovaný s jinými. To nasvědčuje tomu, že jsou exprimovány nezávisle na sobě a každý z nich používá odlišné transkripční regulační mechanismy, které následně vedou k transkripčním vzplanutím (Kapitola 3.1.1). To znamená, že ne vždy každá buňka musí exprimovat oba geny zároveň a při hledání určitého typu buněk, je nutné testovat více markerů specifických pro daný typ buněk. Dále 11 % buněk exprimovalo *Gfap*, *Cspg4* nebo *Pdgfrα* na úrovni mRNA (typické markery pro astrocyty a NG2 glie). Ve stádiu P10 a P20 byla nalezena ko-exprese GFAP a markerů pro NG2 glie na proteinové úrovni. NG2 glie by tedy mohly být prekurzory pro astrocyty i v postnatálním vývoji. Ačkoliv je GFAP považován za hlavní marker astrocytů a všechny analyzované buňky byly pozitivní na EGFP, jehož gen je pod stejným promotorem jako gen pro *Gfap*, mRNA hladina *Gfap* byla velmi nízká. Stejně tak jsou známy buňky, které vykazují znaky astrocytů a nejsou pozitivní na imunochemickou detekci GFAP proteinu (Benesova et al., 2009). Tato data ukazují, že identifikace astrocytů založená pouze na jednom markeru je problematická. Během svého vývoje mají astrocytální markery nejen zralé astrocyty, ale i jejich prekurzory a NG2 glie. Obzvláště u mladých zvířat je nutné použít k identifikaci astrocytů více markerů a nejlépe i přímou vizualizaci buněk.

Data z vývojových stádií P10 – P50 byla rozdělena pomocí SOM a PCA (Kapitoly 3.2.3.1 a 3.2.3.3) do tří subpopulací (A1, A2, A3), (Rusnakova et al., 2013, Figure 3) a jednotlivé subpopulace byly charakterizovány. Buňky subpopulace A3 (tvořené hlavně stádii P30 a P50) vykazovaly znaky plně zralých astrocytů s vysokou expresí *Eaat1*, *Glul* a *Aqp4* (Regan et al., 2007). Buňky subpopulace A1, která je tvořena hlavně buňkami ze stádií P10 a P20, mimo *Eaat1* exprimovaly vysoké hladiny *Cspg4* a *Pdgfrα*,

což jsou znaky tradičně spojované s NG2 gliemi (Butt et al., 2005), přesto byly již dříve také nalezeny buňky, které byly označeny za astrocyty a exprimovaly markery pro NG2 glie a astrocyty dohromady (Cahoy et al., 2008; Trotter et al., 2010; Honsa et al., 2012). Subpopulace buněk A1, je dále charakterizována zvýšenou expresí podjednotek glutamátových ionotropních receptorů *Gria* a *Grik*. Přítomnost podjednotky *Gria2* naznačuje, že AMPA receptory subpopulace A1 mohou být nepropustné pro Ca^{2+} ionty na rozdíl od zralých astrocytů A3. Stejně tak odlišná exprese draselných kanálů ve zralých (hlavně *Kcnj10*) a nezralých (oboje, *Kcnj10* i *Kcnj16)* astrocytech vede k rozdílným schopnostem propouštět K^+ ionty dovnitř během zvýšené neuronální aktivity nebo za patologických podmínek při poranění. Analýza korelačních koeficientů neprokázala přímou interakci mezi *Kcnj10* a *Aqp4* a potvrdila se tak jejich rozdílná regulace exprese publikovaná již dříve (Liu et al., 2012).

Subpopulace buněk A2, jež je tvořena převážně z buněk P20, se odlišovala od buněk A1 vysokou expresí *Glul, Hcn2, Gria2* a *Grin2a*, zatímco exprese ostatních genů byla nízká. Exprese *Hcn2* byla později nalezena i u astrocytů po ischemickém poranění a může naznačovat podobný vývoj. Předpokládáme, že tato subpopulace představuje přechod mezi zralými a nezralými astrocyty, což je také podpořeno její skladbou. Otázkou zůstává, zda jsou tyto astrocyty odvozeny z *Cspg4/Pdgfrα* pozitivních buněk, podobně jako u poraněných zvířat, kde vznikají nové astrocyty i se aktivují původní.

V souladu s již publikovanými daty (Cahoy et al., 2008; Sun et al., 2013) bylo ukázáno, že zralé astrocyty exprimují sníženou hladinu glutamátového metabotropního receptoru *Grm5* a zvýšenou hladinu *Grm3*. Předpokládá se, že *Grm5* podporuje glutamátem indukovanou Ca^{2+} signalizaci pouze v nezralých astrocytech. Ve zralých astrocytech je Ca^{2+} signalizace umožněna pomocí transmiterů jako jsou endokanabinoidy, puriny, norepinephin, acetylcholin a extracelulární koncentrací Ca^{2+} iontů. *Grm3* nehraje v této signalizaci pravděpodobně žádnou roli, ale účastní se regulačních procesů jako je proliferace astrocytů, exprese glutamátových receptorů a *Gfap* (Sun et al., 2013).

