Univerzita Hradec Králové

Přírodovědecká fakulta

Katedra biologie

Evoluce GC architektury savců s důrazem na letouny (Chiroptera)

Diplomová práce

Autor: Bc. Kateřina Rodrová

Studijní program: Biologie a ekologie

Studijní obor: Biologie živočichů

Vedoucí práce: Mgr. Radka Symonová, Ph.D.

Konzultanti: RNDr. Michal Andreas, Ph.D., Ing. Dominik Matoulek

Hradec Králové 2021

Univerzita Hradec Králové
Přírodovědecká fakulta

Zadání diplomové práce

Autor:	**Bc. Kateřina Rodrová**
Studium:	S19BI008NP
Studijní program:	N0588A030001 Biologie a ekologie
Studijní obor:	Biologie a ekologie - spec. biologie živočichů
Název diplomové práce:	**Evoluce GC architektury savců s důrazem na letouny (Chiroptera)**
Název diplomové práce AJ:	Evolution of the GC Architecture in Mammals with Focus on Bats (Chiroptera)

Cíl, metody, literatura, předpoklady:

Diplomová práce se bude zabývat stanovením základních parametrů sekvenovaných genomů letounů (velikost a GC% celkové cDNA a cds) a jejich analýzou v kontextu všech dostupných genomů savců. Zdrojem dat (FASTA sekvencí) bude databáze Ensembl a pomocí jazyka Python budou data získána, analyzována, vyhodnocena a vizualizována.

1. V jazyce Python vytvořit automatickou pipeline ke stahování a analýze sekvencí ve formátu FASTA s výstupem ve formátu JSON
2. Analýza GC% a délky jednotlivých readů ze souborů cDNA a cds všech druhů savců dostupných v databázi Ensembl
3. Vizualizace výsledků s pomocí jazyka Python, statistické vyhodnocení výsledků u letounů a savců pomocí dalších nástrojů (MS Excel, jazyk R).
4. Vyhodnocení v kontextu dosavadních známých výsledků o základní GC architektuře savců.

Andreas et al., in prep. Transposons drive chiropteran quantitative cytogenomic evolution and follow bats phylogeny.

Nam and Ellegren. 2012. Recombination Drives Vertebrate Genome Contraction. *PLoS Genet* 8(5): e1002680. doi:10.1371/journal.pgen.1002680

Mugal CF, Weber CC, Ellegren H. 2015. GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *BioEssays* 37(12):1317-26

Teeling et al., 2017. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu. Rev. Anim. Biosci.* 6:23?46 https://doi.org/10.1146/annurev-animal-022516-022811

Tsagkogeorga et al., 2017. Comparative genomics reveals contraction in olfactory receptor genes in bats. *Sci Rep* 7: 259 | DOI:10.1038/s41598-017-00132-9

https://genome.cshlp.org/content/20/8/1001.full.pdf

Garantující pracoviště:	Katedra biologie, Přírodovědecká fakulta
Vedoucí práce:	Mgr. Radka Symonová, Ph.D.
Datum zadání závěrečné práce:	23. 1. 2020

Prohlášení:

Prohlašuji, že jsem diplomovou práci vypracovala samostatně a že jsem v seznamu použité literatury uvedla všechny prameny, ze kterých jsem vycházela.

V Hradci Králové dne Jméno a příjmení

Anotace

RODROVÁ, K. Evoluce GC architektury savců s důrazem na letouny (Chiroptera). Hradec Králové, 2021. Diplomová práce na Přírodovědecké fakultě Univerzity Hradec Králové. Vedoucí diplomové práce Radka Symonová. 63 s.

Savčí genom se liší v GC architektuře od ostatních obratlovců, protože vykazuje AT/GC heterogenitu, tedy střídání AT- a GC-bohatých oblastí patrné v sekvenci DNA. Ektotermní obratlovci a bezobratlí jsou AT/GC homogenní. Příčiny a mechanizmy vzniku AT/GC heterogenity savců zůstávají nejasné. Jedním z významných faktorů ovlivňujících GC architekturu je velikost genomu a jeho jednotlivých frakcí (kódující a nekódující). Letouni jsou zajímavou skupinou, která může pomoci pochopit kompoziční evoluci obratlovců. U letounů je známé výrazné zmenšení genomu a fylogeneticky podmíněné zvýšení obsahu GC (GC%) na úrovni celé DNA. Cílem této diplomové práce je vytvořit pipeline (tedy nástroj, který pomáhá automatizovat pracovní postupy) v jazyce Python k analýze základních parametrů osekvenovaných genomů letounů v kontextu všech savců a ověření hypotézy, zda jsou letouni GC bohatší i na ostatních úrovních genomu (cDNA a cds). Zdrojem genomických je databáze Ensembl, kde jsou v době psaní této práce 3 genomy letounů a 89 genomů ostatních savců. I přes nízký počet letounů v datasetu výsledky ukázaly, že letouni jsou GC bohatší i na úrovni cDNA a cds. To ukazuje, že u letounů nemusí být hlavní příčinou GC bohatosti pouze transpozony, jak naznačoval dosavadní výzkum. Cíle této práce byly splněny a práce položila základy dalšímu výzkumu kompoziční biologie savců. Výsledky budou součástí připravované publikace.

Klíčová slova

Genom, GC%, letouni, savci

Annotation

RODROVÁ, K. Evolution of the GC Architecture in Mammals With Focus on Bats (Chiroptera). Hradec Králové, 2021. Diploma Thesis at Faculty of Science University of Hradec Králové. Thesis Supervisor Radka Symonová. 63 p.

The mammalian genome is fundamentally different from other vertebrates because it exhibits AT/GC heterogenity, i. e. alternation of AT- and GC-rich regions evident in the DNA sequence. Extothermic vertebrates and invertebrates are AT/GC homogeneous. Exact causes and mechanisms of mammalian AT/GC heterogenity remain unclear. One of the significant factors affecting the GC architecture is the size of the genome and its individual fractions (coding and non-coding). Bats are an interesting group that may help to understand the compositional evolution of vertebrates. There is a known simultaneous significant genome reduction and phylogenetical increase in GC content (GC%) at the whole DNA level. The aim of this diploma thesis is to create a Python pipeline (utility to help automate workflows) to analyze the basic parameters of sequenced chiropteran genomes in the context of all mammals, and to test the hypothesis that bats are GC-rich at other genome levels (cDNA and cds). The genomic source is the Ensembl database, which, at the time of writing this thesis, contains 3 bat genomes and 89 genomes of other mammals.

Despite the low numbers of bats in the dataset, the results showed that bats are also GC rich at the cDNA and cds levels. This indicates that in bats, transposons alone may not be the main cause of GC richness, as previous research has suggested. The goals of this work were fulfilled and lay the foundation for further research in mammalian compositional biology. The results will be included in a forthcoming publication.

Keywords

Genome, GC, bats, mammals

Obsah

1. Úvod 6
1.1 Velikost genomů eukaryot 8
1.2 cDNA 10
1.3 cds 10
1.4 Odhady GC % na úrovni DNA, cDNA a cds 11
1.5 C-value paradox a C-value enigma 12
1.6 Fylogeneze žijících savců 13
1.7 Letouni jako jediní létající savci 15
2 Materiál a metodika 17
2.1 Online databáze Ensembl a NCBI 17
2.2 Úvod k jazyku Python 18
2.3 Sada vlastních nástrojů v Pythonu 19
2.3.1 Analýza dat dostupných z databáze Ensembl 19
2.3.2 Data z NCBI a Animal Genome Size Database 30
2.4 Statistické testování výsledků 31
3 Výsledky 31
3.1 Základní statistické vyhodnocení GC% letounů v rámci savců 31
3.2 Histogramy GC% pro sekvence cds a cDNA 33
3.3 Tabulka s výpočty a grafy 34
3.4 Kvantitativní pohled na získané výsledky 35
4 Diskuse 38
4.1 Výsledky v kontextu současného stavu znalostí o kompoziční evoluci obratlovců 38
4.2 Kvalita genomových assembly 39
4.3 Vztah počtu cds sekvencí a velikosti exomu 40
4.4 Využití výsledků a výhled do budoucna 40
5 Závěr 41
6 Literatura 42
7 Přílohy 48

Poděkování:

Ráda bych poděkovala především Mgr. Radce Symonové, Ph.D., za motivaci, přátelský přístup, obohacující konzultace, možnost zapojení do projektu Erasmus+ KA 203-061433 na Katedře biologie PřF UHK a zprostředkování programátorských online kurzů. Nemalý dík patří Ing. Dominiku Matoulkovi, který mi vždy trpělivě a s ochotou pomohl s technickými záležitostmi týkající se práce v Pythonu. Děkuji také RNDr. Michalu Andreasovi, Ph.D., za cenné připomínky. Dále děkuji Mgr. Veronice Borůvkové za pomoc se statistickým testováním mých výsledků v jazyce R a výběr vhodné testovací metody. Rovněž děkuji komunitě hradeckých PyLadies zodpovědné za mé první programovací krůčky a Mgr. Ivaně Brabcové, Ph.D., za organizaci těchto kurzů v Hradci Králové.

1. Úvod

Předkládaná práce se zaměřuje na kompoziční analýzu genomických dat letounů (Chiroptera) a dalších zástupců savců z online dostupné genomové databáze Ensembl (*http://www.ensembl.org/).* Letouni (Chiroptera) byli vybráni jako významná modelová skupina savců se zajímavými a dosud ne zcela pochopenými a zdaleka neprobádanými kompozičními vlastnostmi genomu (Andreas et al., *in prep*).

Makromolekulární komplex jaderné DNA (nuDNA) má podobu pravotočivé dvoušroubovice díky dvěma antiparalelním vláknům DNA (3‘→5‘ a 5‘→3‘) a vodíkovým můstkům mezi jejich dusíkatými bázemi, které obě vlákna spojují. Na základě pravidla komplementarity (Watson et Crick, 1953) se báze párují typicky A=T (adenin-thymin) a G≡C (guanin-cytosin). Existuje však celá řada výjimek v podobě alternativního párování bází – kromě zmíněného tzv. watson-crickovského např. hoogsteenovské párování, kdy se jedná o vazby vodíkových můstků mezi bázemi už spárovanými i volnými, které spojují nekomplementární báze nebo komplementární báze v jiných polohách než obvyklé vazby (kanonické). Vznikají tak tzv. nekanonické struktury, jež mohou vytvářet i triplexy a kvadruplexy, například guaninu (Buček, 2008; Takahashi et Sugimoto, 2021). Navíc se struktura molekul DNA může měnit v závislosti na okolním prostředí – přesná konformace určité molekuly DNA nebo jejího úseku závisí i na povaze molekul, s nimiž interaguje (Snustad et Simmons, 2009). Byla zjištěna existence i levotočivé DNA, tzv. (Z-DNA), která se vyskytuje u dvoušroubovic bohatých na GC, u níž se usuzuje, že by mohla souviset s vývojem a rozvojem některých onemocnění (Snustad et Simmons, 2009; Ravichandran et al., 2019). V literatuře se poslední desetiletí diskutuje nad významem obsahu GC (tedy GC%) v genomech a jeho evolucí, a to jak u rostlin (např. Šmarda et al., 2014) a živočichů (např. Symonová et al., 2017) tak u virů (Matyášek a Kovařík, 2020), bakterií (Hildebrand et al., 2010) a dalších organizmů. V dostupných publikacích se uvádí, že tři vodíkové můstky mezi G a C jsou příčinou vyšší stability DNA bohaté na GC (např. Snustad et Simmons, 2009; Vinogradov et Anatskaya, 2017). Galtier et al. (2001) uvádí, že oblasti bohaté na GC obsahují geny s krátkými introny, kdežto v oblastech chudých na GC se sekvencí kódující proteiny vyskytuje podstatně méně (Bernardi, 2005). Také se předpokládalo, že tendence k vyššímu obsahu GC (tj. GC%) koreluje s termostabilitou. Souhrn velkého množství studií na toto téma podává Bernardi (2005). To však bylo vyvráceno jak u prokaryot

(Hurst et Merchant, 2001), tak u eukaryot (Ream et al., 2003). Stabilita dvoušroubovice DNA je však dána především interakcemi mezi bázemi – tzv. base-stacking (patrové interakce) jsou dominantní složkou celkové stability složitějších struktur DNA (Yakovchuk et al., 2006). V rámci obratlovců existují značné rozdíly v uspořádání na GC bohatých a na GC chudých oblastí v genomu mezi savci a ptáky na jedné straně a ostatními skupinami na straně druhé. U vyšších obratlovců (savci a ptáci) byla zjištěna tzv. AT/GC heterogenita, kdy se na chromozomech střídají oblasti DNA bohaté či bohatší na GC s oblastmi DNA na GC chudými či chudšími. Genomy nižších obratlovců a bezobratlých jsou pak z tohoto pohledu tzv. AT/GC homogenní, protože se u nich výše popsané zásadní rozdíly v GC% nevyskytují (Bernardi, 2005). Tato AT/GC heterogenita je patrná na chromozomech při cytogenetické analýze stejně jako na sekvenci DNA při bioinformatické analýze (např. Symonová et al., 2017). Přesné důvody a mechanizmy vzniku AT/GC heterogenity u vyšších obratlovců dosud nejsou známy. Toto téma v letošním roce detailně rozpracovala Symonová, 2021. Jedním ze současně nejpřijímanějších navrhovaných vysvětlení tohoto fenoménu je tzv. GC-vychýlená genová konverze (GC-biased gene conversion, gBGC) související s meiotickou rekombinací a řídící evoluci bází, kdy je u AT/GC heterozygotů preferována fixace alely s G nebo C, bez ohledu na výslednou fitness organizmu (Mugal et al., 2015; Weber et al., 2014; Kostka et al., 2012).

V průběhu evoluce druhů dochází v jaderných genomech k neustálým změnám, přičemž letouni vykazují znatelné zmenšení genomu (Kapusta et al., 2017) a výraznou dynamiku chromozomové evoluce (Soitero et al., viz Andreas et al, *in prep*). Změna na molekulární úrovni však nezaručuje, že dojde i k fenotypové změně – jedním z důležitých poznatků je skutečnost, že se geny během evoluce často duplikují a duplikované geny někdy získávají odlišné funkce (Snustad et Simmons, 2009). K získání nových zajímavých informací nejen z genetiky se tak postupně dostává do popředí bioinformatika, tedy obor s vyváženou kombinací znalostí jak z biologie, tak z informatiky, která získává a analyzuje zejména molekulárně-biologická data z osekvenovaných genomů a může tak významně přispět k osvětlení zatím nezodpovězených otázek.

1.1 Velikost genomů eukaryot

Pojmem jaderný genom rozumíme celou genetickou informaci uloženou v buněčných jádrech organismu (nuDNA; u některých virů v RNA) zahrnující kódující i nekódující oblasti. Eukaryota mají nejen mnohem více nuDNA oproti prokaryotům, ale jejich nuDNA je navíc během metafáze sbalena do různého počtu chromozomů, z nichž je každý obvykle přítomen ve dvou (diploidie) nebo více kopiích (polyploidie) (Snustad et Simmons, 2009). Eukaryota mají další, dodatečné, genomy ve svých organelách – mitochondriální a plastidový, vzniklé endosymbiózou. O velikosti genomu je již dlouho známo, že koreluje s velikostí buňky, rychlostí dělení buněk a dalšími vlastnostmi organismu, např. metabolismem, rychlostí individuálního vývoje (čili vývinu) či velikostí těla (Elliott et Gregory, 2015). Velikosti genomů zjištěné na základě genomických dat se uvádí v jednotkách *bp* (base-pairs), tzn. jednoho páru dusíkatých bází, tedy proti sobě stojící báze z každého komplementárního řetězce, od čehož se odvíjejí další úrovně výpočtu – kbp (1 000 bp), Mbp (1 000 000 bp) a Gbp (1 000 000 000 bb). Ačkoliv může jaderný genom dosáhnout ohromných rozměrů a liší se u různých druhů organismů, pouze jeho malá část, tzv. exom, kóduje sekvenci aminokyselin. Exom je soubor exonů všech genů, které jsou odděleny introny. Exom je zatím nejlépe prozkoumán především u člověka, kde mu však náleží pouze 1,15-1,17 % z celkového genomu (Zhao, 2012; Hatje et al., 2019). Napříč živočišnou říší jsou dostupná kvalitní data o procentuálním zastoupení exomu v celém genomu pro 5 dalších druhů: *Amphimedon queensladnica* (Porifera, 24,45 %), *Sycon ciliatum* (Porifera, 9,95 %), *Mnemiopsis leidyi* (Ctenophora, 17,37 %), *Nematostella. vectensis* (Cnidaria, 11,68 %) a danio pruhované (*Danio rerio*, 4,67 %) (de Mendoza et al., 2019). Např. masožravá rostlina bublinatka opomíjená (*Utricularia gibba*) je přesným opakem – díky svému výjimečně malému eukaryotickému genomu nekódující oblast představuje pouze 3 % (Ibarra-Laclette et al., 2013). Dalším příkladem budiž axolotl mexický (*Ambystoma mexicanum*), oblíbený modelový organismus známý svou neotenií, jehož haploidní genom čítá 32 Gbp, což je ve srovnání s člověkem přibližně 10x víc (lidský genom dosahuje velikosti přes 3 Gbp; Nowoshilow et al., 2018, Piovesan et al., 2019). Ovšem počet repetitivních sekvencí u axolotla dosahuje až 65,6 % celkového genomu (18,6 Gbp). U člověka se udávaná proporce repetitivních sekvencí zvyšuje téměř s každou verzí nově zveřejněného referenčního genomu (nyní je to verze GRCh38.p13, NCBI), zejm. pak s použitím nových long-read technologií (diskutováno v Symonová, 2021). Současná literatura uvádí pro lidský genom

50 % repetic (Healton et al., 2020, Nowoshilow et al., 2018; Treangen et Salzberg, 2012). Avšak existují náznaky, že frakce lidského genomu, která je repetitivní nebo odvozená od repetitivních sekvencí dosahuje velikosti až 66-69 % (de Koninng et al., 2011). Nedávno první příčku největšího osekvenovaného živočišného genomu po axolotlovi obsadil bahník australský (*Neoceratodus forsteri*) s velikostí genomu 43 Gbp (asi o 30 % větší než axolotl; Meyer et al., 2021).

Na různých úrovních DNA probíhá evoluce, ať už mutací, rekombinací či duplikací a ukazuje se, že velkou roli v ní hrají transpozony, které v dnešní vědě představují rozsáhlé a důležité téma biomedicínského výzkumu. Transpozibilní elementy (TE, nebo také mobilní genetické elementy, „skákající geny“) jsou sekvence DNA se schopností měnit v genomu svou pozici a počet kopií (Bourque et al., 2018). Zkoumá se jejich eliminace (degradace)/přírůstky (inzerce) během evoluce (Kapusta et al., 2017), jejich funkce, distribuce či epigenetické projevy na fenotypu (Bourgeois et Boissinot, 2019). Kromě jejich přemisťování, tzv. transpozice, odtud jejich název, se také ukázalo, že jsou z velké míry epigeneticky umlčovány, čímž je redukován jejich efekt na genom hostitelského organismu. Tyto epigenetické modifikace však mohou zpětnovazebně vést k reaktivaci TE, neboť metylovaný cytosin je hypermutabilní, což vede v konečném důsledku ke snížení proporce cytosinu, a tudíž snížení proporce substrátu pro epigenetické regulace TE (Zhou et al., 2020). Pokud dojde např. vlivem působení dlouhodobého a závažného stresu k mobilizaci transpozonů v genomu, záleží na tom, kam se integrují, resp. kde dojde k jejich inzerci (Pray, 2008). Jak již bylo nastíněno výše, eukaryota vykazují v genomech velké rozdíly ve velikostech – studie Zhou et al. (2020) popisuje, že transpozony za toto nesou podstatný díl zodpovědnosti, která však do značné míry závisí na působení DNA metyltransferázy, která potlačuje transkripci metylovaných CpG ostrůvků. Dopad evoluce na velikost genomu je široce diskutován, jelikož se zdá, že velká část organizmů prošla jeho více či méně dramatickým zvětšení či naopak zmenšením. Např. hlodavec *Tympanoctomys barrerae* z čeledi osmákovitých (Octodontidae) vykazuje ze všech známých savců nejvyšší počet chromozomů (2n = 102, ca 6,19 Gbp) (Evans et al., 2017); naproti tomu např. ryby z rodu čtverzubců (*Tetraodon* sp.) se vyznačují nejmenším genomem ze všech obratlovců (340-385 Mbp) (Crollius, 2006). Genomy letounů a ptáků prodělaly znatelnou redukci DNA (Kapusta et al., 2017), připisovanou mj. jejich adaptaci k aktivnímu letu (Wright, Gregory, Witt, 2014).

1.2 cDNA

Komplementární DNA (cDNA) je produkována při sekvenaci transkriptomu reverzní transkriptázou, což je enzym katalyzující proces transkripce mRNA do DNA. Její přirozené využití je známé především u retrovirů pro přenos jejich genetické informace (Ryabov, 2008). U eukaryot se reverzní transkripce využívá ve formě telomerázy, která pracuje jako reverzní transkriptáza a je schopna prodlužovat konce eukaryotických chromozomů (telomery) tvorbou telomerické DNA, kterou obsahuje a používá jako templát. Reverzní transkriptázu lze vyrobit i laboratorně, např. pro potřebu exprese specifického proteinu. Díky DNA-polymeráze je možné přenést jednovláknové molekuly cDNA do dvouvláknové molekuly cDNA, které nachází využití v cDNA knihovnách (Stekel et al., 2000). To, které ze svých možných složek laboratorně získaná cDNA obsahuje, záleží na technologii její přípravy – v laboratoři nebo *in silico* – čili zda je templátem primární transkript (pre-mRNA, příprava před sestřihem, splicingem) nebo zralá mRNA (příprava po splicingu). Data cDNA dostupná z databáze Ensemblu obsahují celou sekvenci včetně kódujících a nepřekládaných sekvencí (tj. exony, introny a oba nepřekládané regiony 5´UTR a 3´UTR (Ensembl)). Podle Browna (2021) se při porovnávání cDNA s genomickou DNA vymezují pozice relevantních genů a odhalují hranice mezi introny a exony.

