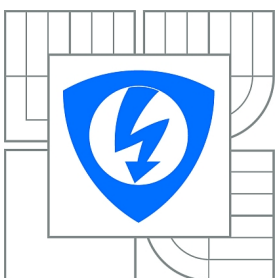


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH
TECHNOLOGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

ANALÝZA DAT ZE SEKVENOVÁNÍ PŘÍŠTÍ GENERACE KE STUDIU AKTIVITY TRANSPOSONŮ V NÁDOROVÝCH BUŇKÁCH

ANALYSIS OF NGS DATA FOR STUDY OF TRANSPOSON ACTIVITY IN CANCER CELLS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

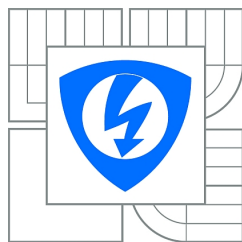
AUTOR PRÁCE
AUTHOR

Bc. IVANA HRAZDILOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. Kejnovský Eduard CSc.

BRNO 2013



VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Studentka: Bc. Ivana Hrazdilová

ID: 115095

Ročník: 2

Akademický rok: 2012/2013

NÁZEV TÉMATU:

Analýza dat ze sekvenování příští generace ke studiu aktivity transposonů v nádorových buňkách

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši zabývající se systematikou lidských transposonů a jejich aktivitou se zvláštním zaměřením na aktivaci transposonů v souvislosti s lidskými chorobami, zejména nádorovými onemocněními. 2) Z veřejně přístupných databází i experimentálních dat z pracoviště vedoucího získajte expresní data pocházející z nádorových i normálních buněk. 3) Pomocí dostupných programů namapujte data z NGS (next generation sequencing) na referenční sekvenci lidského genomu. 4) Analýzou expresních dat určete transkripční aktivitu jednotlivých rodin transposonů. Zvolte vhodnou formu vizualizace výsledků provedených analýz. 5) Proveďte diskusi získaných výsledků a shrňte hlavní závěry práce ohledně aktivity transposonů v nádorových buňkách a jejich případné využití pro diagnostiku pacientů či dokonce potenciální terapii.

DOPORUČENÁ LITERATURA:

- [1] LEE, E., ISKOW, R., YANG, L., GOKCUMEN, O. et al. Landscape of somatic retrotransposition in human cancers. *Science* 337, 967-971, 2012.
- [2] BAILIE, J.K., BARNETT, M.W., UPTÁŇ, K.R. et al. Somatic retrotransposition alter the genetic landscape of the human brain. *Nature* 479, 534-537, 2011.

Termín zadání: 11.2.2013

Termín odevzdání: 24.5.2013

Vedoucí práce: doc. RNDr. Kejnovský Eduard CSc.

Konzultanti diplomové práce: prof. Ing. Ivo Provazník, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

Abstrakt

Teoretická část diplomové práce poskytuje stručnou charakteristiku lidských mobilních genetických elementů (transposonů), které tvoří přibližně 50% lidského genomu a jsou schopny “skákat” z místa na místo. Jsou zde popsány základní rozdělení a typy transposonů přítomné v lidském genomu, mechanismy jejich šíření, aktivace a umlčování. Práce se také věnuje tzv. domestikaci transposonů, popisuje způsoby jakými TE přispívají k poškození DNA a shrnuje nemoci způsobené mutagenní aktivitou transposonů v lidském genomu. Závěr teoretické části je věnován technologiím sekvenace příští generace (NGS). V praktické části byla analyzována data z RNA-seq experimentu, pomocí kterých byla srovnána aktivita transposonů v normálních a rakovinných buňkách prostaty a tlustého střeva. K analýze byly použity jak veřejně dostupné sofistikované nástroje (TopHat), tak vlastní skripty. Výsledky dokazují, že u rakovinných buněk dochází ke zvýšené expresi transposonů, což koresponduje s publikovanými výsledky a naznačuje souvislost aktivity transposonů se vznikem rakoviny.

Klíčová slova

transposony, mobilní genetické elementy, LINE, L1, SINE, Alu, sekvenace nové generace, NGS, masivní paralelní sekvenování, RNA-seq, rakovina, single-end, paired-end, TopHat, Bowtie

Abstract

Theoretical part of this diploma thesis gives a brief characteristic of human mobile elements (transposons), which represents nearly 50% of human genome. It provides basic transposon classification and describes types of transposons present in human genome, as well as mobilization, activation and regulation mechanisms. The work also deals with the domestication of transposons, describes the ways in which TE contribute to DNA damage and summarizes the diseases caused by mutagenic activity of transposons in the human genome. Conclusion of theoretical part describes next-generation sequencing technologies (NGS). As practical part, data from RNA-seq experiment were analyzed in order to compare different transposon activity in normal and cancer cells from prostate and colorectal tissues. As well as publicly available sophisticated tools (TopHat), new scripts were created to analyze these data. The results show that cancer cells exhibit overexpression of transposons. This corresponds with the published results and suggests a connection of transposon activation with cancer development.

Keywords

transposons, mobile genetic elements, LINE, L1, SINE, Alu, next generation sequencing, NGS, massively parallel signature sequencing, RNA-seq, cancer, single-end, paired-end, TopHat, Bowtie

HRAZDILOVÁ, I. *Analýza dat ze sekvenování příští generace ke studiu aktivity transposonů v nádorových buňkách: diplomová práce*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 79 s, 3 příl. Vedoucí diplomové práce doc. RNDr. Kejnovský Eduard CSc.

PROHLÁŠENÍ

Prohlašuji, že svoji diplomovou práci na téma Analýza dat ze sekvenování příští generace ke studiu aktivity transposonů v nádorových buňkách jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne 24. května 2013

Bc. Ivana Hrazdilová

PODĚKOVÁNÍ

Děkuji vedoucímu diplomové práce doc. RNDr. Eduardu Kejnovskému CSc. a prof. Ing. Ivo Provazníkovi, Ph.D., za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne 24. května 2013

Bc. Ivana Hrazdilová

Obsah

1	ÚVOD	8
2	Mobilní genetické elementy	9
2.1	Rozdělení TE	9
2.2	Mechanismy šíření TE	10
2.2.1	DNA transposony	11
2.2.2	LTR retrotransposony	12
2.2.3	Non-LTR transposony	13
2.2.4	Helitrony	13
2.3	TE v lidském genomu	14
2.3.1	LINE elementy	15
2.3.2	SINE elementy	16
2.4	Distribuce TE v lidském genomu	17
2.5	Aktivace TE	19
2.6	Umlčování TE	20
2.7	Poškození DNA způsobené TE	22
2.7.1	Inzerční mutagenese	22
2.7.2	Ektopická rekombinace	22
2.7.3	Dvouřetězcové zlomy	24
2.8	Domestikace TE	24
3	Nemoci způsobené aktivitou TE	26
3.1	Inzerční mutagenese TE	26
3.2	Post-inzerční mutagenese TE	27
4	Sekvenační metody nové generace	29
4.1	Sangerova metoda	29
4.2	Masivně paralelní sekvenování	29
4.3	Platformy NGS	30
4.3.1	Roche/ 454 GS-FLX	30
4.3.2	Illumina/Genome Analyzer	31
4.3.3	Applied Biosystems/ SOLiD	32
4.3.4	Life Technologies/ Ion Torrent	34
5	Sekvenace transkriptomu pomocí RNA-seq	35
6	Příprava dat pro namapování	37
6.1	FASTQ formát	37
6.1.1	Phred skóre	38
6.2	Single-end, paired-end a mate-pair	39
6.3	Získání dat	40
6.4	Zhodnocení kvality dat	40

6.4.1	Odstranění adaptorů	41
6.4.2	Odstranění nekvalitních konců	42
6.4.3	Odstranění duplikací	44
7	Namapování readů	46
7.1	Přehled metod pro namapování readů	46
7.2	Burrowsova Wheelerova transformace (BWT)	46
7.3	Strategie pro multi-ready	48
7.4	SAM formát	49
7.5	Nastavení a implementace TopHat	52
8	Analýza aktivity transposonů	55
8.1	Normalizace RPKM a FPKM	55
8.2	Úprava anotace	57
8.3	Postup pro single-end knihovnu	58
8.4	Postup pro paired-end knihovnu	59
9	Zhodnocení výsledků	63
9.1	Rozdíly v aktivitě hlavních tříd TE	63
9.2	Rozdíly v aktivitě rodin TE	65
9.3	Diskuze o použití RPKM	65
9.4	Diskuze o použití TopHat	68
10	ZÁVĚR	69
	PŘÍLOHY	80
	A Délky transposonů v anotaci	1
	B Rozdělení readů	6
	C Rozdíly v expresi rodin lidských transposonů	12

1 ÚVOD

Mobilní genetické elementy, neboli transposony, jsou sekvence DNA schopné se v genomu přemisťovat z místa na místo. Ačkoliv obecně přijímaný názor je, že mobilní genetické elementy jsou významnou hybnou silou evoluce a hrají důležitou roli v mnoha biologických procesech, dříve se řadily spolu s dalšími sekvencemi do tzv. “junk DNA”, tedy DNA pro kterou se zatím nenašla žádná funkce. Označení transposonů jako “junk DNA” vyplývá už z jejich distribuce v genomu. Tato distribuce není náhodná, ale je zvýšená v heterochromatinu a v centromerových oblastech. Přesněji řečeno, různé třídy transposonů obývají různé oblasti v genomu. U savců se transposony ze třídy SINE přednostně nacházejí na GC-bohaté regiony (bohaté na geny), zatímco LINE transposony mají tendenci se hromadit v AT-bohatých oblastech (chudých na geny).

To že geny mohou “skákat” z místa na místo a replikovat se, bylo v rozporu s obecně přijímaným modelem statického genomu. Pohled na genom jako na dynamický systém nezměnil ani průlomový objev transposonů u kukuřice Barbary McClintockové v roce 1948. Její práce byla oceněna Nobelovou cenou až zpětně roku 1983.

Výsledky Human Genome Project v roce 2001 odhalily, že lidský genom je téměř z 50% tvořen mobilními genetickými elementy. Ačkoliv většinou jde o nefunkční fragmenty poškozené silnou akumulací mutací, přibližně 0.05% transposonů zůstává aktivní [53]. Jejich mutagenní aktivita je spojena se vznikem mnoha geneticky podmíněných onemocnění. Kromě inzerční a post-inzerční mutageneze se spekuluje o spojení aktivity transposonů a vznikem nádorových onemocnění. Rakovina představuje problém komplexních genomických změn, ne mutací jednoho genu. Schopnost transposonů indukovat velké změny ve struktuře, velikosti a organizaci genomu je proto přivádí do popředí studia rakoviny.

Mnoho aspektů jejich fungování nicméně zůstává stále neobjasněno. Příchod masivního paralelního sekvenování však v posledních letech způsobil nárůst zájmu o oblast mobilních genetických elementů, protože umožňuje obdržet věrná kvantitativní data rychleji a s nižšími náklady, než tomu je u klasické Sangerovy metody. Jelikož jsou transposony zdrojem genetické rozmanitosti, jejich studium by nám mohlo pomoci lépe pochopit evoluci, stejně jako pomoci porozumět vzniku rakoviny a jiných nemocí.

2 Mobilní genetické elementy

V minulosti se odhadovalo, že lidská DNA obsahuje v rozmezí 50,000-100,000 genů. Jedním z překvapení Human Genome Project byl revidovaný odhad asi 30,000 genů, což tvoří pouze 1.5% z celkového množství (uvažujeme pouze exony). Jak je toto číslo významné vynikne při srovnání s genomem *C.elegans*, ve kterém bylo objeveno přibližně 20,000 genů [3]. Tyto dva odhady jsou si velmi blízké, navzdory skutečnosti, jak moc jsou tyto dva organismy od sebe evolučně vzdáleny. Obecně se ukázalo, že rozdíl mezi počty genů spojených s diverzifikací obratlovců není zase tak závažný jak se původně předpokládalo. Zajímavý je také nesoulad mezi velikostí genomů a komplexností organismu, nazývaný “paradox hodnoty C” [19], který říká, že neexistuje korelace mezi velikostí genomu a komplexností organismu. Ačkoli je počet genů u obou organismů relativně podobný, velikost lidského genomu téměř 33-krát přesahuje velikost genomu *C.elegans*. Co tedy stojí za tak velkým rozdílem ve velikosti genomů?

Dnes již víme, že hlavní složkou převládající v lidském i mnoha jiných genomech jsou repetitivní sekvence DNA, které mohou mít dvě hlavní podoby. Buď mohou být uspořádány jedna za druhou do podoby tandemových repetit, které se také označují satelitní DNA, a nebo se mohou vyskytovat rozptýleně. Rozptýlené repetice označujeme jako mobilní genetické elementy, neboli transposony (TE, z angl. Transposable Elements). Transposony jsou repetitivní a mobilní sekvence DNA, které jsou schopné přemísťovat se (“skákat”) v genomu z místa na místo.

Transposony se mohou šířit jak vertikálním přenosem, tedy z rodičů na potomky, tak i horizontálně mezi druhy, jako retroviry.

2.1 Rozdělení TE

Podle způsobu jejich transposice rozdělujeme transposony do dvou hlavních skupin [1]. První skupina se přemísťuje prostřednictvím molekuly DNA mechanismem vyjmi a vlož (z angl. ‘cut and paste’) a její členy nazýváme DNA transposony (třída II). Elementy spadající do druhé skupiny označujeme jako retroelementy (třída I). Retroelementy využívají molekuly mRNA ke své transposici a šíří se mechanismem zkopíruj a vlož (z angl. ‘copy and paste’). Retroelementy se dále dělí na retrotransposony obsahující dlouhé koncové repetice LTR (z angl. Long Terminal Repeat) a retroposony co LTR neobsahují, tedy non-LTR. K retroelementům se řadí i virové elementy, především retro-

viry a jim podobné elementy.

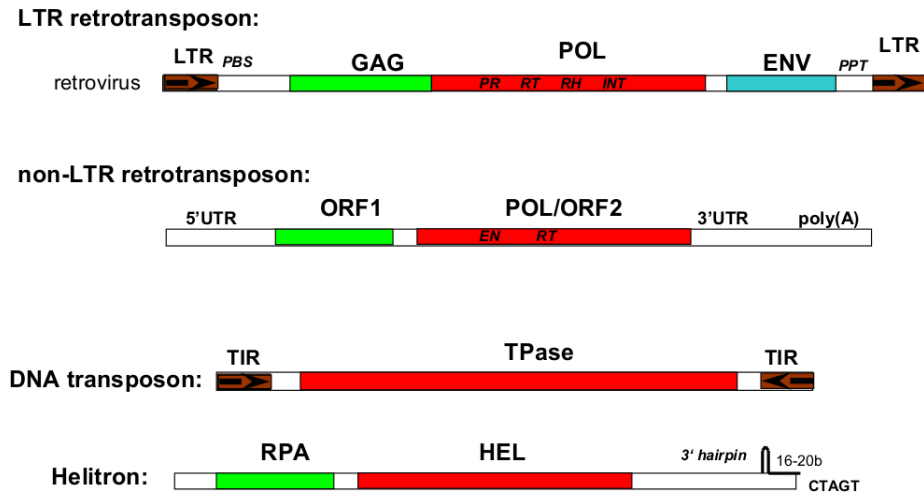
DNA transposony převládají u bakteriálních genomů, kde nalézáme především IS elementy (z angl. Insertion Sequences) [13], složené transposony, Tn a Tn3 elementy. Poslední dva se řadí mezi ty složitější a obsahují navíc geny, které kódují produkty funkčně nesouvisející s procesem transposice. DNA TE jsou rozšířené i mezi eukaryotami, mezi kterými jsou TE rozděleny do dvou podtříd. První podtřída obsahuje hlavně transposony s obrácenými koncovými repeticemi (TIR, z angl. Terminal Inverted Repeat) a mezi nejznámější případy patří Ac/Dc elementy u kukuřice, P elementy u octomilky a evolučně starý Mariner element. Druhá podtřída zahrnuje dva řády Helitron a Maverick.

Retrotransposony s LTR nacházíme zejména u rostlin. Mezi ně patří rodiny Gypsy a Copia (také označované jako Ty1 a Ty3). Nejvíce se vyskytují u obilnin. Nejznámější případ maximálního pomnožení LTR TE je u kukuřice, kde tvoří přibližně 80% genomu. LTR TE jsou velmi podobné retrovirům. Oba obsahují geny *gag* a *pol* kódující strukturní a enzymatický aparát potřebný k jejich transposici. Na rozdíl od retrovirů jim však chybí gen *env* (z angl. envelope, obálka), který kóduje jednu ze složek virové kapsuly umožňující retrovirům opustit buňku. Z tohoto důvodu je pravděpodobné, že se LTR retrotransposony vyvinuly z endogenních retrovirů, nebo se naopak endogenní retroviry vyvinuly z transposonů [14]. Transposony bez LTR byly nalezeny ve velkém množství u všech eukaryot. Do této skupiny především patří LINE a SINE elementy, které jsou významně zastoupeny u savců a více jsou popsány v kapitole 2.3.

Autonomními se označují ty elementy, které jsou schopny samostatné transposice. Obvykle si samy kódují enzymatický aparát nezbytný k transposici. Neautonomní elementy jsou závislé na pomoci okolních autonomních elementů.

2.2 Mechanismy šíření TE

Transposony si osvojily rozličné strategie svojí transposice, specifická místa integrace a modely provázející jejich aktivitu. Stále více důkazů naznačuje, že u savců dochází k transposici hlavně během raného vývoje, nebo přímo v zárodečných buňkách. Tento model aktivace jim zaručí přenos i do dalších generací. K transposici dochází ale i v somatických buňkách, jak dokazují studie z oblasti neurogeneze a některých forem rakoviny.

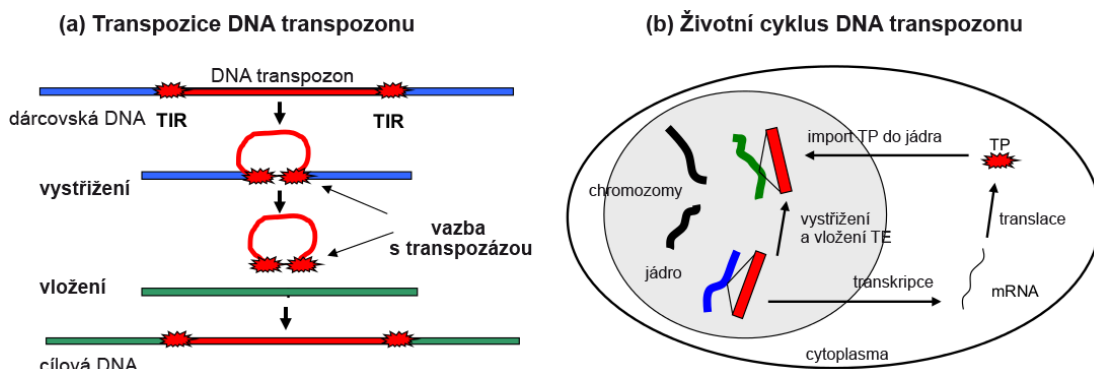


Obrázek 1: Typy mobilních genetických elementů rozdělené podle jejich struktury. GAG-protein virové částice, POL-polyprotein (reverzní transkriptáza a restrikční endonukleáza), ENV-obal virové částice, TPase-transpozáza, RPA-replikační protein A, HEL-helikáza. Převzato z [17].

U většiny transposonových inzercí můžeme pozorovat zdvojení cílového místa, neboli TSD (z angl. Target Site Duplication), která je výsledkem štěpených lepivých konců hostitelské DNA. Transposony I a II třídy obsahují přilehlé přímé repetice (angl. flanking direct repeat), které nejsou součástí transposonu ale při jeho vystřížení zůstávají na původním místě jako stopy a umožňují zpětnou identifikaci tohoto místa v genomu. Byly evidovány případy, kdy tyto repetice také mohou ovlivňovat expresi genu, v níž byly zanechány.

2.2.1 DNA transposony

Protože se DNA transposony pohybují prostřednictvím mechanismu vyjmi a vlož, počet jejich kopií zůstává v genomu konstantní. Mnoho DNA transposonů jsou na obou okrajích lemovány koncovými obrácenými repeticemi TIR (z angl. Terminal Inverted Repeat) o délce cca 9 až 40 bp, které jsou významné pro jejich transposici, jak bude vysvětleno dále. DNA transposony kódují enzym transpozázu, která je klíčová pro jejich transposici. Transpozáza rozeznává TIR repetice na které se naváže, rozštěpí hostitelskou DNA a integruje transposon na cílové místo. V průběhu transposice mohou k sobě TIR konce ligovat a tím stabilizovat celou strukturu transposon + transpozáza. Inzerce DNA transposonu má za TSD přibližně 4 - 8 bp dlouhou.

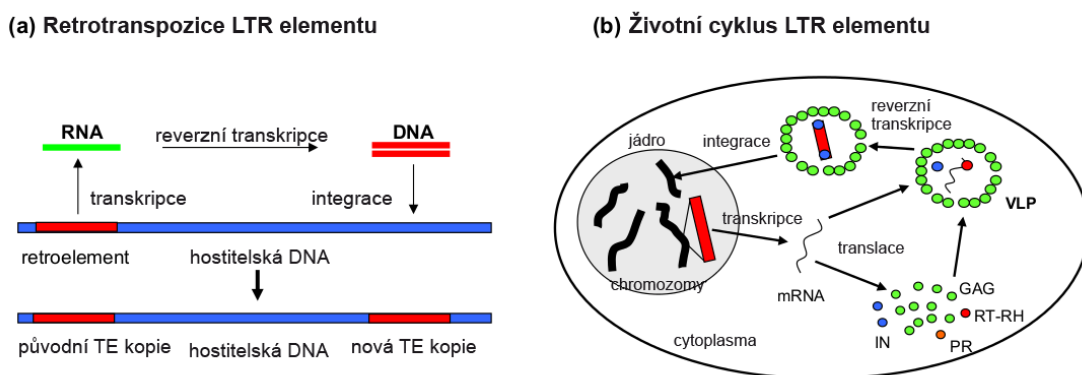


Obrázek 2: Transpozice a životní cyklus DNA transposonu. Převzato z [17].

2.2.2 LTR retrotransposony

Tento typ retrotransposonů je charakteristický tím, že při přemístění zachovává na místě svoji původní kopii (mechanismus zkopíruj a vlož). Dochází tedy k jejich replikaci a díky tomu mohou genom zahltit velkým počtem svých kopií. Kromě LTR konců, které z obou stran lemují retrotransposon, obsahují LTR TE geny GAG a POL, což jsou strukturní a enzymatické proteiny nezbytné k jejich mobilizaci. GAG kóduje protein virové částice a POL (polyprotein) kóduje proteázu, reverzní transkriptázu, integrázu a RNázu H.

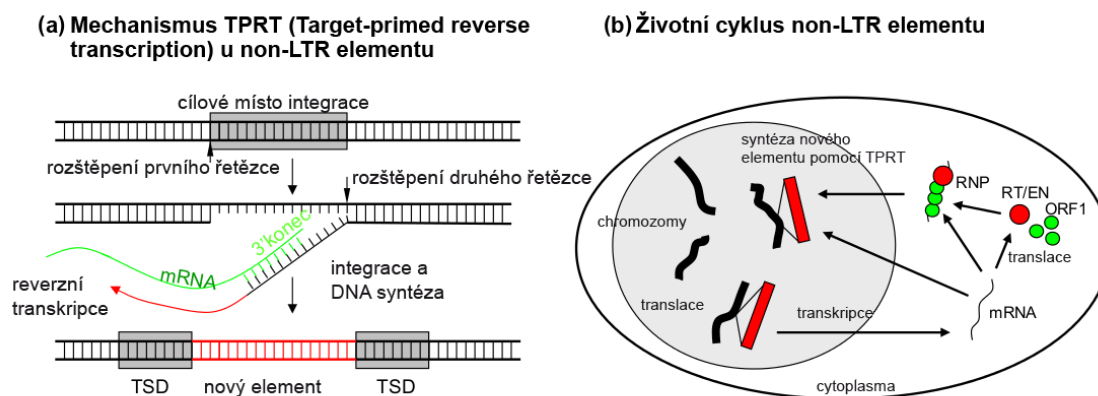
Transpozice proběhne tak, že se RNA polymeráza II naváže na promotor umístěný na 5' konci LTR a vytvoří mRNA přepis transposonu. Následně se z proteinu GAG vytvoří VLP částice (z angl. Virus Like Particle), která obsahuje vytvořenou mRNA, reverzní transkriptázu a integrázu. Reverzní transkripcí se mRNA přepíše do cDNA a tu poté integráza vloží do nové místo v hostitelském genomu.



Obrázek 3: Transpozice a životní cyklus LTR transposonu. Převzato z [17].

2.2.3 Non-LTR transposony

Non-LTR elementy obsahují na svém 5' konci promotor, následovaný dvěma čtecími rámci (ORF1 a ORF2) a polyadenylačním signálem na 3' konci. Jejich přemístění se děje na základě mechanismu TPRT (z angl. Target Site Primed Reverse Transcription). Nejprve je transposon přepsán polymerázou II do mRNA, tak jak tomu bylo u LTR transposonů. Poté endonukleáza vytvoří na jednom řetězci hostitelské DNA zářez, uvolňující 3'OH konec, který je využit jako počátek pro reverzní transkripci. Mnohé non-LTR transposony jsou na 5' konci poškozeny z důvodu předčasného ukončení reverzní transkripce, která není tak efektivní jako transkripce dopředná. Vytvářejí se tak většinou pouze zkrácené nefunkční kopie, lemované 7-20 bp dlouhými TSD.



Obrázek 4: Transpozice a životní cyklus non-LTR transposonu. Převzato z [17].

2.2.4 Helitrony

Helitrony jsou 5-15 kbp dlouhé transposony, které se replikují mechanismem otáčivé kružnice (angl. Rolling Circle), podobně jako plazmidy. Společnými vlastnostmi helitronů jsou integrace do míst bohatých na AT. Také postrádají zdvojení cílového místa (TSD) a na rozdíl od DNA transposonů neobsahují obrácené koncové repetice. Na 5' konci obsahují TCT nukleotidy a 3' konec CTAG, kterým předchází 18-25 bp dlouhá palindromická sekvence, schopná vytvářet vlásenkovou strukturu. Helitrony se často nacházejí v blízkosti genů.

Molekuly transpozázy vytvoří jednořetězcové zářezy na DNA molekule donorové i cílové na jejich 5' konci. Replikace helitronu začíná na původní donorové molekule od

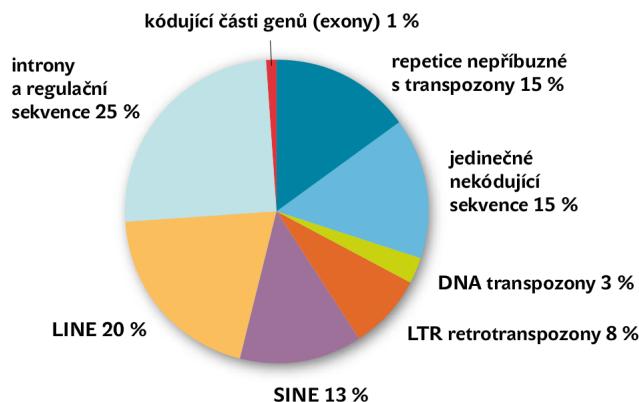
jejího volného 3' konce. Pokud je správně rozeznána koncová palindromická sekvence na 3' konci, dojde k odštěpení druhého konce helitronu po sekvenci CTRR. Helitron je poté přemístěn na cílovou molekulu kde vytvoří heteroduplex. Někdy také dochází ke špatnému rozeznání 3' konce a část sekvence donorové molekuly může být přenesena spolu s Helitronem [15].

2.3 TE v lidském genomu

Transposony tvoří přibližně 45% lidského genomu [3]. Některé současné odhady jsou však ještě odvážnější. Podle nich je více jak 2/3 našeho genomu výsledkem nedávné i evolučně vzdálené aktivity transposonů [24]. Obsahuje zástupce obou tříd DNA transposonů i retroelementů, ačkoli díky výrazné akumulaci mutací je většina inaktivní a jedná se o genetické fosílie. Elementy, které v lidském genomu převažují a zároveň jsou stále aktivní, jsou retrotransposony (non-LTR), přemísťující se mechanismem TPRT. Hlavními zástupci non-LTR elementů jsou LINE elementy L1, L2 a L3 tvořící 20% genomu (500,000 kopií), dále SINE elementy představující 13% genomu (Alu 10%, 1,000,000 kopií) a SVA elementy, které v lidském genomu vytvořily přibližně 3,000 kopií [4].

Lidský genom obsahuje tři hlavní SINE rodiny, a to Alu, MIR a MIR3. SVA elementy byly poprvé zmíněny byly roku 1994 a vzhledem jejich fylogenetické distribuci se odhaduje, že jsou relativně evolučně mladé v porovnání s L1 nebo Alu. Skládají se z SINE-R elementu, VNTR sekce (z angl. Variable Number Tandem Repeats) a Alu komponenty. LTR elementy v lidském genomu už tak hustě zastoupeny nejsou. Tvoří 8% genomu a jejich nejznámějšími zástupci jsou lidské endogenní retroviry HERV (z angl. Human Endogenous RetroViruses), mezi které patří HERV-I, HERV-K a HERV-L. Ačkoli je v lidském genomu více než 400,000 sekvencí odvozených z elementů podobných retrovirům, většina je neschopná transposice. DNA transposony jsou zastoupeny pouze 3% a všechny jsou považovány za inaktivní. Hlavními zástupci DNA transposonů jsou transposony z nad-rodin TC-1/mariner, hAT, a PiggyBac.

Odhady naznačují, že k nové retrotransposici Alu elementu dojde každé 20-té narození, u L1 každé 100-té a SVA každé 900-sté narození [4]. Elementy SINE a SVA nekódují enzymatický aparát nutný k jejich transposici a jsou tedy neautonomní. LINE elementy, jako jediní zástupci autonomních elementů v lidském genomu, jsou hlavním zdrojem



Obrázek 5: Složky lidského genomu, převzato z [20].

reverzní transkriptázy i pro neautonomní elementy a jejich aktivita tedy řídí rozsáhlé změny v architektuře celého genomu. Je tedy zřejmé, že elementy SINE a SVA fungují jako paraziti parazitů.

2.3.1 LINE elementy

V lidském genomu jsou z LINE elementů (z angl. Long Interspersed Nuclear Elements) aktivní pouze tři rodiny, a to L1 představující 17% genomu, L2 a L3. FL (z angl. Full Length) L1 elementy jsou dlouhé přibližně 6-7.5 kb a obsahují dva čtecí rámce ORF1 a ORF2, schopné produkovat funkční proteiny nezbytné pro jejich transposici, i pro transposici na nich závislých SINE a Alu elementů. Na obou koncích LINE elementů se nacházejí netranslatované oblasti UTR (z angl. Untranslated Region). V nich se na 5' konci kromě regulační sekvence může nacházet metylguanositová čepička, která je chrání před rozkladem buněčnými enzymy a usnadňuje iniciaci translace. Na 3' konci obsahují polyadeninový úsek, který je co do délky velmi variabilní. LINE elementy jsou transkribovány RNA polymerázou II. Lidské L1 elementy mohou být rozděleny do několika podrodin na základě obsahu specifických diagnostických sekvencí mezi 5' a 3' UTR. Každá z L1 podrodin se v průběhu evoluce primátů amplifikovala v jinou dobu. Například podrodina Ta byla aktivní docela nedávno. L1 elementy přispívají ke genetické variabilitě mezi jednotlivci. L1 inserční polymorfismus také je užitečný zdroj pro sledování populační historie [23].

LINE elementy se v lidském genomu nachází většinou jen jako nefunkční, na 5' konci

zkrácené kopie, neschopné další retrotransposice. Zkrácené kopie však mohou obsahovat regulační sekvence, které ovlivňují genovou expresi v přilehlých oblastech. Z 500,000 L1 kopií, které se nashromáždily v lidském genomu jich je přibližně jen 3,000 plné délky [8]. Z nich pouze cca 80-100 obsahuje oba neporušené ORF a jsou schopné transposice [7].

