Universidade Federal do Rio Grande do Sul
Instituto de Biociências
bacharel em Biotecnologia - Ênfase em Bioinformática

Luís Dias Ferreira Soares

MPSBase: Banco de dados compreensivo de genes diferencialmente expressos em Mucopolissacaridoses

Porto Alegre
2021

Luís Dias Ferreira Soares

MPSBase: Banco de dados compreensivo de genes diferencialmente expressos em Mucopolissacaridoses

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de bacharel em Biotecnologia - Ênfase em Bioinformática da/do Instituto de Biociências da Universidade Federal do Rio Grande do Sul. Artigo formatado para o periódico *Molecular Genetics and Metabolism.*
Orientadora: Ursula da Silveira Matte

Porto Alegre
2021

FOLHA DE APROVAÇÃO

Luís Dias Ferreira Soares

MPSBase: Banco de dados compreensivo de genes diferencialmente expressos em Mucopolissacaridoses

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de bacharel em Biotecnologia - Ênfase em Bioinformática da/do Instituto de Biociências da Universidade Federal do Rio Grande do Sul.

Orientadora: Ursula da Silveira Matte

Aprovado em: Porto Alegre,19 de maio de 2021.

BANCA EXAMINADORA:

Ursula da Silveira Matte, PhD, Msc

Universidade Federal do Rio Grande do Sul

Guilherme Baldo, PhD, Msc

Universidade Federal do Rio Grande do Sul

Thayne Woycinck Kowalski, PhD, Msc

Universidade Federal do Rio Grande do Sul

Rafaella Mergener, PhD, Msc

Universidade Federal do Rio Grande do Sul

RESUMO

As Mucopolissacaridoses (MPSs) são doenças de acúmulo lisossomal causadas pela deficiência de enzimas necessárias para o metabolismo de glicosaminoglicanos (GAGs). Essas e outras doenças raras têm ganhado uma significativa melhoria nos processos de pesquisa e diagnóstico com o advento das tecnologias de análise de DNA em larga escala. Para doenças monogênicas, como as MPSs, diferentes pesquisas têm focado na identificação de perfis de transcrição gênica em resposta à deficiência enzimática. Atualmente, há 13 estudos publicamente disponíveis avaliando os perfis de transcrição para 6 tipos diferentes de MPS. A análise de dados de transcriptoma requer conhecimentos em programação, o que pode dificultar a análise.Para mitigar essa barreira de acesso à informação, foi desenvolvido um banco de dados de interface amigável, abrangendo todos esses estudos e foi disponibilizado para acesso pela internet. Combinando diferentes grupos experimentais par-a-par, chegamos a mais de 50 comparações entre diferentes condições da doença, assim como grupos controle. As análises foram realizadas pelo R usando os pacotes limma, affy e oligo para análises de microarranjo e edgeR para RNA-Seq. Todos os dados de microarranjo foram normalizados com o método média robusta multi-arranjo (RMA) e o valor p foi corrigido pela taxa de falsas descobertas (FDR). Utilizando esses métodos, foram identificados os genes diferencialmente expressos (DEGs) e com esses foi feito o enriquecimento de vias e termos ontológicos com os pacotes enrichKEGG e ClusterProfiler. Com o intuito de melhorar a experiência do usuário, foram incluídas representações gráficas para ambos DEGs e vias e termos enriquecidos.

Palavras-chave: Mucopolissacaridoses. Análise de expressão diferencial. Bancos de dados para genética. Distúrbios raros do metabolismo.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

DEG	Gene diferencialmente expresso
FDR	Taxa de falso descobrimento
GAG	Glicosaminoglicano
GEO	Gene expression omnibus
GO	Ontologia Gênica
IEM	Erros inatos do metabolismo
logFC	logarítmo na base 2 do valor de Fold Change
MPS	Mucopolissacaridose
NGS	Sequenciamento de próxima geração

SUMÁRIO

1 INTRODUÇÃO

1.1 ERROS INATOS DO METABOLISMO

Os erros inatos do metabolismo (IEMs) constituem um grupo com mais 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; GARCIA-CAZORLA, 2018).

Os mesmos autores relatam uma sub-representatividade dos IEMs dentro das ciências médicas, o que retarda a descoberta de novas terapias, além de dificuldades diagnósticas. Em parte isso se deve a barreira entre a biologia molecular e a prática clínica, que carece de ferramentas adequadas para um fluxo e compatibilidade de informação.

A classificação proposta pelos autores divide os IEMs em três grandes grupos. O grupo 1 abrange distúrbios do metabolismo intermediário afetando pequenas moléculas. Esse é o maior grupo e engloba os IEMs incluídos no Programa Nacional de Triagem Neonatal (teste do pezinho): fenilcetonúria e deficiência de biotinidase (MINISTÉRIO DA SAÚDE, 2016). Outros distúrbios também bem conhecidos nesse grupo estão relacionados ao metabolismo de aminoácidos, açúcares simples, metais e neurotransmissores, alguns exemplos são a doença da urina do xarope de bordo, homocistinúria, galactosemia, hemocromatose (SAUDUBRAY; GARCIA-CAZORLA, 2018).

O grupo 2 inclui distúrbios relacionados às deficiências no metabolismo energético. Essas doenças estão relacionadas a funções mitocondriais ou metabolismo de ácidos graxos e corpos cetônicos. Os quadros clínicos provocados podem ser severos, como no caso das disfunções mitocondriais que são intratáveis, em geral. Ou mais brandos, como na maioria das anomalias energéticas citoplasmáticas, como alterações no metabolismo de glicose e glicogênio que são tratáveis.

No último grupo estão os erros no metabolismo relacionados a moléculas complexas. Esses distúrbios envolvem processos que ocorrem em organelas e geralmente são progressivos. As doenças desse grupo podem ser classificadas pela

organela acometida, como os distúrbios peroxissomais, de transporte de membranas e as doenças de depósito lisossômico. Esse último grupo engloba 66 condições clínicas (POSWAR et al., 2019).

1.2 DOENÇAS DE DEPÓSITO LISOSSÔMICO E MPSS

As doenças de depósito lisossômico são categorizadas de acordo com o substrato acumulado, podendo ser oligossacaridoses, esfingolipidoses, mucolipidoses, lipofuscinoses ceróides neuronais e mucopolissacaridoses (POSWAR et al., 2019).

As mucopolissacaridoses (MPSs), que serão abordadas neste trabalho, são divididas em 11 tipos, cada um relacionado a uma deficiência enzimática e acúmulos de GAGs diferentes (Fig. 1). O que todos eles têm em comum é o distúrbio na degradação de glicosaminoglicanos (GAGs) (SHAPIRO; EISENGART, 2021). Os GAGs são componentes da matriz extracelular, e interagem com uma grande diversidade de proteínas e receptores. Quando acumulados nos lisossomos, levam a complicações multissistêmicas (HARPER et al., 1998; SCRIVER et al., 2001).

As MPSs, quase em sua totalidade, têm padrão de herança autossômico recessivo. A exceção é a MPS II, ou síndrome de Hunter, que é recessiva ligada ao cromossomo X (GIUGLIANI et al., 2014). O quadro clínico de um portador de MPS pode ser muito diferente de acordo com o gene afetado, ou então, dois portadores de alterações no mesmo gene podem ter manifestações distintas, por motivos não bem elucidados ainda (CAMPOS; MONAGA, 2012). Na maioria dos casos, há anomalias cerebrais detectadas por ressonância magnética, assim como disfunções neurocognitivas e comportamentais (SHAPIRO; EISENGART, 2021).

Com relação ao diagnóstico, nenhuma das MPSs estão incluídas no Programa Nacional de Triagem Neonatal (MINISTÉRIO DA SAÚDE, 2016), diferentemente dos Estados Unidos, onde o teste de MPS I está incluso na triagem neonatal federal, mas cuja implementação depende dos estados (FISHER; KLEIN, [s.d.]). No Brasil, diante de suspeita clínica, realizam-se testes bioquímicos para a dosagem de GAGs na urina, principalmente (DIETER et al., 2002; “Doenças raras”, 2020). Para as MPSs em que a enzima deficiente é do tipo sulfatase, a dosagem dos GAGs deve ser acompanhada de um ensaio de atividade enzimática, para excluir a possibilidade de diagnóstico de uma deficiência múltipla de sulfatases

(WOOD et al. 2012; GIUGLIANI et al., 2014).

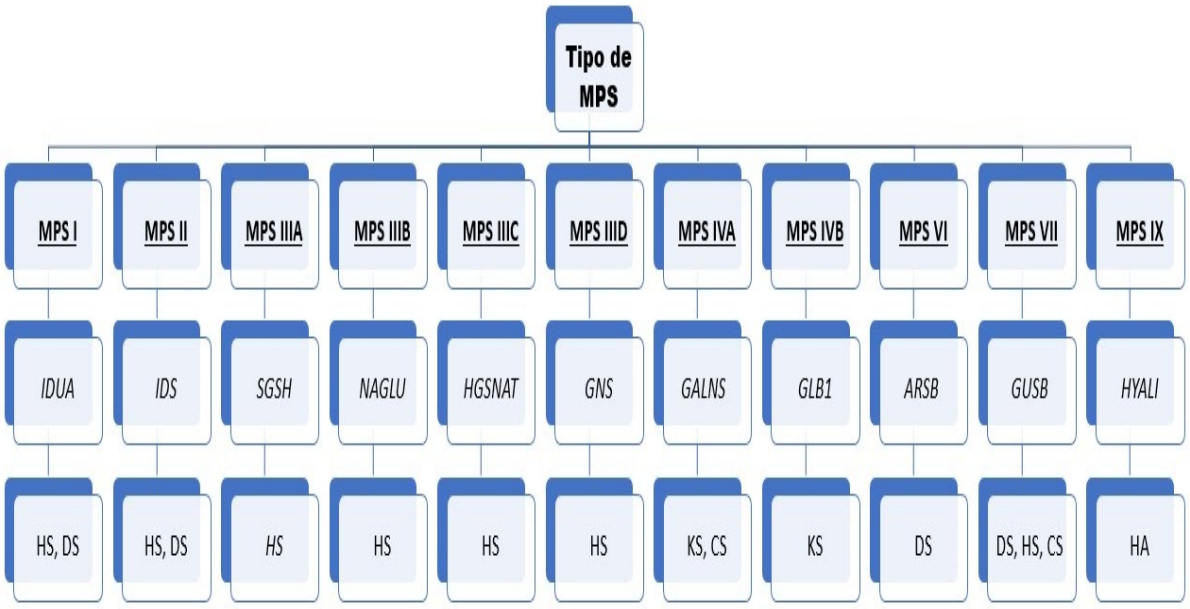


Figura 1 – Classificação das Mucopolissacaridoses. Na primeira fileira de blocos estão os subtipos de MPS, na segunda, os genes afetados, e na terceira, os glicosaminoglicanos acumulados. Adaptado de Clarke (2008). MPS = Mucopolissacaridoses; AH = ácido hialurônico; CS = condroitina sulfato; DS = dermatan sulfato; HS = heparan sulfato; KS = queratan sulfato.

