UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

THIERES TAYRONI MARTINS DA SILVA

AVALIAÇÃO DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS EM GENES VARIÁVEIS EM TETRAPODA E SUAS LINHAGENS

BELO HORIZONTE
2023

THIERES TAYRONI MARTINS DA SILVA

AVALIAÇÃO DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS EM GENES VARIÁVEIS EM TETRAPODA E SUAS LINHAGENS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr Francisco Pereira Lobo

Belo Horizonte
2023

043 Silva, Thieres Tayroni Martins da.
Avaliação de processos biológicos enriquecidos em genes variáveis em Tetrapoda e suas linhagens [manuscrito] / Thieres Tayroni Martins da Silva. – 2023.
110 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Francisco Pereira Lobo.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Anfíbios. 3. Répteis. 4. Aves. 5. Mamíferos. 6. Genômica. 7. Filogenia. I. Lobo, Francisco Pereira. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

FOLHA DE APROVAÇÃO

"Avaliação de processos biológicos enriquecidos em genes variáveis em Tetrapoda e suas linhagens"

Thieres Tayroni Martins da Silva

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Francisco Pereira Lobo - Orientador
UFMG

Prof. Aristóteles Góes Neto
UFMG

Prof. Alexandre Liparini Campos
UFMG

Belo Horizonte, 05 de outubro de 2023.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Coordenador(a) de curso de pós-graduação**, em 05/10/2023, às 16:08, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alexandre Liparini Campos**, **Professor do Magistério Superior**, em 05/10/2023, às 16:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 16/10/2023, às 12:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2665568** e o código CRC **F9D6DA2D**.

Referência: Processo nº 23072.259227/2023-16 SEI nº 2665568

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

THIERES TAYRONI MARTINS DA SILVA

Às quatorze horas do dia **05 de outubro de 2023**, reuniu-se, no aplicativo Zoom, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Avaliação de processos biológicos enriquecidos em genes variáveis em Tetrapoda e suas linhagens"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Francisco Pereira Lobo**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	**Instituição**	**Indicação**
Dr. Francisco Pereira Lobo	UFMG	**Aprovado**
Dr. Aristóteles Góes Neto	UFMG	**Aprovado**
Dr. Alexandre Liparini Campos	UFMG	**Aprovado**

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 05 de outubro de 2023.

Documento assinado eletronicamente por **Aristoteles Goes Neto**, **Coordenador(a) de curso de pós-graduação**, em 05/10/2023, às 16:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alexandre Liparini Campos**, **Professor do Magistério Superior**, em 05/10/2023, às 16:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 16/10/2023, às 12:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2665529** e o código CRC **4BDB85F9**.

Referência: Processo nº 23072.259227/2023-16 SEI nº 2665529

Dedico a todos os que acreditam que formas infinitas, as mais belas e as mais maravilhosas, têm sido, e estão sendo, evoluídas.

AGRADECIMENTOS

Este trabalho nunca chegaria a vós leitores se não fosse pela ajuda, direta ou indireta, de diversas pessoas. Agradeço profundamente a todos os que contribuíram para tornar esta jornada possível, seja através de suas orientações, apoio emocional, discussões inspiradoras ou simplesmente por estarem ao meu lado durante esse desafio.

Primeiramente gostaria de agradecer ao meu orientador, Francisco Pereira Lobo, que nos últimos 6 anos vem me ensinando como ser um bioinformata. Obrigado por me aceitar como membro do seu grupo, por me permitir participar de diversos projetos, onde pude amadurecer como pesquisador e cientista, por me mostrar que os erros fazem parte do processo e não deixar que eles me abalassem.

Agradeço ao meu grande amor, Bianca, por estar comigo durante estes dois anos de mestrado, por me apoiar e incentivar, acreditando no meu potencial muitas vezes mais do que eu mesmo acreditei.

Aos meus pais, Evaldo e Maria, muito obrigado por me incentivarem a me manter nos estudos, mesmo no momento em que tive que mudar de cidade e ficar longe de vocês nunca me senti distante, pois vocês estavam comigo há todo momento, me apoiando em cada decisão.

A minha tia Rosilene, que me acolheu em sua casa quando me mudei para Belo Horizonte, e que cuidou de mim como um filho. Infelizmente você não pode estar comigo neste momento especial, mas sei que estaria muito orgulhosa.

Ao meu irmão mais velho Bráulio, que serviu de modelo para mim, que me ajudou com meus estudos quando mais novo, e que me incentivou a me manter na pós-graduação e seguir nesta carreira científica. Obrigado por acreditar no meu potencial e sempre deixar isso claro. E a minha irmã mais nova Eduarda, que tento cada dia mais ser uma pessoa melhor para servir de modelo para você.

Ao meu amigo, Leonardo, pelo grande apoio durante minha graduação, por seu mentor Linux, e por me receber na sua casa durante a pandemia, sendo o local onde fiz a seleção e dei início ao meu mestrado.

A todos os atuais e antigos membros do LAB, Agnello, Aline, Alison, Amanda, Dalbert, Diogo, Giovanni, Henry, Igor, Maycon, Raul (Anderson), Thais e Zandora. Obrigado por cada ajuda, cada conversa, por compartilharem seus conhecimentos e tornar o LAB um grupo maravilhoso de fazer parte.

Aos secretários do Programa de Pós-Graduação em Bioinformática, Tiago e Sheila, por sempre resolverem todas as demandas e dúvidas que tive durante o mestrado, sempre com prontidão e clareza.

Aos professores que compartilharam seus conhecimentos nas disciplinas, que foram essenciais para minha formação.

Todos vocês contribuíram para que eu chegasse a este momento, portanto, muito obrigado!

"*De pequenas sementes nascem e crescem as maiores árvores, que as sementes da sua mente sejam igualmente férteis*."

(Visionário Élfico, *Magic The Gathering*, 2012)

RESUMO

Os tetrápodes compreendem um importante grupo de animais com distribuição global, ocupando diferentes nichos e desempenhando diversos papéis biológicos. Possuem diversas adaptações, como membros modificados que proporcionam formas distintas de locomoção, juntamente com órgãos sensoriais complexos. Embora se espere que a maioria dos genes homólogos seja conservada entre as espécies devido à seleção purificadora, os genes envolvidos nos processos de adaptação estão frequentemente sujeitos à seleção diversificadora, o que favorece mutações que alteram a função do gene. Neste trabalho, utilizamos uma nova abordagem de genômica comparativa para detectar os processos biológicos enriquecidos em genes variáveis, utilizando tetrápodes como estudo de caso, para obter uma melhor compreensão dos mecanismos moleculares envolvidos em sua evolução fenotípica. Para tal, baixamos todos os genomas completos de tetrápodes disponíveis no *NCBI RefSeq* e construímos proteomas não redundantes selecionando a isoforma mais longa de cada gene codificador de proteína. Usamos o BUSCO para avaliar a integridade do proteoma e selecionamos somente aqueles com completude acima de 80%. Prosseguimos usando o OrthoFinder para estabelecer relações de homologia entre as sequências proteicas. Usamos MAFFT para alinhamento múltiplo de sequências, seguido por trimAl para calcular a identidade de alinhamento de cada grupo de homólogos. Avaliamos apenas grupos com pelo menos um gene da espécie modelo *Homo sapiens* ou *Gallus gallus* para explorar a anotação de genes individuais e realizar análises de enriquecimento usando WebGestalt. Como esperado, observamos que várias funções *housekeeping* foram significativamente enriquecidas em genes conservados. Entre os conjuntos significativamente enriquecidos em genes variáveis, podemos observar diversos destes relacionados com processos imunes, percepção sensorial e componentes do citoesqueleto. Concluímos que nossa metodologia foi capaz de encontrar categorias enriquecidas em genes variáveis compatíveis com análises de seleção positiva, como é o caso dos genes imunes, além de genes com possível evolução convergente. Adicionalmente, mecanismos de percepção sensorial parecem ter desempenhado um papel crucial na diversificação dos tetrápodes.

Palavras chave: Bioinformática; Genômica comparativa; Evolução; Tetrápodes.

ABSTRACT

Tetrapods are an important group of metazoans with a global distribution, occupying different niches and playing various biological roles. They possess various adaptations such as modified limbs that provide distinct forms of locomotion along with complex sensory organs. While most homologous genes are expected to be conserved among species due to purifying selection, genes involved in adaptation processes are often subject to diversifying selection, favoring mutations that alter gene function. In this study, we employed a novel approach in comparative genomics to detect enriched biological processes in variable genes using tetrapod species as a case study, to gain a better understanding of the molecular mechanisms involved in their phenotypic evolution. We downloaded all available complete tetrapod genomes from the NCBI RefSeq and constructed non-redundant proteomes by selecting the longest isoform for each protein-coding gene. We used BUSCO to assess proteome completeness and selected only those with completeness above 80%. We proceeded to use OrthoFinder to establish homology relationships among protein sequences. We utilized MAFFT for multiple sequence alignment, followed by trimAl to calculate alignment identity for each group of homologs. We evaluated only groups containing at least one gene from the model species *Homo sapien*s or *Gallus gallus* to explore individual gene annotation and perform enrichment analyses using WebGestalt. As expected, we observed that several housekeeping functions were significantly enriched in conserved genes among the analyzed species. Among the significantly enriched sets of variable genes, we identified various terms and genes related to immune processes, sensory perception, and cytoskeletal components. We conclude that our methodology was capable of identifying enriched categories in variable genes consistent with positive selection analyses, as seen in immune genes, as well as genes potentially undergoing convergent evolution. Additionally, mechanisms of sensory perception appear to have played a crucial role in the diversification of tetrapods.

Keywords: Bioinformatics; Comparative genomics; Evolution; Tetrapods.

LISTA DE ILUSTRAÇÕES

LISTA DE TERMOS E ABREVIATURAS

BUSCO – Do inglês “Benchmarking Universal Single-Copy Orthologs” (“Comparação de Ortólogos Universais de Cópia Única”), programa de computador utilizado na avaliação da qualidade de montagem e anotação de genomas.

CentOS – Distribuição gratuita do sistema operacional Linux baseada na distribuição proprietária RedHat.

DNA – do ingles "deoxyribonucleic acid" (“ácido desoxirribonucleico”).

e.g. – Do latim “exempli gratia” (“por exemplo”).

EE – Escore de Enriquecimento. Um dos resultados da análise de enriquecimento de conjunto de genes.

EEN – Escore de Enriquecimento Normalizado. Um dos resultados da análise de enriquecimento de conjunto de genes.

ENHYDRA –

FDR – Do inglês “False Discovery Rate” (“Taxa de Descoberta Falsa”), proporção de resultados positivos identificados incorretamente como significativos.

GB – Gigabytes, 1024 megabytes em contagem binária. Medida de capacidade de armazenamento e memória computacional.

GO – Do inglês "Gene Ontology", estrutura padronizada para anotação e categorização de genes e proteínas.

GSEA – Do inglês "Gene Set Enrichment Analysis" (“Analise de enriqueicmento de conjuntos de genes”).

Housekeeping – conjunto de proteínas essenciais na manutenção das funções celulares básicas e vitais.

ID – Identificador.

KEGG - Do inglês "Kyoto Encyclopedia of Genes and Genomes", banco de dados de representação e interpretação de vias metabólicas, genes, proteínas e doenças.

NCBI – Do inglês “National Center for Biotechnology Information” (Centro Nacional de Informação Biotecnológica).

NGS – Do inglês "Next Generation Sequence", método de sequenciamento automatizado, paralelo e de alto rendimento.

Pathways maps – mapas de vias ou rotas metabólicas que representam visualmente as séries de reações químicas e processos biológicos que ocorrem em um organismo ou sistema biológico específico.

Perl – (Practical Extraction and Reporting Language) Linguagem de programação com especial capacidade de manipulação de texto.

Pipeline – Cadeia de etapas de processamento onde a saída de cada etapa é a entrada do próxima.

RefSeq – Banco de dados de sequências de referência do NCBI.

Script – Arquivo de instruções computacionais escrito em alguma linguagem de computação e comumente utilizado para execução de tarefas curtas.

SUMÁRIO

1 - INTRODUÇÃO

Os tetrápodes (animais terrestres com quatro membros e seus descendentes) constituem uma superclasse que teve origem na linhagem dos Sarcopterygii, peixes com barbatanas lobadas, no final do Devoniano, há 360 milhões de anos. Possuem grande importância ecológica, atuando como polinizadores (Ratto *et al.,* 2018; De-Oliveira-Nogueira *et al.,* 2023) e dispersores de sementes (Tiffney, 2004), além de uma grande importância agropecuária, servindo como controle natural de pragas, como fonte de alimentação e produtos para a utilização humana (*e.g.* carne, ovos, couro).

1.1 - Tetrapoda

Os tetrápodes são comumente classificados em diversos grupos com diferentes origens evolutivas. Dentre os grupos classicamente reconhecidos, destacam-se: 1) anfíbios (classe Amphibia), onde estão as ordens Anura (sapos, pererecas e rãs), Gymnophiona (cecílias) e Urodela (salamandras); 2) répteis (classe Reptilia), que formam um grupo parafilético contendo os Testudines (tartarugas), Lepidosauria (lagartos, serpentes e tuataras) e Crocodilia (crocodilos e jacarés); 3) aves (classe Aves); e 4) mamíferos (classe Mammalia). Possuem distribuição global, ocupando diferentes nichos e desempenhando diversos papéis biológicos. Durante o curso da evolução dos tetrápodes, diversas adaptações foram selecionadas e permitiram a ocupação e sobrevivência no meio terrestre, tais como um esqueleto interno robusto para suportar seu peso em terra, bem como diversas modificações em seus membros, os quais evoluíram para estruturas especializadas com dedos e articulações que lhes permitem andar, correr, escalar, nadar ou voar; destaca-se também a evolução de órgãos sensoriais complexos, como olhos especializados e ouvidos como mecanismos para a percepção e resposta ao ambiente (Pough, 2008).

1.2 - Análises comparativas em Tetrapoda

Nos últimos anos, houve um aumento significativo no número de genomas de alta qualidade de tetrápodes disponíveis em bancos de dados públicos. Enquanto em 2014 existia apenas um genoma disponível de anfíbios (Sun *et al.,* 2015), hoje esse número já supera a marca de 67 genomas de referência no NCBI. Destes, 11 possuindo anotações de referência, tais como a predição de regiões codificadoras, produzidas pelo RefSeq. O aumento no número de genomas pode ser explicado

pelas tecnologias de NGS (*next generation sequencing*, também conhecidas como sequenciamento massivamente paralelo de DNA), que reduziram o custo de sequenciamento de DNA e produziram uma imensa quantidade de genomas de alta qualidade em um curto período (Shendure & Aiden, 2012), além de consórcios de sequenciamento como o *Birds10K* (Zhang, 2015) e o *Vertebrate Genome Project* (Koepfli *et al.*, 2015), que têm contribuído particularmente para o aumento do número de genomas de alta qualidade disponíveis.

Uma abordagem para extrair informações biologicamente relevantes do crescente número de genomas disponíveis é a genômica comparativa. A genômica comparativa compreende o ramo da bioinformática que visa extrair conhecimento biológico a partir dos padrões de conservação e variação dos diferentes elementos estruturais e funcionais que compõem os genomas de organismos. Em tetrápodes, estudos de genômica comparativa de tetrápodes já contribuíram para elucidar diversas questões evolutivas e funcionais, tais como a origem e evolução dos pulmões (Bi *et al.*, 2021), e a busca por genes associados ao processo de aprendizado vocal e à dieta (Zhang *et al.,* 2014) (Kosiol *et al., 2008*).

Um método tradicional da genômica comparativa consiste na análise dos padrões de conservação e variação de genes codificadores homólogos, ou seja, de genes que codificam proteínas e que evoluíram de um gene ancestral comum. Adicionalmente, genes homólogos usualmente compartilham similaridade de sequência e função. Estes genes podem ser encontrados nos genomas de diferentes espécies, e sua homologia pode ser determinada por comparação de sequências em um arcabouço evolutivo (Liu *et al.*, 2004; Miller *et al.*, 2004). Assim, genes são comumente anotados em função do grupo de homólogos ao qual eles pertencem. Do ponto de vista computacional, esse processo consiste na atribuição de um identificador único para cada grupo de homólogo. Este identificador é, então, associado a cada gene pertencente ao grupo, de modo a indicar sua pertinência ao conjunto definido pelo identificador.

Os genes homólogos fornecem diversas informações sobre as relações evolutivas e a conservação funcional dos mesmos entre as espécies. A presença de genes homólogos compartilhados entre genomas de múltiplas espécies é um forte indicativo da sua importância para a evolução fenotípica destes organismos. Adicionalmente, espera-se que vasta maioria dos genes homólogos apresentem pressão seletiva para sua conservação. Mutações não-sinônimas, que ocasionam a

troca do aminoácido codificado, usualmente diminuem a funcionalidade da proteína codificada. Consequentemente, tais alelos são removidos das populações por seleção purificadora (Yang, 2007).

Entretanto, uma minoria dos genes homólogos pode apresentar pressão seletiva para sua variação, apresentando uma taxa evolutiva consideravelmente maior do que o esperado (Hongo *et al.*, 2015). Nesse cenário, considera-se que a troca do aminoácido codificado aumenta a aptidão evolutiva dos indivíduos que possuem o alelo mutante. Consequentemente, a frequência desse alelo tende a aumentar nas gerações subsequentes.

Um exemplo clássico de genes com evidência de seleção positiva são as proteínas imunes de hospedeiros e as proteínas de virulência do parasita, as quais apresentam altas taxas evolutivas como consequência da sua coevolução. Genes de parasitos que sofram mutações que eventualmente permitam explorar os recursos dos hospedeiros de maneira mais eficiente são selecionadas positivamente. Entretanto, estas mutações se tornam uma pressão seletiva nos genes de hospedeiros. Hospedeiros que possuam mutações em genes que eventualmente conferem resistência à infecção do parasito também terão sua aptidão evolutiva aumentada. Consequentemente, os genes de parasitas e hospedeiros encontram-se em uma corrida armamentista , constituindo um exemplo clássico da Teoria da Rainha Vermelha, "*É preciso correr o máximo possível, para permanecermos no mesmo lugar*" (Van Valen, 1974; Carrol, 1980).

Assim, a busca por genes que apresentem altas taxas evolutivas compreende uma importante estratégia para a identificação de genes responsáveis por processos adaptativos em diferentes grupos de organismos. Usualmente, buscas dessa natureza envolvem a utilização de modelos estatísticos sofisticados para a busca por seleção positiva Darwiniana, definida formalmente como a busca por grupos de homólogos com uma taxa de substituição não-sinônimas significativamente maior que a taxa de substituições sinônimas (Hongo *et al.*, 2015).

1.3 - Detecção de convergências moleculares funcionais na genômica comparativa

Uma outra vertente da genômica comparativa visa avaliar a ocorrência de convergências moleculares funcionais, definida como a ocorrência de genes não-homólogos realizando a mesma função molecular (Hongo *et al.,* 2023). Para tal,

faz-se necessária a utilização de um esquema de anotação de genes que reflita os diferentes papéis biológicos dos genes, tais como suas funções moleculares ou as vias bioquímicas das quais participam. Uma alternativa atraente para anotar os genes funcionalmente consiste na utilização de bancos de dados que sumarizam informações de produtos gênicos em listas de processos biológicos. Nesse tipo de anotação funcional, um determinado gene recebe um identificador que determina a sua contribuição em um determinado processo biológico, independentemente da sua sequência e origem evolutiva. Duas alternativas para realizar tal anotação são os bancos de dados *Gene Ontology* (GO) (Ashburner *et al.*, 2000) e a *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa & Goto, 2000).

O banco de dados Gene Ontology compreende uma estrutura padronizada para anotação e categorização de genes e proteínas, com o propósito de elucidar suas funções biológicas, processos e componentes celulares. Sua concepção teve o intuito de facilitar a interpretação e análise de informações genômicas, especialmente em contextos de investigações em larga escala, como estudos comparativos de genômica, proteômica e transcriptômica.

O GO é estruturado em três principais categorias hierárquicas de termos:

- Componente Celular: Descreve as partes ou estruturas celulares em que uma proteína está localizada ou atua, como núcleo, membrana celular ou ribossomos.
- Função Molecular: Descreve as atividades moleculares das proteínas, como funções enzimáticas, atividades de ligação a moléculas ou outras proteínas, e assim por diante.
- Processo Biológico: Refere-se a sequências de eventos ou etapas que ocorrem em uma célula ou organismo para cumprir uma função específica. Isso pode incluir processos como a divisão celular, sinalização celular ou metabolismo de substâncias.

Cada uma destas categorias é desdobrada em termos mais precisos, estabelecendo um conjunto controlado de vocabulário que representa distintas atividades biológicas, processos e localizações celulares. Adicionalmente, os termos utilizados no GO encontram-se interligados em uma estrutura hierárquica de relações entre os termos: conceitos/termos mais gerais podem possuir um ou mais

conceitos/termos específicos (figura 1), o que viabiliza anotações mais minuciosas e contextualizadas.

Figura 1 - Relação entre termos pai e filhos do Gene Ontology

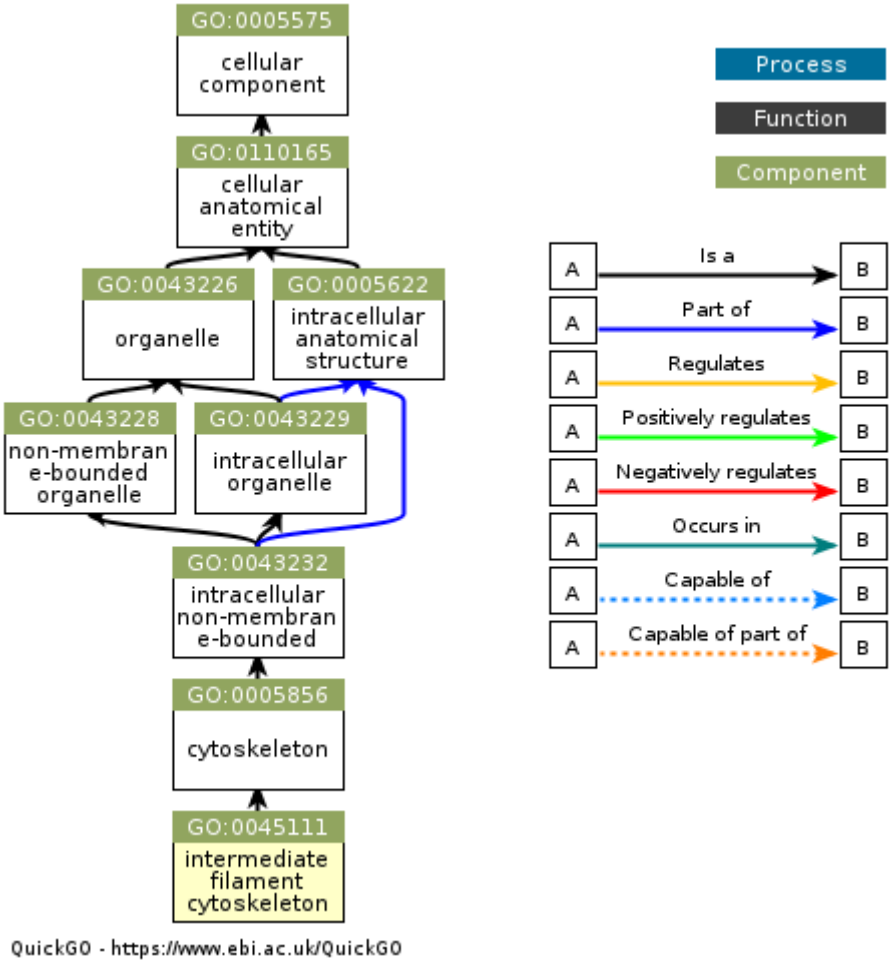


Gráfico de termos do GO:0045111 - Intermediate filament cytoskeleton. O termo GO:0045111 destacado em amarelo é mais específico e as setas indicam que é um termo filho do GO:0005856 - Cytoskeleton, que por sua vez é um termo filho do termo que sua seta aponta, chegando até o termo mais geral do qual faz parte. Um gene anotado com o GO:0045111 é automaticamente anotado como realizando todas as funções biológicas dos termos pais, os quais compreendem conceitos mais gerais que englobam o conceito mais específico. Fonte : https://www.ebi.ac.uk/QuickGO, 2023

Um gene ou produto gênico pode utilizar quantos termos GO forem requeridos, de diversas ontologias, visando de modo a descrever sua função biológica de maneira mais precisa. Adicionalmente, Um único termo do GO pode estar relacionado a múltiplos genes ou proteínas que participam do mesmo processo biológico, sem necessariamente refletir relações de homologia entre eles.

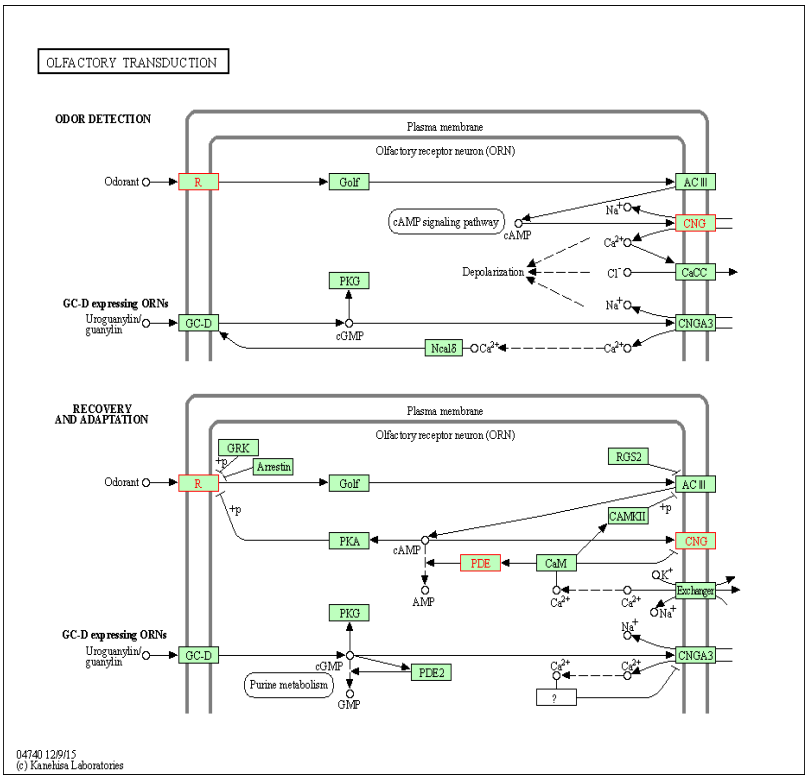
1.4 - Kyoto Encyclopedia of Genes and Genomes (KEGG)

O banco de dados KEGG é composto por várias bases de dados interconectadas, cada uma abrangendo diferentes aspectos da biologia molecular e da genômica funcional. Consideradas conjuntamente, as bases de dados do KEGG fornecem uma visão integrada das diferentes vias bioquímicas já caracterizadas, fornecendo um importante arcabouço conceitual para a modelagem bioquímica de sistemas biológicos. Uma das bases de dados do KEGG é o KEGG *Pathway Database*, uma coleção de *pathways maps* desenhados curados manualmente que representam o conhecimento das redes moleculares de interação, reação e relação para:

- Metabolismo;
- Processamento de Informação Genética;
- Processamento de Informações Ambientais;
- Processos Celulares;
- Sistemas do Organismo;
- Doenças Humanas;
- Desenvolvimento de Medicamentos.

Os *pathways maps* (figura 2) consistem em uma rede de interação/reação molecular representada em termos dos grupos KEGG Orthology (KO), de modo que evidências experimentais em organismos específicos possam ser generalizadas para outros organismos por meio de informações genômicas.

Figura 2 - Mapa de vias do KEGG.



Aqui está representado o mapa da via “Olfactory Transduction”. Os retângulos são os produtos gênicos, os círculos representam moléculas no geral e as setas indicam processos de interação/reação Os retângulos coloridos em verde representam que aquele produto gênico é encontrado na espécie de interesse, que no exemplo é *Homo sapiens. Fonte: https://www.https://www.genome.jp/pathway/hsa04740, 2023*

1.5 - Análise de Enriquecimento de Conjunto de Genes

Análise de Enriquecimento de Conjunto de Genes, do inglês *Gene Set Enrichment Analysis* (GSEA), foi proposta pela primeira vez por Mootha e colaboradores (2003) com a proposta de avaliar funcionalmente, e em um contexto biológico, conjuntos de genes diferencialmente expressos. Nesse procedimento, ao invés de focar as análises em um ou mais genes específicos, busca-se avaliar grupos de genes funcionalmente associados que estejam enriquecidos no início ou fim de uma lista de genes ordenados por algum critério estatístico (*e.g.* valor-p em um teste estatístico de expressão gênica diferencial). Em 2005, a metodologia de GSEA foi publicada como um programa por Subramanian e colaboradores, onde os métodos matemáticos dessa abordagem estatística foram aperfeiçoados e mais detalhados. A metodologia de GSEA é amplamente aceita para fornecer contexto biológico para listas gênicas, e conta com mais de 40.000 citações (09/2023).

A abordagem de GSEA busca por conjuntos de genes que, simultaneamente, estejam enriquecidos nas extremidades de listas ordenadas de genes e que também façam parte de uma lista de genes definidos *a priori* S que compartilham uma mesma função biológica (*e.g.* genes que compartilham o mesmo GO ou fazem parte da mesma via do KEGG). O objetivo do GSEA é determinar se os membros da lista S são distribuídos aleatoriamente ao longo de uma lista ranqueada de genes de interesse L ou se estes são encontrados principalmente no topo ou na base desta lista L (Subramanian *et al.,* 2005).

