Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Departamento de Genética, Ecologia e Evolução

Programa de Pós-Graduação em Genética

Dissertação de Mestrado

A evolução da complexidade biológica em *Eukarya*: funções biológicas e domínios de proteínas associados aos número de tipos celulares

Autor: Dalbert Benjamim da Costa

Orientador: Prof. Dr. Francisco Pereira Lobo

Belo Horizonte

Setembro/2019

Dalbert Benjamim da Costa

A evolução da complexidade biológica em *Eukarya*: funções biológicas e domínios de proteínas associados aos número de tipos celulares

Dissertação apresentada ao Programa de Pós-Graduação em Genética do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito para obteção do título de Mestre em Genética.

Orientador: Prof. Dr. Francisco Pereira Lobo

Belo Horizonte

Departamento de Genética, Ecologia e Evolução

Instituto de Ciências Biológicas – UFMG

Setembro/2019

043 Costa, Dalbert Benjamim.
A evolução da complexidade biológica em *Eukarya*: funções biológicas e Domínios de proteínas associados ao número de tipos celulares [manuscrito] / Dalbert Benjamim Costa. – 2019.

121 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Francisco Pereira Lobo.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Departamento de Biologia Geral.

1. Genética. 2. Eucariotos. 3. Genômica. 4. Biologia computacional. 5. Ontologia genética. I. Lobo, Francisco Pereira. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título

CDU: 575

Ficha catalográfica elaborada por Cleide Vieira de Faria – CRB 6/2314 BC/UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

ATA DA DEFESA DE DISSERTAÇÃO

293/2019
entrada
2°/2017
CPF:
067.812.796-45

DALBERT BENJAMIM DA COSTA

Às nove horas do dia **26 de setembro de 2019**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"A evolução da complexidade biológica em Eukarya: funções biológicas e domínios de proteínas associados ao número de tipos celulares"**, requisito para obtenção do grau de Mestre em **Genética.** 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:

Prof./Pesq.	Instituição	CPF	Indicação
Francisco Pereira Lobo	UFMG	012273736-34	Aprovado
Romeu Cardoso Guimarães	UFMG	000679826-00	Aprovado
Gustavo Campos e Silva Kuhn	UFMG	260136648-62	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, 26 de setembro de 2019.

Francisco Pereira Lobo - Orientador Francisco P. Lobo

Romeu Cardoso Guimarães

Gustavo Campos e Silva Kuhn Gust. Kuhn

Pós-Graduação em Genética
Departamento de Genética, Ecologia e Evolução, ICB
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - C.P. 486 - Pampulha - 31270-901 - Belo Horizonte - MG
e-mail: pg-gen@icb.ufmg.br FAX: (+31) - 3409-2570

"A evolução da complexidade biológica em Eukarya: funções biológicas e domínios de proteínas associados ao número de tipos celulares"

DALBERT BENJAMIM DA COSTA

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Francisco Pereira Lobo - Orientador
UFMG

Romeu Cardoso Guimarães
UFMG

Gustavo Campos e Silva Kuhn
UFMG

Belo Horizonte, 26 de setembro de 2019.

Dedico esta dissertação aos meus super-heróis de verdade, às forças motrizes José Beijamim e dona Enilza, que permitiram que sonhos como este se realizassem.

À minha alma gêmea siamêsa, Danielle Agnes, por ser minha âncora no olho do furacão e por me fazer autosuperar todos os dias.

Ao meu orientador, Prof. Dr. Francisco Pereira Lobo, pelo exemplo extremo de controle positivo do que é ser um orientador e por todo aprendizado; por fazer milagres e pelo socorro em momentos difíceis.

AGRADECIMENTOS

Antes de tudo agradeço a aqueles que desde o início me deram apoio, portanto à minha família. Valeuzaço mamuskita dona Enilza pelos lanches da tarde e da noite, pelas palavras de carinho e de força, os conselhos enfim por ser a digníssima melhor primeira do mundo. Obrigado grande super-herói sô Benjamim, The "Big Ben" (deveria tatuar isto no peito e sair voando por entre os prédios heim haha) por ser meu modelo de conduta de ser humano, um dos caras mais corretos no mundo.

Agradeço à minha companheira durante esta longa jornada, Danielle, por todo esforço e carinho para que as coisas aconteçam, por toda experiência de vida que estamos agregando, essencial para afirmar que estamos crescendo como seres humanos e como família. Valeu demais Danny por segurar esta barra (que é gostar de vc rsrs).

Agradeço ao frango d'agua, parceria e muleta, o big brother Jeandersom. Agradeço à manola Luciana, sempre que foi preciso estava lá de prontidão para ajudar e com um puxãozinho de orelha. Valeu Jeanzinho, pequeno Enzo, Gilmara; Marcelo e Gabriel. Qualquer vitória também é de vocês.

Agradeço à galera do Vale do Aço, o Dom Evair e sua senhora Ivanete, Daynara e Thiago pelo apoio, conselhos, carinho e cuidado, tanto comigo quanto com Danielle. Valeu também tia Graça e a Vania pelos conselhos, aos amigos irmãos camaradas de tantas décadas, Digão e Leonardo, se não fossem vocês, a saga não teria iniciado (to decidindo se lhes pago com um rim se precisarem rs) ao Alexsandro, o grande Jay e Fabrício Antônio Conselheiro pela referência em como ser biólogo, pela amizade e parceria de décadas, ao casal XXI Giovanni e Aline por ampliar a amizade contada no dedo para as duas mãos (também to decidindo acerca do outro rim kk).

Um ode ao LAB (Laboratório de Algoritmos em Biologia) e aos Franc1scan0s, nesta época muito bem frequentado pelos parçxs prodígixs da genética e bioinformática Giovanni, Thieres, Leo, Anderson Raul, Igor, Thiago, Zandora, Alison, Agnello, Maycon, Amanda e Aline. Vlwzaço pelas interações, parcerias, amizade e o bom papo, um estado da arte pura e simplesmente. Obrigado pelos bons conselhos acadêmicos, profissionais, pessoais e todos aqueles helps na hora do aperto. Obrigado a todos professores (Marcelo, Mônica e Diana) e amigos, em especial à Daniela, Iza, Ju, Malu e Rahysa que todos os dias compartilham este espaço do laboratório conosco pela boa convivência, amizade e conselhos.

Um muito obrigado à família da Pós-Graduação em Genética, que ficaria gigante o nome de todos aqui, mas guardo um bom sentimento de todos. Agradeço a todos os

professores e coordenadores do Programa de Pós-Graduação em Genética pelo incentivo e formação, agradeço à CAPES pela bolsa de estudos e por fazer a mágica acontecer. Agradeço à galera da secretaria (Raíssa, Danny e Vitória) pela competência em todas as ocasiões necessárias. Agradeço à banca examinadora pela dedicação, contribuição e atenção prestada.

Agradeço especialmente e essencialmente ao professor e orientador Dr. Fancisco Lobo por permitir que tudo aconteça, pela oportunidade, formação, amizade, preparação e orientação. Valeu mestre Chico por todas as vezes que foi solícito em saciar minhas dúvidas, atendendo mesmo fora de horários ou quando estava extremamente ocupado (ou seja sempre kk). Danielle e eu somos enormemente gratos e não temos palavras para descrever o seu suporte e compreensão durante as adversidades de dezembro de 2018 até os dias de hoje. Agradeço por toda a logística que me proporcionou, tanto aquele "tomo cafezinho bão logo existo" aos equipamentos do LAB utilizados, ao seu tempo dispensado em cada linha de código e algoritmos e pela dedicação e empenho por este trabalho. Considero genial a ideia desta dissertação, de modo que ficou evidente a nossa necessidade de propor tecnologias e conhecimento como forma de mudar o mundo, não apenas em prol da felicidade da nossa espécie e entes queridos, mas da biodiversidade em geral e porque não pelo planeta. Enfim valeu por atiçar todos os dias nossa curiosidade no pioneirismo da descoberta. Um muito valeu demais por todas as oportunidades que o LAB me ofereceu não apenas para a confecção desta dissertação, como também pela amizade e pelo crescimento como pesquisador e pessoal.

Vlw demais galerê!!!

(...)."As montagens multicelulares converteram-se em indivíduos animais, vegetais e fúngicos. Portanto, a vida não é toda feita de divergência e discórdia, mas é também a junção de entidades díspares em novos seres. E ela não se deteve nas células complexas e nos seres multicelulares. Seguiu adiante, forjando sociedade, comunidades e a própria biosfera viva."

Lynn Margulis e Dorion Sagan – O que é vida ? p 255

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

GLOSSÁRIO

Ape – pacote do R para análises de filogenética e evolução; utilizado para o cálculo dos contrastes filogeneticamente independentes.

BUSCO - *Benchmarking Universal Single-Copy Orthologs*; utilizado para avaliação da qualidade de proteomas não-redundantes.

CDS - (*Coding Sequence* – região codificadora do gene).

Contigs – trecho de *reads* contínuas geradas a partir do alinhamento *De Novo* de *reads* menores sobrepostas.

CTK – proteína tirosina quinase não receptoras no citoplasma.

de novo - à partir do princípio; na biologia, utilizado para denotar eventos (e.g. síntese de macromoléculas) à partir de moléculas mais simples.

DNA – ácido desoxiribonucleico

Domain shuffling – embaralhamento de domínios proteicos.

duplicates – genes duplicados no BUSCO.

Estatística N50 – estatística de um conjunto de reads ou scaffolds. O N50 é semelhante a uma média ou mediana de comprimentos, mas tem maior peso dado aos contigs mais longos.

Exon - regiões codificantes de proteínas.

Exon shuffling – embaralhamento de exons.

Fingerprints - múltiplos motivos conservados.

fragmented – genes fragmentados no BUSCO.

FT – Fatores de transcrição.

GBFF – formato do NCBI genbank de genoma contendo anotações.

GCA – prefixo de identificadores de genomas que não são refseq.

GenBank – banco de dados genéticos associado ao *NCBI* e parte integrante do *International Nucleotide Sequence Database Collaboration*.

GO – Gene Ontology.

HD – homeodomínio, domínios de genes homeobox.

HMMs – Hidden Markov Models ou Modelos Ocultos de Markov.

in tandem – em série.

InterPro - plataforma com acesso a bancos de dados de famílias, domínios e lugares funcionais de proteínas onde as características identificáveis encontradas em proteínas conhecidas podem ser aplicadas a novas sequências de proteínas.

InterProScan - software que identifica estas assinaturas dos bancos de dados utilizados pelo InterPro.

Introns - regiões não codificantes de proteínas.

KOMODO2 – Software de genômica comparativa que foi usado para calcular as frequências de uma função biológica ou domínio proteíco associado à alguma medida de complexidade biológica.

MAPK - proteínas-quinases ativadas por mitógenos.

Misassemblies – não alinhamento que pode ocasionar falha de montagem de um genoma.

missings - genes ausentes no BUSCO.

MKP - fosfatases de proteínas-quinases ativadas por mitógenos.

motif - assinatura por padrão.

mRNA – RNA mensageiro.

NCBI - O Centro Nacional de Informações sobre Biotecnologia (NCBI) faz parte da Biblioteca Nacional de Medicina dos Estados Unidos (NLM), uma divisão do National Institutes of Health (NIH).

ncDNA – DNA não codificante.

nex – formato nexus, utilizado para representar árvores filogenéticas.

NEXUS – formato nexus, utilizado para representar árvores filogenéticas

NGS – sequenciamento de nova geração; sequenciamento massivamente paralelo de DNA.

Nwk – formato newick, utilizado para representar árvores filogenéticas.

Nw_labels – programa usado para retornar nome as espécies de arquivos de árvores filogenéticas, parte do pacote *newick tools*.

Operons - conjunto de genes nos procariontes e em alguns eucariontes que se encontram funcionalmente relacionados, contíguos e controlados coordenadamente, sendo todos expressos em apenas um RNA mensageiro.

ORF - (Open Reading Frame – Quadros de Leituras Abertas)

OrthoDB - um catálogo de ortólogos, isto é, genes herdados por espécies existentes do seu último ancestral comum.

ρ – Coeficiente de correlação de Spearman.

Pipeline – termo da Ciências da Computação que significa um conjunto de algoritmos, funções ou processos que são construídos para serem rodados rapidamente, e normalmente, de forma linear para filtrar ou transformar dados.

PCM - métodos comparativos filogenéticos (*Phylogenetic Comparative Methods)*.

PERL - acrônimo de *Practical Extraction and Report Language.*

PIC - Contrastes Filogenéticos Independentes (*Phylogenetic Independent Contrasts)*

Pfam – banco de dados de domínios de proteínas.

PTK - Proteína tirosina-quinase.

pTK – Proteína tirosina fosfatase.

r – Coeficiente correlação de Pearson.

R - ambiente computacional e linguagem de programação para manipulação, análise e visualização gráfica de dados.

Reads – sequências fragmentadas de bases nucleotídicas correspondentes ao DNA da amostra gerada pelo NGS.

RefSeq – banco de dados de sequências de referência do *NCBI*.

REVIGO - listas de termos do Gene Ontology e remove termos GO redundantes.

RNA - ácido ribonucleico.

RTK – proteína tirosina quinase receptora transmambrana

Scaffolds - série não contígua de sequências genómicas num suporte, consistindo em sequências separadas por intervalos de comprimento conhecido. As sequências que estão ligadas são tipicamente sequências contíguas correspondendo a sobreposições de leitura.

Shell – uma linguagem de script usada em vários sistemas operativos (operacionais), com diferentes dialetos, dependendo do interpretador de comandos utilizado. Um exemplo de interpretador de comandos é o bash, usado na grande maioria das distribuições GNU/Linux.

Singles - ortólogos de cópia única quase universal.

τ - Coeficiente de Correlação de Kendall.

TimeTree of Life - serviço *web* que, a partir de uma lista de espécies, retorna as relações filogenéticas entre as mesmas em um arquivo no formato newick.

TK – tirosina quinase.

WGD – *whole genome duplication*, duplicação gênica de um genoma inteiro.

RESUMO

Durante o curso da evolução biológica de eucariotos, organismos com diferentes graus de complexidade emergiram. Para fins práticos, o número de tipos celulares distintos tem sido comumente utilizado como um *proxy* para a complexidade biológica. Também durante o curso da evolução, novas proteínas emergiram em *Eukarya* como resultado de evolução de novo, duplicações gênicas seguidas por divergência e, em vários casos, embaralhamento de domínios (domain shuffling). Utilizamos uma abordagem estatística e de genômica comparativa para estudar a evolução da complexidade biológica em eucariotos, pesquisando por funções biológicas (representadas como a frequência de domínios de proteínas e de funções gênicas codificadas em uma ampla gama de genomas eucarióticos) associadas ao seu número de tipos celulares diferentes. Para tal, inicialmente selecionamos 41 proteomas não-redundantes eucarióticos de alta qualidade em termos de completude do repertório gênico, estimado pelo software BUSCO, e que possuam informação sobre o número de tipos celulares. Para os proteomas selecionados, realizamos a anotação dos mesmos usando o programa InterProscan, de modo a detectarmos quais são os domínios protéicos (identificados no banco de dados Pfam) e quais funções biológicas (identificados por termos Gene Ontology) codificados nestes genomas. Buscamos dois tipos de associação entre as frequências de domínios/termos GO em cada proteoma não-redundante e o número de diferentes tipos de células para as espécies correspondentes. Uma das associações consiste na correlação de Spearman, sendo o outro tipo de modelo corrigido de modo a levar em consideração a história filogenética das espécies analisadas, de modo a eliminar possíveis dependências dos dados em função da origem evolutiva comum dos organismos em análise. Para ambos computamos valores p, os quais são posteriormente corrigidos em função do cenário de múltiplas hipóteses (BH). Consideramos como positivos os modelos onde obtivemos valores p corrigidos menores que $p \leq 0.05$. Encontramos 256 domínios Pfam e 304 funções biológicas que desempenham papéis importantes nos processos de matriz extracelular, interação célula-célula, fatores de transcrição, hormônios, processos regulatórios e fatores-chave para diferenciação celular e processos de desenvolvimento corporal. Em conjunto, nossa abordagem destaca importantes processos biológicos associados ao aumento da complexidade em *Eukarya*, sugerindo sua importância para o estabelecimento da complexidade biológica existente.

Palavras chaves: *Eukarya*, genômica comparativa, complexidade biológica, número de diferentes tipos de células, biologia computacional, domínio de proteínas, Pfam, Gene Ontology.

ABSTRACT

During the course of biological evolution, organisms with different degrees of complexity have arisen. For practical purposes, the number of distinct cell types has been commonly used as a proxy for biological complexity. Also during the course of evolution, new proteins emerged in *Eukarya* as the result of de novo gene evolution, gene duplications followed by divergence and, in several cases, functional domain shuffling. We used a statistical comparative genomics approach to study the evolution of biological complexity in *Eukarya* by searching for biological functions (represented as the frequency of protein domains and gene functions coded in a wide range of eukaryotic genomes) associated with their number of cell types. We selected 41 high-quality non-redundant eucaryotic proteomes in terms of gene repertoire completeness as estimated by BUSCO and, for each proteome was annotated to identify protein domains (Pfam) and biological functions (Gene Ontology - GO - terms) using InterProScan. We compute two classes of association metrics for the frequencies of each Pfam/GO term and the number of cell types. One class consists on traditional Spearman correlation, while the other is corrected to take into account the common ancestry relationships across species data, therefore correcting for this bias. For each linear model we computed p-values, and we applied multiple hypothesis correction (BH methods) to take into account the multiple-comparison problem. We considered as positive models with corrected p-values smaller than 0.05 resulting in 256 Pfam domains and 304 GO terms significantly associated with biological complexity. Among these sets we found several domains that play important roles in extracellular matrix processes, cell-cell interaction, transcription factors, hormones, regulatory processes and key factors for cell differentiation and body development processes. Taken together, our approach highlights important biological processes associated with the increase of complexity in *Eukarya*, suggesting their importance for the establishment of extant biological complexity.

Keywords: *Eukarya*, comparative genomics, biological complexity, number of different cell types, computational biology, protein domain, Pfam, Gene Ontology.

1. Introdução

1.1 - Complexidade biológica em *Eukarya* – definições

Um consenso sobre a definição de complexidade biológica nos organismos vivos, de maneira geral, e nos eucariotos, em particular, ainda é uma discussão em aberto na Biologia, mesmo após décadas de debate (Bell e Mooers 1997; Szathmary *et al.* 2001; Carroll, 2001; McShea, 1996, 2005; Lynch, 2007; Pennell *et al.* 2014). Formalizar uma definição única e ideal de complexidade biológica nos eucariotos é particularmente difícil frente ao universo dos vários modos de se pensar a biologia, dos diferentes contextos e dos diferentes níveis de informação comumente utilizados na ciência (conceitos, objetos de estudo, teorias, hipóteses, modelos, definições, termos e ideias). As várias abordagens de complexidade biológica são conceitos distintos e subjetivos usados para descrever algum fenômeno ou processo físico, químico e biológico dentro de um contexto específico. O próprio termo "variação da complexidade biológica" é ambíguo e assume várias interpretações e múltiplos sentidos e contextos (Finlay e Esteban, 2009).

Dentro dessas limitações, várias abordagens, muitas vezes complementares, buscaram explicar a origem ou aumentar a compreensão dos mecanismos por trás da evolução da complexidade biológica: 1) os nichos ecológicos se tornam mais complexos e são preenchidos por organismos mais complexos à medida que aumenta a sua diversidade (Arthur, 1994); 2) todos as forças evolutivas, como mutações, seleção natural e a deriva influenciam a variação da complexidade biológica (Yaeger, 2009); 3) como a evolução do genoma e do repertório proteico confere novas funcionalidades para os organismos (Vogel e Chotia, 2006). Para estes trabalhos e vários outros, um conceito constantentemente utilizado na literatura para as diferentes definições de métricas de complexidade biológica é o "*proxy*", traduzido livremente aqui como "representante". Bons representantes seriam métricas que possam avaliar a complexidade biológica considerando dois aspectos: (i) universalidade (aplicável a qualquer sistema) e (ii) operacionalidade (que indique, sem ambiguidade, como medir a complexidade nos diferentes sistemas) (McShea, 1996).

Neste contexto, alguns trabalhos buscaram descrever e sumarizar as métricas que quantificam a complexidade biológica dos organismos (McShea, 1996). Neste trabalho, estas métricas foram agrupadas em duas categorias principais: (i) Complexidade Estrutural - objetos (número de partes físicas de um sistema; por exemplo número de

genes de um genoma); (ii) Complexidade Funcional - processos (número de interações entre estas partes físicas; por exemplo interações de redes de proteínas). Estas duas categorias podem ser hierarquicas - resultante do número de níveis de aninhamento tanto de objetos (e.g. > órgão > tecido > célula > organelas > proteína) quanto de processos (e.g. genes > vias bioquímicas > redes metabólicas/informacionais) ou não hierarquicas, descrevendo o número de diferentes tipos de elementos (objetos, por exemplo por exemplo o número de tipos celulares diferentes) ou interações (processos, por exemplo contagem de diferentes funções de um organismo) (McShea, 1996; Lang e Rensing, 2015; Suga e Ruiz-Trillo, 2015).

Embora as categorias e ideias destes e de vários outros trabalhos ao longo das décadas intuitivamente contemplem, de maneira variável, os critérios da universalidade e da operacionalidade, existe uma dificuldade que é exatamente validar experimentalmente as hipóteses de modo a demonstrar a utilidade destes critérios (McShea, 1996; Lang e Rensing, 2015). Em virtude da enorme gama de diversidade de formas e funções dos organismos, documentar uma tendência genuína de evolução da complexidade biológica é difícil e desafiante frente o nível exigido de detalhamento dos processos como, por exemplo, a resolução taxonômica (Lang e Rensing, 2015), para medir e descrever esta complexidade (Adami, 2000; Niklas, Cobb e Dunker, 2014). Mesmo nesse cenário, vários estimadores de complexidade, grande parte deles com origem biológica evidente (por exemplo o tamanho do genoma, o número de genes codificadores de proteínas, o tamanho do proteoma) foram utilizadas em estudos ao longo das décadas para tentar explicar a variação da complexidade biológica durante a evolução dos eucariotos.

Conforme descrito acima, há diversos problemas na definição de *proxy* para descrever a complexidade biológica de maneira universal e operacional, uma vez que todas as diferentes métricas ainda são, em alguma escala, arbitrárias, sendo baseadas em estimativas, aproximações, projeções ou, nos piores casos, suposições e simplificações excessivas, todas com deficiências específicas (Bell e Moers, 1997; Adami, 2000; Lang e Rensing, 2015; Suga e Ruiz-Trillo, 2015). Embora ainda não exista um conceito "plenamente ideal", capaz de discriminar todos os processos da forma mais detalhada possível, para fins práticos, estes *proxies* são exemplos de como uma variável quantitativa pode ser utilizada como um indicador ou estimador do fenótipo mensurado (Lang e Rensing, 2015).

1.2 - O número de tipos celulares como um *proxy* de complexidade biológica em *Eukarya*

Durante o processo evolutivo, diversas novidades evolutivas surgiram com a origem dos eucariotos: 1) a cromatina e a diversificação da maquinaria de regulação da expressão gênica e do processamento alternativo de RNA; 2) a compartimentalização celular, com a origem dos sistemas internos de membranas, dos cloroplastos e/ou mitocôndria (assim como interações simbióticas entre as células - permitindo assim o surgimento da primeira célula eucariótica); 3) o núcleo celular; 4) a reprodução sexuada a partir da meiose e os gametas haplóides e 5) a multicelularidade a partir de ancestrais unicelulares, tema de estudo dessa dissertação (Baum e Baum, 2014; Grosberg e Strathmann, 2007).

A estimativa do número de diferentes tipos de células de um organismo eucariótico é um *proxy* comumente usado para mensurar a complexidade biológica nos eucariotos, permitindo a comparação entre diferentes espécies de maneira quantitava e homogênea (Bonner, 1988; Valentine *et al.*, 1994; Bell e Mooers 1997; Hedges *et al.*, 2004; Haygood e Investigators, 2006; Lang *et al*., 2010; Schad *et al.*, 2011, Chen *et al.*, 2014).

A estimativa do número diferente de células é conceitualmente definida como "quantidade de células com mesmo genótipo que expressam fenótipos distintos e se dispõem espacialmente de modo a formar tecidos funcionais dentro de um organismo integrado" (Rokas, 2008). Esse conceito é usado para classificar formas celulares morfologicamente ou fenotipicamente distintas nos organismos (Arendt, 2008, Arendt *et al*., 2016). Um organismo multicelular contém diferentes tipos celulares amplamente diferentes e especializados para realizar diferentes funções. Na grande maioria dos eucariotos multicelulares, embora as células sejam funcionalmente distintas, são genotipicamente similares ou idênticas, sendo as diferenças entre os diferentes tipos celulares usualmente causada pela regulação diferencial da expressão dos genes em seus genomas. Esta habilidade das células em regular e coordenar a síntese de proteínas nos diferentes tipos celulares aparenta um fator chave para a evolução da multicelularidade nos eucariotos (Gregory, 2005).

Uma questão relevante da pesquisa biológica contemporânea é avaliar como ocorreu a evolução da complexidade biológica a partir do surgimento de novas características fenotípicas nos ancestrais unicelulares, que habilitaram os organismos a aumentarem seus planos corporais e variar o desenvolvimento de novos tipos de células

nos diferentes clados dos eucariotos. O surgimento de organismos multicelulares a partir dos ancestrais unicelulares ocorreu de forma independente pelo menos 25 vezes em diferentes linhagens eucariotas, incluindo animais, plantas, fungos, algas verdes e marrons, e vários outros eucariontes (Grosberg e Strathmann, 2007; Niklas, 2014). Entretanto, estes dados podem divergir na literatura. Esta divergência ocorre devido a distintas abordagens em relação aos critérios funcionais e biologicamente relevantes para o advento da multicelularidade, como adesão (conexão), comunicação e cooperação entre as células.

Por exemplo, para Niklas, 2014 a multicelularidade evoluiu notavelmente uma vez em animais, três vezes em fungos, seis vezes em algas e diversas vezes em bactérias. No caso dos eucariotos, a grande maioria se tornou multicelular através da divisão clonal de uma única célula (esporos ou zigotos), onde espécies facultativamente multicelulares formam agregados de diferentes células como, por exemplo, colônias (Sebé-Pedrós *et al.*, 2015).

O incremento independente do número de diferentes tipos de células nos eucariotos constituem eventos antigos na árvore da vida (Hedges *et al.*, 2004). As primeiras evidências oriundas de registros fósseis que indicam a transição da unicelularidade para a multicelularidade foram cianobactérias filamentosas, datando entre 3 a 3,5 bilhões de anos (Knoll 2011) e os primeiros sinais de diferenciação celular, pré-requisito para organismos maiores, datam períodos superiores a 2 bilhões de anos atrás (Grosberg e Strathmann, 2007). Registros fósseis dos primeiros eucariotos multicelulares indicam que estes podem ter surgido há pelo menos 1 bilhão de anos (Knoll *et al*. 2006). Uma expansão considerável de diversidade de espécies de eucariotos como metazoários, fungos, algas vermelhas, algas verdes e um grande clado composto de alveolados, rhizarios e outros protistas ocorreu entre 800 - 100 milhões de anos atrás, relacionado à grandes eventos, dentre outros exemplos, de aumentos dramáticos de oxigênio atmosférico e oceânico (Carroll 2001, King 2004, Knoll 2003, Maynard Smith e Szathmary 1995).

Para fins práticos, o número de tipos celulares pode ser um *proxy* para mensurar a complexidade biológica por duas razões principais:

(i) É uma métrica quantificada de forma razoavelmente objetiva a partir de uma extensa revisão bibliografica da literatura para várias linhagens eucarióticas (Niklas, Cobb e Dunker, 2014). Por consequência, abrange tanto linhagens de organismos unicelulares que alteram seus fenótipos, funcionalidades e morfologias celulares nos diferentes

estágios em seus ciclos de vida (por exemplo alguns protistas) aos organismos obrigatoriamente multcelulares durante a maior parte do seu ciclo de vida, onde que células com mesmo genótipo podem expressar fenótipos distintos com potencial para realizar um determinado número de funções especializadas.

(ii) Vem sendo comumente usado para analisar relações entre a complexidade do organismo e características genômicas e funcionais, o que permite comparar e avaliar diferentes estudos já realizados. Diferentes hipóteses foram associadas à quantificação do tipos celulares diferentes como critérios para explicar os principais mecanismos moleculares associados à variação da complexidade biológica nos eucariotos (Basu *et al*., 2008; Chen *et al*., 2014; Lang e Rensing, 2015; Nam *et al*, 2015; Niklas *et al*., 2018), tais como: por exemplo o tamanho do genoma, número de genes codificadores de proteínas do genoma, tamanho do proteoma, comprimento de proteína, eventos de splicing alternativo e desordem de proteínas (Niklas, Cobb e Dunker, 2014; Yruela *et al.*, 2017).

Neste último contexto, as primeiras análises de genômica comparativa da evolução da complexidade propuseram que a complexidade biológica, mensurada como o número de tipos celulares de um organismo, poderia estar positivamente associada a variáveis estruturais, como o tamanho do genoma, ou ao potencial codificador do mesmo, como o número de genes codificadores de proteínas. Estes conceitos são intuitivamente coerentes, uma vez que genomas maiores ou com mais genes codificadores de proteínas contém, em teoria, mais informação armazenada, sendo, portanto, estimadores de complexidade. Entretanto, a aparente falta de associação entre estas variáveis (tamanho do genoma e número de genes codificadores) e a complexidade deu origem a dois aparentes paradoxos na biologia: paradoxo do valor C (valor definido como tamanho do genoma haplóide, mensurado pela quantidade de DNA no núcleo de células haplóides) e o paradoxo do valor G (número de genes codificadores de proteínas).

Estudos posteriores demonstraram que diversas razões, tais como a quantidade de seqüências de DNA não codificadoras de proteínas (ncDNA) (Gregory, 2005) e o nível de ploidia (Bennett e Leitch, 2005), são parcialmente responsáveis pela variação dos tamanhos de genoma observados em todas as linhagens dos eucariotos, o que explica sugere cautela ao utilizar estas variáveis como *proxies* da complexidade de forma global. Algumas espécies de protozoários e anfíbios, como amebas (Gregory, 2005) e espécies de salamandras (Keinath *et al*., 2015), por exemplo, possuem um número relativamente pequeno de diferentes tipos de células quando comparados à organismos filogeneticamente próximos, mas têm genomas consideravelmente maiores se

comparados à estes (Gregory, 2005; Lang e Rensing, 2015), fato esse causado por acumulo de material repetitivo em seus genomas (Luke, 2005; Keinath *et al*., 2015); no caso de plantas, a variação na ploidia é muito mais ampla que em espécies de animais (Bennett e Leitch, 2005); em conjunto, esses fenômenos são considerados explicações satisfatórias para o paradoxo dos valores C e G, o que sugere que esses *proxies* não são adequados como um bom estimador de complexidade biológica.

1.3 - Diversificação de funções moleculares em genomas eucarióticos

O recente advento da era genômica, com o desenvolvimento e a maturação de técnicas de sequenciamento de próxima geração (NGS), também conhecido como sequenciamento massivamente paralelo de DNA, bem como o rápido crescimento de bancos de dados de anotações das regiões funcionais do genoma e ontologias, tem permitido a comparação detalhada dos padrões de herança, semelhança e divergência entre os genomas de diversos táxons. Estes processos também permitem o detalhamento e compreensão das funções biológicas dos organismos e como elas estão organizadas nos genomas. Na biologia molecular, o termo "função biológica" é comumente associado às proteínas, uma vez que os genes codificadores de proteínas são transcritos em mRNA, que são então traduzidos em proteínas, as quais adotam estruturas tridimensionais particulares, muitas vezes passando por modificações pós traducionais, o que permite a execução de suas funções, tanto a nível molecular (genes e interações entre proteínas) ou em alto nível por exemplo em diferentes componentes do organismo (células, tecidos e órgãos) (Thomas, 2017).

Muitos dos genes essenciais relacionados aos processos moleculares fundamentais, como a replicação e reparo do DNA e a síntese de proteínas, são compartilhados entre as diversas espécies de vida celular observadas na Terra. Entretanto, é extremamente improvável que existam duas espécies com o mesmo conjunto exato de genes (Mushegian e Koonin, 1996). Durante o curso da evolução biológica, o surgimento de novos genes com novas funções potenciais foi um processo biológico que conferiu notável diversidade funcional, estando envolvidos no surgimento de complexos de proteínas, interações regulatórias e processos metabólicos responsáveis pelas propriedades fisiológicas e bioquímicas de um organismo.

Toda esta diversidade molecular é particularmente observada no desenvolvimento dos eucariotos, que, conforme discutido anteriormente, abrange

diferentes linhagens com graus variáveis de complexidade biológica, desde organismos unicelulares estritos, como protistas e algumas leveduras, organismos multicelulares facultativos, como alguns fungos e algas, aos organismos multicelulares obrigatórios, como plantas, animais e alguns fungos (Lespinet *et al*., 2002). Consequentemente, é razoável supor que algumas arquiteturas específicas de proteínas podem estar associadas à evolução da complexidade em *Eukarya*.

Há vários mecanismos responsáveis pela produção de novas estruturas gênicas, tais como retrotransposição, transferência lateral de genes; fissão ou fusão de genes, origem *de novo*, duplicação gênica de um genoma inteiro (WGD), ou poliploidização; duplicações em série ("*in tandem*"); duplicação seguido de divergência e *exon shuffling* (ou embaralhamento de domínios - *domain shuffling*) (Long *et al*., 2003). Eventos de duplicação gênica são bases fundamentais para a evolução de organismos mais complexos (Ohno, 1970). A duplicação gênica é um dos principais mecanismos moleculares responsáveis tanto pela variação da complexidade gênica e fenotípica, assim como responsáveis pela diversificação de funções potencialmente adaptativas, expansões de famílias e superfamílias gênicas para uma grande diversidade de espécies dos eucariotos (Meyer e Schartl, 1999). Por exemplo grandes partes dos proteomas de eucariotos mais complexos como os metazoários evoluíram a partir de eventos de duplicações repetidas de domínios de proteínas (Moore *et al*., 2008; Kawashima *et al*., 2009). Assim como cerca de 90% dos genes nos eucariotos são oriundos de eventos de duplicação (Rosanova *et al*., 2017).

Durante a evolução dos eucariotos, novas famílias de proteínas podem surgir a partir de eventos de duplicação que pode resultar em (i) eventos de especiação em que os ortólogos tem a probabilidade de manter a função ancestral em diferentes espécies ou (ii) eventos de duplicação dentro de um genoma com alguma probabilidade de divergência funcional entre os parálogos (Das, Dawson e Orengo, 2015). Neste segundo processo, após um evento de duplicação, as cópias podem eventualmente divergir acumulando mutações não sinônimas. A deriva e a seleção também podem atuar para criar uma função genética especializada ou nova (Moore *et al*., 2008). Deste modo um dos genes parálogos podem sofrer um processo de neofuncionalização, acumulando mutações livre de pressão seletiva, sem alterar a atividade da célula. Uma vez que a mutação ocorre somente em uma das cópias do gene, um dos genes duplicados ainda mantém a funcionalidade ancestral. Esta mutação pode originar funções *de novo* que não estavam presentes no gene ancestral (Das, Dawson e Orengo, 2015).

Uma típica proteína eucariótica consiste em uma estrutura modular composta de vários domínios (Han *et al*., 2007). Os domínios proteicos são estruturas terciárias tridimensionais e conservadas de sequências proteicas que podem enovelar independentemente do resto da proteína e possuem uma função biológica, tal como interações com outras biomoléculas (e.g. outros domínios, DNA/ RNA) ou uma atividade enzimática (Bashton e Chotia, 2007). Os domínios podem ser considerados as menores unidades estruturais e funcionais para a evolução de genes codificadores de proteínas em *Eukarya* (Chotia *et al*., 2003, Vogel *et al*., 2004).

Outros mecanismos fundamentais que conferem novas arquiteturas de domínios além de eventos de duplicação são a fusão e fissão de domínios (ou genes) (Vogel *et al*., 2004). A fusão é um tipo particular de recombinação em que as regiões codificadoras de dois domínios de proteínas diferentes, ou mesmo de duas proteínas inteiras, são fundidas. A fissão por outro lado, ocorre quando um gene pode se dividir em duas ou mais regiões codificadoras menores. Combinações destes mecanismos podem gerar novas proteínas multidomínios, com novos domínios sendo eventualmente adicionados a uma proteína, modificando sua funcionalidade, ou com a emergência de novas proteínas pela fissão de proteínas de múltiplos domínios (Kummerfeld *et al*., 2005; Fong *et al*., 2007).

Além da fusão e fissão de genes e eventos de duplicação seguido de divergência, o embaralhamento de domínios é outro dos principais mecanismos que conferiu novidade funcional nos eucariotos. O embaralhamento de domínios foi proposto por Gilbert em 1978 para explicar como a recombinação de genes parálogos em que dois ou mais domínios (ou exons) oriundos tanto de eventos de duplicação ou genes diferentes se combinam para formar um novo rearranjo exon-intron. Em eucariotos, os limites dos éxons são aproximadamente equivalentes aos domínios codificados por estes. Adicionalmente, ao longo da evolução, éxons são duplicados, embaralhados dentro de um genoma e combinados de maneiras diferentes (Vogel *et al*., 2004). Desta forma através destes processos de recombinação novos éxons podem ser inseridos dentro de um gene sem gerar interrupções do quadro de leitura da proteína (Kaneko, 2004). Deste modo, o embaralhamento de domínios fornece um mecanismo para a troca de sequências de exons entre genes diferentes (de Souza, 2012) e ocasiona genes híbridos com múltiplos domínios (Patthy 1995, 1999).

Outros conceitos biológicos importantes associados aos domínios da proteína são as famílias de proteínas e as superfamílias. Domínios relacionados por divergência à partir de uma sequência ancestral comum (ou seja, domínios homólogos) são agrupados

em superfamílias com base em evidências de sequência, estrutura e função (Vogel *et al*., 2004). Famílias de proteínas são formadas por conjunto de proteínas que contêm domínios da mesma superfamília (Vogel e Chotia, 2006). Frente estes conceitos, durante a evolução, a combinação e expansão de um conjunto finito de domínios foram responsáveis pelo vasto repertório proteico observado na árvore da vida (Kanapin, Mulder e Kuznetsov, 2010).

A combinação e expansão de domínios proteicos em novas arquiteturas gênicas com novas funcionalidades foi um importante mecanismo molecular responsável pela grande complexidade fenotípica em eucariotos (Babushok, Ostetarg e Kazazian, 2007). Proteínas de organismos distintos com as mesmas arquiteturas de domínios provavelmente compartilham ancestralidade em comum em diferentes organismos e, possívelmente, as mesmas propriedades funcionais (Vogel *et al*., 2004). Entretanto, arquiteturas específicas de proteínas multi-domínios são muitas vezes linhagem específicas nos organismos eucariotos, as quais podem possuir funções específicas em cada linhagem (Lespinet *et al*., 2002). Cerca de dois terços do repertório proteico nos eucariotos são proteínas modulares multidomínios (Han *et al*., 2007).

A identificação da sequência de uma proteína codificada por um gene permite inferir sobre suas funções biológicas. Estas funções geralmente podem ser descritas mediantes três aspectos: (i) a atividade bioquímica desta proteína que ocasiona uma função molecular em vias, células, tecidos e órgãos (ii) o local específico na célula, tecido ou órgão onde essa proteína exerce sua função e (iii) o papel desta proteína envolvida nos processos biológicos (Thomas, 2017). Assim, para buscar funções biológicas, representadas por domínios protéicos e outras anotações de proteínas, associados ao aumento da complexidade em *Eukarya*, faz-se necessário primeiro realizar a anotação funcional das proteínas encontradas nos genomas. Posteriormente à anotação, é necessário também utilizar metodologias de correlação para buscar pela associação entre a frequência de funções biológicas e o aumento da complexidade biológica.

1.4 - Anotação funcional de proteínas

A anotação de um genoma é um conjunto heterogêneo de processos que identifica e caracteriza as regiões funcionalmente importantes dos genomas, tais como genes, centrômeros, telômeros e operons, dentre outras. No caso de genes codificadores de proteínas, a sua anotação pode ser (i) estrutural, quando exatamente identifica as diferentes regiões dos genes, tais como as regiões presentes em mRNAs maduros (exons)

e removidas durante seu processamento (introns), regiões promotoras e terminadoras, dentre outras (Richardson e Watson, 2012); (ii) funcional, quando anexa uma camada adicional de informação biológica (metadados) à estrutura dos genes, por exemplo as funções biológicas oriundas da identificação e classificação de famílias de proteínas e domínios, funções moleculares ou uma ontologia (Yandell e Ence, 2012).

Geralmente uma proteína nova é caracterizada inicialmente com base nas propriedades funcionais e estruturais de famílias de proteínas ou domínios previamente conhecidos os quais, quando localizados em novas proteínas, é um indicativo de que a função realizada pelo domínio/família já caracterizada também deve ser realizada pela nova proteína (Xu e Drunback, 2012). Essa caracterização de novas proteínas usualmente ocorre através do uso de modelos preditivos conhecidos como assinaturas de proteínas, que basicamente são modelos matemáticos construídos a partir de múltiplos alinhamentos de regiões conservadas de famílias de proteínas ou domínios proteicos já conhecidos (Loewenstein *et al*., 2009). Estes modelos matemáticos são utilizados para buscar, na sequência de novas proteínas, regiões suficientemente parecidas com as assinaturas. Diferentes modelos de assinaturas de proteínas utilizam distintas abordagens para identificação e descrição de famílias de proteínas, domínios ou alguma outra característica de uma sequência em particular.

Cada modelo é construído inicialmente com um alinhamento de múltiplas seqüências de proteínas, e dependendo da estratégia adotada pode focar em uma única região de seqüência conservada - *motif* (assinatura por padrão), focar em múltiplos motifs conservados (*fingerprints*), ou considerar todo o alinhamento com todas as proteínas como também apenas um domínio em particular (utilizando *HMMs – Hidden Markov Models* ou Modelos Ocultos de Markov) (Gough *et al*., 2001). Após a etapa do alinhamento das multiplas sequências, o modelo é refinado por meio de pesquisas de forma iterativa com um banco de dados de sequencias de proteínas a procura de sequencias mais distantes que possam ser identificadas até resultar um modelo final e robusto.

