DATABASE
The Journal of Biological Databases and Curation

Database, 2017, 1–7
doi: 10.1093/database/bax071
Database Update

Database Update

SFMetaDB: a comprehensive annotation of mouse RNA splicing factor RNA-Seq datasets

Jin Li[1,2], Ching-San Tseng[3], Antonio Federico[4,5], Franjo Ivankovic[6], Yi-Shuian Huang[3], Alfredo Ciccodicola[4,5], Maurice S. Swanson[6] and Peng Yu[1,2,*]

[1]Department of Electrical and Computer Engineering, [2]TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA, [3]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, [4]Institute of Genetics and Biophysics "Adriano Buzzati Traverso", CNR, Naples, Italy, [5]Department of Science and Technology, University of Naples "Parthenope", Naples, Italy, [6]Department of Molecular Genetics and Microbiology, Center for NeuroGenetics and the Genetics Institute, College of Medicine, University of Florida, Gainesville, Florida, USA.

*Corresponding author: Tel: 1 979 845 7441; Fax: 1 979 845 6259; Email: pengyu.bio@gmail.com

Citation details: Li,J., Tseng,C.-S., Federico,A. *et al.* SFMetaDB: a comprehensive annotation of mouse RNA splicing factor RNA-Seq datasets. *Database* (2017) Vol. 2017: article ID bax071; doi:10.1093/database/bax071

Received 13 April 2017; Revised 10 July 2017; Accepted 15 August 2017

Abstract

Although the number of RNA-Seq datasets deposited publicly has increased over the past few years, incomplete annotation of the associated metadata limits their potential use. Because of the importance of RNA splicing in diseases and biological processes, we constructed a database called SFMetaDB by curating datasets related with RNA splicing factors. Our effort focused on the RNA-Seq datasets in which splicing factors were knocked-down, knocked-out or over-expressed, leading to 75 datasets corresponding to 56 splicing factors. These datasets can be used in differential alternative splicing analysis for the identification of the potential targets of these splicing factors and other functional studies. Surprisingly, only ~15% of all the splicing factors have been studied by loss- or gain-of-function experiments using RNA-Seq. In particular, splicing factors with domains from a few dominant Pfam domain families have not been studied. This suggests a significant gap that needs to be addressed to fully elucidate the splicing regulatory landscape. Indeed, there are already mouse models available for ~20 of the unstudied splicing factors, and it can be a fruitful research direction to study these splicing factors *in vitro* and *in vivo* using RNA-Seq.

Database URL: http://sfmetadb.ece.tamu.edu/

Introduction

Due to the lack of fully structured metadata, the wide use of the valuable RNA-Seq datasets in public repositories such as ArrayExpress (1) and Gene Expression Omnibus (GEO) (2) may be restricted, despite structured metadata having been used elsewhere for raw data usability (3). For example, ArrayExpress is only a repository of datasets, and the completeness of metadata information relies on dataset submitters. Although submission facilities have been improving, metadata information of many datasets in ArrayExpress is still not well structured (1). To fill this gap, manual curation has been devoted to developing and maintaining metadata databases (4). For example, microarray and RNA-Seq datasets have been curated for the downstream analyses in Expression Atlas (5). We previously launched the RNASeqMetaDB (6) database to facilitate the access of the metadata of public available mouse RNA-Seq datasets. Here, we present a new database, SFMetaDB, as an update with metadata of RNA-Seq datasets related with splicing factors with either loss- or gain-of-function experiments.