1.3 cds

Zkratka cds se odvozuje z anglického „coding sequence“ a označuje tu část DNA nebo RNA, která kóduje protein, je čtená jako kodony a skládá se pouze z exonů. Typicky začíná kodonem ATG v jeho funkci start kodonu a končí stop kodonem (UAA, UAG, UGA). Jde tedy o tu frakci genomu, kde se uplatňuje genetický kód a která je obecně na GC nejbohatší (Bernardi, 2005). Soubor všech exonů v genomu se nazývá exom, tj. sekvence vyskytující se ve zralé mRNA po splicingu a odstranění intronů. Když Gilbert et al. (1977) předložili hypotézu o tzv. přesouvání exonů, která se nakonec potvrdila – předpokládali, že by se exony z jednoho genu mohly kombinovat s exony genu jiného za vytvoření kódující sekvence, která by měla některé vlastnosti obou původních genů. Dnes je tato rekombinace zprostředkovaná introny známá jako hlavní doména přeskupování genů u živočichů (Cancherini et al., 2010). Pokročilejší doba umožňuje genomy nově sekvenovat, a tak máme příležitost o jejich evoluci odhalit více informací. Nižší náklady

na sekvenování kódující frakce genomu přispěly k vlně zájmu o sekvenování transkriptomu (RNA-Seq) jako alternativní a technicky celkově jednodušší a levnější strategii, která je navíc tkáňově specifická a má tak důležitou vypovídací hodnotu o funkci genomu a celé buňky (Cirulli et al., 2010). Transkriptom totiž obsahuje molekuly RNA vzniklé z DNA přepisem (transkripcí), tedy hlavně mRNA, ale mohou to být i proteiny nekódující RNA, jako např. ribozomální RNA (rRNA) a transferové RNA (tRNA). Exomové a transkriptomové sekvenování společně s SNP array (single nucleotide polymorphism) se významně uplatňují např. při detekci nádorů (např. Liu et al., 2014, Wyatt et al., 2014) nebo patologických nálezů u plodu (např. Srebniak et al., 2016). Zatímco exom zůstává u jednotlivých typů buněk neměnný, transkriptom se mění v závislosti na typu a funkci buněk. V důsledku toho není celý exom v každé buňce přeložen do bílkovin – různé typy buněk přepisují pouze části exomu a pouze kódující oblasti exonů jsou nakonec přeloženy do proteinů (Ku et al., 2014).

1.4 Odhady GC % na úrovni DNA, cDNA a cds

Na úvod do tzv. kompoziční biologie lze obecně říci, že eukaryotní organizmy s většími populacemi mají GC% vyšší než ty s populacemi malými a velikost chromozomů negativně koreluje s GC % (vlivem již zmíněného mechanizmu gBGC, současně nejrobustnější koncept vysvětlující AT/GC heterogenitu savců a ptáků a AT/GC homogenitu nižších obratlovců; Mugal et al., 2015). U tří výše popsaných úrovní genomu se GC% obsah poměrně liší. Existuje negativní vztah mezi GC% a velikostí DNA sekvence, tzn. obecně kratší geny, kratší exony, kratší introny a menší genomy jsou GC bohatší a naopak (Bernardi, 2005). Totéž se zdá, že přinejmenším do určité míry platí o TE (Symonová et Suh, 2019). Proto je celková DNA na GC nejchudší, cDNA je na GC bohatší a cds na GC nejbohatší frakcí genomu – podobný je vztah tří výše uvedených frakcí a jejich velikosti, tj. DNA → cDNA → cds (Symonová et Matoulek, nepublikovaná data). Existuje množství literatury, kde se autoři snažili použít jako tzv. proxy GC% celého genomu (tam, kde nebyl sekvenován genom) GC% na 3. pozici kodonu, tzv. GC3 (bylo snazší osekvenovat cds/transkriptom než celý genom) (Galtier et Mouchiroud, 1998). GC3 je sice důležitým faktorem, stejně jako GC1 a GC2 (tj. GC% na 1. a 2. pozici kodonu), ale od jeho užití jako proxy GC% celého genomu se upustilo, a to nejen proto, že stále roste počet osekvenovaných celých genomů (Elhaik et al., 2009).

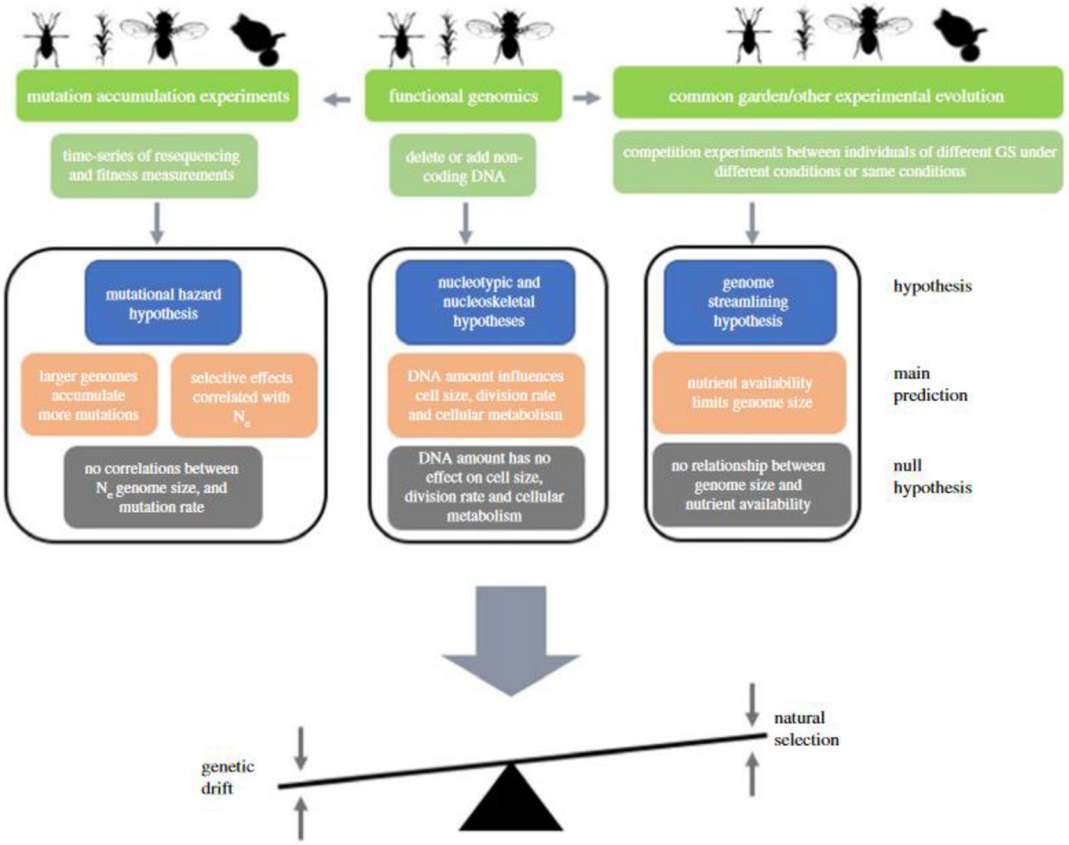
1.5 C-value paradox a C-value enigma

V kapitole 1.1 již bylo nastíněno, že i zdánlivě primitivní eukaryotní organismus může mít velký genom a naopak – znamená to tedy, že velikost genomu eukaryot nekoreluje s komplexitou organismu. Tomuto jevu se začalo říkat „C-value paradox“ („C-value“ míněno jako haploidní cytologická velikost genomu v pikogramech (pg), nebo 1C, množství DNA; Gregory, 2005). Po zjištění, že eukaryotické genomy obsahují spoustu nekódujících oblastí, se od tohoto termínu upustilo a se vznikem nových otázek se přerodil v tzv. „C-value enigma“. Analýza Elliotta a Gregoryho (2015) ukázala nejmenší průměrný počet genů kódujících protein u hub (9 953), téměř dvakrát vyšší počet genů pak u živočichů (18 943) a ještě vyšší počet genů u rostlin (35 577). V porovnání s celkovým genomem však představovaly exony u hub cca 47 %, u živočichů cca 10 % a u rostlin cca 8 %, přičemž repetitivní sekvence měly u hub zastoupení cca ze 14 %, u živočichů cca 27 % a u rostlin cca 51 %, z toho bylo 75 % identifikováno jako transpozony. Pro vysvětlení tohoto jevu bylo navrženo několik teorií o „sekundární DNA“ – teorie mutačního tlaku (zahrnující „odpadní“ a sobeckou DNA a její akumulaci v průběhu času), nukleoskeletální teorie (koevoluční interakce mezi obsahem DNA a velikostí buňky, které se postupně vyrovnávají) nebo nukleotypická teorie (kauzální souvislost mezi obsahem DNA a velikostí buňky – změny velikosti genomu přimějí buňku k přímým změnám) (Gregory, 2001). Další hypotézou o výskytu této „odpadní“ DNA, jak byla dříve repetitivní frakce hojně nazývána, je genomová rovnováha (Gene Balance Hypothesis). Podle této hypotézy by mohly transpozony vyrovnávat exprimované geny poskytnutím jakéhosi pružinového napětí mezi pericentromerickým heterochromatinem a mikrotubuly (Freeling et al., 2015), víceméně na principu stechiometrie, což by mohlo být důvodem k výskytu repetitivních částí genomů a mít celkový vliv na genovou expresi (Birchler et Veitia, 2012). Tento princip byl zatím však testován spíše jen na populárních rostlinných modelech s velkými genomy, a to rýži, kukuřici a čiroku (Freeling et al., 2015; Birchler et Veitia, 2012; Shi et al., 2020).

Review Julie Blommaert (2020) představuje ucelený přehled v současnosti diskutovaných teorií evoluce velikosti genomu – tzv. neutrální a selekční (obr. č. 1). Neutrální teorie (MHH – mutational hazard hypothesis a MEH – mutational equilibrium hypothesis) předpokládají, že velikost genomu je produktem genetického driftu a selekční tlaky v akumulaci či redukci DNA v ní hrají buď minimální, nebo žádnou roli. Selekční teorie (nukleotypická a nukleoskeletální) navrhují, že by velikost genomu mohla být

(mal)adaptivní vlastností s dopadem na fenotyp organismu – tělesné rozměry, vývin a další vlivy související s velikostí buněk.

Velikost genomu a plasticita chromozomů, jakožto dvě rozdílné proměnné, a procesy jejich regulace jsou důležitými aspekty v evoluci genomu (Slijepcevic, 2018). Mnoho dalších faktorů v této oblasti hraje roli, a ještě zdaleka není tak prozkoumaná.



*Obr. č. 1 – Přehled evolučních hypotéz o objasnění velikosti genomu, jejich predikce a nulové hypotézy s navrženými experimenty pro otestování a vypracování nového přehledu vlivu genetického driftu a přírodního výběru na velikost genomu. Jako potenciální příkladové organismy jsou znázorněny kukuřice (*Zea mays*), vířník (*Brachionus plicatilis*), potemník (*Tribolium *sp.*) a octomilka (*Drosophila melanogaster*). (převzato z Blommaert, 2020).*

1.6 Fylogeneze žijících savců

V této práci jsou dostupná genomická data rozřazena do velkých savčích skupin dle Springera et al. (2004), a to na Monotremata, Marsupialia, Afrotheria, Xenarthra, Euarchontoglires a Laurasiatheria (obr. č. 2). Veškeré poznatky jsou čerpány od Zimy a Gaislera (2007).

Skupina Monotremata neboli ptakořitní zahrnuje ptakopyska a ježury, kteří se v rámci savců vyznačují nejednou zvláštností – např. existencí kloaky a vejcorodostí.

Marsupialia (vačnatci) se původně vyskytovali na severních kontinentech, posléze se rozšířili po bývalé Gondwaně, dnes se tedy vyskytují v Jižní Americe (Ameridelphia) a Austrálii (Australidelphia), případně na Novém Zélandu či Tasmánii. Typická je pro ně přítomnost vaku, ve kterém se vyvíjí mláďata, a přítomnost epipubických kostí. V této skupině bychom také našli zástupce se schopností pasivního letu (vakoveverky), v této práci však nejsou zahrnuty.

Ostatní velké skupiny se řadí do nadřádu placentálové (Placentalia), kam patří i člověk. Hlavní charakteristiku a zároveň apomorfii představuje především alantochoriální placenta. Do řádu Afrotheria, jehož zástupci, jak už název napovídá, žijí hlavně v Africe, event. na Madagaskaru, patří např. chobotnatci (sloni, Proboscidea), sirény (Sirenia), hrabáči (Tubulidentata) a damani (Hyracoidea). Morfologicky je tato skupina dosti rozličná – např. drobní bércouni v poměru ke slonům. Jako velice zajímavým řádem bychom mohli označit sirény, které se zdržují ve vodním prostředí a lze na ně narazit i v mělkých vodách Indonésie i Jižní Ameriky, již vyhynulý koroun bezzubý se dokonce vyskytoval v chladných vodách Komandorských ostrovů.

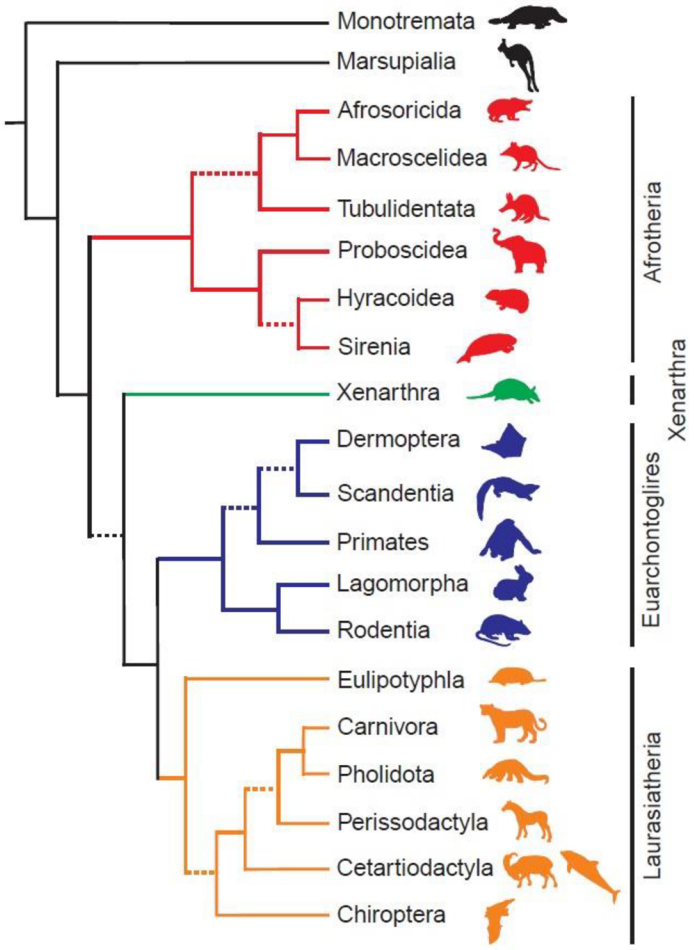
Samostatnou linii gondwanského původu pak tvoří skupina chudozubých (Xenarthra), kteří mohou mít úplně redukovaný chrup – zahrnujeme sem mravenečníkovité (Myrmecophagidae), lenochodovité tříprsté (Bradopodidae) a dvouprsté (Choloepodidae) a pásovcovité (Dasypodidae). U rodu *Dasypus* byla dokázána jako u jediného savce mimo člověka polyembryonie.

Rozsáhlá skupina Euarchontoglires pokrývá pět řádů, a to Dermoptera (letuchy), Scandentia (tany), Primates (primáti), Lagomorpha (zajícovci) a Rodentia (hlodavci). Druhově nejpočetnější a zároveň kosmopolitní řád savců přestavují hlodavci, jejichž systematika je mimořádně složitá a stále není úplně dořešená. Tany se dají považovat za jakýsi přechodný článek mezi hlodavci a poloopicemi. U letuch pozorujeme opět schopnost pasivního letu, kožní blána však už vzdáleně připomíná tu, kterou bychom nalezli u letounů. Pro primáty je typická denní aktivita, rozvoj koncového mozku, prostorové vidění a složitý sociální život. Klasicky (a zjednodušeněji) se rozdělují na poloopice (např. lemuři a outloni), opice Starého Světa (např. paviáni, kočkodani), opice Nového Světa (např. kosmani, chápani) a Hominidy (např. šimpanzi, gorily a člověk).

Poslední nemalou skupinou jsou Laurasiatheria zahrnující hmyzožravce (Eulipotyphla), šelmy (Carnivora), luskouny (Pholidota), sudokopytníky a kytovce

(Cetartiodactyla), lichokopytníky (Perissodactyla) a letouny (Chiroptera). V této nesmírně diverzifikované kategorii živočichů narazíme jak na suchozemské, tak vodní druhy, tělní rozměry v řádech několika málo centimetrů, ale i několika metrů, s délkou života sotva jeden rok, nebo třeba i dvacet let. Různé zástupce nalezneme na všech kontinentech od rovníkové zóny až po polární oblasti.

V dřívějších dobách se druhy rozřazovaly zejména na základě morfologických podobností, ale molekulární genetika je v ohledech fylogeneze naprosto zásadní součástí výzkumu a přináší stále nové výsledky, díky kterým můžeme lépe porozumět příbuznostem a rozšíření živočichů v dobách současných i minulých.



Obr. č. 2 – Fylogenetický strom savců dle Springer et al. (2004)

1.7 Letouni jako jediní létající savci

Letouni jsou nesmírně diversifikovaná skupina obratlovců a po hlodavcích druhým nejpočetnějším řádem savců (přes 1 300 druhů), jejichž rozšíření sahá do všech koutů světa kromě polárních oblastí a odlehlých ostrovů. Uzpůsobená anatomie jim umožňuje aktivní let a dle své potravní niky mají zpravidla i různě tvarovaná křídla. To však není

jedinou jejich vymožeností – známá je jejich orientace v prostoru a lov pomocí echolokace, hibernace u netropických druhů, s výjimkou kaloňů noční aktivita, rozmanitá strava nebo dlouhověkost. V současné době je velmi populární téma jejich výkonná imunita úspěšně odolávající virovým nákazám včetně koronavirů, jichž jsou hostitelé a dle aktuálních informací i původci zoonotického přenosu SARS-Cov-2 na člověka. Ve vypuknutí pandemie, která počala v jižní Číně, jež se současně vyznačuje vysokou biodiverzitou netopýřích obratlovců, může navíc sehrávat důležitou roli klimatická změna (Beyer et al., 2021). Vzhledem k mnoha svým unikátním vlastnostem jsou tak letouni např. vhodnými modelovými organismy pro výzkum v oblastech možného prodloužení života (Teeling et al., 2018), zkvalitnění imunitního systému a vyšší toleranci k chorobám (Jebb et al., 2020).

Na základě dat molekulární genetiky se v posledních letech letouni rozdělují na dva podřády: 1) Yinpterochiroptera, kam se řadí kaloňovití (Pteropodidae, Megabats), vrápencovití (Rhinolophidae, horseshoe bats), pavrápencovití (Hipposideridae, Old World leaf-nosed bats), megadermovití (Megadermatidae, false vampires), netopýrkovití (Craseonycteridae, hog-nosed bats) a víkonosovití (Rhinopomatidae, mouse-tailed bats), a 2) Yangochiroptera se všemi ostatními čeleděmi, kde bychom našli i skupinu netopýrovitých (Vespertilionidae) vyskytující se v ČR a dále např. listonosovitých (Phyllostomidae) (doplnit sem nějakou citaci).

Letouni mají s ptáky kromě schopnosti aktivního letu společných ještě několik věcí – během své evoluce totiž prodělali podstatné zmenšení genomu, což by mohlo být výhodné vzhledem ke zvýšeným nárokům na metabolismus, zejména výměně plynů a zatížení křídel (wing loading). Analýza fosilních záznamů dokonce odhalila, že ptakoještěři společně s ptačí linií theropodních dinosaurů měli ve srovnání se svými nelétavými příbuznými rovněž menší genomy (Organ et Shedlock, 2009). Kasai et al. (2013) měřili pomocí průtokové cytometrie obsah GC netopýra velkého (*Myotis myotis*). Jejich výsledek odhalil vyšší obsah GC na menších chromozomech v porovnání s chromozomy většími – tento rys sdílí letouni s ptáky, u ostatních savců nebylo nic takového dosud zaznamenáno. Na druhou stranu je třeba zdůraznit, že ptáci mají na GC výrazně bohaté mikrochromozomy (Burt, 2002), které se u letounů (a u savců obecně) nevyskytují. Naopak, u letounů často nacházíme chromozomy značných velikostí (Sotero-Caio et al., 2017). To naznačuje, že mechanizmy, které vedly ke zmenšování genomu ptáků a letounů byly zásadně odlišné (Andreas et al., *in prep*). Kapusta et al. (2017) uvádějí u netopýrů (Microbats) i u kaloňů (Megabats) nejvyšší zaznamenanou

eliminaci DNA z 10 porovnávaných savčích linií, těsně následovanou linií myší a krys. Výsledky studie Sotero-Caio et al. (2017) předkládají robertsonské translokace jako převažující způsob chromozomální evoluce letounů. Jedná se o zvláštní typ chromozomové abnormality, při níž dochází k výměně částí nehomologních chromozomů, ale při evoluci mnoha druhů se s ní setkáme poměrně často (Snustad et Simmons, 2009).

2 Materiál a metodika

2.1 Online databáze Ensembl a NCBI

Veškerá sekvenční data zpracovaná v této diplomové práci pochází ze dvou následujících veřejně dostupných online databází.

Začátek projektu Ensembl (*https://www.ensembl.org/index.html*, Flicek et al., 2021) se datuje už rokem 1999 a jeho cílem je poskytnout veřejně a zdarma přístupná genomická data nejrůznějších organismů od bakterií po živočichy včetně člověka. Sekvence DNA jsou do databáze poskytovány různými autory a projekty z celého světa. Umožňuje vyhledávat a stahovat jak jednotlivé geny, tak celé sekvence, na kartě každého druhu lze nalézt souhrnné informace (počet bp, počet kódujících i nekódujících genů, současnou verzi a datum poslední úpravy atd.) a nahlížet na části genomu, stahovat ve FASTA formátu celou sekvenci DNA či pouze cds, cDNA, ncRNA (non-coding RNA). Dále webová stránka disponuje přímo svými online nástroji, např. BLAST (Basic Local Alignment Search Tool; Altschul et al., 1990) pro porovnávání DNA a proteinů nebo Data Slicer pro vybrání a stažení určité oblasti genomu. Databázi Ensembl provozuje Evropský bioinformatický institut ve Spojeném království.

NCBI je zkratkou National Center for Biotechnology Information (*https://www.ncbi.nlm.nih.gov/*), které bylo založeno v roce 1988. NCBI je součástí Národní lékařské knihovny USA a sídlí v areálu agentury National Institutes of Health v Bethsedě v Marylandu. Funguje podobně jako Ensembl, ale je daleko obsáhlejší a mimo shromažďování genomických dat a poskytováním softwarových nástrojů pro jejich analýzu má také na starost přes 40 integrovaných databází s knihami a odbornými články pro lékaře, vědce, ale i širokou veřejnost. Vzhledem k délce existence NCBI tato databáze na rozdíl od Ensembl dlouhodobě shromažďuje i jednotlivé sekvence DNA generované metodou PCR, tedy ne pouze celé genomy. Z tohoto důvodu je NCBI výrazně rozsáhlejší. Navíc obsahuje i celogenomové sekvence početných virů a prokaryot (ve svých počátcích

zejm. medicínsky významných). Tyto typy dat byly k dispozici mnohem dříve, než začaly geometrickou řadou narůstat počty sekvenovaných genomů vyšších organismů. Po zadání příslušného hesla (název druhu) do vyhledávače NCBI projde veškeré databáze a po vyhledání genomu druhu se zobrazí karta shrnující užitečné informace, např. základní statistiku (v případě existence více než jednoho dostupného genomu daného druhu) medián GC%, medián celkové délky v Mbp, medián množství proteinů. Dostupné zde jsou i související publikace a samozřejmě je zde opět možnost stažení genomických dat ve formátu FASTA či využití online nástroj BLAST.