1.4.1. InterProScan

O programa InterProScan (Zdobnov e Apweiler, 2001) é uma ferramenta que combina e sumariza várias assinaturas de proteínas de distintos bancos de dados em um único recurso pesquisável, sendo hoje o estado da arte para a anotação funcional *de novo*

de sequências protéicas (Quevillon *et al*., 2005; Mitchell *et al*., 2019). Estes bancos de dados podem ser de origem diversas e possuir diferentes focos (por exemplo Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMÍLY, Gene3D e PANTHER) (Quevillon *et al*., 2005; Jones *et al*., 2014). Assim, InterProScan garante a universalidade ao acesso aos diferentes bancos de dados de anotação de sequências sem a necessidade de cruzar os dados com cada database de forma individual (Mitchell *et al*., 2019). O software utiliza como base de dados para a anotação as diferentes assinaturas descritas anteriormente e, como saída, sumariza os resultados da pesquisa uma tabela contendo anotações oriundas do Gene Ontology e de outros bancos de dados como os citados acima.

1.4.2. Pfam (Protein families) database

Conforme exposto anteriormente, os domínios proteicos podem ser considerados unidades evolutivas conservadas, associadas tanto com a estrutura tridimensional como com o repertório funcional do genoma de um organismo sendo, portanto, possível inferir sobre as funções biológicas de uma proteína a partir da identificação de seus domínios. De fato, a combinação de vários domínios em várias proteínas em eucariotos é em grande parte responsável pela formação de novas proteínas e, consequentemente, pelo aumento da diversificação funcional destes organismos. O conhecimento das regiões evolutivas altamente conservadas dos genes, através da identificação de domínios proteicos, portanto permitem classificar e identificar proteínas para várias espécies com base nas relações de homologia (Xu e Drunback, 2012).

Estas informações são depositadas posteriormente nos repositórios públicos e bancos de dados de anotação funcional de famílias de proteínas e domínios. O Pfam (Protein families database) (Sonnhammer, Eddy e Durbin, 1997) é um dos bancos de dados de informação sobre famílias de proteínas e domínios proteicos. Metodologicamente, o Pfam utiliza a literatura e assinaturas de proteína para inferir sobre a anotação funcional de sequências não caracterizadas experimentalmente (Finn *et al*., 2016).

Perfis HMM são obtidos previamente com alinhamento de sementes (por exemplo, um conjunto de sequências do mesmo domínio) para cada entrada Pfam, com o objetivo de identificar regiões homólogas entre as sequências que são significativamente semelhantes ao perfil HMM. De posse dos perfis HMM para os diferentes domínios, o

programa InterProScan é capaz de, a partir de uma sequência protéica ainda não caracterizada, predizer possíveis domínios protéicos na mesma, eventualmente predizendo suas funções moleculares (Finn *et al*., 2016).

1.4.3. Gene Ontology

Uma vez que um grande volume de dados genômicos encontram-se disponíveis atualmente, o principal empecilho para analizar estes dados em um arcabouço comparativo e evolutivo é a sistematização da anotação genômica. Grupos de pesquisa distintos podem utilizar termos diferentes para descrever o mesmo processo biológico (e.g. "atividade de degradação de ATP" e "ATPase"). Embora sejam automaticamente interpretáveis por seres humanos, o uso de termos distintos para se referir ao mesmo processo impede que computadores possam analisar estes dados de maneira independente.

Deste modo, é necessário uma linguagem universal que abranja todo o conjunto conceitual de processos biológicos, bem como as interações/relações entre as partes deste conjunto, para anotar de maneira uniforme entidades biológicas que eventualmente realizem a mesma função. Uma ontologia, do ponto de vista das ciências da computação e da informação, é descrita como um “vocabulário formal e controlado que descreve os atributos de um conjunto de elementos (por exemplo classes, termos e conceitos) de um domínio específico (neste caso, o termo “domínio” não se trata de domínio proteíco, e sim de uma determinada área do conhecimento) e seus relacionamentos”. O uso de ontologias permite representar em computadores, de maneira formal e sistematizada, entidades e suas relações.

O Gene Ontology (GO) é uma ontologia que visa padronizar a anotação de produtos gênicos em uma linguagem comum e aplicável para uma enorme gama de organismos, de bactérias a seres humanos (Ashburner *et al*., 2000). Deste modo, o GO proporciona uma representação computacional e funcional dos diferentes produtos gênicos. O GO consiste em um vocabulário formal, padronizado, hierárquico e controlado de termos e as relações entre estes para descrever tanto atributos, funções e características dos genes e seus produtos como também fornecer ferramentas que permitem análisar dados oriundos das anotações dos produtos gênicos. O GO utiliza grandes três ontologias distintas, na qual as funções dos produtos gênicos podem ser descritas em três eixos ortogonais (Munoz-Torres e Carbon *et al*., 2017):

(i) Função Molecular: termos que representam as atividades bioquímicas a nível molecular realizadas por produtos gênicos. As funções moleculares dos termos GO são freqüentemente acrescentadas com a palavra “atividade” para evitar confusão entre os nomes dos produtos gênicos e suas funções moleculares (por exemplo o termo GO de uma proteína quinase é “atividade de proteína quinase”).
(ii) Componente Celular: termos que representam aspectos relativos à anatomia celular, ou seja, as diferentes estruturas celulares onde os produto gênico desempenha uma função. Estes locais abrangem tanto subestruturas celulares (por exemplo o núcleo) e compartimentos celulares (por exemplo uma mitocôndria), quanto complexos macromoleculares estáveis dos quais são partes (por exemplo, o ribossomo).
(iii) Processos Biológicos: termos que representam os diferentes processos biológicos onde o produto gênico atua. Um processo biológico é descrito formalmente como uma série de eventos moleculares realizado por conjuntos de funções moleculares. Embora haja uma certa sobreposição entre funções moleculares e processos biológicos, usualmente os processos biológicos incluem conjuntos de funções moleculares individuais. Novamente, termos GO de processos biológicos abrangem tanto processos específicos (por exemplo metabolismo de pirimidinas) como gerais (por exemplo transdução de sinais, crescimento ou migração celular).

Computacionalmente, os termos GO são representados por um grafo, onde os termos estão representados em nós de uma rede hierarquica, de forma que os termos menos específicos e mais gerais (pai) podem possuir vários termos específicos (filhos) (Thomas, 2017). Suas conexões e interações formam o que é tecnicamente descrito como grafos acíclicos direcionados hierarquicos. A leitura desta rede de conexões deve levar em consideração dois aspectos que definem esta relação entre termos pais e filhos: (i) Quando existem combinações de termos filhos que juntos sempre são integrantes de um mesmo termo pai (part of) e (ii) quando os termo filhos são mais específicos, detalhados e explicativos que o termo pai (is a) (Figura 1).

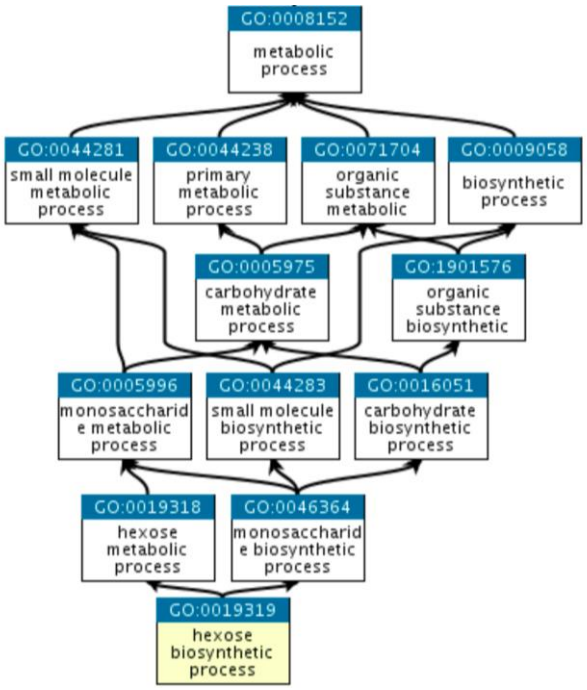

Figura 1 - relações entre termos pais e filhos usados no Gene Ontology.

As setas indicam que o termo filho é parte do termo pai, no exemplo utilizado o termo GO GO:0019319 ("*hexose biosynthetic process*"), destacado em amarelo, é filho dos GOs GO:0019318 ("*hexose metabolic process*") e GO:0046364 ("*monosaccharide biosynthetic process*"). Quando um termo é filho de outro termo mais geral, a existência do termo mais específico implica necessariamente que o gene que também é anotado como todos os termos-pai de maneira recursiva.

Um produto gênico utiliza quantos termos GO forem necessários, de qualquer ontologia, para descrever seu papel biológico da maneira mais clara possível. Diferente de outros bancos de dados como, por exemplo o Pfam para domínios de proteínas, os termos GO levam em consideração apenas o aspecto funcional dos produtos gênicos. Um mesmo termo GO pode estar associado a vários genes ou proteínas que fazem parte do mesmo processo biológico sem refletir necessariamente suas relações de homologia.

1.5 Coeficiente de Correlação

Medidas de associação são métricas que avaliam a correlação entre duas variáveis *x* e *y*, ou seja, avaliam se aumentos de *x* estão associados a aumentos/diminuições de *y*. Dentre as medidas de associação comumente utilizadas, duas destacam-se por seu amplo uso e fácil compreensão: correlação de Pearson, utilizada para detectar associações lineares (onde aumentos de *x* estão associados à mesma taxa de aumento em mesma taxa de aumento em *y*), e correlação de Pearson, utilizada para detectar associações monótonas

(onde aumentos de *x* estão associados a aumento/diminuição de *y*, sejam estes lineares ou não). Ambas as correlações assumem independência entre os dados analisados, o que não é verdadeiro na análise de dados de espécies, onde existe dependência causada por ancestralidade comum.

O coeficiente de correlação de Pearson (r) assume, além da existência de relação linear entre as variáveis, que a distribuição dos dados é normal. O cálculo de r é obtido pela divisão da covariância das duas variáveis pelo produto dos seus desvios-padrão. Como resultado desta divisão pelos desvios, r = [-1,1], sendo que $-1 < r < 1$. Valores negativos indicam correlação negativa (aumentos de *x* estão associados à diminuição de *y*), e valores positivos indicam correlação negativa (aumentos de *x* estão associados ao aumento de *y*). Valores próximos dos extremos de r = -1 ou r = 1 indicam uma forte correlação linear entre as variáveis, e valores próximos a zero indicam correlação fraca ou ausência de correlação.

Como alternativa não-paramétrica ao coeficiente de correlação de Pearson, outros procedimentos estatísticos detectam qualquer tipo de variação monotônica (aumento de x implica em algum aumento/diminuição de *y*, não necessariamente linear). O Coeficiente de Correlação de Spearman (ρ), utilizado como uma generalização do coeficiente de correlação de Pearson, ordena as variáveis em escala ordinal crescente, onde o menor valor de *x* recebe rank 1, o segundo menor, rank 2 e assim sussessivamente. O mesmo é realizado para a variável *y*. À partir do rank produzido para *x* e *y*, o Coeficiente de Correlação de Spearman (ρ) utiliza a fórmula:

$$\rho = 1 - 6\sum d_i^2 \,/\, n\,(n^2-1) \qquad (1)$$

sendo d_i a diferença entre dois valores do *rank* de cada observação e *n* é o número de observações.

1.6 - Assinaturas moleculares associadas ao aumento do número de tipos de células em genomas eucarióticos

Alguns trabalhos têm explorado a associação entre o número de tipos celulares diferentes com a expansão e diversidade do repertório de proteínas, bem como de suas interações e modificações pós tradução, os quais podem eventualmente contribuir para a variação fenotípica do número de tipos celulares. Geralmente as proteínas definem uma considerável parcela da diversidade de funções e estruturas moleculares das células para formarem tecidos especializados nos eucariotos (Di Roberto e Peisajovich, 2013).

Através de análises de genômica comparativa, foi observado que algumas cópias de genes ou famílias de genes podem ter contribuído significativamente para a evolução da multicelularidade, uma vez que são mais frequentes em espécies multicelulares do que em espécies unicelulares (Björklund, Ekman e Elofsson, 2006; Kawashima *et al*., 2009; Miller, 2010). Dentre estas, destacam-se as famílias de proteínas envolvidas em matriz extracelular, adesão, sinalização, diferenciação e receptores celulares de alguns eucariotos mais complexos, como os metazoários, que surgiram originalmente nos ancestrais eucariotos unicelulares e se expandiram nos organimos multicelulares (Anderson *et al*., 2015; Kawashima *et al*., 2009), o que indica possível associação entre essas funções biológicas e a evolução da multicelularidade nas linhagens dos eucariotos (Sebé-Pedrós *et al.*, 2013).

Um estudo clássico conduzido por Patthy, 1999 procurou analisar a distribuição evolutiva de proteínas modulares que evoluíram claramente por embaralhamento de domínios para vários eucariotos. O estudo avaliou a contribuição de embaralhamento de domínios atribuída a categorias funcionais ligadas à multicelularidade em eucariotos superiores como os metazoários e dentre estes os vertebrados. Estas categorias envolvem constituintes da matriz extracelular, processos de remodelação de tecidos, interações e comunicação célula-célula/célula-matriz. Alguns trabalhos demonstraram uma correlação extremamente significativa entre as fronteiras de um exon e as fronteiras de domínios de proteínas conrrepondentes (Liu e Grigoriev, 2004; França *et al*., 2012; Choudhuri, 2014) para vários eucariotos, apoiando a ideia de embaralhamento de domínios mediado por recombinação.

Em vista disto, estudos de genômicas comparativa, onde se avalia como o conteúdo total dos genes codificadores de proteínas, são interessantes para avaliar como a expansão de assinaturas protéicas, tais como domínios, famílias e superfamílias, desempenham um papel fundamental na compreensão das principais categorias funcionais associadas ao aumento da complexidade morfológica nos eucariotos. Um estudo particularmente relevante nessa área foi realizado por Vogel e Chotia, em 2006. Nessa análise, avaliou-se a frequência de famílias / superfamílias codificadas em genomas completos e o número de tipos de células para 38 espécies de eucariotos (incluindo protozoários, fungos, plantas, e animais não-cordados e cordados). Os autores encontraram 194 de 1219 superfamílias de proteínas com frequências associadas ao número de tipos de células (correlação de Pearson > 0,8). Os dados aqui apresentados demostraram que as principais categorias de famílias de proteínas/superfamílias

associadas à multicelularidade em linhagens eucarióticas desempenham papéis no desenvolvimento, processos extracelulares (por exemplo, adesão celular, comunicação celular e resposta imune) e regulação gênica (por exemplo, transdução de sinal e ligação ao DNA).

Embora os estudos anteriores tenham apresentado evidências importantes para permitir a compreensão molecular de eventos genômicos associados ao aumento da complexidade em *Eukarya*, especialmente o trabalho de Vogel e Chotia, em 2006, os mesmos apresentam diversas questões que não foram tratadas adequadamente: 1) não houve a avaliação da qualidade das montagens dos genomas, a qual varia consideravelmente entre espécies e pode interferir na frequência dos elementos genômicos; 2) não houve tratamento estatístico adequado para lidar com o cenário de teste múltiplo de hipóteses, o que infla a frequência de falso-positivos detectados; 3) ao se comparar dados de espécies filogeneticamente relacionadas, deve-se adotar metodologias estatísticas para levar em consideração o fato de que as amostras (espécies) não são independentes; 4) a análise de domínios e/ou superfamílias associados ao aumento da complexidade falha em detectar sequências de origem filogenética distinta, mas que realizam a mesma função molecular (convergência molecular de função), o que é apropriadamente descrito pela anotação utilizando termos GO; 5) esse estudo buscou somente correlações positivas entre o número de tipos de células, não investigando quais funções biológicas e/ou domínios diminuem sua frequência com o aumento da complexidade; 6) esse estudo utilizou a correlação de Pearson, a qual é sensível à presença de *outliers*, enquanto outras metodologias estatísticas de associação (e.g. correlação de Spearman), as quais são baseadas no ordenamento dos pontos, detectam eventualmente associações não-lineares entre variáveis. Nas próximas seções expomos mais detalhadamente os problemas mencionados nos ítens 1-3, uma vez que os demais já foram abordados anteriormente, bem expomos as soluções adotadas nesse projeto para corrigir ou mitigar estas questões.

1.7. Avaliação da qualidade de montagem e a completude dos genomas

Conforme discutimos, o emprego de plataformas de sequenciamento de próxima geração permitiu a rápida produção de genomas completos. Entretanto, a qualidade das montagens destes genomas varia enormemente em função de dois grandes grupos de características: 1) propriedades biológicas do genoma/organismo alvo da montagem, tais

como tamanho do genoma, quantidade de elementos repetitivos, ploidia e heterozigozidade média, dentre outras; 2) características específicas do processo de montagem, tais como o tipo de tecnologia de sequenciamento, cobertura de sequenciamento e algoritmos de montagem, dentre outras. (Parra, Bradnam e Korf, 2007; Gurevich *et al*., 2013; Ou, Chen e Jiang, 2018).

Estudos de genômica comparativa que utilizam informações de funções biológicas são particularmente sensíveis a dados incompletos, uma vez que a detecção da frequência de elementos genômicos associados ao número de tipos celulares depende de uma estimativa fidedigna das frequências. Como exemplo, uma frequência baixa de um domínio em um determinado genoma pode ser devida a um fenômeno biológico real ou porque a qualidade de montagem do genoma é baixa e, eventualmente, algumas regiões genômicas não estão presentes na montagem, causando consequentemente uma anotação funcional incompleta. Conforme demonstrado, a qualidade dos genomas ditos completos pode variar consideravelmente (Waterhouse *et al*., 2017).

Uma alternativa metodológica é o desenvolvimento de softwares que aplicam alguns critérios objetivos para avaliação da qualidade de montagem dos genomas, de modo a selecionar montagens de alta qualidade (Ou, Chen e Jiang, 2018). Um critério simples, porém, com algumas desvantagens consiste em avaliar o tamanho mediano dos *contigs* e dos *scaffolds* montados através da estatística N50. Resumidamente, essa métrica descreve qual é o tamanho mínimo da sequência que cobre 50% da montagem, ou seja, 50% da montagem está contida em sequências maiores do que a sequência cujo tamanho é o N50. A desvantagem dessa metodologia ocorre em função da possibilidade da escolha subjetiva dos parâmetros utilizados pelo procedimento de montagem de genomas para aumentar o N50 sem melhoria objetiva da montagem. Qualquer mudança mínima nos parâmetros utilizados durante a montagem pode ocasionar em fusões errôneas entre as reads para gerar uma contig maior, o que resulta em sequências quiméricas que não existem no cromossomo original (Simão e Waterhouse *et al*., 2015).

Outras estratégias de controle de qualidade de montagem de genomas tem como enfoque a completude do genoma, definida como a porcentagem de genes ortólogos 1-1 táxon-específicos observada dividida pelo total de genes dessa natureza que deveriam estar presentes no genoma. Uma vez que os organismos vivos celulares descendem de ancestrais comuns, divergindo via especiação, eles também compartilham alguns conjuntos de genes entre si. Dessa maneira, pode-se buscar pela presença de tais genes universais em um determinado táxon em genomas recém-montados como critério de

avaliação de montagem de genomas, o qual compreende uma métrica definida por completude genômica. A completude de um genoma é uma interessante métrica que consiste em utilizar um critério biológico e não apenas estrutural (como avaliar o tamanho dos *scaffolds*) para analisar o conteúdo genético esperado para determinada espécie. Estas técnicas, por avaliarem a completude do conteúdo codificador do genoma, capturam de maneira fidedigna a qualidade dos proteomas preditos (Simão e Waterhouse *et al*., 2015).

1.7.1 – BUSCO

O programa BUSCO (*Benchmarking Universal Single-Copy Orthologs*) se tornou extremamente utilizado para realizar o controle de qualidade de montagem de genomas através da detecção de ortólogos 1-1 quase universais táxon-específicos. Os bancos de dados de ortólogos são originados a partir de alguns clados selecionados do *OrthoDB*. O programa BUSCO contém diversos bancos de dados de ortólogos 1-1 de diversos clados, tais como artrópodes, vertebrados, metazoários, fungos e eucariotos, permitindo avaliar objetivamente a qualidade de montagem de diversos genomas dentro de um táxon qualquer em função do seu conteúdo gênico esperado (Kriventseva, *et al*., 2015, Waterhouse, *et al*., 2013).

Estes grupos de ortólogos 1-1 foram selecionados a partir de cada radiação principal da filogenia das espécies com base em alguns critérios, por exemplo a obrigatoriedade que os genes estivessem presentes como ortólogos de cópia única em pelo menos 90% das espécies do clado em questão (Simão e Waterhouse *et al*., 2015). Por essa razão, muitas vezes refere-se aos genes desse programa como ortólogos 1-1 quase universais. Como resultado para tomadas de decisões *a posteriori*, o BUSCO avalia um genoma ou proteoma retornando, para cada ortólogo 1-1 do táxon em questão, qual é o seu status: completo cópia simples (o resultado desejável); fragmentado (tamanho menor do que o observado em outros genomas do mesmo clado); duplicado (mais de uma cópia no genoma em questão) ou ausente. A soma das quatro categorias anteriores totaliza 1, e a porcentagem de cada uma delas permite avaliar objetivamente a qualidade do genoma em função do seu conteúdo gênico esperado (Simão e Waterhouse *et al*., 2015; Waterhouse *et al*., 2017).

1.8 - Teste múltiplo de hipóteses

Uma outra questão que ainda não foi tratada de maneira adequada em estudos de genômica comparativa onde se busca a associação potencial entre a frequência de dezenas de milhares de elementos genômicos (e.g. domínios protéicos) *versus* um fenótipo quantitativo (e.g. números de tipos de células) é o cenário de teste múltiplo de hipóteses que naturalmente emerge dessas análises (Goeman e Solari, 2014). Ao se buscar, no mesmo conjunto de dados (e.g., um mesmo conjunto de genomas em análise), pela associação potencial entre a frequência de milhares de domínios protéicos e o número de tipos de células dos organismos, a cada teste individual realizado há a probabilidade de se observar associações significativas simplesmente pelo fato de que um número suficiente grande de hipóteses foi testado no mesmo conjunto de dados amostral. Como exemplo, ao se estabelecer um valor arbitrário de significância (valor $p < 0,05$), isso significa que, ao se realizar 20 testes onde não há associação entre a frequência de 20 domínios diferentes e o número de tipos de células, espera-se observar em média uma associação ao acaso. Assim, ao se testar milhares de hipóteses no mesmo conjunto, um número inaceitavelmente elevado de associações significativas espúrias são esperados ao acaso.

Esse tipo de situação já é conhecido há décadas em diversos campos da ciência, sendo denominado como o problema das comparações múltiplas (*multiple comparsion problem*), e algumas soluções estatísticas já foram apresentadas. A mais comum chama-se correção de Bonferroni, e consiste em dividir o ponto de corte para aceitar a hipótese como verdadeira (no exemplo acima, 0,05) pelo total de hipóteses sendo testadas. Assim, ao se testar 10.000 hipóteses no mesmo conjunto de dados, o ponto de corte passa a ser 5x10e-6 ao invés do valor original de 5x10e-2. Outras correções menos estringentes que a de Bonferroni também foram propostas, e são amplamente utilizadas em análises genômicas que testam milhares de hipóteses no mesmo conjunto de dados, tais como análises de enriquecimento funcional e estudos de associação em escala genômica. Entretanto, até o momento, nenhum estudo que busca associação significativa entre atributos genômicos e o número de tipos de células em *Eukarya* foi realizado.

1.9 - Métodos comparativos filogenéticos e método de Contrastes Independentes

Desde que Theodosius Dobzhansky, um dos pais da teoria sintética da evolução escreveu um artigo intitulado “Nada em biologia faz sentido exceto à luz da evolução”

(Dobzhansky, 1973), a unificação dos campos da genética, paleontologia, estatística e sistemática, dentre outros, são usadas para explicar como a evolução, através de forças como a seleção natural e as mutações, influenciam a história de vida e a evolução fenotípica nos organismos (Diniz-Filho, 2000). Em especial, diversos estudos visam estudar se duas variáveis quantitativas medidas em diferentes espécies (e.g. peso corporal *versus* área territorial) estão associadas, ou seja, se as suas taxas de variação co-variam. Entretanto, estudos dessa natureza devem considerar a dependência filogenética ao comparar dados de diferentes espécies, uma vez que as similaridades entre espécies podem ser resultantes tanto de herança por ancestralidade comum quanto de processos de convergência (Cornwell e Nakagawa, 2017).

Uma vez que as espécies descendem hierarquicamente de ancestrais comuns e, devido a essa estrutura hierárquica, há dependência entre os dados fenótipicos e genotípicos de diferentes espécies, ocorre a violação da suposição estatística de independência das observações para estudos de associação. Assim, o uso de metodologias como coeficientes de correlação e construção de modelos lineares, dentre outros, não são adequados para a análise de dados de espécies (Felsenstein, 1985).

Para tratar esse fato adequadamente, diversas famílias de métodos computacionais e estatísticos, coletivamente denominados métodos comparativos filogenéticos, foram propostas. Esses métodos utilizam a história filogenética das espécies em análise para corrigir a dependência filogenética dados fenotípicos/genéticos obtidos para as espécies, incorporando assim a informação filogenética aos testes de hipóteses ao se analisar dados de espécies (Cornwell e Nakagawa, 2017). Entre esses métodos, um dos mais populares consiste no cálculo de contrastes filogeneticamente independentes (*phylogenetically independent contrasts* - PIC) (Felsenstein, 1985). Os métodos de PIC sumarizam a quantidade de mudança de um caractere em cada nó interno de uma filogenia. Assim, esses valores são tanto independentes quanto podem ser utilizados para estimar a taxa de mudança de um caractere em uma filogenia, o que passa a ser analizado utilizando modelos lineares simples sem a dependência introduzida pela história filogenética compartilhada.

Após eventos de especiação, pode-se assumir que a evolução genômica das espécies decorrentes desse processo como eventos independentes (embora haja exceções importantes, como a transferência gênica horizontal). Em análises de PIC, os valores fenotípicos/genômicos de cada espécie (e.g. número de tipos de células) são utilizados, em conjunto com uma árvore ultramétrica (onde os tamanhos dos ramos são

proporcionais ao tempo evolutivo), para obter novos valores denominados "contrastes independentes filogeneticamente", os quais correspondem à taxa de mudança em cada nó interno da árvore.

Simplificadamente, o método PIC computa, para cada nó interno de uma filogenia, qual a taxa de mudança do caractere em questão observada desde o evento de especiação assumindo que a evolução do caractere segue movimento browniano, onde ramos maiores das árvores filogenéticas apresentariam uma maior variação mudança do caractere. Se duas espécies "A" e "B" possuem um ancestral comum "C" e desejamos analizar como um caractere "X" evoluiu ao longo da história de "A" e "B", o algoritmo PIC obtém inicialmente o "contraste independente", que consiste na diferença entre os fenótipos/genótipos medidos para as espécies "A" e "B" (X_A - X_B, denominada M_{AB}). Ao assumir a evolução do caractere como um movimento browniano, o valor do contraste obtido tem valor esperado de zero e variância proporcional ao tamanho dos ramos de "A" e "B" na árvore ultramétrica (variância V_{AB} proporcional ao tempo $T_A + T_B$, sendo "T" o tamanho do ramo na árvore). Assim, os valores de contraste normalizado para o nó interno "C" são obtidos dividindo-se o contraste pela variância dos nós terminais "A" e "B" (contraste $C = M_{AB}/V_{AB}$). Cabe observar que esse método também pode ser utilizado para todos os nós mais internos de uma filogenia até a raiz da árvore, produzindo, para N espécies, N-1 pontos independentes, cada um correspondendo à taxa de evolução do caractere "X" nos nós internos da árvore.

Após a obtenção dos valores de PIC para duas variáveis fenotípicas/genotípicas quaisquer "X e "Y", o cálculo de associação entre os PICs consiste no uso de modelos de regressão linear simples. O PIC é uma reformulação de modelos lineares (regressão) clássicos, nos quais os valores fenotípicos originais medidos para espécies individuais não são mais vistos como pontos de dados. Em vez disso, cada ponto de ramificação em uma árvore filogenética de espécies é tratado como um evento independente, o que consiste em uma réplica em um sentido estatístico (Felsenstein, 1985).

2. Hipótese

Os organismos eucarióticos apresentam imensa variação em sua complexidade, a qual pode ser estimada pelo número de tipos celulares observados nesses organismos. Diferentes espécies de eucariotos também apresentam imensa variação em sua composição de elementos genômicos, tais como a frequência de domínios protéicos e de funcões biológicas de genes, representadas por termos GO. Nosso trabalho hipotetiza que os elementos genômicos significativamente associados ao número de tipos de células podem evidenciar eventos importantes para a evolução da multicelularidade, bem como fornecer potenciais candidatos para novos processos biológicos importantes para o surgimento de múltiplos tipos de células.

3. Objetivos

3.1. Objetivos gerais

Detectar funções biológicas codificadas em genomas eucarióticos cujas frequências genômicas estejam significativamente associadas ao número de tipos distintos de células nestes organismos.

3.2. Objetivos específicos

• Busca bibliográfica pelo número de tipos de células e de genomas completos para as espécies de *Eukarya*
• Obtenção dos proteomas não-redundantes para cada espécie
• Avaliação da qualidade dos proteomas não-redundantes via completude genômica
• Anotação *de novo* dos proteomas não-redundantes
• Obtenção da árvore ultramétrica para as espécies analisadas
• Procura por domínios de proteína / termos GO cuja frequência seja significativamente associados ao número de tipos de células, utilizando para isso: 1) a correção para o cenário de testes múltiplos de hipóteses, de modo a obter a robustez estatística necessária; e 2) a utilização do método de contrastes filogeneticamente independentes, de modo a corrigir para a dependência intrínseca nos dados de análise comparativa de espécies filogeneticamente relacionadas

4 - Material e Métodos

A metodologia utilizada nessa dissertação encontra-se representada na Figura 2, a qual será utilizada como guia ao longo dessa seção.

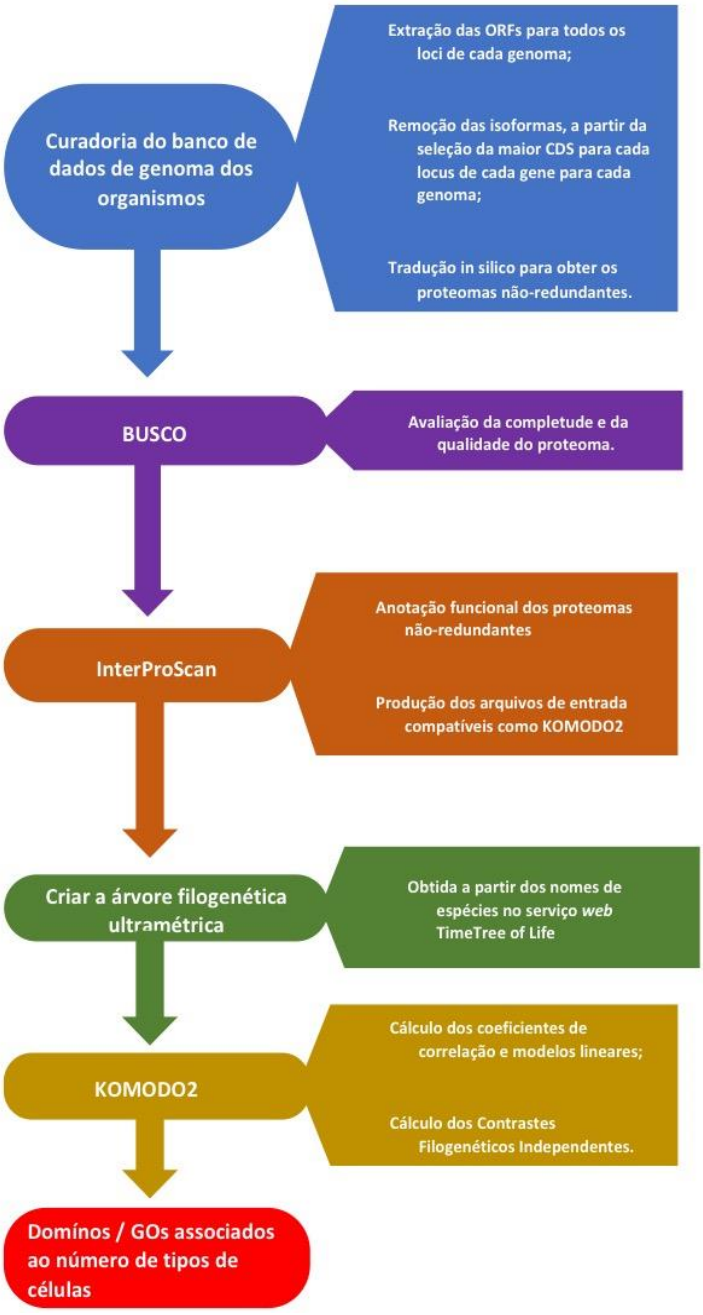

Figura 2 - Protocolo utilizado para calcular as frequências de domínios Pfam e termos Go associados ao aumento da complexidade biológica em Eukarya.

Azul: curadoria dos banco de dados de proteomas não-redundantes e tipos celulares; roxo: avaliação da qualidade genômica em termos do conteúdo gênico esperado; laranja: anotação funcional dos proteomas não-redundantes; verde: obtenção de árvore ultramétrica; amarelo: busca por funções biológicas associadas ao aumento da complexidade em *Eukarya*.

4.1 - Ambiente computacional utilizado no projeto

As análises realizadas nesse projeto foram executadas em um servidor DELL com 2 processadores Intel Xeon E5-4610 v2 2.3GHz totalizando 64 threads; 128GB de memória RAM e sistema operacional CentOS Linux release 7.5.1804 com suporte às linguagens de programação PERL5 versão 7.5 e R versão 3.0.0. Esse servidor também contém os programas descritos nas seções abaixo, os quais foram utilizados para a realização desse projeto.

4.2 - Obtenção do número de tipos celulares para organismos eucarióticos com genomas completos

Inicialmente, selecionamos todas as espécies eucarióticas com genomas completos disponíveis no banco de dados NCBI (pesquisa realizada no dia 05/06/2018). Para cada uma dessas espécies, realizamos uma busca minuciosa por informações sobre o seu número de tipos de células. O estudo realizado por Chen e colaboradores, em 2014, compreende o maior compêndio disponível de números de tipos de células para organismos eucarióticos (Chen *et al*., 2014). Assim, optamos por utilizar esse artigo como base para obtermos o número de tipos de células por espécie. Adcionalmente, estimamos o número de tipos celulares diferentes para seis protistas identificando as suas fases de diferenciação celular durante o ciclo de vida (Black e Boothroyd, 2001; Chen *et al*., 2014; Sebé-Pedrós *et al*., 2013 e Lone e Manohar, 2018).

4.3 - Obtenção dos genomas completos dos organismos utilizados nesse estudo

Cada genoma completo é descrito em um arquivo no formato *genbank*, representado como um arquivo de texto com extensão "gbff". (NCBI, 2013). Arquivos .gbff contém informações sobre a sequência genômica da espécie, bem como diversas informações da anotação genômica disponíveis em vários campos. Especificamente, esses arquivos contém 1) metadados sobre o genoma tais como tamanho, espécie de origem, taxonomia, tipo de molécula (DNA ou RNA), identificadores e versão do genoma; 2) informações sobre os *loci* encontrados no genoma, tais como nome do *locus*, isoformas observadas no *locus*, tamanho de cada isoforma, tipo de gene (codificador ou RNA funcional), localização cromossômica e sequência. As informações disponíveis em arquivos *genbank* permitem que programas de bioinformática extraiam, de maneira automática, porções funcionalmente interessantes dos genomas (e.g., todas as regiões

codificadoras encontradas em um genoma). Uma vez definidas as espécies que possuem tanto genoma sequenciado disponível no NCBI bem como o número de tipos de células, utilizamos ferramentas de bioinformática (*scripts in-house* desenvolvidos na linguagem de programação em PERL) para realizar o *download* do genoma completo a partir do banco de dados público NCBI GenBank (Clark *et al*., 2015).

4.4 - Obtenção dos proteomas não-redundantes

Utilizamos o seguinte algoritmo para a obtenção dos proteomas não-redundantes, definidos como todo o potencial codificador descrito para uma espécie onde cada *locus* encontra-se representado uma vez:

(i) extração das ORFs (Open Reading Frame – janelas abertas de leitura)

Nesse estudo, utilizamos algoritmos que extraem todas as ORFs de todos os *loci* descritas em um arquivo .gbff, as quais representam o transcriptoma redundante já descrito para o genoma da espécie armazenado no arquivo. Adicionalmente, nossos algoritmos também utilizam filtros para verificar a integridade das ORFs. Deste modo, para cada ORF, avaliamos (i) a existência de códons de início e término válidos, respeitando as diferentes tabelas de código genético; (ii) se há a ocorrência de nucleotídeos não-canônicos (nucleotídeos diferentes de “A”, “C”, “T” e “G”) e (iii) se as ORFs são realmente lidas como códons que especificam os aminoácidos, ou seja, se as sequencias codificadoras descritas são múltiplas de três.

Ao final dessa etapa, todas as ORFs que passaram pelos filtros descritos acima foram armazenadas em um arquivo no formato fasta. Para cada ORF, armazenamos também identificadores do seu *locus* de origem, descrito no arquivo genbank através de um nome único (*gene ID* e/ou *locus tag*), bem como identificadores da proteína codificada por cada isoforma (*protein ID*).

ii) Obtenção da maior isoforma por *locus*

Utilizando os arquivos fasta produzidos na etapa anterior, os quais contém, para cada sequência, o seu *locus* e/ou gene de origem, desenvolvemos *scripts* na linguagem PERL para selecionar, para cada *locus* (descrito nos arquivos *genbank* pela flag "locus_tag") e/ou gene (descrito nos arquivos *genbank* pela flag "gene") codificador de

proteína, a maior ORF disonível, a qual contém, em teoria, a maior quantidade de informação de sequência para o gene em questão. No caso de mais de uma isoforma com o mesmo tamanho para um mesmo *locus*, escolhemos aleatoriamente uma sequencia como representante do gene.

(iii) Obtenção das sequencias protéicas dos transcritos

Após extrair a maior CDS para cada todos os *loci* do genoma em questão, utilizamos o identificador da proteína codificada pelo transcrito (*protein ID*) armazenado anteriormente para obtermos a sequência protéica codificada pelo transcrito. O conjunto total das sequências protéicas de cada organismo obtidas através desse procedimento foi denominado proteoma não-redundante do organismo, e foi utilizado nas etapas posteriores conforme descrito nas próximas seções.

4.5 - Avaliação da qualidade dos proteomas através de análise de completude

Os proteomas não-redundantes obtidos na etapa anterior foram analisados quanto à sua completude utilizando o programa BUSCO versão 3.0.2 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). Resumidamente, esse programa busca por ortólogos 1-1 conservados quase universalmente em um determinado táxon como evidência indireta da qualidade da montagem de novos genomas/proteomas não-redundantes de organismos do mesmo táxon. Como resultado, BUSCO retorna quais ortólogos 1-1 foram encontrados no proteoma não-redundante analisado, o número de cópias de cada um deles e a ocorrência de sequências protéicas com tamanho inferior ao tamanho do ortólogo 1-1 analisado. Para a execução de BUSCO, utilizamos os proteomas não-redundantes obtidos na etapa anterior como entrada, e os avaliamos quanto à completude utilizando o banco de dados de ortólogos 1-1 quase universais de *Eukarya*, disponibilizado para *download* na página do programa BUSCO (303 ortólogos 1-1 quase universais encontrados nos genoma de mais de 90% de 100 eucariotos espalhados ao longo da árvore da vida).

Para decidirmos se um proteoma não-redundante está suficientemente completo para as análises posteriores, utilizamos como pontos de corte os seguintes critérios: 1) detecção de 70% ou mais dos 303 ortólogos 1-1 quase universais de *Eukarya*; 2) menos de 10% de ortólogos 1-1 quase universais duplicados; 3) menos de 10% de ortólogos 1-1 fragmentados; 4) menos de 10% de genes ausentes.

4.6 - Anotação funcional dos proteomas não-redundantes

Após a obtenção dos proteomas não-redundantes, procedemos com a anotação *de novo* dos mesmos utilizando o programa InterProScan (Zdobnov e Apweiler, 2001). Resumidamente, esse programa faz uso de diversos bancos de dados, os quais contém domínios, superfamílias e outras assinaturas funcionais encontradas anteriormente em proteínas já caracterizadas funcionalmente, para buscar pelas mesmas assinaturas em proteínas ainda não-anotadas, de modo a predizer as possíveis funções das novas proteínas.

As funções das proteínas analisadas no InterProScan podem ser inferidas ao se detectar, por exemplo, um domínio que possua função conhecida (e.g. "ligação ao DNA"). Adicionalmente, algumas das assinaturas detectadas pelo programa InterProScan possuem termos GO associados às mesmas, o que permite inferir que a função biológica descrita por esse termo está presente na proteína anotada. Ao final dessa etapa, obtivemos arquivos que contém, para cada proteína dos proteomas completos, informações sobre seus domínios e termos GO, quando disponíveis.

4.7 - Obtenção de árvore ultramétrica para as espécies analisadas

Para as análises de contrastes filogenéticamente independentes, faz-se necessário o uso de uma árvore filogenética ultramétrica, onde os tamanhos dos ramos são proporcionais ao tempo evolutivo. Para esse fim, utilizando o serviço *web* TimeTree of Life, o qual recebe como entrada uma lista de espécies e retorna ao final uma árvore ultramétrica das mesmas (Hedges *et al*., 2006; Kumar *et al*., 2017).