RNA splicing is a fundamental biological process in eukaryotes that substantially contributes to the overall protein diversity in a cell. According to GENCODE (Release 25) basic transcript annotation, 19 903 human protein-coding genes encode 54 896 isoforms by alternative splicing. The importance of alternative splicing is underscored by the distinct biological functions played by splicing isoforms. Recently, the splicing isoform function of a number of genes has been tested experimentally in a variety of biological contexts, including cancer. For example, two isoforms of *CD44*, a widely expressed cell surface marker, have recently been shown to be important in cancer development. The first isoform CD44V6 is required for the migration and generation of metastatic tumors in colorectal cancer stem cells and can initiate the metastatic process (7). The second isoform of *CD44*, CD44V8-10, is an important marker for human gastric cancer and increases tumor initiation in gastric cancer cells (8). Another example is *NUMB*, a gene that is critical for cell fate determination. Two splicing isoforms varying in the length of proline-rich region (PRR), PRR^L and PRR^S, were recently found to have opposite roles in hepatocellular carcinoma (HCC), suggesting that the alternative splicing of *NUMB* can serve as an important biomarker for HCC (9). In particular, PRR^L promotes proliferation, migration, invasion and colony formation while PRR^S generally works in the opposite way.

Splicing isoforms may also play some critical roles in biological processes other than cancer. For example, *MICU1* is a gene encoding an essential regulator of mitochondrial Ca^{2+} uptake, a process that is critical for energy production in skeletal muscle. Through the inclusion of a micro-exon (<15 bp) of this gene, an alternative splice isoform named MICU1.1 can be generated. It was found that the exclusion of this microexon causes a ~10× decrease of the Ca^{2+}-binding affinity of MICU1 proteins. Therefore, alternative splicing is essential for the sustainability of Ca^{2+} uptake and ATP production of mitochondria, the energy source of skeletal muscle (10). For another example, FANCE is a part of the Fanconi anemia complex, which functions in DNA interstrand crosslink repair. FANCE plays a critical role to regulate FANCD2, which is required in FANC–BRCA functions. Overexpression of an alternative splicing isoform FANCEΔ4 promotes degradation of FANCD2 and causes dysfunction of DNA repair (11). Furthermore, *VEGF-A* is a gene that functions in angiogenesis, vasculogenesis and endothelial cell growth. Two alternative splicing isoforms, VEGF-A_{xxx}a and VEGF-A_{xxx}b, are critical in nociception (12). VEGF-A_{xxx}a is increased with nerve injury and promotes nociceptive function. On the contrary, the overexpression of VEGF-A_{xxx}b reduces neuropathic pain. In addition, the *Fas*/*CD95* gene is critical in the physiological regulation of programmed cell death. *Fas*/*CD95* has two splicing isoforms with inclusion or exclusion of exon 6, a membrane-bound receptor or a soluble isoform (13). The membrane-bound receptor isoform promotes apoptosis while the soluble isoform inhibits apoptosis.

Alternative splicing is commonly mediated by RNA splicing factors (14). For example, the splicing factor NOVA1 regulates the alternative splicing of a series of genes in pancreatic beta cells, and knockdown of *Nova1* suppresses insulin secretion and promotes apoptosis (15). Moreover, the splicing factor NOVA2 uniquely mediates the alternative splicing of many axon guidance-related genes during cortical development (16). As another example, the splicing factor PTBP1 suppresses *Pbx1* exon 7 and the neuronal PBX1A isoform in embryonic stem cells during neuronal development (17).

In this article, we describe our recent effort in curating the metadata of RNA-Seq datasets from ArrayExpress and GEO, which were derived from studies using cell or animal models with a specific splicing factor being knocked-out, knocked-down or overexpressed. We further launched SFMetaDB to facilitate access to the metadata of these datasets and share them with the biomedical community.

Results and discussion

The launch of SFMetaDB focuses on RNA-Seq datasets with perturbed splicing factors. Users can query a given splicing factor to identify the relevant datasets. A use case

for MBNL splicing factors is shown as follows. MBNL1 is an important RNA splicing factor (18), thus we use MBNL1 to demonstrate the usage of SFMetaDB, which confirms the advantage of SFMetaDB over ArrayExpress. As shown in Figure 1a, a query of MBNL1 on SFMetaDB returns the accurate datasets related with *Mbnl1* loss- or gain-of-function experiments. Figure 1a shows that five datasets could be used for the alternative splicing analysis for MBNL1, and the potential targets of MBNL1 can be concluded from the datasets. For example, the dataset GSE39911 (i.e. E-GEOD-39911) includes biological replicates of various tissues, such as brain, heart and muscle, from *Mbnl1*-knockout mice and *Mbnl1*-knockdown C2C12 mouse myoblasts (Figure 1b).

