UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
DEPARTAMENTO DE GENÉTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Muito além do lisossomo: análise de genes lisossômicos utilizando estratégias de bioinformática

Gerda Cristal Villalba Silva

Tese submetida ao Programa de Pós-graduação em Genética e Biologia Molecular da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do grau de ***Doutora em Ciências (Genética e Biologia Molecular)****.*

Orientadora: Profa. Dra. Ursula Matte

Porto Alegre, 2021

O presente trabalho de tese foi realizado no Hospital de Clínicas de Porto Alegre (HCPA), no Centro de Terapia Gênica do Centro de Pesquisa Experimental (CTG-CPE) e no Núcleo de Bioinformática - HCPA. Os estudos foram financiados pelo Fundo de Incentivo à Pesquisa e Eventos do HCPA (FIPE-HCPA), pela Fundação Médica do Rio Grande do Sul (FMRS) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

AGRADECIMENTOS

No começo do Universo, não havia estrelas, nem galáxias ou seres vivos. Apenas hidrogênio, hélio e muita energia. Primeiro agradeço a essa energia por tornar-nos capazes de existir, pensar e refletir. Agradeço a energia dos meus familiares, por terem sempre me incentivado a estudar, por aguçar minha curiosidade acerca do universo e de tudo mais. Agradeço a todos e todas que pela profissão de professor, que algum dia transmitiram conhecimento para mim, gerando a minha energia e a vontade de estudar e fazer ciência.

Agradeço aos colegas do Centro de Terapia Genica e do Núcleo de Bioinformática, por todo o apoio, as conversas regadas a café, cerveja, lágrimas e preocupações acadêmicas e da vida. A energia que eu sempre tive para ajudá-los e ouvi-los sempre foi convertida em bons momentos. Agradeço com destaque aos colegas Martiela e Bragatte pelos momentos noturnos no nosso arround e google meet elaborando nossas “escrituras sagradas”, codinomes para nossos trabalhos de tese.

Agradeço, principalmente nos momentos da pandemia, minhas duas gatinhas, Marie e Sophie, que foram essenciais para a distração perante todo o caos desses dias obscuros. E ao meu amor, que sempre me incentivou, e que demanda e despende muita energia também.

Meu profundo agradecimento a todas as mulheres cientistas que deixaram suas marcas na minha trajetória acadêmica. Somos mais que guerreiras, somos companheiras, amigas, cientistas, exímias profissionais que estão pouco a pouco mudando essa sociedade patriarcal e machista.

Agradeço à minha orientadora, psicóloga, professora e *coach* Dra. Ursula Matte, por ter me aceitado no seu grupo, por sempre mostrar que eu sou capaz, e por embarcar nas minhas ideias malucas e sempre encontrar soluções viáveis para nossos projetos. Obrigada por podar algumas ideias excipientes, pois assim poupei muita energia.

Agradeço a todos os cafés maravilhosos que o Elmo e a equipe do PPGBM passaram, pois estes me deram muita energia para aguentar as manhãs e as longas tardes no Campus do Vale.

Ao PPGBM, e a representação discente, a qual eu tive a oportunidade de fazer parte, pela oportunidade de estudar, e de fazer com que a voz dos estudantes seja ouvida.

Agradeço também as agências de fomento, que no meio de tantos desmontes sempre nos ajudam a continuar batalhando, visando melhorias para os pacientes com doenças raras. Para os pacientes, agradeço também, por mais que eu tenha utilizado somente dados públicos, é para eles que fazemos pesquisa, para que um dia tenham uma melhoria na qualidade de vida.

Eu não poderia deixar de agradecer ao nosso querido corvo portador da chave do conhecimento, mais conhecido como *Scihub*, que me permitiu poupar muito tempo encontrando os artigos necessários para o embasamento desta tese.

Por fim, agradeço a Vida, o Universo e tudo mais.

SUMÁRIO

Lista de figuras

Lista de tabelas

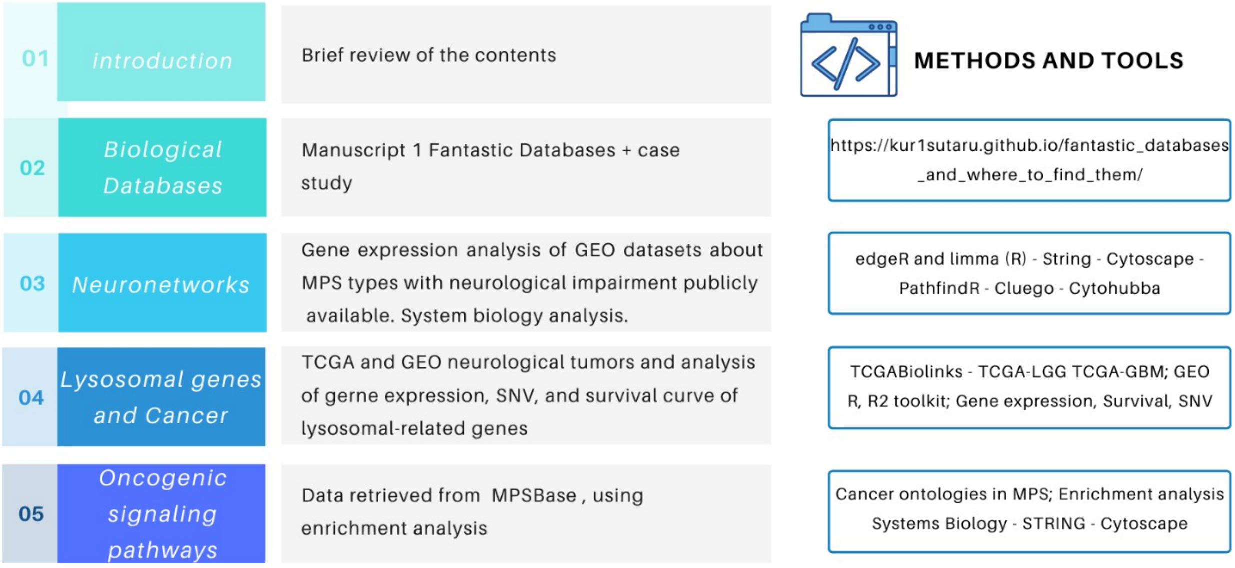
RESUMO

Os lisossomos são responsáveis pela degradação de macromoléculas e participam de diversos processos biológicos. Os lisossomos contêm hidrolases ácidas dentro deles, e defeitos nessas enzimas culminam no acúmulo de metabólitos ou macromoléculas, levando a doenças de depósito lisossomal (DDL). O lisossomo e suas enzimas também participam de diversos processos oncogênicos, principalmente em tumores neurológicos. Explorar genes lisossômicos usando dados ômicos disponíveis em bancos de dados multi-ômicos públicos pode ajudar a entender os mecanismos comuns envolvidos na fisiopatologia de DDLs e tumores neurológicos. O primeiro manuscrito analisa ferramentas da web e bancos de dados de dados ômicos e fornece um repositório da web com um estudo de caso. O manuscrito 2 é uma análise da biologia de sistemas de conjuntos de dados sobre dano neurológico em um grupo particular de DDLs (Mucopolissacaridoses, MPS) disponíveis em um repositório público de dados genômicos. O manuscrito 3 apresenta uma análise de expressão gênica, sobrevivência e análise de variantes de genes lisossomais em tumores neurológicos disponíveis no portal *Genomic data commons* (GDC, consórcio TCGA) e GEO. O manuscrito 4 contém dados recuperados de nossa ferramenta web MPSBase de vias de sinalização oncogênica em MPS, usando análises de enriquecimento. Esses estudos podem ajudar a ampliar as abordagens terapêuticas para distúrbios lisossomais, apontando mecanismos comuns com doenças mais prevalentes.

ABSTRACT

The lysosomes are responsible for the degradation and participate in several biological processes. Lysosomes contain acid hydrolases within them, and defects in these enzymes culminate in the accumulation of metabolites or macromolecules, leading to lysosomal storage diseases (LSD). The lysosome and its enzymes also participate in several oncogenic processes, especially in neurological tumors. To explore lysosomal genes using omics data available in public multi-omic databases may help to understand the common mechanisms involved in the pathophysiology of LSDs and neurological tumors. The first manuscript reviews web tools and databases of omics data and provides a web repository with a case study. Manuscript 2 is a system biology analysis of datasets about neurological impairment in a particular group of LSDs (Mucopolysaccharidoses, MPS) available at a public functional genomics data repository. Manuscript 3 presents a gene expression, survival, and SNV analysis of lysosomal genes in neurological tumors available in the genomic data commons portal (GDC, TCGA consortium) and GEO. Manuscript 4 contains data retrieved from our web tool MPSBase of oncogenic signaling pathways in MPS, using enrichment analysis. These studies may help amplify therapeutic approaches for lysosomal disorders by pointing common mechanisms with more prevalent diseases.

FLUXOGRAMA / GRAPHICAL ABSTRACT



METHODS AND TOOLS

01	introduction	Brief review of the contents	
02	Biological Databases	Manuscript 1 Fantastic Databases + case study	https://kur1sutaru.github.io/fantastic_databases_and_where_to_find_them/
03	Neuronetworks	Gene expression analysis of GEO datasets about MPS types with neurological impairment publicly available. System biology analysis.	edgeR and limma (R) - String - Cytoscape - PathfindR - Cluego - Cytohubba
04	Lysosomal genes and Cancer	TCGA and GEO neurological tumors and analysis of gerne expression, SNV, and survival curve of lysosomal-related genes	TCGABiolinks - TCGA-LGG TCGA-GBM; GEO R, R2 toolkit; Gene expression, Survival, SNV
05	Oncogenic signaling pathways	Data retrieved from MPSBase , using enrichment analysis	Cancer ontologies in MPS; Enrichment analysis Systems Biology - STRING - Cytoscape

Nota aos leitores

O presente trabalho de tese apresenta 4 manuscritos que foram escritos como resultado da pesquisa da autora como aluna de doutorado no programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM). É importante saber que a maior parte do trabalho desta tese foi elaborado durante a pandemia da COVID-19 que nos acometeu mundialmente.

O primeiro capítulo é uma revisão de todos os conteúdos abordados na tese, partindo do lisossomo até bancos de dados especializados em doenças lisossomais. O capítulo 2 apresenta os objetivos da tese, que foram respondidos através de diferentes trabalhos, apresentados nos capítulos 3 a 6.

O primeiro manuscrito é uma revisão de ferramentas web e bancos de dados de dados ômicos, acompanhada de um repositório web com um estudo de caso ilustrando os resultados coletados, de forma a exemplificar algumas usabilidades dos bancos de dados selecionados. Este manuscrito foi publicado na revista GMB - *Genetics and Molecular Biology*.

O segundo manuscrito é uma análise de biologia de sistemas de conjuntos de dados sobre comprometimento neurológico em mucopolissacaridoses (MPS) disponíveis em um repositório público de dados genômicos funcionais. Este trabalho foi submetido para a revista *Neuroinformatics*.

O manuscrito 3 apresenta dados de expressão gênica e de SNV de genes lisossomais em tumores neurológicos disponíveis no portal *genomic data commons* (GDC, consórcio TCGA) e GEO. O manuscrito foi submetido para o *Journal of Neuro-Oncology.*

O manuscrito 4 foi gerado a partir de dados recuperados de nossa ferramenta web MPSBase, sobre de vias de sinalização oncogênica em MPS, usando análises de enriquecimento. O manuscrito foi publicado na *Lecture Notes in Bioinformatics*.

Na parte final, encontram-se os trabalhos e colaborações realizadas durante o período da tese.

Capítulo 1

INTRODUÇÃO

Os Erros Inatos do Metabolismo constituem um grupo de cerca de 750 doenças genéticas que podem afetar a síntese, degradação, processamento e transporte de moléculas no organismo. Individualmente são doenças raras, mas quando reunidas apresentam uma prevalência estimada de 1:800 indivíduos. Em sua maioria, os erros inatos do metabolismo possuem padrão de herança autossômico recessivo (Saudubray & García-Cazorla, 2018).

A classificação recente proposta por Saudubray & García-Cazorla (2019) divide essas doenças em três grandes grupos, dentre eles destacamos o grupo 3, onde estão agrupadas desordens no metabolismo de moléculas complexas que acometem organelas como os lisossomos, peroxissomos ou complexo de golgi, tais como as esfingolipidoses, doença de Fabry, Niemann-pick C, Lipofuscinose ceróide neuronal, doença de Krabbe, as gangliosidoses, mucolipidoses e as mucopolissacaridoses (Beck, 2007; Hoffmann *et al.*,2017). O esquema a seguir mostra desordens que se enquadram nessa classificação:

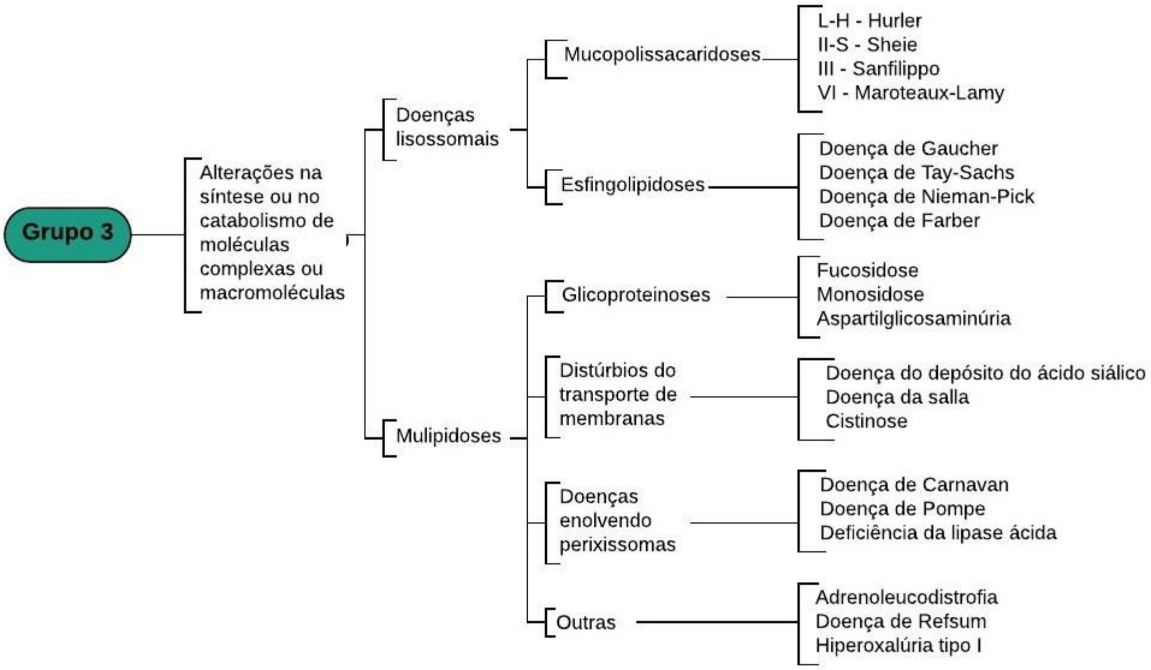


Figura 1: Doenças classificadas no grupo 3 onde observa-se alterações na síntese ou catabolismo de macromoléculas ou moléculas complexas. (Adaptado de Saudubray & García-Cazorla, 2018).

No que diz respeito ao grupo 3, as doenças que acometem os lisossomos, mais conhecidas como doenças de depósito lisossômico ou depósito lisossomal, são caracterizadas por defeitos em genes que codificam enzimas envolvidas na degradação de substratos que culminam no seu acúmulo de macromoléculas no interior dos lisossomos e em defeitos secundários (Donati *et al.*, 2018).

O termo **lisossomo** tem origem nas palavras gregas *lise* (destruição ou dissolução) e *soma* (corpo). A descoberta dessa organela ocorreu em 1949, pelo citologista belga Christian de Duve, que estudava os mecanismos de ação da insulina em células hepáticas. O grupo de de Duve tinha como foco de pesquisa a enzima glicose 6-fosfatase, e ao isolá-la pelo método de fraccionamento celular, identificaram algumas organelas membranosas que foram denominadas de lisossomos. Nessa época, os pesquisadores descreviam os lisossomos como "bolsas suicidas" (Ballabio, 2016).

Os lisossomos, classicamente conhecidos por sua função de degradação, também exercem uma gama de papéis fundamentais para a manutenção dos organismos, tais como listados na figura 2:

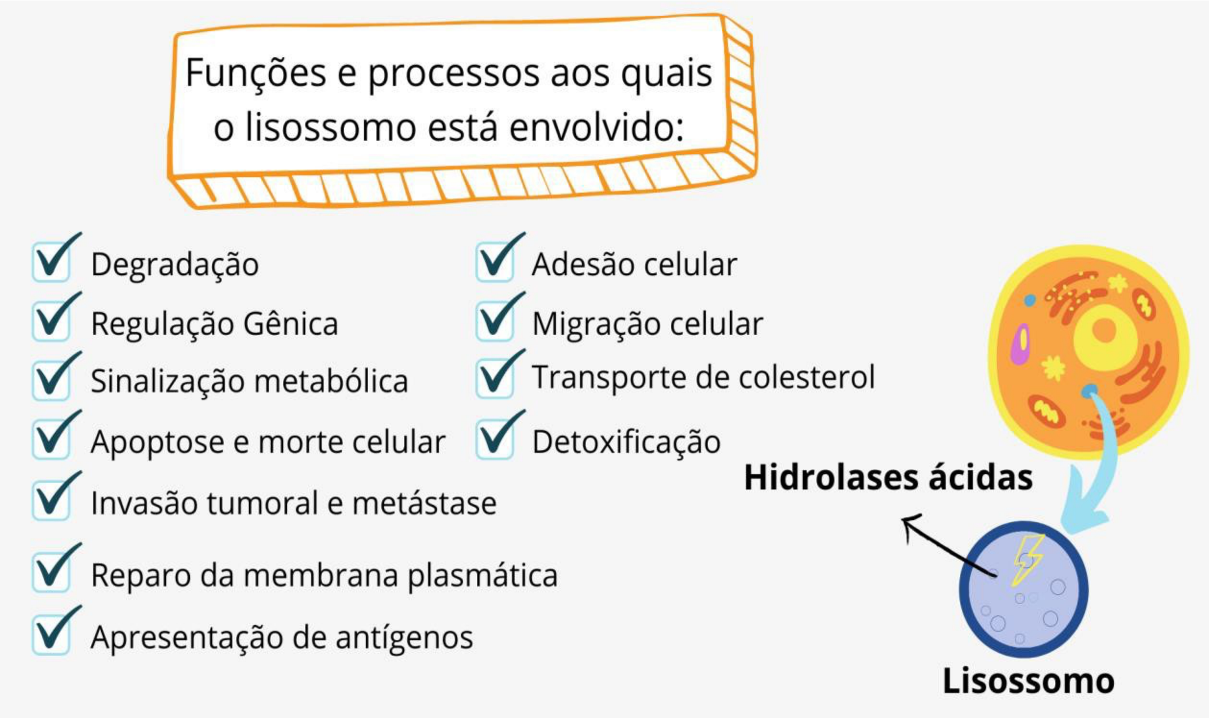

Funções e processos aos quais o lisossomo está envolvido:
✓ Degradação
✓ Regulação Gênica
✓ Sinalização metabólica
✓ Apoptose e morte celular
✓ Invasão tumoral e metástase
✓ Reparo da membrana plasmática
✓ Apresentação de antígenos
✓ Adesão celular
✓ Migração celular
✓ Transporte de colesterol
✓ Detoxificação
Hidrolases ácidas
Lisossomo

Figura 2: Funções do lisossomo. Defeitos em enzimas lisossômicas, tais como as hidrolases ácidas culminam em um grupo de Erros Inatos do Metabolismo denominado de Doenças de Depósito Lisossomal (DDL). Adaptado de Meyer-Schwesinger, 2021.

Devido à vasta gama de processos biológicos nos quais o lisossomo participa, diversos receptores e proteínas interagem com proteínas lisossômicas, sendo assim, as disfunções em lisossomos são descritas em câncer (Tang *et al.*, 2020), e em doenças neurodegenerativas, tais como Alzheimer (Wie *et al.*, 2021) e Parkinson (Navarro-Romero *et al.*, 2020).

No contexto tumoral, a cascata de eventos carcinogênicos consiste em uma série de mecanismos de controle da proliferação celular, onde a desregulação deste processo gera o crescimento demasiado e descontrolado de células, formando assim tumores (Popolin & Cominetti, 2017; Machado *et al.,* 2021). O microambiente tumoral possui uma vastidão de proteínas associadas a fatores de crescimento, os quais desempenham importante papel na sinalização entre as células tumorais e as outras células do seu microambiente (Corn *et al.,* 2013). Um exemplo clássico é a transformação oncogênica ocasionada por alterações em vias metabólicas, como a super ativação de Ras e da via fosfatidilinositol-3-cinase classe I (PI3KI) /Akt (Cairns *et al*., 2011). Enquanto Akt aumenta os níveis transcricionais de enzimas como a hexoquinase e fosfofrutoquinase, K-Ras promove a transcrição de diversas outras enzimas da via glicolítica (Fan *et al.,* 2010). Os lisossomos também atuam nessas vias de sinalização, e em outros processos como a autofagia (Geisslinger *et al.*, 2020).

Dentre as principais vias de sinalização alteradas em tumores, uma das mais estudadas é a via da autofagia. A autofagia é um processo evolutivamente conservado em diversas espécies, sendo ativada em resposta a diferentes situações de estresse, como por exemplo, hipóxia, depleção de nutrientes e fatores de crescimento e danos no DNA. Dentre as diferentes respostas, as que mais se destacam são a hipóxia e ausência de fatores de crescimento, e estas são reguladas através do complexo mTORC1, que se encontra na superfície dos lisossomos (Lum *et al.,* 2005; Jung *et al.*, 2013).

A via de sinalização de mTORC1, além de regular as vias lisossomais, também regula positivamente a família de fatores de transcrição do tipo EB, tais como TFEB, que conhecidamente regula a maior parte dos genes lisossomais (Puertollano, 2014). Em situações celulares normais, mTORC1 sequestra TFEB no citosol, e os lisossomos direcionam-se para a periferia das células. Em situações de estresse, a inativação de mTORC1 promove a translocação de TFEB para o núcleo, a indução de da biogênese lisossomal e posterior ativação da autofagia. Essa inativação de mTORC1 é necessária para que ocorra a formação do autofagolissomo (Zhao *et al.,* 2020). A interação entre mTORC1 e os lisossomos também é essencial para regular diversas vias oncogênicas, tais como Akt, Ras, Mek e outras vias de sinalização reguladas upstream, tais como as vias dos receptores tirosina kinases (Asrani *et al.,* 2019).

A inibição ou super ativação da via de autofagia são fatores determinantes na progressão tumoral e podem nortear o diagnóstico e o tratamento de diferentes tipos de câncer (Swartz *et al.*, 2012). Além disso, a limitação nos níveis de glicose induz aos processos autofágicos, que, nesse caso, podem servir como fonte alternativa de produção energética. Além disso, sabe-se que a autofagia exerce papel na eliminação de moléculas e organelas danificadas.

A autofagia compreende o sequestro de proteínas, organelas citosólicas e patógenos em estruturas chamadas de autofagossomos. Os autofagossomos, por sua vez, são estruturas que fundem-se aos lisossomos e degradam o material celular através da ação das hidrolases lisossomais. O resumo da formação do autofagolisossomo pode ser observado na figura 3:

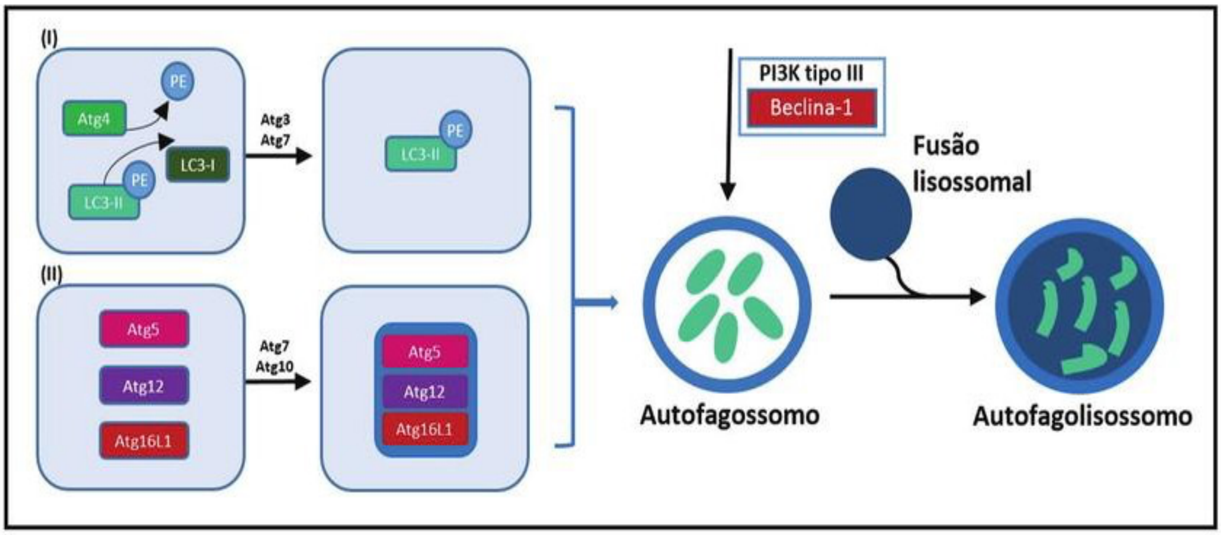


Figura 3: Complexos proteicos envolvidos na formação do Autofagolisossomo. A beclina-1, em conjunto com outras proteínas, participa do complexo da PI3K do tipo III, responsável por recrutar proteínas relacionadas à autofagia (Atg), iniciando assim o processo de formação do autofagossomo. Os complexos ATG iniciam a formação da membrana autofagossômica, que leva à fusão entre o autofagossomo e a membrana lisossômica. (Adaptado de Netea *et al.*, 2015).

O processo de degradação resulta na liberação de aminoácidos livres e macromoléculas, os quais são transportados de volta ao citosol para serem reutilizados (Kung *et al.,* 2011; Seranova *et al.,* 2017). Tendo em vista esses processos, recentemente verificou-se o envolvimento dessas estruturas em processos de proliferação e sinalização celular, indução de angiogênese e processos metastáticos (Pastores & Hughes, 2017; Davidson & Vander Heiden, 2017).

Estudos recentes sobre doenças de depósito lisossomal e câncer envolvem a doença de Gaucher tipo 1 e Mieloma Múltiplo (Pastores & Hughes, 2017), entretanto os mecanismos de interação entre o acúmulo de substratos e os processos de tumorigênese ainda são pouco conhecidos. Outras associações entre genes lisossomais e câncer foram descritos, como compilado na figura a seguir:

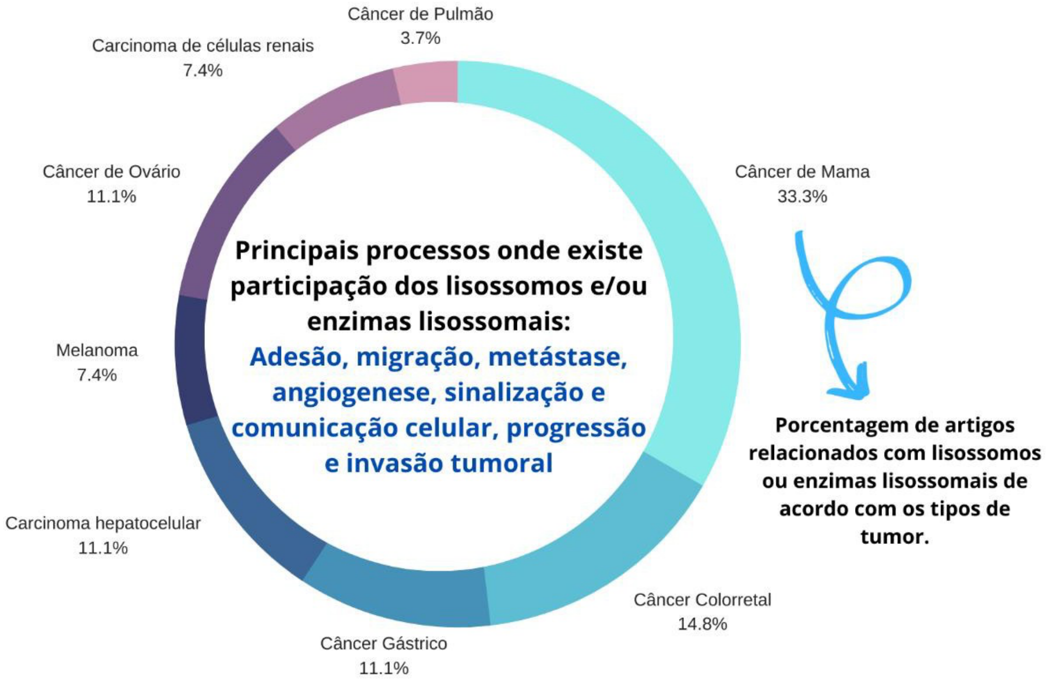


Figura 4: Levantamento de dados da literatura acerca de lisossomos e / ou enzimas lisossomais em câncer. Os dados foram compilados no artigo resultante do capítulo 4 da presente tese. A imagem é uma adaptação da figura 1 e da tabela suplementar 4 do referido artigo. Para mais informações, ler capítulo 4.

Disfunção ou variantes patogênicas podem impactar as hidrolases lisossômicas, culminando nas doenças de depósito lisossomal (DDL). A lista oficial de doenças lisossômicas pode ser encontrada através do link a seguir https://worldsymposia.org/official-list-of-lysosomal-diseases/. A lista de genes e doenças lisossômicas analisadas neste trabalho pode ser vista na Tabela 1.

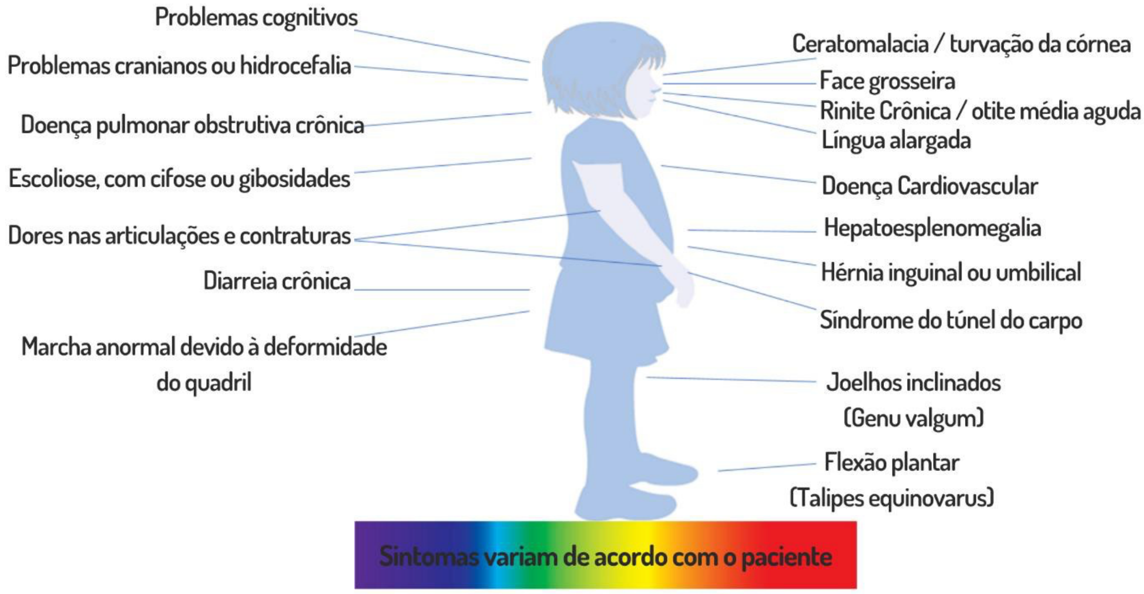
Tabela 1: Lista dos genes lisossômicos analisados neste trabalho de tese. A lista dos genes relacionados a doenças de depósito lisossomal foi retirada do website Worldsymposia.

Gene	Localização	OMIM	Doença
ABHD5	3p21.33	604780	Síndrome de Chanarin-Dorfman
AGA	4q34.3	613228	Aspartilglicosaminúria
ARSA	22q13.33	607574	Leucodistrofia metacromática
ARSB	5q14.1	611542	Mucopolissacaridose tipo VI (Maroteaux-Lamy)
ASAH1	8p22	613468	Lipogranulomatose de Farber ou Deficiência de ceramidase
CLCN5	Xp11.23	300008	Nefrolitíase hipercalciúrica ligada ao X (Doença de Dent)
CLN3	16p12.1	607042	Lipofuscinose Ceróide neuronal tipo 3 (Síndrome de Batten)
CLN5	13q22.3	608102	Lipofuscinose Ceróide Neuronal tipo 5
CLN6	15q23	606725	Lipofuscinose Ceróide Neuronal tipo 6 Adulta (Kuffs)
CLN8	8p23.3	607837	Lipofuscinose Ceróide Neuronal tipo 8, variante da epilepsia do norte
CTNS	17p13.2	606272	Cistinose
CTSA	20q13.12	613111	Galactosialidose
CTSK	1q21.3	601105	Picnodisostose
FIG4	6q21	609390	Doença de Charcot-Marie-Tooth tipo 4J
FUCA1	1p36.11	612280	Fucosidose
GAA	17q25.3	606800	Doença de Pompe
GALC	14q31.3	606890	Leucodistrofia de células globóides (Krabbe)
GALNS	16q24.3	612222	Mucopolissacaridose tipo IVA (Morquio)
GBA	1q22	606463	Doença de Gaucher
GLA	Xq22.1	300644	Doença de Fabry
GLB1	3p22.3	611458	Gangliosidose tipo I
GM2A	5q33.1	613109	Doença de Tay-Sachs

GNPTAB	12q23.2	607840	Mucolipidose tipo II
GNPTAG	16p13.3	607838	Mucolipidose tipo III
GNS	12q14.3	607664	Mucopolissacaridose tipo IIID (Sanfilippo)
GUSB	7q11.21	611499	Mucopolissacaridose tipo VII (Sly)
HEXA	15q23	606869	Gangliosidose tipo II (Juvenil)
HEXB	5q13.3	606873	Doença de Sandhoff (Gangliosidose GM2)
HGSNAT	8p11	610453	Mucopolissacaridose tipo IIIC (Sanfilippo)
HYAL1	3p21.31	607071	Mucopolissacaridose tipo IX (Natowicz)
IDS	Xq28	300823	Mucopolissacaridose tipo II (Hunter)
IDUA	4q16.3	607014	Mucopolissacaridose tipo I (Hurler)
LAMP2	Xq24	309060	Doença de Dannon
LIPA	10q23.31	613497	Deficiência de Lipase ácida lisossomal
MAN2B1	19q13.13	609458	Alfa-manosidose
MANBA	4q24	609489	Beta-manosidose
MCOLN1	19p13.2	605248	Mucolipidose tipo IV
NAGA	22q13.2	609242	Doença de Kanzaki
NAGLU	17q21.2	609701	Mucopolissacaridose tipo IIIB (Sanfilippo)
NEU1	6p21.33	608272	Sialidose
NPC1	18q11.2	607623	Niemann-Pick tipo C1
NPC2	14q24.3	601015	Niemann-Pick tipo C2
OCRL	Xq26.1	300535	Doença de Dent tipo 2
PNPLA2	11p15.5	609059	Doença por armazenamento lipídico neutro
PPT1	1p34.2	600722	Lipofuscinose Ceróide neuronal tipo 1
PSAP	10q22.1	176801	Deficiência de Prosaposina
SGSH	17q25.3	605270	Mucopolissacaridose tipo IIIA (Sanfilippo)
SLC17A5	6q13	604322	Doença de armazenamento do ácido siálico livre (forma infantil)
SLC9A6	Xq26.3	300231	Síndrome de Christianson
SMPD1	11p15.4	607608	Niemann-Pick tipo A and B
SUMF1	3p26.1	607939	Deficiência múltipla de sulfatases (Doença de Austin)
TPP1	11p15.4	607998	Lipofuscinose Ceróide neuronal tipo 2

Adentrando o universo das doenças lisossomais, mais especificamente das mucopolissacaridoses (MPS), este grupo de doenças lisossômicas de depósito são causadas pela deficiência de enzimas essenciais para o metabolismo de componentes da matriz extracelular como os glicosaminoglicanos (GAG). Quando não degradados da maneira correta, os GAGs são estocados nos compartimentos lisossomais das células, acarretando uma série de complicações progressivas e multissistêmicas (Harper *et al.,* 1998; Scriver *et al.,* 2001).

As MPS de modo geral possuem padrão de herança autossômico recessiva, com exceção da MPS tipo II, Síndrome de Hunter, cujo padrão de herança é ligado ao X. Diversos fenótipos clínicos foram descritos com o passar das décadas, mas na maior parte dos casos pode-se constatar acometimentos neurológicos, os quais os mecanismos da fisiopatologia da doença ainda não foram bem elucidados (Campos & Monaga, 2012). A figura 5 elenca alguns dos principais sintomas das MPS.


Problemas cognitivos
Problemas cranianos ou hidrocefalia
Doença pulmonar obstrutiva crônica
Escoliose, com cifose ou gibosidades
Dores nas articulações e contraturas
Diarreia crônica
Marcha anormal devido à deformidade do quadril
Ceratomalacia / turvação da córnea
Face grosseira
Rinite Crônica / otite média aguda
Língua alargada
Doença Cardiovascular
Hepatoesplenomegalia
Hérnia inguinal ou umbilical
Síndrome do túnel do carpo
Joelhos inclinados (Genu valgum)
Flexão plantar (Talipes equinovarus)
Sintomas variam de acordo com o paciente

HTTP://WWW.SANOFIMEDICALAFFAIRS.COM/RAREDISEASES/INDEX.PHP/21-MPS-I

Figura 5: Esquema geral dos principais acometimentos encontrados nos pacientes com Mucopolissacaridoses. Adaptado do site da Sanofi Rare Diseases.

A classificação das MPS baseia-se especificamente na deficiência enzimática propriamente dita, apesar do fato que existem fenótipos distintos dentro do mesmo defeito enzimático, como no caso da MPS do tipo I (Hurler, Hurler-Scheie e Scheie), bem como fenótipos semelhantes para deficiências enzimáticas diferentes, no caso dos acometimentos neurológicos encontrados predominantemente nas MPS do tipo I, II, III e VII (Neufeld & Muenzer, 2001; Myerowitz *et al.,* 2002). A figura 6 resume os tipos de MPS, os genes que codificam as enzimas defeituosas e o tipo de glicosaminoglicano acumulado:

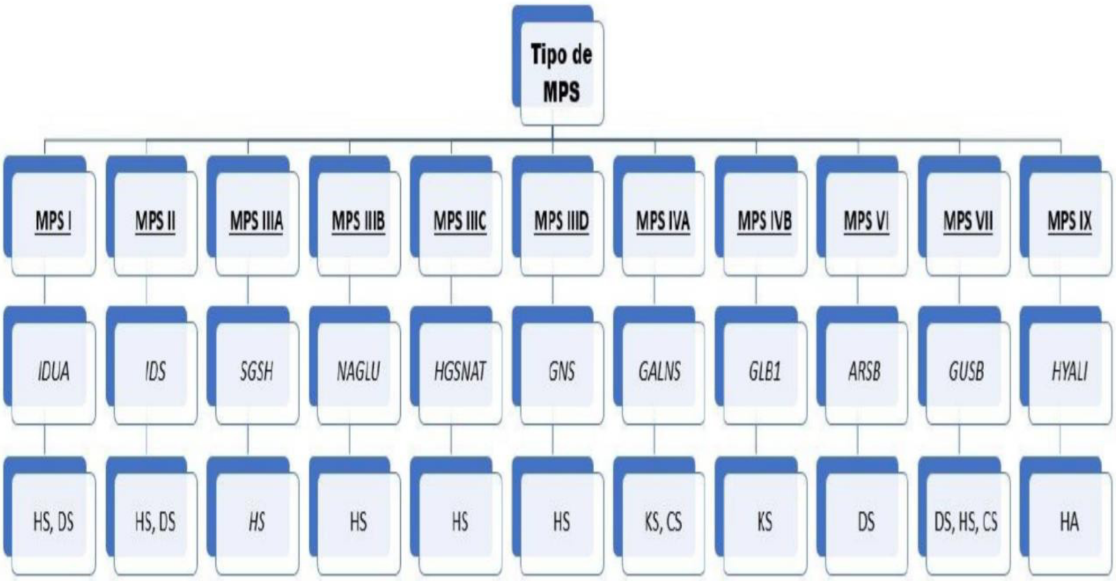


Figura 6: Classificação das Mucopolissacaridoses. Adaptado de Clarke (2008). MPS = Mucopolissacaridoses; AH = ácido hialurônico; CS = condroitin sulfato; DS = dermatan sulfato; HS = heparan sulfato; KS = keratan sulfato.