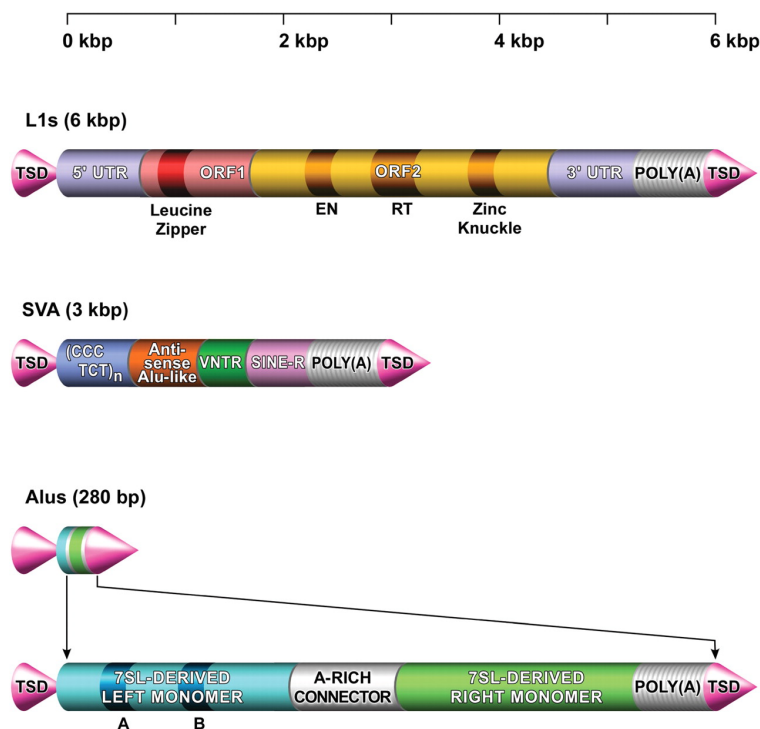
ORF1 kóduje protein, který se in vitro váže k L1 mRNA a vytváří ribonukleo-proteinový komplex RNP. Ten je považován za meziprodukt retrotransposice, který je patrně generován na polyribosomu. Pro retrotransposici je nutný, ale sám o sobě nedostatečný a zatím nevíme přesně jaká je jeho úloha. ORF2 kóduje protein s aktivitou endonukleázy a reverzní transkriptázy. Lidský L1 ORF2 může být vytvořen také alternativním splicingem z ORF1. Ačkoli takto vytvořený protein nemůže přispívat k L1 retrotransposici, může přispívat k mobilizaci neautonomních Alu elementů [5].

Dříve se předpokládalo, že k L1 aktivaci dochází pouze v zárodečných buňkách. Tento předpoklad stavěl na modelu *Mus musculus*, který vykazoval zvýšenou expresi L1 ORF1 v myších zárodečných buňkách. Vysoké hladiny ORF1 však byly detekovány i v lidských somatických buňkách. V některých tkáních byla míra exprese FL L1 elementů na úrovni, která byla detekovaná u rakovinných buněk. Tyto rozličné hodnoty exprese L1 v lidských tkáních mohou být vysvětleny různou úrovní metylace promotoru [4].

2.3.2 SINE elementy

SINE elementy (z angl. Short Interspersed Nuclear Elements) jsou oproti L1 elementům velmi krátké. Délka jejich hlavního zástupce v lidském genomu Alu elementu nepřekročí 400 bp (konsenzus cca 280 bp [9]). Protože neobsahují ani gen pro reverzní transkriptázu, jsou zcela závislé na enzymatické aktivitě L1 ORF2. Lidský genom obsahuje tři hlavní rodiny SINE elementů, a to Alu, MIR a MIR3. Alu elementy jsou nejčastěji se objevující sekvencí v lidském genomu (více jak 1,000,000 kopií tvořící 13% genomu). Jsou složeny ze dvou podobných podjednotek navzájem oddělených regionem bohatým na adenin. Na jejich 3' konci je většinou umístěna polyadeninová sekvence a 5' konec obsahuje promotor pro RNA polymerázu III. Alu jsou svojí délkou dosti heterogenní, protože RNA polymeráza III nerozeznává terminační sekvenci, a místo ní rozeznává čtyři nebo více thyminů. Efektivita Alu transposice je také ovlivněna délkou polyadeninové sekvence.

Alu elementy jsou pravděpodobně odvozeny z 7SL RNA a podobně jako geny pro



Obrázek 6: Aktivní mobilní elementy v lidském genomu. TSD-target site duplication, UTR-untranslated region, EN-endonukleázová doména, RT-doména reverzní transkrip-tázy, VNTR- variable number tandem repeats, převzato z [22].

tRNA jsou transkribovány RNA polymerázou III. To je odlišuje od jiných krátkých rozptýlených repetit například mikrosatelitů, nebo MITE elementů. Zajímavé je, že expanze Alu elementů se shoduje s oddělením linie primátů zhruba před 6 miliony let [21]. Jen zlomek Alu elementů je retrotransposičně aktivní. Ačkoli je charakteristika funkčních Alu nejasná, předpokládá se že hraje roli stáří Alu elementů, integrita promotoru RNA polymerázy III a délka spolu s homogenitou poly(A) sekvence. Jako transposičně nejaktivnější skupina mají Alu hlavní vliv na lidský organismus a jeho nemoci. Alu sekvence jsou také hlavní předlohou pro ektopickou rekombinaci (viz dále).

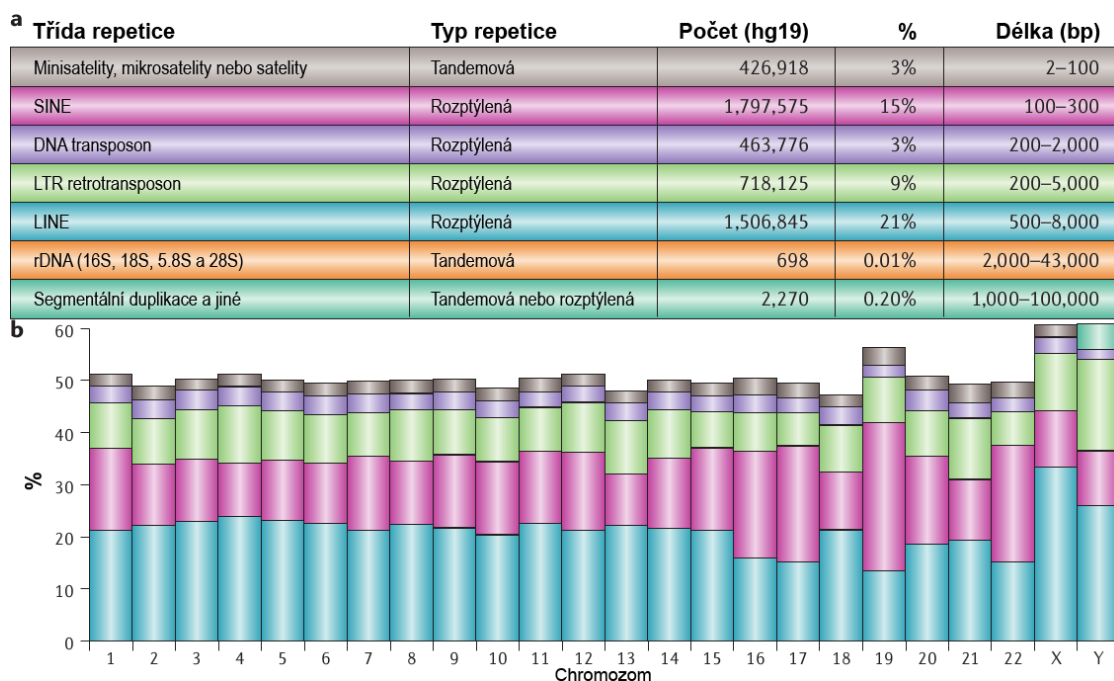
2.4 Distribuce TE v lidském genomu

Retroelementy jsou v lidském genomu přítomny na všech chromozomech. V některých místech je jejich koncentrace zvýšená a někde naopak snížena. Distribuce retroelementů je výsledkem post-integračních selekčních procesů, které utvářely podobu genomu po miliony let. Přírodní výběr fixoval ty inserce, které poskytovaly svému hostiteli určitou

výhodu a eliminoval inserce, které se ukázaly jako škodlivé. Bylo zjištěno, že některé transposony používají elegantní mechanismy pro cílenou integraci do specifických genomických lokusů (ang. gene targeting), zatímco jiné transposony tuto specifitu zřejmě postrádají [17]. Pokud se zaměříme na lidský genom, Alu elementy jsou zastoupeny zejména v intronech. Oproti tomu L1 jsou více rozptýleny, jejich kopie jsou od sebe často více vzdálené a jsou výrazně méně zastoupeny v kódujících oblastech [6]. I přes rozptýlený charakter distribuce L1 elementů, množství důkazů naznačuje existenci jejich cílené integrace. Rozdělení může také odrážet cílenou specifitu endonukleázy L1 elementů, která štěpí sekvenci 5'-TTTT/A-3' a zahajuje TPRT transposici stejně jako Alu preferenci insercí do 5'-TTAAAA motivů [18]. Přesný mechanismus tohoto asymetrického rozdělení není znám. Asymetrické rozdělení Alu a L1 elementů však pravděpodobně vzniklo následkem současného působení více faktorů.

Byl prokázán zvýšený počet transposonů v heterochromatinových oblastech, zejména v subtelomerických a pericentromerických. Při integraci do heterochromatinu je menší šance, že dojde k poškození kódující části DNA a tak je snížena pravděpodobnost eliminace elementů. Většina TE v heterochromatinových oblastech je většinou neaktivní a tyto oblasti představují spíše "hřbitovy" retroelementů. Včleňování do heterochromatinu je pravděpodobně následkem cílené integrace TE.

Transposony jsou nejvíce pomnoženy na pohlavním chromozomu Y, u kterého je potlačena, nebo snížena meiotická rekombinace. Navíc na něm nedochází k selekci proti inserčním mutacím. Na chromozomu X jsou dokonce L1 elementy zastoupeny dvakrát více, než na ostatních autosomech. Tyto elementy zde také pravděpodobně hrají roli jako podpora procesu, který se nazývá inaktivace X chromosomu [12]. Jeden z X chromozomů v 46-XX je genomu umlčen inaktivací, z důvodu kompenzace délky genů na chromozomu X u mužů a u žen.



Obrázek 7: a) Zastoupení různých typů repetice v lidském genomu. Barvy v obrázku a) korespondují s barvami v obrázku b), popisujícím jejich procentuální rozdělení na jednotlivých chromozomech. Převzato z [31].

2.5 Aktivace TE

Aktivace TE se může objevit jako následek fungování rozličných mechanismů. Například demethylace a epigenetické přeprogramování aktivuje L1 elementy a dochází k ní během brzké embryogeneze. Experimenty popisují také transposici vyvolanou mnoha fyzikálními faktory. Transposice může být navozena UV či gamma zářením, teplotním šokem, virovou infekcí, nebo vlivem působení mutagenního faktoru či přímo v rakovinových buňkách. U některých chemikálií byla prokázána až 3x větší aktivita v lidských buněčných kulturách, a to u rtuťi, kadmia a niklu. U rostlin byla aktivace transposonů prokazatelně zapříčiněna různými formami stresu. Zvýšený počet počtu transposonů byl pozorován například u rostliny *Hordeum spontaneum* na svazích v Izraeli. Na sušším svahu, kde byla rostlina vystavena vyšším stresovým podmínkám bylo detekováno více BARE-1 elementů, než na protějším svahu s příznivějšími podmínkami. Dále byl větší počet transposonů detekován u hybridů, vzniklými křížením. Příklad k tomuto tvrzení najdeme opět mezi rostlinami, kdy častěji křížená *Arabidopsis lyrata* vykazuje zvýšenou transposiční frekvenci a výraznější selekci proti novým TE inzercím, stejně jako rychlejší

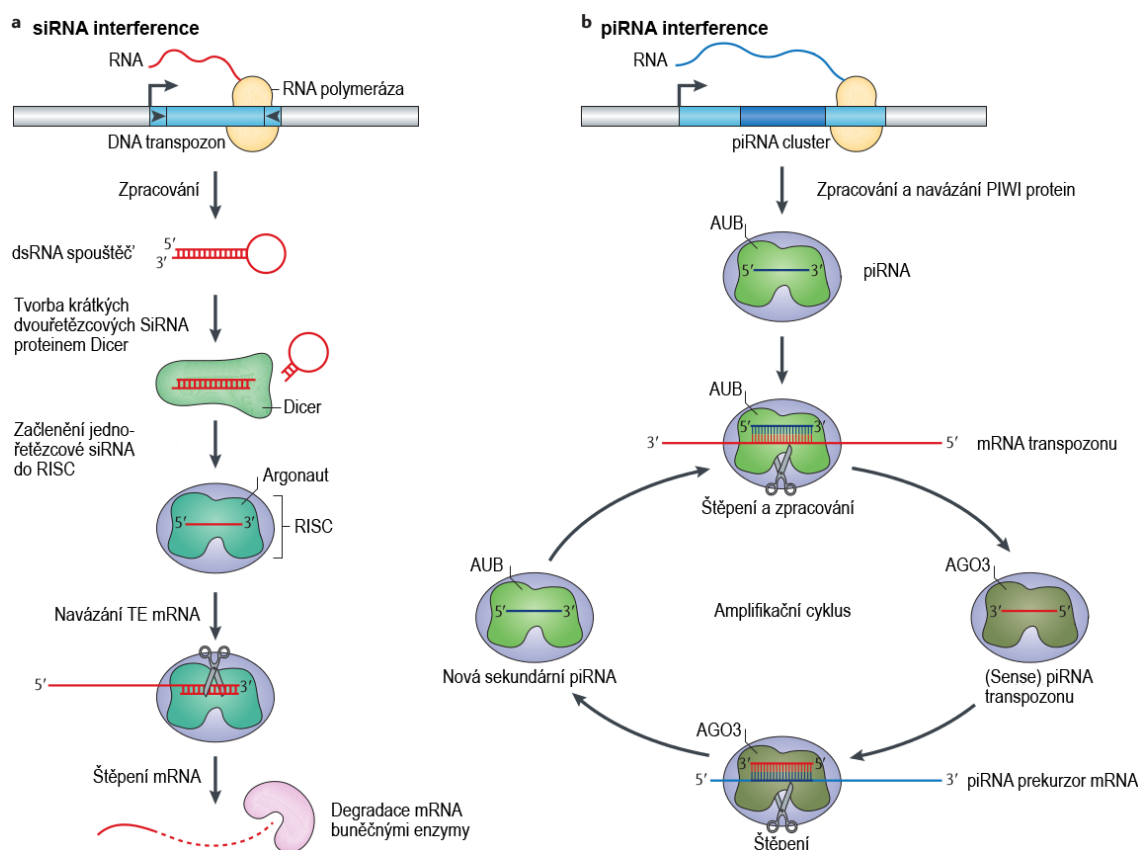
mechanismus jejich odstranění ektopickou rekombinací, než nekřížená *A.thaliana*.

2.6 Umlčování TE

Umlčování transposonů je způsob, jakým se hostující organismus brání proti jejich šíření. Svoji aktivitou transposony mohou generovat množství mutací, jako inserce, delece, duplikace, inverze, translokace, posuny čtecího rámce i větší genomové přestavby. Mezi strategie používané k omezení aktivity TE patří ektopická rekombinace, RNA interference a metylace cytosinů nacházejících se v promotorech transposonů. DNA metylace patří mezi epigenetické modifikace DNA, což je mechanismus, jakým vnější prostředí ovlivňuje genom. Epigenetické modifikace vysvětlují například rozdíly mezi jednovaječnými dvojčaty. Takové modifikace nemění kódující posloupnost bází, ale upravují promotorové části genomové DNA regulující genovou expresi. Mezi epigenetické modifikace kromě DNA metylace patří i modifikace histonů (methylace a acetylace). Tyto epigenetické značky mají přímý efekt na expresi přilehlých genů a transposonů nevýmaje. Jsou považovány za určitý sekundární kód, který doplňuje DNA sekvenci nukleotidů [11]. Vzor pro DNA methylici je vytvořen už během gametogeneze, ale může se během života měnit např. pod vlivem stresu. Nevysvětlenou otázkou zůstává, jak hostující organismus rozezná transposony jako substrát pro methylici.

Další možností umlčující aktivitu transposonů je potranskripční degradace mRNA transposonů RNA interferencí (RNAi). RNA interference může probíhat dvěma cestami, siRNAs (z angl. Small interfering RNA) a nebo piRNA (z angl. PIWI-interacting RNA) [2]. Pro degradaci mRNA TE cestou siRNA je nejdříve potřeba dsRNA 'spouštěče', který je odvozený z DNA transposonu s obrácenou koncovou repeticí (ve formě vlásenky, kdy se konce spojí). Takovýto 'spouštěč' je poté rozštěpen proteinem z rodiny Dicer na 21-24 nt dlouhé siRNA. Jednovláknový úsek siRNA, který je komplementární k mRNA TE, je začleněn do RISC komplexu (z angl. RNA-induced silencing complex), který se váže k TE mRNA a štěpí ji na malé kousky, které jsou poté degradovány buněčnými enzymy.

Pro degradaci transposonové mRNA cestou piRNA je nejdříve vytvořen primární piRNA transkript, vygenerovaný z piRNA clusteru (piRNA cluster je DNA lokus, který obsahuje sense i antisense sekvence, které jsou odvozeny z mobilních genetických elementů). Na tento primární piRNA transkript o délce 24-35 nt se naváže PIWI, nebo



Obrázek 8: Umlčování transposonů RNA interferencí. (a) siRNA cesta, (b) piRNA cesta, převzato z [2].

Aubergine (AUB) protein, který umožní jeho nasměrování ke komplementární mRNA TE sekvenci. Po navázání, je komplementární mRNA TE sekvence přestřížena a uvolněna, čímž z ní vzniká sekundární (angl. sense) piRNA, která se váže k Argonatu 3 (AGO3) proteinem. Takto vzniklý nový komplex se váže k původnímu předchůdci piRNA a je opět následováno štěpením endonukleázou, což regeneruje (angl. anti-sense) piRNA, která je nasměrována k mRNA TE. Kvůli iterativnímu charakteru se tento mechanismus nazývá 'ping-pong' cyklus a často slouží k destrukci transposonové mRNA v zárodečných buňkách. Transposony vedou ke vzniku množství siRNA ve většině organismů a urovně siRNA jsou korelovány s aktivitou transposonů.

2.7 Poškození DNA způsobené TE

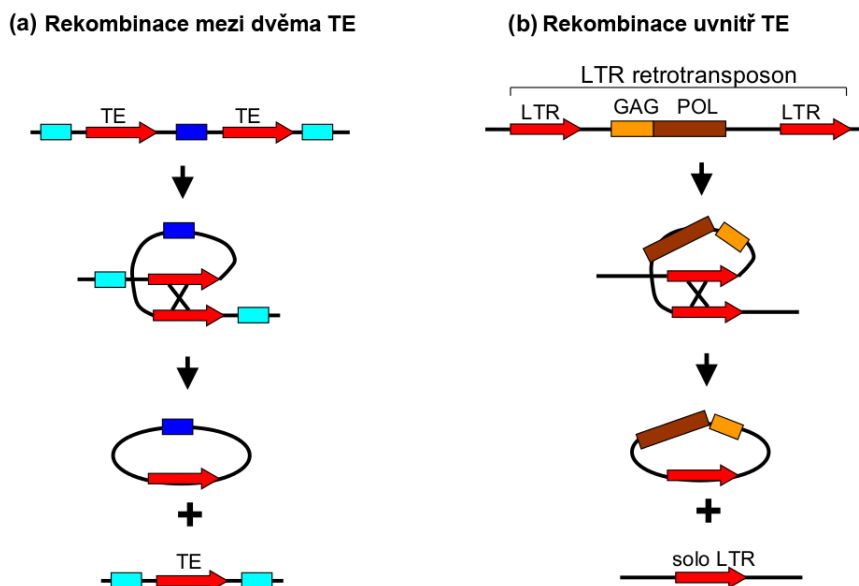
2.7.1 Inzerční mutageneze

Inzerční mutageneze vzniká inzercí transposonu do oblasti v genomu s významnou biologickou funkcí. Pokud dojde k inzerci TE do kódující oblasti, může dojít k narušení funkce příslušného genu, nebo ke změně v jeho expresi. Kvůli usnadnění svojí transkripce obsahují transposony vlastní promotory a regulační sekvence. Pokud dojde k inzerci transposonu do oblasti blízké genu, vyvstává možný potenciál, že promotory a regulační sekvence ovlivní expresi tohoto genu. K inzerční mutagenezi může docházet jak v zárodečných, tak v somatických buňkách a mutace jimi způsobené mohou vést přímo ke vzniku nemocí a maligní transformace, nebo k ní buňky učiní náchylnějšími. Bylo zjištěno, že v tumorech dochází k více TE inzercím, než v jim typově odpovídajících normálních buňkách. Při L1 integraci často dochází k delecí genomické sekvence, které mohou být těžko detekovány jako způsobené L1 aktivitou. Proto je složité určit hranice inzerční mutageneze, za kterou nesou zodpovědnost L1 transposony. Ačkoli by se mohlo očekávat, že inzerční mutageneze je hlavní mutagenní faktor spojený s TE aktivitou, ektopická rekombinace způsobuje ještě vyšší genomovou nestabilitu.

2.7.2 Ektopická rekombinace

Dokonce i pokud dojde k inzerci transposonu do zdánlivě neškodného místa, mohou jejich nové kopie přispět k narušení genetické stability post-inzerčně, protože TE často působí v genomu jako zdroj pro ektopickou rekombinaci NAHR (z angl. Non Allelic Homologous Recombination), tedy rekombinaci mezi sekvencemi které jsou silně homologní, ale nejsou alelami. Při ektopické rekombinaci dochází k delecím, nebo duplikacím sekvence, nacházející se mezi zdrojovými TE sekvencemi. Může tak docházet i k větším přestavbám genomu, protože rekombinace ovlivňuje i lokusy TE nacházející se ve značné vzdálenosti.

Alu elementy obsahují určité charakteristické znaky, které je činí náchylnějšími k rekombinacím. Těmito znaky jsou relativní blízkost jednotlivých Alu elementů, které činí Alu/Alu rekombinaci více snesitelnou, dále jejich sekvenční identita (v průměru větší jak 75 %), která zajistí efektivní sekvenční párování v průběhu crossoveru, chi-like motiv přítomný u Alu sekvencí, který stimuluje rekombinaci, a také jejich vysoký počet v genomu, který zvyšuje rekombinační pravděpodobnost. Ačkoli jsou geny ob-



Obrázek 9: Ektopická rekombinace, (a) vzniklá rekombinací mezi dvěma TE, (b) mezi LTR koncovými sekvencemi jednoho TE. Převzato z [17].

sahující větší množství Alu náchylnější k rekombinaci, některé Alu elementy mají větší sklon přispívat k rekombinaci, i když nejsou tak hustě zastoupeny. Pravděpodobně je to způsobené různou mírou homologie mezi Alu sekvencemi. Oproti tomu rekombinace mezi vzdálenými L1 sekvencemi by znamenala kritickou škodu v genomu. Ektopická rekombinace je nejúčinnějším mechanismem regulujícím počty TE.

Ektopická rekombinace byla rozpoznána jako hlavní zdroj poškození DNA vedoucí k duplikaci nebo delecí sekvenční mezi dvěma zúčastněnými Alu. Alu elementy, které jsou blízko sebe v obrácené orientaci jsou nejvíce náchylné k rekombinaci. Bylo také dokázáno, že Alu jsou 10,000 krát více nestabilní při absenci funkčního p53, což naznačuje že významně přispívají ke genetické nestabilitě během tumorigeneze. Bylo publikováno mnoho konkrétních Alu/Alu případů rekombinace přispívajících k nádorovým onemocněním, jejichž počet dalece přesahuje počet nemocí způsobených Alu insercemi. Z tohoto důvodu mají rekombinace zahrnující Alu elementy dalekosáhlejší efekt na genomickou nestabilitu, než inserce těchto elementů. Ektopická rekombinace mezi jednotlivými elementy také patří mezi důležité regulátory počtu TE.

2.7.3 Dvouřetězcové zlomy

Aktivita L1 elementů přispívá k nestabilitě genomu nejen svými insercemi, ale také vede k dvouřetězcovým zlomům DSB (z angl. Double Strand Breaks). Bylo dokázáno, že poškození DNA DBS je způsobeno endonukleázovou aktivitou, která je kódovaná L1 ORF2. Nicméně stále nebylo dokázáno, zda jsou tyto dvouřetězcové zlomy výsledkem nezdařených L1 insercí, nebo jestli se endonukleáza uvolí a začne neřízeně napadat genom. Jako obranu proti DBS si buňka vytvořila mechanismy jako NHEJ (angl. Non-homologous end joining), nebo HDR (angl. Homology-driven repair pathways), které umožňují opravu poškozené dvoušroubovice. Neopravené DBS DNA jsou vysoce toxické a často mají za následek vznik mutací. DBS vede ke gamma fosforylaci histonu 2AX, které mohou být detekovány imunohistochemicky ve formě gamma-H2AFX ložisek.

2.8 Domestikace TE

Pokud hostitelský organismus převezme transposony k vykonávání funkce pro něj prospěšné, nazýváme to domestikací transposonů. Protože jsou transposony v genomu zastoupeny v mnoha kopiích, mohly v rámci evoluce sloužit jako ideální stavební materiál pro tvorbu nových nebo modifikovaných forem genů. To v důsledku vede k diverzifikaci nových druhů. V přírodě najdeme mnoho příkladů kdy TE vykonávají nějakou hostiteli prospěšnou funkci. Ty zahrnují například vliv transposonů na regulaci genové exprese, na evoluci v sekvencích kódujících proteiny, na genomovou plasticitu hostitele, speciaci, a mnohé jiné. Souvislost transposonů a speciace naznačuje i studie, která zjistila že oddělování jednotlivých větví savců koreluje s periodickými explozivními amplifikacemi transposonů.

Například 25% promotorů u člověka obsahují sekvence, které byly původně odvozeny z mobilních DNA elementů [12, 10]. Jak již bylo zmíněno výše, TE se podílejí na inaktivaci chromozomu X. Také centromerové sekvence jsou často tvořeny různě degenerovanými transposony. Transposony se také podílejí na reparaci dvouřetězcových zlomů, které představují ideální cíl pro inserci elementu. Příkladem, že transposony slouží jako genetický materiál pro formování nových genů, je i vznik genu SETMAR, který vznikl insercí transposonu typu Mariner (MAR) do blízkosti genu SET kódujícího histon metyltransferázu. Vznikl tak gen s funkcí histon methyltransferázy se schopností se současně vázat na cílové sekvence TIR Mariner transposonů. Jedinečný přík-

lad domestikace transposonů najdeme u *D.melanogaster*, kde LINE elementy nahrazují funkci telomerázy, která regeneruje konce chromosomů zkracované během DNA replikace.

Působivé je také spojení mezi transposony a V(D)J rekombinací, která je základem pro imunitní systém obratlovců [16]. Imunoglobulinové receptory jsou syntetizovány a vylučovány diferencovanými B-lymfocyty. Jsou kódovány lokusem V(D)J genů, skládající se ze série diskrétních kódujících segmentů (L-V, D, J a C). Během diferenciací lymfocytů dochází k V(D)J rekombinaci. Jedná se o proces vytvářející množství různých kombinací genů, jejichž široký repertoár má za následek specifickou imunitní odpověď v podobě rozličných typů Ig. Tato specificita je rozhodující součástí imunitního systému obratlovců. Rekombinace V(D)J lokusu je katalyzována dvěma enzymy RAG1 a RAG2, které společně tvoří enzym, u kterého se zjistila příbuznost s transpozázou (enzym umožňující transposici DNA transposonů) [10]. Tyto objevy naznačují, že enzymatická výbava transposonů stála za evolucí V(D)J rekombinace.

3 Nemoci způsobené aktivitou TE

Mutace, vzniklé aktivitou mobilních genetických elementů v lidském organismu, mohou přímo způsobit propuknutí různých nemocí, včetně těch nádorových. Demonstrovat, že aktivita transposonů způsobuje rakovinu je obtížné, hlavně kvůli problematické sekvenaci repetitivních sekvencí. Stále více studií na zvířecích modelových organismech ale nabízí argumenty podporující aktivní roli TE na vzniku nádorových onemocnění. Bylo zjištěno, že retrotransposice probíhá ve velmi vysokých frekvencích v somatických buňkách, obzvláště tedy v mozku. Hlavním mutagenním nástrojem TE jsou jejich inserce. Jako mutagenní faktor ale působí i post-inzerčně, prostřednictvím ektopické rekombinace.

L1 transposony způsobují choroby především prostřednictvím inzerční mutagenese, která je často doprovázená 3' transdukci, tedy přenosem části DNA umístěné za L1 elementem, jako výsledek slabého transkripci ukončujícího poly(A) signálu. Alu elementy škodí především tím, že fungují jako častý zdroj pro ektopickou rekombinaci, ale také způsobují inzerční mutagenesi a změny v sestřihu RNA molekul. SVA elementy také mají schopnost narušovat geny prostřednictvím inzerční mutagenese a odlišného sestřihu RNA. Jelikož DNA transposony jsou v lidském genomu neaktivní, nejsou zaznamenány žádné nemoci jimi způsobené. Některé lidské endogenní retroviry HERV si ale zachovaly schopnost transposice, jako například K113, který vykazuje inzerční polymorfismus napříč lidskou populací. Transposony typu HERV jsou navrhovány jako potenciální přispěvatel k autoimunitním chorobám.

Prozatím bylo zaznamenáno 96 TE insercí způsobujících nemoci. Z nich 25 patřilo L1 elementům a zbylých 71 jsou L1 zprostředkované, mezi kterými je 60 případů připsáno Alu elementům, 7 SVA elementům a 4 inserce ze kterých zbyla pouze poly(A) sekvence. Celkem transposony pokrývají 0.4% ze všech nemoc způsobujících mutací [24].

3.1 Inzerční mutagenese TE

L1 elementy jsou svými insercemi zodpovědné za méně než 20% lidských nemocí. Tento malý počet může vyplývat z preference insercí do oblastí bohatých na AT (které neobsahují tolik kódujících sekvencí), nebo také hraje roli silná negativní selekce (jejich délka cca 20 krát překračuje délku Alu elementů, proto jejich inserce způsobí větší škodu). L1 inserce mohou také způsobovat delece v cílovém místě o délce 1-70,000 bp [26].

K lidským L1 elementy způsobeným somatickým mutacím patří například inserce L1 elementu do proto-onkogenu *c-myc* nalezená v buňkách karcinomu prsu, způsobující přeskupení jednoho *myc* lokusu a amplifikaci druhého lokusu. Dalším příkladem je inserce L1 do tumor supresorového genu *Apc*, nalezeného u pacientů s rakovinou tlustého střeva.

Inserce Alu elementu byla pozorovaná například do genu *Apc* u desmoidních nádorů, nebo také v souvislosti s neurofibromatózou, kdy inserce Alu elementu do intronu genu NF-1 vedla k deleci a posunu čtečího rámce. Známá je také souvislost insercí Alu elementů do genů BRCA1 a BRCA2 vedoucí k rakovině prsu. Alu elementy mají také souvislost s onemocněním očí. Polymorfismus Alu elementu u *ACE* genu je spojován s lepší ochranou proti věkem podmíněné makulární degeneraci. Ačkoli se Alu sekvence častěji vyskytují v intronech, jsou známy i případy, kdy inserce Alu do exonu způsobila chorobu. Jak již bylo zmíněno, u Alu sekvencí inzerovaných do intronů dochází často k jejich zadržení a přepsání do mRNA procesem Alu exonizace, kde vytváří nová štěpná místa pro RNA splicing.

Retrotransposony jsou také původci populačně specifických chorob [24]. Mezi ně patří inserce Alu do exonu genu *MAK* identifikovaná u pacientů židovského původu s diagnostikovanou retinitis pigmentosa, kogenitální svalová dystrofie typu Fukuyama u japonských pacientů způsobená insercí SVA elementů, nebo L1 způsobená 3' transdukce genu *dytrophin* vedoucí k Duchenneově svalové dystrofii u japonských chlapců.

3.2 Post-inzerční mutageneze TE

Ektopická rekombinace mezi dvěma inzerovanými Alu elementy, způsobí buď deleci nebo duplikaci sekvence umístěné mezi nimi. Gen BRCA1 obsahuje neobvykle vysoký počet Alu elementů, což vede taktéž k častým ektopickým rekombinacím a tedy i delecím/duplikacím na tomto lokusu. U genu BRCA2 delece nebyly detekovány tak často. Alu/Alu rekombinace byla pozorována také u hepatomu a bylo také zjištěno, že rekombinace mezi Alu elementy vyvolává translokace v *Tre-2* onkogenu a způsobuje vývoj Ewingova sarkomu. Dalším příkladem může být opakující se duplikace *MYB* lokusu (kóduje transkripční faktor), které jsou výsledkem homologní rekombinace mezi Alu elementy lemujícími *MYB* gen [25]. Tyto duplikace přispívají ke vzniku akutní lymphoblastické leukemii u T-buněk (T-AML). Opakující se duplikace *MLL* genu (myeloidní/lymphoidní

nebo mixed-lineage leukemia) mohou zase mít za následek vznik akutní myeloidní leukemie (AML). Alu elementy také hrají roli v chronické myeloidní leukemii, kde jsou přítomny v chromozomových zlomech, produkující translokace mezi chromozomy 9 a 22.