However, as shown in Figure 1c, ArrayExpress returned a total of 13 mouse RNA-Seq datasets with the query Mbnl1, and 8 of them were not from *Mbnl1* gain- or loss-of-function experiments. Therefore, these datasets were eliminated in SFMetaDB. For example, the dataset E-GEOD-76222 is retrieved by ArrayExpress because of the appearance of Mbnl1 in its description, 'Changes in the expression of alternative splicing factors Zcchc24, Esrp1, Mbnl1/2 and Rbm47 were demonstrated to be key contributors to phase-specific AS.' However, this dataset is about an *ESRP* knock-out, thus it is not suitable for MBNL1 related alternative splicing analysis (Figure 1d). The rest of eight retrieved datasets were considered not appropriate for RNA splicing analysis of MBNL1 by our manual curation of metadata information. In summary, no irrelevant datasets of a given splicing factor are shown in SFMetaDB, and SFMetaDB returned more specific results than ArrayExpress.

Guided by SFMetaDB, users can perform potential target identification for a specific splicing factor. In addition, by integrating multiple datasets curated on SFMetaDB, users can form a more comprehensive view on how a splicing event is regulated across different biological contexts. As another use case, we show below a Pfam domain analysis among splicing factors (see Materials and methods).

Only ∼15% of known splicing factors have been studied with loss- or gain-of-function RNA-Seq experiments. Because splicing factors sharing similar domains tend to regulate common splicing targets, we determined what additional splicing factors may be prioritized for study by investigating the domain structures of the splicing factors using UniProt (19). Among the 353 splicing factors, 299 of them contained one or multiple conservative domains. Of these 299 splicing factors, 190 have a single domain that belongs to a Pfam domain family, and the rest have domains that belong to more than one Pfam domain family.

RNA splicing factors have highly conserved functional domains, and some domains are dominant among all the splicing factors. In Figure 2, the domain families are ranked by their number of occurrences in all the splicing factors. Pfam family PF00076 (RNA recognition motif) is the most dominant, and the splicing factors with domains from this family are relatively well studied (25 over the total 87). Splicing factors from five additional Pfam families are fairly well-studied ($\geq$3 splicing factors annotated), consisting of PF00271 (Helicase conserved C-terminal domain), PF00270 (DEAD/DEAH box helicase), PF00013 (KH domain), PF00642 (Zinc finger C–x8–C–x5–C–x3–H type) and PF12414 (Calcitonin gene-related peptide regulator C terminal). However, three highly dominant families are not. Specifically, none of the 17 splicing factors with the Pfam family PF01423 (LSM domain) (Figure 2) have been studied yet (20), and these splicing factors provide feasible candidates for future studies. For example, the splicing factor SNRPN has two mouse models from the International Mouse Strain Resource (IMSR) (21) that can be used for splicing analysis. In fact, 25 unstudied splicing factors (Supplementary Table S1) have been identified with more than one mouse model from IMSR. Therefore, splicing factors that are non-homologous with already studied ones constitute promising candidates for comprehensive studies of splicing regulation.

Materials and methods

RNA-Seq dataset curation and SFMetaDB web server deployment

We extracted 353 RNA splicing factors annotated in Gene Ontology (GO) (accession GO:0008380) (22) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (entry mmu03040) (23) for mice. Then, we queried ArrayExpress (1) and GEO (2) using the official symbol of each splicing factor to search for related mouse RNA-Seq datasets and obtained a total of 214 datasets. Note that due to the limitation of the search function in ArrayExpress and GEO, many of these datasets were not directly relevant to the manipulation of these splicing factors despite that the symbols were mentioned in the metadata of these datasets. We chose to manually curate each dataset, providing a total of 75 datasets that have biological replications in which at least one splicing factor was knocked-out, knocked-down or overexpressed (along with the corresponding wild types/controls) (Supplementary Table S1). Because some splicing factors were studied in more than one dataset, a total of 56 splicing factors were found (Supplementary Table S1).