2.2 Úvod k jazyku Python

Python (v době psaní této diplomové práce verze 3) je v bioinformatice široce užívaným programovacím jazykem díky své snadné čitelnosti, přehlednosti a uživatelské příjemnosti i pro úplné začátečníky (*https://www.python.org/*). Pro účely analýzy genomických dat představuje vhodné webové rozhraní Jupyter Notebook (JN, *https://jupyter.org/*), ve kterém jsem prováděla veškeré potřebné úkony. Tento open-source produkt vycházející z iPython Notebooku disponuje interaktivním prostředím, v němž se zapsaný kód ihned vyhodnocuje s téměř okamžitou zpětnou vazbou. Zprovozní ho naprosto bez problémů každý v jakémkoliv operačním systému, ať už Windows, Linux či Mac.

Aktuální verzi Pythonu v JN lze zjistit jednoduchým příkazem:

```
from platform import python_version
print(python_version())
```

(V době psaní této práce verze 3.7.6)

Pro cíle této práce bylo však zapotřebí do JN dodatečně importovat knihovny pro realizaci genomových analýz, zejména BioPython (*http://biopython.org*, verze 1.78), Matplotlib (*https://matplotlib.org/*, verze 3.1.3) a Pandas (*https://pandas.pydata.org/*, verze 1.0.1). JN lze vyzkoušet i bez instalace, ale je potřeba počítat s určitými limity. Na podobném principu funguje i Google Colab (*https://colab.research.google.com*), avšak s omezenou pamětí a vyžadující účet Google. Analýza genomických dat je takovouto cestou nejvhodnější a nejjednodušší, protože pokud bychom si zkusili otevřít pouze

FASTA cds ve Wordu, zabralo by to více než 10 tisíc stran, což by bylo vzhledem k našim potřebám velice kontraproduktivní.

2.3 Sada vlastních nástrojů v Pythonu

Sadu vlastních nástrojů v jazyce Python jsem vytvořila ve spolupráci s Dominikem Matoulkem a za podpory projektu Erasmus+ KA203 řešeného na UHK. Z Ensemblu bylo potřeba stáhnout požadované sekvence, tedy DNA, cDNA a cds, přes protokol FTP (file transfer protocol), což je starší a jednoduchá technologie s nezašifrovaným přenosem dat mezi koncovými stanicemi pomocí sítě. Zabalený soubor se poté rozbalí a ideálně konvertuje do formátu *.csv (comma-separated values, hodnoty oddělené čárkami). S takto připraveným souborem dokáže dobře pracovat knihovna Pandas, která umožňuje provádět potřebné analýzy, např. základní statistické vyhodnocení, s dalšími knihovnami lze poté i vykreslovat grafy apod. (např. Matplotlib, Seaborn; https://seaborn.pydata.org/).

2.3.1 Analýza dat dostupných z databáze Ensembl

K datům příslušného druhu se lze dostat buď přes kartu daného druhu, např. *Pteropus vampyrus*: *https://www.ensembl.org/Pteropus_vampyrus/Info/Index* a poté přes odkaz „Download FASTA files“, nebo přes FTP Ensemblu: *ftp://ftp.ensembl.org/pub/release-103/fasta/* a zde najít příslušný druh (v době psaní práce *release 103*). Ensembl se několikrát ročně aktualizuje a je záhodno vědět, z jaké verze pochází zpracovaná data a současně stahovat aktuální verze (release). V jednotlivých složkách se nabízí možnosti výběru genomických DNA sekvencí ke stažení ve formátu *.fa.gz, je ale potřeba se orientovat ve všech jejich typech, jak jsou sekvence nemaskované, (hard) maskované a soft-maskované. Soubor nemaskované DNA obsahuje genomovou sekvenci včetně veškerých repetic a méně komplexních oblastí bez jakékoliv změny. Naproti tomu maskovaná DNA projde např. nástrojem RepeatMasker (*https://www.repeatmasker.org/*), který oblasti transpozonů a méně komplexních repetic (např. mikrosatelity) detekuje a následně „tvrdě zamaskuje“ písmenem „N“ – dle IUPAC (International Union of Pure and Applied Chemistry) znamená jakoukoliv dusíkatou bázi (Tab. č. 1). U soft-maskované DNA jsou veškeré zmíněné atributy pouze nahrazeny malými písmeny jednotlivých bází, zatímco unikátní sekvence, tedy nerepetitivní, jsou

ponechány velkými písmeny. Primárně tedy stahujeme soubory obsahující v názvu „toplevel“ (zahrnující veškeré sekvence) a „sm“ (soft-masked).

Tab. č. 1 – Značení sekvence DNA dle IUPAC
(převzato z: http://www.bioinformatics.org/sms/iupac.html)

IUPAC kód	Báze
A	Adenin
C	Cytosin
G	Guanin
T (U)	Tymin (Uracil)
R	A nebo G
Y	C nebo T
S	G nebo C
W	A nebo T
K	G nebo T
M	A nebo C
B	C nebo G nebo T
D	A nebo G nebo T
H	A nebo C nebo T
V	A nebo C nebo G
N	jakákoliv
. or -	mezera

Nejprve si tedy importujeme veškeré potřebné balíčky:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import gzip
from ftplib import FTP
from Bio import SeqIO
from collections import Counter
```

Modul `gzip` poskytuje jednoduché rozhraní pro kompresi a dekompresi souborů s koncovkou *.gz, které se na rozdíl od souborů typu *.rar apod. vyznačují značnou kompaktností. Pomocí modulu `ftplib` se lze napojit na úložiště serveru a poté stáhnout požadované soubory. `SeqIO` z BioPythonu představuje rozhraní Sequence Input/Output pro manipulaci se sekvencemi a umožňuje čtení, zápis a indexování se soubory ve

formátu *.fa apod. Modul `Counter` vytváří prostředí pro výpočet jednotlivých prvků ve slovníku. Přihlásíme se prvně na FTP server Ensemblu:

```
ftp = FTP('ftp.ensembl.org')
ftp.login()
```

Pak potřebujeme upřesnit serveru, ze které složky a který soubor chceme stáhnout:

```
ftp.cwd('/pub/current_fasta/myotis_lucifugus/cds')
```

Soubor stahujeme jako binární kód a zapisujeme do *.fa:

```
with open('file.fa.gz', 'wb') as file:
ftp.retrbinary('RETR
Myotis_lucifugus.Myoluc2.0.cds.all.fa.gz',file.write)
handle = gzip.open('file.fa.gz')

with open('myo_luc _cds.fa', 'wb') as out:
    for line in handle:
        out.write(line)
```

BioPython má metodu to_dict, která z iterátoru nebo seznamu vytvoří rovnou slovník. Z něj pak můžeme celou sekvenci uložit do formátu DataFrame:

```
record_dict = SeqIO.to_dict(SeqIO.parse("myo_luc_cds.fa",
"fasta"))
record_dict = {record_id: Counter(record_seq) for record_id,
record_seq in record_dict.items()}
df = pd.DataFrame.from_dict(record_dict, orient='index')
```

V rámci průběžné kontroly je záhodno se podívat, jestli se nám vše naimportovalo a zapsalo správně:

```
df.head()
```

Tab. č. 2 – Prvních pět řádků DataFrame

	A	T	G	C	N
ESMLUT00000027040.1	191	190	296	277	NaN
ESMLUT00000030196.1	258	242	290	428	NaN
ESMLUT00000001598.2	376	317	491	613	NaN
ESMLUT00000007188.2	134	126	92	137	NaN
ESMLUT00000012029.2	405	454	719	681	NaN

Místa, kde se objevuje `NaN`, nahradíme nulou, protože pandas potřebuje pracovat s numerickými daty:

```
df = df.fillna(0)
```

Pro vypočtení GC%:

```
df['cds GC%'] = (df['G'] + df['C']) / (df['A'] + df['G'] + df['C'] + df['T'] - df['N'])
```

Tab. č. 3 – Prvních pět řádků DataFrame s výpočtem GC%

	A	T	G	C	N	cds GC%
ESMLUT00000027040.1	191	190	296	277.0	0.0	0.60
ESMLUT00000030196.1	258	242	290	428.0	0.0	0.58
ESMLUT00000001598.2	376	317	491	613.0	0.0	0.61
ESMLUT00000007188.2	134	126	92	137.0	0.0	0.46
ESMLUT00000012029.2	405	454	719	681.0	0.0	0.61

Pro základní statistické vyhodnocení:

```
df.describe()
```

Tab. č. 4 – Základní statistika GC% DataFrame

	A	G	C	T	N	cds GC%
count	44934	44934	44934	44934	44934	44934
mean	514.31	521.43	511.78	427.63	0.07	0.52
std	615.83	529.79	512.06	477.03	5.78	0.1
min	9	4	7	7	0	0.25
25%	205	226	225	186	0	0.46
50%	354	388	380	311	0	0.52
75%	637	644	632.75	516	0	0.59
max	33038	24087	21760	24051	1000	12

S takto připraveným DataFrame si již můžeme zobrazit histogram (viz Graf č. 1 v kap. Přílohy):

```
df['cds GC percentage'].plot.hist(bins=200)
plt.xlim([0, 1])
plt.xlabel("cds GC%")
plt.title("Myotis lucifugus", style="italic")
```

Od dusíkatých bází odečítáme ještě „N“, které značí neznámé báze. U některých druhů se jich vyskytuje méně, u některých více, záleží na kvalitě dostupných dat, která se mohou v těchto poměrech i výrazně lišit. Stejný postup uplatníme i při psaní kódu pro cDNA (i celkové DNA). Pro lepší představu a praktičnost lze proložit hodnoty cds i cDNA do jednoho grafu, je však potřeba importovat další knihovnu – NumPy (https://numpy.org/, verze 1.18.1). Tato knihovna disponuje funkcemi pro práci s vícerozměrnými poli, maticemi, vektory apod. Kód by mohl vypadat následovně (výstup viz Graf č. 2 v kap. Přílohy):

```
import numpy as np

x = df1['cds GC%].plot.hist(bins=200, alpha=0.5, color="red")
y = df2['cdna GC%'].plot.hist(bins=200, alpha=0.5, color="green")
bins = np.linspace(0, 1)
plt.xlim([0, 1])
plt.legend(loc='upper right')
plt.ylabel('Frequency')
plt.xlabel('GC %')
```

```
plt.title("Myotis lucifugus", style="italic")
plt.show()
```

Pokud se nám mají v grafu nějaká data prolínat, je vhodné nastavit větší průhlednost (alpha) a především barvy – ideálně barvy ne vedle sebe stojící v barevném spektru, v našem případě červená a zelená.

Uvedený postup je vhodný pro nárazovou analýzu jednoho druhu, pro větší objem dat využijeme nástroj dna_puller, který je ke stažení se všemi potřebnými proprietami z veřejného úložiště kódů Bioinformatického centra HK zde: *https://github.com/bioinfohk/dna_puller*. Nutno říci, že tento nástroj nemusí fungovat v JN na OS Windows, pro účely této diplomové práce tak probíhala většina analýz na OS Linux Ubuntu 20.04 LTS. Před spuštěním kódu je nutné si vedle složky s nástrojem dna_puller také vytvořit ještě složku „jsons“. Nástroj funguje tím způsobem, že po zadání požadovaných druhů organismů a typů sekvencí se stáhnou zabalené soubory FASTA jeden po druhém a ukládají do slovníku ve formátu *.json. JSON je zkratkou pro „JavaScript Object Notation“ a reprezentuje univerzální způsob zápisu dat, se kterým je následně možné pracovat v jakémkoli programovacím jazyku. Při stahování dalšího balíku se souborem FASTA se předchozí balík smaže, aby zbytečně nezabíral místo na disku, jelikož nabývají velikostí i několik Gb.

Např. potřebujeme analyzovat zástupce nadřádu Afrotheria a jejich cds a cDNA:

```
import dna_puller.dna_puller as puller

species = ['Loxodonta_africana', "Procavia_capensis",
"Echinops_telfairi"]
puller = puller.DnaPuller(species, True, True, ['cds', 'cdna'])
puller.download_and_parse_data()
```

Takto máme připraveny soubory pro hromadnou analýzu. Všechna data potřebujeme sloučit, abychom přes ně mohli iterovat (opakovat proces v měnícím se kontextu), v našem případě GC% u cds a cDNA:

```
import os, json

species_files = os.listdir('jsons')
```

```
dna_types = ['cdna', 'cds']

files = []

for species_file in species_files:
    if species_file[-5:] == '.json':
        files.append(species_file)

species_data = {}

for species_file in files:
    with open('jsons/' + species_file) as file:
        species_name = species_file[0:-5]
        species_data[species_name] = {}
        for type in dna_types:
            species_data[species_name][type] = {}
        data = json.load(file)
        for type in dna_types:
            for gene_key, gene_data in data[type].items():
          species_data[species_name][type][gene_key] =
          gene_data

aggregated_data = {}
for type in dna_types:
    aggregated_data[type] = {}
    for species_key, datas in species_data.items():
        aggregated_data[type][species_key] = []
        for key, data in datas[type].items():
            all_count = float(data['all']) - float(data['N'])
      aggregated_data[type][species_key].append(float(data['G']
      + data['C']) / all_count)
```

Nyní už jen potřebujeme umístit vše do jednoho souboru. Rozměry a rozmístění grafů upravíme dle obsahu a potřeb a uložíme ve formátu *.png:

```
import matplotlib.pyplot as plt

fig, axs = plt.subplots(2, 3, figsize=(30,50))

for index in range(0, len(aggregated_data['cdna'])):
```

```
        species_key = list(aggregated_data['cdna'].keys())[index]
        y = index % 3
        x = int(index / 3)

        axs[x, y].set_title(species_key[0].upper() +
        species_key[1:], style='italic')
        axs[x, y].set_ylim(0, 500)
        axs[x, y].set_xlim(0, 1)

        cdna_data = aggregated_data['cdna'][species_key]
        cds_data = aggregated_data['cds'][species_key]

    axs[x, y].hist(cds_data, bins=200, range=(0.2, 0.8),
    color='red', alpha=0.5)
        axs[x, y].hist(cdna_data, bins=200, range=(0.2, 0.8),
        color='green', alpha=0.5)

      plt.savefig('Afrotheria_cdna_cds.png')
```

Následuje tvorba vlastního datasetu, kterou provádíme opět v JN s daty, které máme ve formátu *.json v připraveném adresáři. Nejefektivnější metodou je iterace nad soubory, tzn. cyklus, smyčka, která se neustále opakuje, dokud nedojde na konec slovníku, souboru či rovnou celé složky souborů. Do datasetu potřebujeme hodnoty pro GC% celkové DNA, cDNA a cds, počet readů (genů), počet neznámých bází (N), počet bází A, T, G, C a jejich procento z celkových genomových assemblies.

Nejprve je potřeba si načíst veškeré soubory ve složce, se kterými budeme pracovat, takže např. opět zástupce skupiny Afrotheria:

```
import os
files = os.listdir('jsons/Afrotheria')
jsons_files = []
for file in files:
    if file[-5:] == '.json':
        jsons_files.append(file)
```

Pro kontrolu, jestli se nám načetly soubory správně provedeme:

```
jsons_files
```

Nyní si spočteme neznámé báze (N) pro cDNA i cds jednotlivých druhů, k čemuž potřebujeme tři cykly. První na vytvoření slovníku se všemi soubory a jejich načtení, druhý pro typy sekvencí (cDNA, cds) a ve třetím se postupně přičítají požadované báze, iteruje se v podstatě po klíči a hodnotě jako v klasickém slovníku:

```
import json
data = {}
types = ['cds', 'cdna']

for file in jsons_files:
    species_name = file[:-5]
    data[species_name] = {}
    with open('jsons/Afrotheria/' + file) as input_file:
        data[species_name] = json.load(input_file)
    for typ in types:
        print(species_name)
        counter = 0
        for key, values in data[species_name][typ].items():
            counter += values['N']
        print(typ, counter)
```

Ve výstupu se načtou postupně všechny druhy a jim odpovídající počty N. Obdobně můžeme sečíst i známé báze, pouze zaměníme `values['N']` za `values['A'] + values['T'] + values['G'] +values['C']`.

Pro zjištění počtu readů, tedy sekvencí v jednotlivých souborech použijeme funkci len, která umí spočíst počet položek v daném seznamu hodnot:

```
for file in jsons_files:
    species_name = file[:-5]
    data[species_name] = {}
    with open('jsons/Laurasiatheria/' + file) as input_file:
        data[species_name] = json.load(input_file)
        print([species_name])
    for key, values in data[species_name].items():
        print(key, len(values))
```

Pro výpočet GC% přidáme několik proměnných, které uložíme do seznamu a počet G + C vydělíme všemi hodnotami bez neznámých sekvencí a vynásobíme stem:

```
gc_val = []
all_val = []

for file in jsons_files:
    species_name = file[:-5]
    data[species_name] = {}
    with open('jsons/Euarchontoglires/' + file) as input_file:
        data[species_name] = json.load(input_file)
        print(species_name)
    for typ in types:
        for key, values in data[species_name][typ].items():
            gc = values['G'] + values['C']
            total = values['all'] - values['N']
            perc = gc/total*100
        print(typ, perc)
```

Pro tvorbu korelačních diagramů (scatter plot) délky cds versus jejich GC% je třeba předešlý kód patřičně upravit, ze souborů *.json budeme potřebovat pouze cds:

```
import os, json
      species_files = os.listdir('jsons/Afrotheria')
      dna_types = ['cds']
      files = []
      for species_file in species_files:
    if species_file[-5:] == '.json':
      files.append(species_file)
```

Další buňka bude shodná s kódem pro tvorbu histogramů:

```
species_data = {}

for species_file in files:
    with open('jsons/Afrotheria/' + species_file) as file:
        species_name = species_file[0:-5]
        species_data[species_name] = {}
        for type in dna_types:
            species_data[species_name][type] = {}
        data = json.load(file)
```

```
            for type in dna_types:
                for gene_key, gene_data in data[type].items():
              species_data[species_name][type][gene_key] =
               gene_data
```

Nyní je potřeba udělat nový cyklus, kde si vytvoříme seznamy ve slovníku, a to jeden pro G + C včetně „S“ (G nebo C) a jeden pro celkový počet bází, opět bez neznámých bází (N), plus výpočet pro GC%:

```
aggregated_data = {}
for type in dna_types:
    aggregated_data[type] = {}
    for species_key, datas in species_data.items():
         aggregated_data[type][species_key] = {'gc': [], 'all':
         []}
        for key, data in datas[type].items():
             gc = float(data['G']) + float(data['C']) +
             float(data['S'])
             all_count = float(data['all']) - float(data['N'])
             gc_percent = (float(gc) / (all_count)*100)
aggregated_data[type][species_key]['gc'].append
(gc_percent)
aggregated_data[type][species_key]['all'].append
(all_count)
```

Při vizualizaci GC% je třeba nastavit vyšší limit frekvence pro osu *y* a průhlednost vlastního grafu (alpha) tentokrát na nejvyšší pro lepší výpovědní hodnotu. Osy ještě patřičně pojmenujeme a opět uložíme:

```
import matplotlib.pyplot as plt

plt.rcParams.update({'font.size': 14})

fig, axs = plt.subplots(2, 2, figsize=(20,20))

for index in range(0, len(aggregated_data['cds'])):
    species_key = list(aggregated_data['cds'].keys())[index]
    y = index % 2
    x = int(index / 2)
```