Para cada conjunto de genes S *(e.g.* cada lista de genes anotados com um determinado termo GO ou via bioquímica KEGG), realiza-se inicialmente um cálculo de escore de enriquecimento (EE), o qual reflete o grau em que um conjunto S é super-representado nos extremos (superior ou inferior) de toda a lista classificada L. A pontuação é calculada percorrendo a lista L, aumentando conforme a seguinte função:

$$P_{\mathrm{hit}}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{onde } N_R = \sum_{g_j \in S} |r_j|^p$$

quando é encontrado um gene em S e diminuindo conforme a seguinte função:

$$P_{\mathrm{miss}}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

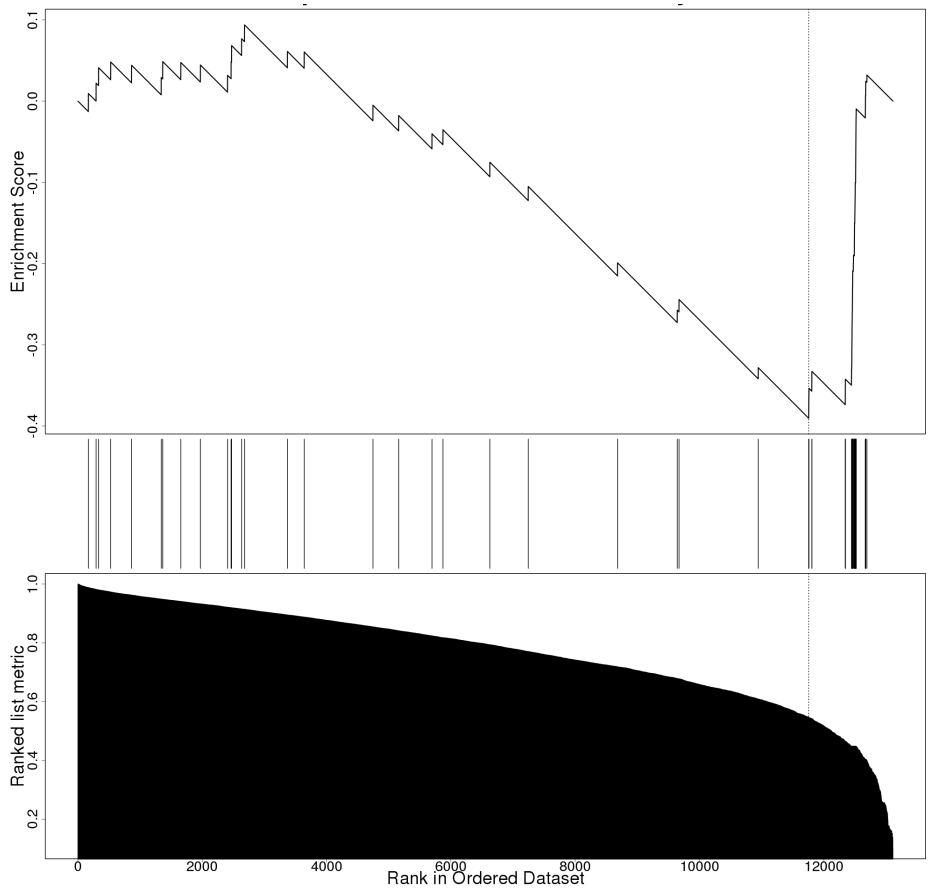
quando são encontrados genes que não estão em S, sendo g_j um gene qualquer na lista ordenada, r_j o valor de g_j na lista ordenada, P o expoente de ponderação de enriquecimento, o valor de N o número de genes na lista L e N_H o número de genes na lista de interesse S. O EE é definido formalmente como o desvio máximo de zero encontrado na caminhada aleatória, dado por P_{hit} - P_{miss}.

Após o cálculo de EE, realiza-se um procedimento de permuta dos rótulos dos genes, onde a pertinência ao conjunto S é aleatorizada ao longo da lista L. Após este procedimento, realiza-se um novo cálculo de EE de modo a gerar uma distribuição nula de EE para a lista S. Usando esta distribuição nula, é possível calcular um valor *p* nominal empírico para o EE observado. Por fim, é feito uma correção para considerar o teste de múltiplas hipóteses da seguinte forma: inicialmente, é feita a normalização do EE para cada conjunto de genes, levando em consideração o tamanho do conjunto, o que gera um escore de enriquecimento

normalizado (EEN). Posteriormente, a proporção de falsos positivos é controlada através do cálculo da taxa de descoberta falsa (*false discovery rate,* FDR) correspondente a cada valor de EEN. O FDR representa a probabilidade estimada de que um conjunto com um EEN específico seja um resultado falso positivo. Essa estimativa é obtida ao comparar as caudas das distribuições observadas e nulas para os valores de EEN.

Os membros que mais contribuem para o EE formam o subconjunto de ponta (Figura 3). O subconjunto de ponta de um conjunto de genes com EE positivo são os genes que aparecem na lista classificada antes do valor da pontuação máxima, enquanto nos grupos com EE negativo são os genes que aparecem após o valor de pontuação máxima.

Figura 3 - Gráfico de enriquecimento de conjunto de genes



Exemplo de uma análise de enriquecimento. Eixo x: lista ordenada de genes (lista L). A parte superior do gráfico mostra o escore de enriquecimento em execução para o conjunto de genes no eixo X à medida que a análise percorre a lista ranqueada L. A pontuação no pico do gráfico (a pontuação mais distante de 0,0) é o EE para o conjunto de genes S. A parte intermediária do gráfico (barras verticais) mostram a localização (*rank*) dos genes que pertencem ao conjunto S em função da sua ocorrência na lista de classificação de genes L. Quanto maior a intensidade da barra, maior a quantidade de

genes encontrados naquele ponto da lista. No caso, observa-se uma concentração destes genes na porção direita do gráfico, próxima a uma das extremidades da lista. A parte inferior do gráfico mostra o valor da métrica de classificação à medida que se percorre a lista de genes ranqueados. O subconjunto de ponta de um conjunto de genes é o conjunto de membros que mais contribuem para o EE. Para um EE positivo, o subconjunto de ponta é o conjunto de membros que aparecem na lista classificada antes da pontuação máxima (linha vertical pontilhada). Para um EE negativo (como o do exemplo), é o conjunto de membros que aparecem após a pontuação máxima. Fonte: Elaborado pelo autor, 2023.

A metodologia de GSEA provê uma interpretação imediata de análises de transcriptoma, uma vez que permite explorar quais são os processos biológicos enriquecidos em listas de genes diferencialmente expressos, definidos como as listas de genes S que encontram-se enriquecidas nas extremidades da lista L.

Entretanto, há uma surpreendente ausência da aplicação da metodologia de GSEA para a análise de outras listas de genes que podem ser produzidas em outras análises ômicas. Nessa dissertação, desenvolvemos uma nova metodologia de genômica comparativa que utiliza GSEA (ENHYDRA), utilizando genomas de Tetrapoda como estudo de caso. Para tal, inicialmente estimamos os grupos de homólogos do táxon de interesse. Em seguida, alinhamos cada grupo de homólogos, e utilizamos o valor de identidade de alinhamento dos diferentes grupos de homólogos fazer produzir uma lista ordenada de genes em função de sua variação biológica, de modo que os extremos dessa lista contém respectivamente os grupos de genes mais conservados e os mais variáveis. Desse modo, pretendemos avaliar quais são as funções biológicas que apresentam genes mais variáveis, indicando a contribuição destas funções para a adaptação de Tetrapoda aos diferentes nichos onde estes são encontrados.

2 - OBJETIVOS

2.1 - Objetivos Gerais

Identificar funções biológicas (termos GO e vias KEGG) significamente enriquecidas em genes variáveis nas linhagens Tetrapoda, Aves e Mammalia.

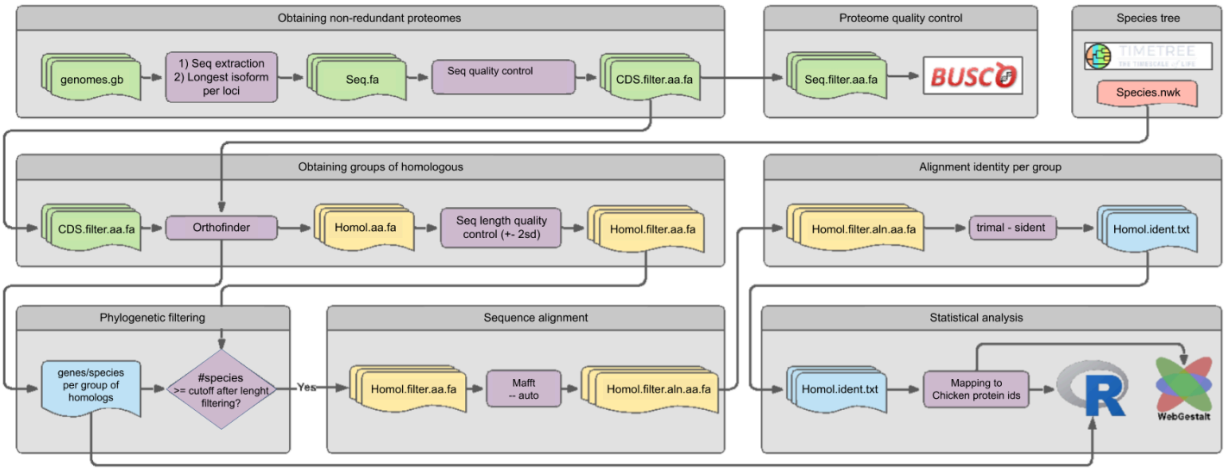
2.2 - Objetivos Específicos

- Obtenção e tratamento de todos os genomas de Tetrapoda disponíveis no banco de dados *Reference Sequence* (RefSeq) do *National Center for Biotechnology Information* (NCBI), a fim de produzir o equivalente aos proteomas não redundantes;
- Avaliação da qualidade de montagem e completude dos proteomas não redundantes;
- Obtenção da árvore filogenética das espécies de Tetrapoda que possuam proteomas de alta qualidade.
- Construção dos grupos de homólogos das espécies de Tetrapoda que possuam proteomas de alta qualidade;
- Desenvolvimento de um programa *python* para o pré-processamento dos dados: 1) filtragem de sequências de baixa qualidade; 2) remoção de grupos de homólogos com baixo suporte/qualidade; 3) alinhamento dos grupos de homólogos restantes; 4) obtenção das métricas de alinhamento; e 5) construção da lista ordenada de genes utilizando os valores de identidade de alinhamento, de maneira a representar os grupos de genes conservados e variáveis nas extremidades.
- Realizar a análise de enriquecimento de conjunto de genes para verificar quais categorias estão enriquecidas nas linhagens de Tetrapoda, Aves e Mammalia.

3 - MATERIAIS E MÉTODOS

A Figura 4 contém a representação esquemática do fluxo de trabalho desenvolvido e utilizado neste projeto.

Figura 4 - Workflow da metodologia.



Os genomas (genomes.gb) foram obtidos do NCBI e submetidos a diversos protocolos computacionais. Inicialmente, obtivemos as maiores regiões codificadoras por lócus, seguido de controle de qualidade e tradução automática, de modo a obtermos proteomas não-redundantes, onde cada gene codificador é representado uma vez (Obtaining non-redundant proteomes). Avaliamos a qualidade dos proteomas não-redundantes utilizando BUSCO para avaliar a completude de cada proteoma (Proteome quality control). Utilizamos o programa OrthoFinder para estabelecer relações de homologia para os proteomas de alta qualidade (Obtaining groups of homologous). Cada grupo foi avaliado para a remoção de sequências parciais e quimeras (Seq length quality control). Usamos o programa MAFFT para o alinhamento múltiplo de cada grupo de homólogos, seguido do programa trimAl para calcular as estatísticas de identidade de cada alinhamento de grupo. A árvore de espécies foi obtida no site TimeTree. Utilizamos a linguagem de programação R para a construção de gráficos de métricas dos grupos. O site WebGestalt foi utilizado para as análises de enriquecimento. Fonte: elaborado pelo autor, 2022.

3.1 - Infraestrutura computacional

As análises computacionais para a obtenção e controle de qualidade dos proteomas não-redundantes foram realizadas em um servidor DELL com 1 processador Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz 12 threads, 32GB de memória RAM e sistema operacional Red Hat Enterprise Linux Server 7.2 (Maipo). As análises de relações de homologia foram realizadas no Servidor SAGARANA em um nó de processamento com 16 processadores Intel(R) Xeon(R) CPU E5-4640 0 @ 2.40GHz totalizando 256 threads, 2048GB de memória RAM compartilhada e sistema operacional CentOS Linux release 7.7.1908. As demais análises foram realizadas em um servidor DELL com 2 processadores Intel(R) Xeon(R) E5-4610 v2 2 @ 2.3GHz totalizando 64 threads, 128GB de memória RAM e sistema operacional CentOS Linux release 7.5.1804.

3.2 - Obtenção dos dados

Para obtenção e tratamento dos dados utilizamos o *script* Perl "calanguize_genomes.pl" (Hongo *et al*., 2023), previamente desenvolvido por mim, e adaptado para obter cópias locais dos arquivos dos genomas, extrair as proteínas descritas, sumarizar as proteínas de maneira a obter somente a maior isoforma por *locus*, realizar filtros para a remoção de sequências de baixa qualidade (sequências com caracteres que não representam os 20 aminoácidos proteinogênicos) e realizar a análise de completude utilizando o *software* BUSCO (Simão *et al.*, 2015). O programa *calanguize_genomes.pl* requer como entrada 1) um arquivo texto contendo os nomes das espécies de interesse para *download*; 2) um arquivo de *assembly summary*, obtido do NCBI descrevendo as informações de disponibilidades de genomas e 3) um arquivo de banco de dados do BUSCO, indicando quais genes serão utilizados no controle de qualidade. Os bancos de dados BUSCO compreendem ortólogos 1:1 conservados quase universalmente no táxon de interesse. Dessa maneira, a avaliação da presença/ausência destes genes em um genoma de interesse fornece um estimador indireto da qualidade da montagem em relação ao seu conteúdo gênico esperado.

A lista de espécies foi obtida utilizando a plataforma do *NCBI Taxonomy* (Schoch *et al.*, 2020), para isto realizamos uma busca, utilizando os seguintes termos: "species[rank] & tetrapoda[org]" e "subspecies[rank] & tetrapoda[org]". Com esta busca obtivemos um arquivo-texto com os nomes de todas as espécies e subespécies de tetrápodes. O arquivo de *assembly summary* escolhido foi o "*assembly summary refseq*", obtido no dia 25 de abril de 2023 em 'ftp.ncbi.nlm.nih.gov/genomes/refseq/assembly_summary_refseq.txt. Este arquivo contém informações sobre os genomas disponíveis no RefSeq (coluna 20), bem como suas espécies de origem (coluna 8). O banco de dados BUSCO foi "tetrapoda_odb10"[1], e compreende 5310 grupos de ortólogos 1-1 presentes em mais de 90% dos genomas de Tetrapoda utilizados para construir este banco.

Calanguize_genomes.pl verifica quais as espécies no arquivo de entrada possuem genoma disponível através do arquivo de *assembly summary*. Para cada espécie que possui genoma disponível, este programa realiza seu *download* no formato *GenBank flat file* (extensão .gbff), e realiza a extração das sequências

[1] Disponível em https://BUSCO-data.ezlab.org/v5/data/lineages/tetrapoda_odb10.2021-02-19.tar.gz

proteicas descritas no campo “translate” em um arquivo no formato fasta, utilizando nesse momento um filtro para remover todas as sequências de baixa qualidade que eventualmente possuam caracteres que não representem os 20 aminoácidos proteinogênicos, tais como “X” que pode representar qualquer aminoácido ou “*” no meio das sequências, que representa um códon de parada interno, que usualmente não são reconhecidos pelos programas de alinhamento múltiplo de sequências. Após esse filtro, este programa realiza uma etapa que seleciona a isoforma que contém a maior região codificadora descrita para cada *locus*, de modo a obter proteomas não-redundantes, definidos como todas as regiões codificadoras descritas na espécie. Essa abordagem visa evitar possíveis vieses introduzidos pelo maior número de isoformas descritas em organismos-modelo (Vogel & Chothia, 2006).

Ao final desta etapa, *calanguize_genomes.pl* retornou um total de 356 proteomas não redundantes de Tetrapoda.

3.3 - Avaliação de completude

Avaliamos os 356 genomas obtidos na etapa anterior quanto à sua qualidade utilizando o *software* BUSCO versão 5.2.2 (Manni *et al.*, 2021). BUSCO é baseado na expectativa evolutiva de que ortólogos de cópia única encontrados em quase todas as espécies de um determinado táxon devem estar presentes e de cópia única em qualquer espécie recém-sequenciada do mesmo clado. Os bancos de dados BUSCO são construídos para várias linhagens taxonômicas, identificando grupos quase universais de ortólogos de cópia única do OrthoDB (Waterhouse *et al.*, 2013). Para que um grupo de homólogos seja considerado um identificador BUSCO, é necessário que ele esteja presente em mais de 90% das espécies da linhagem utilizadas para a construção dos bancos de dados BUSCO.

O programa BUSCO foi executado com a seguinte chamada de sistema: "*$busco_path -i $input -o $id -l $busco_database -c $cpu -m prot*”, sendo os parâmetros definidos como se segue: 1) ‘$busco_path’ é o caminho para o executável busco; 2) ‘*-i $input*’ é um proteoma não-redundante a ser avaliado; 3) ‘-o $id’ indica o diretório para gravar os arquivos de resultados; 4) ‘-l $busco_database’ é o banco de dados BUSCO utilizado na análise(tetrapoda_odb10); 5) ‘-c $cpu’ é o número de cpu’s utilizados na etapa de paralelismo; e 6) ‘-m prot’ é o tipo de sequência utilizada no arquivo de entrada (arquivos de sequências de proteínas).

Como resultado do BUSCO, obtivemos as métricas referentes à porcentagem de completude referente aos identificadores BUSCO totais encontrados (cópia única e duplicados), porcentagem de identificadores BUSCO cópia única, porcentagem de identificadores BUSCO duplicados, porcentagem de identificadores BUSCO fragmentados e a porcentagem de identificadores BUSCO ausentes.

Foram selecionados como de alta qualidade os proteomas que apresentaram um valor de completude BUSCO acima de 80%, além de um máximo de 10% de valor de BUSCOs faltantes. Um total de 315 proteomas de diferentes espécies de Tetrapoda foram selecionados para as análises subsequentes.

3.4 - Obtenção de árvore filogenética de espécies

Obtivemos uma a árvore de espécies para as 315 espécies de Tetrapoda no *site TimeTree of Life* (Hedges *et al.*, 2006, Kumar *et al*., 2022). Como entrada para o *TimeTree*, utilizamos uma lista com as espécies selecionadas na etapa anterior, com exceção da subespécie *Canis lupus dingo,* por já possuir um representante da subespécie (*Canis lupus familiaris*), e da espécie *Harpia harpyja,* por não possuir nenhuma entrada no *site*. Como arquivo de saída, *TimeTree* retorna uma árvore ultramétrica de espécies no formato *newick* (Felsenstein *et al.,* 1986).

3.5 - Construção dos grupos de homólogos

Utilizamos o programa OrthoFinder versão 2.5.4 para a construção dos grupos de homólogos para as 313 espécies de Tetrapoda (Emms & Kelly, 2019). Este programa recebe como entrada um diretório contendo arquivos fasta contendo os proteomas individuais de cada espécie. A construção dos grupos de homólogos utiliza uma árvore de espécies para reconciliar a árvore de genes e produzir resultados levando em consideração a história evolutiva das linhagens (Emms & Kelly, 2017). A inferência dos grupos de homólogos é, portanto, afetada pela árvore de espécies utilizada. Este programa pode fazer a inferência da árvore de espécies automaticamente. Alternativamente, o usuário pode utilizar uma árvore de espécies com uma topologia previamente definida como parâmetro adicional para o programa.

Como resultado, este programa fornece um arquivo associando cada gene a um identificador do grupo de homólogo ao qual o gene pertence, além de um arquivo fasta para cada grupo de homólogos contendo todas as sequências pertencentes ao grupo. Estes arquivos foram utilizados nas análises seguintes.

3.6 - Seleção dos conjuntos de dados para análise

Para nossa análise dos padrões de conservação e variação de sequência em Tetrapoda e suas linhagens, decidimos por analisar os clados Tetrapoda, Mammalia e Aves. Para cada grupo, selecionamos até duas espécies por família, de modo a privilegiar a diversidade natural dos grandes grupos (ao nível de família) e impedir que o excesso de genomas de determinados grupos criasse possíveis vieses na análise. Esta seleção foi feita levando em consideração as espécies que possuem o maior valor de completude BUSCO e menor valor de BUSCO ausentes dentro de cada família. Utilizamos como âncora a espécie *Homo sapiens* para o conjunto de Tetrapoda e Mammalia e *Gallus gallus* para o conjunto de aves, de modo que os identificadores dos genes destas espécies foram utilizados para a produção das listas de genes ordenados em cada análise.

Ao final, selecionamos 198 espécies de tetrápodes, 108 de mamíferos e 50 de aves. Para cada arquivo fasta contendo as sequências de um determinado grupo de homólogos, produzimos arquivos contendo somente as sequências das espécies selecionadas, os quais foram utilizados como entrada para o programa ENHYDRA, cujo funcionamento descrevemos.

3.7 - ENHYDRA - ENriched Homology group anaIYsis Ranked by IDentity of Alignment

Desenvolvemos um *pipeline* Python chamado ENHYDRA (disponível em *https://github.com/ThieresTMS/ENHYDRA*) para produzir as listas de genes ordenados pela sua identidade de alinhamento e verificar quais categorias funcionais estão enriquecidas em genes mais conservados e mais variáveis nas diferentes linhagens em análise. ENHYDRA recebe como entrada um arquivo de configuração onde o usuário define os seguintes parâmetros:

- inputdir: caminho para o diretório de entrada com os arquivos fasta dos grupos de homólogos.
- outdir: caminho para o diretório de saída onde vai ser escrito os resultados
- min_species: número mínimo de espécies nos grupos de homólogos para um grupo ser considerado válido e ser analisado.

- anchor: identificador da espécie de referência escolhida para a análise de enriquecimento.
- max_process: número máximo de processadores.

Nosso *pipeline* começa as análises utilizando um filtro de tamanho de sequência onde removem-se as sequências que possuam um tamanho menor ou maior do que dois desvios padrão do tamanho médio da sequência do grupo pré-alinhamento. Essa etapa visa remover sequências particularmente pequenas ou grandes dos grupos de homólogos, as quais comumente representam fragmentos de genes, quimeras e outros possíveis artefatos de sequência produzidos durante as etapas de montagem e anotação. Assim, pretendemos evitar possíveis vieses a jusante na estimativa da identidade média das sequências por grupo causadas por sequências de baixa qualidade.

ENHYDRA prossegue utilizando um filtro para remover todos os grupos de homólogos que não possuam sequências no genoma da espécie de referência, ou que não estejam presentes em um número mínimo suficiente de espécies satisfaça o valor estipulado pelo usuário no campo “min_species”. Esta etapa remove grupos que ocorram em poucos genomas, os quais não representariam a maior parte dos genomas do táxon em análise. Adicionalmente, esta etapa também remove genes que não possuam evidência de existência no genoma âncora, uma vez que estes não poderiam ser utilizados em análises de enriquecimento.

Em seguida, ENHYDRA realiza o alinhamento múltiplo de sequências de cada grupo de homólogo utilizando o programa MAFFT (Katoh & Standley, 2013) com a opção –auto, de modo a detectar o melhor modelo de alinhamento para cada grupo, seguido da ferramenta trimAl (Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009) com a opção *-sident* para calcular as estatísticas de identidade de alinhamento para cada grupo. De posse dos valores de identidade de alinhamento por grupo, ENHYDRA construi uma tabela com duas colunas, onde a primeira coluna é o identificador da sequência do grupo de homologia referente a espécie de referência (genoma âncora) e a segunda coluna é o valor da identidade do alinhamento do grupo de homólogo ao qual a sequência pertence, ordenada do maior valor de identidade de alinhamento até o menor valor. Utilizamos as listas ordenadas produzidas para os três *taxa* em análise para análises de enriquecimento dos conjuntos de genes variáveis e conservados utilizando a plataforma *web*

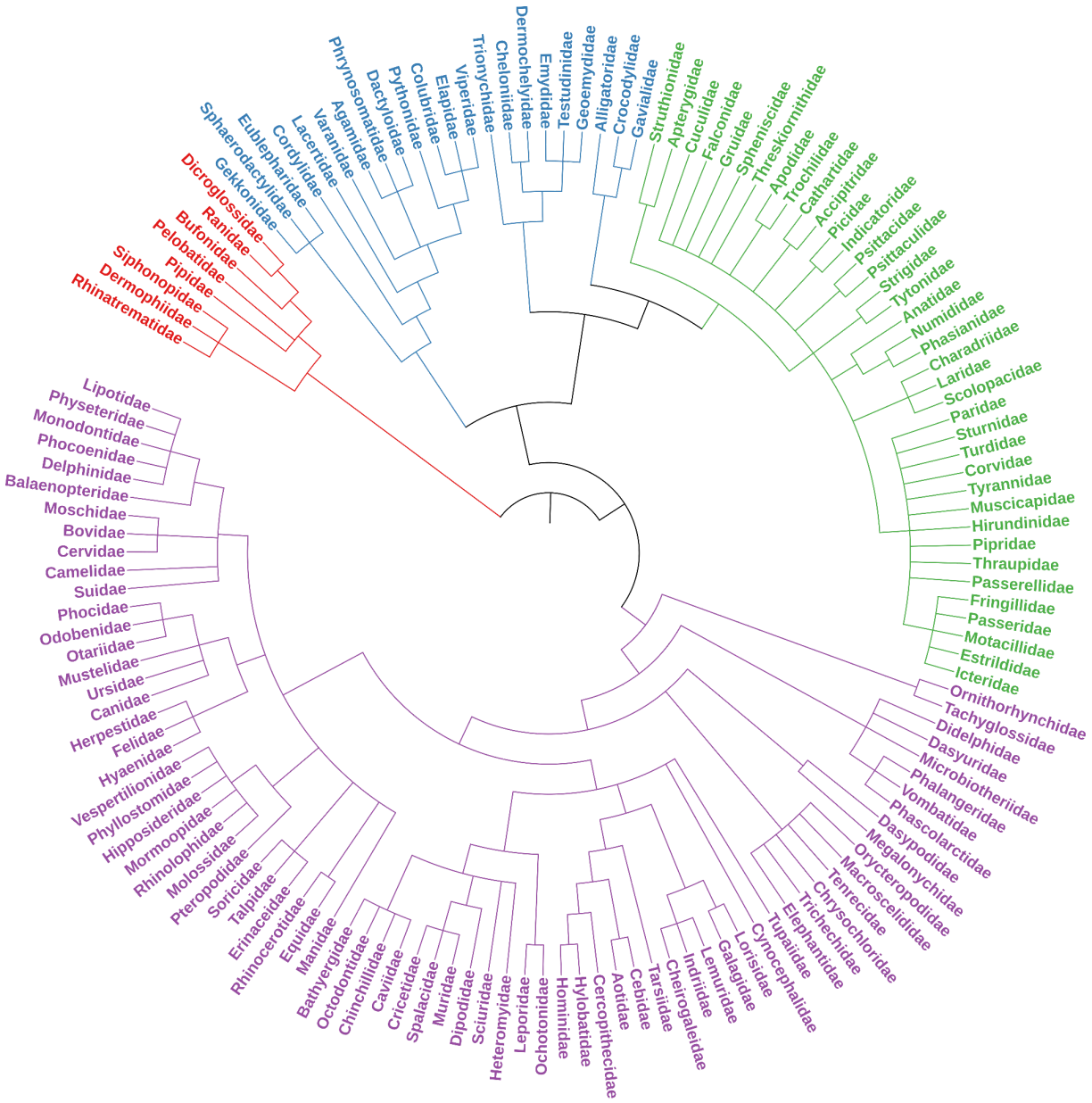
WebGestalt (Liao *et al.,* 2019) para os bancos de dados do Gene Ontology e das vias bioquímicas do KEGG.

4 - RESULTADOS

4.1 - Genomas escolhidos para a análise

Após as análises de qualidade dos proteomas não-redundantes via BUSCO seguida da curadoria manual para selecionar, dentro de cada família de Tetrapoda, as duas espécies com maior valor BUSCO. A árvore filogenética representando as famílias de Tetrapoda selecionadas pode ser visualizada na Figura 5.

Figura 5 - Árvore filogenética das famílias de Tetrapoda selecionadas para análise



Árvore filogenética representando as famílias selecionadas para análise. Vermelho: famílias de Amphibia; Azul: famílias de Reptilia; Verde: famílias de Aves; Roxo: famílias de Mammalia. Fonte: elaborado pelo autor, 2022.

