

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-graduação em Genética

Zandora Celeste Hastenreiter Ferreira Nunes

**DOMÍNIOS PROTEICOS E SUBSTITUIÇÕES DE AMINOÁCIDOS ASSOCIADOS
COM A EVOLUÇÃO DA MASSA CEREBRAL EM PRIMATAS**

Belo Horizonte

2025

Zandora Celeste Hastenreiter Ferreira Nunes

**DOMÍNIOS PROTEICOS E SUBSTITUIÇÕES DE AMINOÁCIDOS ASSOCIADOS
COM A EVOLUÇÃO DA MASSA CEREBRAL EM PRIMATAS**

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

Orientador: Prof. Dr. Francisco Pereira Lobo
Coorientador: Prof. Dr. Fabrício Rodrigues dos Santos

Belo Horizonte

2025

043

Nunes, Zandora Celeste Hastenreiter Ferreira.

Domínios proteicos e substituições de aminoácidos associados com a evolução da massa cerebral em primatas [manuscrito] / Zandora Celeste Hastenreiter Ferreira Nunes. – 2025.

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

Orientador: Prof. Dr. Francisco Pereira Lobo. Coorientador: Prof. Dr. Fabrício Rodrigues dos Santos.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genômica. 2. Primatas. 3. Encéfalo. 4. Evolução Biológica. I. Lobo, Francisco Pereira. II. Santos, Fabrício Rodrigues dos. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575



UNIVERSIDADE FEDERAL DE MINAS GERAIS
ICB - COLEGIADO DE PÓS-GRADUAÇÃO EM GENÉTICA - SECRETARIA

ATA DE DEFESA DE DISSERTAÇÃO

ATA DA DEFESA DE DISSERTAÇÃO

371/2025

entrada

Zandora Celeste Hastenreiter Ferreira Nunes

2º/2022

CPF: 093.877.706-80

Às quatorze horas do dia **24 de abril de 2025**, reuniu-se a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Domínios Proteicos e Substituições de Aminoácidos Associados Com A Evolução da Massa Cerebral em Primatas**", 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 à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	Indicação
Dr. Francisco Pereira Lobo	UFMG	APROVADA
Dr. Fabrício Rodrigues dos Santos	UFMG	APROVADA
Dr. Gustavo Campos e Silva Kuhn	UFMG	APROVADA
Dra. Maria Cátira Bortolini	UFRGS	APROVADA

Pelas indicações, a candidata foi considerada: **APROVADA**

O resultado final foi comunicado publicamente à candidata 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, 24 de abril de 2025.

Dr. Francisco Pereira Lobo

Dr. Fabrício Rodrigues dos Santos

Dr. Gustavo Campos e Silva Kuhn



Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos, Professor do Magistério Superior**, em 25/04/2025, às 09:24, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn, Professor(a)**, em 25/04/2025, às 14:30, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maria Cátira Bortolini, Usuário Externo**, em 28/04/2025, às 11:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 06/05/2025, às 10:54, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4152059** e o código CRC **7614F41E**.



UNIVERSIDADE FEDERAL DE MINAS GERAIS

"Domínios Proteicos e Substituições de Aminoácidos Associados Com A Evolução da Massa Cerebral em Primatas"

Zandora Celeste Hastenreiter Ferreira Nunes

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Francisco Pereira Lobo
UFMG

Prof. Fabrício Rodrigues dos Santos
UFMG

Prof. Gustavo Campos e Silva Kuhn
UFMG

Profa Maria Cátira Bortolini
UFRGS

Belo Horizonte, 24 de abril de 2025.



Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos, Professor do Magistério Superior**, em 25/04/2025, às 09:24, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gustavo Campos e Silva Kuhn, Professor(a)**, em 25/04/2025, às 14:29, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maria Cátira Bortolini, Usuário Externo**, em 28/04/2025, às 11:34, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 06/05/2025, às 10:54, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4152070** e o código CRC **F9ED97A3**.

Para Maria Eugênia e Zoey

AGRADECIMENTOS

Não sei exatamente como estruturar estes agradecimentos, mas acredito que uma forma de começar é descrever um cadinho o que foram esses 2 anos (e mais alguns meses) de mestrado. Tanta coisa aconteceu, tanto academicamente e pessoalmente, e alívio é o que eu sinto ao pensar que esse ciclo vai se encerrar. Perdi uma amiga muito amada, a Maria Eugênia e também perdi minha neném canina, a Zoey. As duas não estavam bem de saúde durante o período do mestrado e eu me preocupava diariamente. Essas duas são quem eu mais queria agradecer só por terem me dado a honra de tê-las na minha vida, queria poder abraçar as duas mais que tudo.

Esse mestrado é dedicado principalmente a Maria Eugênia. Anos atrás, o CMZ - boiolinha's club (grupo composto pela Clara, Mareu e essa que vos fala), planejava fazer o mestrado junto, aqui na genética. Apesar da Mareu ter optado por fazer outra graduação depois da biologia e não entrar no mestrado, conversando com a mãe dela um tempo atrás, ela mencionou que Mareu queria ter feito o mestrado sim, mas que por n razões, não se sentia fisicamente capaz. Por esse motivo, a minha dedicatória deste trabalho é para ela.

Aos vivos, gostaria de agradecer minha irmã, Zaira Clarice, por existir na minha vida. Tenho por ela um amor incondicional e vê-la crescer me enche de alegria. O fato de que ela fala muito de mim para as pessoas com as quais ela convive — fato este que talvez ela não admita facilmente — me faz acreditar que sou importante para ela e isso, literalmente, me dá vontade de viver. Agradeço também à minha mãe, Ediana, pelo apoio ao longo da minha longa jornada tanto na UFMG como fora dela. Mãe, obrigada por ser quem você é e por me aceitar como eu sou.

Agradeço aos velhos e novos amigos. Túlio, obrigada por ser minha dupla pro que der e vier. A vida é engraçada, num momento estamos voltando para casa na mesma van e no outro já se passaram 9 anos de amizade. Por mais 9, e mais 9, e mais 9 anos. Luís, boa sorte na sua mais nova empreitada, espero que tudo dê certo. Clara, queria que as circunstâncias fossem melhores para nossa amizade no momento, mas ainda penso sempre em você. Henry, agradeço por todo o amparo e pedidos de socorro atendidos, principalmente nos últimos meses — obrigada também por me proporcionar a chance de ser sua madrinha de casamento! Agradeço também ao Maycon e a Thais, por me permitirem chorar as pitangas em conjunto.

Agradeço ao Chico, pela mentoria e atenção durante todos os anos em que estive no LAB. Tenho grande admiração pela pessoa que você é. Obrigada por me aceitar no LAB lá

nos primórdios e me ajudar a realizar meu sonho inocente de caloura de “trabalhar com genética”.

Agradeço também ao Fabrício, por quem tenho enorme apreço, e que aceitou me co-orientar no meio do mestrado. Com a nova co-orientação pude ir me agregando aos poucos ao LBEM e eu sou imensamente grata a todos os membros do LBEM, principalmente Davi, Isaac, Laura e Pedro que me receberam com muito carinho e que, provavelmente sem saber, me fizeram repensar em continuar na UFMG e eu pude lembrar como gosto daqui. Eu não me via continuando nessas terras, mas me mudar para um laboratório habitado trouxe brilho aos meus olhos novamente. Meu profundo obrigada!

Um abraço pro meu psiquiatra Bernardo que finalmente conseguiu acertar o combo de medicação perfeito para mim! E pra minha psicóloga Isabella, que ainda há de escutar muito com o doutorado.

Menção honrosa às minhas irmãs e irmão felinos, Zuri, Ziggy e Zico, em quem posso dar um cheirinho revitalizante no cangote.

Por fim, gostaria de agradecer novamente a todos — incluindo aqueles aqui não nomeados — mas que estiveram e passaram pela minha vida nesses últimos dois anos e tornaram ela um pouco menos amarga e mais agradável de se viver.

٩(◡̈)٩

RESUMO

Primates compreendem uma ordem de mamíferos conhecidos por sua capacidade cognitiva, um fenótipo complexo que resulta, ao menos parcialmente, de uma massa cerebral maior do que a esperada, dada sua massa corporal. Este fenótipo é recorrentemente descrito por valores numéricos que visam caracterizar a direção e intensidade de sua variação. Dois estimadores são frequentemente utilizados em vertebrados para medir a variação da massa cerebral observada, dada a massa corporal: o quociente de encefalização (EQ) e os resíduos da massa cerebral. A massa cerebral dos primatas apresenta enorme variação, abrangendo mais de três ordens de grandeza, e este fenótipo é herdável, ou seja, definido geneticamente. A genômica comparativa consiste em analisar características genótípicas entre espécies para nos ajudar a compreender a história evolutiva de diversos fenótipos. Uma abordagem recorrente na genômica comparativa é a busca por relações de correlação entre genótipos e fenótipos. Ao realizar análises desta natureza, é imprescindível considerar a relação filogenética entre as espécies, pois as observações são dependentes do padrão de ancestralidade compartilhada entre cada par de espécies. Outra metodologia possível é a procura por substituições individuais de aminoácidos compartilhadas entre as espécies com os diferentes valores dos fenótipos avaliados. Neste trabalho, integramos dados genômicos, fenotípicos e filogenéticos para identificar domínios proteicos e funções biológicas cuja expansão esteja associada com a variação da massa cerebral em primatas. Procuramos também por substituições de aminoácidos compartilhadas entre os primatas com maiores valores de EQ. Para tal, obtivemos, a partir de genomas de primatas disponíveis publicamente, proteomas não-redundantes anotados e de alta qualidade, bem como informação sobre a história evolutiva e as massas corporal e cerebral do grupo. Tais dados foram integrados para procurar correlações entre a frequência dos termos de anotação e os valores de EQ e dos resíduos. Entre os domínios significativamente associados ao EQ ou aos resíduos, destacamos o domínio FAM72, com alta expressão em células tronco neuronais. Outro domínio notável é o Olduvai, com expansão marcante no número de cópias nos humanos e implicado no aumento do tamanho cerebral nesta linhagem. Quanto às funções biológicas, diversas delas estão relacionadas com o metabolismo de proteínas. Há ainda domínios e funções que carecem de informação para entender sua possível relação com a massa cerebral. Entre os genes nos quais existem substituições convergentes compartilhadas pelas espécies com maiores EQ (e não pelas demais), houve enriquecimento de funções como *neuron projection morphogenesis*, *neuron differentiation* e *generation of neurons*. Em suma, este estudo traz informações sobre alterações genéticas associadas ao tamanho cerebral de primatas contribuindo para uma maior compreensão dos fundamentos moleculares da encefalização em nossa linhagem.

Palavras-chave: genômica comparativa, evolução do cérebro de primatas, encefalização

ABSTRACT

Primates form an order of mammals known for their cognitive abilities, a complex phenotype that results, at least partially, from a larger-than-expected brain mass given their body mass. This phenotype is frequently described using numerical values that aim to characterize the direction and intensity of its variation. Two frequently used estimators are the encephalization quotient (EQ) and the residuals of brain mass. Brain mass in primates exhibits enormous variation, spanning more than three orders of magnitude, and this phenotype is a heritable trait, therefore defined genetically. Comparative genomics is the comparison of genotypic traits across species to identify similarities and differences and it helps us understand the evolutionary history of various traits. A common approach in this field is searching for correlations between genotypes and phenotypes within certain groups. When conducting such analysis it is advisable to take into account the phylogenetic relationships between species, given that they share common ancestry and are not independent entities — an assumption often made in many statistical analyses. Another possible methodology is the search for individual amino acid substitutions shared among species with the highest values of the evaluated phenotypes. In this study, we integrated genomic, phenotypic, and phylogenetic data to identify both protein domains and biological functions whose expansion is associated with the variation in brain mass among primates. We also searched for amino acid substitutions shared among primates with higher EQ values. To achieve this, we obtained high-quality, non-redundant annotated proteomes from publicly available primate genomes, along with information on the evolutionary history and body/brain masses of the group. These data were integrated to investigate correlations between annotation term frequencies and EQ and residual values. Among the domains significantly associated with EQ or residuals, we highlight the FAM72 domain, which is highly expressed in neural stem cells. Another notable domain is Olduvai, which shows a marked expansion in copy number in humans and has been implicated in brain size increase in this lineage. Regarding biological functions, several are related to protein metabolism. There are also domains and functions for which further information is needed to understand their potential relationship with brain mass. Among the genes in which convergent substitutions are shared by species with higher EQ (but not by others), we found an enrichment of functions such as neuron projection morphogenesis, neuron differentiation, and generation of neurons. In summary, this study provides insights into genetic factors associated with primate brain size, contributing to a better understanding of the molecular foundations of encephalization in our lineage.

Keywords: comparative genomics, primate brain evolution, encephalization

LISTA DE FIGURAS

Figura 1: Controle de qualidade dos genomas utilizados.	25
Figura 2: Dados fenotípicos e imputação de dados faltantes.	27
Figura 3: Correlação entre a frequência do domínio FAM72 (PF14976) e o EQ.	31
Figura 4: Correlação entre a frequência do domínio Olduvai (PF06758) e o EQ.	32
Figura 5: Correlação entre a frequência do Alpha Amylase, catalytic domain (PF00128) e o EQ.	33
Figura 6: Correlação entre a frequência do domínio Lipid Desaturase (PF10520) e os resíduos de massa cerebral.	35
Figura 7: Correlação entre o Rho GTPase-activating protein 11 (IPR042869) e EQ.	38
Figura 8: Correlação entre Musculoskeletal embryonic nuclear protein 1 (IPR031394) e os resíduos de massa cerebral.	40
Figura 9: Categorias funcionais enriquecidas.	43

LISTA DE TABELAS

Tabela 1: Espécies do grupo de descoberta de CAAS.	22
Tabela 2: Dados fenotípicos dos primatas utilizados nas análises.	28
Tabela 3: Exemplos de domínios de proteínas associados com a variação do EQ.	30
Tabela 4: Exemplos de domínios de proteínas associados com a variação dos resíduos da massa cerebral.	34
Tabela 5: Exemplos de funções biológicas associadas com a variação do EQ.	35
Tabela 6: Funções biológicas associadas com a variação dos resíduos da massa cerebral.	37
Tabela 7: Exemplos de termos IPR de anotação associados com a variação do EQ.	37
Tabela 8: Exemplos de termos IPR de anotação associados com a variação dos resíduos da massa cerebral.	39

ABREVIACOES

BG	<i>Background</i>
CAAS	Substituioes Convergentes de Aminocidos
CNV	<i>Copy Number Variation</i>
DNA	cido Desoxiribonucleico
DUF	<i>Domain of Unknown Function</i>
EQ	Quociente de Encefalizaao
ER	<i>Enrichment Ratio</i>
FDR	<i>False Discovery Rate</i>
FG	<i>Foreground</i>
GO	<i>Gene Ontology</i>
HAR	<i>Human Accelerated Regions</i>
IUCN	<i>International Union for Conservation of Nature</i>
MSA	Alinhamento Mltiplo de Sequncias
NCBI	<i>National Center for Biotechnology Information</i>
NPC	<i>Neuronal Progenitor Cell</i>
NSC	<i>Neuronal Stem Cell</i>
PIC	<i>Phylogenetic Independent Contrasts</i>
RefSeq	<i>Reference Sequence</i>
RNA	cido Ribonucleico
WebGestalt	WEB-based GEne SeT AnaLysis Toolkit

SUMÁRIO

1. INTRODUÇÃO.....	11
1.1. A Ordem Primata.....	11
1.2. O cérebro dos Primatas.....	12
1.3. O método comparativo e suas aplicações em genômica.....	14
2. OBJETIVOS.....	16
2.1. Objetivo geral.....	16
2.2. Objetivos específicos.....	16
3. MATERIAIS & MÉTODOS.....	17
3.1. Obtenção de genomas.....	17
3.2. Controle de qualidade.....	17
3.3. Anotação genômica.....	18
3.4. Dados fenotípicos e imputação de dados faltantes.....	18
3.5. Árvores filogenéticas.....	19
3.6. Processamento e adequação de arquivos para uso do CALANGO.....	19
3.7. Alinhamento de genes ortólogos.....	21
3.8. Análise de sítios de substituições convergentes de aminoácidos.....	21
3.9. Análise de enriquecimento.....	22
4. RESULTADOS & DISCUSSÃO.....	23
4.1. Obtenção de genomas.....	23
4.2. Controle de qualidade e anotação dos genomas.....	23
4.3. Dados fenotípicos.....	26
4.4. Imputação de dados faltantes.....	26
4.5. Árvores filogenéticas.....	29
4.6. Análise de correlações CALANGO.....	29
4.6.1. Domínios de proteínas associados com a variação do EQ.....	30
4.6.2. Domínios de proteínas associados com a variação dos resíduos da massa cerebral.....	33
4.6.3. Funções biológicas associadas com a variação do EQ.....	35
4.6.4. Funções biológicas associadas com a variação dos resíduos da massa cerebral.....	36
4.6.5. Termos IPR associados com a variação do EQ.....	37
4.6.6. Termos IPR associados com a variação dos resíduos da massa cerebral.....	38
4.7. Alinhamento de genes ortólogos.....	41
4.8. Análise de sítios de substituições convergentes de aminoácidos.....	41
4.9. Análise de enriquecimento.....	42
5. CONCLUSÃO.....	43
REFERÊNCIAS.....	45
APÊNDICE A - Termos de anotação significativamente associados com os fenótipos avaliados de acordo com o CALANGO.....	53
APÊNDICE B - Genes onde existem substituições convergentes de aminoácidos em espécies nos extremos de variação da massa cerebral.....	59

1. INTRODUÇÃO

Os Primatas, a terceira ordem mais diversa dentro da classe Mammalia, são popularmente conhecidos como tárrios, macacos, lêmures e símios. A maioria dos membros desta ordem habitam regiões tropicais, especialmente florestas, nas Américas, Ásia e África (ORKIN; KUDERNA; MARQUES-BONET, 2021). O interesse científico e público por esse grupo é notável, sendo amplamente atribuído ao fato de que a nossa espécie também pertence a essa ordem (VERACINI; WOOD, 2024). Nesse contexto, o estudo da diversidade morfológica, comportamental e genética dos primatas oferece conhecimentos fundamentais para a compreensão das origens evolutivas e das particularidades biológicas da nossa própria espécie.

1.1. A Ordem Primata

A ordem Primata tem sua origem estimada em mais 65 milhões de anos atrás, no período do Cretáceo (JANECKA *et al.*, 2007). O fóssil mais antigo atribuído ao grupo é o *Purgatorius*, com aproximadamente 63 milhões de anos e encontrado onde hoje é a América do Norte (MANTILLA *et al.*, 2021). Atualmente, existem cerca de 500 espécies descritas de primatas, sendo que cerca de 70% dessas estão sob risco de extinção (ESTRADA *et al.*, 2017).