To facilitate the access to these datasets, we launched the database SFMetaDB (http://sfmetadb.ece.tamu.edu/).

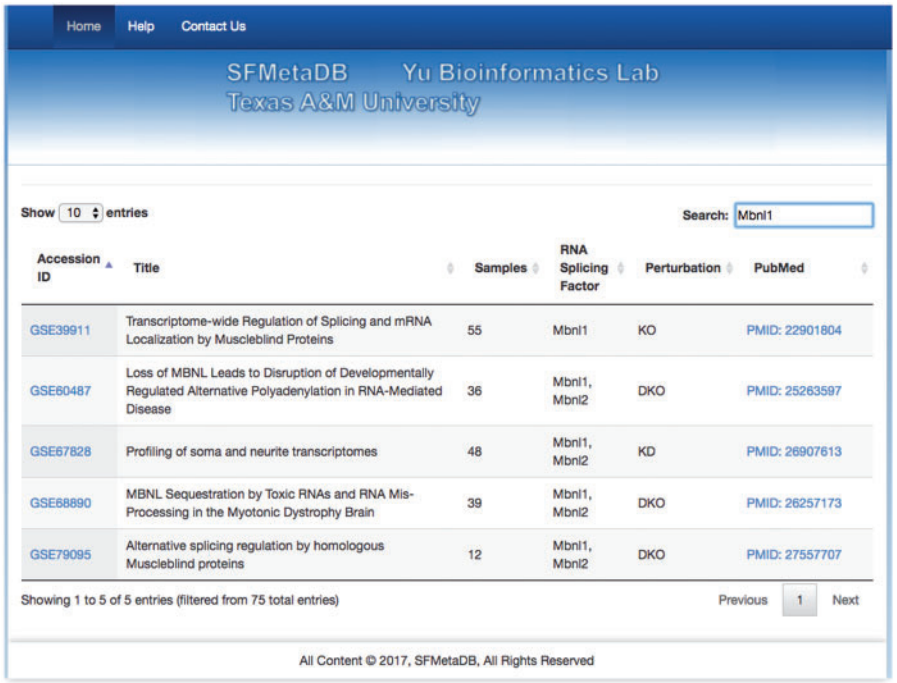

Figure 1. A use case of SFMetaDB for the splicing factor Mbnl1. We showed a use case of the splicing factor Mbnl1 to demonstrate the advantage of SFMetaDB over ArrayExpress. By using the same keyword, Mbnl1, SFMetaDB returned five accurate datasets that can be used for the downstream alternative splicing analyses. On the contrary, ArrayExpress returned 13 datasets with 8 that could not be used for the downstream alternative splicing analyses for Mbnl1. (**a**) The result page in SFMetaDB of the query Mbnl1. (**b**) The description page of the dataset GSE39911 in GEO. (**c**) The result page in ArrayExpress of the query Mbnl1. (**d**) The description page of the dataset E-GEOD-76222 in ArrayExpress.