Com relação ao diagnóstico, as MPS não estão incluídas na triagem neonatal brasileira (Kubaski *et al*, 2020), mas diante de suspeita clínica realizam-se testes bioquímicos quantitativos ou qualitativos, a partir da dosagem de GAGs na urina, plasma, leucócitos ou fibroblastos. Já para aquelas MPS nas quais a enzima deficiente é da categoria de sulfatase, a dosagem deve ser realizada de modo comparativo com outra sulfatase, para excluir a possibilidade de diagnóstico de uma deficiência múltipla de sulfatases (Donati *et al.*, 2018). O diagnóstico genético pode contribuir para diferenciar as formas mais atenuadas dentro de uma mesma MPS, para detecção de portadores e, em alguns casos, relacionar genótipo com fenótipo bioquímico, apesar de não existir uma correlação totalmente certa entre as variantes genéticas encontradas e a atividade enzimática dentro das diferentes MPS (Muenzer, 2011).

Com o avanço das tecnologias de sequenciamento de larga escala e a redução relativa de custo de análises de exomas e transcriptomas, a tendência é que tais metodologias tornem-se recorrentes e frequentemente úteis no diagnóstico e estudo de doenças genéticas (Cummings *et al.*, 2017). Para isso, o desenvolvimento de estratégias e pipelines de análises acessíveis para os profissionais da saúde é essencial para o bom planejamento e escolha das metodologias adequadas, permitindo assim a rápida identificação de mutações novas e recorrentes, melhorando de forma global o diagnóstico, prognóstico e desenvolvimento de estratégias terapêuticas para o tratamento de inúmeras doenças e síndromes. Além do mais, a busca automatizada dos genes com expressão diferencial serve para otimizar o tempo de análise das informações (Pereira *et al.,* 2015).

Estudos envolvendo RNA-seq podem ser utilizados para desvendar mecanismos fisiopatológicos de acometimentos complexos, como dano neurológico em modelo murino de MPS II (Salvalaio *et al.*, 2017); imunossupressão do sistema nervoso central e progressão da doença em MPS IIIB (DiRosario *et al.,* 2008); mecanismos envolvidos na neurodegeneração em MPS IIIA, MPS IIIB e MPS IIIC (Moskot *et al.*, 2019), e edição de transcritos mutantes em Síndrome de Hunter - MPS II (Lualdi *et al.*, 2010). Também podem ser usados para avaliar níveis de citocinas, neurotrofinas e estresse oxidativo em MPS IIIB (Villani *et al.*, 2006); desenvolvimento de biomarcadores de diagnóstico, prognóstico e descobertas de novas drogas para MPS I (Swaroop *et al.*, 2018), e análise e descobrimento de biomarcadores relacionados a cardiopatias em MPS IIIB (Schiatarrela *et al.*, 2015).

Para organizar e armazenar informações como os estudos supracitados, de modo geral são necessários os bancos de dados. Por definição, um banco de dados biológico é uma coleção de dados organizados de forma sistemática, onde informações biológicas podem ser rapidamente acessadas, organizadas e manipuladas (Toomula *et al.,* 2012). Os bancos de dados geralmente são construídos em uma linguagem padrão com estrutura de queries, chamada de *Standard Query Language* - ou SQL (Kriegel *et al.,* 2004).

O primeiro banco de dados biológicos foi desenvolvido pela "Mãe" da Bioinformática, a Professora Margaret Dayhoff, em 1965. Esse banco foi criado na era dos cartões perfurados. Desse modo, Margaret criou o A*tlas of Protein Sequence* *and Structure*, que anos mais tarde virou o *Protein Data Bank*, semelhante ao *Genbank* (Benson *et al*., 2013). Um fato curioso: ela também desenvolveu a primeira matriz de substituição de mutações de ponto (PAM) e o código de uma letra para os aminoácidos (Strasser, 2012). Desde então, os bancos de dados biológicos são cruciais para acelerar o progresso das descobertas nos campos da Bioinformática e ciências da saúde (Imker, 2018).

O primeiro banco de dados para genes lisossômicos é o *The Human Lysosome Gene Database* (Brozi *et al.,*2013, <http://lysosome.unipg.it/index.php>), porém ele combina e analisa informações de genes lisossomais de maneira geral, com dados de miRNAs (microRNAs) e reguladores mestre da transcrição, como o TFEB. O mesmo contempla alguns estudos de proteoma, mas o banco não apresenta muitos estudos. Além disso, o banco conta com uma parte onde o usuário pode buscar artigos relacionados com biologia de sistemas, mas a base possui poucas informações. Para dados dos diferentes tipos de MPS, nosso grupo desenvolveu o MPSBase <https://www.ufrgs.br/mpsbase/ > que engloba apenas um grupo de doenças lisossomais, as Mucopolissacaridoses (MPS). Nesse banco de dados é possível consultar informações sobre genes diferencialmente expressos, e suas perspectivas vias biológicas (Soares *et al*., 2021).

Para consultar informações clínicas acerca de erros inatos do metabolismo, existe o *IEMBase*, < http://iembase.org/ >, um banco de dados que concentra diversas fontes de pesquisa, auxiliando no diagnóstico de doenças raras. O banco retorna uma lista classificada de possíveis distúrbios que correspondem ao perfil de entrada, por exemplo, um fenótipo ou sintoma. Além disso, o sistema pode sugerir possíveis testes que ajudariam a estreitar o diagnóstico diferencial e fornecer acesso a informações bioquímicas, moleculares e clínicas, além de evidências experimentais, tais como artigos (Lee *et al*., 2018).

RareLSD < https://webs.iiitd.edu.in/raghava/rarelsd/ > é um banco de dados com curadoria manual que apresenta informações sobre enzimas lisossomais associadas a doenças raras (Akhter *et al.*, 2019). O usuário pode buscar por informações completas sobre o distúrbio, incluindo o nome da doença, órgão afetado, idade de início, medicamento disponível, padrão de herança, enzima defeituosa e informações sobre variantes. Também disponibiliza as estruturas tridimensionais das enzimas lisossomais. Assim como o *IEMBase*, possui informações sobre ensaios bioquímicos e informações obtidas no *PubChem*. *RareLSD* possui integração com ferramentas de pesquisa de similaridade de sequência (por exemplo, BLAST e algoritmo Smith-Waterman).

Por último, até o momento, específico para descoberta de novos fármacos em erros inatos do metabolismo, citamos o *DDIEM: Drug Database for Inborn Errors of Metabolism* (Banco de dados de drogas para erros inatos do metabolismo, website: http://ddiem.phenomebrowser.net/). Essa ferramenta utiliza uma estrutura de queries diferentes das convencionais, o SPARQL (semelhante ao SQL, mas este utiliza grafos e linguagem semântica para consultar o banco de dados, através de arquivos com o RDF - *Resource Description Framework*). O *DDIEM* utiliza ontologias para categorizar os tratamentos sugeridos para um dado erro inato do metabolismo. Para contemplar essa busca, o banco possui informações acerca de 338 erros inatos, 643 drogas associadas, 1615 fenótipos, e 1899 artigos de referência (Abdelhakim *et al*., 2020).

Tendo em vista as amplas aplicações do uso de bancos de dados para o estudo de doenças genéticas raras, enfatizamos a importância de se estabelecer pipelines e workflows de estudos *in silico* para automatizar e organizar essas análises. Perante isso, o objetivo dessa tese é utilizar diversos bancos de dados para aprimorar o entendimento da fisiopatologia das doenças de depósito lisossomal e dos seus genes relacionados, mais especificamente do acometimento neurológico apresentado nas MPS, e identificar e estabelecer relações entre tumores neurológicos e vias oncogênicas nessas doenças.

Capítulo 2

OBJETIVOS

Objetivo Geral

Aprimorar o entendimento da fisiopatologia das doenças de depósito lisossomal e dos seus genes relacionados, mais especificamente do acometimento neurológico apresentado nas MPS, e identificar e estabelecer relações entre tumores neurológicos e vias oncogênicas nessas doenças através da utilização de dados depositados em bancos públicos.

Objetivos específicos:

Bancos de dados biológicos

- Fazer um levantamento dos bancos de dados biológicos existentes, classificando-os de acordo com diferentes subáreas da bioinformática;
- Elaborar uma lista de ferramentas para que pessoas sem muito conhecimento de linguagem de programação sejam capazes de conduzir pesquisas em bioinformática;
- Criar um estudo de caso para ilustrar a usabilidade dos bancos listados

Biologia de sistemas e dano neurológico em MPS

- Identificar possíveis genes e vias que sejam biomarcadores de dano neurológico nas diferentes MPS que possuem dados de transcriptomas disponíveis publicamente

Doenças lisossomais e tumores neurológicos

- Investigar a expressão genica de enzimas lisossomais em gliomas;
- Identificar variantes patogênicas em genes lisossômicos;
- Avaliar o impacto da expressão dessas enzimas na sobrevida dos pacientes

Vias de sinalização oncogênica em Mucopolissacaridoses

- Identificar vias biológicas de sinalização de câncer presentes nas diferentes MPS

Capítulo 3

O uso de bancos de dados biológicos em pesquisas biomédicas - como a bioinformática pode nos ajudar a resolver nossos problemas?

Para tentar responder essa pergunta, realizamos um levantamento de bancos de dados biológicos que foram compilados no seguinte manuscrito, publicado na revista Genetics and Molecular Biology - GMB. Para a leitura eletrônica, acesse

< https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8022358/>

Genetics and Molecular Biology, 44, 2, e20200203 (2021)
Copyright © Sociedade Brasileira de Genética.
DOI: https://doi.org/10.1590/1678-4685-GMB-2020-0203

Research Article
Genomics and Bioinformatics

Fantastic databases and where to find them: Web applications for researchers in a rush

Gerda Cristal Villalba[1,2,3] and Ursula Matte[1,2,3,4]

[1]*Hospital de Clínicas de Porto Alegre, Laboratório de células, tecidos e genes, Porto Alegre, RS, Brazil.*
[2]*Universidade Federal do Rio Grande do Sul, Programa de Pós-Graduação em Genética e Biologia Molecular, Porto Alegre, RS, Brazil.*
[3]*Hospital de Clínicas de Porto Alegre, Bioinformatics Core, Porto Alegre, RS, Brazil.*
[4]*Universidade Federal do Rio Grande do Sul, Departamento de Genética, Porto Alegre, RS, Brazil.*

Abstract

Public databases are essential to the development of multi-omics resources. The amount of data created by biological technologies needs a systematic and organized form of storage, that can quickly be accessed, and managed. This is the objective of a biological database. Here, we present an overview of human databases with web applications. The databases and tools allow the search of biological sequences, genes and genomes, gene expression patterns, epigenetic variation, protein-protein interactions, variant frequency, regulatory elements, and comparative analysis between human and model organisms. Our goal is to provide an opportunity for exploring large datasets and analyzing the data for users with little or no programming skills. Public user-friendly web-based databases facilitate data mining and the search for information applicable to healthcare professionals. Besides, biological databases are essential to improve biomedical search sensitivity and efficiency and merge multiple datasets needed to share data and build global initiatives for the diagnosis, prognosis, and discovery of new treatments for genetic diseases. To show the databases at work, we present a a case study using *ACE2* as example of a gene to be investigated. The analysis and the complete list of databases is available in the following website <https://kur1sutaru.github.io/fantastic_databases_and_where_to_find_them/>.

Keywords: Human databases; bioinformatics tools; web application; data mining; big data.

Received: June 30, 2020; Accepted: February 22, 2021.

Introduction

The advent of sequencing technologies has motivated the development of public databases initiatives. Recently, there has been an exponential growth in generation of biological data, and these require information technology tools to store, organize and analyze biological data that is available in the form of raw data, annotated sequences, tables, and other archive files generated by omics technologies (Toomula *et al.*, 2012).

The first reported biological database was a protein sequence database developed by Margaret Dayhoff in 1965. She also created the first substitution matrix for point accepted mutations (PAM) and the one-letter code for amino acids (Strasser, 2012). In the early 1980s, the EMBL Data Library (currently European Nucleotide Archive, <https://www.ebi.ac.uk/ena>) created a catalog of published biological data. A timeline with some historical facts about biological databases are provided in Figure 1.

A biological database is a collection of data organized in systematic contents that can quickly be accessed, managed, and updated (Toomula *et al.*, 2012). Biological databases usually use relational management frameworks and the Standard Query Language (SQL), which allows data definition as well as data manipulation statements (Kriegel and Schonauer, 2004).

According to the level of data curation, biological databases are divided into primary and secondary databases. Primary databases store experimental data from nucleotide sequence, protein sequence, or molecular structure. Secondary databases comprise data from the results of analyzing primary data and has become a reference library for just about any gene or gene product that has been investigated by the research community (Selzer *et al.*, 2008; Zou *et al.*, 2015).

The National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov/>) classifies the biological databases in Comprehensive and Specialized. Comprehensive databases store data from many organisms and many different types of sequences, for example, nucleotide, protein, and genomes. Specialized databases contain data from specific organisms (e.g., humans or mice), with functional or sequence information, and data generated by specific sequencing technologies.

Here we present an overview of human databases with web applications. The databases and tools allow to search biological sequences, genes and genomes, gene expression patterns, epigenetic variation, protein-protein interactions, variant frequency, regulatory elements, and comparative

Send correspondence Ursula Matte. Hospital de Clínicas de Porto Alegre, Laboratório de células, tecidos e genes, Rua Ramiro Barcelos, 2350, 90035-903, Porto Alegre, RS, Brazil. Email: umatte@hcpa.edu.br.

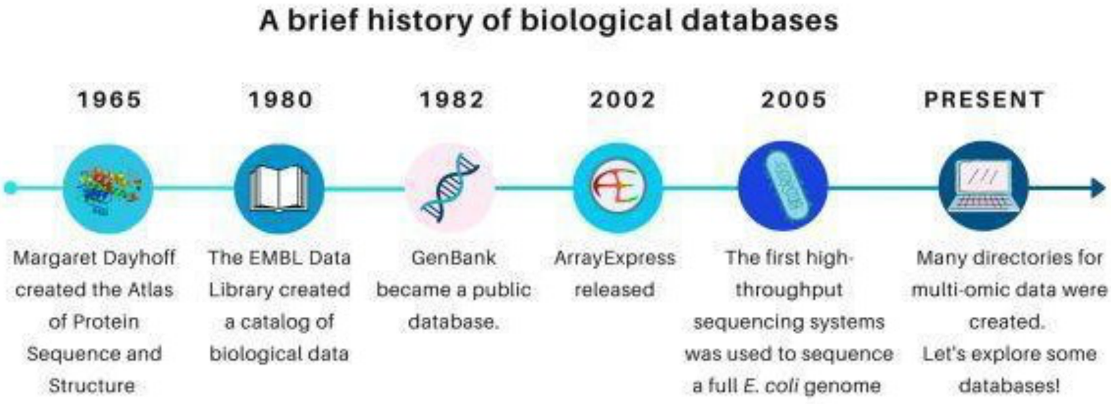

A brief history of biological databases
1965
1980
1982
2002
2005
PRESENT
Margaret Dayhoff created the Atlas of Protein Sequence and Structure
The EMBL Data Library created a catalog of biological data
GenBank became a public database.
ArrayExpress released
The first high-throughput sequencing systems was used to sequence a full *E. coli* genome
Many directories for multi-omic data were created.
Let's explore some databases!

Figure 1 – A brief history of the biological databases.

analysis between human and model organisms. Our goal is to provide an overview of available tools for exploring large datasets and analyzing the data for users with little or no programming skills.

Methods

We reviewed the databases from two platforms of multi-omic data: Biotools: Bioinformatics Tools and Services Discovery Portal (Ison *et al.*, 2013, <https://bio.tools/>) and OMICtools: an informative directory for multi-omic data analysis (Henry *et al.*, 2014, <https://omictools.com/>). We revised the tools from 15 March 2020 to 15 April 2020.

In the OMICtools portal, we used the keywords "human resources", in the "Databases" tab, sorted by "A to Z" and Taxonomy: "*Homo sapiens*". Subsequently, we discarded the results found that were of non-human species. In the Biotools portal, we used the keywords "human" and "Web application", the categories were found in the description of the filtering tools. We discarded the tools with the 'Temporarily unavailable' flag.

Results

The Newt Scamander and its magic creatures

The Omictools portal includes more than 4400 tools, classified by software, protocols, datasets, operation system, distribution when web user interface was selected. When filtered by the organisms represented in the tools, we obtained 1390 tools (Figure 2). When we chose only human results, we obtained 762 tools, divided in to Genomic Databases (91), Gene expression (81), miRNA (50), Protein-Protein Interaction (45), Rare/low frequency variation (28), Variant-disease association (28), Disease-specific variation (24), Proteome (23), LncRNA (17), miRNA target (17), Transcription Factor (17), DNA methylation (16), Metabolic network (15), Alternative Splicing (14), Cis-regulatory DNA element (12), Genetic association (12), Sequence databases (12), Comparative Genome (10), Enzyme databases (10), Gene regulatory network (10). Some tools were found in more than two of the classifications above. Besides, some of the tools were with the sites under maintenance or unavailable. For human databases, we revised 505 tools, detailed in Tables S1-S9.

In the Biotools portal, we found a total of 17,234 tools, divided into eight popular terms: Genetics, Proteins, Nucleic acids, Sequence analysis, Structure analysis, Omics, Virology and vaccine design, and Other. In order to facilitate the count of tools, we subdivided the classification in six terms presented in Figure 3. Filtered by human tools, we found 838 tools, and restricting to only "human Web Applications", we found 235 tools. When excluded the unavailable tools, we obtained 178 tools listed in Tables S1-S9.

There are only 23 tools present in the two portals: Clinvar, COSMIC, dbGAP, dbMAE, dbSNP, DisGeNET, DO, GENCODE, GeneCards, GIANT, GTEx, Gencode, GeneCards, GTEx, HumanMine, HuPho, InvFEST, LOVD, OMIM, Pickle, ProteomicsDB, Pseudomap, Varsome (Figure 4). In the next topic, we provide a brief description of the functionality and types of analysis that we can perform with the tools presented in this work. To facilitate, we grouped the tools into twelve general terms.

Databases and resources highlights

To illustrate the use of the databases reviewed here, we provided a Case Study with omic analysis using the *ACE2* gene, the receptor of coronavirus SARS-Cov-2. The Case Study is available on <https://kur1sutaru.github.io/fantastic_databases_and_where_to_find_them/>.

Alternative splicing

In summary, we found 24 tools for splice site analysis, to analyze the effect of alternative splicing on protein interaction and network through alteration of protein structure, and to predict a splicing consequence of an SNV at intron positions in the human genome (Table S1). Other databases search for constitutively and alternatively spliced introns and exons in humans and compare with other species. They also predict occurrences of alternative-splicing (AS) modes in the human genome, including exon skipping, 5'-alternative splicing,

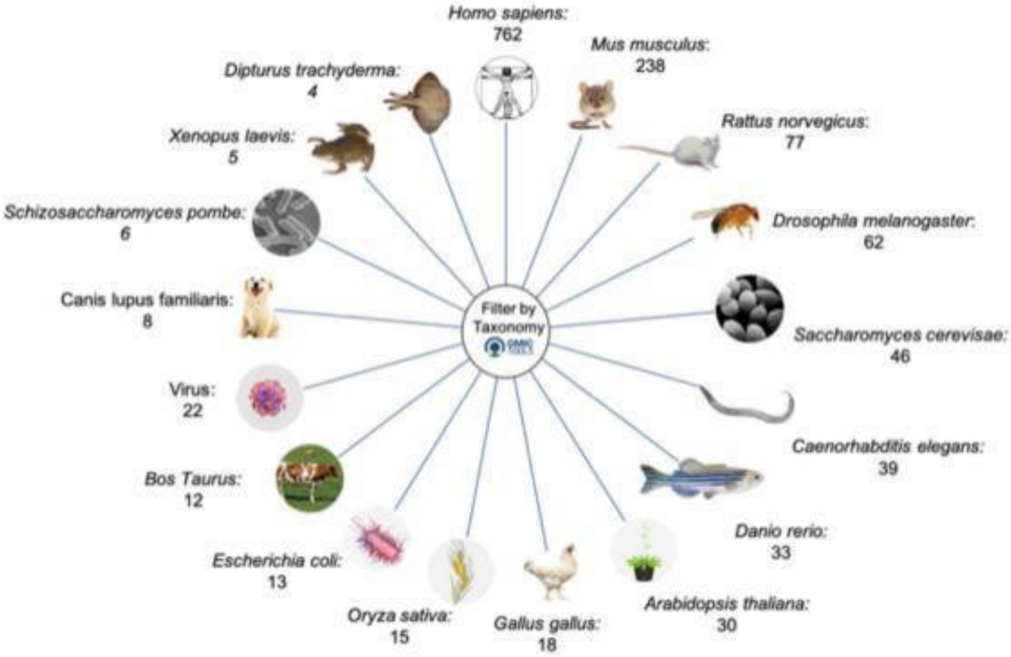


Figure 2 – Omictools databases and tools ordered by species. In this work, we used only the tools and databases for humans.

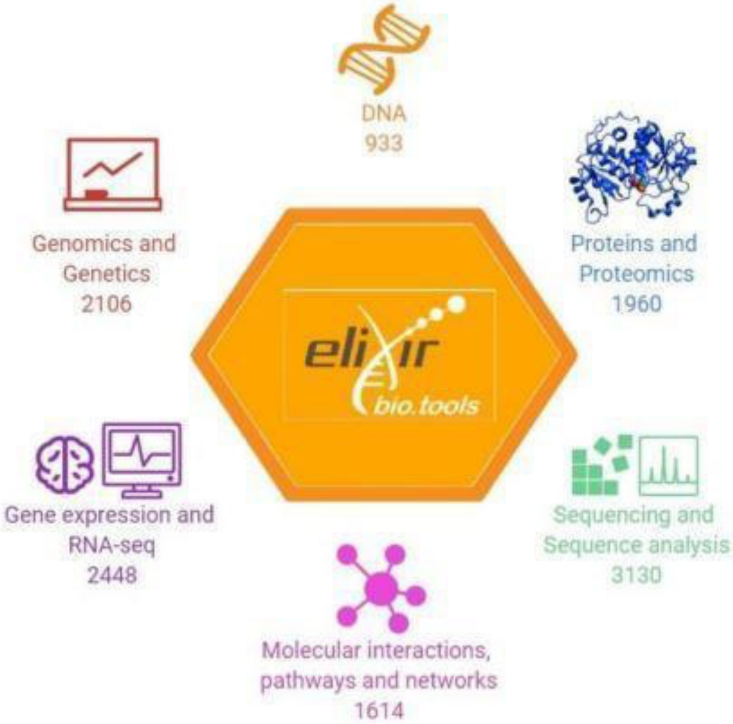

DNA 933
Proteins and Proteomics 1960
Sequencing and Sequence analysis 3130
Molecular interactions, pathways and networks 1614
Gene expression and RNA-seq 2448
Genomics and Genetics 2106
elixir bio.tools

Figure 3 – Bio.tools databases ordered by type.

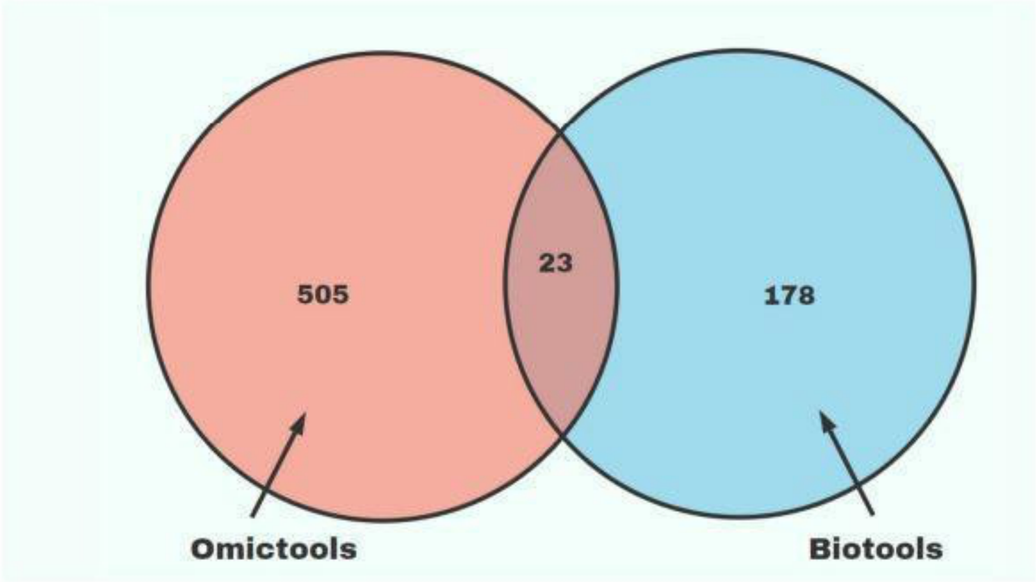


Figure 4 – Venn diagram of the resources used for the review of databases.

3'-alternative splicing and intron retention. Some databases provide a curated catalog of Alternative Splicing sites and gene transcripts, and to assign the function to the human protein-coding splice variants (PCSVs). Most of the results are obtained in molecular studies and based on Machine Learning predictions. Usually, the input is the Gene ID or Symbol, or a Fasta file containing the interest sequence. Some tools present the results in the browser with graphical reports and results are available for download.

Cancer databases

We found 56 tools (Table S2) and databases which provide genomic, transcriptomic, epigenomic profiles of a large group of tumor types with available tumor and normal tissue samples for comparisons, tools to predict the recognition of cancer epitopes by human T cell receptors (TCRs), literature mining of cancer-related genes in human, resources for exploring the impact of somatic mutations in human cancer, tools for inferring human and mouse gene expression patterns in various normal and cancerous tissues, and information about isoform-level expression analysis. Also, there are systematic analysis of mutations affecting cancer drug sensitivity based on individual genomic profiles from large-scale chemical screening using human cancer cell lines. Human cancer-specific microRNA- (miRNA) target interactions, protein-protein interactions (PPI), and functionally synergistic miRNA pairs are also available. aCGH data is analyzed to yield gene-specific copy numbers in different types of tumors.

Beyond this, many databases provide a compilation of cancer cell lines, driver genes, cancer metastasis, metastasis suppressor and histological characteristics, and complex queries built on the Boolean logic rules. Many tools are disease-specific, such as Human Gastric Cancer or Pediatric Cancer, for example. The user can provide Gene ID, miRNA, type of tumor, tissue, or cell line to search in the databases. The results allow us to explore, compare, and analyze all available cancer data (Clinical data, Gene Mutation, Gene Methylation, Gene Expression, Protein Phosphorylation, Copy Number Alteration, and so on). Outputs include barplots, heatmaps, principal component analysis plots, volcano plot, and survival analysis possible to view in the browser or to download.

Comparative databases

We found 31 tools and databases on this topic (Table S3). These tools give a comparative phenome-genome cross-species identification of genes associated with orthologous phenotypes, evolutionary analysis of genes, and comparative genomic analysis. Also comparisons between human and animal models' genomes are available to investigate links between disease genes, experimental data dealing with antibody and T-cell epitopes studied in various animal species encompassing humans and non-human primates. Data about interacting regions (IRs) in human and mouse proteins, miRNA target sites across species, non-human primate reference transcriptomes, allow to analyze and compare human nonsynonymous SNPs (nsSNP) in protein structures, protein complexes, protein-protein interfaces, and metabolic networks. Users can browse a genome, gene or region, phenotype term, disease name, or tissue for comparison. The output may present alignment across the genomes of different animal models and gene information (species, chromosome, gene name, accession number). Other output options are alternative transcripts, phylogeny plots, and gene regulatory networks.

Disease-specific and variant-disease association

We found 84 tools and specialized databases (Table S4) related to monogenic and complex diseases such as Alzheimer's, Autism, mendelian genetic diseases, and variant databases, such as repositories of human mutations, and single nucleotide polymorphisms (SNV). The input, in general, is the gene ID or symbol or name of the disease. The output, in general, is a list of annotated variants with the position number (genomic or chromosomal position), and the impact of the variation (missense, nonsense, frameshift, among others). Many databases provide germline and somatic variants of any size, type or genomic location, and the possible interpretations of the clinical significance of variants for reported conditions. The user may also search for genes based on user-specific disease/phenotype terms.

Gene expression

In this specific topic, we found 116 tools and databases with data about transcript counts, gene expression in normal and diseased tissues (Table S5). The databases comprehend data on age, apoptosis, gene expression in particular tissues or situations (such as allergy or immune response). Also, RNA-seq data from human and animal cells, single cell studies, and datasets for chromatin regulators and histone modifications are also available. The input can be a Gene ID, Pubmed ID, tissue cell or cell line, and phenotype information. The output is diverse across the databases, with tables, diagrams, heatmaps, barplot, and dotplot. Some databases present information about the gene structure, subcellular localization, primary function, cellular processes, and pathways. Many databases also inform the literature reference on the experimental data about the selected gene.

Genomic and sequence databases

We found 217 tools (Table S6), related to post-translational (PTM) modification sites and mutations (both germline and somatic) from multiple sources, mitochondrial sequences coming from ancient DNA samples (aDNA), disease-specific genomic data, genomic histone marks, chromatin states and motifs, chromosome annotation, evolutionary relationship of human proteins and animal models, analyses of allelic imbalance in clonal cell populations based on sequence polymorphisms. General information on genes, such as orthologs and paralogs, exon, intron and UTRs, gene classification, transcript sequences, protein sequences, mutations and SNPs, transcript cluster or selected publications can also be found. The general input is a Gene Symbol, Chromosomal position or region. The output is presented in browser visualization, tables with the genomic sequence of the interest, gene report, and heatmaps with pathogenicity of the mutations found in a sequence of the gene or region of interest.

LncRNA and miRNA databases

We found 73 tools (Table S7) related to miRNA and long non-conding RNAs (lncRNAs) in different organisms and pathogenic conditions. Of special interest may be data on putative antagomirs-miRNA heterodimers, protein-protein interactions (PPI) and functionally synergistic miRNA pairs (including transcription factors). The input of the search can be a Gene symbol, miRNA or LncRNA ID or target (hsa-let-7a, for example), Fasta file of the region of interest, or chromosomal location (any position in hg38 version, for example). In the case of LncRNA, the output is a table with the basic information of sequence, strand, class (antisense, for example), and conservation across species, with or without prediction scores. For miRNA, the output show information about the seed region, conservation score, accessibility, and secondary structures, minimum free energy estimation, correlations about expression, and start/end of the miRNA sequences. In disease-specific databases, the table may present the disease name, information about causality, and related literature.

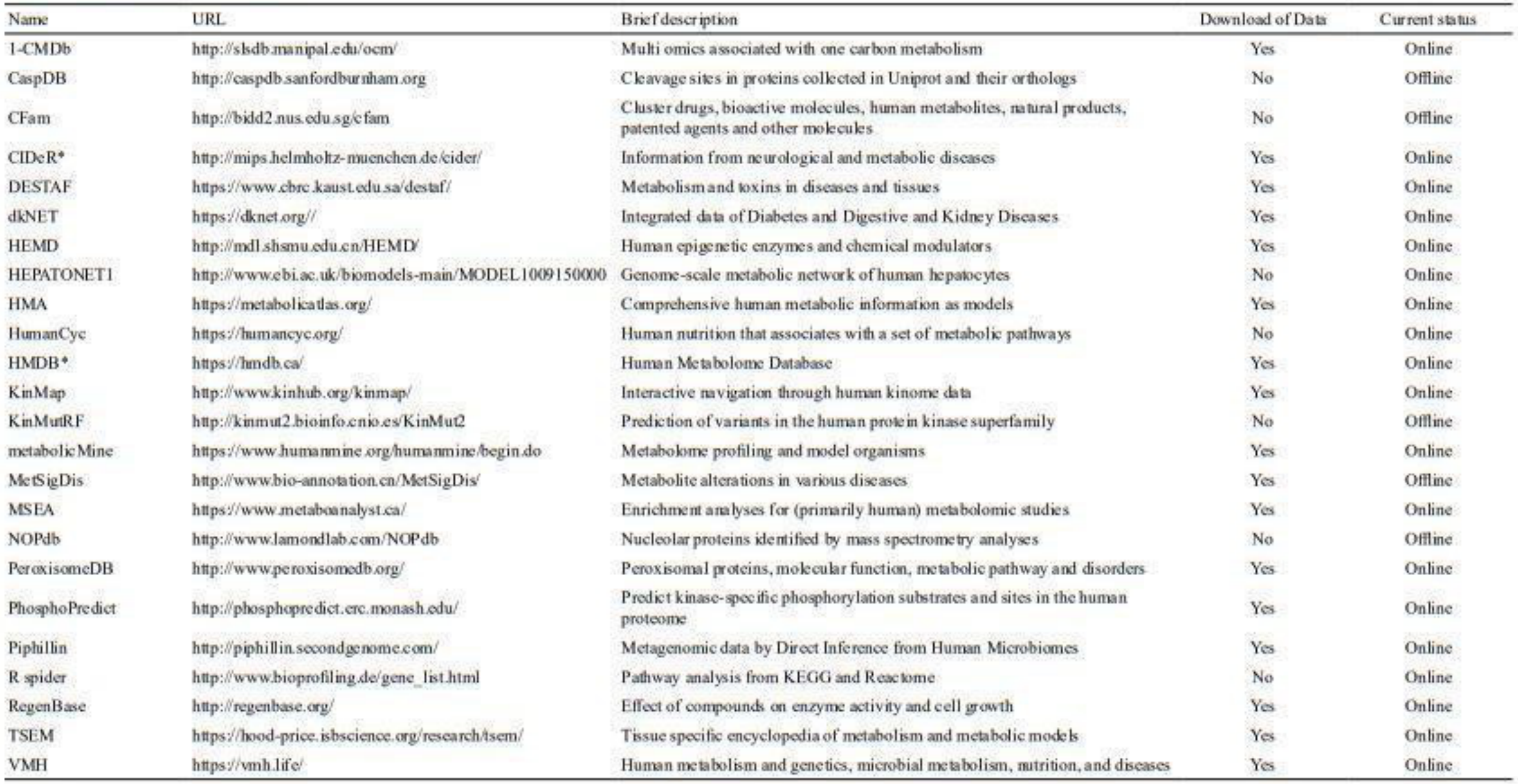
Metabolic and enzyme databases

We found 24 tools (Table 1) involving databases with information about potential cleavage sites in datasets of all human proteins collected in Uniprot and their orthologs, allowing for tracing of cleavage motif conservation, bioactive molecules, human metabolites, natural products, patented agents and other molecules. Databases are organized as disease-specific, tissue-specific, organelle-related or by protein families. Epigenetic enzymes and chemical modulators focused on epigenetic therapeutics, as well as functional databases including information from cells and tissues at a variety of physiological conditions, are also available. The input of the analyses is either a GO term, gene symbol or name, chromosomal location, metabolite or molecule name, or enzyme family name. Compartment and subsystem 2D maps, tables with literature related, enzyme domains, Uniprot ID, plots of classic or multivariate ROC curves, and images illustrating predicted small molecules are present in the output.

Methylation databases

We found 16 tools (Table 2) related to methylation patterns in aging-related diseases, normal or tumoral tissues, CpG islands and sites, and epigenome-wide studies. Information about the tissue-specific variation of methylation in the human central nervous system and matched blood samples collected from multiple donors are also available. Human cancer-specific DNA methylation, imprinting and gametogenesis-related methylation changes, and associations between RNAs and methylation databases are also available.

Table 1 – Metabolic and enzyme databases.



Name	URL	Brief description	Download of Data	Current status
1-CMDb	http://slsdb.manipal.edu/ocm/	Multi omics associated with one carbon metabolism	Yes	Online
CaspDB	http://caspdb.sanfordburnham.org	Cleavage sites in proteins collected in Uniprot and their orthologs	No	Offline
CFam	http://bidd2.nus.edu.sg/cfam	Cluster drugs, bioactive molecules, human metabolites, natural products, patented agents and other molecules	No	Offline
CIDeR*	http://mips.helmholtz-muenchen.de/cider/	Information from neurological and metabolic diseases	Yes	Online
DESTAF	https://www.cbrc.kaust.edu.sa/destaf/	Metabolism and toxins in diseases and tissues	Yes	Online
dkNET	https://dknet.org//	Integrated data of Diabetes and Digestive and Kidney Diseases	Yes	Online
HEMD	http://mdl.shsmu.edu.cn/HEMD/	Human epigenetic enzymes and chemical modulators	Yes	Online
HEPATONET1	http://www.ebi.ac.uk/biomodels-main/MODEL1009150000	Genome-scale metabolic network of human hepatocytes	No	Online
HMA	https://metabolicatlas.org/	Comprehensive human metabolic information as models	Yes	Online
HumanCyc	https://humancyc.org/	Human nutrition that associates with a set of metabolic pathways	No	Online
HMDB*	https://hmdb.ca/	Human Metabolome Database	Yes	Online
KinMap	http://www.kinhub.org/kinmap/	Interactive navigation through human kinome data	Yes	Online
KinMutRF	http://kinmut2.bioinfo.cnio.es/KinMut2	Prediction of variants in the human protein kinase superfamily	No	Offline
metabolicMine	https://www.humanmine.org/humanmine/begin.do	Metabolome profiling and model organisms	Yes	Online
MetSigDis	http://www.bio-annotation.cn/MetSigDis/	Metabolite alterations in various diseases	Yes	Offline
MSEA	https://www.metaboanalyst.ca/	Enrichment analyses for (primarily human) metabolomic studies	Yes	Online
NOPdb	http://www.lamondlab.com/NOPdb	Nucleolar proteins identified by mass spectrometry analyses	No	Offline
PeroxisomeDB	http://www.peroxisomedb.org/	Peroxisomal proteins, molecular function, metabolic pathway and disorders	Yes	Online
PhosphoPredict	http://phosphopredict.erc.monash.edu/	Predict kinase-specific phosphorylation substrates and sites in the human proteome	Yes	Online
Piphillin	http://piphillin.secondgenome.com/	Metagenomic data by Direct Inference from Human Microbiomes	Yes	Online
R spider	http://www.bioprofiling.de/gene_list.html	Pathway analysis from KEGG and Reactome	No	Online
RegenBase	http://regenbase.org/	Effect of compounds on enzyme activity and cell growth	Yes	Online
TSEM	https://hood-price.isbscience.org/research/tsem/	Tissue specific encyclopedia of metabolism and metabolic models	Yes	Online
VMH	https://vmh.life/	Human metabolism and genetics, microbial metabolism, nutrition, and diseases	Yes	Online

*Databases present in the case study.

Table 2 – Methylation databases.