6.2.3. Vývoj a přeměna astrocytů po ischemii

Po poranění CNS jako je trauma, ischemie nebo při onemocnění CNS neurodegenerativními chorobami se vlastnosti astrocytů mění a stávají se reaktivními (Kapitola 3.3.3). Podle získaných expresních profilů byly jednotlivé buňky rozděleny stejně jako buňky z vývoje (Kapitola 6.2.2) do tří subpopulací (B1, B2 a B3), (Rusnakova et al., 2013, Figure 4). B1 subpopulace obsahovala hlavně buňky nereaktivních astrocytů z kontroly a částečně i 3 a 7 dní po poranění. B2 subpopulace se skládala hlavně z buněk 3 a 7 dní po ischemii a subpopulace B3 obsahovala pouze reaktivní astrocyty 14 dnů po poranění.

B1 subpopulace vykazovala stejné chování jako subpopulace A3 (Kapitola 6.2.2). B2 subpopulace, která má ve shodě s publikovanými pracemi (Pekny and Nilsson, 2005) zvýšenou expresi *Gfap*, *S100β*, vimentinu a markerů NG2 glií, se skládá z tzv. „astrocytů v přechodovém stádiu“. Exprese *Cspg4* a *Pdgfrα* naznačuje, že tyto astrocyty mohou mít za prekurzory NG2 glie, které se mohou plně vyvinout později v reaktivní astrocyty. Tato varianta původu astrocytů je podporována podobným expresním profilem buněk ze stádia B2 jako mají buňky v nezralé kůře.

Subpopulace B2 a B3 vykazovaly vysokou expresi mRNA pro *Trpv4*, což odpovídá i datům naměřených u potkanů (Butenko et al., 2012). *Trpv4* má roli v buňkách jako osmosenzor a účastní se i Ca^{2+} signalizace a pravděpodobně zprostředkovává vstup Ca^{2+} do buňky a tím se může účastnit patogenních mechanismů v astrocytech. Dalším společným genem se zvýšenou expresí pro subpopulace B2 a B3 je *Gria3*, což může znamenat, že AMPA receptory se mohou podílet na vpouštění Na^{+} a Ca^{2+} do astrocytu a tím zvyšují jeho patologický stav.

Buňky, které byly 14 dní po ischemii, vykazovaly expresní profil velmi odlišný od ostatních. Astrocyty jsou aktivovány a plní svojí ochranou funkci při vytváření gliové jizvy (Kapitola 3.3.3). V souladu s předchozími pracemi zde byla pozorována zvýšená exprese genů *Aqp9* a *Aqp1* (Zelenina, 2010), *Grm5* (Ulas et al., 2000) i dosud nepopsaných genů typických pro reaktivní astrocyty jako jsou glutamátové ionotropních receptorů *Grin2a*, *Gria2*, *Gria3*, draselného kanálu *Kcna3*, *Trpv4*, *Snap25*, kódující t-SNARE protein, a hyperpolarizací aktivované nespecifické kationtové kanály *Hcn1*, *2* a *3*. *Aqp1* a *Aqp9* jsou obecně přítomny hlavně v hypertrofických, GFAP exprimujících astrocytech během patologických stavů jako je ischemie, Alzheimerova nebo Creutzfeldt-Jakobova choroba. V jednotlivých buňkách byly aquaporiny detekovány až 14 dní po ischemii. To nasvědčuje tomu, že se pravděpodobně neúčastní přenosu molekul vody během vzniku edému, ale mají roli jako kanály pro metabolity a umožňují difusi glycerolu a laktátu (Badaut et al., 2004). Vysoká exprese *Kcna3* ukazuje, že buňky proliferují (Anderova et al., 2011). *Gria3* spolu s vysokou expresí *Grm1* a *Grm5* pravděpodobně zprostředkovává Ca^{2+} oscilace. Zvýšená exprese *Grm5*, která je společná reaktivním astrocytům a nezralým astrocytům může také naznačovat jejich společný původ.

Exprese genu pro *Snap25*, který je v ostatních subpopulacích přítomen minimálně, nebyla očekávána, vzhledem k tomu, že byl původně označen jako marker buněk se schopností exocytózy za fyziologických podmínek (Cahoy et al., 2008). Reaktivní astrocyty uvolňují glutamát a aspartát ve váčkách jako odpověď na zvyšující se intracelulární množství Ca^{2+}, což může vysvětlit zvýšené množství mRNA pro protein SNAP25, který je běžně součástí t-SNARE komplexu účastnícího se exocytózy.

