Univerzita Karlova v Praze

Přírodovědecká fakulta

DOKTORSKÁ DISERTAČNÍ PRÁCE

Uspořádání genů v genomech eukaryot:

analýza s využitím sekvenačních genových expresních dat

Petr Divina

Školitel: Prof. MUDr. Jiří Forejt, DrSc.

Praha 2007

Tuto doktorskou disertační práci jsem vypracoval samostatně a nepředložil jsem ji ani její podstatnou část k získání jiného ani stejného akademického titulu.

Poděkování

Rád bych poděkoval svému školiteli Jiřímu Forejtovi za to, že jsem v jeho laboratoři mohl pracovat na své disertační práci. Děkuji zejména za vědeckou svobodu, kterou mi poskytl a která mi umožnila rozvíjet své schopnosti směrem k bioinformatice. Děkuji také všem kolegům, pracovníkům a přátelům, které jsem během práce na oddělení myší molekulární genetiky ÚMG poznal. Zvláště si dovolím poděkovat Honzovi Pačesovi, který mě dokázal pro bioinformatiku nadchnout a s jehož pomocí jsem vytvořil svoji první myší databázi. Děkuji také Radce Storchové, díky níž jsem poznal, jak skvěle se dají využít bio-informatické přístupy při testování vědeckých hypotéz. Dále chci poděkovat Lence Piherové a Robertovi Ivánkovi, s jejichž pomocí jsem pronikl do analýzy dat z DNA čipů. Jiřímu Forejtovi, Zdeňku Trachtulcovi a Radce Storchové děkuji také za cenné komentáře k mé disertační práci. Velký dík patří také mé mamince a mé rodině za podporu během celé doby mého studia.

OSNOVA

1. ÚVOD

1.1. Uspořádání genů v genomech eukaryot

Uspořádání genů v genomech eukaryotických organismů bylo dlouho považováno za náhodné. Existovala představa o *trans*-regulaci genové exprese, při níž si transkripční faktory, které rozhodují o tom, jaké geny budou v určité tkáni exprimovány, vyhledají příslušné regulační sekvence nezávisle na tom, v jakém místě genomu tyto sekvence leží. Některé odchylky od náhodného uspořádání byly sice popsány, např. u genů vzniklých tandemovými duplikacemi (Shen *et al.*, 1981) nebo u genů, které sdílejí *cis*-působící regulační elementy (Levine *et al.*, 1983; Scott *et al.*, 1983), avšak tyto případy byly považovány spíše za výjimky. Postupem času však přibylo studií, které ukázaly, že původní představu o náhodném rozložení genů v genomech je potřeba přehodnotit. Díky dostupnosti kompletní genomové sekvence řady organismů (Adams *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001; Waterston *et al.*, 2002; Hillier *et al.*, 2004) a velkých souborů genových expresních dat (Boguski *et al.*, 1993; Caron *et al.*, 2001; Su *et al.*, 2002) bylo možné provést analýzy uspořádání genů na úrovni celých genomů. S využitím těchto dat byly nalezeny důkazy o nenáhodném uspořádání genů v každém eukaryotickém genomu, který byl dosud studován (Hurst *et al.*, 2004). Pochopení tohoto fenoménu má význam pro odhalení mechanismů, kterými byl obsah genomů formován během evoluce, ale také pro objasnění, jakým způsobem je regulována a koordinována exprese genů v různých částech genomu. Z nových poznatků je zřetelné, že genomy eukaryotických organismů neslouží pouze jako statická úložiště pro genetickou informaci, ale jejich genový obsah podléhá dynamickým změnám. Uspořádání genů v genomech eukaryotických organismů je možné posuzovat ve dvou úrovních: *mezi chromosomy* a *podél chromosomů*.

1.1.1. Nenáhodný obsah genů na pohlavních chromosomech

Při studiu uspořádání genů mezi chromosomy je obvykle porovnávána distribuce genů exprimovaných v určité tkáni na jednotlivých chromosomech, případně mezi autosomy a

pohlavními chromosomy. U různých organismů bylo zjištěno, že jednotlivé autosomy jsou si svým genovým obsahem navzájem velmi podobné a každý z autosomů obsahuje vesměs náhodný soubor genů z celého genomu (Vallender & Lahn, 2004). Přestože některé výjimky z tohoto pravidla byly zaznamenány (Bortoluzzi *et al.*, 1998; Khil *et al.*, 2004), nebyl pozorován systematický trend v souvislosti se zastoupením určitých typů genů na jednotlivých autosomech. Naproti tomu genový obsah pohlavních chromosomů je u většiny organismů odlišný od autosomů.

Pohlavní chromosomy hrají klíčovou roli při určování pohlaví u většiny živočichů i některých rostlin. Vyvinuly se z původního páru autosomů v několika nezávislých evolučních liniích (Ohno, 1967; Matsubara *et al.*, 2006). U různých taxonů existují odlišné systémy chromosomového určení pohlaví. Mezi nejvíce prostudované systémy určení pohlaví patří systém XX:XY u savců a drozofily, systém XX:X0 u *Caenorhabditis elegans* a systém ZZ:ZW u ptáků (Vallender & Lahn, 2006). Pohlaví, které má dva stejné pohlavní chromosomy se označuje jako homogametické a pohlaví, které má dva různé pohlavní chromosomy je heterogametické. Zatímco u savců jsou samčí jedinci heterogametičtí (XY) a samice homogametické (XX), u ptáků je tomu naopak – samice jsou heterogametické (ZW) a samci homogametičtí (ZZ).

Pro nepárové pohlavní chromosomy (Y a W) je typické, že obsahují velmi malý počet funkčních genů a naopak velké množství nefunkčních pseudogenů, transpozonů a jiných repetitivních sekvencí, které jsou obvykle soustředěny v rozsáhlých blocích heterochromatinu (Vallender & Lahn, 2004; Fukova *et al.*, 2007). Genový obsah nepárového pohlavního chromosomu Y byl nejlépe prostudován u člověka (Skaletsky *et al.*, 2003). Podstatnou část lidského chromosomu Y (asi 95%) zaujímá oblast specifická pro samce (MSY, male-specific region of the Y chromosome), která není homologní k chromosomu X a proto v ní nedochází ke crossing-overu. Oblast MSY je složena z mozaiky heterochromatinu a tří typů euchromatinových sekvencí a v porovnání s autosomy je genově chudá. Geny ležící v této oblasti chromosomu Y lze rozdělit do dvou funkčních skupin. První skupinu tvoří geny exprimované v širokém spektru tkání (housekeeping geny), které mají homologní gen na chromosomu X. Druhá skupina zahrnuje geny s výlučnou expresí ve varleti (Lahn & Page, 1997). Tyto geny leží v ramenech osmi velkých obrácených repetic a podléhají genové konverzi. Na obou koncích lidského chromosomu Y jsou umístěny pseduautosomální oblasti (PAR), které umožňují párování a rekombinaci chromosomů X a Y během meiózy. Pseudoautosomální oblasti se svým nukleotidovým složením, hustotou a obsahem genů podobají autosomům. Geny obsažené

v PAR se stejně jako autosomální geny vyskytují v obou pohlavích. Genový obsah chromosomu Y na úrovni DNA sekvence byl také částečně charakterizován u šimpanze (Hughes *et al.*, 2005). Na šimpanzím chromosomu Y však bylo nalezeno mnohem více nefunčních genů, které zdegenerovaly během evoluce šimpanzí linie, než bylo zjištěno na lidském chromosomu Y. Chromosom Y myši je v současnosti teprve sekvenován (Alfoldi *et al.,* 18th International Mouse Genome Conference, 2004). Z předchozích studií je však známo, že myší chromosom Y obsahuje geny nezbytné pro spermatogenezi (Kay *et al.*, 1991; Agulnik *et al.*, 1994; Mazeyrat *et al.*, 1998). U drozofily je chromosom Y velmi malý, téměř celý vyplněný heterochromatinem (Adams *et al.*, 2000) a bylo na něm identifikováno několik genů, které ovlivňují plodnost samců (Carvalho *et al.*, 2001; Charlesworth, 2001).

Párové pohlavní chromosomy X a Z jsou svou velikostí, počtem a hustotou genů spíše podobné autosomům. Řada studií však dokládá, že některé funkční skupiny genů jsou na chromosomu X zastoupené více a jiné naopak méně než na autosomech (Hurst, 2001). Pozoruhodný je zejména nenáhodný obsah genů, které jsou preferenčně nebo výlučně exprimované v jednom pohlaví. Studie různých organismů však přinesly protichůdné poznatky ohledně zastoupení těchto genů na chromosomu X. Lidský chromosom X je například obohacen o geny, které souvisejí s rozmnožováním nebo s určením pohlaví (Hurst & Randerson, 1999; Saifi & Chandra, 1999). Geny, které jsou specificky exprimované v prostatě člověka, jsou na chromosomu X zastoupené dvakrát častěji než na autosomech (Lercher *et al.*, 2003b). Chromosom X člověka je také obohacen o geny exprimované v mozku, jejichž mutace způsobují mentální poruchy (Zechner *et al.*, 2001) a dále o geny exprimované ve svalech (Bortoluzzi *et al.*, 1998). Podobně jako na lidském chromosomu Y bylo také na lidském chromosomu X odhaleno několik velkých obrácených repetic, které ve svých ramenech obsahují geny preferenčně exprimované ve varleti a umožňují genovou konverzi (Warburton *et al.*, 2004). U myši bylo zjištěno, že geny specificky exprimované ve spermatogoniích jsou řádově vícekrát zastoupeny na chromosomu X (Wang *et al.*, 2001). Také geny s preferenční expresí v samičích tkáních (vaječnících a placentě) jsou na myším chromosomu X zastoupeny častěji (Khil *et al.*, 2004). V rozporu s těmito poznatky je však zjištění, že myší chromosom X je ochuzen o geny exprimované ve varleti (Khil *et al.*, 2004). U drozofily bylo zjištěno, že se na chromosomu X téměř nevyskytují geny s preferenční expresí v samcích (Parisi *et al.*, 2003) a u *Caenorhabditis elegans* je chromosom X ochuzen o geny exprimované ve spermatogenních a oogenních buňkách (Reinke *et al.*, 2000; Reinke *et al.*, 2004). Přesto, že

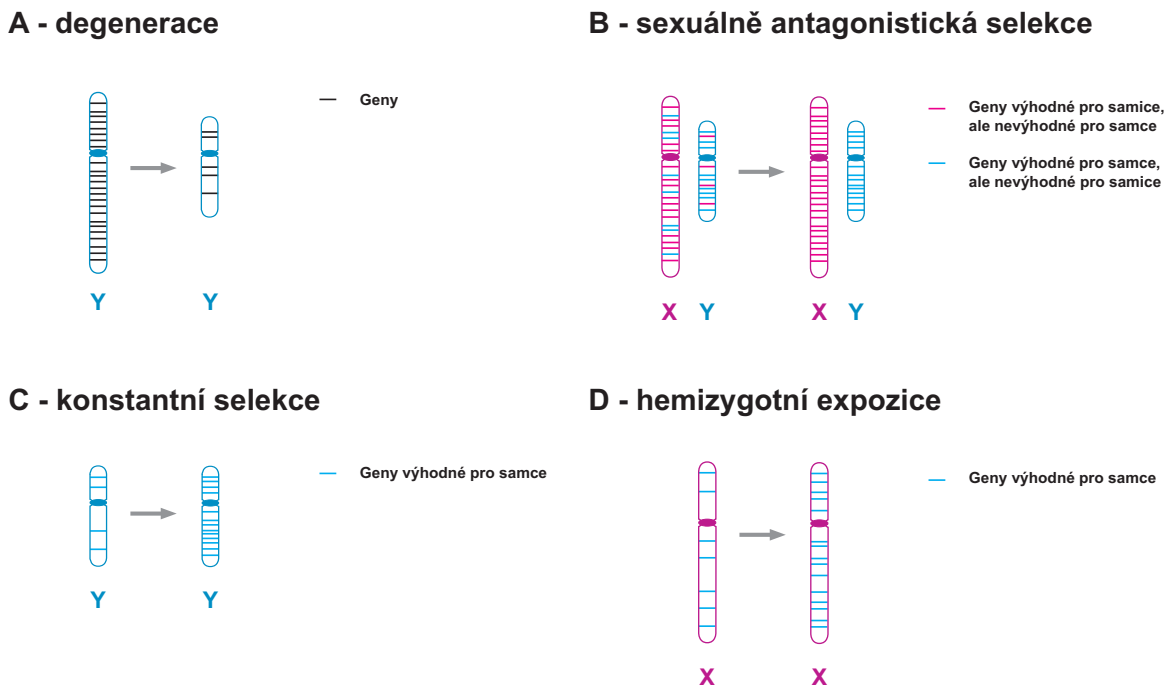
je k dispozici kompletní genomová sekvence kura domácího (Hillier *et al.*, 2004), genový obsah párového pohlavního chromosomu Z nebyl zatím uspokojivě prostudován.

1.1.2. Mechanismy, které ovlivňují genový obsah pohlavních chromosomů

Pro vysvětlení genového obsahu pohlavních chromosomů je třeba pochopit, jak pohlavní chromosomy vznikly a jak probíhala jejich evoluce. Vznik pohlavních chromosomů úzce souvisí se změnou způsobu určení pohlaví z negenetického na genetický. U negenetického způsobu určení pohlaví, který je rozšířen u mnoha živočichů, je pohlaví jedinců určeno působením faktorů vnějšího prostředí (např. teplotou během embryonálního vývoje plazů). Při genetickém způsobu určení pohlaví je pohlaví jedinců určeno v okamžiku oplození podle kombinace pohlavních chromosomů. Pohlavní chromosomy vznikly z původního páru autosomů, který neměl ničím výjimečný genový obsah (Vallender & Lahn, 2004). Spouštěcím mechanismem pro vznik pohlavních chromosomů byla mutace, která změnila jeden z genů ležících na tomto autosomu v gen určující pohlaví. Pohlavní chromosomy savců a ptáků vznikly z odlišného páru autosomů a v důsledku mutací odlišných genů určujících pohlaví (Graves & Shetty, 2001). U savců pravděpodobně došlo ke vzniku pohlavních chromosomů mutací genu *SOX3*, která změnila jeho kopii na původním chromosomu Y v gen *SRY* určující samčí pohlaví (Stevanovic *et al.*, 1993; Foster & Graves, 1994). U ptáků byl na chromosomu Z identifikován kandidátní gen *DMRT1,* jehož mutace mohla být určující pro vznik samčího pohlaví, a dva kandidátní geny (*ASW* a *FET1*) ležící na chromosomu W, jejichž mutace mohla dát vznik samičímu pohlaví (Smith *et al.*, 1999). Ačkoliv pohlavní chromosomy vznikly během evoluce několikrát nezávisle (Ohno, 1967; Matsubara *et al.*, 2006), jejich následný vývoj má ve všech liniích podobné rysy (Graves & Shetty, 2001).

Podle současných evolučních teorií je genový obsah pohlavních chromosomů X a Y výsledkem působení nejméně čtyř evolučních procesů (Obr. 1) (Vallender & Lahn, 2004). (1) V důsledku potlačení meiotické rekombinace došlo na chromosomu Y k *degeneraci* a redukci počtu genů (Ohno, 1967). (2) Geny ležící na chromosomu Y byly vystaveny *konstantní selekci*, která vedla k fixaci genů, které jsou výhodné pro samce (Vallender & Lahn, 2004). (3) *Sexuálně antagonistická selekce* působila na geny, které jsou výhodné pro jedno pohlaví a zároveň škodlivé pro druhé pohlaví a umožnila jejich rozdílné hromadění na obou pohlavních chromosomech (Rice, 1984; Hurst, 2001).

(4) Na chromosomu X došlo v důsledku jeho *hemizygotní expozice* také k nahromadění recesívních mutací genů výhodných pro samce.



Obr. 1. Důsledky působení čtyř evolučních sil na pohlavní chromosomy.

(A) Redukce počtu genů na chromosomu Y v důsledku jeho degenerace. (B) Sexuální antagonismus způsobuje feminizaci chromosomu X a maskulinizaci chromosomu Y. (C) Konstantní selekce udržuje na chromosomu Y geny výhodné pro samce. (D) V hemizygotním stavu se na chromosomu X projeví účinek recesívních alel. (Podle Vallender & Lahn, 2004, upraveno).

Genový obsah pohlavních chromosomů mohou ovlivňovat také jejich epigenetické modifikace, ke kterým dochází v souvislosti s meiotickou inaktivací pohlavního chromosomu X ve spermatogenezi (Lifschytz & Lindsley, 1972; Kelly *et al.*, 2002; McCarrey *et al.*, 2002; Khil *et al.*, 2004) a kompenzací dávky genů v somatických buňkách (Meller, 2000). Předpokládá se, že právě meiotická inaktivace chromosomu X u samců je nejpravděpodobnější příčinou ochuzení chromosomu X o geny exprimované během spermatogeneze. Jedním z mechanismů, kterým k tomuto ochuzení dochází, mohou být retropozice. V genomech savců i drozofily bylo nalezeno signifikantně více retropozic z chromosomu X na autosomy (Betran *et al.*, 2002; Emerson *et al.*, 2004). Velké množství těchto retrogenů navíc vykazuje expresi právě během spermatogeneze.

Evoluční procesy, které formují genový obsah pohlavních chromosomů XY a ZW se mohou trochu lišit. Jelikož chromosom Z se vyskytuje častěji u samců než chromosom X, je vystaven vyšší mutační rychlosti, neboť během spermatogeneze dochází k více buněčným dělením než během oogeneze (Ellegren & Fridolfsson, 1997; Montell *et al.*, 2001; Axelsson *et al.*, 2004; Kirkpatrick & Hall, 2004). Vyšší mutační rychlost vytváří více genetické variability pro působení selekce. Chromosom Z je také citlivější k účinkům pohlavního výběru než chromosom X (Reeve & Pfennig, 2003). U organismů s heterogametickými samicemi nebyl genový obsah pohlavních chromosomů prozkoumán. Nelze proto spolehlivě rozhodnout zda pohlavní chromosomy Z a W ovlivňují stejné evoluční či epigenetické mechanismy jako u chromosomů X a Y.

1.1.3. Nenáhodné uspořádání genů podél chromosomů

Analýzy genomové sekvence u různých organismů ukázaly, že hustota genů a obsah GC nejsou rovnoměrné po celé délce chromosomů (Lercher *et al.*, 2003c; Paces *et al.*, 2004). Díky dostupnosti velkých souborů genových expresních dat bylo navíc zjištěno, že některé skupiny genů jsou podél chromosomů rozmístěny nenáhodně a mají tendenci vyskytovat se ve shlucích (klastrech). Při analýze shlukování (klastrování) genů je obvykle srovnáváno, jak často se genové klastry pozorované ve skutečném genomu vyskytují v mnoha náhodně vygenerovaných genomech. Není přitom podstatné, zda v náhodných genomech klastrují vybrané konkrétní geny, ale zda vždy klastrují geny, které patří do určité předem vybrané skupiny genů (např. koexprimované geny, geny s podobnou funkcí). Většina studií se také při analýze klastrování snaží eliminovat vliv tandemových duplikací jako jednu z možných příčin nenáhodného uspořádání genů v genomu.

Ve všech dosud analyzovaných genomech eukaryotických organismů bylo zjištěno, že koexprimované geny jsou určitým způsobem uspořádány a mají tendenci se vyskytovat v klastrech (Kosak & Groudine, 2004). S využitím genových expresních dat získaných pomocí metody SAGE (sériová analýza genové exprese) z 12 lidských tkání byla sestavena transkripční mapa lidského genomu, která ukázala nerovnoměrnou distribuci genové exprese podél chromosomů (Caron *et al.*, 2001). Úseky s vysokou aktivitou genů (RIDGEs, regions of increased gene expression) byly oddělené dlouhými úseky s nízkou aktivitou genů (anti-RIDGEs) a jejich rozmístění korespondovalo s genově bohatými a genově chudými oblastmi lidského genomu. Úseky s vysokou aktivitou genů byly tvořeny

převážně provozními (housekeeping) geny, které byly vysoce exprimované ve všech tkáních (Lercher *et al.*, 2002). Bylo také zjištěno, že geny mají tendenci klastrovat v závislosti na šířce jejich exprese (počtu tkání v nichž jsou exprimovány) a že dominantní trend pro klastrování projevují právě geny exprimované v mnoha tkáních (Hurst *et al.*, 2004). V úsecích s vysokou i nízkou aktivitou genů mohou být také zastoupeny geny s tkáňově specifikou či preferenční expresí (Versteeg *et al.*, 2003). V lidském genomu však zatím nebyly identifikovány souvislé úseky více genů, které by byly výlučně exprimované ve stejném typu tkáně (Hurst *et al.*, 2004). Velikost úseků s vysokou aktivitou genové exprese může také souviset se udržováním syntenií v genomech. Srovnání lidské a myší genomové sekvence ukázalo, že alespoň polovina syntenních oblastí mezi oběma genomy má délku ~ 20 Mbp, což přibližně odpovídá velikosti RIDGEs (Waterston *et al.*, 2002). Kromě uspořádání genů na úrovní RIDGEs přinesly genomické přístupy také poznatky o nenáhodném uspořádání genů na jemnější úrovni (Kosak & Groudine, 2004). Analýza genové exprese u drozofily s využitím DNA čipů odhalila, že asi 20 % koexprimovaných genů je uspořádáno ve skupinách o 10-30 genech, které se vyskytují v úsecích o délce 20 až 200 kb (Spellman & Rubin, 2002). Důkazy o klastrování genů v genomu drozofily přinesla také analýza veřejně dostupných EST knihoven. Asi 45 % genů exprimovaných výlučně ve varlatech drozofily se nacházelo v nepřerušovaných úsecích obsahujících 4 a více genů (Boutanaev *et al.*, 2002). Také u genů koexprimovaných v embryu drozofily a v oblasti hlavy dospělých jedinců byla zjištěna tendence ke klastrování (Boutanaev *et al.*, 2002). V genomu drozofily byly také nalezeny geny uspořádané v operonech, které jsou přepisovány do polycistronní pre-mRNA (Trachtulec, 2004; Ben-Shahar *et al.*, 2007). V genomu *C. elegans* se přibližně 15 % genů vyskytuje v operonech dvou až osmi genů přepisovaných do polycistronní pre-mRNA (Blumenthal *et al.*, 2002). Tyto operony se liší od operonů bakterií a drozofily tím, že transkripty jednotlivých genů jsou z pre-mRNA sestříhávány *trans*-sestřihem (Blumenthal *et al.*, 2002). Ačkoliv operonové geny představují většinu koexprimovaných genů u červa, signifikantní tendence ke klastrování koexprimovaných genů byla popsána i u genů, které nebyly součástí operonů (Lercher *et al.*, 2003a). Pomocí DNA čipů a mRNA tagging bylo zjištěno, že geny koregulované v určitých buněčných typech červa (svalech, spermiích, oocytech a germinálních buňkách) tvoří klastry o dvou až pěti genech (Roy *et al.*, 2002). Nenáhodné uspořádání genů bylo také popsáno u jednobuněčného eukaryotického organismu, kvasinky *Saccharomyces cerevisiae*. Geny účastnící se buněčného cyklu, sporulace a feromonové odpovědi u pučící kvasinky byly uspořádané v párech roztroušených po celém genomu (Cohen *et al.*, 2000).

U řady těchto genů byla také identifikována společná aktivační sekvence ovlivňující transkripci (UAS, upstream activating sequence).

Další studie se zabývaly tím, zda tendenci ke klastrování vykazují geny s podobnou funkcí. Na rozdíl od koexprimovaných genů, však definice podobné funkce genů není zcela jednoznačná. Tyto geny mohou být například součástí stejné metabolické dráhy, kódovat navzájem interagující proteiny nebo ovlivňovat společný fenotyp (Hurst *et al.*, 2004). S využitím databáze KEGG (Ogata *et al.*, 1999) byla v genomech různých organismů analyzována distribuce genů, jejichž produkty byly součástí stejné metabolické dráhy. V genomech člověka, červa, drozofily a kvasinky byla zjištěna signifikantní tendence ke klastrování u genů ze stejných metabolických drah (Lee & Sonnhammer, 2003). Množství metabolických drah, jejichž geny vykazovaly klastrování bylo u jednotlivých organismů variabilní (od 30 % u drozofily po 98 % u kvasinky), avšak větší než by se dalo očekávat v případě náhodného výběru genů (Lee & Sonnhammer, 2003). Studie vazebné nerovnováhy u 60 myších inbredních kmenů přinesla důkazy, že nejméně čtvrtina myšího genomu obsahuje klastry genů s podobnou funkcí (Petkov *et al.*, 2005). Potomci, kteří zdědili alely těchto genů v určitých kombinacích měli výhodu při přežívání v dalších generacích příbuzenského křížení. U kvasinky bylo popsáno, že geny, které kódovaly podjednotky stabilních proteinových komplexů, měly tendenci vyskytovat se v genomu v těsné blízkosti, ve vzdálenosti 10-30 kb (Teichmann & Veitia, 2004). U kvasinky bylo rovněž zjištěno, že koexprimované geny umístěné v klastrech často patří do stejné funkční kategorie podle databáze Gene Ontology (Harris *et al.*, 2004); v genomu člověka však podobný trend nebyl zaznamenán (Fukuoka *et al.*, 2004).

1.1.4. Mechanismy transkripční regulace

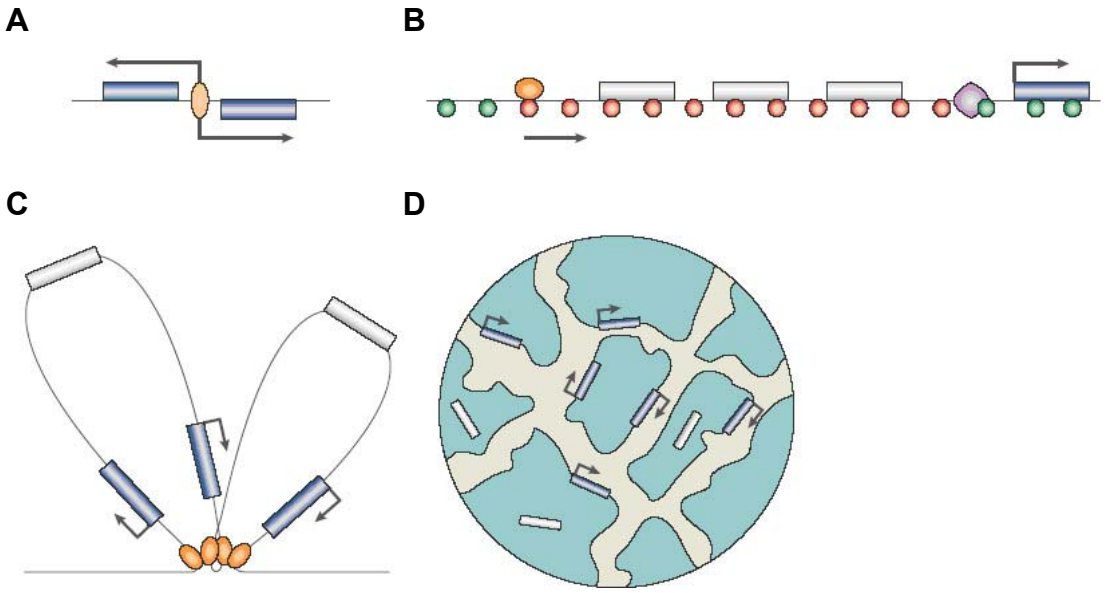
Z předchozích studií vyplývá, že nenáhodné uspořádání koexprimovaných genů podél chromosomů existuje na několika různých úrovních. Koexprese byla zjištěna u malých klastrů obsahujících jen několik málo genů, ale i u velkých úseků genomu. Mechanismy, které regulují transkripci genů působí ve třech hierarchických úrovních (van Driel *et al.*, 2003) (Obr. 2). Na *lokální* úrovni je exprese určitého genu ovlivněna promotorem, který řídí transkripci přilehlých genů a případně zesilovačem (enhancerem), který zvyšuje aktivitu promotoru. Obousměrný promotor může zajišťovat koexpresi přilehlých genů na obou řetězcích DNA (Kruglyak & Tang, 2000; Trinklein *et al.*, 2004). U genů vzniklých

tandemovými duplikacemi dochází ke koexpresi v důsledku toho, že geny mají podobné promotory (Lercher *et al.*, 2003a; Papp *et al.*, 2003). Některé tandemově uspořádané geny mohou být přepisovány do jedné pre-mRNA spolu se sousedním genem a následně sestříhány do jednoho transkriptu (Akiva *et al.*, 2006; Parra *et al.*, 2006). V případě operonových genů u *C. elegans* zajišťuje jeden promotor koexpresi všech genů z operonu do jedné molekuly polycistronní pre-mRNA (Blumenthal *et al.*, 2002). Koexprese několika genů ve vazbě může být také zajištěna fúzí těchto genů do jednoho proteinového produktu (Zhang & Smith, 1998).

Pro vysvětlení koexprese genů ve velkých úsecích genomu je třeba uvažovat dvě vyšší úrovně organizace chromosomů: *stav chromatinu* a *umístění chromatinu uvnitř buněčného jádra* (Hurst *et al.*, 2004). Tyto dvě úrovně není vždy možné zcela jednoznačně rozlišit. Kondenzovaný chromatin (heterochromatin), v němž jsou geny nepřístupné působení transkripčních faktorů, má tendenci se obvykle vyskytovat při okraji jádra (Cremer & Cremer, 2001; Hurst *et al.*, 2004). Změny kondenzace chromatinu jsou spojeny se změnou genové exprese (Eberharter & Becker, 2002) a obvykle doprovázeny modifikacemi histonů (Strahl & Allis, 2000). Podle současného modelu specifické proteiny modifikují histony od kontrolního lokusu podél chromosomu až k hraničnímu elementu (Labrador & Corces, 2002; Turner, 2002) a tyto modifikované histony pak potlačují tranksripci přilehlých genů v tomto úseku. Tento model však nemusí mít univerzální platnost. Zda bude gen skutečně exprimován závisí také na dalších faktorech, např. metylaci DNA, pozici v buněčném jádře, dostupnosti transkripčních faktorů či přítomnosti *cis*-působících regulačních elementů (Cohen *et al.*, 2000). Nová studie ukázala, že v lidském genomu existují chromatinové domény o délce několika desítek až stovek kilobází, které mají podobné histonové modifikace a jsou oddělené úseky DNA (insulátory), na jejichž hranicích se váže insulátorový protein *CTCF* (Barski *et al.*, 2007).

Pro transkripci genů má význam také prostorové uspořádání genomu uvnitř jádra (Hurst *et al.*, 2004). Geny umístěné v blízkosti okraje jádra jsou obvykle transkripčně neaktivní (Andrulis *et al.*, 1998). Naproti tomu během interfáze zaujímá každý chromosom v jádře určitou pozici, v níž jsou geny přístupné transkripci a sestřihu (Cremer & Cremer, 2001). Genově bohaté chromosomy mají tendenci se vyskytovat spíše ve vnitřní části jádra, zatímco genově chudé chromosomy lokalizují u jeho okraje (Boyle *et al.*, 2001). Geny, jejichž transkripce je v genomu kvasinky řízena stejným transkripčním faktorem, jsou v DNA rozmístěny v pravidelných intervalech (Kepes, 2003). Stočením vlákna DNA do pravidelných smyček se vzdálené promotory těchto genů dostanou do vzájemné

blízkosti a mohou vytvořit tzv. *aktivní chromatinové centrum („hub“)* (de Laat & Grosveld, 2003). Exprese genů může poté probíhat v blízkosti těchto regulačních elementů, zatímco geny ležící na smyčkách DNA nejsou přístupné transkripci. Nové poznatky ukazují, že úseky DNA mohou být do transkripčních center aktivně umísťovány, což následně způsobí zahájení či umlčení transkripce příslušných genů (Osborne *et al.*, 2004). Genovou expresi a prostorové umístění genů v jádře mohou také ovlivňovat epigenetické regulace prostřednictvím změny struktury chromatinu (van Driel *et al.*, 2003).



Obr. 2. Znázornění různých úrovní transkripční regulace.

(A) Na primární úrovni ovlivňují transkripci okolních genů *cis*-působící regulační elementy. Obousměrný promotor reguluje transkripci genů na obou řetězcích DNA v úseku několika kilobází (~ 10kb). (B) Na sekundární úrovni dochází k modifikaci histonů v úseku mezi kontrolním lokusem (zobrazen oranžově) a hraničním elementem (zobrazen fialově). Modifikované histony (zobrazené červeně) potlačují transkripci přilehlých genů (šedé boxy). Transkripce genů (modré boxy) probíhá až za hraničním elementem, kde nemodifikované histony (zobrazené zeleně) udržují otevřenou strukturu chromatinu. Tento typ regulace transkripce ovlivňuje geny v úseku několika stovek kilobází (~ 100 kb). (C) Na terciální úrovni se sdružují *cis*-působící regulační elementy z několika chromatinových smyček a tvoří transkripční centrum (tzv. aktivní chromatinový hub). Geny ležící v blízkosti tohoto centra (modré boxy) jsou přístupné transkripci, vzdálené geny (šedé boxy) transkribovány nejsou. Na této úrovni mohou regulační elementy ovlivňovat transkripci genů až na vzdálenost několika megabází (~ 1000 kb). (D) Alternativní zobrazení terciální úrovně transkripční regulace ukazuje uspořádání chromatinu do kompaktních chromosomových domén. Transkripce genů probíhá pouze na povrchu domén (modré boxy); geny umístěné uvnitř domén nejsou transkribovány (šedé boxy). (Hurst *et al.*, 2004).

1.1.5. Vznik a udržování genových klastrů

O příčinách vzniku genových klastrů je možné spekulovat. Řídící silou pro klastrování genů by mohla být *pozitivní selekce na koregulované geny* (Hurst *et al.*, 2004). Tuto představu podporuje zjištění, že u řady organismů byly nalezeny klastry koexprimovaných genů s podobnou funkcí (Lee & Sonnhammer, 2003). Nenáhodné uspořádání genů však nemusí být výsledkem působení selekce, neboť při transkripci může náhodně docházet také k mírné expresi okolních genů (Spellman & Rubin, 2002). Úseky koexprimovaných genů také ne vždy obsahují geny s podobnou funkcí (Spellman & Rubin, 2002). Genové klastry by mohly také vzniknout v důsledku *pozitivní selekce na vázané geny* (Hurst *et al.*, 2004). Studie ukázaly, že esenciální geny v genomu kvasinky a *C. elegans* byly uspořádány v klastrech, které se vyskytovaly v oblastech genomu s nízkou frekvencí rekombinace (Kamath *et al.*, 2003; Pal & Hurst, 2003). Také v genomu myši byly identifikovány velké oblasti vazebné nerovnováhy, v nichž se preferenčně vyskytovaly určité alelické kombinace (Petkov *et al.*, 2005). Řada těchto oblastí obsahovala funkčně podobné geny. Zatímco vznik genových klastrů není příliš objasněn, ukazuje se, že na udržování genových klastrů má selekce významný podíl. Vysoce koexprimované geny v genomu kvasinky mají tendenci být uspořádány v párech. Udržování genů v tomto uspořádání však může být jen částečně vysvětleno na základě jejich vzájemně blízké fyzické vzdálenosti na chromosomu (Hurst *et al.*, 2002). Důkazy pro udržování některých genových klastrů pomocí selekce byly také nalezeny v genomech člověka a myši (Semon & Duret, 2006).