A ordem dos primatas é dividida em duas subordens: **Haplorrhini** e **Strepsirrhini** — Haplorrhini significa “nariz simples” ou “nariz seco”, nome que reflete o fato de que o nariz desse grupo é, de fato, “seco”, uma vez que é recoberto por pelos. Já Strepsirrhini, por sua vez, significa “nariz molhado”, e o nariz desses animais é úmido, se assemelhando ao rinário dos cães (HERITAGE, 2014). Os Haplorrhini são divididos em duas infraordens: Tarsiiformes e Simiiformes. O táxon Tarsiiformes é formado por uma única família de tárrios nativos do sudeste asiático. Duas parvordens formam o grupo Simiiformes: Platyrrhini e Catarrhini. Os Platyrrhini são popularmente conhecidos como macacos do Novo Mundo e são nativos das Américas do Sul, Central e do Norte (sul do México). Os Catarrhini são popularmente conhecidos como macacos do Velho Mundo, e são nativos da África e Ásia. Duas superfamílias compõem o grupo dos Catarrhini: Cercopithecoidea — macacos com cauda — e Hominoidea, formada pelos gibões e demais grandes primatas com cóccix (cauda vestigial), incluindo *Homo sapiens*. Por fim, os Strepsirrhini são compostos pelos lêmures de Madagascar, lorisídeos do sul e sudeste da Ásia e da África subsaariana e pelos gálagos,

também encontrados na África subsaariana (ORKIN; KUDERNA; MARQUES-BONET, 2021).

Além da sua considerável diversidade taxonômica e ampla distribuição, há ainda duas características marcantes dos primatas quando comparados aos demais mamíferos: sua inteligência geral e o tamanho de seus cérebros (MELCHIONNA, M *et al.*, 2019).

1.2. O cérebro dos Primatas

Primatas apresentam uma estrutura corporal geral conservada entre seus membros e sem grandes especializações, diferentemente de outros grupos de mamíferos (morcegos, por exemplo). Primatas, entretanto, apresentam considerável variação no tamanho corporal. O maior dos primatas, o gorila (*Gorilla gorilla*), pode chegar a pesar mais de 200 kg, enquanto o menor deles, o lêmure-rato (*Microcebus murinus*) pode vir a pesar menos de 100 g (FLEAGLE; BADEN; GILBERT, 2024). De forma concomitante, o tamanho cerebral da ordem também apresenta uma enorme variação, com massas cerebrais variando entre 1,6 g em *M. murinus* até aproximadamente 1,3 kg em *Homo sapiens* (ROTH; DICKE, 2012). Além disso, quando considerado o peso corporal dos organismos, os primatas apresentam o maior tamanho cerebral relativo dentre os mamíferos (TSUBOI *et al.*, 2018). Ainda sobre o tamanho cerebral, este é um tema recorrente de estudos, visto que o tamanho do cérebro parece ser um dos principais fatores associados à inteligência, característica notória dos primatas e em especial o *Homo sapiens* (MELCHIONNA, MARINA *et al.*, 2025).

Existem diferentes métricas comumente utilizadas para descrever o tamanho relativo cerebral em vertebrados. Uma medida amplamente utilizada é o quociente de encefalização (EQ, do inglês *encephalization quotient*) que leva em conta as relações alométricas entre o aumento da massa corporal e da massa cerebral, sendo capaz de descrever o quanto a massa cerebral real desvia do esperado dado um tamanho corporal (JERISON, 1973). Inicialmente, propôs-se que o expoente que descreveria o crescimento da massa cerebral em relação ao aumento da massa corporal seria de aproximadamente $\frac{2}{3}$. Isso significa que, à medida que a massa corporal aumenta, a massa cerebral aumentaria em uma proporção de $\frac{2}{3}$ da mesma (GOULD, 1975; JERISON, 1985). Trabalhos posteriores, no entanto, reportaram que o expoente no valor de $\frac{3}{4}$ melhor descreveria a relação de aumento entre as duas massas (CAPELLINI; VENDITTI; BARTON, 2010). Ainda, o trabalho de Grabowski *et al.*, 2016, descreve um expoente de escalonamento de $\frac{3}{5}$ para o aumento específico da massa cerebral dos primatas. Para alcançar esse valor, os pesquisadores fizeram uso de métodos comparativos

filogenéticos, descritos na próxima seção, os quais visam considerar a história filogenética compartilhada entre espécies para realizar estudos de associação em organismos filogeneticamente relacionados (FELSENSTEIN, 1985).

Outra medida tradicionalmente utilizada para descrever o tamanho cerebral relativo é definida pela estimativa dos resíduos da massa cerebral observada. Para obter esse valor, estima-se a regressão linear entre os valores logarítmicos da massa cerebral e corporal para as espécies de interesse. O resíduo consiste na distância entre o valor real da massa cerebral de uma espécie e o valor predito da mesma de acordo com a equação do modelo linear. Espécies com valores negativos têm uma massa cerebral menor do que a esperada, enquanto espécies com valores positivos possuem massa cerebral maior do que a esperada (DEANER *et al.*, 2007; SHULTZ, 2010).

Dentro do clado dos primatas, existe uma tendência ao aumento do tamanho cerebral, tendência esta que é ainda mais exacerbada na linhagem dos hominíneos — composta pelos gêneros existentes atualmente *Homo* e *Pan* (MELCHIONNA, M *et al.*, 2019; PÜSCHEL *et al.*, 2024; VENDITTI; BAKER; BARTON, 2024). Quanto às possíveis razões que levaram ao aumento da massa cerebral, existem algumas hipóteses. Entre elas está a hipótese do cérebro social, que postula que a complexidade de diferentes estruturas sociais encontradas na ordem levou à seleção de cérebros maiores capazes de lidar satisfatoriamente com os requisitos de convivência de grupos sociais complexos (DUNBAR; SHULTZ, 2017). Uma outra hipótese postula que uma dieta frugívora implicaria na seleção de cérebros maiores em resposta ao desafio de encontrar frutos para se alimentar comparada à uma dieta herbívora dependente de folhas, um alimento mais facilmente encontrado e portanto menos “cognitivamente exigente” (DECASIEN; WILLIAMS; HIGHAM, 2017). Entretanto, levando em conta a complexidade que o cérebro apresenta e que uma hipótese não necessariamente exclui a outra, atualmente não há consenso sobre a influência que cada um desses possíveis cenários teve sobre o fenótipo que observamos hoje (CHAMBERS; HELDSTAB; O’HARA, 2021; DECASIEN; BARTON; HIGHAM, 2022; GRABOWSKI *et al.*, 2023).

O desenvolvimento do cérebro humano segue essencialmente o mesmo padrão de desenvolvimento cerebral dos demais mamíferos durante a sua embriogênese. Entretanto, observa-se que quase todas as etapas se apresentam prolongadas em relação às demais espécies. Um exemplo consiste no maior período de proliferação de células progenitoras, resultando em um número maior de células geradas e, conseqüentemente, um maior volume final (POLLEN *et al.*, 2023; SILBEREIS *et al.*, 2016). Estágios de diferenciação e maturação

de neurônios também se mostram prolongados, levando tanto a neurônios maiores como ao aprimoramento de neuritos e sinapses (ZHOU; SONG; MING, 2024).

Muitos dos genes descritos como tendo papel no desenvolvimento cerebral dos humanos estão relacionados com a proliferação celular (TYNIANSKAIA; HEIDE, 2024). Entre estes, ressaltamos os genes ARHGAP11B (FLORIO *et al.*, 2016; NAMBA *et al.*, 2020), NOTCH2NL (FIDDES *et al.*, 2018; SUZUKI *et al.*, 2018) e TBC1D3 (HOU *et al.*, 2021). Todos estes genes levam ao aumento da proliferação celular de células progenitoras neurais (NPC) através de diferentes mecanismos. Estes são alguns exemplos de elementos descritos e validados quanto a sua relação com a variação da massa cerebral, no entanto, certamente ainda há elementos a serem identificados. Na busca por estes elementos, o método comparativo de maneira geral, e a genômica comparativa, em particular, são importantes ferramentas para se estudar associações entre fenótipos e genótipos em contextos históricos evolutivos.

1.3. O método comparativo e suas aplicações em genômica

Estudos comparativos são lugar comum na biologia, uma vez que seres vivos compartilham ancestrais comuns entre si — sejam ancestrais que viveram há milhares, milhões ou bilhões de anos (CORNWALLIS; GRIFFIN, 2024). O artigo *Phylogenies and the Comparative Method* de Felsenstein revolucionou a metodologia comparativa na biologia evolutiva ao introduzir o conceito de contrastes filogenéticos independentes (FELSENSTEIN, 1985). Publicado em 1985, esse método passou a exigir que os pesquisadores considerassem as relações evolutivas entre as espécies ao analisar correlações entre características, em vez de ignorá-las, como faziam os métodos estatísticos tradicionais. O trabalho destacou a importância de incorporar informações filogenéticas em estudos comparativos, a fim de evitar vieses estatísticos decorrentes da ancestralidade comum.

O método proposto por Felsenstein, os contrastes filogenéticos independentes (*Phylogenetically Independent Contrasts - PIC*), transforma os dados observados em espécies atuais em diferenças (ou, na nomenclatura estatística, em contrastes) entre pares de ramos adjacentes em uma árvore filogenética. Esses contrastes são calculados para cada evento de especiação (isto é, em cada nó da árvore) e são normalizados pelo comprimento dos ramos de uma filogenia de espécies onde estes representam tempo ou divergência evolutiva. Por exemplo, se duas espécies irmãs, A e B, possuem valores de uma característica quantitativa iguais a 10 e 14, respectivamente, e ramos de comprimento igual em uma árvore filogenética

de espécies, então o contraste desse fenótipo desde o evento de especiação seria a diferença entre os valores ($14 - 10 = 4$), padronizada pela variância esperada (proporcional ao comprimento dos ramos), o que permite comparabilidade entre contrastes de diferentes partes da árvore.

Cada contraste deve ser interpretado como uma estimativa da mudança associada a um único evento de especiação, que, por definição, é considerado estatisticamente independente dos demais, uma vez que os eventos de especiação em si são independentes. Ao analisar essas diferenças ao longo de toda a filogenia, é possível aplicar testes estatísticos tradicionais (como regressão linear) de forma apropriada, sem violar pressupostos de independência entre observações, removendo assim o viés causado pelo uso dos métodos comparativos tradicionais que ignoram a filogenia. Assim, o PIC permite investigar associações entre características de forma estatisticamente robusta ao respeitar a história evolutiva compartilhada entre as espécies. Com o surgimento e posterior aprimoramento de técnicas moleculares que permitem o sequenciamento de genomas com maior acurácia e de forma cada vez mais rápida, veio o advento do campo da genômica comparativa. A genômica comparativa é um campo de estudo que consiste essencialmente em comparar genomas de diferentes espécies de modo a caracterizar semelhanças e divergências entre as sequências — e se possível estabelecer qual a significância funcional dessas variações, considerando o padrão de ancestralidade comum. Ao comparar diferentes genomas é possível, por exemplo, inferir a função biológica de um gene em uma espécie baseado em sua anotação proveniente de uma outra espécie, caso estes sejam suficientemente similares. Outra possibilidade é inferir quais mudanças podem estar relacionadas a variações fenotípicas em diferentes espécies. A comparação pode ser feita entre diferentes elementos do genoma, entre eles regiões codificadoras de proteínas, domínios de proteínas, regiões não codificantes e elementos repetitivos (ALFÖLDI; LINDBLAD-TOH, 2013; NIH, 2020; HARDISON, 2003).

Por meio da genômica comparativa, é que houve a descrição das “regiões humanas aceleradas” (HAR, do inglês *Human Accelerated Regions*), que se mostram conservadas nos demais vertebrados, mas que apresentam acelerada taxa de substituições nucleotídicas nos seres humanos (POLLARD *et al.*, 2006). Houve também a descrição de inúmeros genes sob seleção positiva, famílias gênicas expandidas e regiões aceleradas na linhagem ancestral dos Simiiformes (SHAO *et al.*, 2023). Naturalmente, ferramentas de genômica comparativa foram utilizadas para se estudar o cérebro dos primatas (POLLEN *et al.*, 2023), e os genes citados mais acima (ARHGAP11B, NOTCH2NL e TBC1D3) são um bom exemplo do conhecimento derivado destes métodos.

A anotação funcional, que visa atribuir significado biológico às sequências, é altamente relevante para a identificação de genes associados a fenótipos complexos, como o desenvolvimento cerebral (STEIN, 2001). Um popular *software* para anotação de sequências codificadoras é o InterProScan (BLUM *et al.*, 2025), que integra em si 13 bancos de dados, com informações de famílias e superfamílias de proteínas e informações afins. Entre estes banco de dados, está o Pfam (MISTRY *et al.*, 2021), banco de dados que traz informações sobre domínios de proteínas — unidades básicas que constituem as proteínas e são capazes de se enovelar e funcionar independentemente (WANG *et al.*, 2021). A anotação também retorna informações sobre os termos de ontologia do Gene Ontology (GO), (THE GENE ONTOLOGY CONSORTIUM *et al.*, 2023) que descrevem a característica funcional atribuída a um certo gene independente da similaridade das sequências compartilhadas.

Por fim, os estudos de genômica comparativa vêm fornecendo conhecimentos inestimáveis acerca da evolução do cérebro. Com o sequenciamento e disponibilização de novos genomas realmente completos e de altíssima qualidade, os estudos tendem a se tornarem cada vez mais robustos — ainda que o estabelecimento da conexão entre um determinado genótipo e o fenótipo que ele gera continua a ser um desafio, especialmente para fenótipos complexos onde diversos fatores interagem entre si. Combinar genomas de alta qualidade com dados fenotípicos e informações filogenéticas pode se mostrar promissor e trazer novos conhecimentos acerca da evolução da massa cerebral dos Primatas.

2. OBJETIVOS

2.1. Objetivo geral

Buscar associações entre variações genômicas (número relativo de cópias de domínios proteicos e substituições convergentes de aminoácidos) associadas à variação da massa cerebral relativa em primatas.

2.2. Objetivos específicos

- Obter dados fenotípicos de massa cerebral e corporal para espécies de primatas.
- Obter genomas de alta qualidade de primatas disponíveis de maneira pública.
- Obter árvores filogenéticas que reflitam a história evolutiva de primatas.
- Integrar dados filogenéticos, genotípicos e fenotípicos para a avaliação de correlações entre as variáveis genótipo e fenótipo.

- Verificar se existem domínios de proteínas cuja frequência apresenta variação de forma linear com a variação de fenótipos quantitativos.
- Verificar se existem funções biológicas cuja frequência apresenta variação de forma linear com a variação de fenótipos quantitativos.
- Identificar substituições convergentes de aminoácidos presentes em primatas com alto EQ e ausentes nas demais.

3. MATERIAIS & MÉTODOS

3.1. Obtenção de genomas

Os genomas utilizados provêm majoritariamente do banco de dados *RefSeq* do NCBI. O RefSeq (abreviação de *Reference Sequences* - sequências de referência) é um banco de dados público e disponibiliza sequências de referências que, para serem consideradas tal, devem cumprir requerimentos relacionados a sua qualidade (O'LEARY *et al.*, 2016). Também obtivemos genomas a partir do artigo de Shao *et al.*, 2023, que reportou o sequenciamento de 27 novos genomas de primatas até então não publicados no *Refseq*.

A partir dos genomas encontrados, obtivemos os proteomas não-redundantes preditos, onde somente a isoforma mais longa para cada locus codificador é mantida. Essa sumarização por locus é feita para evitar possíveis vieses decorrentes do fato de que locus são mais ou menos caracterizados e, portanto, podem apresentar discrepância no número de isoformas descritas.

3.2. Controle de qualidade

O controle de qualidade dos genomas é uma etapa que visa garantir a integridade e qualidade geral dos dados, e mesmo aqueles originados a partir do *RefSeq* foram também submetidos a esta etapa. Fizemos uso do software BUSCO v5.2.2 (*Benchmarking Universal Single-Copy Orthologs*) que avalia a presença ou ausência de genes ortólogos de cópia única quase universais nos genomas analisados. Sua métrica de qualidade leva em conta não somente aspectos técnicos da montagem de um genoma e sim parte da história evolutiva dos organismos avaliados. Para distinguir os ortólogos de cópia única, o BUSCO conta com bancos de dados curados destes genes que devem estar presentes em grupos específicos (e.g. genes presentes em quase todos os vertebrados, artrópodes ou primatas). O descritivo “quase universais” é referente ao fato de que, durante a construção dos datasets pelos

desenvolvedores do BUSCO, não era necessário que os ortólogos de cópia única estivessem presentes em absolutamente todas as espécies analisadas, bastava que estivessem presentes em pelo menos 90% destas espécies. Esta é uma forma de considerar eventuais ausências de um ortólogo devido, por exemplo, a erros técnicos no momento de sequenciamento e montagem de um genoma, e não a sua real ausência (MANNI *et al.*, 2021).

3.3. Anotação genômica

A anotação *de novo* dos proteomas não-redundantes foi feita utilizando o InterProScan (BLUM *et al.*, 2025), uma ferramenta capaz de prover informações relevantes quanto à função e classificação das proteínas. O InterProScan integra diferentes bancos de dados relacionados a anotação funcional e faz uso de modelos preditivos — descritos também pelo termo “assinaturas” — para, com base nas informações providas por cada banco de dados, classificar proteínas em diferentes famílias, superfamílias, encontrar diferentes domínios proteicos, e também descrever demais sítios de importância.

A anotação funcional pelo InterProScan pode levar a uma certa redundância nas informações providenciadas visto que um mesmo sítio pode ser anotado por diferentes bancos de dados. A forma com a qual o *software* lida com tal circunstância é a atribuição de um termo unificador que agrega as diferentes informações de uma mesma sequência sob um mesmo identificador — o código IPR seguido de uma sequência numérica.

3.4. Dados fenotípicos e imputação de dados faltantes

Os dados fenotípicos fundamentais para a nossa análise (a massa corporal e cerebral das espécies) foram obtidos através da curadoria da literatura científica. As informações sobre a massa corporal foram obtidas no banco de dados AnAge (DE MAGALHÃES *et al.*, 2023) e as massas cerebrais obtidas a partir do artigo de Herculano-Houzel, 2019.

É razoável esperar que, ao se trabalhar com um grupo relativamente grande de organismos, não encontramos informações fenotípicas disponíveis para todas as espécies que possuam o genoma disponível. Uma possível solução seria excluir das análises as espécies para as quais os dados de interesse não existem. Porém, uma outra alternativa existente e que seguimos é a imputação de dados faltantes de modo a evitar a exclusão de espécies nas análises comparativas. (PENONE *et al.*, 2014)

A imputação foi feita utilizando o pacote *Rphylopars*, desenvolvido na linguagem de programação R (GOOLSBY; BRUGGEMAN; ANE, 2024; R CORE TEAM, 2024). Este

pacote executa a imputação dos dados faltantes de acordo com um determinado modelo evolutivo — *Brownian Motion*, no caso deste estudo. Como parâmetro para a estimativa dos dados faltantes e entrada para a função utilizada, fornecemos os dados fenotípicos de cerca de 200 espécies de primatas que recuperamos na literatura já mencionada e uma árvore filogenética que reporta a relação entre estas espécies e aumenta a robustez da imputação.

De posse dos dados fenotípicos (massa corporal e cerebral), utilizamos ambos os valores para calcular o EQ, obtido a partir da equação 1 (GRABOWSKI; VOJE; HANSEN, 2016), e os valores residuais do tamanho cerebral, obtidos ao calcular uma regressão linear da massa cerebral em relação à massa corporal (valores logaritmizados no caso da regressão linear).

