DATA MINING FOR IDENTIFYING KEY GENES IN BIOLOGICAL PROCESSES USING GENE EXPRESSION DATA

A Dissertation

by

JIN LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Peng Yu
Committee Members,	Ulisses Braga-Neto
	James Cai
	Jiang Hu
Head of Department,	Miroslav M. Begovic

December 2018

Major Subject: Electrical Engineering

Copyright 2018 Jin Li

ABSTRACT

A large volume of gene expression data is being generated for studying mechanisms of various biological processes. These precious data enabled various computational analyses to speed up the understanding of biological knowledge. However, it remains a challenge to analyze the data efficiently for new knowledge mining. These data were generated for different purposes, and their heterogeneity makes it difficult to consistently integrate the datasets, slowing down the reuse of these data and the process of biological discovery for new knowledge. To facilitate the reuse of these precious data, we engaged biology experts to manually collected RNA-Seq gene expression datasets for perturbed splicing factors and RNA-binding proteins, resulting in two online databases, SFMetaDB and RBPMetaDB. These two databases hold comprehensive RNA-Seq gene expression data for mouse splicing factors and RNA-binding proteins, and they can be used for identify key genes or regulators in biological processes or human diseases. Beside showing an importance of two databases, these two projects also demonstrated an efficient way to collect data. In my dissertation, we also engaged biology collaborators to collect comprehensive regulate genes in cold-induced thermogenesis supported by *in vivo* experiments with key genes deposited to CITGeneDB. This database is the first to offer comprehensive list of regulators in cold-induced thermogenesis in a higher regulatory hierarchy. In addition to build data resources, my dissertation also worked on analyze RNA-Seq gene expression data to gain biological insights. To study the mechanism of human skin disease psoriasis, we analyzed mouse and human public psoriasis datasets, and compared to splicing factor perturbed datasets in SFMetaDB, resulting in candidate genes for psoriasis. Our computational predictions provide candidate factors to follow to study fundamental processes underlying psoriasis. In addition, we introduced a data processing paradigm

to identify key genes in biological processes via systematic collection of gene expression datasets, primary analysis of data, and evaluation of consistent signals. Our paradigm was applied to two applications of epidermal development and cold-induced thermogenesis, and revealed many key genes in the two applications. By collaborating with web labs, we experimentally validate a novel gene suprabasin (SBSN) in epidermal development. These findings enable a better understanding of the mechanisms underlying epidermal development and cold-induced thermogenesis, and also demonstrate the effectiveness of our paradigm by combining data collection and integrated analysis. My dissertation has mainly investigated a biological data process paradigm, consisting of systematic data collection, data analysis and hypothesis generation. By intensive works, we demonstrated the effectiveness of this novel biological data process approach, and this approach can be readily generalized to other biological processes or human diseases.

DEDICATION

To my wife Qin Huang and my son Yuxuan Li.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Peng Yu, for his thorough guidance, encouragement and support over the past three years. This dissertation would not have been possible without his guidance and support. Besides showing me how to work efficiently and productively, he also taught me the underlying logic to come up with better solutions. His high research standards have a long-standing deep impact on my life.

I am also grateful to my committee members, Dr. Ulisses Braga-Neto, Dr. James Cai, and Dr. Jiang Hu, for their advice and support.

Thanks also go to Yu Lab members for broad discussions and collaborations.

I dedicate this dissertation to my beloved wife Qin Huang. Without her endless love, support and encouragement, my work would not have been accomplished. I also dedicate this dissertation to my lovely son Yuxuan Li. His birth is the happiest thing ever happened to me.

Thanks to my parents for their unconditional love, support and encouragement throughout my life.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Peng Yu, Prof. Ulisses Braga-Neto and Prof. Jiang Hu of the Department of Electrical & Computer Engineering and Prof. James Cai of the Department of Veterinary Integrative Biosciences. All work for the dissertation was completed independently by the student, under the supervision of Dr. Peng Yu of the Department of Electrical & Computer Engineering.

Funding Sources

This work was supported by startup funding to Peng Yu 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.