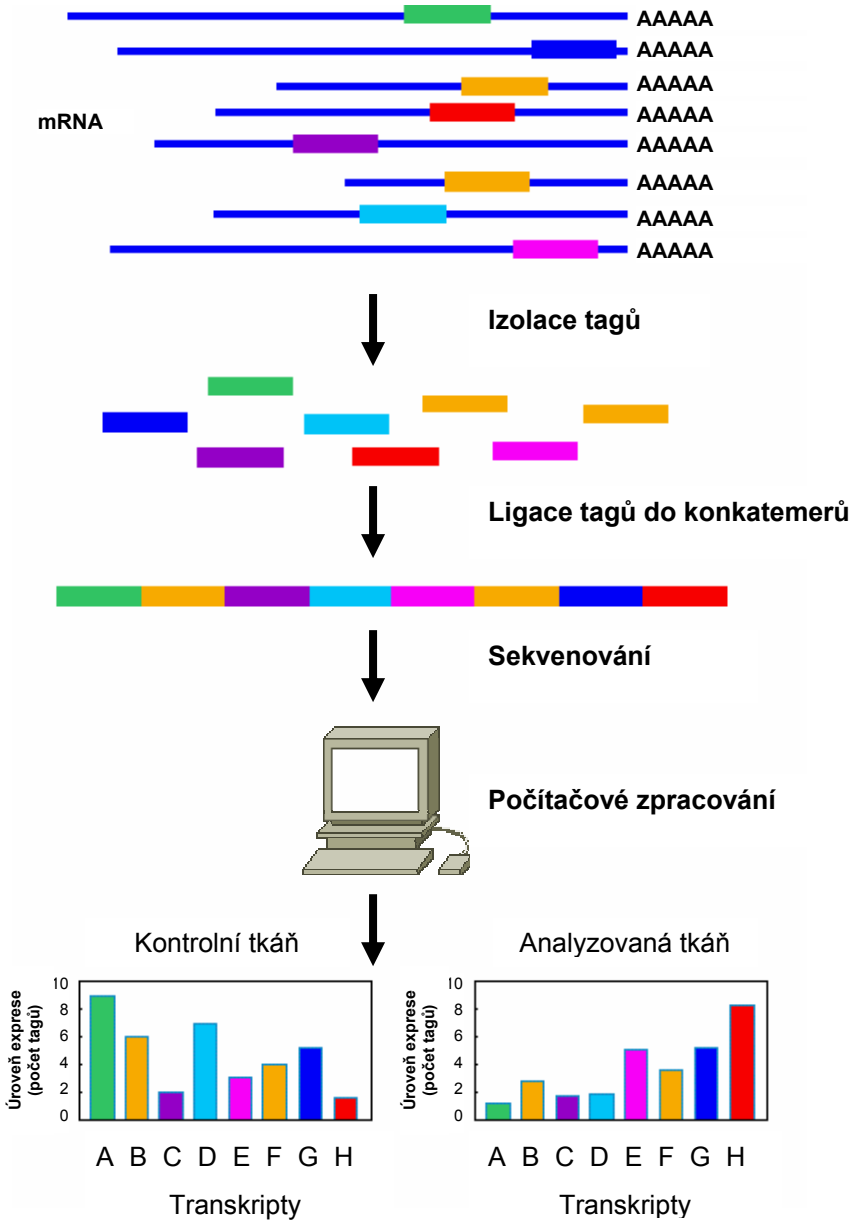
1.2. Globální analýza genové exprese

Předpokladem pro studium uspořádání genů v genomech je nejen znalost sekvence genomů různých organismů, ale také dostupnost genových expresních dat. V posledním desetiletí zaznamenaly velký rozvoj metody, které umožňují analyzovat genovou expresi na úrovni celého transkriptomu. Na rozdíl od tradičných metod, které jsou zaměřeny na analýzu transkripce jednoho nebo několika málo genů, dovolují tyto metody stanovit aktivitu tisíců genů současně. Velké množství dat, které tyto metody produkují, vyžaduje následné počítačové zpracování a statistické vyhodnocení. Vytvořená data bývají obvykle uložena ve veřejně dostupných databázích. Existují dvě skupiny metod pro globální analýzu genové exprese. První skupina je založena na *sekvenování* fragmentů cDNA z jednotlivých transkriptů. Druhá skupina využívá *hybridizace* transkriptů po jejich převedení do cDNA k próbám umístěných na pevném podkladu (DNA čipu). Předmětem této disertační práce bylo využití dat získaných pomocí sekvenačních metod pro analýzu uspořádání genů v genomech a proto jsou sekvenační metody v následujícím textu probrány podrobněji.

Základní informaci o genové expresi v určité tkáni umožňují získat *expresní sekvenční tagy* (EST, expressed sequence tags) (Adams *et al.*, 1991). EST představují krátké sekvence cDNA (300-500 nukleotidů) získané z 5' nebo 3' konců molekul mRNA. Každý EST je sekvenován pouze jednou a jeho sekvence může obsahovat velké množství chyb. Tento nedostatek je však nahrazen velkým množstvím EST, které jsou dostupné ve veřejných databázích. Databáze EST (dbEST - http://www.ncbi.nlm.nih.gov/dbEST/) v současné době obsahuje přes 43 miliónů EST z více než 1300 organismů (Boguski *et al.*, 1993). Soubor EST získaný sekvenováním molekul cDNA z určité tkáně představuje knihovnu EST, která poskytuje zejména kvalitativní informaci o genové expresi (tj. které geny jsou v této tkáni exprimovány). Pro kvantitativní analýzu genové exprese je možné využít pouze ty knihovny EST, u nichž pro sekvenování nebyly cíleně vybírány jen některé cDNA (Bonaldo *et al.*, 1996). Sekvence EST a cDNA jsou rovněž dostupné v databázi UniGene (http://www.ncbi.nlm.nih.gov/UniGene/), která představuje soubor všech transkriptů daného organismu a pro každý transkript definuje reprezentativní sekvenci (Pontius *et al.*, 2003).

Sériová analýza genové exprese (SAGE) umožňuje kvantitativní analýzu velkého množství transkriptů současně (Velculescu *et al.*, 1995). SAGE je založena na dvou základních principech (Obr. 3). Z přesně definované pozice cDNA je izolován krátký úsek

(tag) o délce 10-17 bp, který slouží pro identifikaci transkriptu. Aby bylo možné analyzovat velké množství tagů současně (v sérii), jsou tagy ligovány do konkatemerů a ty pak sekvenovány. Při počítačovém zpracování jsou tagy vystříhány ze sekvencí konkatemerů a počty jednotlivých typů tagů uspořádány do tabulky. Soubor tagů a jejich počtů získaných v jednom experimentu z molekul cDNA z určité tkáně představuje knihovnu SAGE. Knihovny SAGE vytvořené v různých laboratořích je možno vzájemně srovnávat za předpokladu, že pro izolaci tagů byly použity stejné restrikční enzymy. K porovnávání SAGE knihoven a hledání rozdílně zastoupených tagů je možné využít řadu



Obr. 3. Základní princip sériové analýzy genové exprese (http://www.sagenet.org/).

statistických testů (Ruijter *et al.*, 2002). Pro identifikaci tagů je obvykle využívána databáze SAGEmap (Lash *et al.*, 2000), která přiřazuje jednotlivé typy tagů k transkriptům definovaným v databázi UniGene. Data získaná pomocí SAGE bývají dostupná ve veřejných internetových databázích (např. GEO, http://www.ncbi.nlm.nih.gov/geo/) (Edgar *et al.*, 2002). K dispozici jsou také specializované databáze, které poskytují rovněž nástroje pro analýzu SAGE dat (SAGE Genie, http://cgap.nci.nih.gov/SAGE/, Mouse Atlas of Gene Expression, http://www.mouseatlas.org/) (Boon *et al.*, 2002; Khattra *et al.*, 2006). K využívání metody SAGE přispěla také volná dostupnost podrobného protokolu pro akademický výzkum (http://www.sagenet.org/), což také umožnilo vyvinutí vylepšených a specializovaných variant metody, např. MicroSAGE, LongSAGE (Saha *et al.*, 2002), SuperSAGE (Matsumura *et al.*, 2003), miRAGE (Cummins *et al.*, 2006).

Masívní paralelní sekvenování signatur (MPSS) je podobně jako SAGE založeno na sekvenování krátkých signatur (16 bp) z určité pozice v cDNA. Sekvenování velkého množství signatur probíhá současně na proprietárních kuličkách uspořádaných v jedné vrstvě. V každém sekvenačním kroku je snímán obraz, který zachycuje fluorescenční signál všech kuliček. Sada obrazů je poté převedena do sekvence nukleotidů a je spočítáno zastoupení jednotlivých typů signatur. MPSS byla vyvinuta a patentována firmou Lynx Therapeutics, Inc. (http://www.lynxgen.com/) (Brenner *et al.*, 2000), která také sama provádí konstrukci knihoven MPSS na zakázku. Tato metoda byla využita například v projektu zaměřeném na analýzu myšího transkriptomu (Mouse Transcriptome Project, http://www.ncbi.nlm.nih.gov/genome/guide/mouse/MouseTranscriptome.html).

V současné době dominuje pro globální analýzu genové exprese použití hybridizačních metod (DNA čipů). Jsou dostupné DNA čipy od řady výrobců (Affymetrix, Agilent, Codelink, Illumina), které jsou zpracovávány ve specializovaných servisních laboratořích. Na rozdíl od DNA čipů představují sekvenační metody otevřený systém, neboť nejsou závislé na zhotovení prób pro jednotlivé geny a umožňují identifikovat také nové transkripty. V budoucnu by měly být k dispozici další dvě metody pro masívní sekvenování: 454 (Roche, http://www.454.com/) (Margulies *et al.*, 2005) a Solexa (Illumina, http://www.solexa.com/) (Bennett *et al.*, 2005). Dostupnost těchto rychlých a levných sekvenačních technik umožní také jejich využití pro globální analýzu genové exprese.

2. CÍLE PRÁCE

2.1. Analýza genů exprimovaných v myším varleti a jejich uspořádání v genomu

Prvním cílem této disertační práce bylo charakterizovat transkriptom varlete myši domácí (*Mus musculus domesticus*) pomocí metody SAGE (sériová analýza genové exprese), s využitím získaných dat provést analýzu genomového uspořádání genů exprimovaných ve varleti, vyhodnotit genový obsah chromosomu X a zjistit, zda existuje poziční klastrování genů na chromosomech. Tyto poznatky bylo dále třeba srovnat s existujícími studiemi o nenáhodném uspořádání genů v genomech jiných organismů. Vedlejším cílem bylo vytvořit databázi veřejně dostupných SAGE knihoven připravených z myších tkání a buněčných linií, která poslouží pro výše uvedené srovnávací analýzy a bude veřejně přístupná na Internetu.

2.2. Analýza genového obsahu chromosomu Z kura domácího

Z předchozích studií je známo, že chromosom X má nenáhodný obsah genů, které jsou preferenčně či výlučně exprimovány v jednom pohlaví. Cílem této disertační práce bylo provést analýzu genového obsahu pohlavního chromosomu Z, který dosud nebyl uspokojivě prostudován. Porovnání genového obsahu párového pohlavního chromosomu u organismů s heterogametickými samci a heterogametickými samicemi by mohlo přispět k objasnění mechanismů, které vedou k nenáhodnému obsahu genů na pohlavních chromosomech. Pro tuto analýzu byl jako modelový organismus využit kur domácí (*Gallus gallus*), jehož genomová sekvence je známá a zároveň jsou k dispozici veřejně dostupná genová expresní data.

3. MATERIÁL A METODY

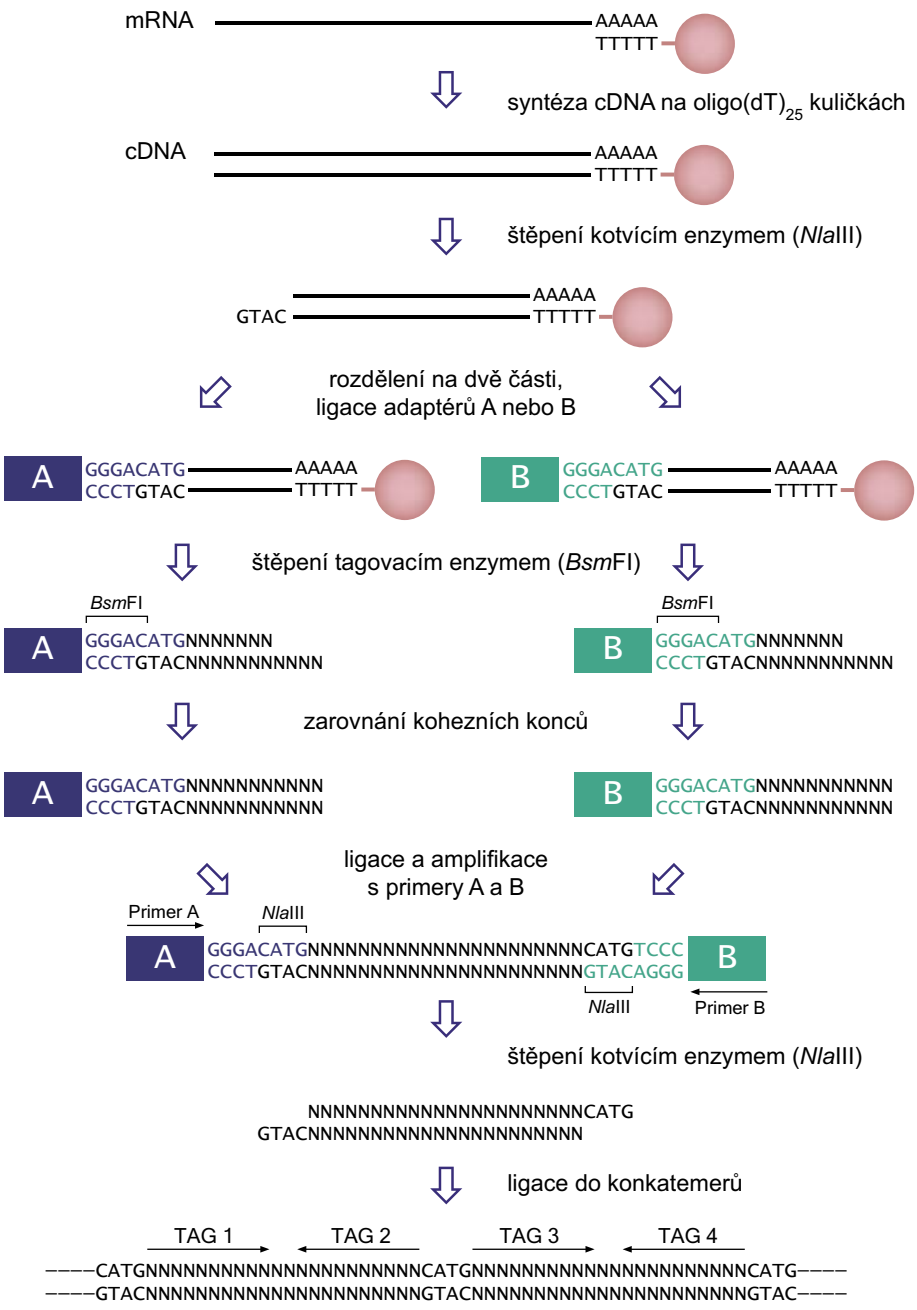
3.1. Analýza genů exprimovaných v myším varleti a jejich uspořádání v genomu

3.1.1. Myši, odběr tkání a izolace RNA

Varlata byla získána z 11-týdenních myších samců z inbredního kmene C57BL/6J (dále jen B6), kteří byli chováni v prostředí bez specifických patogenních organismů v Ústavu molekulární genetiky (ÚMG, Akademie věd České republiky). S myšmi bylo zacházeno v souladu se zákonem č.264/1992 Sb., na ochranu zvířat proti týrání. Myši byly zabity zlomením vazu, varlata rychle vyjmuta z těla, zbavena tuniky a homogenizována. Celková RNA byla vyizolována pomocí TRIzolu (Invitrogen) podle návodu od výrobce.

3.1.2. Konstrukce SAGE knihoven, sekvenování

SAGE knihovny byly připraveny podle protokolu pro MicroSAGE verze 1.0e, který je dostupný na webové stránce http://www.sagenet.org/. Do vzorku celkové RNA byly přidány magnetické kuličky s kovalentně navázaným oligo$(dT)_{25}$ (Dynal) a vyizolována poly(A) RNA za pomoci magnetu (Obr. 4). Poly(A) RNA navázaná na magnetických kuličkách byla přeměněna na dvouřetězcovou cDNA pomocí SuperScript Choice System kitu (Invitrogen). Poté byla cDNA štěpena kotvícím enzymem *Nla*III (New England Biolabs), takže na magnetických kuličkách zůstal navázán úsek cDNA od posledního restrikčního místa pro kotvící enzym do 3' konce. Dále byl vzorek rozdělen na dvě stejné části a pomocí T4 ligázy (Invitrogen) připojeny adaptéry A nebo B, které obsahovaly rozpoznávací místo pro tagovací enzym. V následujícím kroku byly tagovacím enzymem *Bsm*FI (New England Biolabs) odštěpeny adaptéry spolu s krátkým úsekem cDNA („tagem“) a odstraněn zbytek cDNA navázaný na magnetické kuličky. Po zarovnání kohezních konců pomocí Klenowovy polymerázy (Pharmacia) byly obě části vzorku smíchány a působením T4 ligázy vytvořeny „ditagy“, tj. dva krátké úseky cDNA spojené



Obr. 4. Schéma konstrukce SAGE knihovny.

k sobě v protilehlé orientaci s navázanými adaptéry. V dalším kroku byly ditagy amplifikovány PCR s primery navrženými podle sekvence adaptérů a produkt PCR byl rozdělen na polyakrylamidové gelové elektroforéze. Fragment odpovídající ditagům (~ 102 bp) byl vyizolován z gelu a po štěpení kotvícím enzymem *Nla*III byly z ditagů odděleny adaptéry. Směs byla rozdělena na další polyakrylamidové gelové elektroforéze a z gelu byl vyizolován fragment o velikosti ~ 26 bp představující ditagy bez adaptérů. Následně byly ditagy působením T4 ligázy pospojovány do konkatemerů. Konkatemery o velikosti 750-1000 bp byly zaklonovány do plazmidu pZero (Invitrogen), který byl transfekován elektroporací do buněk ElectroMAX DH10Bs (Invitrogen). Klony obsahující plazmid pZero s konkatemerem byly selektovány na plotnách obsahujících antibiotikum zeocin. Sekvenování konkatemerů bylo provedeno pomocí univerzálních primerů M13 na přístroji CEQ 2000 DNA Analysis System (Beckmann Coulter).

3.1.3. Extrakce tagů ze sekvencí konkatemerů

Tagy byly vyextrahovány pomocí Perlového skriptu (sagetag_parser.pl), který byl pro tento účel napsán (Divina, nepublikováno). Skript nejprve v sekvenci konkatemeru lokalizoval restrikční místa pro kotvící enzym *Nla*III (CATG) a vyextrahoval všechny ditagy o délce 22-26 bp. Ditagy, které byly v sekvencích konkatemerů nalezeny opakovaně, byly z dalšího zpracování vyloučeny. (Každý ditag by totiž měl být zastoupen právě jedenkrát, aby byla eliminována případná preferenční amplifikace některých ditagů.) V dalším kroku byly z ditagů vyextrahovány jednotlivé tagy o délce 10 bp a odstraněny tagy odvozené ze sekvencí adaptérů. Výsledkem byl seznam tagů a jejich počtů označovaný jako SAGE knihovna.

3.1.4. Identifikace tagů ke genům

Pro identifikaci tagů ke genům byla použita databáze SAGEmap (Lash *et al.*, 2000) dostupná ke stažení z FTP serveru NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/Mm/NlaIII/). Pomocí databáze SAGEmap byly tagy přiřazeny k transkriptům definovaným v databázi UniGene (Pontius *et al.*, 2003), která představuje soubor všech známých mRNA a EST sekvencí daného organismu uspořádaných do neredundantních transkripčních klastrů. Transkripční klastry z databáze UniGene byly poté přiřazeny ke genům z databáze LocusLink (Pruitt & Maglott, 2001).

3.1.5. Párové porovnávání SAGE knihoven

Párovým porovnáváním SAGE knihoven byly určeny tagy, které mají signifikantně odlišné počty v jedné a druhé SAGE knihovně. Pro každý porovnávaný tag byla testována nulová hypotéza H0, že se jeho zastoupení v obou porovnávaných SAGE knihovnách neliší. K porovnání počtů tagů byly použity simulace Monte Carlo. Vstupním souborem byla tabulka o třech sloupcích, která obsahovala seznam porovnávaných tagů a jejich pozorované počty v první a druhé SAGE knihovně. S použitím popsaného algoritmu (Patefield, 1981) bylo vygenerováno 100 000 tabulek, které byly náhodně vyplněny počty tagů tak, aby byl zachován stejný součet tagů v řádcích a sloupcích jako ve vstupním souboru. Pro každý tag byl stanoven počet náhodných tabulek v nichž byl rozdíl v počtu tagů v obou SAGE knihovnách stejný nebo větší jako ve vstupním souboru. Proporce těchto tabulek z celkového počtu náhodných tabulek udávala hodnotu pravděpodobnosti (označované „p-chance“) chyby α při zamítnutí H0.

Pro každý porovnávaný tag by stanoven koeficient změny exprese (f_i), který vyjadřuje kolikanásobně se změnil počet tagů v porovnávaných SAGE knihovnách, podle vzorce

$$f_{i,12} = \frac{\frac{n_{i,1}}{N_1}}{\frac{n_{i,2}}{N_2}}, \qquad (1)$$

kde ($f_{i,12}$) je koeficient změny exprese pro tag i, ($n_{i,1}$), ($n_{i,2}$) jsou pozorované počty tagu i v SAGE knihovnách 1 a 2, a (N_1), (N_2) jsou celkové počty tagů v SAGE knihovnách 1 a 2. Pro hodnoty koeficientů $f_i < 1$ byla uváděna jejich záporná reciproká hodnota ($-1/f_i$). V případě, že byl počet tagů v jedné z knihoven nulový ($n_{i,1} = 0$ nebo $n_{i,2} = 0$), byl uvažován počet tagů 1.

3.1.6. Výběr myších SAGE knihoven pro analýzu uspořádání genů v genomu

Pro analýzu uspořádání genů v genomu byly vybrány SAGE knihovny vytvořené z myších tkání, které byly veřejně dostupné ke dni 1. července 2004 (Tabulka 1). Celkem 27 SAGE knihoven vytvořených ze somatických tkání bylo uspořádáno do 7 skupin podle typu tkáně

Tabulka 1. Seznam SAGE knihoven použitých k analýze uspořádání genů v genomu.

Většina SAGE knihoven byla získána z databáze Gene Expression Omnibus (odkazy označené GSM, http://www.ncbi.nlm.nih.gov/geo/) (Edgar *et al.*, 2002), zbývající byly staženy z uvedených internetových stránek. Podrobnější informace o SAGE knihovnách je možno nalézt v databázi Mouse SAGE Site (http://mouse.biomed.cas.cz/sage/). Vysvětlivky: E = embryonální den, P = postnatální den.

Název SAGE knihovny	Počet tagů	Popis - tkáň, fyziologický stav (věk, pohlaví)	Odkaz
Brain_Male	51270	celý mozek, normální (P30, samec)	1)
Brain_Female	50594	celý mozek, normální (P30, samice)	1)
Brain_Ts65Dn	50414	celý mozek, Ts65Dn translokace (P30, samec)	1)
Neocortex_E15	20107	neokortex, normální (E15.5)	2)
Neocortex_P1	20626	neokortex, normální (P1.1)	2)
Hypothalamus	55184	hypotalamus, normální	3)
Cerebellum_P23	19429	cerebelum, normální (P23, samec+samice)	GSM2451
Hippocampus_SAL	31190	hipokampus, normální (15-16 týdnů, samec)	GSM12540
Hippocampus_LAL	30931	hipokampus, normální (15-16 týdnů, samec)	GSM12541
Retina	53282	sítnice, normální (dospělá)	3)
Retina_Crx+/+	51954	sítnice, normální (P10.5)	3)
Retina_Crx-/-	53709	sítnice, Crx mutant (P10.5)	3)
Outer_Nuclear_Layer	52053	vnější nukleární vrstva sítnice, normální	3)
Cornea	62093	rohovka, normální (6 týdnů, samec)	GSM13195
Cornea_P9	63308	rohovka, normální (P9)	GSM13196
Heart	84275	srdce, normální (3 měsíce, samice)	GSM1681
Liver	18742	játra, normální (10 týdnů, samice)	GSM12777
Liver_E3L	18376	játra, ApoE3L transgen (10 týdnů, samice)	GSM12757
Kidney_Saline	11197	ledviny, solná infuze (dospělé, samec)	GSM9194
Kidney_AngII	11337	ledviny, infuze Angiotenzinu II (dospělé, samec)	GSM9195
Kidney_1	19793	ledviny, normální (9-11 týdnů, samec)	GSM24251
Kidney_1_UN_Acute	23343	ledviny, akutní poškození (9-11 týdnů, samec)	GSM24255
Kidney_2	12571	ledviny, normální (25-28 týdnů, samec)	GSM24256
Kidney_2_UN_Chronic	9569	ledviny, chronické poškození (25-28 týdnů, samec)	GSM24257
Forelimb	68302	přední končetina, normální (E11.5)	GSM55
Hindlimb	68348	zadní končetina, normální (E11.5)	GSM56
Adipose_Retroperitoneal	44974	tuková tkáň retroperitoneální (12-14 týdnů, samec)	GSM17227
Total_Testis_1	24975	varle, normální (9 týdnů, samec)	GSM34767
Total_Testis_2	51879	varle, normální (9 týdnů, samec)	GSM34768
Testis_Somatic_Cells_Adult	81478	varle, bez germinálních buněk (dospělé, samec)	GSM5435

1) http://medgen.unige.ch/research/projects.html

2) http://www.hfi.unimelb.edu.au/research/developmental_biology/sage/

3) http://genetics.med.harvard.edu/~cepko/SAGE/

a tagy z jednotlivých knihoven byly sloučeny (Tabulka 2). Další dvě skupiny tvořily 2 SAGE knihovny zkonstruované z celého myšího varlete v této práci (GSM34767, GSM34768) a SAGE knihovna připravená ze somatických buněk varlete (GSM5435) (O'Shaughnessy *et al.*, 2003).

Tabulka 2. Skupiny SAGE knihoven podle typu tkáně.

Typ tkáně	Počet SAGE knihoven	Celkový počet tagů
Mozek	9	329745
Oko	6	336399
Srdce	1	84275
Játra	2	37118
Ledviny	6	87810
Končetiny	2	136650
Tuková tkáň	1	44974
Varle, celé	2	76854
Varle, somatické buňky	1	81478

3.1.7. Výběr tkáňově specificky exprimovaných genů

Tkáňově specificky exprimované geny byly vybrány podle počtu jejich tagů v jednotlivých skupinách SAGE knihoven (Tabulka 2). Za geny s tkáňově specifickou expresí byly považovány ty geny, které měly tagy přítomné pouze v jednom typu tkáně a v této tkáni byly zjištěny alespoň dva jejich tagy.

3.1.8. Analýza zastoupení tkáňově specifických genů na chromosomu X

Při této analýze byla testována nulová hypotéza H0, že zastoupení tkáňově specifických X-vázaných genů exprimovaných ve varleti se neliší od zastoupení tkáňově specifických X-vázaných genů exprimovaných v ostatních somatických tkáních. Paralelně byly porovnány tkáňově specifické geny zjištěné v celém varleti a v somatických buňkách varlete s tkánově specifikými geny, které byly zjištěny ve skupině obsahující sedm somatických tkání. Porovnání počtu X-vázaných tkáňově specificky exprimovaných genů mezi dvěma skupinami tkání bylo provedeno permutačním testem. Vstupním souborem byla tabulka o třech sloupcích, která obsahovala seznam tkáňově specificky exprimovaných genů a každému z nich přiřazené dvě hodnoty: chromosom, na který byl gen lokalizován podle databáze LocusLink (Pruitt & Maglott, 2001) (autosom nebo chromosom X) a tkáň, ve které byl gen specificky exprimován (varle nebo ostatní tkáň).

Počet X-vázaných genů exprimovaných ve varleti pozorovaných v tomto vstupním souboru byl zaznamenán. Bylo vygenerováno 100 000 permutací vstupního souboru tak, že pokaždé byly promíchány chromosomy a zůstal zachován konstantní počet genů na autosomech a chromosomu X. Byl stanoven počet permutací, u nichž byl počet X-vázaných genů exprimovaných ve varleti a) větší nebo roven, b) menší nebo roven pozorovanému počtu X-vázaných genů exprimovaných ve varleti ve vstupním souboru. Menší z těchto dvou hodnot dělená celkovým počtem permutací byla vynásobena dvěma a výsledná hodnota považována za dvoustrannou hodnotu pravděpodobnosti („p-value“) chyby α při zamítnutí H0. Alternativně byla H0 testována Fisherovým exaktním testem s použitím čtyřpolní tabulky.

3.1.9. Výběr genů preferenčně exprimovaných ve varleti

Geny preferenčně exprimované ve varleti byly vybrány na základě hodnoty PEM (preferential expression measure) (Huminiecki *et al.*, 2003), která udává míru preferenční exprese určitého genu v dané tkáni a byla vypočítána podle vzorce

$$PEM_{g,t} = \log_{10} \frac{o_{g,t}}{e_{g,t}}, \tag{2}$$

kde ($PEM_{g,t}$) je koeficient preferenční exprese genu g v tkáni t, ($o_{g,t}$) je pozorovaný počet tagů genu g v tkáni t a ($e_{g,t}$) je očekávaný počet tagů genu g v tkáni t za předpokladu jeho stejnoměrné exprese ve všech tkáních stanovený podle vzorce

$$e_{g,t} = N \frac{G}{T}, \tag{3}$$

kde (N) je celkový počet tagů v dané tkáni, (G) je celkový počet tagů genu g ve všech tkáních a (T) je celkový počet tagů ve všech tkáních. PEM nabývá pozitivních hodnot pro geny, které jsou v dané tkáni preferenčně exprimované a negativních hodnot pro geny v dané tkáni preferenčně reprimované. Čím větší (příp. menší) je hodnota PEM, tím více je gen v dané tkáni preferenčně exprimován (příp. reprimován). Geny, které jsou preferenčně exprimované v dané tkáni je možné vybrat porovnáním jejich hodnoty PEM s maximální hodnotou $PEM_{(max)}$ dosaženou v této tkáni. Pro výběr genů preferenčně exprimovaných v celém varleti a somatických buňkách varlete bylo zvoleno kritérum PEM >= ½ $PEM_{(max)}$.

Maximální hodnoty PEM byly pro celé varle $PEM_{(max)} = 1.169$ a pro somatické buňky varlete $PEM_{(max)} = 1.145$.

3.1.10. Analýza pozičního uspořádání genů preferenčně exprimovaných ve varleti

3.1.10.1. Identifikace pozičních klastrů na chromosomech

Při této analýze byly identifikovány dva typy klastrů („volné“ a „těsné“), které obsahovaly geny s preferenční expresí v celém varleti a v somatických buňkách varlete. Vstupním souborem byla tabulka obsahující seznam 16 858 genů v genomu, jejich pozice na chromosomech a stav exprese vyjádřený hodnotami: exprimován, preferenčně exprimován a neexprimován (popř. informace o expresi není známa). Každý chromosom byl procházen pomocí okna o velikosti 3 genů, které bylo posouváno po 1 genu a okna obsahující 3 geny s preferenční expresí byla považována za klastr (těsné klastry). Podobně byl každý z chromosomů procházen pomocí okna o velikosti 6 genů, které bylo posouváno po 1 genu a okna obsahující alespoň 3 geny s preferenční expresí byla považována za klastr (volné klastry). Překrývající se klastry byly sjednoceny do jednoho klastru zahrnující všechny preferenčně exprimované geny (zvlášť pro těsné a volné klastry). Byl stanoven počet preferenčně exprimovaných genů v klastrech nalezených ve vstupním souboru (zvlášť ve volných a těsných klastrech, a zvlášť v celém varleti a v somatických buňkách varlete).

3.1.10.2. Statistické vyhodnocení obsahu preferenčně exprimovaných genů v klastrech

Byla testována nulová hypotéza H0, že celkový počet preferenčně exprimovaných genů zahrnutých v klastrech ve vstupním souboru se neliší od počtu preferenčně exprimovaných genů identifikovaných v klastrech v náhodném genomu z něhož byly předem odstraněny tandemově duplikované geny (viz 3.1.10.3). Hypotéza H0 byla testována permutačním testem, při němž bylo vygenerováno 100 000 permutací vstupního souboru tak, že pokaždé byly promíchány hodnoty expresních stavů genů, zatímco pozice genů na chromosomech zůstaly zachovány. Při každé permutaci byl stanovem počet preferenčně exprimovaných genů v těsných a volných klastrech pomocí procházení posuvným oknem (viz 3.1.10.1). Byl stanoven počet permutací, u nichž byl zjištěn větší nebo stejný počet preferenčně

exprimovaných genů v klastrech jako ve vstupním souboru. Proporce těchto permutací z celkového počtu permutací byla považována za jednostrannou hodnotu pravděpodobnosti („p-value“) chyby α při zamítnutí H0.

3.1.10.3. Odstranění tandemově duplikovaných genů z genomu

Předpokladem pro výše uvedenou analýzu bylo odstranění tandemově duplikovaných genů z genomu proto, aby mohl být vyloučen vliv genových duplikací na množství preferenčně exprimovaných genů v klastrech. Tandemově duplikované geny byly identifikovány a odstraněny následujícím způsobem. Nejprve byly z databáze LocusLink (Pruitt & Maglott, 2001) vybrány všechny známé myší geny a byla určena jejich pozice na chromosomech. Pro každý gen byla z databáze RefSeq (Pruitt & Maglott, 2001) získána sekvence proteinu, který je tímto genem kódován a byl proveden proteinový BLAST (standardní nastavení) proti sekvencím všech známých myších proteinů z databáze RefSeq. Hity, které dosahovaly hodnot e-value < $1e^{-10}$, 30 % identity sekvencí a alignment pokrýval alespoň 50 % délky sekvencí byly vybrány a nalezené proteiny identifikovány zpět ke genům z databáze LocusLink. Pokud se takto identifikované geny vyskytovaly v okolí původního genu (uvažováno 10 genů v obou orientacích) byly oba geny (původní a nalezený) považovány za tandemově duplikovaný pár genů a vyloučeny z další analýzy. Z původního počtu 19 684 známých genů z databáze LocusLink zbylo po odstranění tandemově duplikovaných párů celkem 16 858 genů.