Name	URL	Brief description	Download of Data	Current status
ANCOGeneDB	https://bioinfo.uth.edu/ancogenedb/	Epigenomic, enhancers, and expression quantitative trait loci	Yes	Online
BECon*	https://redgar598.shinyapps.io/BECon/	Interpreting methylation findings from blood in the context of brain	Yes	Online
CMS	http://cbbiweb.uthscsa.edu/KMethylomes/	Analytic functions for cancer methylome datasets	No	Offline
DBCAT	http://dbcat.cgm.ntu.edu.tw/	Methylation profiles of DNA alteration in human cancer	Yes	Online
DiseaseMeth*	http://bio-bigdata.hrbmu.edu.cn/diseasemeth/	Aberrant methylomes of human diseases	No	Online
GED	http://gametsepi.nwsuaflmz.com/	Epigenetic modification of gametogenesis in mammals	Yes	Online
Geneimprint	http://www.geneimprint.com/site/home	Gene imprinting and which allele is expressed	Yes	Online
Lnc2Meth	http://bio-bigdata.hrbmu.edu.cn/Lnc2Meth/	Informs about RNAs and DNA methylation of transcripts	Yes	Online
MeInfoText	http://bws.iis.sinica.edu.tw:8081/MeInfoText2/	Gene methylation and cancers, protein-protein interactions, and biological pathways	Yes	Offline
MethHC	http://methhc.mbc.nctu.edu.tw/	Focuses on aberrant methylomes of human diseases	No	Offline
MethylomeDB	http://epigenomics.columbia.edu/methylomedb/index.html	DNA methylation profiles for human and mouse brains	Yes	Offline
mPod	www.genome.org	Genome-wide tissue-specific DNA methylation profiles	No	Online
PhenoScanner	http://www.phenoscanner.medschl.cam.ac.uk/	Methylation and human genotype-phenotype associations	Yes	Online
ROADMAP	https://egg2.wustl.edu/roadmap/web_portal/index.html	Epigenetic modifications and mRNA expression of human cell types and tissues	Yes	Online
TCGA	https://www.cancer.gov/	Cancer methylation and expression	Yes	Online
TSGene	http://bioinfo.mc.vanderbilt.edu/TSGene/	Methylation status of tumor suppressor genes	No	Offline

*Databases present in the case study.

The input generally is an Entrez ID or Gene Symbol, SNP ID (e.g., rs1001098), Chromosome Location, CpG island ID or the DNA sequence in Fasta file. Also, a target tissue can be informed. The output shows tables with gene and methylated positions, CpG location, and in some cases, the literature related to the study. Some tools show the CpG islands in the genome browser, in a graph with the content of the CpG island region, CpG sites, and the DNA sequence.

Proteome and protein-protein interaction

In this topic, we found 97 databases (Table S8) that investigate the behavior of protein subunits in known complexes by comparing their abundance profiles across cell types, also tools for the identification of protein hydroxylation sites, to classify and score human coding variants based on the probability to damage their protein-related function. Databases of molecular-level putative protein-drug interactions, explorable and interactive human proteome database including MS/MS data, databases of protein interaction information pre-computed from existing structural and experimental data were also found. Fasta protein sequences, or Uniprot and Gene ID, are the input for these tools. Interactions between metabolites and molecules, small compounds, and enzyme's families and characterization are the output.

Regulatory elements

The 29 databases (Table 3) allow to visualize modified ribosomal nucleotides of human and several major model organisms, present resources to the identification of transcription factors, functional elements, cis-regulatory elements, interferon regulated genes, large intergenic non-coding RNAs (lincRNAs) and miRNA regulatory cascades in human diseases, Triplex Target DNA Site (TSS) with genomic regulatory sequences and signals, and RNA binding elements. The data is either from text-mining-assisted workflow, chromatin immunoprecipitation (ChIP), high-throughput datasets, Genome-Wide Association Studies (GWAS), next-generation sequencing techniques and/or predicted by computational models with annotations obtained by expert review of the scientific literature. Gene names, accession numbers, Fasta sequences, ligand ID (e.g., G4L0021), ligand name (e.g., TMPyP4), ligand activity or binding properties (e.g., Cytotoxicity), author name of ligand related literature, Ensembl ID List, tissue or cell type are examples of the possible inputs. Databases for Triplex target DNA sites provide specific search criteria, such as percent guanine content and pyrimidine interruption. The result shows a list of genes, tables, venn diagrams, scatter plot, position weight matrix for a selected motif, navigation across the motifs, and heatmaps with the target elements and biosamples.

Other specialized databases

We found 64 databases (Table S9) related to immunogenetics and antigen tumor receptors, genome-wide studies, rare diseases, cell culture, experimental design, genetic traits, and human copy number variations. The immunogenetic databases focus in cytokine receptors, their ligands, their involvement in diseases and their use in clinical treatments, human Major Histocompatibility Complex (MHC) genes, peptides, predictions, and proteins about human leukocyte antigen (HLA) I and HLA II-restricted peptides, *in silico* prediction of epitopes, tumor T cell antigens and literature about immunoproteins. Genome-wide databases compiled various public resources dealing with summary-level genome-wide association studies (GWAS) results. Rare disease databases comprehend specialized databases in rare diseases, such as the DECHIPER database, LungMap, and MARRVEL. Cell line databases integrated molecular authentication and identification tools for human and animal cell lines available from some of the main European cell banks, with curated literature. Experimental design databases contain experimental data and results from studies that have investigated the interaction of genotype and phenotype in humans, and CRISPR knockout libraries to target custom subsets of genes in the human or mouse genome. Genetic traits databases contain information about the global distribution of genetic traits. Copy number variations databases provide curated reference databases and bioinformatics resources targeting copy number profiling data in human diseases, especially in cancer. In general, the input is the Gene ID, DNA sequence in a Fasta format, organism, antigen name, haplotype, disease name, or ontology. For immunogenetic databases, the output is epitope sequences with scores, a list of antigens, assays, and receptors with the respective literature reference. Genome-wide databases present tables with the GWAS summarization of the study, SNP id, hits, allelic p-value, genotypic p-value, and Manhattan plot with the significant SNPs. Experimental design databases contain lists and tables with information about disease profiles and conditions with curated literature reviews. Genetic traits databases provide lists with SNP-trait associations in different human populations, and functional regions overlapping with SNPs in high linkage disequilibrium. Also, demographic information such as Country and Ethnicity is displayed within other GWAS information according to GWAS catalog guidelines.

Future and Perspectives

Biological databases are essential to provide information on normal and disease conditions, to search for information about DNA sequences, RNA, protein, and all possible data from different species and animal models. Public user-friendly web-based databases facilitate data mining and the search for information applicable to healthcare professionals. Besides, biological databases are essential to improve biomedical search sensitivity and efficiency and merge multiple datasets needed to share data and build global initiatives for the diagnosis, prognosis, and discovery of new treatments for genetic diseases. Gathering information about primary and secondary biological databases in a single review centralizes the search for recent and easy-to-use bioinformatics tools that can help to address some of the challenges in (Big) Data-driven research.

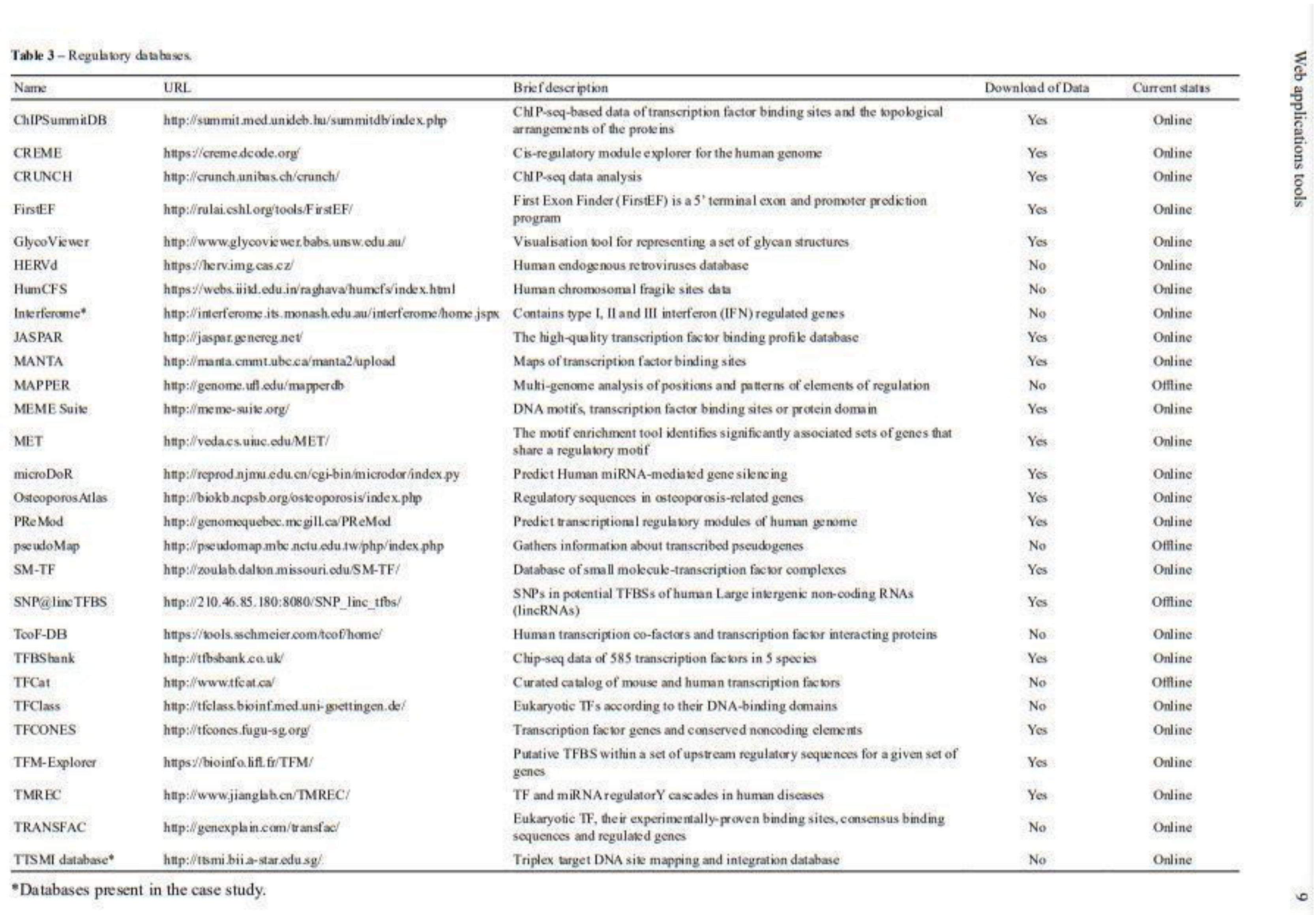
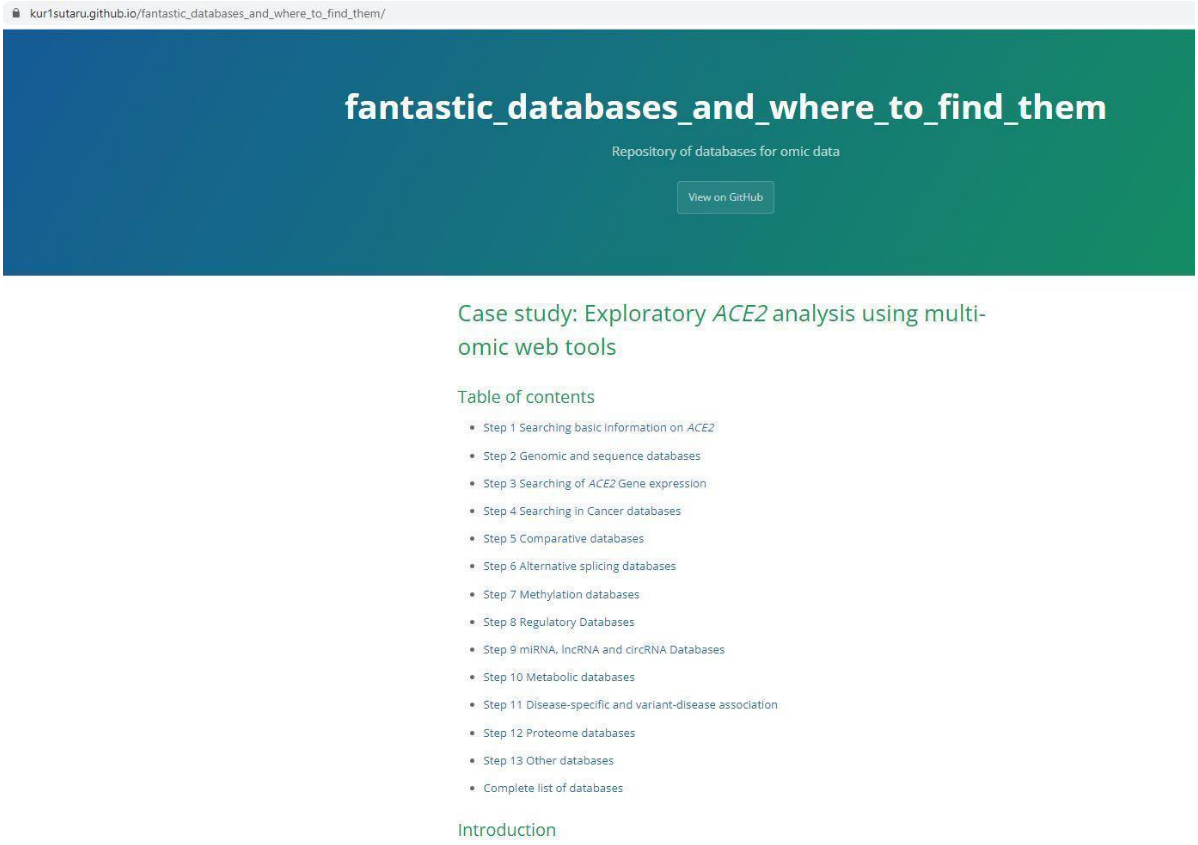


Table 3 – Regulatory databases.

Name	URL	Brief description	Download of Data	Current status
ChIPSummitDB	http://summit.med.unideb.hu/summitdb/index.php	ChIP-seq-based data of transcription factor binding sites and the topological arrangements of the proteins	Yes	Online
CREME	https://creme.dcode.org/	Cis-regulatory module explorer for the human genome	Yes	Online
CRUNCH	http://crunch.unibas.ch/crunch/	ChIP-seq data analysis	Yes	Online
FirstEF	http://rulai.cshl.org/tools/FirstEF/	First Exon Finder (FirstEF) is a 5' terminal exon and promoter prediction program	Yes	Online
GlycoViewer	http://www.glycoviewer.babs.unsw.edu.au/	Visualisation tool for representing a set of glycan structures	Yes	Online
HERVd	https://herv.img.cas.cz/	Human endogenous retroviruses database	No	Online
HumCFS	https://webs.iiitd.edu.in/raghava/humcfs/index.html	Human chromosomal fragile sites data	No	Online
Interferome*	http://interferome.its.monash.edu.au/interferome/home.jspx	Contains type I, II and III interferon (IFN) regulated genes	No	Online
JASPAR	http://jaspar.genereg.net/	The high-quality transcription factor binding profile database	Yes	Online
MANTA	http://manta.cmmt.ubc.ca/manta2/upload	Maps of transcription factor binding sites	Yes	Online
MAPPER	http://genome.ufl.edu/mapperdb	Multi-genome analysis of positions and patterns of elements of regulation	No	Offline
MEME Suite	http://meme-suite.org/	DNA motifs, transcription factor binding sites or protein domain	Yes	Online
MET	http://veda.cs.uiuc.edu/MET/	The motif enrichment tool identifies significantly associated sets of genes that share a regulatory motif	Yes	Online
microDoR	http://reprod.njmu.edu.cn/cgi-bin/microdor/index.py	Predict Human miRNA-mediated gene silencing	Yes	Online
OsteoporosAtlas	http://biokb.ncpsb.org/osteoporosis/index.php	Regulatory sequences in osteoporosis-related genes	Yes	Online
PReMod	http://genomequebec.mcgill.ca/PReMod	Predict transcriptional regulatory modules of human genome	Yes	Online
pseudoMap	http://pseudomap.mbc.nctu.edu.tw/php/index.php	Gathers information about transcribed pseudogenes	No	Offline
SM-TF	http://zoulab.dalton.missouri.edu/SM-TF/	Database of small molecule-transcription factor complexes	Yes	Online
SNP@lincTFBS	http://210.46.85.180:8080/SNP_linc_tfbs/	SNPs in potential TFBSs of human Large intergenic non-coding RNAs (lincRNAs)	Yes	Offline
TcoF-DB	https://tools.sschmeier.com/tcof/home/	Human transcription co-factors and transcription factor interacting proteins	No	Online
TFBSbank	http://tfbsbank.co.uk/	Chip-seq data of 585 transcription factors in 5 species	Yes	Online
TFCat	http://www.tfcat.ca/	Curated catalog of mouse and human transcription factors	No	Offline
TFClass	http://tfclass.bioinf.med.uni-goettingen.de/	Eukaryotic TFs according to their DNA-binding domains	No	Online
TFCONES	http://tfcones.fugu-sg.org/	Transcription factor genes and conserved noncoding elements	Yes	Online
TFM-Explorer	https://bioinfo.lifl.fr/TFM/	Putative TFBS within a set of upstream regulatory sequences for a given set of genes	Yes	Online
TMREC	http://www.jianglab.cn/TMREC/	TF and miRNA regulatorY cascades in human diseases	Yes	Online
TRANSFAC	http://genexplain.com/transfac/	Eukaryotic TF, their experimentally-proven binding sites, consensus binding sequences and regulated genes	No	Online
TTSMI database*	http://ttsmi.bii.a-star.edu.sg/	Triplex target DNA site mapping and integration database	No	Online

*Databases present in the case study.

Exemplo do site construído para a apresentação do estudo de caso:



kur1sutaru.github.io/fantastic_databases_and_where_to_find_them/

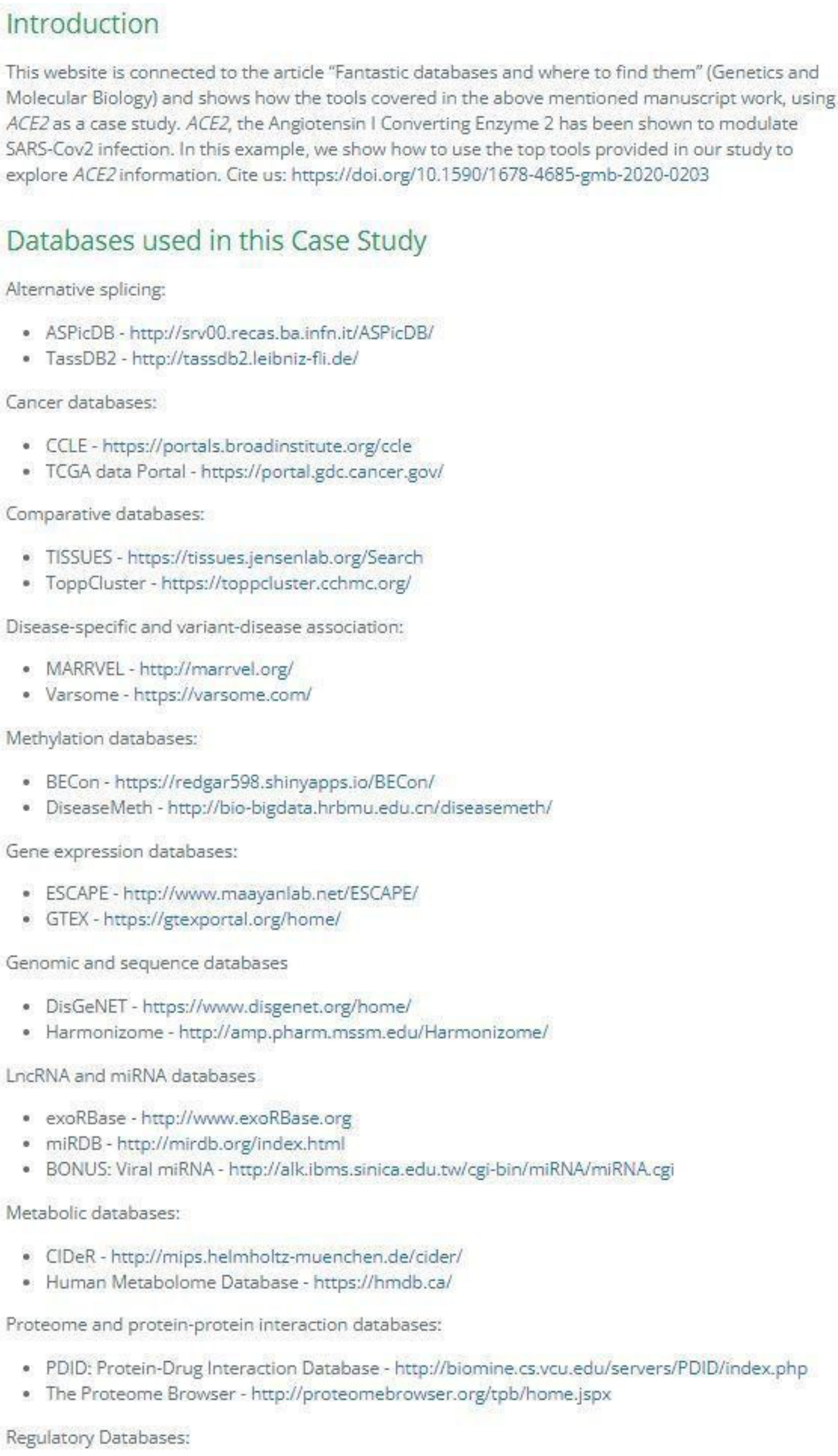
fantastic_databases_and_where_to_find_them

Repository of databases for omic data

Case study: Exploratory *ACE2* analysis using multi-omic web tools

Table of contents

Introduction



Introduction

This website is connected to the article "Fantastic databases and where to find them" (Genetics and Molecular Biology) and shows how the tools covered in the above mentioned manuscript work, using *ACE2* as a case study. *ACE2*, the Angiotensin I Converting Enzyme 2 has been shown to modulate SARS-Cov2 infection. In this example, we show how to use the top tools provided in our study to explore *ACE2* information. Cite us: https://doi.org/10.1590/1678-4685-gmb-2020-0203

Databases used in this Case Study

Alternative splicing:

- ASPicDB - http://srv00.recas.ba.infn.it/ASPicDB/
- TassDB2 - http://tassdb2.leibniz-fli.de/

Cancer databases:

- CCLE - https://portals.broadinstitute.org/ccle
- TCGA data Portal - https://portal.gdc.cancer.gov/

Comparative databases:

- TISSUES - https://tissues.jensenlab.org/Search
- ToppCluster - https://toppcluster.cchmc.org/

Disease-specific and variant-disease association:

- MARRVEL - http://marrvel.org/
- Varsome - https://varsome.com/

Methylation databases:

- BECon - https://redgar598.shinyapps.io/BECon/
- DiseaseMeth - http://bio-bigdata.hrbmu.edu.cn/diseasemeth/

Gene expression databases:

- ESCAPE - http://www.maayanlab.net/ESCAPE/
- GTEX - https://gtexportal.org/home/

Genomic and sequence databases

- DisGeNET - https://www.disgenet.org/home/
- Harmonizome - http://amp.pharm.mssm.edu/Harmonizome/

LncRNA and miRNA databases

- exoRBase - http://www.exoRBase.org
- miRDB - http://mirdb.org/index.html
- BONUS: Viral miRNA - http://alk.ibms.sinica.edu.tw/cgi-bin/miRNA/miRNA.cgi

Metabolic databases:

- CIDeR - http://mips.helmholtz-muenchen.de/cider/
- Human Metabolome Database - https://hmdb.ca/

Proteome and protein-protein interaction databases:

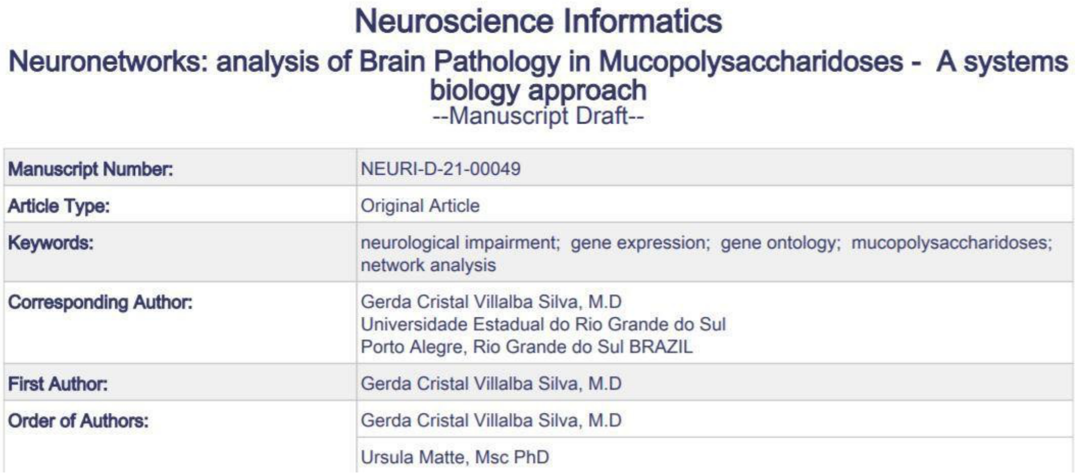
- PDID: Protein-Drug Interaction Database - http://biomine.cs.vcu.edu/servers/PDID/index.php
- The Proteome Browser - http://proteomebrowser.org/tpb/home.jspx

Regulatory Databases:

Capítulo 4

O uso de biologia de sistemas para entender o dano neurológico em diferentes tipos de Mucopolissacaridoses

(Manuscrito submetido para a *Neuroscience Informatics*)



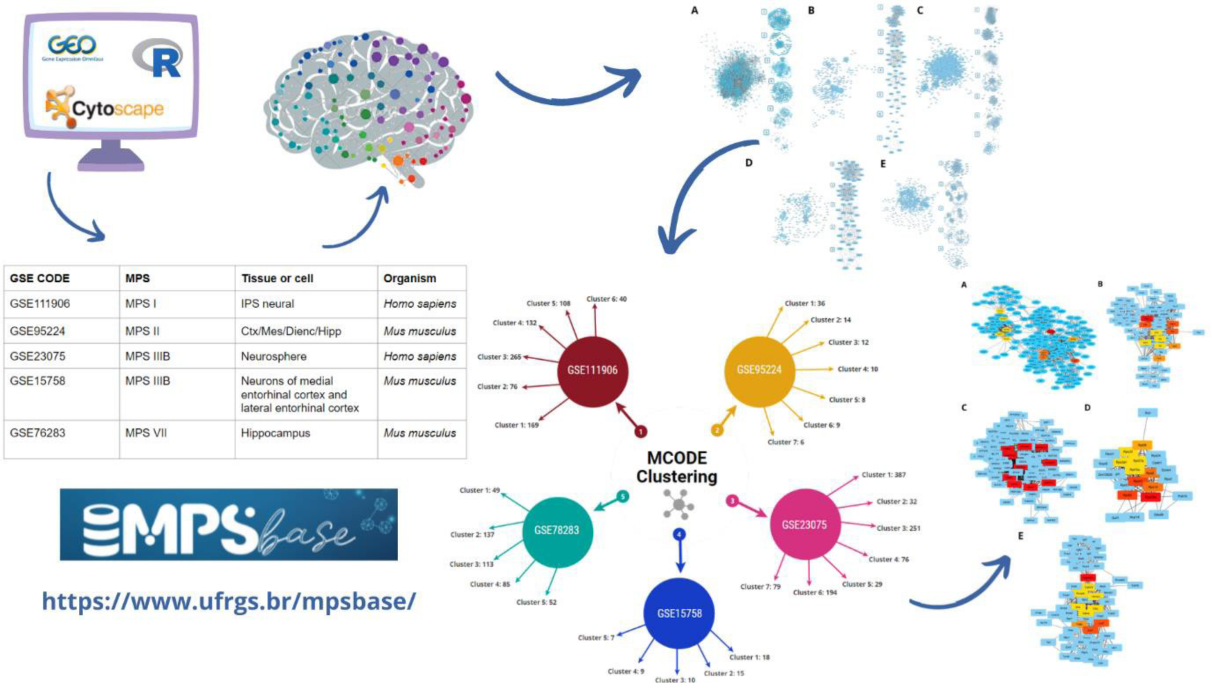
Neuroscience Informatics

Neuronetworks: analysis of Brain Pathology in Mucopolysaccharidoses - A systems biology approach

--Manuscript Draft--

Manuscript Number:	NEURI-D-21-00049
Article Type:	Original Article
Keywords:	neurological impairment; gene expression; gene ontology; mucopolysaccharidoses; network analysis
Corresponding Author:	Gerda Cristal Villalba Silva, M.D Universidade Estadual do Rio Grande do Sul Porto Alegre, Rio Grande do Sul BRAZIL
First Author:	Gerda Cristal Villalba Silva, M.D
Order of Authors:	Gerda Cristal Villalba Silva, M.D
	Ursula Matte, Msc PhD

Neuronetworks: analysis of Brain Pathology in Mucopolysaccharidoses - A systems biology approach



GSE CODE	MPS	Tissue or cell	Organism
GSE111906	MPS I	IPS neural	*Homo sapiens*
GSE95224	MPS II	Ctx/Mes/Dienc/Hipp	*Mus musculus*
GSE23075	MPS IIIB	Neurosphere	*Homo sapiens*
GSE15758	MPS IIIB	Neurons of medial entorhinal cortex and lateral entorhinal cortex	*Mus musculus*
GSE76283	MPS VII	Hippocampus	*Mus musculus*

https://www.ufrgs.br/mpsbase/

Highlights

- Mucopolysaccharidoses (MPS) presents a wide range of deranged signaling pathways
- Systems biology may help us understand the mechanisms of neurological impairment in MPS
- Processes related to Calcium signaling, GPCR and immune pathways are present in all the analyzed datasets
- Due to our results, we hypothesize that the primary accumulation of GAGs leads to perturbation of several biological processes and pathways, and are related to the neurological manifestations presents in the MPS disease

Abstract

Mucopolysaccharidoses (MPS) are rare lysosomal storage diseases characterized by defects in the activity of lysosomal hydrolases that degrade glycosaminoglycan, with progressive multisystemic involvement. Neurological damage is present in several MPS types. The relationship between the accumulation of glycosaminoglycans and the neurological disorder presented in these diseases remains unknown. For this purpose, we analyzed distinct types of MPS using publicly available transcriptomic data with a systems biology approach to search for clues about the pathophysiological mechanisms involved in the brain pathology of MPS. The most relevant proteins in the networks and ontology terms related to neurological damage in MPS were identified and compared among diseases. We performed the clustering analysis for GSE111906 (MPSI), GSE95224 (MPSII), GSE23075 (MPSIIIB), GSE15758 (MPSIIIB), and GSE76283 (MPSVII). Regarding biomarker discovery analysis, the top 10 genes were ranked according to the maximal clique centrality. Different ontologies were present in the different types of MPS. Ontologies were also present in all the MPS types, like axon guidance, Calcium signaling, PI3K-Akt signaling pathway, and Wnt signaling pathway. We hypothesize that these pathways are deranged because glycosaminoglycans play an essential role in the extracellular matrix composition, helping to regulate several processes. Systems biology approaches may help to understand the mechanisms of neuropathology in the different types of Mucopolysaccharidoses.

Keywords: neurological impairment, gene expression, gene ontology, mucopolysaccharidoses, network analysis.

1. Mucopolysaccharidoses with Neuronopathic Impairment

Mucopolysaccharidoses (MPS) are lysosomal storage diseases (LSD) characterized by defects in the activity of glycosaminoglycan degrading enzymes (Muenzer, 2011). Neurological damage is present in MPS I and II (severe cases), III (all subtypes), and VII (Viana et al., 2020). In all these MPS, the major brain-accumulated glycosaminoglycan is heparan sulfate (HS), although MPS I, II, and VII also accumulate dermatan sulfate. HS interacts with many molecules is involved with multiple ligands, receptor interactions and signaling (Dreyfuss et al., 2009; O'Callaghan et al., 2018). For this reason, HS storage elicits a wide range of pathogenic cascades, leading to disturbances in neurons and glial cells (Morimoto et al., 2021). Impaired HS degradation induces the neuropathological progression, and symptoms like optic atrophy, retinopathy, hearing impairment, seizures, cognitive and behavioral symptoms, neurodevelopmental delays, and sleep disturbances (Sato &

Okuyama, 2020). HS also interacts with many cytokines, such as FGF, TGF-β family, HGF, VEGF, hedgehog (Hh) and Wnt (Xie & Li, 2019), leading to neuroinflammatory processes (Killedar et al., 2010; Vitner et al., 2010; Baldo et al., 2015).

Networks facilitate the representation and modeling of biological data (Gaudelet, & Prˇzulj, 2019). A network view of neurological diseases facilitates accurate quantitative characterizations and biomarkers discovery of complex nervous system disorders with graph theory approaches (Medaglia & Bassett, 2017), as they help identifying genes (nodes) with functional relevance. For LSD, specifically MPS, this approach is gaining momentum, especially with metabolomic and proteomic studies (Salazar et al., 2016; Tebani et al., 2019; De Pasquale et al., 2020), as the mechanisms underlying neurological impairment in MPS are not completely understood (Kubaski et al., 2020). Here we analyzed different types of MPS using publicly available transcriptomic data to search for clues about the pathophysiological mechanisms involved in the brain pathology of MPS, beyond the primary storage. The selected studies for this work include MPSI Hurler (OMIM#607014); MPS II (OMIM#309900); MPS IIIB (OMIM#252920), and MPS VII (OMIM#253220), which were the ones with available data.

2. Methods

2.1 Download and Analysis of Transcriptome Data

The transcriptome datasets were retrieved from Gene Expression Omnibus (GEO – Edgar et al., 2002; Barrett et al., 2013) with the accession numbers GSE111906 (MPS type I, human neural iPS), GSE95224 (MPS II, mouse cerebral cortex and midbrain/diencephalon/hippocampus), GSE23075 (MPS IIIB, human neural stem cells cultivated in neutrosphere's), GSE15758 (MPS IIIB, mouse neurons of medial entorhinal cortex and lateral entorhinal cortex), and GSE76283 (MPS VII, mouse hippocampus).

We performed gene expression analysis in edgeR v.3.28.1(Robinson et al., 2010), in the case of RNA-seq datasets. We used limma package v.3.44.3 for the microarray datasets (Ritchie et al., 2015), with the appropriate functions according to the GeneChip requirements. More information about the datasets may be found in our database for differential expressed genes in Mucopolysaccharidoses, MPSBase <https://www.ufrgs.br/mpsbase/>. Transcriptome data was analyzed as differentially expressed genes (DEG), compared to normal tissue in each dataset, filtered by False Discovery Rate (FDR) adjust method, as summarized in table 1.

2.2 Network design

Gene network was primarily employed in metasearch engine STRING v.11.0 (Szklarczyk et al., 2019), using up to 2000 differentially expressed genes for each dataset. This cut-off value was determined as this was the maximum number of DEGs for all datasets when using a fold-change of 2 and FDR 0.05. We used confidence network edges with high confidence scores, aiming to obtain only experimental data. Text mining interactions were excluded from the analysis. Only the query proteins were considered, without first and second shell interactors. The analyses were performed in Cytoscape v.3.8, with curated plugins (Shannon et al., 2003).

2.3 Clustering and Centrality Metrics

For clustering analysis, the Molecular Complex Detection - MCODE v.1.6.1 was used to determine the densely connected regions in the networks (Bader & Hogue, 2003). The analysis was based on vertex weighting by the local neighborhood density and outward traversal from a locally dense seed protein to isolate the highly clustered regions (Bader & Hogue, 2003; Corrêa et al., 2009). We chose the following parameters: node score cutoff = 0.2; fluff = 0.5, and no haircut.

To identify candidate genes for biomarker discovery we used Cytohubba v.0.1 (Chin et al., 2014) with the local based method Maximal Clique Centrality (MCC), when $MCC (v) = \sum C \in S(v) (|C| - 1)!$ where $S(v)$ is the collection of maximal cliques that contain v, and $(|C| - 1)!$ is the product of all positive integers less than $|C|$.

To identify the most topologically relevant nodes in the network, we used Centiscape v.2.2 (Scardoni et al., 2014), and we chose five parameters of centrality: one for the network - (Diameter) and four for the nodes: Betweenness, Closeness, Degree, and Stress. The network diameter influences how proteins communicate and influence the function of each other.

The degree is the simplest topological index, corresponding to the number of nodes adjacent to a given node, where "adjacent" means directly connected. The degree allows evaluating the regulatory relevance of the node.

The closeness in a protein-protein interaction network can be interpreted as the "probability" of a protein to be functionally relevant for many other proteins. A protein with high closeness, when compared to the average closeness of the network, may be essential to the regulation of other proteins (Scardoni & Laudanna, 2012).

Stress is calculated by measuring the number of shortest paths passing through a node. It means the importance of a protein and its capability of holding together communicating nodes. The higher the value, the relevance of the protein increases in connecting regulatory pathways (Scardoni & Laudanna, 2012).

The betweenness is like stress but provides a more robust and informative centrality index. This is crucial to evaluate how it maintains the functionality and consistency of signaling mechanisms or a given biological function. Also, it may indicate the relevance of a protein as capable of holding together communicating hubs with a similar function (Scardoni & Laudanna, 2012; Scardoni et al., 2014). We combined these parameters to provide biologically meaningful node identification and functional classification.

2.4 Functional enrichment analysis

The functional enrichment was quantitatively assessed (p-value) using a hypergeometric distribution. Multiple test correction was also implemented by applying the FDR algorithm (Benjamini and Hochberg 1995) at a significance level of $p<0.05$. We used the Biological Network Gene Ontology (BiNGO) plugin v.3.0.4 (Maere et al., 2005) and CluePedia: A ClueGO plugin for pathway insights using integrated experimental and in silico data v.2.5.7 (Bindea et al., 2009; Bindea et al., 2013), with the terms or pathways consulted in the database of GO Immune System, KEGG, Reactome, and Wikipathways (Ashburner et al., 2000; Kanehisa et al., 2002; Kelder et al., 2012; Jassal et al., 2020). A summary of the methodology is presented in Figure 1.

3. Results

3.1 Clustering analysis and topological findings

We performed the clustering analysis for the five datasets: GSE111906 (MPS type I), GSE95224 (MPS II), GSE23075 (MPS IIIB), GSE15758 (MPS IIIB), and GSE76283 (MPS VII). The number of nodes varied from 2000 (GSE111906) to 1003 nodes (GSE95224). Differences between the number of genes in the input dataset and in the final output network are explained by the exclusion of nodes with no interactors. The GSE111906 network shows six clusters; whereas GSE95224 is divided into seven clusters as well as GSE23075, and GSE15758 and GSE78283 had five clusters each (Figure 2). We summarize the centiscape topological network indices and the five datasets' statistical results in Table 1.

According to centiscape centralities, the hub genes were ordered by degree, as shown in Table 2. Results indicate that the GSE111906 hub genes are markedly associated with integrin, and signaling pathways like ERK, G-protein coupled receptor (GPCR), and RET. The top 5 genes ordered by degree for GSE95224 are related to clathrin-mediated endocytosis, ERK, MAPK, TGF-Beta and signaling by GPCR. For the GSE23075 the top 5 genes are related to the immune system and MAPK cascade involved in innate immune response, and signaling pathways like ATM, FGFR, insulin, and TLR4. The GSE15758 top hub genes were related to rRNA processing in the nucleus and cytosol, signaling by ERK and the ribosome. The top 5 genes in GSE76283 are related to ATM pathway, adaptive and innate immune system, antigen processing and presentation, G-protein mediated regulation, and Calcium pathways.

Regarding the biomarker discovery analysis, the top 10 hub genes were ranked according to the maximal clique centrality (MCC, Table 3). For the GSE111906 (MPS I), integrin, clathrin-mediated endocytosis, and signaling pathways like ERK, GPCR, and RET were found. Other pathways are related to the immune system and cell cycle. The GSE95224 (MPS II) cytohubba top genes are related to ERK, cell adhesion molecules, neuropeptide hormone activity (PENK), neuroactive ligand-receptor interaction, and signaling by GPCR. For GSE23075 (MPS IIIB), top genes comprises the ontologies acetylation and gene expression, chromatin regulation, lysosome, mRNA splicing - major pathway, and immune processes such as activated TLR4 signaling and innate immune system. For GSE15758 (MPS IIIB), top 10 genes were related to the activation of the mRNA upon binding the cap-binding complex and eIFs, and subsequent binding to 43S, rRNA processing in the nucleus and cytosol, RNA binding, and viral mRNA translation. Finally, the GSE76283 (MPS VII) top 10 genes provided by cytohubba were present in the ontologies cytochrome P450, glycosaminoglycan degradation, innate immune system, NF-KappaB Pathway, signaling by GPCR, and transport to the Golgi and subsequent modification. Figure 3 shows the nodes with the best ranks, according to Maximal Clique Centrality and its first neighbors.