```
        axs[x, y].set_title(species_key[0].upper() +
        species_key[1:], style="italic")
        axs[x, y].set_ylim(0, 15_000)
        axs[x, y].set_xlim(0, 100)
        axs[x, y].set_xlabel("GC%")
        axs[x, y].set_ylabel("Cds genes length")

        cds_data = aggregated_data['cds'][species_key]

axs[x, y].scatter(cds_data['gc'], cds_data['all'], color='red',
alpha=0.1)

  plt.rcParams['figure.constrained_layout.use'] = True
  plt.savefig('Afrotheria_scatters_cds.png')
```

2.3.2 Data z NCBI a Animal Genome Size Database

Pro potřebu práce s celogenomovými hodnotami GC% a velikostmi genomu je vhodné využít databázi NCBI. K tomu je možné potřebné taxony vyfiltrovat v odkazu *https://www.ncbi.nlm.nih.gov/genome/browse#!/eukaryotes/* a následně exportovat do formátu *.csv. Tato data lze poté otevřít a editovat v MS Excelu. V případě této diplomové práce tedy bylo třeba zaškrtnout „Mammals“ jako podskupinu (subgroup) a u „Choose Columns“ vybrat potřebná data, a to název organismu, velikost v Mbp, GC% a počet cds a poté už jen stáhnout a následně po manuální kontrole ručně odstranit nadbytečná
a nekompletní data. Tyto GC% hodnoty jsou mediánem všech sekvenovaných genomů, proto je třeba ještě ručně upravit hodnoty dle nejnovější assembly a vyšší kvality sekvenace. Pro kontrolu kompletnosti assembly je nezbytné využít etablovaný online nástroj Animal Genome Size Database (www.*genomesize.com*). Ten umožňuje dohledat k požadovaným živočichům C-value, tzn. haploidní velikost genomu v pikogramech (pg; viz kap. 1.5), kterou si v případě potřeby můžeme převést na počet párů bází (bp) dle vzorce od Doležela et. al (2003):

Hodnota v pg * 0,978 * 10^9, přičemž 1 pg = 978 Mbp

2.4 Statistické testování výsledků

Výsledky GC% na úrovních cDNA a cds mezi letouny a ostatními savci získané výše popsanou metodikou byly podrobeny statistickému testování jejich významnosti v programu R (R Core Team, 2014). Vzhledem k nízkému početnímu zastoupení letounů (3 druhy) vůči 89 druhům ostatních savců, bylo nutné zvolit metodu intervalového odhadu středních hodnot GC%. Tyto intervalové hodnoty byly nejprve vypočteny pro všechny ostatní savce dohromady pro každou úroveň DNA (tj. cDNA a cds) na 90% hladině pravděpodobnosti pomocí jednovýběrového t-testu. Následně bylo porovnáno, zda hodnoty GC% vypočtené pro letouny do tohoto intervalu spadají, či nikoliv.

3 Výsledky

3.1 Základní statistické vyhodnocení GC% letounů v rámci savců

I s ohledem na nízké početní zastoupení letounů (3 druhy ve srovnání s 89 druhy ostatních savců) v našem datasetu mají letouni vyšší průměrné GC% na úrovni cDNA (51,16 %) v porovnání s hodnotou 47,6 % u ostatních savců (viz Tab. č. 5). Stejně tak medián hodnot GC% na úrovni cDNA je u letounů 49,47 % a tudíž vyšší než u ostatních savců, kde činí 46,35 %. Podobně i na úrovni cds mají letouni výrazně vyšší hodnoty průměru GC% (57,37 %) než ostatní savci (50,76 %) a stejně tak i medián (57,26 % letouni, 49,18 % ostatní savci). Vyšší rozptyl (Tab. č. 5) hodnot GC% na obou úrovních u ostatních savců je dán výrazně vyšším počtem zastoupených druhů v porovnání s letouny. Detailní testování GC% hodnot jednovýběrovým t-testem odhalilo, že dva druhy letounů (*M. myotis* a *R. ferrumequinum*) mají hodnoty GC% výrazně vyšší než je intervalový odhad střední hodnoty GC% ostatních savců. GC% třetího druhu letounů (*P. vampyrus*) se pohybuje těsně na horní hranici intervalového odhadu středních hodnot ostatních savců.

Naopak proporce cDNA a cds vůči celkové DNA jsou u letounů a ostatních savců srovnatelné (viz Tab. č. 5).

Tab. č. 5 – Základní statistické vyhodnocení GC% na úrovni cDNA a cds s důrazem na letouny

	GC% cDNA (ostatní savci)	GC% cDNA (letouni)	GC% cds (ostatní savci)	GC% cds (letouni)	% cDNA z DNA (ostatní savci)	% cDNA z DNA (letouni)	% cds z DNA (ostatní savci)	% cds z DNA (letouni)
count	89	3	89	3	89	3	89	3
mean	**47.6**	**51.16**	**50.76**	**57.73**	**3**	**2.45**	**2.16**	**2.01**
std	7.38	5.56	9.10	5.5	1.57	1.44	0.83	0.8
min	31.79	46.64	32.41	52.48	0.85	1.53	0.06	1.49
25%	42.09	48.05	43.57	54.87	1.98	1.62	1.61	1.55
50%	46.35	49.47	49.18	57.26	2.92	1.71	2.31	1.62
75%	52.8	53.42	57.73	60.36	3.56	2.91	2.66	2.28
max	**71.43**	**57.37**	**73.82**	**63.46**	**11.71**	**4.11**	**4.5**	**2.94**

Tab. č. 6 – Základní statistické vyhodnocení GC% na úrovni cDNA a cds u hlavních skupin savců

	GC% cDNA Mono-tremata	GC% cDNA Marsu-pialia	GC% cDNA Afrotheria	GC% cDNA Xenarthra	GC% cDNA Euarchon-toglires*	GC% cDNA Laurasia-theria
count	1	5	3	2	43	36
mean	**52.98**	**42.58**	**55.53**	**45.71**	**46.99**	**48.56**
std	NaN	7.52	6.84	1.64	7.78	6.83
min	52.98	31.79	49.17	44.55	33.33	37.75
25%	52.98	40.42	51.91	45.13	41.42	42.94
50%	**52.98**	**42.81**	**54.65**	**45.71**	**44.44**	**48.7**
75%	52.98	45.45	58.71	46.29	52.62	52.84
max	52.98	52.44	62.77	46.88	67.3	71.43
	GC% cds Mono-tremata	**GC% cds Marsu-pialia**	**GC% cds Afrotheria**	**GC% cds Xenarthra**	**GC% cds Euarchon-toglires***	**GC% cds Laurasia-theria**
count	1	5	3	2	43	36
mean	**61.51**	**47.71**	**40.91**	**47.6**	**52.1**	**50.64**
std	NaN	4.1	2.24	8.01	9.43	9.23
min	61.51	44.5	38.77	41.93	36.72	32.41
25%	61.51	44.77	39.4	44.76	45.17	43.06
50%	**61.51**	**46.11**	**40.72**	**47.6**	**50**	**50.61**
75%	61.51	48.66	41.98	50.43	59.04	57.92
max	61.51	54.52	43.25	53.27	73.82	67.71

Pro získání představy, jaký obsah GC existuje u hlavních fylogenetických linií savců bylo GC% vypočítáno na úrovních cDNA a cds u šesti hlavních skupin savců (viz Tab. č. 6). Na úrovni cDNA se průměrné hodnoty GC% pohybují v rozmezí 42,58 % u skupiny Marsupialia a 55,53 % u skupiny Afrotheria. Hodnoty mediánu GC% na úrovni cDNA jsou u těchto skupin srovnatelné (42,81 % u Marsupialia a 54,65 % u Afrotheria). Na úrovni cds se průměrné hodnoty GC% pohybují v rozmezí 40,91 % u skupiny Afrotheria a 61,51 % u skupiny Monotremata. Hodnoty mediánu GC% na úrovni cds jsou u těchto skupin srovnatelné (61,51 % u Afrotheria a 40,72 % u Monotremata). Nižší GC% na úrovni cds než na úrovni cDNA u skupiny Afrotheria je poněkud překvapivý výsledek a bude mu věnována pozornost v další fázi výzkumu.

3.2 Histogramy GC% pro sekvence cds a cDNA

Celkem bylo pomocí Pythonu zanalyzováno 92 druhů savců dostupných z databáze Ensembl v březnu roku 2021 (verze 103). Tyto druhy byly pro lepší přehlednost a s ohledem na své fylogenetické vztahy rozděleny podle hlavních savčích skupin na Monotremata, Marsupialia, Afrotheria, Xenarthra, Euarchontoglires a Laurasiatheria (dle Springer et al., 2004). Histogramy hodnot GC% pro jednotlivé sekvence na úrovních cds a cDNA jsou na *Grafech 3-7*, kde cds je zobrazena červenou barvou a cDNA barvou zelenou.

Jediný zástupce bazální skupiny ptakořitných (Monotremata), ptakopysk podivný (*Ornitorhynchus anatinus*), je na *Grafu 3* společně s vačnatci (Marsupialia). U vačnatců můžeme vidět téměř symetrické unimodální rozdělení s mírným zešikmením z pravé strany a frekvence výskytu dosahující 600-700 téměř 50% obsahu GC. Výjimkou v rámci těchto druhů je klokan dama (*Notamacropus eugenii*), který je využíván jako modelový organismus při výzkumu vačnatců a savců obecně, u něhož hodnoty frekvence nepřesahují 300 ca 45 – 55 % hodnot GC. Frekvence hodnot GC% je u ptakopyska rozdělena rovnoměrněji, GC% u něj zato dosahuje vyšších hodnot (ca 45 – 65 %).

Ze skupiny Afrotheria (*Graf 4*) jsou v databázi v době psaní této práce zastoupeny pouze tři druhy. Ty vykazují jednak unimodální symetrické rozdělení (*Procavia capensis* a *Echinops telfairi*) až náznak bimodálního rozdělení hodnot GC% (*Loxodonta africana*).

Oba dva zástupci chudozubých (Xenarthra, *Graf 5*) mají histogramy zprava výrazně zešikmené.

Zde nejobsáhleji zastoupená skupina Euarchontoglires čítá 40 druhů a zahrnuje primáty, tany, hlodavce a zajícovce bez zastoupení letuch. Z technických důvodů byli člověk a myš z *Grafu 6* a *7* pro svou výrazně vyšší kvalitu zpracování genomu vyřazeni a analyzováni a vizualizováni zvlášť (*Graf 9*), jelikož by snižovali rozlišení histogramů ostatních druhů. Pro lepší přehlednost byla skupina Euarchontoglires rozdělena na dvě obrazové tabule, kde jsou zvlášť primáti a jim příbuzné tany (*Graf 6*) a hlodavci se zajícovci (*Graf 7*). Histogramy většiny primátů se vyznačují náznaky bimodálního (dvouvrcholového) rozdělení, které lze vidět i u králíka divokého (*Oryctolagus cuniculus*). U tany (*Tupaia belangeri*), komby Garnettovy (*Otolemur garnettii*), kočkodana zeleného (*Chlorocebus sabaeus*), orangutana sumaterského (*Pongo abelii*) a sifaky Coquerelovy (*Propithecus coquereli*) nepřesahuje frekvence GC% hodnotu 400 a je podobná spíše hlodavcům, ačkoliv u tany je opět náznak bimodality a zároveň má nejnižší GC% frekvenci.