4 Sekvenační metody nové generace

4.1 Sangerova metoda

V roce 1977 byly nezávisle na sobě popsány dva postupy pro sekvenaci molekul DNA. F. Sanger a A. R. Coulson vyvinuli Sangerovu metodu a nezávisle na nich přišli američtí vědci A. Maxam a W. Gilbert s podobnou Maxam-Gilbertovou metodou. Oba týmy byly oceněny Nobelovou cenou v roce 1980, avšak pro svou jednoduchost se stala kapilární automatizovaná Sangerova metoda standardem, který dominoval trhu přes téměř dvě desetiletí a vedl k dosažení monumentálních výsledků, jako například osekvenování kompletního lidského genomu. Sangerova metoda je v současnosti nejpoužívanější metodou pro klinické použití a pro sekvenaci *de novo*.

Sekvence probíhá začlenováním směsi deoxyribonukleotidů spolu s dideoxynukleotidy, které po přidání ukončí elongaci řetězce. Výsledkem je směs různě dlouhých fragmentů, ze kterých je po elektroforetickém rozdělení možno rekonstruovat sekvenci DNA. Tímto postupem je možné dosáhnout sekvenace fragmentů délky až 1000 bp s náklady \$500 za 1Mb [32]. Navzdory mnohým technickým vylepšením které zvýšily účinnost a přesnost o tři řády, limitace Sangerovy metody odkryly potřebu pro nové a lepší technologie. Automatizovaná Sangerova metoda je považovaná za technologii první generace (z angl. “first-generation”) a novější metody jsou označovány jako nová generace (z angl. NGS-Next Generation Sequencing) [28].

4.2 Masivně paralelní sekvenování

Jako první z řady NGS postupů byla vyvinuta metoda masivně paralelního sekvenování (z angl. MPSS-Massively Parallel Signature Sequencing) firmou Lynx Therapeutics, [27]. Novátorský přístup MPSS metody spočívá v rozdělení DNA molekuly na miliony kratších fragmentů, které jsou analyzovány současně. Tím dojde k snížení nákladů a zvýšení výkonosti. Sekvence metodou MPSS probíhala amplifikací vzorku emulzní PCR (emPCR, viz dále) na mikrokuličkách po které následovala detekce fluorescenčního signálu CCD kamerou, produkovaného přidáváním fluorescenčně značených nukleotidů. Očekává se, že MPSS si najde své uplatnění v klinické osobní diagnostice. MPSS umožňuje celogenomový screening pro nové mutace a nebo současný screening stovek různých lokusů pro geneticky heterogenní onemocnění. Pomocí MPSS bylo

odhaleno obrovské množství zárodečných a somatických variant u normálních jedinců.

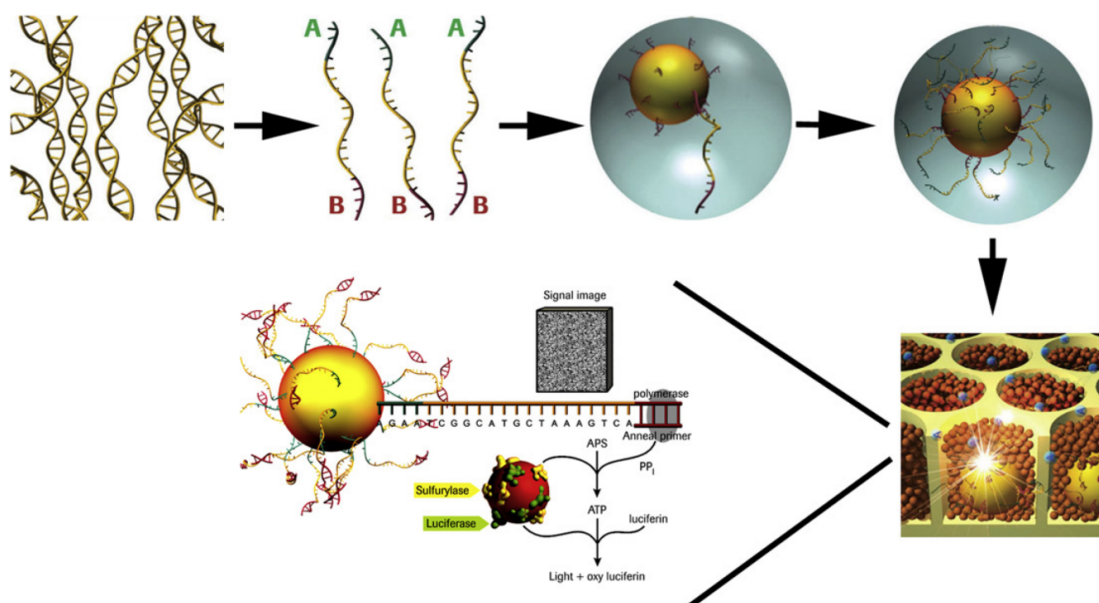
Metoda MPSS se stala základním kamenem pro rozrůstající se množství komerčně dostupných platforem. Sekvenační metody nové generace se vyznačují kombinací různých strategií od přípravy vzorku, sekvenování, zobrazování a zarovnávacích či shlukovacích metod. V současnosti trhu dominují čtyři hlavní představitelé Roche, Illumina, Applied Biosystems a Ion Torrent. Dalšími zástupci jsou firmy Helicos BioSciences, Polonator instrument, Pacific Biosciences a Oxford Nanopore, využívající nanopóry jako biosenzory k detekci DNA sekvence nukleotidů.

Ačkoli je metoda MPSS velmi komplexní, má i své omezení. Většina dostupných platforem NGS vykazuje vyšší chybovost oproti Sangerově sekvenaci. Další nevýhodou je obvyklá délka fragmentů, která není vyšší jak 50-100 bp. To dělá *de novo* sekvenování více obtížným, hlavně pokud se DNA skládá z repetitivních sekvencí.

4.3 Platformy NGS

4.3.1 Roche/ 454 GS-FLX

Sekvenátor typu 454 od společnosti Roche byl první dostupný NGS přístroj [29]. Kombinuje techniku amplifikace vzorku prostřednictvím emulzní PCR s pyrosekvenací. Emulzní PCR (emPCR) slouží k amplifikaci DNA fragmentů přibližné délky 450 bp přichycených na mikrokuličkách. Přichycení DNA fragmentů na mikrokuličky je umožněno přidáním adaptérů na oba konce fragmentu. Adaptéry obsahují také primery RNA polymerázy II pro iniciaci PCR reakce. Adaptéry ohraničená DNA je současně přichycena k povrchu mikrokuličky. Mikrokuličky jsou izolovány uvnitř kapiček na bázi voda-olej a vytvářejí tak uzavřený mikro-bioreaktor, ve kterém probíhá samotná PCR reakce. V každé kapce může být umístěna pouze jedna mikrokulička s jednou vzorovou DNA molekulou. Tím se zajistí, že amplifikace bude probíhat současně. Po proběhnutí PCR reakce obsahuje každá mikrokulička miliony kopií cDNA molekul. V dalším kroku je emulze voda-olej rozbita a cDNA přichycená na mikrokuličkách je podrobena sekvenaci. Ta probíhá uvnitř pikolitrových jamek čipu PicoTiterPlate (PTP), způsobem jedna kulička na jamku. Do jamek jsou dále přidány DNA polymeráza, sulfuryláza a luciferáza. Při přidání každého dNTP polymerázou zprostředkovanou reakcí je uvolněno světlo ze vzniklého pyrofosfátu. Světlo je zachyceno na podložním sklíčku složeného z optických vláken, které vede světelný signál až na vstup CDD kamery. Počet přidávaných stejných



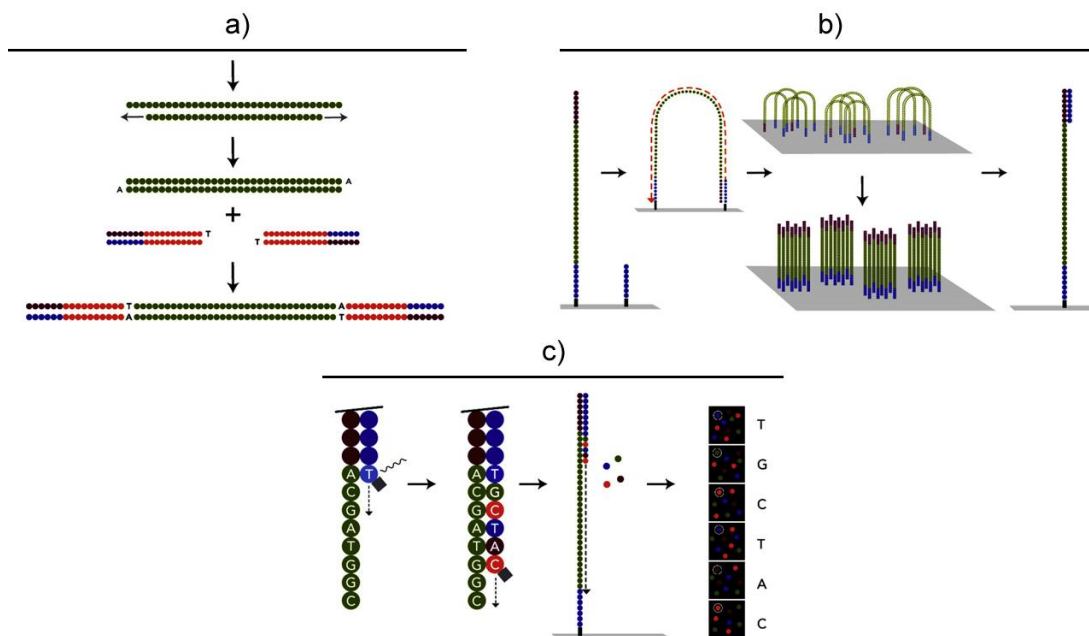
Obrázek 10: Pyrosekvenace pomocí přístroje Roche 454 GS FLX. Převzato z [32].

dNTP je přímo úměrný detekovanému signálu.

Hlavní nevýhodou této metody je cena, která je v porovnání s dalšími metodami vyšší. Na druhou stranu velkou výhodou je, že generuje fragmenty vyšší délky. Za jeden běh přibližné délky 10 hodin generuje přístroj 454 GS-FLX v titanové verzi PTP ~400-600 Mb dat fragmentů délky ~450 bp, s náklady ~\$85/Mb. Uváděná přesnost je 99.5% [32].

4.3.2 Illumina/Genome Analyzer

Metoda je založena na sekvenování syntézou (z angl. sequence by synthesis), která je doplněna o detekci fluorescenčně značených nukleotidů podobající se tradiční Sangerově metodě. Illumina sekvenace používá k amplifikaci fragmentů techniku tzv. “můstkovou” PCR (z angl. bridge). Zpracovávané fragmenty jsou 36-125 bp dlouhé a jsou obohaceny připojením adaptérů, které jsou komplementární k oligonukleotidovým sekvencím kovalentně umístěným na amplifikační destičce. Adaptory jsou spolu s jednovláknovými DNA fragmenty hybridizovány k ukotveným oligonukleotidům a cyklickou izotermickou PCR dochází k vytvoření shluků. Každý shluk se skládá přibližně z 1000 identických molekul. Můstková se tato metoda označuje proto, že při ní dochází ohybu molekuly DNA do podoby mostu. Značené nuklotidy používané pro detekci signálu obsahují modi-



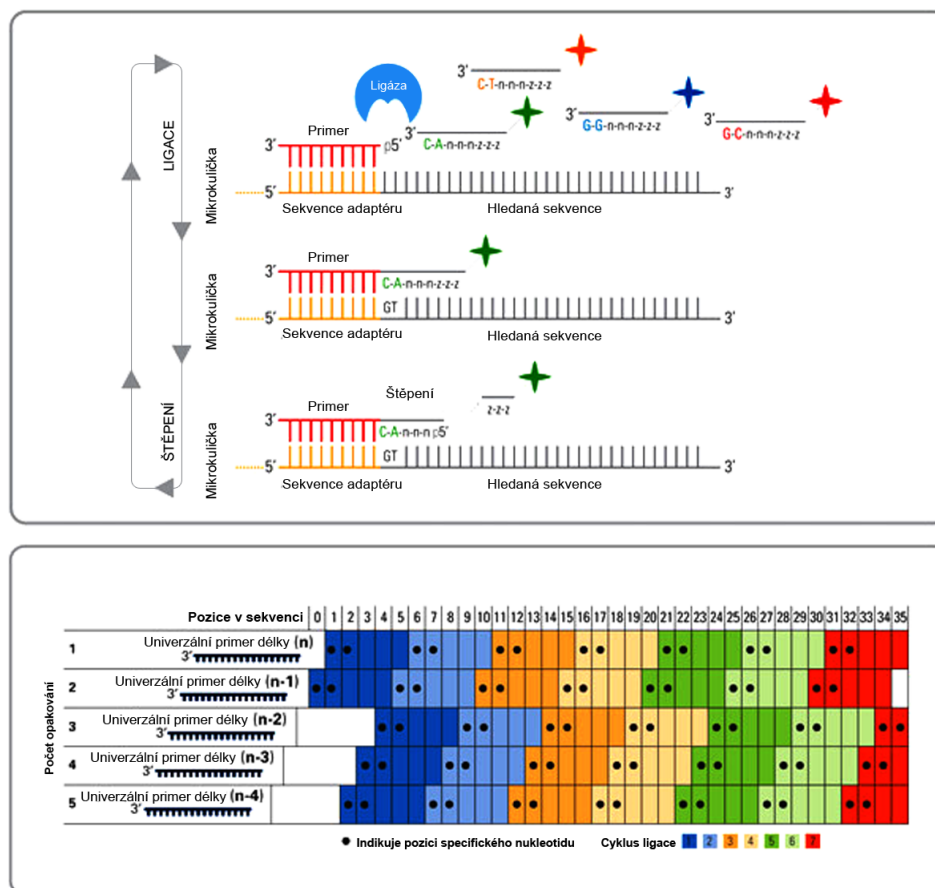
Obrázek 11: Sekvence přístroji Illumina. a) Příprava knihovny-navázání adaptérů na fragmenty. b) Můstková PCR amplifikace, vytvoření shluků a následné navázání primeru. c) Detekce fluorescenčního signálu laserem. Převzato z [32].

fikovanou 3'-OH skupinu, zajišťující zařazení pouze jednoho nukleotidu v jednom cyklu. Každé zařazení jednoho nukleotidu je následované laserovou detekcí fluorescenčního signálu z každého shluku. Při jednom cyklu spolu všechny nukleotidy soutěží o začlenění do řetězce. Po zopakování 125 cyklů vygenerujeme 125 bp dlouho sekvenci.

Společnost Illumina v současnosti dominuje NGS trhu, hlavně díky robustnosti provedení a přesnosti. Illumina Genome Analyzer produkuje za 7 dní běhu přibližně 17 Gb sekvencí délky 75+ bp za cenu ~\$6/Mb. Uváděná přesnost je více jak 99.5% [32].

4.3.3 Applied Biosystems/ SOLiD

Platforma SOLiD (z angl. Sequencing by Oligo Ligation and Detection) je založena na principu sekvenování prostřednictvím ligace (z angl. sequencing by ligation). Příprava knihovny může probíhat dvěma způsoby. Buď se vytváří klasická fragmentová DNA knihovna se dvěma adaptéry na každém konci fragmentu, nebo se produkuje párová (z angl. mate-paired) knihovna, která navíc ke dvěma krajním adaptérům obsahuje ještě jeden interní. Analyzované fragmenty mají délku 50 bp a jsou amplifikovány metodou emulzní PCR, podobně jako tomu bylo u sekvenace 454 Roche. Sekvenování na prin-



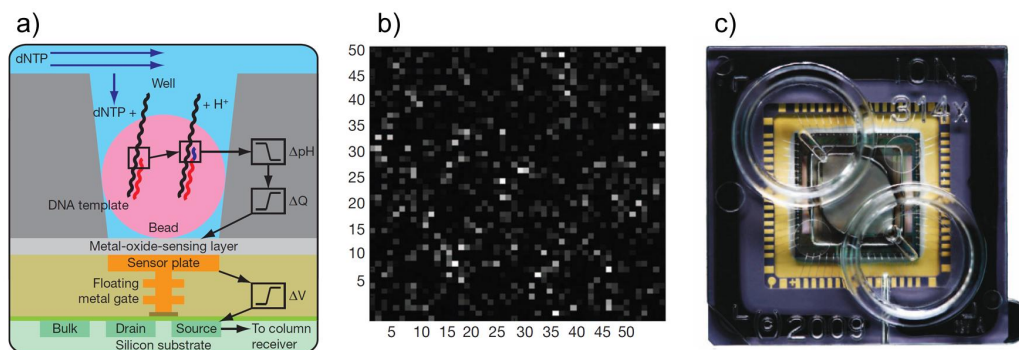
Obrázek 12: Sekvenace přístroji SOLiD. Značené sondy soutěží o navázání za primer, poté následuje štěpení a celý proces se opakuje (nahore). Posouváním primeru o jeden dNTP se restauruje celá sekvence (opakuje se 5x). Převzato z [32].

cipu ligace používá systém nukleotidových sond kódovaných dvěma bázemi. Sondy tvoří oligonukleotidové oktamery, ale specifické jsou pouze první dva nukleotidy. Po navázání primeru je do reakce přidána sada fluorescenčně značených sond, která soutěží o zařazení do řetězce za primer. Po ligaci sondy k primeru je detekován fluorescenční signál a poslední tři nukleotidy jsou odštěpeny, což zanechá na místě sondu pokrývající 5 nukleotidů. Hybridizace sondy, ligace, detekce signálu a štěpení sondy se musí zopakovat 10-krát, aby se sondami pokryl celý 50 bp dlouhý fragment. Pro určení sekvence nukleotidů musíme celý proces opakovat celkem 5-krát, a to vždy s primerem o jeden nukleotid kratším než u předchozího cyklu. Díky systému sond kódovaných dvěma bázemi je každá pozice charakterizovaná dvěma fluorescenčními signály, což dělá SOLiD system velmi přesným.

Největší výhodou této metody je zvýšený výkon a přesnost. SOLiD sekvenátor produkuje za 3-7 dní běhu přibližně 10-15 Gb sekvencí délky 50 bp za cenu ~\$5.80/Mb. Uváděná přesnost je více jak 99.94% [32].

4.3.4 Life Technologies/ Ion Torrent

Technologie Ion Torrent byla představena na začátku roku 2011. Jedná se o první metodu která není založená na detekci uvolněného světelného záření, jako všechny předchozí zmiňované metody, [30]. Místo toho měří ionty H^+ produkované při syntéze řetězce DNA polymerázou, která přidává nukleotidy v jejich základní neupravené podobě. Masivně paralelní sekvenování probíhá na iontově senzitivním tranzistorovém poli ISFET (z angl. Ion Sensitive Field Effect Transistor). Architektura čipu ISFET využívá elektronické adresování běžné v CMOS senzorech. Nad elektronikou a senzitivní vrstvou je umístěna vrstva s jamkami, do kterých jsou rozmístěny mikrokuličky s DNA fragmenty amplifikovanými pomocí emulzní PCR. Čip ISFET je pokrytý 1.2 miliony senzorů (novější Proton II čip dokonce 660 milionů), které zaznamenají každou změnu v gradientu pH způsobenou začleněním nukleotidu a následným uvolněním protonu H^+ . Změna gradientu pH indukuje změnu potenciálu na kov-oxid senzitivní vrstvě, která je dále digitalizována a zpracována. Typický běh trvá 2 hodiny a generuje až ~1Gb dat. Zpracovávají se fragmenty délky ~100bp produkuje přibližně 25 milionů bází.

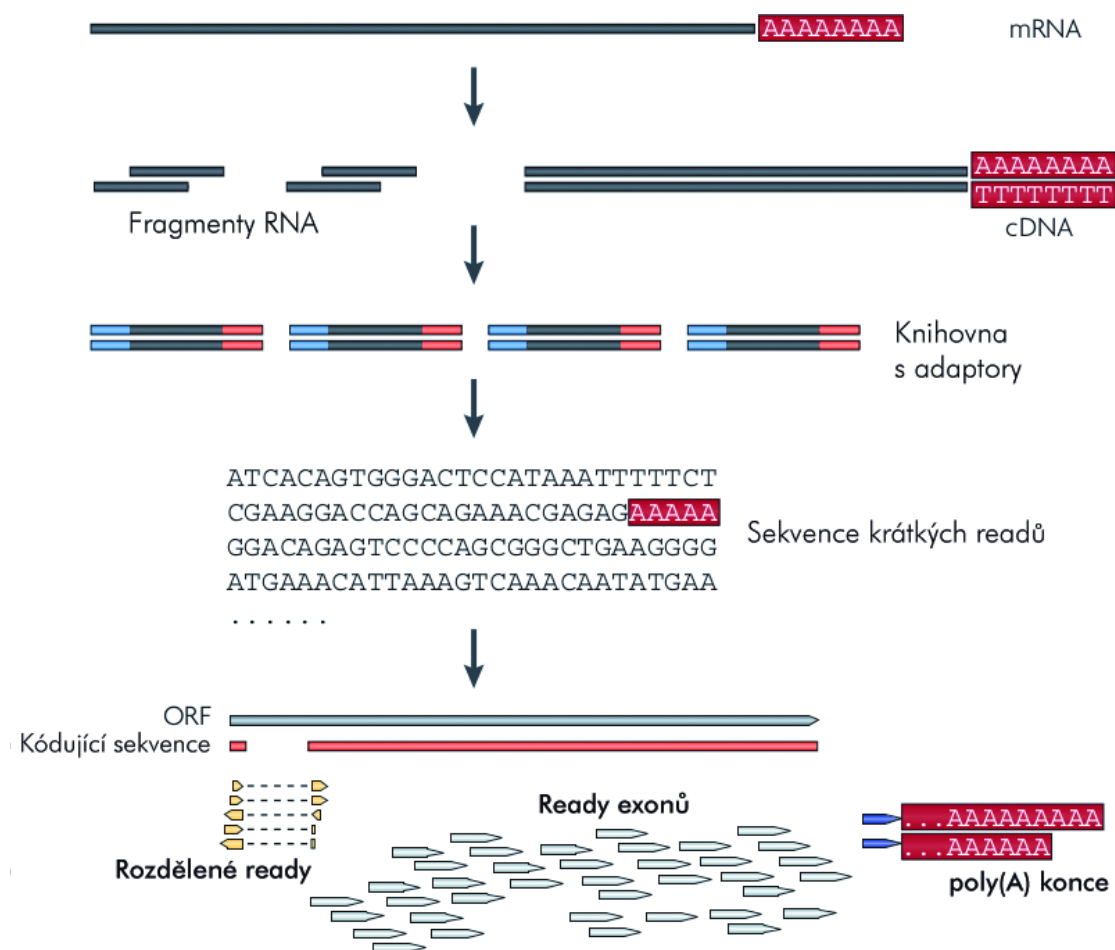


Obrázek 13: Ion Torrent technologie. a) Jamka s DNA fragmentem. b) 50x50 region senzoru. c) Čip v keramickém balení. Převzato z [30].

5 Sekvenace transkriptomu pomocí RNA-seq

Sekvenování mRNA s použitím NGS technologií umožňuje měření genové exprese celého transkriptomu. Postup a provedení RNA-seq experimentu je znázorněn na obr. 14. Prvním úkolem je vyčistit zkoumaný vzorek o rRNA, tRNA a mitochondriální RNA, které u prokaryot i eukaryot tvoří přibližně 75% všech RNA molekul. Navzdory použití purifikačních metod, mezi které patří například poly(A)purifikace a DNS normalizace, sekvenační data mohou obsahovat menší množství těchto RNA molekul [59]. Ty mohou být odfiltrovány v následujících krocích bioinformatickými postupy. Zbývá mRNA je poté nastříhána na menší části, a je z ní připravena knihovna krátkých fragmentů s navázanými adaptory. Ty jsou poté sekvenovány sekvenačním přístrojem a jako výsledek získáme tzv. *reads*. Anglické slovo 'read' značí datovou reprezentaci krátké sekvence DNA obvykle 50-150 bp dlouhou, která byla vyprodukována sekvenačním přístrojem. Samotné *reads* však nemají žádnou vypovídající hodnotu, a proto jsou dále bioinformaticky zpracovány. Namapováním na referenční sekvenci zjistíme jejich genomickou pozici, ze které byly odvozeny. Většina *readů* je namapována na exony, což jsou transkripčně aktivní jednotky, a pouze malé množství *readů* je namapováno na transposony. *Ready* které nejde namapovat v celku, jsou rozděleny na menší části a ty jsou namapovávány zvlášť. Rozdělené *reads* umožňují jednodušší identifikaci mezer mezi exony (angl. splice junctions) [54, 56].

RNA-seq metoda není zatížena téměř žádným šumem, je mnohem přesnější a má větší dynamický rozsah oproti DNA microarray technologiím. Microarray totiž postrádá senzitivitu pro velmi malé, nebo naopak velmi velké úrovně exprese, z důvodu laserové detekce. RNA-seq má nekonečně mnoho úrovní exprese, protože postrádá horní limit pro kvantifikaci exprese. Záleží pouze na počtu namapovaných *readů*. RNA-seq metoda je výhodná pro sekvenaci *de novo*, protože (oproti microarray) není omezena znalostí genomické sekvence. RNA-seq umožňuje objevování nových genů, jejich transkriptů, měření genové exprese a detekci zárodečných i somatických mutací v oblasti transkribované části genomu. RNA-seq se také hodí pro analýzu diferenciální exprese (zdravý/nemocný, léčený/neléčený atd.) z důvodu jednoduché reprodukovatelnosti experimentu. V neposlední řadě je RNA-seq vhodná k analýze alternativního sestřihu mRNA, prostřednictvím kterého jsou generovány odlišné transkripty z jednoho genu. Detekce těchto isoform pomocí mRNA je velmi přesná a odhaluje jak známé, tak nové isoformy.



Obrázek 14: Průběh RNA-seq experimentu. Rozdělené ready jsou namapovány přes nekódující sekvence intronů a umožňují snazší identifikaci mezer mezi exony. Upraveno podle [36].

Z důvodu velkého množství dat produkovaných sekvenačními přístroji (v současnosti více jak 500 Gb na jeden běh) bylo nutné vyvinout robustní a efektivní bioinformatické algoritmy, které takovou analýzu umožní. Obecně můžeme postup při analýze RNA-seq dat rozdělit na tři části. Jedná se o namapování readů, klasifikaci a anotaci transkriptu (angl. Transcript assembly), a také kvantifikaci exprese genů a jejich transkriptů. Existuje několik sofistikovaných nástrojů pro každou z těchto částí. Většinou jsou volně dostupné a jsou vyvíjeny laboratořemi na celém světě. Nástroje poskytující klasifikaci, kvantifikaci a anotaci transkriptu, se obecně zaměřují na genovou expresi a proto je pro analýzu transposonů nebylo možné použít. Ze všech dostupných nástrojů byl použit jeden nástroj sloužící k namapování readů, ale následná kvantifikace úrovně exprese

transposonů je řešena vlastními postupy a skripty.

Celkový objem dat použitých v této práci je v součtu 66.8 GB ve FASTQ formátu, který bude popsán dále. Z důvodu velkého objemu dat a značné výpočetní síle, která je pro tuto práci vyžadována především pro namapování readů, byla veškerá analýza prováděna prostřednictvím výpočetního centra MetaCentrum [40], ke kterému lze přistupovat vzdáleně přes příkazovou řádku. MetaCentrum je vedeno virtuální organizací MetaVO, která poskytuje studentům a akademickým pracovníkům bezplatné využití výpočetní a úložné kapacity, stejně jako řadu programů (například Matlab, Maple, TopHat, Bowtie, Cufflinks, atd.).

Protože anglické názvy jsou v oboru zažitě a jejich počest'ování by bylo matoucí, používám v textu několik anglických termínů, které jsou vždy vysvětleny, např. zmíněné ready.

6 Příprava dat pro namapování

6.1 FASTQ formát

Sekvenační přístroje produkují data ve formátu FASTQ [38], který kromě samotné sekvence obsahuje informaci o kvalitě každého nukleotidu (Q jako quality, angl. kvalita). FASTQ formát se stal standardem pro uchování a sdílení sekvenovaných dat. Existuje však několik variant FASTQ souborů, které se od sebe mírně liší. Všechna data použitá pro tuto diplomovou práci pocházejí z Illumina Genome Analyzeru, a proto zde podrobněji popíši tento typ formátu.

Jak je vidět na obr. 15, každý read je popsán čtyřmi řádky. První řádek obsahuje hlavičku. Podobně jako začátek FASTA hlavičky začíná znakem '>', začátek FASTQ začíná znakem '@'. Hlavička má tři pole, jejichž délka nemá žádný limit. V prvním poli dat pocházejících z NCBI SRA databáze (viz dále) se nachází identifikátor readu. Na obrázku je vidět že na jeho konci je '.1', což značí, že ready tvoří páry (angl. paired-end). Druhý read z páru má na konci '.2'. Na prostředním poli se nachází Illumina identifikátor. 'HWUSI-EAS230-R' je unikátní název přístroje, za kterým následují identifikátory konkrétních souřadnic, kde se sekvenovaný DNA fragment na čipu nacházel. Jsou jimi číslo pruhu (angl. lane) '5', číslo dlaždice (angl. tile) '1', x-souřadnice shluku (angl. cluster) '16' a y-souřadnice '884' uvnitř dlaždice. Poslední pole obsahuje informaci o

```

1.ř.    @SRR057654.1 HWUSI-EAS230-R:5:1:16:884 length=36
2.ř.    GCGGGGCCGGAGCGAGGCTGAGANCNGNNNTTCCT
3.ř.    +SRR057654.1 HWUSI-EAS230-R:5:1:16:884 length=36
4.ř.    AA@AA@>@9<??>-9;#####!#!#!!!!#####

```

Obrázek 15: První 4 řádky FASTQ Sanger/Illumina formátu steženého ze NCBI SRA databáze, popisující jeden read.

délce fragmentu, tedy 'length=36'. Druhý řádek obsahuje samotnou sekvenci readu, jejíž znaky odpovídají jednopísmenným kódům dle IUPAC nomenklatury. Třetí řádek začíná znakem '+' a může obsahovat dodatečné informace. V našem příkladu je zde zopakovaná hlavička. Tento řádek je volitelný může zde být pouze povinný znak '+', čímž dojde ke výraznému snížení velikosti souboru.

6.1.1 Phred skóre

Každé bázi je přístrojem přiřazeno ohodnocení kvality správné identifikace, tzv. Phred skóre. Poslední řádek na obr. 15 obsahuje jednomístný ASCII kód pro Phred skóre každé báze. Z toho důvodu je každý čtvrtý řádek stejně dlouhý, jako délka sekvence na druhém řádku. Existuje několik typů kódování Phred skóre dle druhu přístroje. Použitá data jsou ve formátu Sanger/Illumina1.9. Pro tyto data je použit klasický typ kódování Phred+33. Je důležité si všimnout, že se mezi znaky kvality na čtvrtém řádku může vyskytnout i znak '@' (viz obr. 15). Není proto možné použít řádek začínající na '@' pro identifikaci řádku hlavičky, například pro parser.

Kvalita správné identifikace báze přístrojem je posuzována na základě Phred skóre. K výpočtu kvality Q je použit vzorec z [38]

$$Q = -10 * \log(P_e), \quad (1)$$

kde P_e značí pravděpodobnost, že při volání báze (angl. base call) došlo k chybě. Vyšší skóre tedy implikují nižší pravděpodobnost, že došlo k chybě. Například skóre $Q=10$ odpovídá pravděpodobnosti, že 1 báze z deseti je špatně, tedy 90% přesnost. Pro $Q=20$ je 1 báze ze 100 špatně (přesnost 99%), a pro $Q=30$ je to 1 z 1000 bází (přesnost 99.9%). Sanger/Illumina1.9 formát dokáže zakódovat Phred skóre od 0 do 93 použitím ASCII znaků 33-126:

!"#\$%&'()*+,-./0123456789;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`a

bcdefghijklmnopqrstuvwxyz{|}~.

Většinou se však skóre pohybuje od 0 do 40, a jen zřídka překročí hodnotu 60. Kódování Phred+33 značí, že kvůli posunu vznikajícím při převodu skóre do ASCII, je nutné ke skóre přičíst číslo 33.

6.2 Single-end, paired-end a mate-pair

Knihovnu readů je možné připravit několika způsoby. Prvním je metoda single-end, při které je read vygenerován z jednoho fragmentu DNA. Single-end strategie však má jisté nevýhody, především pokud se jedná o krátké ready pro *de novo* sekvenenci, tedy pro sekvenenci bez známé publikované referenční genomické sekvence. Díky velkým segmentálním duplikacím, repetitívám a jiným málo komplexním oblastem v genomu, je obtížné ze single-end readů vytvořit pro každý chromozom jeden tzv. contig. Anglické slovo 'contig' znamená datovou reprezentaci dlouhé sekvence DNA, která neobsahuje mezery a byla zrekonstruovaná z namapovaných readů. U NGS sekvenace se vždy snažíme o co nejdelší contigy. U single-end knihovny je i pro malé genomy třeba velkého pokrytí (angl. coverage).