Fonte: (VILLALBA, 2019).

A partir da confirmação do diagnóstico é feito o tratamento, que, no Brasil, está incluso na política de doenças raras do SUS, sendo disponíveis medicamentos para o tratamento de MPSs I, II, IV A, VI e VII (MINISTÉRIO DA SAÚDE, 2019; CONITEC, 2020b). Além disso, o mesmo sistema possibilita a utilização de transplante de células-tronco hematopoiéticas para os tipos I, II, IV A e VI (CONITEC, 2020a).

1.3 EMPREGO DE SEQUENCIAMENTO E MICRO-ARRANJO NA PESQUISA EM MPS

Tecnologias de análise de DNA em larga escala, genoma e transcriptoma, se tornaram mais acessíveis e precisas no diagnóstico de doenças genéticas (PEREIRA *et al*., 2015, CUMMINGS *et al*., 2017). Mais recentemente, foi demonstrado que as análises de transcriptômica também podem ser úteis no diagnóstico de doenças raras, principalmente avaliando formas anômalas de *splicing*

mas também se utilizando de análises de expressão diferencial (CUMMINGS *et al*.,2017). Atualmente, a análise de expressão diferencial obtida por microarranjo e, principalmente RNA-Seq, é amplamente empregada na busca por novos alvos terapêuticos e para compreender melhor os mecanismos de fisiopatologia das doenças (DHARSHINI *et al*., 2021; BROKOWSKA *et al*., 2021). Essas tecnologias têm a capacidade de dar um resultado preciso, buscando em uma diversidade de possíveis genes acometidos em um único exame.

Aplicados ao estudo de MPSs, métodos de transcriptômica têm contribuído para desvendar mecanismos fisiopatológicos 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*., 2009); neurodegeneração em MPS IIIA, MPS IIIB e MPS IIIC (MOSKOT *et al*., 2019); edição de transcritos mutantes em Síndrome de Hunter - MPS II (LUALDI *et al*., 2010); expressão de citocinas, neurotrofinas e estresse oxidativo em MPS IIIB (VILLANI *et al*.,2006); e apoptose em todas as MPSs (BROKOWSKA *et al*., 2021). Além disso, também contribuem para o 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).

1.4 BANCOS DE DADOS NA PESQUISA BIOLÓGICA.

A primeira base de dados biológicos é datada de 1965, ela foi criada por Margaret Dayhoff e contemplava estrutura de proteínas (VILLALBA; MATTE, 2021). A partir da virada do século XXI, em concomitância com a publicação de diversos projetos genoma e o desenvolvimento de conexões de internet mais eficientes, houve um crescimento exponencial na quantidade de bancos de dados biológicos (BAXEVANIS; BATEMAN, 2015). A disponibilidade de dados sistematizados tornou-se uma demanda de pesquisadores que trabalham com enorme quantidade de informação. Hoje, há áreas de pesquisa intrinsecamente dependentes do acesso a essas coleções (IMKER, 2018). Considerando exclusivamente a pesquisa com humanos, há inúmeras bases de dados publicamente disponíveis, contemplando diferentes experimentos, moléculas e doenças de forma especializada (VILLALBA; MATTE, 2021). Essas ferramentas acessíveis de bioinformática possibilitaram ao

grupo avaliar potenciais biomarcadores e RNAs longos não codificantes relacionados a diferentes tipos de câncer (DA CUNHA JAEGER et al., 2020; FALCON et al., 2018).

Entretanto, a principal limitação no uso de dados genômicos publicamente disponíveis é, muitas vezes, o desconhecimento de técnicas de bioinformática para análise dos resultados. Por isso, desenvolvemos um banco de dados de estudos de expressão em MPSs que possibilita, de forma intuitiva, a identificação de genes diferencialmente expressos e a comparação entre diferentes grupos.

A seguir, apresentamos o artigo submetido ao periódico Molecular Genetics and Metabolism, que se encontra na fase de resposta aos revisores. O artigo descreve o banco de dados, disponível em: ufrgs.br/mpsbase/ e suas ferramentas, apresentando um estudo de caso.

2 ARTIGO - MPSBASE: COMPREHENSIVE REPOSITORY OF DIFFERENTIALLY EXPRESSED GENES FOR MUCOPOLYSACCHARIDOSES

MPSBase: Comprehensive repository of differentially expressed genes for mucopolysaccharidoses

Luís Dias Ferreira Soares [1,2,3], Gerda Cristal Villalba Silva [2,3,4], Francyne Kubaski [5,6,7] , Roberto Giugliani [2,5,6,7], Ursula Matte* [2,3,4,5]

1 Graduation Program on Biotechnology/Bioinformatics, UFRGS, Porto Alegre, 91501-970, Brazil

2 Cells, Tissues and Genes Laboratory, HCPA, Porto Alegre 90035903, Brazil

3 Bioinformatics Core, HCPA, Porto Alegre 90035903, Brazil

4 Postgraduate Program in Genetics and Molecular Biology, UFRGS, Porto Alegre 91501970, Brazil

5 Department of Genetics, UFRGS, Porto Alegre 91501970, Brazil

6 Medical Genetics Service, HCPA, Porto Alegre 90035903, Brazil

7 Biodiscovery Lab, HCPA, Porto Alegre 90035903, Brazil

*Corresponding author, umatte@hcpa.edu.br

Address: Ramiro Barcelos, 2350, 90035903, Santa Cecilia, Porto Alegre, Brazil.

Word Counts (abstract): 235

Abbreviations: GAGs, Glycosaminoglycans; MPS, Mucopolysaccharidoses; LSDs, Lysosomal Storage Diseases; DEGs, Differentially expressed genes; FDR, False Discovery Rate; BP, Biological Process; GEO, Gene expression omnibus.

Authors and contributions:

Luis Dias Ferreira Soares – website design and construction, data analysis, manuscript writing.

Gerda Cristal Villalba Silva – website design, data analysis, manuscript writing.

Francyne Kubaski – manuscript writing and review.

Roberto Giugliani – manuscript writing and review.

Ursula Matte – website design, manuscript writing and review.

Corresponding Author:

Ursula Matte – umatte@hcpa.edu.br

Competing interest statement:

The authors declare they have no competing interests.

Ethics approval:

This work does not require Ethics approval because the used data are fully available at omic databases, like Gene Expression Omnibus - GEO.

Funding:

We would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the fellowships provided to GCVS (grant # 148615/2018-0), UM (grant # 312714/2018-1), and RG (grant # 30.3219-2019-0). FK has a postdoctoral fellowship from CAPES (grant # 88887.136366/2017-00). We also thank FIPE/HCPA (Project Number 2019-0761) for financial support and the fellowship provided to LDFS, and FUNDMED. The authors confirm independence from the sponsors.

Keywords:

Lysosomal storage diseases; Mucopolysaccharidoses; gene expression analysis; ontology analysis; biomarkers; biological databases.

Availability of data statement:

All the data presented on the article is available on MPSBase <https://www.ufrgs.br/mpsbase/>.

2.1 INTRODUCTION

Defects in lysosomal enzymes and transport proteins lead to the accumulation of undegraded or partially degraded metabolites, causing lysosomal storage diseases - LSDs (SAUDUBRAY; GARCIA-CAZORLA, 2018). The mucopolysaccharidoses (MPS) are diseases caused by disturbances in glycosaminoglycans' (GAGs) degradation . These enzyme defects lead to severe and multisystemic effects and characteristic clinical symptoms (BELLETTATO et al., 2018; KUBASKI et al., 2020; SCRIVER, 2001). The MPS types are classified according to the affected gene. A spectrum of disease severity has been described. However, in many aspects, MPS physiopathology mechanisms are still unclear (FECAROTTA et al., 2020).

The emergence of next-generation sequencing (NGS) technologies has improved genomic research's range and accuracy to provide diagnostics and physiopathological insights into genetic diseases. Transcriptomic studies have been used to elucidate the mechanisms of complex phenotypes such as neurological damage in the MPS II (OMIM: 309900) murine model (DIROSARIO et al., 2009), central nervous system immunosuppression, and disease progression in MPS IIIB (OMIM: 252920) (SALVALAIO et al., 2017), among others in MPSs.

There are currently 13 publicly available transcriptomic studies, with microarray or RNA-Seq methods, of several MPS types. They comprehend six MPS types and four species. However, there is no database to integrate these data. The first database for lysosomal genes is The Human Lysosome Gene Database (BROZZI et al., 2013). It combines and analyzes information on lysosomal genes with miRNA data, transcription key regulators, and proteomic studies. However, there are few studies included and no recent update. Another database is RareLSD (AKHTER et al., 2019), which is restricted to genomic data and analysis for the 63 enzymes related to lysosomal diseases. Here we describe the first online database to encompass every available transcriptomic study about MPS. Also, IEMbase (LEE et al., 2018) presents data related to disease information, clinical symptoms and characteristics (stages like neonatal, infancy, childhood), biomarkers, and gene information.

2.2 MATERIALS AND METHODS

2.2.1 Data collection

MPSBase was designed to integrate all the available transcriptomics data about MPS. We used data from two public genome resources databases: GEO and ArrayExpress. The datasets included were generated with RNA-Seq or microarray technologies. A query for ‘Mucopolysaccharidosis” resulted in 13 studies that included 6 MPS types and four species. We used the chip non-normalized quantification (for microarrays) or raw counts (for RNA-Seq) data for all datasets available.