Exprese genů pro hyperpolarizací aktivované nespecifické kationtové kanály propustné pro K^+ a Na^+ ionty byla doposud popsána pouze v neuronech a jejich funkce je v astrocytech neznámá. Předpokládáme, že mohou mít funkci při zachovávání klidového membránového potenciálu, případně mohou interagovat s proteiny účastnících se uvolňování glutamátu, tak jak bylo popsáno pro konce synapsí (Neitz et al., 2011).

Výpočtem korelací mezi jednotlivými geny se ukázal silný vztah mezi *Aqp4* a *Eaat1*, což potvrzuje jejich blízký vztah za patologického stavu CNS spojeného s uvolňováním glutamátu (Zelenina, 2010). Pozitivní korelace mezi *Snap25* a *Eaat1* potvrzuje hypotézu o zahrnutí proteinu SNAP25 do exocytózy glutamátu. Výpočet Spearmanových parciálních korelací neukázal na jeden gen, který by řídil celou expresi v reaktivních astrocytech. Přesto se ukázalo, že velký význam mají geny *Hcn2* a *Gria3*, které budou dále analyzovány.

Heterogenita astrocytů po poranění byla očekávaná. Proto byla subpopulace buněk B3 byla dále rozdělena na dvě skupiny (B3A a B3B) na základě pozitivity na markery pro NG2 glie. Skupina B3B s vysokou expresí *Cspg4* a *Pdgfrα* vykazovala celkově nižší expresi ostatních genů, včetně astrocytálních markerů jako je *Eaat1*, *Aqp4* a *Glul* než skupina B3A. Vzhledem k tomu, že všechny astrocyty byly izolovány z místa ischemického poranění, se skupina B3 pravděpodobně skládá z reaktivních glií, které jsou v různém stupni vývoje, nebo mohly také být různě vzdáleny od místa poranění. Přítomnost NG2 markerů naznačuje možný původ astrocytů a jejich vývoj z NG2 glií. Tato teorie je i podporována PCA analýzou, kdy jsou buňky skupiny B3B jsou blíže ke skupině B2 než buňky B3A. Tyto buňky mohou být astrocyty v „přechodovém stádiu“, jak již bylo popsáno dříve (Honsa et al., 2012).

7. Závěr

Genové expresní profilování jednotlivých buněk přináší další rozměr k poznatkům o jejich fungování a funkci tkání a orgánů jako celku. Protokol pro jejich analýzu musí být plně optimalizován, od sběru samotných buněk, přes reverzní transkripci a preamplifikaci, až po samotné qPCR. Bez tohoto není možné získat spolehlivé výsledky. Byl navržen postup pro statistické zpracování dat získaných měřením genové exprese v jednotlivých buňkách. Dále je zdůrazněn rozdíl od analýzy dat z populačních studií, a to především v doplnění chybějících dat a v normalizaci. Na několika biologických systémech jsme ukázali použití shlukových analýz, analýzy hlavních komponent a arteficiálních neuronových sítí. Pro tyto analýzy byl použit software GenEx.

Druhá část dizertační práce je zaměřena na analýzu astrocytů pomocí výše uvedených metod. Byly nalezeny rozdílné astrocytální populace, které se liší v expresi draselných a chloridových kanálů majících účast na změně objemu astrocytů. Nalezené rozdíly podpořily již dříve publikovanou studii popisující dva typy kortikálních astrocytů, které odpovídají rozdílně na stresové podmínky, jako je nedostatek kyslíku a glukózy (Benesova et al., 2009). Dále bylo ukázáno, že kortikální astrocyty během postnatálního vývoje i po poranění tvoří vysoce heterogenní populaci vzhledem k jejich genové expresi. Použitím genového expresního profilování jednotlivých buněk a multivariantní analýzy bylo identifikováno a charakterizováno několik skupin astrocytů. Zatímco obecně je transkripční aktivita nezralých astrocytů velmi vysoká, plně vyvinuté astrocyty exprimují hlavně *Eaat1*, *Glul*, *Aqp4* a *Kcnj10*. Po poranění se astrocyty aktivují a jejich exprese se rychle změní. Mimo známé exprese *Gfap*, *Aqp1* a *Aqp9* jsme nalezli i expresi zatím neznámých genů pro ischemické poranění jako jsou *Snap25*, hyperpolarizací aktivované kationtové kanály *Hcn1, Hcn2* a *Hcn3*, glutamátové receptory *Grin2a*, *Gria2*, *Gria3*, *Grm5* a draselný kanál *Kcna3*, které mohou být v budoucnu použity jako znaky reaktivity.

Cílem této dizertační práce bylo vypracování metody pro analýzu jednotlivých buněk pomocí qPCR a její aplikace na biologický systém. Podařilo se ukázat, jak pomocí shlukových analýz a vizualizace dat vytvořit hypotézu o chování jednotlivých buněk. Matematické modelování se ukázalo jako silný nástroj pro nalezení různých podskupin vzorků a nových populací. U jednotlivých metod je potřeba znát jejich silné a slabé stránky. Nutnou podmínkou k hledání nových podtypů buněk je stanovení hypotézy a její následné ověření pomocí dalších metod. V předložených publikacích bylo ukázáno, jak využít již známá data pro kontrolu kvality experimentu a spolu s novými je využít k nalezení nových podtypů buněk.