Z tohoto důvodu je velmi oblíbená strategie paired-end typu knihovny. U paired-end metody jsou z fragmentu DNA sekvenovány pouze jeho konce. Vzniklé dva ready jsou ve výstupním souboru označeny jako pár (buď jako 'nazev.1' a 'nazev.2', nebo 'nazev/1' a 'nazev/2'). Při tomto typu přípravy knihovny obdržíme nejen informaci o genomické sekvenci, ale navíc získáme i informaci o vzdálenosti ve které se párové ready nachází. Druhá zmíněná informace je užitečná například v případě, kdy jeden read pochází z repetitivní oblasti genomu a je velmi obtížné ho správně namapovat a druhý read obsahuje unikátní sekvenci. Unikátní read poslouží jako 'kotva' při namapování repetitivního readu. Dále je možné pomocí paired-end readů detekovat strukturní varianty u mutovaných genomů, jako jsou inserce, delece a translokace. Pokud se například ready namapují výrazně blíže oproti očekávané vzdálenosti, značí to možný výskyt delece ve zkoumaném vzorku.

Poslední metodou přípravy knihovny je mate-pair, který je podobný paired-end. Liší se v tom, že párové ready nepocházejí z fragmentu dlouhého cca 200-300 bp, ale z mnohem delších úseků od 20 kb do 10 kb. Mate-pair ready se na genom namapovávají v opačné orientaci než paired-end ready. Důvodem je složitější příprava knihovny readů.

Mate-pair ready se používají například pro spojení velkých mezer u *de novo* sekvenace, nebo pro identifikaci velkých strukturních variant.

6.3 Získání dat

Data pro tuto práci byla stažena z NCBI archivu SRA (Sequence Read Archive) [39], kde jsou uložena jak data volně přístupná, tak data s omezeným přístupem. Z databáze SRA byly vybrány dvě studie, které se zabývaly diferencíální analýzou transkripomu normálních a rakovinných buněk. První studie je dohledatelná na stránkách SRA pod kódem SRP002628. Vzorky pocházejí od pacientů s adenokarcinomem prostaty. Zveřejněná data k této studii tvoří dva páry zdravý/nemocný (2 pacienti). Druhá studie s kódem SRP006900 se zabývá adenokarcinomem tlustého střeva. Studii tvoří 10 párů zdravý/nemocný. Ke své práci jsem z nich použila polovinu. Podrobný souhrn jednotlivých dat naleznete v tabulce 1.

SRA databáze neumožňuje stáhnout data z ftp serveru přímo v požadovaném formátu (například '.fastq') kvůli nedostatku finančních prostředků, které by byly potřeba na vývoj nástrojů provádějících konverzi, a poskytuje data pouze ve formátu '.sra'. Z tohoto důvodu byl vytvořen SRAToolkit, což je knihovna nástrojů pro práci se '.sra' soubory. Po stáhnutí '.sra' souboru umožňuje konverzi do požadovaného formátu. Konverze ze '.sra' do '.fastq' u paired-end dat provádí nástroj `./fastq-dump` s volbou `--split-3`, která paired-end data uloží ve dvou samostatných souborech. Například po stáhnutí souboru SRR057658.sra obdržím soubory SRR057658_1.fastq pro ready pocházející z levé strany původního fragmentu a SRR057658_2.fastq pocházející ze strany pravé.

6.4 Zhodnocení kvality dat

NGS data trpí na systematické chyby, které jsou důsledkem nedokonalé přípravy knihovny a sekvenace. Zhodnocení kvality dat je proto důležitým krokem, který by se měl provádět vždy před použitím dat k další práci, aby se zamezilo možným problémům a odchylkám, které by v důsledku mohly ovlivnit biologickou interpretaci. Velmi oblíbený je nástroj pro ohodnocení kvality dat FastQC [42], který umožňuje komplexnější pohled na data prostřednictvím grafického rozhraní. Při načtení dat se provede sada analýz, která poskytne možnost identifikace problémových oblastí, kterými je třeba se při

ID běhu	Původ tkáně	Typ	Počet readů	Délka readů
SRR222176	tl.střevo/rak.	single-end	8542144	65
SRR222178	tl.střevo/rak.	single-end	11461875	65
SRR222175	tl.střevo/norm.	single-end	9037384	65
SRR222177	tl.střevo/norm.	single-end	11308009	65
SRR057639	prostata/rak.	paired-end	31370536	36
SRR057640	prostata/rak.	paired-end	31979850	36
SRR057641	prostata/rak.	paired-end	32614990	36
SRR057642	prostata/rak.	paired-end	30425120	36
SRR057643	prostata/rak.	paired-end	33158138	36
SRR057654	prostata/norm.	paired-end	29523906	36
SRR057655	prostata/norm.	paired-end	29495276	36
SRR057656	prostata/norm.	paired-end	28473964	36
SRR057657	prostata/norm.	paired-end	23829402	36
SRR057658	prostata/norm.	paired-end	29352538	36

Tabulka 1: Přehled analyzovaných datových souborů.

úpravách kvality dat zabývat. Kromě základní statistiky udávající počet readů, obsah GC, typ kódování a délky readů, obsahuje FastQC dalších 9 modulů. Detailní popis všech by byl příliš zdlouhavý, je uveden v referenci[42]. Těmi nejdůležitějšími parametry, kterými je vhodné se zabývat jsou přítomnost adaptorů (jejich případné odstranění), Phred kvalita bází na každé pozici readu, GC obsah, přítomnost duplikovaných sekvencí a průměrná Phred kvalita sekvencí.

Pokud v datech objevíme problém, kterým se rozhodneme zabývat, můžeme využít FASTX-Toolkit [41], což je knihovna nástrojů pro práci s '.fastq' soubory. Obsahuje mnoho užitečných funkcí určených například k ořezání readů, filtraci nekvalitních readů, formátování dat, a jiné. Umí podobně jako FastQC graficky vykreslit základní statistické údaje.

6.4.1 Odstranění adaptorů

Nejdříve je nutné odstranit sekvence adaptorů, které byly sekvenovány spolu s žádanými fragmenty DNA. Pokud známe sekvenci adaptoru, můžeme k tomuto účelu použít nástroj `fastx_clipper` od FASTX-Toolkit, kterému je zadána sekvence adaptoru a on ji z každého readu odstraní. Pokud sekvenci adaptoru neznáme je možné použít modul 'Per Base Sequence Content' od FastQC (obr. 17), který umožní přítomnost a délku adaptoru odhadnout. Tento graf znázorňuje míru bází na každé pozici

readu. U knihovny readů se očekává, že obsah bází bude náhodný a tedy horizontální linie nebudou vykazovat větší odchylky. Na obrázku je vidět že prvních 14 bází vykazuje odchylky, které vypovídají o nadměrném zastoupení určitých sekvencí. Omezení tohoto biasu na začátek sekvence odpovídá sekvenci adaptoru, které je třeba z readu odstranit. K tomu je možné použít nástroj `fastx_trimmer`, kterému stačí zadat rozsah bází, které si přejeme zachovat. Pro všechna použitá data bylo odstraněno prvních 13 bází adaptoru.

6.4.2 Odstranění nekvalitních konců

Vykreslení box-whisker grafu poskytne přehled o rozložení kvality bází na jednotlivých pozicích v readu. Klasická situace kvality bází je znázorněna na obr. 16, kdy s přirůstajícím počtem bází klesá kvalita Phred skóre. Nekvalitní konce readu je proto vhodné odstranit. U readů analyzovaných v této práci byly odstraněny konce u kterých klesla Phred kvalita pod hodnotu 10. Pro upřesnění grafu, modrá křivka znázorňuje průměrnou kvalitu, červené čáry jsou pro medián, žlutý box reprezentuje kvartil pro 25-75% a horní a spodní whiskery představují spodních 10% a horních 90%.

U dat pocházejících z prostaty a z tlustého střeva, byly odstraněny stejně dlouhé konce, aby zůstala délka readů u všech datasetů konzistentní. Kromě odstranění 13 bází adaptoru (14 báze je první bází nového readu), byly u prostaty dále odstraněny poslední 2 báze, i když nekvalitních bází na konci bylo více. Je to z toho důvodu, že minimální délka readu pro namapování programem TopHat (viz dále) je 20 bp, a při dalším zkracování bychom dostali méně a nebylo by možné ready namapovat. U dat pocházejících z tlustého střeva jsme si mohli dovolit odstranit z konce více readů. Bylo odstraněno 22 bp (poslední bází nového readu je 43 báze původního).

Ořezáním jsem tedy ready u tlustého střeva zkrátily na 20 bp a u prostaty na 29 bp. Použili jsme nástroj `fastx_trimmer` pro tlusté střevo v podobě:

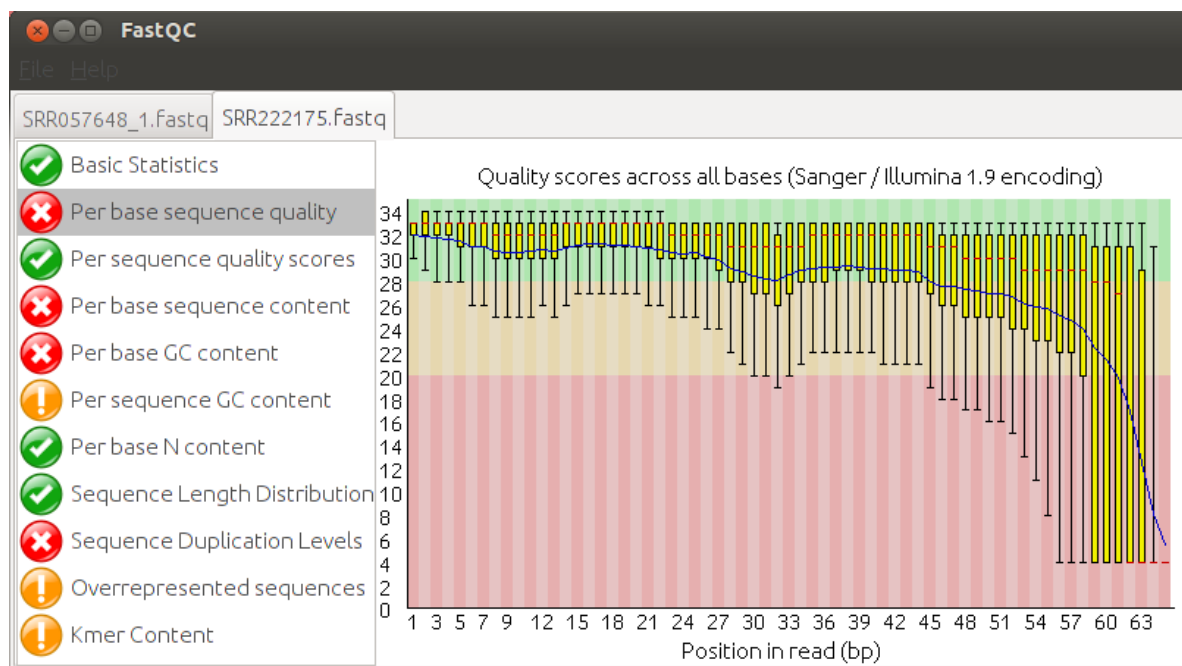
```
fastx_trimmer -f 14 -l 43 -Q 33 -i vstup.fastq -o vystup.fastq
```

A pro prostatu (paired-end):

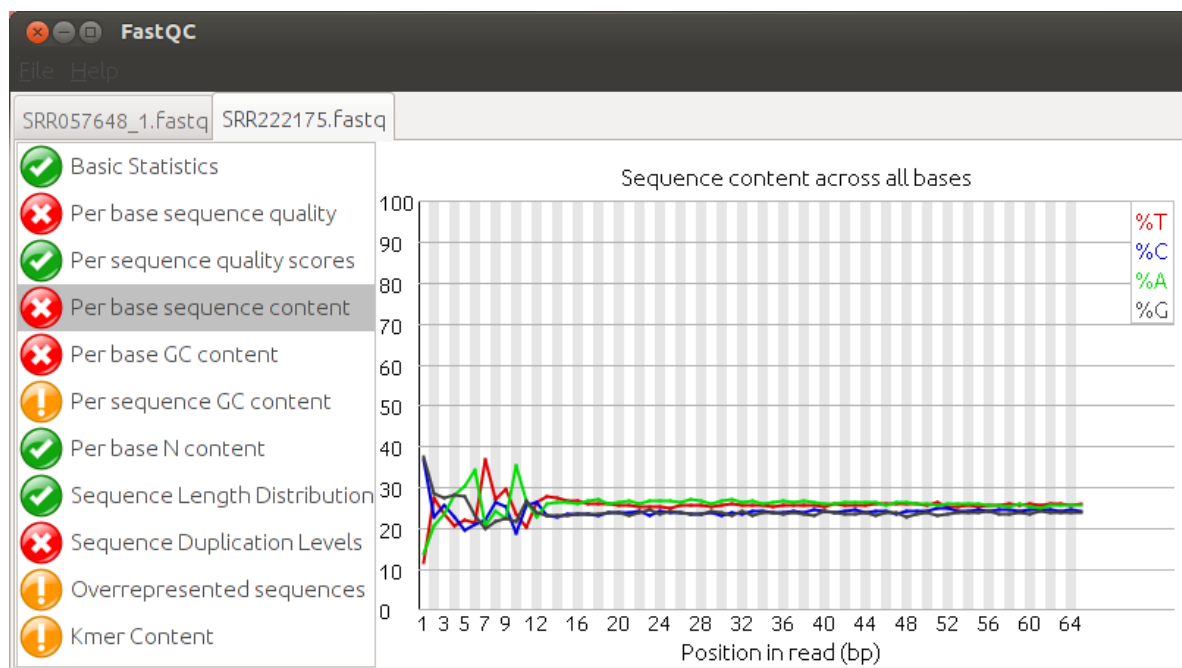
```
fastx_trimmer -f 14 -l 36 -Q 33 -i vstup_1.fastq -o vystup_1.fastq
```

```
fastx_trimmer -f 14 -l 36 -Q 33 -i vstup_2.fastq -o vystup_2.fastq
```

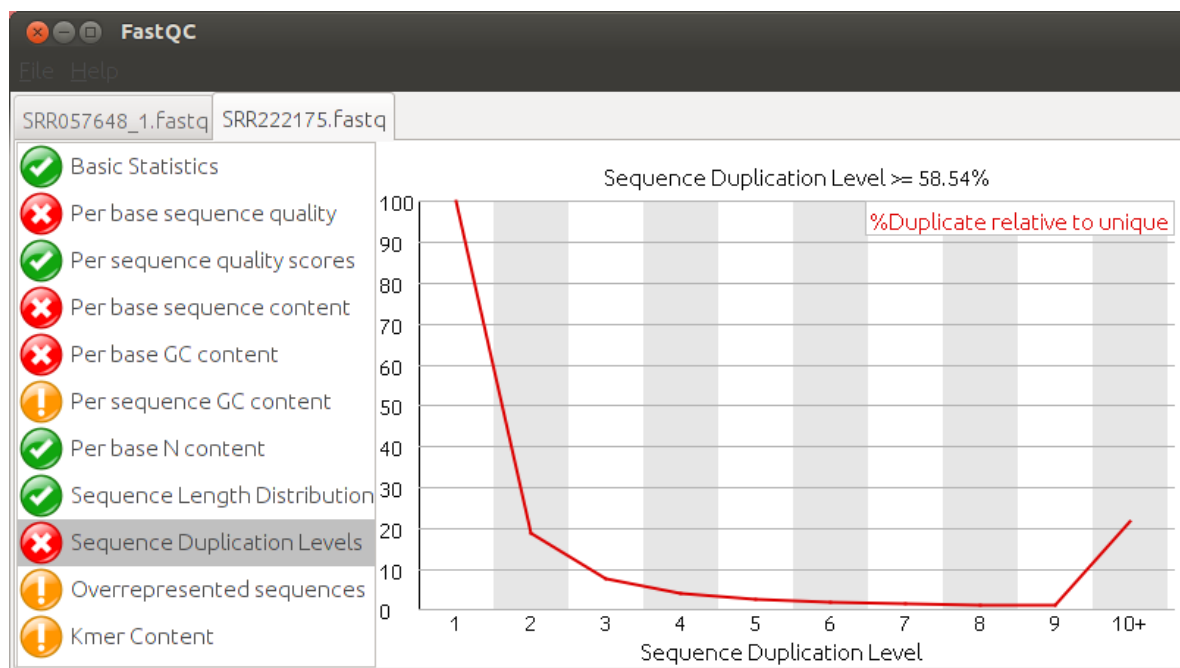
Parametr `-Q 33` nastavuje kódování kvality na Phred+33.



Obrázek 16: FastQC program. Box-whisker graf reprezentující kvalitu bazí.



Obrázek 17: FastQC program. Poměrné zastoupení bazí na každé pozici.



Obrázek 18: FastQC program. Procentuální zastoupení duplikací.

6.4.3 Odstranění duplikací

Většina sekvencí by se v celé knihovně readů měla objevit pouze jednou. Malý počet duplikovaných readů (tj. readů, jejichž stejná sekvence je v knihovně přítomna více než jednou) může být způsoben velmi vysokým pokrytím. Vysoká úroveň duplikace readů může být známkou jejich nadměrného zastoupení, často zapříčiněného PCR amplifikací. Na obr. 18 je procentuálně vyjádřeno zastoupení duplikovaných sekvencí. Je patrné, že většina sekvencí je zastoupena jedenkrát. Rostoucí tendence je přítomna pro 10+ duplikací. Malý nárůst v kategorii 10+ je očekávaný, protože kategorie sdružuje všechny duplikace zastoupení 10 a vícekrát. Podle dokumentace FastQC programu je překročení 20% v kategorii 10+ duplikací pomyslná hranice, kdy by tomuto jevu měla být věnována pozornost a duplikace by měly být odstraněny. Rozhodně ale není na škodu odstranit duplikace při každé analýze automaticky.

Existuje několik nástrojů které lze k tomuto účelu použít. FASTX-Toolkit obsahuje příkaz `fastx_collapse` sloužící pro odstranění duplikací. Ten se však ukázal být velmi pomalý pro použití na lidský genom. Další alternativou je nástroj `rmdup` od SAMTools, který na těchto datech nefungoval. Dokumentace obsahovala informaci o tom, že v některých případech nástroj nemusí pracovat dobře a odkázali uživatele na knihovnu

nástrojů Picard. Picard je knihovna předkompilovaných .jar funkcí. Bohužel ani jejich nástroj `MarkDuplicates.jar` nefungoval podle očekávání. Odstraňoval přibližně polovinu všech readů nehledě na typ dat. Tak velký počet dat neodpovídá grafu na obr. 18 a přišli bychom o velký počet dat. Jednou možností jak si tento jev vysvětlit je to, že oproti předchozím nástrojům na vstup `MakeDuplicates.jar` přichází '.bam' soubor, tedy již namapované ready. Dokumentace k programu provádějícímu namapování (TopHat) udává, že pokud je více readů možné namapovat na více míst, vybírá místo náhodně. Tento efekt spíše naznačuje, že všechny ready jsou namapovány na první pozici na kterou program narazí. Tímto způsobem může docházet k tomu, že na určité genomické pozici (která je dostatečně neunikátní, což repetitivní oblasti jsou) se namapuje převážná většina readů, které jsou poté vyhodnoceny jako duplikace a program Picard je odstraní.

Úroveň duplikace u použitých dat se pohybovala kolem 20%. Proto byl tento krok kvůli výše zmíněným komplikacím přeskočen.

7 Namapování readů

7.1 Přehled metod pro namapování readů

Namapovat ready znamená nalézt v genomu původní pozici, ze které byl read odvozen. Existují dva hlavní přístupy. První přístup při namapování readů nepovoluje velké mezery (angl. *unsplice read aligners*) a používá se například u genomických dat. Mezi nejpoužívanější programy tohoto typu patří velmi známé BWA, Bowtie (využívající Burrowsovu wheelerovu transformaci) a Shrimp (fungující na principu Hash tabulek). Druhý přístup, navržený pro analýzu transkriptomu, velké mezery povoluje (angl. *spliced aligners*). Mezi hlavní dvě metody realizující tento přístup patří *exon first* (*exon první*) a *seed and extend* (*zasej a prodluž*). Pravděpodobně nejoblíbenější nástroj používaný pro namapování dat z RNA-seq je TopHat. Metoda *exon first* se skládá ze dvou kroků. Jak její název napovídá, nejdříve se namapují ready, které jde namapovat bez přerušení v celku. V druhém kroku se nenamapované redy rozdělí na menší fragmenty, které jsou namapovány zvlášť. Následně se genomická sekvence kolem namapovaných fragmentů readů prohledá o možná spojení exonů (angl. *splice connection*). Rozdělené ready umožňují jednodušší identifikaci mezer mezi exony (angl. *splice junctions*). Algoritmus *seed and extend* například používají programy GSNAP a QPALMA. [55]

V porovnání se *seed-extend* metodou je *exon first* rychlejší a méně výpočetně náročnější, zejména pokud se v prvním kroku namapuje pouze malá část readů (druhý krok je více výpočetně náročnější). V této práci jsou ready namapovány pomocí programu TopHat, který k prvnímu 'předmapování' využívá nástroj Bowtie, a až potom identifikuje možná spojení. Namapování readů na referenční sekvenci je výpočetně nejnáročnější část diplomové práce.

7.2 Burrowsova Wheelerova transformace (BWT)

Program Bowtie nevyhledává ready přímo v genomu, ale vytvoří si indexovanou verzi genomické sekvence pomocí Burrowsovy wheelerovy transformace (BWT). Metoda BWT data transformuje (tzn. nedochází ke ztrátě informace) do struktury, která je mnohem vhodnější pro kompresi. Samotná komprese poté může být realizována například jednoduchým RLE (Run Length Encoding), nebo MTF (Move To Front) algoritmem. Transformace pomocí BWT je vratný proces.

1	ema_ma_maso	1	_ma_masoema
2	ma_ma_masoe	2	_masoema_ma
3	a_ma_masoem	3	a_ma_masoem
4	_ma_masoema	4	a_masoema_m
5	ma_masoema_	5	asoema_ma_m
6	a_masoema_m	6	ema_ma_maso
7	_masoema_ma	7	ma_ma_masoe
8	masoema_ma_	8	ma_masoema_
9	asoema_ma_m	9	masoema_ma_
10	soema_ma_ma	10	oema_ma_mas
11	oema_ma_mas	11	soema_ma_ma

Tabulka 2: BWA transformace. Tabulka vlevo obsahuje řetězce s posunutým začátkem, v tabulce vpravo jsou řádky uspořádány podle abecedy a jsou u nich zvýrazněny první a poslední sloupec.

Jako příklad poslouží řetězec 'ema_ma_maso'. Účinnost BWA však roste s délkou vstupního řetězce, tudíž u takto krátké věty neočekáváme výrazné zlepšení. Nejdříve si vytvoříme N opakujících se řetězců (kde N je zároveň délka řetězce), které jsou posunuty způsobem uvedeným v seznamu vlevo (viz levý sloupec tabulka 2). Následně jsou řetězce lexikograficky setříděny. Výstupem BWA je řetězec složený z posledních písmen (zeleně v tabulce 2) a z indexu řádku s původními daty, tedy 'aammoe__sa6'. Seskupení stejných znaků u sebe je důvodem účinné komprese.

Pro dekompresi je nutné sestavit transformační vektor. To jest vektor čísel o délce N, který určuje posloupnost písmen v původním řetězci. Náš transformační vektor je '8 9 1 2 11 7 3 4 5 6 10'. Známe index řádku prvního písmena, tedy 6. V tabulce na 6. řádku je písmeno 'e' (červeně v tabulce 2), které je prvním znakem hledaného řetězce. Na šesté pozici v transformačním vektoru je číslo 7, odpovídající písmenu 'm' (červeně). Dekodovali jsme tedy první dva znaky řetězce 'em'. Dále na sedmé pozici ve vektoru je číslo 3, tedy nám přibude písmeno 'a' v tabulce. Tímto způsobem se dekóduje celá zpráva 'ema_ma_maso'.

Transformační vektor lze vytvořit podle jednoduchého pravidla. Začínáme na prvním řádku v pravém sloupci v tabulce 2 a hledáme první výskyt symbolu prvního červeného sloupce (tedy '_') v posledním zeleném sloupci. Symbol '_' se nachází na osmém řádku. Do transformačního vektoru proto napíšeme jako první číslo 8. Pokračujeme druhým řádkem, kde se opět nachází symbol '_'. Najdeme ho v posledním sloupci na deváté pozici. Napíšeme proto do transformačního vektoru další číslo 9. Tímto způsobem pokračujeme pro celý

vektor. Symboly jsou v tabulce uspořádány podle abecedy, proto vždy platí pravidlo prvního výskytu.

7.3 Strategie pro multi-ready

Program se při namapování často dostane do situace, kdy najde více vhodných genomických pozic pro jeden read. Tyto ready nazýváme multi-ready. Výběr vhodné strategie pro práci s multi-ready patří mezi největší výzvy sekvenace, především sekvenace repetitivních oblastí, které jsou dost nejednoznačné z hlediska namapování readů. Existují tři možnosti jak s takovými ready naložit [34]. První možností je, že tyto ready z analýzy úplně vyřadíme. Druhou možností je vybrat pouze jednu pozici pro namapování s nejméně neshodami (angl. mismatches). Pokud je více stejně vhodných pozic, může program vybrat jednu náhodně, nebo nahlásit všechny stejně vhodné. Třetí možností je nechat program nahlásit maximální počet pozic d , nehledě na celkový počet nalezených pozic. Další variantou této třetí možnosti je ignorovat ready, které se namapují více jak d krát. Paired-end knihovna do jisté míry řeší problém s multi-ready, jelikož se oba konce fragmentu DNA, ze které ready pocházejí, nacházejí ve známé vzdálenosti. Informace o tom, jak daleko by se ready od sebe měly nacházet je poté využita pro výběr správného namapování.

Pokud se rozhodneme multi-ready ignorovat a odstranit je, omezíme analýzu pouze na unikátní oblasti v genomu. Protože se práce věnuje repetitivním oblastem u kterých se předpokládá že budou zdrojem mnoha multi-readů, není tato strategie vhodná. Aktivita (exprese) transposonů bude měřena podle počtu readů namapovaných na element. Proto ani povolení namapování readu na více pozic by nebylo vhodné, neboť by do měření aktivity zanášelo chyby. Z toho důvodu byla zvolena možnost, kdy je pro každý read vstupující do analýzy nahlášena pouze jedna nejvhodnější pozice. TopHat pro to používá volbu `-g/--max-multihits <int>` [35], která ho instruuje povolit jen zvolený maximální počet zarovnání pro jeden read. Tento parametr je tedy nastaven na `-g -1`. Pokud TopHat narazí na více stejně dobrých pozic, vybere jednu náhodně. Více v dokumentaci TopHat [46].

7.4 SAM formát

Soubory s příponou '.sam' jsou produktem většiny nástrojů pro namapování readů (TopHat, SpliceMap, BWA, Bowtie, atd.) a také jsou obecně přijímaným vstupem pro následnou analýzu exprese (Cufflinks). SAM formát byl vyvinut z důvodu sdílení dat napříč laboratořemi a pro ukládání velkého objemu sekvenačních dat, pro které bylo nutné vytvořit standardizovaný formát. Tím se stal textový, tabulátorem oddělený SAM formát (z angl. Sequence Alignment Map format), respektive jeho binárně komprimovaná verze BAM (z angl. Binary Alignment Map format). Specifikace SAM formátu najdete v referenci [45]. Existuje také mnoho volně dostupných nástrojů pro práci s těmito soubory. Nejpopulárnějšími jsou pravděpodobně SAMTools [33] a Picard tools [43].

SAM soubor se skládá z hlavičky, která není povinná, a samotného namapování. Pokud je hlavička přítomna, nachází se zpravidla před zarovnáním. Každý řádek hlavičky začíná znakem '@' a dodržuje formát 'značka: hodnota', kde 'značka' je dvouznakový řetězec definující obsah a formát informace nacházející se na řádku. Každý řádek hlavičky tedy musí začínat jedním z následujících kódů: @HD (z angl. header), @SQ (z angl. sequence, myšleno referenční sekvence), @RG (z angl. read group), @PG (program), @CO (comment). Kromě značky @CO jsou řádky hlavičky rozděleny tabulátorem (angl. tab-delimited). V ukázce na obr. 19 jsou přítomny pouze značky @HD a @SQ. Informace na každém řádku hlavičky dodržují specifické uspořádání, jehož popis by byl nad rozsah této práce, a proto je zde rozebrán pouze formát hlavičky ukázkového souboru na obr. 19. Za značkou @HD následuje informace o verzi SAM formátu ('VN:1.3', aktuální verze je 1.4 ze 7.září 2011) a o způsobu seřazení ('SO: coordinate', ready jsou seřazené podle jejich souřadnic). Za @SQ je uvedené jméno referenční sekvence ('SN:ref') a její délka ('LN:45'). Pokud se počítalo s více referencemi, např. pro každý chromozom zvlášť, musí mít každý @SQ řádek unikátní SN označení.

Za hlavičkou následuje část samotného namapování readů na referenční sekvenci. Zde se každý řádek vztahuje k namapování jednoho readu a obsahuje všechny informace, které byly nasbírány sekvenačním přístrojem. Tyto informace jsou rozděleny do 11-ti povinných sloupců, u kterých je vždy dodržováno pořadí. Pokud je některá z jedenácti informací nedostupná, musí být v souboru nahrazena '0' nebo '*' (podle typu). Řádky začínají názvem konkrétního readu, QNAME (z angl. query name). Pokud mají ready stejné QNAME, patří do stejného páru readů (mate-paired/paired-end).

```

@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *

```

Obrázek 19: Ukázka části jednoduchého SAM souboru. Převzato z [45].

Sl.	ID	Typ	Rozsah/Regexp	Krátký popis
1	QNAME	String	[!-?A-~]{1,255}	Jméno dotazovaného readu
2	FLAG	Int	[0,2 ¹⁶ -1]	Binárně zakódovaná vlajka
3	RNAME	String	* !-()+-<>~ !-~*	Název referenční sekvence
4	POS	Int	[0,2 ²⁹ -1]	Počáteční (levá) pozice readu
5	MAPQ	Int	[0,2 ⁸ -1]	Kvalita namapování
6	CIGAR	String	* ([0-9]+MIDNSHPX=))+	CIGAR značka
7	RNEXT	String	* !-()+-<>~ !-~*	Jméno dalšího/mate readu
8	PNEXT	Int	[0,2 ²⁹ -1]	Pozice dalšího/mate readu
9	TLEN	Int	[-2 ²⁹ +1,2 ²⁹ -1]	Vypozorovaná délka reference
10	SEQ	String	* [A-Za-z=.]+	sekvence readu
11	QUAL	String	[!-~]+	ASCII kód Phred kvality + 33

Tabulka 3: Typy informací o namapování readu, odpovídající jednotlivým sloupcům v SAM formátu. Upraveno podle [45]

Následuje vlajka FLAG, ve které jsou binárně zakódované různé doplňující informace: má-li například read více segmentů (paired-end/mate paire), jestli je správně namapován, nebo není nenamapován, zda-li jde o primární nebo sekundární namapování, jestli prošel, nebo neprošel kontrolou kvality a v neposlední řadě jestli se jedná o duplikaci vzniklou při PCR reakci. Výpis všech je uveden v tabulce 3. Zkontrolovat význam vlajky je možné interaktivně prostřednictvím k tomu určené aplikace na oficiálních stránkách Picard [44]. Pro první a poslední read v ukázce SAM souboru na obr. 19 zjistíme informaci zakódovanou ve vlajce následujícím způsobem. Vlajku 163 můžeme rozložit jako 1+2+32+128, kdy 1 znamená že read pochází z páru (paired-read/mate-pair), 2 že oba ready z páru jsou správně namapované, 32 že druhý read z páru je namapován na obráceném '-' řetězci a 128 znamená že read je druhým readem z páru (viz tabulka 4). S tímto readem (s vlajkou 163) tvoří pár read s vlajkou 83 o čemž

Bitově	Decimálně	Popis
0x1	1	read pochází z páru
0x2	2	read je v páru správně namapován
0x4	4	read je nenamapován
0x8	8	druhý z páru/mate je nenamapován
0x10	16	read je namapován na '-' řetězci
0x20	32	druhý z páru/mate je namapován na '-' řetězci
0x40	64	read je prvním z páru
0x 80	128	read je druhý z páru
0x100	256	namapování není primární
0x200	512	neprošel kontrolou kvality
0x400	1024	read je PCR nebo optická kopie

Tabulka 4: Význam jednotlivých bitů u vlajky FLAG v SAM formátu.

napovídá i stejné jméno 'r001'. Číslo 83 můžeme rozepsat jako 1+2+16+64, kde 16 značí, že jde o read namapovaný na obrácený '-' řetězec a 64, že jde o první read z páru.

Za FLAG vlajkou následuje pole se jménem referenční sekvence RNAME, které pokud je uvedeno (není místo něj '*'), musí být stejné jako některé ze jmen uvedené v @SQ části hlavičky. Nenamapovaný read má na tomto poli '*'. Následuje pole POS, které informuje o pozici první namapované báze, myšleno zleva a indexováno od čísla 1. Pro nenamapované ready toto pole obsahuje '0'. Kvalita namapování MAPQ, uvedená v pátém sloupečku je rovna Phred skóre. Hodnota 255 říká že MAPQ není dostupné.