Equação 1:

$$EQ = \frac{ECV}{e^{\frac{3}{5}(BM) - 1.40}}$$

Todo o processamento dos dados fenotípicos foi executado por meio da linguagem R (R CORE TEAM, 2024).

3.5. Árvores filogenéticas

Para as diferentes etapas do presente trabalho, diferentes árvores filogenéticas de primatas compostas por números distintos de espécies foram utilizadas e as recuperamos a partir da ferramenta online TimeTree of Life. O TimeTree of Life compila informações filogenéticas a partir de milhares de artigos e nele é possível recuperar árvores filogenéticas ultramétricas — onde o comprimento dos ramos equivale ao tempo evolutivo das linhagens, a partir de uma lista de espécies (KUMAR *et al.*, 2022). Para eventuais tratamentos e processamento das árvores obtidas, fizemos uso do pacote *phytools* do R (REVELL, 2024).

3.6. Processamento e adequação de arquivos para uso do CALANGO

Na etapa de análise de correlações, utilizamos o CALANGO, um *software* para análises de genômica comparativa desenvolvido por nosso grupo de pesquisa. Com ele é possível calcular correlações entre genótipos e diferentes fenótipos de relevância em modelos que integram dados evolutivos, genômicos e fenotípicos, de modo a incorporar a ancestralidade comum no procedimento de inferência estatística (HONGO *et al.*, 2023).

Os arquivos de entrada do CALANGO consistem em 1) uma árvore filogenética ultramétrica; 2) arquivos de anotação funcional referentes a cada espécie; 3) um arquivo com os valores dos dados fenotípicos de interesse. A criação destes arquivos é feita utilizando scripts internos que têm como entrada os arquivos de anotação de cada genoma e uma *flag* que indica qual *feature* (e.g. termo Pfam, GO ou IPR) deve ser extraída destes arquivos. O dicionário com a identificação dos termos Pfam ou IPR mais a descrição desses termos também é criado utilizando os arquivos de anotação.

Um procedimento importante executado pelo CALANGO compreende estabelecer a independência estatística entre os dados fenotípicos e genômicos analisados. Grande parte das análises estatísticas tradicionais têm como pressuposto a independência entre os dados, suposição essa que não é real visto que as espécies apresentam ancestralidade compartilhada e portanto são relacionadas entre si, sendo que essas relações são representadas pelas filogenias (IVES; ZHU, 2006). Para contornar este impasse o CALANGO computa as correlações considerando a filogenia através da metodologia dos contrastes filogeneticamente independentes (PIC), proposta por Felsenstein (FELSENSTEIN, 1985). O método essencialmente transforma os valores fenotípicos dos ramos — dados não independentes — em contrastes, valores que derivam da diferença entre os valores observados em espécies irmãs e são independentes entre si. Além disso, o CALANGO também calcula os usuais coeficientes de correlação como pearson e spearman.

Dada a natureza das análises de busca por associações entre centenas ou milhares de elementos e um fenótipo quantitativo, surge o cenário do teste de múltiplas hipóteses (GOEMAN; SOLARI, 2014). Incontáveis testes individuais são realizados durante a busca, e em decorrência do enorme número de hipóteses testadas simultaneamente, espera-se que, ao acaso, possam surgir associações significativas que na realidade não são (resultados falso-positivos). A título de exemplo: ao serem executados 20 testes distintos (cada teste corresponde a um termo analisado) para sabermos se há ou não correlação entre o termo e a variação da massa cerebral, e fazendo uso de um valor $p < 0,05$, pelo menos uma associação será falso-positiva, ou seja, uma associação significativa onde não há associação de fato. Ao considerar a enorme escala dos dados genômicos utilizados, onde dezenas de milhares de genes são testados para associações, espera-se que um número expressivo de termos erroneamente significativos sejam encontrados. Uma das correções mais habitualmente utilizadas para lidar com este cenário é a correção de Bonferroni, que é relativamente simples. Em um cenário onde 1000 hipóteses estão sendo testadas e o valor considerado de significância para rejeitar a hipótese nula é 0,05, a correção é feita de modo que o novo valor

de significância seja o valor de significância dividido pelo número de hipóteses testadas, neste caso, o novo valor de significância seria $0,05/1000$, que resulta em 5×10^{-5} . Entretanto, a correção de Bonferroni se mostra exageradamente estridente em estudos de associação genômica, uma vez que ela corrige considerando a possibilidade de uma única associação falso-positiva. Outra abordagem menos estridente é o FDR (do inglês *False Discovery Rate*). Essa é uma medida estatística utilizada para controlar a proporção de resultados falso-positivos (erros do Tipo I) em testes múltiplos de hipóteses. Ela é especialmente útil em contextos nos quais diversos testes são realizados simultaneamente, como em estudos de expressão gênica em larga escala ou em estudos de associação genômica ampla (*genome-wide association studies*). O objetivo do FDR é equilibrar o compromisso entre o controle de erros do Tipo I e a manutenção do poder estatístico. Nesse estudo, utilizamos a abordagem FDR conforme implementada no CALANGO. Após a correção para o cenário de teste múltiplo de hipóteses, consideramos como associados os termos que possuem valores p corrigidos < 0.05 para as análises que utilizam os contrastes filogeneticamente independentes.

3.7. Alinhamento de genes ortólogos

A etapa de análise de substituições convergentes, descrita na seção seguinte, necessita de alinhamentos de múltiplas sequências de aminoácidos dos genes ortólogos. Para a descoberta dos ortólogos utilizamos o *software* Orthofinder (EMMS; KELLY, 2019), que como *input* recebe sequências proteicas das espécies a serem avaliadas (proteomas não-redundantes), e produz como resultados grupos de genes homólogos. Descobertos os genes ortólogos, partimos para o alinhamento destes, utilizando o *software* MAFFT com os parâmetros padrões (KATO; STANDLEY, 2013).

3.8. Análise de sítios de substituições convergentes de aminoácidos

A análise de sítios de substituições convergentes de aminoácidos que utilizamos é baseada na metodologia de Farré *et al.*, 2021. O estudo busca por substituições convergentes de aminoácidos compartilhadas entre os mamíferos mais longevos e têm como foco genes associados com a longevidade. No presente estudo, procuramos por genes que apresentam substituições convergentes de aminoácidos (ou CAAS, do inglês *convergent amino acid substitutions*) compartilhadas entre os primatas com maiores valores de EQ. O *software* CAAStools (BARTERI *et al.*, 2023), publicado pelo grupo do trabalho mencionado, funciona procurando em alinhamentos múltiplos de sequência (MSA, do inglês *multiple sequence*

alignment) de genes ortólogos codificadores de proteínas por substituições que são compartilhadas entre espécies de dois grupos distintos pré-estabelecidos pelo usuário. Uma posição é considerada como uma substituição convergente de aminoácido quando ela é compartilhada pelas espécies de um grupo e não estão presentes no outro grupo.

A seleção pelas espécies que compõem os grupos da etapa de descoberta foi feita ordenando as espécies de acordo com o seu quociente de encefalização e escolhendo 5 valores nos dois extremos (totalizando 10 espécies). No entanto, como os valores extremos do EQ geralmente se concentram em determinadas famílias — por exemplo, os maiores valores de EQ se encontram na família Hominidae — decidimos por permitir que apenas um representante de cada família estivesse presente no grupo de descoberta. A tabela 1 mostra quem compõe os dois grupos utilizados, o grupo das espécies com os maiores EQ é referido como o *foreground* (FG) e o grupo com os menores valores de EQ é referido como *background* (BG).

Tabela 1: Espécies do grupo de descoberta de CAAS.

Foreground: espécies com maiores valores de EQ	Background: espécies com menores valores de EQ
<i>Homo sapiens</i>	<i>Nycticebus bengalensis</i>
<i>Sapajus apella</i>	<i>Galago moholi</i>
<i>Macaca nemestrina</i>	<i>Lemur catta</i>
<i>Hylobates moloch</i>	<i>Propithecus coquereli</i>
<i>Ateles geoffroyi</i>	<i>Microcebus murinus</i>

Executado o CAAStools, um arquivo de resultado relativo a cada um dos arquivos MSA é gerado, a etapa seguinte envolveu o processamento destes arquivos a fim de obter um único arquivo com a informação do nome dos genes onde há a presença de CAAS.

3.9. Análise de enriquecimento

As análises de enriquecimento são uma forma comum de se integrar e interpretar dados gerados em experimentos de ômicas, uma vez que estes experimentos usualmente geram uma grande quantidade de listas gênicas, e focar em genes únicos pode se mostrar um grande desafio. Com as análises de enriquecimento, é possível identificar processos e vias bioquímicas com maior representação do que o esperado — enriquecidas — em um certo

conjunto de genes. Esse tipo de análise também se mostra relevante quando diversos genes, desempenhando funções semelhantes, contribuem individualmente com efeitos pequenos no fenótipo de interesse, porém, como atuam em grande número e por mecanismos parecidos, seu impacto combinado pode ser significativo, algo que uma análise focada em genes isolados poderia não detectar (HUNG *et al.*, 2012; SUBRAMANIAN *et al.*, 2005).

Para a análise de enriquecimento utilizamos a ferramenta *online WebGestalt* (WEB-based GENE SeT AnaLysis Toolkit) (ELIZARRARAS *et al.*, 2024). Os parâmetros utilizados foram: *Method of Interest: Over-Representation Analysis; Organism of Interest: Homo sapiens; Functional Database: geneontology, Biological Process; Select Reference Set: genome-protein coding e Significance Level: FDR < 0.05.*

4. RESULTADOS & DISCUSSÃO

4.1. Obtenção de genomas

A busca por genomas de referência de primatas nos retornou, ao todo, 57 genomas: 31 provenientes do banco de dados *RefSeq* e os demais 26 novos genomas sequenciados por SHAO *et al.*, 2023, e que até então não estavam classificados como dados *RefSeq*. Estes pertencem a 15 diferentes famílias de primatas e é possível visualizar as espécies obtidas na figura 1. O próximo passo consistiu na obtenção dos proteomas não redundantes a partir dos genomas encontrados utilizando um script em *Perl* rotineiro em nosso laboratório, todos os 57 apresentaram por volta de 20 mil sequências únicas. Com os proteomas não-redundantes disponíveis, partimos para a análise de completude usando o BUSCO.

4.2. Controle de qualidade e anotação dos genomas

Utilizamos o BUSCO versão 5.2.2, no modo *protein* e com o dataset de ortólogos *primates_odb10* que conta com 13780 ortólogos de cópia única comum a (quase) todos os genomas de primatas utilizados na construção do dataset.

Dos 57 genomas encontrados, 38 apresentaram completude maior que 90% (*Homo sapiens* com o maior índice de completude com 99%), 11 apresentaram completude entre 87 e 90%, os 8 genomas restantes apresentaram completude menor que 85%, sendo que apenas um destes tinha a completude abaixo dos 80% (especificamente 78,6% da espécie *Pithecia pithecia*, um primata da América do Sul). Na etapa das análises de correlação pelo

CALANGO, o *cutoff* de qualidade escolhido foi o de 85% de completude, totalizando 49 genomas.

A figura 1 foi construída utilizando — com alterações — o script *busco_figure.R* originalmente fornecido pelos desenvolvedores do BUSCO, e sumariza o resultado da avaliação de qualidade. Ela mostra a porcentagem de ortólogos encontrados em cópia única e os ortólogos duplicados, assim como o valor dos ortólogos faltantes e fragmentados para cada genoma obtido.

Em relação à anotação funcional dos genomas, o InterProScan nos retornou ~6900 termos Pfam únicos, ~15800 termos IPR únicos e ~3500 termos GO únicos.

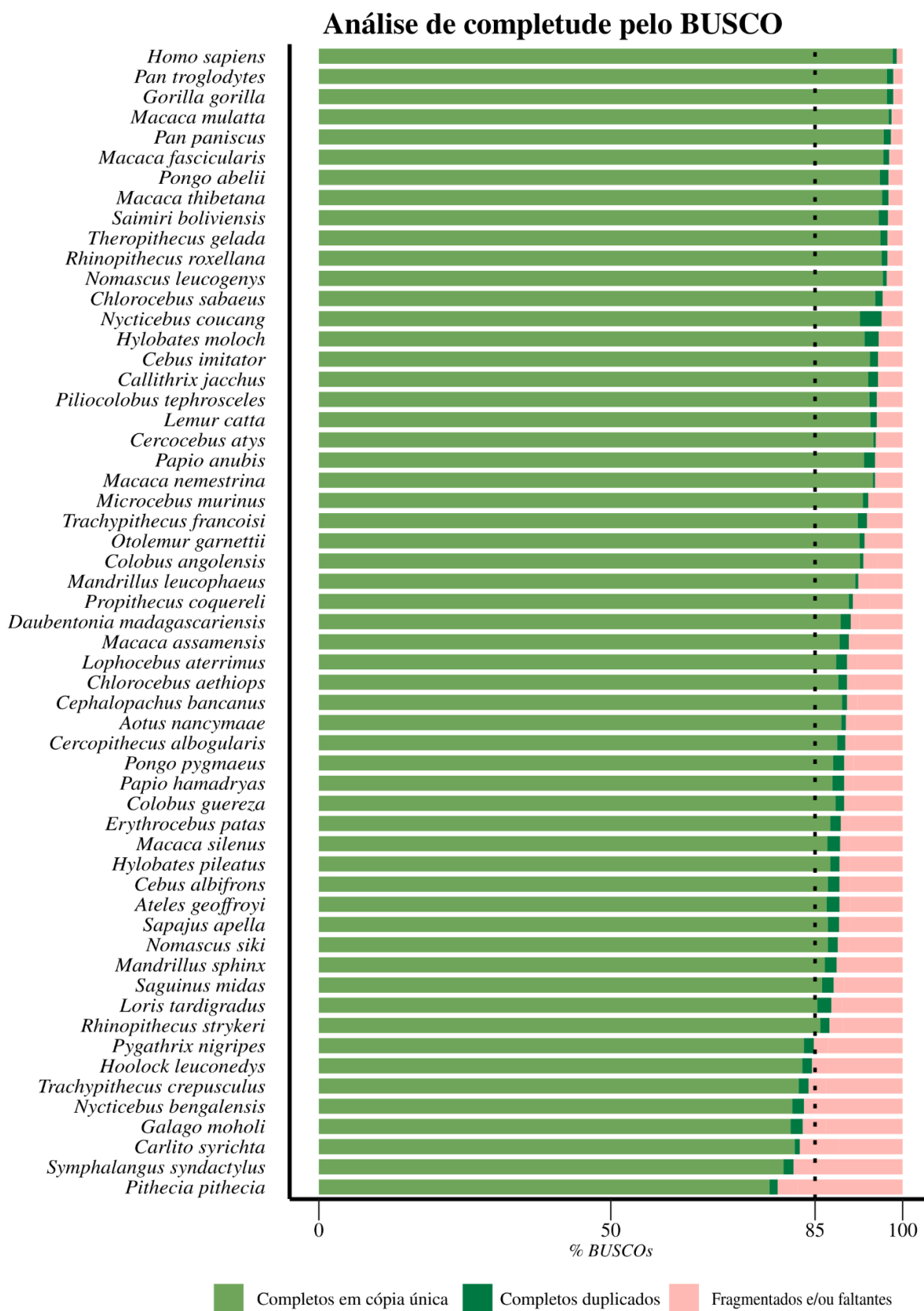


Figura 1: Controle de qualidade dos genomas utilizados. O valor de completude escolhido no trabalho foi de 85%, indicado pela linha pontilhada.

4.3. Dados fenotípicos

As massas cerebral e corporal foram obtidas a partir de dois bancos de dados que reúnem informações compiladas a partir de diferentes referências. O conjunto de dados que continha as informações referentes às massas cerebrais (HERCULANO-HOUZEL, 2019) listava um total de 174 espécies de primatas, das quais 84 apresentavam informação sobre a massa cerebral. O conjunto de dados com os valores de massa corporal (DE MAGALHÃES *et al.*, 2023) listava também coincidentemente 174 espécies, porém não as mesmas. Ao integrarmos os dois conjuntos de dados em um único conjunto, obtivemos um total de 218 espécies de primatas para as quais possuímos valores de massa cerebral e/ou massa corporal.

4.4. Imputação de dados faltantes

De posse das 218 espécies com informações fenotípicas, obtivemos uma árvore filogenética para estas espécies. A árvore se faz necessária durante o processo de imputação de dados faltantes, porque o Rphylopars leva em conta a relação entre as espécies no momento da imputação de modo a melhorar sua acurácia. Em relação a estas espécies, 177 delas estavam presentes na ferramenta do TimeTree of Life e foram utilizadas nas análises posteriores. Destas, 24 delas foram excluídas do dataset, pois não havia dados de massa corporal, cerebral e nem genoma disponível (a razão destas espécies estarem nos nossos dados compilados se dá pelo fato de que os dois bancos de dados inicialmente utilizados as listava junto com informações que não fizemos uso). Também excluimos as espécies para as quais tínhamos informações sobre a massa corporal, mas não sobre a massa cerebral e nem o genoma, visto que estas não contribuiriam para as análises posteriores. Ao final dessa etapa, nosso conjunto de dados a ser utilizado para a imputação dos dados faltantes contava com 102 espécies. Para as espécies com genomas disponíveis, imputamos as massas corporais e cerebrais para 16 e 28 destas, respectivamente. De maneira geral, consideramos os resultados da imputação coerentes quando comparados às espécies filogeneticamente próximas com dados reais disponíveis (figura 2).