1.1.1 Data analysis

The methods for background correction and normalization in all microarray datasets were RMA and quantile. The packages were affy and oligo for Affymetrix genomic and exonic array, respectively, and limma for Illumina and Agilent arrays. For the RNA-Seq analysis, we chose the normalization procedure with the calcNormFactors function of the edgeR package, responsible also for the differentially expressed genes (DEGs) discrimination. We also applied the False Discovery Rate (FDR) for the differential test, with the cutoff FDR<0.05 and log2FC>1.5 considered for the enrichment analysis and insertion on the database. For each dataset, all possible pairwise combinations, within the same study, were performed, generating 53 comparisons.

To provide a representative and graphic overview of the DEGs relationship, we performed gene ontology and KEGG pathways enrichment with ClusterProfiler and EnrichKEGG packages. For automating this process, we developed an R package that integrates these analyses and allows the user to make sequential enrichments. This package outputs the results on tabular tables and dot plot charts in a one-step and flexible pipeline that we called autoGO. AutoGO can be downloaded from GitHub (github.com/ldiass/MPSBase/tree/master/autoGO). We created the dot plot and dendrogram heatmaps with ggplot2 and heatmap2 packages, respectively.

We obtained the annotation of the genes and their ontologies with the OrgDB R packages. To fill some gaps on the same annotation, we query the blank lines on the BiomaRt package using the Ensembl Gene ID as the primary key, supporting the

DEGs relationships on the database. For the dog microarray chip, exceptionally, we had to create our annotation due to the lack of a good one. We then performed a BLAST analysis with the probes against the dog cDNA genome and increased the number of annotated genes by 12%. This annotation table is available on the MPSBase website.

1.1.2 Overview of computational resources

The MPSBase is currently hosted at the UFRGS data processing center (CPD) and built mainly based on the following components: the WebNGINX system, with a virtual shared server composed of 16 VCPUs and 12Gb of RAM. The DBMS was developed in MySQL v. 5.5.31, supporting six tables containing the biological data and another 6 for front-end resources. The back-end application was developed with PHP v.5.6, using PDO API for database access. The front-end was built on CSS3, HTML, and Javascript frameworks.

1.1.3 Implementation

We developed a user-friendly web interface to help users browse the available experiments, search genes of interest or enriched terms. Thus, both possibilities are displayed on the main page. The web interface was split into the following three main pages: i) Species, ii) MPS type, iii) Tissue type. The user can also insert a specific term after the species definition, such as MGI (only Mus musculus), Ensembl ID, Entrez ID, GO or KEGG ID, and Gene Symbol. The OMIM gene ID and the Gene Symbol, according to the HGNC guidelines, are also available for the human datasets.

1.1.4 Browsing experiments by species

On a point-and-click interface, the user can browse through the experiment with three criteria. The first one is browsing by species. When clicked, the user will be redirected to a visual table with all four species comprehended on the MPS base. We designed the selection screen with figures provided by Biorender, representing the organisms. Once the species is defined, the system loads a table with all available datasets. This table will describe the experiment's platform, MPS type, and GEO

accession number. The user may choose one MPS type to be redirected to the MPS type selector. Otherwise, the user can click on the selected line on Dataset ID to access the table of available comparisons of that study. In doing so, a screen shows the comparison tables with the tissue, treatment (if available), and the number of replicates.

1.1.5 Browsing experiments by MPS types

The second option for starting to browse the experiments is by MPS type. Once clicked, the system loads a table with the six available MPS types on MPSBase. We tried to integrate more information about the diseases on the page, so we provide the diseases alias, external links to clinical description databases, such as OMIM and Orphanet, and the genes affected on each one. When the user selects one disease, a page with the available datasets is loaded with the particular species and platform.

1.1.6 Browsing experiments by Tissues

The last option is to select the samples' organ or cell origin. Due to the variety of tissue specificity, we chose to group samples into comprehensive options: blood, blood vessel, brain, *in vitro*, and liver. Some datasets comprehend more than one organ or cell origin. Just as with species selection, figures from Biorender are used to make a visual table. Once this option is defined, the next screen table, different from the two previous paths, will directly present the comparisons. The screen loads the specific tissue or cell of the comparisons, such as the cell lineage, the species, and the dataset code. Clicking on the latter, the user may retrieve all comparisons of the dataset.

The tables show the organization of the datasets in the MPSBase and can be downloaded. On the final step of the three previous browsing methods, a comparison table is available.

1.1.7 DEGs table

On the comparison table, the user has two options of results overview: genes and ontologies. If the user selects the first option, the screen loads the summary of

DEGs. This page shows the groups' information on the comparison, a heatmap dendrogram with the 50 top DEGs when possible, and a table describing these genes. The table contains the gene name, the gene symbol, the FDR, and the log2FC of the comparison. It is possible to download even the heatmap as the whole DEGs table on the website. The user can also interact on the page querying for a specific gene or enriched term on the comparison, such as one gene that is not on this top 50 table. Besides, it is possible to check the enriched terms related to each gene on the table or query this gene for the whole database.

1.1.8 Enriched terms table

The other option on the comparisons page is the Gene Ontology overview. When selected, the program will load a header with the info about the comparison, just as for the genes. Below, instead of the heatmap, four dot plot charts will appear. There is one chart for each of the three gene ontology categories and another one for enriched pathways. The dot plot seems a good alternative of representation, as it shows not only the terms and the number of genes, how representative they are compared to the whole set of genes on the term, and the adjusted p-value for the enrichment.

Below the graphs, there are two tabs; one loads the Gene Ontology terms and the other the KEGG pathways. In the GO enrichment table, we provide a description and identification of the ontology, the FDR value for the comparison, and a pair of options. The first is to query whether this term is enriched in all datasets, and the other is “related gene,” to query which genes of that term are differentially expressed in the comparison. Selecting the KEGG, the user will face the same data and options. Just as on the DEG table, the enriched terms table can be downloaded too.

1.1.9 String query

Another option to browse the database is to search for a gene or enriched term on MPSBase, only searching the query on the index page. First, the species must be selected, as the MPSBase comprehends different gene identifiers that may be restrained to one species. For example, only when Mus musculus is previously selected will the MGI option appear on the query’s chosen box. However, the selection with Gene Symbol, Entrez ID, Ensembl ID, GO ID, and KEGG pathway are

common for all species.

When a gene identifier is entered, the page loads some gene data, their full description, and the identifiers on the databases cited above. A larger table below this header is shown with all the comparisons where the gene was differentially expressed. Each row will also present the tissue, FDR, and log2FC values and a pair of links. Each link leads to the page of the comparison's DEGs or enriched terms overview described in the previous sections.

When a gene ontology identifier is entered, the page will load its description, classification, and links to KEGG pathways or AmiGO2 databases. Below, all comparisons in which the term is enriched will be listed. In each row will be the links for the dataset and comparison summaries, besides the FDR value. Just like with others on MPSBase, the user can download these tables.

2.3 CASE STUDY: USE OF MPSBASE APPLYED TO CARDIAC DISEASE IN ANIMAL MODELS OF MPS I AND MPS VII

Examples of applications may be related to biomarker discovery, modifier genes, and pathway analysis related to the disease. For instance, we performed a case study based on Gonzalez and colleagues (GONZALEZ et al., 2018). They showed the effect of losartan treatment in the MPS I mouse model, suggesting a role for ERK1/2 and G protein-coupled receptors (GPCR) signaling in cardiac disease. We then searched MPSBase for aortic tissue data to investigate DEGs in these pathways. GSE78889 (MPS I (OMIM: 607014), dog) and GSE30657 (MPS VII (OMIM: 253220), mouse) datasets were the only two with selected studies.

We gathered the orthologous dog's genes from the mouse ones to compare the genes on mouse and dog. We used the biomart package v.2.46.0 that presented the shared genes instead of performing a de novo alignment. We obtained 854 shared unique DEGs. Due to the microarray's dual-channel, the DEG analysis is less sensitive than NGS and single-channel chips (SÎRBU et al., 2012). This dataset is unique with dual-channel technology present in the MPSBase. For the mice dataset, there were 14238 DEGs and 12198 of these presented annotated homologous to dog genes (Fig. 2).

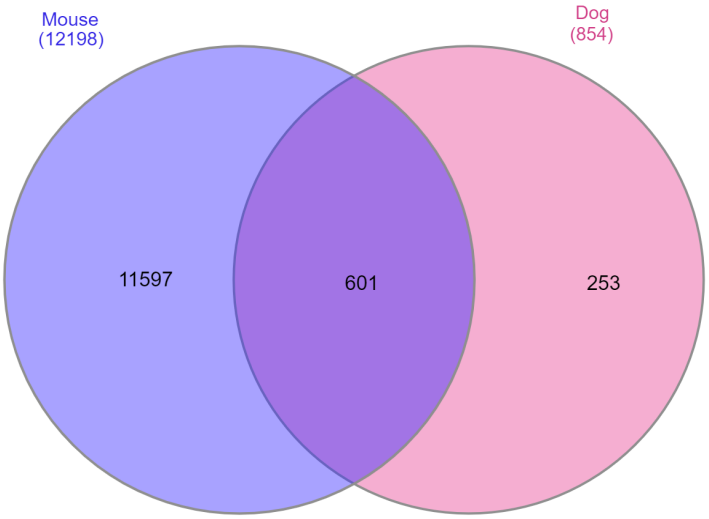


Figure 2 – Venn diagram of DEGs for both mouse and dog microarray experiments.
Source: Own authorship.

For the mouse ontologies, TNF signaling pathway, dilated cardiomyopathy, relaxin signaling, and autophagy. For the dog dataset, Alzheimer and Parkinson disease, AGE-RAGE signaling, prion disease, lysosome and axon guidance (Fig. 3). The ontologies related to phosphorylation, regulation of apoptosis, and the cell cycle's negative regulation are the top enriched ontologies for mice. For biological function, the most enriched pathways related to the dog dataset are axonogenesis, extracellular matrix organization, and cell activation. For the intersection of both datasets, the only KEGG pathways enriched were lysosome and apoptosis. The biological processes enriched are related to immune system processes.

1.2 DISCUSSION

The MPSBase is the first database dedicated to gene expression studies about MPS. It is a user-friendly platform where physicians and scientists can find new insights about MPS physiopathology, improving their research.