V šestém sloupečku se nachází CIGAR značka. Tento řetězec znaků v sobě kóduje informaci o inzercích/delecích namapovaného readu oproti referenci a také o bázích, které se namapovaly správně a které špatně (match/mismatch). Například takto namapovaný read

```
Pozice reference: 12345678901234 5678901234567890123456789012345
Reference:      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
Read:          TTAGATAAAGGATA*CTG
```

koresponduje s 8M2I4M1D3M CIGAR značkou, která značí, že prvních osm bází je namapováno na referenci, další dvě báze na referenci neexistují, čtyři báze se opět namapovaly, jedna báze neexistuje na readu a nakonci jsou 3 správně namapované báze.

Další sloupeček RNEXT udává jméno referenční sekvence, ke které se další paired-end/mate-pair read namapoval. Znak '=' znamená, že jméno je stejné jako RNAME

a znak '*' je uveden pokud tato informace není dostupná. Osmý sloupeček PNEXT informuje o pozici dalšího paired-end/mate-pair readu. Hodnota je nastavena na '0' pokud je informace nedostupná. Devátý sloupeček TLEN je tzv. vypočítaná délka reference. TLEN značí délku, kterou pokrývá daná skupina readů (paired-end/mate-pair) počítaná od první báze readu namapovaného nejvíce vlevo k bázi posledního readu namapovaného nejvíce vpravo. Pokud je číslo kladné, jde o read umístěný nejvíce vlevo na referenci. V případě záporného čísla jde o read umístěný nejvíce vpravo. Pokud read nepochází z páru, nebo je informace nedostupná, je tato hodnota opět nastavena na '0'.

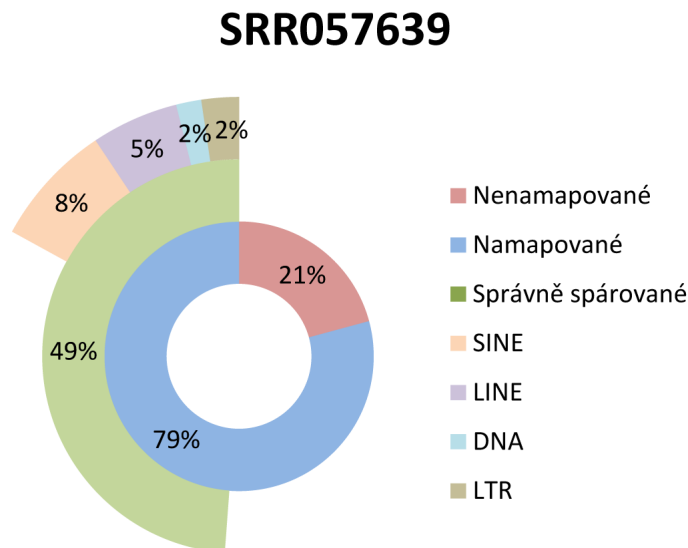
7.5 Nastavení a implementace TopHat

Program TopHat je jedním z nejpoužívanějších nástrojů pro namapování RNA-seq NGS dat a identifikaci mezer mezi exony (angl. splice junctions). Umožňuje nastavit celou řadu parametrů, jejich kompletní výčet je k nalezení v dokumentaci [46]. Protože se TopHat stále vyvíjí a ladí, přibližně každé 3 měsíce vychází nová verze programu. Hlavním výstupem z programu je BAM soubor, binárně zakódovaná verze SAM souboru (viz kapitola 7.4), která obsahuje namapované ready. TopHat vrací nenamapované ready zvlášť v dalším souboru. Tyto dva BAM soubory doplňuje ještě několik dalších BED souborů (např. se souřadnicemi exonových mezer, atd.), které však pro tuto práci nejsou podstatné. Pro práci byl použit TopHat verze 2.0.8 využívající program Bowtie2, verzi 2.1.0. Bowtie2 obsluhuje první část namapování readů, než dojde k jejich rozdělení na části (viz kapitola 7.1 Přehled metod pro namapování readů). Volání TopHatu je následující:

```
TopHat [options]* <index_base> <reads1_1[,...,readsN_1]>
[reads1_2,...readsN_2] ,
```

kde [options]* jsou volitelné nastavení, <index_base> je tzv. 'basename' (název bez přípon) indexovaného referenčního genomu pro Bowtie2. Na konci příkazu jsou zadány '.fastq' soubory s daty. Pokud jde o paired-end ready, musí být vstupní data zakončena *_1 a *_2. Při použití více datasetů (například typu zdravý/nemocný) musí být tyto datasety odděleny čárkou. Indexovaný referenční genom hg19 byl stažen spolu s anotací na stránkách UCSC [47] Dále je dostupný také na stránkách Bowtie2 [48].

Pro práci bylo použito následující nastavení programu (první pro single-end, druhý pro paired-end).



Obrázek 20: Ukázka rozdělení namapovaných readů. Pro ostatní datasety jsou tyto rozdělení uvedena v příloze.

```
TopHat -p 4 --max-multihits 1 --no-coverage-search -G rmsk.gtf genome
single.fastq
```

```
TopHat -p 4 --max-multihits 1 --no-coverage-search -G rmsk.gtf genome
paired_1.fastq paired_2.fastq
```

První parametr `-p/--num-threads <int>` nastavuje paralelní zpracování ve 4 vláknech. Druhý parametr `-g/--max-multihits <int>` nastavuje maximální počet namapování pro jeden read. Standardně je tento parametr nastaven na hodnotu 20. Pro účely diplomové práce byl parametr nastaven na hodnotu 1, aby bylo po namapování možné z počtu namapovaných readů počítat míru exprese pro transposony. Parametr `--no-coverage-search` vypíná hledání mezer mezi exony, které je založené na hustotě pokrytí ready. Toto hledání je časově náročné a pro naše účely redundantní. Čtvrtý parametr `-G/--GTF <GTF/GFF3 file>` poskytuje TopHatu soubor s anotací, ve formátu GTF 2.2, nebo GFF3. Pokud byla anotace poskytnuta, bude se TopHat přednostně snažit namapovat ready na souřadnice definované v anotaci. Následuje basenome indexu referenčního genomu `genome`. Nakonci příkazu jsou uvedeny vstupní FASTQ soubory.

Průměrně se programu TopHat podařilo namapovat 84% readů. Většina nenamapovaných readů byla pravděpodobně vyřazena při dodatečné kontrole kvality, kterou provádí TopHat. Přibližně 43% z namapovaných paired-end readů bylo správně spárováno

(tedy označeno jako angl. properly-paired, což odpovídá vlajce 0x2 ve FLAG vlajce) a 13% paired-end readů bylo namapováno jako singleton, tedy byl namapován pouze jeden read z páru. Ze zbylých readů sice byly namapovány oba z páru, ale ne správně podle nastavených parametrů. Tyto ready se mohly namapovat mimo očekávaný rozsah, nebo také mohlo dojít k namapování párového readu na jiný chromozom. Informace o statistickém rozdělení readů byly získány z BAM souborů použitím nástroje **flag-stat** od SamTools a vlastními awk skripty. V příloze je k nalezení celková statistika namapovaných readů, ukázka pro data SRR057639 je na obr. 20.

8 Analýza aktivity transposonů

Zatímco analýza exprese genů je oblastí, pro kterou existuje řada sofistikovaných nástrojů, analýza aktivity transposonů se dostala do popředí až v posledních letech a existuje málo specializovaných nástrojů k tomuto určených. Z tohoto důvodu bylo třeba vytvořit vlastní postup. Na druhou stranu, odhad exprese genů vyžaduje (oproti odhadu exprese transposonů) složitější postup, u kterého musíme brát v úvahu produkci různých isoform jednoho genu. Isoformy vznikají díky alternativnímu stříhání mRNA transkriptu. To do značné míry komplikuje proces výpočtu aktivity transkriptu, protože read namapovaný na sdílený exon může pocházet z více isoform. Cufflinks je nejpoužívanějším nástrojem, který odhaduje míru exprese transkriptu statistickým zpracováním [57, 58]. Transposony alternativnímu splicingu nepodléhají, a proto není takového zpracování potřeba.

Základem výpočtu pro odhad exprese transposonů je počet readů, které se na transposony namapovaly. Pokud by se ale TE elementy v genomu navzájem překrývaly, mohly by se některé namapované ready počítat vícekrát a do analýzy by bylo zanášeno zkreslení. Z tohoto důvodu byla použita anotace repetitivních sekvencí z programu RepeatMasker [49] testována na poměr překrývajících se elementů. Pokud bychom z anotace vybrali pouze transposony (tedy všechny elementy zařazené do jedné z tříd LINE, SINE, DNA, LTR), které v součtu pokrývají 1 398 596 178 bp, z čehož 1 342 971 bp je překrývajících se s jinými elementy, tedy přibližně 0.096%. Přítomnost překrývajících se transposonů může být zanedbána, protože nezanáší do analýzy výraznou chybu.

K výpočtům prováděným v této části práce byly použity nástroje BedTools [50], SamTools [51] a unixový programovací jazyk awk. Typy readů (single-end a paired-end) si vyžadovaly odlišný přístup k následující analýze aktivity transposonů. Míra exprese byla počítána jak pro jednotlivé rodiny (viz příloha C), tak pro celé třídy transposonů.

8.1 Normalizace RPKM a FPKM

Celkový počet readů v RNA-seq knihovně pocházejících z jednoho transkriptu je přímo úměrný množství transkriptu přítomného ve vzorku. Pokud ale potřebujeme porovnávat úroveň exprese transkriptu mezi dvěma vzorky, nebo úroveň exprese dvou transkriptů u jednoho vzorku, je nutné provést normalizaci dat. Normalizace je prováděna dvěma

způsoby [37]. Prvním způsobem je normalizace na délku transkriptu. Pokud se na delší transkripty namapovalo více readů než na ty krátké, neznámá to nutně, že u nich pozorujeme vyšší expresi. Mějme například dva transkripty A a B, které jsou ve vzorku stejně hojně zastoupeny, ale transkript B je dvakrát tak delší než transkript A. Výsledná knihovna bude obsahovat dvakrát tolik readů pocházejících z B jak z A. Druhým způsobem je normalizace na celkový počet readů. Při porovnání exprese u dvou vzorků, přičemž jeden má 25 milionů readů a druhý 50 milionů, je jasné že, i při reálně stejné expresi transkriptu, bude u druhého vzorku po namapování exprese zvýšená, protože obsahuje 2x více vstupních dat.

Aby nedocházelo ke zkreslování výsledků, byla zavedena metrika nazvaná RPKM (angl. Reads Per Kilobase of Transcript Per Million Mapped Reads), která normalizuje data jak z hlediska délky transkriptů, tak z hlediska velikosti knihovny readů. Pro paired-end sekvenování se používá podobná metrika FPKM (angl. Fragments Per Kilobase of Transcript Per Million Mapped Reads). Tento přístup normalizace transkriptů u RNA-seq byl zaveden Mortazaviho skupinou [35].

RPKM se vypočítá jako

$$RPKM_i = \frac{C_i * 10^9}{L_i * N}, \quad (2)$$

kde C_i značí počet readů namapovaných na transkript, L_i počet bází transkriptu a je celkový počet readů v experimentu. Za N by bylo možné dosadit počet readů vyprodukovaných sekvenačním přístrojem, nebo i počet readů prošlých kontrolou kvality. V této práci bylo za N dosazen počet správně namapovaných readů TopHatem. Pokud bychom ale v jednom případě pracovali s nekvalitními daty, ze kterých po kontrole kvality a namapování zůstal jen zlomek, a výsledek srovnávali s kvalitními daty, jejichž počet se v průběhu analýzy tak razantně nezmenšuje, zanášelo by to do výpočtu zkreslení. Je proto vhodné volit dosazení za hodnotu N dle konkrétního experimentu a v závislosti na typu analyzovaných dat.

Vzorec pro výpočet FPKM zůstává stejný, pouze místo readů dosazujeme vždy informace vztahující se k celému DNA fragmentu, ze kterého byly paired-end ready odvozeny. Tedy C_i značí počet fragmentů namapovaných na transkript a L_i délku fragmentů. Celkový počet fragmentů je pro paired-end ready roven $N/2$.

8.2 Úprava anotace

Anotace použitá při namapování programem TopHat nevyhovovala výpočtům rozdílné exprese z namapovaných readů, protože neobsahovala klasifikaci transposonů do tříd a do rodin. Z tohoto důvodu byl soubor s anotací upraven tak, aby klasifikace byla v souboru zahrnuta. Původní anotace pochází z databáze programu RepeatMasker, který vyhledává repetice a jiné málo komplexní sekvence, viz [49]. Použitý soubor 'rmsk.gtf' byl stažen z UCSC Table Browseru [47] a kromě transposonů obsahuje i jiné repetitivní sekvence lidského genomu (pseudogeny, jednoduché repetice, satelity, aj.) Nepotřebné repetitivní sekvence jsou z analýzy vyřazeny. Stažená anotace je velká 545 MB a její formát vypadá následovně.

```
$ head -5 rmsk.gtf
chr1 hg19_rmsk exon 16777161 16777470 2147.000000 + . gene_id "AluSp"; transcript_id "AluSp";
chr1 hg19_rmsk exon 25165801 25166089 2626.000000 - . gene_id "AluY"; transcript_id "AluY";
chr1 hg19_rmsk exon 33553607 33554646 626.000000 + . gene_id "L2b"; transcript_id "L2b";
chr1 hg19_rmsk exon 50330064 50332153 12545.000000 + . gene_id "L1PA10"; transcript_id "L1PA10";
chr1 hg19_rmsk exon 58720068 58720973 8050.000000 - . gene_id "L1PA2"; transcript_id "L1PA2";
```

Tento soubor je ve formátu GTF a skládá se z osmi sloupečků definujícími různé informace o elementu a jednoho sloupečku s atributy. Tyto sloupečky jsou seřazené způsobem <sekvence> <zdroj> <vlastnost> <začátek> <konec> <skóre> <vlákno> <čtecí rámec> [atributy] [komentáře] a jsou navzájem odděleny tabulátory. Podrobná specifikace GTF formátu a jeho dalších verzí je k nalezení na stránkách [52].

Jak je vidět, anotace obsahuje rozdělení repetitivních sekvencí podle jména, ale chybí rozdělení do tříd a do rodin. Nová anotace byla vytvořena pomocí awk skriptu a z důvodu snížení velikosti souboru byly nadbytečné sloupečky nahrazeny "." Upravená anotace v úsporné verzi vypadá následovně.

```
$ head -5 rmskNEW
chr1 . exon 16777161 16777470 . + . AluSp;SINE;Alu
chr1 . exon 25165801 25166089 . - . AluY;SINE;Alu
chr1 . exon 33553607 33554646 . + . L2b;LINE;L2
chr1 . exon 50330064 50332153 . + . L1PA10;LINE;L1
chr1 . exon 58720068 58720973 . - . L1PA2;LINE;L1
```

Poslední devátý sloupeček obsahuje informace pro klasifikaci elementu do skupin, které jsou ve formátu < název repetice>; <třída repetice>; <rodina repetice> . Záznam pro devátý sloupeček byl opět stažen z UCSC Table Browseru [47].

8.3 Postup pro single-end knihovnu

K analýze aktivity TE u single-end knihovny byl nejdříve použit nástroj `coveragebed` od BedTools [50], který vypočítá pokrytí namapovanými ready obsažených v BAM souboru napříč všemi elementy definovanými v anotaci. K následné klasifikaci do skupin a k výpočtu RPKM byl vytvořen awk skript. K započítání readu stačí aby se s elementem v anotaci překrýval na 1 bp. Situace je nastíněna v následujícím příkladu:

Chromozom	~~~~~	~~~~~	~~~~~	~~~~~
Anotace	*****	*****	*****	*****
BAM	~~~~~	~~~~~	~~~~~	~~~~~
Výsledek	[3 překryvy]	[1 překryv]	[1překryv]	[6 překryvů]

Nevýhodou tohoto nástroje je, že pro větší soubory je příliš pomalý. Nástroj `coveragebed` spočítá pro každý element v anotaci 4 informace, které jsou přidány na konce každého řádku anotace (poslední čtyři sloupce, ukázka dále). V prvním přidaném sloupečku se nachází počet překryvů ('hloubky pokrytí') pro daný element v anotaci, ve druhém je počet překrývajících se bází ('šířka pokrytí'), ve třetím je délka elementu a na posledním místě je poměr překrývajících se bází ku celkové délce elementu. Nástroj `coveragebed` byl volán následujícím způsobem

```
coverageBed -abam accepted_hits.bam -b rmsk.gtf > coverage,
```

kde `-abam accepted_hits.bam` je soubor s namapovanými ready a `-b rmsk.gtf` je soubor s anotací. Výsledek je uložen do souboru 'coverage', který vypadá následovně.

```
$ head -5 coverage
chr1 . exon 25165801 25166089 . - . AluY;SINE;Alu 3 289 289 1.0000000
chr1 . exon 33553607 33554646 . + . L2b;LINE;L2 0 0 1040 0.0000000
chr1 . exon 50330064 50332153 . + . L1PA10;LINE;L1 0 0 2090 0.0000000
chr1 . exon 58720068 58720973 . - . L1PA2;LINE;L1 0 0 906 0.0000000
chr1 . exon 75496181 75498100 . + . L1MB7;LINE;L1 0 0 1920 0.0000000
```

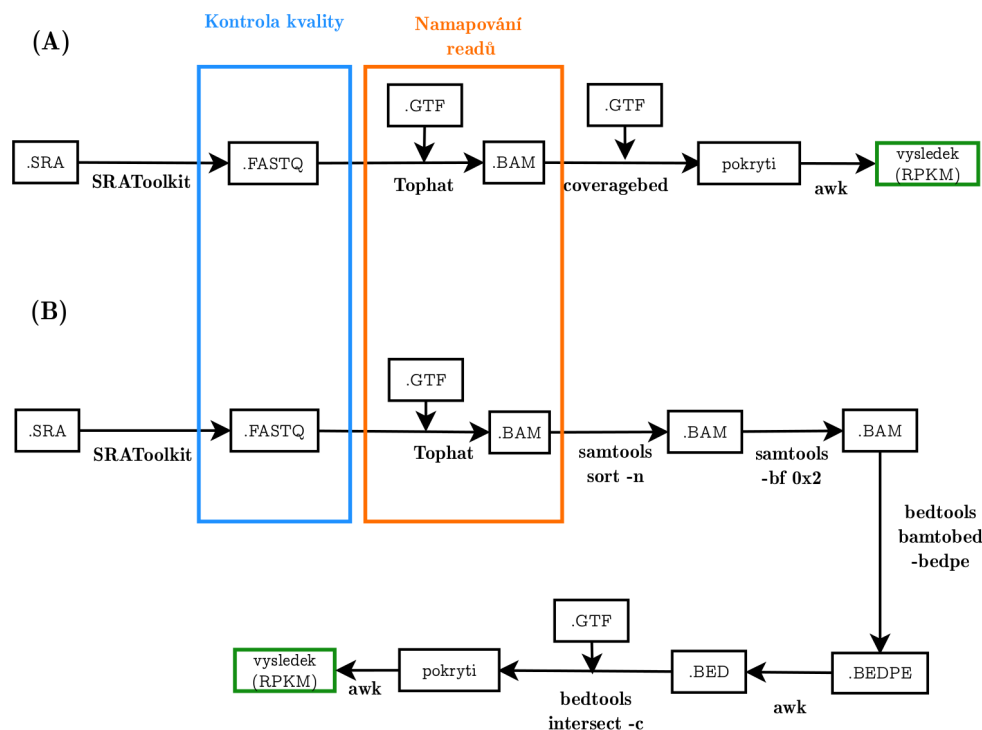
V souboru jsou za každým definovaným elementem anotace přidány 4 sloupce, z nichž nás zajímá pouze první (desátý sloupeček celkem) obsahující počet namapovaných readů. K tomuto souboru byla přidána RPKM normalizace jako 14. sloupeček. Nakonec byly všechny elementy roztrženy do skupin a pro každou skupinu byl vrácen počet namapovaných readů, hodnota RPKM a počet elementů ve skupině. Rozdělením

do skupin byly ze souboru s anotací vybrány pouze transposony, ostatní repetitivní sekvence byly tímto z analýzy vyřazeny. K práci byl vytvořen krátký awk skript, který počítá RPKM, klasifikuje transposony do hlavních 5-ti skupin (Alu, MIR, LINE, DNA a LTR) a pro tyto skupiny počítá počet readů, které se na ně namapovaly a sumu RPKM. Následuje ukázka.

```
awk 'BEGIN {
while (getline < "coverage")
{
#přidám sloupeček s RPKM hodnotami
$14 = $10*1000000000/($12*7338751);
#rozdělím 9-tý sloupeček podle ";"
split($9, ret, ";");
    if (ret[3] == "Alu") {
        #pro skupinu sčítám počet namapovaných readů
        AluR += $10;
        #pečet elementů
        AluC++;
        # a hodnoty RPKM pro všechny Alu
        AluRPKM += $14;
    }
    else if (ret[3] == "MIR") {MIRR += $10; MIRC++; MIRRPKM += $14;}
    else if (ret[2] == "LINE") {LINER += $10; LINEC++; LINERPKM += $14;}
    else if (ret[2] == "DNA") {DNAR += $10; DNAC++; DNARPKM += $14;}
    else if (ret[2] == "LTR") {LTRR += $10; LTRC++; LTRRPKM += $14;}
}
celkem = AluR+MIRR+LINER+DNAR+LTRR;
```

8.4 Postup pro paired-end knihovnu

Výpočet pro paired-end ready je podobný předchozímu výpočtu u single-end s tím rozdílem, že nejprve musíme vytvořit z paired-end readů úseky původních DNA fragmentů a až poté lze počítat pokrytí těchto fragmentů namapovanými ready. Aby bylo možné přepočítat paired-end ready na fragmenty, je nutné z BAM souboru vybrat pouze správně spárované ready (angl. properly-paired). Konkrétně se jedná o ty případy, u kterých jsou oba ready z páru namapované a navíc jsou namapované ve správné vzdálenosti. Tyto ready jsou v BAM souboru označeny bitovou vlajkou 0x2, kterou použijeme k jejich filtraci. Před filtrací je nutné seřadit BAM soubor podle názvů readů (tedy podle 1. sloupceku BAM souboru), protože programem TopHat je seřazen podle



Obrázek 21: Schéma toku analyzovaných dat. (A) - pro single-end knihovnu, (B) - pro paired-end knihovnu

souřadnic. Toto seřazení si vyžaduje nástroj `bamtools`. Seřazení provedeme nástrojem `sort` od SamTools s volbou `-n` (značí angl. name, jméno).

```
samtools sort -n accepted_hits.bam sorted_accepted_hits
```

Poté je provedena filtrace příkazem `view` od Samtools.

```
samtools view -bf 0x2 sorted_accepted_hits.bam
```

Následně využijeme nástroje `bamtools` od BedTools [50]. `Bamtools` s volbou `-bedpe` umožňuje konverzi formátu z BAM na BEDPE.

```
bedtools bamtoBEDPE -bedpe -i sorted_accepted_hits.bam > bedpe
```

Formát BEDPE bude výhodný pro spojování paired-end readů do fragmentů. Ve vytvořeném souboru 'bedpe' jsou na jednom řádku zapsány oba ready z páru následujícím způsobem.

```
$ head -5 bedpe
chr16 15841444 15841467 chr16 15841505 15841528 SRR057639.31 50 + -
chr20 42177074 42177097 chr20 42177127 42177150 SRR057639.34 50 + -
chrX 148685682 148685705 chrX 148685716 148685739 SRR057639.45 50 + -
chr1 151221077 151221100 chr1 151221090 151221113 SRR057639.51 50 + -
chr7 100613344 100613367 chr7 100613386 100613409 SRR057639.52 50 + -
```

kde 1 a 4 sloupeček je označení chromozomů, na kterých se ready nachází. Sloupečky 2 ,3 ,5 ,6 jsou souřadnice readů. Závěrem spojíme takto uspořádané paired-end ready do jednoho fragmentu následujícím skriptem.

```
awk 'BEGIN {
while (getline < "bedpe")
{
    # pokud ready pocházejí ze stejného chromozomu
    if ($1 == $4){
        # pokud první read se nachází před druhým readem, uložím
        počáteční a konečnou pozici pro fragment
        if($2<$5) {start = $2; end = $6;}
        else {start = $5; end = $3;}
        # výsledek zapíšu do souboru
        {printf $1 "\t" start "\t" end "\n" >> "bed";}
    }
}
}'
```

Tímto způsobem byl formát BEDPE převeden na formát BED, který již můžeme použít pro výpočet pokrytí transposonů fragmenty a pokračovat způsobem podobným u single-end readů. BED formát vypadá následovně.

```
$ head -5 bed
chr1 204095144 204095212
chrM 13895 13960
chr1 22853680 22853745
chr5 133938103 133938169
chr20 34241601 34241668
```

Problém nastal při použití `coveragebed`, kdy výpočet trval velmi dlouho, jelikož soubory pro paired-end ready byly cca 4x větší než u single-end knihovny. Proto byl jako náhrada použit jiný nástroj s názvem `intersect` (také `BedTools`), který pracoval mnohem rychleji. Příkaz `intersect` v základním nastavení vrací pro 2 vstupní soubory (BED a anotaci) souřadnice společného překryvu. Proto jsme využili parametru `-c`,

který vrací pouze počet překryvů (angl. count, počet). Příkaz `intersect` byl volán následujícím způsobem.

```
bedtools intersect -a rmsk.gtf -b fragments.bed -c > coverage
```

Výstupem je stejný formát souboru jako u nástroje `coveragebed`, s tím rozdílem, že `intersect` doplní na konec každého řádku (odpovídající jednomu elementu) anotace počet překrývajících se readů v BED souboru. Na konec souboru se připojí pouze 1 místo 4 sloupečků.

Podobně jako u single-end readů, je dalším krokem výpočet normalizované FPKM hodnoty pro každý element a klasifikace elementů do skupin. Protože již pracujeme s celými fragmenty, můžeme do vzorce na výpočet FPKM rovnou za C dosadit počty překryvů, uložené v posledním sloupečku. Za hodnotu N dosadíme počet readů vyfiltrovaných pomocí FLAG vlajky 0x2 a podělíme 2 (abychom z paired-end obdržely fragmenty). Celý další postup klasifikace transposonů do skupin je stejný jako u single-end readů.

9 Zhodnocení výsledků

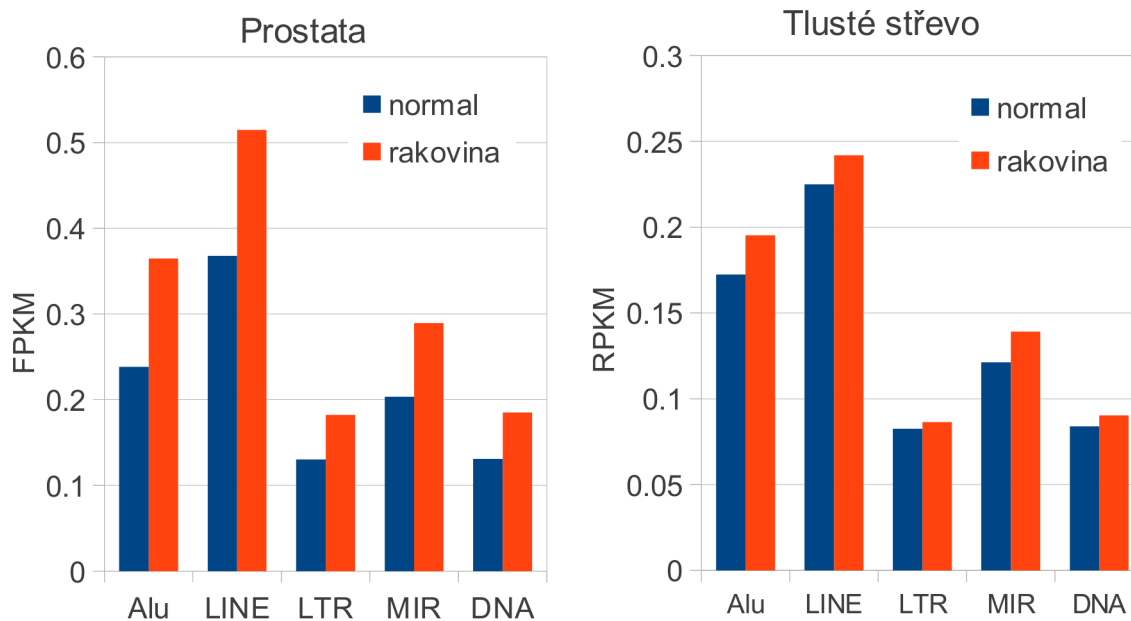
Pro srovnání a ověření správnosti byl celý postup analýzy aplikován na data, použitá při tvorbě článku zabývající se mimo jiné aktivitou transposonů u člověka [45]. Zmíněná data jsou v případě potřeby k nalezení na SRA NCBI databázi (odkazy v článku) a pochází z krevních buněk typu PBMC (angl. zkr. peripheral blood mononuclear cells, tedy periferní krevní mononukleární buňky). Výsledná aktivita transposonů, vyjádřená počtem namapovaných readů, byla shodná s výsledky, kterých bylo dosaženo aplikací postupu zpracování dat, vytvořeným v této práci. Tím došlo k ověření správnosti postupu. Analyzovaná data byla normalizována pomocí RPKM.

Lidské transposony jsou rozděleny do čtyř hlavních tříd, LINE, SINE, DNA a LTR. Protože třída SINE zahrnuje zajímavou skupinu Alu elementů, byla tato třída pro účely analýzy dále rozdělena do dvou hlavních skupin Alu a MIR. Těchto pět skupin (LINE, DNA, LTR, Alu, MIR) tvoří základní klasifikaci transposonů. Detailnější pohled na aktivitu transposonů je poskytnut rozdělením na jednotlivé rodiny, pro které byla také použita RPKM normalizace. Příslušné grafy jsou z důvodu prostorové náročnosti uvedeny v příloze C. Použité datasety (pro tlusté střevo 4 datasety - 2 rakovina a 2 normal, pro prostatu 10 datasetů - 5 rakovina a 5 normal) byly pro ohodnocení a vykreslení grafů v rámci své skupiny zprůměrovány.

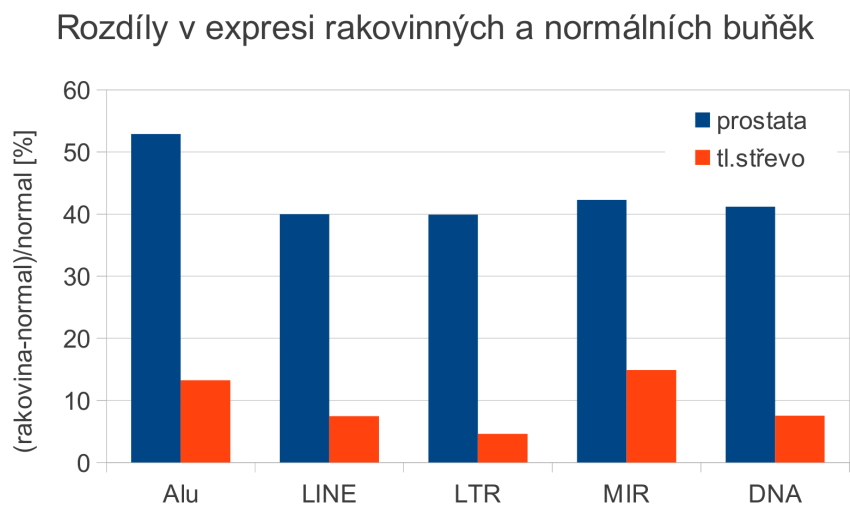
9.1 Rozdíly v aktivitě hlavních tříd TE

Aktivitu transposonů u normálních a rakovinných buněk tlustého střeva a prostaty shrnuje obr. 22. Při použití RPKM normalizace byly potvrzeny publikované výsledky, tedy že v rakovinných buňkách dochází ke zvýšení aktivity transposonů [4]. Ta byla detekovaná ve všech skupinách TE. Nejvyšší aktivity zdravé a nemocné tkáně dosahovaly LINE a Alu elementy. Na obr. 23 je vidět rozdíl mezi expresí TE v prostatě a tlustém střevě, kde u prostaty dochází k výrazně vyšší aktivitě, než je tomu u tlustého střeva. Pokud bereme v úvahu úroveň exprese v za normálního stavu, u Alu došlo k nejvyššímu rozdílu v expresi.

Jedním z hlavních mechanismů způsobujících zvýšení aktivity transposonů, je ztráta metylace v oblasti promotoru. Výskytem změn v DNA methylaci by bylo možné nárůst v aktivitě transposonů vysvětlit. Tyto výsledky bychom mohli zpřesnit analýzou většího množství datasetů. Rozdílná aktivita transposonů u prostaty a tlustého střeva naz-



Obrázek 22: Míra exprese transposonů u normálních a rakovinných buněk s použitou RPKM normalizací.