Pro sesterskou skupinu Laurasiatheria (*Graf 8*) je vyobrazeno 35 druhů a spadají do ní letouni, hmyzožravci, kytovci, sudokopytníci a lichokopytníci, šelmy a luskouni. U většiny druhů je alespoň z části opět patrná bimodalita jejich histogramů. Výraznou výjimku tvoří domestikovaná lama alpaka (*Vicugna pacos*), jejíž histogram vykazuje jediný vrchol a zprava je mírně zešikmený. Pouze nevýrazný náznak bimodality je vidět u ježka západního (*Erinaceus europaeus*) a rejska obecného (*Sorex araneus*), přičemž exom těchto druhů patří k nejmenším (viz příloha *Tab. S1*). Naopak nejvýraznější bimodalita je patrná u obou druhů medvěda, obou turů a prasete. Z letounů projevuje bimodalitu vrápenec velký (*Rhinolophus ferrumequinum*), jenž má ze tří zkoumaných druhů letounů největší exom a současně je na rozdíl od ostatních dvou jediným zástupcem podřádu Yinpterochiroptera.

3.3 Tabulka s výpočty a grafy

Tabulka S1 předkládá výpočty GC% pro jednotlivé frakce savčích genomů, jejich délky (počet genů), počet bází A, T, G a C a počet nespecifikovaných dusíkatých bází (N) z dat z databáze Ensembl. Z těchto dat byly následně vytvořeny grafy v programu Microsoft Excel (viz níže).

Jelikož jsou soubory genomů z databáze Ensembl z různých zdrojů, lze předpokládat i odlišné kvality osekvenování, o čemž vypovídají především počty neznámých bází (N). Tyto počty mnohdy dosahují i milionových hodnot a je třeba je ve

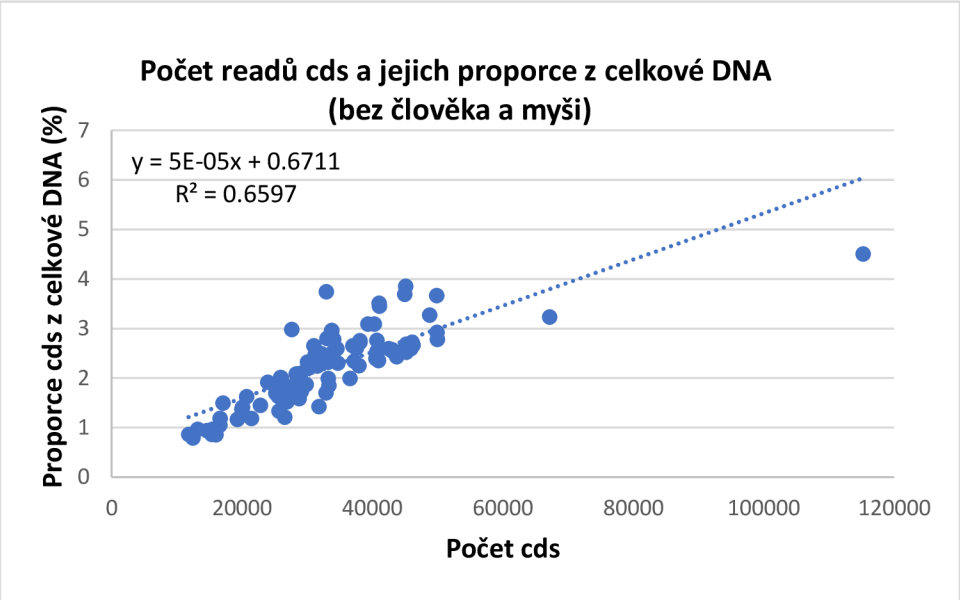
výsledcích zohledňovat, protože mohou vysvětlit např. nebývale vysoké hodnoty GC% u některých druhů. Vzhledem k těmto rozdílům nebylo GC% na úrovni celého genomu věnována velká pozornost a toto téma tak bude řešeno v budoucnu.

Do zobrazení dat nebyli opět ze stejných důvodů jako pro histogramy zařazeni člověk a myš. Dle předpokladu je na GC% bohatá především kódující část genomu (cds), kde se průměrné hodnoty pohybují těsně pod 50 % (viz *Graf 17*), u cDNA je to pak zhruba 44 % s odlehlou hodnotou náležící ježkovi evropskému (*Erinaceus europaeus*). Průměrné GC% cDNA se pohybuje okolo 40 %, přičemž pět druhů vykazuje odlehlé hodnoty. Jedná se o netopýra hnědavého (*Myotis lucifugus*, 58,09 %), langura Bietova (*Rhinopithecus bieti*, 60 %), kaloně malajského (*Pteropus vampyrus*, 64.56 %), *Dipodomys ordii* (66 %) a tygra ussurijského (*Panthera tigris altaica*, 67,5 %). Po manuální úpravě hodnot GC% genomu podle již zveřejněných hodnot v databázi NCBI se mezikvartilové rozpětí hodnot výrazně zúžilo, což ukazuje velmi úzké rozpětí těchto hodnot u savců a potenciální selekci ve prospěch konkrétních hodnot (*Graf 18*). Dále byly porovnány délky cds a cDNA vůči GC% (*Graf 19 a 20*), GC% cds a cDNA vůči počtu neznámých bází (Graf 21 a 22) a GC% cds a cDNA vůči celé sekvenci cds a cDNA (ATGC).

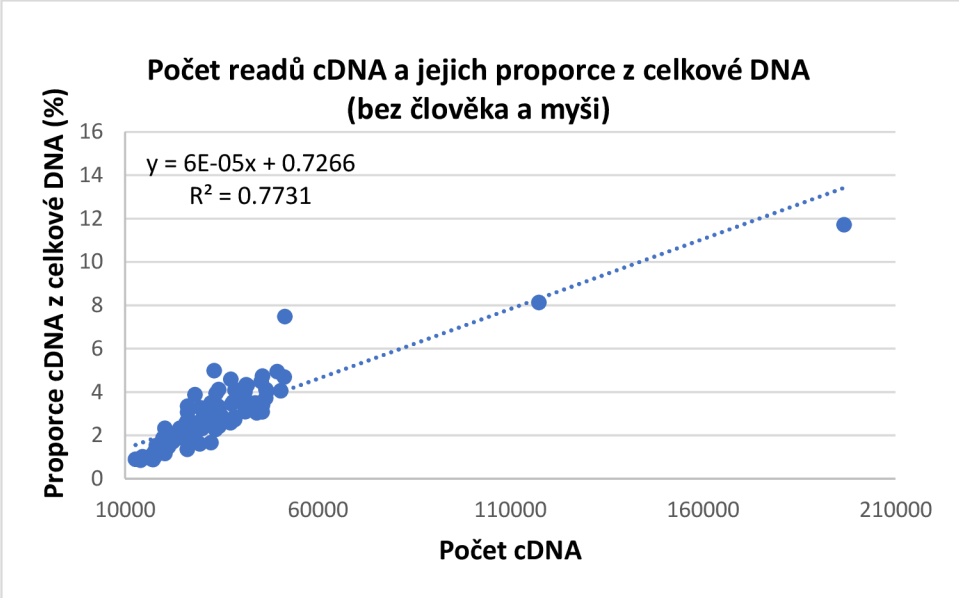
3.4 Kvantitativní pohled na získané výsledky

Korelační diagram (*scatter plot)* slouží k porovnání dvou proměnných. Ve své práci jsem využila v jazyce Python získané výsledky pro následující typy porovnání v rámci všech savců hromadně:

1) Počet všech sekvencí cDNA a cds (tedy jednotlivých oblastí DNA kódujících protein) a jejich souhrnné proporce z celého genomu (Grafy 25 a 26, z obou grafů byl vynechán člověk a myš kvůli jejich diametrálně odlišné kvalitě genomových assembly); Zde existuje těsný pozitivní lineární vztah na obou úrovních genomu.

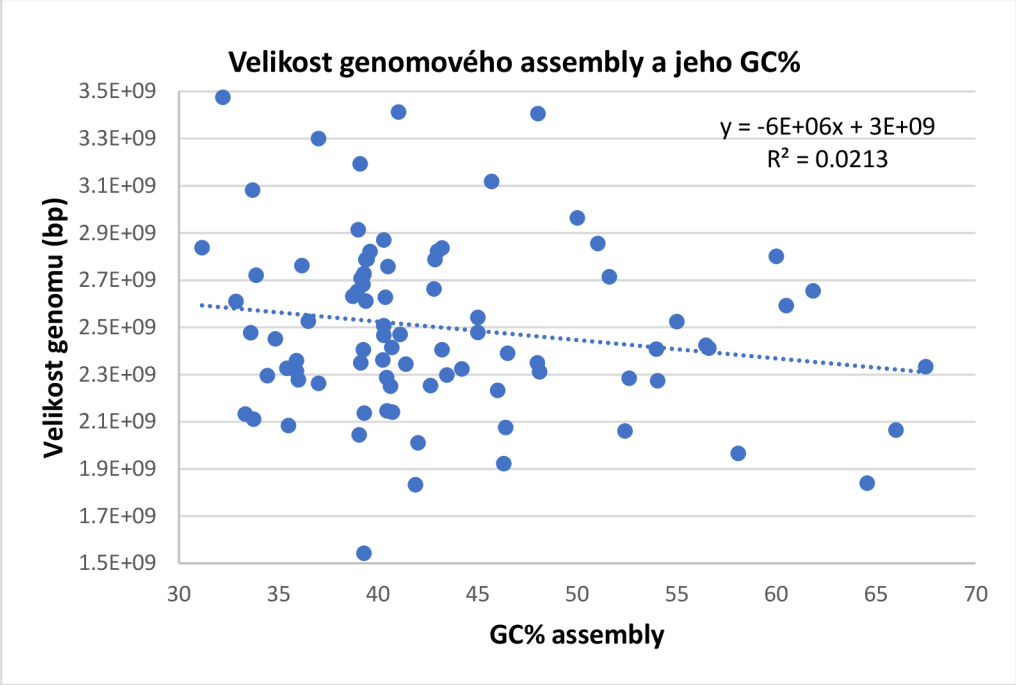


Graf č. 25 – Souhrnné proporce genomů cds z celkové DNA



Graf č. 26 – Souhrnné proporce genomů cDNA z celkové DNA

2) Vztah mezi velikostí genomového assembly a jejich GC% u všech savců. Se současnou kvalitou dat a při zahrnutí všech současně dostupných druhů savců nelze vztah mezi těmito veličinami vyjádřit lineárním modelem (viz Graf č. 27).



Graf č. 27 - Vztah mezi velikostí genomových assembly a GC%

3) Porovnání délky readů cds a jejich GC% pro každý druh savců

Společným rysem všech těchto grafů je nahloučení většiny bodů v dolní části s postupným rozptylováním směrem nahoru, kde je ručně nastavený maximální horní limit pro osu *y* na počet 15 000 bází z důvodu lépe vypovídající hodnoty (Graf 10 – 16). Pro člověka a myš opět platí jiný limit, a to 20 000 bází, přičemž některé geny mají nejspíše téměř nulovou hodnotu GC%. Nejkratší exony v porovnání s GC% má klokan dama (*Notamacropus eugenii*), tana severní (*Tupaia belangeri*), lenochod krátkokrký (*Choloepus hoffmanni*), daman skalní (*Procavia capensis*), ježek evropský (*Erinaceus europaeus*) a delfín skákavý (*Tursiops truncatus*). Rejsek má sice jeden z nejkratších exomů, ale dle dat z Ensemblu je velmi bohatý na GC% (60,81 %).

4 Diskuse

Výpočet GC% u FASTA souborů a tvorba grafů GC% z *celého genomu* savců (tedy ne pouze na úrovních cDNA a cds) savců již představuje výpočetně a hlavně technicky náročnější proceduru, která vyžaduje práci na výkonném výpočetním klastru (např. Linuxovém serveru nebo virtuálním serveru). Tato etapa je předmětem našeho dalšího směřování v rámci Bioinformatického centra HK, kde bude řešena. Z těchto důvodů tato úroveň genomu není v této diplomové práci zahrnuta, protože bych výpočty nemohla provádět na svém laptopu.

4.1 Výsledky v kontextu současného stavu znalostí o kompoziční evoluci obratlovců

Vyšší GC% u cDNA i cds letounů je v porovnání s ostatními savci zajímavým výsledkem, protože nárůst GC% byl připisován spíše na transposonům, které jsou u letounů vzhledem k savcům nebývale aktivní, a tudíž i vysoce zastoupeny (Andreas et al., *in prep*.). Navíc skutečnost, že k horní hranici GC% cDNA a cds ostatních savců se přibližuje druh *P. vampyrus*, zatímco další dostupný zástupce skupiny Yinpterochiroptera (*R. ferrumequinum*) je na GC ještě bohatší, podobně jako zástupce Yangochiroptera (*M. myotis*), je opět v rozporu s dosud robustními výsledky o GC% transpozonů signifikantně rozdílnými právě mezi těmito dvěma hlavními liniemi letounů (Andreas et al., *in prep*.). To vše naznačuje ještě komplexnější situaci v GC evoluci genomu letounů, než předpokládá předchozí studie.

Nižší GC% na úrovni cds než na úrovni cDNA u skupiny Afrotheria je poněkud překvapivý výsledek, protože cds se alespoň v rámci obratlovců považuje za frakci genomu, která je na GC nejbohatší, a to i u ryb a obojživelníků (např. Symonová, 2021; Matoulek et Symonová, *pers. comm.*). Tento výsledek tak bude podroben dalšímu zkoumání v následné fázi výzkumu kompoziční evoluce genomu savců.

Zde zdokumentovaná rozmanitost tvarů histogramů mezi savci je zajímavým výsledkem přinejmenším z následujících důvodů:

1) Výskyt jednotlivých tvarů histogramů se jeví jako alespoň do určité míry fylogeneticky podmíněný, protože bimodalita se vyskytuje zejména u skupin Euarchontoglires a Laurasiatheria, zatímco zešikmený histogram u vačnatců (Marsupialia) a chudozubých (Xenarthra). Skupina Afrotheria zatím nemůže být dostatečně vyhodnocena vzhledem k nízkému počtu druhů a jejich rozdílných tvarů histogramů.

2) Zjištěná bimodalita se nevyskytuje u jiných skupin obratlovců dosud analyzovaných jinými autory (Matoulek et Symonová, *in prep.*) a naznačuje roli GC% na úrovni transkripce. Nabízí se tudíž budoucí analýza GC% transkriptomu v kontextu genové ontologie, jak prozatím pouze u 12 druhů savců provedli autoři studie Berná et al., (2012), avšak bez dalšího pokračování, jakmile byla dostupná data dalších druhů.

4.2 Kvalita genomových assembly

Kvalita dostupných genomů je zásadní faktor, od kterého se odvíjí i kvalita všech výstupů a výsledků. Kvalita genomů se liší podle jejich úrovně assembly, tedy podle sestavení genomu (protože se i v česky psané literatuře vžil anglický termín *assembly*, používám ho rovněž v této práci): úroveň *contig* (česky kontig, souvislý úsek DNA bez mezer sestavený z překrývajících se čtení, tzv. *readů*, získaných při sekvenování), úroveň *scaffold* (sady uspořádaných kontigů sestavených dle překrývaných sekvencí, delší než kontigů, ale kratší než celé chromozomy) a úroveň *chromosome* (téměř úplná sekvence DNA sestavená z menších sekvencí představující chromozom avšak s možným výskytem mezer). Dalšími faktory ovlivňujícími kvalitu genomů jsou využité sekvenační technologie, respektive generace sekvenování. V současné době poskytují nejvyšší kvalitu hybridní sekvenování, která kombinují méně chybové short-read sekvenování s chybovějším long-read sekvenováním a navzájem tak kompenzují nedostatky obou přístupů. Data, která máme nyní k dispozici, jsou však stále převážně založená na short-read sekvenování, ačkoliv genomy sekvenované hybridně přibývají; kvůli finanční náročnosti prozatím pomalu, ačkoliv co se týče lidského genomu, náklady postupně klesají (Adewale, 2020). Během zpracování dat v rámci této práce vyvstala otázka jejich rozdílné kvality a také zda by zde provedených analýz bylo možné paralelně využít ke zhodnocení kvality těchto dat. Jak již bylo naznačeno v kapitole Výsledky, jen několik organismů má nulový počet neznámých bází (N). Určitý standard kvality vstupních dat

zaručuje skutečnost, že data pro tuto práci jsou výhradně z databáze Ensembl, jejíž genomové assembly jsou převážně na úrovni *chromosome*.

