吴玥,王珏,薛婧,等. 心脏病动物模型比较转录组学数据库的构建[J]. 中国比较医学杂志, 2023, 33(3): 75-81.
Wu Y, Wang J, Xue J, et al. Establishment of a comparative transcriptomics database of heart disease animal models [J]. Chin J Comp Med, 2023, 33(3): 75-81.
doi: 10.3969/j.issn.1671-7856.2023.03.010

心脏病动物模型比较转录组学数据库的构建

吴 玥,王 珏,薛 婧,魏 强,杨志伟,孔 琪*

(中国医学科学院医学实验动物研究所,国家人类疾病动物模型资源库,国家卫生健康委员会人类疾病比较医学重点实验室,新发再发传染病动物模型研究北京市重点实验室,北京市人类重大疾病实验动物模型工程技术研究中心,北京 100021)

【摘要】 目的 整合心脏病动物模型与心脏病患者基因表达谱,系统表征基因表达差异,筛选差异基因,分析基因表达异同,为动物实验与比较医学研究提供依据。**方法** 从公共数据库下载心脏病动物模型与患者的基因芯片、转录组数据。按照物种、病种、病程等对数据进行整理、标准化处理、批次校正,转换为数据库可用格式,搭建在线数据库。**结果** 心脏病动物模型比较转录组学数据库实现基因、功能通路全局检索查询,分析基因在不同物种(人与小鼠)、不同病种、不同组织、不同病程的表达。交互式展示样品表达谱,可视化基因表达量,进行比较分析。在线差异基因挖掘,功能通路富集鉴定,探索差异、易感基因参与调控的生物过程,解释易感机制,提供在线数据分析。**结论** 本文建立了心脏病动物模型比较转录组学数据库,为心脏病基因水平研究提供数据资源与在线分析工具。

【关键词】 心脏病;数据库;动物模型;基因表达;比较分析
【中图分类号】 R-33 **【文献标识码】** A **【文章编号】** 1671-7856 (2023) 03-0075-07

Establishment of a comparative transcriptomics database of heart disease animal models

WU Yue, WANG Jue, XUE Jing, WEI Qiang, YANG Zhiwei, KONG Qi*

(Institute of Laboratory Animal Sciences, CAMS & PUMC, National Human Diseases Animal Model Resource Center, NHC Key Laboratory of Human Disease Comparative Medicine; Beijing Key Laboratory for Animal Models of Emerging and Reemerging Infectious Diseases, Beijing Engineering Research Center for Experimental Animal Models of Human Critical Diseases, Beijing 100021, China)

【Abstract】 Objective To integrate the gene expression profiles of heart disease animal models and patients, systematically characterize the differences in gene expression, screen differentially expressed genes, analyze the similarities and differences of gene expression, and provide a basis for animal experiments and comparative medicine research. **Methods** Microarray and transcriptome data of heart disease animal models and patients were downloaded from a public database. Data were sorted by species and disease type and stage, standardization and removal of batch effects were performed, data were converted in the available format for the database, and an online database was built. **Results** The comparative transcriptomics database of heart disease animal models realized global searching and queries of genes and functional pathways, and analysis of gene expression in humans and mice, diseases, tissues, and stages. It also

[基金项目]中国医学科学院医学与健康科技创新工程项目(2021-I2M-1-034);北京市自然科学基金资助项目(M21027);国家重点研发计划项目(2021YFF0702800)。
[作者简介]吴玥(1993—),女,助理研究员,研究方向:比较医学数据库建设。E-mail:wuyue@cnilas.org
[通信作者]孔琪(1978—),男,研究员,研究方向:比较医学,生物信息学。E-mail:kongqi@cnilas.org

interactively displayed sample expression profiles, visualized gene expression, and facilitated comparative analysis. The database allowed online differential gene mining, enrichment and identification of functional pathways, exploration of the biological processes of susceptibility genes involved in regulation, explanation of the susceptibility mechanism, and online data analysis. **Conclusions** A comparative transcriptomic database of heart disease animal models was established to provide data resources and online analysis tools to study heart disease.

【**Keywords**】 heart disease; database; animal model; gene expression; comparative analysis

Conflicts of Interest: The authors declare no conflict of interest.