Obrázek 23: Rozdíly v expresi normálních a rakovinných buněk s RPKM normalizací.

načuje, že je aktivita transposonů tkáňově specifická. Z toho důvodu by bylo zajímavé provést analýzu u jiného typu tkání.

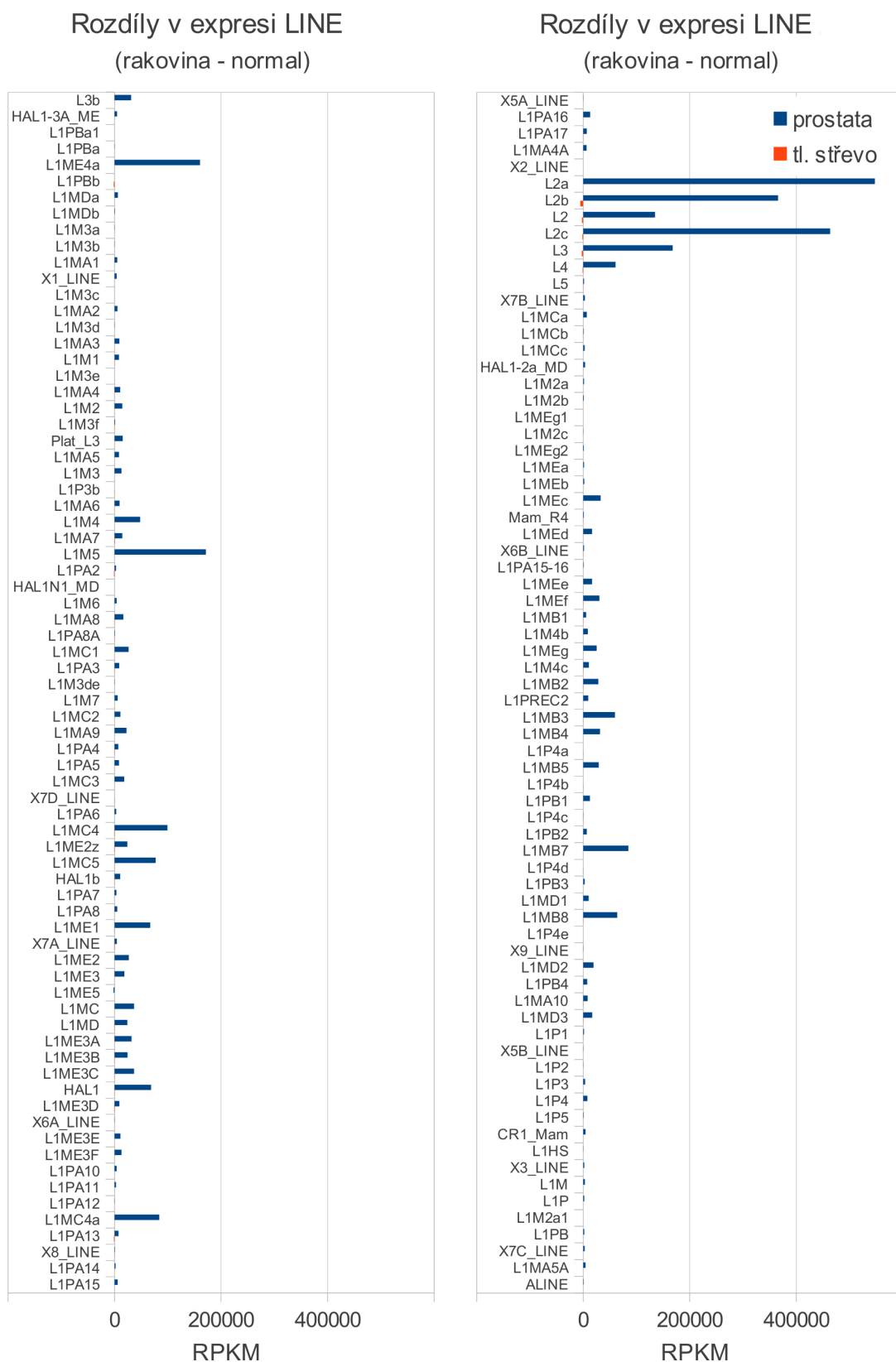
9.2 Rozdíly v aktivitě rodin TE

Detailnější pohled na rozdíly v expresi transposonů lidského genomu poskytne jejich podrobnější rozdělení do jednotlivých rodin. Toto rozdělení je celé k nalezení v příloze, část C. Je patrné, že pouze u několika rodin transposonů dochází k výrazné změně v expresi mezi zdravými a nemocnými buňkami. Například Alu elementy jsou rozděleny do tří hlavních skupin AluJ, AluS a AluY, podle stáří (AluY nejmladší, AluJ nejstarší). Ze skupiny AluY je pouze stejnojmenná rodina AluY výrazně aktivnější u rakovinných než u zdravých buněk. Podobný rozdíl je přítomen jak u prostaty, tak u tlustého střeva. Další takové příklady jsou k nalezení ve všech třídách transposonů.

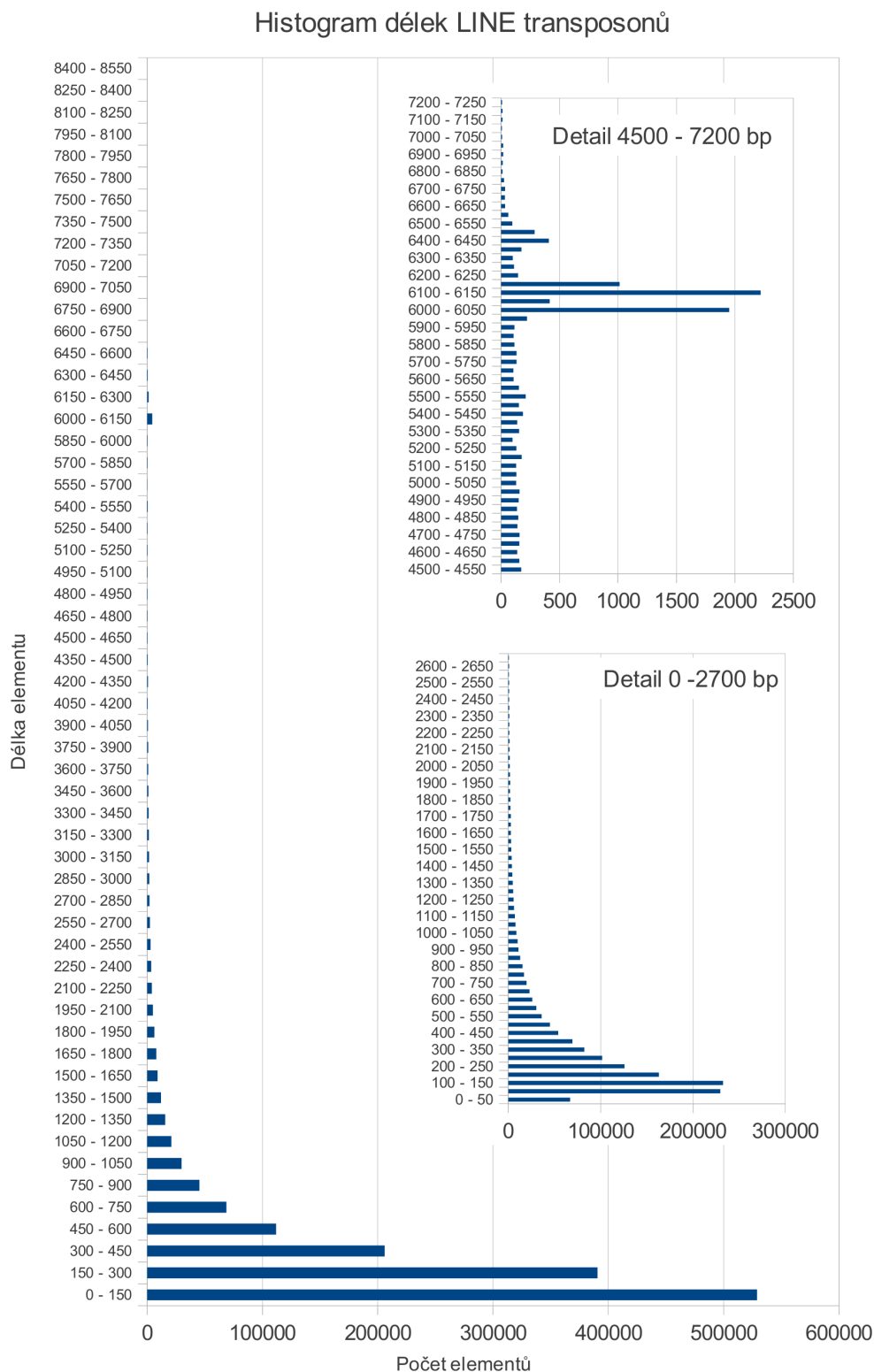
Za povšimnutí také stojí vysoký rozdíl v expresi dlouhých LINE z rodin L2, L3 na obr. 24. Ze třídy DNA transposonů dosahují výrazných rozdílů rodiny MER5A a MER5B, ze třídy LTR rodiny MLT1B a MLT1A1.

9.3 Diskuze o použití RPKM

Pro výpočet RPKM potřebujeme tři hodnoty (C = počet namapovaných readů, L = celkový počet readů v experimentu a N = délku elementu). Při bližším pohledu na délku transposonů v anotaci zjistíme, že se v ní vyskytují i velmi krátké fragmenty, které pravděpodobně ani nemohou být aktivní. Chyba nevznikne namapováním readu na krátký transposon, z důvodu jeho aktivity, ale proto, že jeho sekvence je homologní s ostatními elementy dané skupiny. Po přepočtu do RPKM získá tento fragment vyšší ohodnocení než aktivní dlouhé fragmenty. Vyplývá to ze vzorce (2) na výpočet RPKM. Z důvodu bližšího prozkoumání byly vykresleny histogramy s délkovým rozložením transposonů pro všech pět hlavních skupin (LINE, DNA, LTR, Alu a MIR), viz příloha A. Pro LINE, DNA a LTR obsahuje anotace velké množství krátkých fragmentů, které do normalizace zanášejí chyby. Situace je znázorněna na obr. 25., kde je patrné nejvyšší zastoupení krátkých elementů LINE menších než 150 bp. Malý nárůst na cca 6kbp značí přítomné LINE transposony plné délky. Oproti tomu transposony Alu a MIR vykazují z obou stran ohraničené píky, které korespondují s jejich konsenzuální délkou. U Alu transposonů je v histogramovém rozložení vidět výrazný pík na 300 bp. V úvahu je



Obrázek 24: Detailní pohled na rodiny transposonů pro transposony třídy LINE. Ostatní třídy jsou k nalezení v příloze.



Obrázek 25: Histogram délek LINE transposonů přítomných v anotaci, ostatní histogramy v příloze. Pík na cca 6kbp značí LINE transposony plné délky (cca 6,2 kbp).

také nutné brát to, že metoda RPKM byla vytvořena z důvodu potřebné normalizace exprese genů, více v referenci [46].

Možným východiskem pro provedení potřebné normalizace by bylo odstranit z anotace elementy, které se svojí délkou výrazně liší od průměrné délky aktivního elementu pro danou skupinu. Tímto zobecněním bychom však o část informace přišli. Nebo by také bylo možné do výpočtu RPKM dosadit za N přímo průměrnou délku elementu dané skupiny, čímž by mělo dojít k mírnému zpřesnění (znevýhodnili bychom problémové krátké fragmenty).

9.4 Diskuze o použití TopHat

Při ověřování výsledků vyvstala otázka o vhodnosti použití programu TopHat pro namapování readů. Při snaze nalézt konkrétní transposony v genomu, které vykazují výrazný rozdíl v expresi (a jejich zobrazení na IGV prohlížeči [60]) byly nalezeny transposony, které se nacházejí mezi dvěma exony. Ready přitom nebyly namapované přímo na tyto transposony, tudíž tyto transposony nevykazovaly žádný rozdíl v aktivitě. Jednalo se o ready, které se nepodařilo namapovat v celku. Z toho důvodu byly ready TopHatem rozděleny na dva kratší celky, které byly namapovány zvlášť. Tento přístup umožňuje efektivní identifikaci mezer mezi exony. Rozdělené ready však nejsou rozděleny na 2 části i v SAM souboru, ale jsou reprezentovány jedním readem, který má na startovní pozici počáteční pozici první části readu a na konečné pozici konečnou pozici druhé části readu. Informace o mezeře mezi těmito dvěma částmi je uložena v CIGAR řetězci, se kterým jsme při výpočtu pokrytí transposonů ready nepracovali.

Protože se TopHat snaží nenamapované ready namapovat pomocí jejich rozdělení na menší úseky, vzniká v následné analýze odchylka. Tuto informaci bychom mohli využít pro další upřesnění výsledku a vyzkoušet použití jiného nástroje, který namapovává ready v celku. Například zmiňovanou Bowtie. Namapováním readů v celku bychom nejenom odstranili tuto odchylku, ale došlo by ke snížení počtu namapovaných readů. To by vysvětlilo, proč bylo na TE namapováno cca 19,6% readů, když je v lidském genomu aktivních transposonů méně než 1%. Program TopHat byl použit, protože je nejpoužívanějším programem pro namapování RNA-seq readů, jehož hlavním cílem je analýza genů. Výše uvedené důvody však naznačují, že pro analýzu transposonů není příliš vhodný.

10 ZÁVĚR

Transposony jsou schopny svojí aktivitou způsobovat větší či menší změny v genomu vedoucí ke genetické nestabilitě, která byla prokázána v rakovinných buňkách. Jejich aktivita je v normálních buňkách omezena, aby se zabránilo škodám. Regulaci transposonů mají na starost tři hlavní mechanismy: metylace promotoru, umlčování transposonové mRNA interferencí a transkripční faktory nutné pro aktivaci transposonů. Abnormální vzor metylace DNA je často pozorován v rakovinných buňkách. Hypometylace promotoru transposonu umožní jeho aktivaci. Poté je nutná přítomnost správných transkripčních faktorů a utlumení mechanismu mRNA interference. V případě aktivní mRNA interference by byl transkript transposonu zničen. Otázkou zůstává, zda zvýšená aktivita transposonů v rakovinných buňkách je důsledkem změn, kterými buňka prochází, nebo se jedná o jednu z příčin vzniku rakoviny.

Spojení vzniku rakoviny se zvýšenou aktivitou transposonů je aktuální téma, především díky rozvoji nových technologií umožňujících sekvenaci repetitivních oblastí genomu. RNA-seq technologie patří mezi sekvenační metody nové generace a umožňuje přesnou analýzu celého buněčného transkriptomu. Oproti microarray technologiím poskytuje řadu výhod, zejména umožňuje objevovat nové transkripty. RNA-seq produkuje velké množství dat, které je třeba výpočetně zpracovat. V této práci byly srovnány exprese jednotlivých rodin lidských transposonů u rakovinné a normální tkáně pomocí dat z RNA-seq experimentu. Z SRA NCBI databáze byla stažena experimentální data, která byla použita ke srovnání rozdílné exprese transposonů tlustého střeva a prostaty. Pro zpracování dat byly použity jak volně dostupné sofistikované nástroje, tak vlastní skripty. U obou dvou analyzovaných tkání byl detekován nárůst aktivity transposonů u rakovinných buněk. Výrazně vyšší nárůst aktivity byl pozorován u buněk prostaty. Z hlavních skupin transposonů byla nejvyšší aktivita a zároveň nejvyšší rozdíly detekovány u LINE a Alu transposonů. Transposony byly dále klasifikovány do jednotlivých rodin. Pro rodiny byly vykresleny rozdíly v expresi, což poskytlo detailnější náhled na aktivitu lidských transposonů. Výsledky dosažené v této práci korespondují s publikovanými i přes rizika analýzy, diskutovaná v kapitolách 9.3 a 9.4, .

Tato práce poskytuje základní náhled do perspektivní oblasti aktivity transposonů v rakovinných buňkách. Pro objasnění souvislosti zvýšené aktivity transposonů se vznikem rakoviny je třeba rozsáhlého výzkumu. Techniku sekvenace nové generace je možné

použít na detekci změn v metylaci DNA (angl. Bisulfite Sequencing) a zkoumání nových transposonových inzercí. Ze srovnání exprese transposonů u tlustého střeva a prostaty je vidět, že dochází ke tkáňově specifickým rozdílům. Pro více komplexní pohled je třeba dalších analýz jiných typů tkáně.

Z důvodu vysoké výpočetní náročnosti, při které je nutné zpracovávat velké objemy dat (celkový objem vstupních dat je 67 GB ve formátu FASTQ), je téměř nemožné provádět analýzu RNA-seq dat na běžném počítači. Veškerá analýza proto byla prováděna na výpočetním centru MetaCentrum [40]. Na testování i samotnou analýzu bylo na MetaCentru spuštěno 528 úloh s celkovou spotřebou 73,8 dnů procesorového času.

Protože stále dochází ke snižování nákladů na sekvenaci DNA, v budoucnosti by mohla být sekvenace genomu a transkriptomu prováděna v rámci preventivní péče. Identifikace nových transposonových inzercí a detekce jejich aktivity v různých rakovinových buňkách by mohla pomoci při diagnostice rakoviny. Transposony by se také mohly uplatnit při léčbě geneticky podmíněných chorob a tudíž by sloužily jako základ pro genovou terapii. Použití transposonů slibuje nižší imunogenicitu, vyšší bezpečnost a snížené operační náklady než je tomu při použití virových vektorů. Dva DNA transposony byly již rekonstruovány jako nástroj genové terapie. Jedná se o transposon *Sleeping Beauty*, který je původně neaktivní element z lososovité ryby. Název poukazuje na to, že byl transposon zaktivován po dlouhém evolučním spánku. Dalším je PiggyBac element z genomu baciloviru. Použití o transposonů jako strážců našeho genomu je jednou z budoucích vizí genové terapie.

Reference

- [1] PRAY, Leslie. *Transposons: The Jumping Genes: Nature Education 1*. 2008. Dostupné z: <http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518>
- [2] LEVIN, Henry L. a John V. MORAN. *Dynamic interactions between transposable elements and their hosts*. Nature Reviews Genetics. 2011-8-18, roč. 12, č. 9, s. 615-627. ISSN 1471-0056. DOI: 10.1038/nrg3030. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg3030>
- [3] BOWEN, Nathan J. a I. King JORDAN. *Transposable elements and the evolution of eukaryotic complexity*. Curr Issues Mol Biol. 2002, roč. 4, č. 3, 65–76.
- [4] BELANCIO, Victoria P., Astrid M. ROY-ENGEL a Prescott L. DEININGER. *All y'all need to know 'bout retroelements in cancer*. Seminars in Cancer Biology. 2010, roč. 20, č. 4, s. 200-210. ISSN 1044579x. DOI: 10.1016/j.semcancer.2010.06.001. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S1044579X1000043X>
- [5] BELANCIO, V. P., D. J. HEDGES a P. DEININGER. *Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health*. Genome Research. 2008-02-06, roč. 18, č. 3, s. 343-358. ISSN 1088-9051. DOI: 10.1101/gr.5558208. Dostupné z: <http://www.genome.org/cgi/doi/10.1101/gr.5558208>
- [6] MEDSTRAND, P. *Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes*. Genome Research. roč. 12, č. 10, s. 1483-1495. ISSN 10889051. DOI: 10.1101/gr.388902. Dostupné z: <http://www.genome.org/cgi/doi/10.1101/gr.388902>
- [7] BECK, Christine R., Pamela COLLIER, Catriona MACFARLANE, Maika MALLIG, Jeffrey M. KIDD, Evan E. EICHLER, Richard M. BADGE a John V. MORAN. *LINE-1 Retrotransposition Activity in Human Genomes*. Cell. 2010, roč. 141, č. 7, s. 1159-1170. ISSN 00928674. DOI: 10.1016/j.cell.2010.05.021. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S009286741000557X>
- [8] BELANCIO, V. P., A. M. ROY-ENGEL, R. R. POCHAMPALLY a P. DEININGER. *Somatic expression of LINE-1 elements in human tissues*. Nucleic Acids Research. 2010-07-03, roč. 38, č. 12, s. 3909-3922. ISSN 0305-1048. DOI: 10.1093/nar/gkq132. Dostupné z: <http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkq132>
- [9] BENNETT, E. A., H. KELLER, R. E. MILLS, S. SCHMIDT, J. V. MORAN, O. WEICHENRIEDER a S. E. DEVINE. *Active Alu retrotransposons in the human genome*. Genome Research. 2008-10-03, roč. 18, č.

- 12, s. 1875-1883. ISSN 1088-9051. DOI: 10.1101/gr.081737.108. Dostupné z: <http://www.genome.org/cgi/doi/10.1101/gr.081737.108>
- [10] BROOKFIELD, John F. Y. *The ecology of the genome - mobile DNA elements and their hosts*. Nature Reviews Genetics. 2005-1-10, roč. 6, č. 2, s. 128-136. ISSN 1471-0056. DOI: 10.1038/nrg1524. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg1524>
- [11] BIÉMONT, Christian a Cristina VIEIRA. *Genetics: Junk DNA as an evolutionary force*. Nature. 2006-10-5, roč. 443, č. 7111, s. 521-524. ISSN 0028-0836. DOI: 10.1038/443521a. Dostupné z: <http://www.nature.com/doifinder/10.1038/443521a>
- [12] SLOTKIN, R. Keith a Robert MARTIENSSEN. *Transposable elements and the epigenetic regulation of the genome*. Nature Reviews Genetics. 2007, roč. 8, č. 4, s. 272-285. ISSN 1471-0056. DOI: 10.1038/nrg2072. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg2072>
- [13] KAZAZIAN, H. H. *Mobile Elements: Drivers of Genome Evolution*. Science. 2004-03-12, roč. 303, č. 5664, s. 1626-1632. ISSN 0036-8075. DOI: 10.1126/science.1089670. Dostupné z: <http://www.sciencemag.org/cgi/doi/10.1126/science.1089670>
- [14] SNUSTAD, D a Michael J SIMMONS. *Genetika. 5th ed. Překlad Jiřina Relichová*. Brno: Masarykova univerzita, 2009, xxi, 871 s. ISBN 978-802-1048-522.
- [15] FESCHOTTE, C. *Treasures in the attic: Rolling circle transposons discovered in eukaryotic genomes*. Proceedings of the National Academy of Sciences. roč. 98, č. 16, s. 8923-8924. ISSN 00278424. DOI: 10.1073/pnas.171326198. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.17132619>
- [16] RAMSDEN, Dale A., Brett D. WEED a Yeturu V.R. REDDY. *V(D)J recombination: Born to be wild*. Seminars in Cancer Biology. 2010, roč. 20, č. 4, s. 254-260. ISSN 1044579x. DOI: 10.1016/j.semcancer.2010.06.002. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S1044579X10000441>
- [17] KEJNOVSKY, Eduard, Jennifer S HAWKINS a Cédric FESCHOTTE. *Plant transposable elements: biology and evolution, Plant Genome Diversity Volume 1*. 2012. vyd. Springer-Verlag Wien 2012, s. 17-34. ISBN 978-3-7091-1129-1.
- [18] ZHANG, Wensheng, Andrea EDWARDS, Wei FAN, Prescott DEININGER a Kun ZHANG. *Alu distribution and mutation types of cancer genes*. BMC Genomics. 2011, roč. 12, č. 1, s. 157-. ISSN 1471-2164. DOI: 10.1186/1471-2164-12-157. Dostupné z: <http://www.biomedcentral.com/1471-2164/12/157>
- [19] KEJNOVSKÝ, Eduard a Roman HOBZA. *Evoluční genomika (skripta)*. 2009. Dostupné z: http://www.evolucnigenomika.cz/Skripta/Evolucni_genomika_skripta_2008.pdf

- [20] KEJNOVSKÝ, Eduard. *Skákající geny a evoluce: Paraziti, nebo pomocníci?* Vesmír. 2009, roč. 88, č. 4.
- [21] MILLS, Ryan E., E. Andrew BENNETT, Rebecca C. ISKOW, Christopher T. LUTTIG, Circe TSUI, W. Stephen PITTARD a Scott E. DEVINE. *Recently Mobilized Transposons in the Human and Chimpanzee Genomes*. The American Journal of Human Genetics. 2006, roč. 78, č. 4, s. 671-679. ISSN 00029297. DOI: 10.1086/501028. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0002929707637045>
- [22] MUOTRI, A. R., M. C.N. MARCHETTO, N. G. COUFAL a F. H. GAGE. *The necessary junk: new functions for transposable elements*. Human Molecular Genetics. 2007-07-31, roč. 16, R2, R159-R167. ISSN 0964-6906. DOI: 10.1093/hmg/ddm196. Dostupné z: <http://www.hmg.oxfordjournals.org/cgi/doi/10.1093/hmg/ddm196>
- [23] MYERS, Jeremy S., Bethaney J. VINCENT, Hunt UDALL, W. Scott WATKINS, Tammy A. MORRISH, Gail E. KILROY, Gary D. SWERGOLD, Jurgen HENKE, Lotte HENKE, John V. MORAN, Lynn B. JORDE a Mark A. BATZER. *A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements*. The American Journal of Human Genetics. 2002, roč. 71, č. 2, s. 312-326. ISSN 00029297. DOI: 10.1086/341718. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0002929707604777>
- [24] SNULL, Solyom a KAZAZIAN HH. *Mobile elements in the human genome: implications for disease*. Genome Medicine. 2012, roč. 12, č. 4.
- [25] CHÉNAIS, Benoît. *Transposable elements and human cancer: A causal relationship?* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2013, roč. 1835, č. 1, s. 28-35. ISSN 0304419x. DOI: 10.1016/j.bbcan.2012.09.001. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0304419X12000583>
- [26] CALLINAN, PA a MA BATZER. *Retrotransposable Elements and Human Disease*. Genome and Disease. 2006, 104–115. DOI: 10.1159/000092503.
- [27] BRENNER, Sydney, Maria JOHNSON, John BRIDGHAM, George GOLDA, David H. LLOYD, Davida JOHNSON, Shujun LUO, Sarah MCCURDY, Michael FOY, Mark EWAN, Rithy ROTH, Dave GEORGE, Sam ELETR, Glenn ALBRECHT, Eric VERMAAS, Steven R. WILLIAMS, Keith MOON, Timothy BURCHAM, Michael PALLAS, Robert B. DUBRIDGE, James KIRCHNER, Karen FEARON, Jen-i MAO a Kevin CORCORAN. *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays*. Nature Biotechnology. 2000, roč. 18, č. 6, s. 630-634. ISSN 1087-0156. DOI: 10.1038/76469. Dostupné z: <http://www.nature.com/doifinder/10.1038/76469>

- [28] METZKER, Michael L. *Sequencing technologies - the next generation*. Nature Reviews Genetics. 2009-12-8, roč. 11, č. 1, s. 31-46. ISSN 1471-0056. DOI: 10.1038/nrg2626. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg2626>
- [29] TAYLOR, K.H., C.W. CALDWELL a H. SHI. *Next Generation Sequencing: Advances in Characterizing the Methyloome*. Genes. 2010, roč. 2, č. 1, s. 143-165.
- [30] ROTHBERG, Jonathan M., Wolfgang HINZ, Todd M. REARICK, Jonathan SCHULTZ, William MILESKEI, Mel DAVEY, John H. LEAMON, Kim JOHNSON, Mark J. MILGREW, Matthew EDWARDS, Jeremy HOON, Jan F. SIMONS, David MARRAN, Jason W. MYERS, John F. DAVIDSON, Annika BRANTING, John R. NOBILE, Bernard P. PUC, David LIGHT, Travis A. CLARK, Martin HUBER, Jeffrey T. BRANCIFORTE, Isaac B. STONER, Simon E. CAWLEY, Michael LYONS, Yutao FU, Nils HOMER, Marina SEDOVA, Xin MIAO, Brian REED, Jeffrey SABINA, Erika FEIERSTEIN, Michelle SCHORN, Mohammad ALANJARY, Eileen DIMALANTA, Devin DRESSMAN, Rachel KASINSKAS, Tanya SOKOLSKY, Jacqueline A. FIDANZA, Eugeni NAMSARAEV, Kevin J. MCKERNAN, Alan WILLIAMS, G. Thomas ROTH a James BUSTILLO. *An integrated semiconductor device enabling non-optical genome sequencing*. Nature. 2011-7-20, roč. 475, č. 7356, s. 348-352. ISSN 0028-0836. DOI: 10.1038/nature10242. Dostupné z: <http://www.nature.com/doifinder/10.1038/nature10242>
- [31] TREANGEN, Todd J. a Steven L. SALZBERG. *Repetitive DNA and next-generation sequencing: computational challenges and solutions*. Nature Reviews Genetics. 2011-11-29, s. -. ISSN 1471-0056. DOI: 10.1038/nrg3117. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg3117>
- [32] TUCKER, Tracy, Marco MARRA a Jan M. FRIEDMAN. *Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine*. The American Journal of Human Genetics. 2009, roč. 85, č. 2, s. 142-154. ISSN 00029297. DOI: 10.1016/j.ajhg.2009.06.022. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0002929709002833>
- [33] MANDAL, A. K., R. PANDEY, V. JHA a M. MUKERJI. Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. Nucleic Acids Research. 2013-02-18, vol. 41, issue 4, s. 2121-2137. DOI: 10.1093/nar/gks1457. Dostupné z: <http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gks1457>

- [34] TREANGEN, Todd J. a Steven L. SALZBERG. Repetitive DNA and next-generation sequencing: computational challenges and solutions. DOI: 10.1038/nrg3117. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg3117>
- [35] MORTAZAVI, Ali, Brian A WILLIAMS, Kenneth MCCUE, Lorian SCHAEFFER a Barbara WOLD. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008-5-30, vol. 5, issue 7, s. 621-628. DOI: 10.1038/nmeth.1226. Dostupné z: <http://www.nature.com/doifinder/10.1038/nmeth.1226>
- [36] WANG, Zhong, Mark GERSTEIN a Michael SNYDER. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. roč. 10, č. 1, s. 57-63. ISSN 1471-0056. DOI: 10.1038/nrg2484. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg2484>
- [37] OSHLACK, Alicia, Mark D ROBINSON a Matthew D YOUNG. From RNA-seq reads to differential expression results: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. roč. 10, č. 1, s. 57-63. ISSN 1471-0056. DOI: 10.1186/gb-2010-11-12-220. Dostupné z: <http://genomebiology.com/2010/11/12/220>
- [38] COCK, P. J. A., C. J. FIELDS, N. GOTO, M. L. HEUER a P. M. RICE. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*. roč. 38, č. 6, s. 1767-1771. ISSN 0305-1048. DOI: 10.1093/nar/gkp1137. Dostupné z: <http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkp1137>
- [39] Sequence Read Archive (SRA). NCBI. [online]. [cit. 2013-04-24]. Dostupné z: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=download_reads
- [40] METAVO. MetaCentrum [online]. [cit. 2013-04-24]. Dostupné z: <http://metavo.metacentrum.cz/cs/>
- [41] FASTX-TOOLKIT. [online]. [cit. 2013-04-25]. Dostupné z: http://hannonlab.cshl.edu/fastx_toolkit/index.html
- [42] FASTQC. [online]. [cit. 2013-04-25]. Dostupné z: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- [43] PICARD TOOLS. [online]. [cit. 2013-04-27]. Dostupné z: <http://picard.sourceforge.net/>
- [44] PICARD TOOLS. Explain flags [online]. [cit. 2013-04-27]. Dostupné z: <http://picard.sourceforge.net/explain-flags.html>
- [45] SAMTOOLS. The SAM Format Specification (v1.4-r985): The SAM Format Specification Working Group. 2011. Dostupné z: <http://samtools.sourceforge.net/SAM1.pdf>
- [46] TopHat. [online]. [cit. 2013-04-23]. Dostupné z: <http://TopHat.cbcb.umd.edu/index.shtml>
- [47] UCSC. UCSC Genome Bioinformatics [online]. [cit. 2013-05-13]. Dostupné z: <http://genome.ucsc.edu/>
- [48] BOWTIE 2. Bowtie 2: Fast and sensitive read alignment [online]. [cit. 2013-05-13]. Dostupné z: <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>
- [49] REPEATMASKER. Institute for Systems Biology [online]. [cit. 2013-05-13]. Dostupné z: <http://www.repeatmasker.org/>
- [50] BEDTOOLS. The genome arithmetic suite. [online]. [cit. 2013-05-13]. Dostupné z: <http://bedtools.readthedocs.org/en/latest/index.html>
- [51] SAMTOOLS. Manual Reference Pages [online]. [cit. 2013-05-13]. Dostupné z: <http://samtools.sourceforge.net/samtools.shtml>
- [52] GTF. GTF2.2: A Gene Annotation Format [online]. [cit. 2013-05-13]. Dostupné z: <http://mblab.wustl.edu/GTF22.html>
- [53] MILLS, Ryan E., E. Andrew BENNETT, Rebecca C. ISKOW a Scott E. DEVINE. Which transposable elements are active in the human genome?. Trends in Genetics. 2007, vol. 23, issue 4, s. 183-191. DOI: 10.1016/j.tig.2007.02.006. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0168952507000595>
- [54] TRAPNELL, C., L. PACHTER a S. L. SALZBERG. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009-04-23, vol. 25,