8. Seznam citované literatury

Alonso, G. (2005). NG2 proteoglycan-expressing cells of the adult rat brain: possible involvement in the formation of glial scar astrocytes following stab wound. *Glia* **49**, 318-338.

Anderova, M., Vorisek, I., Pivonkova, H., Benesova, J., Vargova, L., Cicanic, M., Chvatal, A. and Sykova, E. (2011). Cell death/proliferation and alterations in glial morphology contribute to changes in diffusivity in the rat hippocampus after hypoxia-ischemia. *J Cereb Blood Flow Metab* **31**, 894-907.

Andersen, C. L., Jensen, J. L. and Orntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* **64**, 5245-5250.

Andrew, A. M. (2000). STATISTICAL PATTERN RECOGNITION, by Andrew Webb, Arnold, London (Cambridge University Press, New York, for USA), 1999, xviii+454 pp., ISBN 0-340-74164-3 (pbk, £29.99). *Robotica* **18**, 219-223.

Araque, A., Parpura, V., Sanzgiri, R. P. and Haydon, P. G. (1999). Tripartite synapses: glia, the unacknowledged partner. *Trends Neurosci* **22**, 208-215.

Azevedo, F. A.C., Carvalho, L. R.B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E.L., Leite, R. E.P., Filho, W. J., Lent, R. and Herculano-Houzel, S. (2009), Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol* **513**, 532–541

Badaut, J., Petit, J. M., Brunet, J. F., Magistretti, P. J., Charriaut-Marlangue, C. and Regli, L. (2004). Distribution of Aquaporin 9 in the adult rat brain: preferential expression in catecholaminergic neurons and in glial cells. *Neuroscience* **128**, 27-38.

Benesova, J., Hock, M., Butenko, O., Prajerova, I., Anderova, M. and Chvatal, A. (2009). Quantification of astrocyte volume changes during ischemia in situ reveals two populations of astrocytes in the cortex of GFAP/EGFP mice. *J Neurosci Res* **87**, 96-111.

Benesova, J., Rusnakova, V., Honsa, P., Pivonkova, H., Dzamba, D., Kubista, M. and Anderova, M. (2012). Distinct Expression/Function of Potassium and Chloride Channels Contributes to the Diverse Volume Regulation in Cortical Astrocytes of GFAP/EGFP Mice. *PLoS One* **7** (1):e29725.

Bengtsson, M., Stahlberg, A., Rorsman, P. and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* **15**, 1388-1392.

Bengtsson, M., Hemberg, M., Rorsman, P. and Stahlberg, A. (2008). Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. *BMC Mol Biol* **9**, 63.

Bergkvist, A., Rusnakova, V., Sindelka, R., Garda, J. M., Sjogreen, B., Lindh, D., Forootan, A. and Kubista, M. (2010). Gene expression profiling--Clusters of possibilities. *Methods* **50**, 323-335.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H. and Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. *Mol Cell* **2**, 437-445.

Boggy, G. J. and Woolf, P. J. (2010). A mechanistic model of PCR for accurate quantification of quantitative PCR data. *PLoS One* **5**, e12355.

Brenner, M. and Messing, A. (1996). GFAP Transgenic Mice. *Methods* **10**, 351-364.

Brevik, A., Lindeman, B., Rusnakova, V., Olsen, A. K., Brunborg, G. and Duale, N. (2012). Paternal Benzo a pyrene Exposure Affects Gene Expression in the Early Developing Mouse Embryo. *Toxicological Sciences* **129**, 157-165.

Brevik, A., Rusnakova, V., Duale, N., Slagsvold, H. H., Olsen, A. K., Storeng, R., Kubista, M., Brunborg, G. and Lindeman, B. (2011). Preconceptional paternal glycidamide exposure affects embryonic gene expression: Single embryo gene expression study following in vitro fertilization. *Reproductive Toxicology* **32**, 463-471.

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* **28**, i626-i632.

Bushong, E. A., Martone, M. E., Jones, Y. Z. and Ellisman, M. H. (2002). Protoplasmic astrocytes in CA1 stratum radiatum occupy separate anatomical domains. *J Neurosci* **22**, 183-192.

Bustin, S. A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* **25**, 169-193.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L. et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611-622.

Butenko, O., Dzamba, D., Benesova, J., Honsa, P., Benfenati, V., Rusnakova, V., Ferroni, S. and Anderova, M. (2012). The increased activity of TRPV4 channel in the astrocytes of the adult rat hippocampus after cerebral hypoxia/ischemia. *PLoS One* **7**, e39959.

Butt, A. M., Hamilton, N., Hubbard, P., Pugh, M. and Ibrahim, M. (2005). Synantocytes: the fifth element. *J Anat* **207**, 695-706.