4.8. Análises estatísticas - KOMODO2

De posse dos arquivos de anotação dos proteomas completos não-redundantes, obtidos através do programa InterProScan, e da árvore ultramétrica de espécies, obtida através do serviço *web* TimeTree of Life, procedemos com a análise de genômica comparativa visando detectar domínios protéicos e funções biológicas significativamente associadas ao aumento da complexidade em *Eukarya*. Para tal, utilizamos o programa KOMODO2 (não-publicado), desenvolvido em nosso grupo de pesquisa.

4.8.1 - Arquivos de entrada de KOMODO2

O programa KOMODO2 foi desenvolvido para detectar elementos genômicos (domínios protéicos, superfamílias de proteínas, termos GO etc.) significativamente associados à um fenótipo quantitativo, tal como o número de tipos de células de uma espécie. Adicionalmente, KOMODO2 também permite que eventuais viéses causados pela estrutura hierárquica produzida pela especiação seja removido da análise, utilizando para isso o método de contrastes filogeneticamente independentes (PIC) (Felsenstein, 1985). KOMODO2 necessita de dois tipos de arquivos de entrada para realizar sua análise: 1) uma árvore filogenética de espécies, utilizada para o cálculo dos valores de PIC (a qual obtivemos no serviço *web* TimeTree of Life); 2) arquivos de anotação dos proteomas não-redundantes (obtidos através do programa InterProScan). Em nosso estudo, utilizamos duas anotações distintas: uma compreende a frequência de todos domínios protéicos detectados em um determinado proteoma, e a outra consiste na frequência de todos os termos GO encontrados em um proteoma.

4.8.2 - Análises estatísticas utilizando KOMODO2

Após obtermos os arquivos de entrada do programa KOMODO2, procedemos com as análises descritas abaixo para identificarmos as funções moleculares significativamente associadas ao aumento do número de tipos de células em *Eukarya*.

(i) cálculo dos coeficientes de correlação e correção para testes múltiplos de hipóteses

De posse dos arquivos contendo a árvore filogenética ultramétrica de espécies e os arquivos de anotação de cada proteoma não-redundante, procedemos com as análises estatísticas visando computar a correlação entre as frequências dos elementos genômicos (domínios protéicos ou termos GO) e o número de tipos celulares das espécies em análise. Para obter os resultados de correlação o programa KOMODO2 realiza os seguintes procedimentos: 1) cálculo das correlações de Pearson e Spearman entre a frequência dos elementos genômicos e o número de tipos celulares; 2) cálculo dos valores de significância para cada correlação, utilizando a função "cor.test" da linguagem de programação R (R Development Core Team, 2001); 3) correção para teste múltiplo de hipóteses utilizando a função "BH" da linguagem de programação R; 4) produção dos arquivos de resultados (tabelas contendo os resultados para cada elemento genômico:

domínios Pfam ou termos GO), bem como seus valores p-corrigidos. Ressaltamos que as correlações de Pearson foram computadas somente para permitir a sua comparação com as correlações de Spearman, não sendo levadas em consideração para estabelecer associação significativa.

(**ii) análise de contrastes independentes e correção para testes múltiplos de hipóteses**

O método "pic", implementado no pacote "ape" (Paradis & Schliep, 2018), disponível na linguagem de programação R, é uma reformulação de modelos lineares (regressão linear) clássicos, nos quais os valores fenotípicos medidos para espécies individuais (por exemplo, número de tipos de células ou a frequência de um domínio de proteínas) não são mais vistos como pontos de dados. Em vez disso, a taxa de variação após cada evento de especiação em uma árvore filogenética é tratado como uma réplica em um sentido estatístico (Felsenstein, 1985).

Para computar modelos lineares para termos GO e domínios Pfam, procedemos da seguinte maneira: 1) Cálculo das frequências de cada termo GO/domínio Pfam; 2) execução da função "pic", fornecendo como entrada as frequências de termos GO/domínios Pfam e o número de tipos celulares de cada genoma, bem como a árvore filogenética de espécies utilizada; 3) Obtenção dos valores de contrastes independentes pra as frequências de termos GO/domínios Pfam e o número de tipos de células; 4) construção dos modelos lineares para cada termo GO/domínio Pfam, e obtenção de seus valores de significância (probabilidade da regressão possuir inclinação diferente de zero); correção dos valores-p obtidos utilizando o método BH, considerando o cenário de teste múltiplo de hipóteses; 5) produção dos arquivos de resultados (tabelas contendo os resultados para cada elemento genômico: domínios Pfam ou termos GO), bem como seus valores p-corrigidos.

(iii) Filtro e obtenção dos resultados finais

Para considerarmos um determinado elemento genômico como significativamente associado ao número de tipos de células, utilizamos três filtros distintos: 1) valores-p corrigidos menores que 0,05 para o teste de correlação de Spearman; 2) valores-p corrigidos menores que 0,05 para os modelos lineares construídos utilizando a metodologia dos contrastes filogeneticamente independentes; 3) ocorrência mínima de 40

para um determinado elemento genômico, ou seja, o elemento em questão precisa ser observado pelo menos 40 vezes em um ou mais genomas para ser considerado em nossas análises; 4) desvio-padrão acima de um para os dados brutos. A representação gráfica dos resultados de correlação de Spearman e Pearson foi obtida através da linguagem de programação R e do pacote ggplot2 (Wickham, 2016).

4.9 Redução da redundância dos termos GO

Devido à estrutura da ontologia representada pelo GO, diversos termos pais e filhos são parcialmente redundantes, acrescenteando pouca informação. Assim, utilzamos o programa REVIGO (Supek *et al*., 2011) para sumarizar nossos resultados para termos GO, produzindo ao final uma representação visual dos termos GO mais únicos (definição baseada em medidas de semelhança semântica) e mais significativos em uma determinada análise (utilizamos os valores de correlação de Spearman corrigidos para fins de significância). Não utilizamos critérios para sumarizar os domínios Pfam.

5. Resultados

5.1 - Obtenção do número de tipos celulares para organismos eucarióticos com genomas completos

Iniciamos nossa análise com uma revisão completa da literatura buscando espécies com genomas completos disponíveis e informações sobre o número de tipos de células. Encontramos inicialmente 60 espécies com ambas as informações disponíveis (Tabela 1, Figura 3). Deste total, obtivemos o número médio de tipos de células para 54 espécies no bancos de dados do trabalho de Chen *et al.*, 2014. Para as 6 espécies de protozoários remanescentes, encontramos o número de tipos de células nos trabalhos de Black e Boothroyd, 2001; Chen *et al*., 2014; Sebé-Pedrós *et al*., 2013 e Lone e Manohar, 2018. Observamos que os metazoários apresentam o maior número de tipos de células, o qual aumenta nos cordados; em particular, a espécie *Homo sapiens* apresenta o maior número de tipos de células, cerca de 56% maior que a espécie mais próxima relacionada filogeneticamente (*Pan troglodytes*).

Tabela 1: Estimativa do número de tipos de células diferentes para os 60 organismos eucariotos utilizados nesse estudo. As duas primeiras colunas indicam a classificação taxonômica (reino e espécie), a terceira coluna indica os nomes comuns de cada espécie e a quarta coluna indica a média das estimativas do número de tipos celulares diferentes para cada espécie.

Reino	Nome científico	Nome comum	N° tipos celulares (média)
Animalia (non-chordate)	*Trichoplax adhaerens*	Trichoplax	4
Animalia (non-chordate)	*Amphimedon queenslandica*	Sponge	14
Animalia (non-chordate)	*Nematostella vectensis*	Starlet sea anemone	21
Animalia (non-chordate)	*Hydra vulgaris*	Hydra	22
Animalia (non-chordate)	*Caenorhabditis briggsae*	Nematode	28,5
Animalia (non-chordate)	*Caenorhabditis elegans*	Nematode	28,5
Animalia (non-chordate)	*Daphnia pulex*	Water flea	50
Animalia (non-chordate)	*Drosophila melanogaster*	Fruit fly	57
Animalia (non-chordate)	*Apis mellifera*	Homey bee	59
Animalia (non-chordate)	*Anopheles gambiae*	malaria mosquito	64
Animalia (chordate)	*Ciona intestinalis*	Vase tunicate	74
Animalia (chordate)	*Branchiostoma floridae*	Anfioxo	100
Animalia (chordate)	*Danio rerio*	Zebrafish	119,5
Animalia (chordate)	*Takifugu rubripes*	Japanese pufferfish	119,5
Animalia (chordate)	*Tetraodon nigroviridis*	Green spotted pufferfish	119,5
Animalia (chordate)	*Canis lupus*	Dog	129
Animalia (chordate)	*Xenopus tropicalis*	Western clawed frog	129,5
Animalia (chordate)	*Mus musculus*	Mouse	130,5
Animalia (chordate)	*Bos taurus*	Cow	138,5
Animalia (chordate)	*Rattus norvegicus*	Brown rat	138,5
Animalia (chordate)	*Gallus*	Chicken	154
Animalia (chordate)	*Felis catus*	Cat	159
Animalia (chordate)	*Tupaia belangeri*	Northern treeshrew	159
Animalia (chordate)	*Pan troglodytes*	Chimpanzee	169
Animalia (chordate)	*Homo sapiens*	Human	264,5
Fungi	*Kluyveromyces lactis*	Yeast	3,05
Fungi	*Yarrowia lipolytica*	Yeast	3,05
Fungi	*Encephalitozoon cuniculi*	microsporidium	3,35
Fungi	*Saccharomyces cerevisiae*	Baker's yeast	4,05
Fungi	*Phanerochaete chrysosporium*	white rot fungus	4,35
Fungi	*Schizosaccharomyces pombe*	Fission yeast	4,35
Fungi	*Ustilago maydis*	Corn smut	4,35
Fungi	*Aspergillus nidulans*	filamentous fungi	5,55
Fungi	*Neurospora crassa*	filamentous fungi	5,55
Plantae	*Chlamydomonas reinhardtii*	Green alga	1
Plantae	*Ostreococcus lucimarinus*	marine green alga	1
Plantae	*Ostreococcus tauri*	marine green alga	1

Plantae	*Volvox carteri*	Green algae	2
Plantae	*Physcomitrella patens*	Spanish moss	21
Plantae	*Selaginella moellendorffii*	Spikemoss	25
Plantae	*Brachypodium distachyon*	Purple false brome	27,25
Plantae	*Sorghum bicolor*	Sorghum	27,25
Plantae	*Vitis vinifera*	Common grape vine	27,25
Plantae	*Populus trichocarpa*	Eastern balsam poplar	28,5
Plantae	*Oryza sativa*	Rice	35
Plantae	*Arabidopsis thaliana*	Thale cress	59
Plantae	*Zea mays*	Maize	63,625
Protist	*Capsaspora owczarzaki*	Amoeboid holozoan	3
Protist	*Dictyostelium discoideum*	Slime mould	4,65
Protist	*Entamoeba histolytica*	Amoeba	4,65
Protist	*Toxoplasma_gondii*	toxoplasma	5
Protist	*Saprolegnia_parasitica*	oomycete water moulds fish parasite	6
Protist	*Achlya hypogyna*	Oomycete water mold	6
Protist	*Leishmania major*	leishmaniasis parasites	7,85
Protist	*Phytophthora ramorum*	oomycete plant pathogen	7,85
Protist	*Plasmodium falciparum*	oomycete plant pathogen	7,85
Protist	*Theileria annulata*	theileriosis parasites	7,85
Protist	*Trypanosoma brucei*	sleeping sickness parasites	7,85
Protist	*Phytophthora_sojae*	oomycete plant pathogen	7,85
Protist	*Phytophthora infestans*	oomycete plant pathogen	7,85
Protist	*Phytophthora parasitica*	oomycete plant pathogen	7,85

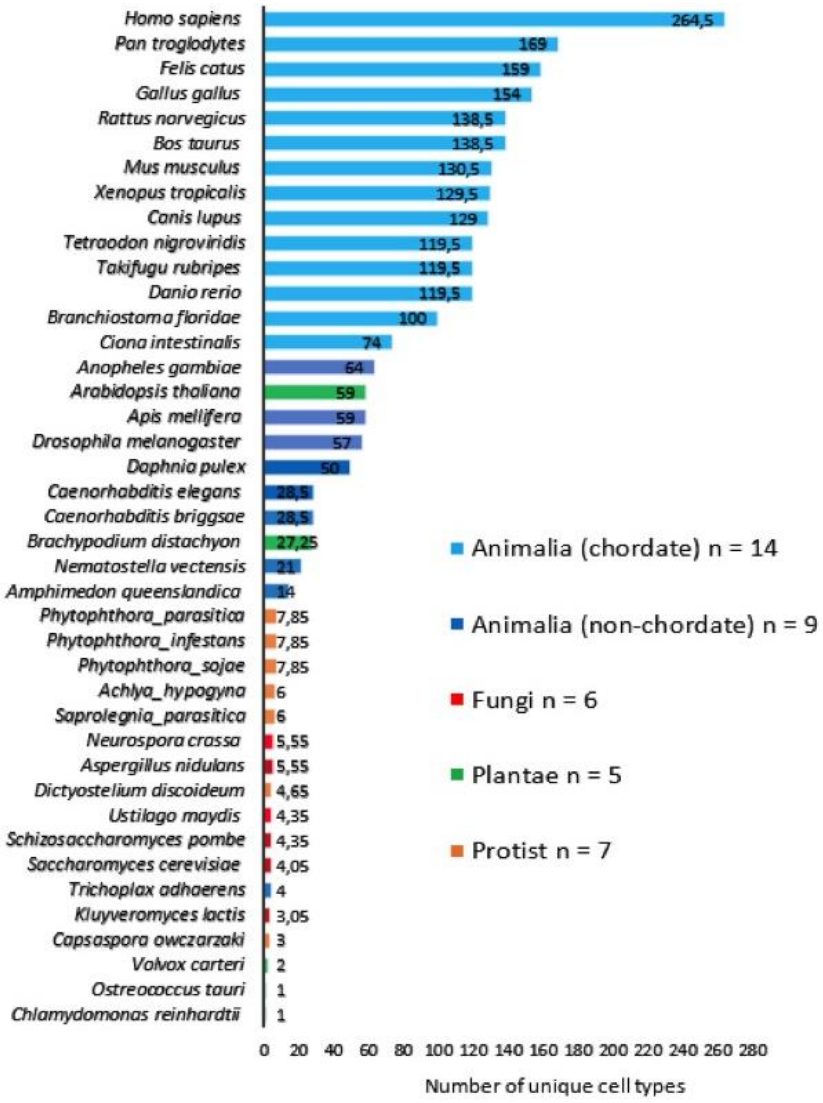

Figura 3 - Números de tipos celulares diferentes para os 41 genomas de eucariotos com alta completude genômica.

As cores na legenda indicam os reinos em que as espécies pertencem e os valores indicam o número de espécies por reino. Os valores para os tipos celulares diferentes para cada espécie estão após as barras coloridas no gráfico de barras**.**

5.2 - Obtenção dos genomas completos dos organismos utilizados nesse estudo

Após obtermos, a partir do banco de dados de genomas do NCBI, os arquivos genbank contendo o genoma das espécies para as quais obtivemos também os números de tipos de células distintas, observamos que, dentre as 60 espécies, três destas (*Tupaia belangeri*, *Phanerochaete chrysosporium* e *Phytophthora ramorum*) não continham anotações das regiões codificadoras em seus respectivos arquivos. Consequentemente, estas espécies foram excluídas as análises posteriores, e prosseguimos nossas análises com as 57 espécies remanescentes.

5.3 - Obtenção dos proteomas não-redundantes

De posse dos genomas das 57 espécies que também possuem número de tipos de células estabelecidos na literatura científica, procedemos com a extração de todas as isoformas descritas para cada espécie, denominada por nós como "transcriptoma codificador redundante". Essas sequências foram obtidas a partir dos arquivos *genbank* através identificação de regiões genômicas que possuem uma *flag* "translation" a flag "gene" ou "locus_tag". Todas estas regiões as regiões codificadoras foram então sumarizadas para obtermos somente uma única sequência protéica representativa por *locus* e/u gene, o qual denominamos como proteoma não redundante de cada espécie (Tabela 2). Para cada espécie, calculamos também a porcentagem de redução dos proteomas, a qual foi obtida dividindo-se o número de sequências do proteoma não-redundante pelo transcriptoma codificador redundante, subtraíndo-se um do resultado e multiplicando-se o número obtido por 100. Observamos que as maiores reduções foram observadas nos organismos-modelo mais estudados.

Tabela 2: Avaliação dos proteomas utilizados nesse estudo. As colunas "proteoma redundante" e "proteoma não-redundante" contém, respectivamente, o número total de transccritos produzidos por *loci* transcricionalmnte ativos e o número total de proteínas dos *loci* codificadores, onde a maior sequência disponível foi escolhia como representante do *locus*.

Nome comum	Espécie	Transcriptoma codificador redundante	Proteoma não-redundante	% redução
Cow	*Bos taurus*	49107	19613	60,06
Anfioxo	*Branchiostoma floridae*	28623	28226	1,39
Dog	*Canis lupus familiaris*	58776	18970	67,72
Vase tunicate	*Ciona intestinalis*	21099	12948	38,63
Zebrafish	*Danio rerio*	57100	25831	54,76
Cat	*Felis catus*	54726	18378	66,42
Chicken	*Gallus gallus*	46393	18064	61,06
Human	*Homo sapiens*	114419	20026	82,50
Mouse	*Mus musculus*	78443	21949	72,02
Chimpanzee	*Pan troglodytes*	79940	21975	72,51
Brown rat	*Rattus norvegicus*	56113	21525	61,64
Japanese pufferfish	*Takifugu rubripes*	31052	19817	36,18
Green spotted pufferfish	*Tetraodon nigroviridis*	27918	26711	4,32
Western clawed frog	*Xenopus tropicalis*	39715	20166	49,22
Sponge	*Amphimedon queenslandica*	24116	19506	19,12
Malaria mosquito	*Anopheles gambiae*	14102	12325	12,60
Honey bee	*Apis mellifera*	22456	9860	56,09
Nematode	*Caenorhabditis briggsae*	21959	21399	2,55
Nematode	*Caenorhabditis elegans*	28420	20063	29,41
Water flea	*Daphnia pulex*	30611	29873	2,41
Fruit fly	*Drosophila melanogaster*	30493	13862	54,54
Hydra	*Hydra vulgaris*	21993	16391	25,47
Starlet sea anemone	*Nematostella vectensis*	24780	24328	1,82
Trichoplax	*Trichoplax adhaerens*	11520	11417	0,89
Filamentous fungi	*Aspergillus nidulans*	9556	9546	0,10
Microsporidium	*Encephalitozoon cuniculi*	1996	1996	0,00
Yeast	*Kluyveromyces lactis*	5085	5084	0,02
Filamentous fungi	*Neurospora crassa*	10812	9748	9,84
Baker's yeast	*Saccharomyces cerevisiae*	6002	5980	0,37
Fission yeast	*Schizosaccharomyces pombe*	5132	5122	0,19

Corn smut	*Ustilago maydis*	6782	6735	0,69
Yeast	*Yarrowia lipolytica*	6472	6465	0,11
Thale cress	*Arabidopsis thaliana*	48350	27475	43,17
Purple false brome	*Brachypodium distachyon*	33944	24719	27,18
Green alga	*Chlamydomonas reinhardtii*	14488	13814	4,65
Rice	*Oryza sativa*	28555	28329	0,79
Green algae	*Ostreococcus lucimarinus*	7603	7572	0,41
marine green alga	*Ostreococcus tauri*	7766	7557	2,69
Spanish moss	*Physcomitrella patens*	35934	35709	0,63
Eastern balsam poplar	*Populus trichocarpa*	45942	40992	10,77
Spikemoss	*Selaginella moellendorffii*	34746	34679	0,19
Sorghum	*Sorghum bicolor*	39248	28093	28,42
Common grape vine	*Vitis vinifera*	41208	25306	38,59
Green algae	*Volvox carteri*	14436	14201	1,63
Maize	*Zea mays*	58411	37148	36,40
Slime mould	*Dictyostelium discoideum*	14406	13781	4,34
Oomycete water mold	*Achlya hypogyna*	8318	8261	0,69
Amoeba	*Entamoeba histolytica*	13315	13119	1,47
Leishmaniasis parasite	*Leishmania major*	8163	8010	1,87
Oomycete plant pathogen	*Phytophthora ramorum*	8316	8311	0,06
Plasmodium	*Plasmodium falciparum*	5339	5290	0,92
Theileriosis	*Theileria annulata*	17797	17566	1,30
Sleeping sickness	*Trypanosoma brucei*	27942	22557	19,27
Toxoplasma	*Toxoplasma gondii*	8318	8318	0
Oomycete plant pathogen	*Phytophthora sojae*	5339	5304	0,66
Oomycete plant pathogen	*Phytophthora infestans*	3795	3789	0,16
Oomycete plant pathogen	*Phytophthora parasitica*	9822	9790	0,33

5.4 - Avaliação da qualidade dos proteomas através de análise de completude

De posse dos proteomas não-redundantes obtidos na etapa anterior, utilizamos o software BUSCO para avaliar a completude dos mesmos como controle de qualidade dos procedimentos anteriores. Selecionamos os seguintes pontos de cortes para a tomada de decisão ao avaliar a qualidade dos proteomas, sendo que a falha em observar qualquer um dos mesmos foi motivo para exclusão das espécies das análises posteriores: 1) ≥ 70% de genes completos (cópias simples + duplicações); 2) ≥ 70% de cópias simples; 3) ≤ 10% de genes duplicados; 4) ≤ 10% de genes fragmentados e 5) ≤ 10% de genes ausentes. Os resultados da execução do programa BUSCO nos proteomas não-redundantes está descrito na tabela 2 e pode ser visualizado na figura 4. Algumas espécies foram excluídas por mais de um critério.

Em relação ao critério "1", três espécies de organismos unicelulares (*Encephalitozoon cuniculi, Entamoeba histolytica* e *Theileria annulata*) foram excluídas, apresentando completude total (cópias simples + duplicações) de 60,1%, 65,7% e 64,7%, respectivamente. O critério "2" excluiu nove espécies (*Encephalitozoon cuniculi* (59,1%), *Selaginella moellendorffii* (7,9%), *Physcomitrella patens* (65,7%), *Zea mays* (55,1%), *Entamoeba histolytica* (54,1%), *Leishmania major* (67,7%), *Plasmodium falciparum* (69%), *Theileria annulata* (62,7%) *e Trypanosoma brucei* (67,7%)).

Oito espécies de organismos apresentaram mais de 10% de ortólogos 1-1 quase universais com duplicações (*Oryza sativa, Ostreococcus tauri, Physcomitrella patens, Populus trichocarpa, Selaginella moellendorffii, Sorghum bicolor, Vitis vinífera, Zea mays e Entamoeba histolytica*), motivo pelo qual foram excluídas das análises posteriores. Nenhuma espécie foi excluída por apresentar ortólogos 1-1 fragmentados. Finalmente, nove espécies (*Hydra vulgaris, Encephalitozoon cuniculi, Ostreococcus lucimarinus, Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Theileria annulata, Trypanosoma brucei e Toxoplasma_gondii*) apresentaram valores superiores a 10% de genes ausentes, sendo excluídas por essa razão.

Após a análise de qualidade dos proteomas utilizando o programa BUSCO, observamos que o grupo com a menor quantidade de cópias simples e genes ausentes são os protistas, por outro lado o grupo com a maior quantidade de genes com duplicações foram as plantas e as algas. Deste modo, eliminamos de nossas análises um total de 15 organismos: cinco protozoários (*Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Theileria annulata, Trypanosoma brucei e Toxoplasma_gondii*), oito plantas

e algas (*Oryza sativa*, *Ostreococcus tauri*, *Physcomitrella patens*, *Populus trichocarpa*, *Selaginella moellendorffii*, *Physcomitrella patens*, *Populus trichocarpa* e *Zea mays*), um fungo (*Encephalitozoon cuniculi*) e um animal não-cordado (*Hydra vulgaris*).

Prosseguimos nossas análises com um total de 42 espécies. Nesse momento, cabe ressaltar que, dentre as espécies com maior número de ortólogos 1-1 quase universais em cópia simples e, portanto, com maior qualidade do proteoma, observamos organismos dos mais diversos grupos taxonômicos, tais como plantas (*Arabidopsis thaliana* e *Brachypodium distachyon*), Fungos (*Neurospora crassa*) e animais (*Caenorhabditis elegans*, *Drosophila melanogaster Pan troglodytes*, *Felis catus* e *Homo sapiens*), indicando que nossa metodologia foi capaz de manter a diversidade taxonômica dos proteomas. De maneira geral, observamos também que os organismos mantidos após a análise do BUSCO compreendem principalmente organismos-modelo.

Tabela 3: Resultados do BUSCO para os 57 organismos. M = ortólogos ausentes (missings); F = ortólogos considerados fragmentados (fragmented); D = ortólogos duplicados (duplicated); S = Ortólogos de Cópias simples (singles) e C = completos (soma de ortólogos de cópia simples e duplicados). Os resultados estão em porcentagem.

Scientific name	commom name	BUSCO Results (%, n = 303)				
		BUSCOS cutoffs: 70% S, 10% D, 10% F and 10% M				
		M	**F**	**D**	**S**	**C**
Drosophila melanogaster	fruit fly	0	0	2,3	97,7	100
Mus musculus	mouse	0	0	5	95	100
Neurospora crassa	filamentous fungi	0	0	0	100	100
Felis catus	cat	0,3	0	3,3	96,4	99,7
Pan troglodytes	chimpanzee	0	0,3	3,3	96,4	99,7
Brachypodium distachyon	purple false brome	0,1	0,2	0,9	98,8	99,7
Arabidopsis thaliana	thale cress	0,1	0,5	0,5	99,1	99,6
Caenorhabditis elegans	nematode	0,4	0,3	1,3	98	99,3
Bos taurus	cow	0	0,7	5,6	93,7	99,3
Canis lupus	dog	0	0,7	3,6	95,7	99,3
Apis mellifera	honey bee	0,7	0,3	1,3	97,7	99
Homo sapiens	human	0	1	2,6	96,4	99
Ustilago maydis	corn smunt	0,7	0,3	0	99	99
Populus trichocarpa	eastern balsam poplar	0,3	0,7	20,2	78,8	99
Yarrowia lipolytica	yeast	0,6	0,7	0,7	98	98,7
Vitis vinifera	common grape vine	0,7	1	14,5	83,8	98,3
Rattus norvegicus	brown rat	1,2	0,7	8,3	89,8	98,1
Physcomitrella patens	spanish moss	1	1	32,3	65,7	98
Zea mays	maize	0,7	1,3	42,9	55,1	98
Anopheles gambiae	malaria mosquito	1,3	1	2,3	95,4	97,7
Daphnia pulex	water flea	2,3	0	3,6	94,1	97,7
Danio rerio	zebrafish	2	0,3	4,6	93,1	97,7
Sorghum bicolor	sorghum	1	1,3	23,1	74,6	97,7
Caenorhabditis briggsae	nematode	0,9	2	4	93,1	97,1
Trichoplax adhaerens	trichoplax	1,7	1,3	0,3	96,7	97
Schizosaccharomyces pombe	fission yeast	3,6	0,3	3	93,1	96,1

Selaginella moellendorffii	spikemoss	1,4	2,6	88,1	7,9	96,1
Dictyostelium discoideum	slime mould	3,6	0,7	3,6	92,1	95,7
Phytophthora parasitica	oomycete plant pathogen	3,9	0,7	2,3	93,1	95,4
Amphimedon queenslandica	sponge	2,9	2	5,3	89,8	95,1
Kluyveromyces lactis	yeast	4,3	0,7	1,3	93,7	95
Saccharomyces cerevisiae	baker's yeast	4,7	0,3	3,6	91,4	95
Takifugu rubripes	japanese pufferfish	2,9	2,3	5	89,8	94,8
Nematostella vectensis	starlet sea anemone	2,7	2,6	3	91,7	94,7
Gallus gallus	chicken	5	1	2,3	91,7	94
Branchiostoma floridae	anfioxo	3,7	2,6	4,6	89,1	93,7
Achlya hypogyna	oomycete water mold	4,6	2	3,3	90,1	93,4
Capsaspora owczarzaki	amoeboid holozoan	6,3	0,3	0,3	93,1	93,4
Phytophthora infestans	oomycete plant pathogen	5,7	1,3	6,9	86,1	93
Ciona intestinalis	vase tunicate	7,3	0,3	0,3	92,1	92,4
Xenopus tropicalis	western clawed frog	6,9	1,3	1,7	90,1	91,8
Saprolegnia parasitica	oomycete water moulds fish parasite	5,6	3,3	6,9	84,2	91,1
Aspergillus nidulans	filamentous fungi	5	4,3	0,3	90,4	90,7
Volvox carteri	green algae	2,9	6,6	1,7	88,8	90,5
Phytophthora sojae	oomycete plant pathogen	7,6	2,3	3,6	86,5	90,1
Oryza sativa	rice	4	6,9	14,5	74,6	89,1
Chlamydomonas reinhardtii	green alga	3,3	8,3	1,3	87,1	88,4
Tetraodon nigroviridis	green spotted pufferfish	5,3	6,6	4,6	83,5	88,1
Ostreococcus tauri	marine green alga	9,5	3	1,7	85,8	87,5
Hydra vulgaris	hydra	10,5	4,3	3	82,2	85,2
Ostreococcus lucimarinus	marine green alga	11,9	3,6	4	80,5	84,5
Toxoplasma gondii	toxoplasma	14,9	5,9	3	76,2	79,2
Leishmania major	leishmaniasis parasites	19,8	6,6	5,9	67,7	73,6
Trypanosoma brucei	sleeping sickness parasites	21,8	5,9	4,6	67,7	72,3
Plasmodium falciparum	malaria parasites	26,4	2,6	2	69	71
Entamoeba histolytica	amoeba	27,4	6,9	11,6	54,1	65,7
Theileria annulata	theileriosis parasites	29	6,3	2	62,7	64,7
Encephalitozoon cuniculi	microsporidium	35,9	4	1	59,1	60,1

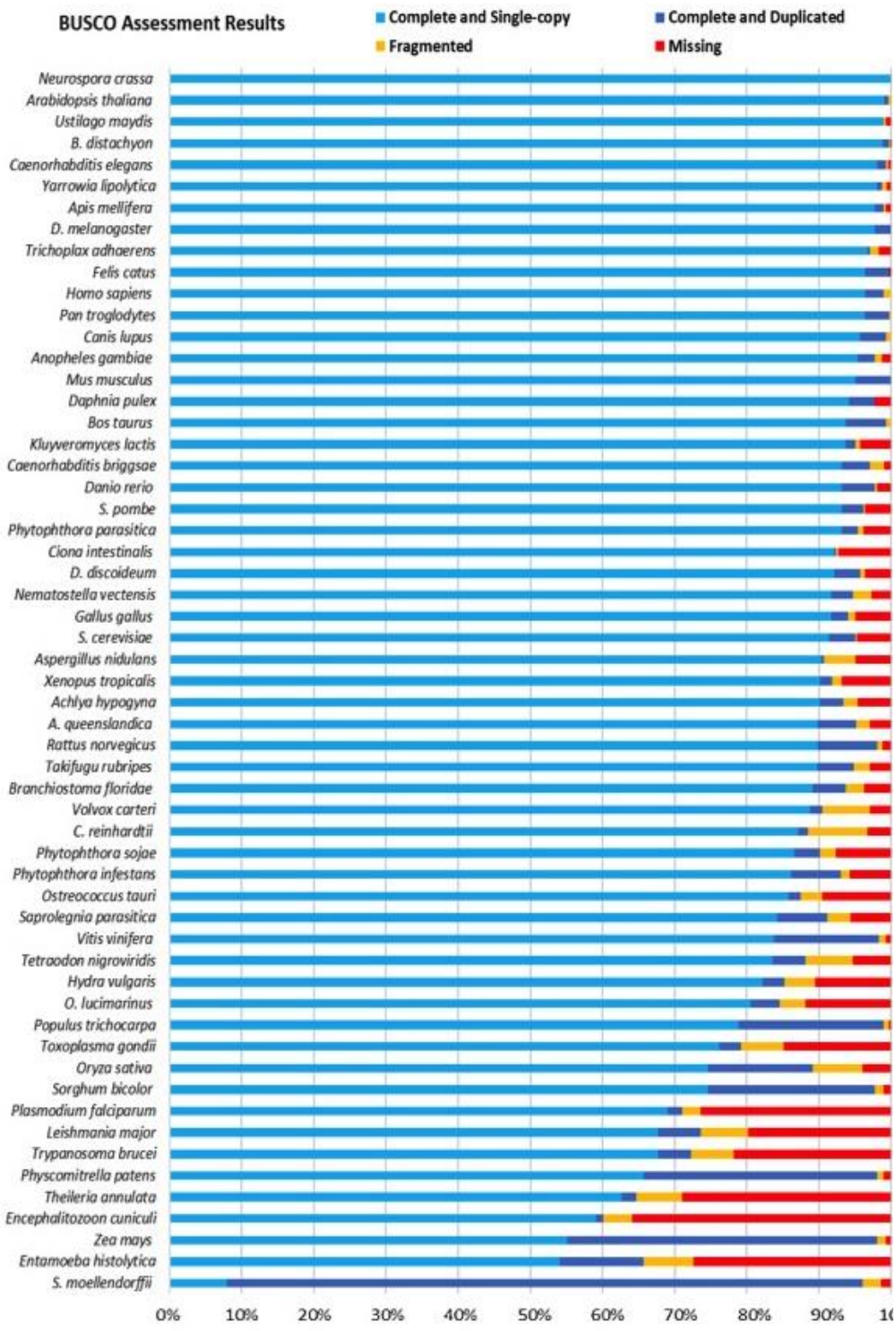

Figura 4 - Resultados do BUSCO para os 60 organismos.

M = ortólogos ausentes (missings); F = ortólogos considerados fragmentados (fragmented); D = ortólogos duplicados (duplicated); S = Ortólogos de Cópias simples (singles) e C = completos (soma de ortólogos de cópia simples e duplicados). Os resultados estão em porcentagem.

5.5 - Obtenção de árvore ultramétrica para as espécies analisadas

De posse das 42 espécies de organismos que apresentam proteoma não-redundante de alta qualidade, utilizamos o serviço *web* TimeTree of Life para produzir uma árvore ultramétrica dos mesmos (Hedges *et al*., 2006), a qual é um dos dados necessários para a obtenção dos dados de contrastes independentes. Nessa etapa, excluímos um organismo (*Yarrowia lipolytica*) das nossas análises, uma vez que essa espécie não se encontra representada no *site* do The Timetree of Life. A árvore final produzida nessa análise encontra-se na figura 5.

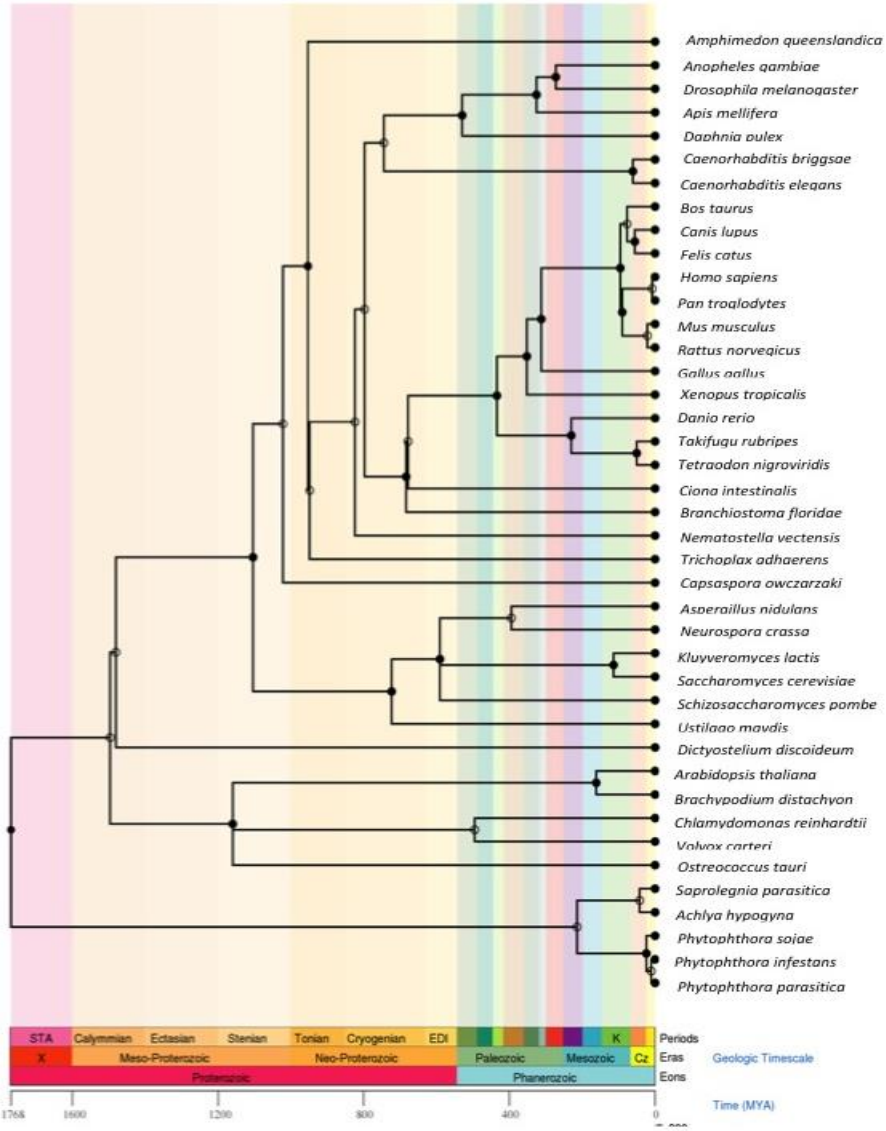

Figura 5 - Árvore filogenética ultramétrica com as 41 espécies disponíveis no site TimeTree of life.

Portanto, do total de 60 espécies para as quais localizamos seus genomas completos, bem como a informação sobre o número de tipos de células que as mesmas

possuem, chegamos a 41 espécies que possuem proteoma não-redundante de alta qualidade e representação no serviço *web* TimeTree of Life: seis protozoários, seis fungos, cinco plantas e algas, nove animais invertebrados não-cordados, incluindo dois placozoários, um porifera, um cnidário, dois nematóides e três insetos. Dentre os 14 animais cordados observamos dois urocordatos, três peixes teleostei, um anfíbio, um pássaro e seis mamíferos (figura 3).

5.6 - Anotação funcional dos proteomas não-redundantes

Os proteomas não-redundantes das 41 espécies de eucariotos foram anotados utilizando o programa InterProScan, o qual busca por domínios, superfamílias e outras assinaturas moleculares em sequencias protéicas primárias, de modo a determinar as possíveis funções biológicas codificadas em cada um dos proteomas. Assim, cada proteína analisada pelo programa InterProScan poderá ser anotada como possuíndo zero ou mais domínios protéicos, bem como zero ou mais termos GO. No caso da análise de domínios, computamos, para cada domínio, sua frequência, dividindo a contagem do domínio no proteoma em questão pelo total de domínios no proteoma, obtendo assim a frequência relativa de cada domínio em cada proteoma. Um procedimento análogo foi realizado para cada termo GO. Esse passo é importante, pois permite comparar proteomas de tamanhos e porcentagem de anotação distintos, de modo a evitar possíveis viéses causados por proteomas particularmente maiores ou com maior porcentagem de anotação que outros.

5.7. Funções biológicas associadas ao aumento da complexidade biológica em eucariotos

Para investigar quais são as funções biológicas, aqui representadas pelas frequências de domínios Pfam ou termos GO, estão significativamente associadas com a complexidade biológica, utilizamos o programa KOMODO2 para primeiramente computar os valores-p para os coeficientes de correlação de Spearman e para os modelos lineares após a correção para o viés filogenético (contrastes filogeneticamente independentes). Após extrairmos os valores-p individuais e executamos a correção de múltiplas hipóteses, consideramos como modelos significativos com p-valores corrigidos menores que 0,05 para ambos os testes. Adicionalmente, somente consideramos os domínios/termos GO que 40 ou mais ocorrências em nossas análises, de modo a evitar

possíveis associações espúrias causadas por funções pouco frequentes nos proteomas não-redundantes.

5.7.1 Domínios Pfam associados ao número de tipos de células

Em nossa avaliação dos domínios protéicos preditos utilizando o banco de dados Pfam, detectamos um total de 1043783 domínios preditos, totalizando 9193 tipos de domínios Pfam distintos com ao menos uma cópia detectada em pelo menos um dos genomas analisados. Avaliamos as frequências de famílias de domínios de proteínas que estão significativamente associados com o número de tipos celulares diferentes utilizando os mesmos critérios de corte utilizados para a análise de GO (valores-p corrigidos para correlação de Pearson e PIC < 0,05, um valor mínimo de ocorrência de 40 domínios, desvio-padrão da contagem de domínios superior a um), obtendo ao final um total de 216 Pfams (~ 2,3%). Observamos um total de 183 domínios com correlação positiva, e 33 com correlação negativa (Tabela Suplementar 1). Essas correlações se expandem principalmente em vertebrados, apresentando abundância intermediária em outros animais e, eventualmente, embora com menor frequência de ocorrência, em plantas, sendo de baixa frequência ou ausentes em fungos e protozoários. Examinamos as funções biológicas dos domínios de proteínas encontradas no banco de dados da Pfam para entender como os eventos de expansão dos mesmos teriam favorecido o surgimento de novos tipos de células, tecidos e órgãos.