NOMENCLATURE

A5SS	alternative 5' splice site
A3SS	alternative 3' splice site
AD	Atopic dermatitis
AFE	alternative first exons
ALE	alternative last exons
BAT	brown adipose tissue
BP	Biological process
CIT	cold-induced thermogenesis
CLIME	Clustering by Inferred Models of Evolution
DAS	differential alternative splicing
DEG	differential gene expression
ES	exon skipping
GEO	Gene Expression Omnibus
GO	gene ontology
IMQ	imiquimod
IR	intron retention
KD	knockdown
KI	knockin
KO	knockout
ME	mutually exclusive exons
MSA	multiple sequence alignment

MSF	master splicing factor
OE	overexpression
PSI	percent spliced in
RBP	RNA-binding protein
SBSN	Suprabasin
SF	splicing factor
VST	variance-stabilizing transformation
WAT	white adipose tissue

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION AND LITERATURE REVIEW

High-throughput expression profiling has been used to identify transcriptional changes associated with many diseases and biological processes (BPs). However, the mechanism underlying the associated changes remains mostly unclear. To uncover the underlying molecular mechanism, complementary approaches such as ChIP-Seq[1] and CLIP-Seq[2] have been incorporated to identify direct interactions among proteins and DNAs/RNAs. These two immunoprecipitation-oriented approaches are limited because high-quality antibodies against specific proteins may not be readily available, the protocols are complex, and only a limited number of specialized labs can perform them well. In addition, direction binding relations do not necessarily represent upstream regulation. Thus, approaches that bypass these limitations are needed to prioritize upstream regulators in diseases and BPs.

Given the large scale of high-throughput expression profiling data publicly available, any method that can utilize these data to identify upstream regulators of transcription in diseases or BPs will be of great value. High throughout expression profiling has become routine and much of the resulting data are available from online repositories, such as Gene Expression Omnibus (GEO)[3]. Up to the second quarter of 2018, GEO hosted over 97,000 data series comprising over 2,400,000 samples. As a popular method for transcriptome analysis, RNA-sequencing (RNA-Seq)[4] has enabled genome-wide analyses of RNA molecules at a high sequencing depth with high accuracy. It has been successfully used on many mouse models[5, 6], and thousands of RNA-Seq datasets have been generated and released to the public. This massive amount of biological data brings great opportunities for generating prominent biological hypotheses[7, 8]. However, these data were produced for diverse purposes and are not friendly to large-scale data integration. Therefore, substantial work is needed to build well-organized resources using these data to enable

efficient and extensive integrated analysis. In this dissertation, we developed an integrated analysis to reveal upstream regulators of transcription in diseases or BPs using these public RNA-Seq data.

We focused on datasets related to splicing factors (SFs), as approximately 95% of human multi-exonic genes are alternatively spliced, and RNA splicing is a fundamental process controlling gene expression in eukaryotes[9]. We previously curated the metadata of a comprehensive and accurate list of mouse RNA-Seq data with perturbed SFs, which are hosted on our SFMetaDB[10]. Using these metadata, corresponding RNA-Seq data were used to compute alternative splicing changes and gene expression changes related to perturbed SFs, represented in RNA splicing signatures and gene expression signatures, respectively. The generated signature data were used to determine the biological relevance of SFs to a disease or a BP using signature comparison[11]. Highly relevant SFs were considered key regulators in the disease or BP.

Our approach is a general high-level regulator discovery method that can be used to identify master splicing factors (MSFs) that are key regulators at the top of the transcriptome regulatory hierarchy affecting a large number of downstream genes in a specific disease or BP[12]. To demonstrate our approach, we conducted several studies, consisting of identifying key genes in psoriasis using transcriptome data (section 2), and revealing underlying biological processes in epidermal development and cold-induced thermogenesis (section 3). To facilitate the first part of our data analysis paradigm, we engaged biologists to manually collect several databases. Particularly, we collected a comprehensive regulatory genes in cold-induced thermogenesis supported by *in vivo* experiments (section 4), and two metadata databases of public RNA-Seq datasets for splicing factors in SFMetaDB (section 5) and for RNA-binding proteins in RBPMetaDB (section 6).