心脏病(heart disease)是心脏发生病变的疾病总称,由心脏结构受损或功能异常引起。心脏病已成为影响全球人口健康的重大复杂疾病之一[1]。使用动物模型能够帮助我们更好理解心脏病的发病机制,评价治疗策略,寻找更高效和可靠的治疗手段。在过去的几十年里,已经建立了数百种心脏病动物模型。常用的实验动物包括小鼠、大鼠、豚鼠、兔、猪等。多采用外科手术、基因工程或药物诱导等造模方法[2-3]。

比较转录组学(comparative transcriptomics)是从 RNA 水平研究不同物种或品系特定组织基因表达和相互关系的一种生物信息学研究方法。对于致病基因的识别、揭示基因在疾病中的作用、分析药物的药效等方面很有价值[4-5]。心脏病的发病与基因表达功能障碍相关,差异基因表达分析、聚类分析在心脏病发病机制研究中具有重要作用。已经发现很多与心脏病发病相关的潜在靶点,通过不同程度的表达导致心脏病[6-8]。

测序技术的发展催生了大量心脏病动物模型相关的转录组学数据,为深入了解和挖掘基因功能创造了条件,但缺乏相关的数据库把分散的心脏病动物模型相关数据集进行收集与整合,并形成心脏病动物模型与患者的比较转录组学分析数据[9]。我们建立了心脏病动物模型比较转录组学数据库,基于单位自有的心脏病动物模型转录组学数据,同时从基因表达数据库(gene expression omnibus, GEO)、芯片表达数据库(ArrayExpress)等公共数据平台收集了与心脏病动物模型相关的基因表达芯片数据集和 RNA-seq 数据集,对基因表达谱进行可视化展示并进行跨物种比较分析,对于心脏病相关研究具有重要作用。

1 实验方法

1.1 数据采集和校正

从 NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/)、EBI ArrayExpress (https://www.ebi.ac.uk/arrayexpress/)公共数据库检索并获取心脏病动物模型与患者的基因芯片、转录组数据,所选数据集符合以下标准:(1)Microarray 或 RNA-seq 研究;(2)数据集样本量>10;(3)样品来源明确,背景信息注释清晰。下载符合标准的数据集后使用 SVA 自动检测批次,ComBat 校正批次,去除批次效应,校正 SVA 鉴定的批次,基于这三者,构建自动在线批次校正工具用于网站中的数据分析。

采用标准化处理方法,获取校正后的基因表达数据,并对芯片探针进行重注释。同时下载实验背景信息,程序脚本与人工相结合校对数据一致性(浓度计量单位、时间计量单位、大小写等方面),例如数据分组信息是否完整明确、有无错漏数据等,缺失的数据依据实际情况填入 NULL 或 Unrecorded,并对照相应文献人工校正或添加品系、年龄、基因型信息。

1.2 数据清洗和质量控制

使用 GEO_metadata.r 对数据表进行清洗,获得包括物种信息、品系信息、年龄、组织、数据类型等实验背景信息。根据 fastq 的质量值(Phred Qulity Score)对数据进行质量控制、标准化处理后系统整合基因表达谱数据,并识别心脏病动物模型和患者的差异表达基因。

1.3 数据库平台的构建及访问

前端使用 HTML5 和 Plotly/ECharts 组件,使用 nginx(v 1.16.1)提供网站服务,后端采用 Django 框架,MySQL(v 8.0)储存数据,编程语言采用 python(v 3.6),部署在 CentOS Linux server(v 7.7)服务器。

数据库网址:https://cvd.com-med.org.cn。数据库为英文版,可免费公开访问使用。

2 结果

2.1 数据库结构与内容

本数据库提供了用户友好的界面,包含检索、数据分析、数据统计、帮助等功能(图 1)。分析工具

可进行差异基因挖掘和多个目标基因表达分析研究。帮助页面为用户提供了详细的介绍和使用教程。数据库包括 2 个物种(人、小鼠)、5 种病种、6 个年龄、7 个组织,共 311 个样本,最终获得的样品属性信息见表 1。

2.2　检索方式

本数据库提供全局检索(global search)与高级检索(advanced search)两种检索方式。首页提供全局检索功能,类似谷歌的全局模糊匹配搜索,用户在输入检索关键词后即可搜索,并获取最为可能的匹配结果,供用户进一步聚焦查询。用户可通过基因名称、功能、通路等信息检索,结果显示所有匹配内容。详情页展示 Gene symbol、Alias、Entrez ID、Gene description、GO term、Reactome pathway 信息。