Após extensa curadoria manual dos domínios, e com o auxílio das análises de processos biológicos utilizando termos GO, detectamos alguns temas comuns aos mesmos. Dentre as correlações positivas, detectamos 37 domínios associados ao desenvolvimento de sistemas, órgãos e tecidos específicos, atuando como determinadores de forma, função e expressão gênica em nível de sistema-órgão-tecido (Figura 6, A-D). Dentre estes, detectamos onze domínios associados ao desenvolvimento do sistema nervoso (e.g. *Folate receptor family* (PF03024), *Teneurin Intracellular Region* (PF06484)), dez associados ao sistema imune (e.g. *TLR4 regulator and MIR-interacting MSAP* (PF11938), *Beta defensin* (PF13841)), seis associados ao sistema muscular (*Myosin N-terminal SH3-like domain* (PF02736), *Tropomyosin* (PF00261)), quatro associados à formação de pele (e.g. *Filaggrin* (PF03516)), três associados ao sistema sanguíneo (*Alpha-2-macroglobulin family* (PF00207)), dois associados ao sistema reprodutório (e.g. *FAM75 family* (PF14650)) e um associado ao sistema digestivo

(*Proline-rich* (PF15240)). Observamos também 27 domínios associados à processos de matrix extracelular e de interações célula-célula, necessários para o estabelecimento de órgãos e tecidos (e.g. *Plectin repeat* (PF00681) e *Tissue factor* (PF01108)) (Figura 7 A).

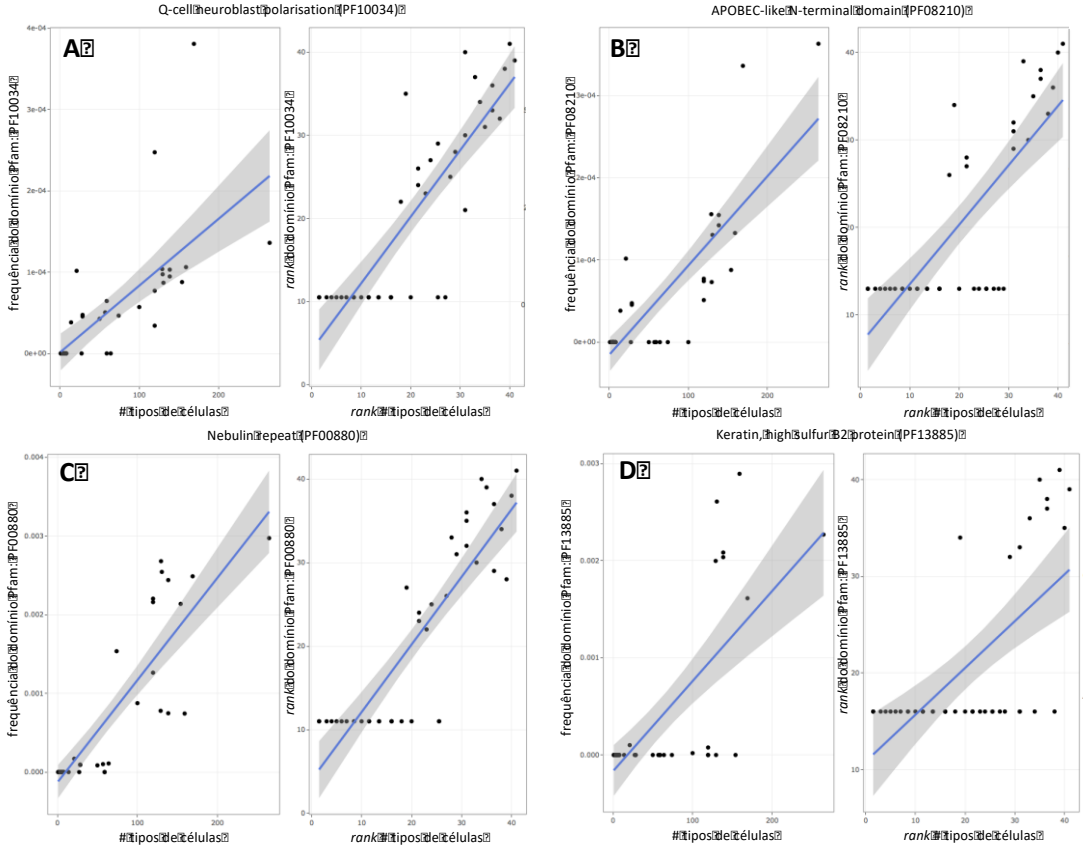

Figura 6 - Exemplos de domínios Pfam com correlação positiva significativa com o número de tipos celulares associados à emergência de sistemas, órgãos e tecidos.

A) Sistema nervoso; B) Sistema imune; C) sistema muscular; D) pele.

Em nível de processos celulares, detectamos 95 domínios, dentre os quais destacam-se as vias de sinalização celular (27 domínios, como *JNK_SAPK-associated protein-1* (PF09744), *N-terminal of Homeobox Meis and PKNOX1* (PF16493), *Retinol binding protein receptor* (PF14752) e), fatores de transcrição (23 domínios, como *SOCS box* (PF07525), *C-myb, C-terminal* (PF09316), *Cystine-knot domain* (PF00007) e *Forkhead domain* (PF00250)), canais de membrana (10 domínios, como *Voltage-gated calcium channel subunit alpha, C-term* (PF16885) e *Calcium-activated potassium channel, beta subunit* (PF03185)), citoesqueleto (10 domínios, como *Tektin family* (PF03148) e *Repeat in HS1/Cortactin* (PF02218)), processos de modificação de cromatina (7 domínios, como *Nucleoplasmin/nucleophosmin domain* (PF03066) e *C-terminal domain of methyl-CpG binding protein 2 and 3* (PF14048)), biologia de RNA (6

domínios, como *PurA ssDNA and RNA-binding protein* (PF04845) e *5'-nucleotidase* (PF06189)), sistemas de endomembranas (5 domínios, como *Putative golgin subfamily A member 2-like protein 5* (PF15070)), biologia do núcleo eucariótico (3 domínios, como *Nuclear pore complex interacting protein* (PF06409)), degradação de proteínas (2 domínios, como *Proteasome activator pa28 alpha subunit* (PF02251)), recombinação de DNA (1 domínio - *Holliday junction regulator protein family C-terminal repeat* (PF12347)), transcrição basal (1 domínio - *Zinc knuckle* (PF15288)), e homeostase do sistema redox (1 domínio -*Thioredoxin-like domain* (PF13848)). Detectamos ainda oito domínios com pouca informação biológica, o que não prmitiu classificá-los nas categorias acima, e 15 domínios de função desconhecida (DUFs) (Figura 8 D).

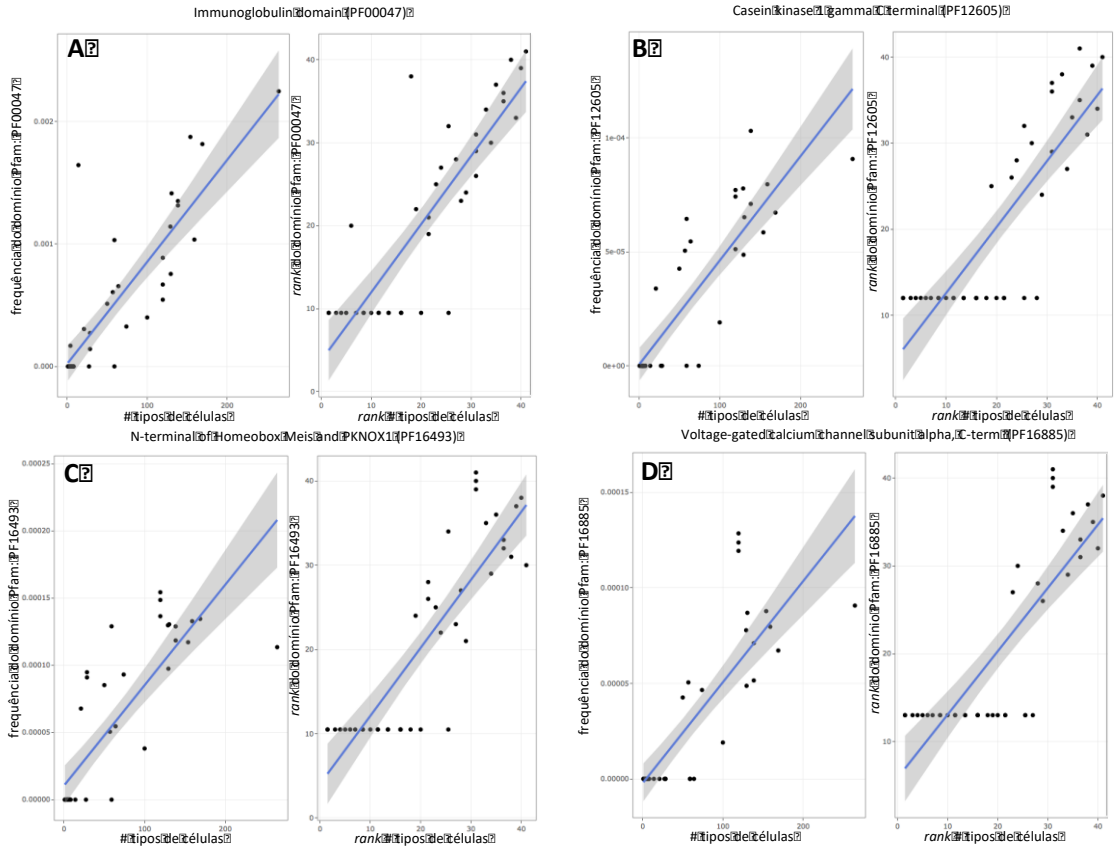

Figura 7 - Exemplos de domínios Pfam com correlação positiva significativa com o número de tipos de células e associados à processos de matrix extra-clular, cascatas de sinalização, fatores de transcrição e canais de membrana.

A) Matriz extracelular; B) via de sinalização; C) Fator de transcrição; D) canais de membrana.

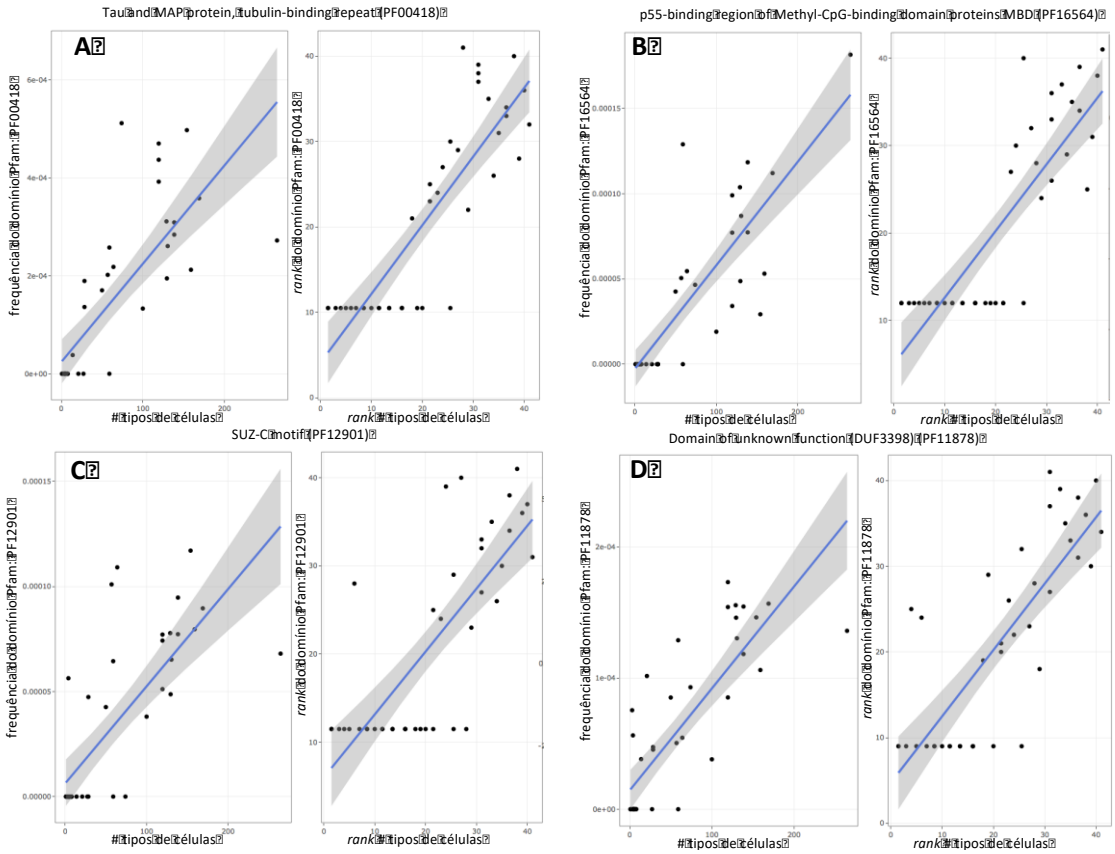

Figura 8 - Exemplos de domínios Pfam com correlação positiva significativa com o número de tipos de células e associados ao citoesqueleto, estrutura de cromatina, biologia de mRNA e de função desconhecida.

A) Citoesqueleto; B) Estrutura de cromatina; C) Biologia do RNA; D) Domínio de função desconhecida.

Dentre os 33 domínios com correlação negativa significativa com o número de tipos de células, detectamos 11 domínios relacionados à diversas vias metabólicas anabólicas e catabólicas (e.g. *Transketolase, thiamine diphosphate binding domain* - (PF00456); *CTP synthase N-terminus* - (PF06418); *Glutamate/Leucine/Phenylalanine/Valine dehydrogenase* (PF00208)), oito domínios de proteínas ribossomais (e.g. *Ribosomal protein L16p/L10* (PF00252); *Ribosomal protein S2* (PF00318); *Ribosomal protein S17* (PF00366) (Figura 9 A)), três domínios associadas à biologia de proteínas (e.g *Ubiquitin carboxyl-terminal hydrolase* (PF00443)), três domínios de sistemas de endomembranas (e.g. *BRO1-like domain (vacuole/lysosome targeting)* (PF03097)), dois dominios de biossíntese de tRNA (e.g. *tRNA pseudouridine synthase D (TruD)* (PF01142) (figura 9 B), dois domínios de componentes do citoesqueleto (e.g. *Spc97 / Spc98 family* (PF04130)), e dois domínios de metabolismo de ácidos nucléicos (e.g. *HRDC domain* (PF00570)). Além destes, detectamos também dois

domínios não classificados nas categorias anteriores (*Cell differentiation family, Rcd1-like* (PF04078) e *Component of IIS longevity pathway SMK-1* (PF04802)).

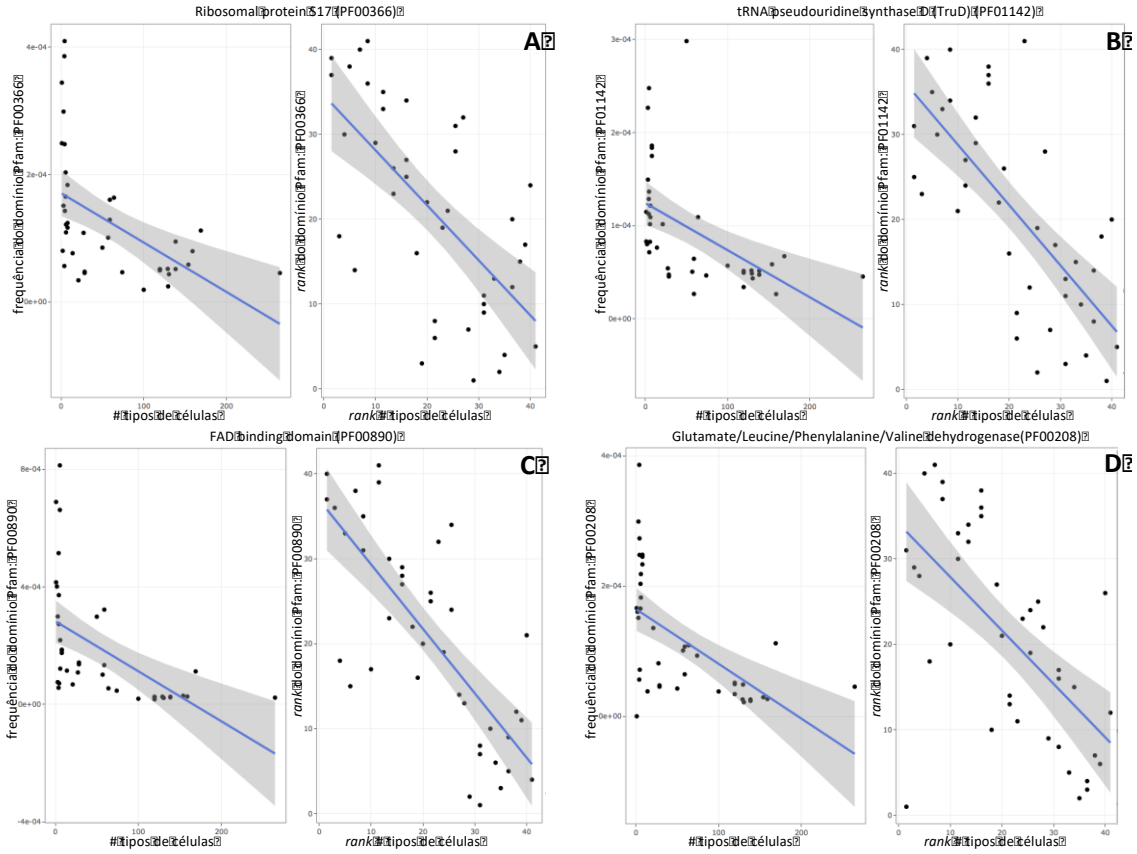

Figura 9 - Exemplos de domínios Pfam com correlação negativa significativa com o número de tipos de células e associados ao ribossomo, produção de tRNAs, e metabolismo biossintético e energético.

A) Ribossomo; B) síntese de tRNAs; C) e D) Metabolismo.

5.7.2 Termos GO associados ao número de tipos de células

Para o banco de dados de 41 genomas de eucariotos com alta completude genômica e baixo número de duplicações, encontramos um total de 407.357 proteínas distintas com pelo menos um termo GO associado pelo programa InterProScan, e um total de 1.075.171 termos GO associados. Dentre o total de 45.049 termos GO disponíveis, um total de 6.608 foi observado anotando ao menos uma proteína nos proteomas não-redundantes analisados. Após aplicarmos os critérios de filtro mencionados anteriormente, encontramos 304 termos GOs, os quais foram então definidos como estando significativamente associados ao número de tipos de células em nosso conjunto de dados. Estes termos GO compreendem aproximadamente 4,6% dos 6.608 GOs com ao

menos uma observação nos proteomas analisados. Os resultados destes GOs encontram-se sumarizados na Tabela Suplementar 2.

Dentre os 304 termos GO significaticamente associados ao aumento da complexidade, 266 apresentam correlação positiva, indicando que sua frequência aumenta à medida em que aumenta o número de tipos celulares nas diferentes espécies, enquanto 38 apresentam correlação negativa, indicando que estas funções biológicas diminuem sua frequência nos genomas dos organismos com o aumento de sua complexidade (Tabela Suplementar 2). De maneira geral, os processos detectados com correlação positiva significativa representam o estabelecimento e a regulação de processos e estruturas de reconhecida importância para o estabelecimento de diferentes tipos de células e para o estabelecimento de diferentes órgãos e estruturas dos metazoários e vertebrados, em particular. Os processos com correlação negativa detectados compreendem principalmente processos básicos do dogma central da biologica molecular, bem como diversas vias bioquímicas catabólicas e anabólicas.

Devido à própria estrutura do GO, diversos termos descrevem funções semelhantes, havendo grande sobreposição parcial entre os mesmos. Adicionalmente, possuímos funções biológicas que aumentam ou diminuem sua frequência nos genomas em função do número de tipos de células. Assim, de modo a permitir a visualização mais geral dos processos biológios representados pelos termos GO associados ao aumento da complexidade em *Eukarya*, utilizamos o programa REVIGO para sumarizar nossos resultados. Os resultados para essa sumarização encontram-se nas figuras 10 e 11 (processo biológico, correlação positiva e negativa, respectivamente), 12 e 13 (componente celular, correlação positiva e negativa, respectivamente) e 14 e 15 (função molecular, correlação positiva e negativa, respectivamente).

Nas figuras acima mencionadas, o tamanho das caixas é proporcional aos valores-p corrigidos para a correlação de Spearman, de modo que as maiores caixas indicam os processos com maiores valores de correlação com o número de tipos de células. Em cada figura, as caixas com a mesma cor representam os processos considerados relacionados pelo programa REVIGO. O resultado de REVIGO foi utilizado como guia na discussão. Quando necessário selecionamos outros processos interessantes, porém mais específicos e, consequentemente, não representados nas Figuras com resultados de REVIGO, a partir da Tabela Suplementar 2.

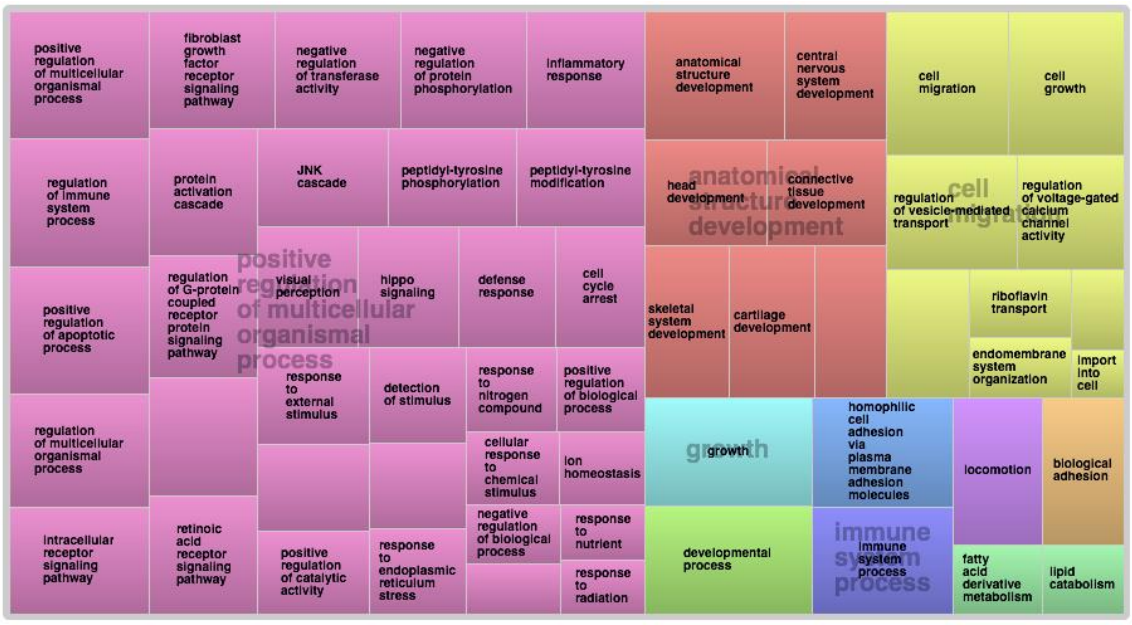

Figura 10 - Categorias GO (processo biológico) com correlação positiva e associadas ao número de tipos de células em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

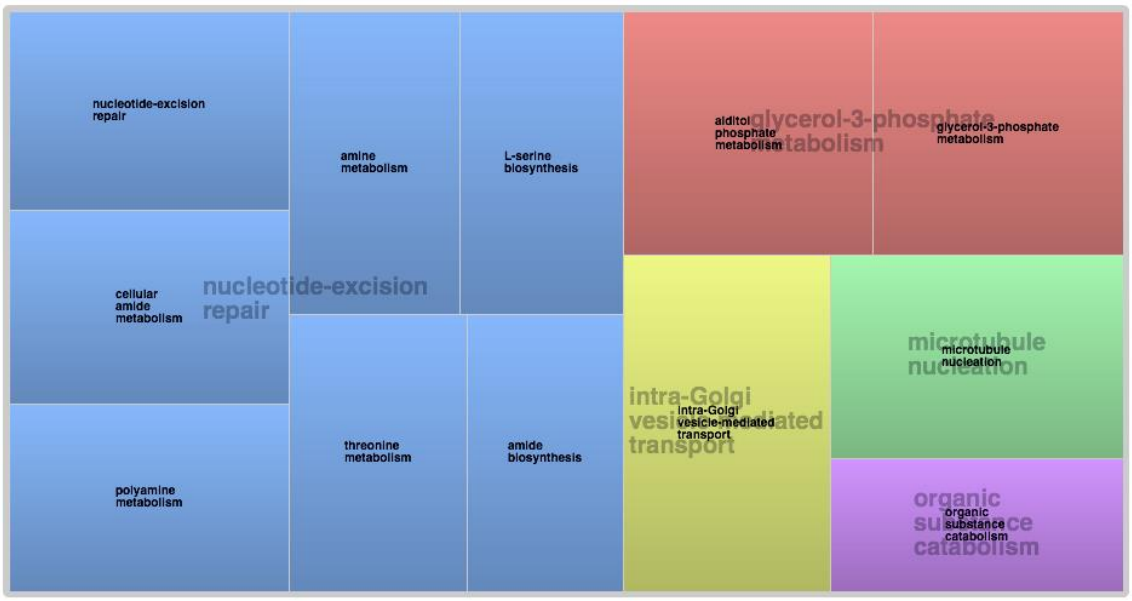

Figura 11 - Categorias GO (processo biológico) com correlação positiva e associadas ao aumento da complexidade em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

As categorias GO de processos biológicos que apresentam maior correlação positiva com o aumento da complexidade são compostas majoritariamente de fenômenos gerais que retratam mecanismos gerais importantes para o desenvolvimento de células especializadas, tecidos, órgãos e sistemas (Figura 10, caixa rosa). Dentre estes termos GO, em uma análise hierárquica, destacam-se inicialmente alguns dos processos biológicos mais amplos relacionados à emergência de organismos com múltiplos sistemas (GO:2000026 - *regulation of multicellular organismal development*, GO:0048731, *system development*, GO:0050793 - *regulation of developmental process*, GO:0022610 - *biological adhesion*, GO:0043085 - *positive regulation of catalytic activity*, GO:0009653 - *anatomical structure morphogenesis*, GO:0051239 - *regulation of multicellular organismal process*). Outros aspectos gerais para o desenvolvimento de diferentes tipos de tecidos (GO:0009888, *tissue development*) e órgãos (GO:0048513 - *animal organ development*) podem ser observados análise da Tabela Suplementar 2.

REVIGO também reportou diversos GO que representam vias de sinalização importantes para a proliferação e estabelecimento de identidade celular: 1) sinalização pela quinase *Hippo* (GO:0035329 - *Hippo signaling*); 2) cascata de sinalização mediada pela quinase JNK (GO:0007254 - *JNK cascade*); 3) via de sinalização mediada por fator de crescimento de fibroblasto (GO:0008543 - *fibroblast growth factor receptor signaling pathway*); 4) vias de sinalização mediadas por receptores acoplados à proteína G (GO:0008277 - *regulation of G protein-coupled receptor signaling pathway*) e 5) via de sinalização mediada por receptor retinóico (GO:0048384 - *retinoic acid receptor signaling pathway*).

Em análise da Suplementar 2, detectamos as seguintes vias de sinalização não reportadas por REVIGO: 1) diversas vias de sinalização por quinases (quinase A (GO:0051018 - *protein kinase A binding*), quinase C (GO:0004697 - *protein kinase C activity*), serina/treonina quinase ciclina-dependente (GO:0004861 - *cyclin-dependent protein serine/threonine kinase inhibitor activity*) e MAP quinase (GO:0043408 - *regulation of MAPK cascade*)); 2) via de sinalização por Ras (GO:0007265 - *Ras protein signal transduction*), 3) via de sinalização mediada por integrinas (GO:0007229 - *integrin-mediated signaling pathway*), 4) via de sinalização mediada por fatores de crescimento *insuline-like* (GO:0005520 - *insulin-like growth factor binding*), 5) sinalização mediada por TOR (GO:0032006 - *regulation of TOR signaling*) e 6) sinalização mediada por fator de crescimento (GO:0019838 - *growth factor binding*).

Outro grupo de termos GO detectados pelo programa REVIGO compreendem o desenvolvimento de estruturas anatômicas, sistemas, órgãos e tecidos que são observados espcificamente em metazoários, de maneira geral, e em vertebrados em particular, os quais constituem os organismos com maior número de células (Figura 10, caixa vermelha, Tabela Suplementar 2). Dentre os sistemas detectados, destacam-se os sistemas nervoso (GO:0007399 - *nervous system development*, GO:0007417 - *central nervous system development*, GO:0060322 - *head development*), circulatório (GO:1903522 - *regulation of blood circulation*), esquelético (GO:0001501 - *skeletal system development*), epitelial (GO:0008544 - *epidermis development*), bem como de tecidos cartilaginosos (GO:0051216 - *cartilage development*), e conectivos (GO:0061448 - *connective tissue development*).

Um outro grupo de GOs sugerido por REVIGO contém majoritariamente processos que ocorrem em nível celular (Figura 10, caixa amarela, Tabela Suplementar 2). Aqui, observa-se diversos GOs associados aos diferentes processos celulares necessários para o funcionamento de organismos com múltiplos tipos de células diferentes, tais como diferenciação (GO:0048869, *cellular developmental process*, GO:0030154 - *cell differentiation*, GO:0008360 - *regulation of cell shape*), migração (GO:0016477, *cell migration*, GO:0006935 - *chemotaxis*, GO:0042330 - *taxis*), adesão (GO:0007155 - *cell adhesion*), crescimento (GO:0016049 - *cell growth*) e morte celular programada (GO:0010942, *positive regulation of cell death*). Além disso, obiservou-se também e diversos sistemas moleculares específicos, como os sistemas internos de membranas (GO:0010256 - *endomembrane system organization*) e transporte por vesículas (GO:0060627 - *regulation of vesicle-mediated transport*, GO:0098657 - *import into cell*). Abaixo do nível celular, análises da Tabela Suplementar 2 detectaram também padrões de controle gerais das redes de atividades moleculares (e.g. GO:0031401 - *positive regulation of protein modification process*, GO:0098772 - *molecular function regulator*, GO:0043085 - *positive regulation of catalytic activity*).

Outros processos detectados por REVIGO, embora com menos relevância, são o desenvolvimento de um sistema imune e de sua regulação (e.g. GO:0002682 - *regulation of immune system process*, GO:0006955 - *immune response*, GO:0050900 - *leukocyte migration*, GO:0006952 - *defense response*), bem como processos gerais para o estabelecimento de multicelularidade, como crescimento (GO:0040007, *growth*), adesão celular (GO:0007155 - *cell adhesion*) e processo de desenvolvimento (GO:0032502 - *developmental process*).

Na análise dos processos que apresentam correlação negativa significativa com o aumento da complexidade em *Eukayra*, detectamos um grande grupo (Figura 11, caixa azul) que representa majoritariamente mecanismos ancestrais em organismos celulares, tais como vias de reparo de DNA (GO:0006289 - *nucleotide-excision repair*), transcrição (GO:0090575 - *RNA polymerase II transcription factor complex*), tradução (GO:0009308 - *amine metabolic process*), sintese de aminoácidos (GO:0006566 - *threonine metabolic process*, GO:0006564 - *L-serine biosynthetic process*). Além destes, detectamos também diversas vias de biossíntese de aminoácidos e a via de reparo de DNA por excisão de nucleotídeos (GO:0006289 - *nucleotide-excision repair*). Outro grupo detectado (Figura 11, caixa vermelha) compreende processos do metabolismo energético de carboidratos simples (GO:0006072 - *glycerol-3-phosphate metabolic process*, GO:0052646 - *alditol phosphate metabolic process*)). Detectamos ainda dois grupos de processos intracelulares: o transporte mediado por Golgi (GO:0006891 - *intra-Golgi vesicle-mediated transport*) e o sistema de polimerização de microtúbulos (GO:0007020 - *microtubule nucleation*). Chama também a atenção a existência de um termo geral do GO, o qual descreve processos catabólicos de substâncias orgânicas, indicando a existência de diversos componentes desse processo biológico em organismos com poucos tipos de células (GO:1901575 - *organic substance catabolic process*).

A análise das categorias GO de componentes celulares associadas positivamente ao aumento da complexidade em *Eukarya* detectou que produtos protéicos associados com a a membrana celular e suas partes(Figura 12, caixa verde). Dentro do ambiente da superfície celular, detectamos a associação de diversas regiões funcionais, tais como a periferia celular (GO:0071944 - *cell periphery*) e ambas as porções interna e externa da membrana plasmática (GO:0098562 - *cytoplasmic side of membrane*, GO:0019897 - *extrinsic component of plasma membrane*).

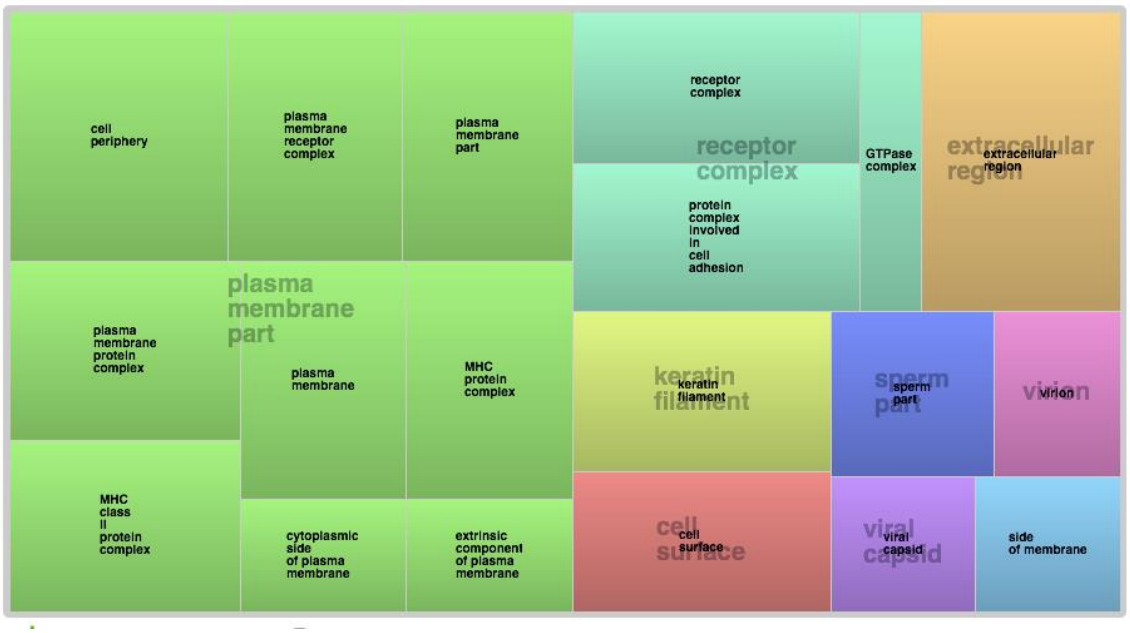

Figura 12 - Categorias GO (componente celular) com correlação positiva e associadas ao aumnto da complexidade em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

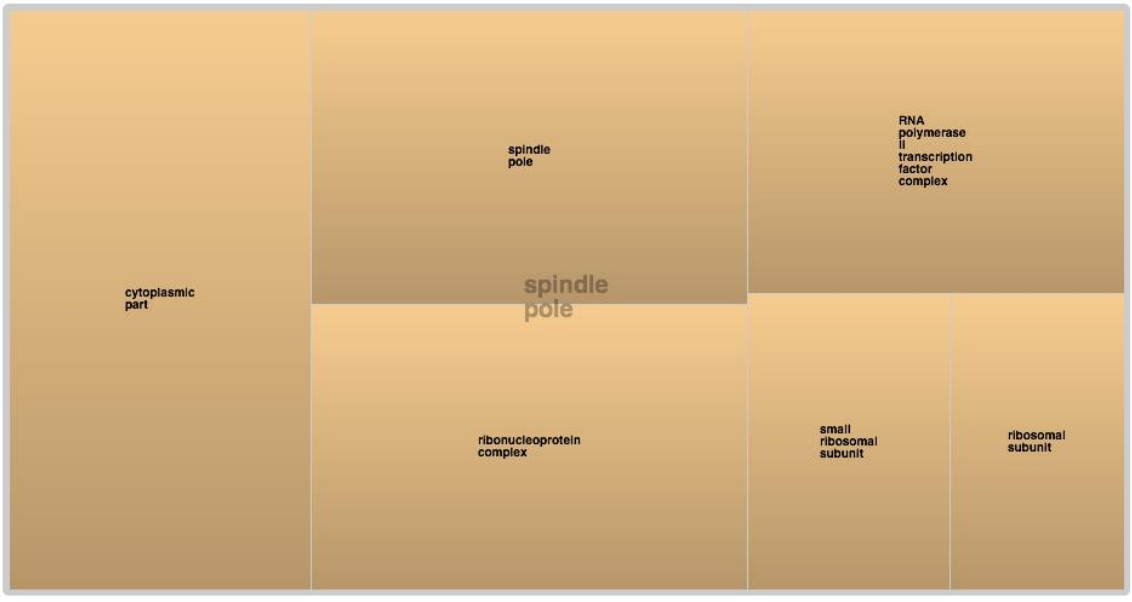

Figura 13 - Categorias GO (componente celular) com correlação negativa e associadas ao aumento da complexidade em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

Um outro conjunto de componentes detectado por REVIGO compreende termos associados ao conceito de receptor celular (GO:0043235 - *receptor complex*, GO:0098802 - *plasma membrane receptor complex*) (Figura 12, caixa verde clara, Tabela Suplementar 1). Dentro deste conceito, destacaram-se os receptores acoplados à proteína G (GO:0005834 - *heterotrimeric G-protein complex*, GO:1905360 - *GTPase complex*). Obsrva-se também uma associação significativa entre o ambiente extracelular e o aumento da complexidade (GO:0005576, *extracellular region*). Detectamos também componentes específicos de metazoários, tais como a queratina (GO:0045095 - *keratin filament*). Curiosamente, detectamos também um aumento nos termos anotados como "esperma" (GO:0097223 - *sperm part*) bem como um aumento no termo vírus (GO:0019012 - *virion*).

Na análise de localizações celulares negativamente associadas ao aumento da complexidade (Figura 13), detectamos uma associação entre o corpo polar do fuso (GO:0000922 - *spindle pole*), bem como entre componentes dos processos de transcrição (GO:0090575 - *RNA polymerase II transcription factor complex*) e tradução (GO:0044391 - *ribosomal subunit*, GO:0005840 - *ribosome*). Também observa-se uma associação negativa entre a fração do proteoma predito não redundante com localização citoplasmática e o número de tipos de células (GO:0044444 - *cytoplasmic part*).

A análise das funções moleculares positivamente associadas à complexidade em *Eukarya* detectou uma série de termos associados à sinalização mediada por receptores (Figura 14, caixas vermelha, azul, azul clara e verde, Tabela Suplementar 2), bem como a sua regulação (Figura 14, caixa roxa). Dentre os receptores detectados, a vasta maioria compreende receptores de vias de sinalização, como as previamnte destacadas, detectamos também diversas vias de proliferação de sistemas de vertebrados, tais como hormônios intestinais (GO:0004999 - *vasoactive intestinal polypeptide receptor activity*), sinalização por interleucina 1 (GO:0004908 - *interleukin-1 receptor activity*), neuropeptídeos (GO:0008188 - *neuropeptide receptor activity*) e sistema adrenérgico (GO:0004935 - *adrenergic receptor activity*). Além destes, detectamos também processos que representam aspectos mais gerais das funções moleculares, como mecanismos regulatórios (GO:0098772 - *molecular function regulator*, GO:0043085 - *positive regulation of catalytic activity*).

Nas funções moleculares negativamente associadas ao aumento da complexidade, detectamos atividades de metabolismo biossintético e energético (GO:0016866 -

intramolecular transferase activity, GO:0009055 - *electron transfer activity*), bem como constituíntes dos ribossomos (GO:0003735 - *structural constituent of ribosome*) e dos microtúbulos (GO:0043015 - *gamma-tubulin binding*) (Figura 15).

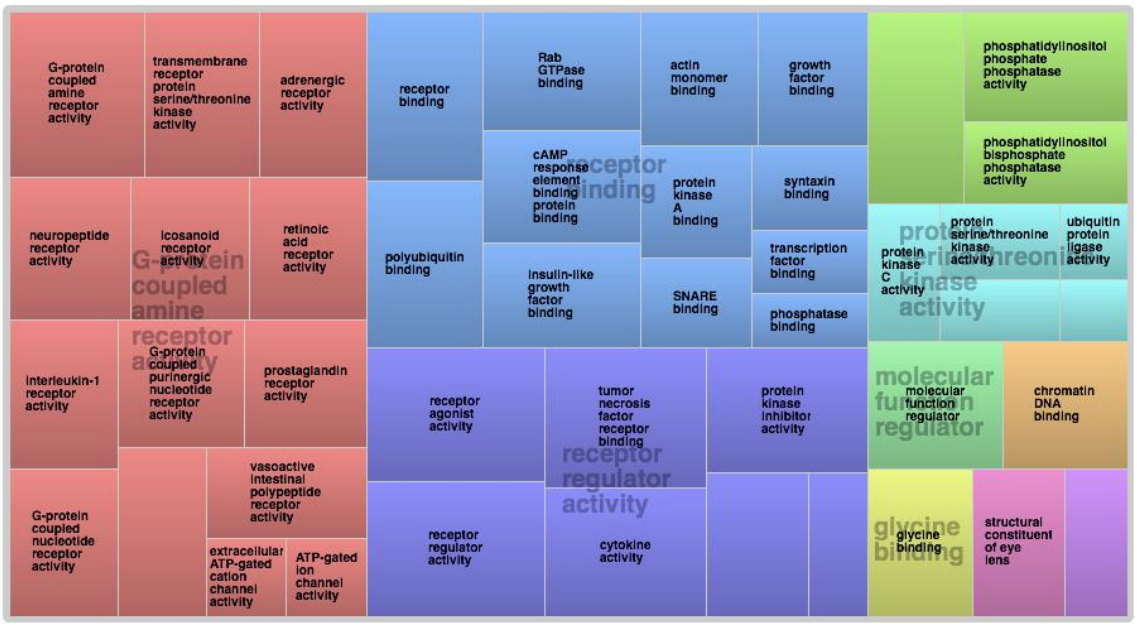

Figura 14 - Categorias GO (função molecular) com correlação positiva e associadas ao aumento de complexidade em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

Figura 15 - Categorias GO (função molecular) com correlação negativa e associadas ao aumnto da complexidade em Eukarya.

Os componentes são agrupados em cores em função de sua similaridade no espaço semântico, não representando, necessariamente, termos com redundância funcional. O tamanho das caixas é proporcional os valores p-corrigidos das correlações de Spearman.