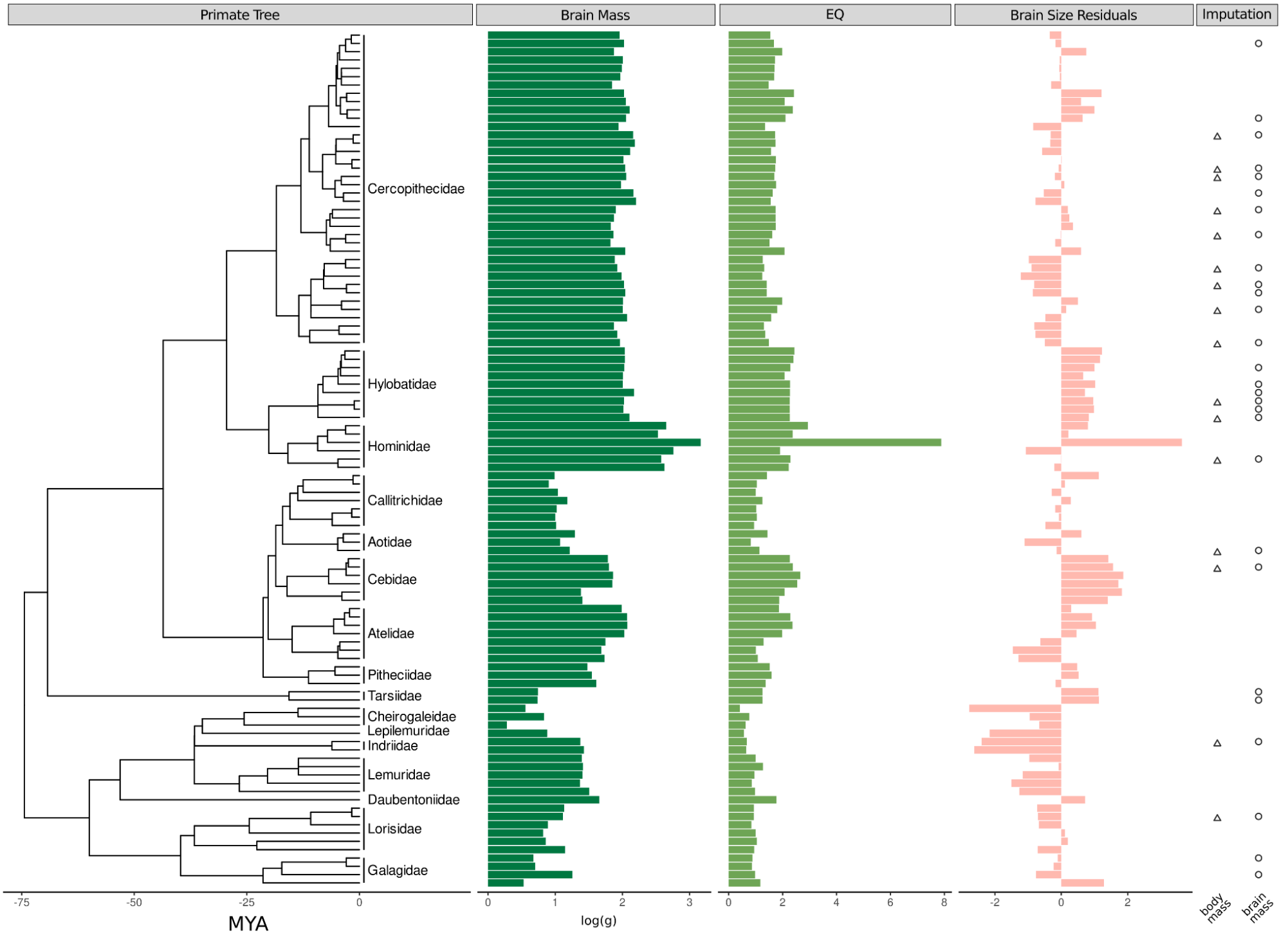


Figura 2: Dados fenotípicos e imputação de dados faltantes. O painel *Primate Tree* representa as espécies utilizadas na etapa de imputação. As espécies presentes na árvore são aquelas com informação sobre massa corporal e cerebral e espécies sem essas informações, mas com o genoma disponibilizado. Os painéis *Brain Mass*, *EQ* e *Brain Size Residuals* representam, respectivamente, a massa cerebral (em log(gramas)), o valor do quociente de encefalização e os resíduos da massa cerebral correspondente às espécies representadas nos ramos da árvore do primeiro painel. O último painel, *Imputation*, traz a informação se as informações de massa cerebral e corporal — necessárias para calcular os índices fenotípicos — são reais ou imputadas (Δ massa corporal imputada; \circ massa cerebral imputada).

Tabela 2: Dados fenotípicos dos primatas utilizados nas análises.

Espécie	resíduos	EQ	Espécie	resíduos	EQ
<i>Aotus nancymaae</i>	-0.14	1.14	<i>Macaca silenus</i>	0.64	2.11
<i>Ateles geoffroyi</i>	0.92	2.29	<i>Macaca thibetana</i>	-0.17	1.67
<i>Callithrix jacchus</i>	1.13	1.42	<i>Mandrillus leucophaeus</i>	-0.53	1.63
<i>Cebus albifrons</i>	1.42	2.27	<i>Mandrillus sphinx</i>	-0.77	1.56
<i>Cebus imitator</i>	1.56	2.37	<i>Microcebus murinus</i>	-0.66	0.63
<i>Cephalopachus bancanus</i>	1.12	1.25	<i>Nomascus leucogenys</i>	0.99	2.27
<i>Cercocebus atys</i>	-0.19	1.69	<i>Nomascus siki</i>	0.97	2.27
<i>Cercopithecus albogularis</i>	0.20	1.74	<i>Nycticebus coucang</i>	-0.72	0.94
<i>Chlorocebus aethiops</i>	-0.18	1.51	<i>Otolemur garnettii</i>	-0.76	0.98
<i>Chlorocebus sabaeus</i>	-0.02	1.62	<i>Pan paniscus</i>	0.21	2.37
<i>Colobus angolensis</i>	-0.81	1.31	<i>Pan troglodytes</i>	0.80	2.94
<i>Colobus guereza</i>	-0.78	1.36	<i>Papio anubis</i>	-0.32	1.72
<i>Daubentonia madagascariensis</i>	0.72	1.77	<i>Ptilocolobus tephrosceles</i>	-0.50	1.49
<i>Erythrocebus patas</i>	0.59	2.07	<i>Papio hamadryas</i>	-0.33	1.74
<i>Gorilla gorilla</i>	-1.07	1.90	<i>Pongo abelii</i>	-0.01	2.29
<i>Homo sapiens</i>	3.63	7.88	<i>Pongo pygmaeus</i>	-0.21	2.23
<i>Hylobates moloch</i>	1.00	2.29	<i>Propithecus coquereli</i>	-2.40	0.67
<i>Hylobates pileatus</i>	1.02	2.28	<i>Rhinopithecus roxellana</i>	-0.85	1.41
<i>Lemur catta</i>	-1.50	0.85	<i>Rhinopithecus strykeri</i>	-0.82	1.41
<i>Lophocebus aterrimus</i>	-0.08	1.73	<i>Saguinus midas</i>	-0.48	0.94
<i>Loris tardigradus</i>	0.11	1.00	<i>Saimiri boliviensis</i>	1.82	2.07
<i>Macaca assamensis</i>	-0.35	1.55	<i>Sapajus apella</i>	1.72	2.54
<i>Macaca fascicularis</i>	-0.30	1.48	<i>Theropithecus gelada</i>	-0.58	1.57
<i>Macaca mulatta</i>	-0.04	1.69	<i>Trachypithecus francoisi</i>	0.50	1.99
<i>Macaca nemestrina</i>	1.00	2.38			

4.5. Árvores filogenéticas

Além da árvore já mencionada, mais duas árvores filogenéticas foram utilizadas ao longo da execução do presente estudo. A primeira delas era composta pelas espécies de primatas das quais encontramos informações da massa cerebral e/ou corporal, independente se estas tinham o genoma disponível ou não, somadas com as espécies que encontramos o genoma — independente se elas tinham informação fenotípica ou não. A segunda árvore, era composta pelas espécies que tinham a porcentagem de completude de BUSCO maior que 85%, nosso filtro de qualidade, e que seriam utilizadas nas análises do CALANGO.

4.6. Análise de correlações CALANGO

De posse dos arquivos de entrada utilizados pelo CALANGO, realizamos as análises de correlação genótipo-fenótipo. Foram feitas, ao todo, 6 análises: três delas empregando como fenótipo quantitativo o quociente de encefalização e outras três utilizando os resíduos. O programa CALANGO oferece uma grande flexibilidade na modelagem da informação codificada em um genoma. Com esse programa é possível representar diferentes regiões de um genoma, tais como regiões promotoras, domínios conservados ou grupos de genes homólogos. Adicionalmente, estes elementos genômicos podem ser anotados utilizando dicionários que refletem diversos aspectos biológicos, tais como relações de homologia ou similaridade funcional. Assim, o CALANGO utiliza pares de elementos genômicos e seus termos anotadores na busca por associações. Para cada um dos dois fenótipos utilizados, representamos a informação genômica em três níveis: 1) domínios de proteína (representados por pares de domínios conservados distintos identificados nos proteomas não-redundantes anotados utilizando o banco de dados Pfam), funções biológicas (representadas por genes anotados para seus termos GO, quando disponíveis) e identificadores IPR (representando expansões de genes contendo regiões homólogas conservadas).

As seções seguintes trazem alguns dos termos encontrados associados aos fenótipos, bem como a sua contextualização biológica. Os resultados completos de termos significativamente associados com a variação dos fenótipos avaliados se encontram no apêndice A ao final do documento.

4.6.1. Domínios de proteínas associados com a variação do EQ

Encontramos 17 domínios de proteínas cuja frequência estavam associadas à variação do EQ (a tabela 3 contém as associações utilizadas na discussão abaixo).

Tabela 3: Exemplos de domínios de proteínas associados com a variação do EQ.

Identificador	q-value	Anotação
PF10520	9.37E-16	Lipid desaturase domain
PF10574	9.37E-16	Arp2/3-interacting proteins Arpin
PF15803	4.54E-10	Zinc-finger of sodium channel modifier 1
PF15682	1.07E-08	Musculoskeletal, temporally activated-embryonic nuclear protein 1
PF07683	2.38E-05	Cobalamin synthesis protein cobW C-terminal domain
PF00128	1.43E-04	Alpha amylase, catalytic domain
PF06758	7.12E-04	Olduvai domain
PF14976	2.22E-03	FAM72 protein
PF02806	5.22E-03	Alpha amylase, C-terminal all-beta domain

O domínio PF14976 (figura 3), descrito como FAM72 protein (*family with sequence similarity 72*), compõe um gene codificador de proteína que apresenta quatro cópias no genoma humano, enquanto os demais primatas apresentam uma única cópia — as cópias duplicadas já foram propostas como um possível biomarcador do *Homo sapiens* (HO *et al.*, 2017). Interessantemente, FAM72 exibe expressão específica em células tronco neuronais (RAMESH *et al.*, 2024). Dada a sua natureza de expressão, este domínio já foi sugerido, junto ao gene SRGAP2 (*SLIT-ROBO Rho GTPase-activating protein 2*), como um *master gene neural* (|-SRGAP2-FAM72A-|) devido ao compartilhamento de um promotor localizado em sua região intergênica. Este par de genes funcionaria como um regulador da proliferação e migração de células tronco neurais (NSC) (HO; KUTZNER; HEESE, 2017, 2019).

A hipótese corrente para o mecanismo de ação do master gene é de que FAM72 estaria ativo e vinculado à proliferação e renovação de células progenitoras neuronais (NPC) (enquanto SRGAP2 estaria inativo), já o SRGAP2 estaria ativo durante a migração e diferenciação das NPC, além de impactar na maturação neural e realizar importante papel durante a sinaptogênese (FOSSATI *et al.*, 2016; HO *et al.*, 2017; RAMESH *et al.*, 2024).

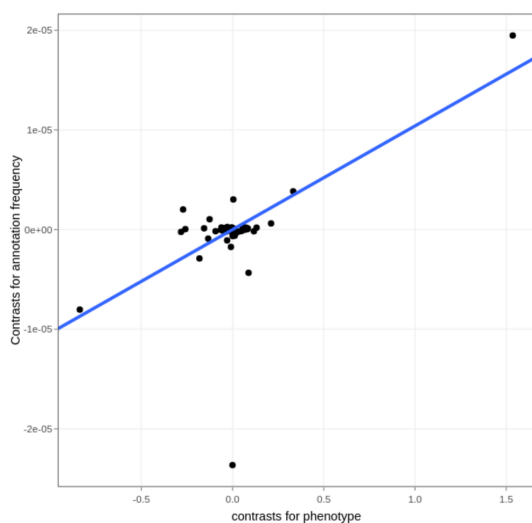


Figura 3: Correlação entre a frequência do domínio FAM72 (PF14976) e o EQ.

O domínio Olduvai (figura 4), anteriormente nomeado de DUF1220 (do inglês *Domain of Unknown Function 1220*), ganhou seu nome definitivo em referência ao sítio arqueológico Olduvai Gorge, no leste do continente africano (SIKELA; VAN ROY, 2018). O domínio Olduvai apresenta uma expansão no número de cópias marcante na linhagem dos grandes símios: *Homo sapiens* apresenta entre 250 e 300 cópias do domínio, chimpanzés cerca de 130, gorilas e orangotangos entre 90 e 100, e os demais primatas por volta de 30 cópias. Fora da ordem dos primatas o número de cópias é ainda menor, com espécies de mamíferos apresentando uma ou nenhuma cópia do domínio (DUMAS *et al.*, 2012; HEFT *et al.*, 2020).

Além da grande variação no número de cópias entre primatas, existe também variação do número de cópias (CNV) dentro da espécie humana. A dosagem do domínio já foi vinculada a diversos caracteres cerebrais, principalmente aqueles que têm alguma relação com o tamanho cerebral (KEENEY, JONATHON G.; DUMAS; SIKELA, 2014). Um número maior de cópias estaria implicado, por exemplo, com maior severidade do transtorno do espectro autista (DAVIS *et al.*, 2014, 2019), enquanto diferentes quantidades de determinados subtipos do domínio podem estar associados a severidade de diferentes sintomas da esquizofrenia (SEARLES QUICK *et al.*, 2015). Há ainda indicativos de uma relação entre a dosagem e fenótipos como micro e macrocefalia (DUMAS *et al.*, 2012).

Embora o mecanismo de ação do domínio Olduvai ainda não seja totalmente elucidado, um estudo de 2015 (KEENEY, J. G. *et al.*, 2015) analisou os padrões de expressão deste domínio ao longo do período gestacional e demonstrou fortes sinais de expressão entre 11 e 15 semanas, período associado com a formação das camadas corticais. Além disso, foi demonstrado que células tronco neurais humanas, quando expostas a expressão exógena do

domínio, mostraram um aumento marcante de proliferação, em comparação ao controle utilizado. O mesmo grupo de pesquisadores, em 2024, descreveu em um artigo ainda em *preprint*, que a função do domínio envolve a regulação negativa da atividade mitocondrial (KEENEY, JONATHON G. *et al.*, 2024).

O domínio Olduvai faz parte da família gênica NBPF (*Neuroblastoma Breakpoint Family*), que recebeu este nome pois sua identificação inicial ocorreu em um paciente com neuroblastoma (VANDEPOELE *et al.*, 2005). De maneira interessante, a localização da maioria dos genes NBPF e seus domínios Olduvai é adjacente a cópias humano-específicas do NOTCH2NL, na região 1q21.1, local no cromossomo 1 do genoma humano que é reconhecidamente instável (DUMAS *et al.*, 2012; FIDDES *et al.*, 2019) e trabalhos recentes demonstraram ação coordenada entre NBPF14 e NOTCH2NLB na manutenção e proliferação de células progenitoras basais (EŞIYOK *et al.*, 2025).

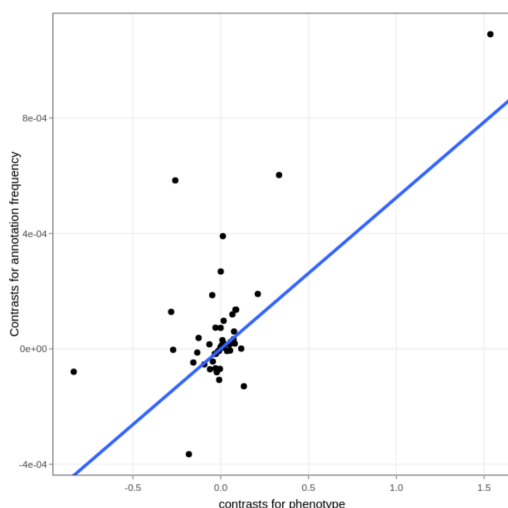


Figura 4: Correlação entre a frequência do domínio Olduvai (PF06758) e o EQ.

A variação no número de cópias (CNV) do gene da α -amilase (figura 5) é documentada tanto entre diferentes espécies de primatas, quanto entre diferentes indivíduos e populações de *Homo sapiens* (JANIAK, 2016). A amilase é uma enzima capaz de digerir amido, um importante carboidrato presente em diversas dietas humanas. Existem duas variações do gene: AMY1, a amilase salivar — presente na saliva e a principal quando se menciona a variação entre populações humanas; AMY2, gene da amilase pancreática (PERRY *et al.*, 2007).

A dosagem de AMY1 está diretamente ligada aos níveis de amilase encontrados na saliva e populações que dependem majoritariamente de dietas derivadas de práticas de

agricultura, ao contrário de, por exemplo, sociedades com dietas pastorais e/ou pesqueiras, apresentam um número maior de cópias da amilase (PERRY *et al.*, 2007). Ainda que o surgimento da agricultura e posterior difusão da mesma tenha contribuído para a fixação do número maior de cópias, o *Homo sapiens* já apresentava cópias duplicadas de amilase antes mesmo de sua expansão fora da África (BOLOGNINI *et al.*, 2024; YILMAZ *et al.*, 2024).

Uma das hipóteses sobre a expansão do tamanho cerebral — não só em primatas — postula que, levando em conta o alto custo energético do cérebro, qualquer aumento da massa cerebral deve estar acompanhado do aumento da obtenção de energia, i.e. obtenção de alimento ou de realocação de energia de outras funções corporais para a manutenção do cérebro. Esta hipótese é denominada *Expensive Brain Hypothesis* (ou Hipótese do “cérebro caro”) (ISLER; VAN SCHAIK, 2009). Apesar da não popularidade desta hipótese, o cérebro é tradicionalmente considerado um órgão altamente custoso do ponto de vista metabólico (HELDSTAB *et al.*, 2022), e há evidências da importância de uma dieta rica em amido como um fator importante a se considerar durante a evolução humana (HARDY *et al.*, 2015). A expansão da alfa-amilase também poderia ser vista sob a ótica da hipótese do aumento do cérebro baseada na dieta frugívora.

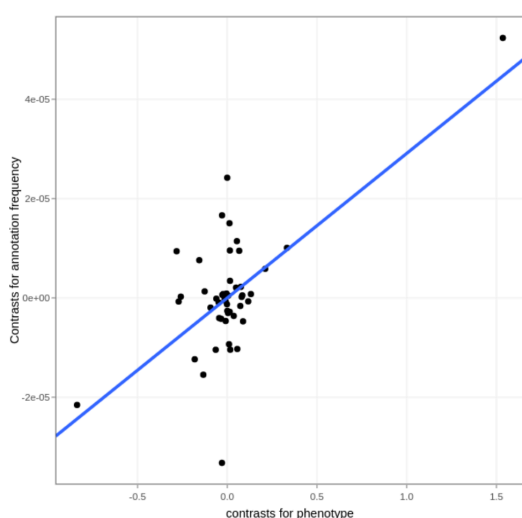


Figura 5: Correlação entre a frequência do Alpha Amylase, catalytic domain (PF00128) e o EQ.

4.6.2. Domínios de proteínas associados com a variação dos resíduos da massa cerebral

Encontramos 10 domínios de proteínas associados com a variação nos resíduos da massa cerebral, todos encontrados também na análise anterior. A tabela 4 traz alguns destes

domínios — dois deles, FAM72 e amilase, foram descritos na seção anterior, e os demais serão discutidos a seguir.

Tabela 4: Exemplos de domínios de proteínas associados com a variação dos resíduos da massa cerebral.