Processes related to Calcium signaling, GPCR and immune pathways are present in all datasets and were identified by both cytohubba and centiscape.

3.2 Gene-Set Enrichment Results

We identified the over-represented gene ontology (GO) pathways related to the nervous system and functions that were significantly enriched in the different gene sets. Figure 4 presents the top 10 enriched GO results, and the number of genes present in the three GO categories. The analysis reveals that some components, like Cytoplasm (GO Cellular Component) and Binding (GO Biological Process) appear in datasets from all MPS types. Others, like Cell (GO Cellular Component) and Membrane (GO Cellular Component) are absent only in the dataset from MPS I, whereas Cytoskeleton (GO Cellular Component) and Signaling (GO Molecular Function) are absent only in both MPS IIIB datasets.

Next, we analyzed the pathways present in Clinvar (only for the human datasets), GO Immune processes, KEGG, Reactome and Wikipathways. Figures 5 and 6 show the enrichment results for the datasets of neurological transcriptomic analysis across the different databases. Since most results were redundant, we will focus the discussion on the KEGG results, which showed the largest number of pathways. Fourteen KEGG terms (36.84%) were present in all the MPS types, such as Axon guidance, Calcium signaling pathway, Focal adhesion, FoxO signaling pathway, Hippo signaling pathway, MAPK signaling pathway, Metabolic pathways, Neuroactive ligand-receptor interaction, Pathways in cancer, PI3K-Akt signaling pathway, Proteoglycans in cancer, Rap1 signaling pathway, Ras signaling pathway, and Wnt signaling pathway. Also, 17 KEGG terms (44.74%) appears in three types of MPS, such as Apelin signaling pathway, Apoptosis, Autophagy, Breast cancer, cAMP signaling pathway, Choline metabolism in cancer, Colorectal cancer, ErbB signaling pathway, Gastric cancer, Hepatocellular carcinoma, Lysosome, Melanogenesis, mTOR signaling pathway, Oxytocin signaling pathway, Phosphatidylinositol signaling system, Relaxin signaling pathway, and Th17 cell differentiation. Finally, 7 terms (18.42%) appear in two types of MPS: AMPK signaling pathway, cGMP-PKG signaling pathway, Chemokine signaling pathway, Estrogen signaling pathway, Phosphonate and phosphinate metabolism, Prostate cancer, and Tight junction. We hypothesize that these pathways are deranged because glycosaminoglycans play an essential role in the extracellular matrix composition, helping to regulate several processes.

To find out which genes are shared between the datasets from MPS types, we combined all the networks and obtained 14 genes shared between three types of MPS (Table 4). Of these genes, 1 (7.14%) is shared by datasets of MPS I, MPS II and MPS VII; 3 (21.43%) by MPS I, MPS II and MPS IIIB; 3 (21.43%) by MPS II, MPS IIIB and MPS VII, and 7 (50%) genes appear in MPS I, MPS IIIB, and MPS VII. We hypothesize that these genes shared between the different types of MPS may have functional relevance to the neurological aspects of these diseases.

4. Discussion

The network-based analysis was implemented using public transcriptomic datasets. To the best of our knowledge this is the first study to carry out such an analysis of Mucopolysaccharidoses with neurological manifestations using systems biology approaches. Our networks include human neural IPS cells (MPS I), mouse hippocampus (MPS II), mouse neurons (MPS IIIB, GSE15758), human neural stem cells in neurospheres (MPS IIIB, GSE23075), and mouse hippocampus tissue (MPS VII). We opted to use only one tissue type in each dataset for analysis, to improve the prediction performance, as suggested by Guan et al. (2012).

We constructed various types of networks and topology analysis to identify critical molecular players and mechanisms involved in pathophysiology of neurological MPS types. As different diseases, with different datasets from various sources were analyzed, results will be discussed separately by the type of MPS.

4.1 MPS Type I

In the study of Swaroop and collaborators (2018) the transcriptomic analysis revealed several deranged pathways in the patient-derived iPS neural stem cells, like GAG biosynthesis and degradation pathways, lysosomal function pathways, and autophagy. Golgi transport, endoplasmic reticulum stress, autophagy and vacuolum organization are also impaired. Those authors suggest investigating the TGFβ signaling pathway. In our analysis, the hub genes also participate in several signaling pathways related to the lysosome and extracellular matrix.

The TGOLN2, the top-ranked hub gene, encodes an integral membrane protein related to the trans-Golgi network, which plays an essential role in vesicle formation and clathrin-mediated endocytosis. Other LSD also show disturbances in this pathway, as Pompe disease (Fukuda et al., 2006), Niemann-Pick A, Niemann-Pick C, Fabry, and Gaucher (Rappaport et al., 2016). The other 10 top genes are related to adaptive immune system, interleukin-11 signaling pathway, regulation of mitotic cell cycle, RET signaling, and signaling by GPCR. Castaneda and colleagues (2007) describe several LSD with early neuroimmune responses. HS accumulation in the brain may disturb immune regulators, many of which possess HS-binding motifs, as TLR4 and TLR2 (Parker & Bigger, 2019). Another deranged pathway in MPS I is the Oncostatin M signaling, that acts in neuronal proliferation and viability. This pathway also induces the mitogen activated protein kinases (MAPK) cascade (Houben et al., 2019). The different MAPKs involved are extracellular signal-regulated kinases 1 and 2 (ERK1/2), p38 and c-jun N-terminal kinases (JNK). All these signaling pathways are deranged in the neurological MPS I network.

Regarding the enrichment analysis, several signaling pathways are disturbed. Some of these are related to proteins which interact with HS, such as RET, PDGF and integrins, and can contribute to the diversity of clinical symptoms secondary to the accumulation of GAGs. Perturbations in autophagy, axon guidance, and vesicle trafficking have been described as leading to neurodegeneration in many MPS types (Fecarotta et al., 2020). Another impaired pathway is the Relaxin signaling pathway, that mediates anti-fibrotic, angiogenic, anti-inflammatory, and anti-apoptotic processes, having organ protective effects across a range of tissues, including the brain. It promotes the activation of downstream signal transduction pathways, such as cAMP signaling, GPCR, VEGF, PI3K/Akt, p38 MAPK, and Notch1 (Valkovic et al., 2018). The Notch1 pathway is related to synaptic plasticity (Ables et al., 2011) and may be involved in pathological modifications in stroke, Alzheimer's disease and CNS tumors (Lathia et al., 2008). Finally, neurotrophin signaling is related to hippocampal plasticity and neurodegeneration (Hennigan et al., 2007). Other functions of neurotrophins are correlated with survival, proliferation and maturation of affected neurons in Alzheimer's disease and Parkinson's disease (Sampaio et al., 2017).

4.2 MPS Type II

In the MPS II network, the top biological processes are signal transduction, transport, and synapse transmission, according to the study of Salvalaio et al. (2017). For cellular components, the ontologies are cell projection, membrane, cytoplasm, cytoskeleton, dendrite, and neuron projection. In addition, we found two new ontologies, synaptosome and ion channel complex. For molecular function ontologies, we found protein binding, voltage ion channel activity, hormone activity, kinase activity, and receptor binding. GPCR binding was found only in our top 10 ontologies enriched in the MPS II network.

The top enriched ontology present in the MPS II network is the Calcium signaling pathway, which is also impaired in Niemann-Pick C, Gaucher Type 1, Fabry, and Mucolipidosis type IV (Feng & Yang, 2016). Also, the FC gamma receptors are part of the immunoglobulin superfamily that contribute to the immune system processes. The FcR activation induces the p38 and NF Kappa Beta cascade in neurons, glial cells, endothelial cells and in infiltrating leukocytes. The endothelial cell dysfunction leads to the instability of the blood-brain barrier permeability (Okun et al., 2010).

Regarding the gene hub analysis, the top ranked gene, G Protein Subunit Gamma 7 plays a role in the regulation of adenylyl cyclase signaling in certain regions of the brain (Sadana & Dessauer, 2009). The related pathways of this gene are G-protein complex, and EPO-induced MAPK pathway, inhibiting the mTOR pathway to induce autophagy and cell death, and also inhibiting cell division by the deregulation of actin cytoskeleton (Liu et al., 2016). Other top ranked gene, the Proenkephalin

(PENK), a component of synaptic vesicles, released into the synapse, helps to modulate the perception of pain, and is involved in brain processes relevant to hippocampal functions (Cabrera-Reyes et al., 2019). In neurological diseases, the aberrant expression of PENK is related to dementia in Parkinson disease (Henderson-Smith et al., 2016), and with cognitive impairments in the murine model of Alzheimer's (Meilandt et al., 2008).

In MPS II, as in MPS VII, the neurological impairment may be severe, with rapidly progressing phenotypes or even mild to absent, presenting slowly progressing phenotypes (Bigger et al., 2018). Alzheimer disease shares several pathways deranged in MPS II, like neuroactive ligand receptors interaction pathway, calcium and insulin signaling. These pathways are molecular biomarkers for the cognitive decline in the hippocampus of Alzheimer (Gomez-Ravetti et al., 2010). We hypothesize that in MPS II these mechanisms also participate in cognitive function and could be biomarkers of neurological disease in MPS II.

4.3 MPS Type IIIB

Two MPS IIIB datasets (GSE23075 and GSE15758) were included in our analysis and hence are discussed together. Lemonnier et al. (2011) used patient-derived iPSC to model neuronal defects in MPS IIIB (GSE15758). The authors showed perturbations in pathways related to Golgi organization, cell migration, inflammation and neuritogenesis. In the MPS IIIB network, we found several processes related to immune response, such as activated TLR4 signaling, innate immune response, and complement activation. Several studies have described the role of immune system processes in the brain pathogenesis of MPS IIIB (DiRosario et al., 2009; Killedar et al., 2010; Parker & Bigger, 2019; Heon-Roberts et al., 2020).

Regarding the hub genes for the GSE23075 network, the top genes participate in Acetylation and Gene expression processes, Innate immune system pathways, Lysosomal pathways, and mRNA splicing – major pathway. GPKOW encodes a putative RNA-binding protein which interacts directly with protein kinase -A and -X and is also found associated with the spliceosome. Mutations in this gene cause congenital microcephaly and growth retardation (Carroll et al., 2017). In the MPS IIIB network, this gene is also associated with neuron differentiation, and may serve as a biomarker for neuron and brain damage. Other hub gene, CTSF, a cysteine lysosomal protease, was identified with an altered expression in Alzheimer and Parkinson patients and is supposed to play an important role in the autophagic and endo-lysosomal pathway and in the ubiquitin-proteasome system in neurodegenerative diseases (Sjödin et al., 2019).

The most enriched pathways of GSE25075 reveal perturbations in the AKT signaling pathway, Oxytocin and Ras signaling pathway. The PI3K/AKT/mTOR signaling pathways are responsible for the regulation of signal transduction, apoptosis and several cellular processes in neurodegenerative diseases (Xu et al., 2020). In MPS III, the accumulated HS in neurons and glial cells leads to neuronal apoptosis and microglia-mediated phagocytosis caused by oxidative stress and neuroinflammation (Baldini et al., 2020). The oxytocin signaling pathway and the oxytocin receptors (OTR) are GPCR which activate the MAPK/ERK1-2 and Ras pathways. In the brain, the oxytocin signaling mediates and controls social behaviors (Busnelli & Chini, 2018). One of the many symptoms of MPS IIIB patients are impaired social skills and aggressive behavior, probably associated with speech delay and hearing loss (Escolar et al., 2020). The perturbation of oxytocin signaling pathway and downstream pathways may contribute to explain the behavioral aspects of MPS IIIB.

The hub genes of GSE15758 belong to the same protein family, the Ribosomal Protein S and L, that encodes ribosomal 40S and 60S subunits. Altered ribosomal proteins lead to defects in ribosome biogenesis, resulting in ribosomal stress, and activation of the p53 signaling pathway, leading to p53-dependent cell cycle arrest and apoptosis (Wang et al., 2015). Ribosomal Proteins (RPs) play a role in several biological processes, like regulation of programmed cell death, modulation of DNA repair, modulation of cell migration, regulation of angiogenesis (Wang et al., 2015; Zhang et al., 2016). The activation of p53 pathway mediated by RP causes decreased body size, skeletal anomalies, and central nervous system malformations in the murine model. Besides, cardiovascular and metabolic diseases are also associated with perturbations in RPs (Wang et al., 2015).

Indeed, in the original work, MPS IIIB in mice is described as a tauopathy (Ohmi et al., 2009). Hyperphosphorylated tau was found in neurons of the medial entorhinal cortex, and in the dentate gyrus. The authors suggest that lysozymes induce the hyperphosphorylation of tau, but this mechanism is not well understood. In fact, HS plays an important role in protein aggregation in neurodegenerative diseases, like Alzheimer's and Parkinson (Maïza et al., 2018). Another important relationship is between the tau and ribosomes. Koren and colleagues propose that tau pathology impacts translation, thus disturbing synaptic plasticity, cellular metabolism, and memory formation. These tau-mediated impairments in translation could explain the role of tau in neurological diseases, including MPS IIIB.

Zhou et al. (2015) described the involvement of ribosomal proteins in various pathways and phenotypes. They suggest that ribosomal stress provokes accumulation of ribosome-free ribosomal proteins, which act in ribosome-independent functions like tumorigenesis, immune signaling, and development. Figure 7 summarizes the functions and biological pathways of ribosomal proteins in the literature, focusing on the genes present in all datasets analyzed in our study.

4.4 MPS Type VII

Parente and collaborators (2016) showed that the top differentially expressed genes participate in apoptosis, immune response, GPCR signaling pathway, and vesicle mediated transport. We also found these pathways in our analysis of the MPS VII network. Furthermore, the analysis of hub genes demonstrated Serpin Family A Member 3, and Lysozyme genes as hub genes of the MPS VII network. Upregulation of SERPINA3N, member of the serine protease inhibitor class, induces neuroinflammation and astrocyte activation, and causes hippocampal neuron loss, epilepsy-like seizures, and memory deficits in mice (Xi et al., 2019). Lysozymes show a protective role against amyloid-β in patients with Alzheimer disease (Helmfors et al., 2015).

GO analysis reveals pathways common to the MPS II network, like Insulin and Calcium signaling. These pathways are deranged in Alzheimer and are related with neurodegeneration (Gomez-Ravetti et al., 2010). The insulin growth factor regulates cell growth, mitochondrial processes, autophagy, oxidative stress, synaptic plasticity, and cognitive function (Hölscher, 2019). Lysosomes can store Calcium and participate in the Calcium signaling. Medina et al. (2015) demonstrated that lysosome controls autophagy via calcineurin-mediated induction of TFEB, a master transcriptional regulator of lysosomal biogenesis and autophagy.

Vesicle mediated transport in neurons is associated with the SNARE complex, and plays an important role in synaptic plasticity, in addition to mediating the exocytosis of neurotransmitter receptors (Verdú et al., 2018). Another pathway related to the MPS VII network is FC gamma receptor signaling pathway. These receptors act as a crosslink between the adaptive and innate immune system, generating signals to the lysosomes and proteasomes to initiate molecule degradation (Molfetta et al., 2014).

Regarding the GPCR signaling pathway, a ligand of these receptors is apelin, a neuropeptide responsible for diverse physiological and pathological processes. The apelin signaling is deranged in the MPS VII network. The Apelin-13 activates the PI3K/Akt and ERK1/2 signaling pathways to ameliorate brain lesions in mouse (Yang et al., 2014), and has a neuroprotective effect against apoptosis activating AMP-kinase pathway in an ischemia animal model (Yang et al., 2016). These pathways are also deranged in the MPS VII network and suggest clues about using Apelin to treat the neurological damage, not only in MPS VII, but in all the MPS with neurological symptoms.

4. Conclusion

Glycosaminoglycans play an essential role in the extracellular matrix composition, helping to regulate several processes, and interact with a wide range of receptors important to cell communication, recognition, and maintenance of cell integrity. We hypothesize that the primary accumulation of GAGs leads to perturbation of several biological processes and pathways, as demonstrated in this work.

Network analysis can be useful to discover the impaired pathways underlying the several processes deranged in the mucopolysaccharidoses. Besides, the detection of subnetworks in disease conditions can provide valuable insight into disease etiology or therapeutic responses (Zhang & Itan, 2019). Systems biology approaches may help us to understand the mechanisms of neuropathology in the several types of Mucopolysaccharidoses, and to discover novel biomarkers and treatments for the neurological symptoms of these diseases.

Funding

We would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the fellowships provided to GCVS (Process no.: 148615/2018-0) and UM (Process no.: 312714/2018-1). We also thank FIPE/HCPA (Project Number 2019-0761) for financial support.

Declaration of Competing Interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Data Availability Statement

The datasets used in this study were available in the Gene expression Omnibus <https://www.ncbi.nlm.nih.gov/geo/>. All the code used is available at <https://github.com/Kur1sutaru/system_biology_mps> .

References

Ables, J. L., Breunig, J. J., Eisch, A. J., Rakic, P, Not(ch) just development: Notch signalling in the adult brain. Nature reviews. Neuroscience, 12 (2011) 269–283.

Ashburner, M., Ball, C. A. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25 (2000) 25–29.

Bader, G. D., Hogue, C. W., An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics, 4 (2003) 2.

Baldini, G. et al., Sanfilippo Syndrome: The Tale of a Challenging Diagnosis. J. inborn errors metab. screen. [online]. 2020, vol.8 [cited 2020-11-19], e20200005. Epub Oct 05, 2020. ISSN 2326-4594.

Baldo, G., Lorenzini, D. M., et al., Shotgun proteomics reveals possible mechanisms for cognitive impairment in Mucopolysaccharidosis I mice. Mol Genet Metab 114 (2015) 138–145.

Barabási, A. L., Oltvai, Z., Network biology: understanding the cell's functional organization. Nature reviews. Genetics, 5 (2004) 101–113.

Barrett T, Wilhite SE, et al., NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 D991-5.

Bassett, D. S., Gazzaniga, M. S., Understanding complexity in the human brain. Trends in cognitive sciences, 15 2011 200–209.

Bigger, B. W., Begley, D. J., Virgintino, D., Pshezhetsky, A. V., Anatomical changes and pathophysiology of the brain in mucopolysaccharidosis disorders. Mol Genet Met 125(4) 2018 322–331.

Bindea, G., Mlecnik, B. et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics (Oxford, England), 25(8) 2009 1091–1093.

Bindea, G., Galon, J.,Mlecnik, B., CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics (Oxford, England), 29(5) 2013 661–663.

Busnelli, M., Chini, B.,. Molecular Basis of Oxytocin Receptor Signalling in the Brain: What We Know and What We Need to Know. Curr Top Behav Neurosci, 2018 3–29.

Cabrera-Reyes, E. A., Vanoye–Carlo, A. et.al., Transcriptomic analysis reveals new hippocampal gene networks induced by prolactin. Scientific Reports, 9 2019.

Carroll, R., Kumar, R., et al., Variant in the X-chromosome spliceosomal gene GPKOW causes male-lethal microcephaly with intrauterine growth restriction. Eur J Hum Genet 25, 2017 1078–1082.

Castaneda, J. A., Lim, M. J., Cooper, J. D., Pearce, D. A., Immune system irregularities in lysosomal storage disorders. Acta Neuropathologica, 115(2), 2007 159–174.

Chin, C. H., Chen, S. et al., cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC systems biology, 2014 S11.

Corrêa, T., Feltes, B. C., Riegel, M., Integrated analysis of the critical region 5p15.3-p15.2 associated with cri-du-chat syndrome. Gen Mol Biol 2019 186–196.

De Pasquale, V., Costanzo, M., et al., Proteomic Analysis of Mucopolysaccharidosis IIIB Mouse Brain. Biomolecules, 10(3) 2020 355.

DiRosario, J., Divers, E., et al., Innate and adaptive immune activation in the brain of MPS IIIB mouse model. J Neurosci Res, 2009 978–990.

Dreyfuss, J. L., Regatieri, C. V., et al., Heparan sulfate proteoglycans: structure, protein interactions and cell signaling. Anais da Academia Brasileira de Ciências, 2009 409-429.

Edgar R., Domrachev M., Lash A.E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository
Nucleic Acids Res. 2002 207-10.

Escolar, M., Bradshaw, J., et al., Development of a Clinical Algorithm for the Early Diagnosis of Mucopolysaccharidosis III. J. Inborn Errors Metab. Screen, 2020 8.

Fecarotta, S., Tarallo, A., Damiano, C., Minopoli, N., Parenti, G., Pathogenesis of Mucopolysaccharidoses, an Update. International journal of molecular sciences, 2020, 2515.

Feng, X., Yang, J., Lysosomal Calcium in Neurodegeneration. Messenger (Los Angeles, Calif. : Print) 2016, 56–66.

Fukuda, T., Ewan, L., et al., Dysfunction of endocytic and autophagic pathways in a lysosomal storage disease. Annals of neurology, 2006 700–708.

Fuller, M., Rozaklis, T., Ramsay, S. L., Hopwood, J. J., Meikle, P. J., Disease-specific markers for the mucopolysaccharidoses. Pediatr Res, 2004, 733–738.

Gaudelet, T. & Prˇzulj, N., Introduction to Graph and Network Theory. In Analyzing Network Data in Biology and Medicine. Cambridge University Press. 2019, 111-150.

Gómez Ravetti, M., Rosso, O. A., Berretta, R., Moscato, P., Uncovering Molecular Biomarkers That Correlate Cognitive Decline with the Changes of Hippocampus' Gene Expression Profiles in Alzheimer's Disease. PLoS ONE, 2010 e10153.

Guan, Y., Gorenshteyn, D.,et al., Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS computational biology, 2012 e1002694.

Helmfors, L., Boman, A., et al., Protective properties of lysozyme on β-amyloid pathology: implications for Alzheimer disease. Neurobiology of Disease, 2015, 122–133.

Henderson-Smith, A., Corneveaux, J. J., et al., Next-generation profiling to identify the molecular etiology of Parkinson dementia. Neurology Genetics, 2016, e75.

Hennigan, A., O'Callaghan, R. M., & Kelly, Á. M., Neurotrophins and their receptors: roles in plasticity, neurodegeneration and neuroprotection. Biochemical Society Transactions, 2007, 424–427.

Heon-Roberts, R., Nguyen, A., Pshezhetsky, A. V., Molecular Bases of Neurodegeneration and Cognitive Decline, the Major Burden of Sanfilippo Disease. Journal of clinical medicine, 2020, 344.

Hölscher C., Insulin signalling impairment in the brain as a risk factor in Alzheimer’s Disease. Front. Aging Neurosci., 2019.

Houben, E., Hellings, N., Broux, B., Oncostatin M, an Underestimated Player in the Central Nervous System. Frontiers in immunology, 2019, 1165.

Jassal, B., Matthews, L., et al., The Reactome pathway knowledgebase. Nucleic acids research, 2020 D498–D503.

Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., The KEGG databases at GenomeNet. Nucleic acids research, 2002, 42–46.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., Pico, A. R., WikiPathways: building research communities on biological pathways. Nucleic acids research, 2012, D1301–D1307.

Killedar, S., Dirosario, J., Divers, E., Popovich, P. G., McCarty, D. M., Fu, H., Mucopolysaccharidosis IIIB, a lysosomal storage disease, triggers a pathogenic CNS autoimmune response. Journal of neuroinflammation, 2010, 39.

Koren, S. A., Hamm, M. J. et al., Tau drives translational selectivity by interacting with ribosomal proteins. Acta neuropathologica, 2019, 571–583.

Kubaski, F., de Oliveira Poswar, F., Michelin-Tirelli, K., Matte, U., Horovitz, D. D., Barth, A. L., Baldo, G., Vairo, F., Giugliani, R., Mucopolysaccharidosis Type I. Diagnostics (Basel, Switzerland), 2020, 161.

Lathia, J. D., Mattson, M. P., Cheng, A., Notch: from neural development to neurological disorders. Journal of neurochemistry, 2008, 1471–1481.

Lemonnier, T., Blanchard, S., Toli, D., Roy, E., Bigou, S., Froissart, R., Rouvet, I., Vitry, S., Heard, J. M., Bohl, D., Modeling neuronal defects associated with a lysosomal disorder using patient-derived induced pluripotent stem cells. Human molecular genetics, 2011, 3653–3666.

Liu, J., Ji, X., Li, Z., Yang, X., Wang, W., Zhang, X., G protein γ subunit 7 induces autophagy and inhibits cell division. Oncotarget, 2016, 24832–24847.

Lv, S.-Y., Chen, W.-D., Wang, Y.-D., The Apelin/APJ System in Psychosis and Neuropathy. Frontiers in Pharmacology, 11, 2020.

Medaglia, J. D., Bassett, D. S., Network analyses and nervous system disorders. arXiv preprint 2017.

Medina, D. L., Di Paola, S., et al., Lysosomal calcium signalling regulates autophagy through calcineurin and TFEB. Nature cell biology, 2015, 288–299.

Maere, S., Heymans, K., Kuiper, M., BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics (Oxford, England), 2005, 3448–3449.

Maïza, A., Chantepie, S., Vera, C., Fifre, A., Huynh, M. B., Stettler, O., Ouidja, M. O., Papy-Garcia, D., The role of heparan sulfates in protein aggregation and their potential impact on neurodegeneration. FEBS Letters, 2018, 3806–3818.

Meilandt, W. J., Yu, G.-Q., Chin, J., Roberson, E. D., Palop, J. J., Wu, T., Scearce-Levie, K., Mucke, L., Enkephalin Elevations Contribute to Neuronal and Behavioral Impairments in a Transgenic Mouse Model of Alzheimer’s Disease. Journal of Neuroscience, 2008, 5007–5017.

Molfetta, R., Quatrini, L., Gasparrini, F., Zitti, B., Santoni, A., Paolini, R., Regulation of fc receptor endocytic trafficking by ubiquitination. Frontiers in immunology, 2014, 449.

Morimoto, H., Kida, S., Yoden, E., Kinoshita, M., Tanaka, N., Yamamoto, R., Koshimura, Y., Takagi, H., Takahashi, K., Hirato, T., Minami, K., & Sonoda, H. (2021). Clearance of heparan sulfate in the brain prevents neurodegeneration and neurocognitive impairment in MPS II mice. Molecular therapy : the journal of the American Society of Gene Therapy, 29(5), 1853–1861. https://doi.org/10.1016/j.ymthe.2021.01.027.

Muenzer J., Overview of the mucopolysaccharidoses. Rheumatology (Oxford, England), 2011, v4–v12.

Myerowitz, R., Lawson, D., Mizukami, H., Mi, Y., Tifft, C. J., Proia, R. L., Molecular pathophysiology in Tay-Sachs and Sandhoff diseases as revealed by gene expression profiling. Human molecular genetics, 2002, 1343–1350.

O'Callaghan, P., Zhang, X., Li, J. P, .Heparan Sulfate Proteoglycans as Relays of Neuroinflammation. The journal of histochemistry and cytochemistry: official journal of the Histochemistry Society, 2018, 305–319.

Okun, E., Mattson, M. P., Arumugam, T. V., Involvement of Fc receptors in disorders of the central nervous system. Neuromolecular medicine, 2010, 164–178.

Ohmi, K., Kudo, L. C., Ryazantsev, S., Zhao, H.-Z., Karsten, S. L., Neufeld, E. F., Sanfilippo syndrome type B, a lysosomal storage disease, is also a tauopathy. PNAS, 2009, 8332–8337.

Parente, M. K., Rozen, R., Seeholzer, S. H., Wolfe, J. H., Integrated analysis of proteome and transcriptome changes in the mucopolysaccharidosis type VII mouse hippocampus. Mol Genet Metab, 2016, 41–54.

Parikshak, N. N., Gandal, M. J., Geschwind, D. H., Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nature reviews. Genetics, 2015, 441–458.

Parker, H., Bigger, B. W., The role of innate immunity in mucopolysaccharide diseases. Journal of neurochemistry, 2019, 639–651.

Rappaport, J., Manthe, R. L., Solomon, M., Garnacho, C., Muro, S., A Comparative Study on the Alterations of Endocytic Pathways in Multiple Lysosomal Storage Disorders. Molecular pharmaceutics, 2016, 357–368.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G. K.,limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015, e47.

Robinson, M. D., McCarthy, D. J., Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 2010, 139–140.

Sadana, R., Dessauer, C. W., Physiological roles for G protein-regulated adenylyl cyclase isoforms: insights from knockout and overexpression studies. Neuro-Signals, 2009, 5–22.

Salvalaio, M., D'Avanzo, F., Rigon, L., Zanetti, A., D'Angelo, M., Valle, G., Scarpa, M., Tomanin, R., Brain RNA-Seq Profiling of the Mucopolysaccharidosis Type II Mouse Model. International journal of molecular sciences, 2017, 1072.

Salazar, D. A., Rodríguez-López, A., Herreño, A., Barbosa, H., Herrera, J., Ardila, A., Barreto, G. E., González, J., Alméciga-Díaz, C. J., Systems biology study of mucopolysaccharidosis using a human metabolic reconstruction network. Mol Genet Metab, 2016, 129–139.

Sampaio, T.B., Saval, A.S., Gutierrez, M.E., Pinton, S., Neurotrophic factors in Alzheimer's and Parkinson's diseases: implications for pathogenesis and therapy. Neural Regen Res 2017;12:549-57.

Sato, Y., & Okuyama, T. (2020). Novel Enzyme Replacement Therapies for Neuropathic Mucopolysaccharidoses. International journal of molecular sciences, 21(2), 400. https://doi.org/10.3390/ijms21020400.

Scardoni, G., Laudanna, C., Centralities Based Analysis of Complex Networks. In: New Frontiers in Graph Theory. Submitted: May 16th 2011. Reviewed: November 7th 2011.

Scardoni, G., Tosadori, G., Faizan, M., Spoto, F., Fabbri, F., Laudanna, C., Biological network analysis with CentiScaPe: centralities and experimental dataset integration. 2014.

Shannon, P., Markiel, A., et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 2003, 2498–2504.

Sjödin, S., Brinkmalm, G., et al., Endo-lysosomal proteins and ubiquitin CSF concentrations in Alzheimer's and Parkinson's disease. Alzheimer's research & therapy, 11(1), 2019, 82.

Stam C. J., Modern network science of neurological disorders. Nature reviews. Neuroscience, 2014, 683–695.

Swaroop, M., Brooks, M. J., Gieser, L., Swaroop, A., Zheng, W., Patient iPSC-derived neural stem cells exhibit phenotypes in concordance with the clinical severity of mucopolysaccharidosis I. Hum Mol Genet, 2018, 3612-3626.

Szklarczyk, D., Gable, A. L., et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research, 2019, D607–D613.

Tebani, A., Abily-Donval, L. et al., Analysis of Mucopolysaccharidosis Type VI through Integrative Functional Metabolomics. International journal of molecular sciences, 2019, 446.

Valkovic, A. L., Bathgate, R. A., Samuel, C. S., Kocan, M.,Understanding relaxin signalling at the cellular level. Mol Cell Endocr, 2018, 24-33.

Verdú, M. P. M., Portalés, A., SanJuan, M. P., Jurado, S., Postsynaptic SNARE proteins: Role in synaptic transmission and plasticity. Neurosci. , 2018, 12-21.

Viana, G. M., Priestman, D. A., Platt, F. M., Khan, S., Tomatsu, S., & Pshezhetsky, A. V. (2020). Brain Pathology in Mucopolysaccharidoses (MPS) Patients with Neurological Forms. Journal of clinical medicine, 9(2), 396. https://doi.org/10.3390/jcm9020396.

Vitner, E. B., Platt, F. M., Futerman, A. H.,Common and uncommon pathogenic cascades in lysosomal storage diseases. The Journal of biological chemistry, 2010, 20423–20427.

Xi, Y., Liu, M.,et al.,Inhibition of SERPINA3N-dependent neuroinflammation is essential for melatonin to ameliorate trimethyltin chloride–induced neurotoxicity. Journal of Pineal Research, 2019, 67(3).

Xie, M., Li, J. P., Heparan sulfate proteoglycan - A common receptor for diverse cytokines. Cellular signaling, 2019, 115–121.

Xu, F., Na, L., Li, Y., Chen, L., Roles of the PI3K/AKT/mTOR signalling pathways in neurodegenerative diseases and tumours. Cell & bioscience,2020, 54.

Wang, W., Nag, S., Zhang, X., Wang, M. H., Wang, H., Zhou, J., Zhang, R., Ribosomal proteins and human diseases: pathogenesis, molecular mechanisms, and therapeutic implications. Medicinal research reviews, 2015, 225–285.

Yang, Y., Zhang, X., Cui, H., Zhang, C., Zhu, C., Li, L., Apelin-13 protects the brain against ischemia/reperfusion injury through activating PI3K/Akt and ERK1/2 signaling pathways. Neurosci. Lett. 2014, 44–49.

Yang, Y., Zhang, X.-J., et al., Apelin-13 protects against apoptosis by activating AMP-activated protein kinase pathway in ischemia stroke. Peptides 2016, 96–100.

Zhang, C., Fu, J., et al., Knockdown of ribosomal protein S15A induces human glioblastoma cell apoptosis. World journal of surgical oncology, 2016, 129.

Zhang, P., Itan, Y., Biological Network Approaches and Applications in Rare Disease Studies. Genes,2019, 797.

Zheng, M., Ambesi, A., J. McKeown-Longo, P., Role of TLR4 Receptor Complex in the Regulation of the Innate Immune Response by Fibronectin. Cells, 9(1),2020, 216.

Zhou, X., Liao, W. J., Liao, J. M., Liao, P., & Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. Journal of molecular cell biology, 7(2), 92–104. https://doi.org/10.1093/jmcb/mjv014.

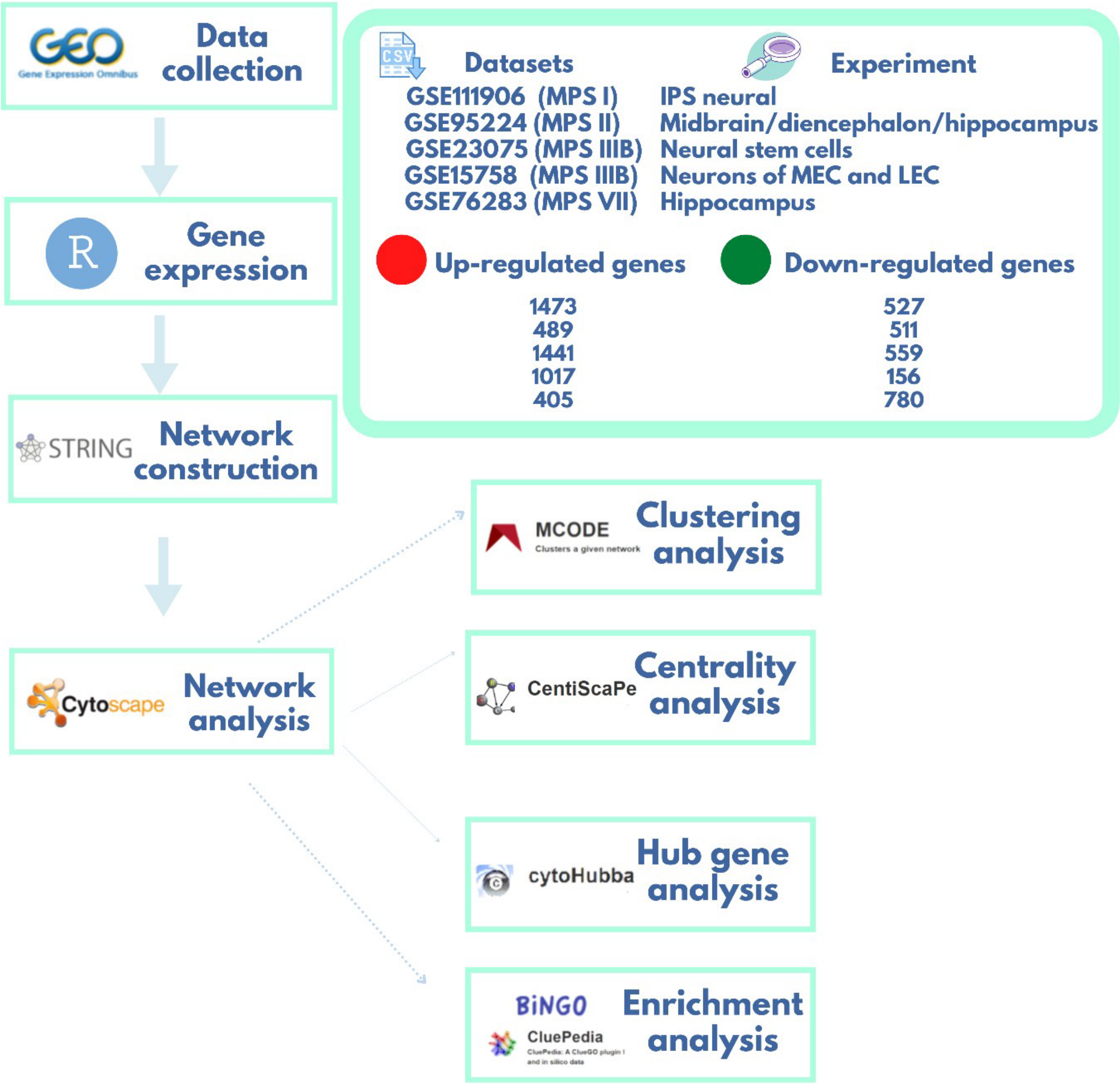


Figure 1: Workflow of the analysis.

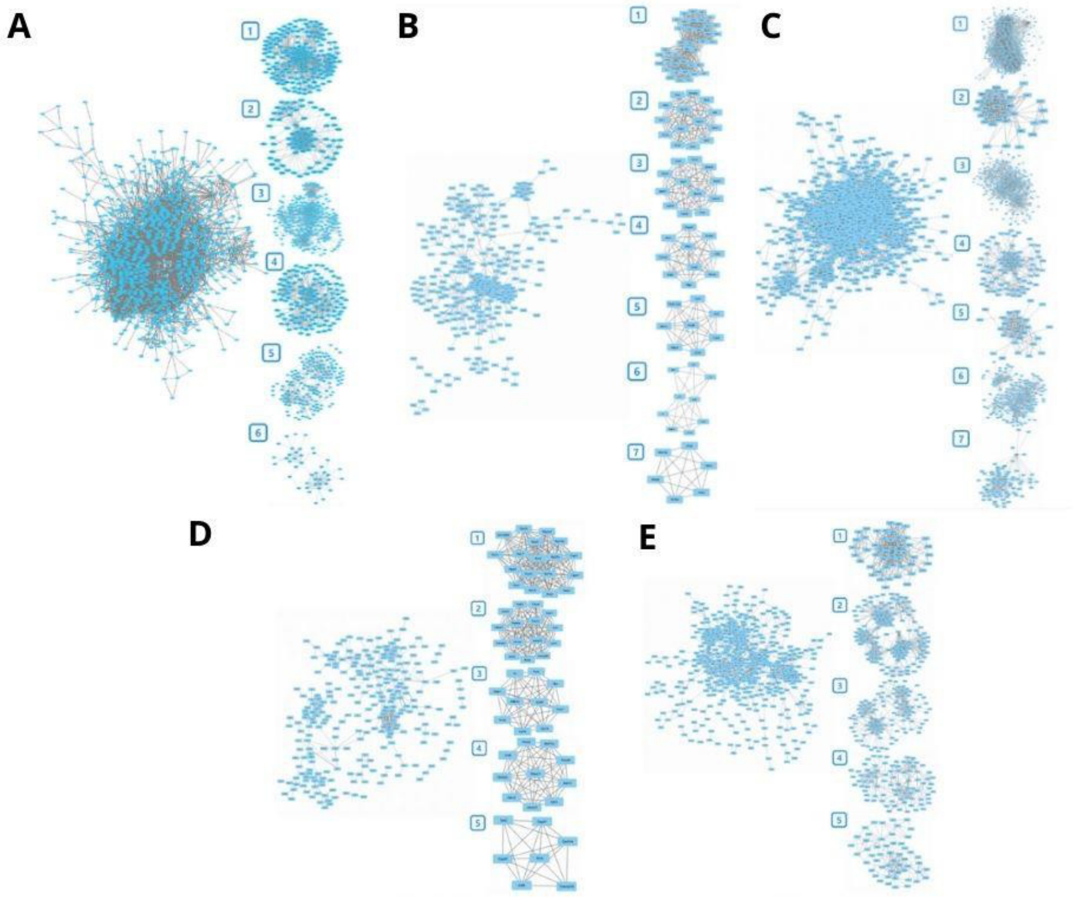


Figure 2: MCODE clustering results. A = GSE111906, MPS I; B = GSE95224, MPS II; C = GSE23075, MPS IIIB, D = GSE15758, MPS IIIB, E = GSE76283, MPS VII.