Para permitir uma visualização de algumas das associações detectadas, construímos alguns gráficos ilustrativos das correlações de Spearman e Pearson para alguns dos termos GO detectados como associados significativamente ao número de tipos celulares em *Eukarya* (Figuras 16 e 17).

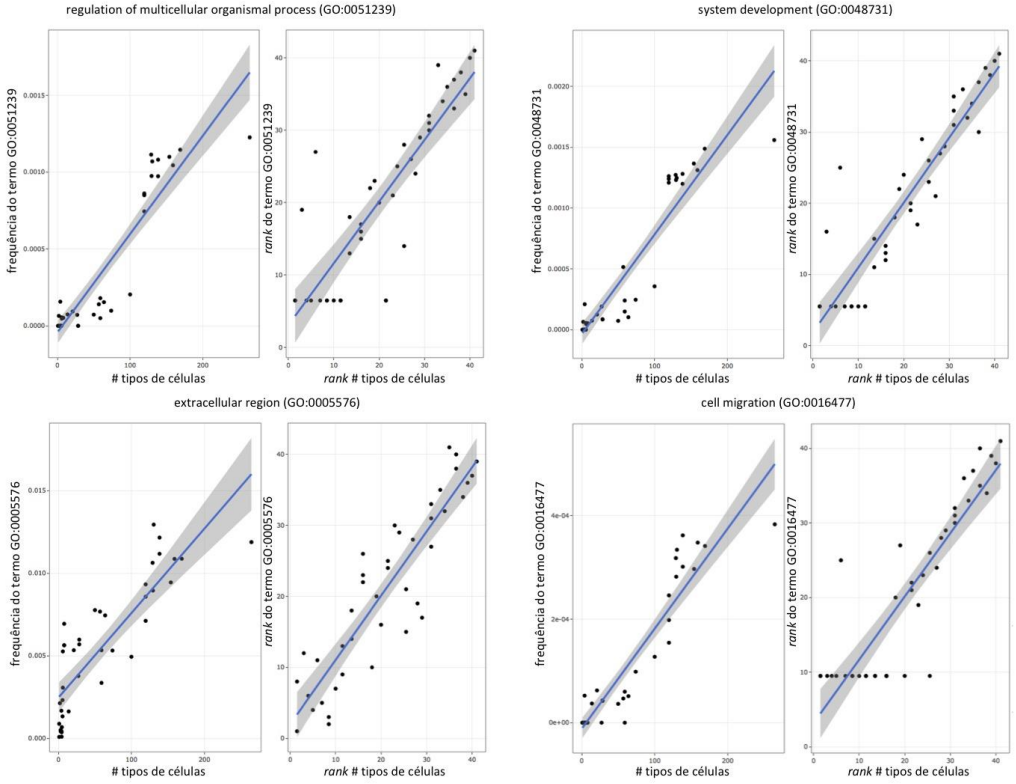

Figura 16 - Exemplos de termos GO com correlação positiva significativa com o número de tipos de células.

Os gráficos de correlação de Pearson contém o número de tipos de células e a frequência relativa do termo GO como eixos X e Y, respecticamente. Os gráficos de correlação de Spearman contém o *rank* do número de tipos de células e da frequência relativa do termo GO como eixos X e Y, respectivamente. As barras azuis indicam os modelos lineares para os valores, e a área cinza, o intervalo de confiança.

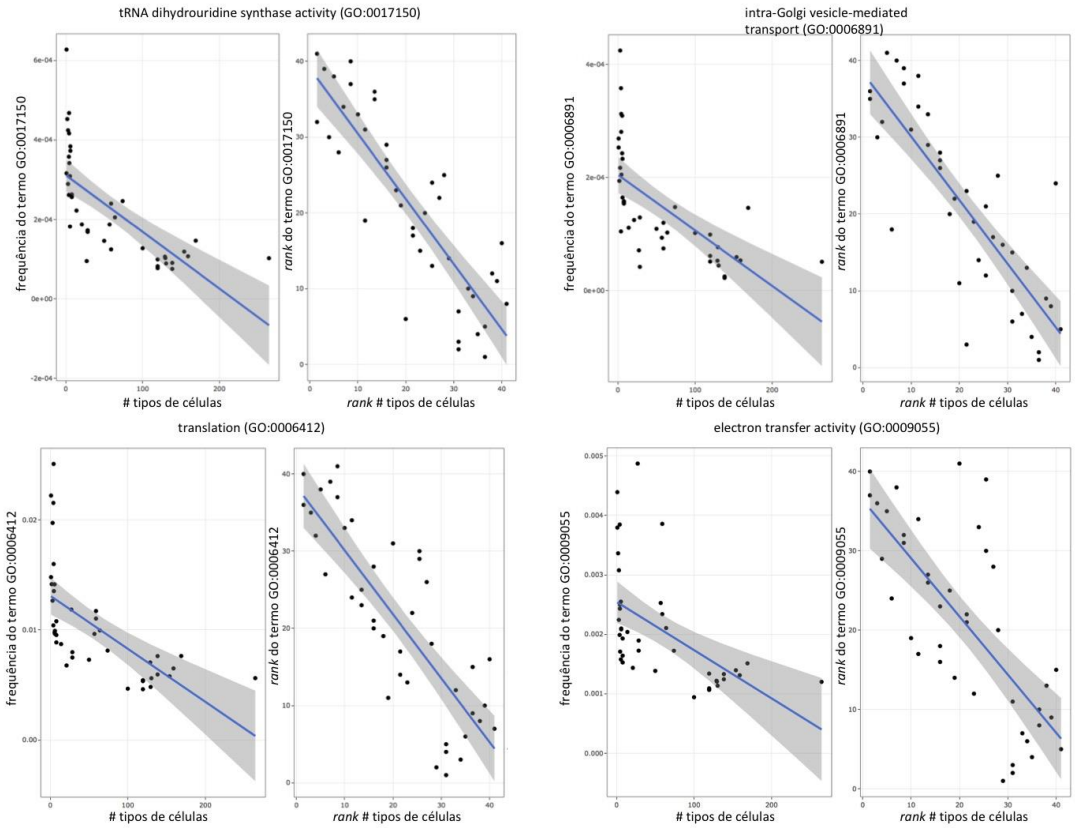

Figura 17 - Exemplos de termos GO com correlação negativa significativa com o número de tipos de células.

Os gráficos de correlação de Pearson contém o número de tipos de células e a frequência relativa do termo GO como eixos X e Y, respecticamente. Os gráficos de correlação de Spearman contém o *rank* do número de tipos de células e da frequência relativa do termo GO como eixos X e Y, respectivamente. As barras azuis indicam os modelos lineares para os valores, e a área cinza, o intervalo de confiança.

6. Discussão

Durante a evolução, a transição para multicelularidade a partir do surgimento independente de novas características e funções biológicas nos ancestrais unicelulares eucarióticos permitiu o surgimento de novos tipos de células especializadas, e compreende uma das maiores e importantes transições evolutivas na árvore da vida (Baum e Baum, 2014; Grosberg e Strathmann, 2007). Simultaneamente, observou-se o surgimento de várias inovações evolutivas tais como adesão, comunicação e diferenciação celular, o que fornece os componentes moleculares e funcionais para a emergência de tecidos, órgãos e sistemas especializados (Tong, Wang e Wu, 2017).

As origens evolutivas da multicelularidade nos eucariotos foram transições evolutivas relativamente antigas (Hedges *et al.*, 2004) e independentes (Grosberg e Strathmann, 2007; Niklas, 2014) na árvore da vida. Um organismo eucarioto pode se tornar multicelular através de dois mecanismos principais: (i) na multicelularidade clonal, todas as células do organismo surgem a partir de sucessivas rodadas de divisão celular de uma célula fundadora, deste modo forma-se um aglomerado de células geneticamente idênticas. Por outro lado (ii) a multicelularidade agregativa, células geneticamente distintas formam uma entidade multicelular se aderindo umas às outras (Sebé-Pedrós, Degnan e Ruiz-Trillo, 2017). Os primeiros organismos multicelulares possivelmente formaram agregados coloniais, com a posterior evolução de sinais de especialização e separação em células somáticas e germinativas, o que ocasionou a produção de tecidos diferenciados através da invaginação destas células germinativas (Niklas e Newman, 2013; Herron e Neldecu, 2015).

De uma forma similar, o desenvolvimento embrionário de muitos organismos multicelulares clonais mais complexos, como os metazoários, ocorre a divisão e o crescimento de um zigoto totipotente a partir de uma única célula e, mais tarde, por meio da diferenciação celular, permite que alterações anatômicas e fisiológicas pós-embrionárias resultem um organismo adulto constituído por tecidos específicos e diversos tipos celulares distintos. O desenvolvimento embrionário de várias linhagens de eucariotos multicelulares mais complexos, como as plantas terrestres e os metazoários, está associado ao desenvolvimento de planos corporais maiores e mais complexos (Sebé-Pedrós, Degnan e Ruiz-Trillo, 2017).

Programas de desenvolvimento regulados espacialmente e temporalmente estão entre os principais requisitos para o surgimento de diferentes tipos de células

espécializadas nos eucariotos mais complexos, os quais envolvem a combinação de um complexo e dinâmico conjunto de funções, como processos celulares que coordenem todos os eventos de migração celular, multiplicação e apoptose, dentre outros, necessários para o estabelecimento da forma, tamanho e composição relativa dos diferentes órgãos, tecidos e sistemas dos organismos eucariotos complexos (Sebé-Pedrós *et al*., 2016). Adicionalmente, diversos experimentos clássicos de embriologia detectaram que os órgãos transplantados entre espécies distintas durante a sua embriogênese possuem informação intrínseca sobre o seu tamanho final na espécie de origem, indicando que programas de desenvolvimento espécie-específicos ocorrem durante a embriogênese para controlar o tamanho dos órgãos (Kragl *et al*., 2009).

Em organismos com muitos tipos celulares, suas células individuais apresentam movimentos de migração, adesão, comunicação, diferenciação morfológica, genômica, epigenômica e transcricional, dentre outros, de modo a dividir tarefas e executar diversas funções biológicas especializadas em organismos (Arendt *et al*., 2016), o que permite que eucariotos multicelulares exibam uma uma notavel variedade de aspectos ecológicos, morfológicos, comportamentais, histórias de vida, aparências e tamanhos nunca anteriomente presenciados durante o curso da evolução.

Dada a relativa facilidade de obtenção e a universalidade de sua aplicação, a estimativa de tipos diferentes de células de um organismo é uma medida frequentemente utilizada para estimar e estudar a variação da complexidade biológica nos eucariotos e sua associação com padrões fenotípicos e genômicos (Bonner, 1988; Valentine *et al.*, 1994; Bell e Mooers 1997; Hedges *et al.*, 2004; Haygood e Investigators, 2006; Vogel e Chotia, 2006; Lang *et al*., 2010; Schad *et al.*, 2011, Chen *et al.*, 2014, Niklas, Cobb e Dunker, 2014; Cardoso e Sharpe, 2017, Sebé-Pedrós *et al*., 2018).

Do ponto de vista estritamente das sequências genômicas e de seus diferentes elementos funcionais, o aumento do número de tipos de células em *Eukarya* não está diretamente associado com parâmetros genômicos que inicialmente foram cogitados como potenciais agentes causais do aumento da complexidade, tais como o tamanho do genoma ou seu número de genes codificadores (paradoxo dos valores C e G, respectivamente) (Gregory, 2005). Entretanto, alguns estudos já obtiveram sucesso na busca por padrões específicos de expansão de domínios e famílias protéicas codificados em genomas e o número de tipos celulares, detectando associação significa positiva entre o número de tipos de células e expansões de domínios, famílias e superfamílias de

domínios de proteínas em *Eukarya* e *Verbetrata* (Vogel e Chotia, 2006, Kawashima *et al*., 2009).

Em nosso estudo, dada a sua ampla adoção e as demais questões expostas anteriormente, utilizamos o número de tipos celulares como *proxy* para a busca por elementos genômicos que repesentassem unidades evolutivas mínimas e as diferentes funcões biológicas representadas nos genes codificadores de proteínas (descritos por domínios Pfam de proteínas e por termos GO, respectivamnte) significativamente associados ao aumento da complexidade.

Para tal, procuramos melhorar e expandir os conceitos e metodologias utilizados pelos autores dos estudos anteriores, especialmente o estudo proposto por Vogel e colaboradores, em 2006 (Vogel e Chotia, 2006). Especificamente, visamos 1) realizar o controle de qualidade dos genomas utilizados; 2) observar o rigor estatístico, levando em consideração o cenário de ocorrência de possíveis *outliers*, da não-independência de dados relativos à espécies e da ocorrência de cenário de múltiplas hipóteses; 3) realizar a busca por possíveis convergências funcionais moleculares, nas quais genes não-homólogos executam o mesmo papel celular (o que é representado apropriadamente por termos GO, mas não por domínios de proteínas ou superfamílias).

Para iniciar nossa análise, realizamos um levantamento de estudos que compilam dados sobre o número de tipos de células em diferentes organismos, de modo a detectar para quais destes há genomas completos disponíveis. Os primeiros trabalhos que utilizaram bancos de dados de tipos celulares diferentes como medida de complexidade biológica datam as décadas de 80 e 90 (por exemplo Bonner, 1988; Valentine *et al.*, 1994 e Bell e Mooers 1997). Dentre estes três trabalhos, Bell e Mooers se destaca por possuir o maior banco de dados curado de tipos celulares, representando 134 espécies distintas, abrangendo os principais grupos de eucariotos conhecidos à época, tais como protistas, plantas e algas. Entretanto, somente dez genomas completos anotados estão disponíveis para esses organismos. Outros trabalhos, como, por exemplo Valentine *et al*. 1994,ß analisaram um total de 13 organismos metazoários. Trabalhos mais atuais, como Cardoso e Sharpe, 2017, utilizaram somente, nove espécies, e todas elas organismos modelos.

O banco de dados de Vogel e Chotia, 2006, onde os autores buscaram por expansões de superfamílias de proteínas associadas ao número de tipos de células, utilizou um total de 38 espécies de eucariotos para os quais há genomas completos disponíveis e número de tipos de células, apresentando um banco de dados considerável de genomas completos. Mais recentemente, o trabalho de Chen *et al.*, 2014, buscou

correlação entre o número de isoformas e o número de tipos celulares, e compilou, a partir de vários trabalhos anteriores, incluíndo Vogel e Chotia, 2006, o número médio de tipos celulares diferentes para 54 eucariotos com ampla distribuição taxonômica à partir dos valores mínimos e máximos de tipos celulares diferentes encontrados em cada um destes trabalhos anteriores (Valentine *et al*. 1994; Bell e Mooers 1997; Hedges *et al*. 2004; Haygood e Investigators, 2006; Vogel e Chotia, 2006; Vickaryous e Hall, 2006; Lang *et al*. 2010; Schad *et al*. 2011). Assim, iniciamos nossos estudos com a informação disponibilizada por Chen *et al*., 2014.

À partir desse banco de dados, obtivemos o genoma completo para os 54 organismos a partir do NCBI, o que garante uma uniformidade das espécies analisadas em termos de sua anotação e qualidade mínima para iniciarmos nossa análise. Entretanto uma análise preliminar detectou que os critérios de qualidade genômica que utilizamos removeram diversos protozoários, o que nos levou a buscar outros genomas de organismos desses táxons e estimarmos seu número de tipos de células para garantirmos a representatividade da maior fração possível de organismos com poucos tipos de células das mais diferentes linhagens filogenéticas. Desta forma, estimamos o número de tipos celulares diferentes para seis protistas oriundos de dados da literatura a partir da diferenciação celular durante seus ciclos de vida (Black e Boothroyd, 2001; Chen *et al*., 2014; Sebé-Pedrós *et al*., 2013; Lone e Manohar, 2018), de acordo com o proposto por Niklas, 2014, totalizando 60 organismos ao nosso banco de dados inicial de organismos que possuem genoma completo disponível no NCBI e número de tipos de células disponível.

Conforme já descrito anteriormente, em nosso banco de dados também detectamos que fungos, protozoários e algas são os organismos com a menor complexidade, possuíndo sete ou menos tipos diferentes de células; plantas vasculares têm uma amplitude de cerca de 30 a 50 tipos diferentes de células; metazoários não-vertebrados possuem entre 14 e 70 tipos de células os vertebrados são de maior complexidade com cerca de 100 a 260 tipos diferentes de células (figura 3). Percebe-se que o ser humano poderia ser classificado como de distribuição atípica em relação aos outros vertebrados (*outlier*), uma vez que este possui 56% de tipos de células do que o Chimpanzé e os demais o que pode ser um resultado real ou simplesmente um viés causado pelo excesso de estudos de seres humanos quando comparado às demais espécies.

O processamento alternativo de mRNA é um processo pós-transcricional nos eucariotos através do qual múltiplos transcritos, que eventualmente codificam várias

proteínas com diferentes funções, são produzidos a partir de uma única região codificadora. Desta forma, eventos de processamento geram muitas isoformas por *locus*, o que pode inflar a frequência de domínios de proteínas em um determinado organismo. Um procedimento comum em análises que visam comparar o proteoma predito de diferentes espécies é a remoção das diversas isoformas descritas para um mesmo *locus*, representando cada região genômica uma única vez (Vogel e Chotia, 2006). Esse procedimento visa eliminar o viés que existe em organismos-modelo, os quais são mais bem estudados e, consequentemente, possuem mais isoformas descritas por *locus* do que organismos com menos estudos transcriptômicos.

Para obtermos uma representação fidedigna do conteúdo gênico não-redundante de cada espécie, optamos por estabelecr uma rotina computacional na qual, para cada locus codificador de proteínas descrito em arquivos *gbff*, recuperarmos a maior ORF possível. Esse procedimento visa recuperar, dentre as diversas isoformas potencialmende descritas para um genoma, aquela que possui o maior conteúdo informacional potencial para cada *locus*.

Acreditamos que nossa metodologia tenha sido satisfatória para remover o excesso de isoformas observado para os organismos eucarióticos que são modelos em suas áreas. Os valores de contagens dos diferentes transcriptomas codificadores redundantes apresentavam-se amplamente variáveis, mesmo entre organismos de linhagens filogenéticamente próximas (e.g. *C. elegans* e *C. briggsae*). Adicionalmente, os mesmos apresentavam altas contagens particularmente em organismos-modelo, o que sugere que os valores são decorrentes de um excesso relativo de estudo de organismos-modelo do que de variação biológica real.

Nossa metodologia de redução de regiões codificadoras redundates (em função da presença de isoformas) obteve proteomas não-redundantes que apresentavam valores mais uniformes entre si para organismos e com número de genes compatível com o número de genes em organismos como *H. sapiens* (aproximadamente 21000 genes codificadores de proteínas (Willyard, 2006), enquanto observamos 20026).

Uma questão que não foi tratada adequadamente nos estudos anteriores compreende a avaliação da qualidade do genoma, a qual, conforme demonstramos, variou consideravelmente em nossos dados. Genes ortólogos, ou seja, genes que descendem, via especiação, de um único gene existente último ancestral comum entre os genomas comparados, tendem a manter similaridade das sequências e funções ao longo do tempo evolutivo (Koonin, 2005). Assim, ortólogos de cópia única (denominados ortólogos 1-1)

usualmente executam funções essenciais em um determinado táxon e são ditos universais ou quase universais, uma vez que são observados na vasta maioria dos organismos daquela linhagem evolutiva. Consequentemente, ortólogos 1-1 podem ser utilizados como um *proxy* de qualidade de montagem genômica, ao permitir avaliar a completude do genoma de uma espécie baseado no conteúdo gênico esperado (Simão e Waterhouse *et al*., 2015). Genomas com baixa completude podem ser um indicativo de montagem ou predição gênica aquém do desejável, o que pode potencialmente enviesar estudos onde se deseje associar a frequência de elementos genômicos com alguma característica fenotípica.

O software BUSCO utiliza genes ortólogos clado-específicos de cópia única e quase universais para avaliar a qualidade de proteomas em função do conteúdo gênico esperado para o táxon. Em nossa análise, avaliamos a completude dos 57 genomas de *Eukarya* selecionados utilizando o banco de dados de ortólogos cópia única altamente conservados e quase universais do BUSCO para eucariotos, construído originalmente a partir de 303 genes encontrados em mais de 90% de 65 espécies de eucariotos (Simão e Waterhouse *et al*., 2015).

Assumimos nossos pontos de corte para avaliar a completude dos genomas de uma forma arbitrária (≥ 70% de genes completos - cópias simples + duplicações, ≥ 70% de cópias simples, ≤ 10% de genes duplicados, ≤ 10% de genes fragmentados e ≤ 10% de genes ausentes), uma vez que não há um critério de ponto de corte formalmente descrito para este tipo de metodologia e análise. Cabe ressaltar que há possíveis viéses na criação dos bancos dos bancos de dados de ortólogos 1-1 quase universais utilizados por BUSCO, os quais são elaborados utilizando majoritariamente organismos-modelo, nos quais existe uma tendência de pontuações maiores que 90% para organismos (e.g. humano e o camundongo), enquanto genomas de organismos não-modelo podem apresentar valores consideravelmente menores, variando entre 50% e 95% (Simão e Waterhouse *et al*., 2015).

Assim, optamos por utilizar uma abordagem menos conservadora, uma vez que os protozoários apresentaram os menores valores de BUSCO, o que poderia indicar tanto a ausência real destes genes nestes organismos, uma vez que o próprio banco de dados BUSCO é enviesado, quanto eventuais problemas na montagem dos genomas. Se utilizássemos um ponto de corte de ≥ 90% de genes completos e de cópias simples, por exemplo, isso nos permitiria abranger somente 50% das espécies do nosso banco de dados (30 de 60 espécies), na sua grande maioria genomas de metazoários (tabela 2). Após

avaliarmos a completude do genoma descartamos cerca de um terço das espécies dos nossos bancos de dados (42 espécies de 60 espécies no início deste estudo). A partir desse ponto, consideramos que possuímos proteomas de alta qualidade. Deste total de 42 genomas de alta qualidade, 40% (17 genomas) são exclusivos do nosso estudo enquanto os outros 60% (25 genomas) também estão no trabalho de Vogel e Chotia, 2006.

O serviço *web* The Timetree of Life (Hedges *et al*., 2006) fornece a possibilidade de reconstruir árvores filogenéticas utlramétricas (onde o tamanho dos ramos é proporcional ao tempo de divergência) para 50632 espécies (Hedges *et al*., 2015). Dentre as 42 espécies com valores de BUSCO acima do ponto de corte, uma (*Yarrowia lipolytica*, Fungi) não foi encontrada nesse serviço, sendo excluída das análises posteriores. Desta forma, chegamos ao total de 41 organismos utilizados para a identificação das funções biológicas asssociadas ao aumento da complexidade em *Eukarya*.

Após consideramos haver abordado de maneira adequada a questão da avaliação da qualidade dos genomas, assunto não tratado nos estudos anteriores de genômica comparativa (e.g. Vogel *et al*., 2016), procedemos com a detecção de termos GO e domínios Pfam associados ao número de tipos de células dos organismos utilizados m nosso estudo. Nesse ponto, utilizamos três procedimentos estatísticos que visam corrigir viéses que sabidamente influenciam estudos dessa natureza: 1) o teste múltiplo de hipóteses; 2) a análise de dados com possíveis extremos de observação (*outliers*); 3) a análise de dados não-independentes, como é o caso de dados fenotípicos/genotípicos de espécis relacionadas por ancestralidade comum.

Para controlar a taxa de falso-positivo inflada que ocorre em cenários de teste múltiplo de hipóteses, utilizamos a correção para testes múltiplos, que visa controlar os valores-p, de modo a permitir uma taxa de falso-positivos máxima ao se testar diversas hipóteses no mesmo conjunto de dados (Yekutieli e Benjamini, 2001). Para lidar com possíveis *outliers*, como pode ser o caso do número de tipos de células da espécie *H. sapiens*, o qual pode ser artificialmente elevado dado o maior número de estudos envolvendo essa espécie, bem como para detectarmos possíveis associações não-lineares entre nossos dados, utilizamos métodos não-paramétricos como a correlação de Spearman, a qual mitiga problemas dessa natureza (Zar, 1999). Cabe ressaltar que o estudo original de Vogel *et al*., 2006, faz uso de correlação de Pearson, que é indicado para a busca por relações lineares entre variáveis indepndentes. Em nossas análises, representamos também a correlação de Pearson para evidenciar como essa análise poderia ser facilmente enviesada pela ocorrência de *outliers* como *H. sapiens*.

Para tratarmos da não-independência dos dados de espécies relacionadas por ancestralidade comum, um aspecto também não abordado em Vogel *et al*., 2006, utilizamos o método dos contrastes filogeneticamente independentes proposto por Felsenstein 1985, o qual já foi amplamente utilizado em estudos que fazem uso de dados fenotípicos, embora ainda careça de adoção generalizada em estudos de genômica comparativa (Felsenstein, 1985). Dessa maneira, as associações observadas nas análises aqui expostas não podem ser atribuídas à possíveis viéses nos dados causado pela ancestralidade comum das espécies, sendo necessário o estabelecimento de hipóteses alternativas para explicar os fatos.

Finalmente, outro fator que consideramos importante, já abordado em estudos anteriores e também adotado nesse estudo, compreende o uso de frequências relativas de elementos genômicos (valores de contagem de cada elemento dividido pelo total de elementos) ao invés de valores absolutos de contagem dos mesmos. Essa normalização visa corrigir possíveis viéses causados por, por exemplo, proteomas não-redundantes de tamanhos distintos (Vogel e Chotia, 2006). Assim, acreditamos haver tratado os dados da maneira adequada do ponto de vista estatístico, ao detectar possíveis fontes de viéses nos mesmos e realizar os procedimentos adequados para mitigá-los.

Nossa estratégia de detecção de elementos genômicos associados ao aumento da complexidade utilizou duas classes de elementos complementares. O primeiro elemento analisado foram os domínios protéicos, os quais são tanto relacionados por eventos de ancestralidade comum como funcionam como unidades funcionais distintas que eventualmente foram perdidas, expandidas e embaralhadas ao longo da evolução de eucariotos. Cabe ressaltar que os estudos anteriores analisaram somente a associação de sequências relacionadas por ancestralidade comum, como domínios, famílias e superfamílias protéicas. Assim, esses estudos não avaliam, portanto, possíveis eventos de convergência molecular, onde proteínas de origem evolutiva distinta realizam a mesma função. O segundo elemento analisado foi a frequência de categorias GO a nível de gene, o que permite utilizar a estrutura do GO e sua extensa curadoria para tentar detectar padrões que vão além das relaçõs de homologia entre as sequências biológicas.

Nossos resultados corroboram vários trabalhos que avaliaram as funções biológicas que estão fortemente associadas com a complexidade biológica nos eucariotos, bem como detecta diversas outras tendências não observadas anteriormente. O estudo de Vogel e Chotia, em 2006, classificou as associações detectadas entre o número de tipos de células e as superfamílias protéicas nas seguintes grandes categorias e subcategorias:

- "Regulação" (subdividido em "transdução de sinais", "atividade de receptor", "outras funções regulatórias", "quinases e fosfatases" e "ligação à DNA")
- "Processos intracelulares" (subdividido em "transporte", "modificação de proteínas", "Proteases", "modificação e transporte de fosfolipídeos", "modificação e transporte de íons", "mobilidade celular" e "ciclo celular / apoptose")
- "Processos extracelulares" (subdividido em "toxinas / defesa", "resposta imune", "adesão celular" e "coagulação")
- "Metabolismo" (subdividido em "transferases", "metabolismo secundário", "redox", "metabolismo e transporte de polissacarídeos", "fotossíntese", "outras enzimas", "metabolismo e transporte de nucleotídeo", "metabolismo e transporte de lipídeos", "energia", "metabolismo e transporte de carboidratos")
- "Informação" (subdividido em "tradução", "processamento de RNA", "replicação/reparo de DNA", "estrutura de cromatina")
- "Geral" (subdividido em "proteína estrutural", "ligação à pequenas moléculas", "interação de proteínas").

A análise dos domínios Pfam que apresentaram correlação positiva realizada em nosso estudo detectou diversas das categorias detectadas anteriormente no trabalho de Voguel e Chotia, 2006, tais como fatores de transcrição associados à proliferação celular, mobilidade celular (tratada em nosso estudo como "citoesqueleto"), regulação do estado redox e elementos de matriz extracelular. Ambos os estudos também detectaram algumas categorias exclusivas, como algumas poucas categorias de plantas em Voguel e Chotia, 2006, não detectadas em nosso estudo, ou os domínios desconhecidos (DUFs), detectados somente em nosso estudo, o que em parte pode representar variações entre os estudos, como o conjunto de genomas utilizados, o nível de resolução analisado (superfamília *versus* domínios protéicos) e os critérios para significância e para o agrupamento de domínios em categorias funcionais, discutido abaixo.

Ambos os trabalhos dependeram de curadoria manual extensa, laboriosa e, de certa maneira, arbitrária, para a categorização dos domínios e superfamílias em categorias mais amplas. Em nosso caso, diversas das categorias biológicas que utilizamos para

classificar os nossos resultados de domínios (e.g. "proteínas de citoesqueleto" ou "receptores de membrana", ou "sistema nervoso") são resultado direto dos resultados de GO, os quais fornecem imediatamente categorias biologicamente relevantes associadas ao aumento da complexidade que nos guiaram na construção de nossas categorias.

Dessa maneira, embora discutamos alguns dos exemplos de domínios associados a categorias funcionais mais amplas abaixo, utilizaremos a estrutura do GO para descrever de maneira mais ampla nossos resultados, pois acreditamos que ela permite uma descrição mais rica e objetiva da evolução da complexidade em *Eukarya*. No caso do trabalho de Voguel e Chotia, 2006, os critérios para a classificação dos domínios em categorias não está disponível. Entretanto, ressalta-se que as categorias propostas pelos autores compreende majoritariamente processos em nível celular, não fazendo discriminação de eventuais fatores que contribuam para o estabelecimento de tecidos e sistemas, o que encontra-se imediatamente disponível em nossas análises de termos GO.

Encontramos dezenas de domínios que, através de anotação manual, puderam ser associados ao desenvolvimento de tecidos, órgãos e sistemas específicos, particularmente ao desenvolvimento dos sistemas nervoso, imunológico, muscular, circulatório, reprodutivo, digestivo e da pele. Estes sistemas e órgãos, de maneira coordenada, permitem que organismos multicelulares metazoários possam interagir com o meio e com outras espécies de maneira cada vez mais complexa, mediante instinto e aprendizado, em atividades como a alimentação, a fuga, o combate aos patógenos e a reprodução. Em conjunto, esses sistemas podem ser vistos como um assinatura molecular dos principais sistemas presentes nos organismos metazoários.

No caso do sistema nervoso, detectamos domínios como as teneurinas, que ocorrem em um conjunto de proteínas conservadas em metazoários envolvidos diretamente com o estabelecimento da morfologia neuronal em metazoários (*Teneurin Intracellular Region* (PF06484) (Drabikowski, Trzebiatowska & Chiquet-Ehrismann, 2005; Antinucci *et al*., 2013). Detectamos também o domínio de polarização de neuroblastos, o qual ocorre em proteínas que induzem a migração de um tipo especial de neurônio (neurônios Q) para posições bem definidas ao longo do eixo anteroposterior do embrião para dar origem aos neurônios sensoriais e de associação (*Q-cell neuroblast polarisation*, (PF10034) (Middelkoop e Korswagen, 2014)). Outros domínio interessante detectado foi o receptor de folato, vitamina essencial para a formação do tubo neural durante embriogênese (*Folate receptor family*, (PF03024)). Assim, observamos que diferentes componentes da formação do sistema nervoso, como a morfologia, migração e

desenvolvimento de estruturas, apresenta correlação positiva com o número de tipos de células em *Eukarya*.

O sistema imunológico confere proteção contra vários tipos de agentes exógenos e patógenos ao organismo tais como vírus, bactérias e células estranhas ou aberrantes, como também executa a reparação de danos nos órgãos e tecidos, a manutenção da integridade e regeneração de tecidos (Rimer, Cohen e Friedman, 2014). A imunidade inata é um processo biológico evolutivamente antigo na árvore da vida e está presente em plantas, invertebrados e vertebrados (Buchmann, 2014; Rimer, Cohen e Friedman, 2014). Como exemplo de imunidade inata, mencionamos o domínio PF08210 (*APOBEC-like N-terminal domain*), que está presente no terminal N da apolipoproteína APOBEC e corresponde uma família de citocinas deaminases evolutivamente conservadas que estão relacionadas com o desenvolvimento de diversidade de anticorpos em linfócitos B em vários eucariotos (Wekedin *et al*., 2003; Iyer *et al*., 2011; Vasudevan *et al*., 2013) e, no caso de mamíferos, está envolvida em mecanismos de proteção à infecção viral (Stavrou e Ross, 2015). Também encontramos vários domínios que estão presentes em várias interleucinas, como o domínio PF00340 (*Interleukin-1 / 18*). Este domínio compreende a família da Interleucina-1, com dois ligantes com atividade agonista nesta família, as interleucina-1 alfa e interleucina-1 beta (IL1-alfa e IL1-beta) que participam da regulação de respostas imunes, reações inflamatórias e hematopoiese (Garlanda, Dinarello e Mantovani *et al*., 2013).

O sistema muscular, em conjunto com o sistema nervoso, permite que os metazoários apresentem uma capacidade de deslocamento no espaço não observada em virtualmente nenhum dos outros organismos multicelulares não-metazoários. Embora comportamentos que envolvem deslocamento espacial sejam descritos em outros táxons, como plantas, estes envolvem a movimentação de um órgão (e.g. flores) em função de um estímulo, sem resultar no deslocamento de todo o corpo do organismo.

No caso de animais, estes dedicam um considerável aporte de recursos e energia na composição do sistema muscular e esquelético, com o qual deslocam todo o seu corpo, muitas vezes por longas distâncias, via locomoção ou outros tipos de mecanismos (Jung e Dasen, 2015). Dentre os domínios com envolvimento no sistema muscular, detectamos domínios que compõem diretamente as proteínas das fibras musculares (e.g *Myosin N-terminal SH3-like domain*, (PF02736) e *Tropomyosin* (PF00261)), além de domínios que participam da reparo de danos às fibras musculares (*Ferlin C-terminus* (PF16165)). Detectamos também, em menor frequência, domínios que atuam em outros sistemas

característicos dos metazoários, tais como os sistemas sanguíneo, reprodutivo e digestivo, além da pele.

Detectamos também dezenas de domínios de proteínas que participam de processos extracelulares primordiais para o desenvolvimento de tecidos, órgãos e sistemas, como a adesão celular, a interação célula-célula e a dinâmica de síntese e degradação de diversos componentes de matriz extracelular. Como exemplos de domínios de adesão celular, destacamos *FG-GAP repeat* (PF01839) e *Cadherin C-terminal cytoplasmic tail, catenin-binding region* (PF15974), os quais são componentes conservados das proteínas integrinas e caderinas, respectivamente, responsáveis pela adesão célula-célula observada nos tecidos de metazoários (Loftus *et al*., 1994; Ishiyama *et al*., 2010).

A matriz extracelular exerce várias funções essenciais para a sobrevivência da célula em organismos multicelulares, contribuindo com a comunicação e adesão entre as células através ligação dos receptores de adesão da superfície celular, sendo fundamental para a evolução da complexidade biológica nos eucariotos superiores como os metazoários (Ozbek *et al*., 2010). Outros domínios de adesão celular repesentados nas nossas análises são o PF07679 (*Immunoglobulin I-set domain*) e PF02210 (*Laminin G domain*). O domínio PF07679 faz parte de uma superfamília de domínios amplamente distribuídos em muitos tipos de receptores de superfície celular, sendo observados em várias moléculas de adesão celular e moléculas de sinalização (Chen, Wang e Wu, 2018). O domínio PF02210 é encontrado na porção C-terminal das cadeias do tipo α das lamininas, que são uma família de glicoproteínas multidomínios conservados nos metazoários envolvidos em várias funções na membrana basal, contribuindo para a formação das matrizes extracelulares nestes organismos (Fahey e Degnan, 2012).

Ao analisarmos os eventos que detectamos em nível celular, observamos aproximadamente uma centena de domínios que fazem parte de diversas vias de sinalização importantes para o estabelecimento da multicelularidade, como receptores, componentes intermediários, fatores de transcrição e modificadores de estrutura de cromatina. Como uma lista não-exaustiva de exemplos dessa categoria, mencionamos o domínio PF00631 (*GGL domain*), que representa a subunidade gama (Gγ) da proteína G. O domínio GGL é encontrado em várias proteínas RGS (reguladoras da sinalização por proteína G) (Sondek e Siderovski, 2001), que correspondem um complexo de sete domínios transmembranas que possuem uma região extracelular transmembranar α-helicoidal ligante externo e um domínio citoplasmático que interage com um complexo

de proteínas G heterotrimérica composto por subunidades Gα, Gβ e Gγ (Krishnan *et al*., 2015). Este complexo de proteína G heterotrimérico atua como um interruptor molecular em várias cascatas de sinalização intracelular, sendo conservados nos excravatas e nos animais (de Mendonza, Sebé-pedrós e Ruiz-Trillo, 2014). Os genes que codificam as RGS são encontrados tanto em organismos eucarióticos superiores como animais, plantas e fungos, e também são encontrados em *Dictyostelium* e *Entamoeba* onde regulam a sinalização intercelular (Wilkie e Kinch, 2005).

Observamos também vários fatores de transcrição já associados à processos biológicos diretamente envolvidos no desenvolvimento na embriogênese e no desenvolvimento de tecidos, órgãos e sistemas, como os domínios "Forkhead" (PF00250, "*Forkhead domain*"), PKNOX1 (PF16493, "*N-terminal of Homeobox Meis and PKNOX1*") e POU (PF00157 "*Pou domain - N-terminal to homeobox domain*"). *Forkhead* consiste em uma família de fatores de transcrição que possuem um domínio de ligação ao DNA estruturalmente conservado de ~100 resíduos e é conhecido como o domínio de hélice (Medina *et al*., 2016). Esta família também é capaz de organizar a produção de transcritos temporalmente e espacialmente durante o desenvolvimento (Zhu, 2016).

PKNOX1 está presente na região terminal N dos membros da família de genes homeobox TALE / MEIS, incluindo a proteína *homeobox* PKNOX e Meis. Um gene homeobox codifica um homeodomínio (HD), que consiste num domínio de ligação ao DNA com 60 aminoácidos e são encontrados em fatores de transcrição em quase todas as espécies eucarióticas (Vonk e Ohm, 2018). Nos eucariotos, esta família está envolvida em uma ampla variedade de processos celulares, como a progressão do ciclo celular, embriogênese, organogênese, proliferação, diferenciação, migração, metabolismo e resposta a danos no DNA (Medina *et al*., 2016; Zhu, 2016) o que demonstra evidências da contribuição destes genes para a formação de diferentes tipos de células ao longo da evolução biológica. Esta família de domínios tem uma origem eucariótica relativamente antiga e divergiu em dezenas de subfamílias em várias linhagens de animais, plantas, fungos e protistas ((Nakagawa *et al*., 2013), (Sebé-Pedrós e de Mendonza, 2015), (Vonk e Ohm, 2018)).

A família de proteínas POU são uma classe de fatores de transcrição homeobox caracterizada por um domínio específico de POU conservado em várias linhagens de metazoários com cerca de 70 resíduos na região terminal N de um domínio específico POU e um HD (homeodomínio) que geralmente contém um resíduo de serina na posição

50 da região terminal C do HD (Cosse-Etchepare *et al*., 2015). O acrônimo POU é derivado dos nomes de três fatores de transcrição de mamíferos, o Pit-1 específico de pituitária, as proteínas de ligação de octâmero Oct-1 e Oct-2 e o Unc-86 neural de Caenorhabditis elegans (Petryniak *et al*., 1990). Os vários membros da família POU têm uma ampla variedade de funções, particularmente ligadas ao desenvolvimento (Nakanoh, 2019).

Alguns domínios que realizam modificações na estrutura da cromatina também foram observadas, como o domínio Pfam PF09011 (*HMG-box domain*). Este domínio curto e conservado de 71 resíduos é um domínio HMG-box (high motility goup) que consiste em três alfa hélices. Domínios HMG-box são encontrados em uma ou mais cópias em proteínas HMG-box, que formam uma família grande e diversa envolvida na regulação de processos dependentes de DNA, tais como transcrição, replicação e reparo das fitas, os quais requerem a flexão e desenrolamento da cromatina (Reeves, 2015). Os domínios HMG-box podem ser encontrados em cópias simples ou múltiplas em várias classes de proteínas de fatores de transcrição que estão envolvidos na gonadogênese diferencial (Harley, Clarkson e Argentaro, 2003) na regulação da organogênese e diferenciação de timócitos (Labbé, Letamendia e Attisano, 2000).

Ainda em âmbito celular, detectamos diversos domínios componentes de canais de membrana com associação significativa com o número de tipos celulares. Uma análise desses domínios detectou tanto componentes específicos de sistemas fisiológicos de metazoários, tais como componentes de canais iônicos importantes para a contração muscular (e.g. *Voltage-gated calcium channel subunit alpha, C-term* (PF16885) (Caterall, 2011)) ou para a ocorrência de impulsos nervosos no sistema nervoso central (*Slow voltage-gated potassium channel* (PF02060) (Moran *et al*., 2015)) e nas vias olfatórias e visuais no no sistema nervoso (e.g. *C-terminal leucine zipper domain of cyclic nucleotide-gated channels* (PF16526) (Biel e Michalakis, 2009)).

Observamos também diversos elementos do citoesqueleto cuja frequência aumenta nos proteomas analisados à medida em que observa-se um aumento da complexidade, compreendendo outra categoria de elementos que atuam em nível celular. No caso dos componentes, detectamos diversos elementos que atuam no citoesqueleto de actina, possuindo papel fundamental em eventos de migração celular, manutenção da morfologia celular e diferenciação de órgãos (e.g. *WH1 domain* (PF00568), *Repeat in HS1/Cortactin* (PF02218), *Tau and MAP protein, tubulin-binding repeat* (PF00418), *Thymosin beta-4 family* (PF01290) (Veltman e Insall, 2010; Uruno *et al*., 2003)).