Our work gives a comprehensive gene expression analysis with more combinations of samples than the GEO datasets. In the GEO2R, the samples table shows only the 250 differentially expressed genes, and the option to download the whole table is time-consuming. Our database provides a more robust result and visualization - our heatmap contains the top 50 DEGs and dot plot about the enriched ontologies. We also allow the user to download the whole table of gene expression

and enriched ontologies, as the FDR, Fold Change, and other resources. Besides, we do not impose a 10-minute cutoff on job processing as in GEO2R.

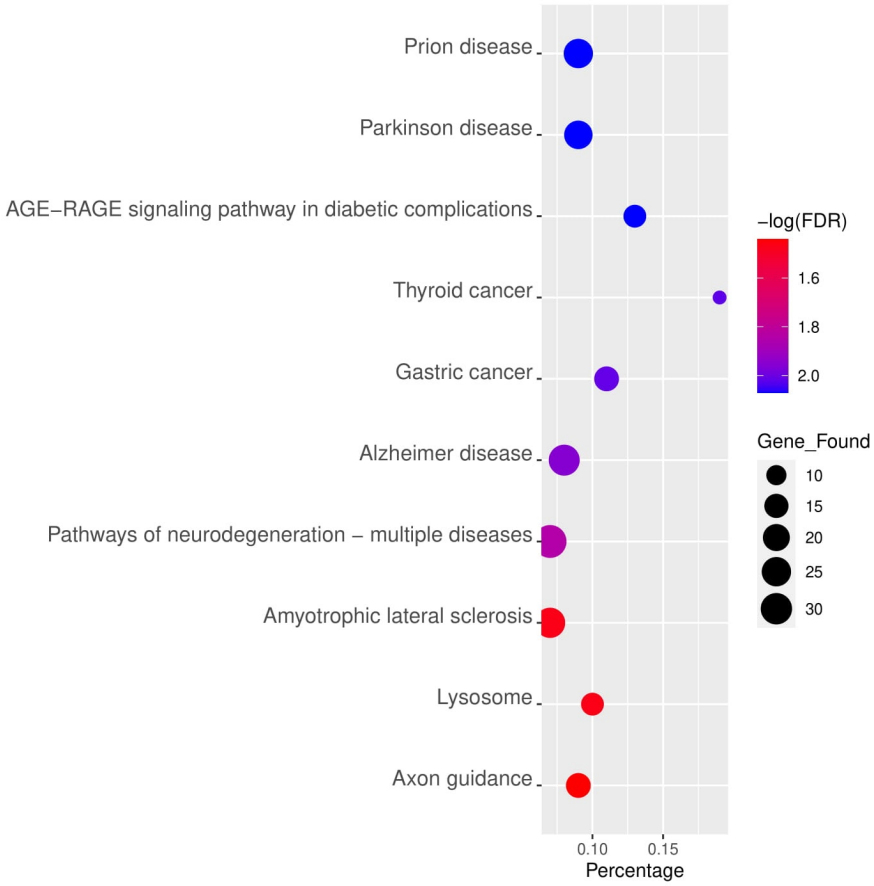


Figure 3 – Kegg Pathways enriched for the DEGs of *C. lupus*

Source: Own authorship.

To the best of the authors' knowledge, there are two works similar to MPSBase. The Human Lysosomal Gene Database (hLGDB) (BROZZI et al., 2013), is a database of human lysosomal genes and their regulation. It includes miRNA and Transcription Factors (TF) information related to lysosomal disorders in humans and with the murine orthology. The database contains a short literature review and links to ontology databases, such as Reactome, Uniprot, KEGG, and Gene Ontology, but only with these resources' citation. RareLSD (AKHTER et al., 2019) is a compendium of lysosomal enzymes and their related diseases. In this database, we can perform a sequence similarity analysis with BLAST to search the enzyme classification and disease information, like clinical symptoms, inheritance patterns, and related bioassays. Gene information contains cytogenetics, related pseudogenes, and paralogs. No graphics or plot visualization are found in this database.

With MPSBase, high-level users can check for differentially expressed genes

on different MPS types and see how these transcriptomic profiles are related to biological terms, such as biological pathways. Immune system-related processes, protein kinase signaling pathways, and extracellular signaling pathways are the most enriched ontologies.

Furthermore, MPSBase was used to identify oncogenic pathways related to cancer in the different MPS datasets (SILVA; SOARES; MATTE, 2020), to search biomarkers of neurological impairment in MPS I, MPS II, MPS IIIB, and MPS VII (SILVA; MATTE, 2020), and to perform systems biology analysis (SILVA; MATTE, 2021). There are several forms of applications of our tool, which can be used both in basic and in translational research to improve the understanding of MPS.

1.3 CONCLUSION

MPSBase is a new comprehensive database with all publicly available transcriptomic studies in MPS. This website allows easy access to search for differentially expressed genes and enriched terms in over 50 experimental conditions with MPS. The 54 comparisons include four species, 13 datasets of 6 MPS types, and two treatments. Our database comprehends annotations for over 50,000 enrichment terms and 92,000 genes for the four species. We intend to provide access to researchers with a low informatics background to this high-level standardized information of high-throughput experiments.

In summary, MPSBase provides a comprehensive collection of differentially expressed genes of MPS studies with enriched biological pathways in the form of tables and graphical representations. No database exists that exclusively addresses researchers' needs in the area of LSDs, specifically MPS. The possibility for the user to access the data by experiment, by species, and by type of MPS can help formulate new research hypotheses in the MPS scope. Additionally, MPSBase allows the user to find information about the disease in OMIM, Orphanet, and Genecards, among other resources, to compare the query results with information retained in specialized databases. The user-friendly web interface enables the researcher or clinician to quickly consult the data and discover biological pathways intrinsically related to multisystemic symptoms found in different tissues *in vivo* and *in vitro* models.

2 DISCUSSÃO GERAL

O presente trabalho apresentou o MPSBase como uma nova ferramenta para o fácil acesso de informações de grande valor para a pesquisa em MPSs. O site, além de integrar dados sobre os estudos de transcriptômica, também pode ser usado como um centro de informações ágil sobre as MPSs, uma vez que nos preocupamos em adicionar outros recursos externos, como informações do gene afetado em cada subtipo, diferentes nomes pelos quais a mesma doença é conhecida e *links* para as bases de dados OMIM <www.omim.org> (AMBERGER *et al*., 2019) e Orphanet <www.orpha.net>. Essas duas bases de dados fazem a síntese de aspectos clínicos, genéticos, farmacológicos e epidemiológicos e trazem de forma breve diversos trabalhos e suas principais descobertas dentro de uma miríade de doenças raras.

Dessa forma, também o MPSBase tem como objetivo dar visibilidade para um grupo de doenças raras, cujo acesso à informação ainda pode ser muito restrito, tanto para pacientes como para profissionais da saúde. Por isso, não se pode deixar de ressaltar que uma das bases de dados citadas, a Orphanet, também disponibiliza informações em português, espanhol e francês.

Cabe destacar que o site está indexado no Google e foi utilizado como base para outros dois estudos. O primeiro estudo é uma análise de vias de sinalização em câncer presentes nas MPS I,MPS II, MPS IIIA, MPS IIIB MPS VI, e MPS VII (SILVA; SOARES; MATTE, 2020). O estudo citado identificou 680 ontologias em 12 estudos. Dessas, foram pontuadas a condução axonal e sinalização por *Wnt* por sua relação com dano neuronal em vários tipos de MPS. O segundo estudo utilizou os dados de expressão gênica para análises de biologia de sistemas dos dados relacionados ao acometimento neurológico nas MPS I, MPS II, MPS IIIB e MPS VII (SILVA; MATTE, 2021). Este identificou diferentes vias imunológicas ativadas em MPSs como apresentação e processamento de antígeno mediados por MHC-I, e também ativação de sinalização por TLR4. Vias relacionadas a tipos específicos de MPSs, como a relação entre a ativação de complexo de ligação de ligação ao *cap* do RNAm, também foram descritas neste trabalho.

Além disso, o site foi divulgado nos seguintes eventos científicos: 40ª Semana Científica do Hospital de Clínicas de Porto Alegre, 5th Brazilian Student Council

Symposium, Brazilian Symposium on Bioinformatics 2020, WORLDSymposium 2021 e Great Lakes Bioinformatics Conference 2021.

No apêndice III podemos observar algumas telas do site, como a página inicial, as diferentes opções de busca por dados, e as páginas com as visualizações de tabelas e gráficos.

Visando ainda contribuir com a escassa disponibilidade de materiais nessa área de análise de transcriptômica, ainda mais quando restrito à língua portuguesa, os códigos para tais análises no R foram disponibilizados no repositório do MPSBase no github <https://github.com/ldiass/MPSbase>. Da mesma forma, apesar do uso de armazenamento de dados em nuvem já ser rotina, a comunicação entre os desenvolvedores e os usuários, nem sempre é feita da melhor forma, o que impede que essas ferramentas sejam exploradas da melhor maneira. Assim, também são trazidos neste trabalho, nos apêndices I e II, o modelo lógico e a linguagem de definição de dados que estruturam as relações entre as entidades do MPSBase. Essas informações são fundamentais para a manutenibilidade de ferramentas de bancos de dados (DA SILVA *et al.*, 2018). Mas, além disso, queremos estimular que mais trabalhos sejam desenvolvidos nessa área cobrindo novos tópicos que ainda são pouco explorados, não só em relação à genética humana, mas também em outros assuntos na área da biotecnologia cuja compreensão pode ser melhorada com similares ferramentas. Como perspectiva, propomos uma metanálise com os dados gerados para o MPSBase para fazer a identificação, de forma sistemática, de similaridades e diferenças nos perfis transcriptômicos entre as diferentes variáveis abordadas na base de dados.

REFERÊNCIAS

AKHTER, S. et al. RareLSD: a manually curated database of lysosomal enzymes associated with rare diseases. **Database**, v. 2019, p. baz112, 1 jan. 2019.

AMBERGER, J. S. et al. OMIM.org: leveraging knowledge across phenotype–gene relationships. **Nucleic Acids Research**, v. 47, n. D1, p. D1038–D1043, 8 jan. 2019.

BAXEVANIS, A. D.; BATEMAN, A. The Importance of Biological Databases in Biological Discovery. **Current Protocols in Bioinformatics**, v. 50, n. 1, jun. 2015.

BELLETTATO, C. M. et al. Inborn Errors of Metabolism Involving Complex Molecules. **Pediatric Clinics of North America**, v. 65, n. 2, p. 353–373, abr. 2018.