Identificador	q-value	Anotação
PF10520	3.92E-05	Lipid desaturase domain
PF10574	3.92E-05	Arp2/3-interacting proteins Arpin
PF15682	1.95E-04	Musculoskeletal, temporally activated-embryonic nuclear protein 1
PF07683	1.26E-03	Cobalamin synthesis protein cobW C-terminal domain
PF14976	2.22E-03	FAM72 protein
PF02806	5.22E-03	Alpha amylase, C-terminal all-beta domain

As dessaturases lipídicas (figura 6) são enzimas que realizam a insaturação das cadeias longas de ácidos graxos das membranas plasmáticas, em seu estado insaturado, essas cadeias são responsáveis pela fluidez das membranas plasmáticas (YEAGLE, 2016). No cérebro, um órgão com enorme conteúdo lipídico, as cadeias poli-insaturadas de ácidos graxos desempenham diversos papéis em diferentes processos — ainda que o mecanismo da maioria deles ainda não esteja totalmente conhecido — desde neurogênese, função sináptica e inflamação no cérebro (BAZINET; LAYÉ, 2014; BJORNEVIK *et al.*, 2023).

Duas patologias degenerativas e inevitavelmente fatais, a esclerose múltipla lateral e a demência frontotemporal (ELA/DFT), demonstraram em análises de lipidômica um padrão reduzido de fosfolipídios contendo ácidos graxos em sua forma insaturada. A indução da expressão das dessaturases lipídicas em modelos de ELA/DFT em *Drosophila*, e em neurônios iP, acarretou, respectivamente, em um aumento na expectativa de vida e redução da morte celular induzida por estresse (GIBLIN *et al.*, 2025).

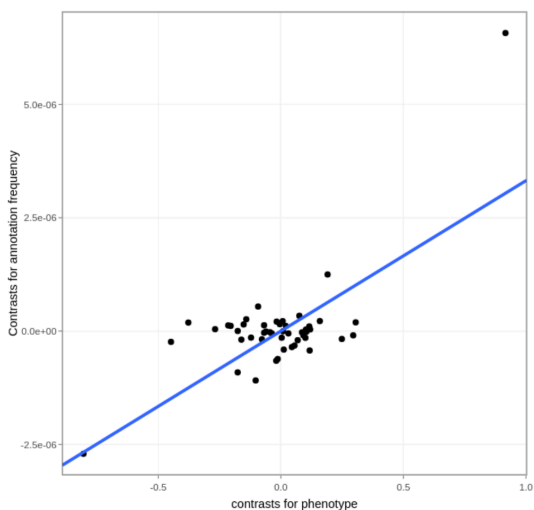


Figura 6: Correlação entre a frequência do domínio Lipid Desaturase (PF10520) e os resíduos de massa cerebral.

4.6.3. Funções biológicas associadas com a variação do EQ

Utilizamos a anotação de pares de genes e seus respectivos termos GO para buscarmos pela expansão de genes funcionalmente equivalentes associada com a variação do EQ, e encontramos 33 associações significativas (tabela 5). Diversas destas são relacionadas com os termos Pfam, tais como os termos GO:0004556 e GO:0016160, respectivamente, *alpha amylase activity* e *amylase activity*, diretamente ligados aos domínios *Alpha amylase, catalytic domain* (PF00128) e *Alpha amylase, C-terminal all-beta domain* (PF02806).

Tabela 5: Exemplos de funções biológicas associadas com a variação do EQ.

Identificador	q-value	Anotação
GO:0018343	1.43E-15	protein farnesylation
GO:0051126	1.43E-15	negative regulation of actin nucleation
GO:0005965	1.43E-15	protein farnesyltransferase complex
GO:0042288	1.81E-09	MHC class I protein binding
GO:0035988	1.05E-08	chondrocyte proliferation
GO:0004556	1.95E-05	alpha-amylase activity
GO:0016160	1.95E-05	amylase activity
GO:0002062	3.72E-03	chondrocyte differentiation
GO:0006325	8.18E-03	chromatin organization
GO:0006338	8.67E-03	chromatin remodeling

Termos relacionados à farnesilação de proteínas surgiram nas análises de funções biológicas e nos termos IPR, tanto associados com o EQ, quanto com os resíduos. O processo de farnesilação de proteínas é uma modificação pós-traducional onde um grupo farnesil é adicionado a uma cisteína. Esse processo, um tipo de prenilação, adiciona moléculas hidrofóbicas às proteínas e são essenciais para o ancoramento destas nas membranas plasmáticas (JUNG; BACHMANN, 2023). No sistema nervoso, a plasticidade sináptica e morfologia dos dendritos são alguns dos processos onde a prenilação têm relevância (QU *et al.*, 2021). A desregulação deste processo é implicada em diferentes doenças e distúrbios, entre eles, a doença de Alzheimer (JEONG *et al.*, 2018, 2021).

Um outro termo que apareceu de forma recorrente entre as análises é relacionado à inibição do complexo Arp2/3. São eles o termo GO:00551126 descrito como *negative regulation of actin nucleation*, o termo PF10574, descrito como Arp2/3-interacting proteins Arpin, e o IPR018889, nomeado Arpin. O complexo Arp2/3 é responsável pela nucleação dos filamentos de actina, componentes de uma rede ramificada de actina — parte do citoesqueleto — que é essencial para a motilidade celular (DING *et al.*, 2022). A Arpin tem papel inibitório sob este complexo (DANG *et al.*, 2013) e um dos efeitos que essa inibição gera é a pausa do movimento de migração celular que então induz uma mudança na direção para a qual a célula migra (FREGOSO *et al.*, 2022). A regulação do citoesqueleto de actina é crucial para a morfogênese neuronal normal (LUO, 2002) e PICK1, um outro inibidor do complexo Arp2/3, quando depletado *in vitro* levou a morfologia neuronal aberrante (ROCCA *et al.*, 2008).

4.6.4. Funções biológicas associadas com a variação dos resíduos da massa cerebral

Em contraste às numerosas funções biológicas encontradas como associadas à variação do EQ, encontramos apenas cinco funções biológicas associadas com a variação dos resíduos da massa cerebral (tabela 6). Assim como nas análises de domínios de proteínas, em que os termos eram comuns aos dois fenótipos avaliados, as funções aqui presentes também estavam presentes na análise com o quociente de encefalização e em parte já descritos anteriormente..

Tabela 6: Funções biológicas associadas com a variação dos resíduos da massa cerebral.

Identificador	q-value	Anotação
GO:0018343	1.20E-04	protein farnesylation
GO:0051126	1.20E-04	negative regulation of actin nucleation
GO:0005965	1.20E-04	protein farnesyltransferase complex
GO:0035988	4.83E-04	chondrocyte proliferation
GO:0042288	5.42E-03	MHC class I protein binding

4.6.5. Termos IPR associados com a variação do EQ

O total de termos IPR associados com a variação do EQ foi de 55 termos (exemplos na tabela 7). Diversos destes compreendem aspectos biológicos já capturados pela anotação utilizando os bancos Pfam e GO. A discussão dos termos inéditos segue abaixo.

Tabela 7: Exemplos de termos IPR de anotação associados com a variação do EQ.

Identificador	q-value	Anotação
IPR042869	3.25E-18	Rho GTPase-activating protein 11
IPR018889	3.01E-17	Arpin
IPR019547	3.01E-17	Lipid desaturase domain
IPR026872	3.01E-17	Protein farnesyltransferase subunit beta
IPR041404	3.01E-17	Protein of unknown function DUF5588
IPR047374	3.01E-17	Ankyrin repeat and KH domain-containing protein 1, type I K homology domain
IPR031394	6.56E-10	Musculoskeletal embryonic nuclear protein 1
IPR039842	8.74E-07	TBC1 domain family member 7
IPR006046	1.25E-05	Alpha amylase
IPR016649	3.59E-05	Glial cell line-derived neurotrophic factor
IPR031319	6.35E-04	Alpha-amylase, C-terminal domain
IPR026768	9.31E-04	Protein FAM72
IPR006048	9.31E-04	Alpha-amylase/branching enzyme, C-terminal all beta
IPR039045	2.79E-03	Schwannomin interacting protein 1
IPR017433	6.96E-03	Dystrophin-related protein 2

O termo IPR042869 (figura 7), descrito como Rho GTPase-activating protein 11, é referente aos genes ARHGAP11A e ARHGAP11B. O segundo deles, ARHGAP11B, é um gene específico do *Homo sapiens* que surgiu a partir da duplicação parcial do ARHGAP11A e posterior divergência funcional (FLORIO *et al.*, 2016). Em humanos, ARHGAP11B, têm papel durante a expansão do neocórtex induzindo o aumento da quantidade e proliferação de progenitoras basais neurais, especificamente as células basais da glia (NAMBA *et al.*, 2020).

Este gene, quando expresso junto de seu promotor humano em fetos de *Callithrix jacchus* (popularmente conhecido como mico-estrela), promoveu a expansão do neocórtex e induziu o dobramento cortical do mico que, naturalmente, não apresenta giros e sulcos (HEIDE *et al.*, 2020). Em organoides cerebrais de chimpanzé, a expressão do ARHGAP11B humano duplicou os níveis de progenitoras basais, enquanto nos organoides cerebrais humanos sob a depleção do gene, a quantidade de células progenitoras basais se equiparavam aos níveis normais do chimpanzé (FISCHER *et al.*, 2022).

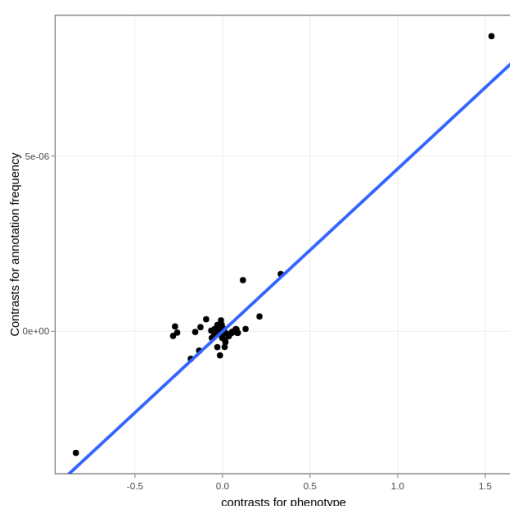


Figura 7: Correlação entre o Rho GTPase-activating protein 11 (IPR042869) e EQ.

4.6.6. Termos IPR associados com a variação dos resíduos da massa cerebral

Foram 24 termos IPR associados com a variação dos resíduos da massa corporal. A tabela 8 lista alguns deles.

Tabela 8: Exemplos de termos IPR de anotação associados com a variação dos resíduos da massa cerebral.

Identificador	q-value	Anotação
IPR042869	1.76E-05	Rho GTPase-activating protein 11
IPR018889	1.76E-05	Arpin
IPR019547	1.76E-05	Lipid desaturase domain
IPR026872	1.76E-05	Protein farnesyltransferase subunit beta
IPR041404	1.76E-05	Protein of unknown function DUF5588
IPR044632	1.76E-05	DnaJ homolog subfamily C member 25-like
IPR047374	1.76E-05	Ankyrin repeat and KH domain-containing protein 1, type I K homology domain
IPR031394	6.82E-05	Musculoskeletal embryonic nuclear protein 1
IPR011629	7.87E-04	Zinc chaperone CobW-like, C-terminal
IPR036627	7.87E-04	CobW-like, C-terminal domain superfamily
IPR042414	8.52E-04	T-cell surface glycoprotein CD8 beta chain
IPR039842	1.67E-03	TBC1 domain family member 7
IPR016649	8.12E-03	Glial cell line-derived neurotrophic factor

Muitos dos termos encontrados entre as diferentes análises têm conexão entre si. É o caso do termo PF15682 (*Musculoskeletal, temporally activated-embryonic nuclear protein 1*), os termos GO:0002062 (*chondrocyte differentiation*) e GO:0035988 (*chondrocyte proliferation*) e o termo IPR031394 (*Musculoskeletal embryonic nuclear protein 1*; figura 8). Todos estes termos estão ligados à família de proteínas Mustang (ou Mustn1). Proteínas desta família são pequenas (~80 aminoácidos) e altamente expressas durante o desenvolvimento, especificamente, em tecidos músculo-esqueléticos. Em adultos, sua expressão é restrita a tecidos ósseos em processo de regeneração devido a fraturas. Embora seu papel na proliferação de condrócitos pareça ser mais evidente nos membros em desenvolvimento, Mustn1 também foi detectado na região craniofacial (DUCOMMUN *et al.*, 2024; GERSCH; HADJIARGYROU, 2009). Em *Xenopus laevis*, um anuro, o knockdown de Mustn1 levou a severos defeitos de desenvolvimento da região craniofacial (GERSCH *et al.*, 2012).

A modificação da musculatura e estrutura óssea craniofacial decerto se faz necessária com a variação da massa cerebral, e agentes que atuam na morfologia destas estruturas são interessantes de serem explorados (LESCIOTTO; RICHTSMEIER, 2019).

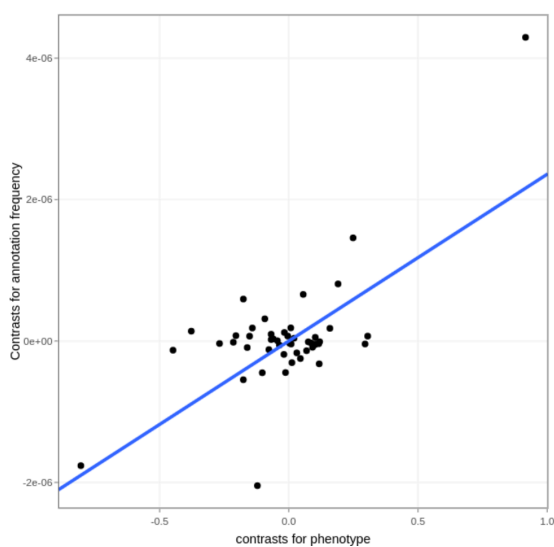


Figura 8: Correlação entre Musculoskeletal embryonic nuclear protein 1 (IPR031394) e os resíduos de massa cerebral.

Outro termos interessantes são o IPR016649, *Glial cell line-derived neurotrophic factor*, que compreende um conjunto de fatores neurotróficos que promovem e mantêm o crescimento, a sobrevivência e funcionamento de diferentes populações neuronais em desenvolvimento e assim como neurônios maduros (KOTLIAROVA; SIDOROVA, 2021; SKAPER, 2018). O termo IPR047374 (*Ankyrin repeat and KH domain-containing protein 1, type I K homology domain*) compreende uma proteína associada às proteínas *Ankyrin* que em neurônios destinam outras proteínas para o segmento axônico inicial ou até as espinhas dendríticas (YOON; PIGUEL; PENZES, 2022).

A duplicação gênica (paralogia) seguida de neofuncionalização é um mecanismo que aparenta ter uma forte importância na evolução humana (KUZMIN; TAYLOR; BOONE, 2022). A busca por correlações entre a abundância de certos elementos com a variação dos fenótipos aqui avaliados, pode mostrar genes e domínios que aumentaram sua frequência junto do aumento da massa cerebral dos primatas.

Interessantemente, diversos dos elementos que descrevemos são explicitamente ligados a fenótipos cerebrais, como é o caso do domínio Olduvai, por exemplo. Outros, em contraste, não têm inicialmente qualquer associação clara com fenótipos exclusivamente cerebrais. Há ainda domínios de proteínas com função desconhecida, os quais compreendem importantes alvos para caracterização experimental futura.

Quanto aos elementos que a princípio não têm função direta sobre o fenótipo cerebral, cabe levantarmos algumas possibilidades. A primeira delas envolve o fato de que muitas vezes a seleção de um certo caractere pode ter efeito sob outro caractere não necessariamente relacionado (FUTUYMA; KIRKPATRICK, 2017). A seleção artificial que ocorreu em

animais domesticados (principalmente o cão) são bons exemplos disto. Inicialmente, a seleção destes animais possivelmente se deu pelo seu comportamento. Entretanto, essa seleção causou a seleção espúria de fenótipos não-relacionados causados por outros loci, como a cor do pelagem e tamanho cerebral (HANSEN WHEAT *et al.*, 2019). Do ponto de vista molecular, diferentes loci podem estar fisicamente ligados (*linkage*), e havendo seleção em uma região, regiões adjacentes ao loci selecionado podem se fixar sem ter um papel no fenótipo alvo da seleção natural (conceito denominado em inglês de *hitchhiking*, pois um loci “pega carona” no loci vantajoso próximo dele e ambos podem eventualmente se fixar) (CHARLESWORTH; JENSEN, 2021). Uma segunda possibilidade seria a existência de genes pleiotrópicos associados ao aumento do tamanho cerebral. A pleiotropia ocorre quando um gene atua e tem efeito em mais de um fenótipo (MACKAY; ANHOLT, 2024). Além disso, há também a possibilidade de que os elementos encontrados sejam correlações espúrias que não representam qualquer correlação causal na realidade. Entretanto, espera-se que a taxa de falso-positivos em nossa análise seja de aproximadamente 5%, uma vez que realizamos a correção para testes múltiplos de hipóteses.

Por fim, um dos motivos que pode explicar o fato que, consistentemente, mais elementos estão significativamente associados com a variação do quociente de encefalização em comparação aos elementos associados com os resíduos, é a própria variação dos índices. Enquanto o EQ tem amplitude de variação entre ~ 0.5 e 8, os resíduos da massa cerebral têm menor amplitude de variação entre -1,5 e 3,5 e, portanto, são menos dispersos, o que pode vir a dificultar o encontro de correlações significativas.

De maneira geral, consideramos estes resultados promissores, uma vez que encontramos tanto elementos sabidamente conhecidos como tendo papel no aumento da massa cerebral, como elementos que podem se tornar interessantes alvos de estudos futuros.

4.7. Alinhamento de genes ortólogos

Utilizamos o *output* do programa OrthoFinder, que compreende, dentre outros arquivos, um arquivo para cada *orthogroup* — ou conjunto de genes ortólogos. O alinhamento, pelo MAFFT, destes ortólogos de 53 espécies de primatas resultou em 11.186 alinhamentos múltiplos de sequências, cada um equivalente a um gene ortólogo. Destes, 10.582 contavam com sequências também encontradas no *Homo sapiens*.

4.8. Análise de sítios de substituições convergentes de aminoácidos

A ferramenta “*discovery*” do CAAStools divide os CAAS em 3 padrões: *pattern 1* consiste em um aminoácido conservado nas espécies com maiores EQ e outro aminoácido conservado nas espécies com menores EQ; *pattern 2* consiste em um aminoácido conservado no grupo das espécies com maiores EQ e aminoácidos variados nas espécies com menores EQ, porém distintos do aminoácido conservado no outro grupo; *pattern 3* corresponde a um aminoácido encontrado no grupo das espécies com menores EQ e aminoácidos variados no grupo dos maiores EQ. Aqui em diante, destacamos os genes homólogos nos quais encontramos os padrões 1 e/ou 2.

Dos 11.186 arquivos de alinhamentos de genes ortólogos, a ferramenta “*discovery*” do CAAStools, encontrou posições consideradas como CAAS em 1634 destes. 1606 contavam com ambos os padrões (listados no apêndice B), 319 contavam com apenas substituições convergentes no padrão 1 e 21 apresentavam apenas o padrão 2.

4.9. Análise de enriquecimento

Utilizando a ferramenta online WebGestalt procuramos por categorias de funções biológicas (definidos como termos GO) enriquecida no conjunto de genes onde existem substituições convergentes de aminoácidos nos grupos com maiores ou menores valores de EQ (1606 genes).