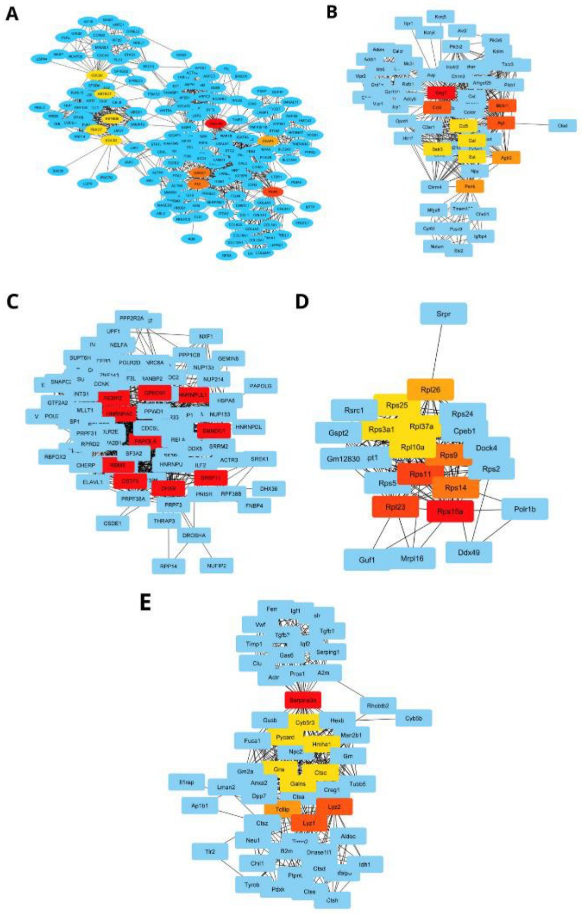


Figure 3: Analysis of gene hubs with Cytohubba, using the MCC algorithm. A = GSE111906, MPS I; B = GSE95224, MPS II; C = GSE23075, MPS IIIB, D = GSE15758, MPS IIIB, E = GSE76283, MPS VII.

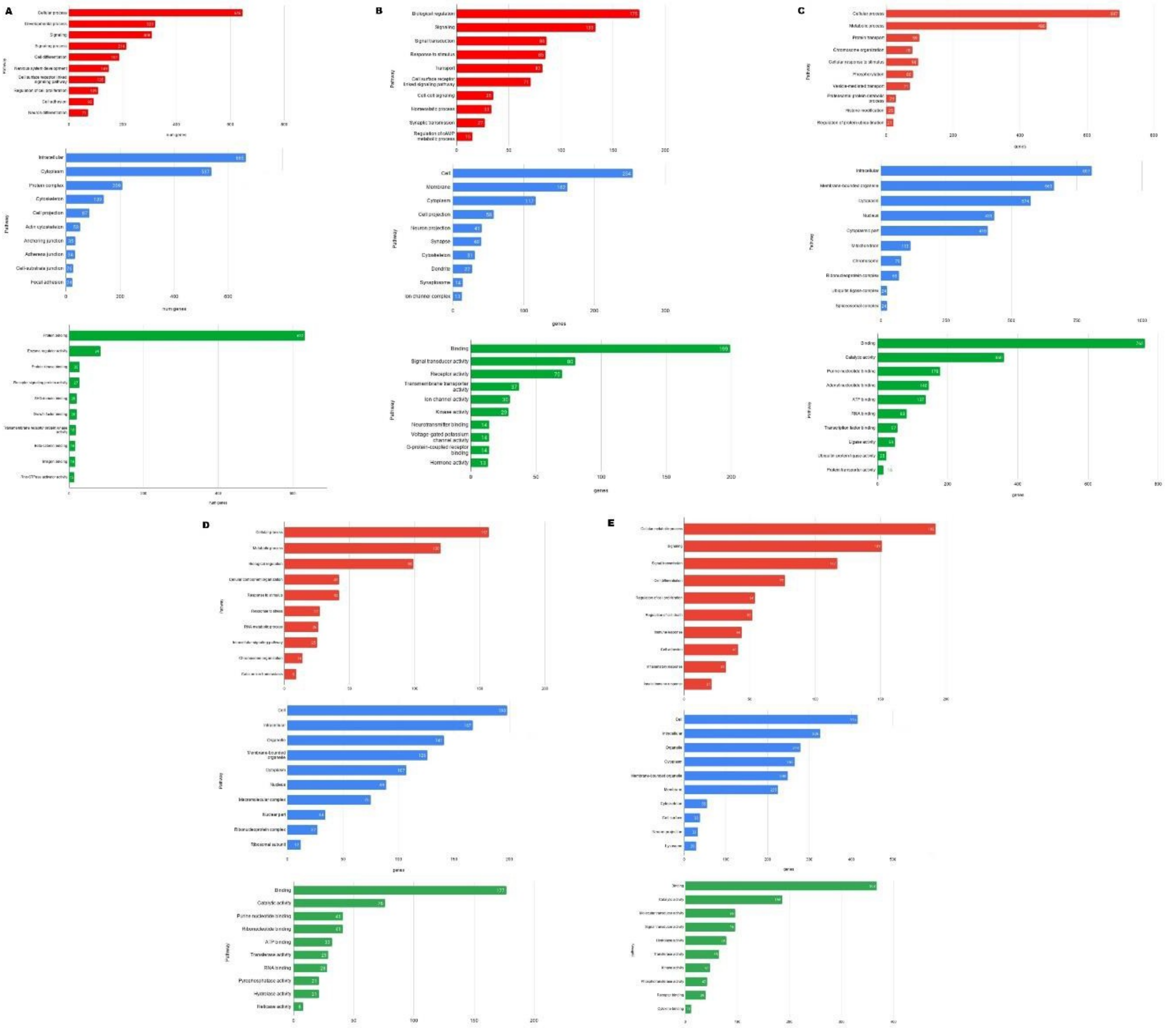


Figure 4: Enrichment results of Gene Ontology. Red = Biological Process; Blue = Cellular Component; Green = Molecular Function.

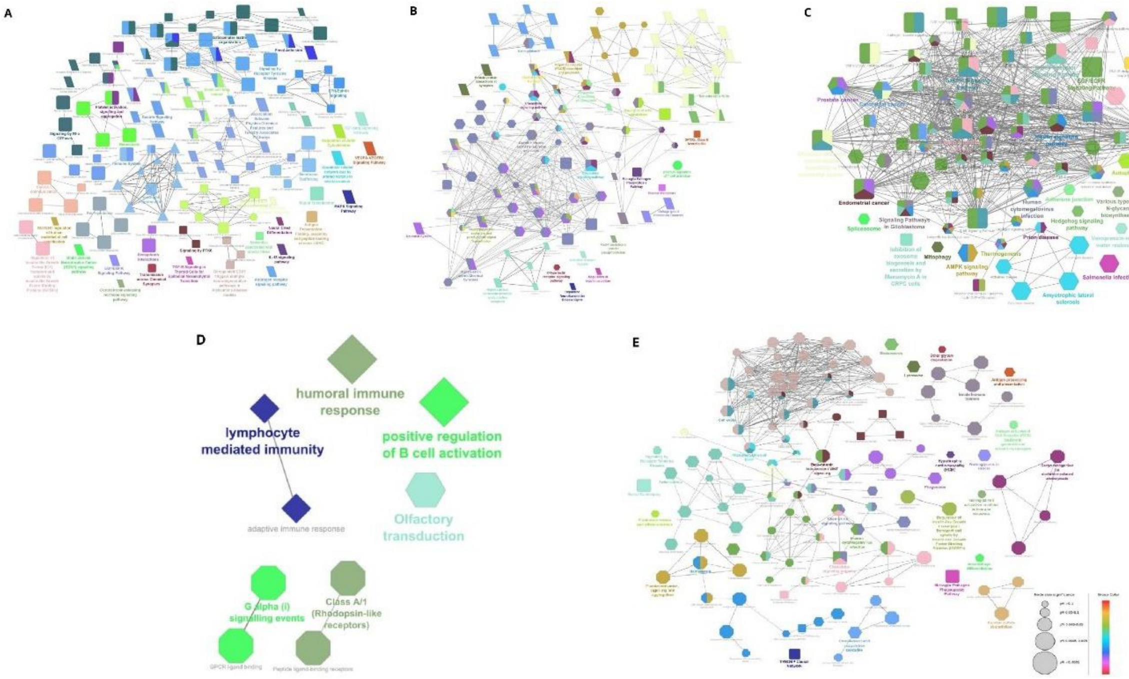

A
B
C
Prostate cancer
Endometrial cancer
Spliceosome
Signaling Pathways in Glioblastoma
Human cytomegalovirus infection
Autophagy
Ubiquitin mediated proteolysis
Various types of N-glycan biosynthesis
Endocytosis
Hedgehog signaling pathway
Prion disease
Mitophagy
Thermogenesis
AMPK signaling pathway
Salmonella infection
Amyotrophic lateral sclerosis
D
humoral immune response
lymphocyte mediated immunity
positive regulation of B cell activation
adaptive immune response
Olfactory transduction
G alpha (i) signalling events
Class A/1 (Rhodopsin-like receptors)
GPCR ligand binding
Peptide ligand-binding receptors
E

Figure 5: ClueGO enrichment results. We used the databases of GO Immune processes, KEGG, Reactome, and Wikipathways. Node size significance is represented by the size. The colors indicate groups of related ontologies.

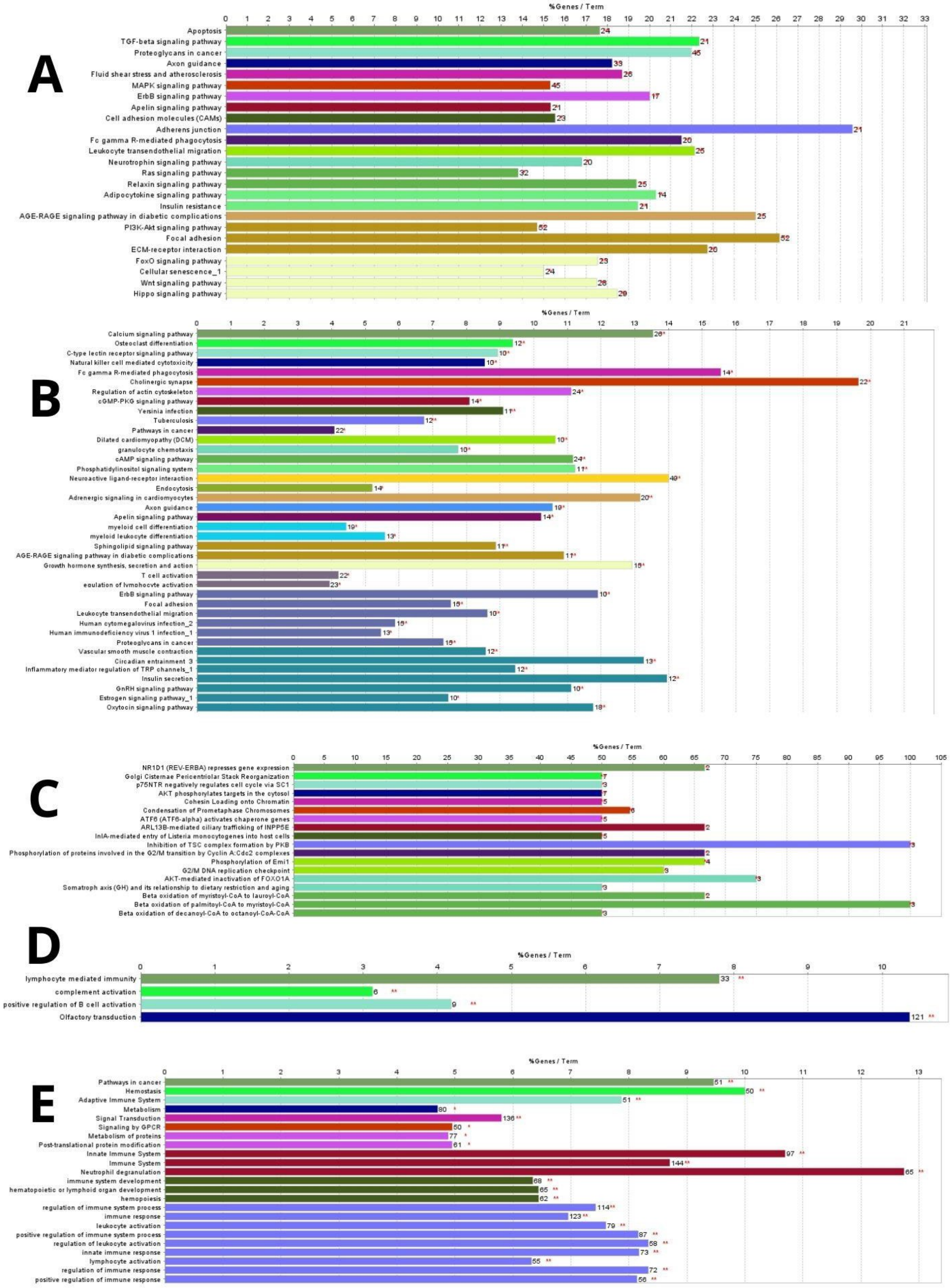


Figure 6: Histogram of the ClueGO results with the % of genes per terms, and the number of genes presented in the networks. The colors indicate groups of related ontologies.

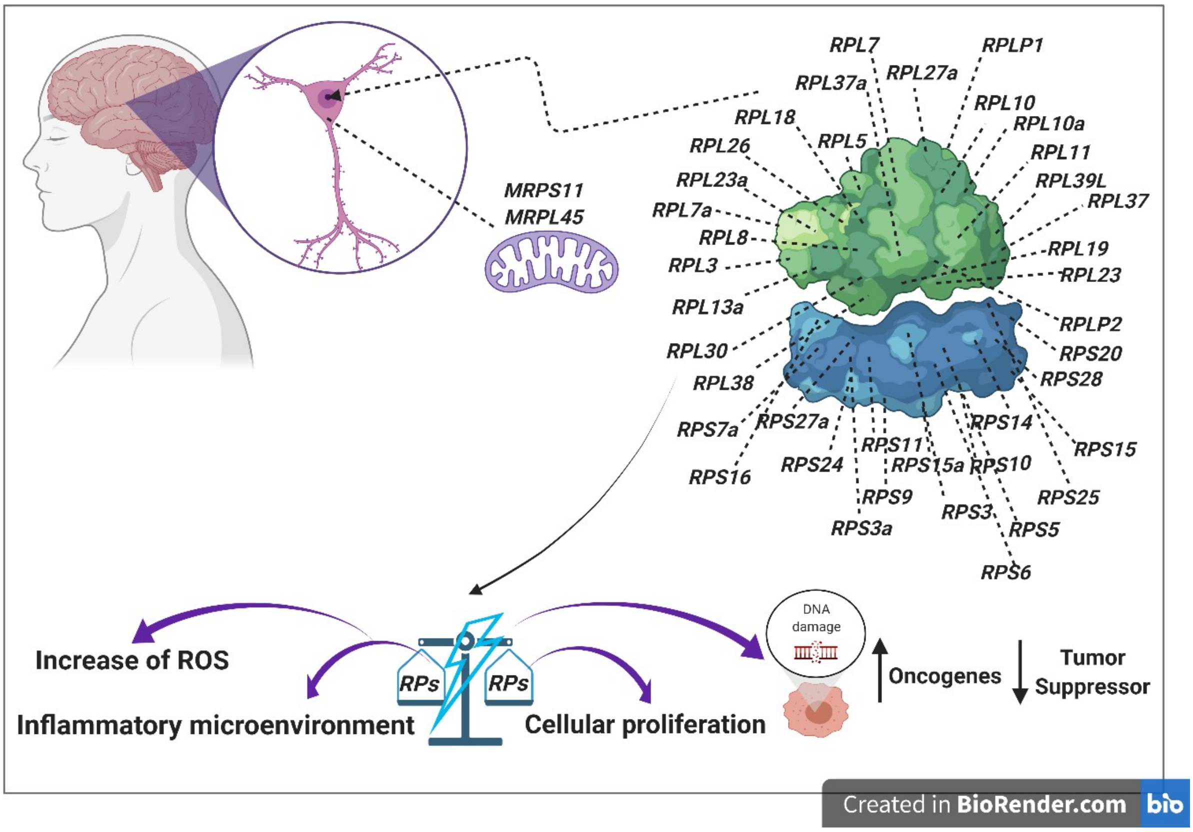


Figure 7: Schematic representation of the localization of ribosomal proteins. On the right, the genes represented in this figure are differentially expressed in the datasets analyzed in our study. On the left, functional groups of the ribosomal proteins identified in the transcriptomes and the related functions. Below this, diagrammatic representation of the consequences of ribosomal protein perturbations in cells. Perturbations in the ribosomal proteins due to increase of reactive oxygen species (ROS), creation of an inflammatory response, disturbances in the proliferation, and DNA damage. This damage leads to increase of concentration of oncogenes and decrease levels of tumor suppressor genes. (Figure create in Bio Render, https://biorender.com/)

Capítulo 5

Doenças lisossomais e tumores neurológicos: O que eles têm em comum?

Manuscrito a ser submetido ao *Journal of Neuro-Oncology*.

Lysosomal genes in neurological tumors: What's that got to do with lysosomal storage diseases?

Gerda Cristal Villalba, Eduardo Chiella, Ursula Matte

Abstract

Introduction: Defects in lysosomal hydrolases activity and transport lead to lysosomal storage diseases (LSD). The involvement of lysosomes also has been described in cell proliferation and signaling processes, microbial killing, cytotoxic killing, induction of angiogenesis, cell adhesion, and metastatic processes. Aims: This work investigates the differential expression and genetic variants in lysosomal genes and autophagy-related genes in brain tumors. We also associated gene expression with the survival status of such patients. Results: In Low-Grade Glioma (LGG) we found 51 variants predicted to be pathogenic by in silico analysis, whereas in Glioblastoma Multiforme (GBM) 100 variants deemed pathogenic were found. All variants were present in heterozygosis in tumor tissue. The lysosomal disorders associated with these variants are Niemann-Pick type C, GM1 Gangliosidosis, Mannosidosis, Mucolipidosis, Neuronal Ceroid Lipofuscinosis, Pompe, Tay-Sachs, Fabry and Mucopolysaccharidoses type IIIC, IVA, VI, and VII. Gene expression analysis discriminated two clusters both in LGG and GBM, but without statistical significance considering gender or vital status. KEGG pathway analysis showed that the top ten down-regulated genes in LGG are related to signaling and cellular processes, glycan biosynthesis and metabolism, and transporter activity. The top ten upregulated genes in those tumors are related to hexosaminidase and hydrolase activity, lysosome organization, glycosphingolipid, ceramide, lipopolysaccharide, glycolipid, and sphingolipid metabolic process. In GBM, in turn, the top ten differentially down-regulated genes participate in transporter activity, ATPase activity, hydrolase activity, synaptic vesicle cycle, and phagosome, while, the ten top upregulated genes are involved in immune processes, carbohydrate derivative catabolic processes, aminoglycan, and glycosaminoglycan catabolic processes. In LGG, increased survival was associated with 11 upregulated and 16 downregulated genes. In GBM, the upregulation of 5 genes and the downregulation of 9 genes were associated with increased survival. Therefore, studying gene signature and prognosis impact of lysosomal genes may help understand the mechanisms by which neurological dysfunction occurs in patients with neurological tumors and with lysosomal storage diseases.

Introduction

Gliomas are neuroepithelial tumors originated from glial derived-cells (Zong et al., 2012). They are classified as pilocytic astrocytoma or gangliogliomas (WHO grade I); diffuse gliomas, such as astrocytic, oligodendroglial, or mixed oligodendroglial-astrocytic in grade II or lower-grade glioma (LGG); anaplastic in WHO grade III; and glioblastoma multiforme (GBM), medulloblastoma, or ependymoblastoma in WHO grade IV. The non-diffuse gliomas are represented by pilocytic astrocytoma and ependymoma (WHO grade I, II, III and IV). In 2016, a new classification was adopted, taking into account molecular aspects of these tumors, especially IDH1 and IDH2 mutation, that occurs more frequently in astrocytomas, which are also the most commonly observed gliomas (Perry & Wesseling, 2016; Wesseling & Capper, 2018).

Cancer cells have enlarged lysosomes, as in most lysosomal storage diseases (Boya et al., 2003). Lysosomes are cellular compartments responsible, among other functions, for the degradation of macromolecules through hydrolases. Defects in these enzymes, or in their transport and in proteins that modulate their activity, culminate in the lysosomal accumulation of macromolecules or intermediate metabolites, known as lysosomal storage diseases (Parenti et al., 2015). Besides, lysosomes are involved in the occurrence or modulation of several hallmarks of cancer, including cell proliferation and signaling processes, angiogenesis induction, and metastatic processes by affecting the composition of extracellular matrix (ECM) (Pastores & Hughes, 2017). Several studies show the interplay between lysosomal genes and cancer (figure 1, supp table 4) (Chi et al.,2010; Sänger et al.,2015). Also, these genes may confer tumor growth advantage by interfering with signal transduction and growth factor distribution (Schaaf et al., 2019; Vijayan et al., 2019). In neurological tumors, lysosomal membrane destabilization leads to vulnerability to glioblastoma invasion (Le Joncour et al., 2019). Some lysosomal markers, such as CD68, are upregulated in GBM and IDHwt gliomas (Wang et al., 2018), are correlated with tumor-associated macrophages (TAMs) (Doan et al., 2017). Other lysosomal enzymes like V-ATPase are related to the progression of IDHwt lower-grade gliomas and are associated with glioma growth (Doan et al., 2017). Moreover, the V-ATPase was identified as a novel therapeutic target for glioblastoma (Halcrow et al., 2019).

Also, altered composition or organization of macromolecules whose turnover involves lysosomal degradation, such as heparan sulfate proteoglycans (HSPG), are associated with tumor progression and promote glioblastoma tumor invasion (Xiong et al., 2014; Tran et al., 2017). There are reports of increased tumor risk in some patients with lysosomal disorders, such as Multiple Myeloma and Gaucher Disease (Choy & Campbell, 2011; Mistry et al., 2013), and two siblings with Gaucher Disease who developed glioblastoma (Lyons et al., 1982).

To identify the impact of genes related to lysosomal storage diseases in the survival of patients with gliomas, we performed gene expression analysis of lysosomal genes, Kaplan Meier multivariate curves, and simple nucleotide variation analysis focusing on LSD related genes in samples from The Cancer Genome Atlas repository.

Methods

Datasets

All data used in this work is available at The Cancer Genome Atlas (TCGA) / GDC Data Portal (<https://portal.gdc.cancer.gov/>). Data from two cohorts of glioma patients were used. One cohort consisted of data from 516 patients of Lower Grade Glioma (LGG), and the other was composed of data from 395 patients with Glioblastoma Multiforme (GBM). We use gene expression datasets retrieved from GEO to identify lysosomal storage diseases related to genes responsible for tumor progression according to the different grades <https://www.ncbi.nlm.nih.gov/geo/>, accession numbers GSE15824, GSE14880, GSE12992, and GSE16155. All analyses were performed in R version 3.5.0 (R Core Team, 2017), except if indicated otherwise.

Single Nucleotide Variant Analysis

For Single Nucleotide Variation (SNV) analysis, all mutation annotation format files (MAF) were downloaded with GDCquery_maf function (Colaprico et al., 2015) in the TCGAbiolinks (Silva et al., 2016) workflow (v.2.16.3), filtered by curated maf, and “access = open”, with the already established muse pipeline (Cibulskis et al., 2013; Mayakonda et al., 2018; Mounir et al., 2019). In further analysis we worked with the curated maf (<LGG_FINAL_ANALYSIS.aggregated.capture.tcga.uuid.curated.somatic.maf>) and (<ucsc.edu_GBM.IlluminaGA_DNASeq_automated.Level_2.1.1.0.somatic.maf>), for LGG and GBM respectively.

For better summarization and plotting of all variants found, we used the maftools (v.2.4) plotmafSummary and oncoplot (Mayakonda et al., 2018). For the classification of SNVs into transitions or transversions, we used the titv function. Finally, for variant annotation, we adopted The Ensembl Variant Effect Predictor (McLaren et al., 2016). The variant effect was analyzed with SilVA (v1.1.1) for the synonymous (silent) variants (Buske et al., 2013), DDIG-in (v.1.0) for the frameshift and nonsense variants (Folkman et al., 2015), and with Human Splice Finder (v.3.1) for the splicing variants (Desmet et al., 2009). The missense variants were analyzed with SIFT (v.6.2.1) (Kumar et al., 2009; Sim et al., 2012), Polyphen2 (v.2.1) (Adzhubei et al., 2010), dbNSFP (v.3.0) (Liu et al., 2016), CADD (v.1.6) (Rentzsch et al., 2018), and Condel algorithms (v.2.0) (González-Pérez & López-Bigas, 2011). The adopted effect was given by the consensual result for at least three algorithms.

All variants deemed non-pathogenic were excluded from further analysis. To find out if the variants found are disease-causing of any lysosomal storage disorders, we performed a search in the Varsome Clinical platform (Zhang et al., 2020), following the American College of Medical Genetics recommendations and guidelines (Kalia et al., 2017). We also examined the population frequency of the variants in a database of healthy individuals (GnomAD) according to the SNP id number (Landrum et al., 2018; Karczewski et al., 2019). Analysis of driver genes was performed with the OncoKB database(Chakravarty et al., 2017).

Gene Expression Analysis

The genes related to LSD were selected based on the official list of lysosomal storage diseases elaborated by WORLD, a research consortium on lysosomal diseases, which can be accessed at <https://worldsymposia.org/official-list-of-lysosomal-diseases/> (supplementary table 2). RSEM / TPM normalization pipeline was used. As this latter was absent for the LGG cohort, the comparison between tumor grades, II and III was performed. The tumor grading system followed the general recommendations, being grade II: Moderately differentiated (intermediate grade); grade III: Poorly differentiated (high grade); and grade IV: Undifferentiated (high grade) (Wesseling & Capper, 2018). For the GBM cohort, gene expression was compared between the tumor and normal adjacent tissue. The analysis was performed on R2: Genomics analysis and visualization platform (Jan et al., 2015). We used the k-means clustering with the centroid value. For the analysis of tumor grade, we used samples of Pilocytic Astrocytoma (WHO grade I, GSE73066), Diffuse Astrocytoma (WHO grade II, GSE68848), Anaplastic Ependymoma (WHO grade III, GSE16155), and Medulloblastoma (WHO grade IV, GSE12992). We retrieved the datasets in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/). As all datasets are derived from the same experimental platform, no batch effect correction was needed. We performed gene enrichment analysis with ClusterProfiler v.3.1.8 (Yu et al., 2012).

Survival Curve

Kaplan Meier multivariate survival analysis was performed in OncoLnc with coxph function from the R survival library (v.3.2), with a p-value cutoff of 0.05 for significance and adjusted by False Discovery Rate (FDR). For the LGG cohort, data from 510 patients were available, whereas for the GMB cohort data from 152 patients was available.

Results

Single Nucleotide Variant Analysis

To identify if pathogenic variants were present in any of the 42 lysosomal disease-related genes were present in glioma patients, mutation annotation files were used to predict deleterious variants. A summary of the types of mutations found in LGG and

GBM can be found in figure 2. In LGG we found 51 variants predicted to be pathogenic by in silico programs, whereas in GBM 100 variants deemed pathogenic were found. All variants were present in tumor tissue and heterozygosis. Clinvar was used to retrieve the rs code for reported variants, and we obtained 55 rs IDs in 22 genes: 17 rs IDs were found for LGG and 38 for GBM, all predicted to be damaging or possibly damaging (Tables 1 and 2). Out of these, there were three variants in LGG and 14 in GBM that have been previously reported in lysosomal storage disease patients. The lysosomal disorders associated with these variants are Niemann-Pick type C, GM1 Gangliosidosis, Mannosidosis, Mucolipidosis, Neuronal Ceroid Lipofuscinosis, Pompe, Tay-Sachs, and Mucopolysaccharidoses type IIIC, IVA, VI, and VII. The low allele frequency of these variants is compatible with the profile of rare alleles found in rare diseases, such as lysosomal storage disorders. In both cohorts, no driver genes were identified in OncoKB. In addition, two missense mutations in the GLA gene were found, one in an LGG patient (p.Leu19Arg), female, 35 years old, and one in a GBM patient (p.Asp234Tyr), female, 77 years old. Both mutations were predicted to be "Deleterious" by SIFT and "Probably Damaging" by PolyPhen and were reported in patients, and p.Asp234Tyr is a pathogenic variant reported in Varsome and the literature. In contrast, the other variant has not been reported before. No clinical information regarding a presumptive diagnostic of Fabry disease was available for these patients.

Gene Expression Analysis

In the 516 samples of LGG (figure 3), 121 genes were found differentially expressed, 75 downregulated and 46 upregulated in cluster 1 (yellow), and 53 downregulated and 68 upregulated in cluster 2 (purple). Cluster 1 (325 samples) comprised 52.61% of the barcodes related to grade II, and 47.09% in grade III and 0.3% were ND (not determined). According to the vital status composition of cluster 1, 84.91% barcodes were alive, 14.79% dead, and 0.3% nd. In cluster 2 (191 samples), 37.17% barcodes were grade II, 62.3% grade III, and 0.52% were discrepant (probably the patient has an Oligoastrocytoma). Regarding the vital status of cluster 2, 76.96 % of barcodes were alive and 23.04% dead. The gender of cluster 1 was 45.84% females, 53.85% males, and 0.3% ND, whereas in cluster 2 there were 42.41% females and 57.59% males. None of these differences was statistically significant. More information about the clinical data are shown in supplementary table 1.

In GBM samples (figure 4), 115 genes were differentially expressed, 84 downregulated and 31 upregulated in cluster 1 (yellow) and 29 downregulated and 86 upregulated in cluster 2 (purple). The gender of cluster 1 was composed of 32.56% females and 67.44% males. The gender of cluster 2 was composed of 38.8% female and 61.19% male. In cluster 1 36.05% of patients were alive, 62.79% were dead, and 1.16% had no information about the vital status. The vital status of cluster 2 comprised 31.34 % of alive patients, 67.16% dead, and 1.5% without information. Again, these differences were not statistically significant.

Biological processes that require the lysosome's functioning and structure involve the integrative activity of dozens of proteins. In an effort to integrate the expression of single genes in patterns of expression, we sum the expression levels of all genes for each individual, generating a global lysosomal gene expression factor (supplementary table 3). As observed in Figure 5A, 82.9% of patients from cluster 1 showed a global reduction in lysosome genes, while 92.7% all patients from cluster 2. In the GBM cohort (Figure 5B), 85.9% of patients from cluster 1 showed a reduction in lysosome genes, while 98.4% all patients from cluster 2. This integrative, individual analysis shows that, despite the heterogeneity, clearly there were two profiles of expression considering lysosomal genes in both LGG and GBM patients.

Then, we analyzed whether the expression levels of the lysosomal genes were associated with the vital status (i.e. death or alive) in the different clusters. In LGG-cluster 1, eight lysosomal genes were significantly associated with death, three down-regulated (GNPTG, IDUA, and LIPA), and five upregulated genes (GLA, GLB1, GNS, HEXB, NAGA). In LGG-cluster 2 17 genes were significantly associated with the vital status, three down-regulated (CLN3, GNPTG, PSAP), and 14 upregulated (CTNS, CTSK, FUCA1, GALNS, GLA, GLB1, GNS, GUSB, HEXB, HGSNAT, MAN2B1, OCRL, SGSH, SLC38A9). For GBM-, 7 upregulated genes were significantly associated with death 6 in Cluster 1 (FUCA1, GUSB, HEXA, SGSH, SLC17A5, SMPD1) and the SLC38A9 gene in Cluster 2. Important to mention, four genes were present in both LGG and GBM: FUCA1 in cluster 2 of LGG and cluster 1 of GBM, GUSB in cluster 2 of LGG and cluster 1 of GBM, SGSH in cluster 2 of LGG and cluster 1 of GBM, and finally, the gene SLC38A9 in the cluster 2 of LGG and cluster 2 of GBM. These results indicate a potential role of these lysosomal genes in the outcome of glioma patients.

In the analysis of differential expression considering tumor grades (Figure 6), 44 lysosomal enzymes were enriched in the comparison of tumor grade I vs. II, 25 in tumor grade II vs. III, 41 in tumor grade III vs. IV, and 28 in tumor grade I vs. IV. Thus, it is plausible to assume that alterations in the lysosome pathway are involved in the progression of gliomas.

We summarized the results of gene expression analysis in the KEGG pathway in Figure 7. The most differentially expressed genes in the lysosome KEGG pathway in LGG are shown in Figure 7A. The top ten down-regulated genes (Table 3) are related to signaling and cellular processes, glycan biosynthesis and metabolism, and transporter activity. The top ten upregulated genes are related to hexosaminidase and hydrolase activity, lysosome organization, glycosphingolipid, ceramide, lipopolysaccharide, glycolipid, and sphingolipid metabolic process. In addition, these genes also participate in neutrophil and myeloid activation involved in immune response and leukocyte degranulation.

In GBM, the top ten differentially down-regulated genes (Figure 7B, Table 4) participate in transporter activity, ATPase activity, hydrolase activity, synaptic vesicle cycle, and phagosome. On the other hand, the ten top upregulated genes are involved in immune processes, carbohydrate derivative catabolic processes, aminoglycan, and glycosaminoglycan catabolic processes.

The most differentially expressed hydrolases in LGG are proteases, such as cathepsin; glycosidases, such as glucosaminidase and mannosidase; sulfatase, represented by arylsulfatase B; ceramidase represented by N-acylsphingosine amidohydrolase 1, and other lysosomal enzymes, for example, GM2 Ganglioside Activator. In the GBM, the most differentially expressed hydrolases are members of proteases, like cathepsin C and S; glycosidases like galactosidase beta, alpha-N-Acetylgalactosaminidase, glucosaminidase and mannosidase, and sulfatases, like glucosamine (N-Acetyl)-6-sulfatase and iduronate 2-sulfatase.

The lysosomal genes analyzed in this study also participate in several biological processes, like glycosaminoglycans degradation, endocytosis, autophagy regulation, vesicular transport, and Golgi network. These processes emphasize the importance of the lysosomal genes and their products to cell organization and survival.

Survival Curve

Out of the 123 lysosomal genes, we selected 42 related to lysosomal diseases (30 lysosomal hydrolases) for survival analysis. In LGG, 11 genes were related to increased survival when upregulated ($p<0.05$) and 16 when downregulated ($p<0.05$) (supplementary table 4). In GBM, 5 up-regulated genes were associated with better prognosis when upregulated ($p<0.05$) and 9 when down-regulated ($p<0.05$) (supplementary table 4). Figure 8 shows representative Kaplan-Meier curves, demonstrating the strong impact of GLA, HEXB, FUCA1, and MANBA in glioma survival.

Discussion

The importance of the lysosomal pathway in cancer development is shown by the different therapeutic strategies focusing on autophagy, lysosomotropic agents, and specific lysosomal protease inhibitors (Morell et al., 2016; Dielschneider et al., 2017; Trejo-Solís et al., 2018; Schaaf et al., 2019). For example, Dielschneider and colleagues reviewed the use of lysosomotropic agents such as Siramesine, Desipramine, Nortriptyline, Amlodipine, and Terfenadine in different cancer cell lines and animal models, including glioblastoma (Dielschneider et al., 2017). Furthermore, the use of specific lysosomal protease inhibitors, as cathepsins, has also been tested (Minchenko et al., 2017; Liang et al., 2019). In addition, cancer cells tend to increase the number and size of lysosomes to cope with the increased demand for recycling macromolecules and growth factors (Dielschneider et al., 2017). Therefore, alterations in the expression of lysosomal-related genes may impact cancer development and progression (Trejo-Solís et al., 2018).

Our results show extensive tumor heterogeneity in LGG and GBM, but both present a gene expression signature when considering genes related to lysosomal function as determined by the KEGG pathway. Interestingly, these signatures are not associated with survival, gender, or histological type, but they could be related to tumor initiation and/or progression, as suggested by our results from differential expression along with tumor progression (Figure 6). The two cohorts shared differentially expressed genes, such as CD68, DNASE2, and MAN2B1 among the top 10 down-regulated genes, and AP3B2, ATP6V1H, NAGPA, and SLC17A6 among the top 10 upregulated (figure 7A and 7B). The CD68 is a transmembrane glycoprotein that is found in human monocytes and tissue macrophages. It is a member of the lysosomal-associated membrane glycoprotein (LAMP) family, localized in lysosomes and endosomes (Chistiakov et al., 2016). The high expression of CD68 is involved in glioma progression, and this gene serves as a prognostic biomarker (Strojnik et al., 2009; Mangogna et al., 2019). Wang and colleagues demonstrated that the high expression of CD68 in tumors was correlated with poor survival in glioma patients, is being a therapeutic promise for cancer immunotherapy (Wang et al., 2018). Deoxyribonuclease II (DNASE2) is primarily found in the lysosome, and its function is related to the degradation of exogenous DNA encountered by phagocytosis (Evans et al., 2003). The intranuclear delivery of recombinant DNASE2 effectively degrades genomic DNA in human breast cancer cells and can be applied to other types of cancer (Malecki et al., 2013). The adaptor protein complex 3 (AP3B2) is a clathrin-associated complex responsible for cargo transport from tubular endosomes to late endosomes (Park et al., 2014). It is responsible for the formation of synaptic vesicles and the transport of lysosomal enzymes to the trans-Golgi network (Blumstein et al., 2001). The AP3B2 gene is related to neurological diseases, such as early-onset epileptic encephalopathy (Ville et al., 2016) and autoimmune cerebellar ataxia (Jarius & Wildemann, 2015). V-ATPases are ATP-dependent proton pumps related to vesicle trafficking, particularly in lysosomes (Futai et al., 2019). The ATP6V1H expression has been implicated in carcinogenesis and metastasis in esophageal squamous cell carcinoma and glioma (Couto-Vieira et al., 2020). Genes of the V-ATPase family have a specific signature and confer glioma aggressiveness by activating oncogenic pathways, such as apoptosis resistance and autophagy, and in the signaling pathways mTOR, Notch, and Wnt (Terrasi et al., 2019). Lysosomal solute carriers (SLCs) are found across the lysosomal membrane and are involved in the lysosomal transport of solutes. Many members of this protein family are involved in the mTOR complex and toll-like receptor (TLR) signaling (Bissa et al., 2016). The SLC17A6 gene, also found down-regulated in glioma, has been suggested as a biomarker for this type of tumor (Geng et al., 2018) and a target for anticancer chemotherapy (Al-Abdulla et al., 2019).

In particular, lysosomal hydrolases were differentially expressed in LGG and GBM, including enzymes involved in glycosaminoglycan and sphingolipid degradation.