3.1.11. Použité verze databází

Pro všechny uvedené analýzy byly použity tyto verze databází: myší UniGene (26. března 2004, sestavení #136), SAGEmap (3. dubna 2004, verze odvozená z UniGene #136), LocusLink (3. dubna 2004), sestavení myšího genomu (NCBI mouse genome assembly, build 32, listopad 2003), RefSeq (3. duben 2004) a Gene Ontology (červenec 2004).

3.2. Analýza genového obsahu chromosomu Z kura domácího

3.2.1. Výběr EST knihoven

Pro analýzu byly použity EST knihovny kura domácího *(Gallus Gallus)*, které byly veřejně dostupné v databázi UniGene (14. října 2004, sestavení #24) (Pontius *et al.*, 2003). Celkem bylo vybráno 68 EST knihoven, které obsahovaly alespoň 500 EST a byly připraveny ze zdravých tkání dospělých jedinců. Vybrané EST knihovny byly roztříděny do 14 skupin podle typu tkáně (Tabulka 3).

Tabulka 3. Seznam EST knihoven použitých k analýze genového obsahu na chromosomu Z a jejich rozdělení do skupin podle typu tkáně.

Tkáň	Počet EST knihoven	Celkový počet EST	Identifikátory EST knihoven v databázi UniGene http://www.ncbi.nlm.nih.gov/UniGene/
Mozek	17	58659	8707, 11179, 11180, 11183, 11192, 11193, 11194, 11195, 11198, 11199, 11218, 11219, 11227, 11228, 15560, 15561, 16171
Lymfocyty	2	40311	3749, 10511
Končetiny	7	28416	11176, 11177, 11178, 11185, 11216, 11223, 13932
Chondrocyty	5	20950	11202, 11203, 11208, 11210, 11230
Vaječník	5	18370	11205, 11206, 11207, 11209, 11232
Srdce	6	14866	11186, 11187, 11196, 11197, 11215, 11229
Ledviny a nadledvinky	3	14211	11181, 11200, 11220
Tenké střevo	3	13551	11182, 11184, 11221
Játra	5	13409	8831, 8874, 11173, 11174, 11222
Sval	5	10607	11204, 11231, 16033, 16034, 16035
Varle	3	10380	15562, 15563, 16173
Pankreas	3	6903	11211, 11212, 11234
Tuková tkáň	3	6171	9526, 10201, 11233
Slezina	1	3060	16172

3.2.2. Lokalizace genů na chromosomy

Pro účely následujících analýz byly za geny považovány trankripční klastry sekvencí mRNA a EST definováné v databázi UniGene (Pontius *et al.*, 2003). Lokalizace genů v genomu byla určena pomocí programu BLAT (Kent, 2002), kterým byly provedeny alignmenty reprezentativních sekvencí každého transkripčního klastru na genomu (chicken draft genome assembly, galGal2, únor 2004, UCSC Genome Browser) (Karolchik *et al.*, 2003; Hillier *et al.*, 2004). Jako parametry pro vyhledání alignmentu byly použity:

minimálně 95% identita sekvencí, alespoň 20% pokrytí délky reprezentativní sekvence, minimálně 96,5% alignment ratio a skóre v rámci 0,2% nejlepších v genomu. Transkripční klastry, které byly lokalizovány na více než 1 místo v genomu byly z analýzy vyloučeny. Stejně tak byly vyloučeny transkripční klastry mapující na chromosom W. S použitím těchto kritérií bylo možné zjistit unikátní polohu v genomu pro 79 % (16 795 z 21 447) transkripčních klastrů reprezentujících jednotlivé geny.

3.2.3. Výběr tkáňově specificky eprimovaných genů

Za tkáňově specificky exprimované geny byly považovány ty, jejichž mRNA a EST sekvence byly zjištěny pouze v jednom typu tkáně (Tabulka 3).

3.2.4. Analýza zastoupení tkáňově specifických genů na chromosomu Z

Byla testována nulová hypotéza H0, že zastoupení tkáňově specifických genů na chromosomu Z v germinálních tkáních (varlatech nebo vaječnících) se neliší od zastoupení tkáňově specifických genů na chromosomu Z ve skupině 12 somatických tkání. Porovnání počtu Z-vázaných tkáňově specificky exprimovaných genů mezi dvěma typy tkání bylo provedeno permutačním testem. Vstupním souborem byla tabulka genů o třech sloupcích, která obsahovala seznam tkáňově specificky exprimovaných genů a každému z nich byly přiřazeny dvě hodnoty: chromosom, na který byl gen lokalizován (autosom nebo chromosom Z) a typ tkáně, v níž byl gen specificky exprimován (germinální nebo somatická). Počet Z-vázaných genů exprimovaných v germinální tkáni zjištěných v tomto vstupním souboru byl zaznamenán. Bylo vygenerováno 100 000 permutací vstupního souboru tak, že pokaždé byly promíchány chromosomy a zůstal zachován konstantní počet genů na autosomech a chromosomu Z. Byl stanoven počet permutací, u nichž byl počet Z-vázaných genů exprimovaných v germinální tkáni a) větší nebo roven, b) menší nebo roven pozorovanému počtu Z-vázaných genů exprimovaných v germinální tkáni ve vstupním souboru. Menší z těchto dvou hodnot dělená celkovým počtem permutací byla vynásobena dvěma a výsledná hodnota považována za dvoustrannou hodnotu pravděpodobnosti („p-value“) chyby α při zamítnutí H0.

3.2.5. Výběr genů preferenčně exprimovaných v jednom pohlaví v mozku

Geny preferenčně exprimované v jednom pohlaví byly identifikovány v mozkové tkáni, pro kterou existovala expresní data z obou pohlaví. K výběru genů byly použity tři nenormalizované EST knihovny připravené ze samčích (celkem 4230 EST sekvencí) a samičích mozků (celkem 8399 EST sekvencí), které byly dostupné v databázi UniGene (identifikátory knihoven: 16171, 15560 a 15561 (http://www.ncbi.nlm.nih.gov/UniGene/). Geny preferenčně exprimované v samčí nebo samičí tkáni byly vytříděny pomocí R statistiky, která se používá pro kvantitativní porovnávání četnosti sekvencí v cDNA a EST knihovnách (Stekel *et al.*, 2000). Pro každý gen exprimovaný v mozku byla spočítána R statistika podle vzorce

$$R_j = \sum_{i=1}^{m} x_{i,j} \log\left(\frac{x_{i,j}}{N_i f_j}\right), \tag{4}$$

kde (R_j) je R statistika pro gen j, (m) je počet tkání či skupin EST knihoven, ($x_{i,j}$) je počet sekvencí genu j ve tkáni i, (N_i) je celkový počet sekvencí ve tkáni i a (f_j) je četnost sekvencí genu j ve všech tkáních stanovená podle vzorce

$$f_j = \frac{\sum_{i=1}^{m} x_{i,j}}{\sum_{i=1}^{m} N_i} \tag{5}$$

Za geny přednostně exprimované v samčí či samičí tkáni byly považovány ty geny, pro něž R statistika překročila daný limit.

3.2.6. Analýza zastoupení genů preferenčně exprimovaných v samčím a samičím mozku na chromosomu Z

Byla testována nulová hypotéza H0, že se zastoupení Z-vázaných genů, které byly preferenčně exprimovány v samčím a samičím mozku od sebe neliší. Dále bylo testováno, zda se proporce Z-vázaných genů preferenčně exprimovaných v samčím nebo samičím mozku liší od proporce Z-vázaných genů ve zjištěných ve skupině 11 somatických tkání.

Statistická signifikance byla stanovena Fisherovým exaktním testem s použitím čtyřpolní tabulky.

3.3. Hardware a software

Zpracování dat bylo prováděno na serveru Sun Ultra 10 (*sun4.biomed.cas.cz*) s hardwarovou konfigurací 1x procesor UltraSPARC-IIi 440MHz, 64-bit, 1 GB operační paměti, 300 GB diskového prostoru a operačním systémem Solaris 5.8. Výpočetně náročnější operace (např. permutace) byly rozděleny na menší dávky a počítány paralelně s využitím dalších serverů na platformě PC i686 s operačním systémem Gentoo Linux: *charon.img.cas.cz* (2x procesor Xeon 2,8 GHz, 2 GB operační paměti, 150 GB diskového prostoru), *lethe.img.cas.cz* (2x procesor Xeon 2,0 GHz, 2 GB operační paměti, 250 GB diskového prostoru) a *medusa.img.cas.cz* (2x procesor Xeon 2,2 GHz, 3 GB operační paměti, 70 GB diskového prostoru). Data byla zpracována pomocí skriptů napsaných v programovacím jazyce Perl a mezivýsledky uloženy v databázi mySQL (http://www.mysql.com/). Statistické vyhodnocení bylo provedeno pomocí skriptů napsaných v prostředí R (http://www.r-project.org/). Pro vytváření alignmentů byly použity programy NCBI BLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/) a BLAT (Jim Kent's Web Page, http://www.soe.ucsc.edu/~kent/).

4. VÝSLEDKY

4.1. Analýza genů exprimovaných v myším varleti a jejich uspořádání v genomu

4.1.1. Charakterizace SAGE knihoven z myších varlat

Pro analýzu genové exprese v myším varleti byly připraveny dvě SAGE knihovny z celých varlat dospělých samců myšího kmene C57BL/6J. Jedna knihovna byla vytvořena z RNA vyizolované z obou varlat jediného samce (Total Testis 1, TT 1). Druhá knihovna byla připravena ze směsi RNA vyizolované z varlat tří samců, kteří pocházeli ze stejného vrhu (Total Testis 2, TT 2). Sekvenováním těchto knihoven bylo získáno celkem 24 975 (TT 1) a 51 879 (TT 2) tagů, které představovaly 10 516 (TT 1) a 18 848 (TT 2) unikátních tagů (Tabulka 4). Tagy s četností > 1 tvořily 69 % (17 244) a 74 % (38 457) z celkového počtu tagů a pouze 26,5 % (2 785) a 29 % (5 426) z počtu unikátních tagů. Podle parametrů, které udávaly vysoký počet tagů v klonech, nízké zastoupení tagů odvozených z adaptérů (< 1 %) a nízký počet duplikovaných ditagů (~ 1 %), bylo možné usoudit na vysokou kvalitu připravených SAGE knihoven. Obě SAGE knihovny také poskytly velmi podobný genový expresní profil varlete (Pearsonův korelační koeficient $R^2 = 0.84$ pro unikátní tagy v obou knihovnách) a potvrdily tak dobrou reprodukovatelnost SAGE. Při srovnání obou knihoven pomocí simulací Monte Carlo vykazovalo 1,3 % tagů (313 z celkového počtu 24 529 unikátních tagů) signifikantně odlišné počty v obou knihovnách (p-chance < 0,05) (Obr. 5, A). Tyto tagy představují biologickou variabilitu v genových expresních profilech a technickou variabilitu vzniklou odchylkami při zpracováním vzorků. Koeficient změny exprese se pro 93,5 % tagů pohyboval v rozmezí od -2,2 do 2,2 a pro 99 % tagů mezi -5 a 5). Pro další analýzy byly tagy z obou knihoven sloučeny do jedné virtuální SAGE knihovny (TT 1+2) s celkovým počtem 76 854 tagů, které představovaly 24 529 unikátních tagů. Primární data obou SAGE knihoven byla uložena v databázi genových expresních dat Gene Expression Omnibus (Edgar *et al.*, 2002), kde jsou dostupná pod identifikátory

GSM34767 (TT 1) a GSM34768 (TT 2). Pro interaktvní prohlížení a analýzu jsou obě knihovny dostupné také v databázi Mouse SAGE Site (http://mouse.biomed.cas.cz/sage/).

Tabulka 4. Parametry vytvořených SAGE knihoven z celého myšího varlete.

SAGE knihovna	Total Testis 1 (TT 1)	Total Testis 2 (TT 2)
Počet sekvenovaných klonů	811	1 510
Celkový počet tagů *	24 975	51 879
Počet unikátních tagů	10 516	18 848
Počet tagů s četností 1	7 731	13 422
Parametry kvality		
Průměrný počet tagů v klonu	30.8	34.4
Počet duplikovaných ditagů	157 (1,2 %)	276 (1,0 %)
Počet tagů odvozených z adaptérů	147 (0,6 %)	223 (0,4 %)

* po odstranění duplikovaných ditagů a tagů odvozených ze sekvencí adaptérů (tj. TCCCTATTAA, TCCCCGTACA a všech možných jednonukeotidových obměn)

Tagy s četností > 1 ve sloučené SAGE knihovně z myšího varlete (TT 1+2) byly přiřazeny ke genům (transkripčním klastrům) definovaným v databázi UniGene třemi způsoby (Tabulka 5). Podle databáze SAGEmap (Lash *et al.*, 2000), která je pro identifikaci tagů používána nejčastěji, bylo možné přiřadit 92,6 % tagů ke genům (54,3 % k jednomu genu a 38,3 % k více genům). S použitím přísnějších kritérií pro spolehlivou identifikaci tagů, které jsou používány v databázi Mouse SAGE Site (viz 4.1.5.1), bylo možné spolehlivě přiřadit 63 % tagů ke genům (47,5 % k jednomu genu a 15,5 % k více genům), zatímco pro 29,6 % tagů nebyla nalezena spolehlivá identifikace. Tento způsob identifikace umožnil více než dvojnásobně snížit počet tagů přiřazených k více genům a eliminovat nespolehlivou identifikaci tagů. Třetí, nejspolehlivější způsob identifikace tagů, který využívá pouze sekvence mRNA dostupné v databázi GenBank (Benson *et al.*, 2003), umožnil přiřadit 51,3 % tagů ke genům (45 % k jednomu genu a 6,3 % k více genům) a pro 41,3 % tagů nebyla nalezena spolehlivá identifikace. Všechny tři způsoby identifikace tagů odhalily malou proporci tagů (7,4 %), které nebylo možné identifikovat a mohou představovat nové transkripty. Seznam 7 481 tagů s četností > 1 ve sloučené SAGE knihovně z celého myšího varlete (TT 1+2) a jejich identifikace ke genům podle databáze Mouse SAGE Site je uveden v on-line příloze k publikaci (http://www.biomedcentral.com/

1471-2164/6/29, Additional file 1). Pro analýzy genomového uspořádání genů exprimovaných ve varleti byl dále v této práci používán nejpřísnější a nejspolehlivější způsob identifikace tagů podle sekvencí mRNA v databázi GenBank.

Tabulka 5. Identifikace tagů v SAGE knihovnách z myšího varlete (TT 1+2) ke genům* při použití různých databází a způsobů pro spolehlivost identifikace.

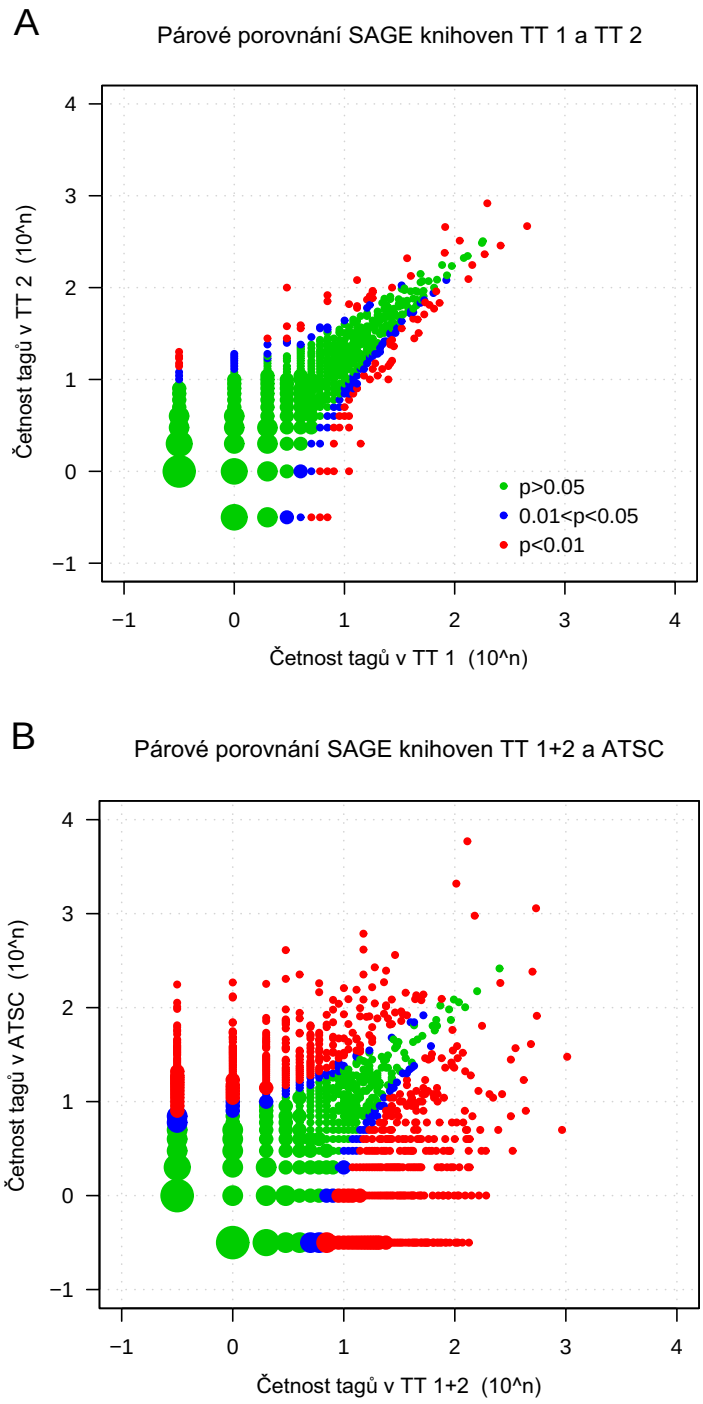
	NCBI SAGEmap		Mouse SAGE Site		GenBank mRNA	
	počet tagů	%	počet tagů	%	počet tagů	%
Identifikace k 1 genu*	4 061	54,3	3 553	47,5	3 367	45,0
Identifikace k více genům*	2 865	38,3	1 157	15,5	472	6,3
Nespolehlivá identifikace	-	-	2 216	29,6	3 087	41,3
Bez identifikace	555	7,4	555	7,4	555	7,4
Počet tagů s četností > 1	7 481	100,0	7 481	100,0	7 481	100,0

* geny představují transkripční klastry definované v databázi UniGene

Genům zastoupeným v transkriptomu myšího varlete byla přiřazena biologická funkce podle databáze Gene Ontology (Harris *et al.*, 2004). Bylo zjištěno, že více než tisíc genů exprimovaných ve varleti má nějakou funkci v metabolismu, zejména v metabolismu proteinů (modifikace proteinů, protein targeting) a nukleových kyselin (sestavování a modifikace chromatinu, replikace DNA, opravy DNA, zpracování RNA, modifikace RNA). Podle očekávání byly mezi vysoce exprimovanými nalezeny geny, které se účastní spermatogeneze (např. protamine 1 a 2, transition protein 1 a 2), ovlivňují uspořádání chromosomů, plní funkci v buněčném cyklu a při buněčné diferenciaci. Velmi exprimované byly také geny, které se podílejí na buněčném transportu (např. diazepam binding inhibitor-like 5, proteasome 26S subunit, ribosomal protein L23), buněčné signalizaci (např. calmodulin 1 a 2, sperm autoantigenic protein 17, A kinase PRKA anchor protein 3, PDZ domain containing 1, WD repeat domain 12), uspořádání cytoskeletu (např. t-complex testis expressed 1, t-complex-associated testis expressed 3, tubulin alpha 7/alpha 3, tubulin alpha 6, thymosin beta 10) a apoptóze (např. Bcl2-associated athanogene 1, Bcl2-like 14, programmed cell death 5, tumor protein translationally-controlled 1). Z mitochondriálního genomu byly vysoce exprimovány geny ATP synthase 6, cytochrome c oxidase I a III.

4.1.2. Srovnání SAGE knihoven z celého myšího varlete a somatických buněk varlete

Myší varle je složeno ze dvou hlavních typů buněk – germinálních a somatických, které se liší svou funkcí i původem. Germinální buňky diferencují ze spermatogonií přes spermatocyty a spermatidy do zralých spermatozoí. Somatické buňky (Sertoliho, Leydigovy, myoidní) naopak vykonávají všechny podpůrné funkce potřebné pro úspěšný průběh spermatogeneze. Varlata dospělých myší obsahují v semenotvorných kanálcích asi 88 % germinálních buněk a 12 % somatických buněk (Sutcliffe *et al.*, 1991). Je tedy možné předpokládat, že většina transkriptů nalezená v SAGE knihovně připravené z celého myšího varlete pocházela z germinálních buněk. Ve veřejné databázi genových expresních dat GEO byla nalezena SAGE knihovna připravená ze somatických buněk varlete dospělých myší (GSM5435, dále označována jako ATSC), která mohla být s výhodou použita ke srovnání s právě vytvořenou SAGE knihovnou z celého varlete a k určení genů přednostně exprimovaných v germinálních a somatických buňkách varlete. Tato knihovna byla vytvořena z varlat dospělých myší, ve kterých byly zlikvidovány germinální buňky varlat působením busulfanu (O'Shaughnessy *et al.*, 2003). Výhodou pro porovnání těchto dvou SAGE knihoven bylo také to, že obě knihovny měly podobné parametry – celkový počet tagů (76 854 a 81 478), počet unikátních tagů (24 529 a 22 809) a proporci tagů s četností > 1 (Tabulka 6). Srovnání obou SAGE knihoven pomocí simulací Monte Carlo potvrdilo značné rozdíly v genové expresi mezi celým myším varletem a somatickými buňkami varlete (Obr. 5, B). Z celkového počtu 42 239 unikátních tagů v knihovnách TT 1+2 a ATSC bylo detekováno 3 258 tagů (7,7 %) se signifikantně odlišnou četností v obou knihovnách (p-chance < 0.05). Koeficient změny exprese se pro 83 % tagů pohyboval v rozmezí od -2.2 do 2.2 a pro 92.5 % tagů mezi -5 a 5. Bylo nalezeno 563 tagů, které měly více než desetinásobně zvýšený počet v somatických buňkách varlete (koeficient změny exprese > 10) a 672 tagů, které dosahovaly více než 10x zvýšeného počtu v celém varleti (koeficient změny exprese < -10). Seznam tagů se signifikantně odlišným počtem v SAGE knihovnách z celého varlete a somatických buněk varlete je uveden v on-line příloze k publikaci (http://www.biomedcentral.com/1471-2164/6/29, Additional file 2).



Obr. 5. Grafy znázorňující párové porovnání SAGE knihoven.

Párové porovnání dvou SAGE knihoven z celého varlete (A) a SAGE knihoven z celého varlete a somatických buněk varlete (B). Tagy se signifikantně odlišnými počty jsou znázorněny modře (0,01 < p-chance < 0,05) a červeně (p-chance < 0,01) (signifikance byla stanovena pomocí simulací Monte Carlo). Tagy přítomné pouze v jedné z porovnávaných knihoven jsou znázorněny v bodech -0,5 na obou koordinátách. Velikost bodu je úměrná počtu tagů, který daný bod zastupuje. SAGE knihovny: TT 1 a TT 2 = jednotlivé knihovny z celého varlete; TT 1+2 = sloučená knihovna z celého varlete; ATSC = knihovna ze somatických buněk varlete dospělých myší.

Tabulka 6. Parametry SAGE knihoven připravených z celého varlete a somatických buněk varlete dospělých myší.

SAGE knihovna	Total Testis (TT 1+2)	Adult Testis Somatic Cells (ATSC)
Celkový počet tagů	76 854	81 478
Počet unikátních tagů	24 529	22 809
Počet tagů s četností > 1	7 481	7 435
Proporce unikátních tagů s četností > 1		
z celkového počtu tagů (%)	77,8	81,1
z počtu unikátních tagů (%)	30,5	32,6

Geny přednostně exprimované v germinální a somatické části varlat byly rozlišeny s použitím striktních kritérií (p-chance < 0.05, Monte Carlo simulace, koeficient změny exprese < -5 nebo > 5), které zohlednily možnou přítomnost transkriptů ze somatických buněk v SAGE knihovně z celého varlete a transkriptů z germinálních buněk v SAGE knihovně ze somatických buněk varlete. Dle uvedených kritérií bylo vybráno 829 genů přednostně exprimovaných v germinálních buňkách a 944 genů přednostně exprimovaných v somatické části varlat. Dále bylo nalezeno 8 genů exprimovaných z mitochondriálního genomu: 6 přednostně exprimovaných v somatických buňkách, 1 přednostně exprimovaný v germinálních buňkách a u 1 genu byly zjištěny dvě isoformy (jedna přednostně exprimovaná v somatických buňkách a druhá exprimovaná výlučně v germinálních buňkách). Seznam genů přednostně exprimovaných v germinálních a somatických buňkách varlete je uveden v on-line příloze k publikaci (http://www.biomedcentral.com/1471-2164/6/29, Additional file 3).

4.1.3. Zastoupení genů exprimovaných ve varleti na chromosomu X

Předchozí studie ukázaly, že chromosom X ve srovnání s autosomy obsahuje signifikantně více genů, které jsou specificky nebo preferenčně exprimovány v samčích tkáních (v prostatě člověka, ve spermatogoniích myši) (Wang *et al.*, 2001; Lercher *et al.*, 2003b). V následujících analýzách bylo proto studováno zastoupení genů identifikovaných v SAGE knihovnách z celého varlete a somatických buněk varlete myši na autosomech a chromosomu X. K těmto analýzám byly s výhodou využity veřejně dostupné SAGE

knihovny získané z databáze GEO (Edgar *et al.*, 2002), které byly připraveny ze 7 typů somatických tkání (Tabulka 1 a Tabulka 2).

Do první analýzy byly zahrnuty všechny geny exprimované ve varleti a v somatických tkáních a byly porovnány proporce genů lokalizovaných na chromosomu X. Z celkového počtu 14 222 genů exprimovaných v celém varleti, somatických buňkách varlete a 7 somatických tkáních (mozek, oko, srdce, játra, ledviny, končetiny a tuková tkáň) byly vybrány pouze ty, které byly v souboru všech tkání reprezentovány alespoň dvěma tagy. Proporce genů exprimovaných z chromosomu X v sedmi somatických tkáních nebyla rovnoměrná (2,4 - 3,2 %), avšak rozdíly mezi jednotlivými somatickými tkáněmi nebyly signifikantní ($p > 0.05$, Chi-square test pro všechny dvojice somatických tkání). Celková proporce genů exprimovaných z chromosomu X ve skupině sedmi somatických tkání byla 3,1 % (371 z 11 903 genů). V somatických buňkách varlete bylo zjištěno 3,2 % X-vázaných genů (133 z 4 216 genů), zatímco v celém varleti bylo pouze 1,4 % genů (48 z 3 338 genů) exprimováno z chromosomu X ($p < 10^{-6}$, Chi-square test). Lze tedy uzavřít, že transkriptom celého varlete myši je ochuzen o geny exprimované z chromosomu X.

Při druhé analýze bylo srovnáváno zastoupení tkáňově specificky exprimovaných genů na autosomech a chromosomu X ve varlatech a ve skupině 7 somatických tkání. Za tkáňově specificky exprimované geny byly považovány ty, jejichž tagy byly zastoupené výhradně v jednom typu tkáně a zároveň byly v této tkáni reprezentovány alespoň dvěma tagy. Proporce X-vázaných tkáňově specifiky exprimovaných genů v celém varleti a somatických buňkách varlete byla srovnávána s proporcí tkáňově specifických genů exprimovaných z chromosomu X ve skupině 7 somatických tkáních (Tabulka 7). Z 395 tkáňově specificky exprimovaných genů v celém varleti mapovalo 3,5 % (14 genů) na chromosom X. Podobná proporce tkáňově specificky exprimovaných genů lokalizovaných na chromosomu X byla nalezana také ve skupině 7 somatických tkání. V somatických buňkách varlete však bylo nalezeno 81 tkáňově specifických genů a 13,6 % (11 genů) bylo exprimovanáno z chromosomu X, což je 3,2 násobně vyšší proporce ve srovnání s tkáňově specifickými geny ve skupině 7 somatických tkání a představuje signifikantní obohacení chromosomu X o geny specificky exprimované v somatických buňkách varlat ($p = 0,0024$, dvoustranný test, 100 000 permutací). Pro všechny X-vázané geny specificky exprimované v celém varleti a somatických buňkých varlete byly provedeny alignmenty programem BLAST na sekvenci celého chromosomu X, které potvrdily, že žádný z těchto genů nebyl na chromosomu X duplikován. Výsledek této analýzy ukazuje, že chromosom X ve

srovnání s autosomy obsahuje signifikantně více genů tkáňově specifiky exprimovaných v somatických buňkách varlete.

Tabulka 7. Proporce genů specificky exprimovaných v celém varleti, somatických buňkách varlete a ostatních somatických tkáních na autosomech a chromosomu X.

Tkáň		Pozorovaný počet genů		Očekávaný počet genů		% chrX	p-value*
		varle/som.b.varl.	ostatní t.	varle/som.b.varl.	ostatní t.	(pozorov.)	
Varle	chrA	381	836	378	839		
	chrX	14	41	17	38	3,5	0,4479
Somatické	chrA	70	885	77	878		
buňky varlete	chrX	11	39	4	46	13,6	0,0024

* dvoustranná hodnota p-value stanovená permutačním testem

4.1.4. Identifikace pozičních genových klastrů, které obsahují geny s preferenční expresí ve varleti

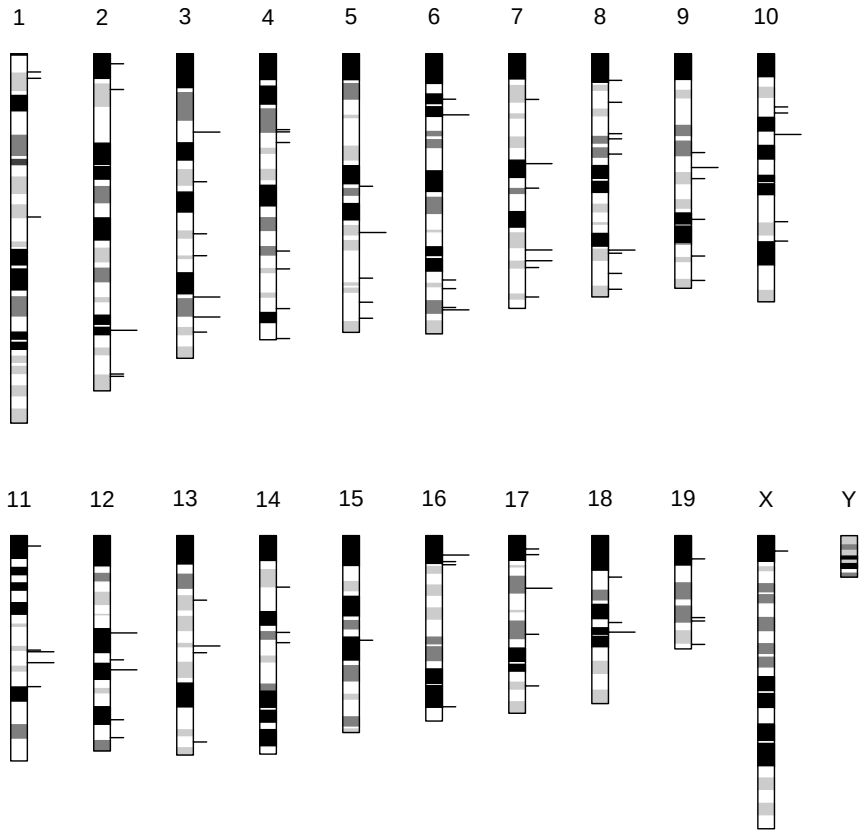
Preferenčně exprimované geny mohou být na rozdíl od tkáňově specificky exprimovaných genů exprimovány také v dalších tkáních, ale podle počtu jejich tagů v různých typech tkání lze stanovit, ve kterých tkáních jsou exprimovány preferenčně. Kritérium pro výběr těchto genů není tak přísné jako pro výběr tkáňově specifických genů, což umožnilo vybrat dostatek genů pro analýzu jejich uspořádání podél chromosomů. Geny preferenčně exprimované ve varleti (zvlášť v celém varleti a v somatických buňkách varlete) byly vybrány pomocí PEM (preferential expression measure) (Huminiecki *et al.*, 2003), které udává míru preferenční exprese v dané tkáni (viz 3.1.9). Hodnoty PEM byly vypočítány pro každý gen z celkového počtu 14 222 genů exprimovaných v celém varleti, somatických buňkách varlete a 7 somatických tkáních (mozek, oko, srdce, játra, ledviny, končetiny a tuková tkáň). S použitím vhodných kritérií pro hodnotu PEM (viz 3.1.9) bylo vybráno 1 300 genů preferenčně exprimovaných v celém varleti a 1 050 genů preferenčně exprimovaných v somatických buňkách varlete (Tabulka 8).

Tabulka 8. Počty preferenčně exprimovaných genů ve varleti identifikovaných v pozičních klastrech na chromosomech.