A figura 9 demonstra as 11 categorias que estavam enriquecidas no conjunto avaliado: *ventricular septum development* (FDR = 0.025, ER = 3.03), *cardiac septum development* (FDR = 5.41E-03, ER = 2.82), *cell morphogenesis* (FDR = 5.41E-03, ER = 1.51), *ventricular septum development* (FDR = 2.55E-02, ER = 3.03), *plasma membrane bounded cell projection morphogenesis* (FDR = 3.31E-02, ER = 1.56), *cell projection morphogenesis* (FDR = 3.56E-02, ER = 1.55), *neuron projection morphogenesis* (FDR = 3.77E-02, ER = 1.55), *cell projection organization* (FDR = 3.77E-02, ER = 1.32), *neuron differentiation* (FDR = 3.77E-02, ER = 1.35), *cardiac chamber development* (FDR = 3.77E-02, ER = 2.15), *plasma membrane bounded cell projection organization* (FDR = 4.08E-02, ER = 1.32), *generation of neurons* (FDR = 4.63E-02, ER = 1.33).

Os resultados sugerem que as substituições convergentes encontradas nos grupos com menores ou maiores EQ, ocorreram em genes com funções relevantes para o cérebro e o tamanho do mesmo. Entre as diferentes categorias encontradas, muitos genes estavam presentes em mais de uma delas, demonstrando certa redundância.

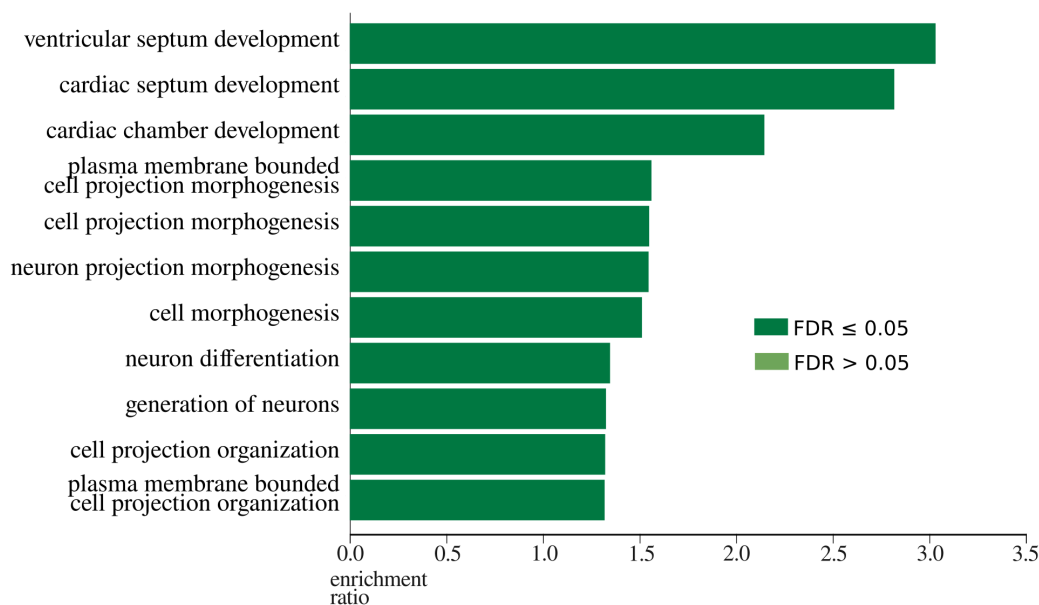


Figura 9: Categorias funcionais enriquecidas.

As diferentes categorias de funções biológicas enriquecidas podem ser divididas em duas categorias maiores: funções relacionadas aos neurônios e funções relacionadas ao coração. Há, nos seres humanos, certa correlação entre transtornos psiquiátricos e doenças cardíacas — estabelecendo um eixo cérebro-coração — e estudos demonstram que existem genes com efeitos pleiotrópicos relacionados a ambos os traços (SONG *et al.*, 2025; ZHANG *et al.*, 2022).

5. CONCLUSÃO

O cérebro é um órgão altamente complexo e apesar dos esforços hercúleos empregados para estudá-lo, ainda sabemos relativamente pouco das suas bases genéticas, que decerto são numerosas. A complexidade funcional e morfológica do cérebro é refletida na complexidade dos fatores genéticos responsáveis pelo seu desenvolvimento embrionário e pela sua evolução biológica. Nosso estudo, utilizando ferramentas de genômica, demonstrou como metodologias *data driven* podem revelar padrões de mudanças moleculares e funcionais relacionados à variação da massa cerebral em Primatas.

Apesar de inicialmente esperarmos encontrar elementos genéticos associados com o cérebro, especialmente elementos que poderiam levar a variação da massa cerebral, visto que este é um caractere herdável, nas metodologias *data-driven* não selecionamos *a priori* os

componentes genéticos a serem estudados. Ainda assim, encontramos diversos fatores genéticos associados ao tamanho cerebral dos primatas e que possivelmente desempenham algum papel causal na evolução deste fenótipo complexo. Cabe salientar que correlações e associações não provam relações de causalidade direta, mas estudos como o presente podem se mostrar como bons direcionadores a futuros estudos de validação funcional.

Dado que uma parcela significativa da diversidade primata encontra-se atualmente ameaçada, estudos que integram aspectos genômicos, fenotípicos e filogenéticos são particularmente valiosos. Compreender a história evolutiva e a biologia desses organismos não apenas contribui para esforços de conservação informados, mas também oferece uma via indireta para aprofundar o conhecimento sobre nossa própria espécie, *Homo sapiens*. Nesse sentido, os resultados aqui apresentados reforçam o potencial da genômica comparativa como ferramenta para explorar os fundamentos moleculares de traços biológicos complexos e evolutivamente relevantes.

REFERÊNCIAS

ALFÖLDI, Jessica; LINDBLAD-TOH, Kerstin. Comparative genomics as a tool to understand evolution and disease. *Genome Research*, v. 23, n. 7, p. 1063–1068, jul. 2013.

BARTERI, Fabio *et al.* CAAStools: a toolbox to identify and test Convergent Amino Acid Substitutions. *Bioinformatics*, v. 39, n. 10, p. btad623, 3 out. 2023.

BAZINET, Richard P.; LAYÉ, Sophie. Polyunsaturated fatty acids and their metabolites in brain function and disease. *Nature Reviews Neuroscience*, v. 15, n. 12, p. 771–785, dez. 2014.

BJORNEVIK, Kjetil *et al.* Association of Polyunsaturated Fatty Acids and Clinical Progression in Patients With ALS. *Neurology*, v. 101, n. 7, p. e690–e698, 15 ago. 2023.

BLUM, Matthias *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, v. 53, n. D1, p. D444–D456, 6 jan. 2025.

BOLOGNINI, Davide *et al.* Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature*, v. 634, n. 8034, p. 617–625, out. 2024.

CAPELLINI, Isabella; VENDITTI, Chris; BARTON, Robert A. Phylogeny and metabolic scaling in mammals. *Ecology*, v. 91, n. 9, p. 2783–2793, 2010.

CHAMBERS, Helen Rebecca; HELDSTAB, Sandra Andrea; O’HARA, Sean J. Why big brains? A comparison of models for both primate and carnivore brain size evolution. *PLOS ONE*, v. 16, n. 12, p. e0261185, 21 dez. 2021.

CHARLESWORTH, Brian; JENSEN, Jeffrey D. Effects of Selection at Linked Sites on Patterns of Genetic Variability. *Annual Review of Ecology, Evolution, and Systematics*, v. 52, n. Volume 52, 2021, p. 177–197, 3 nov. 2021.

Comparative Genomics Fact Sheet. Disponível em:
<<https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet>>.

CORNWALLIS, Charlie K.; GRIFFIN, Ashleigh S. A Guided Tour of Phylogenetic Comparative Methods for Studying Trait Evolution. *Annual Review of Ecology, Evolution, and Systematics*, v. 55, n. 1, p. 181–204, 4 nov. 2024.

DANG, Irene *et al.* Inhibitory signalling to the Arp2/3 complex steers cell migration. *Nature*, v. 503, n. 7475, p. 281–284, nov. 2013.

DAVIS, Jonathan M. *et al.* A Third Linear Association Between Olduvai (DUF1220) Copy Number and Severity of the Classic Symptoms of Inherited Autism. . [S.l: s.n.]. Disponível em:
<<https://psychiatryonline.org/doi/epdf/10.1176/appi.ajp.2018.18080993>>.

DAVIS, Jonathan M. *et al.* DUF1220 Dosage Is Linearly Associated with Increasing Severity of the Three Primary Symptoms of Autism. *PLOS Genetics*, v. 10, n. 3, p. e1004241, 20 mar. 2014.

DEANER, Robert O. *et al.* Overall Brain Size, and Not Encephalization Quotient, Best Predicts Cognitive Ability across Non-Human Primates. *Brain, Behavior and Evolution*, v. 70, n. 2, p. 115–124, 2007.

DECASIEN, Alex R.; BARTON, Robert A.; HIGHAM, James P. Understanding the human brain: insights from comparative biology. *Trends in Cognitive Sciences*, v. 26, n. 5, p. 432–445, 1 maio 2022.

DECASIEN, Alex R.; WILLIAMS, Scott A.; HIGHAM, James P. Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution*, v. 1, n. 5, p. 0112, 27 mar. 2017.

DE MAGALHÃES, João Pedro *et al.* Human Ageing Genomic Resources: updates on key databases in ageing research. *Nucleic Acids Research*, v. 52, n. D1, p. D900–D908, 2 nov. 2023.

DING, Bojian *et al.* Structure of Arp2/3 complex at a branched actin filament junction resolved by single-particle cryo-electron microscopy. *Proceedings of the National Academy of Sciences*, v. 119, n. 22, p. e2202723119, 31 maio 2022.

DUCOMMUN, Serge *et al.* Mustn1 is a smooth muscle cell-secreted microprotein that modulates skeletal muscle extracellular matrix composition. *Molecular Metabolism*, v. 82, p. 101912, 1 abr. 2024.

DUMAS, Laura J. *et al.* DUF1220-Domain Copy Number Implicated in Human Brain-Size Pathology and Evolution. *The American Journal of Human Genetics*, v. 91, n. 3, p. 444–454, 7 set. 2012.

DUNBAR, R. I. M.; SHULTZ, Susanne. Why are there so many explanations for primate brain evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 372, n. 1727, p. 20160244, 3 jul. 2017.

ELIZARRARAS, John M *et al.* WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Research*, v. 52, n. W1, p. W415–W421, 5 jul. 2024.

EMMS, David M.; KELLY, Steven. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, v. 20, n. 1, p. 238, 14 nov. 2019.

EŞIYOK, Nesil *et al.* A dyad of human-specific *NBPF14* and *NOTCH2NLB* orchestrates cortical progenitor abundance crucial for human neocortex expansion. *Science Advances*, v. 11, n. 13, p. eads7543, 28 mar. 2025.

ESTRADA, Alejandro *et al.* Impending extinction crisis of the world's primates: Why primates matter. *Science Advances*, v. 3, n. 1, p. e1600946, 6 jan. 2017.

FARRÉ, Xavier *et al.* Comparative Analysis of Mammal Genomes Unveils Key Genomic Variability for Human Life Span. *Molecular Biology and Evolution*, v. 38, n. 11, p. 4948–4961, 27 out. 2021.

FELSENSTEIN, Joseph. Phylogenies and the Comparative Method. *The American Naturalist*, v. 125, n. 1, p. 1–15, 1985.

FIDDES, Ian T. *et al.* Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*, v. 173, n. 6, p. 1356- 1369.e22, 31 maio 2018.

FIDDES, Ian T. *et al.* Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. *Human Genetics*, v. 138, n. 7, p. 715–721, 1 jul. 2019.

FISCHER, Jan *et al.* Human-specific ARHGAP11B ensures human-like basal progenitor levels in hominid cerebral organoids. *EMBO reports*, v. 23, n. 11, p. e54728, 7 nov. 2022.

FLEAGLE, J.G.; BADEN, A.L.; GILBERT, C.C. *Primate Adaptation and Evolution*. [S.l.]: Academic Press, 2024. Disponível em: <<https://books.google.com.br/books?id=iA9vEAAAQBAJ>>.

FLORIO, Marta *et al.* A single splice site mutation in human-specific *ARHGAP11B* causes basal progenitor amplification. *Science advances*, v. 2, n. 12, p. e1601941, 1 dez. 2016.

FOSSATI, Matteo *et al.* *SRGAP2* and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses. *Neuron*, v. 91, n. 2, p. 356–369, 20 jul. 2016.

FREGOSO, Fred E. *et al.* Molecular mechanism of Arp2/3 complex inhibition by Arpin. *Nature Communications*, v. 13, n. 1, p. 628, 2 fev. 2022.

FUTUYMA, Douglas J.; KIRKPATRICK, Mark. *Evolution*. Fourth edition ed. Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers, 2017.

GERSCH, Robert P *et al.* *Mustn1* is essential for craniofacial chondrogenesis during *Xenopus* development. *Gene expression patterns : GEP*, v. 12, n. 0, p. 145–153, 2012.

GERSCH, Robert P.; HADJIARGYROU, Michael. *Mustn1* is expressed during chondrogenesis and is necessary for chondrocyte proliferation and differentiation *in vitro*. *Bone*, v. 45, n. 2, p. 330–338, 1 ago. 2009.

GIBLIN, Ashling *et al.* Neuronal polyunsaturated fatty acids are protective in ALS/FTD. *Nature Neuroscience*, p. 1–11, 25 fev. 2025.

GOEMAN, Jelle J.; SOLARI, Aldo. Multiple hypothesis testing in genomics. *Statistics in Medicine*, v. 33, n. 11, p. 1946–1978, 2014.

GOOLSBY, Eric W.; BRUGGEMAN, Jorn; ANE, Cecile. *Rphylopars: Phylogenetic Comparative Tools for Missing Data and Within-Species Variation*. [S.l.: s.n.], 2024. Disponível em: <<https://CRAN.R-project.org/package=Rphylopars>>.

GOULD, S.J. Allometry in primates, with emphasis on scaling and the evolution of the brain. *Contributions to primatology*, v. 5, p. 244–292, 1975.

GRABOWSKI, Mark *et al.* Both Diet and Sociality Affect Primate Brain-Size Evolution. *Systematic Biology*, v. 72, n. 2, p. 404–418, 16 jun. 2023.

GRABOWSKI, Mark; VOJE, Kjetil L.; HANSEN, Thomas F. Evolutionary modeling and correcting for observation error support a 3/5 brain-body allometry for primates. *Journal of Human Evolution*, v. 94, p. 106–116, maio 2016.

HANSEN WHEAT, Christina *et al.* Behavioural correlations of the domestication syndrome are decoupled in modern dog breeds. *Nature Communications*, v. 10, n. 1, p. 2422, 3 jun. 2019.

HARDISON, Ross C. Comparative Genomics. *PLOS Biology*, v. 1, n. 2, p. e58, 17 nov. 2003.

HARDY, Karen *et al.* The Importance of Dietary Carbohydrate in Human Evolution. *The Quarterly Review of Biology*, 1 set. 2015. Disponível em: <<https://www.journals.uchicago.edu/doi/10.1086/682587>>.

HEFT, Ilea E *et al.* The Driver of Extreme Human-Specific Olduvai Repeat Expansion Remains Highly Active in the Human Genome. *Genetics*, v. 214, n. 1, p. 179–191, 1 jan. 2020.

HEIDE, Michael *et al.* Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science*, v. 369, n. 6503, p. 546–550, 31 jul. 2020.

HELDSTAB, Sandra A. *et al.* The economics of brain size evolution in vertebrates. *Current Biology*, v. 32, n. 12, p. R697–R708, 20 jun. 2022.

HERCULANO-HOUZEL, Suzana. Longevity and sexual maturity vary across species with number of cortical neurons, and humans are no exception. *Journal of Comparative Neurology*, v. 527, n. 10, p. 1689–1705, jul. 2019.

HERITAGE, Steven. Modeling Olfactory Bulb Evolution through Primate Phylogeny. *PLOS ONE*, v. 9, n. 11, p. e113904, 26 nov. 2014.

HO, Nguyen Thi Thanh *et al.* Cognitive Functions: Human vs. Animal – 4:1 Advantage [FAM72–SRGAP2-]. *Journal of Molecular Neuroscience*, v. 61, n. 4, p. 603–606, abr. 2017.

HO, Nguyen Thi Thanh; KUTZNER, Arne; HEESE, Klaus. A Novel Divergent Gene Transcription Paradigm—the Decisive, Brain-Specific, Neural [Srgap2–Fam72a-] Master Gene Paradigm. *Molecular Neurobiology*, v. 56, n. 8, ago. 2019. Disponível em: <<http://link.springer.com/10.1007/s12035-019-1486-5>>.

HO, Nguyen Thi Thanh; KUTZNER, Arne; HEESE, Klaus. Brain plasticity, cognitive functions and neural stem cells: a pivotal role for the brain-specific neural master gene [SRGAP2–FAM72-]. *Biological Chemistry*, v. 399, n. 1, p. 55–61, 20 dez. 2017.

HONGO, Jorge Augusto *et al.* CALANGO: A phylogeny-aware comparative genomics tool for discovering quantitative genotype-phenotype associations across species. *Patterns*, v. 4, n. 6, p. 100728, jun. 2023.

HOU, Qiong-Qiong *et al.* TBC1D3 promotes neural progenitor proliferation by suppressing the histone methyltransferase G9a. *Science Advances*, v. 7, n. 3, p. eaba8053, 15 jan. 2021.

HUNG, J.-H. *et al.* Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, v. 13, n. 3, p. 281–291, 1 maio 2012.

ISLER, Karin; VAN SCHAİK, Carel P. The Expensive Brain: A framework for explaining evolutionary changes in brain size. *Journal of Human Evolution*, Paleoanthropology Meets Primatology. v. 57, n. 4, p. 392–400, 1 out. 2009.

IVES, Anthony R.; ZHU, Jun. Statistics For Correlated Data: Phylogenies, Space, And Time. *Ecological Applications*, v. 16, n. 1, p. 20–32, fev. 2006.

JANECKA, Jan E. *et al.* Molecular and genomic data identify the closest living relative of primates. *Science (New York, N.Y.)*, v. 318, n. 5851, p. 792–794, 2 nov. 2007.

JANIÁK, Mareike Cora. Digestive enzymes of human and nonhuman primates. *Evolutionary Anthropology: Issues, News, and Reviews*, v. 25, n. 5, p. 253–266, 2016.

JEONG, Angela *et al.* Isoprenoids and protein prenylation: implications in the pathogenesis and therapeutic intervention of Alzheimer's disease. *Critical Reviews in Biochemistry and Molecular Biology*, v. 53, n. 3, p. 279–310, 4 maio 2018.

JEONG, Angela *et al.* Protein farnesylation is upregulated in Alzheimer's human brains and neuron-specific suppression of farnesyltransferase mitigates pathogenic processes in Alzheimer's model mice. *Acta Neuropathologica Communications*, v. 9, n. 1, p. 129, 27 jul. 2021.