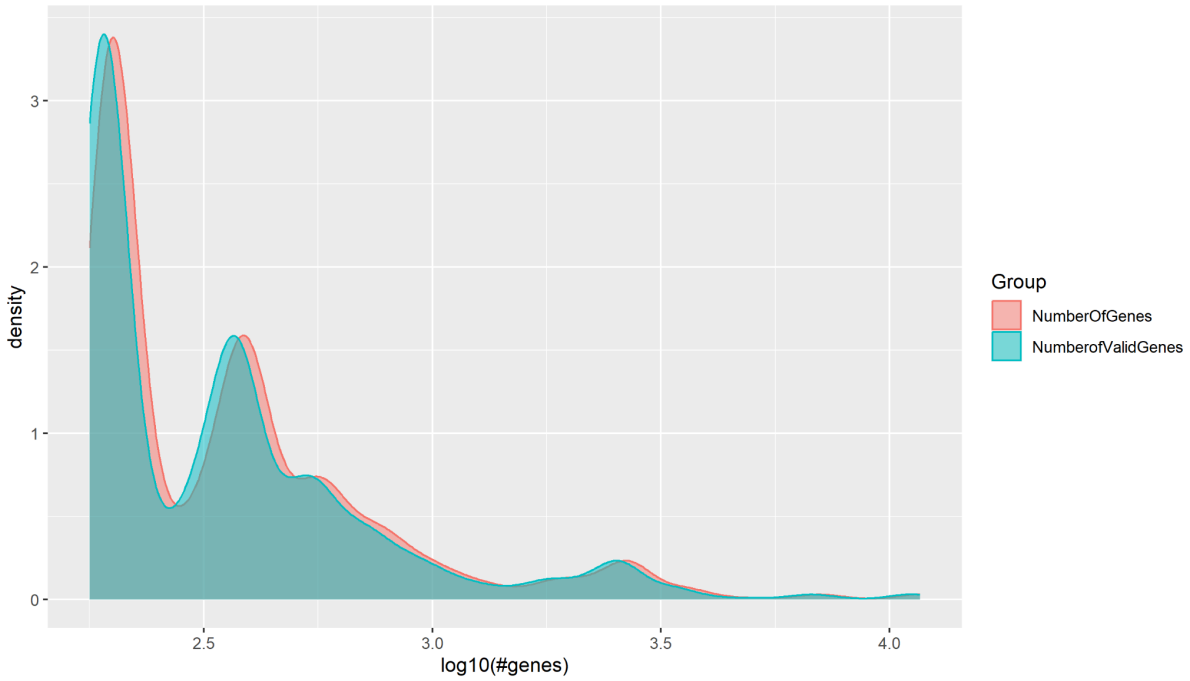
4.2 - Tetrapoda

4.2.1 - Métricas de grupos de homólogos pré e pós-filtros

Iniciamos nossas análises com um total de 3.665.400 sequências proteicas de tetrápodes de 198 espécies, distribuídas em 65.948 grupos de homólogos. O filtro de tamanho de sequência removeu 182.075 sequências (4,96%) que divergiam mais do que dois desvios padrão do tamanho médio das sequências do grupo de

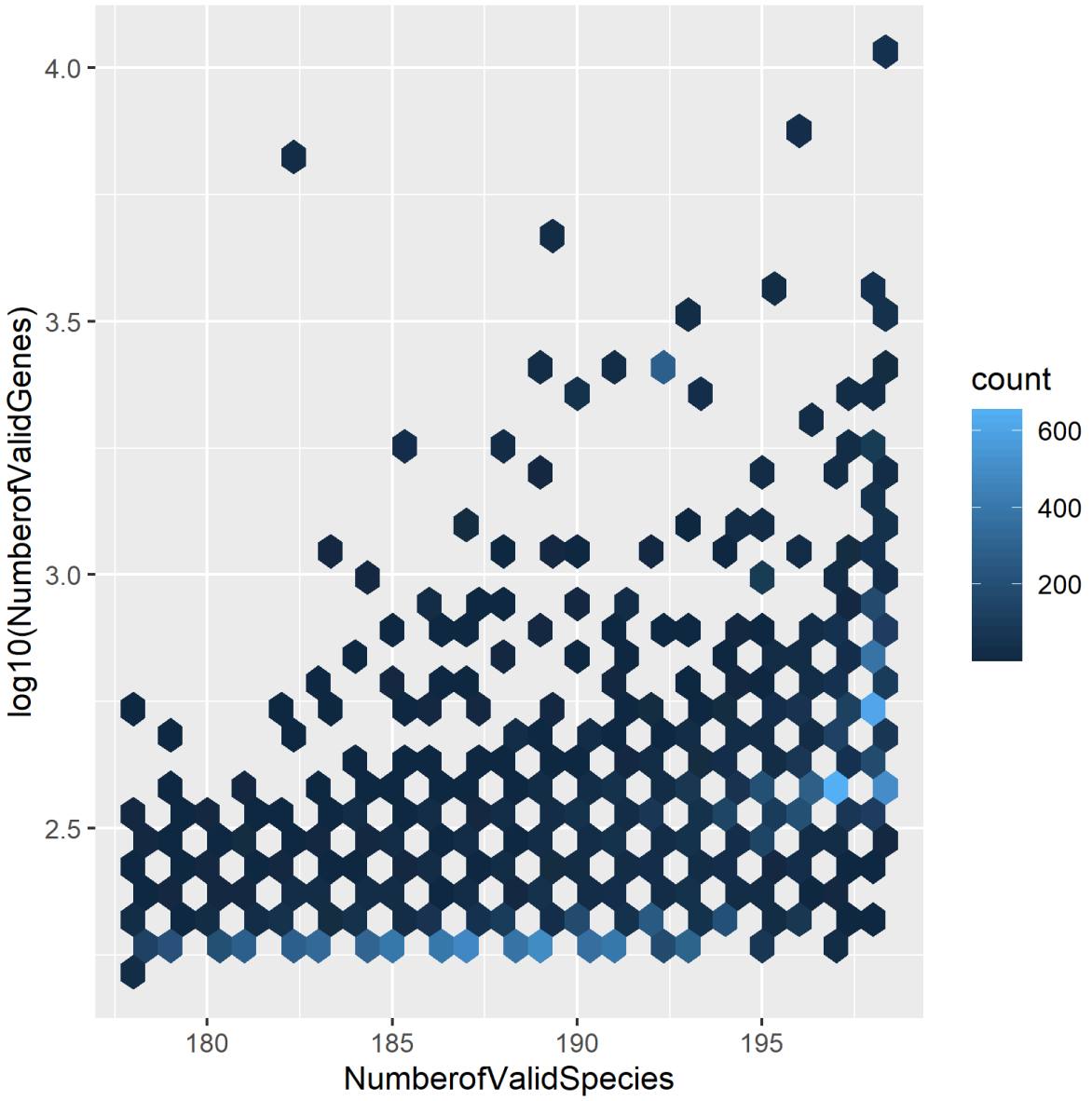
homólogos do qual elas faziam parte (pré-alinhamento). As 3.483.325 sequências remanescentes foram utilizadas nas análises subsequentes. Para que um determinado grupo de homólogos fosse considerado válido para a análise, ele deveria obedecer a duas condições: 1) possuir mais de 90% do total do número de espécies em análise e 2) possuir ao menos uma sequência da espécie-âncora (*Homo sapiens* para Tetrapoda). Após o filtro de qualidade de grupo, removemos 56.657 grupos de homólogos que não obedeciam a estas condições. Ao final do processamento das sequências, foram selecionados 2.557.414 genes, sendo 13.608 do genoma âncora, distribuídos em 9.291 grupos de homólogos válidos. O filtro de tamanho de sequência não gerou perda significativa nos grupos válidos (Figura 6), mantendo a maioria dos grupos com uma sequência de cada espécie (Figura 7), correspondendo aos genes ortólogos de cópia única.

Figura 6 - Gráfico de densidade de genes pré e pós-filtros - Tetrapoda



Densidade do log10 do total de genes antes(vermelho) e após(verde) o filtro de tamanho de sequência em cada grupo de homólogos. O maior pico indica os genes com aproximadamente 198 cópias, sugerindo tratar-se de ortólogos quase universais. Fonte: elaborado pelo autor, 2022.

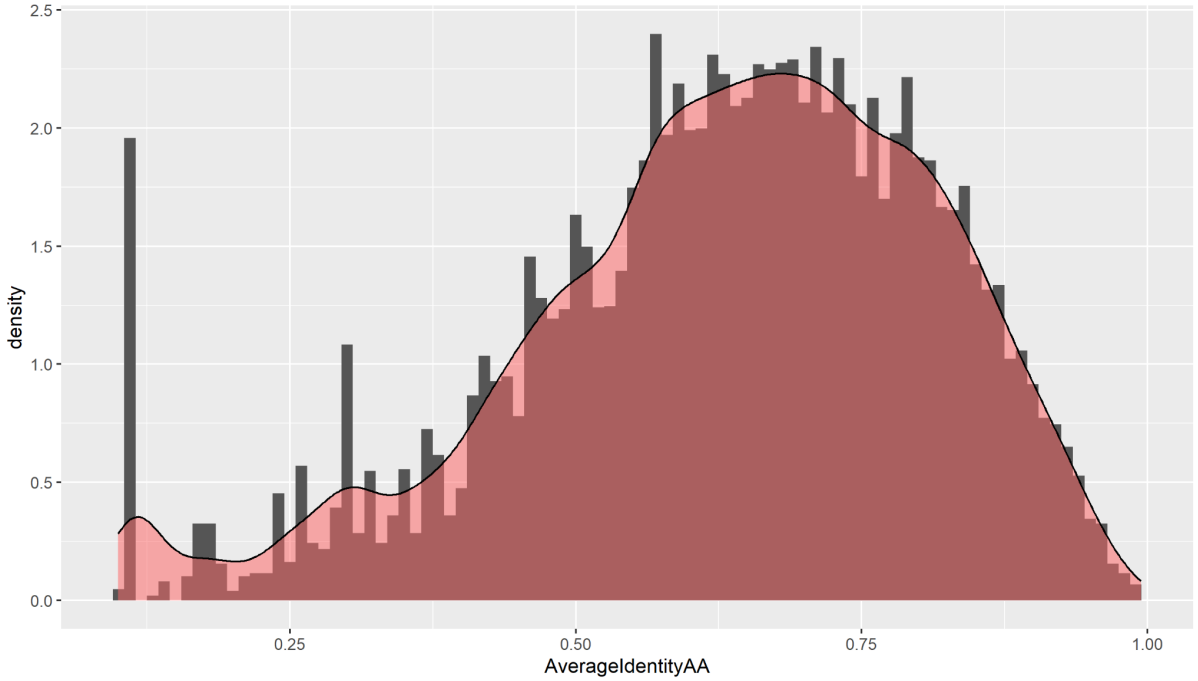
Figura 7 - Número de genes por número de espécies - Tetrapoda



Mapa de calor de hexágonos da contagem dos grupos de homólogos por número de espécies válidas por número de genes válidos. A vasta maioria dos grupos de homólogos foi observada como cópias únicas na vasta maioria das espécies. Fonte: elaborado pelo autor, 2022.

A distribuição das identidades médias de alinhamento de grupo é uma cauda longa, onde a maioria dos grupos são sequências altamente conservadas (identidade média de grupo > 70%), seguido da cauda longa que compreende a pequena fração de homólogos com a maior variação (identidade média de grupo < 30%) (Figura 8).

Figura 8 - Distribuição dos genes ordenados por identidade de alinhamento - Tetrapoda



Densidade de grupos de homólogos por identidade média de alinhamento. A vasta maioria dos grupos de homólogos possui altos valores de identidade, enquanto uma pequena fração dos grupos observada na cauda longa apresenta menores valores de identidade. Fonte: elaborado pelo autor, 2022.

4.2.2 - Análise de enriquecimento

Dentre os 13.608 genes de *H. sapiens* obtidos na etapa anterior, 13.092 (96,2%) foram mapeados na plataforma WebGestalt, utilizada para as análises de enriquecimento nos bancos de dados do Gene Ontology (Função Molecular, Processo Biológico e Componente Celular) e para as vias bioquímicas do KEGG.

Tanto para os termos GO quanto para as vias bioquímicas do KEGG, a grande maioria das categorias enriquecidas com escore de enriquecimento positivo, ou seja, que estão enriquecidos com genes na parte superior da lista (mais conservados) são termos associados com processos *housekeeping*, que são genes necessários para a manutenção das funções celulares básicas (Hounkpe *et al.,* 2021), tais como genes relacionados a processos de duplicação, transcrição e tradução gênica, síntese de aminoácidos, componentes de organelas, entre outros. O mesmo padrão foi observado para os outros dois taxa (Mammalia e Aves). Portanto, a partir desse ponto, nossos resultados terão como foco os componentes enriquecidos com escore de enriquecimento negativo (processos biológicos enriquecidos em genes variáveis nos grupos investigados).

Entre os genes de referência mapeados na plataforma, 7.782 estão anotados para as categorias de GO Componente Celular, e foram utilizados para a análise de enriquecimento. A figura 9 mostra os termos GO enriquecidos com um valor de FDR < 0,05, com os seus respectivos valores de escore de enriquecimento:

Figura 9 - Categorias GO - Componente Celular enriquecidas - Tetrapoda

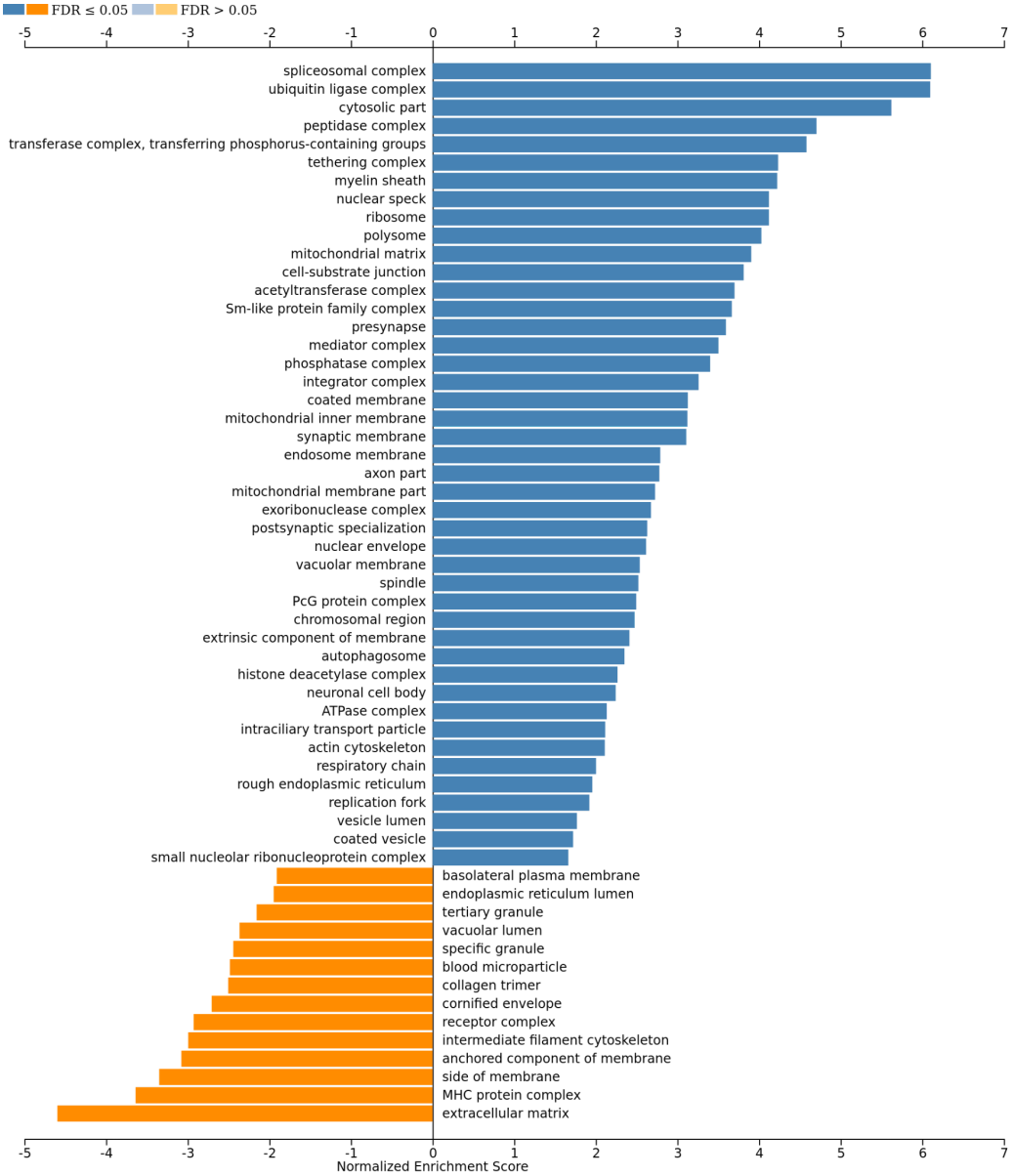


Gráfico de barras de categorias GO - Componente Celular enriquecidas. Valores positivos de escore de enriquecimento indicam um enriquecimento de genes que estão no início da lista ranqueada (mais conservados, barras azuis) enquanto valores negativos indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis, barras laranja). Fonte: elaborado pelo autor, 2022.

Na categoria GO - Função Molecular, 9.462 IDs dos genes mapeados na plataforma estão anotados para esta categoria e foram utilizados para a análise de enriquecimento. Esta categoria foi a que apresentou maior número de termos GO enriquecidos com escore de enriquecimento negativo para os Tetrapoda. A figura 10 mostra os termos GO enriquecidos com um valor de FDR < 0,05, com os seus respectivos valores de escore de enriquecimento:

Figura 10 - Categorias GO - Função Molecular enriquecidas - Tetrapoda

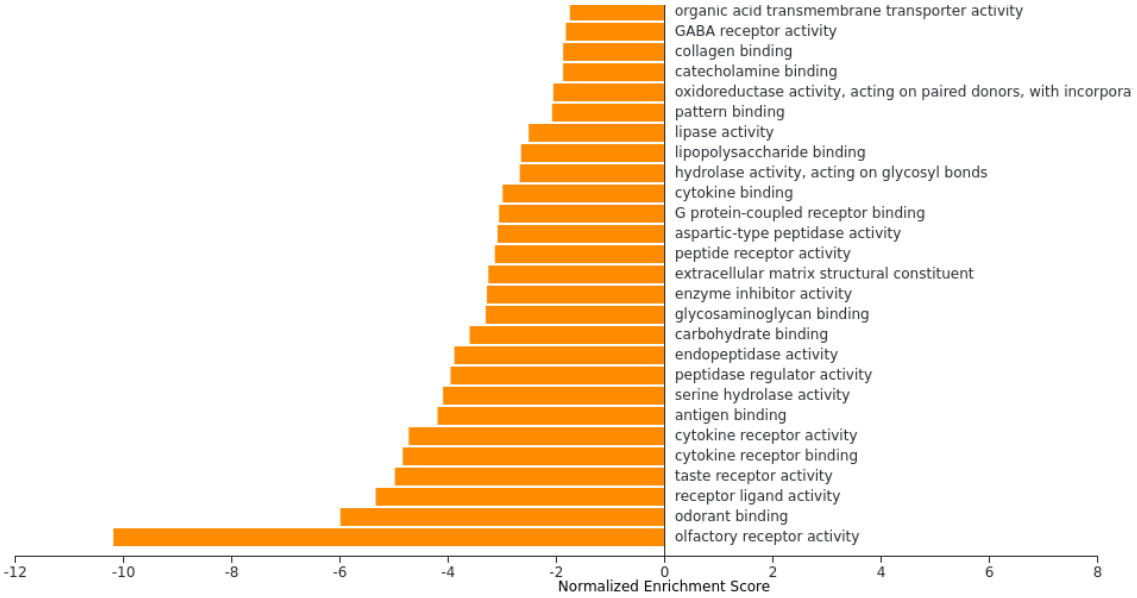


Gráfico de barras de categorias GO - Função Molecular enriquecidas. Valores negativos de escore de enriquecimento indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis). Fonte: elaborado pelo autor, 2022.

Para o GO - Processo Biológico, 11.096 genes de referência estão anotados e foram utilizados na análise de enriquecimento. Os termos GO enriquecidos com escore de enriquecimento foram podem ser vistos na figura 11 abaixo:

Figura 11 - Categorias GO - Processo Biológico enriquecidas - Tetrapoda

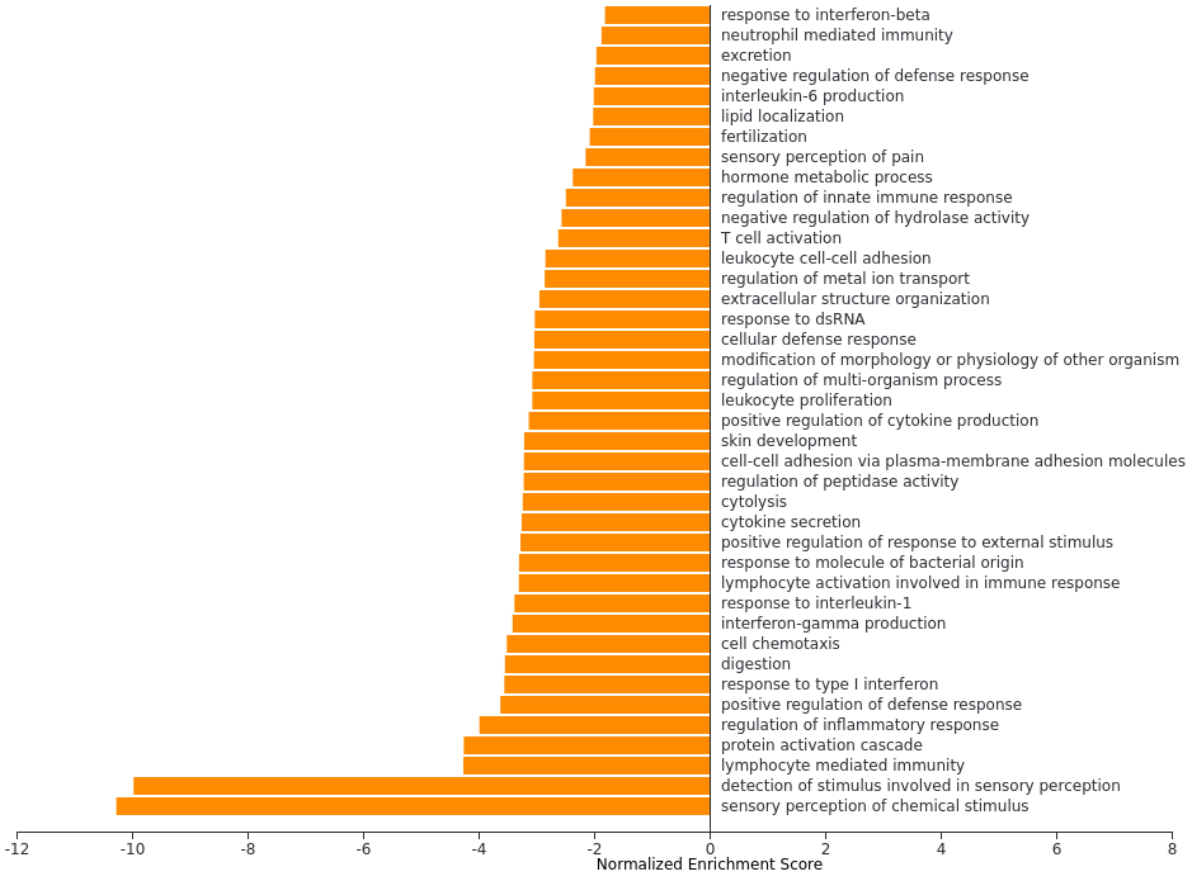


Gráfico de barras de categorias GO - Processo Biológico enriquecidas. Valores negativos de escore de enriquecimento indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis). Fonte: elaborado pelo autor, 2022.

Dos genes mapeados na plataforma, 5.571 estão anotados para vias bioquímicas do KEGG e foram utilizados para análise de enriquecimento, tendo as seguintes vias enriquecidas com EE negativo:

Figura 12 - Vias KEGG enriquecidas - Tetrapoda

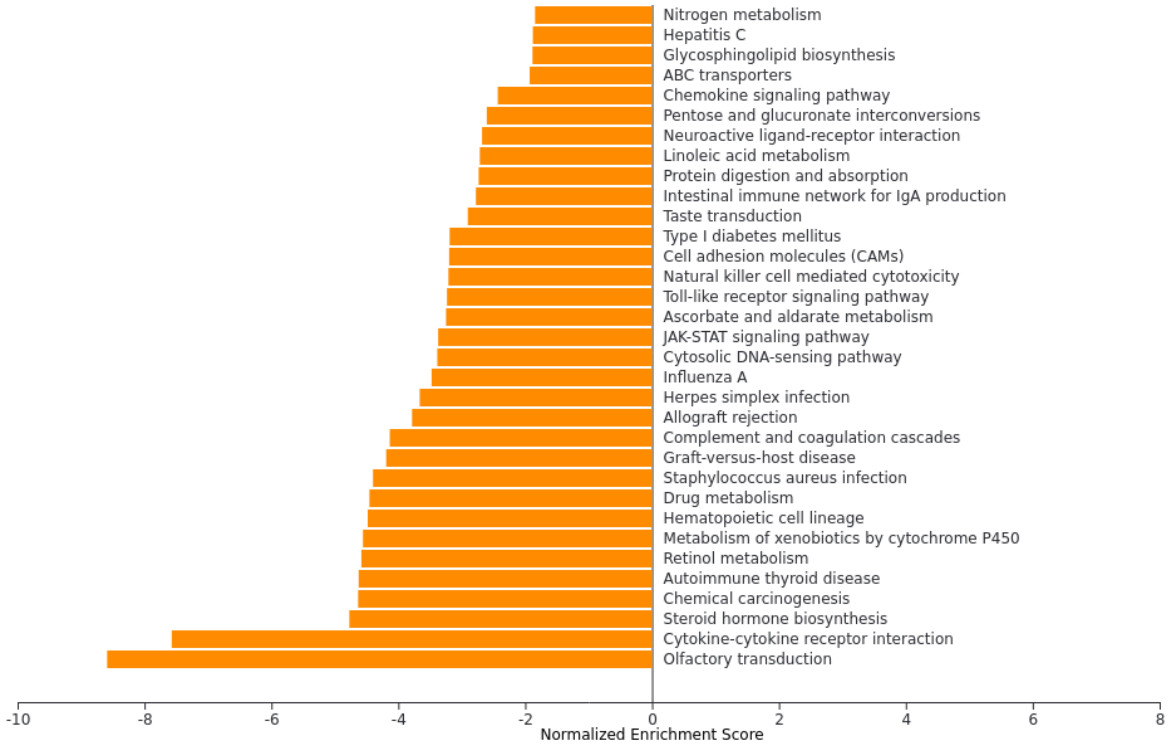


Gráfico de barras de vias KEGG enriquecidas. Valores negativos de escore de enriquecimento indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis). Fonte: elaborado pelo autor, 2022.

4.3 - Mamíferos

4.3.1 - Métricas de grupos de homólogos pré e pós-filtros

Iniciamos nossas análises com um total de 2.103.717 genes de 108 espécies de mamíferos, distribuídos em 46.202 grupos de homólogos. O filtro de tamanho de sequência removeu 103.036 genes (4,89%) que divergiam mais do que dois desvios padrão do tamanho médio de sequência do grupo aos quais elas faziam parte, ficando com um total de 2.000.681 genes. Para um grupo de homólogos ser considerado válido para a análise, ele deve obedecer a duas condições: possuir mais da metade do total do número de espécies e possuir ao menos uma sequência da espécie de referência, que para o conjunto de mamíferos é *Homo sapiens*. Após o filtro de qualidade de grupo, removemos 33.309 grupos de homólogos. Ao final do processamento das sequências, selecionamos 1.886.464 genes, sendo 18.337 presentes no genoma âncora e distribuídos em 12.893 grupos de homólogos.

Novamente, o filtro de tamanho de sequência não gerou perda significativa nos grupos válidos, mantendo a maioria dos grupos com uma sequência de cada espécie, correspondendo aos genes ortólogos de cópia única (Figuras 13 e 14).

Figura 13 - Gráfico de densidade de genes pré e pós-filtros - Mammalia

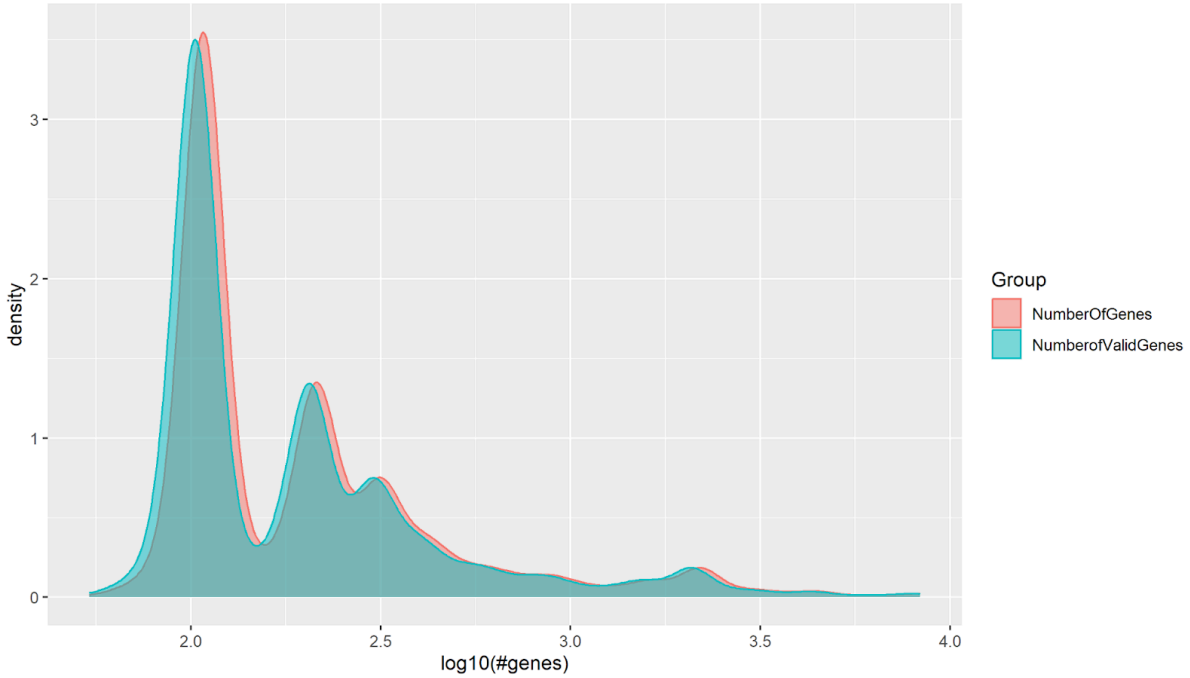
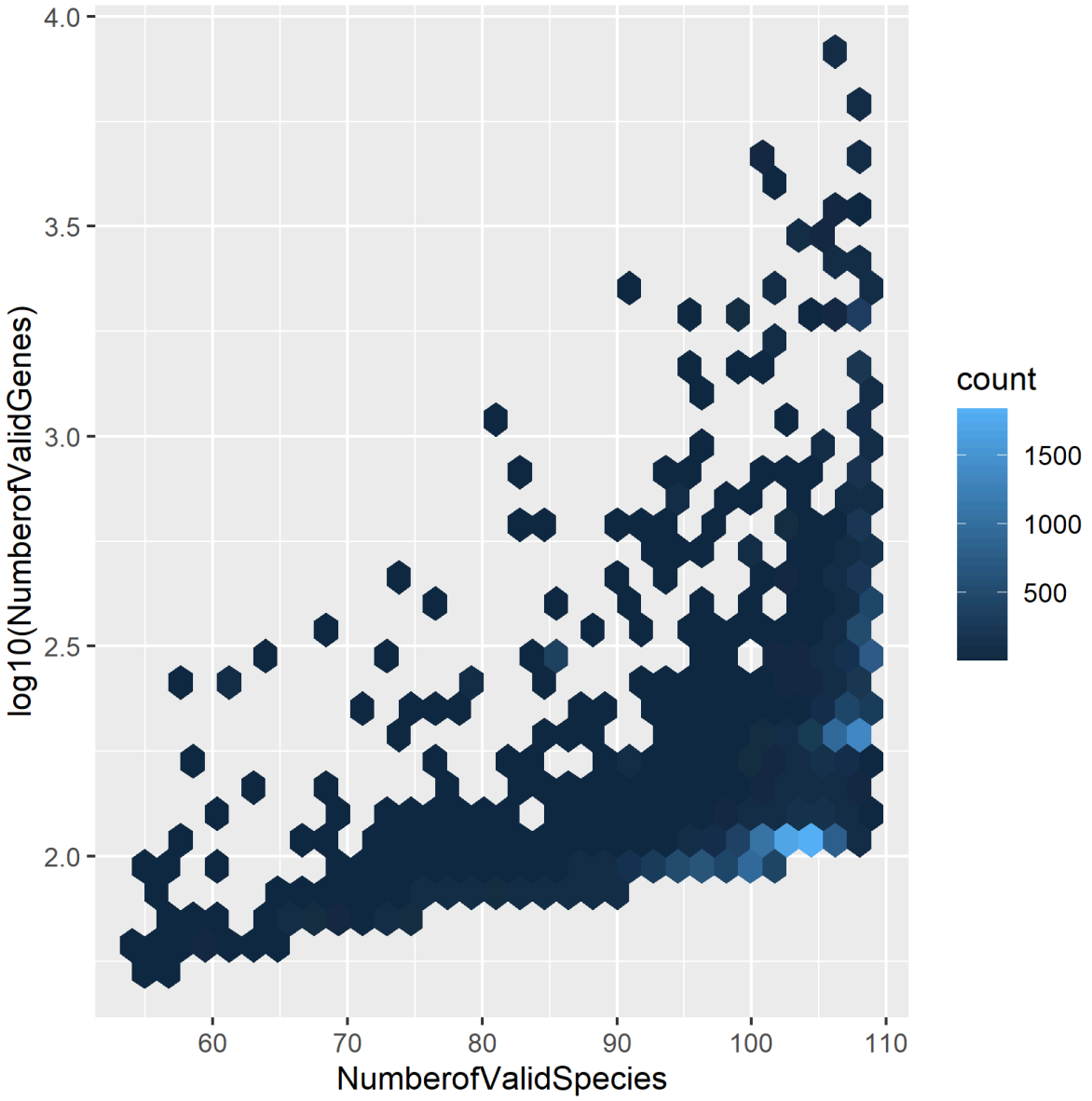


Gráfico de densidade do log 10 do total de genes antes(vermelho) e após(verde) o filtro de tamanho de sequência. O maior pico indica os genes com aproximadamente 108 cópias, sugerindo tratar-se de ortólogos quase universais. Fonte: elaborado pelo autor, 2022.