In summary, our systematic integration of disorganized and unstructured RNA-Seq datasets along with generated signatures provides a novel approach for the identification of the most promising hypotheses for experimental testing. These novel hypotheses will form the basis for new *in vivo* experiments leading to the elucidation of detailed regulatory mechanisms at a molecular level.

2. IDENTIFICATION OF KEY SPLICING FACTORS IN PSORIASIS*

Psoriasis is a chronic inflammatory disease that affects the skin, nails, and joints. For understanding the mechanism of psoriasis, though, alternative splicing analysis has received relatively little attention in the field. We developed and applied several computational analysis methods to study psoriasis [13]. Using psoriasis mouse and human datasets, our differential alternative splicing analyses detected hundreds of differential alternative splicing changes. Our analysis of conservation revealed many exon-skipping events conserved between mice and humans. In addition, our splicing signature comparison analysis using the psoriasis datasets and our curated splicing factor perturbation RNA-Seq database, SFMetaDB, identified nine candidate splicing factors that may be important in regulating splicing in the psoriasis mouse model dataset. Three of the nine splicing factors were confirmed upon analyzing the human data.

2.1 Introduction

Psoriasis is a chronic inflammatory skin disease with symptoms of well-defined, raised, scaly, red lesions on skin. It is characterized by excessive growth and aberrant differentiation of epidermal keratinocytes. A number of known psoriasis susceptibility loci have been identified[14], some of which are shared with other chronic inflammatory diseases[15]. Psoriasis also shares pathways with other diseases. Despite great progress made over the past few years, the exact causes of psoriasis remain unknown[16].

To discover the disease mechanisms, significant effort has been devoted to analyzing psoriasis gene expression. For example, in a study of small and large plaque psoriasis, microarray

---

* Reprinted with permission from "Genome-wide transcriptome analysis identifies alternative splicing regulatory network and key splicing factors in mouse and human psoriasis" by Jin Li and Peng Yu, 2018. *Scientific Reports*, 8(1), 4124, Copyright 2018 by authors

gene expression analysis revealed the up-regulation of genes in the IL-17 pathway in psoriasis. But the expression of genes in this pathway of small plaque psoriasis is significantly higher than that of large plaque psoriasis, and negative immune regulators like CD69 and FAS have been found to be down-regulated in large plaque psoriasis. This result suggests that the down-regulation of these negative immune regulators contributes to the molecular mechanism of large plaque psoriasis subtypes[17].

As high-throughput sequencing becomes the mainstream technology, RNA-Seq has also been used for measuring gene expression to gain biological insights of psoriasis. For example, a recent RNA-Seq–based gene expression study of a large number of samples from lesional psoriatic and normal skin uncovered many differentially expressed genes in immune system processes[18]. The co-expression analysis based on this dataset detected multiple co-expressed gene modules, including a module of epidermal differentiation genes and a module of genes induced by IL-17 in keratinocytes. This study also discovered key transcription factors in psoriasis and highlighted the processes of keratinocyte differentiation, lipid biosynthesis, and the inflammatory interaction among myeloid cells, T-cells, and keratinocytes in psoriasis.

The high resolution of RNA-Seq data allows for study of not only gene expression but also splicing in psoriasis. A recent analysis of psoriasis RNA-Seq data revealed around 9,000 RNA alternative splicing isoforms as a significant feature of this disease[19]. Another study showed that serine/arginine-rich splicing factor 1 (SRSF1) promoted the expression of type-I interferons (IFNs) in psoriatic lesions, and suppression of SRSF1 treated by TNFα in turn suppressed the expression of IFNs[20]. Despite the potentially important role that splicing plays in the mechanism of psoriasis, analyzing alternative splicing in psoriasis has received relatively little attention in the research community. To develop a better understanding of the disease mechanism of psoriasis, this

section seeks to perform an integrated analysis to reveal missing information about splicing in psoriasis that will largely complement previous gene expression analysis.

To reveal the biological functions of the alternative splicing events in psoriasis, we performed multiple sequence alignment (MSA) between the sequences of mouse and human alternative splicing events, as the conserved splicing events are more likely to play similar roles in both species[21, 22]. Our analysis revealed 18 conserved exon-skipping (ES) events between mice and humans. These conserved events are potential candidates for further functional study.