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing, Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A. et al. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* **28**, 264-278.

Chattopadhyay, P. K. and Roederer, M. (2012). Cytometry: Today's technology and tomorrow's horizons. *Methods* **57**, 251-258.

Chubb, J. R., Trcek, T., Shenoy, S. M. and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Curr Biol* **16**, 1018-1025.

da Cunha, A., Aloyo, V. J. and Vitkovic, L. (1991). Developmental regulation of GAP-43, glutamine synthetase and beta-actin mRNA in rat cortical astrocytes. *Brain Res Dev Brain Res* **64**, 212-215.

Elowitz, M. B., Levine, A. J., Siggia, E. D. and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* **297**, 1183-1186.

Emsley, J. G. and Macklis, J. D. (2006). Astroglial heterogeneity closely reflects the neuronal-defined anatomy of the adult murine CNS. *Neuron Glia Biol* **2**, 175-186.

Enyedi, P. and Czirjak, G. (2010). Molecular background of leak K+ currents: two-pore domain potassium channels. *Physiol Rev* **90**, 559-605.

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., Geho, D. H., Petricoin, E. F., 3rd and Liotta, L. A. (2006). Laser-capture microdissection. *Nat Protoc* **1**, 586-603.

Fan, H. C., Wang, J., Potanina, A. and Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* **29**, 51-57.

Fernandez, E. A. and Balzarini, M. (2007). Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. *Computers in Biology and Medicine* **37**, 1677-1689.

Freeman, M. R. (2010). Specification and morphogenesis of astrocytes. *Science* **330**, 774-778.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Stern, T., Harmelin, A., Eilam, R., Rechavi, G. and Shapiro, E. (2008). Cell lineage analysis of a mouse tumor. *Cancer Res* **68**, 5924-5931.

Ge, W. P., Miyawaki, A., Gage, F. H., Jan, Y. N. and Jan, L. Y. (2012). Local generation of glia is a major astrocyte source in postnatal cortex. *Nature* **484**, 376-380.

Gentle, A., Anastasopoulos, F. and McBrien, N. A. (2001). High-resolution semi-quantitative real-time PCR without the use of a standard curve. *Biotechniques* **31**, 502, 504-506, 508.

Gilbert, D. R., Schroeder, M. and van Helden, J. (2000). Interactive visualization and exploration of relationships between biological objects. *Trends in Biotechnology* **18**, 487-494.

Golding, I., Paulsson, J., Zawilski, S. M. and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025-1036.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annu Rev Psychol* **60**, 549-576.

Grundemann, J., Schlaudraff, F., Haeckel, O. and Liss, B. (2008). Elevated alpha-synuclein mRNA levels in individual UV-laser-microdissected dopaminergic substantia nigra neurons in idiopathic Parkinson's disease. *Nucleic Acids Res* **36**, e38.

Hamilton, N. B. and Attwell, D. (2010). Do astrocytes really exocytose neurotransmitters? *Nat Rev Neurosci* **11**, 227-238.

Honsa, P., Pivonkova, H., Dzamba, D., Filipova, M. and Anderova, M. (2012). Polydendrocytes display large lineage plasticity following focal cerebral ischemia. *PLoS One* **7**, e36816.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24**.

Joseph, A. and Gnanapragasam, V. J. (2011). Laser-capture microdissection and transcriptional profiling in archival FFPE tissue in prostate cancer. *Methods Mol Biol* **755**, 291-300.

Kamme, F., Salunga, R., Yu, J., Tran, D. T., Zhu, J., Luo, L., Bittner, A., Guo, H. Q., Miller, N., Wan, J. et al. (2003). Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J Neurosci* **23**, 3607-3615.

Ko, M. S., Nakauchi, H. and Takahashi, N. (1990). The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *Embo J* **9**, 2835-2842.

Kofuji, P. and Newman, E. A. (2004). Potassium buffering in the central nervous system. *Neuroscience* **129**, 1045-1056.

Kohonen, T. and Honkela, T. (2007). Kohonen nework. *Scholarpedia* **2**.

Koziner, B., Gebhard, D., Denny, T. and Evans, R. L. (1982). Characterization of B-cell type chronic lymphocytic leukemia cells by surface markers and a monoclonal antibody. *Am J Med* **73**, 802-807.

Kucheryavykh, L. Y., Kucheryavykh, Y. V., Inyushin, M., Shuba, Y. M., Sanabria, P., Cubano, L. A., Skatchkov, S. N. and Eaton, M. J. (2009). Ischemia Increases TREK-2 Channel Expression in Astrocytes: Relevance to Glutamate Clearance. *Open Neurosci J* **3**, 40-47.

Kurimoto, K., Yabuta, Y., Ohinata, Y. and Saitou, M. (2007). Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc* **2**, 739-752.

Kuznetsov, V. A., Knott, G. D. and Bonner, R. F. (2002). General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics* **161**, 1321-1332.

Lance, G. N. and Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal* **9**, 373-380.