Em menor frequência, detectamos elementos que atuam em nível celular em diversos outros processos, como no sistema de endomembranas, na biologia de ácidos nucléicos, na composição da carioteca e do poro nuclear, na degradação de proteínas. Assim, em nível celular, observamos que a célula dos organismos mais complexos se torna mais rica em elementos de diversos processos celulares, majoritariamente em vias de sinalização, fatores de transcrição, reguladores epigenéticos e elementos do citoesqueleto, além de uma gama de outros processos em menor frequência.

Cabe ressaltar também uma elevada ocorrência de domínios que fazem parte de oncogenes, bem como domínios que estão presentes em genes supressores de tumores e que atuam como reguladores de várias vias de sinalização, cuja alteração está envolvida em vários tipos de câncer (Trigos *et al*., 2018). Alguns trabalhos, como Aktipis *et al*., 2015, sugerem que a proliferação descontrolada, a degradação do meio, a invasão em outros tecidos e órgãos por vários tipos de tumores e células cancerígenas são contrários aos fundamentos da multicelularidade como a cooperação entre as células, a divisão do trabalho, a adesão, a diferenciação e a comunicação celular, portanto seriam uma involução ao estado ancestral unicelular.

Entre os 216 domínios que apresentaram valores significativos de correlação, detectamos oito domínios com pouca informação funcional, insuficiente para classificá-los em uma das categorias descritas acima, além de 15 domínios caracterizados pelo banco de dados Pfam como Domínios de Função Desconhecida (DUF). Vários destes domínios foram predominantes nos eucariotos mais complexos, principalmente nos vertebrados. Apesar de não terem suas funções biológicas muito bem descritas, diversos dos DUFs representados no nosso banco de dados são observados em proteínas que possuem alguma descrição das suas funções biológicas que permite associá-las aos processos envolvidos com o surgimento da multicelularidade. Como exemplo, destacamos os domínios PF06327 (DUF1053), que faz parte de adenilato ciclases que estão onipresente na regulação de atividades enzimáticas e expressão gênica nos eucariotos e PF13281 (DUF4071), detectado na região terminal N de muitas proteínas semelhantes à serina-treonina quinases. Assim, avaliamos que os domínios pouco caracterizados detectados compreendem candidatos interessantes para a validação funcional, pois podem constituir componentes mecanismos importantes e ainda desconhecidos que eventualmente contribuíram para o aumento do número de tipos de células em *Eukarya*.

Ao final de nossa análise dos domínios com correlação positiva, destacamos que o aumento da complexidade está associado a um aumento relativo de domínios que realizam diversas funções indispensáveis para o surgimento dos organismos com o maior número de células conhecidos, os metazoários vertebrados, do nível sistêmico ao celular. Observamos especialmente componentes de órgãos e sistemas específicos, além de vias de sinalização para crescimento e diferenciação celular, componentes de citoesqueleto importantes para a morfologia dos distintos tipos celulares e componentes de matriz extracelular para permitir o estabelecimento de tecidos.

Os domínios com correlação negativa detectados foram observados em frequência menor que os com correlação positiva (33 versus 183), e correspondem majoritariamente à vias bioquímicas catabólicas e anabólicas e componentes do sistema de tradução e de metabolismo de tRNA, ocorrendo também, em menor frequência, domínios que participam do metabolismo de proteínas, sistema de endomembranas, citoesqueleto, ciclo celular e metabolismo de nucleotídeos. Acreditamos que os organismos com menor número de células, especialmente os que apresentam uma única célula ao longo de todo o seu ciclo de vida, compreendem organismos onde uma única célula necessita realizar diversas tarefas, como lidar com um ambiente molecular muito mais heterogêneo do que o encontrado em organismos multicelulares e com muitos tipos de células, onde grupos específicos de células são especializados em localizar, capturar, digerir e distribuir alimentos, criando um ambiente mais estável para todas as células coletivamente, as quais seriam incapazes de sobreviver individualmente no ambiente de uma célula unicelular de vida livre.

Assim, a maior frequência de vias bioquímicas em genomas de organismos unicelulares possivelmente identifica tanto a pressão seletiva de manutenção dessas vias nos organismos unicelulares quanto a ausência de pressão seletiva para sua manutenção em organismos multicelulares metazoários, os quais, como heterotróficos, obtém nutrientes através da captura ativa, através da locomoção, de alimentos de origem orgânica do meio exterior, o que eventualmente resultaria em ausência de pressão seletiva para a manuntenção de vias de síntese de moléculas orgânicas (Guedes *et al*., 2011).

Embora tenham possibilitado uma boa descrição sobre como o panorama geral do proteoma não-redundante dos eucariotos varia em função do número de tipos de células, as análises de unidades que são descritas em função de sua ancestralidade comum podem eventualmente perder relações importantes dos dados, além de eventualmente requerer, conforme já descrito, uma curadoria manual extensa e, de certa forma, arbitrária.

Tentando obter uma análise mais objetiva e sistemática e, ao mesmo tempo, mais rica e biologicamente relevante, utilizamos termos GO para tentar avaliar quais foram os diferentes aspectos funcionais que aumentaram a sua frequência no genoma de eucariotos à medida que aumenta o número de tipos celulares (Thomas, 2017). O banco de dados dos termos GO compreende um dicionário comum, padronizado e elaborado por especialistas para anotar funções gênicas, e possivelmente representa o mais completo modelo computacional de sistemas biológicos disponível, representando, de maneira curada e objetiva, os mais diversos níveis de processos biológicos, do nível molecular ao sistêmico. Os termos GO também descrevem a localização de proteínas e suas funções moleculares, podendo destacar aspectos importantes da evolução da complexidade em *Eukarya* que não foram adequadamente abordadas nos trabalhos anteriores.

Em nossa análise, ao utilizarmos as dezenas de milhares de termos GO disponíveis para anotar os proteomas não-redundantes em nível gênico, esperávamos refletir, de maneira rica e biologicamente relevante, entretanto uniformizada, não-enviesada e objetiva, quais são as diversas funções biológicas que potencialmente exerceram papeis importantes na evolução de organismos multicelulares. Esperávamos também que o uso de termos GO permitiria tanto captar as diferentes facetas das funções gênicas associadas ao aumento da complexidade quanto evitar a necessidade a categorização arbitrária de genes em processos, que eventualmente podem não representar eventos biológicos reais e perder outras camadas de informação biologicamente relevantes. Cabe destacar, entretanto, que termos GO somente detectam tendências de elementos para os quais já se sabe a função gênica, e não podem detectar elementos genômicos sem anotação funcional (como DUFs) signifcativamente associados ao número de tipos de células.

Os termos GO que apresentaram correlação positiva significativa com o aumento da complexidade detectaram algumas das tendências previamente observadas no estudo de Vogel *et al*., em 2006, bem como em nossas análises de domínios, além de diversas novas tendências não detectadas anteriormente e que retratam um perfil de como a composição relativa de diferentes camadas de processos biológicos aumentou e diminuiu em função do número de tipos celulares. Abaixo traçaremos o perfil detalhado descrito pelos termos GO positivamente associados ao aumento número de tipos de células. De maneira geral, à medida em que aumenta o número de tipos celulares, observamos um aumento na fração do proteoma dedicada à interação entre os componentes dos sistemas biológicos, dos sistemas de biomoléculas ao organismo e seu ambiente.

O termo GO com maior correlação com o número de tipos de células nos parece um bom sumário dos demais termos detectados e do que compreende um organismo complexo: desenvolvimento de sistemas (*system development*, GO:0048731), definido como "o processo onde o resultado específico é a progressão de um sistema de um organismo ao longo do tempo, desde sua formação até sua estrutura madura. Um sistema é um grupo de órgãos ou tecidos que interage de maneira regular ou interdependente e que trabalham em conjunto para executar um determinado processo biológico".

O surgimento de sistemas complexos e especializados é consequência direta da existência de diferentes tipos de células e, também, necessário para sua manutenção. Os termos GO com maior significância na correlação positiva (maior área na Figuras 10, 12 e 14, e primeiros termos da Tabela Suplementar 2) descrevem alguns dos processos biológicos mais gerais necessários para o estabelecimento dos diferentes sistemas biológicos, órgãos, tecidos e células que compõem organismos complexos, tais como crescimento, migração e morte programada de células e desenvolvimento de sistemas e tecidos. Embora diversos dos domínios e superfamílias observados em nosso estudo, bem como nos diversos estudos anteriores que fizeram uso desse tipo de anotação, possam eventualmente sugerir que os processos biológicos acima descritos são importantes para o aumento no número de tipos celulares, ressaltamos que os mesmos foram detectados de maneira automática, consistente, biologicamente coerente e estatisticamente sólida como em nossa análise de termos GO.

Além destes termos mais gerais para o estabelecimento de órgãos e tecidos, observamos também o aumento relativo da frequência de termos que descrevem o desenvolvimento de estruturas anatômicas exclusivas dos organismos mais complexos conhecidos (metazoários) ou de algumas de suas linhagens, tais como os sistemas nervoso, esquelético, cartilaginoso, epiderme e conectivo, e circulatório, alguns dos quais já haviam sido observados em nossas análises de domínios, enquanto outras (esquelético e cartilaginoso) foram detectados somente na análise automatizada de termos GO. Outra modificação em nível celular importante para o funcionamento do sistemas muscular e nervoso é a existência de canais especializados para seu funcionamento. Essa faceta desses sistemas foi captada pela detecção dos diversos termos GO de canais de cálcio (e.g. GO:1901386 - *negative regulation of voltage-gated calcium channel activity*). Observamos também diversos GOs associados à detecção de componentes do meio externo, como nutrientes e estímulos luminosos, os quais podem ser subcomponentes do sistema nervoso e indicam um aumento relativo do proteoma dedicado à interação com o

ambiente; essas funções também não foram detectados em nossa análise de domínios Pfam.

Em organismos multicelulares de grande complexidade, a homeostase necessária para a sua existência cria um ambiente estável e rico em nutrientes (seu próprio corpo), o que produz um nicho para a eventual evolução de organismos parasitos, iniciando a competição da fração do componente do proteoma dedicado ao sistema imune ou à patogenicidade. Assim, propomos que o aumento na complexidade pode estar produzindo uma corrida armamentista entre os organismos parasitos e seus hospedeiros também no conteúdo relativo do proteoma do hospedeiro dedicado ao sistema imune, com compoentes importantes, como a imunidade inata e adaptativa em nível celular, os quais são observados em diferentes níveis em plantas e metazoários, especialmente na linhagem dos vertebrados (Buchmann, 2014; van Valen, 1973).

Observamos diversos termos que descrevem a regulação de processos nos mais diversos níveis de um organismo, de órgãos e tecidos às células e às redes moleculares. Exemplos nesse sentido são os termos de regulação do desenvolvimento, regulação da resposta à estímulos externos, regulação de vias sinalização, regulação de migração celular, regulação de morte celular programada e regulação de funções moleculares, dentre outros, o que evidencia que organismos mais complexos possuem um aumento significativo da proporção relativa de seus proteomas dedicados à regulação dos mais diferentes processos biológicos nos mais diferentes níveis, o que é essencial para permitir a coexistência harmônica dos diferentes sistemas que compõem um organismo com múltiplos tipos celulares sem a emergência de doenças que decorrem da multiplicação descontrolada de células.

Em nossas análises, detectamos que a fração relativa do proteoma de *Eukarya* anotada como pertencendo à vias de sinalização também se expande à medida que aumenta o número de tipos celulares (e.g. GO:0043235 - *receptor complex*, GO:0030545 - *receptor regulator activity*). Além desse padrão geral observado, detectamos diversos grandes grupos e classes de vias bioquímicas de sinalização específicas, tais como receptores, componentes intermediários e fatores de transcrição, com associação positiva significativa com o aumento da complexidade em *Eukarya* e que possuem papel de destaque em diversos aspectos necessários para o estabelecimento dos organismos eucarióticos complexos.

Diversos componentes destas vias, detectadas de maneira automática em nossa análise GO, também foram detectadas em nossa curadoria manual de domínios Pfam,

enquanto outras foram detectadas somente na análise de GO, possivelmente por constituírem elementos genômicos não-homólogos que fazem parte da mesma via bioquímica. Descreveremos os diferentes módulos das vias, iniciando pelas que recebem o nome de seus ligantes ou receptores, passando posteriormente à descrição dos componentes intermediários das vias, como quinases.

Como ligantes de vias, detectamos enriquecimento para os ligantes dos fatores de crescimento *insulin-like* (GO:0005520 - *insulin-like growth factor binding*). Essas moléculas compreendem polipeptídeos que são hormônios de estrutura molecular similar à insulina, que promove o crescimento e o desenvolvimento em humanos (Yakar *et al*., 2002). A sinalização dessa via começa pela ligação dos peptídeos ao receptor, o qua sinaliza, via a GTPase de baixo peso molecular Ras (a qual também compreende uma das vias encontradas em nossa análise de GO, discutida abaixo) e posterior ativação de quinases ativadas por mitógenos (MAPK, também detectada e discutida abaixo), o que ocasiona a fosforilação de múltiplos substratos e a transcrição de diversos genes (Laviola, Natalicchio e Giogirno, 2007). Adicionalmente, a ativação cruzada entre esta via e as proteínas acopladas à proteína G (também detectada e descrita abaixo) mediadas pela quinase mTOR (também detectada e descrita abaixo) já foi reportada (Rozengurt, Sinnet-Smith e Klsfalvi, 2010). Embora não tenhamos detectado estudos sobre a evolução desses peptídeos, seus receptores tirosina-quinases já foram observados em esponjas, sugerindo que sua origem pode ter ocorrido nos primeiros metazoários (Skorokhod *et al*., 1999).

A segunda classe de ligantes com correlação positiva com o aumento da complexidade foi a via de sinalização de fator de crescimento de fibroblasto (GO:0008543 - *fibroblast growth factor receptor signaling pathway*), os quais compreendem uma família de peptídeos estruturalmente relacionados, exclusiva de metazoários, que ativa diversas vias de sinalização mediadas por quinases como MAPK, PKC (discutida abaixo) e a GTPase de baixo peso molecular Ras (discutida abaixo). A ativação cruzada entre essa via e as proteínas acopladas à proteína G também já foram descritas (Wang, 2016). Em humanos, essa via de sinalização é essencial na embriogênese, atuando também e como fator homeostático na resposta ao dano tecidual (Itoh e Ornitz, 2011). A evolução dessa via parece ser complexa, com diversos eventos de duplicação e perda de genes nos metazoários, e com um relativo aumento (Oulion, Bertrand e Escriva, 2012).

A evolução de metazoários dos seus ancestrais unicelulares dependeu da emergência de novos mecanismos para adesão e comunicação célula-célula, onde as

integrinas compreenem um dos componentes mais importantes (Hynes, 2002) (GO:0007229 - *integrin-mediated signaling pathway*). Integrinas são receptores transmembrana que participam de interações célula-célula e destas com a matriz extracelular, participando de diversos processos como proliferação, migração e morte celular programada (Cary, Han e Guan, 1999). Estas vias ativam majoritariamente as proteínas Ras e MAPK, embora eventos de sinalização cruzada com a via dos receptores acoplados à proteína G também já tenha sido observada (Teoh, tam e tran, 2012). Embora tenha sido descrita inicialmente como exclusiva dos metazoários, estudos de genômica comparativa reportaram componentes dessa via em grupos basais unicelulares de *Eukarya*, indicando que a mesma foi perdida independentemente em fungos, plantas e coanoflagelados (Sebé-Pedrós *et al*., 2010).

A via mediada por receptores de ácido retinóico (vitamina A) compreende outro receptor que pertence às vias detectadas relacionadas ao desenvolvimento de múltiplos tipos celulares em *Eukarya* (GO:0048384 - *retinoic acid receptor signaling pathway*). Diferentemente das vias descritas anteriormente, as vias vias de sinalização por receptor de ácido retinóico possuem seus receptores intracelulares, uma vez que o ácido retinóico é lipossolúvel e atravessa as membranas biológicas. As vias de sinalização mediadas por ácido retinóico são importantes para os processos de crescimento, diferenciação, proliferação e apoptose, contribuíndo, dentre outros processos, para a determinação do padrão anteroposterior da placa neural, induzindo a formação da porção posterior do embrião (Das *et al*., 2014).

Receptores acoplados à proteína G (GPCRs) (GO:0008277 - *regulation of G protein-coupled receptor signaling pathway*) compreendem a maior diversidade de proteínas transmembrana observadas em eucariotos (Tuteja, 2009; Keshlava *et al*., 2018), e sua origem parece ter ocorrido antes da emergência dos metazoários (Mushegian *et al*., 2012). Nos metazoários, episódios de expansões de famílias de GPCRs por eventos de duplicação e divergência foram essenciais para a evolução de várias funções sensoriais responsáveis pela percepção ambiental que resultaram em vários aspectos ecológicos necessários para a sobrevivência, e já foram descritas em diversas linhagens de organismos multicelulares. (de Mendonza, Sebé-Pedrós e Ruiz-Trillo, 2014). Os GPCRs estão envolvidos na evolução de receptores gustativos em mosquitos transmissores de doenças, quimiorreceptores em nematóides, detecção de feromônio nos vertebrados e receptores olfativos em mamíferos (Ritschard *et al*., 2019). Receptores dessa natureza também estão associados com o controle do desenvolvimento, crescimento celular,

migração, sensor de densidade populacional e transmissão nervosa (Rosenbaum, Rasmussen e Kobilka, 2009).

As GTPases de baixo peso molecular Ras, também detectadas com associação significativa positiva, são componentes intermediários de vias de transdução de sinal para a proliferação, diferenciação e sobrevivência celular, atuando como interruptores de ativação e inativação dessas vias (Wennerberg, 2005). Essas GTPases são componentes intermediários de vias de sinalização mediada por MAPK, determinando os padrões de ativação diferencial dessa quinase no corpo do embrião durante seu desenvolvimento (Shvartsman, Coppey e Berezhkovskii, 2008). Diversas das outras vias detectadas nesse estudo realizam sua sinalização via Ras, conforme descrito anteriormente

A fosforilação de proteínas em resíduos de tirosina pode ter contribuído para o advento da multicelularidade nos eucariotos superiores como plantas terrestres e animais (Cock e Collen, 2015). Expansões e contrações de famílias de quinases receptoras na membrana são processos antigos nos eucariotos, e provavelmente existiam antes da irradiação dos principais clados de eucariotos como as plantas terrestres, os metazoários e os fungos (Suga *et al*. 2012). Quinases receptoras transmembranas participam da comunicação entre as células e da percepção de sinais extracelulares, atuando em diversos dos processos importantes para o surgimento de múltiplos tipos de células, como migração, divisão e diferenciação celular, dentre outros (De Smet *et al*., 2009).

Observamos associação significativa entre a frequência geral de quinases e a complexidade dos organismos (GO:0004675 - *transmembrane receptor protein serine/threonine kinase activity*), além de diversas quinases específicas já descritas como associadas com o desenvolvimento embrionário. O primeiro conjunto detectado compreende diversas quinases, tais como as vias de sinalização via quinase *Hippo*, a via da quinase alvo da rapamicina (mTOR), a quinase ativada por AMP cíclico (PKA), a quinase C (PKC), quinases dependentes de ciclina e as quinases ativadas por mitógeno (MAPK). As quinases controlam os processos de fosforilação, o qual é frequentemente utilizado para ligar e desligar vias de sinalização via fosforilação reversível de proteínas-alvo. A via *Hippo* coordena os processos de apoptose e proliferação celular nos metazoários, tendo sido identificada primeiro em *Drosophila melanogaster* e depois extensivamente estudada em *Drosophila* e mamíferos. Essa via compreende uma quinase central (*Hippo*) e diversos componentes, como coativadores de transcrição e e parceiros de ligação ao DNA, compreendendo mais de 30 fatores conhecidos (Meng, Moroishi e Guan, 2016).

As funções da *Hippo* é o controle do crescimento de tecidos em adultos e a modulação da proliferação celular, diferenciação e migração nos órgãos em desenvolvimento. Estudos mais recentes apontam alguns componentes da via *Hippo* com origem anterior aos metazoários, na espécie *Capsaspora owczarzaki*, de especial importância como grupo-irmão de Choanozoa (composto, por sua vez, por coanoflagelados metazoários) e compreendendo um dos ancestrais unicelulares mais próximos dos metazoários (Sebé-Pedrós *et al*., 2012).

A via da quinase mTOR (*mammalian target of rapamycin*) foi descoberta inicialmente em *Saccharomyces cerevisiae*, sendo posteriormente descrita como uma via existente em todos os eucariotos, regulando diversos processos. Nos eucariotos multicelulares, essa via já foi descrita como um regulador do crescimento, mobilidade e proliferação celular (Saxton e Sabatini, 2017). Embora seja uma via antiga, onde a maior parte dos componentes dessa via surgiu antes do último ancestral comum de *Eukarya*, uma expansão e diversificação dessa via, causada por duplicações e eventos de deleção e subfuncionalização, é descrita nos metazoários (van Dam, *et al*., 2011).

Outras quinases detectadas como associadas com o aumento da complexidade foram as MAPK (*Mitogen-Activated Protein Kinase*, proteína quinase ativada por mitógeno), das quais a quinase JNK faz parte, e que compreendem vias de transdução de sinal evolutivamente conservadas, observadas em virtualmente todos eucariotos estudados até o momento (Ichimura *et al*., 2002). Estas quinases respondem a diversos estímulos extracelulares e, nos organismos com diversos tipos celulares, controlam um vasto número de processos celulares fundamentais, como proliferação, crescimento, diferenciação, resposta à estresse, sobrevivência e apoptose (Plotnikov *et al*., 2011).

As quinases PKA e PKC possuem papel fisiológico na manutenção da homeostase em diversos órgãos e tecidos, possuindo atividades descritas no tecido adiposo, muscular, cardiovascular, fígado, rins, e pele, enquanto as quinases dependentes de ciclinas que possuem papel regulatório no controle do ciclo celular (Nelson e Cox, 2014). Ambas as classes de processos (regulação do funcionamento dos diferentes órgãos e tecidos, regulação da proliferação celular) constituem aspectos necessários para a existência dos organismos multicelulares.

Historicamente, cabe destacar que diversos agentes das vias que descrevemos, tais como as quinases ativadas por mitógenos (Macnamara, Baker e Maini, 2011), a quinase mTOR (Ronch *et al*., 1993), o ácido retinóico (Casci, 2008), o fator de crescimento de fibroblasto (Bokel e Brand, 2013), as vias mediadas por Ras (Shvartsman, Coppey e

Berezhkovskii, 2008), os fatores de crescimento *insulin-like* (Rudman *et al*., 1997), são morfogênicos (*morphogen*). Estas moléculas compreendem alguns dos mais importantes agentes durante a embriogênese, variando sua concentração ao longo dos eixos corporais do embrião e, em função disso, permitindo que as células do embrião em desenvolvimento recuperem informações de sua localização e regulando etapas primordiais do desenvolvimento embrionário, como migração celular, diferenciação e estabelecimento dos eixos corporais (Rogers e Schier, 2011; Sagner e Briscoe, 2017).

Não detectamos nenhum GO que representasse um conjunto específico de fatores de transcrição, como a análise dos domínios detectou. Entretanto, observamos termos GO mais gerais e associados à eventos associados diretamente à transcrição do DNA eucariótico. Especificamente, os componentes da cromatina e os fatores de transcrição, de maneira geral, apresentam correlação positiva significativa com o número de tipos celulares (GO:0031490 - *chromatin DNA binding*, GO:0008134 - *transcription factor binding*). Conforme já exposto na discussão dos domínios Pfam, diversos grupos de fatores de transcrição estão envolvidos no desenvolvimento embrionário das embriófias e metazoários (de Mendoza *et al*., 2013). A cromatina também está estritamente associada à regulação do acesso da maquinaria de transcrição ao DNA nos eucariotos (Koster, Snel e Timmers, 2015).

Do ponto de vista da localização celular, detectamos uma maior frequência de proteínas localizadas na interface celular ou fora da célula à medida em que aumenta a complexidade dos eucariotos (GO:0005576 - *extracellular region*, GO:0009986 - *cell surface*). Conforme extensivamente discutido anteriormente, propomos que isso reflete como o microambiente extracelular (matriz extracelular) e a interação da célula com o seu microambiente local (outras células e os componentes da matriz extracelular) se tornam cada vez mais importantes para o estabelecimento de múltiplos tipos de células.

Curiosamente, observamos também um aumento da frequência de genes possuíndo anotações de elementos de vírus à medida em que aumenta a complexidade. Embora ainda não tenhamos realizado a análise de quais são os genes anotados dessa maneira, há casos descritos de elementos virais exaptados e que contribuírem para o surgimento de novidades evolutivas em mamíferos. Um exemplo de exaptação é o uso de uma proteína de origem viral para a comunicação intercelular no sistema nervoso através da produção de moléculas *capside-like* (Pastuzyn *et al*., 2018). Além disso, o surgimento de uma novidade evolutiva importante na linhagem dos mamíferos, a placenta, deriva da

exaptação de elementos de origem viral para a formação do sinciotrofoblasto, onde ocorre também a produção de *capside-like* (Mi *et al*., 2000).

A análise das funções moleculares associadas ao aumento do número de tipos de células detectou majoritariamente as principais classes de receptores de membrana descritos anteriormente, bem como a função de quinase. Outro componente previamente detectado é a função reguladora dos respectivos receptores. Em nível de DNA, observamos novamente a função de ligação à cromatina.

Os termos GO com correlação negativa descrevem um cenário semelhante ao detectado nos domínios Pfam, com uma maior frequência de vias metabólicas e catabólicas, metabolismo energético e tradução nos eucariotos unicelulares (e.g. GO:0006564 - *L-serine biosynthetic process*, GO:0006072 - *glycerol-3-phosphate metabolic process*, GO:0009055 - *electron transfer activity* e GO:0006412 - *translation*). Acreditamos que esse cenário descreve tanto o perfil dos organismos multicelulares, os quais eventualmente perderam diversas destas vias em função do ambiente mais homeostático proporcionado pela especialização celular e da consequente ausência de pressão seletiva para sua manutenção, quanto o dos unicelulares, os quais lidam com um perfil heterogêneo de componentes externos ao longo do seu dia com uma única célula, generalista por definição.

7. Conclusões

Os eucariotos apresentam uma vasta variação fenotípica em relação ao número de tipos celulares. Os genomas desse táxon também apresentam uma vasta variação natural em sua composição, bem como eventuais viéses decorrentes do procedimento de montagem de genomas e de predição gênica, dentre outros. Nesse projeto, selecionamos espécies distribuídas ao longo do táxon *Eukarya* para os quais três informações encontravam-se disponíveis: um genoma de alta qualidade, o número de tipos celulares e a sua posição na filogenia desse táxon. De posse dessas informações, investigamos como algumas propriedades genômicas, especificamente a frequência de domínios protéicos e de termos GO, variava em função do número de tipos de células. Para remover possíveis viéses causados por dependência dos dados e possíveis *outliers*, fizemos uso de procedimentos estatísticos para considerar a história filogenética compartilhada entre as espécies, bem como para minimizar a influência de pontos com valores atipicamente muito distantes dos demais valores observados.

A análise de domínios protéicos detectou diversos padrões interessantes, mas falhou em detectar como elementos não-homólogos, mas que possuem a mesma função molecular, papel biológico ou localização celular. Esta análise também detectou domínios desconhecidos cuja frequência está associada ao número de tipos de células, os quais constituem alvos interessantes para caracterização funcional.

A análise automática em nível de GO detectou várias das tendências detectadas durante a curadoria manual dos domínios Pfam, bem como diversas outras associações não detectadas. Os termos GO descreveram, de maneira automática e sem a necessidade de curadoria manual, o painel geral necessário para a emergência de sistemas, órgãos e tecidos, como diversas das vias de sinalização, mecanismos regulatórios e os processos dinâmicos de crescimento, morte e sinalização celular. Assim, consideramos que as análises de Pfam e GO detectaram perfis complementares de padrões genômicos associados ao aumento da complexidade em *Eukarya*.

O que faz um eucarioto complexo? Nossa análise sugere que uma fração cada vez maior do proteoma dos organismos multicelulares é dedicada à interação, à especialização e ao controle dos mais diversos tipos de processos necessários para a emergência dos organismos multicelulares complexos, do nível molecular ao sistêmico.

8. Considerações Finais/Perspectivas

Acreditamos que a metodologia aqui apresentada possa constituir uma ferramenta interessante para descobrir domínios (e outros elementos genômicos) associados com outras variáveis fenotípicas quantitativas de interesse que podem ser *proxies* interessantes para avaliar outros fenótipos de interesse. Exemplos de estudos nessa direção abordados por nosso grupo de pesquisa envolvem utilizar o número de neurônios em metazoários para a avaliação da evolução do sistema nervoso central nessa linhagem, bem como o número de indivíduos em colônias de *Hymenoptera* para avaliação da evolução da eussociabilidade.

Os resultados aqui apresentados detectaram elementos genômicos e processos biológicos conhecidos na evolução da complexidade eucariótica, bem como outros elementos ainda não caracterizados que podem potencialmente constituir novos processos biológicos associados ao número de tipos de células. Como trabalhos futuros, planejamos ampliar nossas análises para outras regiões do genoma além dos genes codificadores de proteínas, de modo a investigar como outras entidades biológicas, tais como sítios de ligação a promotores, regiões de DNA repetitivo e número de genes parálogos, variam em função do aumento da complexidade biológica entre os eucariotos.

9. Referências Bibliográficas

Adami C (2002) What is complexity? Bioessays 24(12):1085–1094.

Ahmed, S. M., Nishida-Fukuda, H., Li, Y., McDonald, W. H., Gradinaru, C. C., & Macara, I. G. (2018). Exocyst dynamics during vesicle tethering and fusion. Nature Communications, 9(1). doi:10.1038/s41467-018-07467-5

Aktipis, C. A., Boddy, A. M., Jansen, G., Hibner, U., Hochberg, M. E., Maley, C. C., & Wilkinson, G. S. (2015). Cancer across the tree of life: cooperation and cheating in multicellularity. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1673), 20140219–20140219. doi:10.1098/rstb.2014.0219

Anderson, D. P., Whitney, D. S., Hanson-Smith, V., Woznica, A., Campodonico-Burnett, W., Volkman, B. F. King, N., Thornton, J., Prehoda, K. E. (2015). Evolution of an ancient protein function involved in organized multicellularity in animals. eLife, 5.

Antinucci, P., Nikolaou, N., Meyer, M. P., & Hindges, R. (2013). Teneurin-3 Specifies Morphological and Functional Connectivity of Retinal Ganglion Cells in the Vertebrate Visual System. Cell Reports, 5(3), 582–592.

Aravind, L., Anantharaman, V., Abhiman, S., & Iyer, L. M. (2014). Evolution of Eukaryotic Chromatin Proteins and Transcription Factors. Protein Families, 421–502.

Arenas-Mena, C. (2017). The origins of developmental gene regulation. Evolution & Development, 19(2), 96–107. doi:10.1111/ede.12217

Arendt, D. (2008). The evolution of cell types in animals: emerging principles from molecular studies. Nature Reviews Genetics, 9(11), 868–882.

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., Wagner, G. P. (2016). The origin and evolution of cell types. Nature Reviews Genetics, 17(12), 744–757.

Arthur, B. (1994). On the evolution of complexity. Pp. 65-81 in G. A. Cowan, D. Pines, and D. Meltzer, eds. Complexity: Metaphors, models, and reality. Addison-Wesley, Reading, MA.

Ashburner, M. et. al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29. doi:10.1038/75556.

Babushok DV, Ostertag EM, Kazazian HH Jr. 2007. Current topics in genome evolution: Molecular mechanisms of new gene formation. Cell Mol Life Sci 64: 542–554

Bharathan, G., Janssen, B.-J., Kellogg, E.A., Freeling, M., 1997. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Evolution (N. Y) 94, 13749e13753. https://doi.org/10.1073/pnas.94.25.13749.

Barlow, L. D., Nývltová, E., Aguilar, M., Tachezy, J., & Dacks, J. B. (2018). A sophisticated, differentiated Golgi in the ancestor of eukaryotes. BMC Biology, 16(1). doi:10.1186/s12915-018-0492-9

Baum, D. A., & Baum, B. (2014). An inside-out origin for the eukaryotic cell. BMC Biology, 12(1).

Basu, M. K. et. al. (2008). Evolution of protein domain promiscuity in eukaryotes. Genome Research, 18(3), 449–461. doi:10.1101/gr.6943508

Bell, G., & Mooers, A. O. (1997). Size and complexity among multicellular organisms. Biological Journal of the Linnean Society, 60(3), 345–363.

Bennett, M. D., & Leitch, I. J. (2005). Genome Size Evolution in Plants. The Evolution of the Genome, 89–162. doi:10.1016/b978-012301463-4/50004-8

Bhullar, K. S., Lagarón, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., & Rupasinghe, H. P. V. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. Molecular Cancer, 17(1). doi:10.1186/s12943-018-0804-2

Björklund, Å. K., Ekman, D., & Elofsson, A. (2006). Expansion of Protein Domain Repeats. PLoS Computational Biology, 2(8), e114. doi:10.1371/journal.pcbi.0020114

Bonner, J. T. (1988). The evolution of complexity. Princeton Univ. Press, Princeton, NJ.

Bökel, C., & Brand, M. (2013). Generation and interpretation of FGF morphogen gradients in vertebrates. Current Opinion in Genetics & Development, 23(4), 415–422. doi:10.1016/j.gde.2013.03.002

Buchmann, K. (2014). Evolution of Innate Immunity: Clues from Invertebrates via Fish to Mammals. Frontiers in Immunology, 5. doi:10.3389/fimmu.2014.00459

Bürglin, T. R., & Affolter, M. (2015). Homeodomain proteins: an update. Chromosoma, 125(3), 497–521. doi:10.1007/s00412-015-0543-8

Carroll, S. B. (2001). Chance and necessity: the evolution of morphological complexity and diversity. Nature, 409(6823), 1102–1109. doi:10.1038/35059227

Carroll, S. B. (2005). Evolution at Two Levels: On Genes and Form. PLoS Biology, 3(7), e245. doi:10.1371/journal.pbio.0030245.

Cary LA, Han DC, Guan JL (1999) Integrin-mediated signal transduction pathways. Histol Histopathol 14, 1001–1009.

Casci, T. (2008). Retinoic acid passes the morphogen test. Nature Reviews Genetics, 9(1), 7–7. doi:10.1038/nrg2293

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. International Journal of Systematic and Evolutionary Microbiology, 52(2), 297–354. doi:10.1099/00207713-52-2-297

Chen, L. et. al. (2014). Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. Molecular Biology and Evolution, 31(6), 1402–1413. doi:10.1093/molbev/msu083.

Chen, J., & Wang, N. (2018). Tissue cell differentiation and multicellular evolution via cytoskeletal stiffening in mechanically stressed microenvironments. Acta Mechanica Sinica. doi:10.1007/s10409-018-0814-8

Chen, J., Wang, B., & Wu, Y. (2018). Structural Characterization and Function Prediction of Immunoglobulin-like Fold in Cell Adhesion and Cell Signaling. Journal of Chemical Information and Modeling, 58(2), 532–542. doi:10.1021/acs.jcim.7b00580

Cheng, H.-C., Qi, R. Z., Paudel, H., & Zhu, H.-J. (2011). Regulation and Function of Protein Kinases and Phosphatases. Enzyme Research, 2011, 1–3. doi:10.4061/2011/794089

Cheung, P. P., & Pfeffer, S. R. (2016). Transport Vesicle Tethering at the Trans Golgi Network: Coiled Coil Proteins in Action. Frontiers in Cell and Developmental Biology, 4. doi:10.3389/fcell.2016.00018

Chothia C, Gough J, Vogel C, Teichmann S. 2003. Evolution of the protein repertoire. Science 300: 1701–1703.

Chothia, C., & Gough, J. (2009). Genomic and structural aspects of protein evolution. Biochemical Journal, 419(1), 15–28.

Choudhuri, S. (2014). Fundamentals of Molecular Evolution. Bioinformatics for Beginners, 27–53.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. Nucleic Acids Research, 44(D1), D67–D72. doi:10.1093/nar/gkv1276

Cock, J. M., & Collén, J. (2015). Independent Emergence of Complex Multicellularity in the Brown and Red Algae. Advances in Marine Genomics, 335–361.

Cohen-Gihon, I., Fong, J. H., Sharan, R., Nussinov, R., Przytycka, T. M., & Panchenko, A. R. (2011). Evolution of domain promiscuity in eukaryotic genomes—a perspective from the inferred ancestral domain architectures. Mol. BioSyst., 7(3), 784–792. doi:10.1039/c0mb00182a

Cornwell, W., & Nakagawa, S. (2017). Phylogenetic comparative methods. Current Biology, 27(9), R333–R336. doi:10.1016/j.cub.2017.03.049

Cosse-Etchepare, C., Gervi, I., Buisson, I., Formery, L., Schubert, M., Riou, J.-F., Le Bouffant, R. (2018). Pou3f transcription factor expression during embryonic development highlights distinct pou3f3 and pou3f4 localization in the Xenopus laevis kidney. The International Journal of Developmental Biology, 62(4-5), 325–333. doi:10.1387/ijdb.170260rl

Crowder, S. W., Leonardo, V., Whittaker, T., Papathanasiou, P., & Stevens, M. M. (2016). Material Cues as Potent Regulators of Epigenetics and Stem Cell Function. Cell Stem Cell, 18(1), 39–52. doi:10.1016/j.stem.2015.12.012

Dacks, J. B., Poon, P. P., & Field, M. C. (2008). Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution. Proceedings of the National Academy of Sciences, 105(2), 588–593. doi:10.1073/pnas.0707318105

Das, S., Dawson, N. L., & Orengo, C. A. (2015). Diversity in protein domain superfamilies. Current Opinion in Genetics & Development, 35, 40–49.

Das, B. C., Thapa, P., Karki, R., Das, S., Mahapatra, S., Liu, T.-C., … Evans, T. (2014). Retinoic acid signaling pathways in development and diseases. Bioorganic & Medicinal Chemistry, 22(2), 673–683. doi:10.1016/j.bmc.2013.11.025

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 1;36 (Web Server issue):W465-9. Epub 2008 Apr 19.

De Mendoza, A., Sebé-Pedrós, A., & Ruiz-Trillo, I. (2014). The Evolution of the GPCR Signaling System in Eukaryotes: Modularity, Conservation, and the Transition to Metazoan Multicellularity. Genome Biology and Evolution, 6(3), 606–619. doi:10.1093/gbe/evu038

De Mendoza, A., Sebe-Pedros, A., Sestak, M. S., Matejcic, M., Torruella, G., Domazet-Loso, T., & Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proceedings of the National Academy of Sciences, 110(50), E4858–E4866. doi:10.1073/pnas.1311818110

De Smet, I., Voß, U., Jürgens, G., & Beeckman, T. (2009). Receptor-like kinases shape the plant. Nature Cell Biology, 11(10), 1166–1173. doi:10.1038/ncb1009-1166

Di Roberto, R. B., & Peisajovich, S. G. (2013). The role of domain shuffling in the evolution of signaling networks. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 322(2), 65–72. doi:10.1002/jez.b.22551

Diniz-Filho, J. A. F. 2000. Métodos Filogenéticos Comparativos. Holos, Ribeirão Preto.

Dobzhansky, T. (1973). Nothing in Biology Makes Sense except in the Light of Evolution. The American Biology Teacher, 35(3), 125–129.

Drabikowski, K., Trzebiatowska, A., & Chiquet-Ehrismann, R. (2005). ten-1, an essential gene for germ cell development, epidermal morphogenesis, gonad migration, and neuronal pathfinding in Caenorhabditis elegans. Developmental Biology, 282(1), 27–38.

Drayer, L., & van Haastert, P. J. M. (1994). Transmembrane signalling in eukaryotes: a comparison between higher and lower eukaryotes. Plant Molecular Biology, 26(5), 1239–1270. doi:10.1007/bf00016473

Fahey, B., & Degnan, B. M. (2012). Origin and Evolution of Laminin Gene Family Diversity. Molecular Biology and Evolution, 29(7), 1823–1836. doi:10.1093/molbev/mss060

Felsenstein, J. (1985). Phylogenies and the Comparative Method. The American Naturalist, 125(1), 1–15. doi:10.1086/284325

Finlay, B. J., & Esteban, G. F. (2009). Can Biological Complexity Be Rationalized? BioScience, 59(4), 333–340.

Finn, R. D. et. al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research, 44(D1), D279–D285. doi:10.1093/nar/gkv1344

Fong, J. H., Geer, L. Y., Panchenko, A. R., & Bryant, S. H. (2007). Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony. Journal of Molecular Biology, 366(1), 307–315. doi:10.1016/j.jmb.2006.11.017

França, G. S., Cancherini, D. V., & de Souza, S. J. (2012). Evolutionary history of exon shuffling. Genetica, 140(4-6), 249–257.

Garland Jr., Ives AR (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am Nat 155(3):346–364. doi:10.1086/303327

Gilbert W (1978) Why genes in pieces? Nature 271:501.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. Statistics in Medicine, 33(11), 1946–1978.

Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of

known structure. Journal of Molecular Biology, 313(4), 903–919. doi:10.1006/jmbi.2001.5080

Gregory, T. R. (2005). The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership. Annals of Botany, 95(1), 133–146. doi:10.1093/aob/mci009

Gomperts, B., Kramer, I. & Tatham, P. 2004. Signal Transduction. Elsevier Academic Press, Amsterdam. 424 p.

Grosberg, R. K., & Strathmann, R. R. (2007). The Evolution of Multicellularity: A Minor Major Transition? Annual Review of Ecology, Evolution, and Systematics, 38(1), 621–654.

Gunning, P. W., Ghoshdastider, U., Whitaker, S., Popp, D., & Robinson, R. C. (2015). The evolution of compositionally and functionally distinct actin filaments. Journal of Cell Science, 128(11), 2009–2019. doi:10.1242/jcs.165563

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075.