BROKOWSKA, J. et al. Expression of genes involved in apoptosis is dysregulated in mucopolysaccharidoses as revealed by pilot transcriptomic analyses. **Cell Biology International**, v. 45, n. 3, p. 549–557, mar. 2021.

BROZZI, A. et al. hLGDB: a database of human lysosomal genes and their regulation. **Database**, v. 2013, 1 jan. 2013.

CAMPOS, D.; MONAGA, M. Mucopolysaccharidosis type I: current knowledge on its pathophysiological mechanisms. **Metabolic Brain Disease**, v. 27, n. 2, p. 121–129, jun. 2012.

CLARKE, L. A. The mucopolysaccharidoses: a success of molecular medicine. **Expert Reviews in Molecular Medicine**, v. 10, p. e1, jan. 2008.

CONITEC. **SUS amplia idade para realização de transplante de células-tronco para doenças sanguíneas em idosos**. Disponível em: <http://conitec.gov.br/ultimas-noticias-3/sus-amplia-idade-para-realizacao-de-transplante-de-celulas-tronco-para-doencas-sanguineas-em-idosos>. Acesso em: 30 abr. 2021a.

CONITEC. **Ministério da Saúde amplia tratamento de doença rara no SUS com medicamento para mucopolissacaridose tipo VII**. Disponível em: <http://conitec.gov.br/ultimas-noticias-3/conitec-amplia-tratamento-de-doenca-rara-no-sus-com-medicamento-para-mucopolissacaridose-tipo-vii>. Acesso em: 30 abr. 2021b.

CUMMINGS, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. **Science Translational Medicine**, v. 9, n. 386, p. eaal5209, 19 abr. 2017.

DA CUNHA JAEGER, M. et al. HDAC and MAPK/ERK Inhibitors Cooperate To Reduce Viability and Stemness in Medulloblastoma. **Journal of Molecular Neuroscience**, v. 70, n. 6, p. 981–992, jun. 2020.

DA SILVA, W. M. C. et al. **Graph Databases in Molecular Biology**. (R. Alves, Ed.)Advances in Bioinformatics and Computational Biology. **Anais**...: Lecture Notes in Computer Science.Cham: Springer International Publishing, 2018

DHARSHINI, S. A. P. et al. Exploring Common Therapeutic Targets for Neurodegenerative Disorders Using Transcriptome Study. **Frontiers in Genetics**, v. 12, p. 639160, 19 mar. 2021.

DIETER, T. et al. **Introdução às mucopolissacaridoses**. Porto Alegre: Hospital de Clínicas de Porto Alegre, 2002. Disponível em: <http://www.ufrgs.br/redempsbrasil/sobre/introducao_as_mucopossacaridoses.pdf>. Acesso em: 30 abr. 2021.

DIROSARIO, J. et al. Innate and adaptive immune activation in the brain of MPS IIIB mouse model. **Journal of Neuroscience Research**, v. 87, n. 4, p. 978–990, 2009.

Doenças raras. Government site. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z-1/d/doencas-raras>. Acesso em: 30 abr. 2021.

FALCON, T. et al. Analysis of the Cancer Genome Atlas Data Reveals Novel Putative ncRNAs Targets in Hepatocellular Carcinoma. **BioMed Research International**, v. 2018, p. 1–9, 26 jun. 2018.

FECAROTTA, S. et al. Pathogenesis of Mucopolysaccharidoses, an Update. **International Journal of Molecular Sciences**, v. 21, n. 7, p. 2515, 4 abr. 2020.

FISHER, A.; KLEIN, T. **Newborn Screening**. Disponível em: <https://mps-society.mystagingwebsite.com/learn/education/fact-sheets/newborn-screening/>. Acesso em: 30 abr. 2021.

GIUGLIANI, R. et al. Guidelines for diagnosis and treatment of Hunter Syndrome for clinicians in Latin America. **Genetics and Molecular Biology**, v. 37, n. 2, p. 315–329, jun. 2014.

GONZALEZ, E. A. et al. Cathepsin B inhibition attenuates cardiovascular pathology in mucopolysaccharidosis I mice. **Life Sciences**, v. 196, p. 102–109, mar. 2018.

HEUSER, C. A. **Projeto de banco de dados**. Porto Alegre (RS): Bookman, 2009.

IMKER, H. J. 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. **Frontiers in Research Metrics and Analytics**, v. 3, 2018.

KUBASKI, F. et al. Mucopolysaccharidosis Type I. **Diagnostics**, v. 10, n. 3, p. 161, 16 mar. 2020.

LEE, J. J. Y. et al. Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. **Genetics in Medicine**, v. 20, n. 1, p. 151–158, jan. 2018.

LUALDI, S. et al. Enigmatic In Vivo iduronate-2-sulfatase (IDS) mutant transcript correction to wild-type in Hunter syndrome. **Human Mutation**, v. 31, n. 4, p. E1261–E1285, abr. 2010.

MINISTÉRIO DA SAÚDE. **Triagem Neonatal Biológica Manual Técnico**. 1. ed. Brasília/DF: Ministério da Saúde, 2016.

MINISTÉRIO DA SAÚDE. **Pacientes com doença genética terão tratamento específico no SUS**. Disponível em: <https://www.gov.br/pt-br/noticias/saude-e-vigilancia-sanitaria/2019/12/pacientes-com-doenca-genetica-terao-tratamento-especifico-no-sus>. Acesso em: 30 abr. 2021.

MOSKOT, M. et al. The Role of Dimethyl Sulfoxide (DMSO) in Gene Expression Modulation and Glycosaminoglycan Metabolism in Lysosomal Storage Disorders on an Example of Mucopolysaccharidosis. **International Journal of Molecular Sciences**, v. 20, n. 2, p. 304, jan. 2019.

MURRAY, R. K.; KEELEY, F. W. A Matriz Extracelular. In: **Harper: Bioquímica Ilustrada**. 8. ed. São Paulo: Manole, 1998. p. 667–685.

PEREIRA, P. C. B. et al. Sequenciamento total do exoma como ferramenta de diagnóstico de acidose tubular renal distal. **Jornal de Pediatria**, v. 91, n. 6, p. 583–589, nov. 2015.

POSWAR, F. DE O. et al. Lysosomal diseases: Overview on current diagnosis and treatment. **Genetics and Molecular Biology**, v. 42, n. 1 suppl 1, p. 165–177, 2019.

SALVALAIO, M. et al. Brain RNA-Seq Profiling of the Mucopolysaccharidosis Type II Mouse Model. **International Journal of Molecular Sciences**, v. 18, n. 5, p. 1072, 17 maio 2017.

SAUDUBRAY, J.-M.; GARCIA-CAZORLA, À. Inborn Errors of Metabolism

Overview. **Pediatric Clinics of North America**, v. 65, n. 2, p. 179–208, abr. 2018.

SCHIATTARELLA, G. G. et al. The Murine Model of Mucopolysaccharidosis IIIB Develops Cardiopathies over Time Leading to Heart Failure. **PLOS ONE**, v. 10, n. 7, p. e0131662, 7 jun. 2015.

SCRIVER, C. R. (ED.). **The metabolic & molecular bases of inherited disease**. 8th ed ed. New York: McGraw-Hill, 2001.

SHAPIRO, E. G.; EISENGART, J. B. The natural history of neurocognition in MPS disorders: A review. **Molecular Genetics and Metabolism**, v. 133, n. 1, p. 8–34, maio 2021.

SILVA, G. C. V.; MATTE, U. Cancer pathways are deranged in Mucopolysaccharidosis. 2020.

SILVA, G. C. V.; MATTE, U. Neuro-networks investigating the neurological impairment of mucopolysaccharidoses using a system biology approach. **Molecular Genetics and Metabolism**, v. 132, n. 2, p. S109, fev. 2021.

SILVA, G. C. V.; SOARES, L. D. F.; MATTE, U. Oncogenic Signaling Pathways in Mucopolysaccharidoses. In: SETUBAL, J. C.; SILVA, W. M. (Eds.). . **Advances in Bioinformatics and Computational Biology**. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. v. 12558p. 259–264.

SÎRBU, A. et al. RNA-Seq vs Dual- and Single-Channel Microarray Data: Sensitivity Analysis for Differential Expression and Clustering. **PLoS ONE**, v. 7, n. 12, p. e50986, 10 dez. 2012.

SWAROOP, M. et al. Patient iPSC-derived neural stem cells exhibit phenotypes in concordance with the clinical severity of mucopolysaccharidosis I. **Human Molecular Genetics**, v. 27, n. 20, p. 3612–3626, 15 out. 2018.

VILLALBA, C. **Tipos de MPS** Unpublished, , 2019. Disponível em: <http://rgdoi.net/10.13140/RG.2.2.23287.96162>. Acesso em: 6 maio. 2021

VILLALBA, G. C.; MATTE, U. Fantastic databases and where to find them: Web applications for researchers in a rush. **Genetics and Molecular Biology**, v. 44, n. 2, p. e20200203, 2021.