Levesque, M. J. and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat Methods* **10**, 246-248.

Liss, B., Franz, O., Sewing, S., Bruns, R., Neuhoff, H. and Roeper, J. (2001). Tuning pacemaker frequency of individual dopaminergic neurons by Kv4.3L and KChip3.1 transcription. *Embo J* **20**, 5715-5724.

Liu, H., Yang, M., Qiu, G. P., Zhuo, F., Yu, W. H., Sun, S. Q. and Xiu, Y. (2012). Aquaporin 9 in rat brain after severe traumatic brain injury. *Arq Neuropsiquiatr* **70**, 214-220.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408.

Livak, K. J., Wills, Q. F., Tipping, A. J., Datta, K., Mittal, R., Goldson, A. J., Sexton, D. W. and Holmes, C. C. (2013). Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* **59**, 71-79.

Martinez-Hernandez, A., Bell, K. P. and Norenberg, M. D. (1977). Glutamine synthetase: glial localization in brain. *Science* **195**, 1356-1358.

Matthias, K., Kirchhoff, F., Seifert, G., Huttmann, K., Matyash, M., Kettenmann, H. and Steinhauser, C. (2003). Segregated expression of AMPA-type glutamate receptors and glutamate transporters defines distinct astrocyte populations in the mouse hippocampus. *J Neurosci* **23**, 1750-1758.

Meyer, R. D. and Cook, D. (2000). Visualization of data. *Current Opinion in Biotechnology* **11**, 89-96.

Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., BonDurant, S. S., Richmond, K., Rodesch, M. et al. (2008). Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences* **105**, 1579-1584.

Mongin, A. A. (2007). Disruption of ionic and cell volume homeostasis in cerebral ischemia: The perfect storm. *Pathophysiology* **14**, 183-193.

Nedergaard, M. and Dirnagl, U. (2005). Role of glial cells in cerebral ischemia. *Glia* **50**, 281-286.

Nedergaard, M., Ransom, B. and Goldman, S. A. (2003). New roles for astrocytes: redefining the functional architecture of the brain. *Trends Neurosci* **26**, 523-530.

Neitz, A., Mergia, E., Eysel, U. T., Koesling, D. and Mittmann, T. (2011). Presynaptic nitric oxide/cGMP facilitates glutamate release via hyperpolarization-activated cyclic nucleotide-gated channels in the hippocampus. *Eur J Neurosci* **33**, 1611-1621.

Nilius, B. and Droogmans, G. (2003). Amazing chloride channels: an overview. *Acta Physiol Scand* **177**, 119-147.

Nolte, C., Matyash, M., Pivneva, T., Schipke, C. G., Ohlemeyer, C., Hanisch, U. K., Kirchhoff, F. and Kettenmann, H. (2001). GFAP promoter-controlled EGFP-expressing transgenic mice: a tool to visualize astrocytes and astrogliosis in living brain tissue. *Glia* **33**, 72-86.

Novick, A. and Weiner, M. (1957). Enzyme Induction as an All-or-None Phenomenon. *Proc Natl Acad Sci U S A* **43**, 553-566.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W. et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-1071.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nat Genet* **31**, 69-73.

Palygin, O., Lalo, U. and Pankratov, Y. (2011). Distinct pharmacological and functional properties of NMDA receptors in mouse cortical astrocytes. *Br J Pharmacol* **163**, 1755-1766.

Panickar, K. S. and Norenberg, M. D. (2005). Astrocytes in cerebral ischemic injury: morphological and general considerations. *Glia* **50**, 287-298.

Pearson, K. (1901). {On lines and planes of closest fit to systems of points in space}. *Philosophical Magazine* **2**, 559-572.

Pekny, M. and Nilsson, M. (2005). Astrocyte activation and reactive gliosis. *Glia* **50**, 427-434.

Pekny, M., Wilhelmsson, U., Bogestal, Y. R. and Pekna, M. (2007). The role of astrocytes and complement system in neural plasticity. *Int Rev Neurobiol* **82**, 95-111.

Pesta, M., Fichtl, J., Kulda, V., Topolcan, O. and Treska, V. (2013). Monitoring of circulating tumor cells in patients undergoing surgery for hepatic metastases from colorectal cancer. *Anticancer Res* **33**, 2239-2243.

Pfaffl, M. W., Horgan, G. W. and Dempfle, L. (2002). Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res* **30**, e36.

Pivonkova, H., Benesova, J., Butenko, O., Chvatal, A. and Anderova, M. (2010). Impact of global cerebral ischemia on K+ channel expression and membrane properties of glial cells in the rat hippocampus. *Neurochem Int* **57**, 783-794.

Podgorny, O. V. (2013). Live cell isolation by laser microdissection with gravity transfer. *Journal of Biomedical Optics* **18**, 055002-055002.

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**, e309.