Hajheidari M, Koncz C and Bucher M. (2019) Chromatin evolution - key innovations underpinning morphological complexity. Front. Plant Sci. 10:454. doi:10.3389/fpls.2019.00454

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., & Clarke, J. (2007). The folding and evolution of multidomain proteins. Nature Reviews Molecular Cell Biology, 8(4), 319–330.

Harley, V. R., Clarkson, M. J., & Argentaro, A. (2003). The Molecular Action and Regulation of the Testis-Determining Factors, SRY (Sex-Determining Region on the Y Chromosome) and SOX9 [SRY-Related High-Mobility Group (HMG) Box 9]. Endocrine Reviews, 24(4), 466–487. doi:10.1210/er.2002-0025

Haygood R. 2006. Proceedings of the SMBE Tri-National Young Investigators' Workshop (2005). Mutation rate and the cost of complexity. Mol Biol Evol. 23:957–963.

Hedges S, Blair J, Venturi M, Shoe J. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 4:2.

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics, 22(23), 2971–2972. doi:10.1093/bioinformatics/btl505

Hedges, S. B., Marin, J., Suleski, M., Paymer M., Kumar, S. (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. Mol Biol Evol (2015) 32: 835-845.

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics, 16(1). doi:10.1186/s12859-015-0611-3

Herron, M. D., & Nedelcu, A. M. (2015). Volvocine Algae: From Simple to Complex Multicellularity. Advances in Marine Genomics, 129–152. doi:10.1007/978-94-017-9642-2_7

Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y., & Fostier, J. (2017). Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics, 18.

Hynes, R. O. (2002). Integrins. Cell, 110(6), 673–687. doi:10.1016/s0092-8674(02)00971-6

Ichimura, K., Shinozaki, K., Tena, G., Sheen, J., Henry, Y., … Walker, J. C. (2002). Mitogen-activated protein kinase cascades in plants: a new nomenclature. Trends in Plant Science, 7(7), 301–308. doi:10.1016/s1360-1385(02)02302-6

Ishiyama, N., Lee, S.-H., Liu, S., Li, G.-Y., Smith, M. J., Reichardt, L. F., & Ikura, M. (2010). Dynamic and Static Interactions between p120 Catenin and E-Cadherin Regulate the Stability of Cell-Cell Adhesion. Cell, 141(1), 117–128. doi:10.1016/j.cell.2010.01.017

Itoh, N., & Ornitz, D. M. (2010). Fibroblast growth factors: from molecular evolution to roles in development, metabolism and disease. Journal of Biochemistry, 149(2), 121–130. doi:10.1093/jb/mvq121

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWillian, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador_vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236–1240. doi:10.1093/bioinformatics/btu031

Jung, H., & Dasen, J. S. (2015). Evolution of Patterning Systems and Circuit Elements for Locomotion. Developmental Cell, 32(4), 408–422. doi:10.1016/j.devcel.2015.01.008

Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics, 26(13), 1669–1670.

Kanapin, A. A., Mulder, N., & Kuznetsov, V. A. (2010). Projection of gene-protein networks to the functional space of the proteome and its application to analysis of organism complexity. BMC Genomics, 11(Suppl 1), S4.

Kaneko, S. (2004). Module Shuffling. Protein Engineering, 22–34.

Kawashima, T. et. al. (2009). Domain shuffling and the evolution of vertebrates. Genome Research, 19(8), 1393–1403. doi:10.1101/gr.087072.108.

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Voss, S. R., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. Scientific Reports, 5(1).

Keshelava, A., Solis, G. P., Hersch, M., Koval, A., Kryuchkov, M., Bergmann, S., & Katanaev, V. L. (2018). High capacity in G protein-coupled receptor signaling. Nature Communications, 9(1). doi:10.1038/s41467-018-02868-y

King, N. (2004). The Unicellular Ancestry of Animal Development. Developmental Cell, 7(3), 313–325.

Kiss, Eniko & Hegedis, Botond & Varga, Torda & Merényi, Zsolt & Koszo, Tamas & Balint, Balazs & Prasanna, Arun & Krizsan, Krisztina & Riquelme, Meritxell & Takeshita, Norio & G. Nagy, Laszlo. (2019). Comparative genomics reveals the origin of fungal hyphae and multicellularity. 10.1101/546531.

Kisseleva, T., Bhattacharya, S., Braunstein, J., & Schindler, C. . (2002). Signaling through the JAK/STAT pathway, recent advances and future challenges. Gene, 285(1-2), 1–24. doi:10.1016/s0378-1119(02)00398-0

Knoll, A.H., 2003, The geological consequences of evolution: Geobiology, v.1, p. 3-14.

Knoll, A. H., Javaux, E. J., Hewitt, D., and Cohen, P., 2006, Eukaryotic organisms in Proterozoic oceans: Philosophical Transactions of the Royal Society London Series B, v. 361, p. 1023-1038.

Knoll, A. H. (2011). The multiple origins of complex multicellularity. *Annu. Rev. Earth Planet. Sci.* 39, 217–239.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. Annual Review of Genetics, 39(1), 309–338.

Koonin, E. V. (2010). The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biology, 11(5), 209. doi:10.1186/gb-2010-11-5-209

Koonin, E. V. (2016). Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1678), 20140333. doi:10.1098/rstb.2014.0333

Koster, M. J. E., Snel, B., & Timmers, H. T. M. (2015). Genesis of Chromatin and Transcription Dynamics in the Origin of Species. Cell, 161(4), 724–736. doi:10.1016/j.cell.2015.04.033

Kragl, M., Knapp, D., Nacu, E., Khattak, S., Maden, M., Epperlein, H. H., & Tanaka, E. M. (2009). Cells keep a memory of their tissue origin during axolotl limb regeneration. Nature, 460(7251), 60–65.

Krishnan, A., Mustafa, A., Almén, M. S., Fredriksson, R., Williams, M. J., & Schiöth, H. B. (2015). Evolutionary hierarchy of vertebrate-like heterotrimeric G protein families. Molecular Phylogenetics and Evolution, 91, 27–40. doi:10.1016/j.ympev.2015.05.009

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., Ioannidis, P., Zdobnov, E. M. (2014). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Research, 43(D1), D250–D256.

Kumar, S. et. al. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution, 34(7), 1812–1819. doi:10.1093/molbev/msx116

Kummerfeld, S. K., & Teichmann, S. A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. Trends in Genetics, 21(1), 25–30. doi:10.1016/j.tig.2004.11.007

Labbe, E., Letamendia, A., & Attisano, L. (2000). Association of Smads with lymphoid enhancer binding factor 1/T cell-specific factor mediates cooperative signaling by the transforming growth factor-beta and Wnt pathways. Proceedings of the National Academy of Sciences, 97(15), 8358–8363. doi:10.1073/pnas.150152697

Lang, R., Hammer, M., & Mages, J. (2006). DUSP Meet Immunology: Dual Specificity MAPK Phosphatases in Control of the Inflammatory Response. The Journal of Immunology, 177(11), 7497–7504. doi:10.4049/jimmunol.177.11.7497

Lang D, Weiche B, Timmerhaus G, Richardt S, Riano-Pachon DM, Correa LG, Reski R, Mueller-Roeber B, Rensing SA. (2010). Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. Genome Biol Evol. 2:488–503.

Lang D., Rensing S. (2015) The Evolution of Transcriptional Regulation in the Viridiplantae and its Correlation with Morphological Complexity. In: Ruiz-Trillo I., Nedelcu A. (eds) Evolutionary Transitions to Multicellular Life. Advances in Marine Genomics, vol 2. Springer, Dordrecht.

Lappano, R., & Maggiolini, M. (2011). G protein-coupled receptors: novel targets for drug discovery in cancer. Nature Reviews Drug Discovery, 10(1), 47–60. doi:10.1038/nrd3320

Laviola, L., Natalicchio, A., & Giorgino, F. (2007). The IGF-I Signaling Pathway. Current Pharmaceutical Design, 13(7), 663–669. doi:10.2174/138161207780249146

Lespinet, O., Wolf, Y. I., Koonin, E. V., & Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome research*, *12*(7), 1048–1059. doi:10.1101/gr.174302.

Levine, M., Cattoglio, C., and Tjian, R. 2014. Looping back to leap forward: transcription enters a new era. Cell 157: 13– 25.

Liongue, C., Sertori, R., & Ward, A. C. (2016). Evolution of Cytokine Receptor Signaling. The Journal of Immunology, 197(1), 11–18. doi:10.4049/jimmunol.1600372

Liu M, Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryote genomes: evidence of exon shuffling? Trends Genet 20:399–403.

Liu, Y., Shepherd, E. G., & Nelin, L. D. (2007). MAPK phosphatases — regulating the immune response. Nature Reviews Immunology, 7(3), 202–212. doi:10.1038/nri2035

Lloberas, J., Valverde-Estrella, L., Tur, J., Vico, T., & Celada, A. (2016). Mitogen-Activated Protein Kinases and Mitogen Kinase Phosphatase 1: A Critical Interplay in Macrophage Biology. Frontiers in Molecular Biosciences, 3. doi:10.3389/fmolb.2016.00028

Lobo, F. P., Rodrigues, M. R., Rodrigues, G. O., Hilário, H. O., Souza, R. A., Tauch, A., Miyoshi, A., Franco, G. C., Azevedo, V., Franco, G. R. (2012). KOMODO: a web tool for detecting and visualizing biased distribution of groups of homologous genes in monophyletic taxa. Nucleic acids research, 40(Web Server issue), W491–W497. doi:10.1093/nar/gks490

Loftus JC, Smith JW, Ginsberg MH. (1994). Integrin-mediated cell adhesion: the extracellular face. J Biol Chem 269:25235-25238.

Lone SA, Manohar S (2018). Saprolegnia parasitica, a lethal oomycete pathogen: demands to be controlled. J. Inf. Mol. Biol. 6(2): 36-44.

Long M, Betran E, Thornton K, Wang W. (2003). The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.

Lopes Cardoso, D., & Sharpe, C. (2017). Relating protein functional diversity to cell type number identifies genes that determine dynamic aspects of chromatin organisation as potential contributors to organismal complexity. PLOS ONE, 12(9), e0185409.

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thorton, J., Tramontano, A. (2009). Protein function annotation by homology-based inference. Genome Biology, 10(2), 207. doi:10.1186/gb-2009-10-2-207

Low, H. B., & Zhang, Y. (2016). Regulatory Roles of MAPK Phosphatases in Cancer. Immune Network, 16(2), 85. doi:10.4110/in.2016.16.2.85

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences, 104(Supplement 1), 8597–8604.

MacNamara, S., Baker, R. E., & Maini, P. K. (2011). Distinguishing graded and ultrasensitive signalling cascade kinetics by the shape of morphogen gradients in Drosophila. Journal of Theoretical Biology, 285(1), 136–146. doi:10.1016/j.jtbi.2011.06.012

Malarkey, C. S., & Churchill, M. E. A. (2012). The high mobility group box: the ultimate utility player of a cell. Trends in Biochemical Sciences, 37(12), 553–562. doi:10.1016/j.tibs.2012.09.003

Mallo, M., Wellik, D. M., & Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. Developmental Biology, 344(1), 7–15. doi:10.1016/j.ydbio.2010.04.024

Marsden, R. L., Ranea, J. A. G., Sillero, A., Redfern, O., Yeats, C., Maibaum, M., Lee, D., Addou, S., Reeves, G, A., Dallman, T. J., Orengo, C. A. (2006). Exploiting protein structure data to explore the evolution of protein function and biological complexity. Philosophical Transactions of the Royal Society B: Biological Sciences, 361(1467), 425–440.

Maynard Smith, J., Szathmary, E. (1995). The Major Transitions in Evolution. New York: W. H. Freeman and Company, 346 pp. ISBN 0-7167-4525-9. W. H. Freeman and Company, 41 Madison Ave., E. 26th, 35th Floor, New York, NY 10010, USA.

McShea, D. W. (1996). Perspective metazoan complexity and evolution: is there trend? Evolution, 50(2), 477–492.

McShea, D. W. (2005). The evolution of complexity without natural selection, a possible large-scale trend of the fourth kind. Paleobiology, 31(sp5), 146–156.

Medina, E., Córdova, C., Villalobos, P., Reyes, J., Komives, E. A., Ramírez-Sarmiento, C. A., & Babul, J. (2016). Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1. Biophysical Journal, 110(11), 2349–2360. doi:10.1016/j.bpj.2016.04.043

Mei, K and Guo, W. (2018). The exocyst complex. Current Biology 28, R909–R930

Meng, Z., Moroishi, T., & Guan, K.-L. (2016). Mechanisms of Hippo pathway regulation. Genes & Development, 30(1), 1–17. doi:10.1101/gad.274027.115

Meyer, A., & Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current Opinion in Cell Biology, 11(6), 699–704.

Merabet, S., & Galliot, B. (2015). The TALE face of Hox proteins in animal evolution. Frontiers in Genetics, 6. doi:10.3389/fgene.2015.00267

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., … McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis . Nature, 403(6771), 785–789. doi:10.1038/35001608

Middelkoop, T.C., and Korswagen, H.C. Development and migration of the C. elegans Q neuroblasts and their descendants (October 15, 2014), WormBook, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.173.1, http://www .wormbook.org

Miller, S. M. (2010) Volvox, Chlamydomonas, and the Evolution of Multicellularity. *Nature Education* 3(9):65.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. d., Chang, H-Y, El-Gebali, S., Fraser, M., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Padurangan, A. P., Paysan-lafosse, T., Pesseat, S., Ptter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sirist, C. J. A., Sellitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S-Y., Finn, R. D. (2018). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research. doi:10.1093/nar/gky1100

Moore, A. D., Björklund, Å. K., Ekman, D., Bornberg-Bauer, E., & Elofsson, A. (2008). Arrangements in the modular evolution of proteins. Trends in Biochemical Sciences, 33(9), 444–451.

Moran, Y., Barzilai, M. G., Liebeskind, B. J., & Zakon, H. H. (2015). Evolution of voltage-gated ion channels at the emergence of Metazoa. Journal of Experimental Biology, 218(4), 515–525. doi:10.1242/jeb.110270

Munoz-Torres, M., & Carbon, S. (2016). Get GO! Retrieving GO Data Using AmiGO, QuickGO, API, Files, and Tools. The Gene Ontology Handbook, 149–160. doi:10.1007/978-1-4939-3743-1_11

Mushegian, A R, and E V Koonin. “A minimal gene set for cellular life derived by comparison of complete bacterial genomes.” Proceedings of the National Academy of Sciences of the United States of America vol. 93,19 (1996): 10268-73. doi:10.1073/pnas.93.19.10268.

Mushegian, A., Gurevich, V. V., & Gurevich, E. V. (2012). The Origin and Evolution of G Protein-Coupled Receptor Kinases. PLoS ONE, 7(3), e33806. doi:10.1371/journal.pone.0033806

Nakagawa, S., Gisselbrecht, S. S., Rogers, J. M., Hartl, D. L., & Bulyk, M. L. (2013). DNA-binding specificity changes in the evolution of forkhead transcription factors. Proceedings of the National Academy of Sciences, 110(30), 12349–12354. doi:10.1073/pnas.1310430110

Nakanoh, S., & Agata, K. (2019). Evolutionary view of pluripotency seen from early development of non-mammalian amniotes. Developmental Biology. doi:10.1016/j.ydbio.2019.04.014

Nam, H.-J., Kim, I., Bowie, J. U., & Kim, S. (2015). Metazoans evolved by taking domains from soluble proteins to expand intercellular communication network. Scientific Reports, 5(1).

Nasir, A., Kim, K. M., & Caetano-Anollés, G. (2014). Global Patterns of Protein Domain Gain and Loss in Superkingdoms. PLoS Computational Biology, 10(1), e1003452. doi:10.1371/journal.pcbi.1003452

Nelson, David L.; COX, Michael M. Princípios de bioquímica de Lehninger. Porto Alegre: Artmed, 2011. 6. ed. Porto Alegre: Artmed, 2014.

Niklas, K. J. (2014). The evolutionary-developmental origins of multicellularity. American Journal of Botany, 101(1), 6–25.

Niklas, K. J., & Newman, S. A. (2013). The origins of multicellular organisms. Evolution & Development, 15(1), 41–52. doi:10.1111/ede.12013

Niklas, K. J., Cobb, E. D., & Dunker, A. K. (2014). The number of cell types, information content, and the evolution of complex multicellularity. Acta Societatis Botanicorum Poloniae, 83(4), 337–347.

Niklas, K. J., Dunker, A. K., & Yruela, I. (2018). The evolutionary origins of cell type diversification and the role of intrinsically disordered proteins. Journal of Experimental Botany, 69(7), 1437–1446.

Ohno S. (1970). Evolution by gene duplication. New York: Springer-Verlag.

Oruganty, K., & Kannan, N. (2012). Design principles underpinning the regulatory diversity of protein kinases. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1602), 2529–2539. doi:10.1098/rstb.2012.0015

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research.

Oulion, S., Bertrand, S., & Escriva, H. (2012). Evolution of the FGF Gene Family. International Journal of Evolutionary Biology, 2012, 1–12. doi:10.1155/2012/298147

Özbek, S., Balasubramanian, P. G., Chiquet-Ehrismann, R., Tucker, R. T., & Adams, J. (2010). The Evolution of Extracellular Matrix. Molecular biology of the cell. 21. 4300-5. 10.1091/mbc.E10-03-0251.

Paradis E. & Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526-528.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23(9), 1061–1067.

Pastuzyn, E. D., Day, C. E., Kearns, R. B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., Shepherd, J. D. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell, 172(1-2), 275–288.e18. doi:10.1016/j.cell.2017.12.024

Patthy L (1987) Intron-dependent evolution: preferred types of exons and introns. FEBS Lett 214:1–7.

Patthy L (1996) Exon shuffling and other ways of module exchange. Matrix Biol 15:301–310.

Patthy L. (1999). Genome evolution and the evolution of exon-shuffling—A review. Gene 238: 103–114.

Payne, S. (2017). Virus Interactions With the Cell. Viruses, 23–35.

Pawson, T., & Scott, J. D. (2005). Protein phosphorylation in signaling – 50 years and counting. Trends in Biochemical Sciences, 30(6), 286–290. doi:10.1016/j.tibs.2005.04.013

Pennell MW, Harmon LJ, Uyeda JC (2014). Is there room for punctuated equilibrium in macroevolution? Trends Ecol Evol 29(1):23–32. doi:10.1016/j.tree.2013.07.004

Petryniak, B., Staudt, L. M., Postema, C. E., McCormack, W. T., & Thompson, C. B. (1990). Characterization of chicken octamer-binding proteins demonstrates that POU domain-containing homeobox transcription factors have been highly conserved during vertebrate evolution. Proceedings of the National Academy of Sciences, 87(3), 1099–1103. doi:10.1073/pnas.87.3.1099

Phillips, T. & Shaw, K. (2008) Chromatin Remodeling in Eukaryotes. Nature Education 1(1):209

Plotnikov, A., Zehorai, E., Procaccia, S., & Seger, R. (2011). The MAPK cascades: Signaling components, nuclear roles and mechanisms of nuclear translocation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1813(9), 1619–1633. doi:10.1016/j.bbamcr.2010.12.012

Pollard, T. D., & Borisy, G. G. (2003). Cellular Motility Driven by Assembly and Disassembly of Actin Filaments. Cell, 112(4), 453–465. doi:10.1016/s0092-8674(03)00120-x

Pollard, T. D., & Goldman, R. D. (2018). Overview of the Cytoskeleton from an Evolutionary Perspective. Cold Spring Harbor Perspectives in Biology, 10(7), a030288. doi:10.1101/cshperspect.a030288

Preisner, H., Karin, E. L., Poschmann, G., Stühler, K., Pupko, T., & Gould, S. B. (2016). The Cytoskeleton of Parabasalian Parasites Comprises Proteins that Share Properties Common to Intermediate Filament Proteins. Protist, 167(6), 526–543. doi:10.1016/j.protis.2016.09.001

Preisner, H., Habicht, J., Garg, S. G., & Gould, S. B. (2018). Intermediate filament protein evolution and protists. Cytoskeleton, 75(6), 231–243. doi:10.1002/cm.21443

Pruitt K, Brown G, Murphy M. (2010). RefSeq Frequently Asked Questions (FAQ). In: RefSeq Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.Available from: https://www.ncbi.nlm.nih.gov/books/NBK50679/

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Research, 33(Web Server), W116–W120. doi:10.1093/nar/gki442

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing. ISBN: 3-900051-07-0.

Reeves, R. (2015). High mobility group (HMG) proteins: Modulators of chromatin structure and DNA repair in mammalian cells. DNA Repair, 36, 122–136. doi:10.1016/j.dnarep.2015.09.015

Richardson, E. J., & Watson, M. (2012). The automatic annotation of bacterial genomes. Briefings in Bioinformatics, 14(1), 1–12.

Rimer, J., Cohen, I. R., & Friedman, N. (2014). Do all creatures possess an acquired immune system of some sort? BioEssays, 36(3), 273–281. doi:10.1002/bies.201300124

Ritschard, E. A., Fitak, R. R., Simakov, O., & Johnsen, S. (2019). Genomic signatures of G-protein-coupled receptor expansions reveal functional transitions in the evolution of cephalopod signal transduction. Proceedings of the Royal Society B: Biological Sciences, 286(1897), 20182929. doi:10.1098/rspb.2018.2929

Rogers, K. W., & Schier, A. F. (2011). Morphogen Gradients: From Generation to Interpretation. Annual Review of Cell and Developmental Biology, 27(1), 377–407. doi:10.1146/annurev-cellbio-092910-154148

Rokas, A. (2008). The molecular origins of multicellular transitions. Current Opinion in Genetics & Development, 18(6), 472–478.

Ronch, E., Treisman, J., Dostatni, N., Struhl, G., & Desplan, C. (1993). Down-regulation of the Drosophila morphogen bicoid by the torso receptor-mediated signal transduction cascade. Cell, 74(2), 347–355. doi:10.1016/0092-8674(93)90425-p

Rosanova, A., Colliva, A., Osella, M., & Caselle, M. (2017). Modelling the evolution of transcription factor binding preferences in complex eukaryotes. Scientific Reports, 7(1). doi:10.1038/s41598-017-07761-0

Rosenbaum DM, Rasmussen SGF, Kobilka BK. The structure and function of G-protein-coupled receptors. Nature. 2009;459:356–363.

Rozengurt, E., Sinnett-Smith, J., & Kisfalvi, K. (2010). Crosstalk between Insulin/Insulin-like Growth Factor-1 Receptors and G Protein-Coupled Receptor Signaling Systems: A Novel Target for the Antidiabetic Drug Metformin in Pancreatic Cancer. Clinical Cancer Research, 16(9), 2505–2511. doi:10.1158/1078-0432.ccr-09-2229

Rudman, S. M., Philpott, M. P., Thomas, G. A., & Kealey, T. (1997). The Role of IGF-I in Human Skin and its Appendages: Morphogen as Well as Mitogen? Journal of Investigative Dermatology, 109(6), 770–777. doi:10.1111/1523-1747.ep12340934

Sagner, A., & Briscoe, J. (2017). Morphogen interpretation: concentration, time, competence, and signaling dynamics. Wiley Interdisciplinary Reviews: Developmental Biology, 6(4), e271. doi:10.1002/wdev.271

Salojin, K., & Oravecz, T. (2007). Regulation of innate immunity by MAPK dual-specificity phosphatases: knockout models reveal new tricks of old genes. Journal of Leukocyte Biology, 81(4), 860–869. doi:10.1189/jlb.1006639

Saxton, R. A., & Sabatini, D. M. (2017). mTOR Signaling in Growth, Metabolism, and Disease. Cell, 168(6), 960–976. doi:10.1016/j.cell.2017.02.004

Schad E, Tompa P, Hegyi H. (2011). The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 12: R120.

Schopf, J. W. (1993). Microfossils of the Archean apex chert: new evidence of the antiquity of life. Science 60:640–46

Scita, G., & Di Fiore, P. P. (2010). The endocytic matrix. Nature, 463(7280), 464–473. doi:10.1038/nature08910

Sebe-Pedros, A., Roger, A. J., Lang, F. B., King, N., & Ruiz-Trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. Proceedings of the National Academy of Sciences, 107(22), 10142–10147. doi:10.1073/pnas.1002257107

Sebé-Pedrós, A., Zheng, Y., Ruiz-Trillo, I., & Pan, D. (2012). Premetazoan Origin of the Hippo Signaling Pathway. Cell Reports, 1(1), 13–20. doi:10.1016/j.celrep.2011.11.004

Sebé-Pedrós, A., Irimia, M., del Campo, J., Parra-Acero, H., Russ, C., Nusbaum, C., Blencowe, B. J., Ruiz-Ttrillo, I. (2013). Regulated aggregative multicellularity in a close unicellular relative of metazoan. Genomics and evolutionary biology. *eLife* **2**, e01287.

Sebé-Pedrós, A. & de Mendoza, A. (2015) in *Evolutionary Transitions to Multicellular Life* Vol. 2 Springer (eds Ruiz-Trillo, I. & Nedelcu, A. M.) 379–394.

Sebe-Pedros, A., Roger, A. J., Lang, F. B., King, N., & Ruiz-Trillo, I. (2010). Ancient origin of the integrin-mediated adhesion and signaling machinery. Proceedings of the National Academy of Sciences, 107(22), 10142–10147. doi:10.1073/pnas.1002257107

Sebé-Pedrós, A., Chomsky, E., Pang, K., Lara-Astiaso, D., Gaiti, F., Mukamel, Z., Amit, I., Hejnol, A., Degnan, B. M., Tanay, A. (2018). Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nature Ecology & Evolution, 2(7), 1176–1188.

Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: a unicellular perspective. Nature Reviews Genetics, 18(8), 498–512. doi:10.1038/nrg.2017.21

Sebé-Pedrós, A., Chomsky, E., Pang, K., Lara-Astiaso, D., Gaiti, F., Mukamel, Z., … Tanay, A. (2018). Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nature Ecology & Evolution, 2(7), 1176–1188. doi:10.1038/s41559-018-0575-6

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Gene Prediction, 227–245. doi:10.1007/978-1-4939-9173-0_14

Seternes, O.-M., Kidger, A. M., & Keyse, S. M. (2019). Dual-specificity MAP kinase phosphatases in health and disease. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1866(1), 124–143. doi:10.1016/j.bbamcr.2018.09.002

Shvartsman, S. Y., Coppey, M., & Berezhkovskii, A. M. (2008). Dynamics of maternal morphogen gradients in Drosophila. Current Opinion in Genetics & Development, 18(4), 342–347. doi:10.1016/j.gde.2008.06.002

Sigismund, S., Confalonieri, S., Ciliberto, A., Polo, S., Scita, G., & Di Fiore, P. P. (2012). Endocytosis and Signaling: Cell Logistics Shape the Eukaryotic Cell Plan. Physiological Reviews, 92(1), 273–366. doi:10.1152/physrev.00005.2011

Simão, F. A., Waterhouse, R. M., Ioanidis, P., Kriventseva, E. V, Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. doi:10.1093/bioinformatics/btv351

Skorokhod, A., Gamulin, V., Gundacker, D., Kavsan, V., Muller, I. M., & Muller, W. E. G. (1999). Origin of Insulin Receptor-Like Tyrosine Kinases in Marine Sponges. The Biological Bulletin, 197(2), 198–206. doi:10.2307/1542615

Sondek, J., & Siderovski, D. P. (2001). Gγ-like (ggl) domains: new frontiers in g-protein signaling and β-propeller scaffolding. Biochemical Pharmacology, 61(11), 1329–1337. doi:10.1016/s0006-2952(01)00633-5

Sonnhammer, E. L., Eddy, S. R., Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins, 28, 405–420.

Solé, R. V., & Duran-Nebreda, S. (2015). In Silico Transitions to Multicellularity. Advances in Marine Genomics, 245–266. doi:10.1007/978-94-017-9642-2_13

Stavrou, S., & Ross, S. R. (2015). APOBEC3 Proteins in Viral Immunity. The Journal of Immunology, 195(10), 4565–4570.

Suga, H., Dacre, M., de Mendoza, A., Shalchian-Tabrizi, K., Manning, G., & Ruiz-Trillo, I. (2012). Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic

Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases. Science Signaling, 5(222), ra35–ra35. doi:10.1126/scisignal.2002733

Sugiura, R., Satoh, R., Ishiwata, S., Umeda, N., & Kita, A. (2011). Role of RNA-Binding Proteins in MAPK Signal Transduction Pathway. Journal of Signal Transduction, 2011, 1–8. doi:10.1155/2011/109746

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS ONE, 6(7), e21800.

Synek, L., Sekereš, J., & Žárský, V. (2014). The exocyst at the interface between cytoskeleton and membranes in eukaryotic cells. Frontiers in Plant Science, 4. doi:10.3389/fpls.2013.00543

Szathmary E, Jordan F, Pal C (2001) Molecular biology and evolution. Can genes explain biological complexity? Science 292(5520):1315–1316.

Teoh, C. M., Tam, J. K. C., & Tran, T. (2012). Integrin and GPCR Crosstalk in the Regulation of ASM Contraction Signaling in Asthma. Journal of Allergy, 2012, 1–9. doi:10.1155/2012/341282

The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD). (2013). National Center for Biotechnology Information (US); Available from: https://www.ncbi.nlm.nih.gov/books/NBK143764/

Thomas, P. D. (2017). The Gene Ontology and the Meaning of Biological Function. The Gene Ontology Handbook, 15–24.

Tong, K., Wang, Y., & Su, Z. (2017). Phosphotyrosine signalling and the origin of animal multicellularity. Proceedings of the Royal Society B: Biological Sciences, 284(1860), 20170681. doi:10.1098/rspb.2017.0681

Trigos, A. S., Pearson, R. B., Papenfuss, A. T., Goode D. L. (2019). Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer. eLife 2019;8:e40947 DOI: 10.7554/eLife.40947

Tuteja N. (2009). Signaling through G protein coupled receptors. Plant signaling & behavior, 4(10), 942–947. doi:10.4161/psb.4.10.9530

Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L., & Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends in Plant Science, 14(12), 680–688.

van Dam, T. J. P., Zwartkruis, F. J. T., Bos, J. L., & Snel, B. (2011). Evolution of the TOR Pathway. Journal of Molecular Evolution, 73(3-4), 209–220. doi:10.1007/s00239-011-9469-9

van Valen, L (1973). "A new evolutionary law". Evolutionary Theory. 1: 1–30.

Veeckman, E., Ruttink, T., & Vandepoele, K. (2016). Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. The Plant Cell, 28(8), 1759–1768.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., & Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. Current Opinion in Structural Biology, 14(2), 208–216.

Vogel, C., & Chothia, C. (2006). Protein Family Expansions and Biological Complexity. PLoS Computational Biology, 2(5), e48. doi:10.1371/journal.pcbi.0020048

Vonk, P. J., & Ohm, R. A. (2018). The role of homeodomain transcription factors in fungal development. Fungal Biology Reviews. doi:10.1016/j.fbr.2018.04.002

Wang, Z. (2016). Transactivation of Epidermal Growth Factor Receptor by G Protein-Coupled Receptors: Recent Progress, Challenges and Future Research. International Journal of Molecular Sciences, 17(1), 95. doi:10.3390/ijms17010095

Wang, Y., & Levy, D. E. (2006). C. elegans STAT: evolution of a regulatory switch. The FASEB Journal, 20(10), 1641–1652. doi:10.1096/fj.06-6051com

Wang, Z., Zarlenga, D., Martin, J., Abubucker, S., & Mitreva, M. (2012). Exploring metazoan evolution through dynamic and holistic changes in protein families and domains. BMC Evolutionary Biology, 12(1), 138. doi:10.1186/1471-2148-12-138

Waterhouse, R. M., Zdobnov, E. M., Kriventseva, E. V. (2011). Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi, *Genome Biology and Evolution*, Volume 3, Pages 75–86

Waterhouse, R. M., Tegenfeldt, F. Li, J. Zdobnov, E. M., Kriventseva, E. V. (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs, *Nucleic Acids Research*, Volume 41, Issue D1, Pages D358–D365.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kristentseva, E. V., Zdobnov, E. M. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution, 35(3), 543–548.

Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'Souza, S., Woodruff, R. V., & Walker, G. C. (2009). Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. Microbiology and Molecular Biology Reviews, 73(1), 134–154. doi:10.1128/mmbr.00034-08.

Wennerberg, K. (2005). The Ras superfamily at a glance. Journal of Cell Science, 118(5), 843–846. doi:10.1242/jcs.01660

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wideman, J. G., Leung, K. F., Field, M. C., & Dacks, J. B. (2014). The Cell Biology of the Endocytic System from an Evolutionary Perspective. Cold Spring Harbor Perspectives in Biology, 6(4), a016998–a016998. doi:10.1101/cshperspect.a016998

Wilkie, T. M., & Kinch, L. (2005). New Roles for Gα and RGS Proteins: Communication Continues despite Pulling Sisters Apart. Current Biology, 15(20), R843–R854. doi:10.1016/j.cub.2005.10.008

Willyard, C. (2018). New human gene tally reignites debate. Nature, 558(7710), 354–355.u, Q., & Dunbrack, R. L. (2012). Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. Bioinformatics, 28(21), 2763–2772. doi:10.1093/bioinformatics/bts533

Yaeger, L. S. (2009). How evolution guides complexity. HFSP Journal, 3(5), 328–339.

Yruela I., Oldfield C. J., Niklas K. J., Dunker A. K. (2017). Evidence for a strong correlation between transcription factor protein disorder and organismic complexity. *Genome Biol. Evol.* 91248–1265. 10.1093/gbe/evx073.

Yakar, S., Rosen, C. J., Beamer, W. G., Ackert-Bicknell, C. L., Wu, Y., Liu, J.-L., … LeRoith, D. (2002). Circulating levels of IGF-1 directly regulate bone growth and density. Journal of Clinical Investigation, 110(6), 771–781. doi:10.1172/jci15463

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329–342.

Yekutieli, D., & Benjamini, Y. (2001). under dependency. The Annals of Statistics, 29(4), 1165–1188.

Zar J.H. (1999). Biostatistical Analysis. 4th edition. Prentice-Hall Englewood Cliffs, New Jersey. 929 p.

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17(9), 847–848.

Zhang, F., Yang, H., Wang, Z., Mergler, S., Liu, H., Kawakita, T., Reinach, P. S. (2007). Transient receptor potential vanilloid 1 activation induces inflammatory cytokine release in corneal epithelium through MAPK signaling. Journal of Cellular Physiology, 213(3), 730–739. doi:10.1002/jcp.21141

Zhu, H. (2016). Forkhead box transcription factors in embryonic heart development and congenital heart disease. Life Sciences, 144, 194–201. doi:10.1016/j.lfs.2015.12.001

10. Anexos

Tabela Suplementar 1 - Domínios Pfam significativamente associados ao aumento da complexidade em *Eukarya*

Pfam ID	description	Spearman correlation	Spearman q-value	PIC q-value	size
PF08686	PLAC (protease and lacunin) domain	0,90	8,5E-13	4,0E-02	151
PF01034	Syndecan domain	0,88	4,8E-12	4,5E-04	92
PF07525	SOCS box	0,87	1,0E-11	1,0E-02	508
PF00880	Nebulin repeat	0,87	1,2E-11	1,9E-04	1073
PF00007	Cystine-knot domain	0,87	1,4E-11	7,6E-08	160
PF16493	N-terminal of Homeobox Meis and PKNOX1	0,86	3,6E-11	1,9E-03	81
PF00157	Pou domain - N-terminal to homeobox domain	0,86	3,7E-11	7,8E-05	229
PF00047	Immunoglobulin domain	0,86	3,9E-11	3,0E-09	775
PF00418	Tau and MAP protein, tubulin-binding repeat	0,86	4,7E-11	1,6E-03	210
PF10034	Q-cell neuroblast polarisation	0,85	7,2E-11	3,6E-09	78
PF01839	FG-GAP repeat	0,85	8,5E-11	3,1E-02	477
PF16526	C-terminal leucine zipper domain of cyclic nucleotide-gated channels	0,85	9,4E-11	1,1E-04	69
PF12605	Casein kinase 1 gamma C terminal	0,84	1,2E-10	5,2E-09	42
PF01421	Reprolysin (M12B) family zinc metalloprotease	0,84	1,4E-10	8,1E-04	572
PF05110	AF-4 proto-oncoprotein	0,84	1,8E-10	8,8E-05	75
PF16564	p55-binding region of Methyl-CpG-binding domain proteins MBD	0,84	2,0E-10	1,3E-11	50
PF01390	SEA domain	0,84	2,1E-10	3,1E-07	439
PF00094	von Willebrand factor type D domain	0,84	2,3E-10	1,2E-03	954
PF07531	NHR1 homology to TAF	0,84	2,3E-10	3,1E-02	83
PF08516	ADAM cysteine-rich	0,83	3,4E-10	1,6E-04	271
PF01835	MG2 domain	0,83	3,7E-10	2,9E-02	194
PF09316	C-myb, C-terminal	0,83	3,9E-10	3,2E-04	40
PF00836	Stathmin family	0,83	4,3E-10	1,6E-02	73
PF14709	double strand RNA binding domain from DEAD END PROTEIN 1	0,83	4,7E-10	1,3E-02	69
PF12347	Holliday junction regulator protein family C-terminal repeat	0,83	5,0E-10	6,7E-07	64
PF01562	Reprolysin family propeptide	0,83	5,2E-10	2,8E-02	589
PF16516	Leucine zipper of domain CC2 of NEMO, NF-kappa-B essential modulator	0,83	5,4E-10	1,6E-05	57
PF16885	Voltage-gated calcium channel subunit alpha, C-term	0,82	1,0E-09	3,0E-06	48
PF16034	JAKMIP CC3 domain	0,82	1,1E-09	1,0E-04	40
PF08742	C8 domain	0,82	1,1E-09	7,1E-05	747
PF11261	Interferon regulatory factor 2-binding protein zinc finger	0,82	1,2E-09	2,3E-04	44
PF06021	Aralkyl acyl-CoA:amino acid N-acyltransferase	0,82	1,3E-09	4,2E-02	42
PF05361	PKC-activated protein phosphatase-1 inhibitor	0,81	1,7E-09	1,5E-04	63
PF12884	Transducer of regulated CREB activity, N terminus	0,81	1,8E-09	4,1E-03	40
PF00681	Plectin repeat	0,81	2,0E-09	2,9E-08	861
PF02946	GTF2I-like repeat	0,81	2,4E-09	8,1E-08	158

PF03045	DAN domain	0,81	2,5E-09	2,2E-02	86
PF00079	Serpin (serine protease inhibitor)	0,81	2,8E-09	1,8E-03	627
PF10569	Alpha-macro-globulin thiol-ester bond-forming region	0,81	3,0E-09	1,6E-02	193
PF12714	TILa domain	0,80	3,1E-09	1,8E-03	223
PF12925	E2 domain of amyloid precursor protein	0,80	3,1E-09	7,9E-07	49
PF11878	Domain of unknown function (DUF3398)	0,80	3,5E-09	2,8E-02	87
PF01108	Tissue factor	0,80	3,6E-09	8,0E-04	152
PF00041	Fibronectin type III domain	0,80	3,7E-09	1,1E-03	8981
PF02196	Raf-like Ras-binding domain	0,80	4,3E-09	2,2E-05	129
PF03185	Calcium-activated potassium channel, beta subunit	0,80	4,7E-09	2,4E-04	64
PF02060	Slow voltage-gated potassium channel	0,80	4,8E-09	8,2E-08	51
PF16207	RAWUL domain RING finger- and WD40-associated ubiquitin-like	0,80	5,4E-09	4,7E-02	112
PF13330	Mucin-2 protein WxxW repeating region	0,80	5,5E-09	1,0E-02	280
PF14048	C-terminal domain of methyl-CpG binding protein 2 and 3	0,79	6,8E-09	1,6E-12	52
PF02038	ATP1G1/PLM/MAT8 family	0,79	7,0E-09	8,5E-06	72
PF09294	Interferon-alpha/beta receptor, fibronectin type III	0,79	7,1E-09	3,0E-04	146
PF15229	POM121 family	0,79	7,2E-09	1,4E-10	82
PF02210	Laminin G domain	0,79	7,3E-09	3,9E-03	1471
PF01101	HMG14 and HMG17	0,79	7,8E-09	3,3E-04	61
PF01017	STAT protein, all-alpha domain	0,79	8,0E-09	4,9E-04	96
PF04621	PEA3 subfamily ETS-domain transcription factor N terminal domain	0,79	8,1E-09	4,1E-04	50
PF01023	S-100/ICaBP type calcium binding domain	0,79	8,4E-09	1,1E-02	233
PF08383	Maf N-terminal region	0,79	8,9E-09	4,3E-04	48
PF00357	Integrin alpha cytoplasmic region	0,79	9,3E-09	8,0E-06	71
PF10608	Polyubiquitination (PEST) N-terminal domain of MAGUK	0,79	9,4E-09	1,1E-02	61
PF12001	Domain of unknown function (DUF3496)	0,79	1,1E-08	1,2E-08	57
PF06462	Propeller	0,79	1,1E-08	3,0E-06	256
PF17226	MTA R1 domain	0,79	1,1E-08	1,3E-06	45
PF16159	FOXP coiled-coil domain	0,79	1,2E-08	1,8E-03	60
PF08688	Apx/Shroom domain ASD1	0,78	1,2E-08	1,9E-04	43
PF03529	Otx1 transcription factor	0,78	1,4E-08	2,6E-03	44
PF12045	Protein of unknown function (DUF3528)	0,78	1,4E-08	5,1E-04	46
PF15974	Cadherin C-terminal cytoplasmic tail, catenin-binding region	0,78	1,5E-08	9,3E-03	192
PF01821	Anaphylotoxin-like domain	0,78	1,8E-08	3,9E-06	80
PF00053	Laminin EGF domain	0,78	1,9E-08	1,4E-02	3680
PF09744	JNK_SAPK-associated protein-1	0,78	2,2E-08	5,6E-04	77
PF12901	SUZ-C motif	0,78	2,2E-08	2,6E-02	48
PF03821	Golgi 4-transmembrane spanning transporter	0,77	2,6E-08	5,3E-04	53
PF14707	C-terminal region of aryl-sulfatase	0,77	2,8E-08	2,6E-02	97
PF03066	Nucleoplasmin/nucleophosmin domain	0,77	2,8E-08	2,2E-03	51
PF08210	APOBEC-like N-terminal domain	0,77	2,8E-08	2,9E-03	81
PF02251	Proteasome activator pa28 alpha subunit	0,77	3,1E-08	6,2E-05	49
PF12414	Calcitonin gene-related peptide regulator C terminal	0,77	3,7E-08	5,1E-06	47
PF08474	Myelin transcription factor 1	0,77	3,8E-08	1,2E-05	43
PF14798	Calcium homeostasis modulator	0,77	4,0E-08	1,5E-02	86