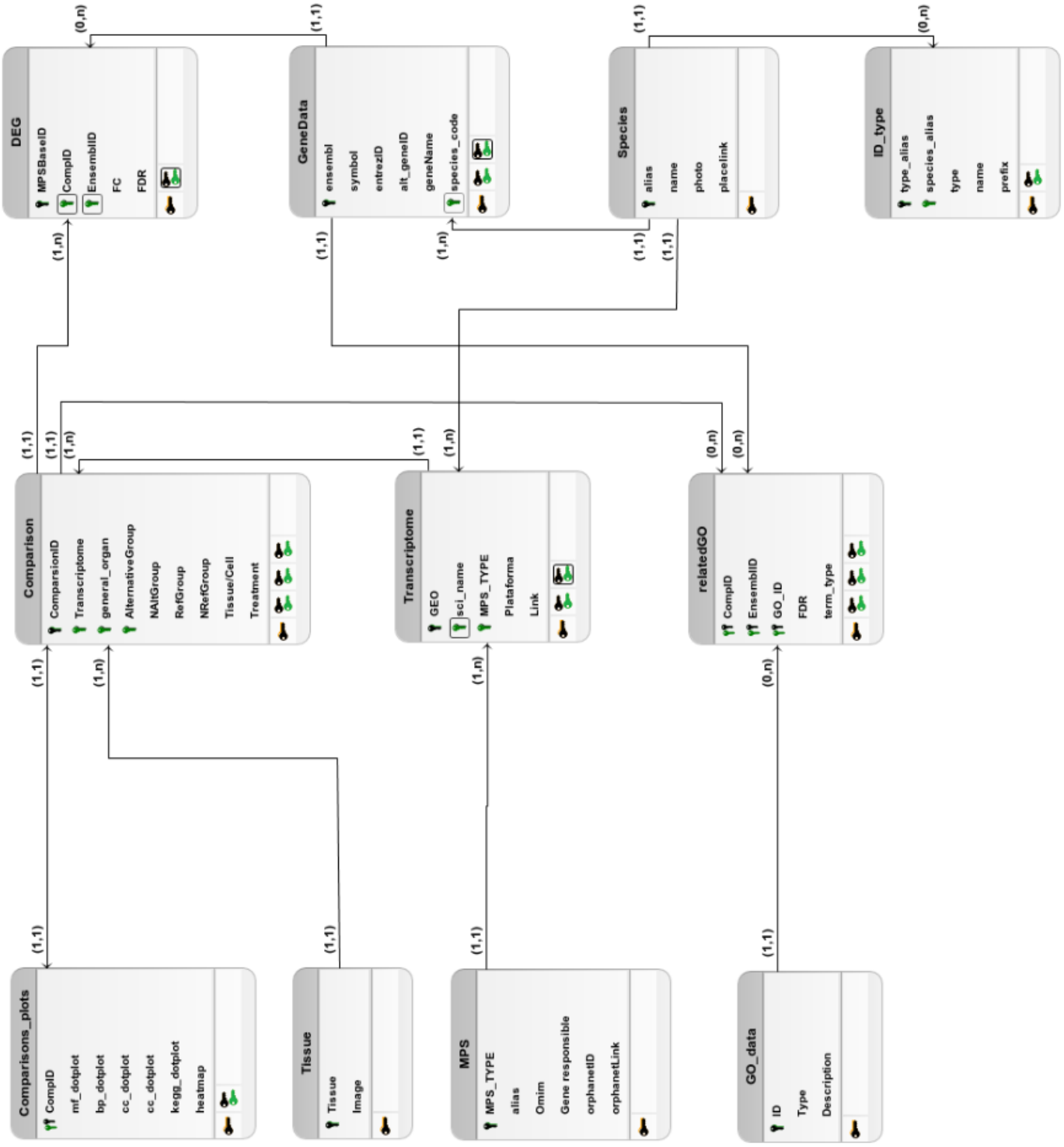
VILLANI, G. R. D. et al. Cytokines, neurotrophins, and oxidative stress in brain disease from mucopolysaccharidosis IIIB. **Journal of Neuroscience Research**, v. 85, n. 3, p. 612–622, 2007.

WOOD, T. et al. Expert recommendations for the laboratory diagnosis of MPS

VI. **Molecular Genetics and Metabolism**, v. 106, n. 1, p. 73–82, maio 2012.

APÊNDICE A – MODELO LÓGICO DO MPSBASE

Modelo lógico do MPSBase desenvolvido no BrModelo, seguindo a convenção de Heuser (2009).



APÊNDICE B – LINGUAGEM DE DEFINIÇÃO DE DADOS (DDL) PARA IMPLEMENTAÇÃO DO MPSBASE

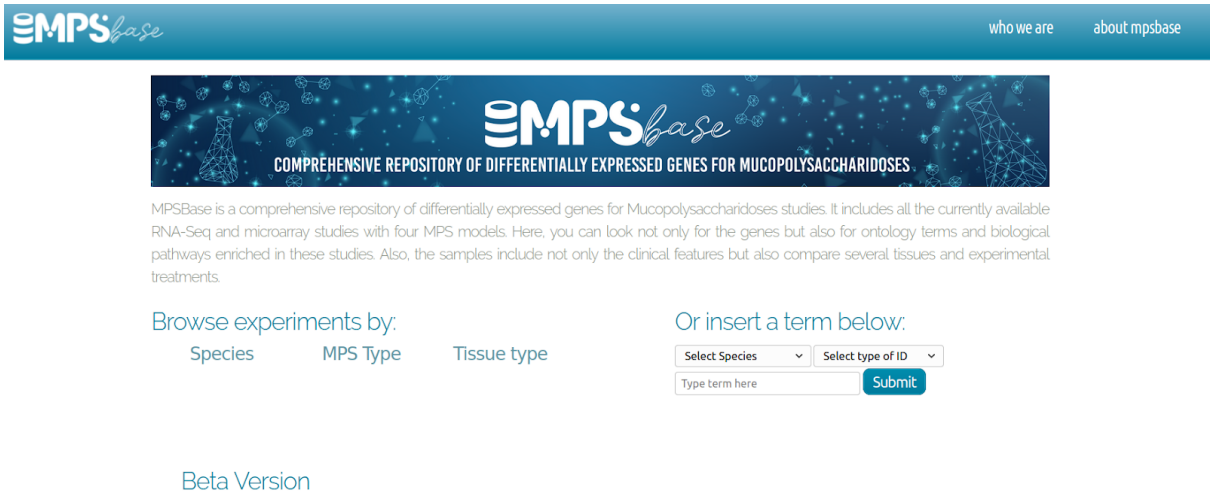
```
CREATE DATABASE `mpsbase`;

USE `mpsbase;

CREATE TABLE `tissue` (
    `tissue` varchar(20) NOT NULL,
    `image` blob NOT NULL,
    PRIMARY KEY (`tissue`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8mb4;

CREATE TABLE `MPS` (
    `MPS_TYPE` varchar(10) NOT NULL,
    `alias` varchar(50) NOT NULL,
    `Omim` varchar(150) NOT NULL,
    `Gene responsible` varchar(15) NOT NULL,
    `orphanetID` varchar(10) NOT NULL,
    `orphanetLink` varchar(300) NOT NULL,
    PRIMARY KEY (`MPS_TYPE`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8mb4;

CREATE TABLE `species` (
    `alias` varchar(5) NOT NULL,
    `name` varchar(20) NOT NULL,
    `photo` blob NOT NULL,
    `placelink` varchar(100) DEFAULT NULL
) ENGINE=MyISAM DEFAULT CHARSET=utf8mb4;

CREATE TABLE `GO_Data` (
    `ID` varchar(15) NOT NULL,
    `Description` varchar(200) NOT NULL,
    `Type` varchar(3) NOT NULL,
    PRIMARY KEY (`ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE `ID_types` (
    `name` varchar(20) NOT NULL,
    `type` varchar(3) NOT NULL,
    `prefix` varchar(7) NOT NULL,
    `species_alias` varchar(5) NOT NULL,
    `type_alias` varchar(12) NOT NULL,
    FOREIGN KEY (species_code) REFERENCES Species(alias) ON DELETE CASCADE
    ON UPDATE CASCADE,
    PRIMARY KEY (`type_alias`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

```
CREATE TABLE `Transcriptome` (
    `GEO` varchar(75) NOT NULL,
    `Link` varchar(75) DEFAULT NULL,
    `MPS_TYPE` varchar(8) DEFAULT NULL,
    `Plataforma` varchar(78) DEFAULT NULL,
    `sci_name` varchar(22) DEFAULT NULL,
    PRIMARY KEY (`GEO`),
    FOREIGN KEY (sci_name) REFERENCES Species(alias) ON DELETE NO ACTION ON
    UPDATE CASCADE,
    FOREIGN KEY (MPS_TYPE) REFERENCES MPS(MPS_TYPE) ON DELETE NO ACTION ON
    UPDATE CASCADE
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE `Comparison` (
    `ComparsionID` int(2) NOT NULL,
    `Transcritpome` varchar(10) DEFAULT NULL,
    `RefGroup` varchar(42) DEFAULT NULL,
    `NRefGroup` int(2) DEFAULT NULL,
    `AlternativeGroup` varchar(42) DEFAULT NULL,
    `NAltGroup` int(2) DEFAULT NULL,
    `Tissue/Cell` varchar(63) DEFAULT NULL,
    `Treatment` varchar(15) NOT NULL DEFAULT 'No',
    `general_organ` varchar(20) DEFAULT NULL,
    PRIMARY KEY (`ComparsionID`),
    FOREIGN KEY (Transcritpome) REFERENCES Transcriptome(GEO) ON DELETE
    CASCADE ON UPDATE CASCADE,
    FOREIGN KEY (general_organ) REFERENCES Tissue(Tissue) ON DELETE NO
    ACTION ON UPDATE CASCADE
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE `Comparsions_plots` (
    `CompID` int(11) NOT NULL,
    `mf_dotplot` mediumblob NOT NULL,
    `bp_dotplot` mediumblob NOT NULL,
    `cc_dotplot` mediumblob NOT NULL,
    `kegg_dotplot` mediumblob NOT NULL,
    `heatmap` mediumblob NOT NULL,
    PRIMARY KEY (`CompID`),
    FOREIGN KEY (CompID) REFERENCES Comparison(ComparsionID) ON DELETE
    CASCADE ON UPDATE CASCADE
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