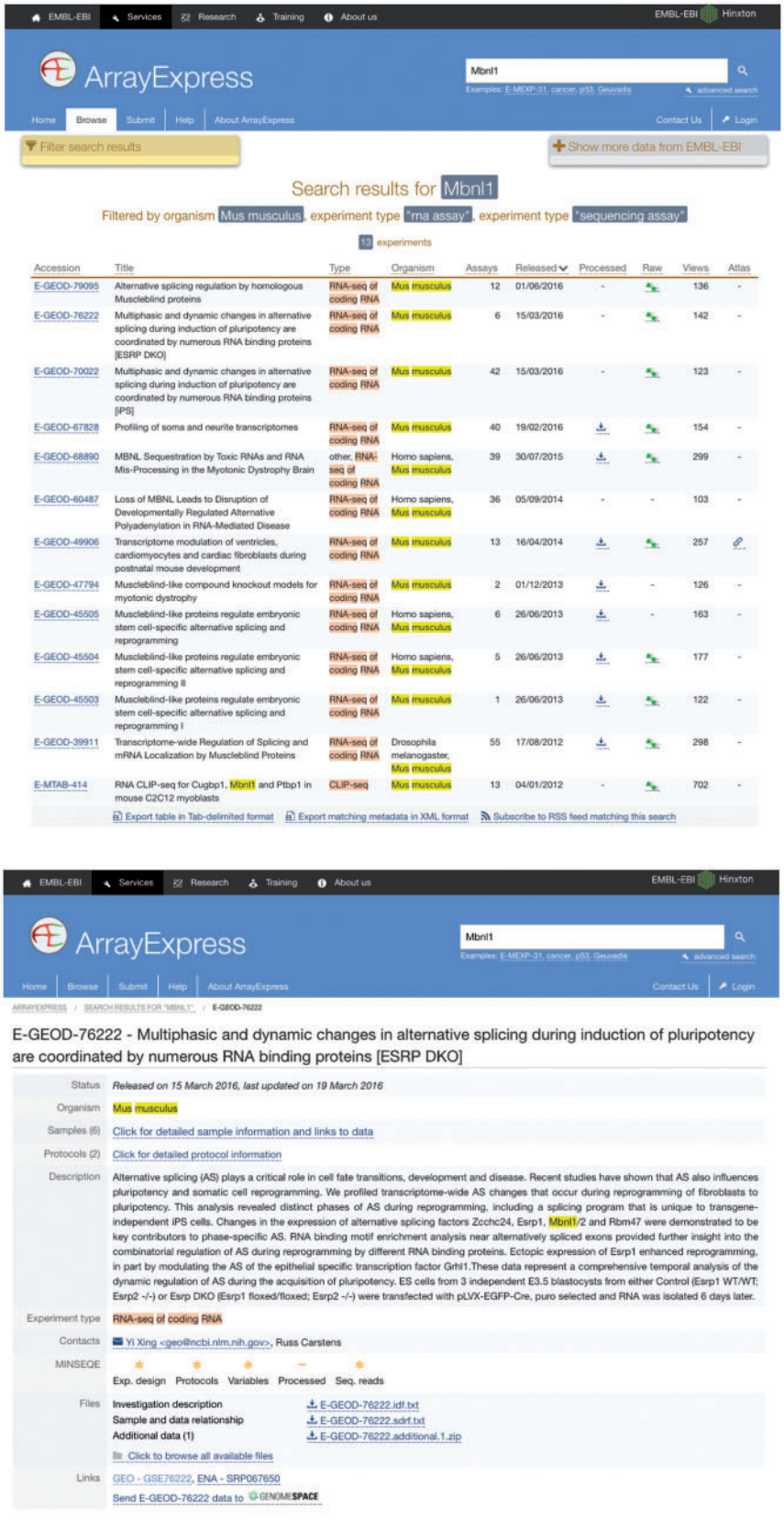

Figure 1. Continued.