Tkáň	Varle		Somatické buňky varlete	
Počet preferenčně exprimovaných genů	1 300		1 050	
	v těsných klastrech	**ve volných klastrech**	**v těsných klastrech**	**ve volných klastrech**
Počet genů v klastrech (pozorovaný)	44	230	36	120
Proporce genů v klastrech z celkového počtu (%)	3,4	17,7	3,4	11,4
Očekávaný počet genů v klastrech stanovený permutacemi genomu (průměr ± sm. odcylka)	21.9 ± 8.1	168.4 ± 20.1	11.7 ± 5.9	94.2 ± 15.6
Poměr pozorovaného k očekávanému počtu genů v klastrech	2.0	1.4	3.1	1.3
Počet permutací genomu, v nichž byl počet genů v klastrech větší než nebo roven pozorovanému počtu	741	180	52	5 722
p-value (jednostranná)	0.0074	0.0018	0.0005	0.0572

Genové klastry na chromosomech byly identifikovány zvlášť pro preferenční geny exprimované v celém varleti a v somatických buňkách varlete. Následně bylo testováno, zda počet těchto genů v klastrech je statisticky významný. Pro tuto analýzu byly navíc z genomu myši odstraněny tandemově duplikované geny (viz 3.1.10.3), aby byl vyloučen jejich případný vliv na počet genů v klastrech. Pomocí posuvného okna byly na chromosomech identifikovány poziční klastry obsahující alespoň 3 přilehlé preferenčně exprimované geny („těsné klastry“) a alespoň 3 preferenčně exprimované geny mezi 6 přilehlými geny („volné klastry“) (viz 3.1.10.1). Těsné klastry tedy tvořily podmnožinu volných klastrů. Celkem bylo nalezeno 44 a 36 genů preferenčně exprimovaných v celém varleti a somatických buňkách varlete, které byly zahrnuty v 13 a 11 těsných klastrech (Tabulka 8, Obr. 6). Ve volných klastrech bylo identifikováno 230 a 120 genů preferenčně exprimovaných v celém varleti a somatických buňkách varlete (v 66 a 37 klastech). Dva těsné klastry a osm volných klastrů sdílelo geny preferenčně exprimované v celém varleti a v somatických buňkách varlete. Statistická analýza ukázala, že pozorovaný počet preferenčně exprimovaných genů zahrnutých v těsných klastrech byl 2,0 a 3,1 násobně vyšší než průměrný počet těchto genů v náhodně vygenerovaných genomech ($p = 0.0074$ a

p = 0.0005, jednostranný test, 100 000 permutací). Ačkoliv byl ve volných klastrech obsažen pouze 1,4 a 1,3 násobně vyšší počet preferenčně exprimovaných genů než v náhodně vygenerovaných genomech, v případě celého varlete byl tento počet signifikantní a v případě somatických buněk varlete téměř signifikantní (Tabulka 8). Nebylo překvapením, že geny vysoce exprimované v celém varleti, které se účastní spermatogeneze (protamine 1, 2, 3, transition protein 2), tvořily jeden z těsných klastrů na chromosomu 16. Tyto výsledky demonstrují nenáhodné uspořádání genů preferenčně exprimovaných v celém varleti a somatických buňkách varlete v pozičních genových klastrech, které nevznikly nedávnou tandemovou duplikací.



Obr. 6. Distribuce klastrů obsahujících geny s preferenční expresí ve varleti na chromosomech.

Pozice 103 klastrů podle fyzikální mapy myšího genomu je zobrazena na ideogramu s odpovídajícími cytogenetickými pruhy. Těsné klastry zahrnují alespoň 3 přilehlé geny s preferenční expresí ve varleti (dlouhé čárky). Volné klastry obsahují alespoň 3 geny s preferenční expresí ve varleti mezi 6 přilehlými geny (krátké čárky). Klastry obsahující geny s preferenční expresí v celém varleti a somatických buňkách varlete nejsou rozlišeny.

4.1.5. Vytvoření databáze veřejně dostupných SAGE knihoven z myších tkání a buněčných linií

Předpokladem pro provedení výše uvedených analýz bylo shromáždění veřejně dostupných SAGE knihoven připravených z myších tkání a buněčných linií a jejich uspořádání do databáze. Primární data většiny SAGE knihoven byla získána z veřejné databáze genových expresních dat GEO (Edgar *et al.*, 2002). Menší část dat byla stažena z webových stránek laboratoří, které jednotlivé knihovny připravily, případně získána z elektronických příloh publikací. Do databáze byly zahrnuty SAGE knihovny vytvořené nejpoužívanější kombinací restrikčních enzymů *Nla*III a *Bsm*FI, které produkují tagy o délce 10 bp. V současné době databáze obsahuje 94 myších SAGE knihoven s celkovým počtem 3 391 887 tagů.

Data všech SAGE knihoven byla zpracována stejným způsobem. Každá SAGE knihovna byla opatřena informacemi o biologickém materiálu z něhož byla připravena, zejména o zdroji tkáně či buněk, fyziologickém stavu a podrobnostmi o jejím zpracování. Každé SAGE knihovně byl poté přidělen unikátní název vystihující zdroj a stav biologického materiálu použitého k její přípravě. Při zpracování tagů byly z každé SAGE knihovny odstraněny tagy odvozené ze sekvencí adaptérů a případně upraven celkový počet tagů.

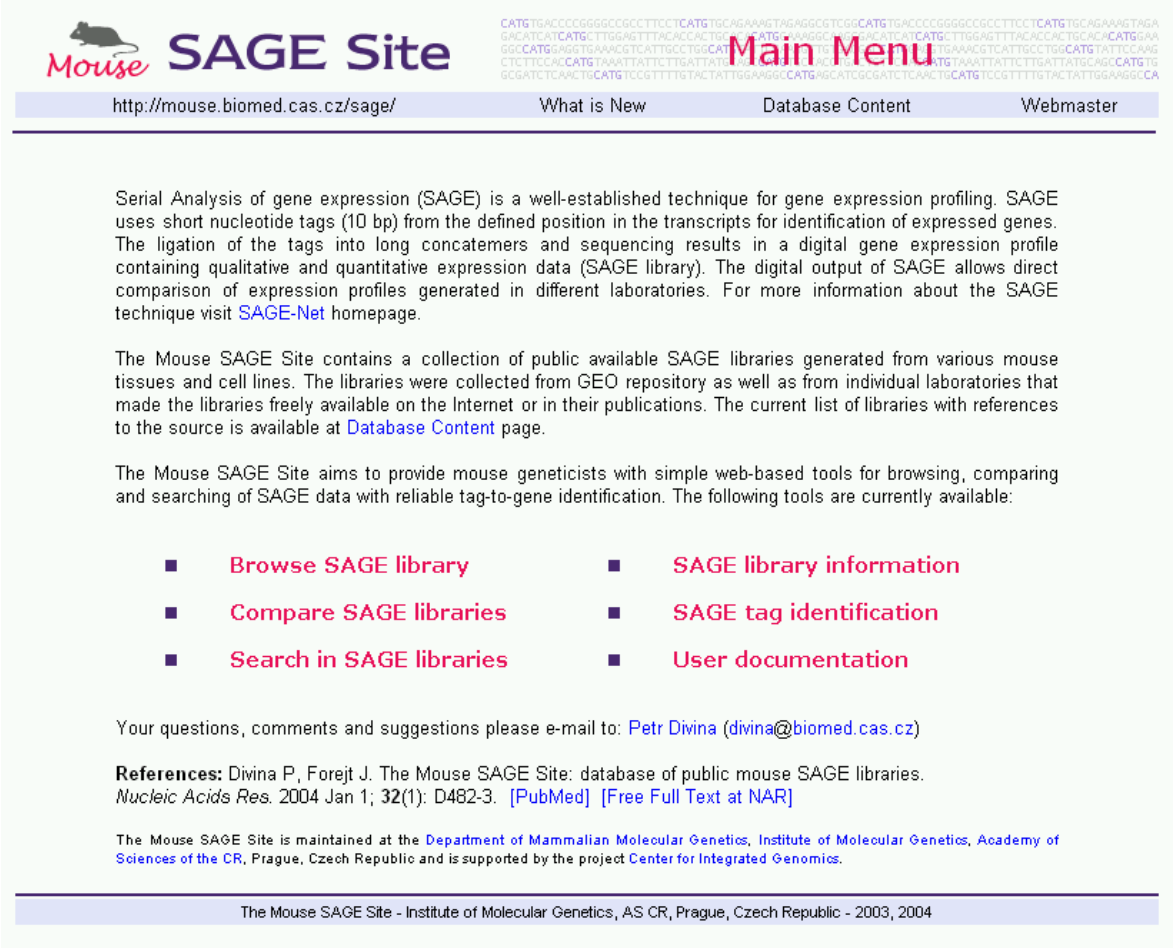
4.1.5.1. Identifikace tagů ke genům

Identifikace tagů ke genům byla založena na databázi SAGEmap (Lash *et al.*, 2000) s použitím vylepšených kritérií pro spolehlivou identifikaci genů. V databázi SAGEmap jsou tagy přiřazovány k jednotlivým transkripčním klastrům definovaným v databázi UniGene (Pontius *et al.*, 2003) na základě skóre, které zohledňuje počet a typ sekvencí mRNA a EST z nichž byl určitý tag vyextrahován. K danému tagu jsou pak přiřazeny dva transkripční klastry z databáze UniGene, které dosahovaly nejvyššího skóre, a ty představují spolehlivou identifikaci ke genům. Tímto způsobem však dochází k tomu, že velké množství tagů má identifikaci k více genům (viz Tabulka 5). Pro databázi Mouse SAGE Site bylo proto kritérium upraveno tak, že za spolehlivou identifikaci k danému transkripčnímu klastru v UniGene bylo považováno, pokud byl tag vyextrahován alespoň z jedné sekvence mRNA, alespoň 3 EST s poly(A) signálem nebo alespoň 8 EST bez poly(A) signálu. Tagy, které byly tímto způsobem přiřazeny ke 12 a více transkripčním

klastrům v UniGene byly pro snadné odlišení označeny jako ‘repetitivní’. Pro identifikaci exprese mitochondriálních genů byly také označeny všechny tagy, které bylo možno vyextrahovat z myší mitochondriální genomové sekvence (GenBank accession no. J01420) jako ‘mitochondriální’.

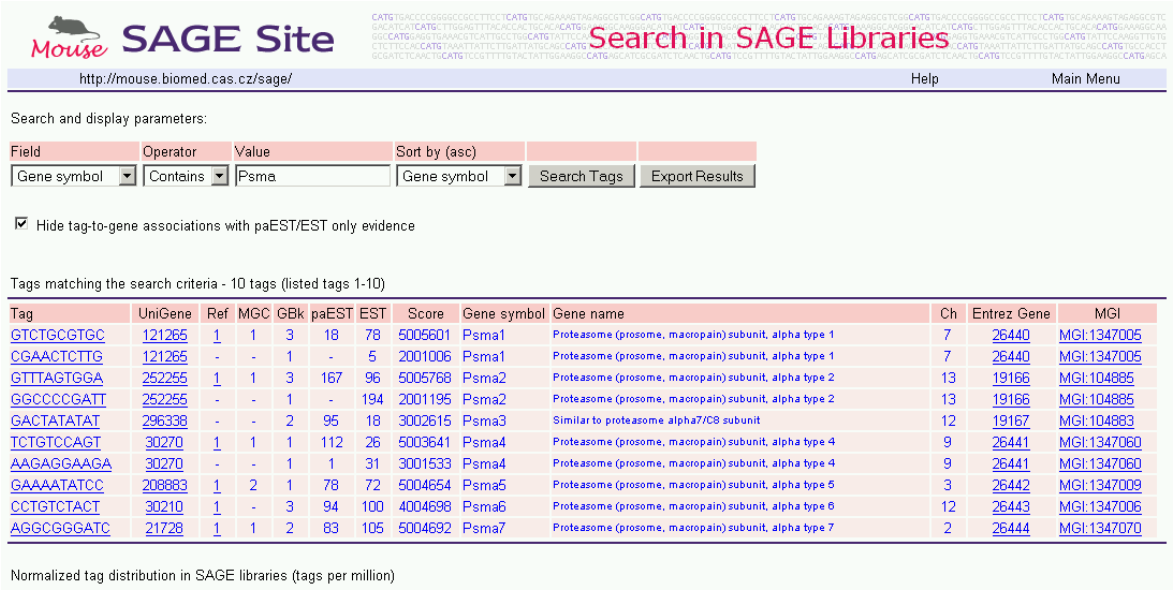
4.1.5.2. Webové rozhraní

Databáze Mouse SAGE Site je veřejně přístupná na internetové adrese: http://mouse.biomed.cas.cz/sage/ (Obr. 7). Webové rozhraní umožňuje uživatelům prohlížení, porovnávání a prohledávání SAGE dat pomocí několika snadno ovladatelných nástrojů. Uživatelé mohou například prohlížet obsah každé SAGE knihovny včetně spolehlivé identifikace tagů k transkripčním klastrům v UniGene a filtrovat seznam tagů s použitím několika parametrů (sekvence tagu, UniGene klastr, symbol genu, pozice na



Obr. 7. Hlavní stránka databáze Mouse SAGE Site.

chromosomu, Entrez Gene ID, MGI accession). Nástroj pro párové porovnávání SAGE knihoven umožňuje definovat dva soubory SAGE knihoven a zobrazit odlišně exprimované geny pro zadanou úroveň signifikance a koeficient změny exprese. Data všech SAGE knihoven je také možné prohledávat pomocí stejných parametrů jako při filtrování. Výsledek prohledávání všech knihoven je zobrazen v podobě expresní matice obsahující normalizované počty tagů (tags per million) v jednotlivých SAGE knihovnách (Obr. 8). Výsledky porovnávání a prohledávání lze uložit do textového souboru pro další analýzy. Ke každému nástroji je k dispozici podrobná on-line dokumentace.

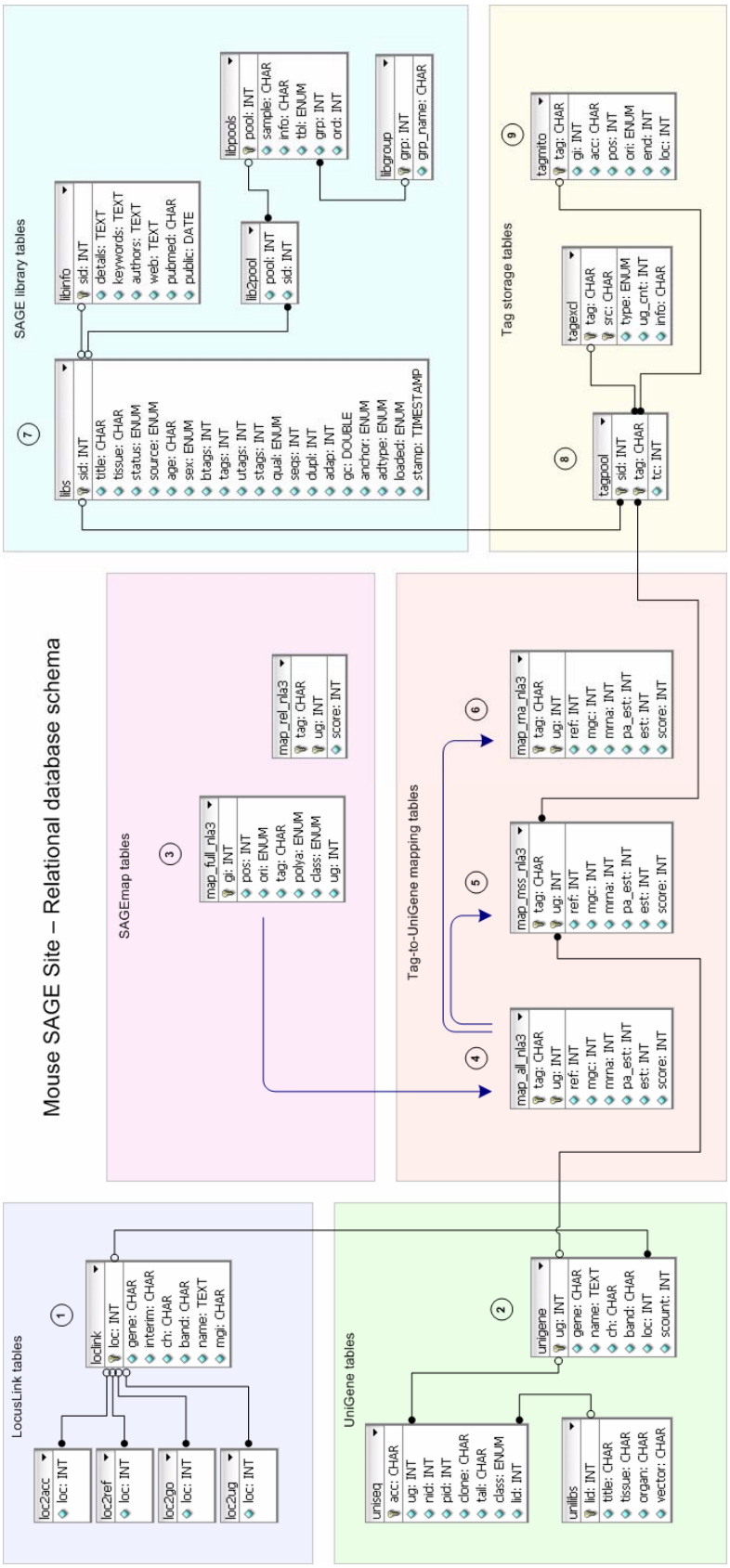


Obr. 8. Databáze Mouse SAGE Site - nástroj pro vyhledávání exprimovaných genů v SAGE knihovnách.

4.1.5.3. Technické provedení

Databáze je provozována na unixovém serveru *sun4.biomed.cas.cz* (Sun Microsystems, Solaris 8), na měmž byl nainstalován relační databázový systém mySQL (http://www.mysql.com/) a webový server apache (http://www.apache.org/). Webové rozhraní bylo vytvořeno pomocí několika cgi skriptů napsaných v programovacím jazyce Perl (http://www.perl.com/).

Vlastní strukturu databáze tvoří několik skupin tabulek, které byly navzájem propojeny relačními vztahy (Obr. 9). Tři skupiny tabulek (1, 2, 3) obsahují zdrojová data z databází LocusLink (EntrezGene), UniGene a SAGEmap. Z databáze SAGEmap byla vytříděna spolehlivá identifikace tagů k transkripčním klastrům v databázi UniGene a genům v databázi LocusLink (EntrezGene) a uložena v samostatné skupině tabulek (4, 5, 6). Další dvě skupiny tabulek obsahují anotace SAGE knihoven (7) a vlastní data – tagy a jejich počty v jednotlivých SAGE knihovnách (8, 9). Proces aktualizace databáze byl částečně automatizován a zahrnuje stažení databází SAGEmap, UniGene, Entrez Gene z FTP serveru NCBI (ftp://ftp.ncbi.nlm.nih.gov), jejich přeformátování pomocí sady perlových skriptů a naplnění příslušných tabulek v databázi. Přidávání nových SAGE knihoven spočívá ve vyplnění informací o konstrukci knihovny (anotace) v příslušné tabulce a vložení tagů a jejich počtů do databáze pomocí perlového skriptu.



Obr. 9. Schéma relační databáze Mouse SAGE Site.

4.2. Analýza genového obsahu chromosomu Z kura domácího

4.2.1. Zastoupení tkáňově specificky exprimovaných genů na chromosomu Z

Na rozdíl od genového obsahu chromosomu X nebyl genový obsah chromosomu Z zatím uspokojivě studován. Cílem této analýzy bylo zjistit, zda se liší zastoupení tkáňově specifických genů exprimovaných z chromosomu Z v somatických tkáních, varlatech a vaječnících kura domácího (*Gallus gallus*) s využitím genových expresních dat z databáze UniGene (Pontius *et al.*, 2003). Za geny s tkáňově specifickou expresí byly považovány ty, jejichž reprezentativní mRNA či EST sekvence se vyskytovaly pouze v jedné ze 14 tkání. Pro každou tkáň byla určena proporce genů s tkáňově specifickou expresí ležících na chromosomu Z (Tabulka 9).

Tabulka 9. Zastoupení tkáňově specificky exprimovaných genů na autosomech a chromosomu Z.

Tkáň	chrA *	chrZ *	% genů na chrZ
Mozek	1135	40	3,40
Končetiny	446	18	3,88
Chondrocyty	527	28	5,05
Srdce	130	3	2,26
Ledviny a nadledvinky	280	6	2,10
Tenké střevo	235	6	2,49
Játra	153	7	4,38
Pankreas	36	2	5,26
Sval	156	8	4,88
Tuková tkáň	42	1	2,33
Lymfocyty	134	7	4,96
Slezina	42	2	4,55
Somatické tkáně	3316	128	3,72
Varle	363	16	4,22
Vaječník	620	11	1,74

* počty genů na autosomech (chrA) a na chromosomu Z (chrZ)

Ve 12 somatických tkáních se proporce genů vázaných na chromosom Z pohybovala mezi 2,10 % (ledviny a nadledvinky) a 5,05 % (chondrocyty). Mezi proporcemi v jednotlivých tkáních však nebyly zjištěny signifikantní rozdíly ($p > 0,05$; Fisherův exaktní test, pro párová porovnání všech tkání) a proto byly v další analýze sloučeny všechny geny vykazující specifickou expresi v jedné z 12 somatických tkání do jedné skupiny. Průměrné zastoupení tkáňově specifických genů na chromosomu Z exprimovaných v somatických tkáních bylo 3,72 %. V porovnání s tím leželo na chromosomu Z 4,22 % genů se specifickou expresí ve varlatech ($p > 0,05$; dvoustranný test, 100 000 permutací), ale pouze 1,74 % genů se specifickou expresí ve vaječnících ($p = 0,027$; dvoustranný test, 100 000 permutací) (Tabulka 10).

Nízký počet tkáňově specifických genů exprimovaných ve vaječnících byl zjištěn pouze u chromosomu Z. U jednotlivých autosomů se proporce genů se specifickou expresí ve vaječnících pohybovala mezi 2,7 % a 5,9 % z celkového počtu genů (byly uvažovány autosomy obsahující více než 500 genů, tj. chr 1-10, 14), zatímco na chromosomu Z byla tato proporce pouze 1,7 % z celkového počtu 653 genů ($p < 0,01$, χ^2 test mezi průměrným zastoupením genů se specifickou expresí ve vaječnících na autosomech a na chromosomu Z). Výsledek této analýzy ukazuje, že chromosom Z obsahuje signifikantně menší počet genů tkáňově specifiky exprimovaných ve vaječnících.

Tabulka 10. Proporce genů specificky exprimovaných ve varlatech, vaječnících a somatických tkáních na autosomech a chromosomu Z.

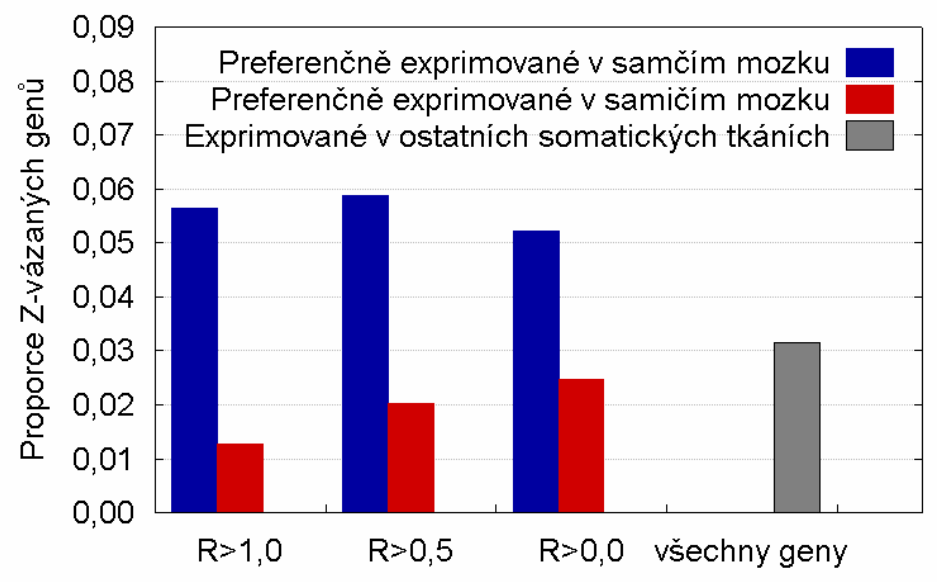
Tkáň		Pozorovaný počet genů		Očekávaný počet genů		% chrZ	p-value*
		varle/vaječník	somatická t.	varle/vaječník	somatická t.	(pozorov.)	
Varle	chrA	363	2676	365,5	2673,5		
	chrZ	16	96	13,5	98,5	4,22	0,5371
Vaječník	chrA	620	2676	611,2	2684,8		
	chrZ	11	96	19,8	87,2	1,74	0,0268

* dvoustranná hodnota p-value stanovená permutačním testem

4.2.2. Zastoupení genů s preferenční expresí v samčí nebo samičí somatické tkáni na chromosomu Z

Druhá analýza byla zaměřena na geny, které jsou preferenčně exprimované v samčí nebo samičí somatické tkáni (mozku) a bylo porovnáváno jejich zastoupení na chromosomu Z. Geny s preferenční expresí byly vybrány podle určitých prahových hodnot R statistiky, která se používá pro kvantitativní srovnávání četnosti transkriptů v cDNA knihovnách (Stekel *et al.*, 2000).

Zastoupení preferenčně exprimovaných genů (R > 1) v samčím mozku na chromosomu Z činilo 5,6 % (58 z 1029). V samičím mozku, byly naproti tomu preferenčně exprimované geny (R > 1) na chromosomu Z výrazně méně zastoupeny – bylo nalezeno pouhých 1,3 % (4 z 314) genů, což představuje vysoce signifikantní rozdíl ($p < 0{,}001$; Fišerův exaktní test). Pro méně přísné prahové hodnoty, R > 0,5 a R > 0 byly rozdíly v zastoupení genů s preferenční expresí v samčím a samičím mozku na chromosomu Z menší, avšak stále vysoce signifikantní (Obr. 10).



Obr. 10. Proporce genů s preferenční expresí v samčím nebo samičím mozku na chromosomu Z.

Geny preferenčně exprimované v samčím mozku byly na chromosomu Z zastoupeny signifikantně více než geny přednostně exprimované v samičím mozku (pro všechny prahové hodnoty R statistiky). Ve srovnání s geny exprimovanými v ostatních somatických tkáních byly geny s preferenční expresí v samčím mozku signifikantně obohacené na chromosomu Z, zatímco pro geny s preferenční expresí v samičím mozku byl patrný pouze náznak jejich ochuzení na chromosomu Z.

Geny přednostně exprimované v samčím mozku (R > 1) se na chromosomu Z vyskytovaly 1,8x častěji než geny exprimované v 11 somatických tkáních, což představovalo vysoce signifikantní obohacení ($p < 0,0001$; Fišerův exaktní test). Naopak geny s preferenční expresí v samičím mozku (R > 1) byly ve srovnání s geny exprimovanými v 11 somatických tkáních 2,5x méně zastoupené na chromosomu Z. Tento rozdíl však nebyl významný na 5% hladině signifikance ($p = 0,052$; Fišerův exaktní test), což by mohlo být způsobeno nízkým počtem genů přednostně exprimovaných v samičím mozku. Tyto výsledky ukazují, že chromosom Z má nenáhodné zastoupení genů, které jsou preferenčně exprimovány v samčím a samičím mozku.

5. DISKUSE

5.1. Transkriptom myšího varlete

Z hlediska evoluce je varle (stejně jako vaječník) velmi důležitý orgán, neboť pouze mutace vzniklé v germinálních buňkách se mohou přenášet na potomstvo. Spermatogeneze slouží také jako kontrolní bod, který eliminuje mnoho „de novo" vzniklých mutací (Cooke & Saunders, 2002; de Rooij & de Boer, 2003) a chromosomových přestaveb (Forejt, 1996; Ashley, 2002) tak, že způsobuje neplodnost jejich nositelů. Meiotická kontrola může také vést k neplodnosti mezidruhových hybridů a napomáhat tak vzniku nových druhů. Podle Haldaneova pravidla, postihuje hybridní sterilita gametogenezi ve varlatech u druhů s heterogametickým pohlavím (XY) (Forejt, 1996; Storchova *et al.*, 2004).

Pro charakterizaci transkriptomu myšího varlete byla v této práci využita sériová analýza genové exprese (SAGE) (Velculescu *et al.*, 1995). SAGE je osvědčená metoda pro globální analýzu genové exprese, která umožňuje sestavit katalog exprimovaných genů a stanovit úroveň jejich exprese. Digitální charakter získaných dat (seznam tagů a jejich počtů) umožňuje knihovny SAGE snadno kombinovat a porovnávat za předpokladu, že byly vytvořeny s použitím stejných restrikčních enzymů a pocházely z tkání či buněk jedinců stejného druhu. Využití těchto katalogů genové exprese je víceúčelové a zahrnuje hledání rozdílně exprimovaných genů v odlišných fyziologických či vývojových stavech tkání (Nacht *et al.*, 1999; Norman *et al.*, 2004), identifikaci kandidátních genů pro lidská onemocnění (Blackshaw *et al.*, 2001), funkční anotaci genomů (Saha *et al.*, 2002; Wei *et al.*, 2004) nebo analýzu uspořádání genů v genomech (Hurst *et al.*, 2004).

V této práci byly připraveny dvě knihovny SAGE z celých myších varlat odebraných dospělým samcům inbredního kmene C57BL/6J (B6). Kmen B6 byl zvolen z důvodu dostupnosti kvalitní genomové sekvence (Waterston *et al.*, 2002) a využití jeho genomu jako pozadí při konstrukci série kongenních a konsomických kmenů (Singer *et al.*, 2004); Gregorova S., Forejt J., osobní sdělení). Obě SAGE knihovny z myších varlat obsahovaly dohromady celkový počet 76 854 tagů, které představovaly 24 529 unikátních tagů. Obě SAGE knihovny byly také vysoce kvalitní a obsahovaly velmi nízký počet kon-

taminujících tagů. Genový expresní profil myšího varlete získaný z obou SAGE knihoven byl velmi podobný, což potvrdilo dobrou reprodukovatelnost SAGE. Sloučená SAGE knihovna z celého myšího varlete má srovnatelnou velikost s publikovanou SAGE knihovnou ze somatických buněk varlat dospělých myší (O'Shaughnessy *et al.*, 2003) a je téměř dvakrát tak velká jako knihovna vytvořená s použitím modifikované metody SAGE z celých myších varlat hybridních myší BDF1 (Yao *et al.*, 2004). Vzhledem k velikosti sloučené SAGE knihovny byly v transkriptomu myšího varlete zachyceny středně až vysoce exprimované geny. S použitím vylepšených kritérií pro spolehlivou identifikaci tagů, které jsou používány v databázi Mouse SAGE Site, bylo možné spolehlivě přiřadit 47,5 % tagů (3 553) k jednomu transkriptu a 15,5 % (1 157) tagů k více transkriptům definovaným v databázi UniGene (Pontius *et al.*, 2003). Na základě anotace v databázi Gene Ontology (Harris *et al.*, 2004) bylo v transkriptomu myšího varlete nalezeno přes tisíc genů, které se účastnily metabolismu proteinů a nukleových kyselin, spermatogeneze, buněčného cyklu, diferenciace, transportu a signalizace. Srovnáním vytvořených SAGE knihoven z celého myšího varlete a veřejně dostupné SAGE knihovny ze somatických buněk varlat dospělých myší (O'Shaughnessy *et al.*, 2003) byly identifikovány geny přednostně exprimované v germinálních a somatických buňkách varlat. Oba typy knihoven z myších varlat tak navzájem komplementují informace o genové expresi ve varleti dospělých myší. V nové studii byla pomocí SAGE analyzována také genová exprese jednotlivých stádií germinálních buněk myšího varlete (spermatogonií, spermatocytů a spermatid) (Wu *et al.*, 2004).

5.2. Nenáhodný obsah genů na myším chromosomu X

Předchozí studie genového obsahu chromosomu X u různých organismů přinesly protichůdné poznatky o zastoupení genů, které jsou preferenčně exprimované v samčích tkáních. Geny výlučně exprimované ve spermatogoniích u myši byly řádově vícekrát zastoupené na chromosomu X než na autosomech (Wang *et al.*, 2001). Také geny výlučně exprimované v prostatě člověka byly dvakrát častěji zastoupené na chromosomu X než na autosomech (Lercher *et al.*, 2003b). V rozporu s tím je však zjištění, že geny preferenčně exprimované v samcích se téměř nevyskytují na chromosomu X u drozofily (Parisi *et al.*, 2003). Podobně, geny exprimované ve spermatogenních buňkách jsou výrazně méně zastoupeny na chromosomu X u *C. elegans* (Reinke *et al.*, 2000; Reinke *et al.*, 2004).

Poznatky získané analýzou transkriptomu celého myšího varlete pomocí SAGE v této práci jsou v souladu s výsledky předchozí studie, která popsala ochuzení myšího chromosomu X o geny exprimované ve varleti na základě analýzy dat EST a DNA čipů (Khil *et al.*, 2004). Obohacení chromosomu X o geny tkáňově specificky exprimované v somatických buňkách varlete myši nebylo dosud zjištěno.

Jelikož germinální buňky z různých stádií spermatogeneze představují téměř 90 % buněk celého varlete u dospělých myší, ochuzení chromosomu X o geny exprimované ve varleti může souviset s potlačením transkripce z chromosomu X v průběhu meiózy. Tyto poznatky podporují představu o inaktivaci chromosomu X během prvního meiotického dělení, pro níž existují převážně nepřímé důkazy u drozofily a myši (Lifschytz & Lindsley, 1972; Forejt, 1984; Handel & Hunt, 1992; Turner *et al.*, 2002). Předpokládá se, že v haploidních germinálních buňkách, které vznikají v pozdějších stádiích spermatogeneze je již většina genů z chromosomu X transkribována. K inaktivaci chromosomu X během meiózy pravděpodobně dochází v primárních spermatocytech. Naproti tomu v Sertoliho buňkách, které tvoří somatickou část semenotvorných kanálků, může být chromosom X transkripčně aktivní. Tomu také nasvědčuje zjištění, že v transkriptomu somatických buněk varlete bylo exprimováno 3,2 % genů z chromosomu X, což je více než dvakrát tak vysoká proporce než v transkriptomu celého varlete (1.4 %). Proporce genů exprimovaných z chromosomu X v somatické tkáni varlete a dalších somatických tkáních byla navíc srovnatelná.

Geny výlučně exprimované ve varleti mohou patřit mezi geny se sexuálně antagonistickými účinky, které by se podle Riceho hypotézy měly vyskytovat častěji na chromosomu X než na autosomech (Rice, 1984). To je způsobeno tím, že na chromosomu X se v hemizigotním stavu může projevit jejich příznivý účinek u jednoho pohlaví (u samců s pohlavními chromosomy XY), zatímco u druhého pohlaví je nepříznivý účinek jejich recesívních alel maskován (u samic s pohlavními chromosomy XX). Hromadění genů, jejichž alely jsou příznivé pro samce na chromosomu X je následně umožněno působením modifikujících genů, které omezí jejich expresi pouze na samčí pohlaví (Rice, 1984). Sexuálně antagonistická selekce a inaktivace chromosomu X mají opačné účinky na genový obsah chromosomu X. Výsledky obou jevů se proto v germinálních buňkách mohou vzájemně vyrušit. Expresní profil chromosomu X v somatických buňkách je naproti tomu ovlivňován pouze sexuálně antagonistickou selekcí. V souladu s tím byla proporce tkáňově specificky exprimovaných X-vázaných genů v celém varleti podobná jako u dalších somatických tkání. Ve srovnání s tím byla u somatických buněk varlat zjištěna

3,2 násobně vyšší proporce tkáňově specifických genů exprimovaných z chromosomu X. Riceho hypotézu podporuje také řádově vyšší zastoupení genů výlučně exprimovaných v raných stádiích spermatogeneze (spermatogoniích) na myším chromosomu X než na autosomech (Wang *et al.*, 2001).