高级检索可以帮助查找用户感兴趣的样本,采用多参数自定义检索方法,用户可自行选择检索字段,自由条件组合检索参数如物种名称、品系名称、年龄、病程、组织类型、数据类型等信息,根据不同层级搜索并浏览样本。搜索结果显示在多功能表格中,并且可跳转至对应的 GSM 数据集,方便用户查阅详细的样本信息及参考文献信息。通过联动式饼图显示统计信息,并可通过单击扇区进一步过滤搜索结果。

2.3　基因表达比较分析

2.3.1　单基因表达比较分析

对于每个基因(以 MYH6 为例),数据库采用箱形图的形式展示基因表达谱并显示 *P* 值,同时可以切换 Raw 或 Log2 两种数据转换方式进行展示。用户可以直观比较疾病动物模型和患者的基因表达差异,也可以查看这个基因在不同病种、不同组织、

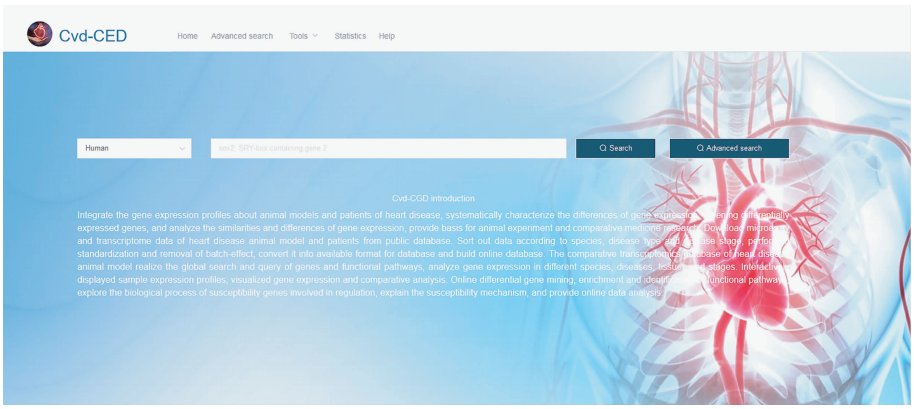

图 1　数据库首页

Figure 1　Homepage of the database

表 1　数据库样品属性信息表

Table 1　Sample property information of database

样本字段 Sample name	格式 Form	获取方式 Obtain access
样本名 Sample	disease_species_tissue_time_GSE number_genotype	Splice group information
物种 Species	Human/mouse...	Metadata
项目 ID Project ID	GSEXXX	Search selection
数据类型 Assay type	Microarray/RNA-seq	Search+filter+metadata
病种 Disease type	Disease type full name	Metadata+literature supplement+manual correction
病程分期 Disease stage	Early/mild/end stage	Metadata+literature supplement+manual correction
组织类型 Tissue	Full name of tissue	Metadata
品系 Strain	Full name of strain	Metadata+literature supplement+manual correction
年龄 Age	1 week/5 weeks/10 weeks...	Metadata+literature supplement+manual correction
时间 Time	12 h/24 h/48 h...	Metadata+manual correction
转基因名称 Transgene	WT or transgene name	Metadata+literature supplement+manual correction
实验分组 Treatment	Treat/mock	Metadata+manual correction
组名 Boxplot subgroup	Description of boxplot dataset	Metadata
是否配对 Paired	True/false	Manual correction
GSM	GSM number	Search selection

不同病程中的表达情况(图 2)。除此之外,当鼠标悬浮疾病名称、组织名称缩写时会展示缩写词的全部名称。

2.3.2 多个目标基因表达分析

用户可以输入多个基因名称(以 SOX2、MTHFR、GATA4、TBX20 为例),并选择组合参数,例如不同物种、不同病种、不同组织、不同病程、不同年龄、不同测序数据类型及绘图类型。选择数据归一化方式(raw、Z 值或 log2 转换),最后可选择折线图、柱形图、热图、箱形图、相关性图等形式可视化展示多基因表达情况(图 3)。所有可视化图为交互式,并且可以使用特定的 toolkit 工具包对颜色、大小和样式等进行修改。