To identify the candidate splicing factors (SFs) that may be key regulators of splicing disruption seen in psoriasis, we created a database—called SFMetaDB—of all RNA-Seq datasets publicly available in ArrayExpress and GEO from gain or loss function studies of SFs in mice. Using the data source from SFMetaDB, we implemented a signature comparison method to infer the critical SFs for psoriasis. The splicing changes in a psoriasis mouse model[23] and the SF perturbation datasets were used to derive the splicing signatures. By comparing the signatures of psoriasis datasets to the splicing signatures of our splicing signature database, we revealed nine candidate SFs that potentially contribute to the regulation of alternative splicing in psoriasis. Genes regulated by such key SFs are involved in a number of critical pathways associated with psoriasis.

2.2 Methods

2.2.1 Differential alternative splicing analysis using RNA-Seq data

To identify the DAS events, we performed DAS analysis for the *Tnip1* KO mouse model dataset (GSE85891), where the *Tnip1* KO mice and controls were treated for two days with imiquimod (IMQ)[23], and for the human psoriasis dataset (GSE54456), where the human lesional psoriatic and normal skins established large-scale gene expression data[18]. We first aligned the raw RNA-Seq reads to mouse (mm9) or human (hg19) genomes using STAR (version 2.5.1b)[24]

with default settings, and only uniquely mapped reads were retained for further analysis. The number of reads for each exon and exon-exon junction in each RNA-Seq file was computed by using the Python package HTSeq[25] with the annotation of the UCSC KnownGene (mm9 or hg19) annotation[26]. DMN was used to model the counts of the reads aligned to each isoform of each event[27], and the likelihood ratio test was used to test the significance of the changes in alternative splicing between psoriasis samples and controls[28]. We calculated the q-values from the p-values in the likelihood ratio test by the Benjamini-Hochberg procedure[29]. The DAS events are classified into seven splicing types: Exon skipping (ES), alternative 5' splice sites (A5SSs), alternative 3' splice sites (A3SSs), mutually exclusive (ME) exons, intron retention (IR), alternative first exons (AFEs) and alternative last exons (ALEs). In addition, PSI was used to evaluate the percentage of the inclusion of variable exons relative to the total mature mRNA in the splicing events[30]. The PSI was originally defined for ES events. Here, its definition is expanded to describe the changes in splicing of all the splicing types in our DAS analysis. Specifically, the splicing event types ES, A5SS, A3SS, ME, and IR involve two isoforms where one isoform is longer. We calculated the PSI as the percentage usage of the longer isoform compared with both isoforms. For the splicing events AFE and ALE, we calculated PSI as the percentage of usage of the proximal isoform (the isoform with the variable exon closer to the constitutive exon) relative to both isoforms of the event. The DAS events are identified under $|\Delta\Psi| > 0.05$ and $q < 0.05$.

2.2.2 Gene ontology analysis

To examine the biological functions of the genes in the *Tnip1* KO mice and the human psoriasis dataset, GO analysis was performed to screen for the enriched GO terms using Fisher's exact test[31] with the null hypothesis H_0: log odds ratio < 1. In the test of enriched GO terms for the genes with DAS events, these genes were taken as the foreground, and the expressed genes

were taken as the background. To reveal the enriched GO terms for differentially expressed genes, specifically up-regulated genes were taken as the foreground and expressed genes were taken as the background. The estimated log odds ratio was also retained for overlapped GO terms. Enriched GO terms were identified under p-value < 0.05.

2.2.3 Splicing conservation analysis

To reveal the biological function of DAS events in psoriasis, we performed splicing conservation analysis between mice and humans. We first checked whether the homologous genes between the two species both had the DAS events. By mapping the human gene symbol to the mouse homologous gene symbol using HomoloGene[32], we constructed a contingency table consisting of the counts of the homologous genes in both species with DAS events or in only one species with DAS events. Taking the homologous genes expressed in both mice and humans as the background genes, the Fisher's exact test was used to test the enrichment of common homologous genes with DAS events in both species.