5.3. Klastrování genů s preferenční expresí v myším varleti

Uspořádání genů v genomech eukaryotických organismů je nenáhodné nejen vzhledem k odlišnému zastoupení určitých skupin genů na chromosomu X a autosomech, ale také v důsledku uspořádání genů do pozičních klastrů na chromosomech. Tyto poněkud nečekané závěry jsou výsledkem analýz globálních transkriptomů řady různých druhů eukaryotických organismů (Hurst *et al.*, 2004). Byly identifikovány různé typy genových klastrů, například rozsáhlé domény genů s podobnou expresí u drosofily a člověka (Spellman & Rubin, 2002; Versteeg *et al.*, 2003), klastry provozních (housekeeping) genů (Lercher *et al.*, 2002), klastry vysoce exprimovaných genů (Caron *et al.*, 2001), či klastry genů s podobnou šířkou exprese v oblastech s podobným obsahem GC (Lercher *et al.*, 2003c). Zdá se také, že existuje korelace mezi fyzickou velikostí klastrů a komplexitou organismu, neboť genové klastry dosahují velikosti od několika kilobází u kvasinky až po několik megabází u savců (Fukuoka *et al.*, 2004; Hurst *et al.*, 2004). Ojedinělý fenomén byl popsán u drozofily, u níž se třetina genů výlučně exprimovaných ve varleti nachází v klastrech (Boutanaev *et al.*, 2002).

V této práci byla využita statistika PEM (preferential expression measure) (Huminiecki *et al.*, 2003) pro výběr genů preferenčně exprimovaných v germinálních i somatických buňkách varlat myší a bylo analyzováno, zda jsou tyto geny uspořádány v klastrech obsahujících alespoň 3 přilehlé preferenčně exprimované geny ve varleti. Výsledky ukázaly, že počet genů preferenčně exprimovaných ve varleti (germinálních i somatických buňkách), které jsou umístěné v klastrech byl 2-3x vyšší než by se dalo očekávat v porovnání s náhodně vygenerovanými genomy. Tento jev navíc nebyl důsledkem tandemových duplikací genů. Přestože se nenáhodné umístění genů v klastrech týkalo pouze malé proporce genů preferenčně exprimovaných ve varleti, byl tento jev statisticky signifikantní. Tento poznatek je také v souladu s předchozím zjištěním, že dominantní trend pro klastrování vykazují geny exprimované v mnoha tkáních, zatímco klastrování tkáňově specificky exprimovaných genů není tak výrazné (Hurst *et al.*, 2004).

V jiné studii byl rovněž odhalen poměrně malý, ale přesto statisticky signifikantní počet genů (3-5 %), který přispívá k nenáhodnému uspořádání genomů člověka a myši (Semon & Duret, 2006). Uspořádání genů s podobnou expresí v klastrech může naznačovat společný mechanismus regulace jejich transkripce a případně jejich podobnou funkci. Důvody uspořádání genů v klastrech však mohou objasnit podrobnější analýzy genů obsažených v jednotlivých genových klastrech.

5.4. Databáze Mouse SAGE Site

S přibývajícím množstvím expresních dat, které byly vytvořeny metodou SAGE vznikaly databáze, které umožňovaly tato data pohodlně prohledávat, zobrazovat a anotovat ke genům. Transkripční mapa lidského genomu (http://bioinfo.amc.uva.nl/HTMseq/) byla jednou z prvních databází, která integrovala expresní data připravená pomocí SAGE z různých tkání člověka s DNA sekvencí lidského genomu (Caron *et al.*, 2001). Vynikající databáze SAGE Genie (http://cgap.nci.nih.gov/SAGE), která vznikla jako součást projektu zaměřeného na analýzu transkriptomu normálních a nádorových tkání (Cancer Genome Anatomy Project), shromáždila SAGE knihovny vytvořené z lidských a později také myších tkání a buněčných linií (Boon *et al.*, 2002). Tato databáze poskytuje také řadu nástrojů pro porovnávání a prohledávání SAGE dat a využívá unikátní a sofistikovaný přístup pro přiřazování tagů ke genům. Další významnou databází je Mouse Atlas of Gene Expression (http://www.mouseatlas.org/) obsahující expresní data získaná pomocí SAGE z tkání v různých stádiích embryonálního vývoje myši (Khattra *et al.*, 2006). K dispozici jsou také menší databáze obsahující SAGE data dalších organismů, např. Chicken SAGE Website (http://www.cgmc.univ-lyon1.fr/Gandrillon/chicken_SAGE.php) či Yeast SAGE (http://db.yeastgenome.org/cgi-bin/SAGE/querySAGE).

V době svého vytvoření v roce 2003 byla databáze Mouse SAGE Site (http://mouse.biomed.cas.cz/sage/) první databází, v níž byla jednotně zorganizována všechna (v té době) veřejně dostupná SAGE data z myších tkání a buněčných linií. Kromě nástrojů pro zobrazování, porovnávání a prohledávání SAGE dat byla v databázi v Mouse SAGE Site použita přísnější kritéria pro spolehlivou identifikaci tagů ke genům, která umožnila výrazně snížit množství tagů přiřazených k více genům a eliminovat nespolehlivou identifikaci tagů ke genům (viz 4.1.1 a 4.1.5.1). Od vlastní publikace byl téměř zdvojnásoben celkový počet knihoven uložených v této databázi, která v současné době

obsahuje 94 SAGE knihoven s celkovým počtem 3 391 887 tagů. Aktualizace anotace tagů ke genům v databázi Mouse SAGE Site byla pozastavena v březnu 2005 v důsledku toho, že bylo pozastaveno vytváření aktualizovaných verzí databáze SAGEmap (Lash *et al.*, 2000), kterou Mouse SAGE Site využívá pro identifikaci tagů ke genům. Během roku 2007 by měly být aktualizace databáze SAGEmap opět vydávány a měla by být vylepšena použitá metoda identifikace tagů ke genům. Předpokládá se, že poté bude databáze Mouse SAGE Site opět aktualizována. Pro budoucí rozvoj bude potřeba také upravit strukturu databáze tak, aby do ní bylo možné uložit SAGE knihovny vytvořené vylepšenou metodou LongSAGE, která produkuje tagy o délce 17 bp, případně SAGE knihovny vytvořené s použitím alternativních kotvících enzymů (např. *Sau*3AI).

5.5. Nenáhodný obsah genů na chromosomu Z kura domácího

Genový obsah párového pohlavního chromosomu u organismů s chromosomovým systémem určení pohlaví ZZ:ZW nebyl zatím uspokojivě prostudován. S využitím dostupných genových expresních dat a genomové sekvence kura domácího (Hillier *et al.*, 2004) bylo v této práci zjištěno, že chromosom Z kura obsahuje nenáhodné zastoupení určitých skupin genů podobně jako chromosom X u savců, drozofily a *C. elegans* (Hurst & Randerson, 1999; Saifi & Chandra, 1999; Reinke *et al.*, 2000; Wang *et al.*, 2001; Lercher *et al.*, 2003b; Parisi *et al.*, 2003; Khil *et al.*, 2004; Reinke *et al.*, 2004). Porovnání genového obsahu párových pohlavních chromosomů X a Z u organismů s odlišným systémem chromosomového určení pohlaví by mohlo objasnit, jaké mechanismy se podílejí na nenáhodném obsahu těchto chromosomů.

Podle teorie sexuálního antagonismu se na pohlavních chromosomech fixují geny, které jsou výhodné pro jedno pohlaví (Rice, 1984) v závislosti na jejich dominanci a na tom, kterému pohlaví přinášejí výhodu. Dominantní mutace, které zvýhodňují homogametické pohlaví by se měly hromadit na párovém pohlavním chromosomu (X či Z), protože tento chromosom se vyskytuje ve dvou kopiích v homogametickém pohlaví, avšak pouze v jedné kopii v heterogametickém pohlaví. Předchozí studie ukázaly, že geny výlučně nebo přednostně exprimované v samičích tkáních u myši a drosofily jsou signifikantně více zastoupené na chromosomu X než na autosomech (Parisi *et al.*, 2003; Khil *et al.*, 2004). Tento jev však nebyl prokázán u člověka (Lercher *et al.*, 2003b). V souladu s teorií sexuálního antagonismu byl na chromosomu Z kura domácího v této

práci zjištěn signifikantně vyšší obsah genů s preferenční expresí v samčím mozku. Proti očekávání však nebyl chromosom Z kura obohacen o geny se specifickou expresí ve varleti.

Genový obsah párových pohlavních chromosomů X a Z mohou ovlivňovat také epigenetické modifikace těchto chromosomů, ke kterým dochází v souvislosti s kompenzací dávky genů nebo s inaktivací chromosomu X během meiózy. K meiotické inaktivaci chromosomu X dochází v germinálních buňkách heterogametických samců (Lifschytz & Lindsley, 1972; Forejt, 1984; Handel & Hunt, 1992; Turner *et al.*, 2002), avšak u organismů s heterogametickými samicemi nebyla pozorována (Itoh *et al.*, 2007). Naproti tomu ke kompenzaci dávky genů v somatických buňkách dochází jak u organismů s heterogametickými samci, tak u organismů s heterogametickými samicemi, pomocí různých mechanismů - inaktivací jednoho ze dvou chromosomů X u samic savců (Lyon, 1961; Avner & Heard, 2001), zdvojnásobením exprese genů na jednom chromosomu X u samců drozofily (Baker *et al.*, 1994; Kelley, 2004) nebo snížením exprese genů na obou chromosomech X u homogametického pohlaví *C. elegans* (Meyer, 2000).

V této práci bylo zjištěno, že geny se specifickou expresí ve vaječnících jsou signifikantně méně zastoupené na chromosomu Z kura domácího a také geny s preferenční expresí v samičím mozku vykazují náznaky ochuzení na chromosomu Z. Pokud by byl mechanismus kompenzace dávky genů realizován u ptáků podobně jako u drozofily, bylo by možné vysvětlit ochuzení chromosomu Z o geny preferenčně exprimované v samičích buňkách kura podobně jako ochuzení chromosomu X o geny preferenčně exprimované v samcích drozofily. Podle hypotézy Rogerse a kolegů by v tomto případě bylo nižší zastoupení genů s preferenční expresí u samců na chromosomu X drosofily způsobeno celkově zvýšenou expresí genů na chromosomu X u samců drozofily (Rogers *et al.*, 2003). Celkové zvýšení genové exprese na chromosomu X u samců drozofily by totiž mohlo limitovat další zvyšování exprese genů, které by umožnilo jejich preferenční expresi v samčích buňkách. K rozdílnému zastoupení genů s preferenční expresí v samčím a samičím mozku na chromosomu Z by mohlo dojít také pokud by u kura domácího nedocházelo ke kompenzaci dávky genů (Scholz *et al.*, 2006). V tom případě by exprese genů vázaných na chromosom Z byla v samčích tkáních dvojnásobná oproti samičím tkáním. Z výsledků nové studie skutečně vyplývá, že mechanismsus kompenzace dávky genů je u ptáků méně efektivní než u savců (Itoh *et al.*, 2007). Podobné poznatky a závěry ohledně zastoupení genů preferenčně exprimovaných v jednom pohlaví na chromosomu Z u kura domácího přinesla také další práce (Kaiser & Ellegren, 2006).

6. ZÁVĚRY

6.1. Analýza genů exprimovaných v myším varleti a jejich uspořádání v genomu

- Pomocí expresního profilování metodou SAGE (sériová analýza genové exprese) byl vytvořen katalog genů exprimovaných ve varleti dospělých myší.

- Byly identifikovány poziční klastry genů na chromosomech obsahující geny s preferenční expresí ve varleti. Tyto klastry obsahovaly signifikantně vyšší počet genů než v náhodně vygenerovaných genomech.

- Geny specificky exprimované v somatických buňkách myšího varlete byly signifikantně obohacené na chromosomu X, což podporuje teorii o hromadění genů preferenčně exprimovaných v samčích tkáních na chromosomu X.

- Geny exprimované z chromosomu X byly ochuzené v transkriptomu celého myšího varlete, což je v souladu s představou o inaktivaci chromosomu X během prvního meiotického dělení.

- Byla vytvořena veřejně přístupná internetová databáze Mouse SAGE Site, která shromažďuje expresní data z myších tkání a buněčných liniích vytvořená pomocí metody SAGE.

6.2. Genový obsah chromosomu Z kura domácího

- Chromosom Z kura domácího byl signifikantně obohacený o geny preferenčně exprimované v samčím mozku.

- Geny s preferenční expresí v samičím mozku vykazovaly náznak ochuzení na chromosomu Z. Podobně, geny specificky exprimované ve vaječnících byly na chromosomu Z signifikantně méně zastoupené.

- Vyšší zastoupení genů s preferenční expresí v samčích tkáních na chromosomu Z by mohlo být způsobeno tím, že chromosom Z se vyskytuje v samcích ve dvou kopiích, zatímco v samicích pouze v jedné kopii.

- Geny preferenčně či specificky exprimované v samičích tkáních by mohly být ochuzené na chromosomu Z v důsledku epigenetických modifikací chromosomu Z, ke kterým dochází v souvislosti s kompenzací dávky genů.

7. SEZNAM PUBLIKACÍ

- **Divina P and Forejt J: The Mouse SAGE Site: database of public mouse SAGE libraries. Nucleic Acids Res. 2004 Jan 1; 32 (Database issue): D482-3.**

Divina P vytvořil databázi dostupných myších SAGE knihoven a psal publikaci. Forejt J vedl projekt.

- **Divina P, Vlcek C, Strnad P, Paces V and Forejt J: Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: evidence for nonrandom gene order. BMC Genomics. 2005 Mar 5; 6(1): 29.**

Divina P připravil SAGE knihovny z myších varlat a analyzoval data. Vlcek C sekvenoval SAGE knihovny. Strnad P se podílel na analýze dat. Paces V koordinoval projekt v rámci Centra Integrované Genomiky. Forejt J vedl projekt. Divina P a Forejt J psali publikaci.

- **Storchova R and Divina P: Nonrandom representation of sex-biased genes on chicken Z chromosome. J Mol Evol. 2006 Nov; 63(5): 676-81. Epub 2006 Oct 6.**

Oba autoři se podíleli stejnou měrou na zpracování dat i na psaní publikace.

8. SEZNAM POUŽITÉ LITERATURY

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F*., et al.* (2000): **The genome sequence of Drosophila melanogaster.** *Science* 287, 2185-95.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F*., et al.* (1991): **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 252, 1651-6.

Agulnik, A.I., Mitchell, M.J., Lerner, J.L., Woods, D.R. and Bishop, C.E. (1994): **A mouse Y chromosome gene encoded by a region essential for spermatogenesis and expression of male-specific minor histocompatibility antigens.** *Hum Mol Genet* 3, 873-8.

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. and Sorek, R. (2006): **Transcription-mediated gene fusion in the human genome.** *Genome Res* 16, 30-6.

Andrulis, E.D., Neiman, A.M., Zappulla, D.C. and Sternglanz, R. (1998): **Perinuclear localization of chromatin facilitates transcriptional silencing.** *Nature* 394, 592-5.

Ashley, T. (2002): **X-Autosome translocations, meiotic synapsis, chromosome evolution and speciation.** *Cytogenet Genome Res* 96, 33-9.

Avner, P. and Heard, E. (2001): **X-chromosome inactivation: counting, choice and initiation.** *Nat Rev Genet* 2, 59-67.

Axelsson, E., Smith, N.G., Sundstrom, H., Berlin, S. and Ellegren, H. (2004): **Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey.** *Mol Biol Evol* 21, 1538-47.

Baker, B.S., Gorman, M. and Marin, I. (1994): **Dosage compensation in Drosophila.** *Annu Rev Genet* 28, 491-521.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007): **High-resolution profiling of histone methylations in the human genome.** *Cell* 129, 823-37.

Bennett, S.T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005): **Toward the 1,000 dollars human genome.** *Pharmacogenomics* 6, 373-82.

Ben-Shahar, Y., Nannapaneni, K., Casavant, T.L., Scheetz, T.E. and Welsh, M.J. (2007): **Eukaryotic operon-like transcription of functionally related genes in Drosophila.** *Proc Natl Acad Sci U S A* 104, 222-7.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003): **GenBank.** *Nucleic Acids Res* 31, 23-7.

Betran, E., Thornton, K. and Long, M. (2002): **Retroposed new genes out of the X in Drosophila.** *Genome Res* 12, 1854-9.

Blackshaw, S., Fraioli, R.E., Furukawa, T. and Cepko, C.L. (2001): **Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes.** *Cell* 107, 579-89.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M.*, et al.* (2002): **A global analysis of Caenorhabditis elegans operons.** *Nature* 417, 851-4.

Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993): **dbEST--database for "expressed sequence tags".** *Nat Genet* 4, 332-3.

Bonaldo, M.F., Lennon, G. and Soares, M.B. (1996): **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome Res* 6, 791-806.

Boon, K., Osorio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Buetow, K.H., Strausberg, R.L., De Souza, S.J.*, et al.* (2002): **An anatomy of normal and malignant gene expression.** *Proc Natl Acad Sci U S A* 99, 11287-92.

Bortoluzzi, S., Rampoldi, L., Simionati, B., Zimbello, R., Barbon, A., d'Alessi, F., Tiso, N., Pallavicini, A., Toppo, S., Cannata, N.*, et al.* (1998): **A comprehensive, high-resolution genomic transcript map of human skeletal muscle.** *Genome Res* 8, 817-25.

Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y. and Nurminsky, D.I. (2002): **Large clusters of co-expressed genes in the Drosophila genome.** *Nature* 420, 666-9.

Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A. and Bickmore, W.A. (2001): **The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells.** *Hum Mol Genet* 10, 211-9.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M.*, et al.* (2000): **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* 18, 630-4.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A.*, et al.* (2001): **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 291, 1289-92.

Carvalho, A.B., Dobo, B.A., Vibranovski, M.D. and Clark, A.G. (2001): **Identification of five new genes on the Y chromosome of Drosophila melanogaster.** *Proc Natl Acad Sci U S A* 98, 13225-30.

Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000): **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 26, 183-6.

Cooke, H.J. and Saunders, P.T. (2002): **Mouse models of male infertility.** *Nat Rev Genet* 3, 790-801.

Cremer, T. and Cremer, C. (2001): **Chromosome territories, nuclear architecture and gene regulation in mammalian cells.** *Nat Rev Genet* 2, 292-301.

Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz, L.A., Jr., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E.*, et al.* (2006): **The colorectal microRNAome.** *Proc Natl Acad Sci U S A* 103, 3687-92.

de Laat, W. and Grosveld, F. (2003): **Spatial organization of gene expression: the active chromatin hub.** *Chromosome Res* 11, 447-59.

de Rooij, D.G. and de Boer, P. (2003): **Specific arrests of spermatogenesis in genetically modified and mutant mice.** *Cytogenet Genome Res* 103, 267-76.

Eberharter, A. and Becker, P.B. (2002): **Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics.** *EMBO Rep* 3, 224-9.

Edgar, R., Domrachev, M. and Lash, A.E. (2002): **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 30, 207-10.

Ellegren, H. and Fridolfsson, A.K. (1997): **Male-driven evolution of DNA sequences in birds.** *Nat Genet* 17, 182-4.

Emerson, J.J., Kaessmann, H., Betran, E. and Long, M. (2004): **Extensive gene traffic on the mammalian X chromosome.** *Science* 303, 537-40.

Forejt, J.: **X-inactivation and its role in male sterility.** In: Chromosomes Today. Edited by Gropp, A., Bennett, M.D. and Wolf, U.; George Allen and Unwin, Hemel Hempstead (England), 1984: 117-127.

Forejt, J. (1996): **Hybrid sterility in the mouse.** *Trends Genet* 12, 412-7.

Foster, J.W. and Graves, J.A. (1994): **An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene.** *Proc Natl Acad Sci U S A* 91, 1927-31.

Fukova, I., Traut, W., Vitkova, M., Nguyen, P., Kubickova, S. and Marec, F. (2007): **Probing the W chromosome of the codling moth, Cydia pomonella, with sequences from microdissected sex chromatin.** *Chromosoma* 116, 135-45.

Fukuoka, Y., Inaoka, H. and Kohane, I.S. (2004): **Inter-species differences of co-expression of neighboring genes in eukaryotic genomes.** *BMC Genomics* 5, 4.

Graves, J.A.M. and Shetty, S. (2001): **Sex from W to Z: evolution of vertebrate sex chromosomes and sex determining genes.** *J Exp Zool* 290, 449-62.

Handel, M.A. and Hunt, P.A. (1992): **Sex-chromosome pairing and activity during mammalian meiosis.** *Bioessays* 14, 817-22.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C.*, et al.* (2004): **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 32 Database issue, D258-61.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E.*, et al.* (2004): **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 432, 695-716.

Hughes, J.F., Skaletsky, H., Pyntikova, T., Minx, P.J., Graves, T., Rozen, S., Wilson, R.K. and Page, D.C. (2005): **Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee.** *Nature* 437, 100-3.

Huminiecki, L., Lloyd, A.T. and Wolfe, K.H. (2003): **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases.** *BMC Genomics* 4, 31.

Hurst, L.D. (2001): **Evolutionary genomics. Sex and the X.** *Nature* 411, 149-50.

Hurst, L.D., Pal, C. and Lercher, M.J. (2004): **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 5, 299-310.

Hurst, L.D. and Randerson, J.P. (1999): **An eXceptional chromosome.** *Trends Genet* 15, 383-5.

Hurst, L.D., Williams, E.J. and Pal, C. (2002): **Natural selection promotes the conservation of linkage of co-expressed genes.** *Trends Genet* 18, 604-6.

Charlesworth, B. (2001): **Genome analysis: More Drosophila Y chromosome genes.** *Curr Biol* 11, R182-4.

Itoh, Y., Melamed, E., Yang, X., Kampf, K., Wang, S., Yehya, N., Van Nas, A., Replogle, K., Band, M.R., Clayton, D.F.*, et al.* (2007): **Dosage compensation is less effective in birds than in mammals.** *J Biol* 6, 2.

Kaiser, V.B. and Ellegren, H. (2006): **Nonrandom distribution of genes with sex-biased expression in the chicken genome.** *Evolution Int J Org Evolution* 60, 1945-51.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M.*, et al.* (2003): **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.** *Nature* 421, 231-7.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J.*, et al.* (2003): **The UCSC Genome Browser Database.** *Nucleic Acids Res* 31, 51-4.

Kay, G.F., Ashworth, A., Penny, G.D., Dunlop, M., Swift, S., Brockdorff, N. and Rastan, S. (1991): **A candidate spermatogenesis gene on the mouse Y chromosome is homologous to ubiquitin-activating enzyme E1.** *Nature* 354, 486-9.

Kelley, R.L. (2004): **Path to equality strewn with roX.** *Dev Biol* 269, 18-25.

Kelly, W.G., Schaner, C.E., Dernburg, A.F., Lee, M.H., Kim, S.K., Villeneuve, A.M. and Reinke, V. (2002): **X-chromosome silencing in the germline of C. elegans.** *Development* 129, 479-92.

Kent, W.J. (2002): **BLAT--the BLAST-like alignment tool.** *Genome Res* 12, 656-64.

Kepes, F. (2003): **Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites.** *J Mol Biol* 329, 859-65.

Khattra, J., Delaney, A.D., Zhao, Y., Siddiqui, A., Asano, J., McDonald, H., Pandoh, P., Dhalla, N., Prabhu, A.L., Ma, K.*, et al.* (2006): **Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines.** *Genome Res*.

Khil, P.P., Smirnova, N.A., Romanienko, P.J. and Camerini-Otero, R.D. (2004): **The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation.** *Nat Genet* 36, 642-6.

Kirkpatrick, M. and Hall, D.W. (2004): **Male-biased mutation, sex linkage, and the rate of adaptive evolution.** *Evolution Int J Org Evolution* 58, 437-40.

Kosak, S.T. and Groudine, M. (2004): **Gene order and dynamic domains.** *Science* 306, 644-7.

Kruglyak, S. and Tang, H. (2000): **Regulation of adjacent yeast genes.** *Trends Genet* 16, 109-11.

Labrador, M. and Corces, V.G. (2002): **Setting the boundaries of chromatin domains and nuclear organization.** *Cell* 111, 151-4.

Lahn, B.T. and Page, D.C. (1997): **Functional coherence of the human Y chromosome.** *Science* 278, 675-80.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001): **Initial sequencing and analysis of the human genome.** *Nature* 409, 860-921.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J. and Altschul, S.F. (2000): **SAGEmap: a public gene expression resource.** *Genome Res* 10, 1051-60.

Lee, J.M. and Sonnhammer, E.L. (2003): **Genomic gene clustering analysis of pathways in eukaryotes.** *Genome Res* 13, 875-82.

Lercher, M.J., Blumenthal, T. and Hurst, L.D. (2003a): **Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes.** *Genome Res* 13, 238-43.

Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002): **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 31, 180-3.

Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2003b): **Evidence that the human X chromosome is enriched for male-specific but not female-specific genes.** *Mol Biol Evol* 20, 1113-6.

Lercher, M.J., Urrutia, A.O., Pavlicek, A. and Hurst, L.D. (2003c): **A unification of mosaic structures in the human genome.** *Hum Mol Genet* 12, 2411-5.

Levine, M., Hafen, E., Garber, R.L. and Gehring, W.J. (1983): **Spatial distribution of Antennapedia transcripts during Drosophila development.** *Embo J* 2, 2037-46.

Lifschytz, E. and Lindsley, D.L. (1972): **The role of X-chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation).** *Proc Natl Acad Sci U S A* 69, 182-6.

Lyon, M.F. (1961): **Gene action in the X-chromosome of the mouse (Mus musculus L.).** *Nature* 190, 372-3.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.*, et al.* (2005): **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature.*

Matsubara, K., Tarui, H., Toriba, M., Yamada, K., Nishida-Umehara, C., Agata, K. and Matsuda, Y. (2006): **Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes.** *Proc Natl Acad Sci U S A* 103, 18190-5.

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H. and Terauchi, R. (2003): **Gene expression analysis of plant host-pathogen interactions by SuperSAGE.** *Proc Natl Acad Sci U S A*.

Mazeyrat, S., Saut, N., Sargent, C.A., Grimmond, S., Longepied, G., Ehrmann, I.E., Ellis, P.S., Greenfield, A., Affara, N.A. and Mitchell, M.J. (1998): **The mouse Y chromosome interval necessary for spermatogonial proliferation is gene dense with syntenic homology to the human AZFa region.** *Hum Mol Genet* 7, 1713-24.

McCarrey, J.R., Watson, C., Atencio, J., Ostermeier, G.C., Marahrens, Y., Jaenisch, R. and Krawetz, S.A. (2002): **X-chromosome inactivation during spermatogenesis is regulated by an Xist/Tsix-independent mechanism in the mouse.** *Genesis* 34, 257-66.

Meller, V.H. (2000): **Dosage compensation: making 1X equal 2X.** *Trends Cell Biol* 10, 54-9.

Meyer, B.J. (2000): **Sex in the wormcounting and compensating X-chromosome dose.** *Trends Genet* 16, 247-53.

Montell, H., Fridolfsson, A.K. and Ellegren, H. (2001): **Contrasting levels of nucleotide diversity on the avian Z and W sex chromosomes.** *Mol Biol Evol* 18, 2010-6.

Nacht, M., Ferguson, A.T., Zhang, W., Petroziello, J.M., Cook, B.P., Gao, Y.H., Maguire, S., Riley, D., Coppola, G., Landes, G.M*., et al.* (1999): **Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer.** *Cancer Res* 59, 5464-70.

Norman, B., Davis, J. and Piatigorsky, J. (2004): **Postnatal Gene Expression in the Normal Mouse Cornea by SAGE.** *Invest Ophthalmol Vis Sci* 45, 429-40.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999): **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 27, 29-34.

Ohno, S.: **Sex Chromosomes and Sex Linked Genes.** Springer, Berlin, 1967.

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, 1065-71.

O'Shaughnessy, P.J., Fleming, L., Baker, P.J., Jackson, G. and Johnston, H. (2003): **Identification of Developmentally-Regulated Genes in the Somatic Cells of the Mouse Testis Using Serial Analysis of Gene Expression.** *Biol Reprod*.

Paces, J., Zika, R., Paces, V., Pavlicek, A., Clay, O. and Bernardi, G. (2004): **Representing GC variation along eukaryotic chromosomes.** *Gene* 333, 135-41.

Pal, C. and Hurst, L.D. (2003): **Evidence for co-evolution of gene order and recombination rate.** *Nat Genet* 33, 392-5.

Papp, B., Pal, C. and Hurst, L.D. (2003): **Evolution of cis-regulatory elements in duplicated genes of yeast.** *Trends Genet* 19, 417-22.

Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S. and Oliver, B. (2003): **Paucity of genes on the Drosophila X chromosome showing male-biased expression.** *Science* 299, 697-700.

Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. and Guigo, R. (2006): **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 16, 37-44.

Patefield, W.M. (1981): **An efficient method of generating R x C tables with given row and column totals.** *Applied Statistics* 30, 91-97.

Petkov, P.M., Graber, J.H., Churchill, G.A., DiPetrillo, K., King, B.L. and Paigen, K. (2005): **Evidence of a large-scale functional organization of mammalian chromosomes.** *PLoS Genet* 1, e33.

Pontius, J.U., Wagner, L. and Schuler, G.D.: **UniGene: a unified view of the transcriptome.** In: The NCBI Handbook. Edited by National Center for Biotechnology Information, 2003.

Pruitt, K.D. and Maglott, D.R. (2001): **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 29, 137-40.

Reeve, H.K. and Pfennig, D.W. (2003): **Genetic biases for showy males: are some genetic systems especially conducive to sexual selection?** *Proc Natl Acad Sci U S A* 100, 1089-94.

Reinke, V., Gil, I.S., Ward, S. and Kazmer, K. (2004): **Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans.** *Development* 131, 311-23.

Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J., Davis, E.B., Scherer, S., Ward, S.*, et al.* (2000): **A global profile of germline gene expression in C. elegans.** *Mol Cell* 6, 605-16.

Rice, W.R. (1984): **Sex-Chromosomes and the Evolution of Sexual Dimorphism.** *Evolution* 38, 735-742.

Rogers, D.W., Carr, M. and Pomiankowski, A. (2003): **Male genes: X-pelled or X-cluded?** *Bioessays* 25, 739-41.

Roy, P.J., Stuart, J.M., Lund, J. and Kim, S.K. (2002): **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.** *Nature* 418, 975-9.

Ruijter, J.M., Van Kampen, A.H. and Baas, F. (2002): **Statistical evaluation of SAGE libraries: consequences for experimental design.** *Physiol Genomics* 11, 37-44.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002): **Using the transcriptome to annotate the genome.** *Nat Biotechnol* 20, 508-12.

Saifi, G.M. and Chandra, H.S. (1999): **An apparent excess of sex- and reproduction-related genes on the human X chromosome.** *Proc Biol Sci* 266, 203-9.

Scott, M.P., Weiner, A.J., Hazelrigg, T.I., Polisky, B.A., Pirrotta, V., Scalenghe, F. and Kaufman, T.C. (1983): **The molecular organization of the Antennapedia locus of Drosophila.** *Cell* 35, 763-76.

Semon, M. and Duret, L. (2006): **Evolutionary origin and maintenance of coexpressed gene clusters in mammals.** *Mol Biol Evol* 23, 1715-23.

Shen, S.H., Slightom, J.L. and Smithies, O. (1981): **A history of the human fetal globin gene duplication.** *Cell* 26, 191-203.

Scholz, B., Kultima, K., Mattsson, A., Axelsson, J., Brunstrom, B., Halldin, K., Stigson, M. and Dencker, L. (2006): **Sex-dependent gene expression in early brain development of chicken embryos.** *BMC Neurosci* 7, 12.

Singer, J.B., Hill, A.E., Burrage, L.C., Olszens, K.R., Song, J., Justice, M., O'Brien, W.E., Conti, D.V., Witte, J.S., Lander, E.S.*, et al.* (2004): **Genetic dissection of complex traits with chromosome substitution strains of mice.** *Science* 304, 445-8.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T.*, et al.* (2003): **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.** *Nature* 423, 825-37.

Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J. and Sinclair, A.H. (1999): **Conservation of a sex-determining gene.** *Nature* 402, 601-2.

Spellman, P.T. and Rubin, G.M. (2002): **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol* 1, 5.

Stekel, D.J., Git, Y. and Falciani, F. (2000): **The comparison of gene expression from multiple cDNA libraries.** *Genome Res* 10, 2055-61.

Stevanovic, M., Lovell-Badge, R., Collignon, J. and Goodfellow, P.N. (1993): **SOX3 is an X-linked gene related to SRY.** *Hum Mol Genet* 2, 2013-8.

Storchova, R., Gregorova, S., Buckiova, D., Kyselova, V., Divina, P. and Forejt, J. (2004): **Genetic analysis of X-linked hybrid sterility in the house mouse.** *Mamm Genome* 15, 515-24.

Strahl, B.D. and Allis, C.D. (2000): **The language of covalent histone modifications.** *Nature* 403, 41-5.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A.*, et al.* (2002): **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 99, 4465-70.

Sutcliffe, M.J., Darling, S.M. and Burgoyne, P.S. (1991): **Spermatogenesis in XY, XYSxra and XOSxra mice: a quantitative analysis of spermatogenesis throughout puberty.** *Mol Reprod Dev* 30, 81-9.

Teichmann, S.A. and Veitia, R.A. (2004): **Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective.** *Genetics* 167, 2121-5.

Trachtulec, Z. (2004): **Eukaryotic operon genes can define highly conserved syntenies.** *Folia Biol (Praha)* 50, 1-6.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P. and Myers, R.M. (2004): **An abundance of bidirectional promoters in the human genome.** *Genome Res* 14, 62-6.

Turner, B.M. (2002): **Cellular memory and the histone code.** *Cell* 111, 285-91.

Turner, J.M., Mahadevaiah, S.K., Elliott, D.J., Garchon, H.J., Pehrson, J.R., Jaenisch, R. and Burgoyne, P.S. (2002): **Meiotic sex chromosome inactivation in male mice with targeted disruptions of Xist.** *J Cell Sci* 115, 4097-105.

Vallender, E.J. and Lahn, B.T. (2004): **How mammalian sex chromosomes acquired their peculiar gene content.** *Bioessays* 26, 159-69.

Vallender, E.J. and Lahn, B.T. (2006): **Multiple independent origins of sex chromosomes in amniotes.** *Proc Natl Acad Sci U S A* 103, 18031-2.

van Driel, R., Fransz, P.F. and Verschure, P.J. (2003): **The eukaryotic genome: a system regulated at different hierarchical levels.** *J Cell Sci* 116, 4067-75.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995): **Serial analysis of gene expression.** *Science* 270, 484-7.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001): **The sequence of the human genome.** *Science* 291, 1304-51.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J. and van Kampen, A.H. (2003): **The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.** *Genome Res* 13, 1998-2004.