JERISON, H. J. Animal intelligence as encephalization. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, v. 308, n. 1135, p. 21–35, 13 fev. 1985.

JERISON, H. J. *Evolution of the Brain and Intelligence*. [S.l.]: Elsevier, 1973. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/B9780123852502X50019>>.

JUNG, Dominik; BACHMANN, Hagen S. Regulation of protein prenylation. *Biomedicine & Pharmacotherapy*, v. 164, p. 114915, ago. 2023.

KATOH, Kazutaka; STANDLEY, Daron M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, v. 30, n. 4, p. 772–780, 1 abr. 2013.

KEENEY, J. G. *et al.* DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. *Brain Structure and Function*, v. 220, n. 5, p. 3053–3060, 1 set. 2015.

KEENEY, Jonathon G. *et al.* Olduvai domain expression downregulates mitochondrial pathways: implications for human brain evolution and neoteny. *bioRxiv*, p. 2024.10.21.619278, 22 out. 2024.

KEENEY, Jonathon G.; DUMAS, Laura; SIKELA, James M. The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. *Frontiers in Human Neuroscience*, v. 8, 24 jun. 2014. Disponível em: <<https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00427/full>>.

- KOTLIAROVA, Anastasiia; SIDOROVA, Yulia A. Glial Cell Line-Derived Neurotrophic Factor Family Ligands, Players at the Interface of Neuroinflammation and Neuroprotection: Focus Onto the Glia. *Frontiers in Cellular Neuroscience*, v. 15, 17 jun. 2021. Disponível em: <<https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2021.679034/full>>.
- KUMAR, Sudhir *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, v. 39, n. 8, p. msac174, 3 ago. 2022.
- KUZMIN, Elena; TAYLOR, John S.; BOONE, Charles. Retention of duplicated genes in evolution. *Trends in Genetics*, v. 38, n. 1, p. 59–72, jan. 2022.
- LESCIOTTO, Kate M.; RICHTSMEIER, Joan T. Craniofacial skeletal response to encephalization: How do we know what we think we know? *American Journal of Physical Anthropology*, v. 168, n. S67, p. 27–46, 2019.
- LUO, Liqun. Actin Cytoskeleton Regulation in Neuronal Morphogenesis and Structural Plasticity. *Annual Review of Cell and Developmental Biology*, v. 18, n. Volume 18, 2002, p. 601–635, 1 nov. 2002.
- MACKAY, Trudy F. C.; ANHOLT, Robert R. H. Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nature Reviews Genetics*, v. 25, n. 9, p. 639–657, set. 2024.
- MANNI, Mosè *et al.* BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, v. 38, n. 10, p. 4647–4654, 27 set. 2021.
- MELCHIONNA, M *et al.* Macroevolutionary trends of brain mass in Primates. *Biological Journal of the Linnean Society*, p. blz161, 7 nov. 2019.
- MELCHIONNA, Marina *et al.* Cortical areas associated to higher cognition drove primate brain evolution. *Communications Biology*, v. 8, n. 1, p. 1–12, 18 jan. 2025.
- MISTRY, Jaina *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research*, v. 49, n. D1, p. D412–D419, 8 jan. 2021.
- NAMBA, Takashi *et al.* Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. *Neuron*, v. 105, n. 5, p. 867–881.e9, 4 mar. 2020.
- O’LEARY, Nuala A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, v. 44, n. D1, p. D733–D745, 4 jan. 2016.
- ORKIN, Joseph D.; KUDERNA, Lukas F. K.; MARQUES-BONET, Tomas. The Diversity of Primates: From Biomedicine to Conservation Genomics. *Annual Review of Animal Biosciences*, v. 9, n. Volume 9, 2021, p. 103–124, 15 fev. 2021.
- PENONE, Caterina *et al.* Imputation of missing data in life-history trait datasets: which approach performs the best? *Methods in Ecology and Evolution*, v. 5, n. 9, p. 961–970, 2014.
- PERRY, George H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, v. 39, n. 10, p. 1256–1260, out. 2007.

POLLARD, Katherine S. *et al.* Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLOS Genetics*, v. 2, n. 10, p. e168, 13 out. 2006.

POLLEN, Alex A. *et al.* Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nature Reviews Genetics*, v. 24, n. 10, p. 687–711, out. 2023.

PÜSCHEL, Thomas A. *et al.* Hominin brain size increase has emerged from within-species encephalization. *Proceedings of the National Academy of Sciences*, v. 121, n. 49, p. e2409542121, 3 dez. 2024.

QU, Wenhui *et al.* Neuronal Protein Farnesylation Regulates Hippocampal Synaptic Plasticity and Cognitive Function. *Molecular Neurobiology*, v. 58, n. 3, p. 1128–1144, 1 mar. 2021.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2024. Disponível em: <<https://www.R-project.org/>>.

RAMESH, Janani *et al.* Deciphering the molecular landscape of the FAM72 gene family: Implications for stem cell biology and cancer. *Neurochemistry International*, v. 180, p. 105853, nov. 2024.

REVELL, Liam J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, v. 12, p. e16505, 2024.

ROCCA, Daniel L. *et al.* Inhibition of Arp2/3-mediated actin polymerization by PICK1 regulates neuronal morphology and AMPA receptor endocytosis. *Nature Cell Biology*, v. 10, n. 3, p. 259–271, mar. 2008.

ROTH, Gerhard; DICKE, Ursula. Evolution of the brain and intelligence in primates. *Progress in Brain Research*. [S.l.]: Elsevier, 2012. v. 195. p. 413–430. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/B9780444538604000209>>.

SEARLES QUICK, V. B. *et al.* DUF1220 copy number is associated with schizophrenia risk and severity: implications for understanding autism and schizophrenia as related diseases. *Translational Psychiatry*, v. 5, n. 12, p. e697–e697, dez. 2015.

SHAO, Yong *et al.* Phylogenomic analyses provide insights into primate evolution. *Science*, v. 380, n. 6648, p. 913–924, 2 jun. 2023.

SHULTZ, S. Brain Evolution in Vertebrates. *Encyclopedia of Behavioral Neuroscience*. [S.l.]: Elsevier, 2010. p. 180–186. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/B9780080453965001032>>.

SIKELA, James M.; VAN ROY, Frans. Changing the name of the NBPF/DUF1220 domain to the Olduvai domain. *F1000Research*, v. 6, p. 2185, 17 jul. 2018.

SILBEREIS, John C. *et al.* The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron*, v. 89, n. 2, p. 248–268, 20 jan. 2016.

SKAPER, Stephen D. Neurotrophic Factors: An Overview. *Methods in Molecular Biology (Clifton, N.J.)*, v. 1727, p. 1–17, 2018.

- SONG, Qifeng *et al.* Exploring the genetic landscape of the brain-heart axis: A comprehensive analysis of pleiotropic effects between heart disease and psychiatric disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, v. 136, p. 111172, jan. 2025.
- STEIN, Lincoln. Genome annotation: from sequence to biology. *Nature Reviews Genetics*, v. 2, n. 7, p. 493–503, jul. 2001.
- SUBRAMANIAN, Aravind *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, v. 102, n. 43, p. 15545–15550, 25 out. 2005.
- SUZUKI, Ikuo K. *et al.* Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. *Cell*, v. 173, n. 6, p. 1370- 1384.e16, 31 maio 2018.
- THE GENE ONTOLOGY CONSORTIUM *et al.* The Gene Ontology knowledgebase in 2023. *GENETICS*, v. 224, n. 1, p. iyad031, 4 maio 2023.
- TSUBOI, Masahito *et al.* Breakdown of brain–body allometry and the encephalization of birds and mammals. *Nature Ecology & Evolution*, v. 2, n. 9, p. 1492–1500, 13 ago. 2018.
- TYNIANSKAIA, Lidia; HEIDE, Michael. Human-specific genetic hallmarks in neocortical development: focus on neural progenitors. *Current Opinion in Genetics & Development*, v. 89, p. 102267, 1 dez. 2024.
- VANDEPOELE, Karl *et al.* A Novel Gene Family NBPF: Intricate Structure Generated by Gene Duplications During Primate Evolution. *Molecular Biology and Evolution*, v. 22, n. 11, p. 2265–2274, 1 nov. 2005.
- VENDITTI, Chris; BAKER, Joanna; BARTON, Robert A. Co-evolutionary dynamics of mammalian brain and body size. *Nature Ecology & Evolution*, v. 8, n. 8, p. 1534–1542, ago. 2024.
- VERACINI, Cecilia; WOOD, Bernard. *Primates in History, Myth, Art, and Science*. 1. ed. Boca Raton: CRC Press, 2024. Disponível em:
<<https://www.taylorfrancis.com/books/9781315270371>>.
- WANG, Yan *et al.* Protein domain identification methods and online resources. *Computational and Structural Biotechnology Journal*, v. 19, p. 1145–1153, 2 fev. 2021.
- WILSON MANTILLA, Gregory P. *et al.* Earliest Palaeocene purgatoriids and the initial radiation of stem primates. *Royal Society Open Science*, v. 8, n. 2, p. rsos.210050, 210050, fev. 2021.
- YEAGLE, Philip L. Chapter 3 - Biogenesis of Membrane Lipids. In: YEAGLE, PHILIP L. (Org.). . *The Membranes of Cells (Third Edition)*. Boston: Academic Press, 2016. p. 57–71. Disponível em:
<<https://www.sciencedirect.com/science/article/pii/B9780128000472000036>>.
- YILMAZ, Feyza *et al.* Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science*, v. 386, n. 6724, p. eadn0609, 17 out. 2024.

YOON, Sehyoun; PIGUEL, Nicolas H.; PENZES, Peter. Roles and mechanisms of ankyrin-G in neuropsychiatric disorders. *Experimental & Molecular Medicine*, v. 54, n. 7, p. 867–877, jul. 2022.

ZHANG, Xinyuan *et al.* Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders. *Nature Communications*, v. 13, n. 1, p. 3428, 14 jun. 2022.

ZHOU, Yi; SONG, Hongjun; MING, Guo-li. Genetics of human brain development. *Nature Reviews Genetics*, v. 25, n. 1, p. 26–45, jan. 2024.

APÊNDICE A - Termos de anotação significativamente associados com os fenótipos avaliados de acordo com o CALANGO

TABELA A1: Domínios de proteínas associados com a variação do quociente de encefalização.

Identificador	q-value	Anotação
PF10520	9.37E-16	Lipid desaturase domain
PF10574	9.37E-16	Arp2/3-interacting proteins Arpin
PF15803	4.54E-10	Zinc-finger of sodium channel modifier 1
PF13483	3.48E-09	Beta-lactamase superfamily domain
PF15682	1.07E-08	Musculoskeletal, temporally activated-embryonic nuclear protein 1
PF02598	1.50E-05	Putative RNA methyltransferase
PF07683	2.38E-05	Cobalamin synthesis protein cobW C-terminal domain
PF02109	1.37E-04	DAD family
PF00128	1.43E-04	Alpha amylase, catalytic domain
PF06758	7.12E-04	Olduvai domain
PF00929	1.43E-03	Exonuclease
PF14976	2.22E-03	FAM72 protein
PF05327	3.29E-03	RNA polymerase I specific transcription initiation factor RRN3
PF04420	3.50E-03	CHD5-like protein
PF02806	5.22E-03	Alpha amylase, C-terminal all-beta domain
PF08608	8.43E-03	Wyosine base formation
PF00956	2.11E-02	Nucleosome assembly protein (NAP)

TABELA A2: Domínios de proteínas associados com a variação dos resíduos da massa cerebral.

Identificador	q-value	Anotação
PF10520	3.92E-05	Lipid desaturase domain
PF10574	3.92E-05	Arp2/3-interacting proteins Arpin
PF15682	1.95E-04	Musculoskeletal, temporally activated-embryonic nuclear protein 1
PF02598	9.28E-04	Putative RNA methyltransferase
PF07683	1.26E-03	Cobalamin synthesis protein cobW C-terminal domain
PF14976	2.22E-03	FAM72 protein
PF05327	3.29E-03	RNA polymerase I specific transcription initiation factor RRN3

Identificador	q-value	Anotação
PF04420	3.50E-03	CHD5-like protein
PF02806	5.22E-03	Alpha amylase, C-terminal all-beta domain
PF08608	8.43E-03	Wyosine base formation

TABELA A3: Funções biológicas associadas com a variação do EQ.

Identificador	q-value	Anotação
GO:0018343	1.43E-15	protein farnesylation
GO:0051126	1.43E-15	negative regulation of actin nucleation
GO:0005965	1.43E-15	protein farnesyltransferase complex
GO:0042288	1.81E-09	MHC class I protein binding
GO:0035988	1.05E-08	chondrocyte proliferation
GO:0004556	1.95E-05	alpha-amylase activity
GO:0016160	1.95E-05	amylase activity
GO:0065004	2.28E-05	protein-DNA complex assembly
GO:0071824	3.51E-04	protein-DNA complex organization
GO:0071816	1.20E-03	tail-anchored membrane protein insertion into ER membrane
GO:0006334	2.25E-03	nucleosome assembly
GO:0034728	2.25E-03	nucleosome organization
GO:0001181	2.52E-03	RNA polymerase I general transcription initiation factor activity
GO:0006901	2.52E-03	vesicle coating
GO:0048207	2.52E-03	vesicle targeting, rough ER to cis-Golgi
GO:0048208	2.52E-03	COPII vesicle coating
GO:0002062	3.72E-03	chondrocyte differentiation
GO:0006361	5.14E-03	transcription initiation at RNA polymerase I promoter
GO:0006325	8.18E-03	chromatin organization
GO:0006338	8.67E-03	chromatin remodeling
GO:0005769	1.83E-02	early endosome
GO:0035299	3.73E-02	inositol pentakisphosphate 2-kinase activity
GO:0003711	3.73E-02	transcription elongation factor activity
GO:0006517	3.73E-02	protein deglycosylation
GO:0035977	3.73E-02	protein deglycosylation involved in glycoprotein catabolic process
GO:0036507	3.73E-02	protein demannosylation
GO:0036508	3.73E-02	protein alpha-1,2-demannosylation

Identificador	q-value	Anotação
GO:0097466	3.73E-02	ubiquitin-dependent glycoprotein ERAD pathway
GO:1904380	3.73E-02	endoplasmic reticulum mannose trimming
GO:1904382	3.73E-02	mannose trimming involved in glycoprotein ERAD pathway
GO:1904587	3.73E-02	response to glycoprotein
GO:0009881	4.87E-02	photoreceptor activity
GO:0009882	4.87E-02	blue light photoreceptor activity

TABELA A4: Funções biológicas associadas com a variação dos resíduos de massa cerebral.

Identificador	q-value	Anotação
GO:0018343	1.20E-04	protein farnesylation
GO:0051126	1.20E-04	negative regulation of actin nucleation
GO:0005965	1.20E-04	protein farnesyltransferase complex
GO:0035988	4.83E-04	chondrocyte proliferation
GO:0042288	5.42E-03	MHC class I protein binding

TABELA A5: Termos de anotação IPR associados com a variação do EQ.

Identificador	q-value	Anotação
IPR017345	1.15E-19	Peptidase S1A, Tysnd1
IPR028341	1.15E-19	Complement factor B
IPR042869	3.25E-18	Rho GTPase-activating protein 11
IPR018889	3.01E-17	Arpin
IPR019547	3.01E-17	Lipid desaturase domain
IPR026872	3.01E-17	Protein farnesyltransferase subunit beta
IPR033570	3.01E-17	Sodium channel modifier 1
IPR039258	3.01E-17	Zinc finger protein 511
IPR041404	3.01E-17	Protein of unknown function DUF5588
IPR044632	3.01E-17	DnaJ homolog subfamily C member 25-like
IPR047374	3.01E-17	Ankyrin repeat and KH domain-containing protein 1, type I K homology domain
IPR045112	3.71E-12	Protein Peter Pan-like
IPR031622	2.76E-11	Sodium channel modifier 1, zinc-finger
IPR042414	5.88E-11	T-cell surface glycoprotein CD8 beta chain
IPR003750	9.85E-11	Putative RNA methyltransferase
IPR022127	6.56E-10	Store-operated calcium entry regulator STIMATE/YPL162C

Identificador	q-value	Anotação
IPR031394	6.56E-10	Musculoskeletal embryonic nuclear protein 1
IPR037339	6.29E-07	Cytidine monophosphate-N-acetylneuraminic acid hydroxylase, Rieske domain
IPR039842	8.74E-07	TBC1 domain family member 7
IPR043039	8.74E-07	TBC1 domain family member 7, domain 2
IPR011629	9.91E-06	Zinc chaperone CobW-like, C-terminal
IPR036627	9.91E-06	CobW-like, C-terminal domain superfamily
IPR006046	1.25E-05	Alpha amylase
IPR006047	2.41E-05	Glycosyl hydrolase, family 13, catalytic domain
IPR016649	3.59E-05	Glial cell line-derived neurotrophic factor
IPR016235	9.27E-05	Tyrosine/threonine-protein kinase, Cdc2 inhibitor
IPR016351	1.12E-04	Plasminogen-related
IPR003038	1.30E-04	DAD/Ost2
IPR046794	3.53E-04	Anaphase-promoting complex subunit 1, middle domain
IPR045137	3.92E-04	RNA-binding protein 26/27
IPR024880	4.85E-04	COPII coat assembly protein, Sec16
IPR031319	6.35E-04	Alpha-amylase, C-terminal domain
IPR028945	9.30E-04	Get1 family
IPR026768	9.31E-04	Protein FAM72
IPR006048	9.31E-04	Alpha-amylase/branching enzyme, C-terminal all beta
IPR002164	2.02E-03	Nucleosome assembly protein (NAP)
IPR027827	2.13E-03	Testis expressed protein 56
IPR013917	2.68E-03	tRNA wybutosine-synthesis
IPR034556	2.68E-03	S-adenosyl-L-methionine-dependent tRNA 4-demethylwyosine synthase
IPR007991	2.70E-03	RNA polymerase I specific transcription initiation factor RRN3
IPR034170	2.71E-03	Retropepsin-like catalytic domain
IPR015649	2.79E-03	Schwannomin interacting protein 1, C-terminal
IPR039045	2.79E-03	Schwannomin interacting protein 1
IPR047460	3.49E-03	TDRD15, fifth Tudor domain
IPR042553	3.49E-03	TSC22 domain family protein 4
IPR037231	6.27E-03	NAP-like superfamily
IPR017433	6.96E-03	Dystrophin-related protein 2
IPR029222	1.08E-02	Protein FAM104
IPR000466	1.74E-02	Adenosine A3 receptor

Identificador	q-value	Anotação
IPR044674	1.92E-02	ER degradation-enhancing alpha-mannosidase-like protein 1/2/3
IPR024990	2.82E-02	Anaphase-promoting complex subunit 1
IPR028290	2.93E-02	WASH1
IPR001356	4.31E-02	Homeobox domain
IPR028155	4.45E-02	RPA-interacting protein, central domain
IPR029363	4.45E-02	Protein of unknown function DUF4575

TABELA A6: Termos de anotação IPR associados com a variação dos resíduos de massa cerebral.