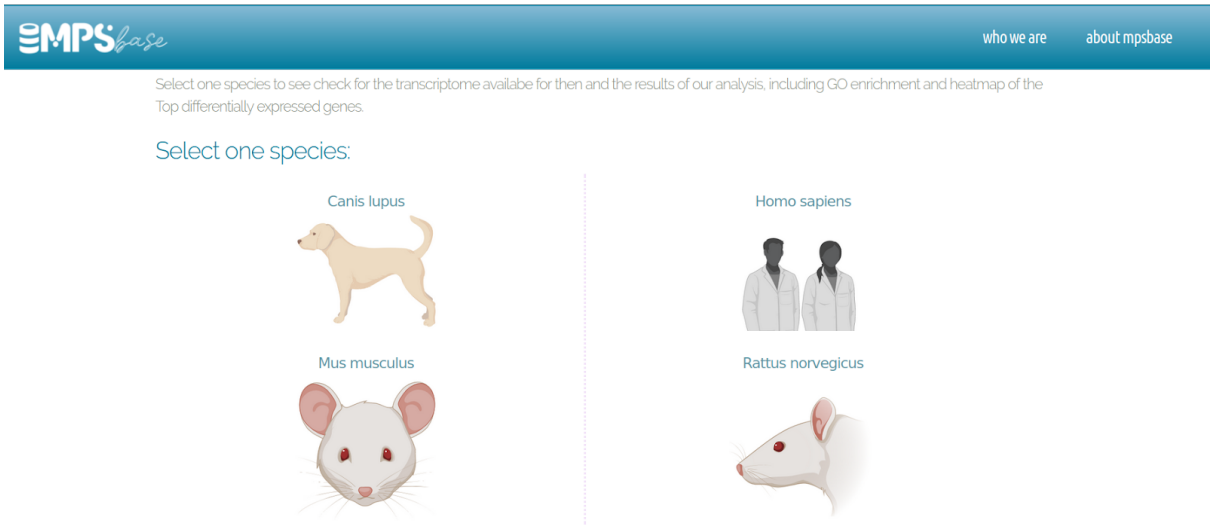
```
CREATE TABLE `DEG` (
    `MPSBaseID` varchar(12) NOT NULL,
    `EnsemblID` varchar(25) NOT NULL,
    `FDR` decimal(16,12) NOT NULL,
    `FC` decimal(10,8) NOT NULL,
    `CompID` tinyint(4) NOT NULL,
    PRIMARY KEY (`MPSBaseID`),
    FOREIGN KEY (CompID) REFERENCES Comparison(ComparsionID) ON DELETE
    CASCADE ON UPDATE CASCADE,
    FOREIGN KEY (EnsemblID) REFERENCES GeneData(ensembl) ON DELETE NO
    ACTION ON UPDATE CASCADE,
    KEY `EnsemblID` (`EnsemblID`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

CREATE TABLE `GeneData` (
    `entrezID` int(11) NOT NULL,
    `symbol` varchar(10) NOT NULL,
    `ensembl` varchar(25) NOT NULL,
    `alt_geneID` varchar(20) NOT NULL,
    `geneName` varchar(70) NOT NULL,
    `species_code` varchar(4) NOT NULL,
    PRIMARY KEY (`ensembl`),
    FOREIGN KEY (species_code) REFERENCES Species(alias) ON DELETE NO ACTION
    ON UPDATE CASCADE,
    KEY `symbol` (`symbol`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

CREATE TABLE `relatedGO` (
    `EnsemblID` varchar(25) NOT NULL,
    `CompID` tinyint(3) NOT NULL,
    `FDR` decimal(16,12) NOT NULL,
    `GO_ID` varchar(15) NOT NULL,
    `term_type` varchar(3) NOT NULL,
    PRIMARY KEY (`EnsemblID`,`CompID`,`GO_ID`),
    FOREIGN KEY (CompID) REFERENCES Comparison(ComparsionID) ON DELETE
    CASCADE ON UPDATE CASCADE,
    FOREIGN KEY (EnsemblID) REFERENCES GeneData(ensembl) ON DELETE NO ACTION
    ON UPDATE CASCADE,
    FOREIGN KEY (GO_ID) REFERENCES GO_data(ID) ON DELETE NO ACTION ON UPDATE
    CASCADE,
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

APÊNDICE C: CAPTURAS DE TELA DO MPSBASE

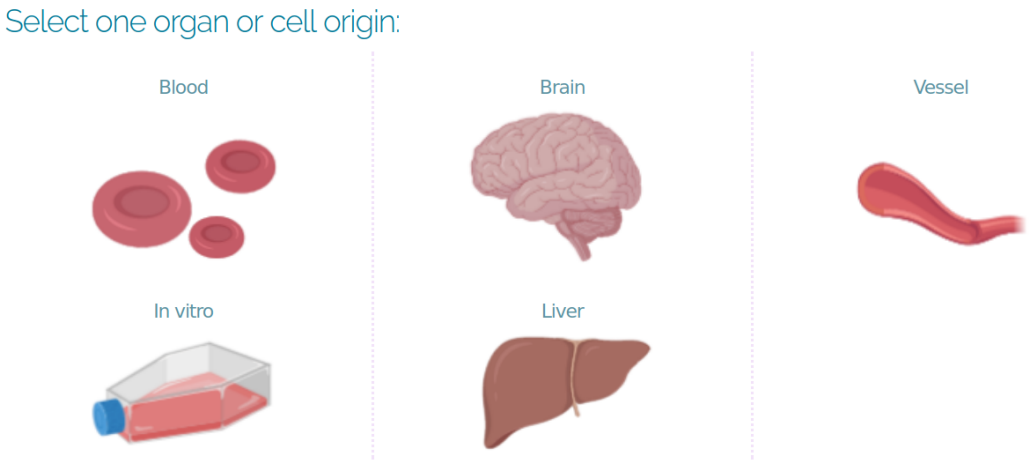
Página inicial:



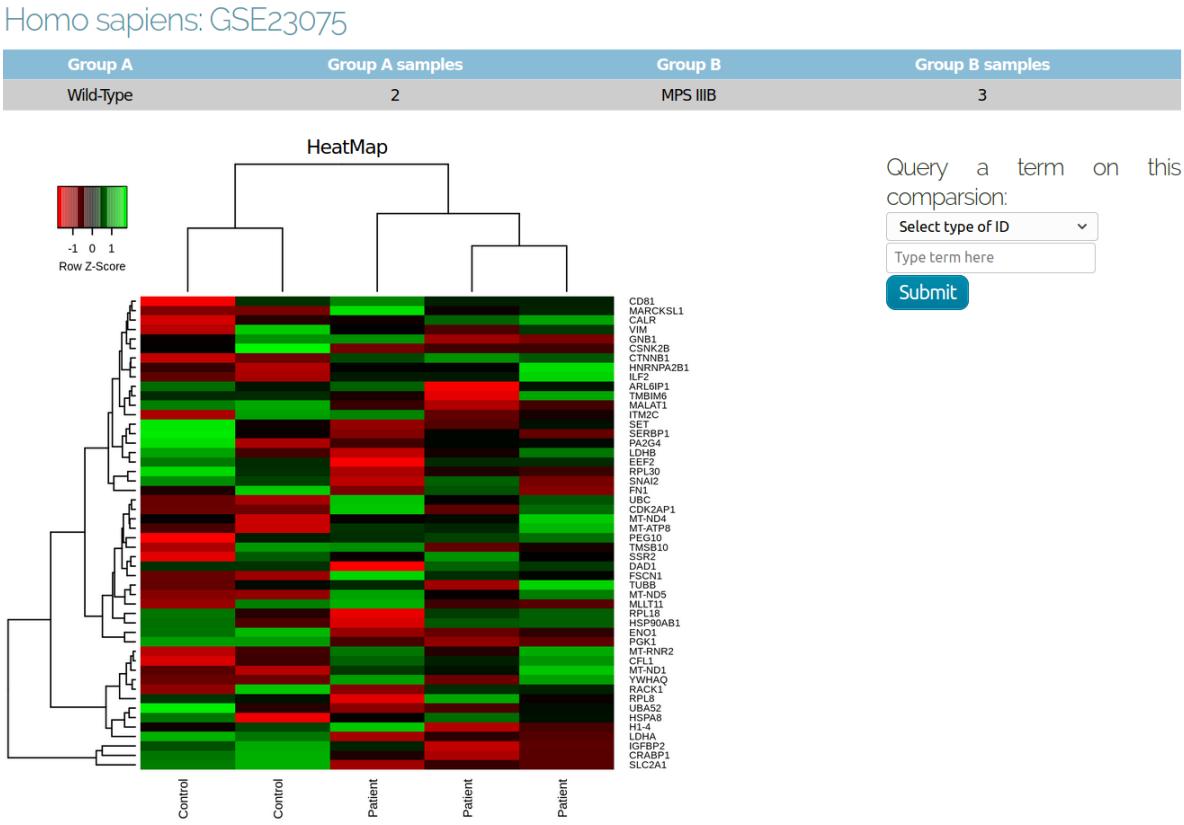
Seleção de experimentos por espécies:



Seleção de experimentos pelos tecidos:

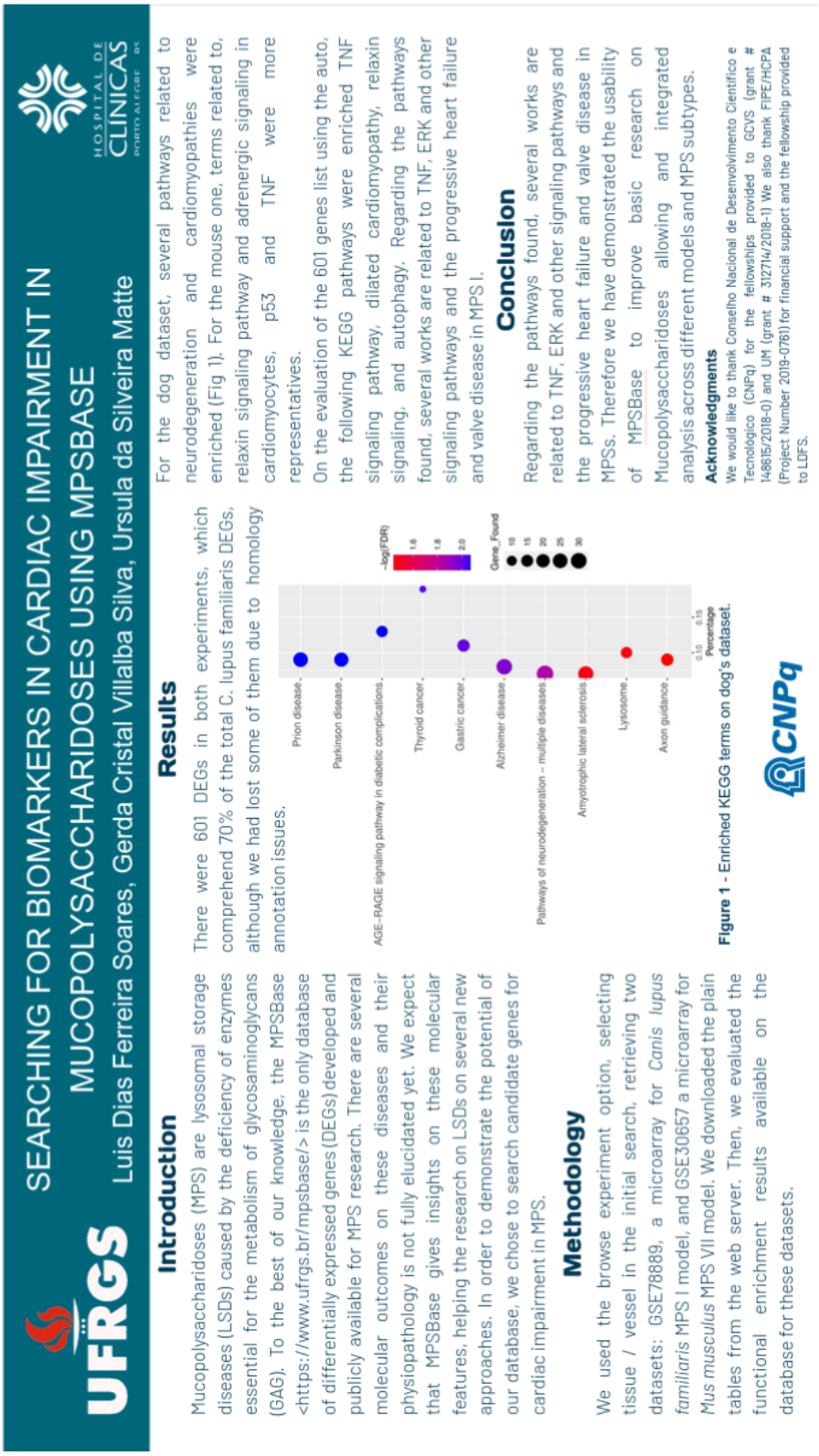


Exibição de heatmap para genes com expressão diferencial:



APÊNDICE D: DIVULGAÇÃO DO TRABALHO

Pôster apresentado no congresso Great Lakes Bioinformatics conference 2021:



SEARCHING FOR BIOMARKERS IN CARDIAC IMPAIRMENT IN MUCOPOLYSACCHARIDOSES USING MPSBASE

Luis Dias Ferreira Soares, Gerda Cristal Villalba Silva, Ursula da Silveira Matte

UFRGS

HOSPITAL DE CLÍNICAS PORTO ALEGRE RS

Introduction

Mucopolysaccharidoses (MPS) are lysosomal storage diseases (LSDs) caused by the deficiency of enzymes essential for the metabolism of glycosaminoglycans (GAG). To the best of our knowledge, the MPSBase <https://www.ufrgs.br/mpsbase/> is the only database of differentially expressed genes (DEGs) developed and publicly available for MPS research. There are several molecular outcomes on these diseases and their physiopathology is not fully elucidated yet. We expect that MPSBase gives insights on these molecular features, helping the research on LSDs on several new approaches. In order to demonstrate the potential of our database, we chose to search candidate genes for cardiac impairment in MPS.

Methodology

We used the browse experiment option, selecting tissue / vessel in the initial search, retrieving two datasets: GSE78889, a microarray for *Canis lupus familiaris* MPS I model, and GSE30657 a microarray for *Mus musculus* MPS VII model. We downloaded the plain tables from the web server. Then, we evaluated the functional enrichment results available on the database for these datasets.

Results

There were 601 DEGs in both experiments, which comprehend 70% of the total C. lupus familiaris DEGs, although we had lost some of them due to homology annotation issues.

Figure 1 - Enriched KEGG terms on dog's dataset.

For the dog dataset, several pathways related to neurodegeneration and cardiomyopathies were enriched (Fig 1). For the mouse one, terms related to, relaxin signaling pathway and adrenergic signaling in cardiomyocytes, p53 and TNF were more representatives.

On the evaluation of the 601 genes list using the auto, the following KEGG pathways were enriched TNF signaling pathway, dilated cardiomyopathy, relaxin signaling, and autophagy. Regarding the pathways found, several works are related to TNF, ERK and other signaling pathways and the progressive heart failure and valve disease in MPS I.

Conclusion

Regarding the pathways found, several works are related to TNF, ERK and other signaling pathways and the progressive heart failure and valve disease in MPSs. Therefore we have demonstrated the usability of MPSBase to improve basic research on Mucopolysaccharidoses allowing and integrated analysis across different models and MPS subtypes.

Acknowledgments

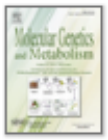
We would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the fellowships provided to GCVS (grant # 148615/2018-0) and UM (grant # 312714/2018-1) We also thank FIPE/HCPA (Project Number 2019-0761) for financial support and the fellowship provided to LDFS.

CNPq

Apresentação do trabalho no *WORLDSymposium* 2021 com publicação na revista Molecular Genetics and Metabolism:

ELSEVIER

Molecular Genetics and Metabolism
Volume 132, Issue 2, February 2021, Pages S102-S103



MPSBase: Comprehensive repository of differentially expressed genes for mucopolysaccharidoses studies

Luis Dias Ferreira Soares, Gerda Cristal Villalba Silva, Ursula da Silveira Matte

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.ymgme.2020.12.249

Previous article in issue Next article in issue