Wang, P.J., McCarrey, J.R., Yang, F. and Page, D.C. (2001): **An abundance of X-linked genes expressed in spermatogonia.** *Nat Genet* 27, 422-6.

Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y. and Benson, G. (2004): **Inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes.** *Genome Res* 14, 1861-9.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P.*, et al.* (2002): **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 420, 520-62.

Wei, C.L., Ng, P., Chiu, K.P., Wong, C.H., Ang, C.C., Lipovich, L., Liu, E.T. and Ruan, Y. (2004): **5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation.** *Proc Natl Acad Sci U S A*.

Wu, S.M., Baxendale, V., Chen, Y., Lap-Yin Pang, A., Stitely, T., Munson, P.J., Yiu-Kwong Leung, M., Ravindranath, N., Dym, M., Rennert, O.M.*, et al.* (2004): **Analysis of mouse germ-cell transcriptome at different stages of spermatogenesis by SAGE: Biological significance.** *Genomics* 84, 971-81.

Yao, J., Chiba, T., Sakai, J., Hirose, K., Yamamoto, M., Hada, A., Kuramoto, K., Higuchi, K. and Mori, M. (2004): **Mouse testis transcriptome revealed using serial analysis of gene expression.** *Mamm Genome* 15, 433-51.

Zechner, U., Wilda, M., Kehrer-Sawatzki, H., Vogel, W., Fundele, R. and Hameister, H. (2001): **A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution?** *Trends Genet* 17, 697-701.

Zhang, X. and Smith, T.F. (1998): **Yeast "operons".** *Microb Comp Genomics* 3, 133-40.

9. PŘÍLOHY

Příloha 1. Publikace: Divina P and Forejt J: The Mouse SAGE Site: database of public mouse SAGE libraries. Nucleic Acids Res. 2004 Jan 1; 32 (Database issue): D482-3.

Příloha 2. Publikace: Divina P, Vlcek C, Strnad P, Paces V and Forejt J: Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: evidence for nonrandom gene order. BMC Genomics. 2005 Mar 5; 6(1): 29.

Příloha 3. Publikace: Storchova R and Divina P: Nonrandom representation of sex-biased genes on chicken Z chromosome. J Mol Evol. 2006 Nov; 63(5): 676-81. Epub 2006 Oct 6.

D482–D483 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh058

The Mouse SAGE Site: database of public mouse SAGE libraries

Petr Divina and Jiří Forejt*

Centre for Integrated Genomics, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Vídeňská 1083, CZ-142 20, Prague 4, Czech Republic

Received August 14, 2003; Revised and Accepted September 24, 2003

ABSTRACT

The Mouse SAGE Site is a web-based database of all available public libraries generated by the Serial Analysis of Gene Expression (SAGE) from various mouse tissues and cell lines. The database contains mouse SAGE libraries organized in a uniform way and provides web-based tools for browsing, comparing and searching SAGE data with reliable tag-to-gene identification. A modified approach based on the SAGEmap database is used for reliable tag identification. The Mouse SAGE Site is maintained on an ongoing basis at the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic and is accessible at the internet address http://mouse.biomed.cas.cz/sage/.

INTRODUCTION

Serial analysis of gene expression (SAGE) is a well-established technique for gene expression profiling (1). SAGE uses short nucleotide tags (10 bp) from the defined position in the transcripts for the identification of expressed genes. The ligation of the tags into long concatemers and their sequencing results in the qualitative and quantitative gene expression profile of a particular tissue.

The main benefits of SAGE include the digital output and the identification of novel genes. The digital output allows direct comparisons of SAGE libraries constructed in different laboratories as long as the anchoring enzyme used in construction of the library is the same. Various tests have been proposed to distinguish significantly different frequencies of tags between SAGE libraries (2–5). The identification of SAGE tags is dependent on the information stored in sequence databases. The SAGEmap database (6) is the commonly used resource for the assigment of tags to transcriptional clusters in the UniGene database (7). Tags without associations to known genes can be further analysed to discover new genes (8).

The SAGE data are usually presented on the web pages of individual laboratories or shared in the Gene Expression Omnibus (GEO) public repository (9), which serves as a central distribution hub of public expression data generated by high-throughput techniques such as microarrays and SAGE.

An excellent website known as SAGE Genie (10) was created as part of the Cancer Genome Anatomy Project. This database contains data from more than 150 human SAGE libraries, predominantly from normal and cancer tissues. Several web-based tools are available for visualization, searching and analysis of human SAGE data. SAGE Genie uses a sophisticated approach for tag-to-gene identification based on the confident tag list and the ranking of sequence databases.

Here we present the Mouse SAGE Site—the database of SAGE libraries generated from various mouse tissues and cell lines that have been publicly available to date and were constructed using the NlaIII anchoring enzyme.

ORGANIZATION OF THE DATABASE

Database construction

The database collection currently consists of 56 publicly available mouse SAGE libraries and is continuously updated. Forty-one libraries were obtained from the GEO repository (9); an additional 15 libraries were added from individual laboratories that published their libraries on the Internet or in their publications. The total of 2 150 000 tags are stored in the database at present. An up-to-date list of the assembled SAGE libraries with a reference to their source is available on the web page http://mouse.biomed.cas.cz/sage/content.

All the SAGE libraries were organized and data processed in a uniform way. The libraries were annotated with information about the tissue origin, tissue histology or pathology status, source type (bulk, cell line, cell culture) and with further information about their construction. Each library was labelled with a unique name best describing the origin and status of the tissue. The preparation of actual SAGE data included removal of the linker-derived tags and all potential 1 bp linker variations. The SAGE library size was then constituted as the total number of tags excluding linker impurities.

A modified approach based on the SAGEmap database (6) was used for reliable tag-to-gene identification. The full list of tags extracted from mRNA and EST sequences is provided as part of the SAGEmap database and is available from the internet address ftp://ftp.ncbi.nlm.nih.gov/pub/sage/map/. This list includes tags extracted from the sequences in the Reference Sequence Project (RefSeq), the Mammalian Gene Collection (MGC), and the GenBank and dbEST databases

*To whom correspondence should be addressed. Tel: +420 24447 2273; Fax: +420 29644 2154; Email: jforejt@biomed.cas.cz

(11–14). In SAGEmap, according to this list, tag-to-gene associations are classified by a reliability score and the tag-to-gene associations with the top two reliability scores are considered reliable. In the Mouse SAGE Site, tag-to-gene associations supported by at least one mRNA sequence from RefSeq, MGC, GenBank or at least three ESTs with a poly(A) signal or eight ESTs with no poly(A) signal were considered as reliable and used for tag identification (see Supplementary Material for detailed information). The tags with reliable associations to 12 or more UniGene clusters were labelled as 'repetitive/low-complexity' to be easily distinguished. All possible tags with associations to the mitochondrial genome were extracted from the mouse mitochondrion genome sequence, accession no. J01420, and labelled as 'mitochondrial'.

The Z-test algorithm described previously (4) was implemented for pairwise comparisons of tag frequencies between SAGE libraries. Performing a lot of pairwise comparisons leads to an accumulation of Type I errors and increases the chance of detecting false positives in significantly different tags. To resolve this issue, the Benjamini–Hochberg correction of false discovery rate (15) was applied.

The database was constructed to allow easy updating of supporting databases and the addition of new public mouse SAGE libraries.

Database description

The Mouse SAGE Site is accessible without restrictions via the world wide web at the address http://mouse.biomed.cas.cz/sage/. The database aims to provide mouse geneticists with easy-to-use web-based tools for exploiting mouse SAGE data.

The tools Browse, Compare and Search are currently available for the SAGE data. Users can browse the content of each SAGE library with reliable tag identification to UniGene clusters and filter the list by several criteria including tag sequence, UniGene cluster, gene symbol, chromosomal location, LocusLink and MGI accession. Separate lists of tags with matches to the mitochondrial genome, repetitive tags and tags with unreliable matches are provided for each SAGE library. The Compare tool allows users to set up two pools of SAGE libraries and display differentially expressed genes at the selected significance level and specified fold factor. The data of all SAGE libraries can be searched by similar criteria to those for the Browse tool. The Search output shows the normalized tag count distribution (tags per million) across all SAGE libraries. The results from the Compare and Search tools can be exported into tab-delimited text format for further analysis by the user. All these tools use the modified approach for reliable tag identification described above and provide direct links from gene identifiers to external databases—UniGene, LocusLink and the Mouse Genome Database (16). Online documentation explains the features of each tool in more detail.

The Mouse SAGE Site is updated as soon as new builds of SAGEmap and UniGene databases are released and new public SAGE libraries from the mouse are available. The site will be improved in accordance with the progress of tag-to-gene identification and requests from the scientific community.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online. The first part of the supplement explains the reliable tag identification approach used in the Mouse SAGE Site. Subsequent parts show sample outputs from the Compare and Search tools.

ACKNOWLEDGEMENTS

We thank J. Novotney and P. Strnad for comments on the manuscript. This work is supported by the project of the Czech Ministry of Education, Youth and Sports No. LN00A079, Centre for Integrated Genomics. J.F. is supported as an International Scholar of Howard Hughes Medical Institute.

REFERENCES

1. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
2. Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Vogelstein,B. and Kinzler,K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.
3. Ruijter,J.M., Van Kampen,A.H. and Baas,F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics*, **11**, 37–44.
4. Kal,A.J., van Zonneveld,A.J., Benes,V., van den Berg,M., Koerkamp,M.G., Albermann,K., Strack,N., Ruijter,J.M., Richter,A., Dujon,B. *et al.* (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell*, **10**, 1859–1872.
5. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
6. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
7. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
8. Chen,J., Sun,M., Lee,S., Zhou,G., Rowley,J.D. and Wang,S.M. (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl Acad. Sci. USA*, **99**, 12257–12262.
9. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
10. Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.
11. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
12. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* (2002) Generation and initial analysis of more than 15 000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
14. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
15. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
16. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.

BMC Genomics

BioMed Central

Research article

Open Access

Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: evidence for nonrandom gene order

Petr Divina, Čestmír Vlček, Petr Strnad, Václav Pačes and Jiří Forejt*

Address: Institute of Molecular Genetics, Academy of Sciences of the Czech Republic and Center for Integrated Genomics, Vídeňská 1083, CZ-142 20, Prague 4, Czech Republic

Email: Petr Divina - divina@biomed.cas.cz; Čestmír Vlček - vlcek@img.cas.cz; Petr Strnad - strnad@biomed.cas.cz; Václav Pačes - vpaces@img.cas.cz; Jiří Forejt* - jforejt@biomed.cas.cz

* Corresponding author

Published: 05 March 2005

BMC Genomics 2005, **6**:29 doi:10.1186/1471-2164-6-29

Received: 06 December 2004

Accepted: 05 March 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/29

Abstract

Background: We generated the gene expression profile of the total testis from the adult C57BL/6J male mice using serial analysis of gene expression (SAGE). Two high-quality SAGE libraries containing a total of 76 854 tags were constructed. An extensive bioinformatic analysis and comparison of SAGE transcriptomes of the total testis, testicular somatic cells and other mouse tissues was performed and the theory of male-biased gene accumulation on the X chromosome was tested.

Results: We sorted out 829 genes predominantly expressed from the germinal part and 944 genes from the somatic part of the testis. The genes preferentially and specifically expressed in total testis and testicular somatic cells were identified by comparing the testis SAGE transcriptomes to the available transcriptomes of seven non-testis tissues. We uncovered chromosomal clusters of adjacent genes with preferential expression in total testis and testicular somatic cells by a genome-wide search and found that the clusters encompassed a significantly higher number of genes than expected by chance. We observed a significant 3.2-fold enrichment of the proportion of X-linked genes specific for testicular somatic cells, while the proportions of X-linked genes specific for total testis and for other tissues were comparable. In contrast to the tissue-specific genes, an under-representation of X-linked genes in the total testis transcriptome but not in the transcriptomes of testicular somatic cells and other tissues was detected.

Conclusion: Our results provide new evidence in favor of the theory of male-biased genes accumulation on the X chromosome in testicular somatic cells and indicate the opposite action of the meiotic X-inactivation in testicular germ cells.

Background

From the selfish DNA perspective [1,2], gonads are fundamentally important organs of an organism. During the first meiotic division of gametogenesis, crossing-over enhances the re-assortment of information carried in parental DNA molecules and virtually immortal genetic information is then transferred to next generations of mortal individuals via the final products of gametogenesis, spermatozoa and eggs. Moreover, testes and ovaries are the only niches where the paternal and maternal DNA interacts with a different environment. The dissimilar gonadal environment enables sex-dependent epigenetic

Table 1: Parameters of constructed SAGE libraries from B6 mouse total testis

SAGE library	Total testis 1 (TT 1)	Total testis 2 (TT 2)
Sequenced clones	811	1 510
Total tags*	24 975	51 879
Unique transcripts	10 516	18 848
Single copy tags	7 731	13 422
Quality parameters		
Average tags per clone	30.8	34.4
Duplicated ditags	157 (1.2 %)	276 (1.0 %)
Linker derived tags	147 (0.6 %)	223 (0.4 %)

* excluding tags from duplicated ditags and linker-derived tags (i.e. two linker tags TCCCTATTAA, TCCCCGTACA and all possible 1-bp linker tag variations)

modifications of paternal and maternal DNA such as reactivation of the X chromosome in female germ cells [3,4], inactivation of a single X chromosome in pachytene spermatocytes [5-7] or differential establishment of imprinting marks on paternally or maternally imprinted genes [8,9]. Spermatogenesis also serves as an important checkpoint filtering out many *de novo* occurring gene mutations [10,11] and chromosomal rearrangements [12,13] by making their carriers sterile. A special form of meiotic checkpoint is represented by hybrid sterility, which facilitates creation of new species. Obeying the Haldane's rule, hybrid sterility preferentially affects gametogenesis in testis in species with heterogametic (XY) sex [13-15]. Molecular analyses of these phenomena are hindered by the fact that testis is a complex organ with many types of intimately intermingled somatic and germline cells. Moreover, the spermatogenic differentiation is almost impossible to achieve *ex vivo*, in a cell culture system. The main cell types can be fractionated, via gravity sedimentation, centrifugal elutriation or fluorescence activated cell sorting, but the time required can be fairly long to exclude possible artificial changes of mRNA levels.

In the present work we used Serial Analysis of Gene Expression (SAGE) [16] to characterize the transcriptome of mouse total testis. We created a catalogue of genes expressed in the adult mouse testis of the C57BL/6J (abbreviated here B6) inbred strain. The B6 inbred strain has been chosen because its genome has been recently sequenced [17] and since it has been selected as a recipient strain for creation of two sets of Chromosome Substitution Strains, C57BL/6J-Chr#$^{A/J}$ [18] and C57BL/6J - Chr#$^{PWD/Ph}$ [Gregorova S, Forejt J et al., in progress]. Except for the characterization of the total testis transcriptome, we compared our data with the publicly available SAGE library from adult testis somatic cells [19] and other SAGE libraries constructed from normal mouse tissues. Furthermore, we were interested in the organization of testicular genes in the mouse genome and we present here a detailed bioinformatic analysis of the distribution of testicular genes between the X chromosome and autosomes, and the positional clustering of genes with preferential expression in testis.

Results

Characterization of the SAGE libraries of B6 mouse testis

We have constructed two high-quality SAGE libraries, TT 1 and TT 2, from the total mouse testis of adult B6 males (Table 1). The libraries contain 24 975 (TT 1) and 51 879 (TT 2) tags corresponding to 10 516 and 18 848 unique tags, respectively. The tags with abundance > 1 comprise 17 244 (69 %) and 38 457 (74 %) of the total tag mass but only 2 785 (26.5 %) and 5 426 (29 %) of the unique tags, respectively. The high average number of tags per clone (> 30) and low contamination with linker-derived tags (< 1 %) and duplicated ditags (~1%) indicate that the SAGE libraries are of high quality. Both total testis SAGE libraries provided similar gene expression profiles ($R^2 = 0.84$ for all unique tags, Pearson correlation), which suggests a good reproducibility of SAGE data. However, a certain variation was observed in the tag abundances when 24 529 unique tags found in both total testis SAGE libraries were compared by Monte Carlo simulations. Three hundred thirteen tags exhibited significant differences in their frequency between TT 1 and TT 2 libraries at $p < 0.05$ (89 tags at $p < 0.01$) representing non-hereditary variations in transcription profiles and variations introduced by the experimental process. The fold factor value (defined as the ratio of normalized tag counts in TT 2 to TT 1 libraries, with ratios < 1 converted to reciprocal negative values) for 93.5 % of the compared tags ranged between -2.2 and 2.2 (for 99% of the tags between -5 and 5). Dot plot comparison and fold factor distribution graphs (Fig. 1A,C) depict the similarity of both total testis libraries. Despite this var-

iation, the SAGE method produced reproducible gene expression profiles and the libraries could be combined into the total testis SAGE library (referred to hereafter as TT 1+2) with the total of 76 854 tags and 24 529 unique tags. The raw data from the total testis SAGE libraries are deposited in the GEO repository [20] under accession numbers GSM34767 (TT 1) and GSM34768 (TT 2). The set of tags with abundance > 1 in TT 1+2 SAGE library with reliable tag identification is listed in Additional file 1. The testis SAGE libraries are also freely available for interactive exploration and analysis in the Mouse SAGE Site database [21].

Tag-to-gene identification in the B6 testis transcriptome

Tag-to-gene identification in the TT 1+2 SAGE library was evaluated using three different criteria applied to the SAGEmap database. The first was the most commonly used SAGEmap reliable mapping [22,23]. The second was a modified approach based on the SAGEmap full mapping file and implemented in the Mouse SAGE Site database [21]. In this approach, the tag-to-gene associations were considered reliable if supported by tags extracted from at least one mRNA sequence (from RefSeq, Mammalian Gene Collection or GenBank) or at least 3 ESTs with a poly(A) signal or at least 8 ESTs with no poly(A) signal [24]. The third approach (referred to here as RNA evidence mapping) was also based on the SAGEmap full mapping file. Tag-to-gene associations were considered reliable if supported by tags extracted from at least one mRNA sequence. The 7 481 tags with tag count > 1 in the TT 1+2 library were subjected to the SAGEmap reliable mapping that could identify 92.6 % tags to UniGene clusters (54.3 % to single and 38.3% to multiple genes; Table 2). When a more restricted reliable mapping from the Mouse SAGE Site was used, only 63% tags were identified to UniGene clusters (47.5% to single and 15.5 % to multiple genes) and about 29.6 % tags had unreliable identification to one or more UniGene clusters. Based only on the tags extracted from mRNA sequences, the RNA evidence mapping identified 51.3 % tags to UniGene clusters (45 % to single and 6.3 % to multiple genes) leaving 41.3% tags with unreliable identification. Using any of the tag identification methods, 7.4 % tags could not be identified to UniGene clusters and may be associated with novel genes. Further in this work, we used Mouse SAGE Site or RNA evidence mapping appropriately for a particular analysis (as indicated in Methods and Additional files).

Functional categories of genes expressed in total testis

We associated genes and their corresponding tag counts to functional categories from the biological process ontology of GO database [25,26] (Fig. 2). In the total testis transcriptome, we observed more than 1000 genes involved in metabolism, particularly in the protein metabolism (protein modification, protein targeting) and nucleic acid metabolism (chromatin assembly and modification, DNA replication, DNA repair, RNA processing, RNA modification). As expected, the genes associated with spermatogenesis (e.g., protamine 1 and 2, transition proteins 1 and 2), chromosome organization, cell cycle and cell differentiation were highly expressed. Notably represented gene functions also included transport (e.g., diazepam binding inhibitor-like 5, proteasome 26S subunit, ribosomal protein L23), signal transduction (e.g., calmodulin 1 and 2, sperm autoantigenic protein 17, A kinase (PRKA) anchor protein 3, PDZ domain containing 1, WD repeat domain 12), cytoskeleton organization (e.g., t-complex testis expressed1, t-complex-associated testis expressed 3, tubulin alpha7/alpha 3, tubulin alpha 6, thymosin beta 10) and apoptosis (e.g., Bcl2-associated athanogene 1, Bcl2-like 14, programmed cell death 5, tumor protein translationally-controlled 1). From the mitochondrial genome, ATP synthase 6, cytochrome c oxidase I and III were the most highly expressed genes.

Comparing the transcriptomes of total testis and adult testis somatic cells

The mouse testis is composed of two main cell types with principally different origin and functions, the germ cells that differentiate from spermatogonia to mature spermatozoa and the somatic cells that carry out all supportive functions to make the spermatogenesis and reproduction possible. Seminiferous tubules of the adult testis consist of approximately 88% germ cells and 12% somatic cells including myoid and Sertoli cells [27]. We compared our total testis SAGE library (TT 1+2) with a SAGE library constructed from the somatic cells of adult testis (GEO, accession GSM5435). This library was created from testes largely devoid of germ cells 60 days after busulphan treatment [19]. The SAGE library sizes are similar for TT 1+2 and the adult testis somatic cells (abbreviated here ATSC) comprising 76 854 and 81 478 tags, respectively. The number of unique tags (24 529 and 22 809) as well as the proportions of tags with abundance > 1 to the total tag mass (77.8% and 81.1 %) and to the number of unique tags (30.5 % and 32.6 %) are also comparable (Table 3). As anticipated, comparison of TT 1+2 and ATSC SAGE libraries using Monte Carlo simulations revealed extensive differences in gene expression between total testis and somatic cells of adult testis. Out of the 42 239 unique tags in TT 1+2 and ATSC libraries, the simulations detected significantly different tag abundances in 3 258 tags at $p < 0.05$. Concerning the fold factor, 83 % of the compared tags stretch in the range between -2.2 and 2.2 (92.5% tags between -5 and 5). At the extreme ends, 563 tags reach > 10-fold increase in tag counts in the ATSC library (fold factor > 10) and 672 tags reach > 10-fold increase in the TT 1+2 library (fold factor <-10) (see Additional file 2).

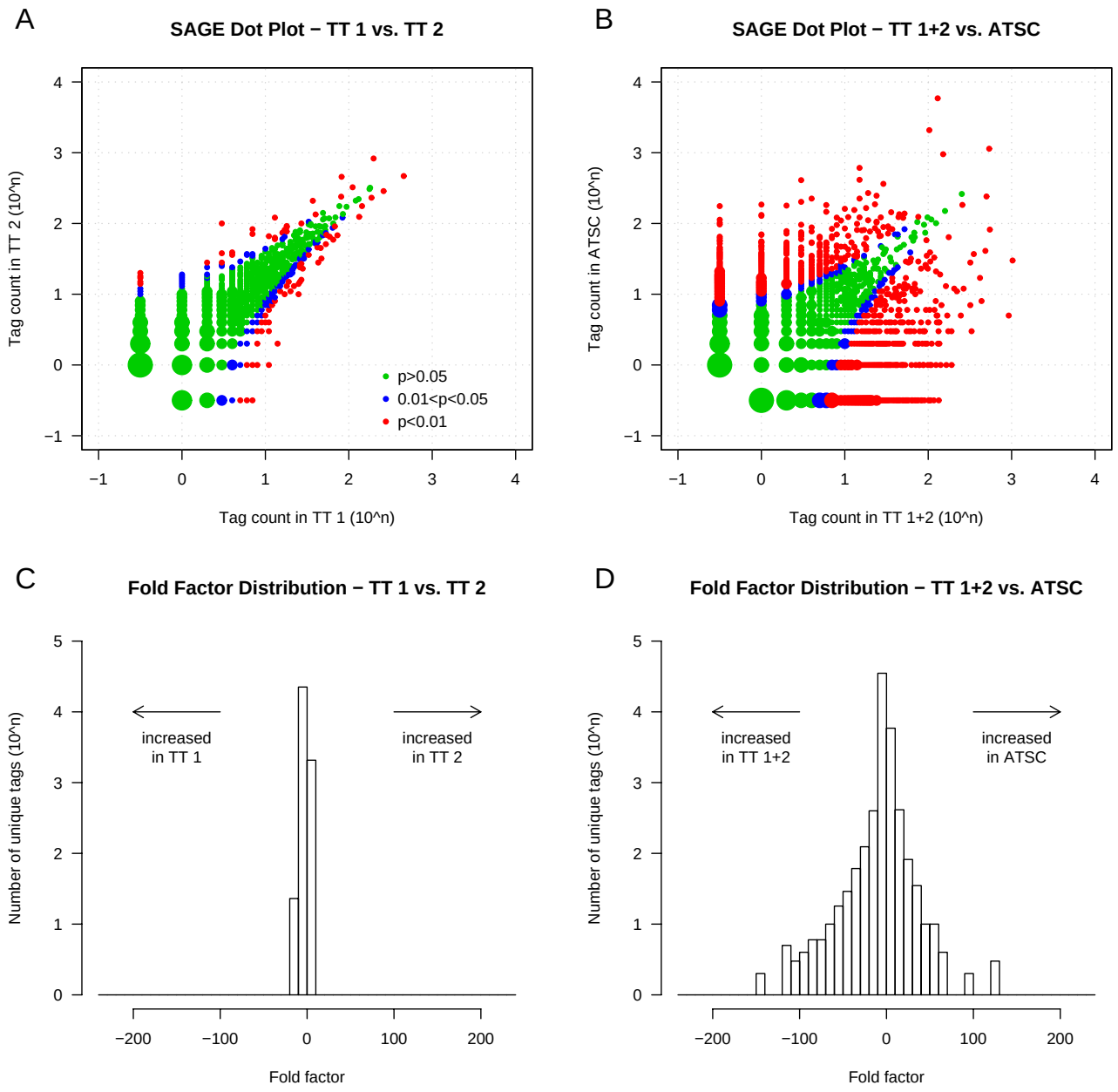


Figure 1

Comparison of mouse testis SAGE libraries represented by dot plots and fold factor distribution graphs. Comparison of tag counts between two total testis libraries (A), and between the combined total testis library and adult testis somatic cells library (B). Tags with significant p-chance are depicted in blue ($0.01 < p < 0.05$) and red ($p < 0.01$). Tags missing in one of the libraries are plotted at -0.5 coordinates. Point size is proportional to the number of represented tags. Distribution of the fold factor between two total testis libraries (C), and between the combined total testis library and adult testis somatic cells library (D). Fold factor is the ratio of normalized tag abundances in two SAGE libraries with ratios < 1 converted to reciprocal negatives. For tags missing in one library, normalized tag count of single copy tags was assumed. Abbreviations: TT 1 = total testis library 1; TT 2 = total testis library 2; TT 1+2 = combined total testis libraries; ATSC = adult testis somatic cells library.

Table 2: Identification of tags in the combined total testis SAGE library (TT 1+2) The tags matching the mitochondrial genome were omitted in this summary. Tags in group "Unreliable matches" (*) are considered not reliable according to Mouse SAGE Site and RNA evidence mappings, because they are not supported by the required number of mRNA and EST sequences. These tags are, however, included in reliable single/multiple match groups in the SAGEmap reliable mapping, which results in a highly increased number of reliable multiple matches and a slightly increased number of reliable single matches.

	NCBI SAGEmap reliable mapping		Mouse SAGE Site reliable mapping		RNA evidence mapping	
	tags	%	tags	%	tags	%
Reliable single match	4 061	54.3	3 553	47.5	3 367	45.0
Reliable multiple matches	2 865	38.3	1 157	15.5	472	6.3
Unreliable match(es)*	-	-	2 216	29.6	3 087	41.3
No match	555	7.4	555	7.4	555	7.4
Total tags (tag count > 1)	7 481	100.0	7 481	100.0	7 481	100.0

Dot plot comparison and fold factor distribution graphs for TT 1+2 and ATSC transcriptomes illustrate their dramatic dissimilarities (Fig. 1B,D).

Genes with predominant expression in the germinal or somatic component of testis

To sort out subsets of genes with predominant expression in either germinal or somatic cells of testis we applied tentative criteria to account for the presence of somatic cells in TT 1+2 and for residues of germ cells in ATSC. Predominant expression of a gene was considered if the corresponding tag was significantly more frequent in one of the libraries ($p < 0.05$, Monte Carlo simulations) and exhibited at least fivefold enrichment of tag counts (fold factor <-5 or > 5). According to this criterion a set of 829 genes is expressed predominantly in germ cells and 944 genes are expressed mainly from the somatic part of the testis (see Additional file 3). Moreover, we identified 12 tags corresponding to 8 genes encoded in the mitochondrial genome (1 gene with increased tag counts in TT 1+2 and 6 genes with increased tag count in ATSC). A gene coding for cytochrome c oxidase III (*mt-Co3*) displayed two tags separated by 87 bp in *mt-Co3* gene mRNA. One isoform was predominantly present in the ATSC library and the other was observed exclusively in the TT 1+2 library. Substantial over-expression of mitochondrial cytochrome c oxidase complexes I, II, III and NADH dehydrogenase 3 and 4 was noted in testicular somatic cells (see Additional file 3).

Exploring the dissimilarity of testis transcriptomes and transcriptomes of other mouse tissues

We examined the similarity of B6 testis transcriptomes to other available mouse SAGE transcriptomes created from normal and diseased bulk tissues by hierarchical clustering. Thirty-two SAGE libraries containing 190 871 unique tags (including single copy tags) were used as input in this analysis (see Additional file 4). We computed pair-wise library distances based on differences between normalized tag counts [28] and used the average agglomerative method for hierarchical clustering due to the highest cophenetic correlation (0.936). In the dendrogram of dissimilarities the two total testis SAGE libraries, TT 1 and TT 2, cluster together in contrast to the library from somatic cells of the adult testis (Fig. 3). The ATSC library is located separately and close to the libraries created from heart, liver and kidney in accord with the somatic origin of all these tissues. Interestingly, another SAGE library created from somatic cells of the fetal testis did not cluster with the ATSC library, but was placed close to the libraries from developing limbs, juvenile retina and whole brains. Another cluster consists of the six libraries generated from the whole adult kidneys. Several specialized brain tissues form a cluster with a brain tumor tissue (cerebellum, hippocampus, hypothalamus, medulloblastoma). An additional small cluster groups three libraries created from whole brain samples (normal male, trisomic Ts65Dn male and normal female).

Nonrandom representation of testis-expressed genes on the X chromosome

Previous works have shown a significant enrichment of prostate- and spermatogonia-specific genes on the X chromosome when compared to autosomes [29,30]. We asked what proportion of testis-expressed genes maps to the X chromosome and compared it with the proportion of X-linked genes expressed in somatic (non-testis) tissues. Furthermore, we examined whether the proportion of testis-specific genes on the X chromosome differs from the pro-

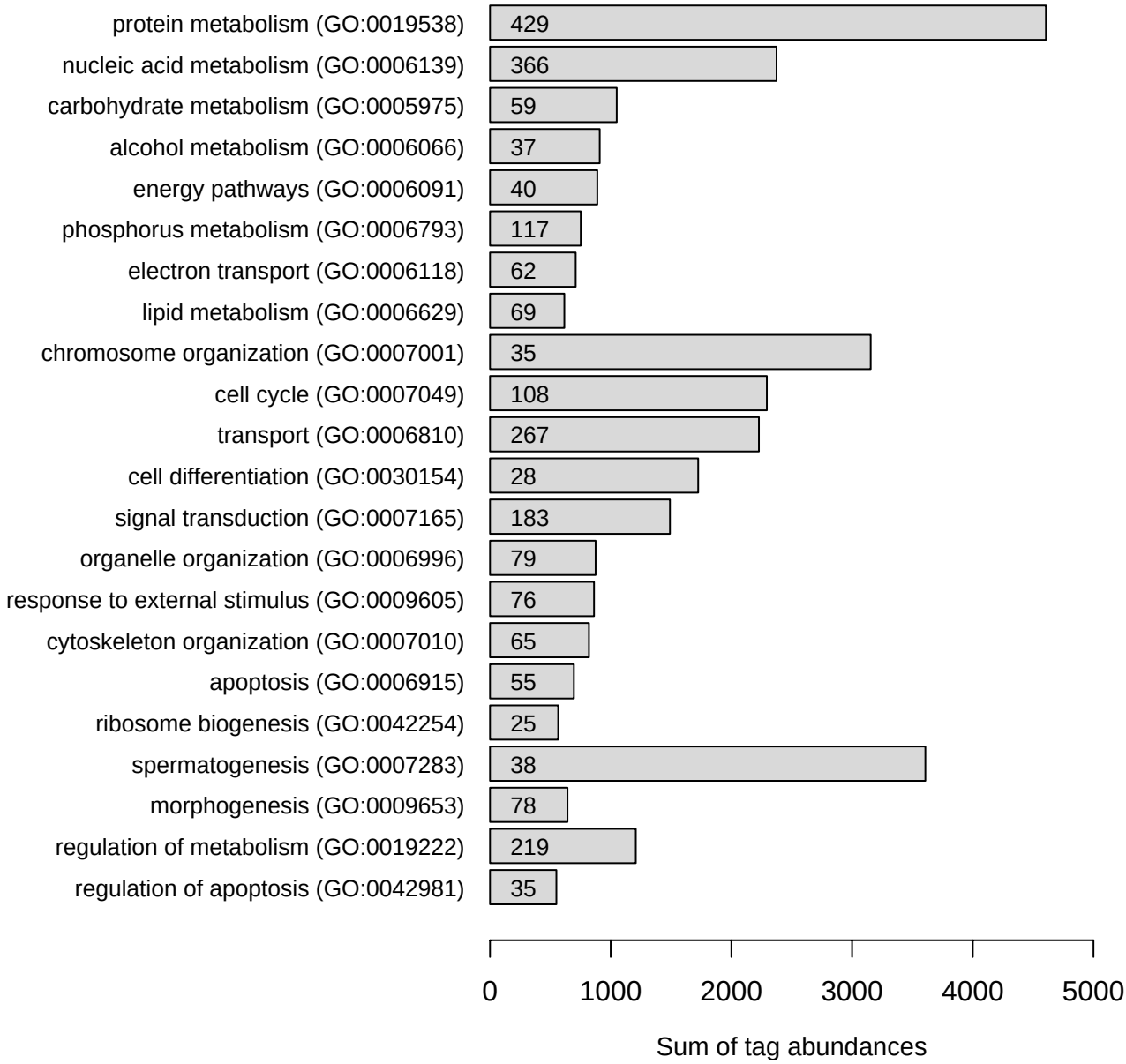


Figure 2
Classification of genes expressed in total testis according to the biological process ontology of the GO database. Bar graphs represent the sum of tag abundances corresponding to genes associated with a particular GO term. Only selected GO terms with the sum of tag abundances > 500 are displayed. The number of genes associated with each GO term is indicated inside the bars.

Table 3: Parameters of the SAGE libraries constructed from total testis and somatic cells of adult testis

SAGE library	Total testis	Adult testis somatic cells
	TT 1+2	ATSC
Total tags	76 854	81 478
Unique tags	24 529	22 809
Unique tags with count > 1	7 481	7 435
Proportions of unique tags with count > 1		
% of total tags	77.8	81.1
% of unique tags	30.5	32.6

portion of X-linked tissue-specific genes in somatic tissues.