- issue 9, s. 1105-1111. DOI: 10.1093/bioinformatics/btp120. Dostupné z: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp120>
- [55] GARBER, Manuel, Manfred G GRABHERR, Mitchell GUTTMAN a Cole TRAPNELL. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*. 2011-5-27, vol. 8, issue 6, s. 469-477. DOI: 10.1038/NMETH.1613. Dostupné z: <http://www.nature.com/doifinder/10.1038/nmeth.1613>
- [56] OZSOLAK, Fatih, Patrice M. MILOS, Matthew D YOUNG a Cole TRAPNELL. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*. 2010-12-30, vol. 12, issue 2, s. 87-98. DOI: 10.1038/nrg2934. Dostupné z: <http://www.nature.com/doifinder/10.1038/nrg2934>
- [57] TRAPNELL, Cole, David G HENDRICKSON, Martin SAUVAGEAU, Loyal GOFF, John L RINN a Lior PACHTER. Differential analysis of gene regulation at transcript resolution with RNA-seq: advances, challenges and opportunities. *Nature Biotechnology*. 2012-12-9, vol. 31, issue 1, s. 46-53. DOI: 10.1038/nbt.2450. Dostupné z: <http://www.nature.com/doifinder/10.1038/nbt.2450>
- [58] TRAPNELL, Cole, Adam ROBERTS, Loyal GOFF, Geo PERTEA, Daehwan KIM, David R KELLEY, Harold PIMENTEL, Steven L SALZBERG, John L RINN a Lior PACHTER. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks: advances, challenges and opportunities. *Nature Protocols*. 2012-3-1, vol. 7, issue 3, s. 562-578. DOI: 10.1038/nprot.2012.016. Dostupné z: <http://www.nature.com/doifinder/10.1038/nprot.2012.016>
- [59] BROWN, Stuart M. Next-generation DNA sequencing informatics. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Presss, 2013, viii, 241 pages. ISBN 978-193-6113-873.
- [60] IGV. Integrative Genomics Viewer [online]. [cit. 2013-05-19]. Dostupné z: <http://www.broadinstitute.org/igv/home>

SEZNAM ZKRATEK

TE	transposon (z angl. Transposable Element)
LTR	typ transposonu obsahující dlouhé koncové repetice (z angl. Long Terminal Repeat), naopak non-LTR
LINE	Long Interspersed Nuclear Elements
SINE	Short Interspersed Nuclear Elements
SVA	typ transposonu složený ze SINE/VNTR/Alu
TSD	krátká duplikace cílového místa (z angl. Target Site Duplication)
TIR	obrácené koncové repetice (z angl. Terminal Inverted Repeats)
VLP	částice typu VLP (z angl. Virus Like Particle)
ORF	čtecí rámec (z angl. Open Reading Frame)
TPRT	mechanismus transpozice non-LTR TE (z angl. Target site Primed Reverse Transcription)
FL	elementy celé délky (z angl. Full Length)
UTR	netranslatovaná oblast (z angl. UnTranslated Region)
VNTR	sekvence tandemových repetit (z angl. Variable Number Tandem Repeats)
siRNAs	způsob RNA interference (z angl. Small Interfering RNA)
piRNA	způsob RNA interference (z angl. PIWI-interacting RNA)
HERV	lidské endogenní retroviry (z angl. Human Endogenous Retrovirus)
NAHR	ektopická rekombinace (z angl. Non Allelic Homologous Recombination)
NGS	sekvenace příští generace (z angl. Next Generation Sequencing)
MPSS	masivně paralelní sekvenování (z angl. Massively Parallel Signature Sequencing)
CCD	typ senzoru (z angl. Charge Coupled Device)
emPCR	emulzní polymerázová řetězová reakce
dNTP	deoxynukleozidtrifosfát
cDNA	komplementární deoxyribonukleová kyselina
SOLiD	platforma pro sekvenaci na základě ligace (z angl. Sequencing by Oligo Ligation and Detection)

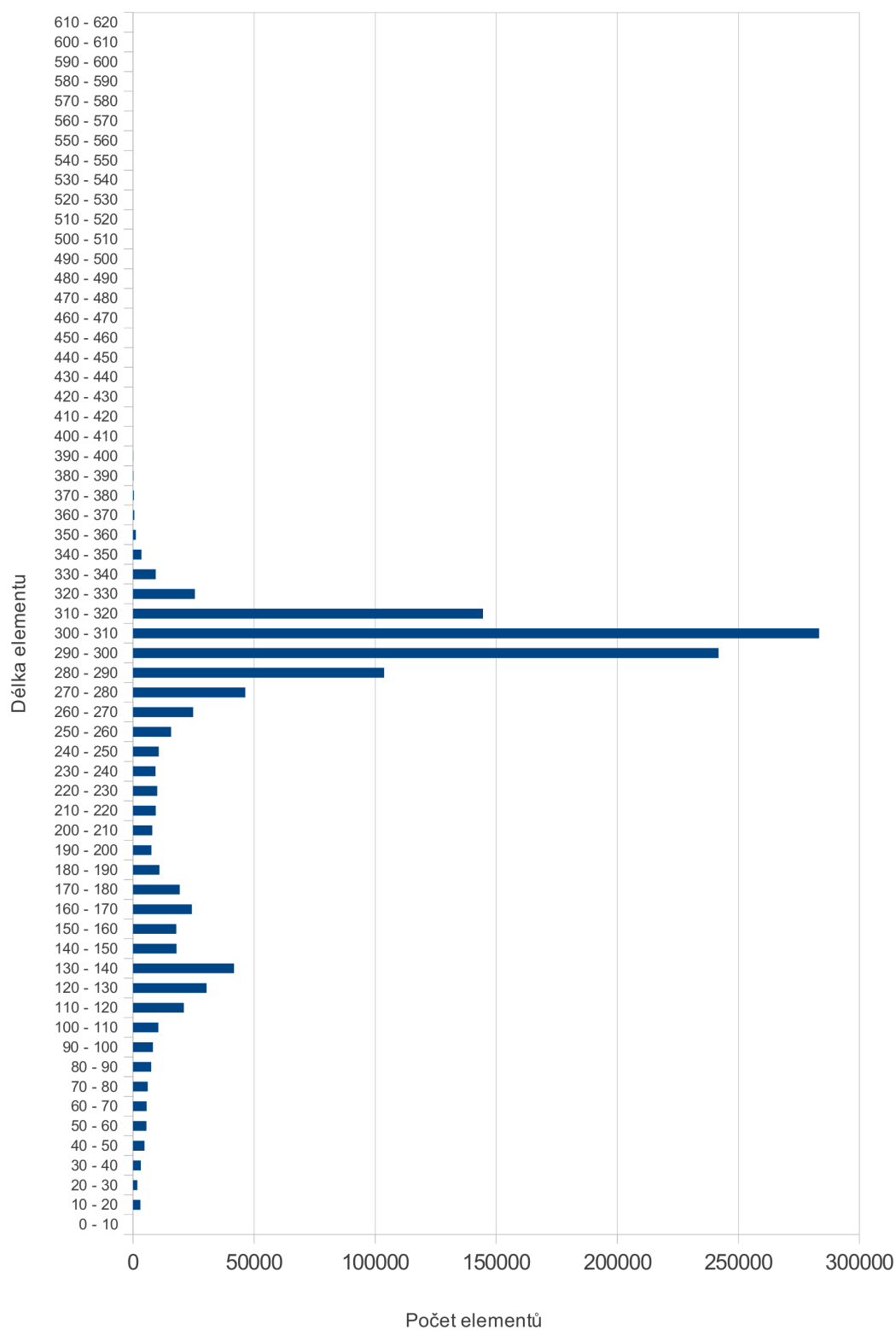
SEZNAM PŘÍLOH

- příloha A - Délky transposonů v anotaci
- příloha B - Rozdělení readů
- příloha C - Rozdíly v expresi rodin lidských transposonů