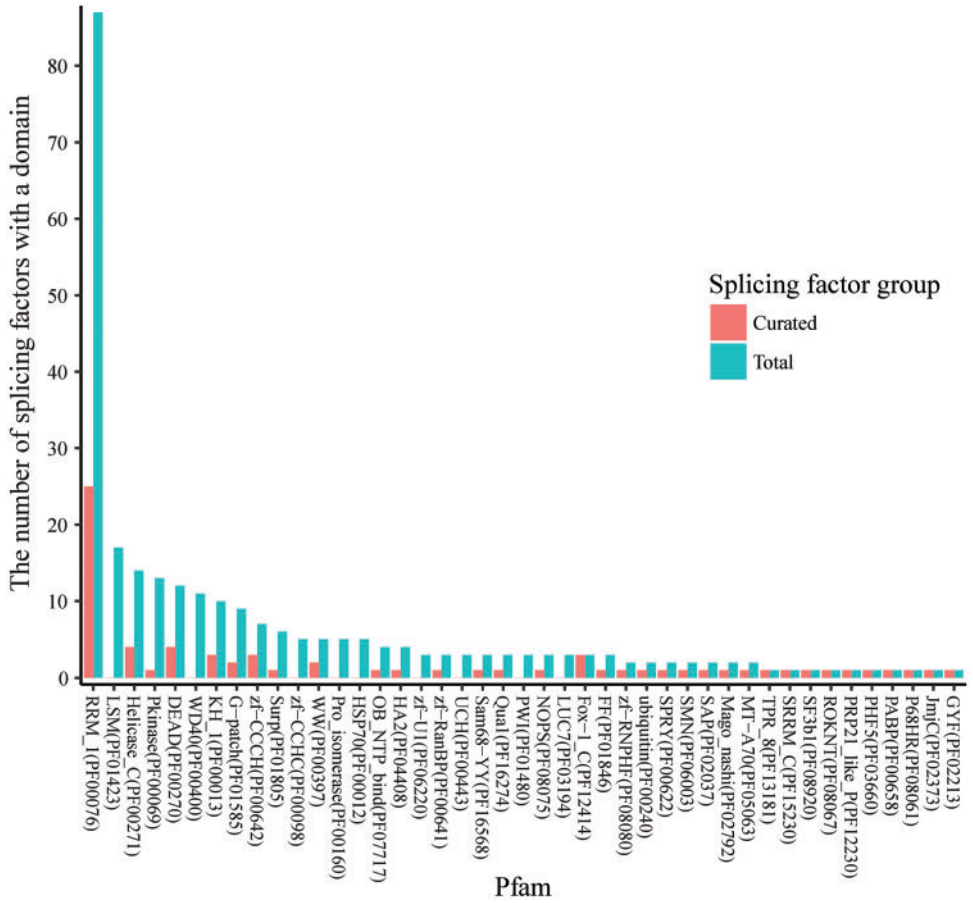

Figure 2. The occurrence of Pfam domain families in splicing factors. The known RNA splicing factors are annotated in UniProt according to the Pfam domain families of the protein domains found in these factors. A splicing factor may have multiple domains that belong to multiple Pfam families, and a Pfam domain family may contain domains in multiple splicing factors. The Pfam annotations were retrieved for each of 353 splicing factors, and the number of splicing factors was calculated for each of the Pfam families. For the 56 splicing factors that have curated datasets in SFMetaDB, the number of splicing factors was also calculated for the associated Pfam families. In the dodged barplots, the Pfam domain families are ranked by the number of the splicing factors which contain domains in the given families. Of the total 217 Pfam domain families annotated in UniProt, 26 Pfam domain families have $\geq$3 splicing factors annotated. The Pfam domain family with the most number of splicing factors is Pfam RRM_1 (PF00076). It contains 87 splicing factors, and 25 of these splicing factors have been studied according to our curation results. However, the splicing factors in the rest of the Pfam domain families have brought relatively less attention in RNA-Seq analysis, and they may be promising candidates for future studies.

When datasets were deposited in GEO, ArrayExpress imported the most metadata information from GEO, and the ArrayExpress description contained the link to the GEO webpage. Therefore, SFMetaDB used GEO accession ID if possible. The web server of SFMetaDB is freely available, and it presents the Accession ID, description, the number of samples, associated curated splicing factors, perturbation and PubMed references of each RNA-Seq dataset.

Domain structures analysis in RNA splicing factors

The domain structures of the RNA splicing factors may guide us to identify the candidate splicing factors for future studies. Known RNA splicing factors are retrieved from GO term (GO:0008380) using R package GO.db (22) and KEGG pathway (entry mmu03040). UniProt annotates the conservative Pfam domain families for the canonical sequences of the splicing factors (19). From these domain annotations, we calculate the numbers of the splicing factors in Pfam domain families. Figure 2 plots the dodged barplots of the number of splicing factors in Pfam domain families using curated splicing factors and the total splicing factors. By comparing the domain families of the splicing factors with RNA-Seq datasets to the families of all the splicing factors, the splicing factors in not well-studied domain families can be the promising candidates for future RNA-Seq studies.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors thank Zhengyu Guo for his contribution to SFMetaDB.