Out of the 14 222 genes expressed in SAGE libraries from total testis, adult testis somatic cells and 7 somatic tissues (brain, eye, heart, liver, kidney, limbs and adipose tissue) (see Additional file 4) we considered only genes identified by corresponding tag count > 1. The proportion of genes expressed from the X chromosome in a pool of 7 somatic tissues was 3.1 % (374 of 11 903 genes). Although the proportions of X-linked genes in somatic tissues were uneven, there were no significant differences among the tissues (3.2 % in brain, 2.7 % in limbs and eye, 2.6 % in liver, 2.5 % in kidney and adipose tissue, 2.4 % in heart; $p > 0.05$, Chi-square test for brain vs. heart). In testicular somatic cells, we observed 3.2% X-linked genes (133 of 4 216 genes), while in total testis only 1.4 % genes (48 of 3 338 genes) were expressed from the X chromosome ($p < 10^{-6}$, Chi-square test). We can conclude that the number of expressed X-linked genes is underrepresented in the transcriptome of total testis.

The same set of 14 222 genes was examined for the distribution of tissue-specific genes on autosomes and the X chromosome. We compared the genes specific for either total testis (Table 4, a) or adult testis somatic cells (Table 4, b) in conjunction with somatic (non-testis) tissue-specific genes. A gene was considered to be tissue-specific if it was expressed only in one tissue type (total testis or adult testis somatic cells, brain, eye, heart, liver, kidney, limbs and adipose tissue). Moreover, the corresponding tag count > 1 was required to guarantee that the gene is truly expressed. The tissue-specific genes were assigned to chromosomes according to the LocusLink database and the significance of their chromosomal distribution was evaluated by permutations (see Methods) and confirmed by Fisher's exact test (Table 4). Out of the 395 genes specific for total testis 3.5% mapped to the X chromosome (see Additional file 5). Essentially the same proportion of X-linked genes was found for genes specific for 7 somatic (non-testis) tissues. In testicular somatic cells, we detected only 81 tissue-specific genes, but 13.6% were X-linked (see Additional file 5). This is a 3.2-fold increase in the proportion of testis somatic cell-specific genes on the X chromosome and represents their significant enrichment ($p = 0.0024$, two tailed, 100 000 permutations) in comparison to the genes specific for other tissues. All the X-linked testis-specific genes were subjected to BLAST against the whole X chromosome, which revealed no duplicated genes. The results from the permutation analysis indicate a significantly increased amount of testis-specific genes on the X chromosome in somatic cells of the testis when compared to autosomal testis-specific genes. The genes specific for 7 somatic tissues did not show a significant preference for the X chromosome. The list of X-linked genes expressed in total testis and testicular somatic cells with indicated testis-specific genes is available in Additional file 6.

Chromosomal clustering of genes with preferential expression in testis

Based on the data from testis and other publicly available SAGE libraries (see Additional file 4) we identified genes with preferential expression in testis by Preferential Expression Measure (PEM) [31]. PEM score controls for the genes that are highly expressed in many tissues (housekeeping genes) and reports positive values for over-expressed genes and negative values for under-expressed genes in a given tissue. Large positive PEM scores for a gene in a particular tissue indicate that the gene is unusually highly expressed in that tissue, relative to its expression in other tissues [31]. We considered a gene to be preferentially expressed if the PEM score reached at least 50 % of the maximum PEM value encountered in that tissue. Using this criterion, we scored expression of genes in total testis or testicular somatic cells in conjunction with

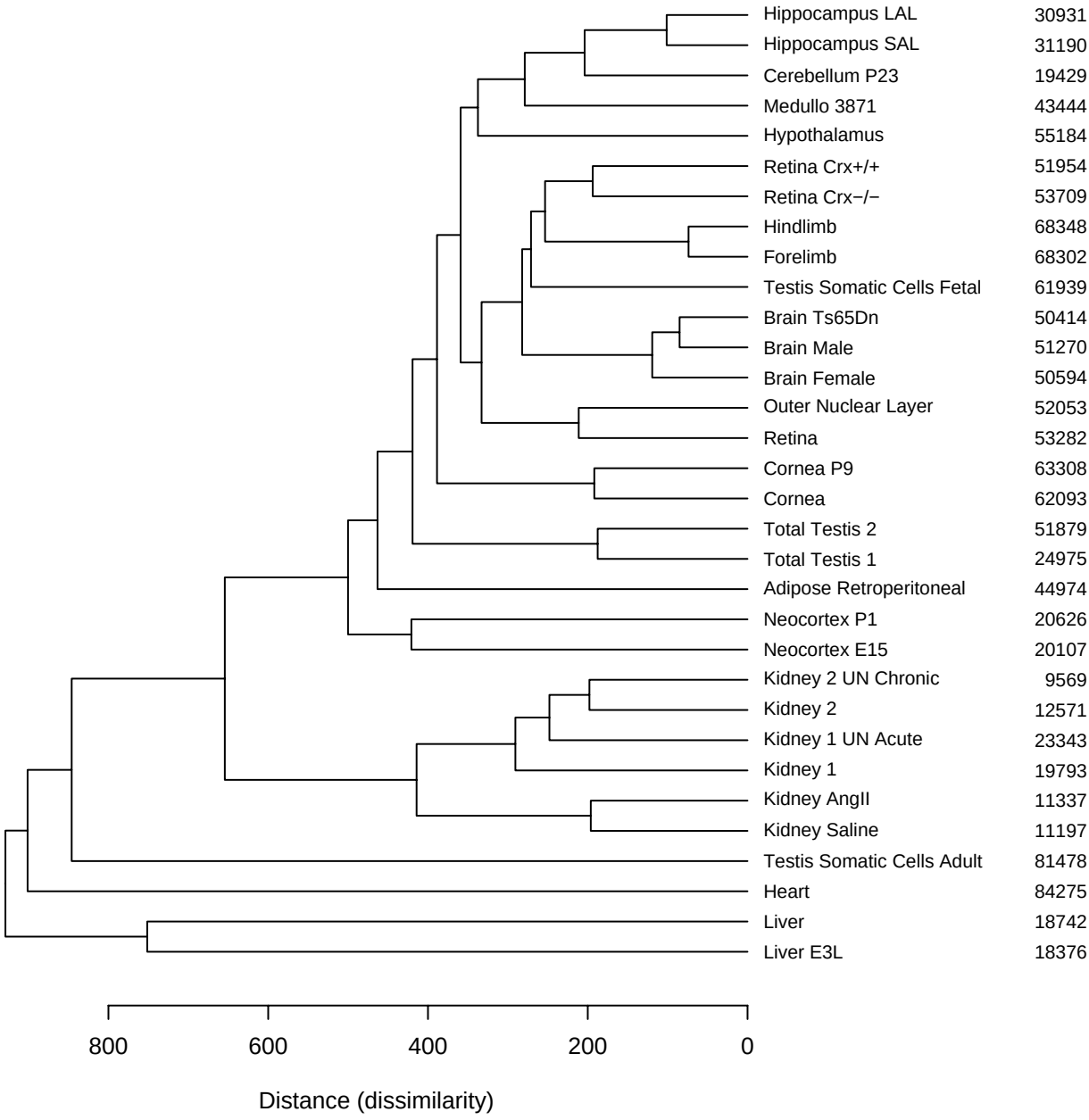


Figure 3

Dissimilarities of mouse SAGE libraries illustrated by a dendrogram. Thirty-two SAGE libraries constructed from bulk tissues containing 190 871 unique tags (including single copy tags) were selected (see Additional file 4). Pairwise library distances based on differences between normalized tag counts were computed according to [28]. The average agglomeration method was used in hierarchical clustering due to the highest cophenetic correlation (0.936) between observed and predicted distances resulting from the dendrogram. The number of tags in each SAGE library is indicated.

Table 4: Distribution of testis-specific genes on autosomes and the X chromosome The total of 14 222 LocusLink genes were identified in total testis, adult testis somatic cells and non-testis tissue SAGE libraries (see Additional file 4) using RNA evidence mapping (tags matching multiple LocusLink genes were discarded). The genes identified by total tag count = 1 were then excluded from analysis. The genes expressed only in one tissue type (total testis, adult testis somatic cells, brain, eye, heart, liver, kidney, limbs and adipose tissue) were considered to be tissue-specific genes. Chromosomal distribution of genes specific for total testis (a) and testis somatic cells (b) in comparison to the non-testis tissue-specific genes was evaluated. The significance was tested by permutations (100 000 random shufflings of the chromosomes while keeping the sum of genes on autosomes and the X chromosome fixed) and confirmed by Fisher's exact test. Abbreviations: total t. = total testis; t. somatic = testicular somatic cells; other = non-testis tissues; ChrA = autosomes; ChrX = X chromosome.

a) Total testis: 395 genes specific for the combined total testis SAGE library (TT 1+2) Other tissues: 877 genes specific for one tissue type in the pool of other SAGE libraries

Chrom	Observed gene counts		Gene counts in randomized genome		% observed gene counts		Ratio of observed proportions
	total t.	other	total t.	other	total t.	other	total t./other
ChrA	381	836	378	839	96.5	95.3	1.0
ChrX	14	41	17	38	3.5	4.7	0.7

Permutations yielding < = observed gene counts in total t. on ChrX	22 395
Permutations, p-value (two tailed)	0.4479
Fisher's exact, p-value (two tailed)	0.4563
Confidence interval (0.95)	0.70 – 2.68

b) Testis somatic cells: 81 genes specific for the adult testis somatic cells SAGE library (ATSC) Other tissues: 924 genes specific for one tissue type in the pool of other SAGE libraries

Chrom	Observed gene counts		Gene counts in randomized genome		% observed gene counts		Ratio of observed proportions
	t.somatic	other	t.somatic	other	t.somatic	other	t. somatic/ other
ChrA	70	885	77	878	86.4	95.8	0.9
ChrX	11	39	4	46	13.6	4.2	3.2

Permutations yielding > = observed gene counts in t. somatic on ChrX	121
Permutations, p-value (two tailed)	0.0024
Fisher's exact, p-value (two tailed)	0.0013
Confidence interval (0.95)	0.13 – 0.64

their expression in 7 other tissues (brain, eye, heart, liver, kidney, limbs and adipose tissue).

Further we analyzed the genome organization of genes preferentially expressed in testis. We evaluated the expression of 14 222 genes among the studied tissues and for 12 331 genes we were able to assign a genomic position according to the NCBI mouse genome assembly (build 32, mapping 19 684 known LocusLink genes). The genomic position was resolved for 5 252 and 5 843 genes

Table 5: Number of preferentially expressed genes in testis located in clusters within tandem duplicate-free mouse genome Out of the 19 684 known genes (LocusLink) mapped on mouse genome assembly (NCBI, build32), 16 858 genes remained in tandem duplicate-free genome, including 1 300 and 1 050 preferentially expressed genes in total testis and testicular somatic cells, respectively. Chromosome search found clusters containing at least three adjacent preferentially expressed genes (tight clusters) or at least three preferentially expressed genes among the six adjacent genes (loose clusters). The tight clusters therefore form a subset of the loose clusters. Observed gene counts were evaluated using permutations (100000 random shufflings of the expression status of genes while keeping the gene positions constant) and the average number of genes located in clusters in the randomized genomes was computed.

	Total testis (TT 1+2)		Adult testis somatic cells (ATSC)	
	1 300		1 050	
	in tight clusters	in loose clusters	in tight clusters	in loose clusters
Observed gene counts	44	230	36	120
Proportion of preferentially expressed genes	3.4 %	17.7 %	3.4 %	11.4 %
Gene counts in randomized genomes (mean ± std. dev.)	21.9 ± 8.1	168.4 ± 20.1	11.7 ± 5.9	94.2 ± 15.6
Ratio observed/mean in randomized genomes	2.0	1.4	3.1	1.3
Permutations yielding > = observed gene counts	741	180	52	5 722
p-value (one tailed)	0.0074	0.0018	0.0005	0.0572

expressed in total testis and testicular somatic cells, respectively, including 1 438 (27.4%) and 1 197 (20.5%) preferentially expressed genes, respectively (see Additional file 7). To evaluate the gene order of preferentially expressed genes in testis and to eliminate the effect of tandem duplications we purged the whole mouse genome of tandemly duplicated genes (see Methods). The tandem duplicate-free genome resulted in total of 16 858 LocusLink genes and preserved 1 300 and 1 050 genes preferentially expressed in total testis and testicular somatic cells, respectively. Using a search with a sliding window (see Methods) we localized chromosomal clusters containing at least three adjacent preferentially expressed genes (tight clusters). Similarly, we searched for clusters with at least three preferentially expressed genes among the six adjacent genes (loose clusters) to include genes that could be preferentially expressed but did not pass the above criterion for preferential expression or their expression was not detected by SAGE. By definition, the tight clusters form a subset of the loose clusters. The chromosomal distribution of clusters with preferentially expressed genes in testis is illustrated in Figure 4. We observed 44 and 36 genes preferentially expressed in total testis and testicular somatic cells located in 13 and 11 tight clusters, respectively. Two hundred and thirty and 120 genes preferentially expressed in total testis and testicular somatic cells resided in 66 and 37 loose clusters, respectively (Table 5; Additional file 8). Two of the tight clusters and eight of the loose clusters shared preferentially expressed genes between total testis and testicular somatic cells. Statistical analysis revealed that the observed number of preferentially expressed genes located in tight clusters was 2.0-fold and 3.1-fold higher for total testis and testicular somatic cells, respectively, than the average number of preferentially expressed genes located in clusters in randomized genomes ($p = 0.0074$ and $p = 0.0005$, one tailed, 100 000 permutations). Although only slighly higher (1.4- and 1.3-fold) than the average in randomized genomes, the observed number of preferentially expressed genes in testis located in loose clusters was still significant in case of total testis and nearly significant in case of testicular somatic cells (Table 5). Not surprisingly, the most highly expressed genes detected in total testis and involved in spermatogenesis (protamine 1, 2, 3 and transition protein 2) formed one of the tight clusters on chromosome 16. The results indicate a nonrandom distribution of the genes preferentially expressed in total testis and testicular somatic cells into chromosomal clusters, which did not arise from tandem duplications.

Comparing the B6 and BDF1 total testis transcriptomes

In a recent study focused on senescence changes in testis, a modified SAGE method was used to generate digital gene expression profiles of total testis from 3- and 29-month-old mice of the BDF1 strain and 14-month-old mice of the SAMP1 strain that exhibits an accelerated senescence [32]. Because of the different anchoring enzyme (*Rsa*I) used in construction of the libraries and the limited availability of data from the BDF1 testis transcriptome, we could perform only a rough manual comparison of our B6 testis transcriptome (76 854 tags) and the combined BDF1 testis transcriptome from 3- and 29-month-old BDF1 mice (41 221 tags). We focused on the most highly expressed testicular genes in GNF Mouse Atlas v2 [33,34] that were detected by Affymetrix Gene-

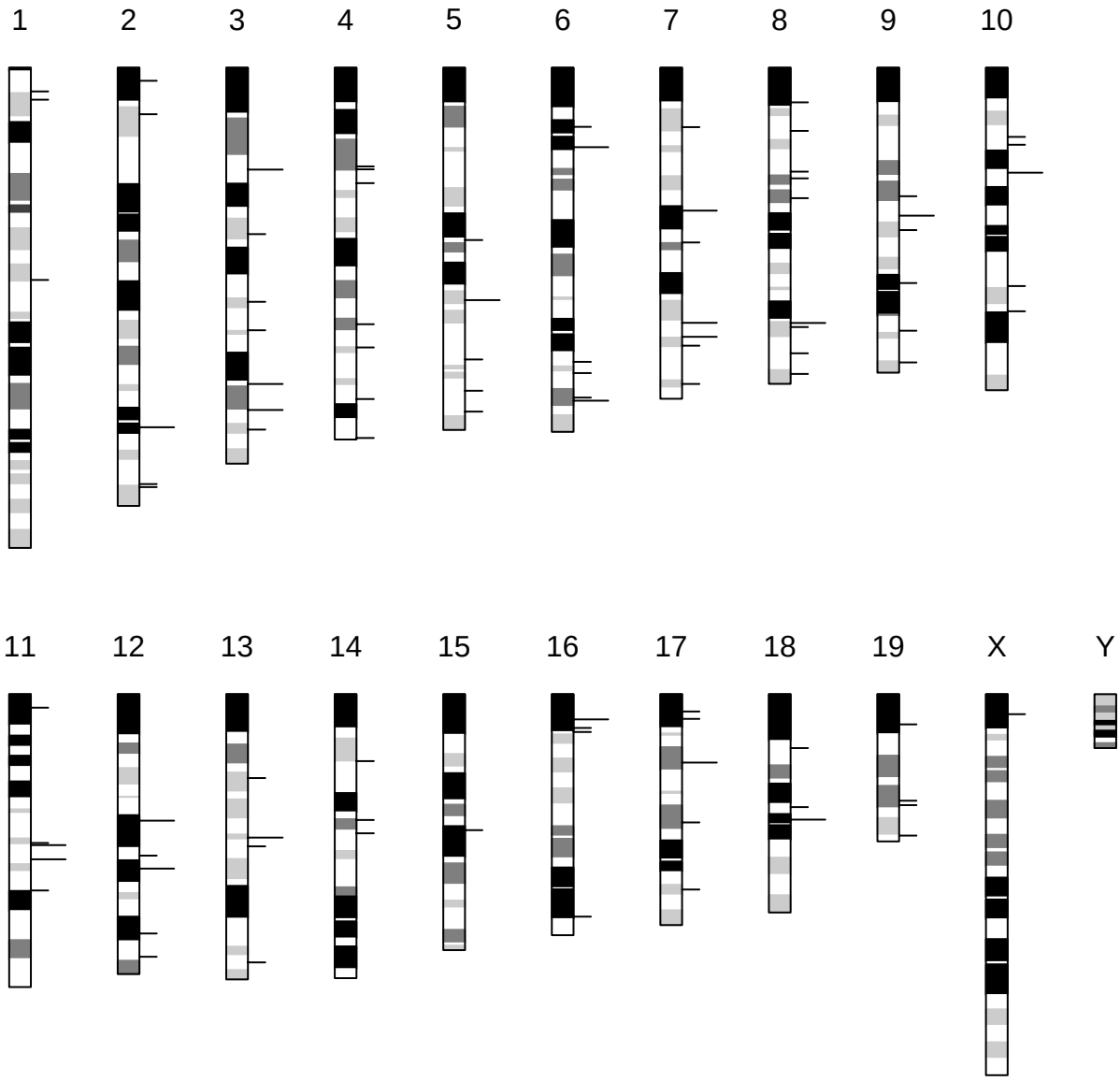


Figure 4

Chromosomal positions of clusters containing preferentially expressed genes in testis. The positions of 103 gene clusters according to the physical map are displayed on an ideogram with corresponding cytogenetic bands on chromosomes. The clusters revealed in total testis and testicular somatic cells are not distinguished. Tight clusters (long dashes) form a subset of loose clusters (short dashes).

Chips. A set of 35 highly expressed genes in testis (average difference > 9 000) was organized with SAGE tag counts from B6 and BDF1 testis (see Additional file 9). In the B6 total testis, we detected 33 out of 35 genes (the *Serf1* gene could not be distinguished because its low complexity tag matches multiple genes and the *Cox7a2* gene is not

detected because its transcript lacks *Nla*III restriction site). In contrast, only 9 genes were detected in the BDF1 testis library, 13 genes were missing due to the absence of *Rsa*I restriction site in the transcript and for 13 other genes the expression data from BDF1 testis were not publicly available. Furthermore, out of the 35 highly expressed genes in testis, 21 genes were among the top 100 most expressed genes in the B6 total testis library, but only 9 genes were among the top 100 most expressed genes in the BDF1 total testis library. It appears that our SAGE data from the B6 testis transcriptome shows better correspondence to the microarray data than the data from the transcriptome of BDF1 testis.

Discussion

Serial analysis of gene expression is a high-throughput method for building a catalogue of expressed genes and their expression levels of "normal" as well as diseased or genetically variant tissues and organs [16]. The digital character of SAGE data enables addition and direct comparison of different SAGE libraries, provided they were built with the same anchoring enzyme and originated from individuals of the same species. The utilization of such global transcriptome databases is multifold, including positional cloning of mutations or quantitative trait loci [35,36], functional genome annotation [37,38] or analysis of a nonrandom gene order [39]. Admittedly, the SAGE, as used in this work, has several limitations, including a significant proportion of repetitive and low complexity tags. The SAGE is obviously more labor-intensive than transcriptome analysis based on microarrays. At present, some of these inconveniencies can be solved by applying LongSAGE or massively parallel signature sequencing technologies [38,40].

In this study we constructed a SAGE library of the total testis of the C57BL/6J (B6) mouse inbred strain, compared it with other public available mouse SAGE libraries and analyzed localization of testis-expressed genes within the mouse genome. The B6 strain was favored for the availability of its high-quality draft genomic sequence [17] and because series of congenics and recently also consomic strains have used the B6 strain as a background strain [18] [Gregorova S, Forejt J, personal communication]. The combined total testis SAGE library, TT 1+2, consisted of 76 854 total tags representing 24 529 unique tags. The tag-to-gene reliable identification method used in Mouse SAGE Site [24] was applied to tags with frequency ≥ 2. Out of these tags, 47.5% (3 553) revealed a reliable match to single and 15.5% (1 157) to multiple UniGene clusters. Considering the size of the total testis SAGE library, medium to highly expressed genes are present in the expression profile. The library size is comparable to the recently published SAGE library of somatic cells of the mouse testis [19] and almost twice the size of a library constructed from the total testis of BDF1 hybrid mice using a modified SAGE method [32].

Contrary to microarrays, SAGE data are platform independent, which permits the use of unrelated datasets coming from various sources to compare gene expression patterns. We analyzed the mouse testis transcriptome by comparing our total testis SAGE library to the adult testis somatic cells library [19] and to additional publicly available SAGE libraries from 7 different tissues. We recognized three different modes of differential expression. (1) Predominant expression of genes in the germinal or somatic part of the testis, which did not consider expression in other tissues. (2) Preferential expression in testis that was defined by comparing the expression of testis to 7 somatic tissues for which SAGE data were available. (3) Testis-specific expression that was defined by null expression (at the resolution of a particular SAGE library) in SAGE libraries of seven tissues or organs other than testis. Complete lists of genes predominantly expressed in germinal or testis-somatic cells, as well as the catalogues of genes preferentially expressed in testis and testis-specific genes are available online in Additional file 3 , 5 and Additional file 7.

Conflicting results have been reported on the representation of male-biased genes on the X chromosome in various species. Spermatogonia-specific genes were found to be an order of magnitude more abundant on the mouse X chromosome [30]. In human, the prostate-specific genes were twice more frequent on the X chromosome, but the female mammary gland- and ovary-specific X-linked genes were not enriched in respective SAGE libraries [29]. On the contrary, under-representation or absence of male-biased genes on the X chromosome was reported in *Caenorhabditis elegans* [41] and in *Drosophila* [42,43]. In the mouse, an under-representation of testis-expressed and testis-enriched genes on the X chromosome was also revealed by the analysis of microarray and EST data [5-7]. Our present data favor under-representation of X-linked genes in the total testis transcriptome but not in testis-somatic cells. Because the germ cells in different stages of differentiation constitute about 90% of the total cell mass of testis, the data indicate that the deficit of X-linked testis-expressed genes may reflect the lack of transcription from the X chromosome in meiotic cells. These results are in agreement with the idea of X-chromosome silencing during the first meiotic division, the phenomenon based mostly on circumstantial evidence in flies and mice [7,44-46]. Thus, transcription at the haploid stage of spermatogenesis is expected for most of the X-linked genes expressed in total testis. The meiotic X chromosome inactivation seems to be restricted to primary spermatocytes, but Sertoli cells, which form the somatic part of seminiferous tubules, may have the X chromosome in the active

state. Indeed, in the transcriptome of adult testis somatic cells the proportion of expressed X-linked genes (3.2 %) was more than twice higher than in total testis (1.4 %) and did not differ from the proportion of X-linked genes expressed in non-testis (somatic) tissues.

Testis-specific genes belong to a wider category of sex-biased genes, which according to the hypothesis of sexually antagonistic genes are more likely to spread on the X chromosome than on autosomes [47]. This is because on the X chromosome they will express their favorable effect in the hemizygous state (XY) while their deleterious effect will be masked by their recessivity in the other sex (XX). Consequently, accumulation of male-specific genes on the X chromosome will be possible by the effect of modifiers that narrow the expression of sex-biased genes only to the male sex [47]. Thus, the evolution of sexually antagonistic genes and X inactivation may act as opposing forces on the germline lineage of testis while accumulation of male-specific genes could be expected in somatic cells of testis. In accord with these assumptions the proportion of X-linked genes specific for total testis did not significantly differ from the proportion of genes specific for other tissues, while we observed a significant 3.2-fold enrichment of the proportion of X-linked genes specific for testicular somatic cells.

The eukaryotic gene order is nonrandom obviously not only due to shifting of sex-biased genes to and from the X chromosome, but also owing to a nonrandom clustering of genes within chromosomes. This somewhat unexpected conclusion (taking into account the relative autonomy of transgene regulation) is gaining gradual support from global transcriptome analyses of various eukaryotic species (see Hurst et al. for review) [39]. The observed examples of clustering are apparently a mixture of several unrelated phenomena, including large domains of similarly expressed genes in *Drosophila* and humans [48,49], clustering of housekeeping genes [50], clustering of highly expressed genes [51] or genes with similar expression breadth in regions of similar GC content [52]. In *Drosophila melanogaster* one third of testes-specific genes occur in clusters [43], a phenomenon not reported in any other species. Using PEM [31] to define preferentially expressed genes we were able to demonstrate that in the mouse, the genes preferentially expressed in germ cells as well as in somatic cells of testis occur in tight clusters with a frequency 2.0-fold and 3.1-fold higher than the expected average frequency in randomized genomes. Moreover, our results indicate that this phenomenon is not merely a consequence of tandem duplications. Further analysis of clustering of testis-expressed genes may reveal new insights into the functional organization of the mammalian genome.

Conclusion

We identified chromosomal clusters of adjacent genes with preferential expression in testis that contain a significantly higher number of genes than expected by chance. This phenomenon is not merely a consequence of tandem duplication. The genes with specific expression in testicular somatic cells are more abundant on the X chromosome, which favors the theory of accumulation of male-biased genes on the X chromosome. In contrast, the X-linked genes are under-represented in the transcriptome of total testis, which is in accordance with the idea of X-chromosome inactivation during the first meiotic division.

Methods

Tissue collection and RNA isolation

Mice were housed in specific pathogen free environment and their manipulation was in accordance with the Czech Animal Protection Act No. 246/92, 162/93, and decrees No. 311/97, fully compatible with the NIH Publication No. 85-23, revised 1985. Testes were obtained from 9-week-old males of the C57BL/6J mouse strain. The animals were killed by cervical dislocation; the testes were quickly removed from the body and released from tunica. The total RNA was extracted from homogenized testes using TRIzol (Invitrogen) according to the manufacturer's protocol. SAGE libraries were constructed from the total RNA isolated from both testes of a single male (TT 1) and from the pool consisting of equal weight amounts of total RNA isolated from both testes of three male littermates (TT 2).

Construction of SAGE libraries, sequencing and tag extraction

SAGE libraries were constructed as described in the Micro-SAGE protocol version 1.0e available from SAGE homepage [53] using *Nla*III as the anchoring enzyme and *Bsm*FI as the tagging enzyme. Two minor modifications of the MicroSAGE protocol were employed: the first strand cDNA synthesis reaction was incubated at 42°C and the amount of linkers used in the linker ligation step was decreased to ~10 ng. Sequencing was performed in a Beckmann Coulter CEQ 2000 DNA Analysis System. The sequence files were processed for the tag extraction using a custom Perl script. Tags were extracted only from clones containing > 2 ditags. Duplicated ditags, linker tags and all 1-bp linker variations were removed. Data of total testis SAGE libraries are available in the GEO repository [20] under accession numbers GSM34767 (TT 1) and GSM34768 (TT 2).

Identification of SAGE tags

Tag identification to UniGene clusters was done using three methods: SAGEmap reliable mapping [22], Mouse SAGE Site reliable mapping [24] and RNA evidence map-

ping. The SAGEmap reliable mapping [23] uses a reliability score to classify tag-to-gene associations and tag-to-gene associations with the top two reliability scores are considered reliable. The Mouse SAGE Site [21] reliable mapping is based on the SAGEmap full mapping file and considers reliable the tag-to-gene associations that are supported by tags extracted from at least one mRNA sequence (from RefSeq, Mammalian Gene Collection, GenBank) or at least 3 ESTs with a poly(A) signal or at least 8 ESTs with no poly(A) signal. The RNA evidence mapping is also based on the SAGEmap full mapping file and considers reliable only tag-to-gene associations supported by tags extracted from at least one mRNA sequence. Mitochondrial tags were identified using all possible tags extracted from the mouse mitochondrial genome reference sequence [GenBank:NC_005089].

Comparison of testis SAGE libraries

Tags significantly different between SAGE libraries were determined by Monte Carlo simulations. Using the described algorithm [54] a set of 100 000 random tables was generated keeping the row and column totals of the observed data fixed. For each tag, the proportion of simulations that produced a difference equal to or greater than the observed difference (p-chance) was computed. The set of 100 000 random tables was generated six times and the average p-chance was calculated. The fold factor was computed as the ratio of normalized tag counts in two SAGE libraries with values < 1 converted to reciprocal negatives. For the tags absent in one library a normalized tag count of single copy tags was assumed.

Data sources

The SAGE library from somatic cells of the adult testis [19] was obtained from GEO repository [20], accession number GSM5435. Other SAGE libraries were obtained from GEO repository or downloaded from Internet sources (see Additional file 4). The data from the BDF1 testis SAGE library were obtained from a printed table in publication [32] (only the top 100 genes expressed in BDF1 testis are listed in publication, the whole library is currently not publicly available). Microarray data of mouse testis, generated by the GNF Mouse Atlas v2 project [33], were obtained from the hgFixed database of the UCSC Genome Browser [55,56].

Hierarchical clustering of mouse SAGE libraries

Thirty-two mouse SAGE libraries constructed from bulk tissues (including normal and diseased) that were publicly available to date (July 1, 2004) were selected (see Additional file 4). For each pair of SAGE libraries a distance based on differences between normalized tag counts was computed [28]. The average agglomeration method was used in hierarchical clustering because of the highest cophenetic correlation (Pearson correlation between the observed distances and the distances calculated from the dendrogram).

Selection and preparation of mouse SAGE libraries for genomic analysis

Twenty-seven SAGE libraries created from bulk tissues (excluding tumors) were organized into 7 groups by tissue type and tag counts from SAGE libraries within each group were combined (see Additional file 4). The groups of SAGE libraries include: brain (9 libraries, 329 745 tags), eye (6 libraries, 336 399 tags), heart (1 library, 84 275 tags), liver (2 libraries, 37 118 tags), kidney (6 libraries, 87 810 tags), limbs (2 libraries, 136 650 tags) and adipose tissue (1 library, 44 974 tags). These groups were analyzed in parallel with total testis (2 libraries, 76 854 tags) and adult testis somatic cells (1 library, 81 478 tags). All tags from prepared tissue groups, total testis and adult testis somatic cells SAGE libraries were identified to UniGene clusters using RNA evidence mapping (tag-to-gene association is supported by at least one mRNA sequence) and linked to LocusLink genes. Only tags with identification to a single LocusLink gene were subjected to further analysis. Tag counts from multiple tags matching the same LocusLink gene were combined.

Distribution of tissue-specific genes on chromosomes

Analysis was done in parallel for testis-specific genes in total testis and somatic cells of adult testis. The tissue-specific genes were selected according to tag counts in the testis tissue and 7 non-testis tissues (see Additional file 4). A gene was considered to be tissue-specific if it was expressed only in one tissue and its expression was supported by tag count > 1. Each tissue-specific gene was then assigned to a chromosome (autosome or X chromosome) according to the LocusLink database and the group (testis or non-testis). The permutations algorithm performed 100 000 random shufflings of the chromosomes while keeping the sum of genes on autosomes and the X chromosome constant. The p-value (two tailed) was computed as doubled number of permutations yielding gene counts above/below (which of this was lower) or equal to the observed gene counts in testis tissue and the X chromosome.

Identification of chromosomal clusters of genes with preferential expression in testis

The preferential expression measure (PEM) [31] was used to score differential expression of genes in testis tissues. PEM for total testis (PEM_{TT}) and adult testis somatic cells (PEM_{ATSC}) were calculated for each gene. The gene was considered to be preferentially expressed in total testis if $PEM_{TT} >= 1/2\ PEM_{TT(max)}$, and in somatic cells of adult testis if $PEM_{ATSC} >= 1/2 PEM_{ATSC(max)}$. $PEM_{(max)}$ values represent the maximum PEM value encountered in the tissue, $PEM_{TT(max)} = 1.169$, $PEM_{ATSC(max)} = 1.145$.

To prepare a tandem duplicate-free mouse genome we considered 19 684 known genes from the LocusLink database that were mapped on the mouse genome assembly (NCBI build 32) [57]. For each LocusLink gene, we obtained a known protein sequence (NP_ accessions) from the mouse RefSeq collection [58] and performed protein BLAST (standard settings) against the RefSeq known protein collection. The hits with expectation value < $1e^{-10}$ and with an alignment of at least 50% length and 30% identity of the query sequence were processed and identified to LocusLink genes. If a LocusLink gene located in the vicinity of the original LocusLink gene was found among the hits (considering 10 adjacent genes in both directions), both genes were considered as a tandem duplicate pair and were excluded from the genome. As a result a tandem duplicate-free genome with 16 858 LocusLink genes was obtained.

Two sets of gene clusters with preferentially expressed genes were identified – for total testis and somatic cells of adult testis. All LocusLink genes from the tandem duplicate-free mouse genome were associated with the expression status (preferentially expressed, expressed, unknown). Each chromosome was searched using a sliding window of three adjacent genes and three consecutive preferentially expressed genes were considered as a cluster (tight clusters). Another search was performed using a sliding window of six adjacent genes and at least three preferentially expressed genes were required to form a cluster spanning from the first to the last preferentially expressed gene (loose clusters). The overlapping clusters were merged into a single cluster encompassing all involved genes (separately for tight or loose clusters). The permutations performed 100 000 random shufflings of the expression status in the genome while keeping the gene positions constant. A search with the above defined sliding windows determined the number of preferentially expressed genes located in clusters in each randomized genome. The p-value (one tailed) was computed as the number of permutations yielding greater than or equal to the observed number of preferentially expressed genes located in clusters.

Statistical evaluation

All statistical analyses, including Monte Carlo simulations, hierarchical clustering, chromosomal and gene permutations were conducted in R statistical environment [59] using custom scripts.