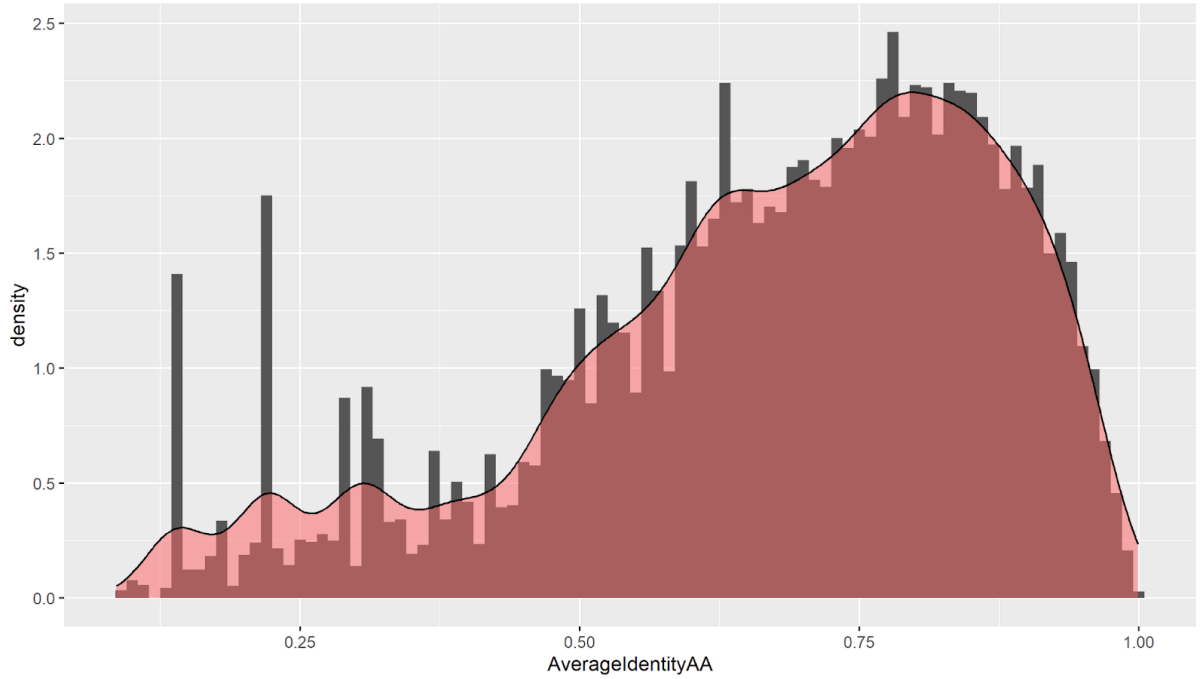
Figura 14 - Número de genes por número de espécies - Mammalia



Mapa de calor de hexágonos da contagem dos grupos de homólogos por número de espécies válidas por número de genes válidos. A vasta maioria dos grupos de homólogos foi observada como cópia simples em todas as espécies. Fonte: elaborado pelo autor, 2022.

A distribuição das identidades médias de alinhamento de grupo foi novamente uma cauda longa, onde a maioria dos grupos compreende sequências altamente conservadas (identidade média de grupo > 70%) e a cauda longa (identidade média de grupo < 30%) sendo composta por uma pequena fração de homólogos com a maior variação (Figura 15).

Figura 15 - Distribuição dos genes ordenados por identidade de alinhamento - Mammalia



Densidade de grupos de homólogos por identidade média de alinhamento. Fonte: elaborado pelo autor, 2022.

4.3.2 - Análise de enriquecimento

Dentre os 18.337 genes de referência selecionados anteriormente, 17.550 (95,7%) foram mapeados na plataforma WebGestalt para a análise de GSEA, novamente utilizando os bancos de dados do Gene Ontology de Função Molecular, Processo Biológico e Componente Celular e para as vias bioquímicas do KEGG.

Tanto para os termos GO quanto para as vias bioquímicas do KEGG, a vasta maioria das categorias enriquecidas com escore de enriquecimento positivo (mais conservados, barras azuis) são termos associados com processos *housekeeping*.

Entre os genes mapeados na plataforma, 9.453 estão anotados para as categorias de GO Componente Celular, e foram utilizados para a análise de enriquecimento (Figura 16). Os seguintes termos estão enriquecidos com ES negativo, com um valor de FDR < 0,05:

Figura 16 - Categorias GO - Componente Celular enriquecidas - Mammalia

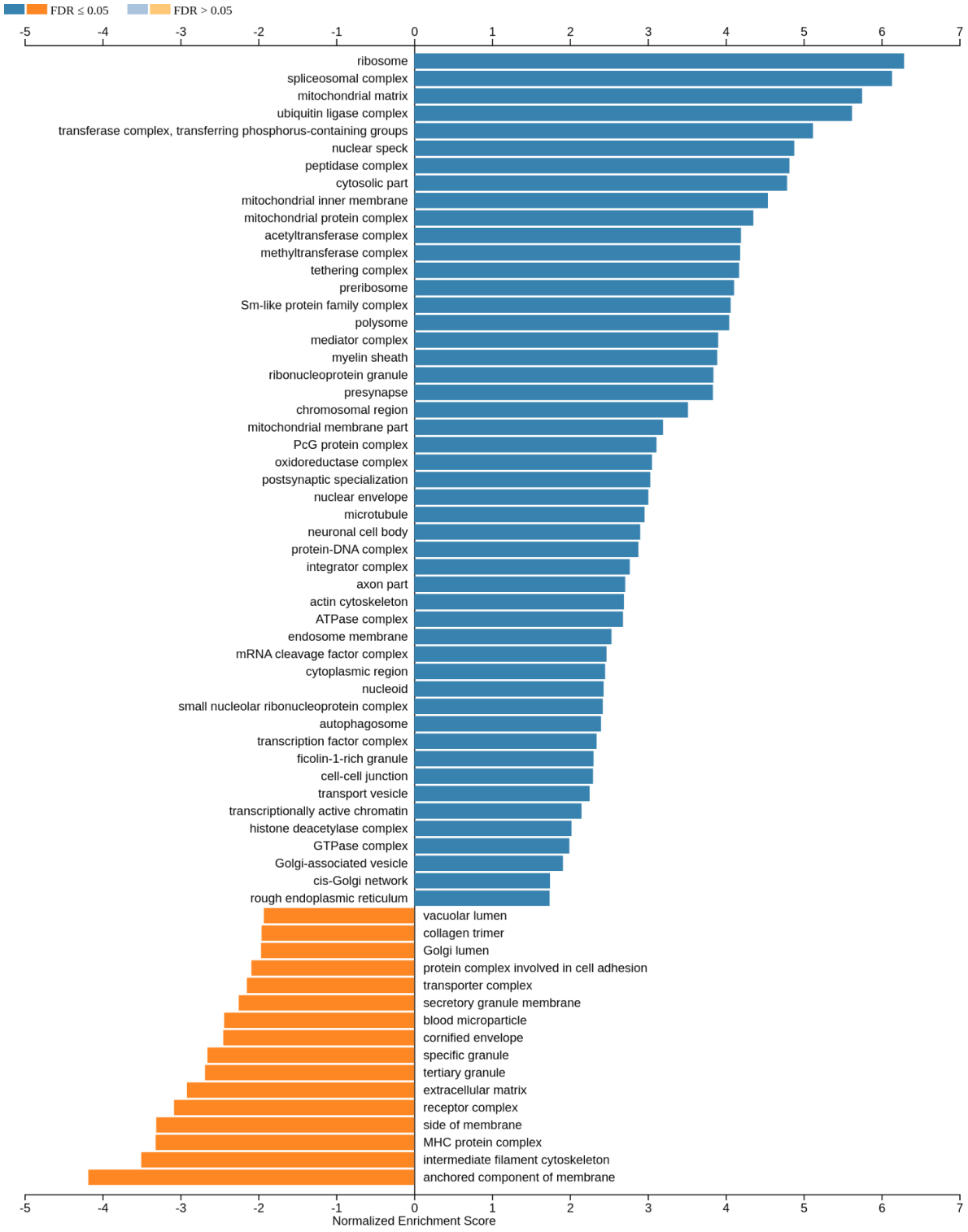


Gráfico de barras de categorias GO - Componente Celular enriquecidas. Valores positivos de escore de enriquecimento indicam um enriquecimento de genes que estão no início da lista ranqueada (mais conservados, barras azuis) enquanto valores negativos indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis, barras laranja). Fonte: elaborado pelo autor, 2022.

Para os termos GO Função Molecular, 11.656 estão anotados e foram utilizados para a análise de enriquecimento, tendo os seguintes termos enriquecidos com EE negativo (Figura 17):

Figura 17 - Categorias GO - Função Molecular enriquecidas - Mammalia

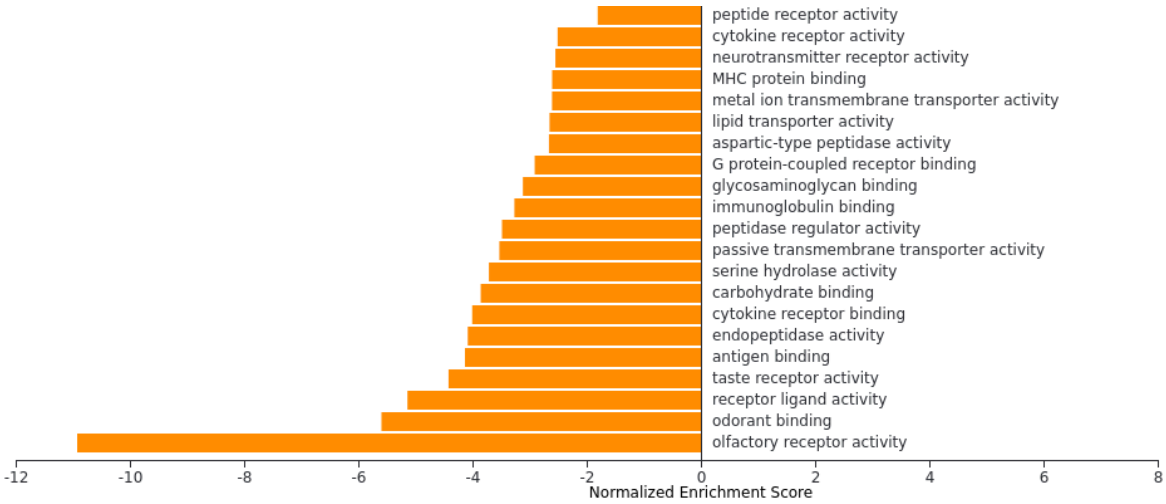


Gráfico de barras de termos GO - Função Molecular enriquecidos, com EE negativo.Fonte: elaborado pelo autor, 2022.

Os genes anotados para os termos GO - Processo Biológico e utilizados na análise de enriquecimento totalizaram 13.879, e revelaram os seguintes termos enriquecidos com EE negativo (Figura 18).

Figura 18 - Categorias GO - Processo Biológico enriquecidas - Mammalia

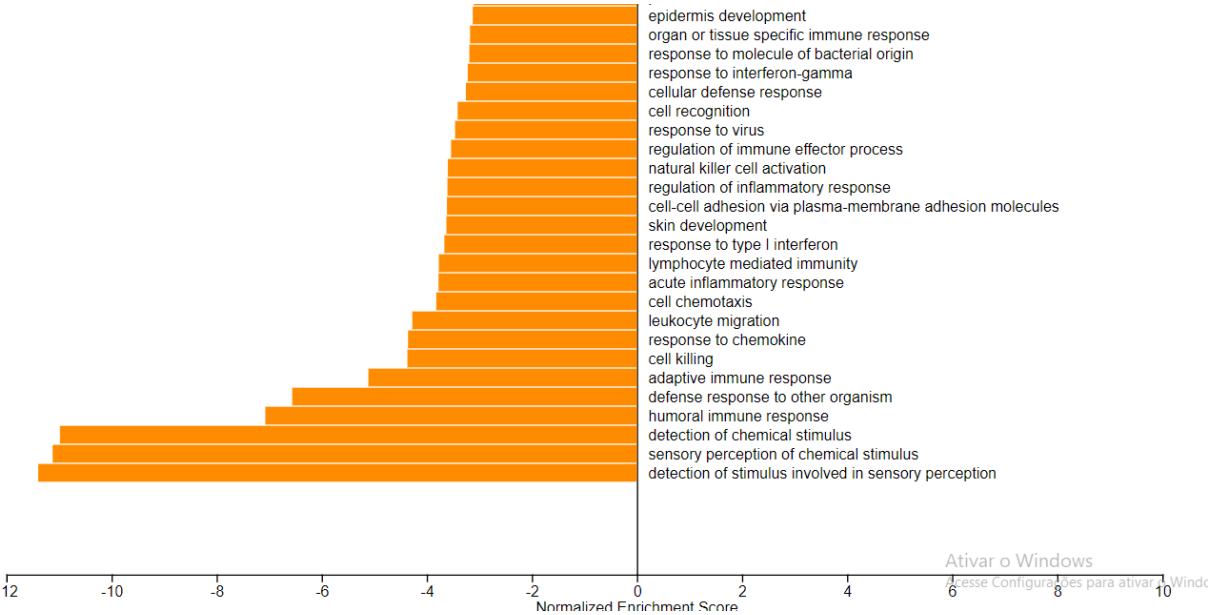


Gráfico de barras de termos GO - Processo Biológico enriquecidos, com EE negativo. Fonte: elaborado pelo autor, 2022.

Em relação as vias bioquímicas do KEGG, um total de 6.833 genes estão anotados e foram utilizados na análise de enriquecimento, tendo os seguintes termos enriquecidos com ES negativo (Figura 19).

Figura 19 - Vias KEGG enriquecidas - Mammalia

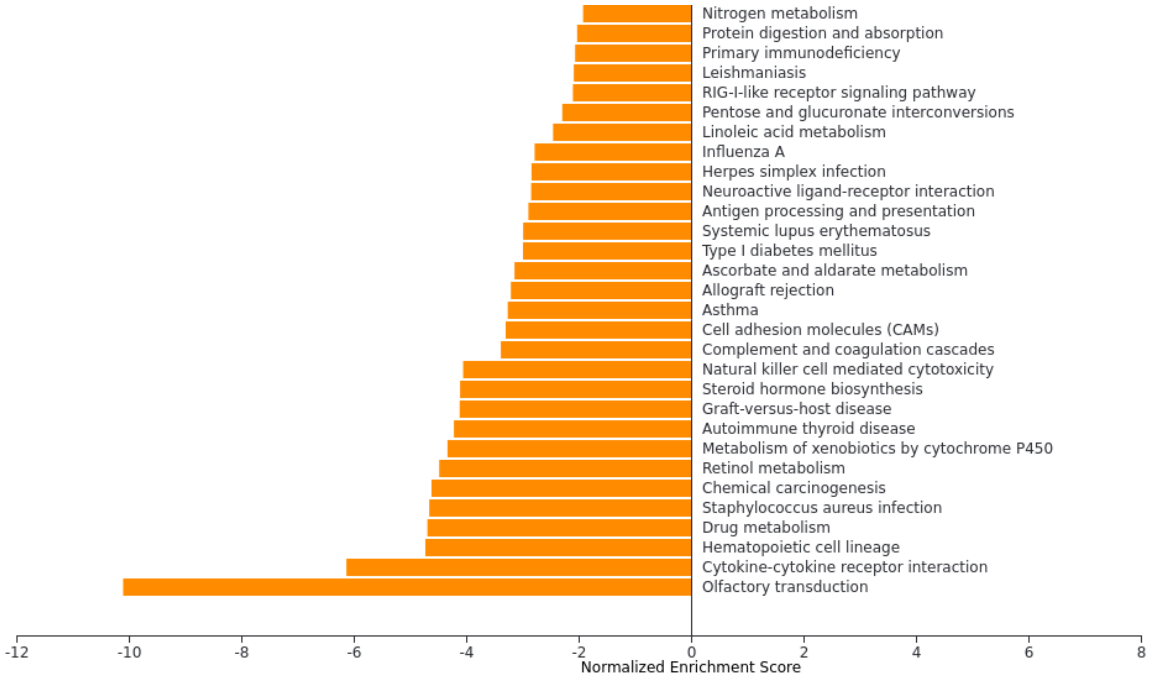


Gráfico de barras das vias bioquímicas do KEGG enriquecidas, com EE negativo. Fonte: elaborado pelo autor, 2022.

4.4 - Aves

4.4.1 - Métricas de grupos de homólogos pré e pós-filtros

Iniciamos nossas análises com um total de 766.774 genes de aves de 50 espécies, distribuídos em 22.975 grupos de homólogos. O filtro de tamanho de sequência removeu 41.460 genes (5,4%) que divergiam mais do que dois desvios-padrão do tamanho médio das sequências de seu grupo de homólogos. Procedemos as análises com um total de 725.314 genes. Para um grupo de homólogos ser considerado válido para a análise, ele deve obedecer a duas condições: possuir mais da metade do total do número de espécies e possuir ao menos uma sequência da espécie-âncora, (*Gallus gallus*). Após o filtro de qualidade

de grupo, removemos 12.102 grupos de homólogos. Ao final do processamento, selecionamos 659.925 genes, sendo 15.035 do genoma âncora, distribuídos em 10.873 grupos de homólogos válidos.

Novamente, o filtro de tamanho de sequência não gerou perda significativa nos grupos válidos (Figura 20), mantendo a maioria dos grupos com uma sequência de cada espécie, correspondendo aos genes ortólogos de cópia única (Figura 21).

Figura 20 - Gráfico de densidade de genes pré e pós-filtros - Aves

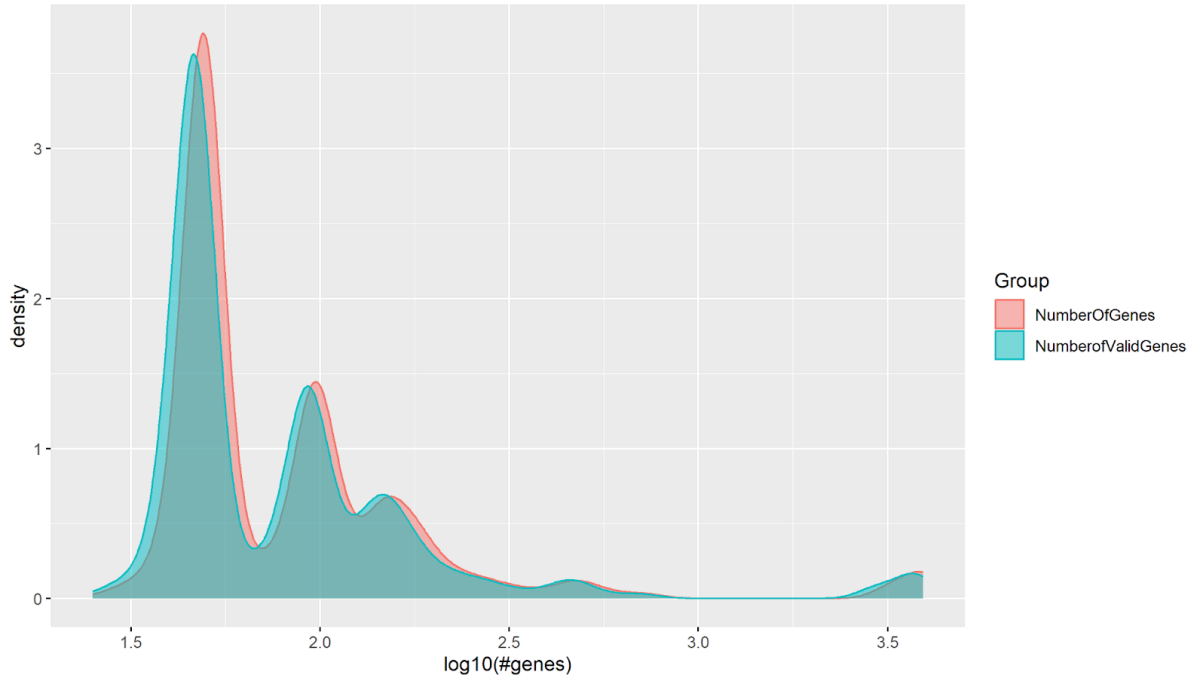
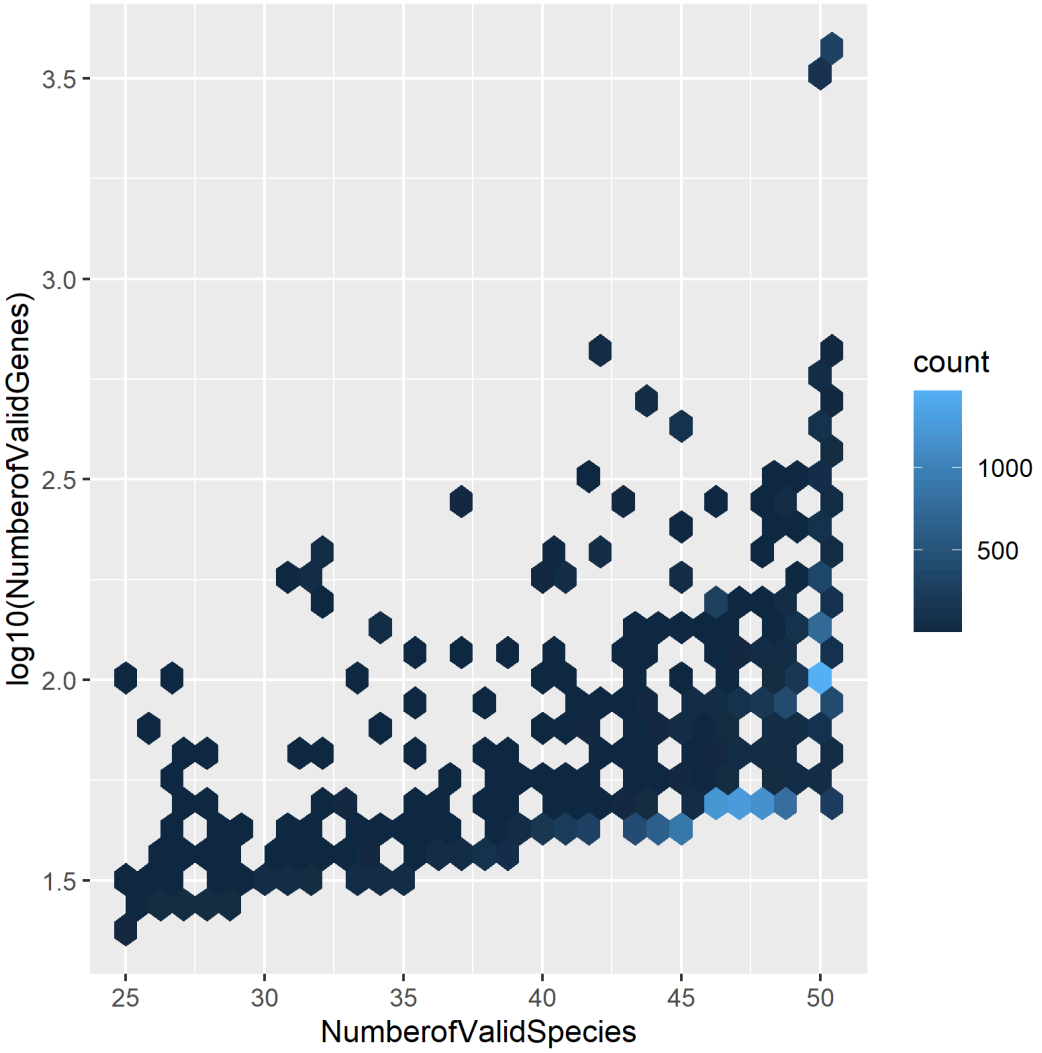


Gráfico de densidade do log 10 do total de genes antes(vermelho) e após(verde) o filtro de tamanho em cada grupo de homólogos. O maior pico indica os genes com aproximadamente 50 cópias, sugerindo tratar-se de ortólogos quase universais. Fonte: elaborado pelo autor, 2022.

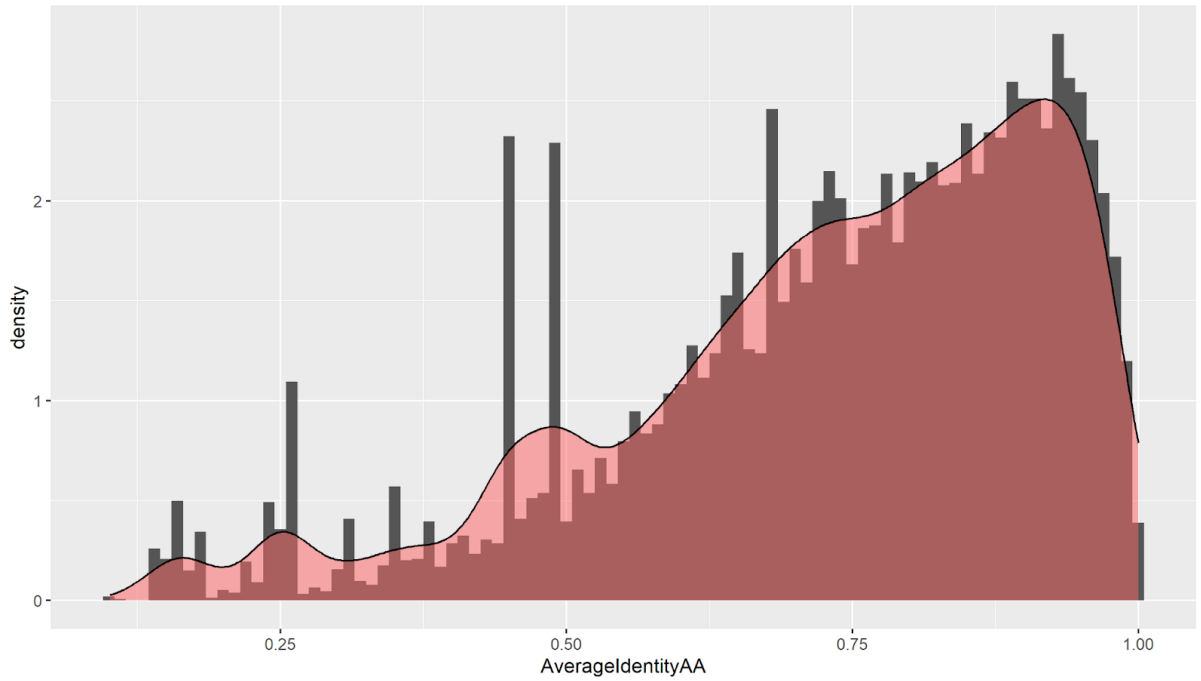
Figura 21 - Número de genes por número de espécies - Aves



Mapa de calor de hexágonos da contagem dos grupos de homólogos por número de espécies válidas por número de genes válidos. A vasta maioria dos grupos de homólogos foi observada como cópia simples em todas as espécies. Fonte: elaborado pelo autor, 2022.

A distribuição das identidades médias de alinhamento de grupo é uma cauda longa, onde a maioria dos grupos são sequências altamente conservadas (identidade média de grupo > 70%) e a cauda longa (identidade média de grupo < 30%) compreende a pequena fração de homólogos com a maior variação (Figura 22).

Figura 22 - Distribuição dos genes ordenados por identidade de alinhamento - Aves



Densidade de grupos de homólogos por identidade média de alinhamento. Fonte: elaborado pelo autor, 2022.

4.4.2 - Análise de enriquecimento

Dos genes de referência, 13.112 (87,2%) foram mapeados na plataforma WebGestalt, onde foram feitas as análises de enriquecimento nos bancos de dados do Gene Ontology de Função Molecular, Processo Biológico e Componente Celular e para as vias bioquímicas do KEGG.

Tanto para os termos GO quanto para as vias bioquímicas do KEGG a grande maioria das categorias enriquecidas com escore de enriquecimento positivo, ou seja, que estão enriquecidos com genes na parte superior da lista (mais conservados), são termos associados com processos *housekeeping*.

Entre os genes mapeados na plataforma, 4.327 IDs estão anotados para as categorias de GO Componente Celular, e foram utilizados para a análise de enriquecimento, tendo os seguintes termos GO enriquecidos com um valor de FDR < 0,05 (Figura 23).