2.4 整合差异性分析

用户可以选择并组合不同参数进行在线实验设计,如组合不同物种、病种、组织、年龄、病程等,动态生成所需的分析样本(图 4)。系统会将所选样本表达数据进行整合,并自动评估批次效应,检测异常样本,最后将去除批次效应的表达数据用于样本相关性分析。用户也可以根据生成的样本聚类相关性热图和 PCA 散点图过滤异常样本进行下一步分析(图 5)。

确认选择的样本和参数后可填写邮箱,接收分析结果。分析结果以在线报告的形式展示,报告包括样本相关性分析、PCA 分析、差异表达基因热图、火山图和功能富集分析气泡图(图 6)。可以进行交互式调整,并且可以将报告导出为 PDF 格式,满足在线分析云平台的基本需求。

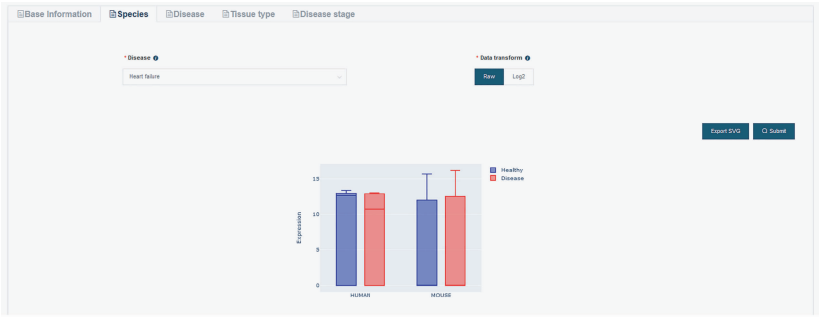

图 2 单基因表达比较分析

Figure 2 Single gene expression comparative analysis

图 3 多基因表达分析

Figure 3 Multiple gene expression analysis

2.5　数据管理后台

数据管理后台对网站栏目及数据进行储存管理。可管理网站名称、网站简介、新闻动态、文献、使用手册等文字信息;首页轮播图、Logo、友情链接等信息;管理数据的更新和修改,如更新物种基因注释信息、同源基因信息、富集分析信息、样本表、基因表达文件、特殊字段简称和全称对照信息。用户也可以通过后台对单基因表达谱横坐标属性进行排序,调整在网页中的展示方式(图 7)。