Database versions

The following database versions were used in all analyses: Mouse UniGene build #136 (March 26, 2004), mouse SAGEmap (April 3, 2004) corresponding to the mouse UniGene #136, LocusLink (April 3, 2004), mouse genome assembly NCBI build 32 (November 2003), mouse Reference Sequence collection (April 3, 2004) and Gene Ontology database (July, 2004).

Authors' contributions

PD constructed the SAGE libraries and performed the bioinformatic analyses. CV carried out sequencing of SAGE libraries. PS participated in bioinformatic analysis of gene order. VP is Head of the Center for Integrated Genomics. JF conceived the study and coordinated work. PD and JF wrote the article. All authors read and approved the final manuscript.

Additional material

Additional File 1

***SAGE tags detected in mouse total testis**. List of 7 481 tags with tag count > 1 in combined total testis SAGE library with reliable tag identification according to the Mouse SAGE Site database.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S1.xls]

Additional File 2

***Tags with significantly different tag counts between total testis and adult testis somatic cells**. List of 3 258 tags with significantly different tag counts between the total testis and adult testis somatic cells SAGE libraries determined by Monte Carlo simulations (1 691 tags have increased tag counts in total testis, 1 567 tags have increased tag counts in adult testis somatic cells at p-chance < 0.05). The reliable tag identification according to the Mouse SAGE Site database is provided.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S2.xls]

Additional File 3

***Genes with predominant expression in germinal and somatic cells of the testis**. List of 924 and 802 genes with predominant expression in germinal and somatic cells of the testis, respectively, based on the comparison of total testis and adult testis somatic cells SAGE libraries. Tags with significantly different tag counts (p-chance < 0.05, Monte Carlo simulation) and at least five-fold increased/decreased tag counts were selected and identified using RNA evidence mapping.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S3.xls]

Additional File 4

***Mouse SAGE libraries used in genomic analysis and hierarchical clustering**. List of mouse SAGE libraries publicly available to date July 1, 2004 that were used in genomic analysis and hierarchical clustering.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S4.xls]

Additional File 5

Genes specific for total testis or adult testis somatic cells. *List of 395 and 81 genes with specific expression in total testis or adult testis somatic cells determined by comparison to SAGE data from seven non-testis tissues (brain, eye, heart, liver, kidney, limbs and adipose tissue). A gene was considered to be testis specific if the corresponding tags were present only in total testis or adult testis somatic cells SAGE libraries and missing in all non-testis libraries. Tags were identified using RNA evidence mapping.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S5.xls]

Additional File 6

Testis expressed genes located on the X chromosome (summary). *Summary of the X-linked genes expressed in total testis and adult testis somatic cells. Tags were identified using RNA evidence mapping.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S6.xls]

Additional File 7

Preferentially expressed genes in total testis and testicular somatic cells. *Preferentially expressed genes were determined separately for total testis and adult testis somatic cells in conjunction with their expression in seven non-testis tissues. Expression of genes was scored using preferential expression measure (PEM). A gene was considered to be preferentially expressed if PEM score was above 50 % of the maximum PEM value encountered in that tissue. Tags were identified using RNA evidence mapping.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S7.xls]

Additional File 8

Genes preferentially expressed in testis located in chromosomal clusters within tandem duplicate-free genome. *Chromosomal clusters of genes preferentially expressed in testis were localized by the search with a sliding window. Two types of clusters were identified: tight clusters (containing at least three adjacent preferentially expressed genes in testis) and loose clusters (containing at least three preferentially expressed genes in testis among the six adjacent genes).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S8.xls]

Additional File 9

Manual comparison of the most highly expressed genes in three total testis transcriptomes (GNF atlas, B6 testis, BDF1 testis). *The list of 35 most highly expressed genes in total testis according to the GNF Mouse Atlas v2 organized with the appropriate* NlaI*II and* RsaI *SAGE tags extracted from their representative mRNA/RefSeq sequences.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-29-S9.xls]

Acknowledgements

We thank Laurence D. Hurst, Adam Pavlíček and Jan Pačes for helpful comments and suggestions, Radka Storchová, Zdeněk Trachtulec and Šárka Takáčová for critically reading the manuscript. This work is supported by the project of the Czech Ministry of Education, Youth and Sports No. LN00A079 – Center for Integrated Genomics and by the project of the Academy of Sciences of the Czech Republic No. K5052113. J.F. is supported as an International Scholar of the Howard Hughes Medical Institute.

References

1. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284:**601-603.
2. Orgel LE, Crick FH, Sapienza C: **Selfish DNA.** *Nature* 1980, **288:**645-646.
3. Watson D, Jacombs AS, Loebel DA, Robinson ES, Johnston PG: **Single nucleotide primer extension (SNuPE) analysis of the G6PD gene in somatic cells and oocytes of a kangaroo (Macropus robustus).** *Genet Res* 2000, **75:**269-274.
4. Handel MA, Hunt PA: **Sex-chromosome pairing and activity during mammalian meiosis.** *Bioessays* 1992, **14:**817-822.
5. McCarrey JR, Watson C, Atencio J, Ostermeier GC, Marahrens Y, Jaenisch R, Krawetz SA: **X-chromosome inactivation during spermatogenesis is regulated by an Xist/Tsix-independent mechanism in the mouse.** *Genesis* 2002, **34:**257-266.
6. Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD: **The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation.** *Nat Genet* 2004, **36:**642-646.
7. Lifschytz E, Lindsley DL: **The role of X-chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation).** *Proc Natl Acad Sci U S A* 1972, **69:**182-186.
8. Davis TL, Yang GJ, McCarrey JR, Bartolomei MS: **The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development.** *Hum Mol Genet* 2000, **9:**2885-2894.
9. Liu J, Yu S, Litman D, Chen W, Weinstein LS: **Identification of a methylation imprint mark within the mouse Gnas locus.** *Mol Cell Biol* 2000, **20:**5808-5817.
10. Cooke HJ, Saunders PT: **Mouse models of male infertility.** *Nat Rev Genet* 2002, **3:**790-801.
11. de Rooij DG, de Boer P: **Specific arrests of spermatogenesis in genetically modified and mutant mice.** *Cytogenet Genome Res* 2003, **103:**267-276.
12. Ashley T: **X-Autosome translocations, meiotic synapsis, chromosome evolution and speciation.** *Cytogenet Genome Res* 2002, **96:**33-39.
13. Forejt J: **Hybrid sterility in the mouse.** *Trends Genet* 1996, **12:**412-417.
14. Orr HA, Presgraves DC: **Speciation by postzygotic isolation: forces, genes and molecules.** *Bioessays* 2000, **22:**1085-1094.
15. Storchova R, Gregorova S, Buckiova D, Kyselova V, Divina P, Forejt J: **Genetic analysis of X-linked hybrid sterility in the house mouse.** *Mammalian Genome* 2004, **15:**515-524.
16. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
17. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor

MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.

18. Singer JB, Hill AE, Burrage LC, Olszens KR, Song J, Justice M, O'Brien WE, Conti DV, Witte JS, Lander ES, Nadeau JH: **Genetic dissection of complex traits with chromosome substitution strains of mice.** *Science* 2004, **304:**445-448.
19. O'Shaughnessy PJ, Fleming L, Baker PJ, Jackson G, Johnston H: **Identification of Developmentally-Regulated Genes in the Somatic Cells of the Mouse Testis Using Serial Analysis of Gene Expression.** *Biol Reprod* 2003, **69:**797-808.
20. **NCBI Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
21. **Mouse SAGE Site** [http://mouse.biomed.cas.cz/sage/]
22. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource.** *Genome Res* 2000, **10:**1051-1060.
23. **SAGEmap database** [ftp://ftp.ncbi.nlm.nih.gov/pub/sage/]
24. Divina P, Forejt J: **The Mouse SAGE Site: database of public mouse SAGE libraries.** *Nucleic Acids Res* 2004, **32:**D482-3.
25. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32 (Database issue):**D258-61.
26. **Gene Ontology Project** [http://www.geneontology.org/]
27. Sutcliffe MJ, Darling SM, Burgoyne PS: **Spermatogenesis in XY, XYSxra and XOSxra mice: a quantitative analysis of spermatogenesis throughout puberty.** *Mol Reprod Dev* 1991, **30:**81-89.
28. Baross A, Schertzer M, Zuyderduyn SD, Jones SJ, Marra MA, Lansdorp PM: **Effect of TERT and ATM on gene expression profiles in human fibroblasts.** *Genes Chromosomes Cancer* 2004, **39:**298-310.
29. Lercher MJ, Urrutia AO, Hurst LD: **Evidence that the human X chromosome is enriched for male-specific but not female-specific genes.** *Mol Biol Evol* 2003, **20:**1113-1116.
30. Wang PJ, McCarrey JR, Yang F, Page DC: **An abundance of X-linked genes expressed in spermatogonia.** *Nat Genet* 2001, **27:**422-426.
31. Huminiecki L, Lloyd AT, Wolfe KH: **Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases.** *BMC Genomics* 2003, **4:**31.
32. Yao J, Chiba T, Sakai J, Hirose K, Yamamoto M, Hada A, Kuramoto K, Higuchi K, Mori M: **Mouse testis transcriptome revealed using serial analysis of gene expression.** *Mamm Genome* 2004, **15:**433-451.
33. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:**6062-6067.
34. **GNF Mouse Atlas v2** [http://symatlas.gnf.org/]
35. Pravenec M, Zidek V, Landa V, Simakova M, Mlejnek P, Kazdova L, Bila V, Krenova D, Kren V: **Genetic analysis of "metabolic syndrome" in the spontaneously hypertensive rat.** *Physiol Res* 2004, **53 Suppl 1:**S15-22.
36. Farrall M: **Quantitative genetic variation: a post-modern view.** *Hum Mol Genet* 2004, **13 Spec No 1:**R1-7.
37. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome.** *Nat Biotechnol* 2002, **20:**508-512.
38. Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y: **5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation.** *Proc Natl Acad Sci U S A* 2004, **101:**11701-6.
39. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5:**299-310.
40. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* 2000, **18:**630-634.
41. Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK: **A global profile of germline gene expression in C. elegans.** *Mol Cell* 2000, **6:**605-616.
42. Betran E, Thornton K, Long M: **Retroposed new genes out of the X in Drosophila.** *Genome Res* 2002, **12:**1854-1859.
43. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI: **Large clusters of co-expressed genes in the Drosophila genome.** *Nature* 2002, **420:**666-669.
44. Forejt J: **X-Y involvment in male sterility caused by autosome translocations - a hypothesis.** In *Genetic control of Gamete Production and Function* Edited by: Fraccaro M and Rubin B. , Academic Press, New York; 1982:135-151.
45. Turner JM, Mahadevaiah SK, Elliott DJ, Garchon HJ, Pehrson JR, Jaenisch R, Burgoyne PS: **Meiotic sex chromosome inactivation in male mice with targeted disruptions of Xist.** *J Cell Sci* 2002, **115:**4097-4105.
46. Handel MA: **The XY body: a specialized meiotic chromatin domain.** *Exp Cell Res* 2004, **296:**57-63.
47. Rice WR: **Sex-Chromosomes and the Evolution of Sexual Dimorphism.** *Evolution* 1984, **38:**735-742.
48. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol* 2002, **1:**5.
49. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH: **The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.** *Genome Res* 2003, **13:**1998-2004.
50. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31:**180-183.
51. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291:**1289-1292.
52. Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD: **A unification of mosaic structures in the human genome.** *Hum Mol Genet* 2003, **12:**2411-2415.
53. **SAGE method homepage** [http://www.sagenet.org/]
54. Patefield WM: **An efficient method of generating random RxC tables with given row and column totals.** *Applied Statistics* 1981, **30:**91-97.
55. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.
56. **UCSC Genome Browser expression database** [http://hgdownload.cse.ucsc.edu/goldenPath/hgFixed/database/]
57. **NCBI mouse genome assembly** [ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/maps/mapview/]
58. **NCBI mouse Reference Sequences** [ftp://ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/]
59. **The R Project for Statistical Computing** [http://www.r-project.org/]

J Mol Evol (2006) 63:676–681
DOI: 10.1007/s00239-006-0022-1

JOURNAL OF MOLECULAR EVOLUTION

Nonrandom Representation of Sex-Biased Genes on Chicken Z Chromosome

R. Storchová, P. Divina

Institute of Molecular Genetics, Academy of Sciences of the Czech Republic and Center for Applied Genomics, Vídeňská 1083, CZ-142 20, Prague 4, Czech Republic

Received: 30 January 2006 / Accepted: 25 July 2006 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. Several lines of evidence suggest that the X chromosome of various animal species has an unusual complement of genes with sex-biased or sex-specific expression. However, the study of the X chromosome gene content in different organisms provided conflicting results. The most striking contrast concerns the male-biased genes, which were reported to be almost depleted from the X chromosome in *Drosophila* but overrepresented on the X chromosome in mammals. To elucidate the reason for these discrepancies, we analysed the gene content of the Z chromosome in chicken. Our analysis of the publicly available expressed sequence tags (EST) data and genome draft sequence revealed a significant underrepresentation of ovary-specific genes on the chicken Z chromosome. For the brain-expressed genes, we found a significant enrichment of male-biased genes but an indication of underrepresentation of female-biased genes on the Z chromosome. This is the first report on the nonrandom gene content in a homogametic sex chromosome of a species with heterogametic female individuals. Further comparison of gene contents of the independently evolved X and Z sex chromosomes may offer new insight into the evolutionary processes leading to the nonrandom genomic distribution of sex-biased and sex-specific genes.

Key words: Comparative genomics — Chicken — Dosage compensation — Evolution — Gene expression — Sex chromosomes — Z chromosome

Correspondence to: R. Storchová; *email:* radkas@biomed.cas.cz

Introduction

Several studies have shown that the X chromosome, a homogametic sex chromosome in male heterogametic organisms, differs from autosomes by a nonrandom content of genes with sex-biased or sex-specific expression (the genes expressed preferentially or exclusively in one sex). However, the direction of the biases in the location of sex-biased and sex-specific genes is not consistent across species. Thus, in humans, the genes related to sex and reproduction as well as the genes connected with brain and muscle functions were enriched on the X chromosome relative to autosomes (Bortoluzzi et al. 1998; Hurst and Randerson 1999; Lercher et al. 2003; Saifi & Chandra 1999; Zechner et al. 2001). In mice, the genes preferentially expressed in ovary, placenta, testicular somatic cells, and premeiotic germinal cells were more abundant on the X chromosome (Divina et al. 2005; Khil et al. 2004; Wang et al. 2001), whereas the genes expressed in the male germ line during meiosis were underrepresented (Divina et al. 2005; Khil et al. 2004). In *Caenorhabditis elegans*, the genes expressed in spermatogenic and oogenic cells were underrepresented on the X chromosome (Reinke et al. 2004), and in *Drosophila*, the male-biased genes were nearly absent from the X chromosome regardless of whether their expression was preponderant in germinal or in somatic tissues (Parisi et al. 2003; Ranz et al. 2003).

Two main hypotheses exist concerning the possible mechanisms causing the nonrandom representation of sex-biased and sex-specific genes on the homogametic sex chromosome. According to the hypothesis of sexual antagonism (Hurst 2001; Rice 1984), an unusual homogametic sex chromosome gene content reflects a

nonrandom accumulation of sexually antagonistic mutations (those favouring one sex, although being detrimental to the other) on this chromosome. This is caused by the different time that the homogametic sex chromosome has spent in the two sexes and by its hemizygous exposure in the heterogametic sex. The other hypothesis concerns the epigenetic modifications of the sex chromosomes associated with meiotic sex chromosome inactivation and dosage compensation (Khil et al. 2005; Parisi et al. 2003; Reinke et al. 2004; Rogers et al. 2003). Although the effect of the meiotic sex chromosome inactivation on the homogametic sex chromosome gene content has been well documented (Betran et al. 2002; Divina et al. 2005; Emerson et al. 2004; Khil et al. 2004; Reinke 2004), the role of dosage compensation and sexual antagonism remains elusive.

The mechanisms responsible for the nonrandom representation of sex-biased and sex-specific genes on the homogametic sex chromosome may be clarified by analysing the gene content of the Z chromosome, a homogametic sex chromosome in heterogametic female organisms. Although the X chromosome occurs more frequently in female individuals, the Z chromosome spends more time in male individuals. If sexually antagonistic selection were the primary mechanism affecting the sex chromosome gene content, we would expect the opposite trend in the representation of sex-biased and sex-specific genes on the X and Z chromosomes. Other evolutionary processes also shape the gene content of the X and Z chromosomes in slightly different ways. For example, because the Z chromosome occurs more frequently than the X chromosome in male individuals, it is exposed to a higher mutation rate, which provides more material for selection to act on (Axelsson et al. 2004; Ellegren & Fridolfsson 1997; Kirkpatrick & Hall 2004; Montell et al. 2001). The Z chromosome has also been suggested to be more responsive to sexual selection than the X chromosome (Reeve and Pfennig 2003). Therefore, the Z chromosome would be expected to show more profound differences in gene composition, relative to autosomes, than the X chromosome. Here we present the first analysis of the gene content of the Z chromosome in chicken. We show that chicken Z chromosome gene content is characterized by underrepresentation of ovary-specific genes and, to a lesser extent, of the brain-expressed female-biased genes, whereas the brain-expressed male-biased genes are significantly overrepresented on the Z chromosome.

Materials and Methods

EST Data

We used publicly available chicken EST data from NCBI UniGene database (build no. 24, October 14, 2004) (Wheeler et al. 2005). To analyse the distribution of tissue-specific genes on autosomes and the Z chromosome, we selected 68 chicken EST libraries containing at least 500 ESTs that were prepared from bulk tissues and were not annotated as diseased or embryonic. These libraries were sorted into 14 groups representing different tissue types (Supplementary Table 1). To get enough data for the analysis, we used both nonnormalized and normalized EST libraries (each tissue type was represented by at least one nonnormalized EST library). In the normalized libraries, the quantitative information about gene expression is biased, but the qualitative information about gene expression, or at least the distribution of genes among the chromosomes, should be preserved, and this was a sufficient condition for our analysis. Indeed, the proportion of Z-linked genes in the nonnormalized and normalized libraries did not show any significant difference for each of the examined tissues ($p > 0.05$, Fisher's Exact test). The tissue-specific genes were defined as the genes present in the libraries from one tissue type but not the others. For subsequent analysis of the distribution of male- and female-biased genes on autosomes and the Z chromosome, we used three nonnormalized EST libraries prepared from male brain (total 4230 ESTs) and female brain (total 8399 ESTs). The corresponding EST library IDs were 16171, 15560, and 15561.

Table 1. The proportion of Z-linked tissue-specific genes in 14 different tissues

Tissue	ChrA	ChrZ	% Z-linked
Brain	1135	40	3.40
Limbs	446	18	3.88
Chondrocytes	527	28	5.05
Heart	130	3	2.26
Kidney and adrenal	280	6	2.10
Small intestine	235	6	2.49
Liver	153	7	4.38
Pancreas	36	2	5.26
Muscle	156	8	4.88
Fat	42	1	2.33
Bursal lymphocytes	134	7	4.96
Spleen	42	2	4.55
All somatic	3316	128	3.72
Testis	363	16	4.22
Ovary	620	11	1.74

ChrA, autosomes; ChrZ, Z chromosome.

Chromosomal Location

For the purpose of our analyses, each UniGene cluster represented a "gene." To determine the chromosomal location we aligned the representative sequence for each UniGene cluster to the chicken draft genome assembly (galGal2, February 2004, University of California Santa Cruz [UCSC] Genome Browser) (International Chicken Genome Sequencing Consortium 2004; Karolchik et al. 2003) using BLAT (Kent Informatics, Santa Cruz, CA, USA) (Kent 2002). BLAT was run with parameters used to build the UniGene track of the UCSC Genome Browser (minimum sequence identity of 95%, at least 20% coverage of the query sequence, at least 96.5% alignment ratio, and scores within 0.2% of the best-in-genome). The hits to multiple locations in the genome were discarded as were the hits to the W chromosome. Using these criteria, we mapped 79% (16,795 of the 21,447) UniGene clusters to unique positions in the genome.

Statistics

To control for the effects of tissue specificity (Lercher et al. 2003), we compared the proportions of Z-linked genes specific for testis

Table 2. The proportion of the tissue-specific genes expressed in testis and ovary on the Z chromosome and comparison with tissue-specific genes expressed in somatic tissues

Tested tissue	Observed gene counts		Expected gene counts		% Z-linked (observed)	p value[a] (two tailed)
	Tested tissue	Somatic tissues	Tested tissue	Somatic tissues		
Testis						
ChrA	363	2676	365.5	2673.5		
ChrZ	16	96	13.5	98.5	4.22	0.537
Ovary						
ChrA	620	2676	611.2	2684.8		
ChrZ	11	96	19.8	87.2	1.74	0.027

[a]The p value (two tailed) corresponds to the proportion of permutations producing the gene counts greater or equal (for testis) and lower or equal (for ovary) to the observed gene counts in the tested tissue and the Z chromosome multiplied by two.
ChrA, autosomes; ChrZ, Z chromosome.

and ovary with the proportions of Z-linked tissue-specific genes in the pool of 12 somatic tissues by permutation test. All genes were randomly reassigned to chromosomes while keeping fixed the total gene count, the total count of genes in each group, and the total number of genes on autosomes and the Z chromosome. To assess significance, fractions of permutations producing the gene counts "lower or equal" and "greater or equal" to the observed gene counts in the tested tissue and the Z chromosome were determined. The p value (two tailed) was defined as the smaller of these fractions multiplied by two.

The genes with preferential expression in male or in female brain were sorted out using the R statistic, which was devised previously for comparison of transcript abundances in cDNA libraries (Stekel et al. 2000). The R statistic was computed for each gene expressed in the brain, and the genes exceeding a given R threshold were considered as preferentially expressed in male or female brain. Fisher's Exact test was used to compare the proportions of male- and female-biased genes located on the Z chromosome.

Results

First, we assessed the allocation of the tissue-specific genes (the genes expressed exclusively in one tissue) between autosomes and the Z chromosome. For this purpose, we compared the proportions of the Z-linked tissue-specific genes in 14 different tissues (12 somatic tissues, testis and ovary; Table 1). Because these proportions did not differ significantly among the 12 somatic tissues ($p > 0.05$, Fisher's Exact test), we combined the EST data of all somatic tissues into one pool. Then we analysed whether the proportions of the Z-linked tissue-specific genes in testis (male specific) and ovary (female specific) differ from those in somatic tissues (Table 2). Of the 379 testis-specific genes, 4.2% mapped to the Z chromosome, a proportion comparable with 3.5% of the Z-linked tissue-specific genes present in the pool of 12 somatic tissues ($p > 0.05$, two tailed, 100,000 permutations). In ovary, 631 tissue-specific genes were detected, but only 11 of them (1.7%) mapped to the Z chromosome, which is a significant decrease compared with the Z-linked tissue-specific genes in the pool of 12 somatic tissues ($p = 0.027$, two tailed, 100,000 permutations). To confirm that the paucity of ovary-specific genes concerns only the Z chromosome, we further examined the distribution of ovary-specific genes among individual autosomes. Although the proportion of ovary-specific genes on the autosomes containing at least 500 genes (Chr: 1 to 10 and 14) was uneven (2.7% to 5.9%), it was in all cases higher than on the Z chromosome, which contained only 1.7% ovary-specific genes of the total number of 653 Z-linked genes. This represents a significant decrease ($p < 0.01$, Chi-square test) compared with the average proportion of ovary-specific genes on autosomes (3.8% or 620 of 16142). A list of the 11 Z-linked ovary-specific genes obtained in this analysis is provided in Supplementary Table 2.

The important question for interpreting our results is whether the nonrandom distribution concerns only the genes expressed in ovary, composed of both somatic and germinal cells, or also the genes expressed in pure somatic tissues. To answer this question, we performed an analysis of the sex-biased genes expressed in brain, for which the nonnormalized EST libraries were created separately from male and female individuals. The genes preferentially expressed in either male or female brain (male biased or female biased) were sorted out using different thresholds of the R statistic (Stekel et al. 2000). When the preferentially expressed genes with $R > 1$ were selected, 5.6% (58 of 1029) of male-biased genes were located on the Z chromosome. In comparison, only 1.3% (4 of 314) of female-biased genes were found on the Z chromosome, representing a highly significant difference ($p < 0.001$, Fisher's Exact test). For more relaxed thresholds ($R > 0.5$ and $R > 0$), the difference in proportion of male-biased and female-biased genes on the Z chromosome was smaller but still highly significant (Fig. 1).

We were also interested whether the male- or female-biased genes expressed in the brain were enriched or impoverished on the Z chromosome. For that reason we compared the proportions of the male-biased and female-biased Z-linked genes to the overall proportion of Z-linked genes expressed in

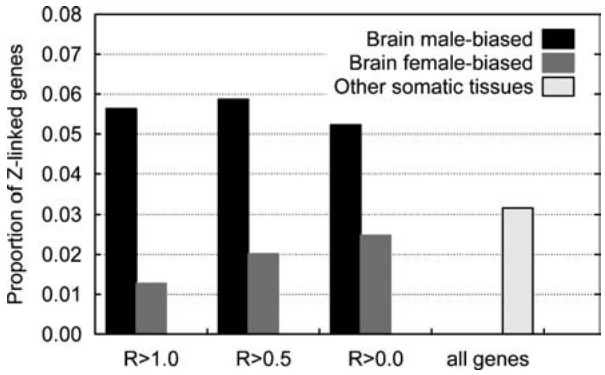


Fig. 1. The proportions of Z-linked genes with male- and female-biased expression in brain for different thresholds of the R statistic. The R statistic (Stekel et al. 2000) was used as a measure of sex-biased expression. For all thresholds of the R statistic, the male-biased genes were significantly more abundant on the Z chromosome than the female-biased genes. Compared with the genes expressed in other somatic tissues, the male-biased genes were significantly enriched on the Z chromosome, whereas the female-biased genes displayed an indication of underrepresentation on the Z chromosome.

the pool of 11 other somatic tissues. The male-biased genes expressed in the brain were 1.8-fold more abundant on the Z chromosome than the genes expressed in other somatic tissues, which is a highly significant difference ($p < 0.0001$, Fisher's Exact test). Significant overrepresentation (1.6-fold) was also observed for more stringent threshold, $R > 2$ ($p < 0.05$, Fisher's Exact test). Using this threshold, we sorted out 421 male-biased genes from which 22 mapped to the Z chromosome. In contrast, the female-biased genes expressed in the brain were 2.5-fold less abundant on the Z chromosome than the genes expressed in other somatic tissues. However, this difference was not significant at the 5% level because of the low number of female-biased genes ($p = 0.052$, Fisher's Exact test) (Fig. 1). A list of the Z-linked genes preferentially expressed in male brain (58 genes) and female brain (4 genes) is available in Supplementary Table 3. Admittedly, the majority of these genes are annotated as unknown transcripts.

Discussion

Previous studies have shown that the X chromosome of various animal species harbours nonrandom proportions of genes with sex-biased or sex-specific expression. However, selective forces responsible for this phenomenon remain mostly elusive. Our results indicate that nonrandom proportions of sex-biased and sex-specific genes also characterise the Z chromosome in chicken. Comparing the gene contents of the independently evolved X and Z chromosomes may help decide which mechanisms are mostly responsible for the nonrandom genomic distribution of sex-biased and sex-specific genes. These mechanisms could involve sexually antagonistic selection and/or epigenetic modifications of the X/Z chromatin such as meiotic sex chromosome inactivation and dosage compensation.

According to the hypothesis of sexual antagonism (Hurst 2001; Rice 1984), dominant mutations favouring the homogametic sex but not the heterogametic sex should accumulate on the X/Z chromosome because this chromosome spends two thirds of its time in the homogametic sex but only one third of its time in the heterogametic sex. Indeed, it was observed that the female-specific and/or female-biased genes are enriched on the X chromosome in mice and *Drosophila* (Khil et al. 2004; Parisi et al. 2003; Ranz et al. 2003), although this was not confirmed in humans (Lercher et al. 2003). In chicken, we have found a significant overrepresentation of the genes expressed preferentially in the male brain on the Z chromosome, which is in agreement with the theory of sexually antagonistic selection. The question still remains why the testis-specific genes are not enriched on the chicken Z chromosome as well.

The role of sexually antagonistic selection in the distribution of genes favouring the heterogametic sex is more complicated because it depends on the proportion of dominant and recessive mutations that emerge in the population (or on the average dominance of new mutations) (Rogers et al. 2003). If the majority of mutations are dominant, we should expect underrepresentation of genes favouring the heterogametic sex on the X/Z chromosome because it spends only one third of its time in the heterogametic sex. However, if the majority of mutations are recessive, the reverse effect should be expected. The reason is that the recessive mutations favouring the heterogametic sex have a greater chance to be fixed on the hemizygous X/Z chromosome where they are exposed to selection. Studies on the X/Z chromosome gene content in different organisms provided conflicting results. The genes that are preferentially expressed in the male somatic tissues are enriched on the X chromosome in mammals (Divina et al. 2005; Khil et al. 2004; Lercher et al. 2003; Wang et al. 2001). In contrast, the male-biased genes in *Drosophila* (Parisi et al. 2003; Ranz et al. 2003) and the female-biased genes in chicken seem to be underrepresented on the X/Z chromosome. Assuming that the proportion of dominant and recessive mutations does not differ among taxa, the sexually antagonistic selection is unlikely to explain the discrepancies in the X/Z chromosome gene content in different species.

Another mechanism affecting sex chromosome gene content concerns the epigenetic modifications of the X/Z chromatin, such as meiotic sex chromosome inactivation and dosage compensation. The meiotic sex chromosome inactivation occurs in germinal cells in heterogametic male individuals, but it has not been

observed in heterogametic female individuals (Jablonka & Lamb 1990) and hence is unlikely to affect the complement of genes on the Z chromosome. In contrast, the dosage compensation occurs in somatic cells and has been observed in both heterogametic male and female organisms. Different species, however, use different mechanisms to achieve dosage compensation (Avner & Heard 2001; Baker et al. 1994; Ellegren 2002; Gupta et al. 2006; Kelley 2004; Meyer 2000). Interestingly, recent findings suggest that dosage compensation in chicken may be achieved by the same mechanism as in *Drosophila*, i.e., by transcriptional upregulation of the single X/Z chromosome in heterogametic sex (Bisoni et al. 2005). According to the hypothesis of Pomiankowski et al. (Rogers et al. 2003), this mechanism of dosage compensation could lead to the paucity of genes upregulated in the heterogametic sex on the X/Z chromosome. The reason is that the overall overactivation of the single X/Z chromosome could constrain further increase of transcription in the heterogametic sex. If dosage compensation is really achieved by the same mechanism in *Drosophila* and birds, our data could support this hypothesis.

Acknowledgments. We thank J. Forejt for continuing support and comments on our work; J. Hejnar, D. Storch, and Z. Trachtulec for valuable suggestions; Š. Takáčová for reading the manuscript; and two anonymous reviewers for insightful comments. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic Grant No. 1M6837805002, Center for Applied Genomics, and by the Grant Agency of the Academy of Sciences Grant No. AVOZ50520514.

References

Avner P, Heard E (2001) X-chromosome inactivation: Counting, choice and initiation. Nat Rev Genet 2:59–67

Axelsson E, Smith NG, Sundstrom H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and turkey. Mol Biol Evol 21:1538–1547

Baker BS, Gorman M, Marin I (1994) Dosage compensation in Drosophila. Annu Rev Genet 28:491–521

Betran E, Thornton K, Long M (2002) Retroposed new genes of the X in Drosophila. Genome Res 12:1854–1859

Bisoni L, Batlle-Morera L, Bird AP, Suzuki M, McQueen HA (2005) Female-specific hyperacetylation of histone H4 in the chicken Z chromosome. Chromosome Res 13:205–214

Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, d'Alessi F, Tiso N, Pallavicini A, Toppo S, Cannata N, Valle G, Lanfranchi G, Danieli GA (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res 8:817–825

Divina P, Vlcek C, Strnad P, Paces V, Forejt J (2005) Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: Evidence for nonrandom gene order. BMC Genomics 6:29

Ellegren H (2002) Dosage compensation: Do birds do it as well? Trends Genet 18:25–28

Ellegren H, Fridolfsson AK (1997) Male-driven evolution of DNA sequences in birds. Nat Genet 17:182–184

Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. Science 303:537–540

Gupta V, Parisi M, Sturgill D, Nuttall R, Doctolero M, Dudko OK, Malley JD, Eastman PS, Oliver B (2006) Global analysis of X-chromosome dosage compensation. J Biol 5:3

Hurst LD (2001) Evolutionary genomics. Sex and the X. Nature 411:149–150

Hurst LD, Randerson JP (1999) An eXceptional chromosome. Trends Genet 15:383–385

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716

Jablonka E, Lamb MJ (1990) The evolution of heteromorphic sex chromosomes. Biol Rev Camb Philos Soc 65:249–276

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31:51–54

Kelley RL (2004) Path to equality strewn with roX. Dev Biol 269:18–25

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Khil PP, Oliver B, Camerini-Otero RD (2005) X for intersection: Retrotransposition both on and off the X chromosome is more frequent. Trends Genet 21:3–7

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD (2004) The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet 36:642–646

Kirkpatrick M, Hall DW (2004) Male-biased mutation, sex linkage, and the rate of adaptive evolution. Evolution Int J Org Evolution 58:437–40

Lercher MJ, Urrutia AO, Hurst LD (2003) Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. Mol Biol Evol 20:1113–1116

Meyer BJ (2000) Sex in the worm-counting and -compensating X-chromosome dose. Trends Genet 16:247–253

Montell H, Fridolfsson AK, Ellegren H (2001) Contrasting levels of nucleotide diversity on the avian Z and W sex chromosomes. Mol Biol Evol 18:2010–2016

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B (2003) Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science 299:697–700

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003) Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300:1742–1745

Reeve HK, Pfennig DW (2003) Genetic biases for showy males: Are some genetic systems especially conducive to sexual selection? Proc Natl Acad Sci U S A 100:1089–1094

Reinke V (2004) Sex and the genome. Nat Genet 36:548–549

Reinke V, Gil IS, Ward S, Kazmer K (2004) Genome-wide germline–enriched and sex-biased expression profiles in Caenorhabditis elegans. Development 131:311–323

Rice WR (1984) Sex-chromosomes and the evolution of sexual dimorphism. Evolution 38:735–742

Rogers DW, Carr M, Pomiankowski A (2003) Male genes: X-pelled or X-cluded? Bioessays 25:739–741

Saifi GM, Chandra HS (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. Proc Biol Sci 266:203–209

Stekel DJ, Git Y, Falciani F (2000) The comparison of gene expression from multiple cDNA libraries. Genome Res 10:2055–2061

Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. Nat Genet 27:422–426

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33:39–45

Zechner U, Wilda M, Kehrer-Sawatzki H, Vogel W, Fundele R, Hameister H (2001) A high density of X-linked genes for general cognitive ability: A runaway process shaping human evolution? Trends Genet 17:697–701