Mucopolysaccharidoses (MPS) are lysosomal diseases caused by the deficiency of enzymes essential for the metabolism of extracellular matrix components such as glycosaminoglycans (GAG). With the advancement of large-scale sequencing technologies, the tendency is for such methodologies to become recurrent and often useful in the diagnosis of rare genetic diseases. For data from different types of MPS, there are no specific databases or software filtering and presenting their results in a one-click approach. In order to comprehend the physiopathology alterations due to the lysosomal accumulation resulting from enzymatic alterations and their secondary outcomes, this work presents a database for all differentially expressed genes from the different MPS study models, called MPSBase. Currently, we included a total of 13 studies previously deposited in a public functional genomics (GEO) data repository being that from MPS type I, MPS II, MPS IIIA, MPS IIIB, MPS VI and MPS type VII. The organisms represented by datasets from *Canis lupus familiaris*, *Homo sapiens* and *Mus musculus*. Significantly differentially expressed genes are related to many processes, such as Immune response, cellular adhesion, calcium regulation, autophagy, golgi vesicle transport, axon guidance, neuroinflammation, B cell activation, cell proliferation, angiogenesis, EGF receptor signaling pathway, Ras and TGF-beta signaling. Therefore, the development of analytical and automation strategies accessible to health professionals is essential for the proper planning and choice of appropriate methodologies, generally improving the diagnosis, prognosis, and development of personalized therapeutic approaches for the treatment of rare disorders. The interactive web-based platform is available on www.ufrgs.br/mpsbase/. We would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the fellowships provided to GCVS and UM and to FIPE/HCPA for the fellowship provided to LDFS.

Pôster apresentado na 40ª semana científica do Hospital de Porto Alegre:

Semana CIENTÍFICA do HCPA

MPSBase: Comprehensive repository of differentially expressed genes for Mucopolysaccharidoses

Luis Dias Ferreira Soares [1,2,3], Gerda Cristal Villalba Silva[1,3], Ursula Matte[1,3,4]

[1] Cell, Tissues and Genes Laboratory, Experimental Research Center, Hospital de Clínicas de Porto Alegre, RS, Brazil
[2] Graduation Program on Biotechnology/Bioinformatics, Federal University of Rio Grande do Sul, RS, Brazil
[3] Bioinformatics Core, Hospital de Clínicas de Porto Alegre, RS, Brazil
[4] Post-Graduation Program on Genetics and Molecular Biology, Federal University of Rio Grande do Sul, RS, Brazil

Introduction

Mucopolysaccharidoses (MPS) are lysosomal storage diseases caused by disturbances in enzymes essential for the metabolism of components of the extracellular matrix called glycosaminoglycans (GAG). To understand the physiopathology and alterations due to the lysosomal accumulation resulting from enzymatic deficiencies and their secondary outcomes can improve the diagnosis and treatment of rare genetic diseases.

Aims

- To catalog all the MPS transcriptomic studies present in public databases;
- To develop a public repository and web interface for differentially expressed genes and ontologies across the different MPS types

Methods

We developed our database including 13 studies previously deposited in a public functional genomics (GEO, https://www.ncbi.nlm.nih.gov/geo/) data repository for six MPS types. The server is hosted in the CPD-UFRGS facility and available at https://www.ufrgs.br/mpsbase/. The site was constructed in PHP and the analysis are implemented on R. The organisms represented by datasets are *Canis lupus familiaris*, *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus* (**Fig 1**). The user can search the differentially expressed genes and ontologies by species, MPS type, or tissue type. For each comparison, a heatmap with the 50 top expressed genes is available and dotplots for the 30 top ontologies divided by biological process, cellular component, KEGG pathways, and molecular function. This data is also fully available in tables.

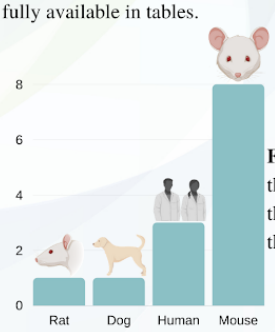


Figure 1: Barchart showing the distribution of species on the experiments covered in this study

Results

Currently, the MPSBase has more than 53 comparisons across different tissues or cells, and several treatments. We classified the tissue type into 5 categories: Blood, Brain, Liver, Vessel and *In vitro*. For the differentially expressed genes, we provided the Ensembl ID, Gene name, FDR and Fold Change values per gene, and other Resources. For each MPS type available, the description, OMIM code, Orphanet code, GeneCards link and the number of datasets in the database are found in the option "Browse experiments by MPS type".

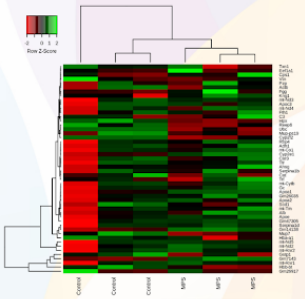


Figure 2: Example of Heatmap of the dataset GSE77689 representing the transcriptomic profile.

Figure 3: Example of dotplot of the dataset GSE77689 representing the KEGG enrichment.

Concluding Remarks

The MPSBase is the first database dedicated to gene expression studies about Mucopolysaccharidoses. It is a user-friendly platform where physicians and scientists can find new insights about MPSs' physiopathology, improving their research. With MPSBase, high-level users can check for differentially expressed genes on different MPS types and also see how these transcriptomic profiles are related to biological terms, such as biological pathways.

For more informations, please visit:
ufrgs.br/mpsbase

Promoção

Finantial Suport
FIPE (Project Number 2019-0761), Hospital de Clínicas de Porto Alegre, and CNPq.