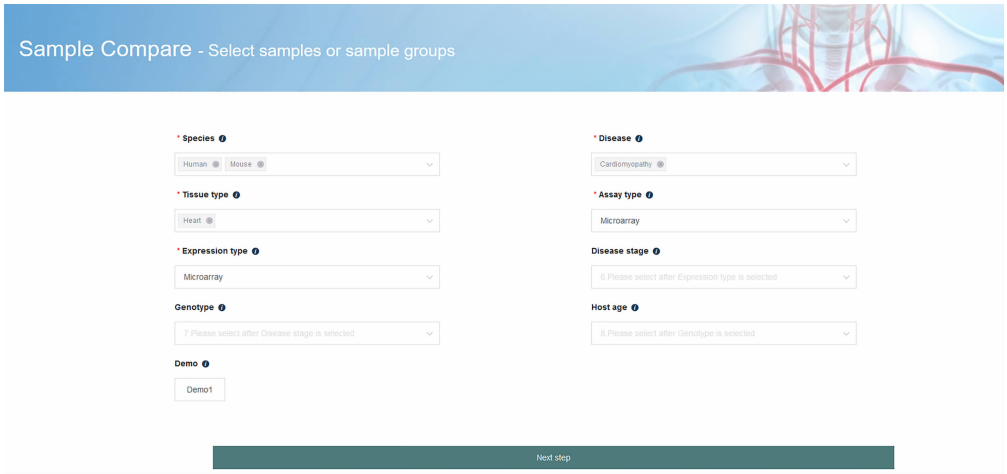

图 4　整合差异性分析参数选择

Figure 4　Parameter selection of integrated difference analysis

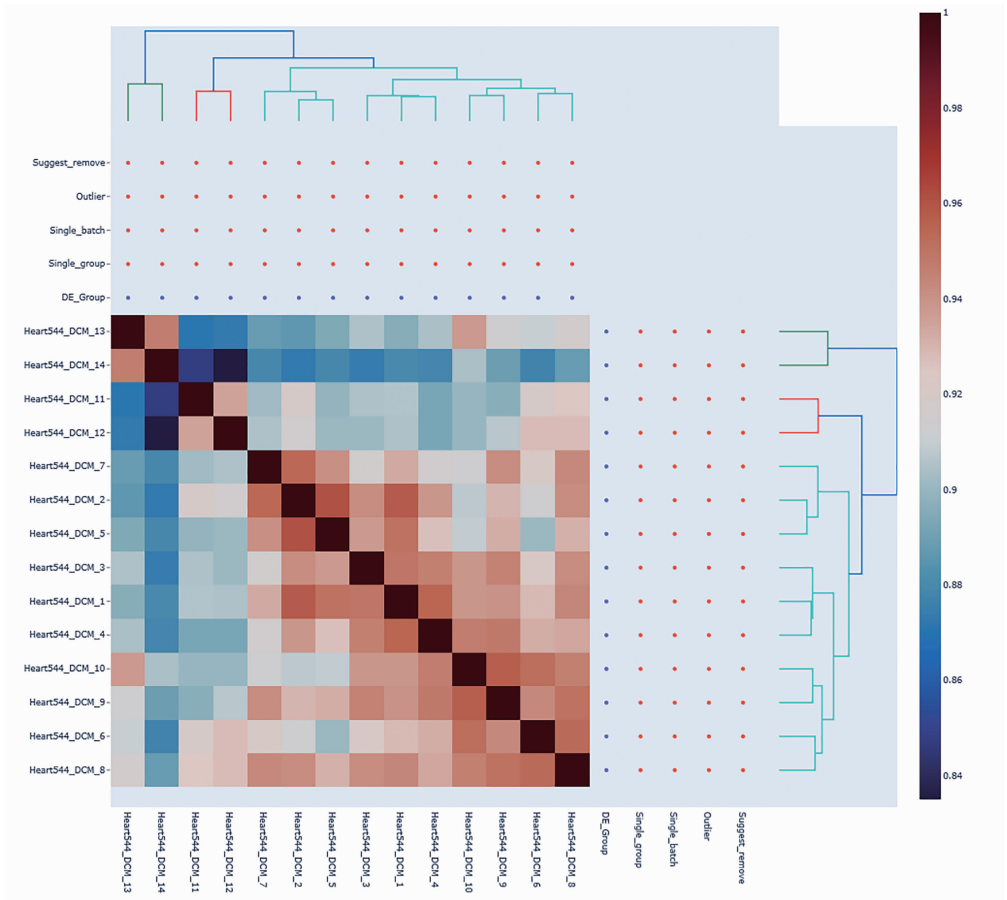

图 5　样品相关性聚类热图

Figure 5　Heatmap of correlation results for samples

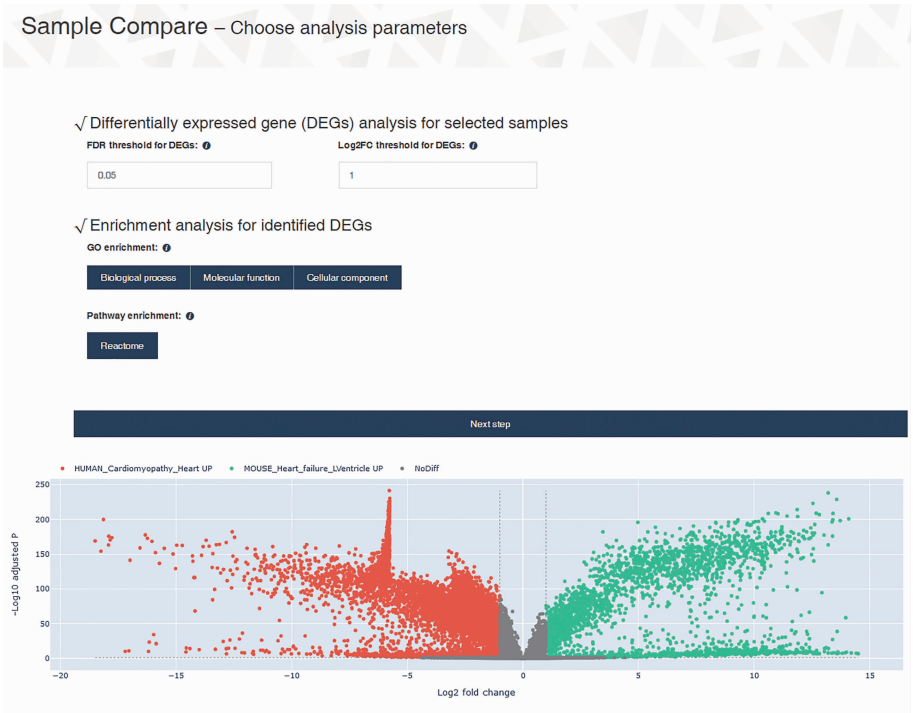

图 6 差异基因火山图

Figure 6 Differential gene volcano map

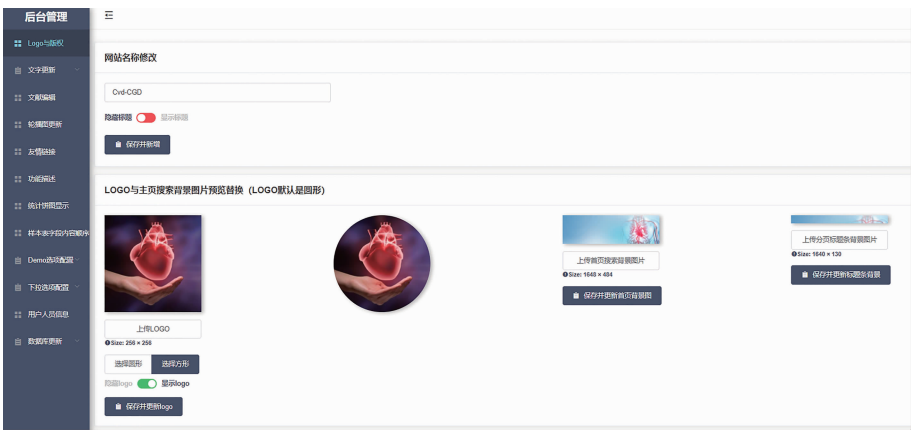

图 7 数据管理后台

Figure 7 Data management background

3 讨论

随着测序技术的迅猛发展,芯片数据、转录组数据猛增。公共数据库积累了大量基因表达谱数据,为我们提供了很好的数据资源,很多大型数据库如 GEO、ArrayExpress、TCGA 等对这些数据进行了广泛收录,但这些数据在使用时仍存在许多挑战[10-13]。第一,主要为原始数据的储存,缺乏定制化的生物信息学工具。第二,不同测序环境下得到的不同样本数据存在批次效应(batch effect)。缺乏合适的方法来消除不同数据集间的批次效应。第三,各数据库、数据集的元数据(metadata)有标准差异。第四,缺乏整合数据集的比较分析(尤其是跨物种比较分析)。