Figura 23 - Categorias GO - Componente Celular enriquecidas - Aves

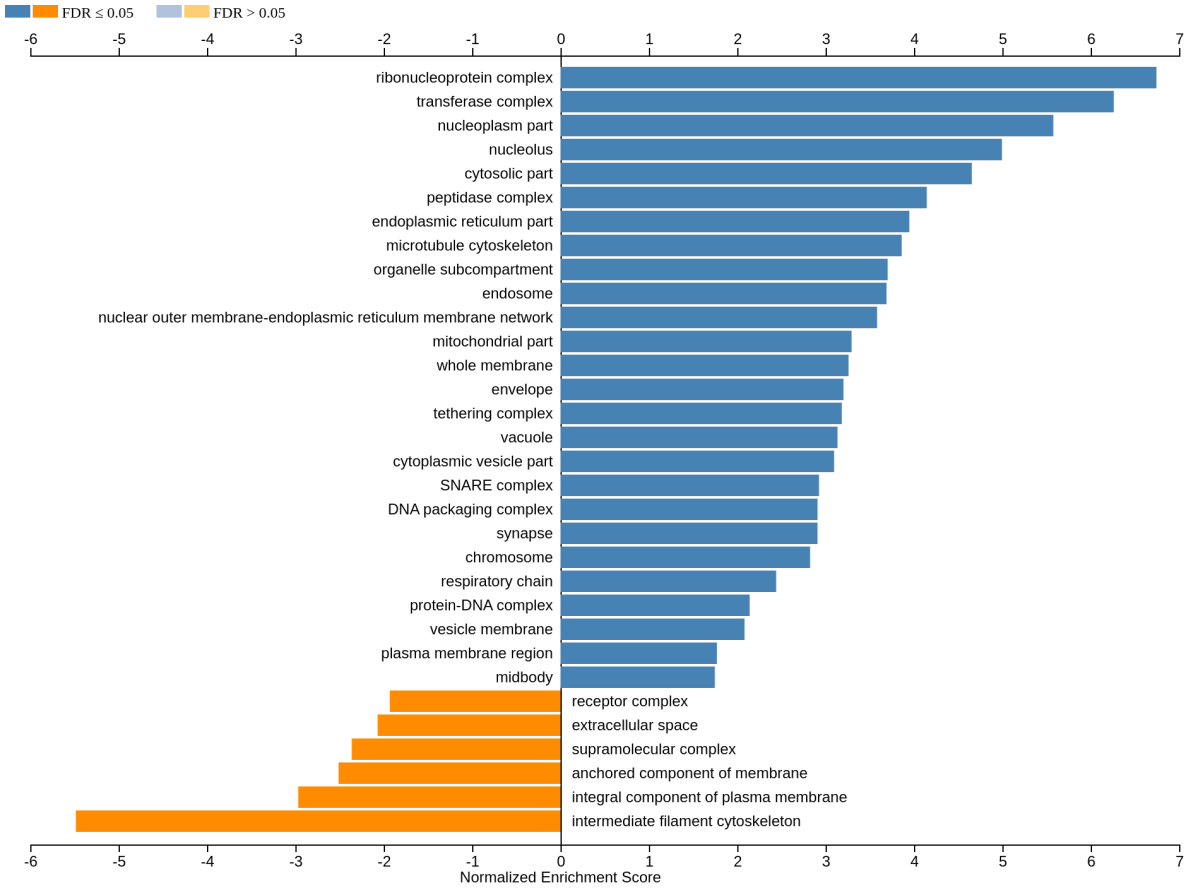


Gráfico de barras de categorias GO - Componente Celular enriquecidas. Valores positivos de escore de enriquecimento indicam um enriquecimento de genes que estão no início da lista ranqueada (mais conservados, barras azuis) enquanto valores negativos indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis, barras laranja). Fonte: elaborado pelo autor, 2022.

Para os termos GO Função Molecular, 5.391 estão anotados e foram utilizados para a análise de enriquecimento, tendo os seguintes termos enriquecidos com EE negativo (Figura 24).

Figura 24 - Categorias GO - Função Molecular enriquecidas - Aves

Gráfico de barras de categorias GO - Função Molecular enriquecidas. Valores negativos de escore de enriquecimento indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis). Fonte: elaborado pelo autor, 2022.

Os genes anotados para os termos GO - Processo Biológico e utilizados na análise de enriquecimento totalizaram 5.873, e revelaram os seguintes termos enriquecidos com EE negativo (Figura 25).

Figura 25 - Categorias GO - Processo Biológico enriquecidas - Aves

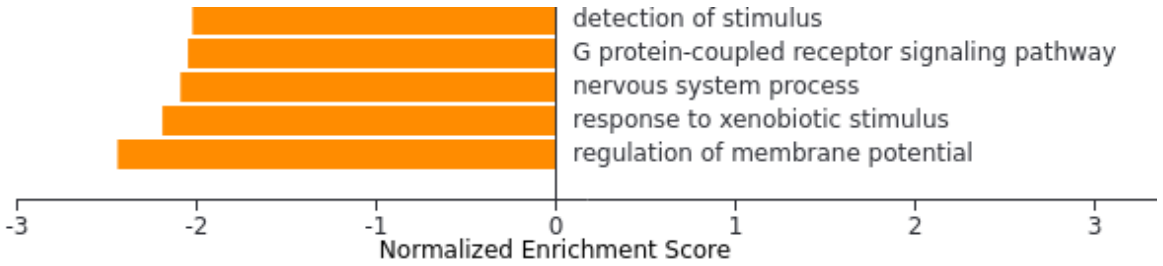


Gráfico de barras de categorias GO - Processo Biológico enriquecidas. Valores negativos de escore de enriquecimento indicam um enriquecimento de genes que estão mais ao final da lista ranqueada (mais variáveis). Fonte: elaborado pelo autor, 2022.

Dos genes mapeados na plataforma, 4.137 estão anotados para vias bioquímicas do KEGG e foram utilizados para análise de enriquecimento, tendo as seguintes vias enriquecidas com EE negativo (Figura 26).

Figura 26 - Vias KEGG enriquecidas - Aves

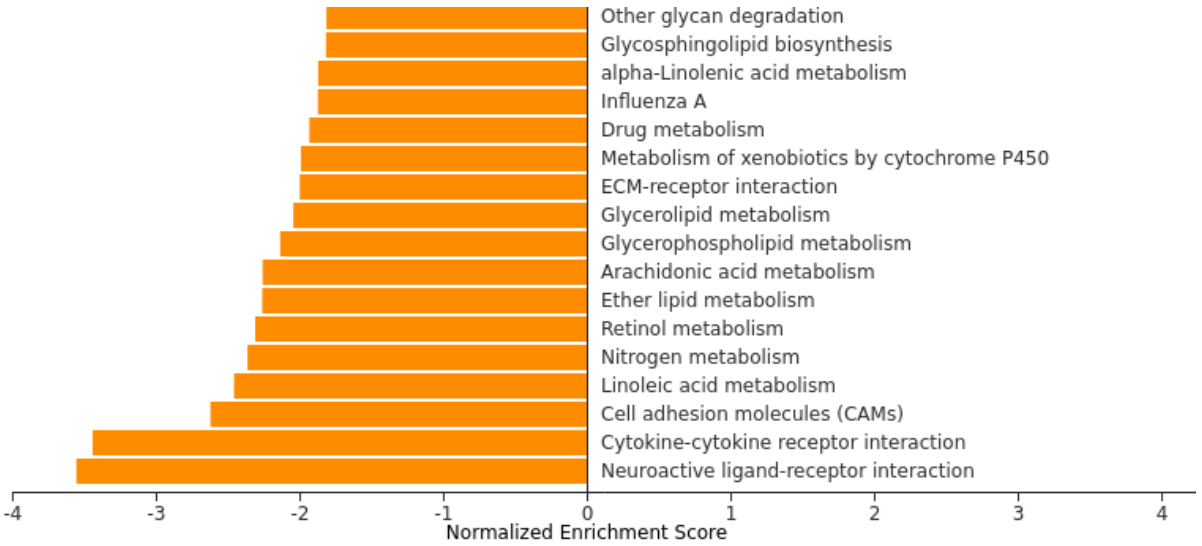


Gráfico de barras de categorias vias bioquímicas do KEGG enriquecidas, com EE negativo. Fonte: elaborado pelo autor, 2022.

5 - DISCUSSÃO

Os resultados obtidos nas etapas de produção dos grupos de homólogos demonstraram que uma fração considerável dos grupos compreendem ortólogos 1-1 presentes na maioria dos genomas avaliados. Adicionalmente, a distribuição da identidade nos diferentes grupos apresenta o padrão esperado onde a vasta maioria das mutações não-sinônimas é deletéria: um grande número de grupos de homólogos com altos valores de identidade, seguido de uma cauda longa contendo os poucos grupos de homólogos que apresentam alta variação de sequência. De maneira geral, estes resultados sugerem uma alta qualidade nos conjuntos de grupos de homólogos produzidos, indicando que os filtros que aplicamos forneceram conjuntos de homólogos possivelmente verdadeiros, uma vez que foram detectados na maior parte dos genomas analisados.

A busca computacional por genes com pressão seletiva para sua variação compreende uma das mais importantes ferramentas para a busca por genes adaptativos (Hongo *et al., 2015*; Sahm *et al.,* 2017; Yang, 2007). Entretanto, os programas necessários para realizar tal busca requerem a instalação de diversas dependências, bem como um elevado uso de recursos computacionais (Álvarez-Carretero, Kapli & Yang, 2023.). Adicionalmente, buscas dessa natureza frequentemente investigam somente ortólogos 1-1, o que inviabiliza a avaliação do cenário de divergência funcional de parálogos via sub e neofuncionalização (Birchler & Yang, 2022.). Finalmente, cabe ressaltar que a busca por genes com evidência de seleção positiva produz ao final uma lista de genes que é comumente interpretada através de análises de enriquecimento como GSEA (Kosiol *et al.*, 2008.).

A metodologia que propomos para buscar por processos biológicos enriquecidos em genes variáveis compreende uma nova abordagem de genômica comparativa que integra a busca por genes adaptativos – definidos em nosso caso como os grupos de genes homólogos mais variáveis – com a detecção de processos biológicos enriquecidos em genes variáveis. Adicionalmente, nossa abordagem também considera a evolução molecular de parálogos, uma vez que eventuais duplicações gênicas presentes no genoma de referência contribuem múltiplas vezes para a lista de genes ordenados.

Nossa premissa assume que a posição relativa de um grupo de homólogos na lista de genes homólogos ordenados por sua similaridade depende da pressão

seletiva relativa nos demais grupos de homólogos da lista. Uma vez que a vasta maioria das mutações não-sinônimas é deletéria, espera-se uma grande quantidade de grupos de homólogos com conservação de sequência. Entretanto, a fração dos homólogos onde mutações não-sinônimas são benéficas – sendo esta definição clássica de genes adaptativos – deve ocupar o outro extremo dessa lista.

Assim, em nossa abordagem, esperamos que os genes com pressão seletiva para sua variação devem se acumular na porção variável da lista e, consequentemente, os processos biológicos que descrevem temas biologicamente relevantes para tais genes também ocorrerão nessa extremidade. Cabe ressaltar que uma lista ordenada de genes é exatamente o tipo de dado necessário para análises de GSEA, o que permite integrar completamente nossa estratégia para a busca por genes com pressão seletiva para sua variação com uma poderosa ferramenta estatística e aplicar na genômica comparativa um conceito que revolucionou a área da transcriptômica.

Discutimos abaixo os principais resultados encontrados nas análises de enriquecimento apresentadas anteriormente. Estes não só corroboram o perfil geral de funções biológicas enriquecidas em genes com evolução acelerada em Tetrapoda, Aves e Mammalia, como também detecta interessantes padrões linhagem-específicos de genes e funções biológicas que podem contribuir para a identidade molecular funcional dessas linhagens.

5.1 - Tetrapoda

5.1.1 - Componente Celular

5.1.1.1 - GO:0031012 *Extracellular Matrix*

Este termo é definido pelo AmiGO (Carbon *et al*., 2009) como “Uma estrutura externa a uma ou mais células, que fornece suporte estrutural, sinais bioquímicos ou biomecânicos para células ou tecidos.”, possui um valor de escore de enriquecimento normalizado de -4,5967 e um valor FDR <2,2e-16 (Figura 27). Anota 370 genes, sendo 244 deles do conjunto de ponta. Dentre os genes desta categoria temos *alpha-1-B glycoprotein* (A1BG), *tectorin alpha* (TECTA), *secretory leukocyte peptidase inhibitor* (SLPI), *EGF like domain multiple 7* (EGFL7) e *angiogenin* (ANG).

Figura 27 - Gráfico de enriquecimento *Extracellular Matrix* - Tetrapoda

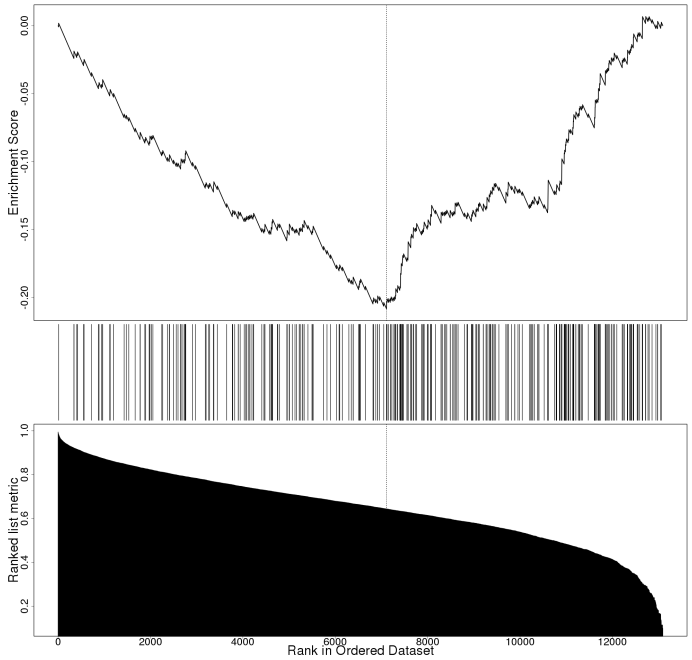


Gráfico de enriquecimento do GO:0031012 *Extracellular Matrix.* Fonte: elaborado pelo autor, 2022.

Alpha-1-B glycoprotein é uma glicoproteína do plasma com função desconhecida, e que possui uma similaridade com algumas proteínas da superfamília das imunoglobulinas. Pode estar associada com processos do sistema imune, o que explicaria sua grande variação, tendo uma identidade de 11% e sendo a proteína mais variável dentro deste GO.

A proteína *tectorin alpha,* possuindo uma identidade de 13%, é um dos principais componentes da membrana tectorial, uma matriz extracelular do ouvido interno que está associada com a percepção auditiva. TECTA está diretamente relacionado com a audição, visto que já foram reportadas deficiências auditivas ou perda total de audição relacionados a perda de função de TECTA (Iwasaki *et al*., 2002.; Verhoeven *et al*.,1998). O processo de percepção sonora é também essencial para os tetrápodes, onde indivíduos comumente se comunicam com outros da mesma espécie emitindo diferentes sons e com diferentes propósitos. Essa comunicação pode encontrar restrições na forma como o som é emitido, propagado e interpretado, podendo atuar como elemento de pressão seletiva nos componentes do sistema auditivo. Outras possíveis pressões no sistema auditivo derivam de sua importância na detecção de presas e predadores (Fay & Popper 2000).

SLPI é um membro das proteínas associadas à imunidade inata, com 18% de identidade de alinhamento. Seu principal papel biológico é a proteção do tecido local

das consequências prejudiciais de processos inflamatórios através da sua atividade antiprotease, uma vez que a manutenção da integridade do tecido requer um equilíbrio entre proteases e antiproteases (Doumas, Kolokotronis & Stefanopoulos, 2005). Sua grande variação pode ser explicada como um processo coevolutivo secundário: a homeostase do sistema proteinase-antiproteinase, em conjunto com uma possível pressão seletiva nas proteases atuantes nos processos inflamatórios por parte dos parasitas. Nesse cenário, a variação das proteinases poderia ser responsável pela co-variação observada nas antiproteinases.

Outros genes interessantes que contribuíram para a significância deste GO são EGFL7 e ANG, com identidades de 20 e 27% respectivamente. Ambos ativam e regulam a angiogênese – processo de formação de novos vasos sanguíneos a partir de vasos sanguíneos preexistentes (Gao & Xu, 2008; Nichol, & Stuhlmann, 2012). Este processo é essencial para fornecer oxigênio e nutrientes aos tecidos, e durante o curso da evolução dos tetrápodes houve o surgimento de novos tecidos mais complexos que necessitam de uma rede vascular eficiente.

5.1.1.2 - GO:0042611 *MHC Protein Complex*

É definido pelo AmiGO como “Um complexo proteico transmembrana composto por uma cadeia alfa do MHC e, na maioria dos casos, uma cadeia beta do MHC classe II ou uma cadeia beta2-microglobulina invariante, e com ou sem um peptídeo ligado, lipídio ou antígeno polissacarídeo”. Possui um valor de FDR <2.2e-16 e um tamanho de conjunto de 13 genes, sendo 11 deles do conjunto de ponta (Figura 28). Anota genes de MHC classe I e II e o gene da beta-2-microglobulina (B2M). Possui um valor de escore de enriquecimento normalizado de -3,6385.

Figura 28 - Gráfico de enriquecimento *MHC Protein Complex* - Tetrapoda

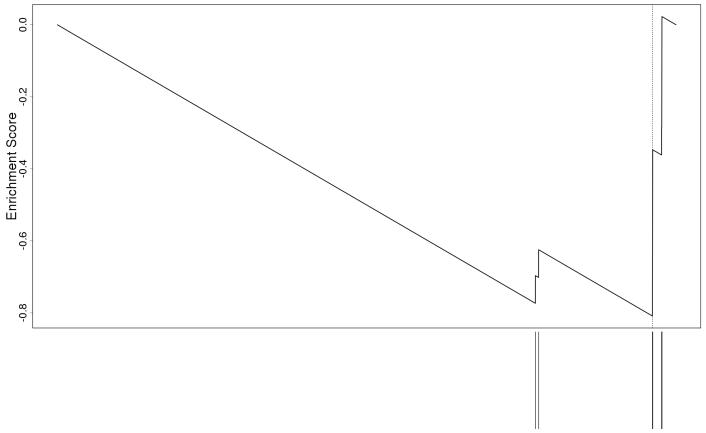


Gráfico de enriquecimento do GO:0042611 MHC *Protein Complex*. Fonte: elaborado pelo autor, 2022.

Genes do complexo MHC constituem parte do sistema imune adaptativo dos tetrápodes, e tem como uma de suas funções reconhecer e apresentar antígenos próprios ou de agentes externos paras as células do sistema imune (Eizaguirre, *et al.*, 2012). Existem duas categorias principais de moléculas de proteínas MHC - classe I e classe II. As moléculas de MHC classe I são formadas por uma associação de três domínios α1, α2 e α3 e beta-2-microglobulina (β2m), e estão presentes na membrana de praticamente todas as células de um organismo, ao passo que as moléculas de classe II são formadas por uma associação de duas cadeias alfa α1 e α2 e duas cadeias beta β1 e β2 e são encontradas apenas em células dendríticas, macrófagos e linfócitos B.

Esta categoria, bem como outras categorias associadas ao sistema imune e seus componentes, funciona como um controle positivo para nossa metodologia, pois sabemos que genes do sistema imune sofrem uma grande pressão seletiva para variação. Espécies parasitas e seus hospedeiros encontram-se em uma contínua ‘corrida armamentista’: parasitas adquirem vantagem adaptativa através da seleção de mecanismos para escapar ou modular o sistema imune hospedeiro, enquanto hospedeiros adquirem vantagem adaptativa através da seleção de mecanismos imunes capazes eliminar os parasitas, conforme postulado pela teoria da Rainha Vermelha e discutido no capítulo 1.

Além de constituintes importantes do sistema imune adaptativo dos tetrápodes, há evidências de que os genes de MHC desempenham um papel fundamental no processo de seleção sexual em diversas espécies (Milinski, 2006; Jan Ejsmond, Radwan & Wilson, 2014.). Os genes MHC estão entre os que

possuem maior diversidade alélica na espécie humana, uma vez que há uma grande vantagem adaptativa em possuir alelos distintos que permitem a apresentação de uma maior gama de antígenos. Em diversas espécies de vertebrados, indivíduos conseguem distinguir a identidade alélica do MHC de possíveis parceiros sexuais através de odores específicos, existindo uma preferência para o acasalamento entre indivíduos que apresentem alelos de MHC diferentes dos seus, aumentando tanto a variabilidade alélica do MHC quanto a heterozigose geral da prole (Penn, 2002).

Os membros do MHC classe II possuem um maior valor de identidade de alinhamento quando comparados com os da classe I (32% x 27%), o que já havia sido observado anteriormente (Kaufman, Salomonsen & Flajnik, 1994) e fornece evidência adicional da robustez de nossa análise. Podemos observar na nossa análise de homólogos que existe uma diminuição no número de cópias da classe I na linhagem das aves, ocorrendo até a perda total em algumas espécies.

A proteína beta-2-microglobulina (β2m), que faz parte do complexo MHC classe I, possui uma maior conservação quando comparado com outros integrantes deste GO, tendo 52% de identidade de alinhamento. β2m se liga de forma não covalente a cadeia α1 do complexo MHC classe I, formando um sítio de ligação de antígeno. É encontrado como cópia única na grande maioria dos genomas de tetrápodes, mas ocorreu um evento de duplicação deste gene no ancestral da linhagem dos Cetartiodactyla, fazendo com que mais de uma cópia possa ser encontrada nestas espécies, podendo ter contribuído para a evolução acelerada de algumas delas (Le *et al.,* 2017).

5.1.1.3 - GO:0001533 *Cornified Envelope*

Definido pelo AmiGO como “Um tipo de membrana plasmática que foi modificada pela adição de componentes intracelulares e extracelulares distintos, incluindo a ceramida, encontrada nas células epiteliais cornificantes (corneócitos).” Tem um tamanho de 26 genes, sendo 18 deles do conjunto de ponta (Figura 29). Possui um valor de FDR de 0,00041821 e um valor de escore de enriquecimento normalizado de -2,7065.

Figura 29 - Gráfico de enriquecimento *Cornified Envelope* - Tetrapoda

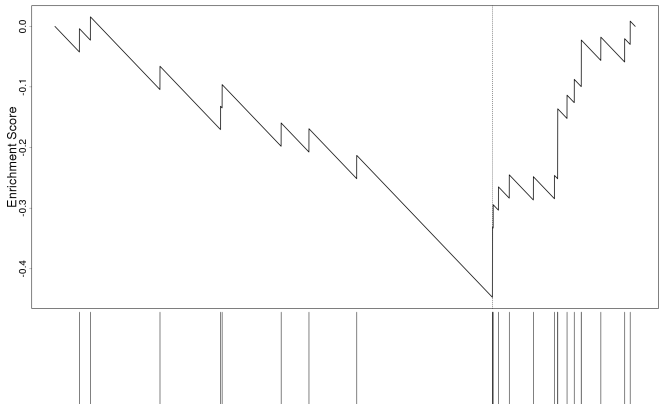


Gráfico de enriquecimento do GO:0001533 *Cornified Envelope*. Fonte: elaborado pelo autor, 2022.

A proteína *Peptidase Inhibitor* 3, também conhecida como *Elafin*, codificada pelo gene PI3, é a que possui maior variação dentro deste GO, possuindo uma identidade de alinhamento de 18%. Tem como principal função inibir as proteínas elastase neutrofílica e a proteinase-3. Através de sua atividade de peptidase, PI3 é um componente do sistema imunológico inato, protegendo as superfícies epiteliais de infecções antimicrobianas contra bactérias gram positivas e gram negativas que atuam no sistema respiratório (Simpson *et al.*, 1999). A sua grande variação dentro de Tetrapoda pode ser explicada pelo processo de seleção de mecanismos do sistema imune.

A proteína *Repetin*, codificada pelo gene RPTN, ocupa a segunda posição de maior variação deste GO, possuindo uma identidade média de alinhamento de 25%. É um membro da família S100 *fused genes* que fazem parte do complexo de diferenciação da epiderme (CDE) (Kypriotou; Huber & Hohl, 2012). A epiderme dos tetrápodes funciona como uma barreira contra danos físicos, perda de água e contra a invasão de possíveis agentes parasitas. Diversas adaptações na epiderme ocorreram durante o curso da evolução dos tetrápodes, que permitiram a ocupação de diferentes nichos. Em particular, modificações da epiderme foram essenciais para a saída do ambiente aquático e a conquista do ambiente terrestre. Adicionalmente, há evidência de seleção positiva de RTPN e outros genes do CDE em mamíferos (Goodwin & De Guzman Strong, 2017).

Sciellin é uma proteína codificada pelo gene SCEL, que possui uma identidade de 37% entre os tetrápodes. É um precursor do envelope cornificado e assim como a *Repetin,* faz parte da construção da epiderme de tetrápodes, o que

pode explicar seu padrão de variação. Interessantemente, estudos demonstraram que existe uma correlação entre o acúmulo de *Sciellin* na epiderme com a ativação de marcadores de diferenciação terminal na epiderme durante o desenvolvimento embrionário (Champliaud *et al*., 2000).

Outro integrante deste GO é a família de proteínas das desmogleínas. Os genes das desmogleínas foram agrupados em dois grupos de homólogos: um grupo contendo os genes *desmoglein* 1, 3 e 4, que possuem uma identidade de 46%, e o outro grupo com o gene *desmogle*in 2, que possui uma identidade de 53%. Desmogleínas são proteínas de adesão celular presente nos desmossomos e que possuem relação com diversas doenças humanas. Amagai & Stanley (2012) demonstram em sua revisão que as desmogleínas estão relacionadas com doenças autoimunes pênfigo foliáceo, pênfigo vulgar e pênfigo paraneoplásico, que são doenças caracterizadas pela formação de bolhas na pele e nas mucosas da boca, garganta, olhos, nariz e região genital. Também estão envolvidas em doenças e síndromes genéticas como a hipotricose - presença reduzida de pelos e cabelo; queratodermia palmoplantar - aumento da queratinização das regiões palmar e plantar e dorsal das mãos e pés; displasia arritmogênica do ventrículo direito -substituição do tecido muscular cardíaco normal por um tecido fibroso, causando interrupções dos sinais elétricos normais no coração, gerando ritmos cardíacos irregulares. Também está associada a algumas doenças infecciosas, sendo um receptor para alguns adenovírus.

O último membro deste GO que vamos abordar é a proteína *Annexin* A1, codificada pelo gene ANXA1, que possui uma identidade média de 54%. É uma proteína pertencente à superfamília *Annexin*: proteínas ligadoras de cálcio e fosfolipídeos. Membros desse grupo de homólogos possuem propriedades anti-inflamatórias, inibindo a liberação de mediadores pró-inflamatórios, como as prostaglandinas, além de reduzir a migração de leucócitos para sítios inflamatórios. Também é um componente regulatório de vias apoptóticas e promove a remoção de células apoptóticas, evitando a liberação de conteúdos celulares prejudiciais que podem desencadear uma resposta imunológica (Lim & Pervaiz, 2007; Perretti & D'acquisto, 2009). ANXA1 pode também desempenhar um papel crucial na fase de implantação do blastocisto, controlando o processo de inflamação necessário para a nidação, induzindo a sinalização necessária para ativar quinases e modulando o citoesqueleto epitelial (Hebeda *et al*., 2020). Em humanos, mutações em ANXA1

estão relacionadas a doenças da gravidez como a pré-eclâmpsia e a diabetes gestacional (Sousa *et al*., 2022).

5.1.1.4 - GO:0045111 *Intermediate Filament Cytoskeleton*

Definido pelo AmiGO como "Estrutura citoesquelética formada por filamentos intermediários, tipicamente organizados no citosol como um sistema estendido que se estende do envelope nuclear à membrana plasmática. Alguns filamentos intermediários correm paralelamente à superfície da célula, enquanto outros atravessam o citosol; juntos, eles formam uma estrutura interna que ajuda a sustentar a forma e a resistência da célula". Possui o escore de enriquecimento de -2,9937, o valor FDR <2,2e-16 e um tamanho de conjunto de 102 genes, sendo 49 deles do conjunto de ponta (Figura 30). Anota os genes da família *spectraplakin*: *Dystomin* (DST, também chamado de *Bullous pemphigoid antigen 1* (BPAG1)) e *microtubule-actin crosslinking factor 1* (MACF1, também chamado de *actin cross-linking factor 7*(ACF7)), o gene *SH3 and multiple ankyrin repeat domains 2* (SHANK2) e diversos genes de queratina. Importantemente, verificamos a importância de vários destes genes para o estabelecimento do tecido epitelial e para o sistema nervoso.

Figura 30 - Gráfico de enriquecimento *Intermediate Filament Cytoskeleton* - Tetrapoda

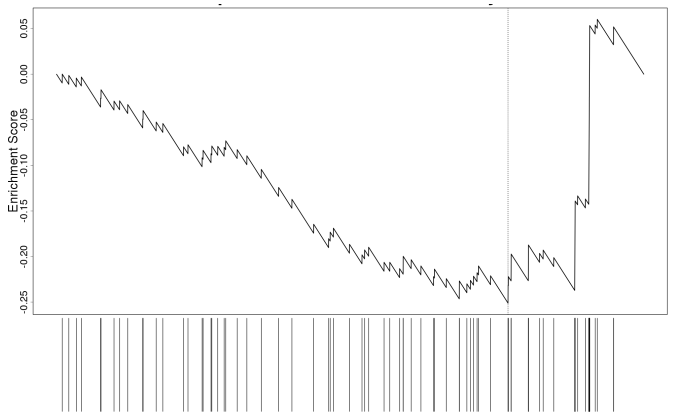


Gráfico de enriquecimento do GO:0045111 - *Intermediate Filament Cytoskeleton.* Fonte: elaborado pelo autor, 2022.

As queratinas podem ser encontradas em todos os tetrápodes, e compreende o maior subconjunto de genes de filamentos intermediários. É encontrada principalmente nos pêlos, unhas, pele, escamas, bicos e penas, onde desempenham

um papel essencial na proteção, resistência e impermeabilidade dessas estruturas. As queratinas podem ser divididas em alfas e beta queratinas, sendo a segunda encontrada somente na linhagem dos Sauropsida (Greenwold *et al.,* 2014). As alfa-queratinas podem ser classificadas como sendo do tipo I ou II, que formam dois grupos de homólogos, possuindo 42% e 44% de identidade de alinhamento, respectivamente.