Arylsulfatase A and B are related to progression and invasion in central nervous system tumors (Kovacs et al., 2019). Acid ceramidase, one of the top 10 upregulated hydrolases present in LGG, may regulate p53-independent apoptosis in human glioma cells (Hara et al., 2004), and the inhibition of ASAH1 decreases the cell growth of glioblastoma (Doan et al., 2017). Levicar and collaborators demonstrated the overexpression of Cathepsin D in Astrocytoma samples and the downregulation of Cathepsin B in Glioblastoma and Meningioma (Levicar et al., 2002). Cathepsin A and K are upregulated in U87 glioma cells and favors cell proliferation of tumor growth (Minchenko et al., 2017). FUCA1 has a tumor-suppressive function and is regulated by p53 in the T98G glioblastoma cell line (Ezawa et al., 2016). Moreover, it is known that inhibition of the Wnt/β-catenin signaling pathway increases the expression of ASAH1, CTSC, DNASE2, GAA, GBA, GM2A, HEXA, MANBA, NAGLU, and TPP1 (Gao et al., 2017), which may explain our results.

The expression of alpha-glucosidase is related to glioblastoma invasiveness (Anji et al., 2015). The GALC gene participates in glycosphingolipid metabolism, and different glycosphingolipids were previously correlated with malignancy grade in gliomas (Becker et al., 2000). According to Thaker and collaborators, GALNS increased glioma survival (Thaker et al., 2009), and other studies show the proliferative function of GALNS in primary and secondary glioblastomas (Tsigelny et al., 2015). GBA amplification is found in astrocytoma, glioblastoma, and oligoastrocytoma and is related to the survival of glioma patients (Gargini et al., 2019). GLB1, which encodes for beta-galactosidase, is upregulated in glioma cells and associated with survival in glioblastoma patients (Tong et al., 2018). In our survival analysis, the high expression of these genes were correlated with better survival in LGG and GBM.

In addition, GNS up-regulation is correlated with survival in GBM patients (Tong et al., 2018), and high expression is unfavorable in glioma (Uhlen et al., 2017). Our analysis, on the contrary, shows that high expression is a better predictor of survival in LGG and the lower expression is better for GBM. In glioblastomas, we observed an increased expression of beta-glucuronidase (GUSB) compared with astrocytomas, as previously described (Uhlen et al., 2017).

Hexosaminidase subunit alpha and beta also are upregulated in GBM and are associated with survival (Tong et al., 2018). For HEXA, the low expression is related to better survival in LGG, and in GBM, the high expression increased the survival. The HEXB low expression in LGG indicates a better survival, and in GBM, the high expression is related to better survival. Hyaluronidase (HYAL1) plays an important role in cancer metastasis (McAtee et al., 2014), especially in gliomas (McAtee et al., 2015). In LGG and GBM, the lower expression of this gene indicated a better survival. Iduronate 2-sulfatase (IDS) was found to be down-regulated in GBM (Wade et al., 2015) and 5 GBM cell lines sensitive to temozolomide (Ujifuku et al., 2010). the analysis here, in both LGG and GBM the lower expression of IDS was associated with better survival. IDUA gene expression is upregulated in GBM flanking tumor and peritumoral areas in patients with long and short survival (Fazi et al., 2015). In LGG, its lower expression was associated with better survival. In contrast, in GBM, the higher expression was correlated with better survival. Lysosomal acid lipase A (LIPA) is upregulated in two cell lines of GBM (Picco et al., 2014). In both LGG and GBM, our analysis showed an association of LIPA high expression with better survival. MANBA is upregulated in glioma cells and associated with survival (Tong et al., 2018), which was corroborated by our results.

In our work, we identified 17 pathogenic variants in LGG and GBM patients that have been previously described in patients with 11 different lysosomal storage diseases. Several studies have shown an increased risk of cancer in various types of LSD patients. For example, Shin et al. (2019) indicated that patients with Gaucher or Fabry disease present an increased cancer risk. They reported pathogenic germline variants enriched in cancer patients, especially in the pancreatic adenocarcinoma cohort (Shin et al., 2019). Therefore, Bird et al. (2017) demonstrated this increased risk in Fabry patients compared to the general population. Fabry Disease is caused by mutations in the GLA gene, which leads to defects in the lysosomal enzyme α-galactosidase A. Their results suggest that Fabry patients have higher rates of melanoma, urological malignancies, and meningiomas than the general population (Bird et al., 2017; Saudubray & Garcia-Cazorla, 2018). We found two missense mutations in the GLA gene, one in an LGG patient (p.Leu19Arg) and one GBM patient (p.Asp234Tyr). Both mutations were predicted to be "Deleterious" by SIFT and "Probably Damaging" by PolyPhen.

In conclusion, here we presented a comprehensive annotation of genomic alterations status and differential expression of lysosomal related genes in which wasGlioblastoma Multiforme and Lower-grade glioma. We also shed some light on the potential roles of lysosome genes in glioma progression, intertumoral heterogeneity, survival, and prognosis. Differentially expressed genes may serve as prognostic biomarkers and therapeutic targets for both lysosomal diseases and neurological tumors.

Funding

We would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the fellowships provided GCVS (Process no.: 148615/2018-0) and UM (Process no.: 312714/2018-1). We also thank FIPE/HCPA (Project Number 2018-0594) for financial support.

Data availability

All data and code used in this work are also available in <https://github.com/Kur1sutaru/TCGA-LGG-GBM >. The TCGA data are available in GDC Data Portal, <https://portal.gdc.cancer.gov/ >

Supplementary data
Supp1 Clinical data information of LGG and GBM cohorts used in this study.
Supp2 Lysosomal genes and the related lysosomal diseases
Supp3 Integrative expression analysis of lysosomal genes both LGG and GBM
Supp4 Kaplan meier results of LGG and GBM
Supp5 Review of lysosomal genes and cancer related articles.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, … Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat Methods, 7(4), 248–249. doi:10.1038/nmeth0410-248.

Al-Abdulla, R., Perez-Silva, L., Abete, L., Romero, M. R., Briz, O., & Marin, J. J. G. (2019). Unraveling “The Cancer Genome Atlas” information on the role of SLC transporters in anticancer drug uptake. Expert Review of Clinical Pharmacology, 12(4), 329–341. https://doi.org/10.1080/17512433.2019.1581605

Anji A, Miller H, Raman C, Phillips M, Ciment G, Kumari M (2015). Expression of α-subunit of α-glucosidase II in adult mouse brain regions and selected organs. J Neurosci Res, 93(1), 82–93. doi:10.1002/jnr.23470

Azad, B., Efthymiou, S., Sultan, T., Scala, M., Alvi, J. R., Neuray, C., Dominik, N., SYNaPS Study Group, Gul, A., & Houlden, H. (2020). Novel likely disease-causing CLN5 variants identified in Pakistani patients with neuronal ceroid lipofuscinosis. Journal of the neurological sciences, 414, 116826.

Becker, R., Rohlfs, J., Jennemann, R., Wiegandt, H., Mennel, H. D., & Bauer, B. L. (2000). Glycosphingolipid component profiles or human gliomas--correlation to survival time and histopathological malignancy grading. Clinical neuropathology, 19(3), 119–125.

Beier UH, Gorogh T (2005). Implications of galactocerebrosidase and galactosylcerebroside metabolism in cancer cells. Int J Cancer, 115,p. 6–10

Bhattacharyya S, Feferman L, Han X, Ouyang Y, Zhang F, Linhardt RJ, Tobacman JK (2018). Decline in arylsulfatase B expression increases EGFR expression by inhibiting the protein-tyrosine phosphatase SHP2 and activating JNK in prostate cells. J biol chem, 293(28), 11076–11087. doi:10.1074/jbc.RA117.001244

Bhattacharyya S, Feferman L, Tobacman JK (2017). Chondroitin sulfatases differentially regulate Wnt signaling in prostate stem cells through effects on SHP2, phospho-ERK1/2, and Dickkopf Wnt signaling pathway inhibitor (DKK3). Oncotarget, 8: p.100242-100260, doi: https://doi.org/10.18632/oncotarget.22152

Bird S, Hadjimichael E, Mehta A, Ramaswami U, Hughes D (2017). Fabry disease and incidence of cancer. Orphan j rare dis, 12(1), 150. doi:10.1186/s13023-017-0701-6

Bissa, B., Beedle, A. M., & Govindarajan, R. (2016). Lysosomal solute carrier transporters gain momentum in research. Clinical pharmacology and therapeutics, 100(5), 431–436. https://doi.org/10.1002/cpt.450

Blumstein, J., Faundez, V., Nakatsu, F., Saito, T., Ohno, H., & Kelly, R. B. (2001). The Neuronal Form of Adaptor Protein-3 Is Required for Synaptic Vesicle Formation from Endosomes. The Journal of Neuroscience, 21(20), 8034–8042. https://doi.org/10.1523/jneurosci.21-20-08034.2001

Bonin S, Parascandolo A, Aversa C et al. (2018). Reduced expression of α-L-Fucosidase-1 (FUCA-1) predicts recurrence and shorter cancer specific survival in luminal B LN+ breast cancer patients. Oncotarget, 9: p.15228-15238, doi:https://doi.org/10.18632/oncotarget.24445

Boya P, Andreau K, Poncet D, et al. (2003). Lysosomal membrane permeabilization induces cell death in a mitochondrion-dependent fashion. J Exp Med, 197(10):1323–1334.

Brandt-Rauf SI, Raveis VH, Drummond NF, Conte JA, Rothman SM (2006). Ashkenazi Jews and breast cancer: the consequences of linking ethnic identity to genetic disease. Am J Public Health, 96(11), 1979–1988. doi:10.2105/AJPH.2005.083014

Brattain MG, Kimball PM, Pretlow TG (1977). β-Hexosaminidase Isozymes in Human Colonic Carcinoma. Cancer Res, 37:3, p.731-735

Buono M, Cosma MP (2010). Sulfatase activities towards the regulation of cell metabolism and signaling in mammals. Cell Mol Life Sci, 67, 769–780 doi:10.1007/s00018-009-0203-3

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. (2013) Identification of deleterious synonymous variants in human genomes. Bioinformatics, doi:10.1093/bioinformatics/btt308.

Cameron C, Greenbaum L, Sato T et al. (2008). Renal cell carcinoma in a patient with cystinosis and inflammatory bowel disease: a case report. Pediatr Nephrol, 23, 1167–1170, doi:10.1007/s00467-008-0773-6

Carvalho, JAD, Barbosa, CCL, Feher, O, Maldaun, MVC, Camargo, VP, Moraes, FY, Marta, GN. (2019). Systemic dissemination of glioblastoma: literature review. Rev Assoc Med Bras, 65(3), 460-468. Epub April 11, 2019.https://dx.doi.org/10.1590/1806-9282.65.3.460

Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, … Schultz N (2017). OncoKB: A Precision Oncology Knowledge Base. JCO precision oncology, 2017, 10.1200/PO.17.00011. doi:10.1200/PO.17.00011.

Chen KJ, Jin RM, Shi CC (2018). The prognostic value of Niemann-Pick C1-like protein 1 and Niemann-Pick disease type C2 in hepatocellular carcinoma. J Canc, 9(3), 556–563. doi:10.7150/jca.19996

Cheng TC, Tu SH, Chen LC et al. (2015). Down-regulation of α-L-fucosidase 1 expression confers inferior survival for triple-negative breast cancer patients by modulating the glycosylation status of the tumor cell surface. Oncotarget, 6(25), 21283–21300. doi:10.18632/oncotarget.4238

Chi C, Zhu H, Han M, Zhuang Y, Wu X, & Xu T (2010). Disruption of lysosome function promotes tumor growth and metastasis in Drosophila. J Biol Chem, 285(28), 21817–21823. doi:10.1074/jbc.M110.131714

Mangogna, A, Belmonte, B, Agostinis, C, Zacchi, P, Iacopino, DG, Martorana, A, Rodolico, V, Bonazza, D, Zanconati, F, Kishore, U, Bulla, R (2019). Prognostic Implications of the Complement Protein C1q in Gliomas. Frontiers in immunology, 10, 2366, DOI: 10.3389/fimmu.2019.02366

Chowdhury, F. A., Hossain, M. K., Mostofa, A. G. M., Akbor, M. M., & Bin Sayeed, M. S. (2018). Therapeutic Potential of Thymoquinone in Glioblastoma Treatment: Targeting Major Gliomagenesis Signaling Pathways. BioMed Research International, 2018, 1–15. https://doi.org/10.1155/2018/4010629

Choy FYM, Campbell TN (2011). Gaucher Disease and Cancer: Concept and Controversy. Intern J Cell Biol, Volume 2011, Article ID 150450, 6 pages

Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol (2013).doi:10.1038/nbt.2514.

Coelho BA, Belo AV, Andrade SP, Amorim WC, Uemura G, da Silva Filho AL (2014). N-acetylglucosaminidase, myeloperoxidase and vascular endothelial growth factor serum levels in breast cancer patients. Biomed Pharmacother ;68(2):185-9. doi: 10.1016/j.biopha.2013.10.009

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research. doi: 10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

Couto-Vieira, J., Nicolau-Neto, P., Costa, E. P., Figueira, F. F., Simão, T. A., Okorokova-Façanha, A. L., Ribeiro Pinto, L. F., & Façanha, A. R. (2020). Multi-cancer V-ATPase molecular signatures: A distinctive balance of subunit C isoforms in esophageal carcinoma. EBioMedicine, 51, 102581

Cybulla M, Kleber M, Walter KN, Kroeber SM, Neumann HP, Engelhardt M (2006). Is Fabry disease associated with leukaemia? Br J Haematol, 35(2):264-5, doi: 10.1111/j.1365-2141.2006.06282.x

Davidson, SM, Vander Heiden, MG (2017). Critical Functions of the Lysosome in Cancer Biology. Annual Review of Pharmacology and Toxicology, 57(1), 481–507. https://doi.org/10.1146/annurev-pharmtox-010715-103101

Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, & Béroud C (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research, 37(9), e67. doi:10.1093/nar/gkp215.

Dielschneider, RF, Henson, ES, Gibson, SB (2017). Lysosomes as Oxidative Targets for Cancer Therapy. Oxidative Medicine and Cellular Longevity, 2017, 1–8. https://doi.org/10.1155/2017/3749157

Doan, N. B., Nguyen, H. S., Al-Gizawiy, M. M., Mueller, W. M., Sabbadini, R. A., Rand, S. D., Connelly, J. M., Chitambar, C. R., Schmainda, K. M., & Mirza, S. P. (2017). Acid ceramidase confers radioresistance to glioblastoma cells. Oncology reports, 38(4), 1932–1940. https://doi.org/10.3892/or.2017.5855

Elgundi, Z., Papanicolaou, M., Major, G., Cox, T. R., Melrose, J., Whitelock, J. M., & Farrugia, B. L. (2020). Cancer Metastasis: The Role of the Extracellular Matrix and the Heparan Sulfate Proteoglycan Perlecan. Frontiers in Oncology, 9. https://doi.org/10.3389/fonc.2019.01482

Evans, C. J., & Aguilera, R. J. (2003). DNase II: genes, enzymes and function. Gene, 322, 1–15. doi:10.1016/j.gene.2003.08.022

Ezawa I, Sawai Y, Kawase T, et al. (2016). Novel p53 target gene FUCA1 encodes a fucosidase and regulates growth and survival of cancer cells. Canc sci, 107(6), 734–745. doi:10.1111/cas.12933

Fazi, B., Felsani, A., Grassi,..., Mangiola, A. (2015). The transcriptome and miRNome profiling of glioblastoma tissues and peritumoral regions highlights molecular pathways shared by tumors and surrounding areas and reveals differences between short-term and long-term survivors. Oncotarget, 6(26), 22526–22552. https://doi.org/10.18632/oncotarget.4151

Fennelly C, Amaravadi RK (2017). Lysosomal Biology in Cancer. Methods in molecular biology (Clifton, N.J.), 1594, 293–308. doi:10.1007/978-1-4939-6934-0_19

Folkman L et al. (2015). DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics, v.31(10), p.1599–1606, doi : https://doi.org/10.1093/bioinformatics/btu862.

Futai, M., Sun-Wada, G. H., Wada, Y., Matsumoto, N., & Nakanishi-Matsui, M. (2019). Vacuolar-type ATPase: A proton pump to lysosomal trafficking. Proceedings of the Japan Academy. Series B, Physical and biological sciences, 95(6), 261–277. https://doi.org/10.2183/pjab.95.018

Gao, J., Arbman, G., He, L., Qiao, F., Zhang, Z., Zhao, Z., Rosell, J., & Sun, X. F. (2008). MANBA polymorphism was related to increased risk of colorectal cancer in Swedish but not in Chinese populations. Acta oncologica (Stockholm, Sweden), 47(3), 372–378. https://doi.org/10.1080/02841860701644052

Gao, L., Chen, B., Li, J., Yang, F., Cen, X., Liao, Z., & Long, X. (2017). Wnt/β-catenin signaling pathway inhibits the proliferation and apoptosis of U87 glioma cells via different mechanisms. PLOS ONE, 12(8), e0181346. https://doi.org/10.1371/journal.pone.0181346

Gargini, R., Segura-Collar, B., & Sánchez-Gómez, P. (2019). Novel Functions of the Neurodegenerative-Related Gene Tau in Cancer. Frontiers in Aging Neuroscience, 11. https://doi.org/10.3389/fnagi.2019.00231

Geng, R. X., Li, N., Xu, Y., Liu, J. H., Yuan, F. E., Sun, Q., Liu, B. H., & Chen, Q. X. (2018). Identification of Core Biomarkers Associated with Outcome in Glioma: Evidence from Bioinformatics Analysis. Disease markers, 2018, 3215958. https://doi.org/10.1155/2018/3215958

González-Pérez A, López-Bigas N (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet, 88(4), 440–449. doi:10.1016/j.ajhg.2011.03.004.

Halaby R (2015). Role of lysosomes in cancer therapy. Research and Reports in Biology, 6 147–155, DOI https://doi.org/10.2147/RRB.S83999

Halcrow, P, Datta, G, Ohm, JE, Soliman, ML, Chen, X, Geiger, JD (2019). Role of endolysosomes and pH in the pathogenesis and treatment of glioblastoma. Cancer Rep e1177. doi:10.1002/cnr2.1177

Hamza A, Khawar S, Ibrahim A, Edens J, Lalonde C, Danforth RD (2017). A second reported malignancy in a patient with Morquio syndrome. Autop rep, 7(2), 9–14. doi:10.4322/acr.2017.019

Hara, S., Nakashima, S., Kiyono, T., Sawada, M., Yoshimura, S., Iwama, T., Banno, Y., Shinoda, J., & Sakai, N. (2004). p53-Independent ceramide formation in human glioma cells during γ-radiation-induced apoptosis. Cell Death & Differentiation, 11(8), 853–861. https://doi.org/10.1038/sj.cdd.4401428

Herroon M, Rajagurubandara E, Rudy D et al. (2013). Macrophage cathepsin K promotes prostate tumor progression in bone. Oncogene, 32, 1580–1593, doi:10.1038/onc.2012.166

Ho ML, Kuo W-K, Chu LJ et al. (2019). N-acetylgalactosamine-6-sulfatase (GALNS), Similar to Glycodelin, Is a Potential General Biomarker for Multiple Malignancies. Int J Canc Res Treat, 39(11), p. 6317-6324, doi: 10.21873/anticanres.13842

Hoogstraat M, de Pagter MS, Cirkel GA et al. (2014). Genomic and transcriptomic plasticity in treatment-naive ovarian cancer. Genom Res, 24(2), 200–211. doi:10.1101/gr.161026.113

Hou G, Liu G, Yang Y et al. (2016). Neuraminidase 1 (NEU1) promotes proliferation and migration as a diagnostic and prognostic biomarker of hepatocellular carcinoma. Oncotarget, 7(40), 64957–64966. doi:10.18632/oncotarget.11778

Hu Z-D, Yan J, Cao KY, Yin ZQ, Xin W-W, Zhang M-F (2019). MCOLN1 Promotes Proliferation and Predicts Poor Survival of Patients with Pancreatic Ductal Adenocarcinoma. Disease Markers, Article ID 9436047, 9 pages, doi: https://doi.org/10.1155/2019/9436047

Jan Koster, Jan J. Molenaar and Rogier Versteeg (2015). R2: Accessible web-based genomics analysis and visualization platform for biomedical researchers. Cancer Res November 15 2015 (75) (22 Supplement 1) A2-45; DOI: 10.1158/1538-7445.

Jarius, S., & Wildemann, B. (2015). “Medusa head ataxia”: the expanding spectrum of Purkinje cell antibodies in autoimmune cerebellar ataxia. Part 3: Anti-Yo/CDR2, anti-Nb/AP3B2, PCA-2, anti-Tr/DNER, other antibodies, diagnostic pitfalls, summary and outlook. Journal of Neuroinflammation, 12(1). https://doi.org/10.1186/s12974-015-0358-9

Journet A, Chapel A, Kieffer S, Roux F, Garin J (2002). Proteomic analysis of human lysosomes: application to monocytic and breast cancer cells. Proteomics. 2002 Aug;2(8):1026-40

Kalia, SS, Adelman, K, Bale, SJ, Chung, WK, Eng, C, Evans, JP, Herman, GE, Hufnagel, SB, Klein, TE, Korf, BR, McKelvey, KD, Ormond, KE, Richards, CS, Vlangos, CN, Watson, M, Martin, CL, Miller, D T(2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genetics in medicine: official journal of the American College of Medical Genetics, 19(2), 249–255, DOI: 10.1038/gim.2016.190

Karczewski KJ et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv, doi: https://doi.org/10.1101/531210

Kasitinon SY , Eskiocak U, Martin M et al. (2019). TRPML1 Promotes Protein Homeostasis in Melanoma Cells by Negatively Regulating MAPK and mTORC1 Signaling. Cell Reports, 28, p.2293–2305, doi: https://doi.org/10.1016/j.celrep.2019.07.086

Kleer CG, Bloushtain-Qimron N, Chen YH, Carrasco D, Hu M, Yao J, Polyak K (2008). Epithelial and stromal cathepsin K and CXCL14 expression in breast tumor progression. J Am Assoc Canc Res, 14(17), 5357–5367. doi:10.1158/1078-0432.CCR-08-0732

Knelson, EH, Nee, JC, Blobe, GC (2014). Heparan sulfate signaling in cancer. Trends in Biochemical Sciences, 39(6), 277–288. https://doi.org/10.1016/j.tibs.2014.03.001

Korbelik, M., Naraparaju, V., & Yamamoto, N. (1998). The value of serum α-N-acetylgalactosaminidase measurement for the assessment of tumor response to radio- and photodynamic therapy. British Journal of Cancer, 77(6), 1009–1014. https://doi.org/10.1038/bjc.1998.166

Kovacs Z, Jung I, Gurzu S (2019). Arylsulfatases A and B: From normal tissues to malignant tumors. Pathol Res Pratic, 215(9), doi: https://doi.org/10.1016/j.prp.2019.152516

Krzeslak A, Pomorski L, Lipinska A (2010). Elevation of nucleocytoplasmic beta-N-acetylglucosaminidase (O-GlcNAcase) activity in thyroid cancers. Int J Mol Med, 25(4):643-8

Kumar P, Henikoff S, Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc, 4(8):1073-1081, doi:10.1038/nprot.2009.86.

Kuzu O, Gowda R, Noory M et al. (2017). Modulating cancer cell survival by targeting intracellular cholesterol transport. Br J Cancer, 117, 513–524, doi:10.1038/bjc.2017.200

Lai M, Realini N, La Ferla M et al. (2017). Complete Acid Ceramidase ablation prevents cancer-initiating cell formation in melanoma cells. Sci Rep, 7:7411, doi:10.1038/s41598-017-07606-w

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Le Joncour V, Filppu P, Hyvönen M, Holopainen M, Turunen SP, Sihto H, Burghardt I, Joensuu H, Tynninen O, Jääskeläinen J, Weller M, Lehti K, Käkelä R, Laakkonen P. Vulnerability of invasive glioblastoma cells to lysosomal membrane destabilization. EMBO Mol Med 2019 Jun;11(6):e9034. doi: 10.15252/emmm.201809034. PMID: 31068339; PMCID: PMC6554674.

Leusink FK, Koudounarakis E, Frank MH, Koole R, van Diest PJ, Willems SM (2018). Cathepsin K associates with lymph node metastasis and poor prognosis in oral squamous cell carcinoma. BMC Cancer, 18(1), 385. doi:10.1186/s12885-018-4315-8

Levicar N, Strojnik T, Kos J, Dewey RA, Pilkington GJ, Lah TT (2002). Lysosomal enzymes, cathepsins in brain tumour invasion. J Neurooncol, 58(1):21-32.

Li L, Wang W, Zhang R, Liu J, Yu J, Wu X, Xu Y, Ma M, Huang J (2017). High expression of LAMP2 predicts poor prognosis in patients with esophageal squamous cell carcinoma. Cancer biomark, Jul 4;19(3):305-311. doi: 10.3233/CBM-160469

Liang W, Wang F, Chen Q et al. (2019). Targeting cathepsin K diminishes prostate cancer establishment and growth in murine bone. J canc res clinic oncol, 145(8), 1999–2012. doi:10.1007/s00432-019-02950-y

Liao YJ , Lin M-W , Yen CH (2013). Characterization of Niemann-Pick Type C2 Protein Expression in Multiple Cancers Using a Novel NPC2 Monoclonal Antibody. PloS One, (10): e77586. doi:10.1371/journal.pone.0077586

Liu X, Wu C, Li C, Boerwinkle E (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Hum Mutat, 37(3), 235-241. doi:10.1002/humu.229326.

Lyons, JC, Scheithauer, BW, Ginsburg, WW (1982). Gaucher's Disease and Glioblastoma Multiforme in Two Siblings. J Neuropathol Exp, 41(1), 45–53. https://doi.org/10.1097/00005072-198201000-00005

Makoukji J, Raad M, Genadry K, El-Sitt S et al. (2015). Association between CLN3 (Neuronal Ceroid Lipofuscinosis, CLN3 Type) Gene Expression and Clinical Characteristics of Breast Cancer Patients. Front Oncol, 5, 215. doi:10.3389/fonc.2015.00215

Malecki, M., Dahlke, J., Haig, M., Wohlwend, L., & Malecki, R. (2013). Eradication of Human Ovarian Cancer Cells by Transgenic Expression of Recombinant DNASE1, DNASE1L3, DNASE2, and DFFB Controlled by EGFR Promoter: Novel Strategy for Targeted Therapy of Cancer. Journal of genetic syndromes & gene therapy, 4(6), 152. https://doi.org/10.4172/2157-7412.1000152

Mangogna, A., Belmonte, B., Agostinis, C., Zacchi, P., Iacopino, D. G., Martorana, A., Rodolico, V., Bonazza, D., Zanconati, F., Kishore, U., & Bulla, R. (2019). Prognostic Implications of the Complement Protein C1q in Gliomas. Frontiers in immunology, 10, 2366. https://doi.org/10.3389/fimmu.2019.02366

Masaki Narita, Naoyuki Taniguchi, Akira Makita, Tetsuo Kodama, Eiji Araki and Kiyoshi Oikawa (1983). Elevated Activity of β-Hexosaminidase and Sulfhydryl Modification in the B-Variant of Human Lung Cancer. Cancer Res, (43) (10) 5037-5042

Mayakonda, A, Lin, DC, Assenov, Y, Plass, C, Koeffler, HP (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genom Res, 28(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

McAtee CO, Barycki JJ, Simpson MA (2014). Emerging roles for hyaluronidase in cancer metastasis and therapy. Adv Cancer Res, 123, 1–34. doi:10.1016/B978-0-12-800092-2.00001-0

McAtee CO, Berkebile AR, Elowsky CG, Fangman et al. (2015). Hyaluronidase Hyal1 Increases Tumor Cell Proliferation and Motility through Accelerated Vesicle Trafficking. J Biol Chem, 290(21), 13144–13156. doi:10.1074/jbc.M115.647446

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, … Cunningham F (2016). The Ensembl Variant Effect Predictor. Gen biol, 17(1), 122. doi:10.1186/s13059-016-0974-4.

Minchenko, O. H., Riabovol, O. O., Halkin, O. V., Minchenko, D. O., & Ratushna, O. O. (2017). IRE1 knockdown modifies hypoxic regulation of cathepsins and LONP1 genes expression in u87 glioma cells. The Ukrainian Biochemical Journal, 89(2), 55–69. https://doi.org/10.15407/ubj89.02.055

Mistry PK, Taddei T, vom Dahl S, Rosenbloom BE (2013). Gaucher disease and malignancy: a model for cancer pathogenesis in an inborn error of metabolism. Crit rev oncogen, 18(3), 235–246. doi:10.1615/critrevoncog.2013006145

Morell, C, Bort A, Vara-Ciruelos D, Ramos-Torres Á, Altamirano-Dimas M, Díaz-Laviada I, Rodríguez-Henche N (2016). Up-Regulated Expression of LAMP2 and Autophagy Activity during Neuroendocrine Differentiation of Prostate Cancer LNCaP Cells. PloS one, 11(9), e0162977. doi:10.1371/journal.pone.0162977

Morelli MB, Nabissi M, Amantini C et al. (2016). Overexpression of transient receptor potential mucolipin-2 ion channels in gliomas: role in tumor growth and progression. Oncotarget, 7(28), 43654–43668. doi:10.18632/oncotarget.9661

Mounir, Mohamed, Lucchetta, Marta, Silva, C T, Olsen, Catharina, Bontempi, Gianluca, Chen, Xi, Noushmehr, Houtan, Colaprico, Antonio, Papaleo, Elena (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS comp biol, 15(3), e1006701.

Nagarajan, A, Malvi, P, Wajapeyee, N (2018). Heparan Sulfate and Heparan Sulfate Proteoglycans in Cancer Initiation and Progression. Frontiers in Endocrinology, 9. https://doi.org/10.3389/fendo.2018.00483

Nakamura T, Nakatsu N, Yoshida Y, Yamazaki K, Dan S, Sadahiro S, Makuuchi H, Yamori T (2009). Identification of candidate genes determining chemosensitivity to anticancer drugs of gastric cancer cell lines. Biol Pharm Bull, 32(11):1936-9

National Cancer Institute (2014) Tumor Grade, accessed 29 January, 2019 < https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet >.

Parenti, G; Andria, G; Ballabio, A (2015). Lysosomal storage diseases: from pathophysiology to therapy. Annu Rev Med, 66:471-86. doi: 10.1146/annurev-med-122313-085916.

Park, S. Y., & Guo, X. (2014). Adaptor protein complexes and intracellular transport. Bioscience Reports, 34(4). https://doi.org/10.1042/bsr20140069

Pastores, GM; Hughes, DA (2017). Lysosomal Storage Disorders and Malignancy. Diseases, 5(1):8.

Perry A, Wesseling P (2016). Histologic classification of gliomas. Histologic classification of gliomas. Handb Clin Neurol, 134:71-95. doi: 10.1016/B978-0-12-802997-8.00005-0.

Picco, R., Tomasella, A., Fogolari, F., & Brancolini, C. (2014). Transcriptomic analysis unveils correlations between regulative apoptotic caspases and genes of cholesterol homeostasis in human brain. PloS one, 9(10), e110610. https://doi.org/10.1371/journal.pone.0110610

R Core Team (2017). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rebecca VW, Nicastri MC, Fennelly C et al. (2019). PPT1 Promotes Tumor Growth and Is the Molecular Target of Chloroquine Derivatives in Cancer. Cancer Discov, 9(2):220-229. doi: 10.1158/2159-8290.CD-18-0706

Rentzsch P, Witten DM, Cooper GM, Shendure J, Kircher M (2018). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research, 47:D886–D894. doi:10.1093/nar/gky1016.

Rosenbloom BE, Weinreb NJ, Zimran A, Kacena KA, Charrow J, Ward E (2005). Gaucher disease and cancer incidence: a study from the Gaucher Registry. Blood, 105 (12): 4569-4572, doi: https://doi.org/10.1182/blood-2004-12-4672

Rylova SN, Amalfitano A, Persaud-Sawin DA, Guo WX, Chang J, Jansen PJ, Proia AD, Boustany RM (2002) The CLN3 gene is a novel molecular target for cancer drug discovery. Cancer Res, 62(3), p801-808

Saburi E, Tavakol-Afshari J, Biglari S, Mortazavi Y (2017). Is α-N-acetylgalactosaminidase the key to curing cancer? A mini-review and hypothesis. J BUON. 2017 Nov-Dec;22(6):1372-1377

Sänger N, Ruckhäberle E, Györffy B et al. (2015). Acid ceramidase is associated with an improved prognosis in both DCIS and invasive breast cancer. Mol oncol, 9(1), 58–67. doi:10.1016/j.molonc.2014.07.016

Saudubray, JM; Garcia-Cazorla, A (2018). Inborn Errors of Metabolism Overview Pathophysiology, Manifestations,Evaluation, and Management. Pediatr Clin N Am, 65:179-208.

Schaaf MB, Houbaert D, Meçe O, To SK, Ganne M, Maes H, Agostinis P (2019). Lysosomal Pathways and Autophagy Distinctively Control Endothelial Cell Behavior to Affect Tumor Vasculature. Front Oncol, 9, 171. doi:10.3389/fonc.2019.00171

Shambhavi A, Salian S, Shah H et al. (2018). Pycnodysostosis: Novel Variants in CTSK and Occurrence of Giant Cell Tumor. J Pediatr Genet, 07(01), p.009-013

Shin J, Kim D, Kim HL, Choi M, Koh Y, Yoon SS (2019). Oncogenic effects of germline variants in lysosomal storage disease genes. Genet Med, Dec;21(12):2695-2705. doi: 10.1038/s41436-019-0588-9

Shin J, Kim G, Lee JW, Lee et al. (2016). Identification of ganglioside GM2 activator playing a role in cancer cell migration through proteomic analysis of breast cancer secretomes. Canc scienc, 107(6), 828–835. doi:10.1111/cas.12935

Silva, C T, Colaprico, Antonio, Olsen, Catharina, D'Angelo, Fulvio, Bontempi, Gianluca, Ceccarelli, Michele, Noushmehr, Houtan (2016). TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research, 5.

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic acids research, 40(Web Server issue), W452–W457. https://doi.org/10.1093/nar/gks539

Singh V, Jha KK, M JK, Kumar RV, Raghunathan V, Bhat R (2019). Iduronate-2-Sulfatase-Regulated Dermatan Sulfate Levels Potentiate the Invasion of Breast Cancer Epithelia through Collagen Matrix. J Clin Med, 8(10), 1562. doi:10.3390/jcm8101562

Staege MS, Hesse M, Max D (2010). Lipases and Related Molecules in Cancer. Cancer Growth Metastasis, doi: https://doi.org/10.4137/CGM.S2816

Tan JX, Wang XY, Su XL, Li HY, Shi Y, Wang L, Ren GS (2011). Upregulation of HYAL1 expression in breast cancer promoted tumor cell proliferation, migration, invasion and angiogenesis. PloS one, 6(7), e22836. doi:10.1371/journal.pone.0022836

Tanemura M, Miyoshi E, Nagano H et al. (2015). Cancer immunotherapy for pancreatic cancer utilizing α-gal epitope/natural anti-Gal antibody reaction. World J Gastroenterol, 21(40), 11396–11410. doi:10.3748/wjg.v21.i40.11396

Teo W, Sekar K, …, Seshachalam, P (2019). Relevance of a TCGA-derived Glioblastoma Subtype Gene-Classifier among Patient Populations. Sci Rep, 9:7442 doi:10.1038/s41598-019-43173-y

Terrasi, A., Bertolini, I., Martelli, C., Gaudioso, G., Di Cristofori, A., Storaci, A. M., Formica, M., Bosari, S., Caroli, M., Ottobrini, L., Vaccari, T., & Vaira, V. (2019). Specific V-ATPase expression sub-classifies IDHwt lower-grade gliomas and impacts glioma growth in vivo. EBioMedicine, 41, 214–224. https://doi.org/10.1016/j.ebiom.2019.01.052

Thaker, N. G., Zhang, F., McDonald, P. R., Shun, T. Y., Lewen, M. D., Pollack, I. F., & Lazo, J. S. (2009). Identification of survival genes in human glioblastoma cells by small interfering RNA screening. Molecular Pharmacology, 76(6), 1246–1255. https://doi.org/10.1124/mol.109.058024

Thurberg BL, Germain DP, Perretta F, Jurca-Simina IE, Politei JM (2016). Fabry disease: Four case reports of meningioma and a review of the literature on other malignancies. Mol Genet Metab Rep, 11, 75–80. doi:10.1016/j.ymgmr.2016.09.005

Tong, L., Yi, L., Liu, P., Abeysekera, I., Hai, L., Li, T., Tao, Z., Ma, H., Xie, Y., Huang, Y., Yu, S., Li, J., Yuan, F., & Yang, X. (2018). Tumor cell dormancy is a contributor to the reduced survival of GBM patients who received standard therapy. Oncology Reports. https://doi.org/10.3892/or.2018.6425

Towers, CG, & Thorburn, A (2017). Targeting the Lysosome for Cancer Therapy. Cancer Discovery, 7(11), 1218–1220. https://doi.org/10.1158/2159-8290.cd-17-0996
Tran VM, Wade A, McKinney A, Chen K, Lindberg OR, Engler JR, Phillips JJ (2017). Heparan Sulfate Glycosaminoglycans in Glioblastoma Promote Tumor Invasion. Mol Canc Res, 15(11), 1623–1633. doi:10.1158/1541-7786.MCR-17-0352

Trejo-Solís, C; Serrano-Garcia, N; Escamilla-Ramírez, Á; Castillo-Rodríguez, RA; Jimenez-Farfan, D; Palencia, G; … Sotelo, J (2018). Autophagic and Apoptotic Pathways as Targets for Chemotherapy in Glioblastoma. Intern J Mol Sci, 19(12), 3773. doi:10.3390/ijms19123773

Tsigelny, I. F., Kouznetsova, V. L., Jiang, P., Pingle, S. C., & Kesari, S. (2015). Hierarchical control of coherent gene clusters defines the molecular mechanisms of glioblastoma. Molecular BioSystems, 11(4), 1012–1028. https://doi.org/10.1039/c5mb00007f

Tsuchida N, Ikeda MA, Ishino Y, Grieco M, Vecchio G (2017). FUCA1 is induced by wild-type p53 and expressed at different levels in thyroid cancers depending on p53 status. Int J Oncol, 50(6):2043-2048. doi: 10.3892/ijo.2017.3968.

Uhlen, M., Zhang, C., Lee, … Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. Science, 357(6352), eaan2507. https://doi.org/10.1126/science.aan2507

Ujifuku, K., Mitsutake, N., Takakura, S., Matsuse, M., Saenko, V., Suzuki, K., Hayashi, K., Matsuo, T., Kamada, K., Nagata, I., & Yamashita, S. (2010). miR-195, miR-455-3p, and miR-10a(*) are implicated in acquired temozolomide resistance in glioblastoma multiforme cells. Cancer letters, 296(2), 241–248. https://doi.org/10.1016/j.canlet.2010.04.013

Van Rappard DF et al. (2016). Gallbladder and the risk of polyps and carcinoma in metachromatic leukodystrophy. Neurology, 87 (1) 103-111; DOI: 10.1212/WNL.0000000000002811

Vijayan Y, Lankadasari MB, Harikumar KB (2019). Acid Ceramidase: A Novel Therapeutic Target in Cancer. Curr Top Med Chem, 19(17):1512-1520. doi: 10.2174/1568026619666190227222930.

Ville, D., Mireskandari, K., … Thevenon, J. (2016). Autosomal-Recessive Mutations in AP3B2, Adaptor-Related Protein Complex 3 Beta 2 Subunit, Cause an Early-Onset Epileptic Encephalopathy with Optic Atrophy. American journal of human genetics, 99(6), 1368–1376. https://doi.org/10.1016/j.ajhg.2016.10.009

Vittner EB, Dekel H, Zigdon H, Shachar T, Farfel-Becker T, Eilam R, Karlsson S, Futerman AH (2010). Altered expression and distribution of cathepsins in neuronopathic forms of Gaucher disease and other sphingolipidoses. Hum Mol Genet, 19: 3583-3590.