Identificador	q-value	Anotação
IPR042869	1.76E-05	Rho GTPase-activating protein 11
IPR017345	1.76E-05	Peptidase S1A, Tysnd1
IPR028341	1.76E-05	Complement factor B
IPR018889	1.76E-05	Arpin
IPR019547	1.76E-05	Lipid desaturase domain
IPR026872	1.76E-05	Protein farnesyltransferase subunit beta
IPR033570	1.76E-05	Sodium channel modifier 1
IPR039258	1.76E-05	Zinc finger protein 511
IPR041404	1.76E-05	Protein of unknown function DUF5588
IPR044632	1.76E-05	DnaJ homolog subfamily C member 25-like
IPR047374	1.76E-05	Ankyrin repeat and KH domain-containing protein 1, type I K homology domain
IPR022127	6.82E-05	Store-operated calcium entry regulator STIMATE/YPL162C
IPR031394	6.82E-05	Musculoskeletal embryonic nuclear protein 1
IPR045112	9.99E-05	Protein Peter Pan-like
IPR003750	9.99E-05	Putative RNA methyltransferase
IPR011629	7.87E-04	Zinc chaperone CobW-like, C-terminal
IPR036627	7.87E-04	CobW-like, C-terminal domain superfamily
IPR042414	8.52E-04	T-cell surface glycoprotein CD8 beta chain
IPR031622	1.09E-03	Sodium channel modifier 1, zinc-finger
IPR039842	1.67E-03	TBC1 domain family member 7
IPR043039	1.67E-03	TBC1 domain family member 7, domain 2
IPR016649	8.12E-03	Glial cell line-derived neurotrophic factor
IPR016235	8.75E-03	Tyrosine/threonine-protein kinase, Cdc2 inhibitor
IPR003038	2.00E-02	DAD/Ost2

ALX4	CNTROB	GADD45GIP1	LPCAT4	PDCD6IP	SCN10A	TOE1
AMDHD1	COA7	GALNT3	LPIN3	PDE12	SCRN2	TOGARAM2
AMER2	COBLL1	GALNT5	LRATD2	PDE1A	SCTR	TOMM34
AMIGO2	COG1	GARIN3	LRFN4	PDE2A	SCUBE3	TOMM70
ANGEL1	COG4	GARRE1	LRFN5	PDE5A	SCYL3	TOR3A
ANGPT4	COG7	GARS1	LRP10	PDE7A	SDR9C7	TOX2
ANGPTL3	COL26A1	GAS2L2	LRP2	PDE7B	SEC24C	TP53BP2
ANGPTL4	COL6A6	GASK1B	LRRC14	PDK4	SEC63	TPCN1
ANGPTL7	COMMD2	GATA4	LRRC17	PDP1	SECISBP2	TPP1
ANK3	COPB1	GATA6	LRRC30	PDP2	SECISBP2L	TPP2
ANKFY1	COQ10A	GATM	LRRC38	PDYN	SELENBP1	TRANK1
ANKIB1	COQ8A	GBA2	LRRC8C	PDZD8	SEMA3A	TRHR
ANKLE1	COX10	GCLM	LRRC8D	PELO	SEMA5A	TRIB1
ANKRA2	COX15	GCNT4	LRRK1	PET117	SEMA5B	TRIM25
ANKRD27	COX7A2L	GDF2	LRRN1	PEX1	SEMA6D	TRIM29
ANKRD52	CPA5	GDPD1	LRRN4	PEX2	SERINC2	TRIM42
ANKRD6	CPM	GEMIN5	LRRTM2	PEX5	SERPINB11	TRIT1
ANKS4B	CPS1	GET1	LSG1	PFAS	SERPINB5	TRMT112
ANO6	CPSF7	GFM1	LTBP1	PFN3	SETD5	TRNT1
ANPEP	CPT2	GFOD2	LURAP1	PGAP4	SETDB1	TRPM1
ANXA1	CREB3L3	GGA1	LURAP1L	PGBD2	SETX	TRPM3
AOC1	CROT	GGA2	LUZP1	PGRMC1	SGCG	TRPM4
AP3B2	CRYAB	GGCT	LXN	PGRMC2	SGSH	TRUB1
AP4B1	CRYBG3	GGT7	LYPD3	PHAX	SH2B1	TSC22D2
AP5Z1	CSF2RB	GHSR	LYPD6	PHF12	SH3BP4	TSGA10IP
APH1B	CSRNP2	GIT2	LYSMD1	PHF20	SH3TC2	TSPAN31
APLP1	CST11	GJA9	LZTS2	PHF3	SHB	TSPYL4
APLP2	CTBS	GJB1	M6PR	PHIP	SHCBP1L	TSPYL5
APOM	CTLA4	GJC1	MAD2L1BP	PHLPP2	SHISA4	TSR2
APTX	CTNNA1	GJD2	MAEL	PHOSPHO2	SHISAL2B	TSSK3
ARHGAP1	CTNS	GKN2	MAG11	PICALM	SHKBP1	TTC14
ARHGAP10	CTR9	GLB1L	MAIP1	PIERCE1	SHLD3	TTC21A
ARHGAP17	CTSA	GLCCI1	MANSC1	PIGC	SHOX2	TTC29
ARHGAP31	CTSB	GLRX	MAP1B	PIGG	SHQ1	TTC3
ARHGDIG	CTSF	GMPPA	MAP3K14	PIGS	SHTN1	TTC32
ARHGEF11	CUL3	GNE	MAP3K19	PIK3C2B	SIAE	TTC39A
ARHGEF39	CXorf66	GNS	MAP3K5	PIK3CD	SIGMAR1	TTC39C
ARHGEF5	CYLD	GOLIM4	MAP3K6	PIK3IP1	SKAP1	TTL12
ARID1A	CYP19A1	GOLPH3	MAP3K9	PIK3R4	SKIL	TTL7
ARL14EP	CYP26B1	GORASP1	MAP4K2	PIK3R6	SLC12A5	TUBD1
ARMC1	DAAM1	GORASP2	MAP4K3	PIP	SLC14A2	TUFM
ARMC7	DARS1	GPC2	MAP4K5	PIWIL4	SLC15A5	TYW1
ARMC8	DAZAP2	GPCPD1	MAPK15	PKNOX2	SLC16A1	UBAC2
ARMH2	DBF4	GPNMB	MAPK4	PLA2R1	SLC16A8	UBAP1
ARNT2	DBH	GPR137C	MAPK7	PLBD1	SLC16A9	UBASH3A
ARSA	DBN1	GPR160	MAPK8IP2	PLBD2	SLC17A2	UBE2O
ASAP2	DCAF5	GPR21	MAPKAPK3	PLCB2	SLC17A3	UBIAD1
ASB10	DCBLD2	GPR22	MAPKBP1	PLCL1	SLC1A1	UBN1
ASCL1	DCHS1	GPR3	MARCHF4	PLCXD2	SLC1A7	UBR3
ASH1L	DCP1A	GPR37	MARCHF7	PLEKHG1	SLC22A8	UBR5
ASIC5	DCP2	GPR37L1	MARCHF8	PLEKHM3	SLC23A2	UBR7
ASZ1	DDHD1	GPR45	MARCKS	PLEKHO2	SLC24A5	UBXN10
ATF4	DDR GK1	GPR52	MAS1	PLGRKT	SLC25A11	UBXN6
ATF7IP2	DDX1	GPR63	MATCAP1	PLK1	SLC25A17	UHRF2
ATG13	DDX42	GPR75	MAVS	PLK4	SLC25A20	ULK3
ATG2A	DDX54	GPR84	MC2R	PLL	SLC25A28	UMPS
ATG4C	DEFB121	GPX5	MCF2	PLPPR4	SLC25A35	UPB1
ATG9B	DENND2B	GREB1L	MCF2L2	PLRG1	SLC25A47	UPF1
ATL2	DENND2D	GRIK4	MCM9	PNLDC1	SLC25A53	UQCC1
ATP23	DEPDC1	GRIN2B	MDM4	PNMA1	SLC27A1	UQCRQ

ATP4B	DFFA	GRN	MED24	PNMA2	SLC27A6	URB2
ATP6V0A2	DGAT2L6	GRWD1	MEF2A	PNO1	SLC30A1	USF3
ATP6V1FNB	DGCR2	GSDME	MEF2D	PNPT1	SLC30A5	USO1
ATP7B	DGCR8	GSG1	MEFV	POC5	SLC35A1	USP13
ATPAF1	DGKD	GTPBP3	MEGF9	POLA2	SLC35C1	USP3
ATPAF2	DGKG	GTPBP8	MEN1	POLE2	SLC35F6	USP30
ATPSCKMT	DHX30	GUCA1A	MEOX1	POLE4	SLC35G2	USP4
AUH	DHX32	GUCA1B	MEPCE	POLG	SLC38A1	USP45
AURKA	DHX37	GZF1	MET	POLH	SLC39A10	USP47
AVPR1A	DHX9	GZMA	METTL3	POLK	SLC3A1	USP5
AVPR2	DIDO1	HAL	MFSD3	POLL	SLC46A2	USP54
AWAT2	DIP2B	HAO1	MFSD6	POLR1B	SLC46A3	USP7
B3GALT2	DISP1	HAUS4	MFSD8	PPFIA4	SLC5A10	USPL1
BAAT	DISP3	HAUS6	MFSD9	PPFIBP2	SLC5A12	UTP25
BAK1	DKKL1	HCAR1	MGARP	PIIG	SLC6A5	UTP3
BATF	DMBX1	HCN3	MGAT4A	PPM1D	SLC6A8	VANGL1
BAZ1B	DMRT2	HDC	MGAT4C	PPM1G	SLC7A11	VAPA
BBOX1	DMTN	HDDC2	MGST2	PPP1R10	SLC7A5	VAT1
BCO1	DMXL1	HEATR5A	MICAL1	PPP1R12B	SLC9A7	VAX1
BEND4	DNAH7	HELZ	MIDEAS	PPP1R3B	SLF1	VAX2
BEST2	DNAJC10	HEPHL1	MIEF1	PPP1R3E	SLFN5	VEZF1
BEST4	DNAJC12	HERC3	MINK1	PPP4R2	SLIT3	VGLL2
BFSP1	DNAJC13	HEXB	MKRN2	PRCP	SLITRK2	VLDLR
BLOC1S3	DNAJC2	HEY1	MLANA	PRDM1	SLITRK3	VMO1
BLOC1S4	DNAJC24	HHLA1	MLH1	PRDM13	SLK	VPS18
BLTP3A	DNAJC30	HIBADH	MLXIPL	PREB	SLMAP	VPS41
BMP7	DNASE1L1	HIBCH	MMACHC	PREX2	SMCR8	VPS54
BMPR2	DNASE2	HID1	MMP20	PRICKLE1	SMDT1	VPS72
BOLA1	DNASE2B	HIVEP1	MMP9	PRICKLE3	SMG7	VSTM2L
BPNT2	DNP1	HIVEP3	MMRN1	PRND	SMO	VSX2
BRAF	DNTTIP1	HK1	MMS22L	PROB1	SMOX	VTN
BRF1	DNTTIP2	HMGCS1	MNT	PROKR1	SMPD2	WASF1
BRF2	DOCK4	HMX3	MOCS1	PRPF40A	SMTN	WASF2
BRINP1	DPAGT1	HNF1A	MOGAT2	PRR15	SMTNL2	WASL
BRINP2	DPCD	HNF1B	MOGAT3	PRR5L	SMYD1	WDFY1
BROX	DPP6	HNRNPUL1	MOGS	PRR9	SNAPC5	WDFY4
BRS3	DPT	HNRNPUL2	MORC2	PRRC1	SNED1	WDR19
BTBD10	DRAM2	HOXA10	MOS	PRSS8	SNRK	WDR31
BTBD17	DRC1	HOXB2	MOSPD2	PSD4	SNRNP200	WDR33
BTBD18	DRC1	HOXC10	MPDZ	PSMD10	SNTA1	WDR43
BTD	DROSHA	HOXC12	MPEG1	PTGES2	SNW1	WDR45
BTLA	DSE	HOXD12	MPHOSPH6	PTGFR	SNX13	WDR6
BUD23	DSEL	HOXD13	MPHOSPH8	PTGFRN	SNX21	WDR62
BVES	DSG1	HOXD4	MPLKIP	PTH1R	SNX33	WDTC1
C10orf88	DSG3	HPS3	MPP4	PTK2	SORL1	WFIKKN2
C14orf93	DTD2	HPS5	MRAS	PTK2B	SOWAHC	WIPI1
C19orf47	DUOX1	HRH1	MRM1	PTPRN	SOX11	WNT8A
C1orf35	DUSP16	HS3ST5	MRPL2	PUS7L	SOX17	WWC2
C1orf43	DZIP1L	HSF5	MRPL37	PWP2	SOX4	WWTR1
C1QBP	EDC4	HSPA4	MRPL54	PXDC1	SP4	XPO5
C2orf69	EDEM1	HTR2B	MRPS10	PXK	SPACA4	XPO6
C2orf74	EEF1AKMT1	HTR6	MRPS30	PXYLP1	SPAG1	XPOT
C3orf62	EFEMP1	HYCC1	MRPS34	PYGO1	SPAG4	YDJC
C3orf70	EFR3A	IAH1	MRTFA	PYROXD1	SPAG8	YIPF2
C6orf47	EGFR	ICA1L	MS4A5	PYURF	SPARCL1	YKT6
C6orf62	EGLN2	ICAM5	MSC	QSER1	SPATA18	YRDC
C7	EIF2AK3	ICE1	MSTN	QTRT2	SPATA22	ZBTB11
C9orf40	EIF2AK4	ICMT	MTF2	R3HDM2	SPATA6	ZBTB44
CA14	EIF2B2	ICOS	MTMR1	RAB11FIP1	SPHKAP	ZBTB7B
CACHD1	EIF3H	IER2	MTMR10	RAB11FIP2	SPRY2	ZBTB9

CACNA1D	EIF4G1	IFNB1	MTMR6	RAB29	SPRY3	ZC3H12B
CACNB1	ELFN1	IFRD1	MTRFR	RAB42	SPRYD3	ZCCHC24
CALCOCO1	ELOVL3	IGF2BP2	MTSS1	RABEP1	SQSTM1	ZCWPW2
CALHM1	ELP1	IGFBP2	MVB12A	RABEP2	SRPRA	ZDHHC13
CALHM2	EMC8	IGFBP5	MYBPC1	RACGAP1	SRXN1	ZDHHC16
CAMK1	EMD	IGSF1	MYC	RAD1	SSMEM1	ZDHHC6
CAMSAP1	EML3	IGSF10	MYCBP2	RAD23A	SSX2IP	ZFC3H1
CAMTA2	EMP3	IKBKB	MYF5	RAD54L2	ST18	ZFHX2
CAP1	EMX1	IL1RL2	MYO1B	RAF1	ST6GALNAC6	ZFHX4
CAP2	EN2	IL2	MYO1G	RAG1	ST8SIA1	ZFPL1
CAPN11	ENPP1	IL23A	MYSM1	RAI14	ST8SIA2	ZFYVE16
CAPRN2	ENTHD1	IL4I1	N4BP1	RALGAPA2	ST8SIA3	ZFYVE27
CAPZB	ENTPD6	IL6ST	NAB2	RALGAPB	ST8SIA5	ZFYVE9
CARMIL1	EOMES	IMMT	NADSYN1	RALGDS	STAC2	ZHX1
CASS4	EP300	INPP5J	NAGPA	RANBP10	STAT2	ZHX2
CAV2	EPHA10	INPPL1	NANOS2	RANBP6	STAT6	ZKSCAN5
CAVIN2	EPN2	INTS10	NAP1L2	RAPGEF2	STEAP3	ZKSCAN8
CBR4	EPN3	INTS11	NCAPD3	RAPGEF6	STIM1	ZMAT5
CBY2	ERCC4	INTS8	NCAPH	RARRES2	STK33	ZMIZ2
CCDC113	ERCC6	INTS9	NCOA2	RASAL1	STOML1	ZMPSTE24
CCDC115	ERI2	INTU	NCOA3	RASGRP3	STPG4	ZMYM2
CCDC121	ERI3	IP6K1	NCOR1	RASSF5	STRA6	ZNF182
CCDC146	ESR1	IP6K3	NDUFAF7	RASSF8	STRIP1	ZNF292
CCDC157	ESR2	IQCB1	NDUFB10	RAVER2	STRN3	ZNF3
CCDC175	ESYT3	IQCG	NDUFB7	RAX	STS	ZNF32
CCDC183	ETFDH	IQCK	NDUFC2	RB1CC1	STX18	ZNF335
CCDC184	ETV1	IQUB	NDUFV1	RBM12B	SUN1	ZNF367
CCDC191	EVI5	IRAK1BP1	NEIL3	RBM20	SUPT16H	ZNF48
CCDC60	EVL	ISCA2	NELFE	RBM28	SUPT7L	ZNF517
CCDC65	EVX1	ISG20L2	NEO1	RBM42	SV2A	ZNF541
CCDC70	EXD1	IST1	NEU4	RBP3	SVBP	ZNF583
CCDC81	EXOC2	ITGA6	NEURL4	RBP4	SWAP70	ZNF592
CCDC87	EXOG	ITGA7	NEXMIF	RBPJL	SYBU	ZNF608
CCN2	EXOSC3	ITGB1	NFASC	RCBTB1	SYDE2	ZNF687
CCND1	EXOSC5	ITGB5	NFAT5	RCE1	SYNM	ZNF750
CCND2	EXOSC9	ITIH4	NFE2L1	RCL1	SYVN1	ZNF789
CCND3	EYA2	ITPRIP	NFIL3	RCVRN	TACR2	ZNF792
CCNI	F13B	ITSN2	NHP2	RD3	TAF6	ZNF804A
CCNJ	F9	JADE2	NHSL1	RD3L	TANC1	ZNF821
CCNT2	FAAH	JAG1	NID2	RDH5	TAOK1	ZNF830
CCPG1	FABP3	JAG2	NIF3L1	RELA	TBC1D10A	ZNF839
CCT7	FADS2	JMJD1C	NIM1K	RELCH	TBC1D31	ZNHIT2
CD1E	FAF1	JRKL	NISCH	REPS1	TBCB	ZP3
CD36	FAH	KANK3	NKAPD1	RETREG2	TBR1	ZPLD1
CD44	FAM124B	KANSL3	NLRC3	REV3L	TBRG1	ZSWIM1
CD68	FAM131A	KATNAL1	NLRP11	RFC4	TBRG4	ZSWIM2
CDC25A	FAM13C	KCNA1	NLRP3	RFT1	TBX18	ZUP1
CDC37	FAM151B	KCNA10	NMT2	RFX2	TBX5	ZYG11B
CDC37L1	FAM163B	KCNA3	NOL4	RFX6	TCAIM	ZZZ3
CDC42BPB	FAM166A	KCNA6	NOL7	RGS9BP	TCHP	
CDC42EP2	FAM171A2	KCNH3	NOL9	RHBDD1	TCOF1	
CDH20	FAM181B	KCNH8	NOS1	RHCG	TDRD9	
CDH8	FAM187B	KCNK1	NPAS2	RHO	TELO2	
CDK15	FAM219B	KCNK17	NR1H4	RHOD	TENT5B	