Z výše uvedeného plyne, že je nezbytné mít vždy na paměti kvalitu analyzovaných dat. Pro letouny existuje od roku 2020 šestice genomů assemblovaných v dosud nejvyšší možné kvalitě (Jebb et al., 2020). Tato situace prozatím představuje spíše výjimku i mezi savci, kde jsou tradičně nejlépe zpracované genomy člověka a myši (NCBI). To tyto dva genomy na druhou stranu činí nesrovnatelné s ostatními druhy, což je důvod, proč byly genomy člověka a myši z některých grafů vynechány a vizualizovány zvlášť. Současně je potřeba zmínit, že kvalita dostupných dat se ukázala jako faktor vyžadující neustálou kontrolu vstupních dat a také výsledků a následnou úpravu grafů atd.

4.3 Vztah počtu cds sekvencí a velikosti exomu

Skutečnost, že proporcionální zastoupení exomu (cds), ale i cDNA roste společně s velikostí celého genomu, může znamenat, že ani u savců neexistuje univerzální hodnota proporce exomu, jak by se mohlo jevit ze současně dostupných dat o lidském exomu (Hatje et al., 2019). Stejně tak je zajímavé, že i mezi savci existují druhy, jejichž exom zaujímá výrazně vyšší frakci celého genomu, než je tomu u člověka (ca. 1,17 %; Hatje et al., 2019). Opět je třeba mít na paměti kvalitu vstupních dat, ačkoliv sekvence exomu reprezentuje podstatně menší celek, u kterého lze tudíž očekávat méně závažné zatížení chybami způsobené výše uvedenými faktory. Naopak úroveň celého reprezentuje již daleko komplexnější a komplikovanější celek, kde je známo, že genomové assembly zůstávají více či méně neprosekvenované a tudíž podhodnocené. To je patrné zejm. při porovnání velikostí assembly s velikostí genomu v pikogramech (Gregory, 2021). Tudíž je možné očekávat, že skutečná proporce exomu je nižší než zde prezentovaná, avšak zatížená podobnou chybou u většiny druhů.

4.4 Využití výsledků a výhled do budoucna

Výsledky této práce budou využity v připravované publikaci, která bude navazovat na práci Andreas et al., *in prep*. Pipeline vytvořená v rámci této diplomové práce bude nadále využívána během aktualizací databáze Ensembl a rozšířena na možnost využití genomů z NCBI. Pipeline bude ještě dále rozšířena tak, aby integrovala analýzu GC% soft-maskované (repetitivní) a nesoft-maskované (nerepetitivně) frakce celého genomu

(úroveň DNA), která je v tuto chvíli již k dispozici z předchozí studie Symonová a Suh, 2019. Výsledky histogramů cDNA a cds s náznakem bimodality u některých skupin savců včetně letounů budou podkladem pro funkční analýzu GC% transkriptomu.

5 Závěr

V rámci této diplomové práce vznikl plánovaný a do budoucna nadále využitelný nástroj v jazyce Python, který automatizuje stahování, analýzu a vizualizaci GC% kompozičních dat, prozatím z databáze Ensembl.

Zásadním zjištěním této práce je, že tři druhy letounů reprezentující obě hlavní linie (Yinptero- a Yangochiroptera) mají výrazně vyšší GC% než ostatní savci jak na úrovni cDNA, tak na úrovni cds. Tato skutečnost vrhá nové světlo na dosavadní poznatky o GC biologii transpozonů u letounů jakožto hlavní faktor určující výsledné GC% hostitelského genomu. Ačkoliv výsledky této diplomové práce přinesly spíše další otazníky než očekávané odpovědi na stávající otázky, potvrzují letouny jako zajímavou modelovou skupinou savců pro kompoziční biologii genomu.

Velmi zajímavým výsledkem této práce je zjištěná bimodalita v distribuci GC% na úrovních cDNA i cds u skupin Euarchontoglires a Laurasiatheria a také u ptakopyska, zatímco ostatní skupiny vykazují unimodální symetrickou nebo zešikmenou distribuci GC% cDNA a cds anebo kombinaci obojího.

6 Literatura

ADEWALE, B. A. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *African Journal of Laboratory Medicine*. 2020, **9**(1): 1340-1345.

ALTSCHUL, S. et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 1990, **215**(3): 403-410.

ANDREAS, A et al., *in prep*. Transposons drive chiropteran quantitative cytogenomic evolution and follow bats phylogeny.

BERNARDI, G. Structural and Evolutionary Genomics: Natural Selection in Genome Evolution. Elsevier Science Ltd, 2005. 442 p.

BERNÁ, L, CHAURASIA, A., ANGELINI, C. et al. The footprint of metabolism in the organization of mammalian genomes. *BMC Genomics*. 2012, **13**: 174.

BEYER, R. M., MANICA, A., MORA, C. Shifts in global bat diversity suggest a possible role of climate change in the emergence of SARS-CoV-1 and SARS-Cov-2. Science of the Total Environment. 2021, 767(3): 145413.

BIRCHLER, J. A. et VEITIA, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *PNAS*. 2012, **109**(37): 14746-14753.

BLOMMAERT, J. Genome size evolution: towards new model systems for old questions. *Proceedings of the Royal Society B*. 2020, **287**: 20201441.

BOURGEOIS, Y. et BOISSINOT, S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes*. 2019, **10**(6): 419.

BORQUE, G., BURNS, K. H., GEHRING, M. et al. Ten things you should know about transposable elements. *Genome Biology*. 2018, **19** (199).

BROWN, T., 2021. *Understanding a Genome Sequence*. [online] Ncbi.nlm.nih.gov. Dostupné z: <https://www.ncbi.nlm.nih.gov/books/NBK21136/> [cit. 2021-02-11].

BUČEK, P. *Použití pokročilých chemometrických technik pro výzkum struktury DNA*. Brno, 2008. Diplomová práce. Masarykova univerzita. Přírodovědecká fakulta.

BURT, D. W. Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*. 2002, **96**: 97-112.

CANCHERINI, D. V., FRANÇA, G. S., de SOUZA, S. J. The role of exon shuffling in shaping protein-protein interaction networks. *BMC Genomics*. 2010, **11**(Suppl 5): S11.

CIRULLI, E. T., SINGH, A, SHIANNA, K. V. et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biology*. 2010, **11**: R57.

CROLLIUS, H. R. The *Tetraodon* Genome. *Genome Dynamics*. 2006, **2**: 154-164.

Ebi.ac.uk. 2021. *Paul Flicek | EMBL's European Bionformatics Institute*. [online] Dostupné z: <https://www.ebi.ac.uk/about/people/paul-flicek> [cit. 2021-03-15].

DOLEŽEL, J., BARTOŠ, J., VOGLMAYR, H., GREILHUBER, J. Letter to the editor. *Cytometry*. 2003, **51A**(2): 127-128.

ELHAIK, E., LANDAN, G., GRAUR, D. Can GC Content at Third-Codon Positions Be Used as a Proxy for Isochore Composition? *Molecular Biology and Evolution*. 2009, **26**(8): 1829–1833.

ELLIOTT, Tyler A. et GREGORY, Ryan T. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B*. 2015, **370**: 20140331.

Ensembl.org. 2021. *Glossary*. [online] Dostupné z: <https://www.ensembl.org/info/website/glossary.html> [cit. 2021-02-12].

EVANS, B., UPHAM, N. S., GOLDING, G. B., OJEDA, R. A., OJEDA, A. A. Evolution of the Largest Mammalian Genome. *Genome Biology and Evolution*. 2017, **9**(6): 1711-1724.

FREELING, M., XU, J., WOODHOUSE, M., LISCH, D. A Solution to the C-Value Paradox and the Function of Junk DNA: The Genome Balance Hypothesis. *Molecular Plant*. 2015, **8**: 899-910.

GAISLER, J., ZIMA, J. Zoologie obratlovců. Praha: Academia, 2007.

GALTIER, N., PIGANEAU, G., MOUCHIROUD, D., DURET, L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*. 2001, **159**(2): 90-911.

GALTIER, N., MOUCHIROUD, D. Isochore evolution in mammals: a human-like ancestral structure. Genetics. 1998, **150**: 1577-1584.

GILBERT, W., de SOUZA, S. J., LONG, M. Origin of Genes. PNAS. 1977, **94**(15): 7698-7703.

GREGORY, T. R. Coincidence, coevolution, or causation? DNAcontent, cell size, and the C-value enigma. *Biological Reviews*. 2001, **76**: 65-101.

GREGORY, T. T. The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. *Annals of Botany*. 2005, **95** (1): 133-146.

HATJE, K., MÜHLHAUSEN, S., SIMM, D., KOLLMAR, M. The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *BioEssays*. 2019, **41**: 1900066.

HEALTON, S. E., PINTO, H. D., MISHRA, L. N., HAMILTON, G. A. et al. H1 linker histones silence repetitive elements by promoting both histone H3K9 methylation and chromatin compaction. *PNAS*. 2020, **117**(25): 14251-14258.

HILDEBRAND, F., MEYER, A., WALKER, A. E. Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genetics*. 2010, **6**(9): E1001107.

HURST, L. D., et MERCHANT, A. R. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceeding. Biological sciences*. 2001, **268**(1466):493-7.

IBARRA-LACLETTE, E., LYONS, E., HERNÁNDEZ-GUZMÁN, G., et al. Architecture and evolution of a minute plant genome. *Nature*. 2013;498(7452):94-98.

JEBB, D., HUANG, Z., PIPPEL, M. et al. Six reference-quality genomes reveal evolution of bat adaptations. *Nature*. 2020, **583**: 578-584.

KAPUSTA, A., SUH, A., FESCHOTTE, C. Dynamics of genome size evolution in birds and mammals. *PNAS*. 2017, **114**(8): E1460-E1469.

KASAI, F., O'BRIEN, P. C. M., FERGUSON-SMITH, M. A. The bat genome: GC-biased small chromosomes associated with reduction in genome size. *Chromosoma*. 2013, **122**: 535-540.

KOSTKA, D., HUBISZ, M. J., SIEPEL, A., POLLARD, K. S. The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution*. 2012, **29**(3): 1047-1057.

KU, CH., WU, M., COOPER, D. N., NAIDOO, N. et al. Exome versus transcriptome sequencing in identifying coding region variants. *Expert Review of Molecular Diagnostics*. 2014, **12**(3): 241-251.

LIU, J., McCLELAND, M., STAWISKI, E. et al. Integrated exome and transcriptome sequencing reveals *ZAK* isoform usage in gastric cancer. *Nature Communications*. 2014, **5**: 3830.