Os genes da família *spectraplakin* estão agrupados no mesmo grupo homólogo, e possuem valor de identidade de 35%, sendo os mais variáveis membros deste GO. Possuem a capacidade de se associar a todos os três elementos do citoesqueleto: filamentos de actina, microtúbulos e filamentos intermediários. São proteínas grandes, podendo chegar a mais de oito mil aminoácidos (Brown, 2008). DST/BPAG1 regula a organização e a estabilização da rede de microtúbulos de neurônios sensoriais axonal, e sua perda pode causar distúrbios letais. Possui também a função de se ligar em redes de queratinas nos queratinócitos basais e os conecta à matriz extracelular. Assim como as Desmogleínas, DST/BPAG1 também está relacionado com o fenótipo de pênfigo, principalmente o bolhoso (Zhang, Yue & Wu, 2017). MACF1/ACF7 desempenha um papel na conexão entre microtúbulos e filamentos de actina, além de se ligar a componentes do filamento intermediário e é importante para vários processos celulares, incluindo migração celular, divisão e manutenção da integridade estrutural geral das células. Estudos em camundongos demonstraram que o desligamento deste gene causa letalidade embrionária pré-implantação (Kodama *et al.*, 2003).

O gene SHANK2, possui 40% de identidade de alinhamento e codifica proteínas que atuam na membrana pós-sináptica das sinapses excitatórias. Possui um domínio SH3, um motivo de repetição de anquirina e um motivo PDZ. O domínio SH3 está envolvido na regulação de importantes vias celulares, como proliferação celular, migração e modificações do citoesqueleto. O motivo de repetição de anquirina é um dos mais abundantes na natureza e sua principal função é mediar interações proteína-proteína (Li, Mahajan & Tsai, 2006). O domínio PDZ também tem como função a interação proteína-proteína, se ligando principalmente na cauda C-terminal das proteínas alvo (Lee & Zheng, 2010). Diversos estudos mostram que alterações em SHANK2 estão associadas ao desenvolvimento do transtorno do espectro autista (Berkel *et al*., 2012; Zaslavsky *et al*., 2019).

Assim, embora tenhamos observado uma variação significativa nas proteínas que compõem e que interagem com os membros do citoesqueleto, há evidências de que a perda de função destas proteínas em humanos está relacionada com diversos fenótipos de graves doenças, muitas vezes letais. Portanto, as modificações destas proteínas, embora aparente ter sido selecionadas via modificação para melhor se adequar aos diferentes fenótipos encontrados nos tetrápodes, não devem possuir modificações em suas funções primárias.

5.1.2 - Função Molecular

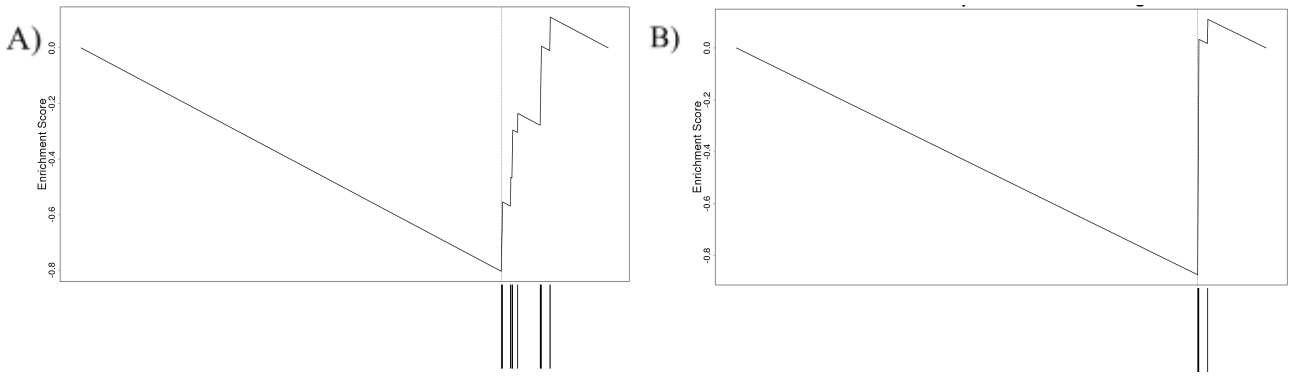
Os GOs de enriquecidos na categoria de função molecular estão relacionados a processos que foram fundamentais para a diversificação dos tetrápodes, como visão e gustação. Entre os GO's de função molecular estão presentes diversos termos relacionados a processos do sistema imune, tais como GO:0048018 *Receptor Ligand Activity*, GO:0004896 *Cytokine Receptor Activity*, GO:0003823 *Antigen Binding,* GO:0005126 *Cytokine Receptor Binding* e GO:0019955 *Cytokine Binding.*

Estes termos imunes anotam diversos genes de quimiocinas, interleucinas, interferons, imunoglobulinas e seus receptores, além de genes do complexo MHC. A grande variabilidade em genes imunes acontece devido à grande pressão seletiva causada pelo processo da Teoria da Rainha Vermelha, já explicada anteriormente, e compreende um importante controle positivo de nossa metodologia.

5.1.2.1 - GO:0004984 *Olfactory Receptor Activity* e GO:0005549 *Odorant Binding*

Estes dois GOs, conceitualmente similares , anotam proteínas da superfamília de receptores olfatórios, a qual é a maior família gênica do genoma de tetrápodes (Olender, Lancet & Nebert, 2008) . No GO de atividade de receptor olfatório, temos membros das famílias de receptores olfatórios 4, 5, 6 8, 10, 11, 51 e 52, enquanto no GO de ligante odorífero estão somente as famílias 5, 6, 8 e 11. Possui valores de identidade que variam entre 44% e 51% (Figura 31).

Figura 31 - Gráfico de enriquecimento *Olfactory Receptor Activity* e *Odorant Binding* - Tetrapoda



A) Gráfico de enriquecimento do GO:0004984 *Olfactory Receptor Activity.* B) Gráfico de enriquecimento do *GO:0005549 Odorant Binding*. Fonte: elaborado pelo autor, 2022.

Tetrápodes utilizam o olfato para diversas tarefas, tais como percepção do ambiente, busca por alimento, detecção de predadores, demarcação de território, comunicação e reprodução, onde a percepção olfatória desempenha um papel essencial na escolha do parceiro sexual em diversas espécies. Nesse cenário, a ocorrência de pressão seletiva para a variação dessa função biológica pode ter contribuído para diversificar a gama de odores percebidos por estes animais e promover a adaptação aos seus estilos de vida.

5.1.2.2 - GO:0008527 *Taste Receptor Activity*

Este GO anota 22 genes, sendo 20 deles do conjunto de ponta, possui um valor de FDR <2,2e-16, um valor de escore de enriquecimento normalizado de -4,9752 e é definido pelo AmiGO como “Combinando com compostos solúveis para iniciar uma mudança na atividade celular. Estes receptores são responsáveis pelo sentido do paladar” (Figura 32).

Figura 32 - Gráfico de enriquecimento *Taste Receptor Activity* - Tetrapoda

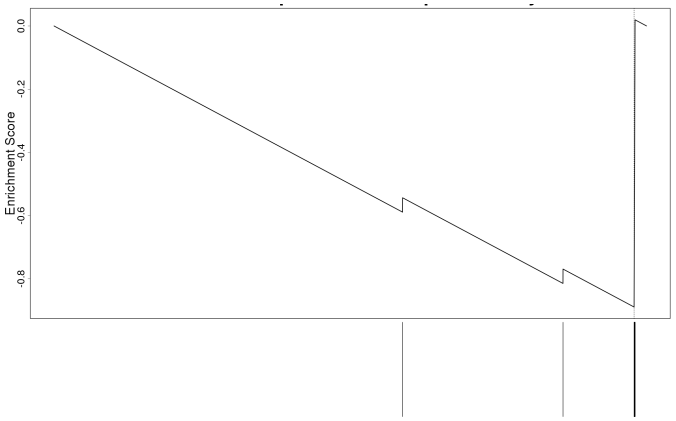


Gráfico de enriquecimento do GO:0008527 *Taste Receptor Activity.* Fonte: elaborado pelo autor, 2022.

Dentre os genes do conjunto de ponta, são encontrados membros da família de receptores *Taste* 2 (TAS2R), que estão agrupados em um único grupo homologo com identidade de alinhamento de 25%. Essas proteínas são responsáveis pela detecção do sabor amargo, e sua grande variação reflete as adaptações às diferentes dietas encontradas nos tetrápodes (A BACHMANOV *et al*., 2014.). Se tomarmos como exemplos animais carnívoros, frugívoros, herbívoros, estes obtém seus diferentes nutrientes a partir de dietas distintas com ampla variação na forma de obtenção destes nutrientes. A escolha dos alimentos em Tetrapoda é sabidamente orientada pelo sabor, tendo o sabor amargo um papel fundamental, pois está associado com alimentos que podem ser tóxicos ou inadequados para o consumo, especialmente em herbívoros. Não surpreendentemente, muitas plantas produzem compostos amargos como um mecanismo de defesa contra predação, causando uma grande pressão seletiva para a variação de TAS2R nos animais herbívoros (Wooding, Ramirez & Behrens, 2021).

5.1.3 - Processo Biológico

Os termos mais enriquecidos nos processos biológicos são GO:0007606 *Sensory Perception of Chemical Stimulus* e GO:0050906 *Detection of Stimulus Involved in Sensory Perception*, que anotam genes de receptores olfatórios e receptores gustativos descritos no tópico 5.1.2. Novamente, observamos diversos termos associados a mecanismos imunes: GO:0002449 *Lymphocyte Mediated Immunity* GO:0006959, GO:0072376 *Protein Activation Cascade* e GO:0050727 *Regulation of Inflammatory Response*. Também observamos um termo que pode

estar relacionado ao processo de diferenciação dos tetrápodes, conforme discutido na próxima seção.

5.1.3.1 - GO:0007586 Digestion

É definido pelo AmiGO como “O conjunto de processos físicos, químicos e bioquímicos realizados por organismos multicelulares para decompor os nutrientes ingeridos em componentes que podem ser facilmente absorvidos e direcionados para o metabolismo.”. Anota 97 genes, sendo 69 do conjunto de ponta (Figura 33). Possui um valor de FDR de <2,2e-16 e o valor de -3,5427 de escore de enriquecimento normalizado. Dentre os genes anotados por esse GO temos: mucin 6, *oligomeric mucus/gel-forming* (MUC6), *gastrokine* 1 (GKN1), *trefoil factor* 1, 2 e 3 (TFF1, TFF2 e TFF3).

Figura 33 - Gráfico de enriquecimento *Digestion* - Tetrapoda

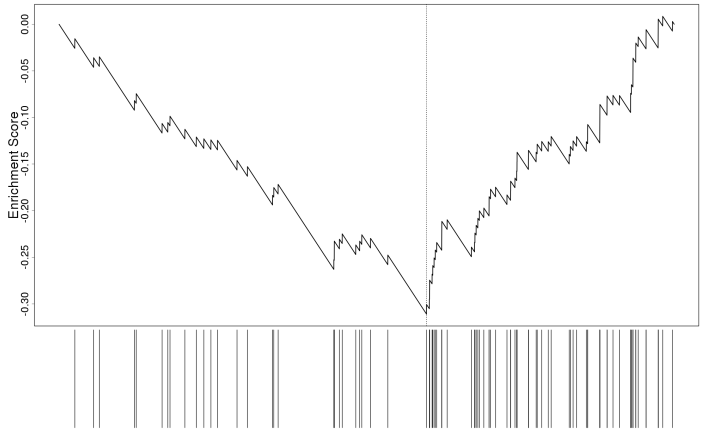


Gráfico de enriquecimento do GO:0007586 *Digestion*. Fonte: elaborado pelo autor, 2022.

Como discutido anteriormente, tetrápodes apresentam ampla variação nas formas de obtenção de alimento, possuindo diversas dietas. O aparelho gastro-intestinal passou por diversas adaptações, com diversas adaptações às diferentes dietas, bem como às injúrias que podem ser provocadas pela alimentação.

MUC6 possui um valor de identidade de alinhamento de 13%, sendo a proteína mais variável deste GO. É um constituinte da camada protetora da superfície gástrica (Cobler, Garrido & DE BOLÓS, 2010), uma mucosa que confere proteção mecânica, além de formar uma barreira biológica contra diversos organismos invasores.

Gastrokine 1 é uma proteína resistente a proteases encontrada na mucosa do estômago e intestinos dos animais. Sua função não ainda não foi totalmente

elucidada, mas Overstreet e colaboradores (2021) acreditam que GKN1 desempenha um papel mediador dos efeitos do estômago no microbioma, no metabolismo ou no ganho de peso corporal. Esta proteína possui um valor de identidade de alinhamento de 25%, evidenciando uma possível pressão seletiva para variação que pode estar relacionada com adaptações às diferentes dietas em tetrápodes.

A família gênica *Trefoil Factor* é formada pelos genes TFF1, TFF2 e TFF3, que formam um grupo homologo com 28% de identidade de alinhamento entre os tetrápodes. Os TFF's também são encontrados nas mucosas, sendo TFF1 e TFF2 mais encontrados no estômago e TFF3 no intestino. Atuam na restituição da mucosa (migração celular), modulando as junções celulares, a apoptose, a angiogênese e os processos de diferenciação da mucosa (Emidio *et al.*, 2019). Não encontramos evidência de que a variação dessa família gênica em Tetrapoda possa estar causalmente associada à imensa variação morfológica e funcional observada no sistema digestivo de Tetrapoda. Os resultados que reportamos sugerem que a avaliação funcional desta família multigênica nesse grupo pode contribuir para a compreensão da evolução funcional do sistema digestivo.

5.1.4 - KEGG

O uso do banco de dados KEGG para análises de GSEA permite encontrar vias bioquímicas enriquecidas em genes variáveis. Observamos a ocorrência de vias conceitualmente similares a termos GO enriquecidas nessa classe de genes, tal como *Olfactory Transduction*, que está relacionada com os GO's de *Odorant Binding* e *Olfactory Receptor Activity*. Também foram encontradas vias relacionadas ao sistema imune, tais como *Cytokine-Cytokine Receptor Interaction, Allograft Rejection, Autoimmune thyroid disease* e *Graft-versus-host disease*, onde encontramos diversos genes do sistema MHC, interleucinas, interferons e imunoglobulinas.

As vias *Steroid Hormone Biosynthesis, Chemical Carcinogenesis*, *Metabolism of Xenobiotics by Cytochrome P450*, *Drug Metabolism e Retinol metabolism* possuem tanto exemplos de enzimas variáveis e exclusivas de cada via como compartilham dois grupos de genes homólogos variáveis observados em todas: membros das famílias *UDP glucuronosyltransferase family* 1 e 2 (UGT1 e 2), e membros da superfamília do citocromo P450 (CYP). Conforme descrevemos abaixo, ambos possivelmente apresentam um importante papel adaptativo para a

detoxificação de xenobióticos abundantes em plantas, e possivelmente compreendem outro exemplo de corrida armamentista entre herbívoros (e seus agentes de detoxificação) e plantas (e seus compostos secundários).

UGTs são responsáveis por realizar o processo de glucuronidação, que envolve a transferência de um ácido glicurônico da molécula ácido UDP-glucurônico para a substância alvo. Isso cria uma molécula conjugada que é mais polar e, consequentemente, mais hidrossolúvel, facilitando a sua excreção. É essencial para a excreção de diversos xenobióticos advindos do ambiente ou da dieta. Enzimas do citocromo P450 são responsáveis pela metabolização de compostos endógenos, como hormônios esteróides e ácidos graxos, além de um importante papel na desintoxicação de xenobióticos. A sua principal função é a oxidação, ou seja, adicionar um átomo de oxigênio na molécula alvo (McDonnell & Dang, 2013). Quantitativamente, as famílias multigênicas do citocromo P450 e das UGTs compreendem as duas enzimas mais importantes para conjugação de xenobiótios (Rowland, Miners & Mackenzie, 2013).

5.1.4.1 - *Retinol Metabolism*

A via de metabolismo de retinol, também conhecido como Vitamina A, anota 60 genes, sendo 54 deles do conjunto de ponta, possui um valor FDR <2,2e-16 e escore de enriquecimento normalizado de -4,4404 (Figura 34). Além dos genes de UGT's e Citocromo P450, essa via KEGG também anota proteínas específicas da via, como as *retinol dehydrogenases 5, 8, 11 e 12* e 16 (RDH*), *alcohol dehydrogenase* 1A, 1B, 1C, 4, 5, 6 e 7 (ADH*), *dehydrogenase/reductase 9* (DHRS9), *hydroxysteroid 17-beta dehydrogenase* 6 (HSD17B6), *short chain dehydrogenase/reductase family 16C member 5* (SDR16C5), *acyl-CoA wax alcohol acyltransferase 2* (AWAT2) e RPE65 - *retinoid isomerohydrolase.*

Figura 34 - Gráfico de enriquecimento *Retinol Metabolism* - Tetrapoda

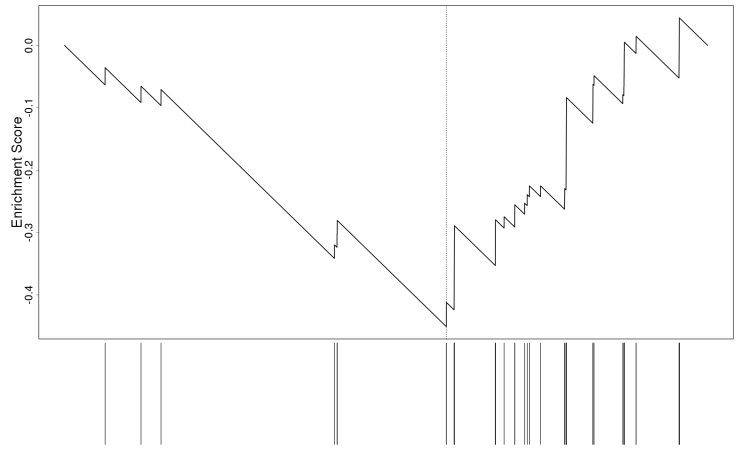


Gráfico de enriquecimento da via *Retinol Metabolism.* Fonte: elaborado pelo autor, 2022.

A detecção de luz pelos animais é feita pela associação entre proteínas opsinas nas celulas bastonetes e fotopsinas nas células cones e um derivado da vitamina A chamado 11-*cis*-retinal. Quando um fóton incide na molécula 11-cis-retinal este causa uma fotoisomerização – que é a mudança conformacional causada pela luz - mudando a molécula para para *11-trans*-retinal. Esta mudança conformacional desencadeia uma mudança conformacional das proteínas opsinas, o que inicia uma cascata de sinalização intracelular que culmina com a produção de um impulso nervoso. A molecula de *all-trans*-retinal é então convertida em 11-*cis*-retinal por processos sequenciais de redução, oxidação e isomeração (Parker & Crouch, 2010). Todas as enzimas envolvidas nesses processos também foram observadas com uma grande variação em Tetrapoda, sendo componentes desta via KEGG.

Especificamente, o processo de redução de *all-trans*-retinal para *all-trans*-retinol pode ser catalisado pelas enzimas ADH 1A, 1B, 1C, 4, 5, 6 e 7, RDH 8, 11, 12 e 16, DHRS9, HSD17B6 e SDR16C5. *All-trans*-retinol é então transformado em *all-trans*-retinil éster pela ação da enzima AWAT2, e depois é hidrolisado em 11-*cis*-retinol pela ação da RPE65. Por fim, 11-*cis*-retinol é oxidado em 11-*cis*-retinal pela ação das RDH 5 e 11.

A ocorrência de homólogos variáveis compreendendo todas as enzimas envolvidas no ciclo de renovação de 11-*cis*-retinal, uma molécula essencial para a visão em todos os tetrápodes, compreendem uma forte evidência da pressão seletiva para a variação dos mecanismos da visão. Além de condizente com o modo de vida desse grupo, acreditamos que nossos resultados compreendem o primeiro relato de seleção para variação desse mecanismo biológico em Tetrapoda.

5.2 - Mammalia

5.2.1 - Componente Celular

Os termos GO de Componente Celular enriquecidos na linhagem de mamíferos são similares ao de Tetrapoda, incluindo alguns termos repetidos como *Intermediate Filament Cytoskeleton,* MHC *Protein Complex e Cornified Envelope*. Uma vez que uma fração considerável dos genomas de Tetrapoda analizados são de mamíferos (108 de 198), esse resultado reflete também a composição relativa dos grupos de genomas utilizados para representar estes taxa.

Na categoria de MHC *Protein Complex,* os mesmos genes encontrados na linhagem dos tetrápodes foram reportados na linhagem dos mamíferos, e sua variação pode ser explicada da mesma forma que nos Tetrapoda. O termo GO - *Intermediate Filament Cytoskeleton* anota diversos genes de queratina e proteínas associadas a queratina (keratin-associated proteins (KRTAPs)) cuja função foi discutida no tópico 5.1.1.4.

5.2.1.1 - GO:0031225 *Anchored Component of Membrane*

Esse GO é definido pelo AmiGO como “O componente de uma membrana que consiste nos produtos gênicos que estão presos à membrana apenas por uma âncora ligada covalentemente, como um grupo lipídico que está embutido na membrana. Os produtos gênicos com sequências peptídicas embutidas na membrana são excluídos deste agrupamento”. Anota 157 genes, sendo 100 deles do conjunto de ponta, possui um valor de FDR <2,2e-16 e um escore de enriquecimento normalizado de -4,1832 (Figura 35). Esse GO anota genes como *tectorin alpha* (TECTA), com uma identidade de 28%, que já foi descrito no tópico 5.1.1.1, *carcinoembryonic antigen related cell adhesion molecule 5, 6, 7 e 8* (CEACAM5, 6, 7 e 8)*, Fc ragmente of IgG receptor III* (FCGR3B) e diversos genes da família serino-proteases, que são a maior família de proteases dos mamíferos e além de possuírem função no sistema imune também atuam nos processos de digestão e coagulação (Heutinck *et al.*, 2010).

Figura 35 – Gráfico de enriquecimento *Anchored Component of Membrane* – Mammalia

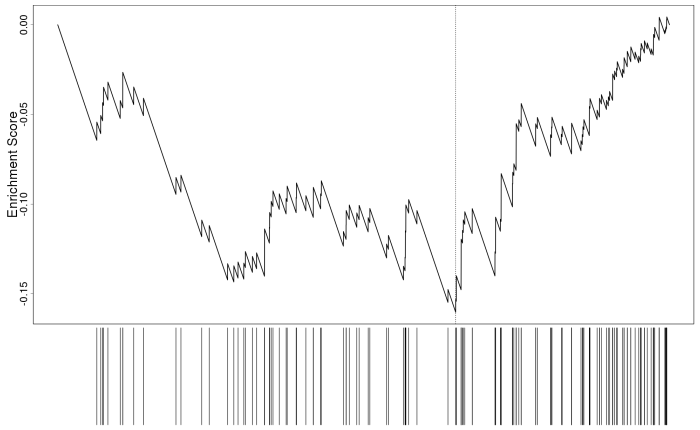


Gráfico de enriquecimento do GO:0031225 Anchored Component of Membrane. Fonte: elaborado pelo autor, 2022.

CEACAM's possuem um valor de identidade de alinhamento de 15% e são proteínas que pertencem a superfamília das imunoglobulinas, e pode atuar como uma molécula de adesão celular e como um componente do sistema imune no reconhecimento de patógenos (Beauchemin & Arabzadeh, 2013).

O gene FCGR3B é um mecanismo do sistema imune, tendo identidade de alinhamento de 16%, que pode desencadear reações de fagositose e o processamento de complexos do sistema imunológico. É expresso principalmente nos neutrófilos e está envolvido no recrutamento de neutrófilos polimorfonucleares para os locais de inflamação. A diminuição no número de cópias de FCGR3B está associada com doenças autoimunes (Lee *et al*., 2015).

5.2.2 - Função Molecular

Virtualmente, todos termos GO encontrados enriquecidos negativamente na linhagem dos mamíferos foram encontrados na linhagem dos tetrápodes, como os termos *Olfactory Receptor Activity, Odorant Binding*, *Taste Receptor Activity*, *Antigen Binding* e *Cytokine Receptor Activity*.

5.2.3 - Processo Biológico

Os termos GO enriquecidos negativamente nesta categoria, majoritariamente estão relacionados a processos do sistema imune, tais como *Cell Killing, Protein Activation Cascade, Organ or Tissue Specific Immune Response, Response to Chemokine* e *Cellular Defense Response.* Estes termos anotam diversas proteínas

do sistema imune, tais como imunoglobulinas, proteínas do sistema complemento, defensinas e receptores de quimiocinas e interleucinas.

5.2.4 - KEGG

As vias KEGG enriquecidas em genes variáveis apresentaram o mesmo padrão observado para os termos GO, e a maioria das vias enriquecidas em tetrápodes estão enriquecidas na linhagem dos mamíferos. Menciona-se as vias *Olfactory Transduction* - que também anota receptores olfatórios -, e as vias *Drug Metabolism, Metabolism of Xenobiotics by Cytochrome P450 e Chemical Carcinogenesis* - que também anotam genes de UDP Glucuronosiltransferase e genes do Citocromo P450. As vias *Cytokine-Cytokine Receptor Interaction*, *Hematopoietic Cell Lineage* e *Graft-Versus-Host Disease* anotam diversos genes do sistema imune, tais como imunoglobulinas, genes de MHC, interleucinas e sistema complemento. Também encontramos a via *Retinol Metabolism* enriquecida em genes variáveis em mamíferos; novamente, todas as enzimas necessárias para a renovação do 11-*cis-retinal* fazem parte da lista.

5.3 - Aves

De maneira geral, encontramos menos termos GO e KEGG enriquecidos com escore negativo na linhagem de aves ao compararmos com as análises de Tetrapoda e Mammalia. Acreditamos que essa observação não é causada por falta de processos biológicos enriquecidos em genes, e sim pela diferença na qualidade das anotações dos genes/proteínas da espécie âncora *Gallus gallus* comparada à espécie-âncora utilizada nas análises anteriores (*Homo sapiens*). De fato, ao utilizarmos os identificadores de galinha para ancorar as análises de Tetrapoda, não encontramos nenhum termo enriquecido com escore negativo (resultados não mostrados), o que sugere que diferenças na qualidade da anotação funcional destes genomas podem ser responsável por algumas das diferenças observadas.

5.3.1 - Componente Celular

5.3.1.1 - GO:0045111 *Intermediate Filament Cytoskeleton*

Este GO, que já foi definido no tópico 5.1.1.4, em aves possui o escore de enriquecimento normalizado de -5,4836 um valor de FDR <2,2e-16 e anota diversos genes de alfa e beta queratinas e *Fas Binding Factor* 1 (FBF1) (Figura 36).

Figura 36 - Gráfico de enriquecimento *Intermediate Filament Cytoskeleton* - Aves

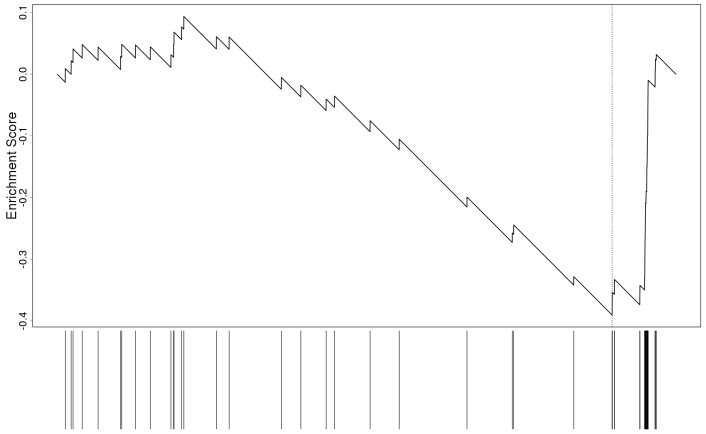


Gráfico de enriquecimento do GO:0045111 - *Intermediate Filament Cytoskeleton.* Fonte: elaborado pelo autor, 2022.

Nas aves, estruturas como bicos, garras, escamas e penas são formados pela junção das famílias multigênicas de alfa e beta queratinas, sendo as beta queratinas uma família gênica exclusiva da linhagem dos sauropsídeos (Greenwold & Sawyer, 2013; Greenwold *et al.,* 2014; Ho *et al.,* 2022). As beta queratinas específicas das penas de aves, como os genes *Feather Keratin* e *Feather Keratin likes,* representam a maioria das β-queratinas totais, e estão entre os homólogos mais variáveis da lista (identidade de 44%). Em algumas linhagens, por exemplo, as aves de rapina, existe uma maior proporção de beta-queratinas de garras. Durante o desenvolvimento das escamas e penas das galinhas, quando observamos as alfa e beta queratinas, percebemos que, embora as alfa queratinas também estejam presentes nesses tecidos, a quantidade e a intensidade das betas queratinas são muito maiores (Greenwold *et.al.*, 2014). Portanto, as expansões e modificações encontradas nas famílias de beta queratinas de aves compreendem um importante processo evolutivo nas aves que permitiu a ocupação de diferentes nichos e possibilitou diferentes estilos de vida (Li *et al., 2013*; Greenwold *et al.,* 2014*)*.

O gene FBF1 possui um valor de identidade de alinhamento de 40%, e compreende aquele com a maior variabilidade dentre os genes anotados pelo termo GO:0045111. FBF1, também chamado de *Albatross*, e um gene ligante de queratina importante para o processo de polarização de células epiteliais em humanos (Sugimoto *et al.,*2008). Além disso, Albatross/FBF1 é uma proteína do apêndice distal de centríolos e é essencial para o processo de ciliogênese em humanos (Tanos *et al.,* 2013). Também já foi demonstrado que, em humanos, o gene FBF1 contribui para os processos tanto de duplicação dos centríolos quanto para separação dos centrossomos (Inoko *et al.,* 2018). Entretanto, cabe ressaltar que nosso estudo demonstrou que, enquanto em aves Albatross/FBF1 é o gene com maior variação dentre deste GO (0.40), em mamíferos ele apresenta uma maior conservação (0.67). Portanto, embora as análises em humanos possam fornecer pistas iniciais para inferir a função de FBF1 em aves, ainda são necessários mais estudos para compreender melhor a função deste gene variável em aves.

5.3.1.2 - GO:0005887 *Integral Component Of Plasma Membrane*

Este termo anota 444 genes, sendo 255 deles do conjunto de ponta (Figura 37). Possui um valor FDR <2,2e-16 e escore de enriquecimento normalizado de -2,9675. Dentre os genes deste GO podemos citar *adenylate cyclase* 3 (ADCY3), *glucagon receptor* (GCGR), *opsin pineal* (OPNP) e *vertebrate ancient opsin* (OPNVA).