Regan, M. R., Huang, Y. H., Kim, Y. S., Dykes-Hoberg, M. I., Jin, L., Watkins, A. M., Bergles, D. E. and Rothstein, J. D. (2007). Variations in promoter activity reveal a differential expression and physiology of glutamate transporters by glia in the developing and mature CNS. *J Neurosci* **27**, 6607-6619.

Rivers, L. E., Young, K. M., Rizzi, M., Jamen, F., Psachoulia, K., Wade, A., Kessaris, N. and Richardson, W. D. (2008). PDGFRA/NG2 glia generate myelinating oligodendrocytes and piriform projection neurons in adult mice. *Nat Neurosci* **11**, 1392-1401.

Roitbak, T. and Sykova, E. (1999). Diffusion barriers evoked in the rat cortex by reactive astrogliosis. *Glia* **28**, 40-48.

Rouach, N., Koulakoff, A., Abudara, V., Willecke, K. and Giaume, C. (2008). Astroglial metabolic networks sustain hippocampal synaptic transmission. *Science* **322**, 1551-1555.

Rusnakova, V., Honsa, P., Dzamba, D., Ståhlberg, A., Kubista, M. and Anderova, M. (2013). Heterogeneity of Astrocytes: From Development to Injury – Single Cell Gene Expression. *PLoS One* **8**, e69734.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523-529.

Seifert, G., Huttmann, K., Binder, D. K., Hartmann, C., Wyczynski, A., Neusch, C. and Steinhauser, C. (2009). Analysis of astroglial K+ channel expression in the developing hippocampus reveals a predominant role of the Kir4.1 subunit. *J Neurosci* **29**, 7474-7488.

Shlens, J. (2005). A tutorial on principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*.

Silbert, S. C., Beacham, D. W. and McCleskey, E. W. (2003). Quantitative single-cell differences in mu-opioid receptor mRNA distinguish myelinated and unmyelinated nociceptors. *J Neurosci* **23**, 34-42.

Skatchkov, S. N., Eaton, M. J., Shuba, Y. M., Kucheryavykh, Y. V., Derst, C., Veh, R. W., Wurm, A., Iandiev, I., Pannicke, T., Bringmann, A. et al. (2006). Tandem-pore domain potassium channels are functionally expressed in retinal (Muller) glial cells. *Glia* **53**, 266-276.

Smith, L. I. (2002). A tutorial on principal components analysis. *Cornell University, USA* **51**, 52.

Sofroniew, M. V. and Vinters, H. V. (2010). Astrocytes: biology and pathology. *Acta Neuropathol* **119**, 7-35.

Stahlberg, A. and Bengtsson, M. (2010). Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods* **50**, 282-288.

Stahlberg, A., Kubista, M. and Aman, P. (2011a). Single-cell gene-expression profiling and its potential diagnostic applications. *Expert Rev Mol Diagn* **11**, 735-740.

Stahlberg, A., Rusnakova, V. and Kubista, M. (2013a). The added value of single-cell gene expression profiling. *Briefings in Functional Genomics* **12**, 81-89.

Stahlberg, A., Hakansson, J., Xian, X., Semb, H. and Kubista, M. (2004). Properties of the reverse transcription reaction in mRNA quantification. *Clin Chem* **50**, 509-515.

Stahlberg, A., Rusnakova, V., Forootan, A., Anderova, M. and Kubista, M. (2013b). RT-qPCR work-flow for single-cell data analysis. *Methods* **59**, 80-88.

Stahlberg, A., Andersson, D., Aurelius, J., Faiz, M., Pekna, M., Kubista, M. and Pekny, M. (2011b). Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res* **39**, e24.

Stahlberg, A., Elbing, K., Andrade-Garda, J. M., Sjogreen, B., Forootan, A. and Kubista, M. (2008). Multiway real-time PCR gene expression profiling in yeast Saccharomyces cerevisiae reveals altered transcriptional response of ADH-genes to glucose stimuli. BMC Genomics 9, 170.

Sun, W., McConnell, E., Pare, J. F., Xu, Q., Chen, M., Peng, W., Lovatt, D., Han, X., Smith, Y. and Nedergaard, M. (2013). Glutamate-dependent neuroglial calcium signaling differs between young and adult brain. *Science* **339**, 197-200.

Swanson, R. A., Ying, W. and Kauppinen, T. M. (2004). Astrocyte influences on ischemic neuronal death. *Curr Mol Med* **4**, 193-205.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382.

Thoren, A. E., Helps, S. C., Nilsson, M. and Sims, N. R. (2005). Astrocytic function assessed from 1-14C-acetate metabolism after temporary focal cerebral ischemia in rats. *J Cereb Blood Flow Metab* **25**, 440-450.

Tichopad, A., Dilger, M., Schwarz, G. and Pfaffl, M. W. (2003). Standardized determination of real-time PCR efficiency from a single reaction set-up. *Nucleic Acids Res* **31**, e122.