PF07894	Protein of unknown function (DUF1669)	0,76	4,3E-08	6,2E-03	108
PF00340	Interleukin-1 / 18	0,76	5,1E-08	1,3E-03	69
PF16739	Caspase recruitment domain	0,76	5,2E-08	6,3E-04	65
PF06189	5'-nucleotidase	0,76	5,5E-08	1,5E-04	45
PF07679	Immunoglobulin I-set domain	0,76	5,5E-08	1,4E-05	8215
PF15010	Putative cell signalling	0,76	5,6E-08	2,7E-03	43
PF03285	Paralemmin	0,76	5,8E-08	3,5E-04	43
PF06484	Teneurin Intracellular Region	0,76	6,0E-08	4,2E-07	83
PF11851	Domain of unknown function (DUF3371)	0,76	6,0E-08	7,8E-03	49
PF01056	Myc amino-terminal region	0,75	9,9E-08	6,8E-04	66
PF06327	Domain of Unknown Function (DUF1053)	0,75	1,1E-07	5,2E-03	94
PF14988	Domain of unknown function (DUF4515)	0,75	1,1E-07	4,2E-02	51
PF00801	PKD domain	0,75	1,1E-07	3,0E-05	269
PF00965	Tissue inhibitor of metalloproteinase	0,75	1,2E-07	9,6E-06	70
PF16165	Ferlin C-terminus	0,74	1,4E-07	1,4E-02	93
PF00631	GGL domain	0,74	1,4E-07	4,6E-04	247
PF13281	Domain of unknown function (DUF4071)	0,74	1,5E-07	1,4E-04	42
PF04516	CP2 transcription factor	0,74	1,7E-07	4,2E-02	103
PF12886	Transducer of regulated CREB activity, C terminus	0,74	2,0E-07	2,9E-04	43
PF16274	Qua1 domain	0,73	2,5E-07	3,5E-04	41
PF10506	PDZ domain of MCC-2 bdg protein for Usher syndrome	0,73	3,7E-07	2,0E-03	41
PF12424	Plasma membrane calcium transporter ATPase C terminal	0,73	3,7E-07	4,3E-03	92
PF00250	Forkhead domain	0,72	4,5E-07	4,0E-02	819
PF02865	STAT protein, protein interaction domain	0,72	4,7E-07	2,8E-03	92
PF03414	Glycosyltransferase family 6	0,72	4,8E-07	1,7E-02	73
PF00200	Disintegrin	0,72	5,0E-07	8,6E-04	340
PF05831	GAGE protein	0,72	5,8E-07	5,0E-05	69
PF03148	Tektin family	0,71	8,7E-07	3,2E-02	109
PF11515	Mouse development and cellular proliferation protein Cullin-7	0,71	8,9E-07	4,4E-04	41
PF02189	Immunoreceptor tyrosine-based activation motif	0,71	9,3E-07	1,2E-02	55
PF15951	MITF/TFEB/TFEC/TFE3 N-terminus	0,71	1,1E-06	1,8E-02	56
PF14915	CCDC144C protein coiled-coil region	0,70	1,2E-06	1,1E-03	74
PF10390	RNA polymerase II elongation factor ELL	0,70	1,4E-06	4,8E-02	58
PF15870	ElonginA binding-protein 1	0,69	2,0E-06	2,4E-09	74
PF04906	Tweety	0,69	2,1E-06	1,2E-02	68
PF00261	Tropomyosin	0,69	2,3E-06	2,8E-03	104
PF04629	Islet cell autoantigen ICA69, C-terminal domain	0,69	2,8E-06	5,7E-03	42
PF02736	Myosin N-terminal SH3-like domain	0,68	3,1E-06	1,8E-03	281
PF15300	INTS6/SAGE1/DDX26B/CT45 C-terminus	0,68	3,2E-06	9,2E-03	58
PF16866	PHD-finger	0,68	3,7E-06	2,2E-03	47
PF15287	KRBA1 family repeat	0,68	4,0E-06	2,3E-09	49
PF16453	PH domain	0,68	4,2E-06	5,1E-05	53
PF09005	Domain of unknown function (DUF1897)	0,68	4,3E-06	6,7E-06	56
PF09815	XK-related protein	0,68	4,4E-06	1,1E-02	185
PF00207	Alpha-2-macroglobulin family	0,67	4,8E-06	6,4E-03	218
PF11357	Cell cycle regulatory protein	0,67	4,8E-06	3,1E-08	104

PF00711	Beta defensin	0,67	5,1E-06	2,8E-03	81
PF07145	Ataxin-2 C-terminal region	0,67	5,1E-06	6,6E-04	122
PF16954	Haem-transporter, endosomal/lysosomal, haem-responsive gene	0,67	5,7E-06	3,3E-07	41
PF10486	Phosphoinositide 3-kinase gamma adapter protein p101 subunit	0,66	7,9E-06	1,5E-02	49
PF02234	Cyclin-dependent kinase inhibitor	0,66	9,5E-06	2,5E-02	71
PF16652	Pleckstrin homology domain	0,65	1,2E-05	1,1E-02	50
PF15275	PEHE domain	0,65	1,2E-05	3,4E-02	51
PF13841	Beta defensin	0,65	1,3E-05	1,1E-06	154
PF13885	Keratin, high sulfur B2 protein	0,64	2,0E-05	1,7E-05	652
PF02257	RFX DNA-binding domain	0,64	2,3E-05	1,8E-02	143
PF06758	Repeat of unknown function (DUF1220)	0,63	3,2E-05	3,5E-23	264
PF01290	Thymosin beta-4 family	0,63	3,4E-05	1,8E-04	88
PF09514	SSXRD motif	0,63	3,6E-05	1,1E-03	45
PF14916	Coiled-coil domain of unknown function	0,63	3,7E-05	1,8E-05	48
PF06237	Protein of unknown function (DUF1011)	0,63	3,8E-05	5,1E-04	45
PF10457	Cholesterol-capturing domain	0,63	3,8E-05	1,1E-02	45
PF14672	Late cornified envelope	0,62	3,9E-05	2,9E-04	92
PF12440	Melanoma associated antigen family N terminal	0,62	4,1E-05	3,4E-02	175
PF11759	Keratin-associated matrix	0,62	4,2E-05	9,9E-03	90
PF03020	LEM domain	0,62	4,4E-05	1,4E-04	113
PF15371	Domain of unknown function (DUF4599)	0,62	4,6E-05	5,9E-13	49
PF14650	FAM75 family	0,62	4,8E-05	6,8E-03	92
PF05287	PMG protein	0,62	5,0E-05	2,9E-05	88
PF12417	Zinc finger protein	0,61	7,7E-05	6,7E-04	52
PF00568	WH1 domain	0,61	7,7E-05	6,0E-03	196
PF02218	Repeat in HS1/Cortactin	0,60	9,3E-05	5,9E-05	171
PF06409	Nuclear pore complex interacting protein (NPIP)	0,59	1,1E-04	2,1E-21	57
PF15309	ALMS motif	0,58	1,9E-04	3,3E-02	45
PF03024	Folate receptor family	0,57	2,3E-04	4,3E-04	121
PF00210	Ferritin-like domain	0,57	2,4E-04	8,0E-08	136
PF14642	FAM47 family	0,57	2,9E-04	9,6E-22	49
PF08976	EF-hand domain	0,56	3,3E-04	3,7E-04	50
PF04845	PurA ssDNA and RNA-binding protein	0,56	3,7E-04	3,1E-03	58
PF09011	HMG-box domain	0,54	5,8E-04	1,2E-02	167
PF05920	Homeobox KN domain	0,54	6,9E-04	2,6E-03	351
PF15070	Putative golgin subfamily A member 2-like protein 5	0,53	7,7E-04	7,0E-03	114
PF00641	Zn-finger in Ran binding protein and others	0,53	8,9E-04	1,2E-03	586
PF15288	Zinc knuckle	0,53	9,0E-04	5,1E-05	45
PF15240	Proline-rich	0,51	1,3E-03	4,8E-06	63
PF11938	TLR4 regulator and MIR-interacting MSAP	0,50	1,9E-03	2,9E-02	128
PF08953	Domain of unknown function (DUF1899)	0,50	2,1E-03	3,4E-02	158
PF08332	Calcium/calmodulin dependent protein kinase II association domain	0,50	2,2E-03	3,7E-04	77
PF02026	RyR domain	0,49	2,7E-03	1,8E-03	243
PF10239	Protein of unknown function (DUF2465)	0,48	3,7E-03	5,6E-06	54

PF03516	Filaggrin	0,46	4,9E-03	6,3E-11	58
PF08726	Ca2+ insensitive EF hand	0,46	5,0E-03	3,6E-02	98
PF13599	Pentapeptide repeats (9 copies)	0,46	5,2E-03	1,6E-03	47
PF14752	Retinol binding protein receptor	0,45	6,1E-03	4,8E-04	52
PF00535	Glycosyl transferase family 2	0,45	6,2E-03	3,1E-03	524
PF00784	MyTH4 domain	0,44	7,4E-03	4,1E-04	276
PF16300	Type of WD40 repeat	0,44	7,8E-03	2,6E-04	164
PF17450	Alpha galactosidase A C-terminal beta sandwich domain	0,43	9,1E-03	2,1E-02	49
PF13848	Thioredoxin-like domain	0,39	2,1E-02	2,0E-02	238
PF15788	Domain of unknown function (DUF4705)	0,37	2,9E-02	5,3E-09	273
PF10409	C2 domain of PTEN tumour-suppressor protein	0,37	3,2E-02	2,0E-02	160
PF02185	Hr1 repeat	0,36	3,7E-02	1,2E-02	189
PF03097	BRO1-like domain	-0,35	4,1E-02	3,2E-05	152
PF00443	Ubiquitin carboxyl-terminal hydrolase	-0,37	2,9E-02	2,0E-06	1383
PF00252	Ribosomal protein L16p/L10e	-0,39	2,0E-02	6,7E-11	128
PF04802	Component of IIS longevity pathway SMK-1	-0,41	1,6E-02	3,6E-05	67
PF00566	Rab-GTPase-TBC domain	-0,44	8,0E-03	1,1E-04	1153
PF13589	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	-0,45	6,3E-03	1,1E-05	263
PF00456	Transketolase, thiamine diphosphate binding domain	-0,46	5,7E-03	1,1E-03	73
PF01268	Formate--tetrahydrofolate ligase	-0,46	5,0E-03	2,6E-02	78
PF04258	Signal peptide peptidase	-0,50	2,2E-03	3,0E-02	148
PF00318	Ribosomal protein S2	-0,50	1,9E-03	3,6E-08	154
PF11976	Ubiquitin-2 like Rad60 SUMO-like	-0,50	1,9E-03	9,9E-03	159
PF01283	Ribosomal protein S26e	-0,52	1,3E-03	7,6E-09	71
PF01564	Spermine/spermidine synthase domain	-0,52	1,0E-03	7,5E-05	115
PF01195	Peptidyl-tRNA hydrolase	-0,57	2,4E-04	2,1E-04	40
PF02782	FGGY family of carbohydrate kinases, C-terminal domain	-0,59	1,5E-04	1,7E-03	173
PF06418	CTP synthase N-terminus	-0,62	5,1E-05	2,6E-02	69
PF00208	Glutamate/Leucine/Phenylalanine/Valine dehydrogenase	-0,62	3,9E-05	8,9E-06	85
PF02879	Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain II	-0,63	2,8E-05	2,2E-04	116
PF00366	Ribosomal protein S17	-0,65	1,5E-05	6,9E-04	93
PF02544	3-oxo-5-alpha-steroid 4-dehydrogenase	-0,67	5,8E-06	4,4E-02	172
PF04950	40S ribosome biogenesis protein Tsr1 and BMS1 C-terminal	-0,70	1,5E-06	5,0E-09	92
PF13774	Regulated-SNARE-like domain	-0,70	1,2E-06	6,6E-03	192
PF01142	tRNA pseudouridine synthase D (TruD)	-0,71	8,8E-07	4,6E-02	78
PF00570	HRDC domain	-0,71	8,8E-07	5,6E-03	111
PF00428	60s Acidic ribosomal protein	-0,72	4,1E-07	2,6E-05	182
PF04078	Cell differentiation family, Rcd1-like	-0,73	2,3E-07	4,4E-02	57
PF04130	Spc97 / Spc98 family	-0,74	2,2E-07	4,4E-03	209
PF04101	Glycosyltransferase family 28 C-terminal domain	-0,75	9,6E-08	1,2E-02	60
PF00890	FAD binding domain	-0,76	6,0E-08	3,6E-03	114
PF00416	Ribosomal protein S13/S18	-0,76	4,3E-08	2,8E-03	66
PF00466	Ribosomal protein L10	-0,81	2,7E-09	2,2E-03	117
PF02878	Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain I	-0,82	7,8E-10	2,7E-06	196
PF00118	TCP-1/cpn60 chaperonin family	-0,88	5,3E-12	2,6E-02	589

Tabela Suplementar 2 - Termos GO significativamente associados ao aumento da complexidade em *Eukarya*

name	Description	Spearman correlation	Spearman q-value	PIC q-value	Size
GO:0048731	system development	0,92	9,40E-14	7,10E-03	703
GO:0005576	extracellular region		4,50E-13	4,60E-02	7397
GO:0030545	receptor regulator activity	0,90	4,50E-13	3,20E-04	2971
GO:0048018	receptor ligand activity	0,90	4,50E-13	3,20E-04	2971
GO:0009888	tissue development	0,90	4,80E-13	7,80E-08	244
GO:0040007	growth	0,90	6,70E-13	2,70E-05	491
GO:0032502	developmental process	0,90	7,10E-13	4,30E-03	2673
GO:0051240	positive regulation of multicellular organismal process	0,89	7,20E-13	3,90E-02	133
GO:0002682	regulation of immune system process	0,89	7,60E-13	7,30E-04	281
GO:0010942	positive regulation of cell death	0,89	7,60E-13	9,30E-04	87
GO:0043065	positive regulation of apoptotic process	0,89	7,60E-13	9,30E-04	87
GO:0043068	positive regulation of programmed cell death	0,89	7,60E-13	9,30E-04	87
GO:0050776	regulation of immune response	0,89	7,60E-13	1,50E-02	189
GO:0048856	anatomical structure development	0,89	7,90E-13	3,30E-03	2372
GO:0048513	animal organ development	0,89	1,50E-12	6,50E-05	346
GO:0016477	cell migration	0,89	1,60E-12	3,30E-07	161
GO:0030334	regulation of cell migration	0,89	1,60E-12	1,10E-02	127
GO:0005164	tumor necrosis factor receptor binding	0,88	2,00E-12	3,30E-05	274
GO:0032813	tumor necrosis factor receptor superfamily binding	0,88	2,00E-12	3,30E-05	274
GO:0008227	G protein-coupled amine receptor activity	0,88	2,30E-12	3,20E-02	560
GO:0071944	cell periphery	0,88	4,20E-12	2,10E-03	6735
GO:0048584	positive regulation of response to stimulus	0,88	5,00E-12	8,50E-03	392
GO:0016049	cell growth	0,87	7,10E-12	5,60E-04	368
GO:0002684	positive regulation of immune system process	0,87	1,00E-11	5,80E-05	177
GO:0050778	positive regulation of immune response	0,87	1,00E-11	1,60E-05	143
GO:0002253	activation of immune response	0,87	1,20E-11	1,80E-05	117
GO:0005125	cytokine activity	0,86	1,80E-11	3,20E-02	711
GO:0006955	immune response	0,86	1,80E-11	2,40E-04	2091
GO:0051239	regulation of multicellular organismal process	0,86	2,20E-11	2,50E-03	530
GO:0045087	innate immune response	0,86	2,40E-11	1,60E-07	269
GO:1903522	regulation of blood circulation	0,86	2,50E-11	4,60E-02	159
GO:0007399	nervous system development	0,86	3,00E-11	2,50E-02	301
GO:0004860	protein kinase inhibitor activity	0,86	3,50E-11	1,00E-05	164
GO:0019210	kinase inhibitor activity	0,86	3,50E-11	1,00E-05	164
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	0,86	4,10E-11	3,90E-02	1639
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	0,86	4,10E-11	3,90E-02	1640
GO:0140030	modification-dependent protein binding	0,85	5,00E-11	1,10E-03	60
GO:0002376	immune system process	0,85	6,60E-11	8,60E-05	2237
GO:0005102	signaling receptor binding	0,85	6,60E-11	1,40E-02	4613

GO:0030155	regulation of cell adhesion	0,85	6,60E-11	1,90E-02	110
GO:0030522	intracellular receptor signaling pathway	0,85	6,80E-11	1,50E-02	99
GO:0060627	regulation of vesicle-mediated transport	0,85	6,80E-11	7,80E-11	60
GO:0031593	polyubiquitin modification-dependent protein binding	0,85	9,10E-11	7,20E-08	40
GO:0070530	K63-linked polyubiquitin modification-dependent protein binding	0,85	9,10E-11	7,20E-08	40
GO:0008543	fibroblast growth factor receptor signaling pathway	0,85	9,30E-11	3,80E-02	66
GO:0044344	cellular response to fibroblast growth factor stimulus	0,85	9,30E-11	3,80E-02	66
GO:0071774	response to fibroblast growth factor	0,85	9,30E-11	3,80E-02	66
GO:0001933	negative regulation of protein phosphorylation	0,85	9,50E-11	2,00E-06	45
GO:0006469	negative regulation of protein kinase activity	0,85	9,50E-11	2,00E-06	45
GO:0033673	negative regulation of kinase activity	0,85	9,50E-11	2,00E-06	45
GO:0042326	negative regulation of phosphorylation	0,85	9,50E-11	2,00E-06	45
GO:0051348	negative regulation of transferase activity	0,85	9,50E-11	2,00E-06	45
GO:0005179	hormone activity	0,85	1,00E-10	1,70E-04	1125
GO:0004675	transmembrane receptor protein serine/threonine kinase activity	0,84	1,50E-10	1,20E-03	214
GO:0017137	Rab GTPase binding	0,84	2,10E-10	3,40E-03	169
GO:0080134	regulation of response to stress	0,84	2,30E-10	1,10E-04	184
GO:0050793	regulation of developmental process	0,84	2,40E-10	4,30E-02	547
GO:0004415	hyalurononglucosaminidase activity	0,83	2,60E-10	4,90E-02	85
GO:0032147	activation of protein kinase activity	0,83	2,90E-10	5,10E-05	79
GO:0006954	inflammatory response	0,83	3,50E-10	1,40E-07	310
GO:0006959	humoral immune response	0,83	4,80E-10	1,00E-07	80
GO:0052866	phosphatidylinositol phosphate phosphatase activity	0,83	5,60E-10	9,00E-06	67
GO:0006956	complement activation	0,82	5,90E-10	9,00E-10	54
GO:0072376	protein activation cascade	0,82	5,90E-10	9,00E-10	54
GO:0004935	adrenergic receptor activity	0,82	6,50E-10	6,90E-04	116
GO:0008140	cAMP response element binding protein binding	0,82	8,00E-10	2,70E-04	40
GO:0098802	plasma membrane receptor complex	0,82	9,40E-10	9,80E-03	296
GO:0043235	receptor complex	0,82	9,60E-10	3,90E-02	319
GO:0008188	neuropeptide receptor activity	0,82	1,00E-09	2,40E-02	593
GO:0030291	protein serine/threonine kinase inhibitor activity	0,82	1,10E-09	1,80E-06	137
GO:0008277	regulation of G protein-coupled receptor signaling pathway	0,82	1,10E-09	1,20E-03	71
GO:0098772	molecular function regulator	0,81	1,30E-09	1,80E-03	10080
GO:0032101	regulation of response to external stimulus	0,81	1,30E-09	1,50E-05	106
GO:0050900	leukocyte migration	0,81	1,30E-09	1,60E-08	63
GO:0040008	regulation of growth	0,81	1,40E-09	2,40E-03	298
GO:0040011	locomotion	0,81	1,40E-09	6,90E-03	694
GO:0007417	central nervous system development	0,81	1,40E-09	7,20E-08	77
GO:0044459	plasma membrane part	0,81	1,40E-09	2,20E-02	4199
GO:0008305	integrin complex	0,81	1,50E-09	9,30E-03	273
GO:0098636	protein complex involved in cell adhesion	0,81	1,50E-09	9,30E-03	273
GO:0004953	icosanoid receptor activity	0,81	1,60E-09	1,40E-03	129
GO:0007420	brain development	0,81	1,70E-09	1,50E-08	74
GO:0060322	head development	0,81	1,70E-09	1,50E-08	74

GO:0004969	histamine receptor activity	0,81	1,90E-09	3,70E-03	60
GO:0001501	skeletal system development	0,81	2,00E-09	6,80E-11	52
GO:0051216	cartilage development	0,81	2,00E-09	6,80E-11	52
GO:0061448	connective tissue development	0,81	2,00E-09	6,80E-11	52
GO:0007229	integrin-mediated signaling pathway	0,81	2,10E-09	2,00E-02	49
GO:2000026	regulation of multicellular organismal development	0,81	2,30E-09	5,30E-04	231
GO:0003708	retinoic acid receptor activity	0,81	2,40E-09	8,20E-05	52
GO:0048384	retinoic acid receptor signaling pathway	0,81	2,40E-09	8,20E-05	52
GO:0045095	keratin filament	0,81	2,50E-09	9,00E-07	562
GO:0002062	chondrocyte differentiation	0,80	2,70E-09	1,20E-15	42
GO:0004954	prostanoid receptor activity	0,80	2,80E-09	1,70E-04	87
GO:0007254	JNK cascade	0,80	2,90E-09	6,30E-09	40
GO:0098797	plasma membrane protein complex	0,80	2,90E-09	1,10E-03	2665
GO:0050953	sensory perception of light stimulus	0,80	3,00E-09	2,50E-03	348
GO:0001558	regulation of cell growth	0,80	3,20E-09	1,50E-03	270
GO:0005520	insulin-like growth factor binding	0,80	3,20E-09	1,50E-03	272
GO:0048585	negative regulation of response to stimulus	0,80	3,30E-09	8,00E-06	313
GO:0018108	peptidyl-tyrosine phosphorylation	0,80	3,50E-09	1,00E-05	43
GO:0018212	peptidyl-tyrosine modification	0,80	3,50E-09	1,00E-05	43
GO:0022898	regulation of transmembrane transporter activity	0,80	3,70E-09	8,00E-06	53
GO:0032409	regulation of transporter activity	0,80	3,70E-09	8,00E-06	53
GO:0032412	regulation of ion transmembrane transporter activity	0,80	3,70E-09	8,00E-06	53
GO:0034762	regulation of transmembrane transport	0,80	3,70E-09	8,00E-06	53
GO:0034765	regulation of ion transmembrane transport	0,8	3,70E-09	8,00E-06	53
GO:1904062	regulation of cation transmembrane transport	0,80	3,70E-09	8,00E-06	53
GO:2001257	regulation of cation channel activity	0,80	3,70E-09	8,00E-06	53
GO:0004908	interleukin-1 receptor activity	0,80	4,50E-09	1,40E-04	62
GO:0002764	immune response-regulating signaling pathway	0,80	4,90E-09	8,90E-03	69
GO:0004936	alpha-adrenergic receptor activity	0,80	4,90E-09	3,60E-10	54
GO:0001608	G protein-coupled nucleotide receptor activity	0,80	5,00E-09	3,60E-06	102
GO:0045028	G protein-coupled purinergic nucleotide receptor activity	0,80	5,00E-09	3,60E-06	102
GO:0007601	visual perception	0,80	5,30E-09	2,90E-03	337
GO:1901019	regulation of calcium ion transmembrane transporter activity	0,79	5,70E-09	1,40E-06	50
GO:1903169	regulation of calcium ion transmembrane transport	0,79	5,70E-09	1,40E-06	50
GO:0031490	chromatin DNA binding	0,79	5,90E-09	9,20E-05	61
GO:0031492	nucleosomal DNA binding	0,79	5,90E-09	9,20E-05	61
GO:0032410	negative regulation of transporter activity	0,79	6,00E-09	7,50E-08	40
GO:0032413	negative regulation of ion transmembrane transporter activity	0,79	6,00E-09	7,50E-08	40
GO:0034763	negative regulation of transmembrane transport	0,79	6,00E-09	7,50E-08	40
GO:0034766	negative regulation of ion transmembrane transport	0,79	6,00E-09	7,50E-08	40
GO:0043271	negative regulation of ion transport	0,79	6,00E-09	7,50E-08	40
GO:0051051	negative regulation of transport	0,79	6,00E-09	7,50E-08	40
GO:0051926	negative regulation of calcium ion transport	0,79	6,00E-09	7,50E-08	40
GO:1901020	negative regulation of calcium ion transmembrane transporter activity	0,79	6,00E-09	7,50E-08	40

GO:1901385	regulation of voltage-gated calcium channel activity	0,79	6,00E-09	7,50E-08	40
GO:1901386	negative regulation of voltage-gated calcium channel activity	0,79	6,00E-09	7,50E-08	40
GO:1901841	regulation of high voltage-gated calcium channel activity	0,79	6,00E-09	7,50E-08	40
GO:1901842	negative regulation of high voltage-gated calcium channel activity	0,79	6,00E-09	7,50E-08	40
GO:1903170	negative regulation of calcium ion transmembrane transport	0,79	6,00E-09	7,50E-08	40
GO:1904063	negative regulation of cation transmembrane transport	0,79	6,00E-09	7,50E-08	40
GO:2001258	negative regulation of cation channel activity	0,79	6,00E-09	7,50E-08	40
GO:0005886	plasma membrane	0,79	6,00E-09	2,50E-03	6062
GO:0019882	antigen processing and presentation	0,79	6,00E-09	6,90E-03	189
GO:0042611	MHC protein complex	0,79	6,10E-09	2,80E-02	128
GO:0042613	MHC class II protein complex	0,79	6,10E-09	2,80E-02	128
GO:0046580	negative regulation of Ras protein signal transduction	0,79	6,20E-09	8,80E-15	51
GO:0051058	negative regulation of small GTPase mediated signal transduction	0,79	6,20E-09	8,80E-15	51
GO:0006935	chemotaxis	0,79	6,40E-09	2,00E-02	457
GO:0042330	taxis	0,79	6,40E-09	2,00E-02	457
GO:0035329	hippo signaling	0,79	6,60E-09	1,80E-02	54
GO:0003785	actin monomer binding	0,79	6,90E-09	1,30E-04	62
GO:0008544	epidermis development	0,79	7,50E-09	1,70E-06	92
GO:0007155	cell adhesion	0,79	7,60E-09	2,50E-02	3283
GO:0022610	biological adhesion	0,79	7,60E-09	2,50E-02	3283
GO:0006952	defense response	0,79	9,30E-09	9,30E-03	1347
GO:0004955	prostaglandin receptor activity	0,79	9,80E-09	3,20E-06	69
GO:0016594	glycine binding	0,78	1,00E-08	3,40E-02	49
GO:1902532	negative regulation of intracellular signal transduction	0,78	1,20E-08	6,10E-13	110
GO:0002252	immune effector process	0,78	1,30E-08	2,10E-08	117
GO:0009968	negative regulation of signal transduction	0,78	1,60E-08	3,60E-06	253
GO:0010648	negative regulation of cell communication	0,78	1,60E-08	3,60E-06	253
GO:0023057	negative regulation of signaling	0,78	1,60E-08	3,60E-06	253
GO:0030154	cell differentiation	0,78	1,70E-08	2,00E-02	611
GO:0038036	sphingosine-1-phosphate receptor activity	0,77	2,00E-08	3,00E-02	62
GO:0019838	growth factor binding	0,77	2,50E-08	1,30E-03	311
GO:0008191	metalloendopeptidase inhibitor activity	0,77	2,70E-08	2,80E-06	68
GO:0004999	vasoactive intestinal polypeptide receptor activity	0,77	3,10E-08	8,00E-06	47
GO:0060284	regulation of cell development	0,76	3,60E-08	1,50E-10	51
GO:0009986	cell surface	0,76	3,60E-08	2,30E-07	45
GO:0007050	cell cycle arrest	0,76	4,70E-08	3,30E-03	99
GO:0009605	response to external stimulus	0,76	4,80E-08	1,10E-02	976
GO:0008053	mitochondrial fusion	0,75	7,10E-08	1,40E-02	74
GO:0005212	structural constituent of eye lens	0,75	8,20E-08	4,40E-07	60
GO:0034593	phosphatidylinositol bisphosphate phosphatase activity	0,74	1,20E-07	1,40E-02	56
GO:1902533	positive regulation of intracellular signal transduction	0,74	1,40E-07	6,70E-04	164
GO:0005149	interleukin-1 receptor binding	0,74	1,40E-07	1,30E-03	45
GO:0043408	regulation of MAPK cascade	0,73	1,90E-07	9,20E-05	122

GO:0007602	phototransduction	0,73	2,70E-07	4,10E-07	126
GO:0022604	regulation of cell morphogenesis	0,73	2,90E-07	2,70E-05	59
GO:0031401	positive regulation of protein modification process	0,72	3,10E-07	1,50E-03	184
GO:0051018	protein kinase A binding	0,72	3,20E-07	7,40E-06	51
GO:0044093	positive regulation of molecular function	0,72	4,00E-07	7,90E-04	434
GO:0008360	regulation of cell shape	0,72	4,60E-07	2,80E-05	44
GO:0004861	cyclin-dependent protein serine/threonine kinase inhibitor activity	0,72	4,90E-07	2,60E-03	100
GO:0043085	positive regulation of catalytic activity	0,71	5,70E-07	6,90E-04	421
GO:0051606	detection of stimulus	0,71	5,80E-07	2,60E-03	206
GO:0033674	positive regulation of kinase activity	0,71	6,30E-07	4,40E-07	113
GO:0045860	positive regulation of protein kinase activity	0,71	6,30E-07	4,40E-07	113
GO:0051347	positive regulation of transferase activity	0,71	6,30E-07	4,40E-07	113
GO:0010562	positive regulation of phosphorus metabolic process	0,69	1,70E-06	2,40E-06	171
GO:0045937	positive regulation of phosphate metabolic process	0,69	1,70E-06	2,40E-06	171
GO:0097223	sperm part	0,69	2,30E-06	8,10E-04	43
GO:0030433	ubiquitin-dependent ERAD pathway	0,68	2,90E-06	3,20E-03	60
GO:0034976	response to endoplasmic reticulum stress	0,68	2,90E-06	3,20E-03	60
GO:0036503	ERAD pathway	0,68	2,90E-06	3,20E-03	60
GO:0048523	negative regulation of cellular process	0,67	5,40E-06	5,40E-03	2242
GO:0097190	apoptotic signaling pathway	0,67	5,40E-06	1,10E-09	50
GO:0022603	regulation of anatomical structure morphogenesis	0,66	6,30E-06	5,10E-03	167
GO:1901698	response to nitrogen compound	0,66	7,20E-06	1,80E-03	245
GO:0001934	positive regulation of protein phosphorylation	0,66	7,30E-06	1,30E-06	148
GO:0042327	positive regulation of phosphorylation	0,66	7,30E-06	1,30E-06	148
GO:0004697	protein kinase C activity	0,66	7,80E-06	2,00E-03	164
GO:0000149	SNARE binding	0,66	7,90E-06	1,90E-02	137
GO:0019905	syntaxin binding	0,66	7,90E-06	1,90E-02	137
GO:0048518	positive regulation of biological process	0,65	9,30E-06	9,00E-03	2080
GO:0032217	riboflavin transmembrane transporter activity	0,64	1,70E-05	1,80E-03	45
GO:0032218	riboflavin transport	0,64	1,70E-05	1,80E-03	45
GO:0010243	response to organonitrogen compound	0,64	1,70E-05	1,00E-03	220
GO:0004674	protein serine/threonine kinase activity	0,64	1,80E-05	3,30E-02	3517
GO:0080135	regulation of cellular response to stress	0,64	1,90E-05	6,70E-05	51
GO:0006879	cellular iron ion homeostasis	0,64	2,00E-05	8,00E-06	177
GO:0055072	iron ion homeostasis	0,64	2,00E-05	8,00E-06	177
GO:0070887	cellular response to chemical stimulus	0,63	2,70E-05	4,80E-02	758
GO:0055065	metal ion homeostasis	0,62	3,50E-05	6,40E-04	453
GO:0019012	virion	0,61	4,60E-05	5,40E-04	89
GO:0044423	virion part	0,61	4,60E-05	5,40E-04	89
GO:0019207	kinase regulator activity	0,61	4,80E-05	3,40E-02	541
GO:0009581	detection of external stimulus	0,61	5,10E-05	2,80E-06	139
GO:0009582	detection of abiotic stimulus	0,61	5,10E-05	2,80E-06	139
GO:0009583	detection of light stimulus	0,61	5,10E-05	2,80E-06	139
GO:0006875	cellular metal ion homeostasis	0,61	5,30E-05	3,60E-04	399
GO:0030003	cellular cation homeostasis	0,61	5,30E-05	1,10E-03	400
GO:0006873	cellular ion homeostasis	0,61	5,40E-05	1,10E-03	402

GO:0043269	regulation of ion transport	0,6	6,60E-05	4,90E-04	91
GO:0050801	ion homeostasis	0,60	6,70E-05	9,10E-04	692
GO:0098771	inorganic ion homeostasis	0,60	6,70E-05	9,10E-04	692
GO:0010256	endomembrane system organization	0,60	6,80E-05	9,40E-03	295
GO:1901568	fatty acid derivative metabolic process	0,60	6,90E-05	3,80E-03	96
GO:0010959	regulation of metal ion transport	0,60	8,00E-05	1,20E-04	82
GO:0055080	cation homeostasis	0,60	8,00E-05	1,00E-03	690
GO:1902531	regulation of intracellular signal transduction	0,60	8,20E-05	3,30E-02	2235
GO:0048869	cellular developmental process	0,60	8,90E-05	1,30E-02	845
GO:0019028	viral capsid	0,59	8,90E-05	5,10E-04	70
GO:0098552	side of membrane	0,59	9,20E-05	1,50E-08	180
GO:0051924	regulation of calcium ion transport	0,59	1,10E-04	3,10E-05	79
GO:0048522	positive regulation of cellular process	0,58	1,30E-04	2,70E-02	1847
GO:0005834	heterotrimeric G-protein complex	0,58	1,30E-04	4,90E-07	167
GO:0009898	cytoplasmic side of plasma membrane	0,58	1,30E-04	4,90E-07	167
GO:0019897	extrinsic component of plasma membrane	0,58	1,30E-04	4,90E-07	167
GO:0031234	extrinsic component of cytoplasmic side of plasma membrane	0,58	1,30E-04	4,90E-07	167
GO:0098562	cytoplasmic side of membrane	0,58	1,30E-04	4,90E-07	167
GO:1905360	GTPase complex	0,58	1,30E-04	4,90E-07	167
GO:0004683	calmodulin-dependent protein kinase activity	0,58	1,40E-04	3,20E-02	135
GO:0048519	negative regulation of biological process	0,58	1,50E-04	1,10E-02	3145
GO:0016042	lipid catabolic process	0,58	1,60E-04	1,80E-02	831
GO:0008134	transcription factor binding	0,58	1,70E-04	2,50E-04	277
GO:0007265	Ras protein signal transduction	0,57	1,90E-04	4,20E-02	1817
GO:0046578	regulation of Ras protein signal transduction	0,56	2,90E-04	4,10E-02	1708
GO:0055082	cellular chemical homeostasis	0,55	3,70E-04	3,30E-02	534
GO:0019887	protein kinase regulator activity	0,55	3,80E-04	4,40E-02	514
GO:0009653	anatomical structure morphogenesis	0,55	4,30E-04	2,00E-02	457
GO:0019902	phosphatase binding	0,54	6,00E-04	2,90E-04	42
GO:0004931	extracellularly ATP-gated cation channel activity	0,53	6,20E-04	9,40E-03	108
GO:0007584	response to nutrient	0,53	6,20E-04	9,40E-03	108
GO:0014074	response to purine-containing compound	0,53	6,20E-04	9,40E-03	108
GO:0033198	response to ATP	0,53	6,20E-04	9,40E-03	108
GO:0035381	ATP-gated ion channel activity	0,53	6,20E-04	9,40E-03	108
GO:0046683	response to organophosphorus	0,53	6,20E-04	9,40E-03	108
GO:0048878	chemical homeostasis	0,53	7,40E-04	3,70E-02	875
GO:0055076	transition metal ion homeostasis	0,53	7,90E-04	2,90E-05	255
GO:0009314	response to radiation	0,52	9,50E-04	8,50E-06	169
GO:0006897	endocytosis	0,50	1,50E-03	3,30E-07	255
GO:0032007	negative regulation of TOR signaling	0,50	1,60E-03	3,40E-04	53
GO:0043086	negative regulation of catalytic activity	0,50	1,80E-03	4,60E-02	319
GO:0061630	ubiquitin protein ligase activity	0,49	2,10E-03	8,40E-04	290
GO:0061659	ubiquitin-like protein ligase activity	0,49	2,10E-03	8,40E-04	290
GO:0046916	cellular transition metal ion homeostasis	0,48	2,70E-03	2,70E-05	216
GO:0004364	glutathione transferase activity	0,43	7,80E-03	1,40E-03	231
GO:0010563	negative regulation of phosphorus metabolic process	0,42	1,00E-02	4,70E-02	105

GO:0045936	negative regulation of phosphate metabolic process	0,42	1,00E-02	4,70E-02	105
GO:0032006	regulation of TOR signaling	0,40	1,40E-02	5,10E-03	107
GO:0098657	import into cell	0,39	1,90E-02	2,70E-07	268
GO:0031400	negative regulation of protein modification process	0,37	2,60E-02	3,20E-02	85
GO:0090575	RNA polymerase II transcription factor complex	-0,40	1,40E-02	3,10E-02	604
GO:0044257	cellular protein catabolic process	-0,48	2,80E-03	3,90E-04	3251
GO:0051603	proteolysis involved in cellular protein catabolic process	-0,48	2,80E-03	3,90E-04	3251
GO:0044265	cellular macromolecule catabolic process	-0,50	1,80E-03	6,30E-03	4221
GO:0004514	nicotinate-nucleotide diphosphorylase (carboxylating) activity	-0,51	1,40E-03	3,50E-05	100
GO:0043015	gamma-tubulin binding	-0,51	1,20E-03	4,60E-04	224
GO:0016638	oxidoreductase activity, acting on the CH-NH2 group of donors	-0,53	6,80E-04	1,20E-02	493
GO:1901565	organonitrogen compound catabolic process	-0,55	3,80E-04	8,40E-03	5205
GO:0006564	L-serine biosynthetic process	-0,56	2,90E-04	2,70E-02	100
GO:0030163	protein catabolic process	-0,57	2,20E-04	6,10E-03	3727
GO:0044444	cytoplasmic part	-0,61	4,50E-05	3,40E-02	18911
GO:0009057	macromolecule catabolic process	-0,62	4,00E-05	2,30E-02	4995
GO:0016639	oxidoreductase activity, acting on the CH-NH2 group of donors, NAD or NADP as acceptor	-0,63	2,10E-05	1,00E-05	96
GO:0006072	glycerol-3-phosphate metabolic process	-0,64	2,00E-05	7,00E-04	194
GO:0052646	alditol phosphate metabolic process	-0,64	2,00E-05	7,00E-04	194
GO:0016868	intramolecular transferase activity, phosphotransferases	-0,64	1,90E-05	5,10E-08	336
GO:0000922	spindle pole	-0,65	1,30E-05	2,00E-04	193
GO:0007020	microtubule nucleation	-0,67	4,20E-06	1,20E-03	244
GO:0046785	microtubule polymerization	-0,67	4,20E-06	1,20E-03	244
GO:0009308	amine metabolic process	-0,68	3,50E-06	3,00E-03	644
GO:0006595	polyamine metabolic process	-0,68	3,10E-06	4,60E-02	221
GO:1901575	organic substance catabolic process	-0,69	1,90E-06	4,80E-02	8603
GO:0009055	electron transfer activity	-0,73	2,00E-07	1,60E-02	2012
GO:0006566	threonine metabolic process	-0,74	1,70E-07	4,90E-02	70
GO:0015935	small ribosomal subunit	-0,74	1,60E-07	8,20E-05	590
GO:0044391	ribosomal subunit	-0,76	5,10E-08	1,20E-02	1111
GO:0005840	ribosome	-0,76	4,40E-08	1,40E-04	5911
GO:0006289	nucleotide-excision repair	-0,77	2,70E-08	2,50E-03	491
GO:0003735	structural constituent of ribosome	-0,78	1,70E-08	1,20E-04	5910
GO:1990904	ribonucleoprotein complex	-0,79	6,00E-09	8,10E-04	7587
GO:0016866	intramolecular transferase activity	-0,82	9,10E-10	6,00E-04	1000
GO:0006412	translation	-0,83	4,50E-10	2,90E-03	9225
GO:0006891	intra-Golgi vesicle-mediated transport	-0,83	4,50E-10	3,50E-07	120
GO:0043043	peptide biosynthetic process	-0,83	3,70E-10	3,00E-03	9353
GO:0006518	peptide metabolic process	-0,84	1,80E-10	4,10E-03	9776
GO:0043603	cellular amide metabolic process	-0,84	1,30E-10	8,20E-03	10971
GO:0043604	amide biosynthetic process	-0,84	1,20E-10	4,10E-03	9918
GO:0017150	tRNA dihydrouridine synthase activity	-0,86	2,60E-11	7,70E-03	190