Figura 37 - Gráfico de enriquecimento *Integral Component Of Plasma Membrane* - Aves

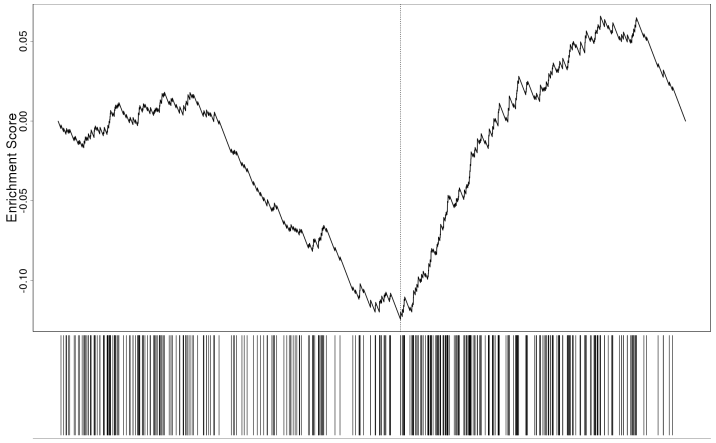


Gráfico de enriquecimento do GO:0045111 - *Integral Component Of Plasma Membrane.* Fonte: elaborado pelo autor, 2022.

A enzima adenilato ciclase 3 catalisa a formação de cAMP, e por isso faz parte da cascata de sinalização desencadeadas por receptores odoríferos e é essencial para a percepção olfatória, possuindo uma identidade de alinhamento de 16% (Sklar, Anholt & Snyder, 1986). Existem evidências que o ADCY3 também pode modular a sensibilidade à insulina, causando um aumento de glicose no sangue (Tong *et al*., 2016). *Glucagon receptor* é ativada pelo glucagon, um hormônio produzido no pâncreas cuja principal função é aumentar o nível de glicose no sangue. A ativação de GCGR causa uma ativação de adenilato ciclases, resultando em uma cascata de ativação que promove a gliconeogênese e o aumento da glicose no sangue (Janah *et al*., 2019). Interessantemente, o nível de glicose no plasma de aves é maior do que outros vertebrados que possuem uma massa corporal similar, apesar de não se saber ao certo quais os eventos evolutivos que selecionaram este fenótipo, mas pode estar relacionado ao grande gasto energético requerido para o voo. Adicionalmente, já foi demonstrado que pombos apresentam aumento no nível de glicose no sangue durante o período de cortejo e acasalamento, podendo estar relacionado ao sucesso reprodutivo desta espécie (Braun & Sweazea, 2008).

Os genes *Opsin pineal* e *vertebrate ancient opsin* codificam opsinas que possuem 50% de identidade de alinhamento. Conforme exposto anteriormente, estas proteínas se ligam com 11-*cis*-retinal e fazem a percepção sensorial de luz. Entretanto, estas opsinas não são expressas nas células cone ou bastonetes, mas na glândula pineal e no hipotálamo. O crânio e o cérebro das aves possuem permeabilidade à luz, permitindo que uma grande quantidade de fótons penetre profundamente no cérebro (Halford *et al*., 2009). OPNVA está relacionado com o fotoperiodismo em aves (Davies *et al*., 2012), alterações no fotoperíodo podem servir como um sinal para aves migratórias realizarem sua migração sazonal, e regular o período reprodutivo de algumas aves (Farner, 1964; Sharp, 2005). Nossos resultados sugerem que este processo biológico possa ser alvo de seleção diversificadora.

5.3.1.3 - GO:003125 *Anchored Component of Membrane*

Este termo, definido no tópico 5.2.1.1, possui um valor de FDR 0,00083412, escore de enriquecimento normalizado de -2,5116 e anota genes como: *lymphocyte antigen 6 family member* E (LY6E, LY6CLEL e LOC101747995), *folate receptor 1* (FOLR1), e alguns genes que atuam no desenvolvimento do sistema nervoso central:

GDNF *family receptor alpha* 1, 2 e 4 (GFRA1, GFRA2, GFRA4), *contactin* 5 e 6 (CNTN5, CNTN6), *neuronal growth regulator* 1 (NEGR1) (Figura 38).

Figura 38 - Gráfico de enriquecimento *Anchored Component of Membrane* - Aves

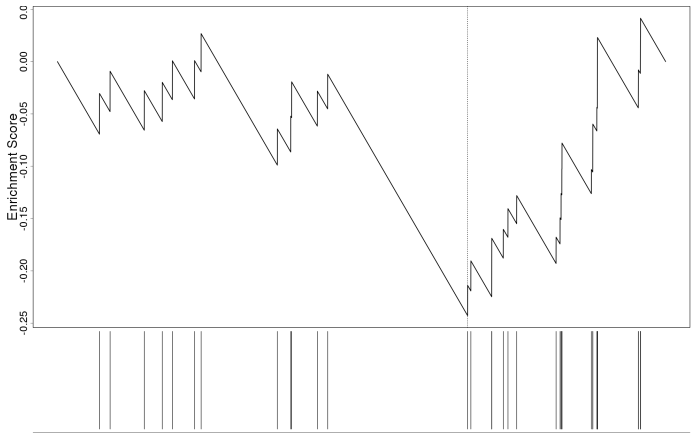


Gráfico de enriquecimento do GO:003125 - Anchored Component of Membrane. Fonte: elaborado pelo autor, 2022.

Em aves, *Lymphocyte antigen 6 family member E* (LY6E, LY6CLEL e LOC101747995) apresentam a maior variabilidade dentre os genes do GO:003125, possuindo uma identidade de alinhamento de 43%, a qual pode ser explicada devido a sua interação com respostas imunes de infecções virais (Yu & Liu, 2020). Em galinhas, já foi demonstrado que LY6E está relacionado à resistência à doença de Marek (Liu *et al.,* 2003), que é uma doença viral altamente contagiosa que afeta aves, especialmente galinhas, causada pelo *Gallid alphaherpesvirus 2*, também chamado de *Marek's disease virus* (MDV), um tipo de herpesvírus que pertence à família Herpesviridae. Essa doença é caracterizada por causar tumores malignos em tecidos nervosos, órgãos internos e, ocasionalmente, na pele das aves infectadas.

Folate Receptor 1 (FOLR1) e *Riboflavin Binding Protein* (RBP) são proteínas que fazem parte do mesmo grupo de homólogos e possuem uma identidade de alinhamento de 44%, sendo as segundas mais variáveis anotadas pelo termo GO:003125. Ambos são ligantes de membros da família da vitamina B (B9 e B2, respectivamente). B9 é um importante doador de grupo metil, portanto é necessário para o processo de manutenção e divisão celular, sendo particularmente importante para células de rápida divisão (Jones *et al.,* 2017). Interessantemente, estudos anteriores demonstraram que FOLR1 é superexpresso em tecidos cancerígenos (Narcela *et al., 2015;* Kim *et al., 2018*; Nawaz & Kipreos, 2022). A variação em aves observada em FOLR1 pode ajudar a explicar o fato de que aves apresentam menor

prevalência de câncer quando comparados a mamíferos e demais répteis (Ujvari, Roche & Thomas, 2017; Boddy, 2020).

GDNF *family receptor alpha* (55% de identidade) são proteínas cruciais para o desenvolvimento e manutenção de conjuntos distintos de neurônios centrais e periféricos (Airaksinen, Holm & Hätinen, 2006), enquanto *contactin* 5 e 6 (61% de identidade) constituem uma família de proteínas de adesão celular que promovem a criação e preservação das conexões entre neurônios. NEGR1 (69% de identidade) regula o crescimento de processos neuronais e a formação de sinapses (Noh *et al*., 2019). As aves apresentam uma maior massa cerebral e uma maior densidade neuronal que outros Sauropsida de igual massa total, e também apresentam comportamento complexos muitas vezes comparados aos de primatas, como comportamentos sociais, reconhecimento de imagens, o reconhecimento de sua própria imagem no espelho, a resolução de problemas e o uso de ferramentas (Emery, 2006; Emery & Clayton, 2009). Estas características podem ter contribuído para a sobrevivência das aves em relação aos demais dinossauros, bem como pela sua diversificação durante a fronteira Cretáceo-Paleogeno (Torres, Norell & Clarke, 2021).

5.3.2 - Função Molecular

Dois termos de função molecular foram encontrados enriquecidos negativamente em Aves: *Structural Molecule Activity* e *Receptor Regulator Activity*. O termo *Structural Molecule Activity* possui como gene mais variável *tectorin alpha* (TECTA), discutido no tópico 5.1.1.1, e também observado em mamíferos, embora em aves ele possua uma variação ainda maior (18% de identidade, versus 28% em mamíferos). Neste GO também observamos diversos genes de alfa e beta queratinas, já discutidos no tópico 5.3.1.1. Diversos genes relacionados a processos do sistema imune são anotados pelo GO *Receptor Regulator Activity*, como ligantes de quimiocinas, interleucinas e defensinas.

5.3.3 - Processo Biológico

Os termos enriquecidos nesta categoria possuem genes que estão anotados com mais de um termo GO de processo biológico enriquecido. Genes das famílias *taste 2*, *olfactory receptor e opsins* são encontrados nos GO's *Nervous System Process, G Protein-Coupled Receptor Signaling Pathway* e *Detection of Stimulus.*

Estes genes relacionados com percepção sensorial estão relacionados com diversos comportamentos, como busca por alimento, percepção e resposta ao ambiente, busca por parceiros e reprodução. Membros da família *taste 2* também são anotados no GO *Response to Xenobiotic Stimulus*, juntamente com diversos genes da superfamília *Cytochrome P450,* essencial para metabolizar xenobióticos adquiridos principalmente durante a alimentação.

5.3.4 - KEGG

As vias KEGG enriquecidas negativamente em aves são similares às encontradas nas linhagens dos tetrápodes e mamíferos. Destacamos as vias relacionadas a metabolismos de xenobióticos como *Drug Metabolism* e *Metabolism of Xenobiotics by Cytochrome P450*, vias relacionadas a sistema imune , *Cytokine-Cytokine Receptor Interaction* e a via de *Retinol Metabolism*, que é essencial para a visão.

6 - CONCLUSÃO

O pequeno conjunto de genes com pressão seletiva para sua variação é comumente denominado de genes adaptativos, uma vez que estes participam dos processos biológicos onde a produção de novidades evolutivas causa aumento de aptidão evolutiva. Entretanto, genes raramente agem sozinhos: ao contrário, estes compreendem agentes moleculares coletivos que participam de funções biológicas e vias bioquímicas comuns. Desse modo, a busca por temas biologicamente relevantes enriquecidos em genes variáveis compreende uma importante ferramenta para a compreensão molecular e funcional dos processos evolutivos.

Nesse trabalho, desenvolvemos e validamos uma metodologia de genômica comparativa (disponível em *https://github.com/ThieresTMS/ENHYDRA*) que permite detectar simultaneamente genes potencialmente adaptativos que possuem temas comuns, tais como a sua participação nos mesmos processos biológicos ou nas mesmas vias bioquímicas. Nas análises de tetrápodes, mamíferos e aves que utilizamos como estudos de caso, encontramos diversos grupos funcionais de genes previamente associados relacionados a processos adaptativos, tais como componentes do sistema imune e elementos dos sistemas sensoriais, tais como receptores olfatórios, de sabor e visão, demonstrando como nossa estratégia produz resultados qualitativamente comparáveis às metodologias disponíveis. Adicionalmente, observamos temas biológicos com evidência de evolução acelerada que, até onde conseguimos averiguar, não haviam sido descritos anteriormente e podem compreender importantes alvos para caracterização experimental, tais como processos digestivos e desenvolvimento de estruturas epiteliais.

7 - PERSPECTIVAS

Como melhorias de nossa abordagem computacional, pretendemos implementar a possibilidade de se realizar as análises de enriquecimento de forma local, bem como de se utilizar uma lista própria de anotações funcionais dos genes. Desse modo, torna-se possível utilizar anotações funcionais customizadas, bem como utilizar organismos não-modelo para ancorar as análises e estudar grupos taxonômicos que não possuam espécies-modelo filogeneticamente próximas. O uso de anotações próprias também permitiria contornar problemas como os descritos anteriormente, onde há diferença considerável na qualidade da anotação funcional de organismos-modelo (*H. sapiens* e *G. gallus*).

Em relação às análises biológicas, pretendemos implementar a possibilidade de análises comparativas entre dois grupos de genomas, de modo a averiguar quais genes/categorias biológicas estão variando mais em determinados grupos de organismos quando a outro grupo (*e.g.* morcegos em relação aos demais mamíferos).

8 - REFERÊNCIAS

A BACHMANOV, Alexander et al. Genetics of taste receptors. **Current pharmaceutical design**, v. 20, n. 16, p. 2669-2683, 2014.

AIRAKSINEN, Matti S.; HOLM, Liisa; HÄTINEN, Tuomas. Evolution of the GDNF family ligands and receptors. **Brain Behavior and Evolution**, v. 68, n. 3, p. 181-190, 2006.

ÁLVAREZ-CARRETERO, Sandra; KAPLI, Paschalia; YANG, Ziheng. Beginner's guide on the use of PAML to detect positive selection. **Molecular Biology and Evolution**, v. 40, n. 4, p. msad041, 2023.

ASHBURNER, Michael et al. Gene ontology: tool for the unification of biology. **Nature genetics**, v. 25, n. 1, p. 25-29, 2000.

AMAGAI, Masayuki; STANLEY, John R. Desmoglein as a target in skin disease and beyond. **Journal of Investigative Dermatology**, v. 132, n. 3, p. 776-784, 2012.

BERKEL, Simone et al. Inherited and de novo SHANK2 variants associated with autism spectrum disorder impair neuronal morphogenesis and physiology. **Human molecular genetics**, v. 21, n. 2, p. 344-357, 2012.

BI, Xupeng et al. Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. **Cell**, v. 184, n. 5, p. 1377-1391. e14, 2021.

BIRCHLER, James A.; YANG, Hua. The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. **The Plant Cell**, v. 34, n. 7, p. 2466-2474, 2022.

BODDY, Amy M. et al. Lifetime cancer prevalence and life history traits in mammals. **Evolution, medicine, and public health**, v. 2020, n. 1, p. 187-195, 2020.

BRAUN, Eldon J.; SWEAZEA, Karen L. Glucose regulation in birds. **Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology**, v. 151, n. 1, p. 1-9, 2008.

BROWN, Nicholas H. Spectraplakins: the cytoskeleton's Swiss army knife. **Cell**, v. 135, n. 1, p. 16-18, 2008.

CAPELLA-GUTIÉRREZ, Salvador; SILLA-MARTÍNEZ, José M.; GABALDÓN, Toni. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, v. 25, n. 15, p. 1972-1973, 2009.

CARBON, Seth et al. AmiGO: online access to ontology and annotation data. **Bioinformatics**, v. 25, n. 2, p. 288-289, 2009.

CARROLL, Lewis. **Aventuras de Alice no país das maravilhas; através do espelho e o que Alice encontrou lá e outros textos**. Grupo Editorial Summus, 1980.

CHAMPLIAUD, Marie-France et al. Gene characterization of sciellin (SCEL) and protein localization in vertebrate epithelia displaying barrier properties. **Genomics**, v. 70, n. 2, p. 264-268, 2000.

COBLER, Lara; GARRIDO, Marta; DE BOLÓS, Carme. MUC6 (mucin 6, oligomeric mucus/gel-forming). **Atlas Genet Cytogenet Oncol Haematol**, v. 14, n. 10, p. 918-922, 2010.

DAVIES, Wayne IL et al. Vertebrate ancient opsin photopigment spectra and the avian photoperiodic response. **Biology letters**, v. 8, n. 2, p. 291-294, 2012.

DE-OLIVEIRA-NOGUEIRA, Carlos Henrique et al. Between fruits, flowers and nectar: The extraordinary diet of the frog Xenohyla truncata. **Food Webs**, v. 35, p. e00281, 2023.

EIZAGUIRRE, Christophe et al. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. **Nature communications**, v. 3, n. 1, p. 621, 2012.

EMERY, Nathan J. Cognitive ornithology: the evolution of avian intelligence. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 361, n. 1465, p. 23-43, 2006.

EMERY, Nathan J.; CLAYTON, Nicola S. Tool use and physical cognition in birds and mammals. **Current opinion in neurobiology**, v. 19, n. 1, p. 27-33, 2009.

EMIDIO, Nayara Braga et al. Trefoil factor family: Unresolved questions and clinical perspectives. **Trends in biochemical sciences**, v. 44, n. 5, p. 387-390, 2019.

EMMS, David M.; KELLY, Steven. STRIDE: species tree root inference from gene duplication events. **Molecular biology and evolution**, v. 34, n. 12, p. 3267-3278, 2017.

EMMS, David M.; KELLY, Steven. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome biology**, v. 20, p. 1-14, 2019.

FARNER, Donald S. The photoperiodic control of reproductive cycles in birds. **American Scientist**, v. 52, n. 1, p. 137-156, 1964.

FAY, Richard R.; POPPER, Arthur N. Evolution of hearing in vertebrates: the inner ears and processing. **Hearing research**, v. 149, n. 1-2, p. 1-10, 2000.

FELSENSTEIN, J. et al. The newick tree format, 1986. **URL: http://evolution. genetics. washington. edu/phylip/newicktree. html**, 2000.

GAO, Xiangwei; XU, Zhengping. Mechanisms of action of angiogenin. **Acta biochimica et biophysica Sinica**, v. 40, n. 7, p. 619-624, 2008.

GREENWOLD, Matthew J.; SAWYER, Roger H. Molecular evolution and expression of archosaurian β-keratins: Diversification and expansion of archosaurian β-keratins and the origin of feather β-keratins. **Journal of Experimental Zoology Part B: Molecular and Developmental Evolution**, v. 320, n. 6, p. 393-405, 2013.

GREENWOLD, Matthew J. et al. Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles. **BMC evolutionary biology**, v. 14, n. 1, p. 1-16, 2014.

GOODWIN, Zane A.; DE GUZMAN STRONG, Cristina. Recent positive selection in genes of the mammalian epidermal differentiation complex locus. **Frontiers in genetics**, v. 7, p. 227, 2017.

HALFORD, Stephanie et al. VA opsin-based photoreceptors in the hypothalamus of birds. **Current Biology**, v. 19, n. 16, p. 1396-1402, 2009.

HEBEDA, Cristina B. et al. Annexin A1/formyl peptide receptor pathway controls uterine receptivity to the blastocyst. **Cells**, v. 9, n. 5, p. 1188, 2020.

HEDGES, S. Blair; DUDLEY, Joel; KUMAR, Sudhir. TimeTree: a public knowledge-base of divergence times among organisms. **Bioinformatics**, v. 22, n. 23, p. 2971-2972, 2006.

HEUTINCK, Kirstin M. et al. Serine proteases of the human immune system in health and disease. **Molecular immunology**, v. 47, n. 11-12, p. 1943-1955, 2010.

HO, Minh et al. Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders. **Human Genomics**, v. 16, n. 1, p. 1-21, 2022.

HONGO, Jorge A. et al. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. **BMC genomics**, v. 16, p. 1-16, 2015.

HONGO, Jorge Augusto et al. CALANGO: A phylogeny-aware comparative genomics tool for discovering quantitative genotype-phenotype associations across species. **Patterns**, v. 4, n. 6, 2023.

HOUNKPE, Bidossessi Wilfried et al. HRT Atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by

mining massive RNA-seq datasets. **Nucleic acids research**, v. 49, n. D1, p. D947-D955, 2021.

INOKO, Akihito et al. Albatross/FBF1 contributes to both centriole duplication and centrosome separation. **Genes to Cells**, v. 23, n. 12, p. 1023-1042, 2018.a

IWASAKI, Satoshi et al. Association of clinical features with mutation of TECTA in a family with autosomal dominant hearing loss. **Archives of Otolaryngology–Head & Neck Surgery**, v. 128, n. 8, p. 913-917, 2002.

JAN EJSMOND, Maciej; RADWAN, Jacek; WILSON, Anthony B. Sexual selection and the evolutionary dynamics of the major histocompatibility complex. **Proceedings of the Royal Society B: Biological Sciences**, v. 281, n. 1796, p. 20141662, 2014.

JONES, RoJenia N. et al. Expression and characterization of the zebrafish orthologue of the human FOLR1 gene during embryogenesis. **Gene Expression Patterns**, v. 25, p. 159-166, 2017.

KANEHISA, Minoru; GOTO, Susumu. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic acids research**, v. 28, n. 1, p. 27-30, 2000.

KATOH, Kazutaka; STANDLEY, Daron M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular biology and evolution**, v. 30, n. 4, p. 772-780, 2013.

KAUFMAN, Jim; SALOMONSEN, Jan; FLAJNIK, Martin. Evolutionary conservation of MHC class I and class II molecules—different yet the same. In: **Seminars in immunology**. Academic Press, 1994. p. 411-424.

KIM, Minsung et al. Folate receptor 1 (FOLR1) targeted chimeric antigen receptor (CAR) T cells for the treatment of gastric cancer. **PloS one**, v. 13, n. 6, p. e0198347, 2018.

KODAMA, Atsuko et al. ACF7: an essential integrator of microtubule dynamics. **Cell**, v. 115, n. 3, p. 343-354, 2003.

KOEPFLI, Klaus-Peter et al. The Genome 10K Project: a way forward. **Annu. Rev. Anim. Biosci.**, v. 3, n. 1, p. 57-111, 2015.

KOSIOL, Carolin et al. Patterns of positive selection in six mammalian genomes. **PLoS genetics**, v. 4, n. 8, p. e1000144, 2008.

KUMAR, Sudhir et al. TimeTree 5: an expanded resource for species divergence times. **Molecular Biology and Evolution**, v. 39, n. 8, p. msac174, 2022.

KYPRIOTOU, Magdalini; HUBER, Marcel; HOHL, Daniel. The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the ‘fused genes’ family. **Experimental dermatology**, v. 21, n. 9, p. 643-649, 2012.

LI, Junan; MAHAJAN, Anjali; TSAI, Ming-Daw. Ankyrin repeat: a unique motif mediating protein− protein interactions. **Biochemistry**, v. 45, n. 51, p. 15168-15178, 2006.

LI, Yang I. et al. Rapid evolution of beta-keratin genes contribute to phenotypic differences that distinguish turtles and birds from other reptiles. **Genome biology and evolution**, v. 5, n. 5, p. 923-933, 2013.

LIAO, Yuxing et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. **Nucleic acids research**, v. 47, n. W1, p. W199-W205, 2019.

LIM, Lina HK; PERVAIZ, Shazib. Annexin 1: the new face of an old molecule. **The FASEB Journal**, v. 21, n. 4, p. 968-975, 2007.

LIU, H.-C. et al. Identification of chicken lymphocyte antigen 6 complex, locus E (LY6E, alias SCA2) as a putative Marek’s disease resistance gene via a virus-host protein interaction screen. **Cytogenetic and genome research**, v. 102, n. 1-4, p. 304-308, 2003.

LIU, Yueyi et al. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. **Genome Research**, v. 14, n. 3, p. 451-458, 2004.

LE, Thong Minh et al. β2-microglobulin gene duplication in cetartiodactyla remains intact only in pigs and possibly confers selective advantage to the species. **PloS One**, v. 12, n. 8, p. e0182322, 2017.

LEE, Ho-Jin; ZHENG, Jie J. PDZ domains and their binding partners: structure, specificity, and modification. **Cell communication and Signaling**, v. 8, n. 1, p. 1-18, 2010.

LEE, Young Ho et al. Association between FCGR3B copy number variations and susceptibility to autoimmune diseases: a meta-analysis. **Inflammation Research**, v. 64, p. 983-991, 2015.

MAK, T. W.; SAUNDERS, M. E. 10—MHC: the major histocompatibility complex. **The Immune Response**, p. 247-277, 2006.

MANNI, Mosè et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic,

and viral genomes. **Molecular biology and evolution**, v. 38, n. 10, p. 4647-4654, 2021.

MCDONNELL, Anne M.; DANG, Cathyyen H. Basic review of the cytochrome p450 system. **Journal of the advanced practitioner in oncology**, v. 4, n. 4, p. 263, 2013.

MILINSKI, Manfred. The major histocompatibility complex, sexual selection, and mate choice. **Annu. Rev. Ecol. Evol. Syst.**, v. 37, p. 159-186, 2006.

MILLER, Webb et al. Comparative genomics. **Annu. Rev. Genomics Hum. Genet.**, v. 5, p. 15-56, 2004.

MOOTHA, Vamsi K. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. **Nature genetics**, v. 34, n. 3, p. 267-273, 2003.

NAWAZ, Fathima Zahra; KIPREOS, Edward T. Emerging roles for folate receptor FOLR1 in signaling and cancer. **Trends in Endocrinology & Metabolism**, 2022.

NECELA, Brian M. et al. Folate receptor-α (FOLR1) expression and function in triple negative tumors. **PloS one**, v. 10, n. 3, p. e0122209, 2015.

NICHOL, Donna; STUHLMANN, Heidi. EGFL7: a unique angiogenic signaling factor in vascular development and disease. **Blood, The Journal of the American Society of Hematology**, v. 119, n. 6, p. 1345-1352, 2012.

NOH, Kyungchul et al. Negr1 controls adult hippocampal neurogenesis and affective behaviors. **Molecular psychiatry**, v. 24, n. 8, p. 1189-1205, 2019.

OLENDER, Tsviya; LANCET, Doron; NEBERT, Daniel W. Update on the olfactory receptor (OR) gene superfamily. **Human genomics**, v. 3, p. 1-11, 2008.

OVERSTREET, Anne-Marie C. et al. Gastrokine-1, an anti-amyloidogenic protein secreted by the stomach, regulates diet-induced obesity. **Scientific Reports**, v. 11, n. 1, p. 9477, 2021.

PARKER, Ryan O.; CROUCH, Rosalie K. Retinol dehydrogenases (RDHs) in the visual cycle. **Experimental eye research**, v. 91, n. 6, p. 788-792, 2010.

PENN, Dustin J. The scent of genetic compatibility: sexual selection and the major histocompatibility complex. **Ethology**, v. 108, n. 1, p. 1-21, 2002.

PERRETTI, Mauro; D'ACQUISTO, Fulvio. Annexin A1 and glucocorticoids as effectors of the resolution of inflammation. **Nature Reviews Immunology**, v. 9, n. 1, p. 62-70, 2009.

POUGH, F. H.; JANIS, C. M.; HEISER, J. B. **A Vida dos Vertebrados**. 4ª edição. São Paulo: Atheneu, 684p, 2008.

RATTO, Fabrizia et al. Global importance of vertebrate pollinators for plant reproductive success: a meta‐analysis. **Frontiers in Ecology and the Environment**, v. 16, n. 2, p. 82-90, 2018.

ROWLAND, Andrew; MINERS, John O.; MACKENZIE, Peter I. The UDP-glucuronosyltransferases: their role in drug metabolism and detoxification. **The international journal of biochemistry & cell biology**, v. 45, n. 6, p. 1121-1132, 2013.

SAHM, Arne et al. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. **Nucleic Acids Research**, v. 45, n. 11, p. e100-e100, 2017.

SCHOCH, Conrad L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database**, v. 2020, p. baaa062, 2020.

SHARP, Peter J. Photoperiodic regulation of seasonal breeding in birds. **Annals of the New York Academy of Sciences**, v. 1040, n. 1, p. 189-199, 2005.

SHENDURE, Jay; AIDEN, Erez Lieberman. The expanding scope of DNA sequencing. **Nature biotechnology**, v. 30, n. 11, p. 1084-1094, 2012.

SIMÃO, Felipe A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210-3212, 2015.

SIMPSON, A. J. et al. Elafin (elastase-specific inhibitor) has anti-microbial activity against gram-positive and gram-negative respiratory pathogens. **FEBS letters**, v. 452, n. 3, p. 309-313, 1999.

SKLAR, P. B.; ANHOLT, R. R.; SNYDER, S. H. The odorant-sensitive adenylate cyclase of olfactory receptor cells. Differential stimulation by distinct classes of odorants. **Journal of Biological Chemistry**, v. 261, n. 33, p. 15538-15543, 1986.

SOUSA, Stefanie Oliveira de et al. ANNEXIN A1: Roles in Placenta, Cell Survival, and Nucleus. **Cells**, v. 11, n. 13, p. 2057, 2022.

SUBRAMANIAN, Aravind et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, v. 102, n. 43, p. 15545-15550, 2005.

SUGIMOTO, Masahiko et al. The keratin-binding protein Albatross regulates polarization of epithelial cells. **The Journal of cell biology**, v. 183, n. 1, p. 19-28, 2008.

SUN, Yan-Bo et al. Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. **Proceedings of the National Academy of Sciences**, v. 112, n. 11, p. E1257-E1262, 2015.

TANOS, Barbara E. et al. Centriole distal appendages promote membrane docking, leading to cilia initiation. **Genes & development**, v. 27, n. 2, p. 163-168, 2013.

TIFFNEY, Bruce H. Vertebrate dispersal of seed plants through time. **Annu. Rev. Ecol. Evol. Syst.**, v. 35, p. 1-29, 2004.

UJVARI, Beata; ROCHE, Benjamin; THOMAS, Frédéric (Ed.). **Ecology and evolution of cancer**. Academic Press, 2017.

VAN VALEN, Leigh. Molecular evolution as predicted by natural selection. **Journal of molecular evolution**, v. 3, p. 89-101, 1974.

VERHOEVEN, Kristien et al. Mutations in the human α-tectorin gene cause autosomal dominant non-syndromic hearing impairment. **Nature genetics**, v. 19, n. 1, p. 60-62, 1998.

VOGEL, Christine; CHOTHIA, Cyrus. Protein family expansions and biological complexity. **PLoS computational biology**, v. 2, n. 5, p. e48, 2006.

WATERHOUSE, Robert M. et al. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. **Nucleic acids research**, v. 41, n. D1, p. D358-D365, 2013.

WOODING, Stephen P.; RAMIREZ, Vicente A.; BEHRENS, Maik. Bitter taste receptors: Genes, evolution and health. **Evolution, Medicine, and Public Health**, v. 9, n. 1, p. 431-447, 2021.

YANG, Ziheng. PAML 4: phylogenetic analysis by maximum likelihood. **Molecular biology and evolution**, v. 24, n. 8, p. 1586-1591, 2007.