例如 AlzData 数据库(www. alzdata. org)收集了 GEO 来源的老年痴呆症(AD)患者和正常人的脑组织基因表达谱,展示基因在 4 个不同脑区(内嗅皮层、海马、颞叶皮层和额叶皮层)的表达图谱。可结合不同证据对 AD 相关基因进行排序[14]。但仅能展示单物种(人的)基因表达谱,无法进行跨种比较分析。并且仅展示特定基因在不同脑区的表达,无基因在不同发病阶段、不同性别、不同年龄的表达图谱。

例如 EWAS 数据库(https://bigd. big. ac. cn/

ewas/datahub)收集不同来源数据集(GEO、TCGA、ArrayExpress 和 Encode 来源),展示基因或探针在不同癌症、不同组织、不同性别、不同血细胞的甲基化水平差别。通过均一化处理和质量控制,使不同平台数据能够进行整合分析[15]。UALCAN 数据库(http://ualcan.path.uab.edu)提供特定基因在癌症患者不同人种、不同性别、不同年龄、不同体重、不同病程的表达差异图谱,可切换不同指标查看对基因表达的影响[15]。

例如 GEPIA 数据库(http://gepia.cancer-pku.cn/index.html)收集来自 TCGA 数据库的数据,展示基因在 33 种肿瘤和正常组织的表达情况,对差异基因进行汇总分析,并在每一条染色体上显示。展示特定基因在特定癌种的正常组织与肿瘤组织的表达差异(箱式图)、特定基因在特定癌种不同病理分期中的表达差异(小提琴图)、多个目标基因的表达分析(热图)[16]。