Wade, A., Engler, J. R., Tran, V. M., & Phillips, J. J. (2015). Measuring sulfatase expression and invasion in glioblastoma. Methods in molecular biology (Clifton, N.J.), 1229, 507–516. https://doi.org/10.1007/978-1-4939-1714-3_39

Wagner J, Damaschke N, Yang B et al. (2015). Overexpression of the novel senescence marker β-galactosidase (GLB1) in prostate cancer predicts reduced PSA recurrence. PloS one, 10(4), e0124366. doi:10.1371/journal.pone.0124366

Wang, L., Zhang, C., Zhang, Z., Han, B., Shen, Z., Li, L., Liu, S., Zhao, X., Ye, F., & Zhang, Y. (2018). Specific clinical and immune features of CD68 in glioma via 1,024 samples. Cancer management and research, 10, 6409–6419. https://doi.org/10.2147/CMAR.S183293

Wesseling P, Capper D (2018). WHO 2016 Classification of gliomas. Neuropathol Appl Neurobiol, 44(2):139-150. doi: 10.1111/nan.12432.

Witzel I, Marx AK, Müller V et al. (2017). Role of HYAL1 expression in primary breast cancer in the formation of brain metastases. Breast Cancer Res Treat. Apr;162(3):427-438. doi: 10.1007/s10549-017-4135-6

Xiong A, Kundu S, Forsberg-Nilsson K (2014). Heparan sulfate in the regulation of neural differentiation and glioma development. FEBS J, 281(22):4993-5008. doi: 10.1111/febs.13097. Epub 2014 Nov 6.

Xu Y, Wang H, ZEng, Y et al. (2019). Overexpression of CLN3 contributes to tumor lysosomal-related progression and predicts poor prognosis in hepatocellular carcinoma. Surg Oncol, 28, p.180-189, doi: https://doi.org/10.1016/j.suronc.2018.12.003

Xu, H., Ma, Y., Zhang, Y., Pan, Z., Lu, Y., Liu, P., & Lu, B. (2016). Identification of Cathepsin K in the Peritoneal Metastasis of Ovarian Carcinoma Using In-silico, Gene Expression Analysis. Journal of Cancer, 7(6), 722–729. https://doi.org/10.7150/jca.14277

Yagi Y, Machida A, Toru S, Kobayashi T, Uchihara T (2011). Sialidosis type I with neoplasms in siblings: the first clinical cases. Neurol Sci, 32(4):737-8. doi: 10.1007/s10072-010-0392-4

Yan C, Zhao T, Du H (2015). Lysosomal acid lipase in cancer. Oncoscience, 2(9), 727–728. doi:10.18632/oncoscience.223

Yao-Hsien T, Yu-Tse T, Wei-Cheng C, Pau-Chung C (2015). Use of an α-Glucosidase Inhibitor and the Risk of Colorectal Cancer in Patients With Diabetes: A Nationwide, Population-Based Cohort Study. Diabetes Care, 38(11): p.2068-2074, doi: https://doi.org/10.2337/dc15-0563

Yu, G, Wang, LG, Han, Y, He, QY (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang C, Zhang M, Song S (2018). Cathepsin D enhances breast cancer invasion and metastasis through promoting hepsin ubiquitin-proteasome degradation. Cancer Lett; 438:105-115. doi: 10.1016/j.canlet.2018.09.021

Zhang, J., Yao, Y., He, H., Shen, J. (2020). Clinical Interpretation of Sequence Variants. Current protocols in human genetics, 106(1), e98. https://doi.org/10.1002/cphg.98

Zhao Y, Guo Y, Wang Z et al. (2015). GALC gene is downregulated by promoter hypermethylation in Epstein-Barr virus-associated nasopharyngeal carcinoma. Oncol Rep, 34(3), p.1369-78. doi: 10.3892/or.2015.4134

Zhao Y, Wang Y, Lou H, Shan L (2017). Alpha-glucosidase inhibitors and risk of cancer in patients with diabetes mellitus: a systematic review and meta-analysis. Oncotarget, 8(46), 81027–81039. doi:10.18632/oncotarget.17515

Zheng H, Yang Y, Ye C, Li P, Wang Z, Xing H, Ren H, Zhou W (2018). LAMP2 inhibits epithelial-mesenchymal transition by suppressing Snail expression in HCC. Oncotarget, 9: 30240-30252, doi: https://doi.org/10.18632/oncotarget.25367

Zou J, Sun R, Xia J et al. (2018). The role of telomere protective protein TPP1 in hepatocellular carcinoma. Mol Cell Biol, DOI: 10.1158/1538-7445.AM2018-341

Zwerschke W, Mannhardt B, Massimi P et al. (2000). Allosteric Activation of Acid a-Glucosidase by the Human Papillomavirus E7 Protein. J biologic chem, 275:13,p. 9534–9541

Zong, H, Verhaak, RG, Canoll, P (2012). The cellular origin for malignant glioma and prospects for clinical advancements. Expert review of molecular diagnostics, 12(4), 383–394, DOI: 10.1586/erm.12.30

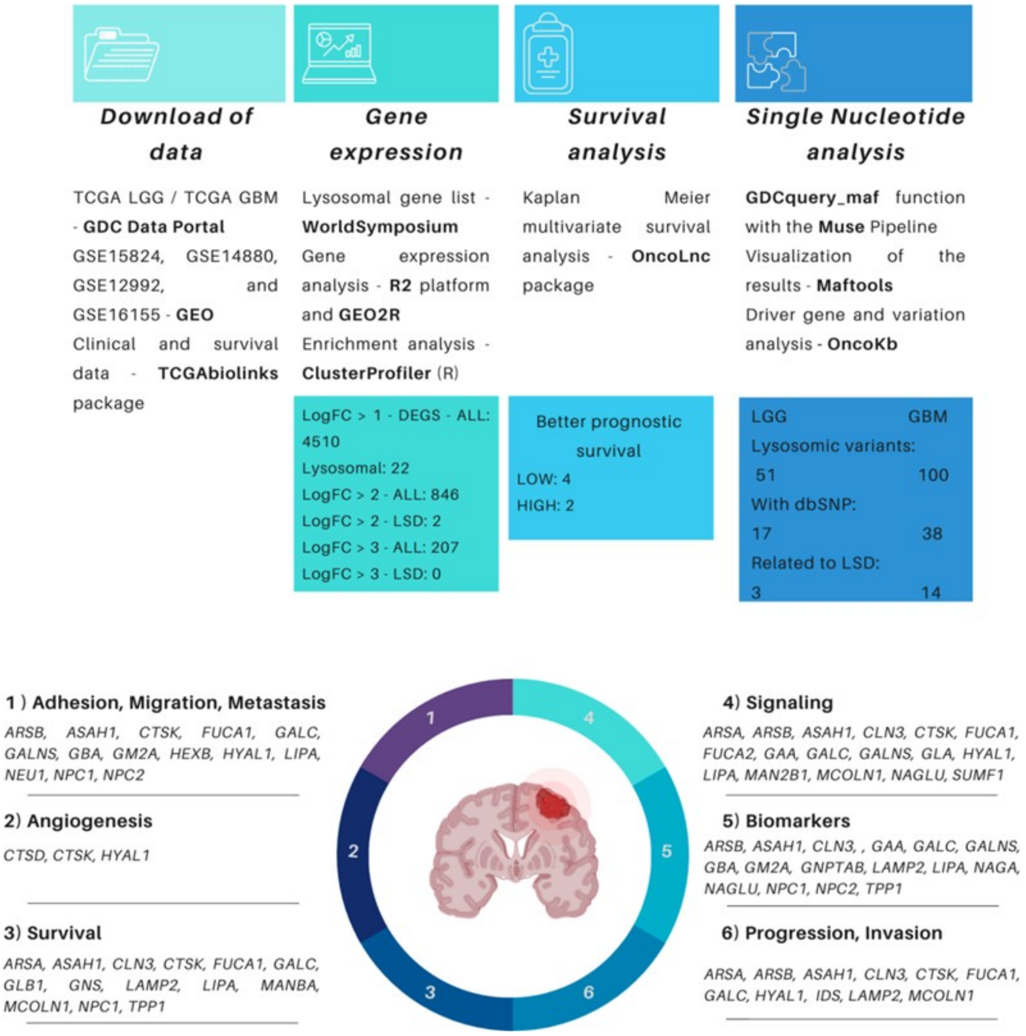

Download of data
TCGA LGG / TCGA GBM - GDC Data Portal
GSE15824, GSE14880, GSE12992, and GSE16155 - GEO
Clinical and survival data - TCGAbiolinks package
Gene expression
Lysosomal gene list - WorldSymposium
Gene expression analysis - R2 platform and GEO2R
Enrichment analysis - ClusterProfiler (R)
LogFC > 1 - DEGS - ALL: 4510
Lysosomal: 22
LogFC > 2 - ALL: 846
LogFC > 2 - LSD: 2
LogFC > 3 - ALL: 207
LogFC > 3 - LSD: 0
Survival analysis
Kaplan Meier multivariate survival analysis - OncoLnc package
Better prognostic survival
LOW: 4
HIGH: 2
Single Nucleotide analysis
GDCquery_maf function with the Muse Pipeline
Visualization of the results - Maftools
Driver gene and variation analysis - OncoKb
LGG GBM
Lysosomic variants:
51 100
With dbSNP:
17 38
Related to LSD:
3 14
1) Adhesion, Migration, Metastasis
ARSB, ASAH1, CTSK, FUCA1, GALC, GALNS, GBA, GM2A, HEXB, HYAL1, LIPA, NEU1, NPC1, NPC2
2) Angiogenesis
CTSD, CTSK, HYAL1
3) Survival
ARSA, ASAH1, CLN3, CTSK, FUCA1, GALC, GLB1, GNS, LAMP2, LIPA, MANBA, MCOLN1, NPC1, TPP1
4) Signaling
ARSA, ARSB, ASAH1, CLN3, CTSK, FUCA1, FUCA2, GAA, GALC, GALNS, GLA, HYAL1, LIPA, MAN2B1, MCOLN1, NAGLU, SUMF1
5) Biomarkers
ARSB, ASAH1, CLN3, , GAA, GALC, GALNS, GBA, GM2A, GNPTAB, LAMP2, LIPA, NAGA, NAGLU, NPC1, NPC2, TPP1
6) Progression, Invasion
ARSA, ARSB, ASAH1, CLN3, CTSK, FUCA1, GALC, HYAL1, IDS, LAMP2, MCOLN1
1
2
3
4
5
6

Figure 1: Workflow of the analysis, and lysosomal related genes with their respective roles in the carcinogenic processes. The respective references of the review of lysosomal-related genes and cancer are shown at supplementary table 5.

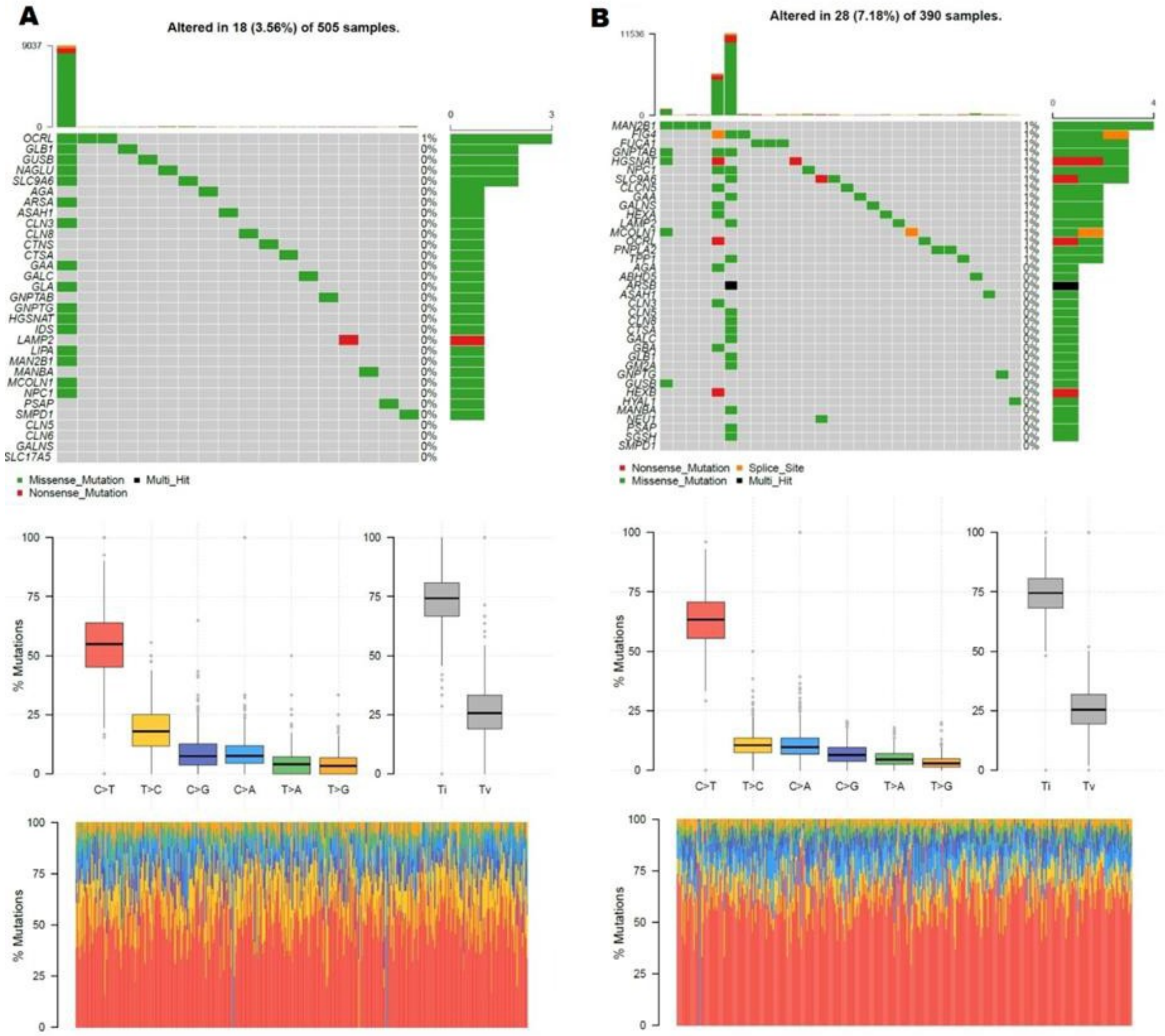


Figure 2: Summary of Maftools variant prioritization. 2A: LGG oncoplot and the number of transitions and transversions found. 2B: Oncoplot of GBM cohort and the respective number of transitions and transversions. In the top, the green color corresponds to the missense variants, the red is the nonsense, and black are multi hit when the same gene has more than one type of variant found. The bottom of the figure shows the percentage of transitions and transversions across the samples.

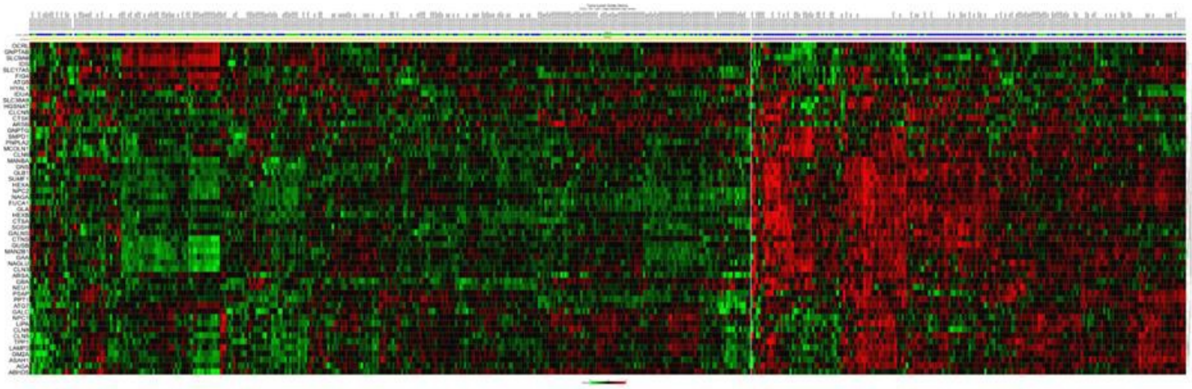


Figure 3: Heatmap of Lower Grade Glioma dataset. K-means analysis has divided the barcodes into multi-hit non-determined 2 groups: yellow and purple In total, 516 samples were clustered. Red = upregulated genes. Green = downregulated genes. The first track represents the tumor grade, the blue represents astrocytoma samples, green represents oligodendroglioma samples, red represents oligoastrocytoma samples, and white the non-determined samples.

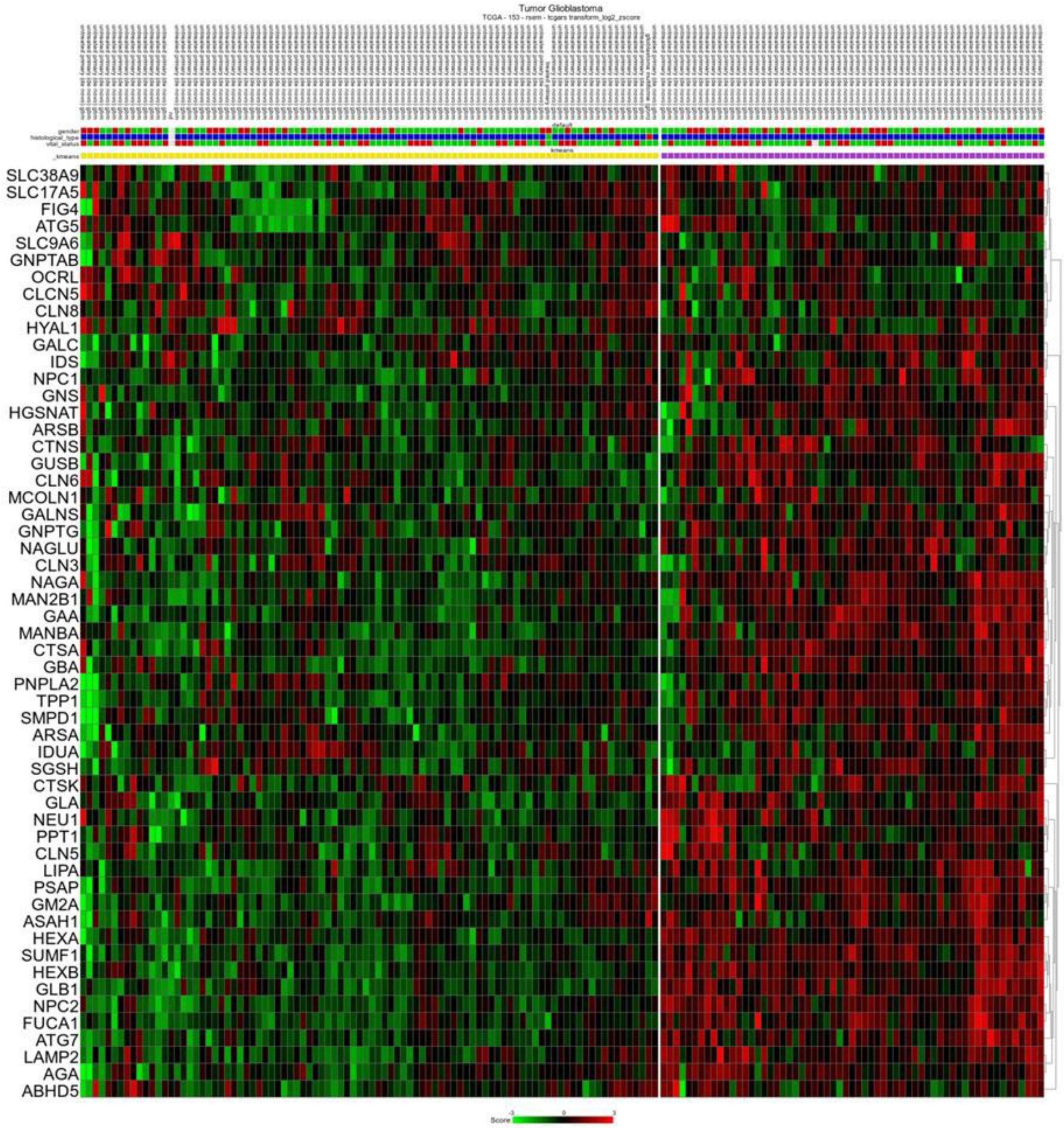


Figure 4: Heatmap of Glioblastoma multiforme dataset. K-means analysis has divided the barcodes into 2 groups: yellow and purple. In total, 153 samples were clustered. Red = upregulated genes. Green = downregulated genes. The first track represents the gender composition of clusters, when red squares are female, and green male. The second track corresponds to the histological type when all the samples are classified by glioblastoma multiforme. The third track corresponds to the vital status when red is alive, and the green represents dead patients.

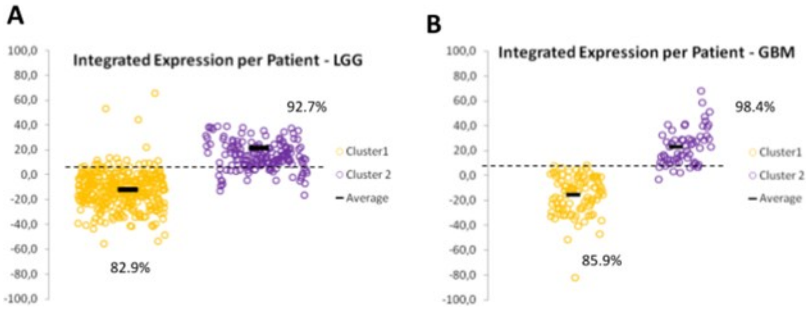


Figure 5: Integrative lysosomal-related gene expression. We summarize the expression levels of all lysosomal genes for each patient, generating a global lysosomal gene expression factor. 5A: LGG patients from cluster 1 and cluster 2. 5B: GBM patients - cluster 1 and cluster 2. The analysis shows two profiles of expression considering lysosomal genes in both LGG and GBM patients.

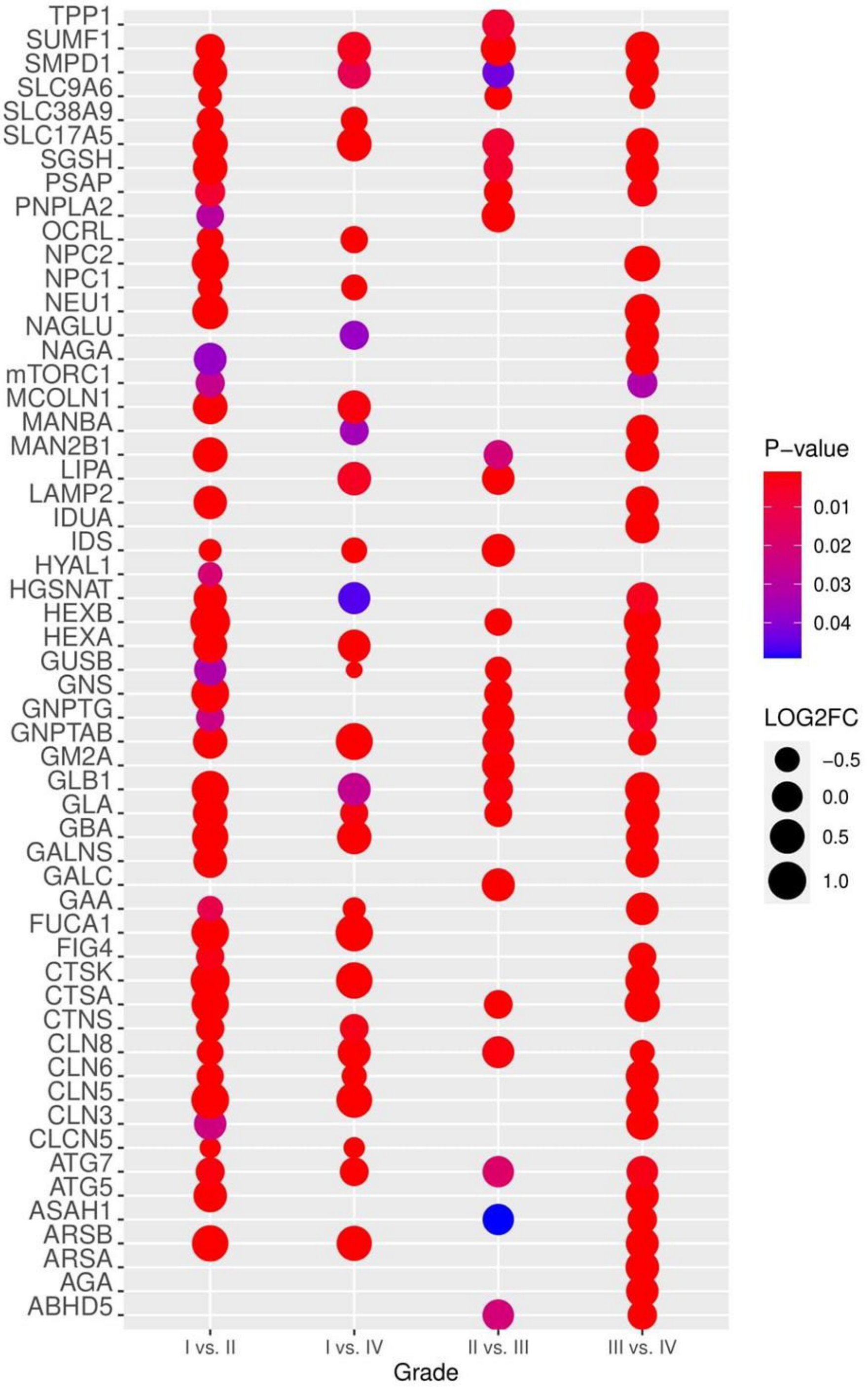


Figure 6: Lysosomal-related genes and progression analysis. Here we evaluated whether the expression of the lysosomal genes impacts the tumor progression in different tumor grades.

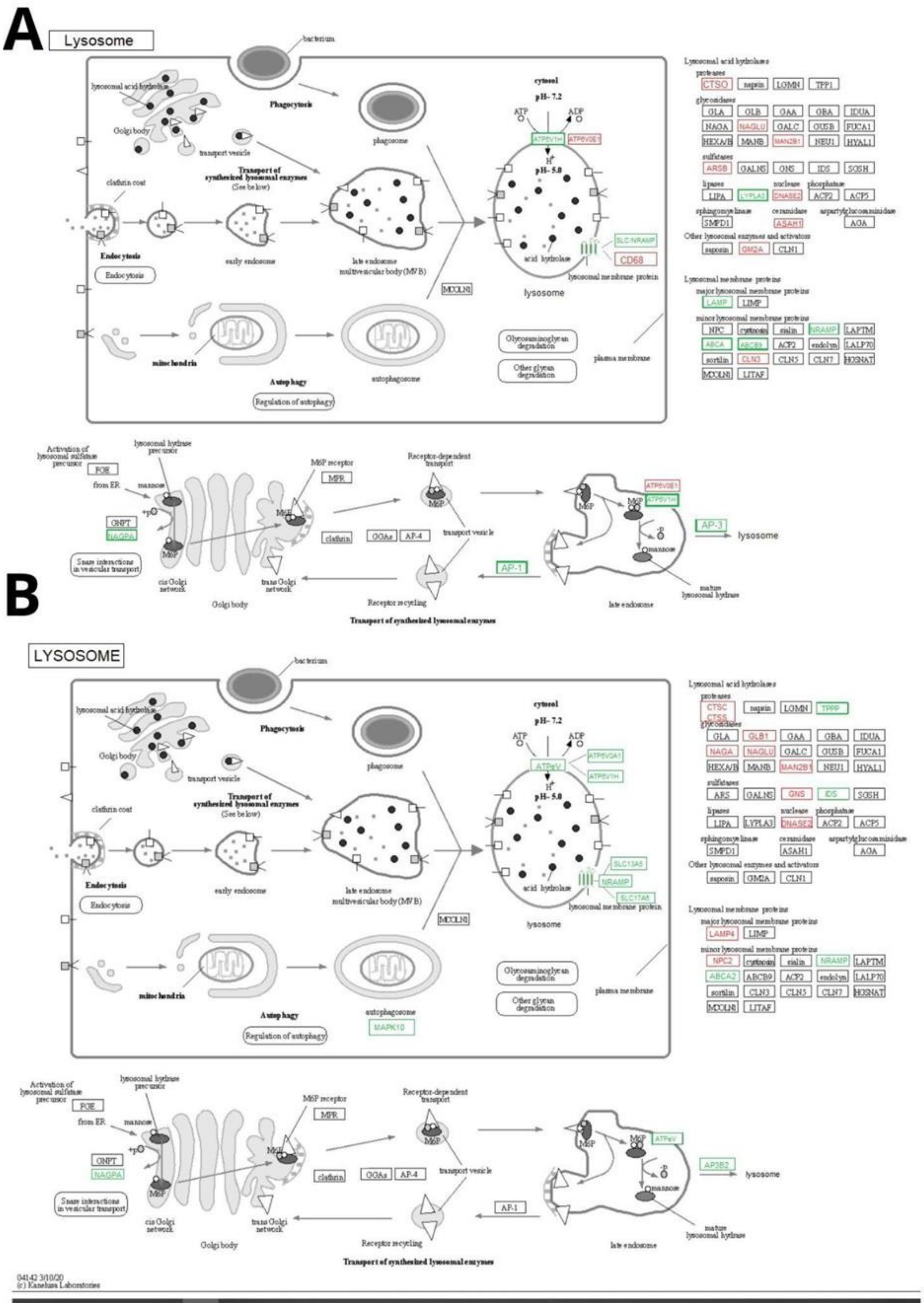


Figure 7: KEGG pathway lysosome and the top 10 down and up-regulated genes. 7A: Lower Grade Glioma lysosome pathway. 7B: Glioblastoma multiforme (GBM) lysosome pathway. Green = down-regulated genes. Red = up-regulated genes. (Adapted of https://www.genome.jp/kegg/).

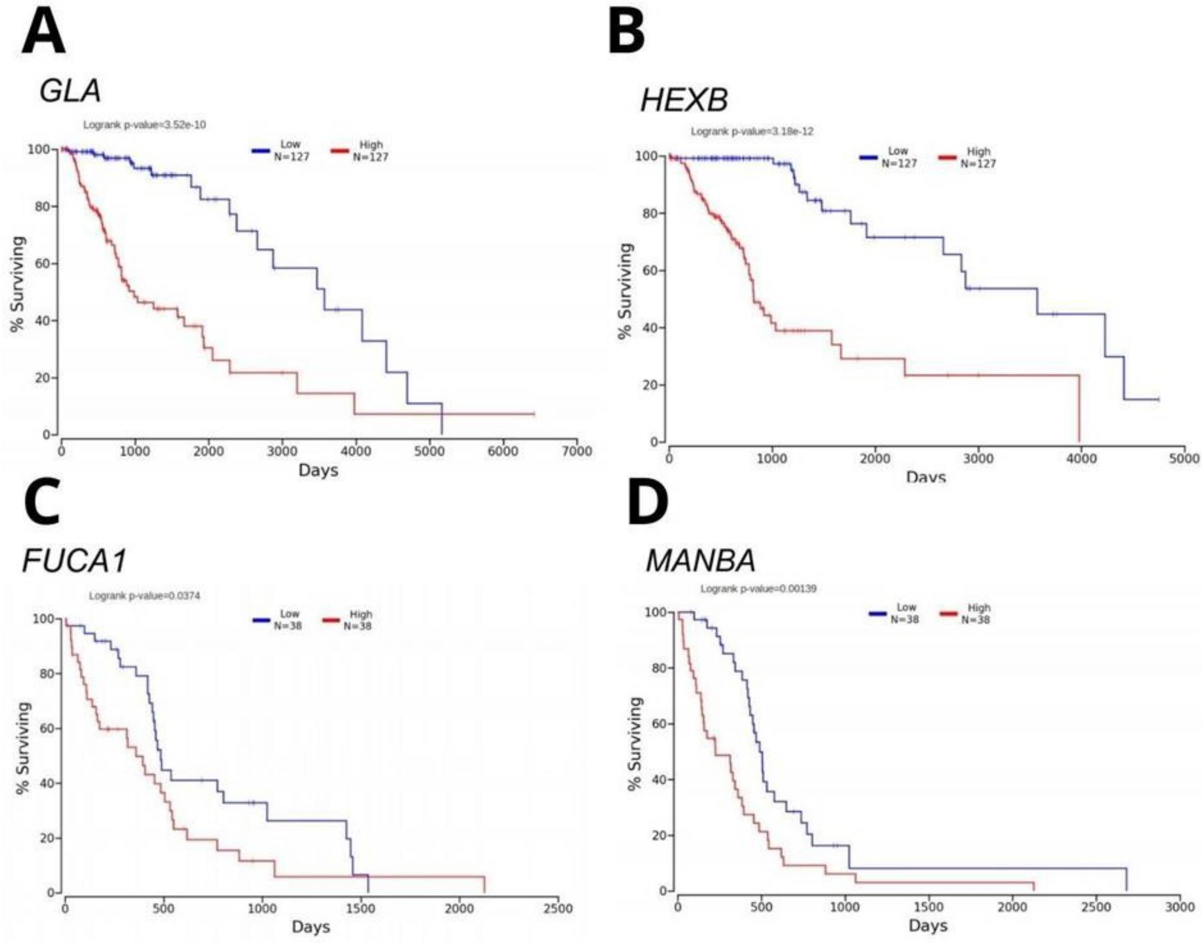


Figure 8: Kaplan-Meier plot of LGG and GBM datasets. 8A: The Galactosidase alpha gene (GLA) was a top p-value in the LGG cohort. In this case, the high expression of the gene increases the survival time of patients with Lower-grade glioma. 8B: The Hexosaminidase Subunit Beta gene (HEXB) was a top p-value in the LGG cohort. In the y-axis, we observed the percentage of survival and in the x-axis the days before the death. In this case, the low expression of the gene increases the survival time of patients with Lower-grade glioma. 8C: The Alpha-L-Fucosidase 1 gene (FUCA1) was a top p-value in the cohort. In the y-axis, we observed the percentage of survival and in the x-axis the days before the dead. In this case, the high expression of the gene increases the survival time of patients with Glioblastoma multiforme. 8D: The Mannosidase Beta gene (MANBA) was a top p-value in the cohort. In the y-axis, we observed the percentage of survival and in the x-axis the days before the dead. In this case, the lower expression of the gene increases the survival time of patients with Glioblastoma multiforme.

Capítulo 6

Vias de sinalização oncogênica em Mucopolissacaridoses

Manuscrito publicado no *Lecture Notes in Bioinformatics*

Silva G.C.V., Soares L.D.F., Matte U. (2020) Oncogenic Signaling Pathways in Mucopolysaccharidoses. In: Setubal J.C., Silva W.M. (eds) Advances in Bioinformatics and Computational Biology. BSB 2020. Lecture Notes in Computer Science, vol 12558. Springer, Cham. https://doi.org/10.1007/978-3-030-65775-8_24

Oncogenic Signaling Pathways in Mucopolysaccharidoses

Gerda Cristal Villalba Silva[1,2,3], Luis Dias Ferreira Soares[3,5], and Ursula Matte[1,2,3,4](✉)

[1] Postgraduate Program in Genetics and Molecular Biology, UFRGS,
Porto Alegre 91501970, Brazil
umatte@hcpa.edu.br
[2] Gene Therapy Center, HCPA, Porto Alegre 90035903, Brazil
[3] Bioinformatics Core, HCPA, Porto Alegre 90035903, Brazil
[4] Department of Genetics, UFRGS, Porto Alegre 91501970, Brazil
[5] Graduation Program on Biotechnology/Bioinformatics, UFRGS,
Porto Alegre 91501-970, Brazil

Abstract. Cancer cells depend on several signaling pathways and organelles, such as the lysosomes. Defects in the activity of lysosomal hydrolases involved in glycosaminoglycan degradation lead to a group of lysosomal storage diseases called Mucopolysaccharidoses (MPS). In MPS, secondary cell disturbance affects pathways common to cancer. This work aims to identify oncogenic pathways related to cancer in the different MPS datasets available in public databases and compare the ontologies across the different types of MPS. For this, we used 12 expression datasets of 6 types of MPS. Statistical analysis was based on hypergeometric distribution followed by FDR correction. We found several enriched pathways across the 12 MPS studies, among being 57.65% were KEGG pathways, 32.5% of GO Biological Process, 2.5% GO Celular Component, and 7.35% GO Molecular Function. Hippo signaling pathway and MAPK signaling pathway appear in all datasets. Proteoglycans in cancer, Rap1 signaling pathway, and Cytokine-mediated signaling pathway appears in 11 of 12 datasets. The lysosome participates in several biological processes, like autophagy, cell adhesion and migration, and antigen presentation. These processes also may affect in several types of cancer and Lysosomal Storage Diseases. Studying the tumor ontology signature in lysosomal disorders may help understand lysosomal storage diseases and cancer's underlying mechanisms. This may help amplify therapeutic approaches for both types of diseases.

Keywords: Cancer pathways · Gene ontology · Lysosomal storage diseases

1 Introduction

Several metabolic pathways are deranged in cancer cells. The proliferation ability of tumors depends on a cascade of signaling pathways in several cancer cells' organelles, such as the lysosomes [1]. Lysosomes are cellular compartments responsible, among

J. C. Setubal and W. M. Silva (Eds.): BSB 2020, LNBI 12558, pp. 259–264, 2020.
https://doi.org/10.1007/978-3-030-65775-8_24

other functions, for the degradation of macromolecules through acid hydrolases contained within them. Defects in these enzymes culminate in the lysosomal accumulation of intermediate metabolites or macromolecules, known as lysosomal storage diseases [2]. Lysosomal Storage Diseases (LSD) are a group of more than 50 rare metabolic diseases, among which we can highlight the Mucopolysaccharidoses (MPS). In MPS, secondary cell disturbance affects pathways common to cancer.

This work aims to identify oncogenic pathways related to cancer in the different datasets of MPS available in public databases and to compare the ontologies across the different types of MPS.

2 Methods

Gene expression analysis considered 12 datasets available at GEO (https://www.ncbi.nlm.nih.gov/geo), from six different MPS types. For RNA-seq data, we used edgeR, and for microarray data, we used R packages according to the experiment's platform. Furthermore, the data present in this work are available in the MPSBase (https://www.ufrgs.br/mpsbase/). Statistical analysis was based on hypergeometric distribution followed by FDR correction. We perform the enrichment analysis in Cytoscape, with Bingo and ClueGo plugins. We search the child terms with QuickGo. We selected 12 datasets, being 2 of MPS I; 1 from MPS II; 1 from MPS IIIA; 3 MPS IIIB; 1 MPS VI; and 4 from MPS VII. These datasets comprise an RNA-seq data of Illumina HiSeq 2500 platform of human iPSC-derived Neuronal Stem Cell (MPS I, GSE111906); Agilent-021193 Canine (V2) microarray of Ascending Aorta, Descending Aorta and Carotid Aorta (MPS I, GSE78889); AB SOLiD 3 Plus System (Mus musculus) of Brain samples (MPS II, GSE95224); Agilent-028005 SurePrint G3 Mouse GE 8×60K Microarray of Brain and Blood samples (MPS IIIA, GSE97759); Agilent-012694 Whole Mouse Genome G4122A of Lateral entorhinal cortex and Medial entorhinal cortex (MPS IIIB, GSE15758); Affymetrix Human Exon 1.0 ST Array of iPSC-derived Neuronal Stem Cell (MPS IIIB, GSE23075); Affymetrix Human Exon 1.0 ST Array of HeLa depleting NAGLU (MPS IIIB, GSE32154); Affymetrix Mouse Gene 1.0 ST Array of ARSB null mouse hepatic cells (MPS VI, GSE77689); Illumina Mouse-8 Expression BeadChip of Descending aorta (MPS VII, GSE30657); Affymetrix Mouse Genome 430A 2.0 Array of six brain regions (MPS VII, GSE34071); Affymetrix Mouse Exon 1.0 ST Array of iPS embryo-derived ES cells with controls derived from B6 Blu ES cells and Mouse embryonic fibroblast (MPS VII, GSE36017); and Affymetrix Mouse Genome 430A 2.0 Array of hippocampus (MPS VII, GSE76283).

3 Results

We found 680 oncogenic enriched ontologies across the 12 MPS studies, among being 392 were KEGG pathways (57.65%), 221 GO Biological Process (32.5%), 17 GO Celular Component (2.5%), and 50 of GO Molecular Function (7.35%). Hippo signaling pathway and MAPK signaling pathway appears in all datasets. Proteoglycans in cancer, Rap1 signaling pathway, and Cytokine-mediated signaling pathway appears in 11 of 12 datasets (see Fig. 1).