Tichopad, A., Kitchen, R., Riedmaier, I., Becker, C., Stahlberg, A. and Kubista, M. (2009). Design and optimization of reverse-transcription quantitative PCR experiments. *Clin Chem* **55**, 1816-1823.

Trask, B. J. (2002). Human cytogenetics: 46 chromosomes, 46 years and counting. *Nat Rev Genet* **3**, 769-778.

Trotter, J., Karram, K. and Nishiyama, A. (2010). NG2 cells: Properties, progeny and origin. *Brain Res Rev* **63**, 72-82.

Tsuchiya, K. D. (2011). Fluorescence In Situ Hybridization. *Clinics in Laboratory Medicine* **31**, 525-542.

Ulas, J., Satou, T., Ivins, K. J., Kesslak, J. P., Cotman, C. W. and Balazs, R. (2000). Expression of metabotropic glutamate receptor 5 is increased in astrocytes after kainate-induced epileptic seizures. *Glia* **30**, 352-361.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034.

Vassar, R., Ngai, J. and Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell* **74**, 309-318.

Veys, K., Labro, A. J., De Schutter, E. and Snyders, D. J. (2012). Quantitative single-cell ion-channel gene expression profiling through an improved qRT-PCR technique combined with whole cell patch clamp. *Journal of Neuroscience Methods* **209**, 227-234.

Volfson, D., Marciniak, J., Blake, W. J., Ostroff, N., Tsimring, L. S. and Hasty, J. (2006). Origins of extrinsic variability in eukaryotic gene expression. *Nature* **439**, 861-864.

Walz, W. (1989). Role of glial cells in the regulation of the brain ion microenvironment. *Prog Neurobiol* **33**, 309-333.

Walz, W. (2000). Controversy surrounding the existence of discrete functional classes of astrocytes in adult gray matter. *Glia* **31**, 95-103.

Wang, D. and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology* **28**, 281-290.

Weinlichova, J. and Fejfar, J. (2010). Využití samoorganizující se neuronové sítě pro hodnocení chovaní spotřebitele. *Acta Universitatis Agriculturae et silviculturae mendelianea brunensinsis* **LVIII**, 8.

Welle, S. (2013). What Statisticians Should Know About Microarray Gene Expression Technology. In *Statistical Methods for Microarray Data Analysis* (eds A. Y. Yakovlev L. Klebanov and D. Gaile), pp. 1-13: Springer New York.

Wernet, M. F., Mazzoni, E. O., Celik, A., Duncan, D. M., Duncan, I. and Desplan, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* **440**, 174-180.

Xiang, C. C., Mezey, E., Chen, M., Key, S., Ma, L. and Brownstein, M. J. (2004). Using DSP, a reversible cross-linker, to fix tissue sections for immunostaining, microdissection and expression profiling. *Nucleic Acids Res* **32**, e185.

Xu, E. Y., Zawadzki, K. A. and Broach, J. R. (2006). Single-cell observations reveal intermediate transcriptional silencing states. *Mol Cell* **23**, 219-229.

Zelenina, M. (2010). Regulation of brain aquaporins. *Neurochem Int* **57**, 468-488.

Další publikace autora

Bibova, I., Linhartova, I., Stanek, O., Rusnakova, V., Kubista, M., Suchanek, M., Vasakova, M. and Sebo, P. (2012). Detection of immune cell response to M. tuberculosis-specific antigens by quantitative polymerase chain reaction. *Diagnostic Microbiology and Infectious Disease***72**, 68-78. (IF=2.260)

Butenko, O., Dzamba, D., Benesova, J., Honsa, P., Benfenati, V., Rusnakova, V., Ferroni, S. and Anderova, M. (2012). The increased activity of TRPV4 channel in the astrocytes of the adult rat hippocampus after cerebral hypoxia/ischemia. *PLoS One***7**, e39959. (IF=3,730)

Kozubenko, N., Turnovcova, K., Kapcalova, M., Butenko, O., Anderova, M., Rusnakova, V., Kubista, M., Hampl, A., Jendelova, P. and Sykova, E. (2010). Analysis of In Vitro and In Vivo Characteristics of Human Embryonic Stem Cell-Derived Neural Precursors. *Cell Transplantation*19, 471-486.(IF=4.422)

Polakova Vymetalkova, V., Vannucci, L., Korenkova, V., Prochazka, P., Slyskova, J., Vodickova, L., Rusnakova, V., Bielik, L., Burocziova, M., Rossmann, P. et al., (2013). Evaluation of tumor suppressor gene expressions and aberrant methylation in the colon of cancer-induced rats: a pilot study. *Mol Biol Rep.* (IF= 2.506.)

Sorg, D., Danowski, K., Korenkova, V., Rusnakova, V., Kuffner, R., Zimmer, R., Meyer, H. H. D. and Kliem, H. (2013) Microfluidic high-throughput RT-qPCR measurements of the immune response of primary bovine mammary epithelial cells cultured from milk to mastitis pathogens. Animal **7**, p. 799-805. (IF=1,648)