MATYÁŠEK, R. et KOVAŘÍK, A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes*. 2020, 11(7): 761.

de MENDOZA, A., HATLEBERG, W. L., PANG, K., LEININGER, S. et al. Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nature Ecology & Evolution*. 2019, **3**(10): 1464-1473.

MEYER, A., SCHLOISSNIG, S., FRANCHINI, P., DU, K. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*. 2021, **590**: 284-289.

MUGAL, C. F., WEBER, C. C., ELLEGREN, H. GC-biased gene conversion links the recombination landscape and demography to genomic base composition. *Bioessays*. 2015, **37**: 1317-1326.

NOWOSHILOW, S., SCHLOISSNIG, S., FEI, J.-F., DAHL, A. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature*. 2018, **554**(7690): 50–55.

ORGAN, CH. L. et SHEDLOCK, A. M. Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. *Biology Letters*. 2009, **5**(1): 47-50.

PRAY, L. Transposons: The Jumping Genes. *Nature Education*. 2008, **1**(1): 204.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

RAVICHANDRAN, S., SUBRAMANI, V. K, KIM, K. K. Z-DNA in the genome: from structure to disease. *Biophysical Reviews*. 2019, **11**(3): 383-387.

REAM, R. A., JOHNS, G. C., SOMERO, G. N. Base compositions of genes encoding alpha-actin and lactate dehydrogenase-A from differently adapted vertebrates show no temperature-adaptive variation in G+C content. *Molecular Biology and Evolution*. 2003, **20**(1): 105-10.

RYABOV, E. V. Construction of infectious cDNA clones for RNA viruses: Turnip crinkle virus. *Methods in Molecular Biology*. 2008, **451**: 491-502.

SHI, X., CHEN, CH., YANG, H., HOU, J., JI, T. et al. The Gene Balance Hypothesis: Epigenetics and Dosage Effects in Plants. *Methods in molecular biology*. **202**, 2093: 161-171.

SLIJEPCEVIS, P. Genome dynamics over evolutionary time: „C-value enigma“ in light of chromosome structure. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*. 2018, 27 p.

SNUSTAD, D. Peter a Michael J. SIMMONS, RELICHOVÁ, Jiřina, ed. *Genetika*. Přeložil Jiří DOŠKAŘ. Brno: Masarykova univerzita, 2009.

SOTERO-CAIO, C. G., BAKER, R. J. Chromosomal Evolution in Chiroptera. *Genes*. 2017, **8**(10): 272.

SPRINGER. M. S., STANHOPE, M. J., MADSEN, O., de JONG, W. W. Molecules consolidate the placental mammal tree. Trends in Ecology and Evolution. 2004, 19(8): 130-438.

SREBNIAK, M. I., DIDERICH, K. EM, JOOSTEN, M., GOVAERTS, L. CP et al. Prenatal SNP array testing in 1000 fetuses with ultrasound anomalies: causative, unexpected and susceptibility CNVs. *European Journal of Human Genetics*. 2016, **24**: 645-651.

STEKEL, D. J., GIT, Y., FALCIANI, F. The Comparison of Gene Expression from Multiple cDNA Libraries. *Genome Research*. 2000, **10**: 2055-2061.

SYMONOVÁ, R., MAJTÁNOVÁ, Z., ARIAS-RODRIGUEZ, L., MOŘKOVSKÝ, L. et al. Genome Compositional Organization in Gars Shows More Similarities to Mammals than to Other Ray-Finned Fish. *J Exp Zool B Mol Dev Evol* **328**(7): 607-619.

SYMONOVÁ, R. 2021. Komposiční evoluce genomu strunatců. Habilitační práce, Přírodovědecká fakulta Jihočeské Univerzity v Českých Budějovicích.

SYMONOVÁ, R. et SUH, A. Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. *Mobile DNA*. 2019, **10**: 49.

ŠMARDA, P., BUREŠ, P., HOROVÁ, L., LEITCH, I. J. et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences USA*. 2014, **111**(39): E4096-102.

TAKAHASHI, S. et SUGIMOTO, N. Watson-Crick versus Hoogsteen Base Pairs: Chemical Strategy to Encode and Express Genetic Information in Life. *Accounts of Chemical Research*. 2021, **54**(9): 2110-2120.

TEELING, E., VERNES, S. C., DÁVALOS, L. M., RAY, D. A. et al. Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for all Living Bat Species. *Annual Review of Animal Biosciences*. 2018, 6: 23-46.

TREANGEN, T. J. et SALZBERG, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature*. 2012, **13**(1): 36-46.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W. et al. The Sequence of the Human Genome. *Science*. 2001, **291**(5507): 1304-1351.

VINOGRADOV, A. E. et ANATSKAYA, O. V. DNA helix: the importance of being AT-rich. *Mammalian Genome*. 2017, **25**: 455-464.

WASKOM, M. L. Seaborn: Statistical data visualization. *Journal of Open Source Software 2021,* **6**(60): 3021. https://doi.org/10.21105/joss.03021

WATSON, J. D. et CRICK, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. 1953, **171**(4356): 737-38.

WEBER, C. C., BOUSSAU, B., ROMIGUIER, J., JARVIS, E. D., ELLEGREN, H. et al. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology*. 2014, **15**: 549.

WRIGHT, N. A., GREGORY, T. R., WITT. CH. C. Metabolic ‘engines’ of flight drive genome size reduction in birds. Proceedings of the Royal Society B. 2014, **287**: 20132780.

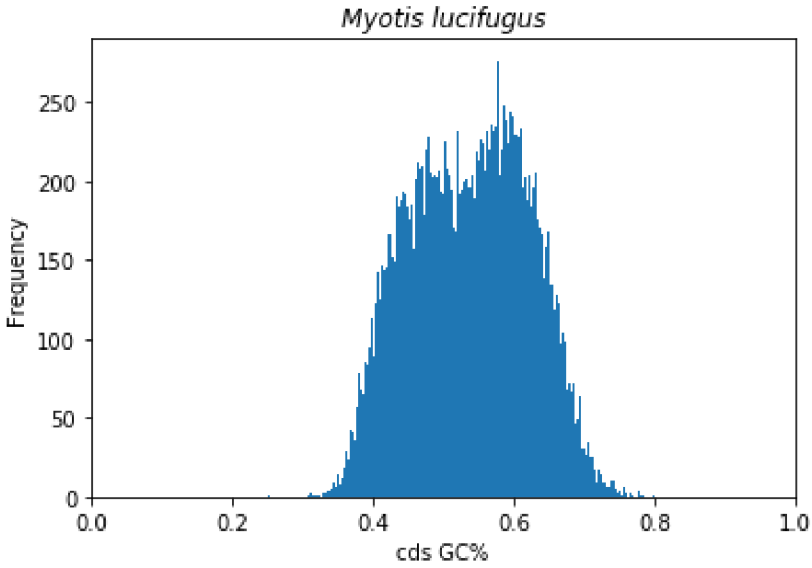
WYATT, A. W., MO, F., WANG, K. et al. Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer. *Genome Biology*. 2014, **15**: 426.

YAKOVCHUK, P., PROTOZANOVA, E., FRANK-KAMENETSKII, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*. 2006, **34**(2): 564-574.

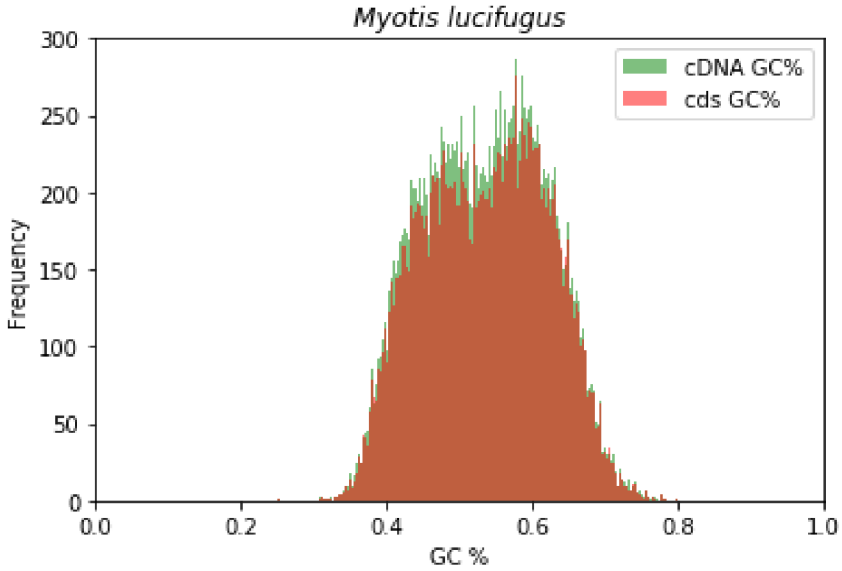
ZHAO, R. F., 2012. *ENCODE: Deciphering Function in the Human Genome*. [online] Dostupné z: <https://www.genome.gov/27551473/genome-advance-of-the-month-encode-deciphering-function-in-the-human-genome> [cit. 2021-02-13]

ZHOU, W., LIANG, G., MOLLOY, P. L., JONES, P. A. DNA methylation enables transposable element-driven genome expansion. *PNAS*. 2020, **177**(32): 19359-19366.

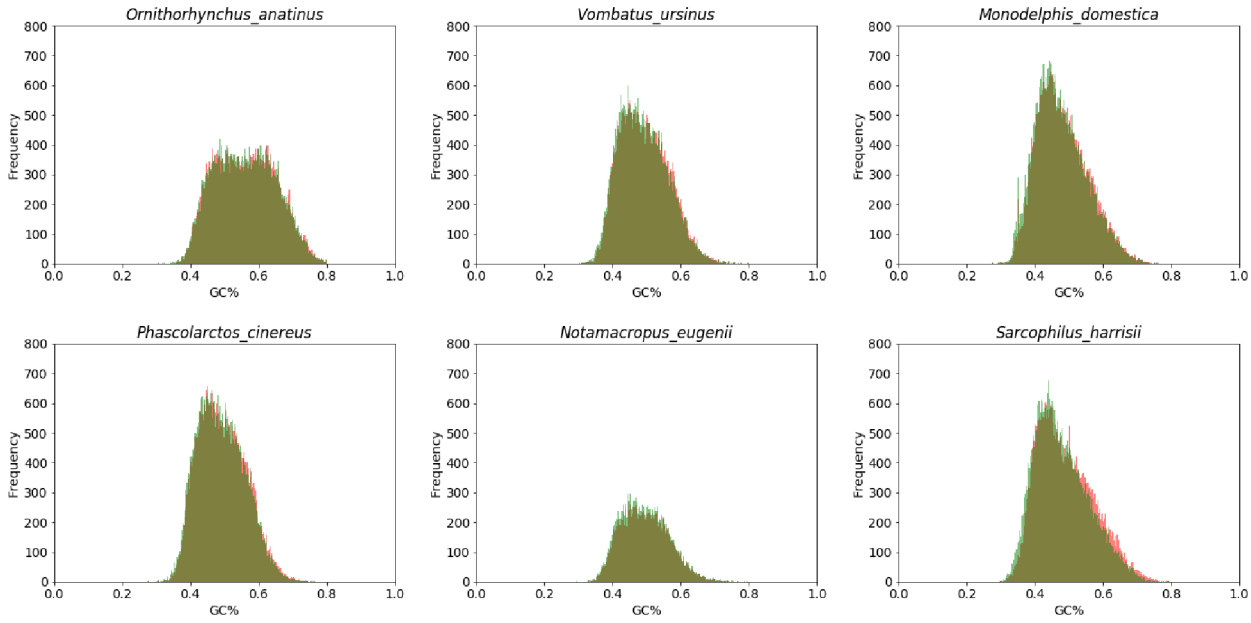
7 Přílohy



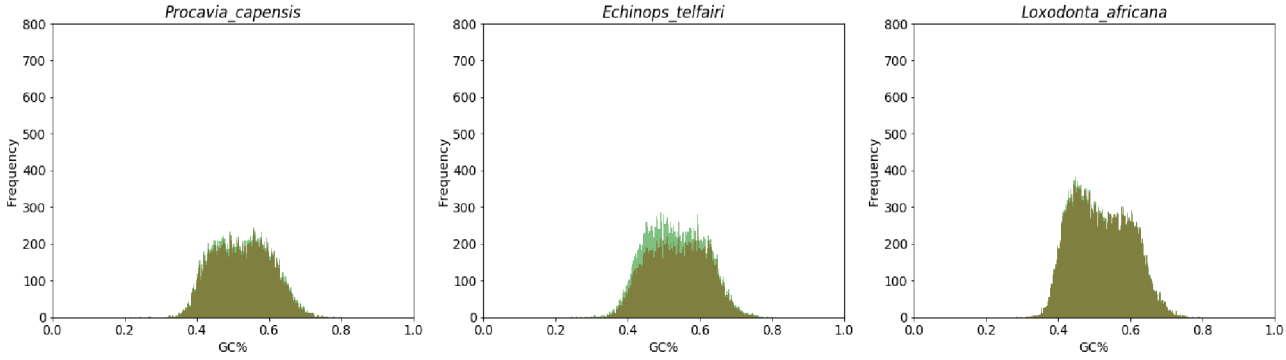
Graf 1 – Histogram frekvence s GC% cds Myotis lucifugus



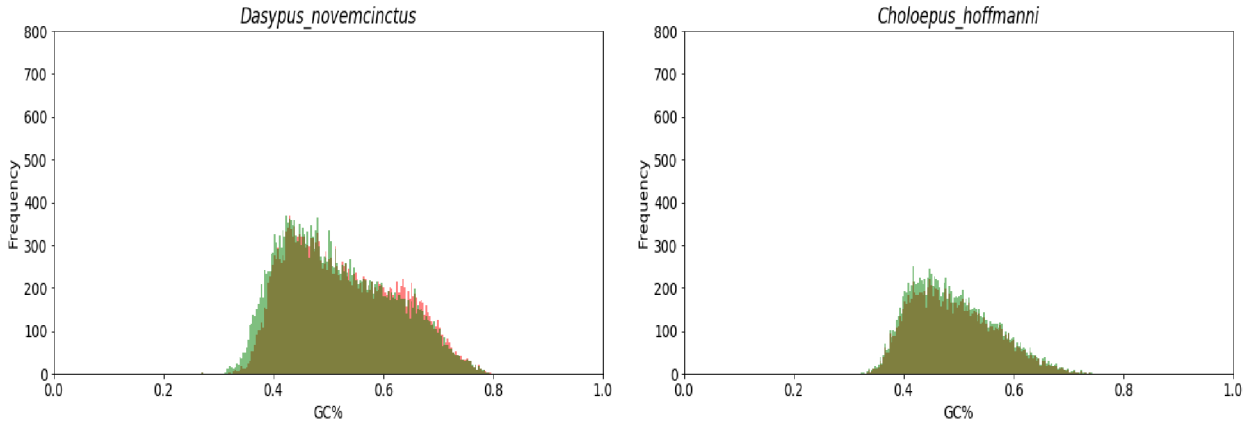
Graf 2 – Histogram s překrývajícími hodnotami cds a cDNA



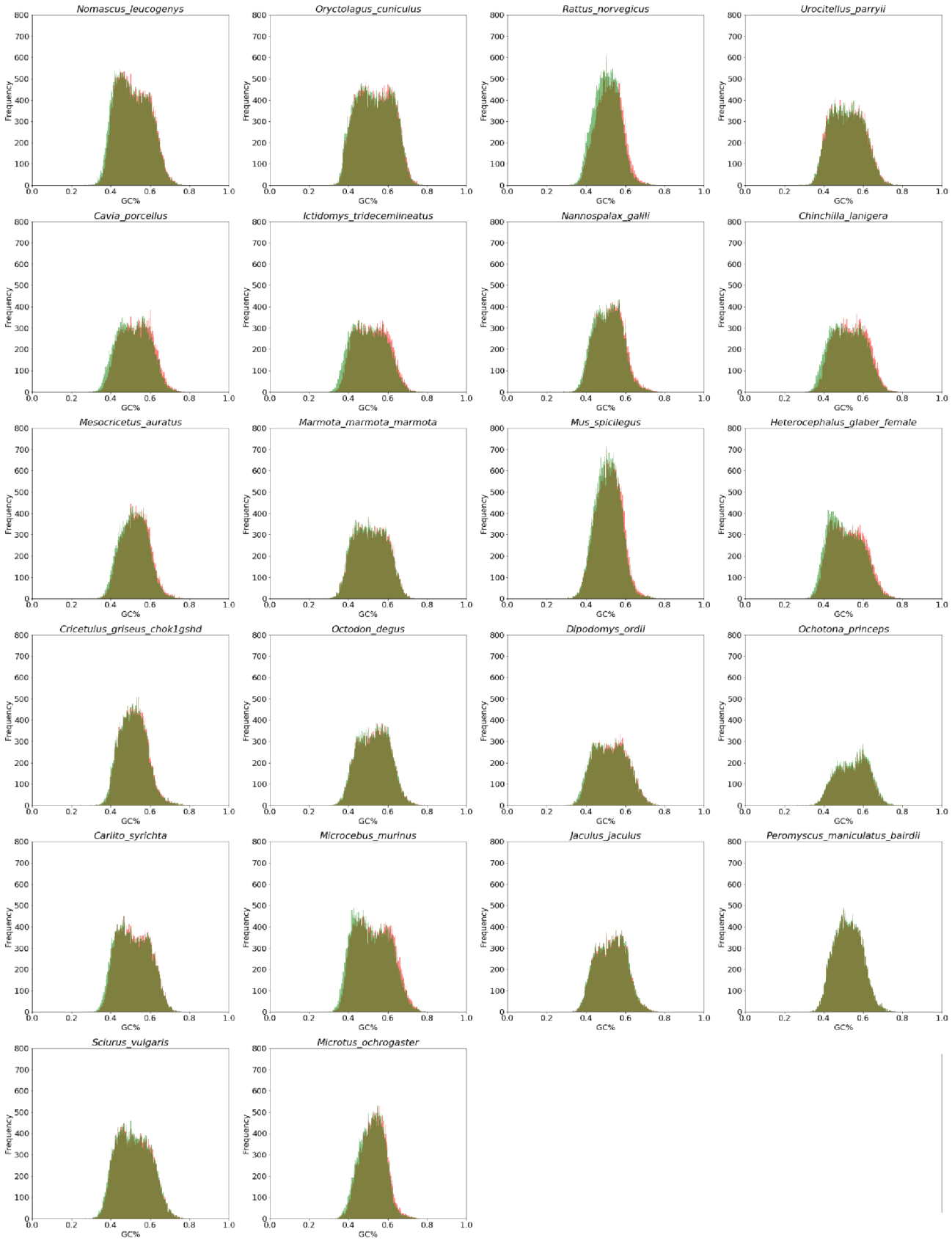
Graf 3 – Histogramy GC% cds a cDNA skupiny Monotremata a Marsupialia



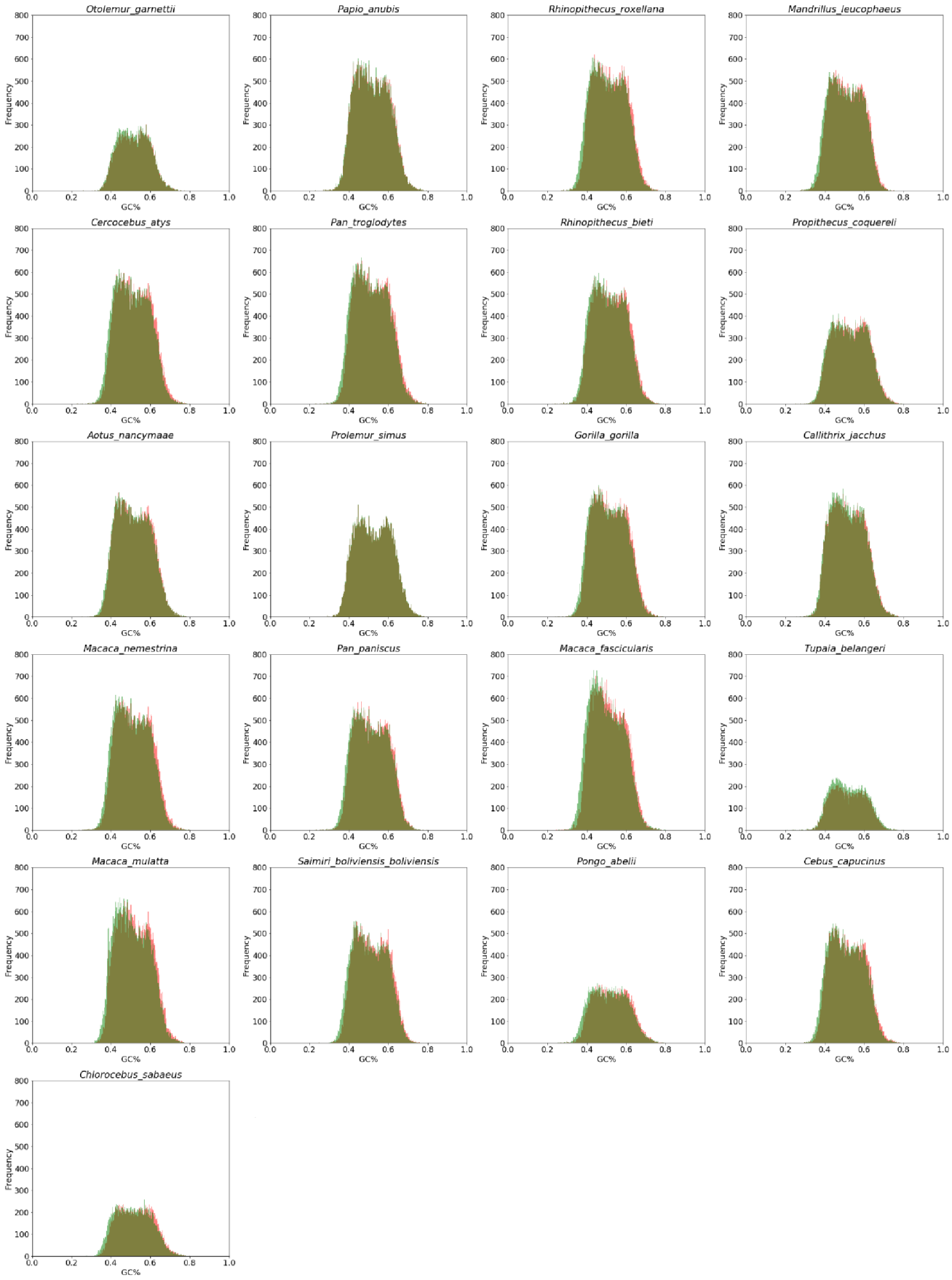
Graf 4 – Histogramy GC% cds a cDNA skupiny Afrotheria



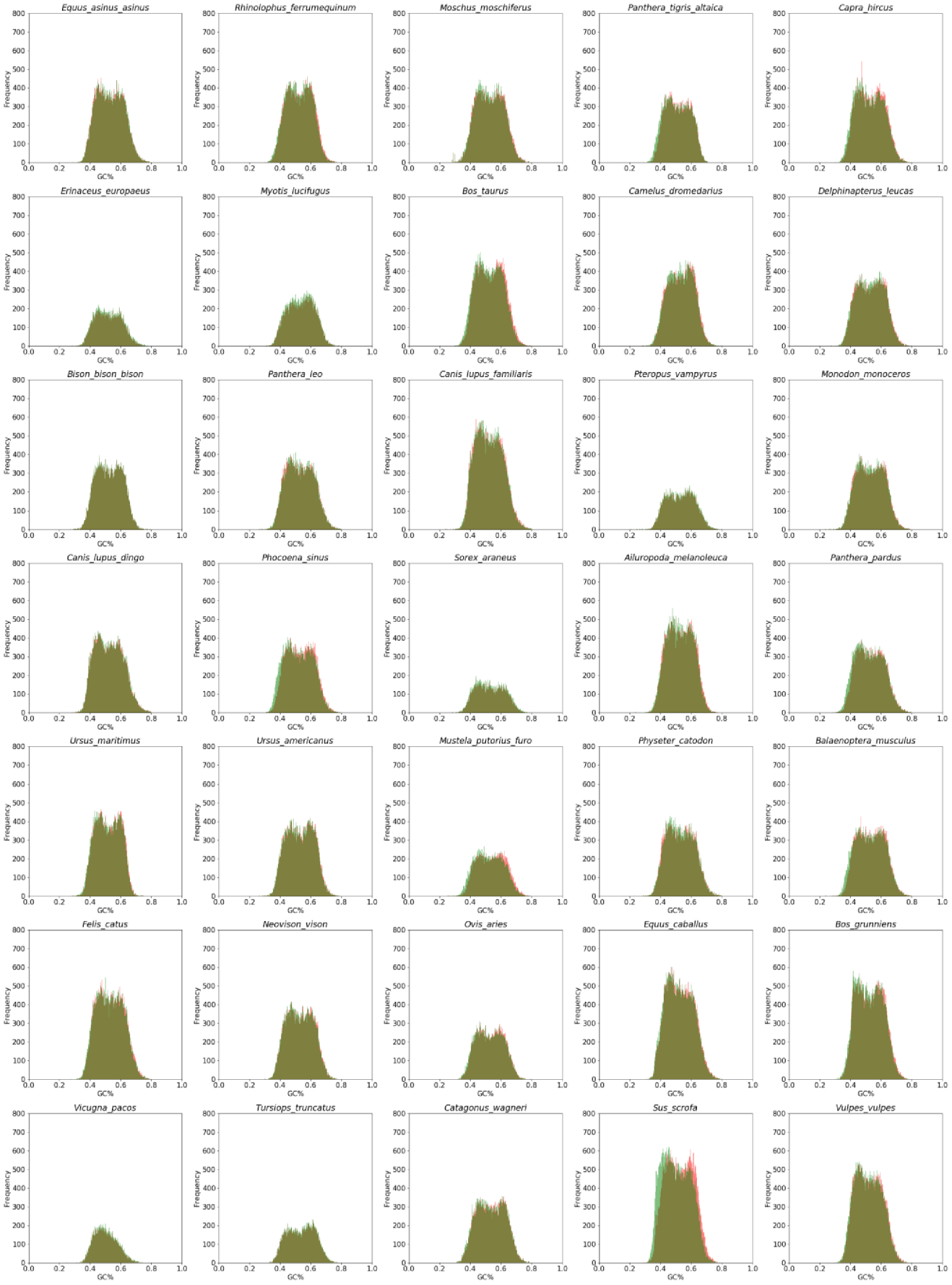
Graf 5 – Histogramy GC% cds a cDNA skupiny Xenarthra



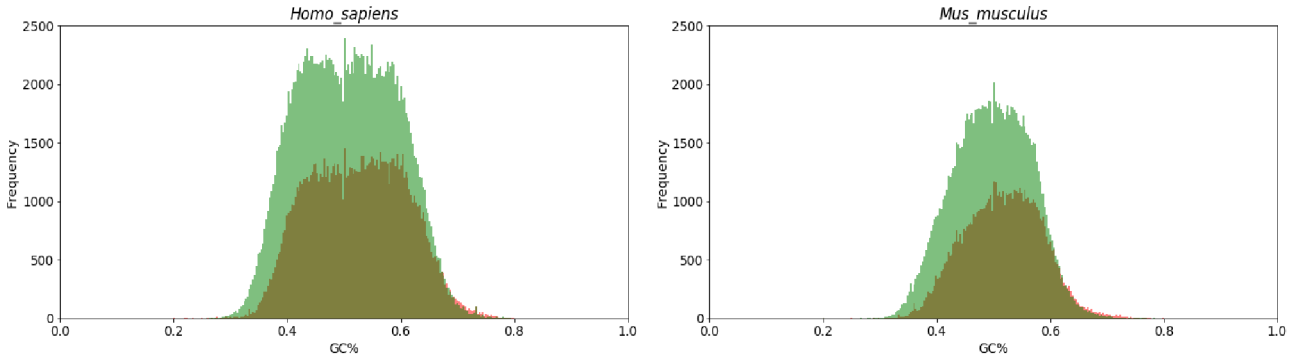
Graf 6 – Histogramy GC% cds a cDNA skupiny Euarchontoglires (Primates, Scandentia)



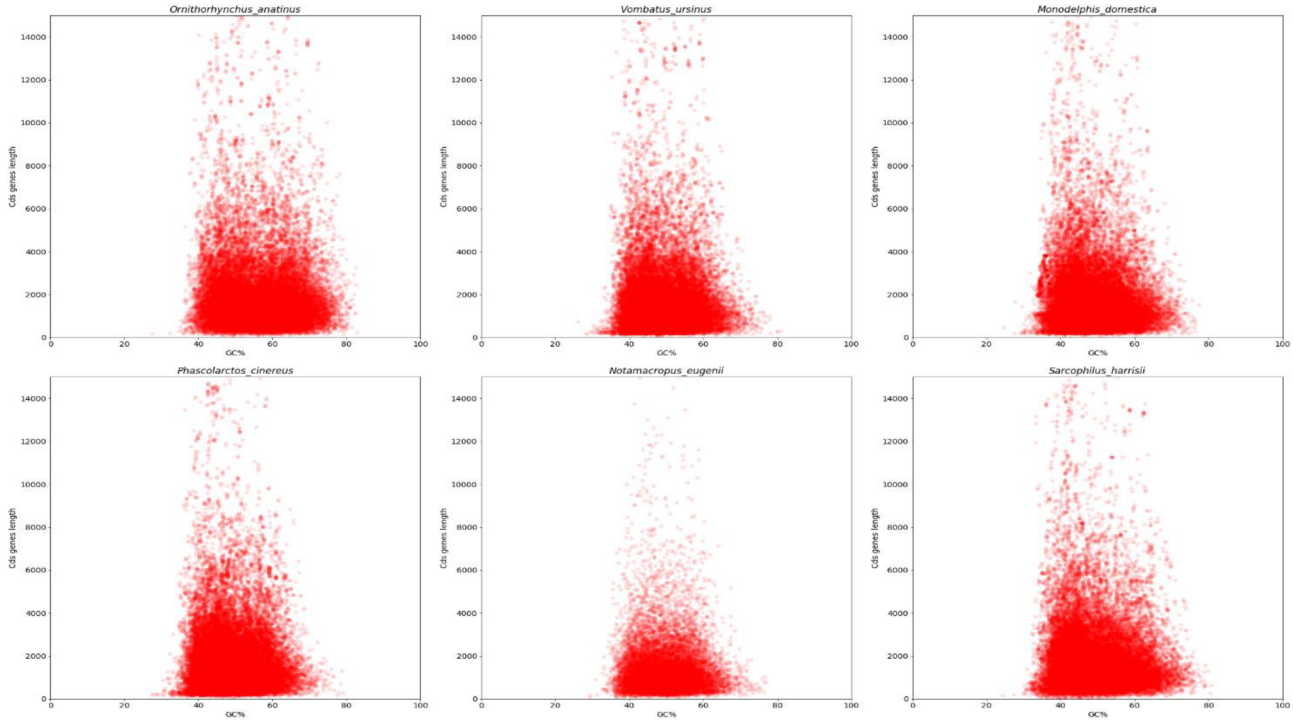
Graf 7 – Histogramy GC% cds a cDNA skupiny Euarchontoglires (Lagomorpha, Rodentia)



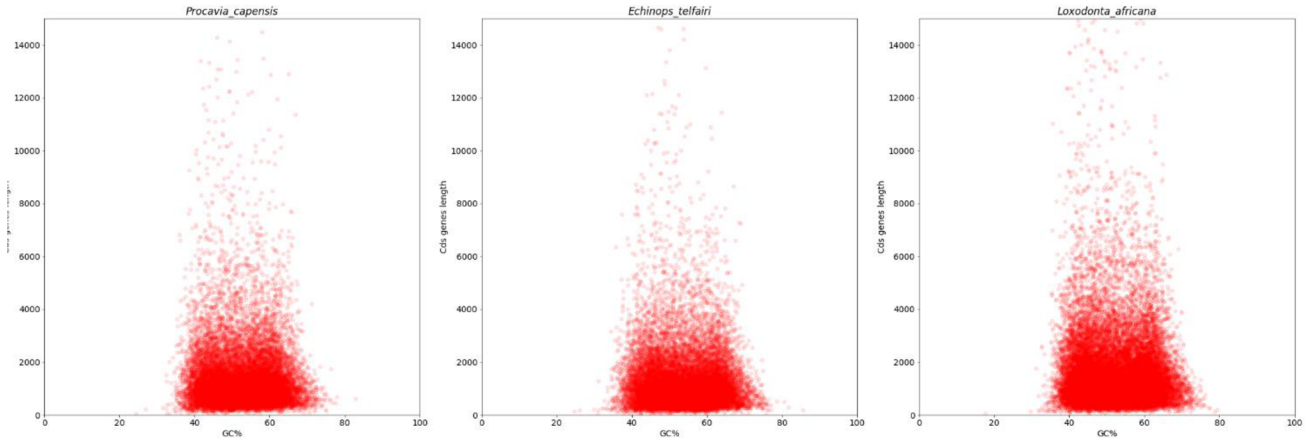
Graf 8 – Histogramy GC% cds a cDNA skupiny Laurasiatheria



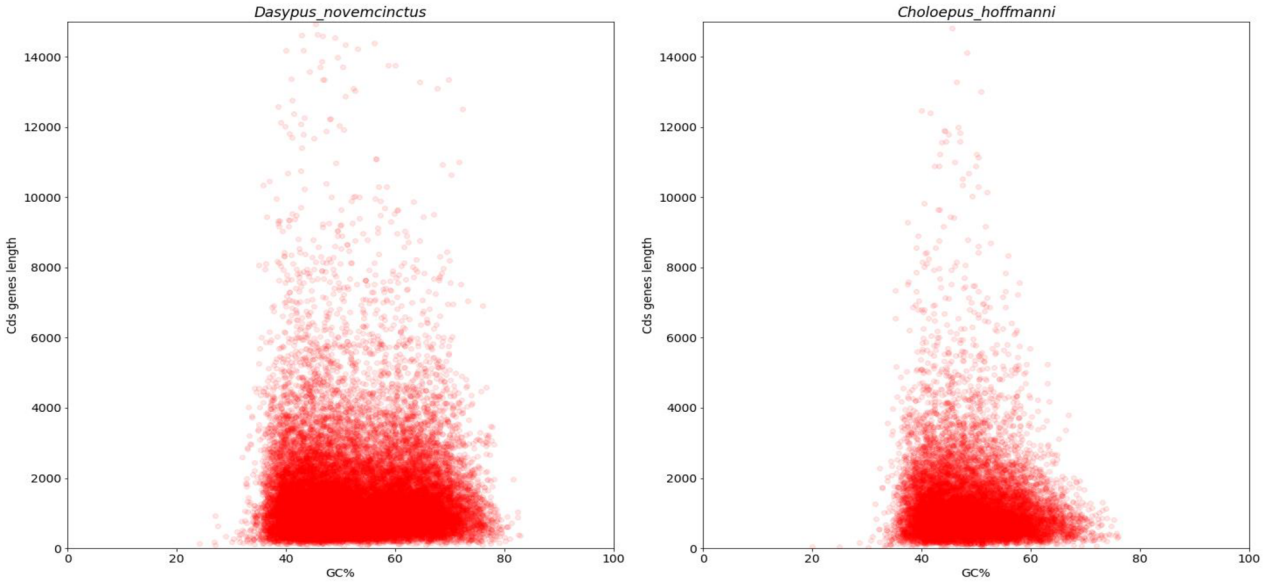
Graf 9 – Histogramy GC% cds a cDNA Homo sapiens a Mus musculus



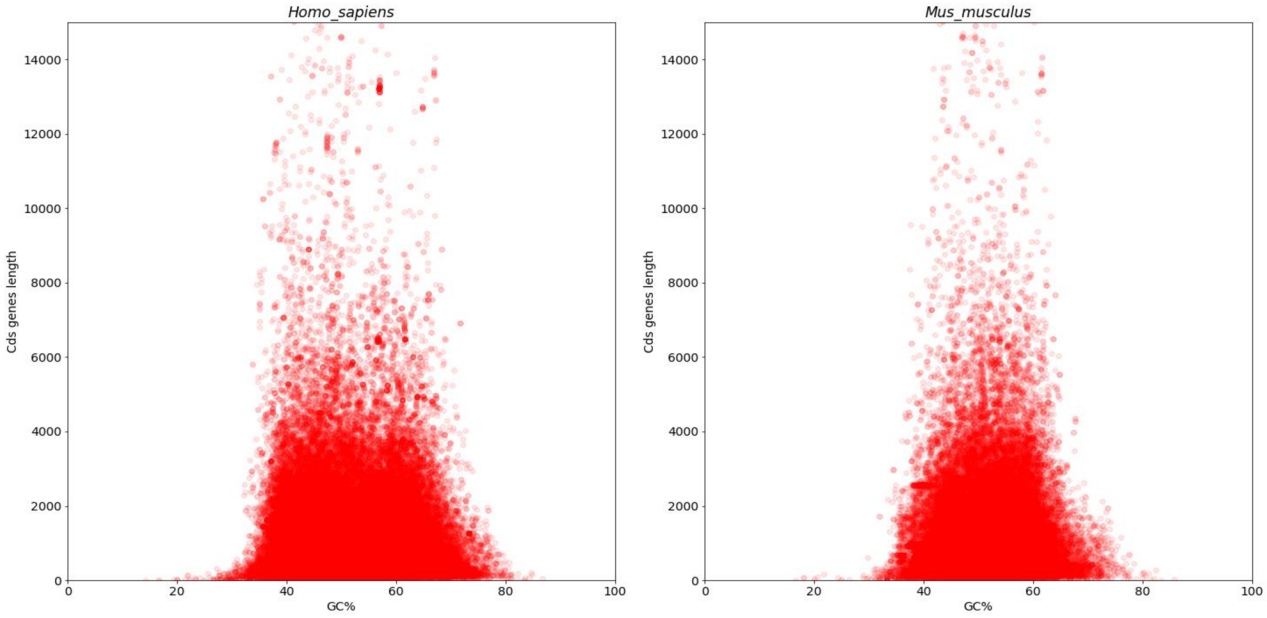
Graf 10 – Scatter ploty GC% vs délka genů cds skupiny Monotremata a Marsupialia



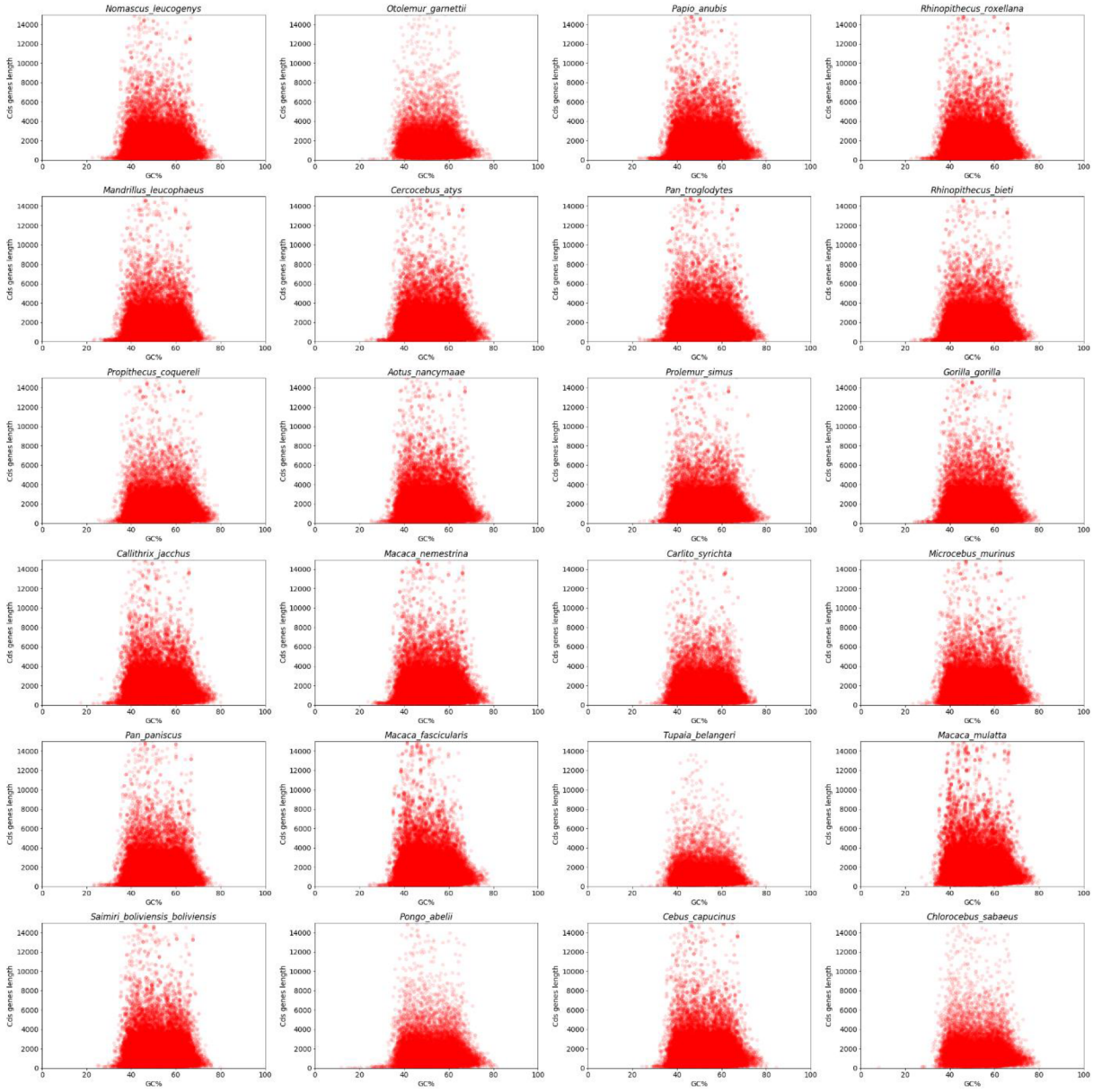
Graf 11 – Scatter ploty GC% vs délka genů cds skupiny Afrotheria



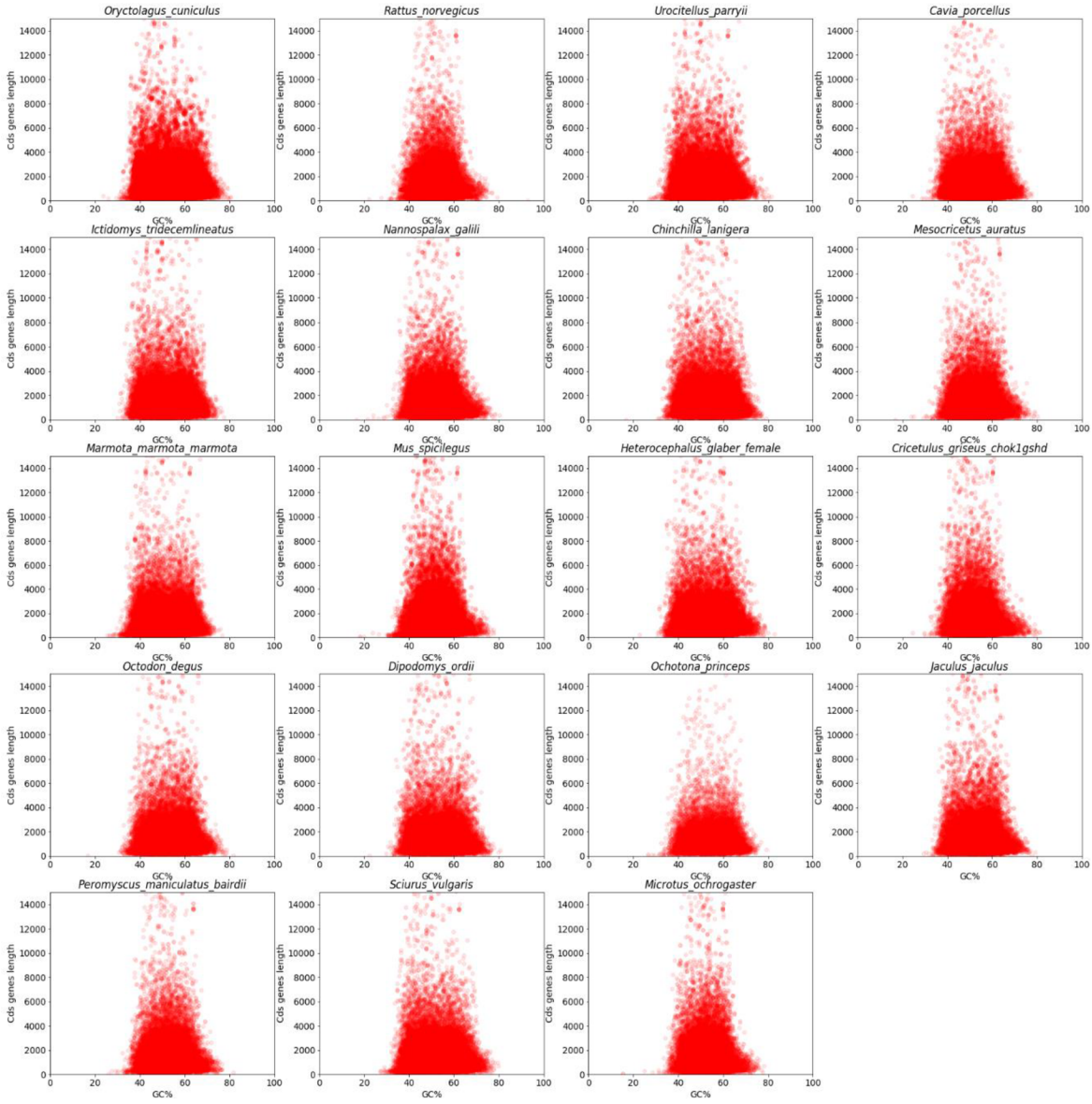
Graf 12 – Scatter ploty GC% vs délka genů cds skupiny Xenarthra



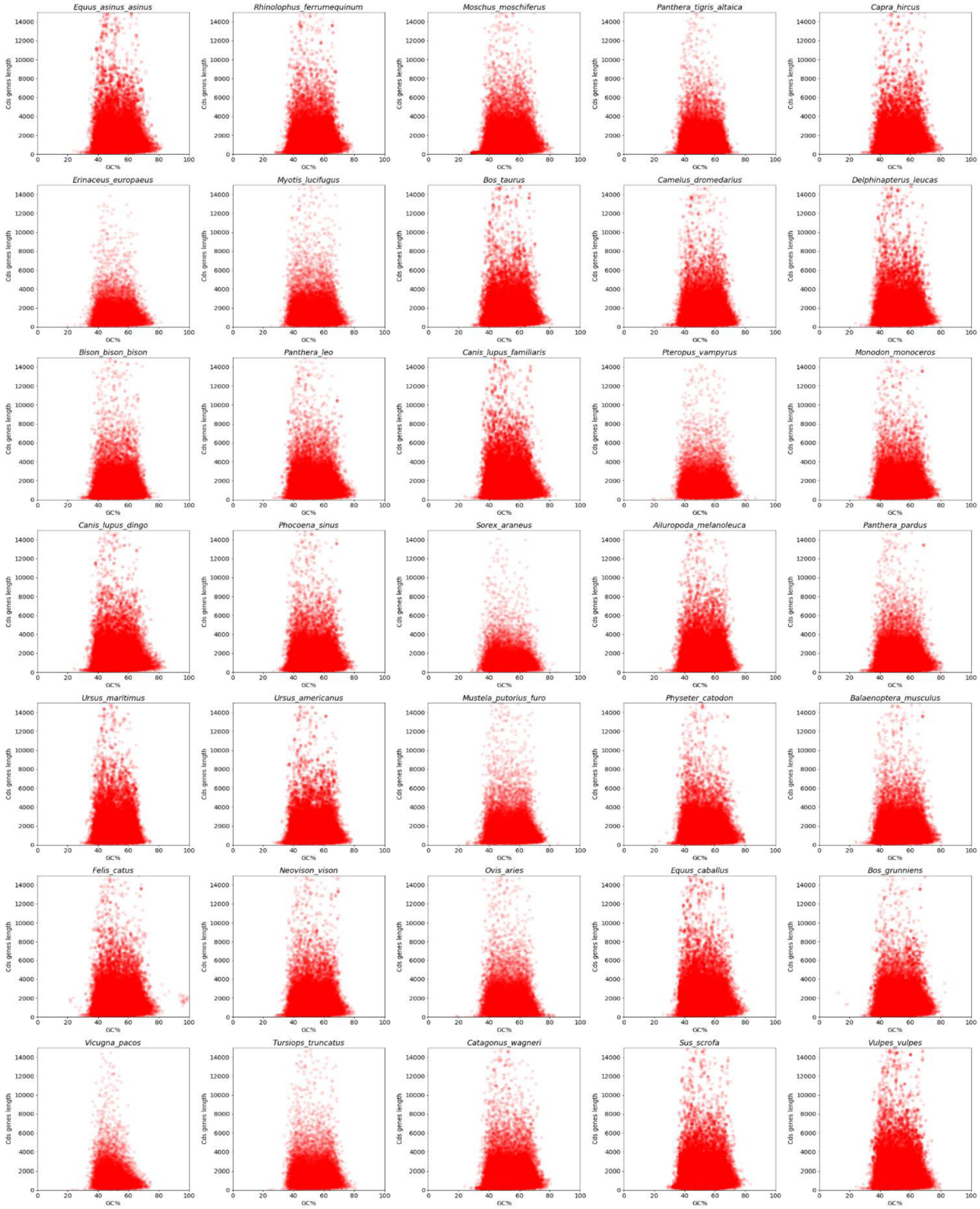
Graf 13 – Scatter ploty GC% vs délka genů cds Homo sapiens a Mus musculus



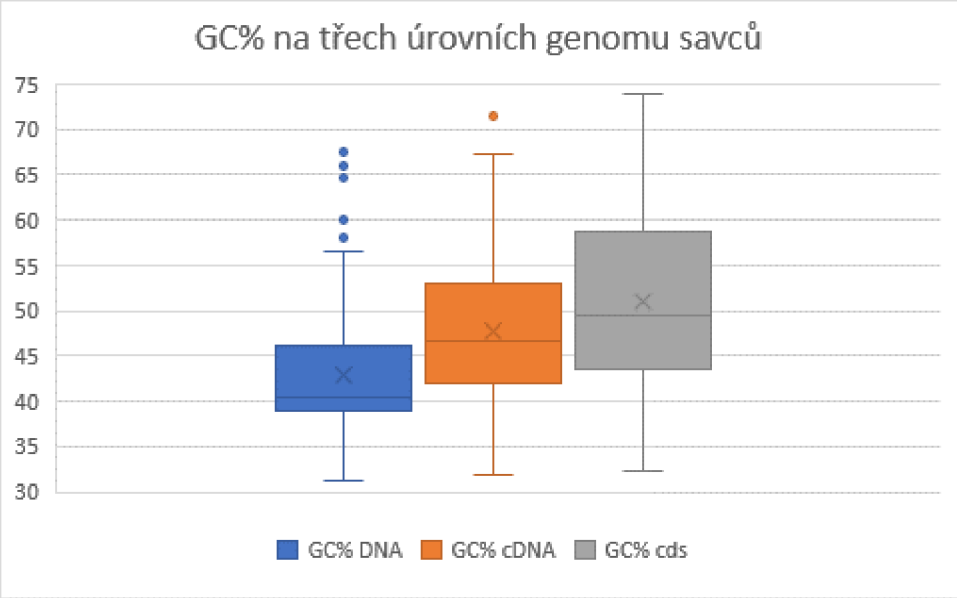
Graf 14 – Scatter ploty GC% vs délka genů cds skupiny Euarchontoglires (Primates, Scandentia)



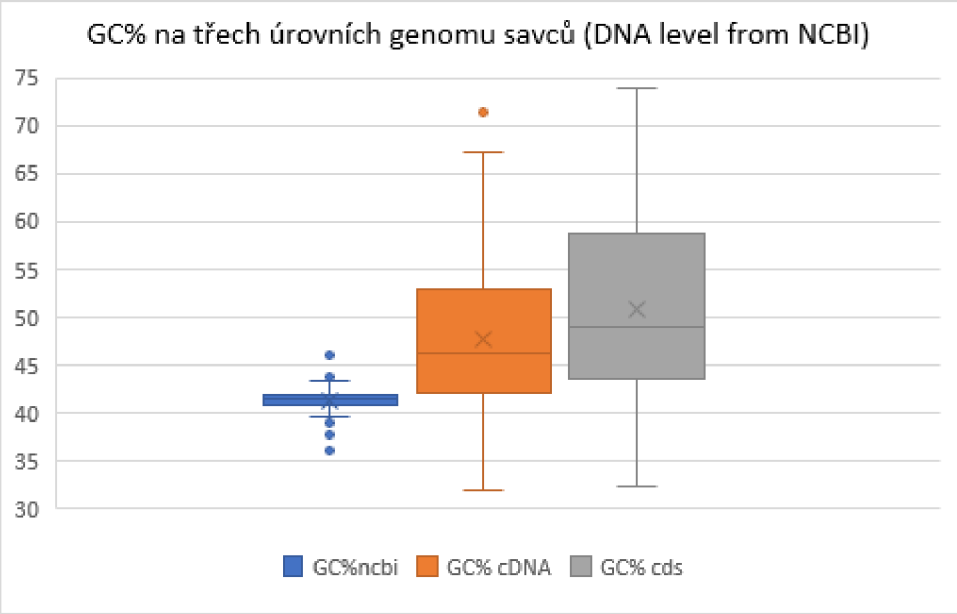
Graf 15 – Scatter ploty GC% vs délka genů cds skupiny Euarchontoglires (Rodentia, Lagomorpha)



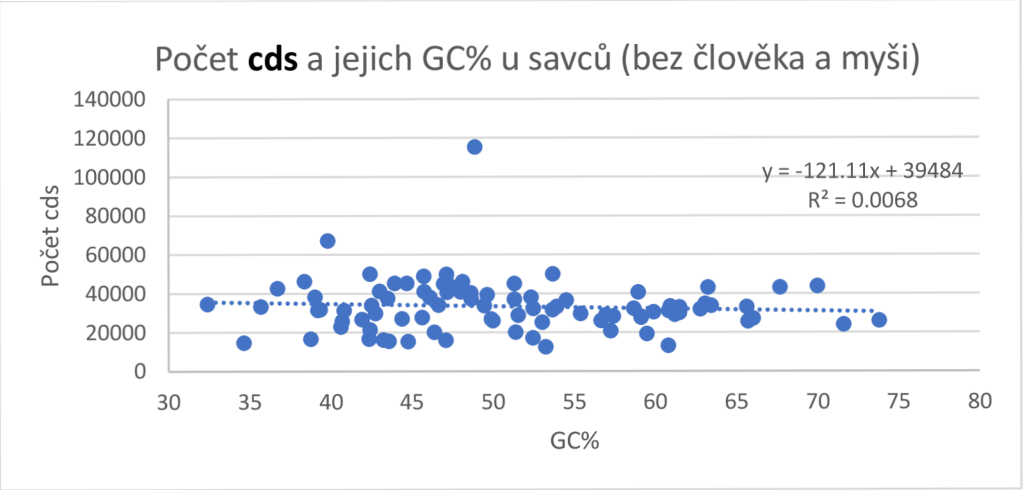
Graf 16 – Scatter ploty GC% vs délka genů cds skupiny Laurasiatheria



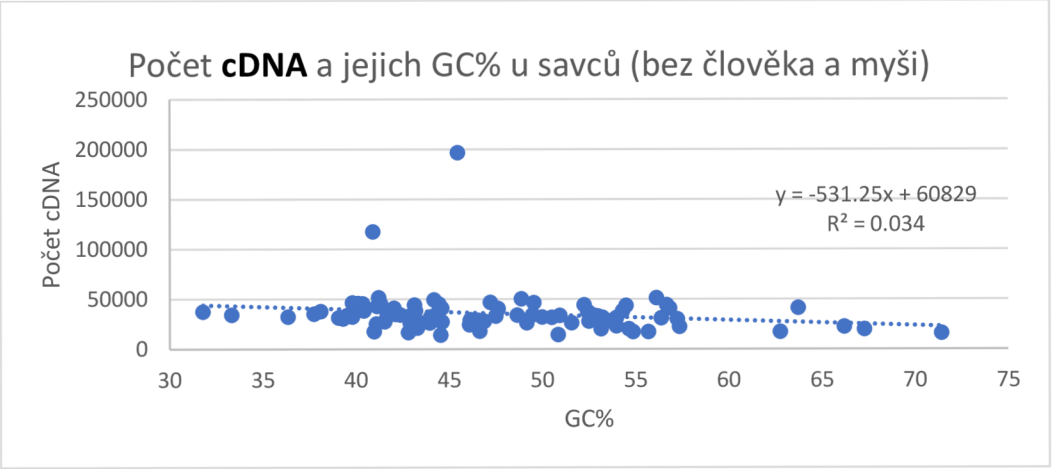
Graf 17 – GC% na třech úrovních genomu (Ensembl)



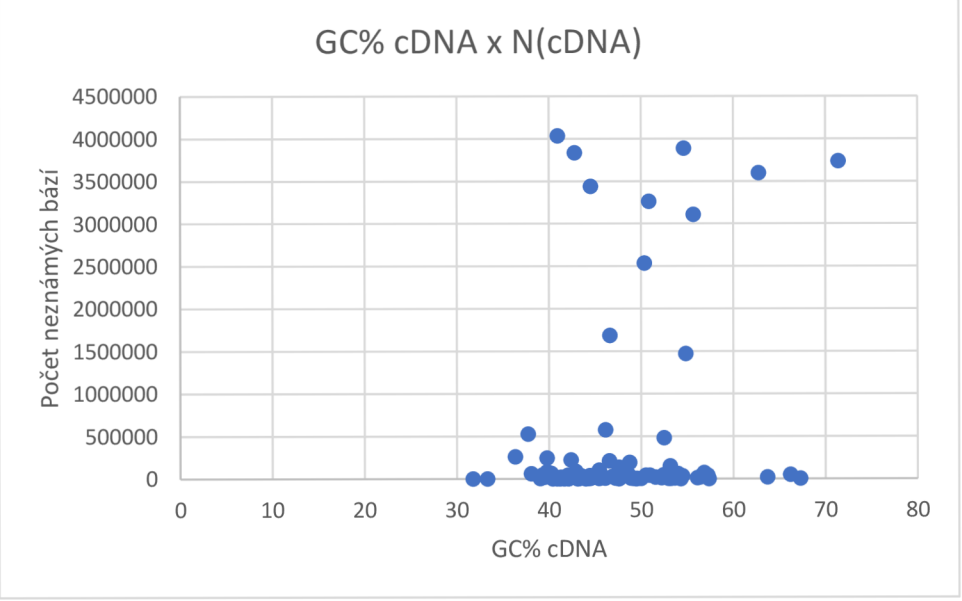
Graf 18 – GC% na třech úrovních genomu (NCBI)



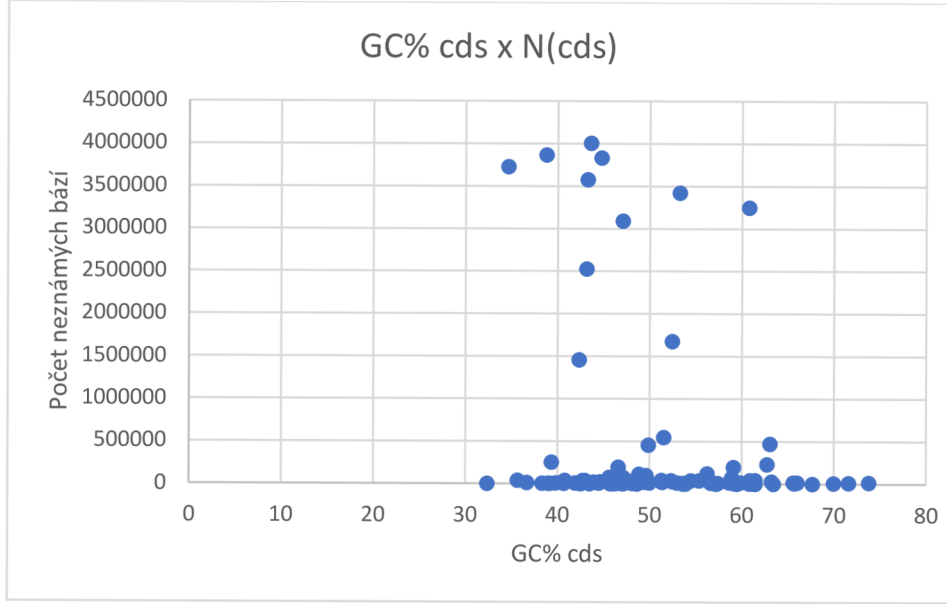
Graf 19 – délky cds vůči GC%



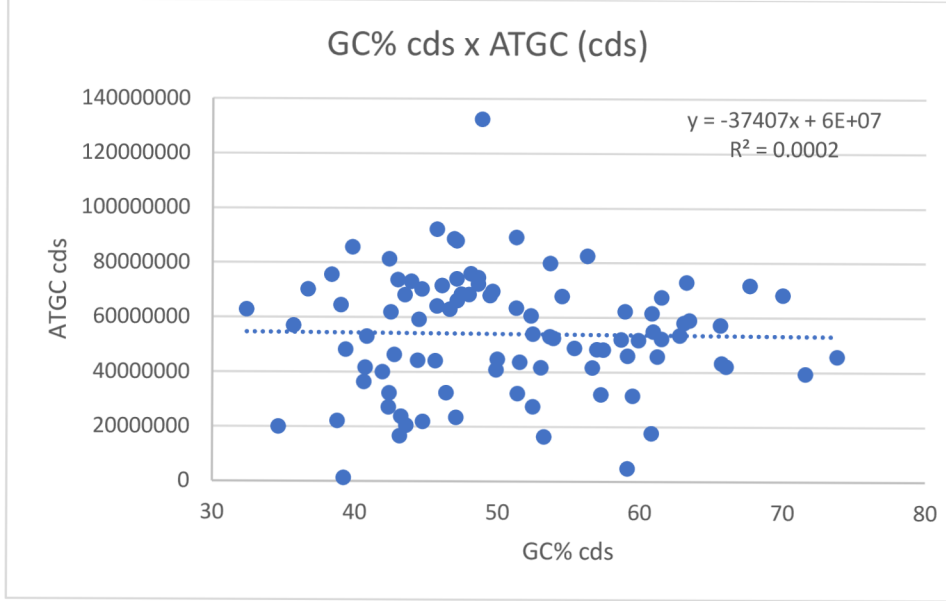
Graf 20 – délky cDNA vůči GC%



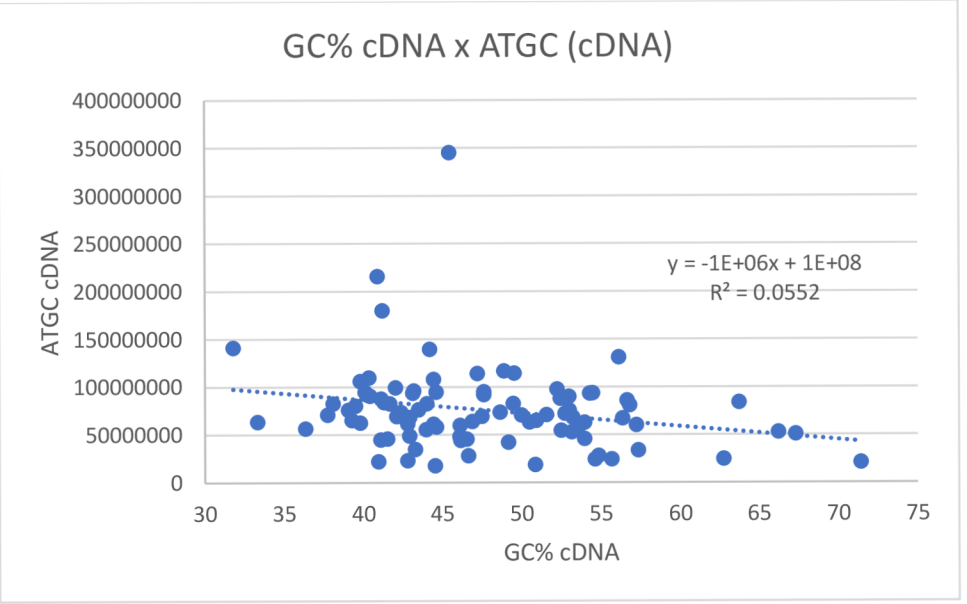
Graf 21 – GC% cDNA vůči počtu neznámých bází



Graf 22 – GC% cds vůči počtu neznámých bází



Graf 23 – GC% vůči celé sekvenci cds



Graf 24 – GC% vůči celé sekvenci cDNA

Tab. S1 – Výpočty GC% pro jednotlivé frakce savčích genomů, jejich délky (počet genů), počet bází A, T, G a C, počet nespecifikovaných dusíkatých bází (N) a procentuální zastoupení cds a cDNA z celkové DNA (Ensembl data)

Druh	GC% DNA	GC% cDNA	GC% cds	len cds	len cDNA	N(DNA)	N(cds)	N(cDNA)	ATGC(DNA)	ATGC (cds)	ATGC (cDNA)	% cDNA z DNA	% cds z DNA
Ornitorhynchus anatinus	46.1	52.98	61.51	32939	33146	10531811	684	1554	1802811296	67423057	89856567	4.98	3.74
Sarcophilus harrisii	33.7	31.79	48.66	37202	37458	46357	300	400	3081590730	72189045	141220233	4.58	2.34
Monodelphis domestica	41.03	52.44	54.52	36557	37423	89779610	35656	45131	3412593428	67748702	87932411	2.58	1.99
Notamactopus eugenii	45	42.81	44.77	15342	16838	414069667	3830323	3835095	2541704270	21849677	23049837	0.91	0.86
Vombatus ursinus	32.2	45.45	44.5	32889	33466	11420771	6186	1464	3475176447	59218507	78944556	2.27	1.7
Phascolarctos cinereus	39.09	40.42	46.11	37921	38432	1	0	0	3192581491	71674219	90476012	2.83	2.25
Choloepus hoffmanni	52.4	44.55	53.27	12435	14072	407073508	3420397	3437856	2060419685	16377944	17536608	0.85	0.79
Dasypus novemcinctus	37	46.88	41.93	26551	28051	331640683	7069	27141	3299882028	39997345	63658588	1.93	1.21