本数据库数据来源于自有数据和 GEO 等公共数据库,实验信息不规范,实验设计复杂,数据整合存在很大不同。故本数据库在以下 3 个方面实现了突破。第一,在数据整合方面,对不同平台测序数据进行质控、标准化、去除批次效应,实现跨平台数据集合并。第二,在背景信息收集方面,Metadata 采集、整合,采用脚本与人工整理相结合的方式。第三,在比较转录组学分析方面,跨物种比较分析一直是难以突破的瓶颈,需扣除自身对照的影响后再进行人与动物的比较。

本数据库实现了不同层面的比较分析,展示特定基因在不同物种、病种、组织、病程的表达差异,实现物种内、物种间差异基因分析,以及图文交互与导出功能。通过整合高通量微阵列及转录组数据,我们为基因表达及差异基因的挖掘提供了一个全面的分析平台,并提供了在线比较分析工具,处理来自不同预处理方法和测序平台的数据,并提供跨物种[17]、组织和病种的标准化基因表达谱。本数据库将持续进行更新,两年内纳入更多物种(如大鼠、猴、黑猩猩、牛、猪等)、更多不同品系的数据集,五年内扩充不同类型(单细胞测序、空间转录组学等)测序数据,以实现更全面、多维度的比较分析。

参考文献:

[1] Sasayama S. Heart disease in Asia [J]. Circulation, 2008, 118(25): 2669-2671.

[2] Riehle C, Bauersachs J. Small animal models of heart failure [J]. Cardiovasc Res, 2019, 115(13): 1838-1849.

[3] Zaragoza C, Gomez-Guerrero C, Martin-Ventura JL, et al. Animal models of cardiovascular diseases [J]. J Biomed Biotechnol, 2011, 2011: 497841.

[4] Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse [J]. Nat Rev Genet, 2017, 18(7): 425-440.

[5] Tucker NR, Chaffin M, Fleming SJ, et al. Transcriptional and cellular diversity of the human heart [J]. Circulation, 2020, 142(5): 466-482.

[6] Gladka MM, Molenaar B, de Ruiter H, et al. Single-cell sequencing of the healthy and diseased heart reveals cytoskeleton-associated protein 4 as a new modulator of fibroblasts activation [J]. Circulation, 2018, 138(2): 166-180.

[7] Hahn VS, Knutsdottir H, Luo X, et al. Myocardial gene expression signatures in human heart failure with preserved ejection fraction [J]. Circulation, 2021, 143(2): 120-134.

[8] Nomura S, Satoh M, Fujita T, et al. Cardiomyocyte gene programs encoding morphological and functional signatures in cardiac hypertrophy and failure [J]. Nat Commun, 2018, 9(1): 4435.

[9] Gluckman PD, Hanson MA, Buklijas T, et al. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases [J]. Nat Rev Endocrinol, 2009, 5(7): 401-408.

[10] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets-update [J]. Nucleic Acids Res, 2013, 41(Database issue): D991-D995.

[11] Parkinson H, Kapushesky M, Shojatalab M, et al. ArrayExpress-a public database of microarray experiments and gene expression profiles [J]. Nucleic Acids Res, 2007, 35(Database issue): D747-D750.

[12] Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project [J]. Nat Genet, 2013, 45(10): 1113-1120.

[13] Xu M, Zhang DF, Luo R, et al. A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease [J]. Alzheimers Dement, 2018, 14(2): 215-229.

[14] Xiong Z, Li M, Yang F, et al. EWAS Data Hub: a resource of DNA methylation array data and metadata [J]. Nucleic Acids Res, 2020, 48(1): D890-D895.

[15] Chandrashekar DS, Bashel B, Balasubramanya SAH, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses [J]. Neoplasia, 2017, 19(8): 649-658.

[16] Tang Z, Li C, Kang B, et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses [J]. Nucleic Acids Res, 2017, 45(W1): W98-W102.

[17] 吴玥, 向志光, 高苒, 等. 冠状病毒感染动物模型比较转录组学数据库的建立 [J]. 中国实验动物学报, 2022, 30(1): 92-99.

〔收稿日期〕2022-03-18