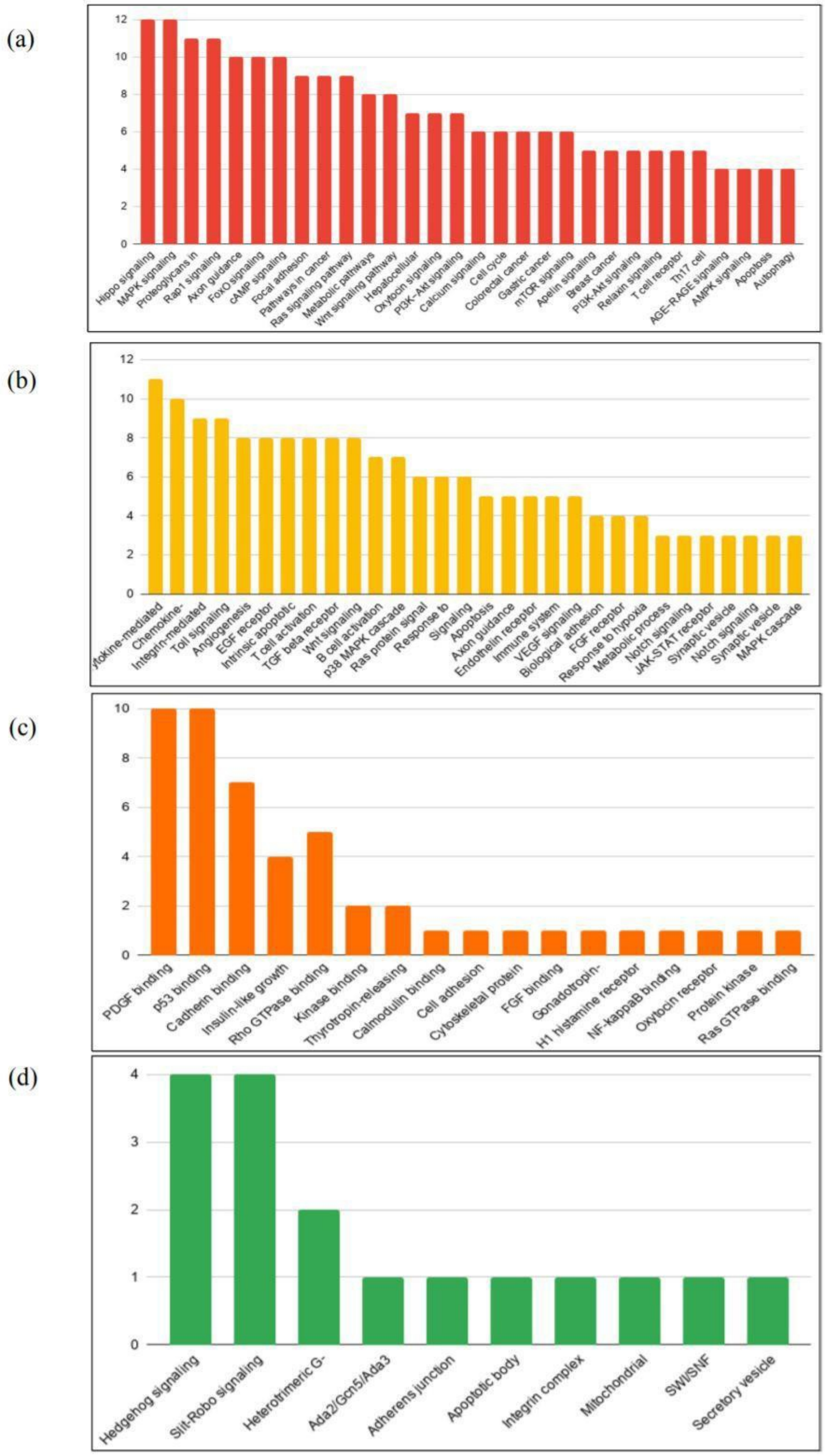


Fig. 1. Top Gene ontology of oncogenic terms of MPS. (a) KEGG pathways; (b) GO Biological Process; (c) Molecular Function; (d) Cellular Component.

Considering only the MPS types, the ontologies Axon guidance, Focal adhesion, Hippo signaling pathway, MAPK signaling pathway, Metabolic pathways, Pathways in cancer, PI3K-Akt signaling pathway, Proteoglycans in cancer, Rap1 signaling pathway, and Ras signaling pathway are present in all the MPS types found in the GEO. The following Table 1 gives a summary of the most frequent oncogenic ontologies according to the MPS type.

Table 1. Prevalent oncogenic enriched pathways of datasets analyzed. In bold, the ontology appears in all MPS types.

Term	Ontology	MPS I	MPS II	MPS IIIA	MPS IIIB	MPS VI	MPS VII
Apelin signaling pathway	KEGG	X	X	X			X
Apoptosis	KEGG	X		X	X	X	X
Autophagy	KEGG	X		X	X		X
Axon guidance	KEGG	X	X	X	X	X	X
Calcium signaling pathway	KEGG	X	X		X	X	X
cAMP signaling pathway	KEGG		X	X	X	X	X
Focal adhesion	KEGG	X	X	X	X	X	X
FoxO signaling pathway	KEGG	X	X	X	X		X
Hepatocellular carcinoma	KEGG	X	X	X			X
Hippo signaling pathway	KEGG	X	X	X	X	X	X
MAPK signaling pathway	KEGG	X	X	X	X	X	X
Metabolic pathways	KEGG	X	X	X	X	X	X
mTOR signaling pathway	KEGG	X		X	X	X	X
Oxytocin signaling pathway	KEGG	X	X	X	X	X	
Pathways in cancer	KEGG	X	X	X	X	X	X
PI3K-Akt signaling pathway	KEGG	X	X	X	X	X	X
Proteoglycans in cancer	KEGG	X	X	X	X	X	X
Rap1 signaling pathway	KEGG	X	X	X	X	X	X
Ras signaling pathway	KEGG	X	X	X	X	X	X
Wnt signaling pathway	KEGG	X	X	X	X	X	X
Chemokine-mediated signaling pathway	GO_BP	X		X	X	X	X
Cytokine-mediated signaling pathway	GO_BP	X		X	X	X	X
EGFR signaling pathway	GO_BP	X	X	X	X		X
Slit-Robo signaling complex	GO_CC			X	X	X	X
p53 binding	GO_MF			X	X	X	X

The dataset with the most enriched pathways is GSE32154 (MPS IIIB, *Homo sapiens*) with 90 ontologies (see Fig. 2). The dataset with the most enriched KEGG terms is GSE30657 (MPS VII, *Mus musculus*) with 60 KEGG terms. The GSE32154 (MPS IIIB, *Homo sapiens*) have the most GO Biological Process enriched terms, with 45 terms. For GO Cellular Component, the dataset with more enriched terms in this category is GSE32154 (MPS IIIB, *Homo sapiens*) with 5 terms. Lastly, in the GO Molecular Function, the GSE32154 is the dataset with more enriched terms (10 terms found).

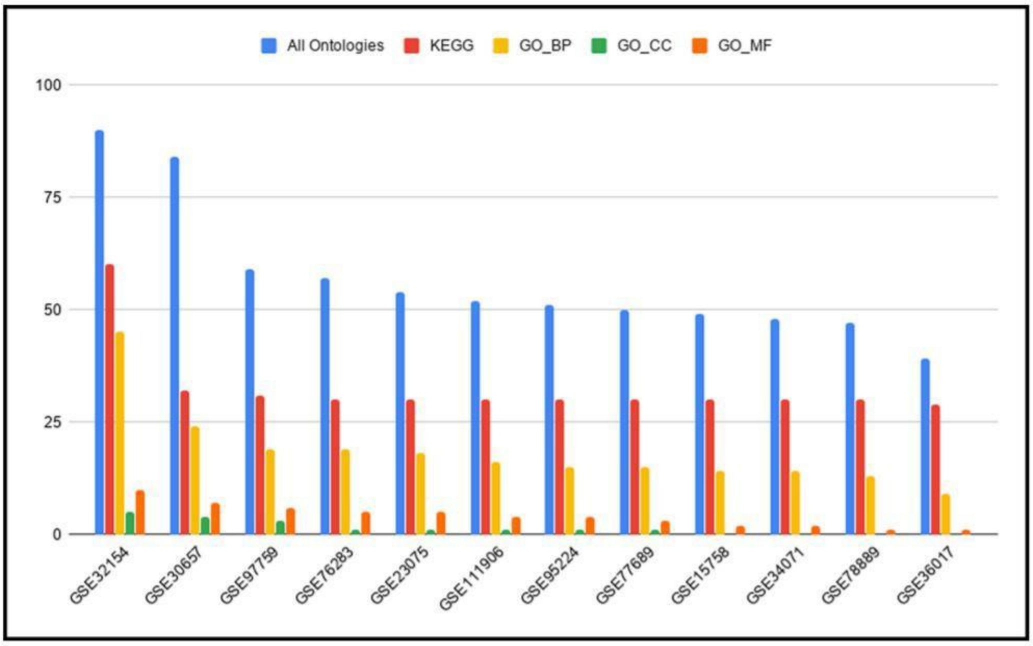


Fig. 2. Number of enriched cancer ontologies across the MPS datasets.

4 Discussion

Oncogenic activation can lead to the destabilization of lysosomal membranes and an increase of lysosomal hydrolases into the cytosol, where they can contribute to the demise of the cancer cell [3].

Axon guidance and Wnt signaling pathway are related to the neurological impairment found in several MPS types [4]. Alterations in autophagy are frequently found in MPS patients with neurodegenerative symptoms [4]. In cancer, autophagy are related to cancer initiation, proliferation, and survival [5].

The ontologies Cell cycle, Hippo, Notch, PI-3-Kinase/Akt, RAS, TGFβ signaling, P53 and β-catenin/WNT signaling pathway are considered canonical oncogenic pathways. Unfortunately, 89% of tumors found in the TCGA consortium had at least one driver alteration in these pathways, and 57% percent of the tumors had at least one alteration potentially targetable by currently available treatments [6]. We hypothesize that these signaling pathways are altered because glycosaminoglycans play an essential role in the composition of the extracellular matrix [7], helping to regulate processes such as metabolic signaling, apoptosis, cell migration, adhesion, and antigen presentation, in both cancer and MPS.

5 Concluding Remarks

The available public data is essential for amplified the multi-omic knowledge of complex and rare diseases. Bioinformatic approaches, such as gene enrichment analysis, may help us understand the complexity of processes deranged in several diseases. Studying the tumor ontology signature in lysosomal disorders may help understand lysosomal storage diseases and cancer's underlying mechanisms. This may help amplify therapeutic approaches for both types of diseases.

References

1. Cairns, R.A., Harris, I.S., Mak, T.W.: Regulation of cancer cell metabolism. Nat. Rev. Cancer **11**(2), 85–95 (2011). https://doi.org/10.1038/nrc2981
2. Matte, U., Pasqualim, G.: Lysosome: the story beyond the storage. J. Inborn Errors Metab. Screen. **4**, e160044 (2016). https://doi.org/10.1177/2326409816679431
3. Kallunki, T., Olsen, O.D., Jäättelä, M.: Cancer-associated lysosomal changes: friends or foes? Oncogene **32**(16), 1995–2004 (2012). https://doi.org/10.1038/onc.2012.292
4. Fiorenza, M.T., Moro, E., Erickson, R.P.: The pathogenesis of lysosomal storage disorders: beyond the engorgement of lysosomes to abnormal development and neuroinflammation. Hum. Mol. Genet. **27**(R2), R119–R129 (2018). https://doi.org/10.1093/hmg/ddy155
5. Martinez-Carreres, L., Nasrallah, A., Fajas, L.: Cancer: linking powerhouses to suicidal bags. Front. Oncol. **7**, 204 (2017). https://doi.org/10.3389/fonc.2017.00204
6. Sanchez-Vega, F., et al.: Oncogenic signaling pathways in the cancer genome atlas. Cell **173**(2), 321–337 (2018). https://doi.org/10.1016/j.cell.2018.03.035
7. Davidson, S.M., Vander Heiden, M.G.: Critical functions of the lysosome in cancer biology. Ann. Rev. Pharmacol. Toxicol. **57**(1), 481–507 (2017). https://doi.org/10.1146/annurev-pharmtox-010715-103101

Capítulo 7

DISCUSSÃO

Com o advento do projeto de sequenciamento do genoma humano, muitas esperanças e expectativas surgiram com relação ao tratamento e a cura de doenças humanas. As tecnologias de sequenciamento massivo paralelo deram oportunidade ao surgimento de novas ferramentas de diagnóstico, o que vem sendo implementado cada vez mais no cotidiano da prática da genética clínica. Com isso, a existência de milhares de tipos de arquivos e dados novos, o que chamamos de *big data*, surge como uma oportunidade para explorar e elaborar novas hipóteses de estudo (Gómez-Vela *et al.*, 2020).

Os estudos computacionais podem apresentar diversas vantagens e desvantagens na sua relação com os estudos em bancada, com modelos animais, linhagens celulares, ou outros. Otimiza-se o tempo de obtenção de dados e os custos da pesquisa. Ao se tratar de doenças raras, que muitas vezes não possuem cura ou tratamento ainda bem estabelecido, a utilização de estudos computacionais, como o reposicionamento de fármacos, ou o uso de biologia de sistemas podem fornecer opções de terapias promissoras (Roesler *et al.*, 2021).

Estudos envolvendo técnicas de sequenciamento de nova geração estão cada vez mais inseridos na prática clínica. O sequenciamento de exomas pode ser utilizado para fornecer um diagnóstico mais assertivo de doenças que ainda não foram diagnosticadas (Turro *et al.*, 2020), diminuindo cada vez mais a odisseia que os pacientes percorrem até saber seu real diagnóstico (Posey, 2019). Para aumentar a sensibilidade e especificidade de diagnóstico, técnicas de RNA-seq podem ser utilizadas de forma integrada com o sequenciamento de exomas e genomas, de modo a melhorar a interpretação de variantes intrônicas e exônicas, como vemos em diversos estudos (Lee *et al.*, 2020).

Ainda tratando-se do uso de transcriptomas, muitas vezes tais técnicas são mais indicadas para identificar alguns fenótipos, devido a possibilidade de se conhecer e identificar os mecanismos biológicos por trás das doenças, o que não pode ser visto com sequenciamento de DNA (Kousi, 2021). Exemplos de aplicação envolvem o estudo de *splicing* aberrante que pode ser tecido-específico, expressão diferencial de genes em tecidos derivados de doenças em comparação com perfis de controles ou indivíduos saudáveis, e casos de expressão alelo-específica ou expressão monoalélica específica, como no caso da Síndrome de Prader-Willi (Curry *et al.*, 2021). Apesar do fato de que as análises transcriptômicas representam uma "*selfie*" da biologia celular naquele contexto, essa técnica oferece uma visão ortogonal de como as informações contidas nas variantes do DNA se traduzem no contexto biológico e nas vias biológicas da doença (Murdock *et al.*, 2021).

A disponibilidade de informações e ferramentas pode ter um grande impacto positivo, reduzindo o número de falhas no desenvolvimento ou na aplicação de novos medicamentos ou terapias (Zloh & Kirton, 2018). Possíveis efeitos adversos podem ser minimizados através da análise dos alvos moleculares (Zloh & Kirton, 2018). No entanto, a disponibilidade desses recursos não garante por si só o sucesso do resultado da técnica ou algoritmo utilizado, pois muitas vezes nem sempre o resultado proveniente das análises *in silico* gera a resposta esperada nos modelos animais ou nas linhagens celulares testadas, apesar da utilização de métodos estatísticos e computacionais extremamente robustos (Camastra *et al.*, 2015).

No que se refere ao tamanho amostral e no número de dados disponíveis publicamente de doenças raras, como no caso desta tese, dados de transcriptomas de MPS, a pouca quantidade de estudos encontrados para realizar as análises computacionais também pode ser um problema ao se treinar um modelo, ou ao realizar análises de redes. A falta desses *datasets* pode ser explicada por se tratar de doenças raras, nas quais o número de estudos e o tamanho amostral de cada um deles é reduzido, dificultando a disponibilidade de dados para as análises ômicas que utilizamos nos pipelines de bioinformática (Schlieben *et al.*, 2021).

Um outro ponto que pode influenciar na falta de dados multi-ômicos é a baixa expectativa de vida pacientes com doenças genéticas, que pode ser de 5 a 10 anos para um terço de todos os pacientes com doenças raras (Kerr *et al.*, 2020). Essa questão enfatiza a importância do diagnóstico precoce e da implementação de pipelines de análises ômicas. Tais necessidades podem ser supridas com o desenvolvimento de pipelines de bioinformática padronizados, para garantir resultados robustos e precisos de biomarcadores (Masica & Karchin, 2016), bem como a necessidade de se estabelecer consórcios para colaboração internacional, o que pode aumentar o número de pacientes dos estudos, e o poder da pesquisa envolvendo doenças raras (Bienstock, 2019).

Estratégias como *data mining* (mineração de texto) podem ser efetivas para correlacionar achados da literatura com dados experimentais, como observado no trabalho de Ehrhart e colaboradores (2021). Desse modo, organizar as informações biológicas por meio de banco de dados, como abordado anteriormente na introdução desta tese, torna-se essencial para incentivar práticas de ciência aberta e reprodutibilidade.

Um bom exemplo de integração de dados em doenças raras é o GARD (The *Genetic and Rare Diseases Information Center*). Zhu e colaboradores (2020) utilizaram os dados do GARD para construir, através de grafos de meta-ontologias, uma ferramenta para integrar dados do FDA, OMIM, Orphanet, HPO, Mondo, e outras bases de busca de informações. Essa ferramenta está vinculada ao Github, o que a torna acessível e reprodutível, de modo que qualquer usuário pode ter contato com os códigos utilizados para construir a ferramenta, e com isso pode implementar boas práticas de programação.

Reprodutibilidade e ciência aberta no campo da bioinformática é um ponto essencial a ser discutido, pois dentro da área não existe um consenso sobre padronizações e *guidelines* de qualidade. Em pesquisas envolvendo animais, podemos citar iniciativas como o Arrive e Arrive2, na metanálise e revisão sistemática, o PRISMA, para estudos randomizados, o CONSORT, mas para bioinformática e biologia computacional não existe um protocolo centralizado para guiar as análises.

A *Plos Computacional Biology* possui uma coleção de artigos chamada 10 regras simples, e nela existem bons exemplos a serem utilizados, tais como:

- 10 regras simples para tirar vantagem de ferramentas como Git e Github;
- 10 regras simples para realizar acessorias e auxílios de qualidade em bioinformática;
- 10 regras simples para aplicar ciência aberta;
- 10 regras simples para escrever e compartilhar códigos no Jupyter notebooks;
- 10 regras simples para armazenar dados em repositórios digitais;

- 10 regras simples para ter experimentos com proveniência;
- 10 regras simples para conduzir experimentos em biologia computacional;
- 10 regras simples para utilizar *big data* de forma responsável;
- 10 regras simples para conduzir análises estatísticas;
- 10 regras simples para biólogos aprenderem a programar.

Atualmente, para reproduzir uma análise de bioinformática, os dados brutos e a lista de ferramentas utilizadas (com a versão dos pacotes descritos) pode não ser suficiente para garantir a reprodutibilidade dos resultados obtidos (Kulkarni *et al.,* 2018). Nem sempre os pesquisadores descrevem com qualidade e rigor as ferramentas que usaram, qual a versão e funções utilizadas em suas análises. Wilson et al. (2014) descreve em seu trabalho de revisão boas práticas para conduzir pesquisas envolvendo computação de forma eficiente, o qual os autores relatam como sendo essencial para a prática de resolução de problemas e execução de projetos.

Tendo em vista todos os fatores supracitados, boas práticas de reprodutibilidade e ciência aberta tornam a bioinformática e a biologia computacional ferramentas valiosas para a aplicação em ciências biológicas e biomédicas. O grande volume de dados disponível pode representar uma boa chance para resolver problemas encontrados em doenças raras, facilitando a prática clínica, e trazendo boas expectativas para os pacientes. No entanto, os achados obtidos através de análises *in silico* precisam ser comprovados através de pesquisas experimentais, mas contribuem para um melhor direcionamento de esforços, demonstrando a complementariedade de ambas as abordagens.

Capítulo 8

CONCLUSÕES

Capítulo 3 - Bancos de dados biológicos

Fazer um levantamento dos bancos de dados biológicos existentes, classificando-os de acordo com diferentes subáreas da bioinformática;

Através da busca nos sites do Omictools e bio.tools, conseguimos encontrar 646 ferramentas e bancos de dados *on line*, classificados em 13 áreas.

Elaborar uma lista de ferramentas para que pessoas sem muito conhecimento de linguagem de programação sejam capazes de conduzir pesquisas em bioinformática

Elaboramos o site com os bancos que estavam online, ou que fazem manutenção periódica dos seus servidores.

Criar um estudo de caso para ilustrar a usabilidade dos bancos listados

Desenvolvemos um estudo de caso utilizando o gene da *ACE2* como exemplo, escolhemos dois bancos de dados em cada categoria. Salientamos que não fizemos ranqueamento dos bancos, a escolha foi subjetiva e levou em consideração a usabilidade das ferramentas.

Capítulo 4 - Neuronetworks – biologia de sistemas e dano neurológico em MPS

Identificar possíveis genes e vias que sejam bons biomarcadores de dano neurológico nas diferentes MPS que possuem dados de transcriptomas disponíveis publicamente

Utilizando análises de centralidade, conseguimos identificar os genes *ITGB1*, *PTK2*, *FN1*, *EGFR*, e *PXN* para a rede de MPS I (GSE11906). Para a rede de MPS II (GSE95224), os genes *GNG7*, *AVP*, *AGT*, *CCL2* e *MCHR1*. Para a rede de MPS IIIB (GSE15758), genes que codificam proteínas ribossomais, tais como *RPS15A*, *RLP7*, *RPS11*, *RPL2*, *RPS9*. Finalmente, para MPS VII (GSE76283), os genes *CDC20*, *CDK1*, *LYZ1*, *SERPINA3N* e *GNB5* foram identificados como centrais nas redes de interação.

Com relação as vias enriquecidas nesses *datasets*, os processos biológicos comum entre os diferentes tipos de MPS foram vias relacionadas a apoptose, direcionamento axonal, sinalização de Cálcio, PI3K-Akt, WNT e vias do sistema imune. Acreditamos que essas vias são promissoras como biomarcadores de dano neurológico nas MPS. Além disso, salientamos que estudos experimentais, como em modelos animais ou linhagens celulares são necessários para validar tais achados.

Capítulo 5 - Doenças lisossomais e tumores neurológicos

Investigar a expressão genica de enzimas lisossomais em gliomas

Para os gliomas de baixo grau, os genes diferencialmente expressos (top diferencialmente expressas no tumor quando comparados ao tecido normal) foram proteases como *CTSO*, glicosidases como *NAGLU* e *MAN2B1*, sulfatases como ARSB, nucleases como *DNASE2*, a ceramidase *ASAH1*, as enzimas *GM2A* e a proteína de membrana *CLN3* como super expressas. Lipases do tipo *LYPLA3*, a proteína de membrana *LAMP*, *NRAMP*, *ABCA* e *ABCB9* foram identificadas como hipoexpressas nesses tumores.

Para glioma de alto grau, como glioblastoma multiforme, os genes diferencialmente expressos foram as proteases *CTSC* e *CTSS*, as glicosidases *GLB1*, *NAGA*, *NAGLU*, *MAN2B1*, a sulfatases *GNS*, nuclease *DNASE2*, e as proteínas lisossomais *LAMP4* e *NPC2* como super expressas. Os genes e proteínas identificados como hipo expressas nesse tumor foram *TPPP*, *IDS*, *NRAMP* e *ABCA2*.

De modo geral, para gliomas de baixo grau foram identificados 1126 genes diferencialmente expressos, e para glioblastoma multiforme, 2280. Destes, 846 genes foram comuns para ambos os tipos tumorais.

Identificar variantes patogênicas em genes lisossômicos

Para gliomas de baixo grau, identificamos 51 variantes patogênicas em genes lisossômicos, destas 17 possuem código do dbSNP, e 3 foram encontradas na literatura como associadas a alguma doença lisossomal.

Para glioblastoma multiforme, identificamos 100 variantes patogênicas, sendo elas 38 descritas com código do dbSNP, e 14 relacionadas a doenças lisossômicas.

Avaliar o impacto da expressão dessas enzimas na sobrevida dos pacientes

As curvas de sobrevida dos gliomas de baixo grau tiveram significância estatística em 16 genes lisossômicos, relacionados a bom prognóstico quando a expressão desses genes era baixa, e 11 foram bons prognósticos quando a expressão dos genes lisossomais era alta. Para os glioblastomas, 8 genes foram identificados como bom prognostico tendo a expressão baixa, e 6 quando a expressão era alta. Para genes prognósticos de expressão baixa, identificamos 3 genes em comum nos dois tumores, e para a expressão alta, apenas 1 gene em comum. Tambem conseguimos construir assinaturas de expressão desses genes lisossômicos, e criamos um score de expressão, correlacionando-o com o status de vida dos pacientes e a progressão dos tumores.

Capítulo 6 - Vias de sinalização oncogênica em Mucopolissacaridoses

Identificar vias biológicas de sinalização de câncer presentes nas diferentes MPS

Através das análises de enriquecimento ontológico, foram identificadas 25 vias de sinalização oncogênica nos dados de transcriptomas de MPSI, MPSII, MPSIIIA, MPSIIIB, MPS VI e MPSVII. Destas, 11 vias foram encontradas em comum em todos os tipos de MPS analisados no estudo.

Capítulo 9

REFERÊNCIAS

Abdelhakim, M., McMurray, E., Syed, A. R., Kafkas, S., Kamau, A. A., Schofield, P. N., & Hoehndorf, R. (2020). DDIEM: drug database for inborn errors of metabolism. Orphanet journal of rare diseases, 15(1), 146. https://doi.org/10.1186/s13023-020-01428-2

Akhter, S., Kaur, H., Agrawal, P., & Raghava, G. P. S. (2019). RareLSD: a manually curated database of lysosomal enzymes associated with rare diseases. Database, 2019. https://doi.org/10.1093/database/baz112

Asrani, K., Murali, S., Lam, B., Na, C.-H., Phatak, P., Sood, A., Kaur, H., Khan, Z., Noë, M., Anchoori, R. K., Talbot, C. C., Jr., Smith, B., Skaro, M., & Lotan, T. L. (2019). mTORC1 feedback to AKT modulates lysosomal biogenesis through MiT/TFE regulation. In Journal of Clinical Investigation (Vol. 129, Issue 12, pp. 5584–5599). American Society for Clinical Investigation. https://doi.org/10.1172/jci128287

Ballabio A. (2016). The awesome lysosome. EMBO molecular medicine, 8(2), 73–76. https://doi.org/10.15252/emmm.201505966

Beck, M (2007). New therapeutic options for lysosomal storage disorders: enzyme replacement, small molecules and gene therapy. Hum Genet, 121(1):1-22

Bienstock, R. J. (2019). Data Sharing Advances Rare and Neglected Disease Clinical Research and Treatments. In ACS Pharmacology & Translational Science (Vol. 2, Issue 6, pp. 491–496). American Chemical Society (ACS). https://doi.org/10.1021/acsptsci.9b00034

Bogart, K. R., & Irvin, V. L. (2017). Health-related quality of life among adults with diverse rare disorders. Orphanet journal of rare diseases, 12(1), 177. https://doi.org/10.1186/s13023-017-0730-1

Cairns, RA; Harris, IS; Mak, TW (2011). Regulation of cancer cell metabolism. Nat Rev Cancer, 11: 85–95

Camastra, F., Di Taranto, M. D., & Staiano, A. (2015). Statistical and Computational Methods for Genetic Diseases: An Overview. In Computational and Mathematical Methods in Medicine (Vol. 2015, pp. 1–8). Hindawi Limited. https://doi.org/10.1155/2015/954598

Corn, PG; Wang, F; McKeehan, WL; Navone, N (2013). Targeting fibroblast growth factor pathways in prostate cancer. Clin Cancer Res., 19(21):5856–66

Curry, P. D. K., Broda, K. L., & Carroll, C. J. (2021). The Role of RNA-Sequencing as a New Genetic Diagnosis Tool. Current Genetic Medicine Reports, 9(2), 13–21. doi:10.1007/s40142-021-00199-x

Donati, M. A., Pasquini, E., Spada, M., Polo, G., & Burlina, A. (2018). Newborn screening in mucopolysaccharidoses. Italian journal of pediatrics, 44(Suppl 2), 126. https://doi.org/10.1186/s13052-018-0552-3

Ehrhart, F., Willighagen, E. L., Kutmon, M., van Hoften, M., Curfs, L. M. G., & Evelo, C. T. (2021). A resource to explore the discovery of rare diseases and their causative genes. In Scientific Data (Vol. 8, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41597-021-00905-y

Fan, Y; Dickman, KG; Zong, WX (2010). Akt and c- Myc differentially activate cellular metabolic programs and prime cells to bioenergetic inhibition. J Biol Chem 285:7324–7333

Geisslinger, F., Müller, M., Vollmar, A. M., & Bartel, K. (2020). Targeting Lysosomes in Cancer as Promising Strategy to Overcome Chemoresistance-A Mini Review. Frontiers in oncology, 10, 1156. https://doi.org/10.3389/fonc.2020.01156

Gómez-Vela, F., Divina, F., & García-Torres, M. (2020). Computational Methods for the Analysis of Genomic Data and Biological Processes. In Genes (Vol. 11, Issue 10, p. 1230). MDPI AG. https://doi.org/10.3390/genes11101230

Hoffmann, GF Zschocke, J, Nyhan, WL (org.) Inherited Metabolic Diseases - A Clinical Approach (2017). Springer, ISBN 978-3-540-74723-9

Jung, CH; Kim, H; Ahn, J; Jung, SK; Um, MY; Son, KH; Kim, TW; Ha, TY (2013). Anthricin Isolated from Anthriscus sylvestris (L.) Hoffm. Inhibits the Growth of Breast Cancer Cells by Inhibiting Akt/mTOR Signaling, and Its Apoptotic Effects Are Enhanced by Autophagy Inhibition. Evid Bas Alter Med, 2013:1-9

Kousi, M. (2021) Chapter 11 - Transcriptomics in rare diseases. In Translational and Applied Genomics, Genomics of Rare Diseases, Academic Press, 2021, Pages 215-228, ISBN 9780128201404, https://doi.org/10.1016/B978-0-12-820140-4.00007-7

Kulkarni, N., Alessandrì, L., Panero, R. et al. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. BMC Bioinformatics 19, 349 (2018). https://doi.org/10.1186/s12859-018-2296-x

Lee J. J. Y., Wasserman W. W., Hoffmann G. F., van Karnebeek C. D. M., Blau N., Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. Genet Med 20 (2018) 151-158

Lee, H., Huang, A.Y., Wang, Lk. et al. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genet Med 22, 490–499 (2020). https://doi.org/10.1038/s41436-019-0672-1

Lum, JJ; Bauer, KM; Kong, M; Harris, MH; Li, C; Lindsen, T; Thompson, CB (2005). Growth factor regulation of autophagy and cell survival in the absence of apoptosis. Cell, 120(2):237-48

Machado, E. R., Annunziata, I., van de Vlekkert, D., Grosveld, G. C., & d'Azzo, A. (2021). Lysosomes and Cancer Progression: A Malignant Liaison. Frontiers in cell and developmental biology, 9, 642494. https://doi.org/10.3389/fcell.2021.642494

Masica DL, Karchin R (2016) Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. PLoS Comput Biol 12(5): e1004725. https://doi.org/10.1371/journal.pcbi.1004725

Meyer-Schwesinger C. Lysosome function in glomerular health and disease. Cell Tissue Res. 2021 Jan 12. doi: 10.1007/s00441-020-03375-7. Epub ahead of print. PMID: 33433692

Murdock, D. R., Dai, H., Burrage, L. C., Rosenfeld, J. A., Ketkar, S., Müller, M. F., Yépez, V. A., Gagneur, J., Liu, P., Chen, S., Jain, M., Zapata, G., Bacino, C. A., Chao, H.-T., Moretti, P., Craigen, W. J., Hanchard, N. A., & Lee, B. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. In Journal of Clinical Investigation (Vol. 131, Issue 1). American Society for Clinical Investigation. https://doi.org/10.1172/jci141500

Navarro-Romero, A., Montpeyó, M., & Martinez-Vicente, M. (2020). The Emerging Role of the Lysosome in Parkinson's Disease. Cells, 9(11), 2399 https://doi.org/10.3390/cells9112399

Platt, FM. (2018) Emptying the stores: lysosomal diseases and therapeutic strategies. Nat Rev Drug Discov. 2018 Feb;17(2):133-150. doi: 10.1038/nrd.2017.214.

Popolin, C; Cominetti, MR. (2017). A review of ruthenium complexes activities on different steps of the metastatic process in breast cancer cells. Mini Rev Med Chem, 17:1-1

Posey, J.E. Genome sequencing and implications for rare disorders. Orphanet J Rare Dis 14, 153 (2019). https://doi.org/10.1186/s13023-019-1127-0

Puertollano R. (2014). mTOR and lysosome regulation. F1000prime reports, 6, 52. https://doi.org/10.12703/P6-52

Saudubray, J. M., & Garcia-Cazorla, À. (2018). Inborn Errors of Metabolism Overview: Pathophysiology, Manifestations, Evaluation, and Management. Pediatric clinics of North America, 65(2), 179–208. https://doi.org/10.1016/j.pcl.2017.11.002

Saudubray, J. M., Mochel, F., Lamari, F., & Garcia-Cazorla, A. (2019). Proposal for a simplified classification of IMD based on a pathophysiological approach: A practical guide for clinicians. Journal of inherited metabolic disease, 42(4), 706–727. https://doi.org/10.1002/jimd.12086

Schlieben, L. D., Prokisch, H., & Yépez, V. A. (2021). How Machine Learning and Statistical Models Advance Molecular Diagnostics of Rare Disorders Via Analysis of RNA Sequencing Data. In Frontiers in Molecular Biosciences (Vol. 8). Frontiers Media SA. https://doi.org/10.3389/fmolb.2021.647277

Soares LDF, Villalba Silva GC, Kubaski F, Giugliani R, Matte U. MPSBase: Comprehensive repository of differentially expressed genes for mucopolysaccharidoses. Mol Genet Metab. 2021 Aug;133(4):372-377. doi: 10.1016/j.ymgme.2021.06.004. Epub 2021 Jun 15. PMID: 34147352

Tang, T., Yang, Zy., Wang, D. et al. The role of lysosomes in cancer development and progression. Cell Biosci 10, 131 (2020). https://doi.org/10.1186/s13578-020-00489-x

Turro, E., Astle, W.J., Megy, K. et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature 583, 96–102 (2020). https://doi.org/10.1038/s41586-020-2434-2

Wie, J., Liu, Z., Song, H., Tropea, T. F., Yang, L., Wang, H., Liang, Y., Cang, C., Aranda, K., Lohmann, J., Yang, J., Lu, B., Chen-Plotkin, A. S., Luk, K. C., & Ren, D. (2021). A growth-factor-activated lysosomal K+ channel regulates Parkinson's pathology. In Nature (Vol. 591, Issue 7850, pp. 431–437). Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-021-03185-z

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. (2014) Best Practices for Scientific Computing. PLoS Biol 12(1): e1001745. https://doi.org/10.1371/journal.pbio.1001745

Zhao, B., Dierichs, L., Gu, J. N., Trajkovic-Arsic, M., Axel Hilger, R., Savvatakis, K., Vega-Rubin-de-Celis, S., Liffers, S. T., Peña-Llopis, S., Behrens, D., Hahn, S., Siveke, J. T., & Lueong, S. S. (2020). TFEB-mediated lysosomal biogenesis and lysosomal drug sequestration confer resistance to MEK inhibition in pancreatic cancer. Cell death discovery, 6, 12. https://doi.org/10.1038/s41420-020-0246-7

Zhu, Q., Nguyen, D.-T., Grishagin, I., Southall, N., Sid, E., & Pariser, A. (2020). An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). In Journal of Biomedical Semantics (Vol. 11, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s13326-020-00232-y

ANEXOS

Artigos publicados com coautorias:

Soares, L., **Villalba Silva, G. C.**, Kubaski, F., Giugliani, R., & Matte, U. (2021). MPSBase: Comprehensive repository of differentially expressed genes for mucopolysaccharidoses. Molecular genetics and metabolism, 133(4), 372–377. https://doi.org/10.1016/j.ymgme.2021.06.004

Eisele, B.S., **Silva, G.C.V.**, Bessow, C. et al. An in silico model using prognostic genetic factors for ovarian response in controlled ovarian stimulation: A systematic review. J Assist Reprod Genet 38, 2007–2020 (2021). https://doi.org/10.1007/s10815-021-02141-0

Matschinske, J., Alcaraz, N., Benis, A., **Silva, G.C.V.** et al. The AIMe registry for artificial intelligence in biomedical research. Nat Methods 18, 1128–1131 (2021). https://doi.org/10.1038/s41592-021-01241-0

Resumos publicados em anais de eventos:

Villalba Silva GC and da Silveira Matte U. Differential expression analysis of lysosomal storage related genes in gliomas [version 1; not peer reviewed]. F1000Research 2019, 8(ISCB Comm J):1958 (poster) (https://doi.org/10.7490/f1000research.1117674.1)

SILVA, G. C. V.; SANTOS, H. S.; BALDO, G.; MATTE, U. S. CHARACTERIZATION OF REPETITIVE ELEMENTS IN GENES ENCODING DIFFERENT MUCOPOLYSACCHARIDOSES. In: XXXI Congresso Brasileiro de Genética Médica, 2019, Salvador. ANAIS CBGM 2019, 2019

Villalba GCV and Matte U. Cancer pathways are deranged in Mucopolysaccharidoses [version 1; not peer reviewed]. F1000Research 2020, 9(ISCB Comm J):820 (poster) (https://doi.org/10.7490/f1000research.1118121.1)

SILVA, Gerda Cristal Villalba; Matte, Ursula. Neuro-networks investigating the neurological impairment of mucopolysaccharidoses using a system biology approach. MOLECULAR GENETICS AND METABOLISM, 2021.

Villalba Silva GC and Matte U. Systems biology gives clues about the neurological impairment in MPSS [version 1; not peer reviewed]. F1000Research 2021, 10(ISCB Comm J):210 (poster) (https://doi.org/10.7490/f1000research.1118525.1)

Villalba Silva GC and Matte U. Drug repositioning for mucopolysaccharidoses based on systems biology data [version 1; not peer reviewed]. F1000Research 2021, 10(ISCB Comm J):845 (poster) (https://doi.org/10.7490/f1000research.1118757.1)

Villalba Silva, GC, Matte, U. Virtual drug screening and repositioning for Mucopolysaccharidoses. 2nd Women in Bioinformatics & Data Science LA Conference, Abstract book available at https://drive.google.com/file/d/1lEikliyXC0qOaL2_KZ5h97IvE3XycfRU/view

Premiações:

Award of virtual ECCB2020 - New trends in Bioinformatics, European Student Council Symposium - 19th European Conference on Computational Biology

Divulgação científica:

BUENO, A. X.; BRAGATTE, M. A. S.; ARENZON, J. J.; FONTES-DUTRA, M.; REALES, G.; BAUM, F.; QUINSANI, D. A.; ALMEIDA, T.; MELO, T. P.; **SILVA, G. C. V.** Pint Of Science Festival. 2018 - 2021. (Festival).

Podcast de Bioinformática da RSG Brasil. Disponível em < https://open.spotify.com/show/6gYtzMr1HtFf91frXIyhNN?si=4b0138fba4df4131>

Durante a pandemia, tive a honra de participar por mais de um ano e meio da equipe de diagnóstico da COVID-19, no Instituto de Ciências Básicas da Saúde, onde realizei extração de RNA e alicotagem das amostras para RT-PCR.