Echinops telfairi	33.75	54.65	38.77	16575	20399	1721987023	3862593	3886527	2111597918	22120505	24399014	1.16	1.05
Loxodonta africana	45.7	49.17	40.72	25635	26203	78195493	3995	3994	3118565340	41611076	42090169	1.35	1.33
Procavia capensis	53.97	62.77	43.25	16114	17234	585475382	3575234	3597313	2407864424	23706028	24584621	1.02	0.98
Erinaceus europaeus	33.33	71.43	34.65	14605	15996	1244406726	3725000	3735630	2133134836	20013639	20941736	0.98	0.94
Sorex araneus	41.87	50.86	60.81	13192	14581	1111328353	3244996	3262842	1832864697	17609732	18566919	1.01	0.96
Myotis lucifugus	58.09	57.37	57.26	20719	22432	68155432	559	574	1966419868	31811562	33557343	1.71	1.62
Rhinolophus ferrumequinum	42.01	49.47	63.46	33717	34345	7297869	0	24	2010135660	59014606	82632781	4.11	2.94
Pteropus vampyrus	64.56	46.64	52.48	17053	18086	160177803	1672623	1685574	1839436660	27378438	28179117	1.53	1.49
Equus caballus	39.26	44.44	46.97	44934	45336	3324287	3504	4077	2405818947	88752066	107844115	4.48	3.69
Equus asinus asinus	34.44	52.74	49.47	33772	34012	26104156	18846	20601	2294940189	68028984	72343739	3.15	2.96
Panthera leo	40.44	50.51	40.83	31178	31668	13956588	37380	39807	2146290051	53056231	63233659	2.95	2.47
Panthera pardus	45	53.97	42.77	29823	31328	99153465	37649	38654	2478865742	46320892	62682181	2.53	1.87
Panthera tigris altaica	67.5	52.5	49.9	26848	27821	58232500	454976	481306	2332832693	40808487	54440004	2.33	1.75
Felis catus	40.68	47.62	48.64	40302	40796	45323815	628	629	2414928095	74551614	94925730	3.93	3.09
Canis lupus famil. reference breed	40.26	40.36	51.33	45094	45707	10040013	23167	32388	2317593971	89263981	109619079	4.73	3.85
Canis lupus dingo	56.46	42.11	32.41	34534	35018	15042436	68	68	2424789071	62890217	69189081	2.85	2.59
Vulpes vulpes	48	48.75	56.3	41026	41501	71714148	120872	189774	2349837111	82461290	101778733	4.33	3.51
Ursus americanus	34.83	50.93	35.71	33197	33546	136542060	36803	40513	2451851856	56938591	64891543	2.65	2.32
Ursus maritimus	37	42.39	46.65	34052	34377	38357410	195065	222292	2263021934	62994956	73409895	3.24	2.78
Ailuropoda melanoleuca	40.62	47.62	49.65	39310	39893	36411305	96458	135049	2250425073	69487583	91752400	4.08	3.09
Mustela putorius furo	36	53.16	51.41	20062	20349	132851443	18874	148409	2277906570	32200188	52923512	2.32	1.41
Neovision vison	54.04	39.81	62.79	31888	32325	173022569	228961	242253	2274181564	53450482	62556800	2.75	2.35
Tursiops truncatus	43.45	54.87	42.37	16611	17536	223463458	1456369	1471408	2298460478	27127090	27875071	1.21	1.18
Delphinapterus leucas	35.42	39.06	60.83	30992	31476	35864186	2347	3944	2326909482	61555576	75959384	3.26	2.65
Monodon monoceros	41.4	56.38	59.9	30379	31102	11468950	14746	17733	2344106029	51692186	67101259	2.86	2.21
Physeter catodon	40.43	47.5	52.49	32206	32966	82107716	26986	30672	2287974538	53901178	69228178	3.03	2.36
Phocoena sinus	40.23	44.01	53.68	31556	32301	8038057	0	0	2361504502	52989461	82663831	3.5	2.24
Balaenoptera musculus	38.32	50	39.22	31354	32176	25925573	2247	2247	2349072746	1297760	70412338	3	0.06
Camelus dromedarius	39.05	39.52	65.64	33007	33669	8484190	13726	19170	2044274219	57222665	80238879	3.93	2.8
Vicugna pacos	46.31	50.38	43.16	11806	12704		2525574	2535794	1922927087	16548597	17186095	0.89	0.86
Sus scrofa	43.21	41.21	47.15	49859	51485	29848884	0	109	2405413179	87985489	179897460	7.48	3.66
Catagonus wagneri	60.5	57.24	55.41	29507	30035	47532400	37278	39679	2592535414	48787960	60000602	2.31	1.88
Capra hircus	40.28	41.68	42.53	33914	34136	31608	63	129	2466159745	61847347	82457230	3.34	2.51
Ovis aries	40.28	53.98	40.64	22823	23113	80755505	20095	58368	2506751578	36423291	45677457	1.82	1.45
Bos taurus	40.36	54.3	43.51	37538	38030	23710	0	0	2628371213	68258226	93161297	3.54	2.6
Bos grunniens	42.99	43.13	67.71	43091	44299	64600	200	279	2823125370	71739924	93642550	3.32	2.54
Bos mutus	55	37.75	63.08	34678	35264	120154638	469898	525743	2525007273	58076918	70992784	2.81	2.3
Bison bison bison	40.5	46.15	51.56	28733	29403	195767988	543199	575600	2757854331	43676324	44212733	1.6	1.58
Moschus moschiferus	50	48.65	60.93	33292	33985	105230400	40489	42959	2964378037	54927363	73371683	2.48	1.85
Otolemur garnettii	35.9	43.29	46.4	19986	21289	160194097	1951	2034	2359210396	32354805	34765588	1.47	1.37
Microcebus murinus	39.12	43.19	39.03	38078	38548	89972405	2168	2504	2349741112	64504186	95954884	4.08	2.75
Propithecus coquereli	35.5	42.8	58.68	32202	32618	714387616	12468	12487	2083764525	51976442	61632871	2.96	2.49
Prolemur simus	46	44.44	52.35	38056	38511	178233908	36014	38429	2233359768	60662101	61073483	2.73	2.72
Callithrix jacchus	43.2	56.67	63.26	43000	44190	9511100	27327	42541	2835864148	72933067	85923986	3.03	2.57
Saimiri boliviensis boliviensis	33.6	42.04	48.01	40695	41134	131440969	28415	39864	2477147568	68302980	99120592	4	2.76
Cebus capucinus	32.86	44.6	47.18	40677	41227	107203577	7650	14355	2610499094	66097527	94465113	3.62	2.53
Aotus nancymaae	51.61	41.13	36.72	42510	42906	147245327	12053	14544	2714423021	70223138	87475471	3.22	2.59
Cercocebus atys	42.86	49.54	48.14	46067	46607	60973514	2628	4340	2787272842	75897388	114460269	4.11	2.72
Macaca fascicularis	39.45	56.13	42.42	49919	51376	36335957	8818	10307	2789706104	81383383	130913688	4.69	2.92
Macaca mulatta	39.6	44.19	45.75	48770	49537	32449625	979	979	2821526716	92267743	139496838	4.94	3.27
Macaca nemestrina	31.16	47.21	38.38	46238	46822	110217031	2380	4974	2838486480	75543870	113857000	4.01	2.66
Papio anubis	39.29	41.31	43.95	45181	45604	9488840	16215	18603	2727993489	73149520	84020935	3.08	2.68
Mandrillus leucophaeus	33.88	63.73	45.76	40903	41317	340585302	17792	17946	2721424085	64076262	83989332	3.09	2.35
Chlorocebus sabaeus	39.14	67.3	59.5	19255	19830	36567272	487	3962	2707548039	31384057	50724581	1.87	1.16
Rhinopithecus bieti	60	52.24	70.01	43730	44293	175194759	7695	9716	2801880187	68182786	97479651	3.48	2.43
Rhinopithecus roxellana	51.03	39.81	47.15	45897	46545	43508003	75101	82074	2856044136	74109074	106031556	3.71	2.59
Pongo abelii	39.27	66.21	42.42	21414	22451	306522543	8159	49393	2722968486	32256837	52815334	1.94	1.18
Gorilla gorilla gorilla	39.38	40.11	44.7	45194	45716	145998322	21237	24124	2788268060	70344475	94607518	3.39	2.52
Pan paniscus	39.24	54.49	47.45	43232	43781	469504276	29737	34270	2682402951	68488732	93484270	3.49	2.55
Pan troglodytes	40.28	48.88	53.7	49949	50434	96495296	5985	6348	2870613227	79771423	116637852	4.06	2.78
Homo sapiens	37.01	45.45	48.9	115262	196722	153901651	111568	100005	2945849067	132548168	345102391	11.71	4.5
Nomascus leucogenys	39.38	56.83	58.96	40527	41094	183586894	71804	72288	2611673151	62299050	80752682	3.09	2.39
Carlito syrichta	48.02	36.36	39.36	31791	32328	48109210	252595	259636	3405755564	48213484	56428880	1.66	1.42
Tupaia belangeri	39.3	40.97	43.6	15475	17788	1533099162	4003076	4032490	2137242230	20473324	21916499	1.03	0.96
Ochotona princeps	36.17	55.69	47.1	16006	17333	2020574260	3090149	3107037	2761329543	23425776	24357605	0.88	0.85
Oryctolagus cuniculus	40.72	40.24	43.02	41055	41711	107161208	36468	64891	2140590896	73750978	92045055	4.3	3.45
Cavia porcellus	42.8	42.92	65.73	25582	25824	59849908	5355	9533	2663369733	43333286	68124325	2.56	1.63
Octodon degus	36.5	41.54	66.04	27047	27628	469628236	15755	15945	2526244269	42137549	45635482	1.81	1.67
Heterocephalus glaber F	35.9	53	56.97	28984	29620	303433536	6227	18732	2314771103	48312952	75631930	3.27	2.09
Chinchilla lanigera	52.6	43.5	73.82	25983	26265	106575991	13574	33268	2284292980	45719399	76213471	3.34	2
Sciurus vulgaris	38.73	33.33	53.93	33234	34421	36299057	0	0	2632067772	52365848	63645292	2.42	1.99
Marmota marmota marmota	56.61	42.92	45.65	27615	28206	98898735	76235	84297	2411688644	44164811	48618060	2.02	1.83
Urocitellus parryii	42.63	39.3	61.5	30023	30456	264570837	40448	43657	2253482840	52180063	65045188	2.89	2.32
Ictidomys tridecemlineatus	48.1	51.57	50	25958	26267	167333470	11450	17968	2311076758	44683051	70949856	3.07	1.93
Dipodomys ordii	66	46.09	71.62	23962	24276	171054779	10868	10979	2065314044	39475984	48114955	2.33	1.91
Jaculus jaculus	41.1	41.1	53.04	25180	25501	364982898	11797	11890	2470267327	41630765	44875561	1.82	1.69
Rattus norvegicus	38.95	53.23	61.21	29107	31715	130297421	1901	2710	2651715181	45748008	68052938	2.57	1.73
Mus musculus	61.84	40.91	39.83	67174	117497	73600668	6457	0	2654621783	85631176	215699291	8.13	3.23
Mus spicilegus	46.5	38.1	51.33	37002	37747	105748500	40975	58511	2390335020	63459346	82458342	3.45	2.65
Cricetulus griseus (CHOK1GSHD)	44.2	46.12	57.42	28327	28433	34242448	5207	8357	2323924942	48259665	59760926	2.57	2.08
Mesocricetus auratus	46.4	43.97	56.67	25910	26216	428748785	10381	10411	2076176254	41663629	55293981	2.66	2.01
Microtus ochrogaster	39.28	53.66	59.14	27611	28140	112230976	7612	7705	1543152531	45953734	59819155	3.88	2.98
Peromyscus maniculatus bairdii	39.13	46.55	59.13	28171	28891	105923603	195142	210878	2330520439	4724010	45291484	1.94	0.2
Nannospalax galili	39	44.62	44.4	26872	27238	147976214	8994	8994	2913431996	44177206	58178212	2	1.52