YU, Jingyou; LIU, Shan-Lu. Emerging role of LY6E in virus–host interactions. **Viruses**, v. 11, n. 11, p. 1020, 2019.

ZASLAVSKY, Kirill et al. SHANK2 mutations associated with autism spectrum disorder cause hyperconnectivity of human neurons. **Nature neuroscience**, v. 22, n. 4, p. 556-564, 2019.

ZHANG, Guojie et al. Comparative genomics reveals insights into avian

genome evolution and adaptation. **Science**, v. 346, n. 6215, p. 1311-1320, 2014.

ZHANG, Guojie. Bird sequencing project takes off. **Nature**, v. 522, n. 7554, p. 34-34, 2015.

ZHANG, Jamie; YUE, Jiping; WU, Xiaoyang. Spectraplakin family proteins–cytoskeletal crosslinkers with versatile roles. **Journal of cell science**, v. 130, n. 15, p. 2447-2457, 2017.

APÊNDICE - Tabela das espécies e valor de completude BUSCO

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Accipiter gentilis	GCF_929443795.1	A/T	Aves	Accipitridae	98,3
Acinonyx jubatus	GCF_027475565.1		Mammalia	Felidae	98,4
Acomys russatus	GCF_903995435.1		Mammalia	Muridae	97,6
Agelaius phoeniceus	GCF_020745825.1	A/T	Aves	Icteridae	96
Ailuropoda melanoleuca	GCF_002007445.2		Mammalia	Ursidae	95,2
Alligator mississippiensis	GCF_000281125.3	T	Reptilia	Alligatoridae	97,2
Alligator sinensis	GCF_000455745.1	T	Reptilia	Alligatoridae	88,7
Anas platyrhynchos	GCF_015476345.1		Aves	Anatidae	96,5
Anolis carolinensis	GCF_000090745.1	T	Reptilia	Dactyloidae	93,3
Anser cygnoides	GCF_002166845.1		Aves	Anatidae	95,2
Antechinus flavipes	GCF_016432865.1	M/T	Mammalia	Dasyuridae	94,7
Aotus nancymaae	GCF_000952055.2	M/T	Mammalia	Aotidae	92,6
Apodemus sylvaticus	GCF_947179515.1	M/T	Mammalia	Muridae	98,9
Aptenodytes forsteri	GCF_000699145.1	A/T	Aves	Spheniscidae	89,5
Apteryx mantelli mantelli	GCF_001039765.1	A/T	Aves	Apterygidae	85
Apus apus	GCF_020740795.1	A/T	Aves	Apodidae	96,1

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Aquila chrysaetos chrysaetos	GCF_900496995.4	A/T	Aves	Accipitridae	97,6
Artibeus jamaicensis	GCF_021234435.1	M/T	Mammalia	Phyllostomidae	98,2
Arvicanthis niloticus	GCF_011762505.1		Mammalia	Muridae	98,7
Arvicola amphibius	GCF_903992535.2		Mammalia	Cricetidae	94,9
Athene cunicularia	GCF_003259725.1	A/T	Aves	Strigidae	88,3
Aythya fuligula	GCF_009819795.1	A/T	Aves	Anatidae	96,7
Balaenoptera acutorostrata scammoni	GCF_000493695.1	M/T	Mammalia	Balaenopteridae	89,3
Balaenoptera musculus	GCF_009873245.2	M/T	Mammalia	Balaenopteridae	95,9
Bison bison bison	GCF_000754665.1		Mammalia	Bovidae	89,7
Bos mutus	GCF_000298355.1		Mammalia	Bovidae	88,2
Bos taurus	GCF_002263795.2		Mammalia	Bovidae	98
Bubalus bubalis	GCF_019923935.1	M/T	Mammalia	Bovidae	98,6
Budorcas taxicolor	GCF_023091745.1	M/T	Mammalia	Bovidae	98,5
Bufo bufo	GCF_905171765.1	T	Amphibia	Bufonidae	93,2
Bufo gargarizans	GCF_014858855.1	T	Amphibia	Bufonidae	90,8
Calidris pugnax	GCF_001431845.1	A/T	Aves	Scolopacidae	96,2
Callithrix jacchus	GCF_011100555.1	M/T	Mammalia	Cebidae	97,2
Callorhinus ursinus	GCF_003265705.1		Mammalia	Otariidae	98,4

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Calypte anna	GCF_003957555.1	A/T	Aves	Trochilidae	92,2
Camarhynchus parvulus	GCF_901933205.1	A/T	Aves	Thraupidae	93,4
Camelus bactrianus	GCF_000767855.1		Mammalia	Camelidae	91
Camelus dromedarius	GCF_000803125.2	M/T	Mammalia	Camelidae	93,7
Camelus ferus	GCF_009834535.1	M/T	Mammalia	Camelidae	97,2
Canis lupus familiaris	GCF_000002285.5	M/T	Mammalia	Canidae	97,8
Capra hircus	GCF_001704415.2		Mammalia	Bovidae	98,3
Caretta caretta	GCF_023653815.1	T	Reptilia	Cheloniidae	97
Carlito syrichta	GCF_000164805.1	M/T	Mammalia	Tarsiidae	87,3
Catharus ustulatus	GCF_009819885.2	A/T	Aves	Turdidae	96,3
Cavia porcellus	GCF_000151735.1	M/T	Mammalia	Caviidae	92
Cebus imitator	GCF_001604975.1		Mammalia	Cebidae	95,6
Centrocercus urophasianus	GCF_019232065.1		Aves	Phasianidae	95,1
Ceratotherium simum simum	GCF_000283155.1	M/T	Mammalia	Rhinocerotidae	97,9
Cercocebus atys	GCF_000955945.1		Mammalia	Cercopithecidae	94,8
Cervus canadensis	GCF_019320065.1	M/T	Mammalia	Cervidae	98
Cervus elaphus	GCF_910594005.1	M/T	Mammalia	Cervidae	99,2
Charadrius vociferus	GCF_000708025.1	A/T	Aves	Charadriidae	85,4

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Chelonia mydas	GCF_015237465.2	T	Reptilia	Cheloniidae	98,4
Chelonoidis abingdonii	GCF_003597395.1		Reptilia	Testudinidae	94,1
Chinchilla lanigera	GCF_000276665.1	M/T	Mammalia	Chinchillidae	98,4
Chiroxiphia lanceolata	GCF_009829145.1		Aves	Pipridae	92,8
Chlorocebus sabaeus	GCF_015252025.1		Mammalia	Cercopithecidae	97
Choloepus didactylus	GCF_015220235.1	M/T	Mammalia	Megalonychidae	96,4
Chrysemys picta bellii	GCF_000241765.5	T	Reptilia	Emydidae	96,2
Chrysochloris asiatica	GCF_000296735.1	M/T	Mammalia	Chrysochloridae	95,5
Colobus angolensis palliatus	GCF_000951035.1		Mammalia	Cercopithecidae	93,7
Condylura cristata	GCF_000260355.1	M/T	Mammalia	Talpidae	90,4
Corapipo altera	GCF_003945725.1		Aves	Pipridae	95,8
Corvus cornix cornix	GCF_000738735.5		Aves	Corvidae	90,1
Corvus hawaiiensis	GCF_020740725.1	A/T	Aves	Corvidae	97,3
Corvus kubaryi	GCF_017639235.1		Aves	Corvidae	94,2
Corvus moneduloides	GCF_009650955.1	A/T	Aves	Corvidae	95,7
Coturnix japonica	GCF_001577835.2		Aves	Phasianidae	96
Cricetulus griseus	GCF_003668045.3		Mammalia	Cricetidae	97
Crocodylus porosus	GCF_001723895.1	T	Reptilia	Crocodylidae	93

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Crotalus tigris	GCF_016545835.1	T	Reptilia	Viperidae	96,9
Cuculus canorus	GCF_017976375.1	A/T	Aves	Cuculidae	96,5
Cyanistes caeruleus	GCF_002901205.1		Aves	Paridae	88,8
Cygnus atratus	GCF_013377495.2		Aves	Anatidae	96,2
Cygnus olor	GCF_009769625.2		Aves	Anatidae	93,4
Dasypus novemcinctus	GCF_000208655.3	M/T	Mammalia	Dasypodidae	86,2
Delphinapterus leucas	GCF_002288925.2	M/T	Mammalia	Monodontidae	97,1
Dermochelys coriacea	GCF_009764565.3	T	Reptilia	Dermochelyidae	93,7
Desmodus rotundus	GCF_022682495.1		Mammalia	Phyllostomidae	97,9
Dipodomys ordii	GCF_000151885.1		Mammalia	Heteromyidae	92,6
Dipodomys spectabilis	GCF_019054845.1	M/T	Mammalia	Heteromyidae	98,4
Dromiciops gliroides	GCF_019393635.1	M/T	Mammalia	Microbiotheriidae	92,3
Dryobates pubescens	GCF_014839835.1	A/T	Aves	Picidae	95,9
Echinops telfairi	GCF_000313985.2	M/T	Mammalia	Tenrecidae	92,6
Elephantulus edwardii	GCF_000299155.1	M/T	Mammalia	Macroscelididae	93,5
Elephas maximus indicus	GCF_024166365.1	M/T	Mammalia	Elephantidae	98,7
Empidonax traillii	GCF_003031625.1	A/T	Aves	Tyrannidae	94,9
Enhydra lutris kenyoni	GCF_002288905.1		Mammalia	Mustelidae	98,2

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Eptesicus fuscus	GCF_027574615.1	M/T	Mammalia	Vespertilionidae	98
Equus asinus	GCF_016077325.2		Mammalia	Equidae	96,7
Equus caballus	GCF_002863925.1	M/T	Mammalia	Equidae	98,2
Equus przewalskii	GCF_000696695.1		Mammalia	Equidae	87,8
Equus quagga	GCF_021613505.1	M/T	Mammalia	Equidae	98,1
Erinaceus europaeus	GCF_000296755.1	M/T	Mammalia	Erinaceidae	95,1
Eublepharis macularius	GCF_028583425.1	T	Reptilia	Eublepharidae	99
Eumetopias jubatus	GCF_004028035.1	M/T	Mammalia	Otariidae	98,1
Falco cherrug	GCF_000337975.1		Aves	Falconidae	86,8
Falco naumanni	GCF_017639655.2	A/T	Aves	Falconidae	97,3
Falco peregrinus	GCF_000337955.1		Aves	Falconidae	86,4
Falco rusticolus	GCF_015220075.1	A/T	Aves	Falconidae	95,6
Felis catus	GCF_018350175.1		Mammalia	Felidae	98,1
Ficedula albicollis	GCF_000247815.1	A/T	Aves	Muscicapidae	93,6
Fukomys damarensis	GCF_012274545.1	M/T	Mammalia	Bathyergidae	93,5
Galeopterus variegatus	GCF_000696425.1	M/T	Mammalia	Cynocephalidae	86,8
Gallus gallus	GCF_016699485.2	A/T	Aves	Phasianidae	97,7
Gavialis gangeticus	GCF_001723915.1	T	Reptilia	Gavialidae	91,6

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Gekko japonicus	GCF_001447785.1	T	Reptilia	Gekkonidae	89,9
Geotrypetes seraphini	GCF_902459505.1	T	Amphibia	Dermophiidae	95
Globicephala melas	GCF_006547405.1		Mammalia	Delphinidae	96,8
Gopherus evgoodei	GCF_007399415.2	T	Reptilia	Testudinidae	94,5
Gopherus flavomarginatus	GCF_025201925.1	T	Reptilia	Testudinidae	97,8
Gorilla gorilla gorilla	GCF_029281585.1		Mammalia	Hominidae	98,7
Gracilinanus agilis	GCF_016433145.1	M/T	Mammalia	Didelphidae	92,6
Grammomys surdaster	GCF_004785775.1		Mammalia	Muridae	97,8
Grus americana	GCF_028858705.1	A/T	Aves	Gruidae	97,9
Gymnogyps californianus	GCF_018139145.2	A/T	Aves	Cathartidae	90,9
Haliaeetus leucocephalus	GCF_000737465.1		Aves	Accipitridae	95,3
Halichoerus grypus	GCF_012393455.1		Mammalia	Phocidae	89,6
Hemicordylus capensis	GCF_027244095.1	T	Reptilia	Cordylidae	98,8
Heterocephalus glaber	GCF_000247695.1	M/T	Mammalia	Bathyergidae	96,5
Hipposideros armiger	GCF_001890085.2	M/T	Mammalia	Hipposideridae	89,2
Hirundo rustica	GCF_015227805.1	A/T	Aves	Hirundinidae	95,9
Homo sapiens	GCF_000001405.40	M/T	Mammalia	Hominidae	99,4
Hyaena hyaena	GCF_003009895.1	M/T	Mammalia	Hyaenidae	97,4

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Hylobates moloch	GCF_009828535.2	M/T	Mammalia	Hylobatidae	96,9
Ictidomys tridecemlineatus	GCF_016881025.1		Mammalia	Sciuridae	95
Indicator indicator	GCF_027791375.1	A/T	Aves	Indicatoridae	93,8
Jaculus jaculus	GCF_020740685.1	M/T	Mammalia	Dipodidae	98,4
Lacerta agilis	GCF_009819535.1		Reptilia	Lacertidae	94,7
Lagenorhynchus obliquidens	GCF_003676395.1	M/T	Mammalia	Delphinidae	97,3
Lagopus leucura	GCF_019238085.1		Aves	Phasianidae	95,9
Lagopus muta	GCF_023343835.1	A/T	Aves	Phasianidae	97
Lemur catta	GCF_020740605.2	M/T	Mammalia	Lemuridae	98
Leopardus geoffroyi	GCF_018350155.1	M/T	Mammalia	Felidae	98,5
Lepidothrix coronata	GCF_001604755.1		Aves	Pipridae	94,9
Lipotes vexillifer	GCF_000442215.2	M/T	Mammalia	Lipotidae	90,1
Lonchura striata domestica	GCF_005870125.1		Aves	Estrildidae	95,3
Lontra canadensis	GCF_010015895.1		Mammalia	Mustelidae	98,2
Loxodonta africana	GCF_000001905.1	M/T	Mammalia	Elephantidae	89,8
Lutra lutra	GCF_902655055.1		Mammalia	Mustelidae	96,2
Lynx canadensis	GCF_007474595.2		Mammalia	Felidae	92,8
Lynx rufus	GCF_022079265.1		Mammalia	Felidae	98,2

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Macaca fascicularis	GCF_012559485.2	M/T	Mammalia	Cercopithecidae	98,1
Macaca mulatta	GCF_003339765.1	M/T	Mammalia	Cercopithecidae	98,5
Macaca nemestrina	GCF_000956065.1		Mammalia	Cercopithecidae	95
Macaca thibetana thibetana	GCF_024542745.1		Mammalia	Cercopithecidae	98,1
Malaclemys terrapin pileata	GCF_027887155.1	T	Reptilia	Emydidae	98,9
Manacus candei	GCF_025592945.1	A/T	Aves	Pipridae	97,6
Manacus vitellinus	GCF_001715985.3		Aves	Pipridae	86,6
Mandrillus leucophaeus	GCF_000951045.1		Mammalia	Cercopithecidae	92,7
Manis javanica	GCF_014570535.1	M/T	Mammalia	Manidae	94,9
Manis pentadactyla	GCF_014570555.1	M/T	Mammalia	Manidae	96,4
Marmota flaviventris	GCF_003676075.3	M/T	Mammalia	Sciuridae	98,1
Marmota marmota marmota	GCF_001458135.2		Mammalia	Sciuridae	93,2
Marmota monax	GCF_021218885.1	M/T	Mammalia	Sciuridae	99,1
Mastomys coucha	GCF_008632895.1		Mammalia	Muridae	97,1
Mauremys mutica	GCF_020497125.1	T	Reptilia	Geoemydidae	98,9
Mauremys reevesii	GCF_016161935.1	T	Reptilia	Geoemydidae	98,8
Meles meles	GCF_922984935.1	M/T	Mammalia	Mustelidae	98,9
Melopsittacus undulatus	GCF_012275295.1	A/T	Aves	Psittaculidae	97,2

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Melozone crissalis	GCF_028551555.1	A/T	Aves	Passerellidae	96,5
Meriones unguiculatus	GCF_002204375.1		Mammalia	Muridae	93,3
Mesocricetus auratus	GCF_017639785.1	M/T	Mammalia	Cricetidae	98,9
Microcaecilia unicolor	GCF_901765095.1	T	Amphibia	Siphonopidae	90,2
Microcebus murinus	GCF_000165445.2	M/T	Mammalia	Cheirogaleidae	95
Microtus fortis	GCF_014885135.2		Mammalia	Cricetidae	96,2
Microtus ochrogaster	GCF_000317375.1		Mammalia	Cricetidae	97
Microtus oregoni	GCF_018167655.1		Mammalia	Cricetidae	99
Miniopterus natalensis	GCF_001595765.1		Mammalia	Vespertilionidae	94,1
Mirounga angustirostris	GCF_021288785.2		Mammalia	Phocidae	96,3
Mirounga leonina	GCF_011800145.1	M/T	Mammalia	Phocidae	97,7
Molossus molossus	GCF_014108415.1	M/T	Mammalia	Molossidae	97,9
Molothrus ater	GCF_012460135.2	A/T	Aves	Icteridae	92,8
Monodelphis domestica	GCF_000002295.2	M/T	Mammalia	Didelphidae	93,3
Monodon monoceros	GCF_005190385.1	M/T	Mammalia	Monodontidae	97
Moschus berezovskii	GCF_022376915.1	M/T	Mammalia	Moschidae	96,3
Motacilla alba alba	GCF_015832195.1	A/T	Aves	Motacillidae	95,1
Mus caroli	GCF_900094665.2		Mammalia	Muridae	95,3

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Mus musculus	GCF_000001635.27	M/T	Mammalia	Muridae	99,5
Mus pahari	GCF_900095145.1		Mammalia	Muridae	97,2
Mustela erminea	GCF_009829155.1		Mammalia	Mustelidae	97,9
Mustela putorius furo	GCF_011764305.1		Mammalia	Mustelidae	98,2
Myiozetetes cayanensis	GCF_022539395.1	A/T	Aves	Tyrannidae	93,9
Myodes glareolus	GCF_902806735.1		Mammalia	Cricetidae	95,1
Myotis brandtii	GCF_000412655.1		Mammalia	Vespertilionidae	90,3
Myotis davidii	GCF_000327345.1		Mammalia	Vespertilionidae	85,3
Myotis myotis	GCF_014108235.1		Mammalia	Vespertilionidae	97,2
Nannospalax galili	GCF_000622305.1	M/T	Mammalia	Spalacidae	96,7
Nanorana parkeri	GCF_000935625.1	T	Amphibia	Dicroglossidae	90,2
Neogale vison	GCF_020171115.1	M/T	Mammalia	Mustelidae	98,6
Neomonachus schauinslandi	GCF_002201575.2		Mammalia	Phocidae	97,4
Neopelma chrysocephalum	GCF_003984885.1	A/T	Aves	Pipridae	96,1
Neophocaena asiaeorientalis asiaeorientalis	GCF_003031525.2	M/T	Mammalia	Phocoenidae	94,7
Nipponia nippon	GCF_000708225.1	A/T	Aves	Threskiornithidae	87,1
Nomascus leucogenys	GCF_006542625.1	M/T	Mammalia	Hylobatidae	98,3
Notechis scutatus	GCF_900518725.1	T	Reptilia	Elapidae	90,9

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Numida meleagris	GCF_002078875.1	A/T	Aves	Numididae	97
Nyctereutes procyonoides	GCF_905146905.1		Mammalia	Canidae	97,2
Nycticebus coucang	GCF_027406575.1	M/T	Mammalia	Lorisidae	98,6
Ochotona curzoniae	GCF_017591425.1	M/T	Mammalia	Ochotonidae	98,6
Ochotona princeps	GCF_014633375.1	M/T	Mammalia	Ochotonidae	91,1
Octodon degus	GCF_000260255.1	M/T	Mammalia	Octodontidae	93,7
Odobenus rosmarus divergens	GCF_000321225.1	M/T	Mammalia	Odobenidae	96,9
Odocoileus virginianus texanus	GCF_002102435.1		Mammalia	Cervidae	93,4
Onychomys torridus	GCF_903995425.1		Mammalia	Cricetidae	96,7
Onychostruthus taczanowskii	GCF_017590055.1	A/T	Aves	Passeridae	95,5
Orcinus orca	GCF_937001465.1	M/T	Mammalia	Delphinidae	97,3
Ornithorhynchus anatinus	GCF_004115215.2	M/T	Mammalia	Ornithorhynchidae	94,6
Orycteropus afer afer	GCF_000298275.1	M/T	Mammalia	Orycteropodidae	94
Oryctolagus cuniculus	GCF_009806435.1	M/T	Mammalia	Leporidae	96,3
Oryx dammah	GCF_014754425.2		Mammalia	Bovidae	98,3
Otolemur garnettii	GCF_000181295.1	M/T	Mammalia	Galagidae	94,6
Ovis aries	GCF_016772045.1		Mammalia	Bovidae	98,4
Oxyura jamaicensis	GCF_011077185.1	A/T	Aves	Anatidae	97,1

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Pan paniscus	GCF_029289425.1		Mammalia	Hominidae	98,4
Pan troglodytes	GCF_028858775.1		Mammalia	Hominidae	98,7
Panthera leo	GCF_018350215.1		Mammalia	Felidae	96,4
Panthera pardus	GCF_024362965.1		Mammalia	Felidae	97,9
Panthera tigris	GCF_018350195.1		Mammalia	Felidae	97,8
Panthera uncia	GCF_023721935.1		Mammalia	Felidae	96,6
Pantherophis guttatus	GCF_001185365.1	T	Reptilia	Colubridae	94,1
Papio anubis	GCF_008728515.1		Mammalia	Cercopithecidae	96,1
Parus major	GCF_001522545.3	A/T	Aves	Paridae	95,1
Passer montanus	GCF_014805655.1		Aves	Passeridae	94,8
Pelodiscus sinensis	GCF_000230535.1	T	Reptilia	Trionychidae	87
Perognathus longimembris pacificus	GCF_023159225.1	M/T	Mammalia	Heteromyidae	96
Peromyscus californicus insignis	GCF_007827085.1	M/T	Mammalia	Cricetidae	98,1
Peromyscus leucopus	GCF_004664715.2		Mammalia	Cricetidae	91,4
Peromyscus maniculatus bairdii	GCF_003704035.1		Mammalia	Cricetidae	97,3
Phacochoerus africanus	GCF_016906955.1	M/T	Mammalia	Suidae	96,4
Phascolarctos cinereus	GCF_002099425.1	M/T	Mammalia	Phascolarctidae	90,8
Phasianus colchicus	GCF_004143745.1		Aves	Phasianidae	96,2

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Phoca vitulina	GCF_004348235.1	M/T	Mammalia	Phocidae	98,3
Phocoena sinus	GCF_008692025.1	M/T	Mammalia	Phocoenidae	95
Phodopus roborovskii	GCF_943737965.1		Mammalia	Cricetidae	98
Phyllostomus discolor	GCF_004126475.2		Mammalia	Phyllostomidae	94,1
Phyllostomus hastatus	GCF_019186645.2	M/T	Mammalia	Phyllostomidae	98,5
Physeter catodon	GCF_002837175.3	M/T	Mammalia	Physeteridae	93,7
Piliocolobus tephrosceles	GCF_002776525.5		Mammalia	Cercopithecidae	96,6
Pipistrellus kuhlii	GCF_014108245.1	M/T	Mammalia	Vespertilionidae	97,4
Pipra filicauda	GCF_003945595.2		Aves	Pipridae	95,9
Podarcis muralis	GCF_004329235.1		Reptilia	Lacertidae	93,9
Podarcis raffonei	GCF_027172205.1	T	Reptilia	Lacertidae	98,7
Pogona vitticeps	GCF_900067755.1	T	Reptilia	Agamidae	96
Pongo abelii	GCF_028885655.1		Mammalia	Hominidae	98,5
Pongo pygmaeus	GCF_028885625.1	M/T	Mammalia	Hominidae	98,9
Prionailurus bengalensis	GCF_016509475.1		Mammalia	Felidae	97,3
Prionailurus viverrinus	GCF_022837055.1	M/T	Mammalia	Felidae	98,8
Propithecus coquereli	GCF_000956105.1	M/T	Mammalia	Indriidae	89,6
Protobothrops mucrosquamatus	GCF_001527695.2	T	Reptilia	Viperidae	92,2

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Pseudonaja textilis	GCF_900518735.1	T	Reptilia	Elapidae	94,6
Pseudopodoces humilis	GCF_000331425.1	A/T	Aves	Paridae	97,3
Pteronotus parnellii mesoamericanus	GCF_021234165.1	M/T	Mammalia	Mormoopidae	98,5
Pteropus alecto	GCF_000325575.1		Mammalia	Pteropodidae	89,8
Pteropus giganteus	GCF_902729225.1	M/T	Mammalia	Pteropodidae	94,6
Pteropus vampyrus	GCF_000151845.1		Mammalia	Pteropodidae	92
Puma concolor	GCF_003327715.1		Mammalia	Felidae	88,5
Puma yagouaroundi	GCF_014898765.1		Mammalia	Felidae	97,5
Pyrgilauda ruficollis	GCF_017590135.1	A/T	Aves	Passeridae	95,5
Python bivittatus	GCF_000186305.1	T	Reptilia	Pythonidae	88,9
Rana temporaria	GCF_905171775.1	T	Amphibia	Ranidae	94,2
Rattus norvegicus	GCF_015227675.2		Mammalia	Muridae	98,2
Rattus rattus	GCF_011064425.1		Mammalia	Muridae	92,7
Rhinatrema bivittatum	GCF_901001135.1	T	Amphibia	Rhinatrematidae	93,5
Rhinolophus ferrumequinum	GCF_004115265.2	M/T	Mammalia	Rhinolophidae	96,6
Rhinopithecus roxellana	GCF_007565055.1		Mammalia	Cercopithecidae	98
Rissa tridactyla	GCF_028500815.1	A/T	Aves	Laridae	97,1
Rousettus aegyptiacus	GCF_014176215.1	M/T	Mammalia	Pteropodidae	98

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Saimiri boliviensis boliviensis	GCF_016699345.2	M/T	Mammalia	Cebidae	98,3
Sapajus apella	GCF_009761245.1		Mammalia	Cebidae	97,1
Sarcophilus harrisii	GCF_902635505.1	M/T	Mammalia	Dasyuridae	96,3
Sceloporus undulatus	GCF_019175285.1	T	Reptilia	Phrynosomatidae	92,6
Sciurus carolinensis	GCF_902686445.1		Mammalia	Sciuridae	93,2
Serinus canaria	GCF_022539315.1	A/T	Aves	Fringillidae	95,1
Sorex araneus	GCF_027595985.1	M/T	Mammalia	Soricidae	97,9
Spea bombifrons	GCF_027358695.1	T	Amphibia	Pelobatidae	95,9
Sphaerodactylus townsendi	GCF_021028975.2	T	Reptilia	Sphaerodactylidae	89,5
Strigops habroptila	GCF_004027225.2	A/T	Aves	Psittacidae	97,3
Struthio camelus australis	GCF_000698965.1	A/T	Aves	Struthionidae	87,4
Sturnira hondurensis	GCF_014824575.3		Mammalia	Phyllostomidae	96,1
Sturnus vulgaris	GCF_001447265.1	A/T	Aves	Sturnidae	95,6
Suncus etruscus	GCF_024139225.1	M/T	Mammalia	Soricidae	93,8
Suricata suricatta	GCF_006229205.1	M/T	Mammalia	Herpestidae	94,8
Sus scrofa	GCF_000003025.6	M/T	Mammalia	Suidae	90,5
Symphalangus syndactylus	GCF_028878055.1		Mammalia	Hylobatidae	94,4
Tachyglossus aculeatus	GCF_015852505.1	M/T	Mammalia	Tachyglossidae	94,9

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Taeniopygia guttata	GCF_003957565.2		Aves	Estrildidae	96,3
Talpa occidentalis	GCF_014898055.3	M/T	Mammalia	Talpidae	95,1
Terrapene carolina triunguis	GCF_002925995.2		Reptilia	Emydidae	95,5
Theropithecus gelada	GCF_003255815.1		Mammalia	Cercopithecidae	97,6
Trachemys scripta elegans	GCF_013100865.1		Reptilia	Emydidae	96,1
Trachypithecus francoisi	GCF_009764315.1		Mammalia	Cercopithecidae	95,2
Trichechus manatus latirostris	GCF_000243295.1	M/T	Mammalia	Trichechidae	93,9
Trichosurus vulpecula	GCF_011100635.1	M/T	Mammalia	Phalangeridae	97,1
Tupaia chinensis	GCF_000334495.1	M/T	Mammalia	Tupaiidae	90,7
Tursiops truncatus	GCF_011762595.1		Mammalia	Delphinidae	96,9
Tympanuchus pallidicinctus	GCF_026119805.1		Aves	Phasianidae	96,7
Tyto alba	GCF_018691265.1	A/T	Aves	Tytonidae	97,5
Ursus americanus	GCF_020975775.1	M/T	Mammalia	Ursidae	98,4
Ursus arctos	GCF_023065955.1	M/T	Mammalia	Ursidae	98,8
Ursus maritimus	GCF_017311325.1		Mammalia	Ursidae	96,1
Varanus komodoensis	GCF_004798865.1	T	Reptilia	Varanidae	96,5
Vicugna pacos	GCF_000164845.4		Mammalia	Camelidae	91,2
Vidua chalybeata	GCF_026979565.1	A/T	Aves	Estrildidae	96,6

Species_Name	Ref_Seq_Genome_ID	Analysis	Class	Family	C
Vidua macroura	GCF_024509145.1	A/T	Aves	Estrildidae	96,3
Vombatus ursinus	GCF_900497805.2	M/T	Mammalia	Vombatidae	96,8
Vulpes lagopus	GCF_018345385.1	M/T	Mammalia	Canidae	98,3
Vulpes vulpes	GCF_003160815.1		Mammalia	Canidae	94,4
Xenopus laevis	GCF_017654675.1	T	Amphibia	Pipidae	97,1
Xenopus tropicalis	GCF_000004195.4	T	Amphibia	Pipidae	96,2
Zalophus californianus	GCF_009762305.2	M/T	Mammalia	Otariidae	97
Zonotrichia albicollis	GCF_000385455.1	A/T	Aves	Passerellidae	89,7
Zootoca vivipara	GCF_011800845.1	T	Reptilia	Lacertidae	97,8