Funding

This work was supported by startup funding to P.Y. from the ECE department and Texas A&M Engineering Experiment Station/ Dwight Look College of Engineering at Texas A&M University, by funding from TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) at Texas A&M University, by TEES seed grant, and by Texas A&M University-CAPES Research Grant Program and by grants from the NIH (NS058901, NS098819 to M.S.S). The open access publishing fees for this article have been covered in part by the Texas A&M University Open Access to Knowledge Fund (OAKFund), supported by the University Libraries and the Office of the Vice President for Research.

Conflict of interest. None declared.

References

1. Kolesnikov,N., Hastings,E., Keays,M. *et al.* (2015) ArrayExpress update-simplifying data submissions. *Nucleic Acids Res.*, 43, D1113–D1116.
2. Edgar,R., Domrachev,M., and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30, 207–210.
3. Mitchell,A., Bucchini,F., Cochrane,G. *et al.* (2016) EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, 44, D595–D603.
4. Qin,B., Zhou,M., Ge,Y. *et al.* (2012) CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. *Bioinformatics*, 28, 1411–1412.
5. Petryszak,R., Keays,M., Tang,Y.A. *et al.* (2016) Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, 44, D746–D752.
6. Guo,Z., Tzvetkova,B., Bassik,J.M. *et al.* (2015) RNASeqMetaDB: a database and web server for navigating metadata of publicly available mouse RNA-Seq datasets. *Bioinformatics*, 31, 4038–4040.
7. Todaro,M., Gaggianesi,M., Catalano,V. *et al.* (2014) CD44v6 is a marker of constitutive and reprogrammed cancer stem cells driving colon cancer metastasis. *Cell Stem Cell*, 14, 342–356.
8. Lau,W.M., Teng,E., Chong,H.S. *et al.* (2014) CD44v8-10 is a cancer-specific marker for gastric cancer stem cells. *Cancer Res.*, 74, 2630–2641.
9. Lu,Y., Xu,W., Ji,J. *et al.* (2015) Alternative splicing of the cell fate determinant Numb in hepatocellular carcinoma. *Hepatology*, 62, 1122–1131.
10. Vecellio Reane,D., Vallese,F., Checchetto,V. *et al.* (2016) A MICU1 splice variant confers high sensitivity to the mitochondrial Ca2+ uptake machinery of skeletal muscle. *Mol. Cell.*, 64, 760–773.
11. Bouffard,F., Plourde,K., Belanger,S. *et al.* (2015) Analysis of a FANCE splice isoform in regard to DNA repair. *J. Mol. Biol.*, 427, 3056–3073.
12. Hulse,R.P., Drake,R.A., Bates,D.O. *et al.* (2016) The control of alternative splicing by SRSF1 in myelinated afferents contributes to the development of neuropathic pain. *Neurobiol. Dis.*, 96, 186–200.
13. Tejedor,J.R., Papasaikas,P., and Valcarcel,J. (2015) Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis. *Mol. Cell.*, 57, 23–38.
14. Cieply,B., and Carstens,R.P. (2015) Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip. Rev. RNA*, 6, 311–326.
15. Villate,O., Turatsinze,J.V., Mascali,L.G. *et al.* (2014) Nova1 is a master regulator of alternative splicing in pancreatic beta cells. *Nucleic Acids Res.*, 42, 11818–11830.
16. Saito,Y., Miranda-Rottmann,S., Ruggiu,M. *et al.* (2016) NOVA2-mediated RNA regulation is required for axonal pathfinding during development. *Elife*, 5, e14371.
17. Linares,A.J., Lin,C.H., Damianov,A. *et al.* (2015) The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife*, 4, e09268.
18. Konieczny,P., Stepniak-Konieczna,E., Taylor,K. *et al.* (2017) Autoregulation of MBNL1 function by exon 1 exclusion from MBNL1 transcript. *Nucleic Acids Res.*, 45, 1760–1775.
19. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.
20. Finn,R.D., Coggill,P., Eberhardt,R.Y. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44, D279–D285.
21. Eppig,J.T., Motenko,H., Richardson,J.E. *et al.* (2015) The International Mouse Strain Resource (IMSR): cataloging worldwide mouse and ES cell line resources. *Mamm. Genome*, 26, 448–455.
22. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25, 25–29.
23. Kanehisa,M., and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28, 27–30.