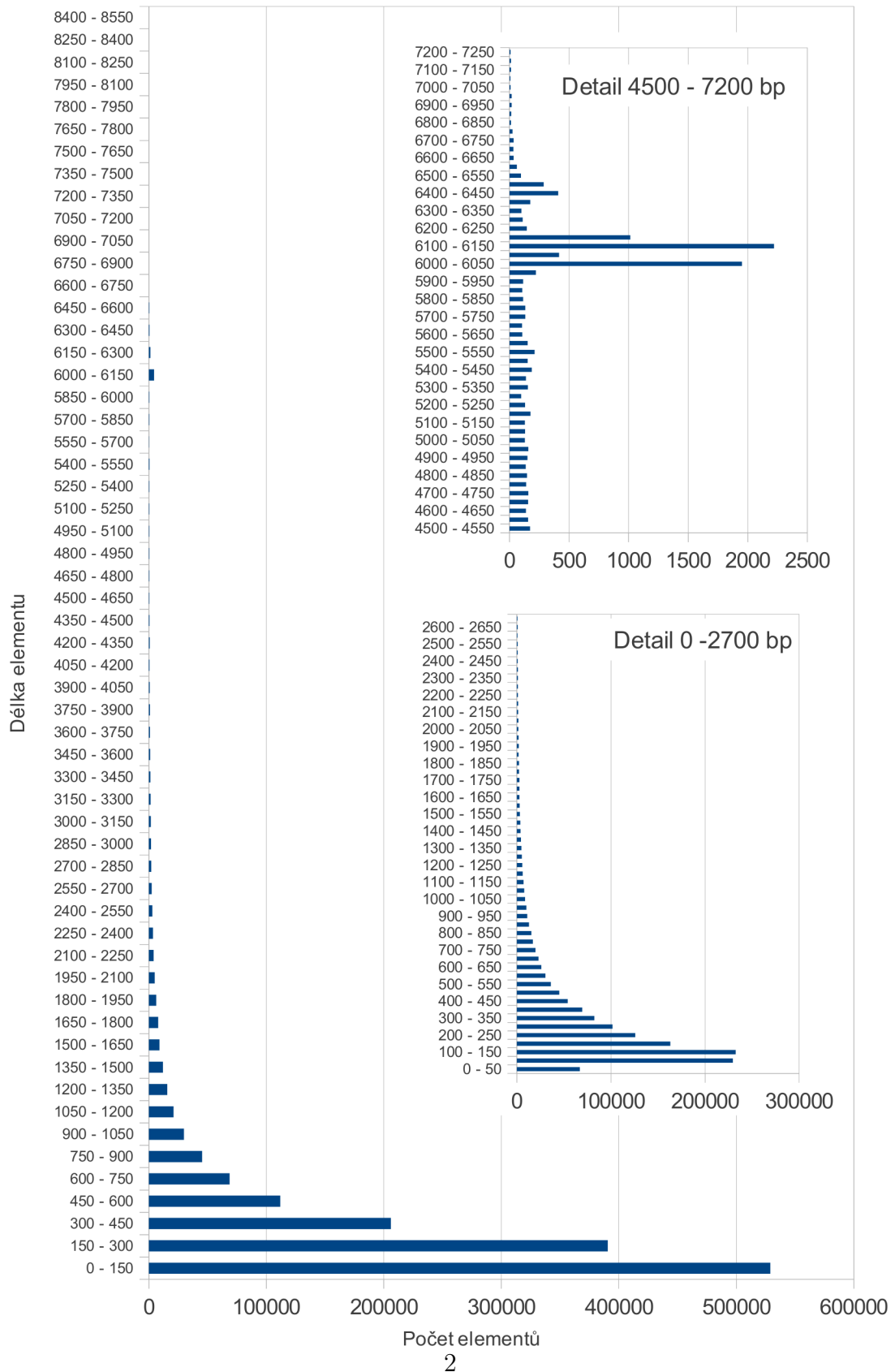
PŘÍLOHY

A Délky transposonů v anotaci

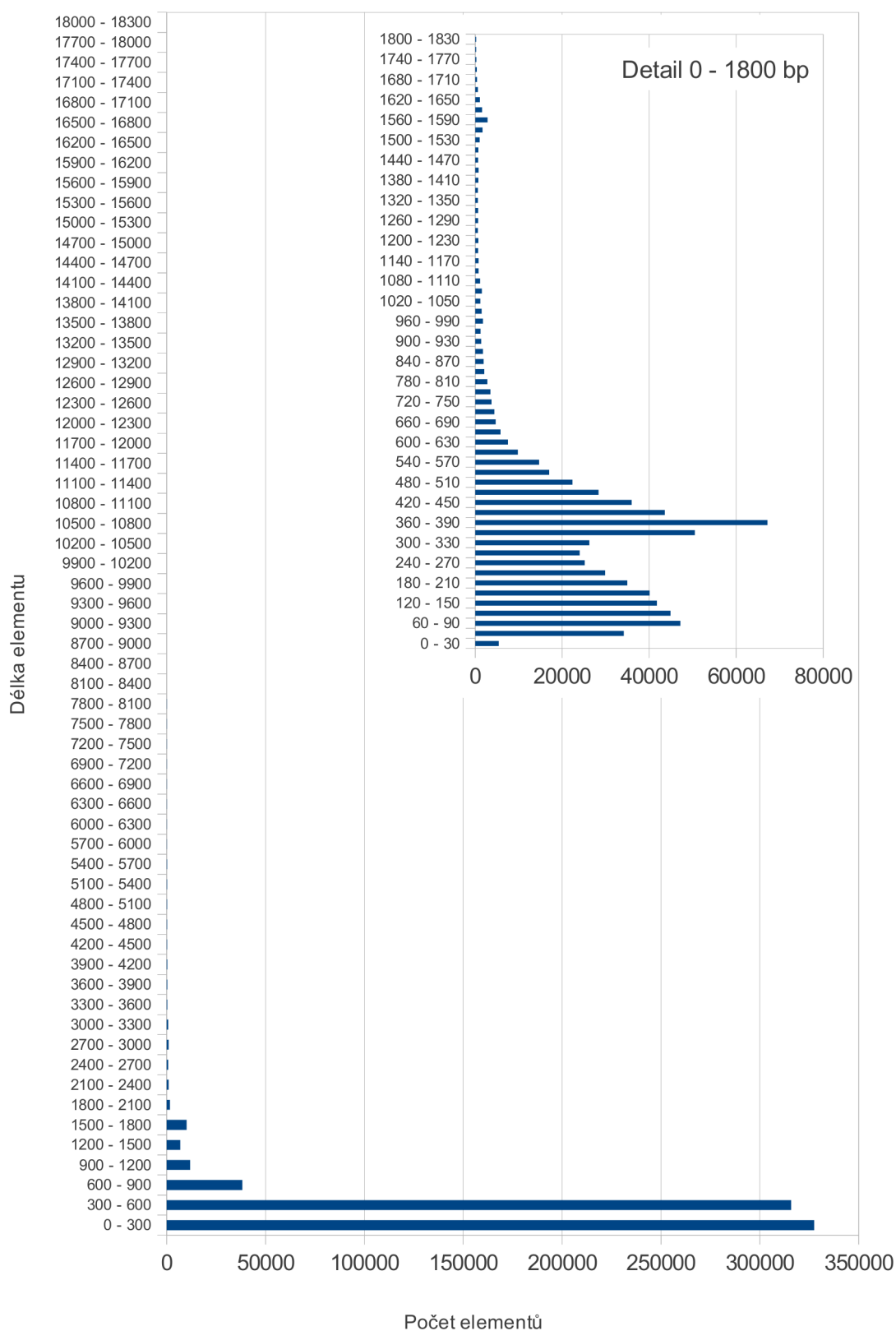
Histogram délek Alu transposonů



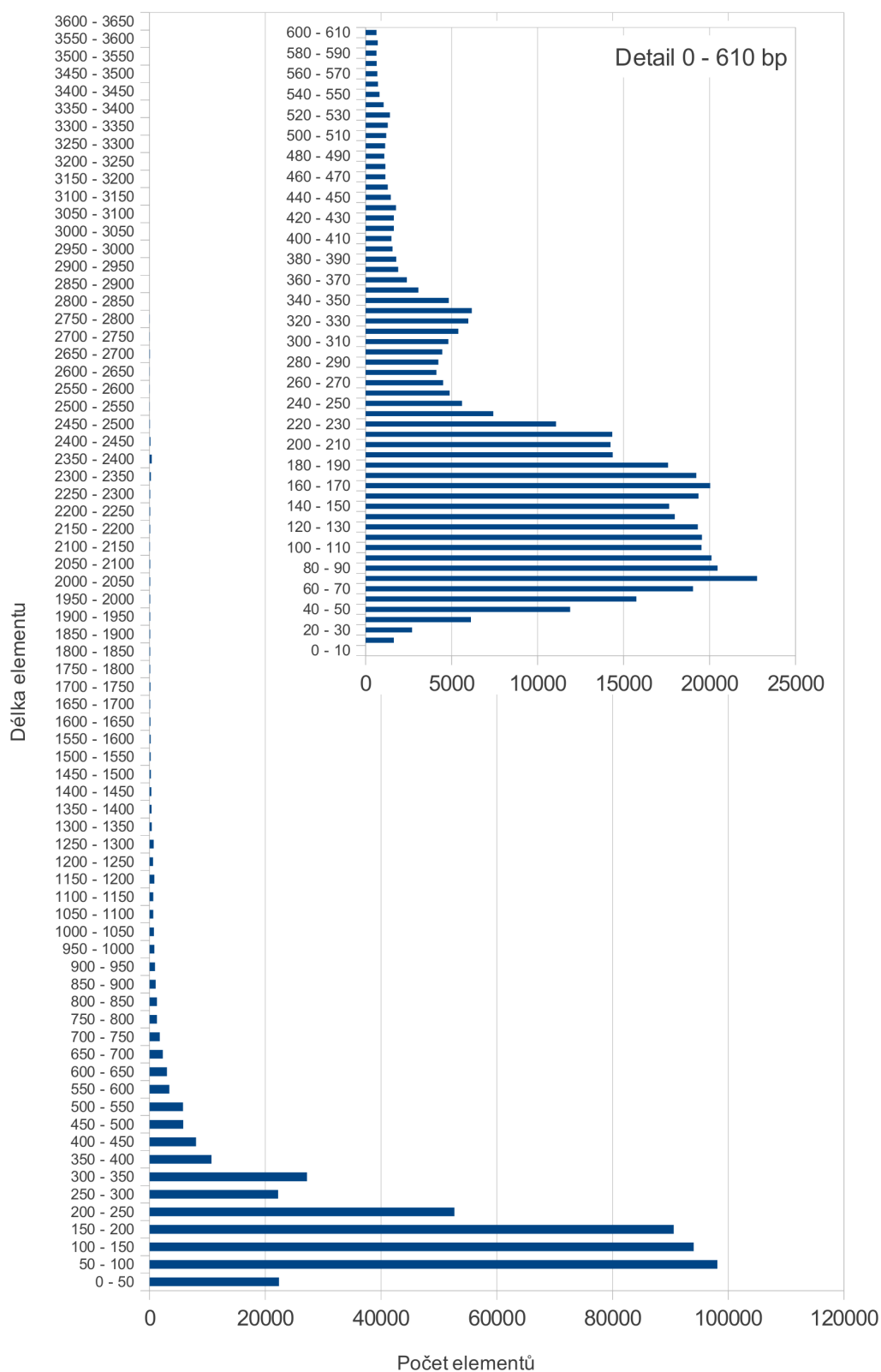
Histogram délek LINE transposonů



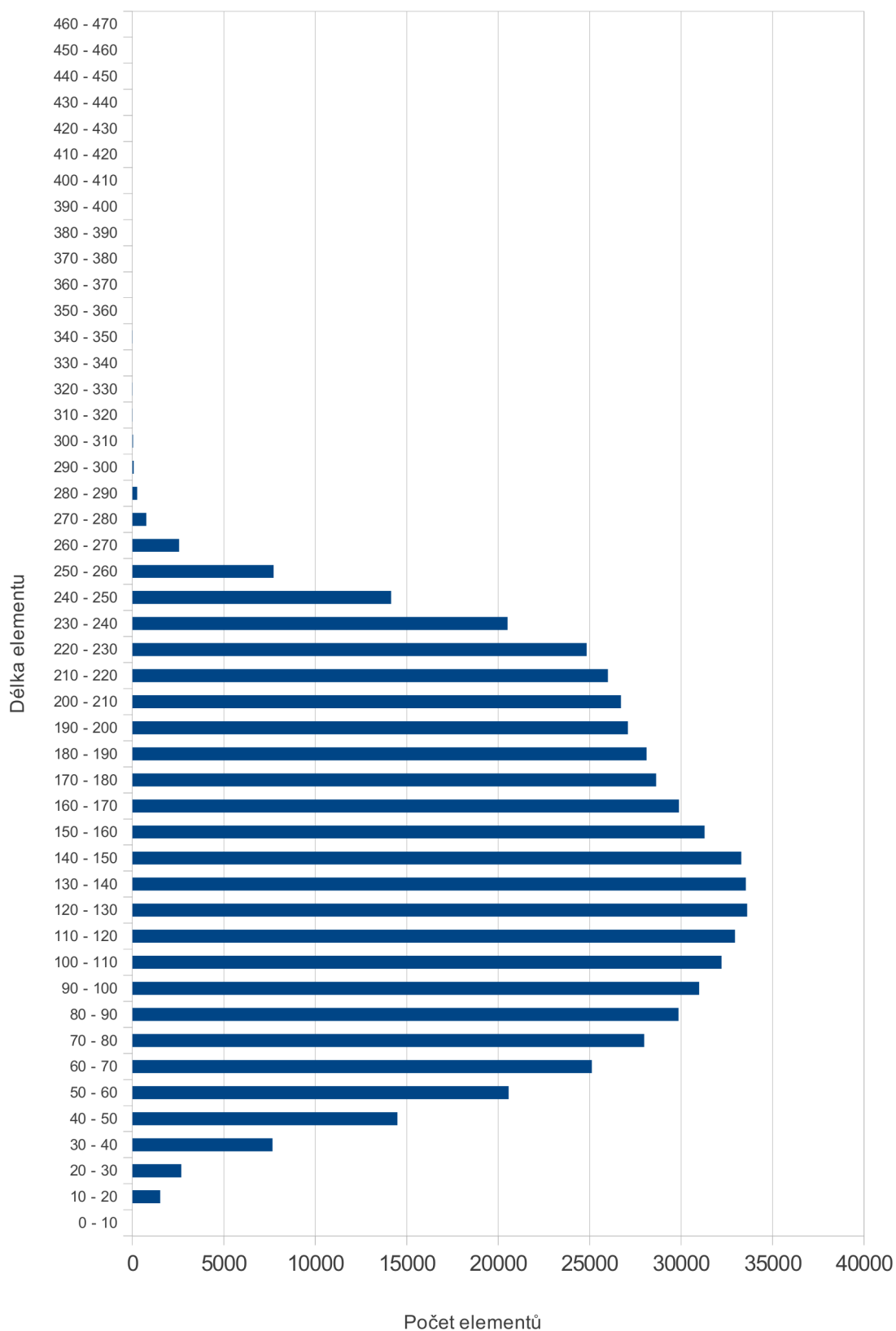
Histogram délek LTR transposonů



Histogram délek DNA transposonů



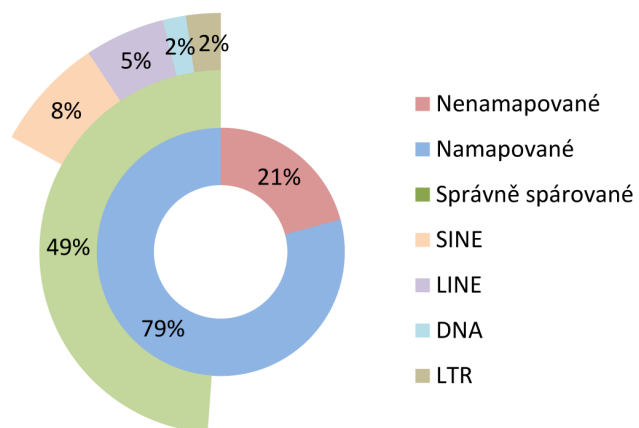
Histogram délek MIR transposonů



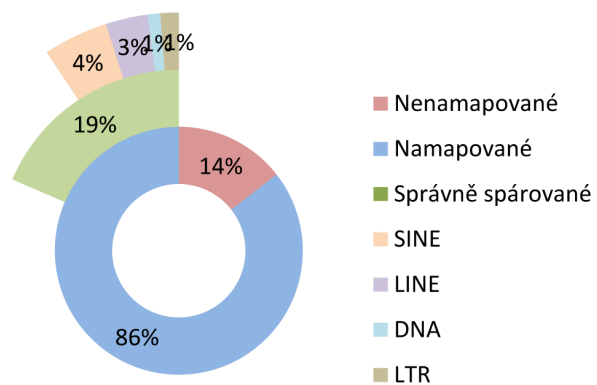
B Rozdělení readů

PROSTATA - RAKOVINA

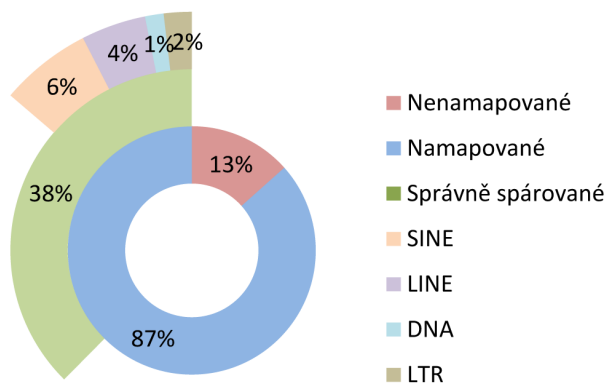
SRR057639



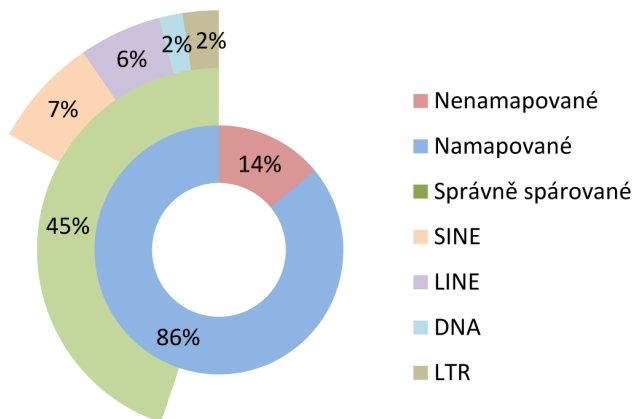
SRR057640



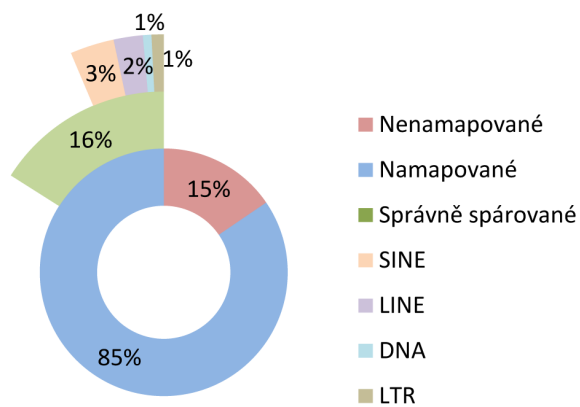
SRR057641



SRR057642

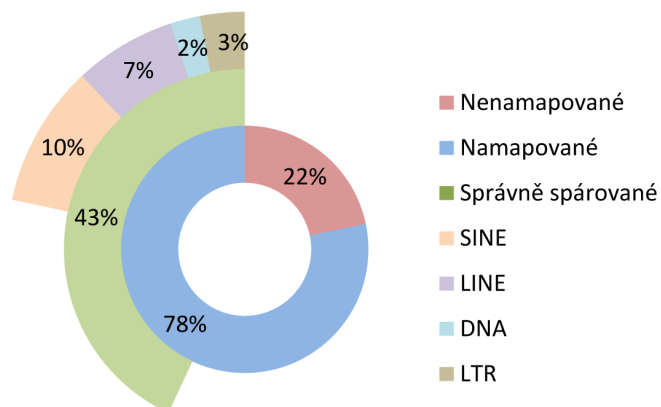


SRR057643

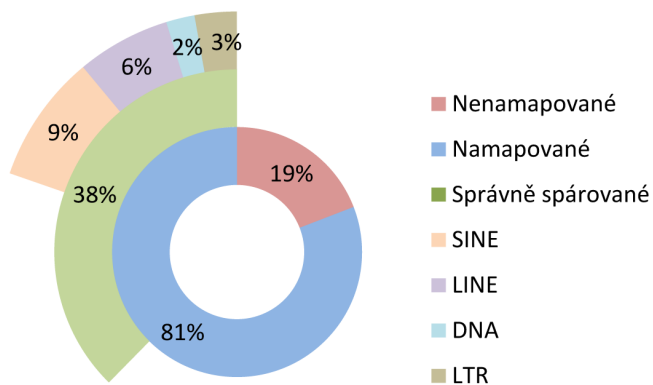


PROSTATA - NORMAL

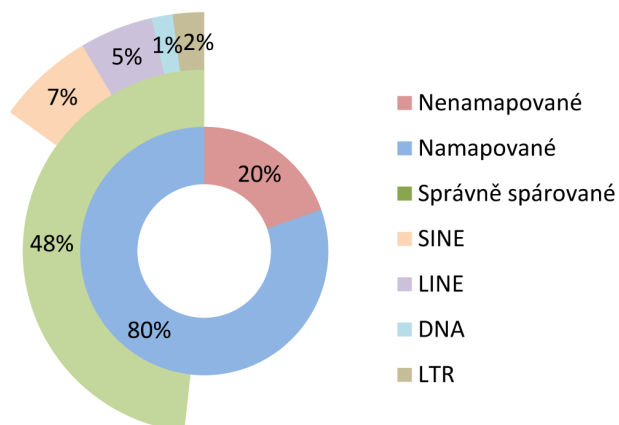
SRR057654



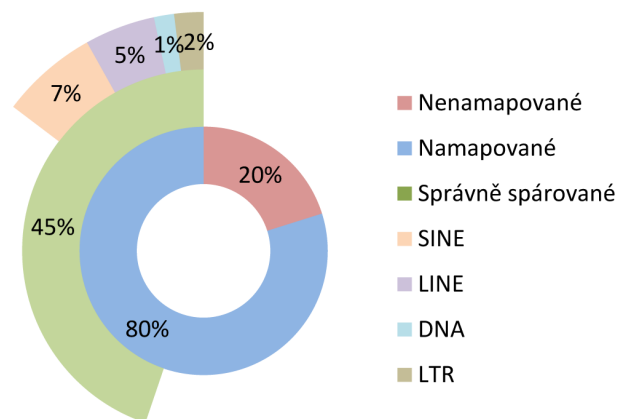
SRR057655



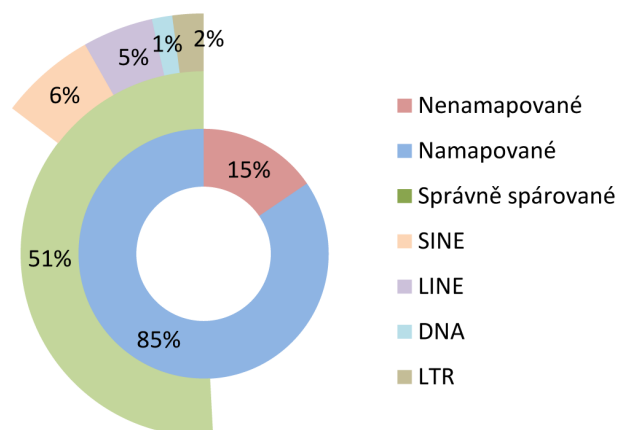
SRR05756



SRR05757

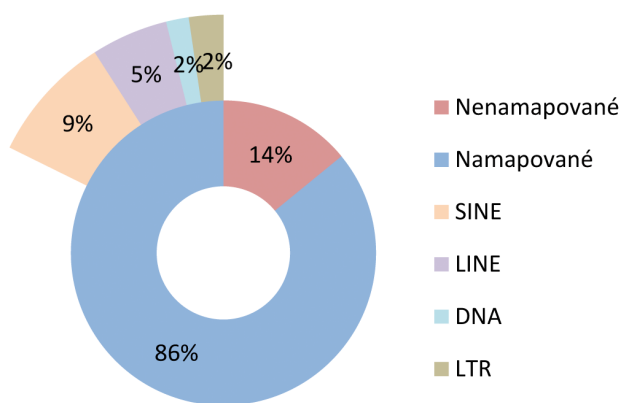


SRR057658

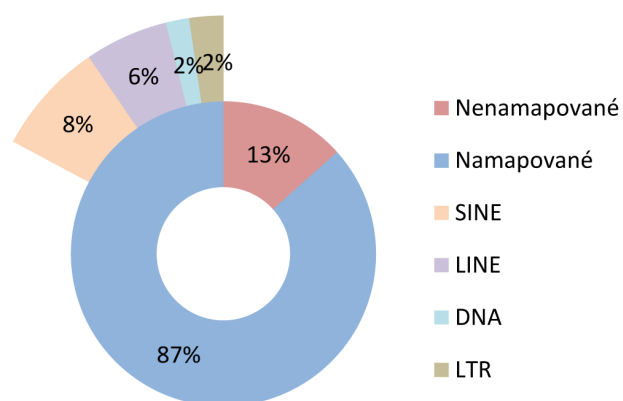


TLUSTÉ STŘEVO - RAKOVINA

SRR222176

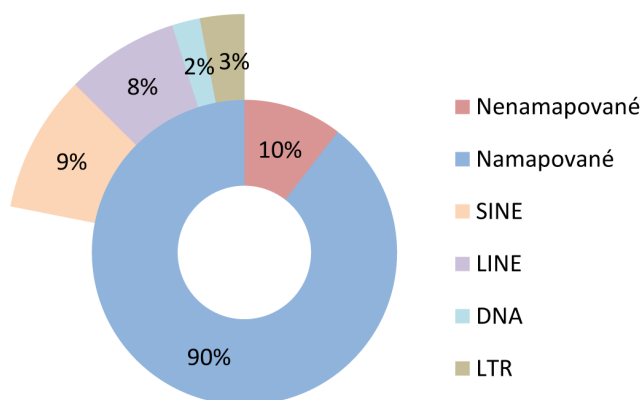


SRR222178

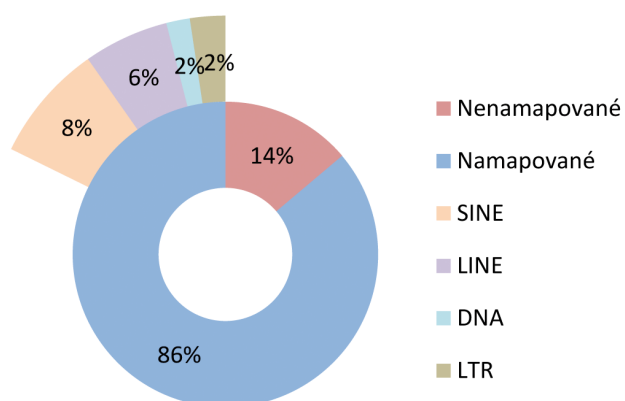


TLUSTÉ STŘEVO - NORMAL

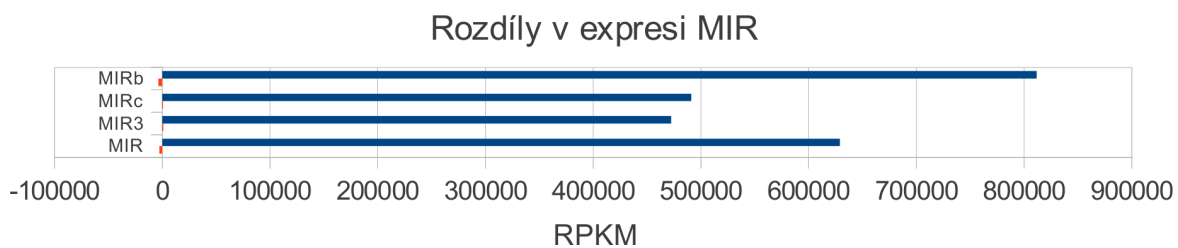
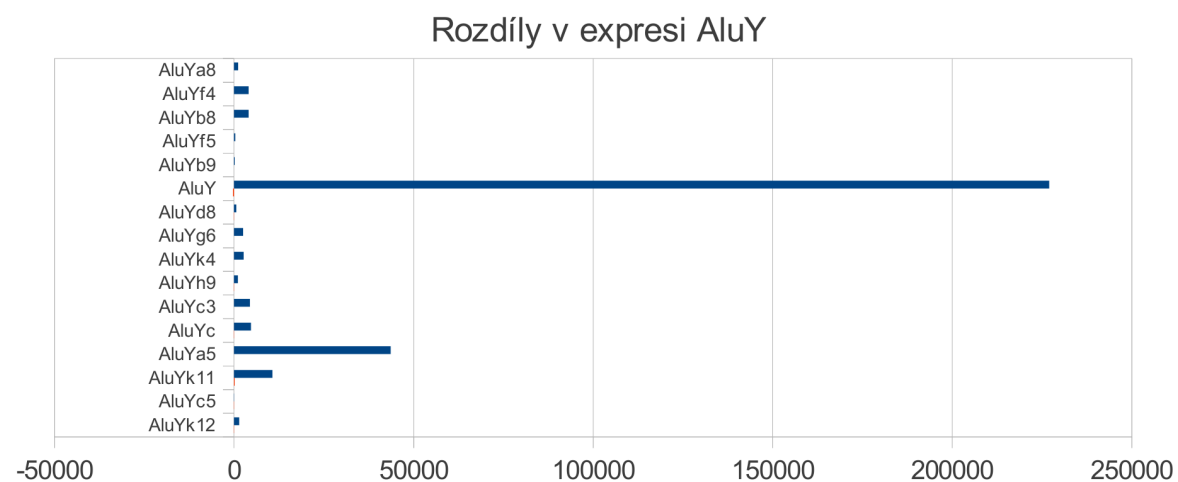
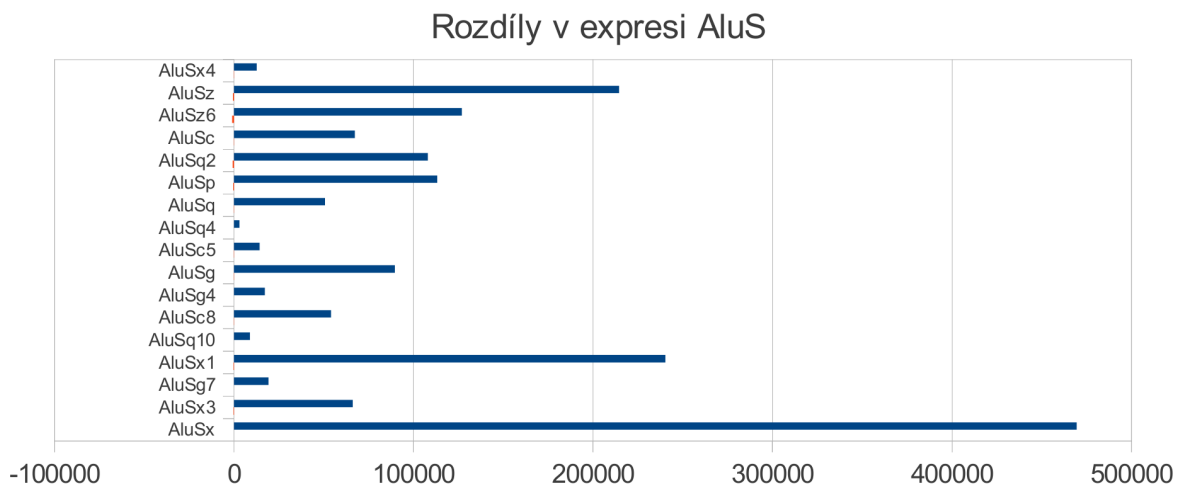
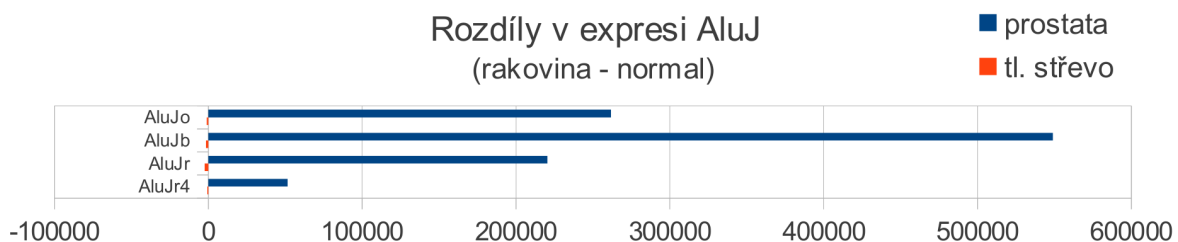
SRR222175



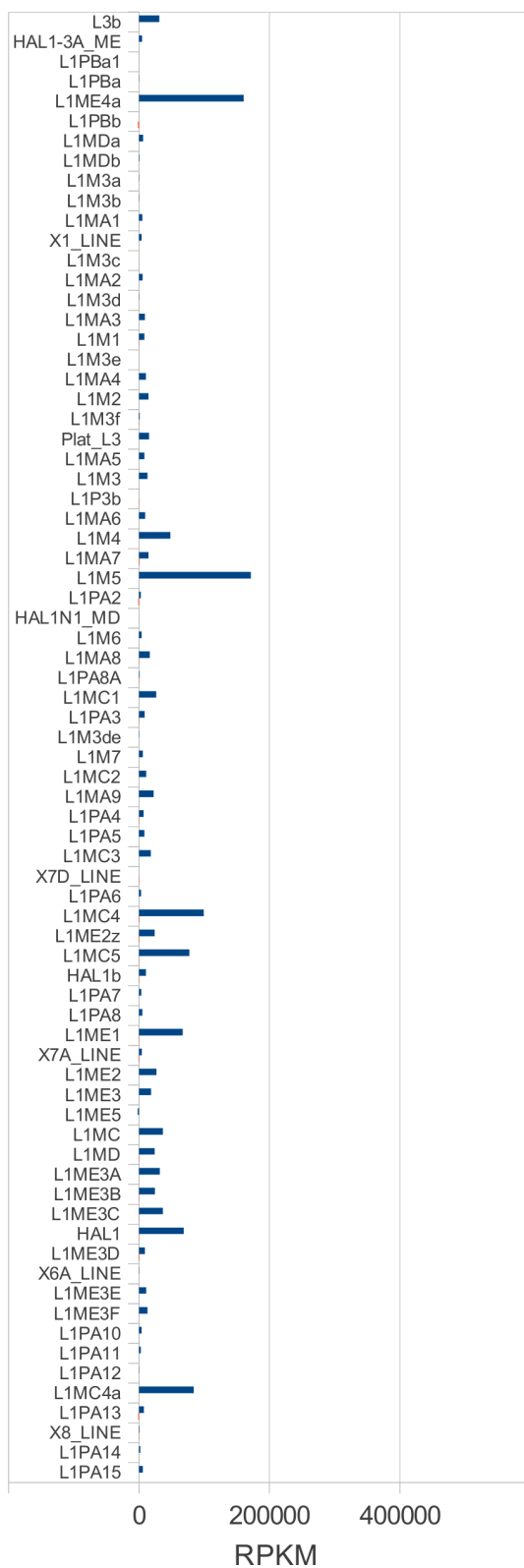
SRR222177



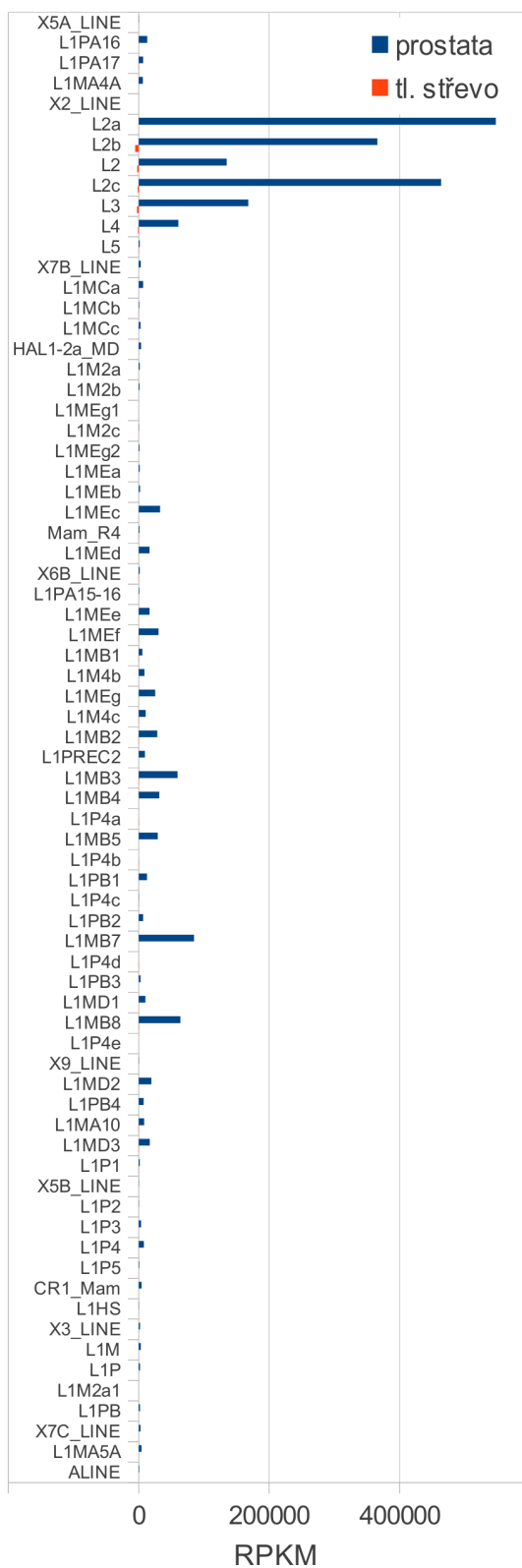
C Rozdíly v expresi rodin lidských transposonů

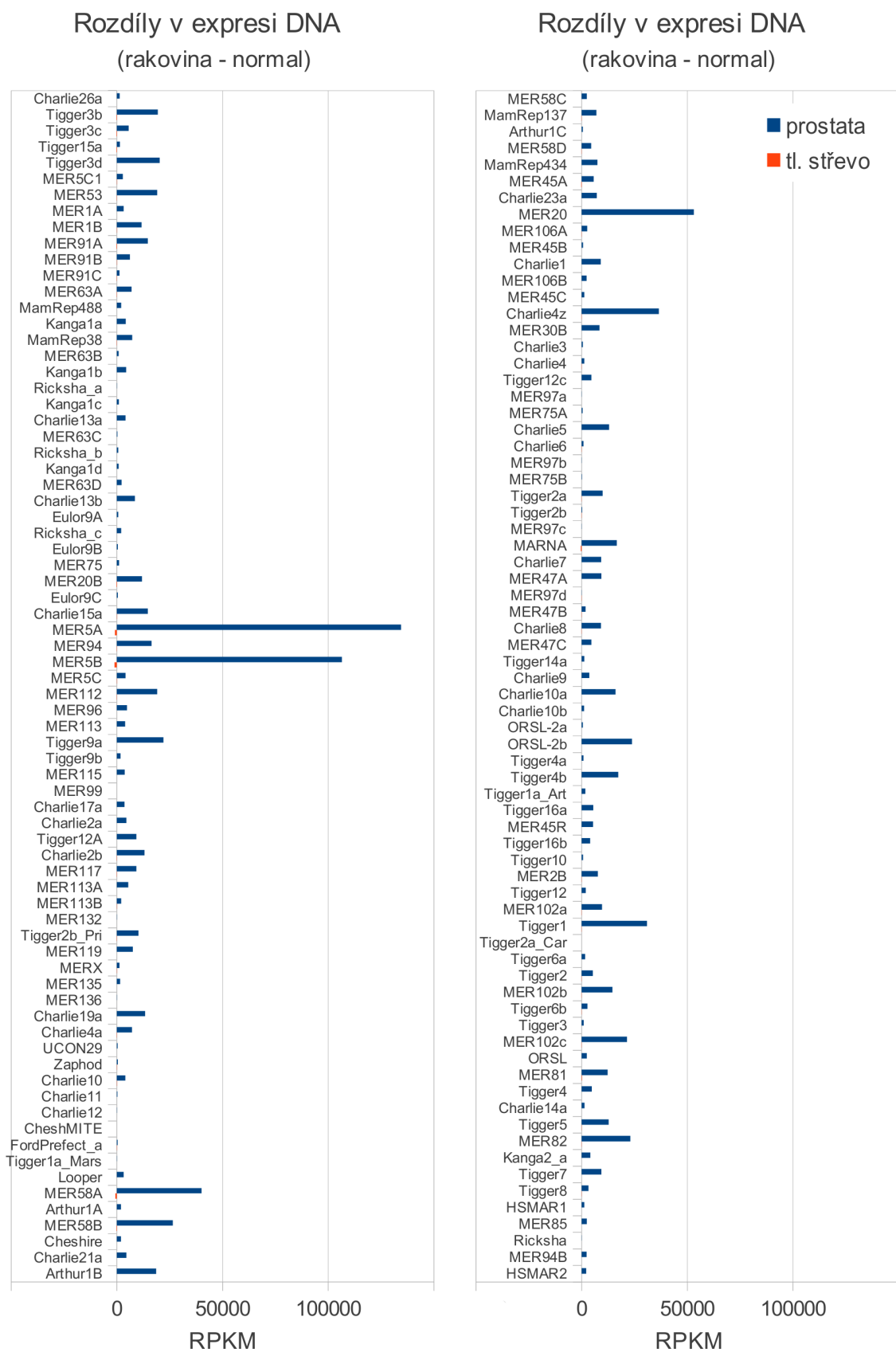


Rozdíly v expresi LINE
(rakovina - normal)

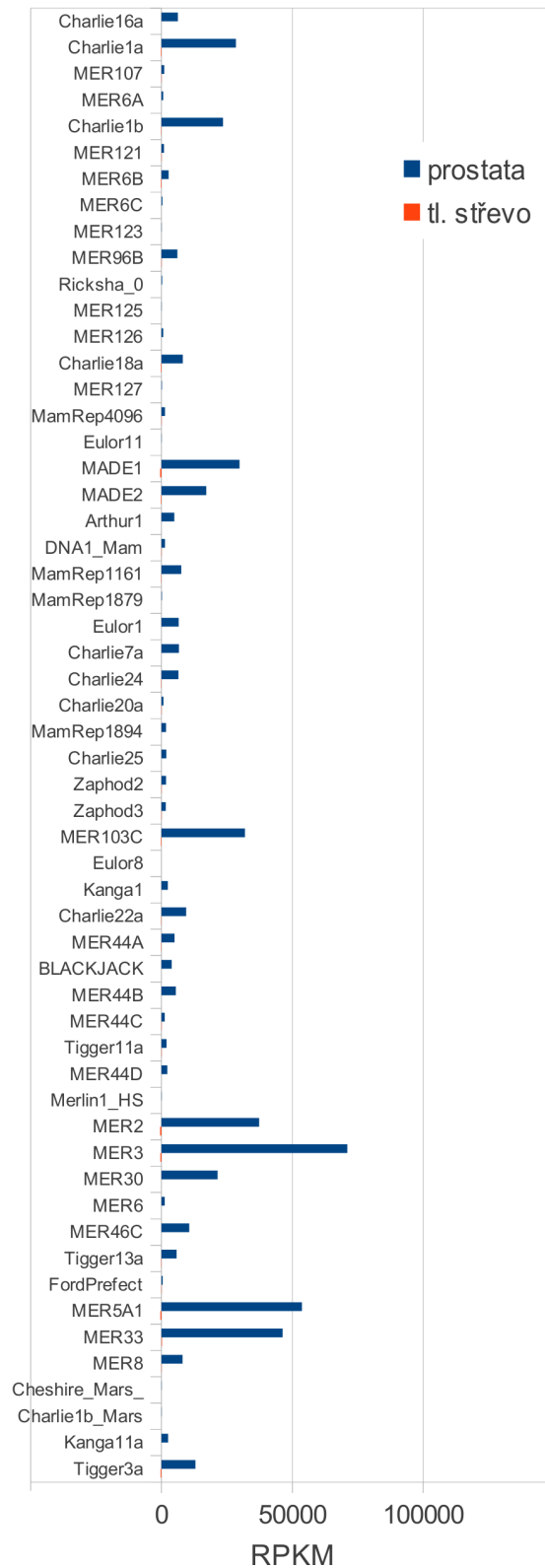


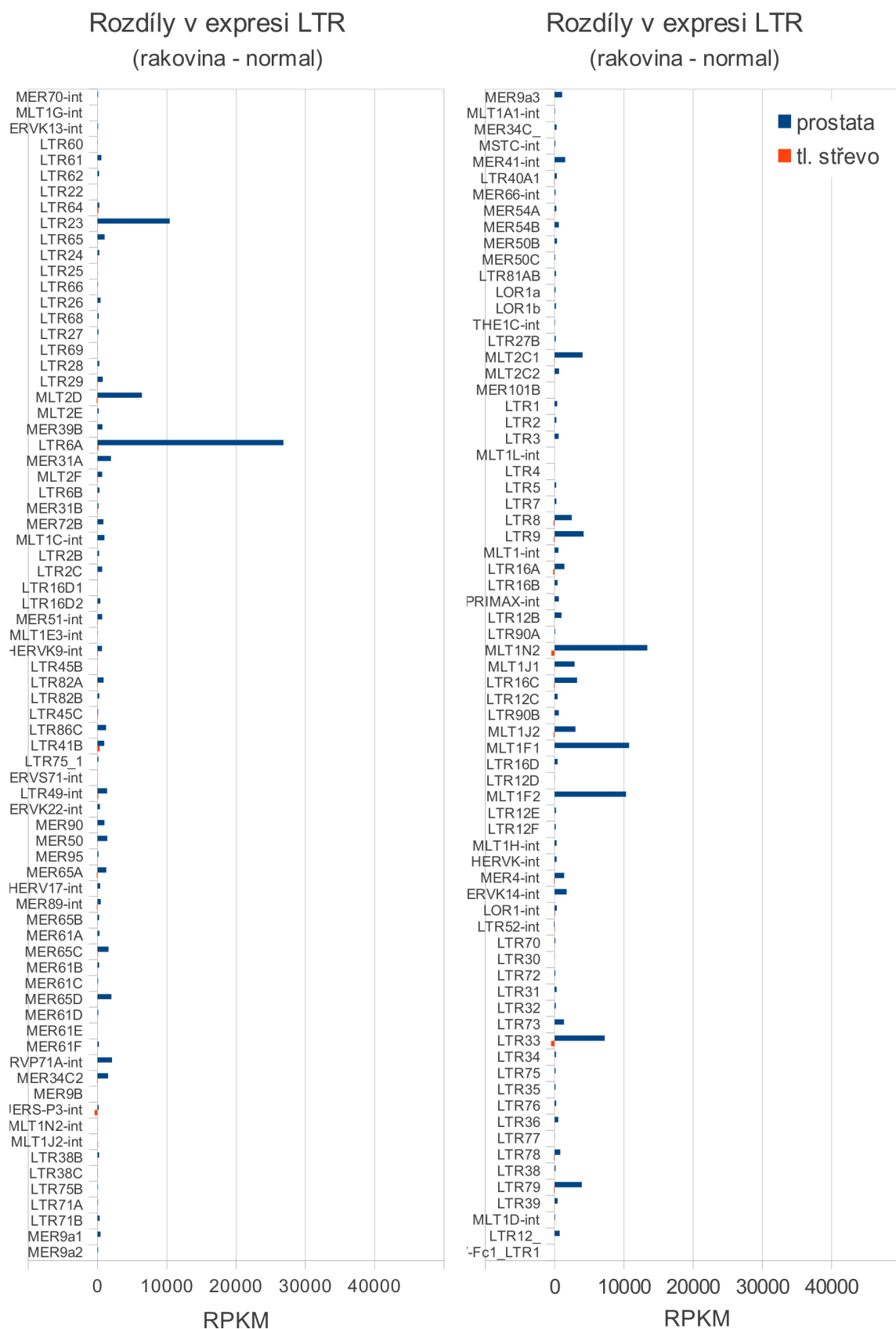
Rozdíly v expresi LINE
(rakovina - normal)

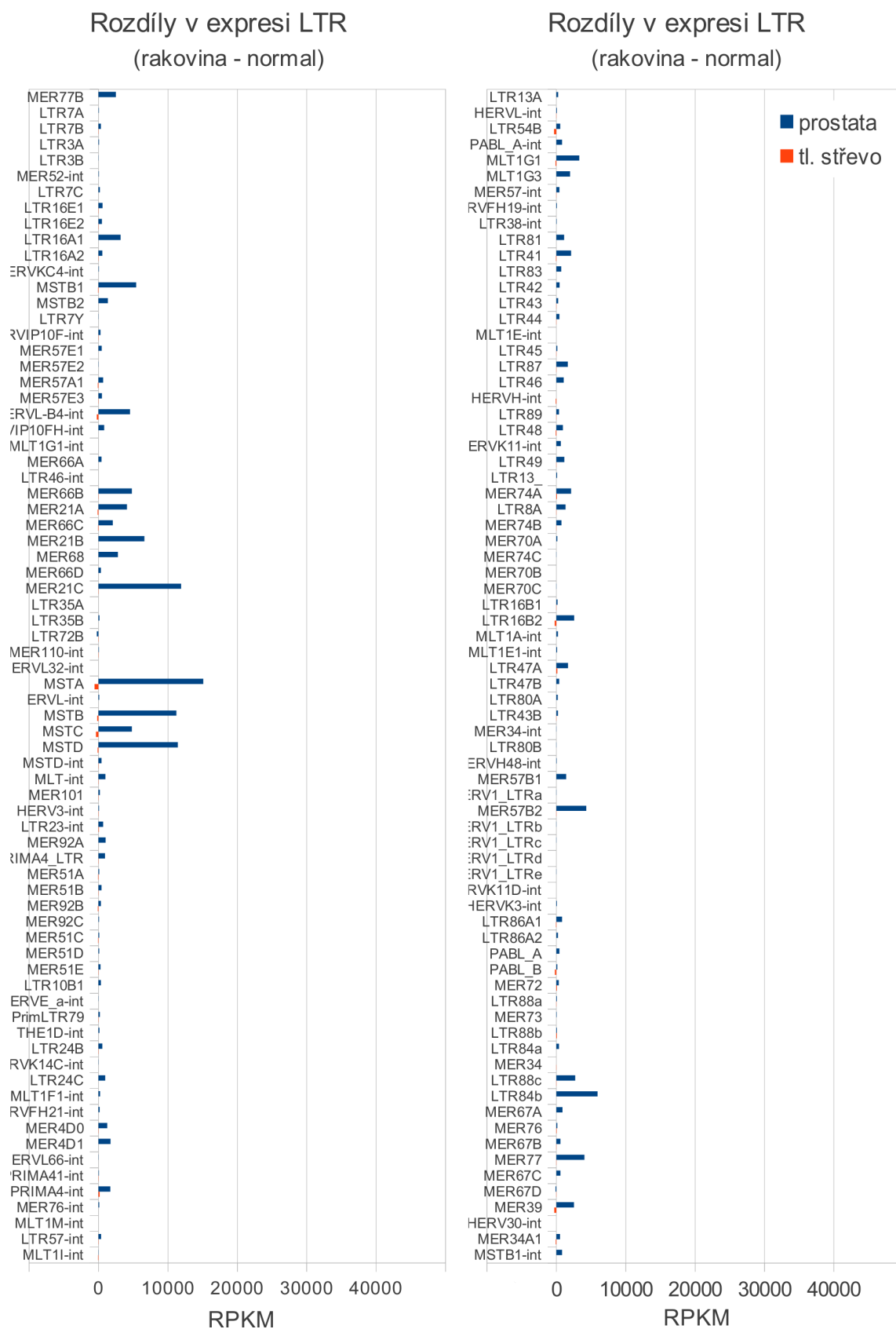




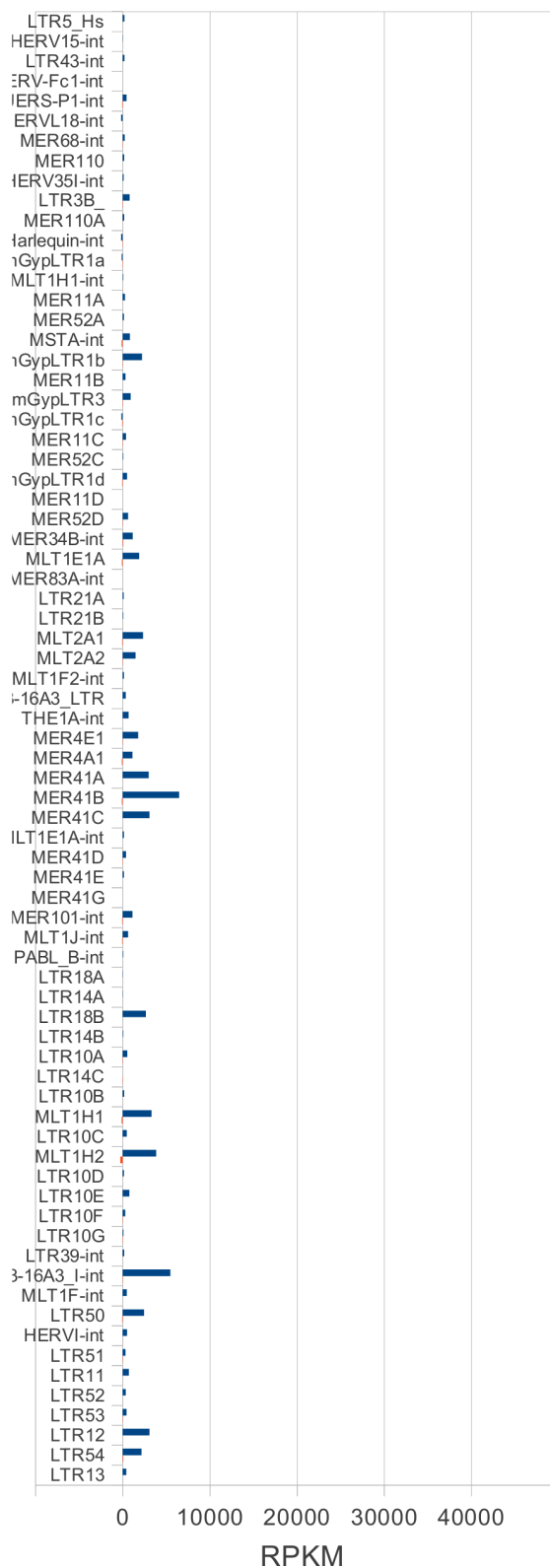
Rozdíly v expresi DNA (rakovina - normal)



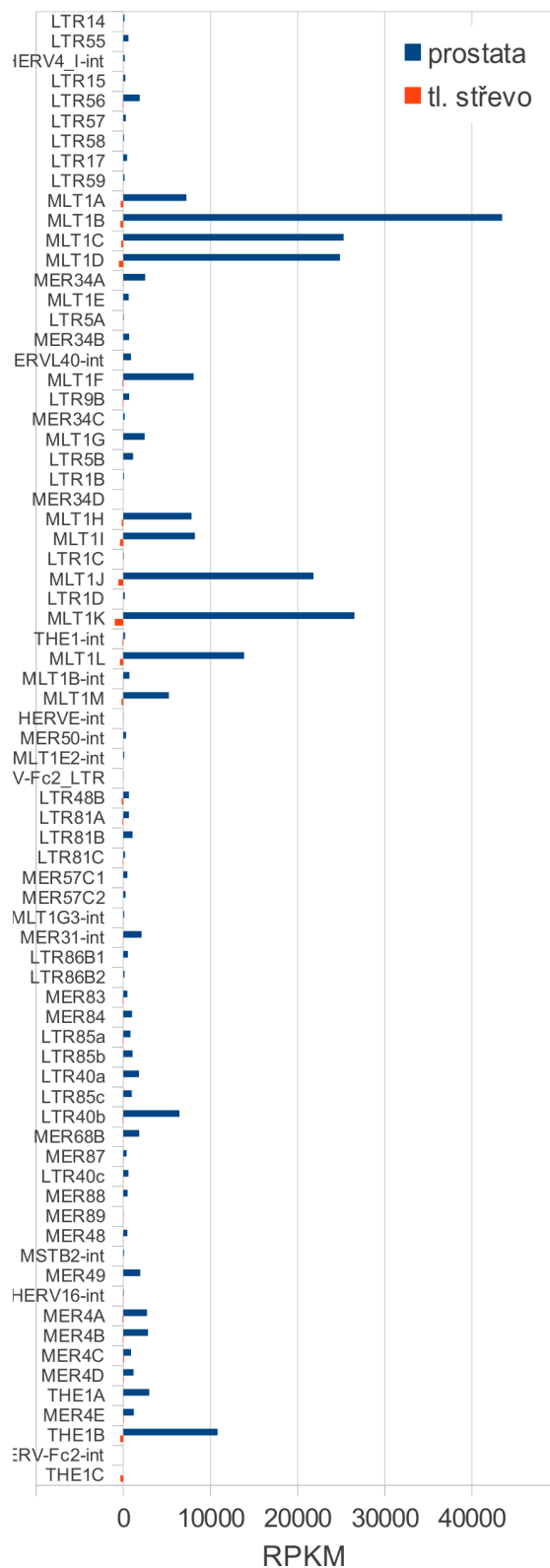




Rozdíly v expresi LTR
(rakovina - normal)



Rozdíly v expresi LTR
(rakovina - normal)



Rozdíly v expresi LTR (rakovina - normal)