Additionally, we compared the isoform sequences between mice and humans. Within the 89 homologous genes with DAS events, 33 showed ES events in both species. To investigate the conservation of splicing changes in these 33 genes, we performed MSA analysis of the ES events in these genes. We first extracted the two isoform sequences that cover each of the ES events—i.e., the upstream and downstream exons in the event are included in both isoforms, but the variable exon is included in only one of the isoforms. Within each homologous gene, we compared all the mouse-human splicing event pairs. In each comparison, we constructed an MSA of the translated protein sequences or the predicted mRNA sequences of the extracted isoforms in mice and humans using MAFFT[33]. For the coding events, we constructed the MSA for the translated protein sequences. Alternatively, for the events with noncoding isoforms, we built the MSA for the

predicted mRNA sequences. The MSA results between the mouse and human isoforms revealed a commonality of splicing events between mice and humans.

2.2.4. Mouse splicing factor perturbation database

To screen for the candidate SFs that may regulate splicing in psoriasis, we curated a set of mouse RNA-Seq datasets with perturbed SFs. Our curated datasets were deployed as a database called SFMetaDB, which hosts the full mouse RNA-Seq datasets with perturbed SFs (knocked-out/knocked-down/overexpressed). To curate the mouse SF perturbation database in SFMetaDB, we extracted 315 RNA SFs in GO (accession GO:0008380) for the mice[31]. For each SF, we used the gene symbol to search against ArrayExpress[34] for mouse RNA-Seq datasets. For the retrieved results from ArrayExpress, we performed manual curation of the dataset to make sure the SF was perturbed in the dataset. We ended up with 34 mouse RNA-Seq datasets for the perturbation of 31 SFs. These 34 SF perturbation datasets provided the precious raw data for us to induce the candidate SFs that regulate splicing in psoriasis.

2.2.5 Splicing signature–based connectivity map

To identify the candidate SFs that regulate splicing events in psoriasis, we first determined whether the expression of SFs increased or decreased in the *Tnip1* KO mouse dataset and the human psoriasis dataset using the following procedure. The raw RNA-Seq reads were aligned to mouse/human genome using STAR, the same as the DAS analysis. The uniquely mapped reads were used to calculate the read-counts for each gene against the UCSC KnownGene annotation (mm9/hg19). A table of read-counts for all the genes and all the samples was created and normalized by DESeq[35]. The fold change calculated from this normalized count table was used to determine whether the expression of an SF increased or decreased.

Then, we checked how the splicing events were regulated by the SFs in the SF perturbation datasets by comparing these splicing events with the events in the psoriasis datasets. For example, if 1) a splicing event was positively regulated by an SF according to an SF perturbation dataset—i.e., the inclusion of the variable exon of the event was increased (Figure 2.1a) upon the overexpression of the SF or the inclusion of the variable exon of the event was decreased upon the knock-down/knock-out of the SF in the SF perturbation dataset (Figure 2.1b), and 2) the same variable exon was more included in psoriasis along with an increased expression of the SF or the same variable exon was less included in psoriasis along with a decreased expression of the SF, this consistency between 1) and 2) suggests that the event is likely regulated by the SF in psoriasis. If this consistency holds across a significantly large number of events, then the SF is likely a key factor responsible for the regulation of large-scale splicing changes in psoriasis. This consistency comparison approach was also used in CMap, a gene expression signature comparison method that has been widely used to detect the consistency between the gene expression signatures of a disease and the small-molecule or drug-treated samples[36]. Such a signature comparison method based on gene expression is powerful because some of the predictions have been validated *in vivo*[37]. However, most signature comparison approaches mainly focus on gene expression data and fail to detect fine-tuning of gene expression by splicing. To obviate the drawback in CMap, we applied a splicing signature–based comparison method using splicing changes in the SF perturbation datasets and the psoriasis datasets (Figure 2.1c). We first calculated the splicing signatures for the 34 SF perturbation datasets, where $+/-$ indicates that an event is positively/negatively regulated by the given SF of the dataset and 0 indicates that no evidence exists that the event is regulated by the SF. Another signature vector made of $+/-/0$ was used to characterize the relation of an SF and the events in the psoriasis dataset. By comparing a signature from the SF perturbation dataset

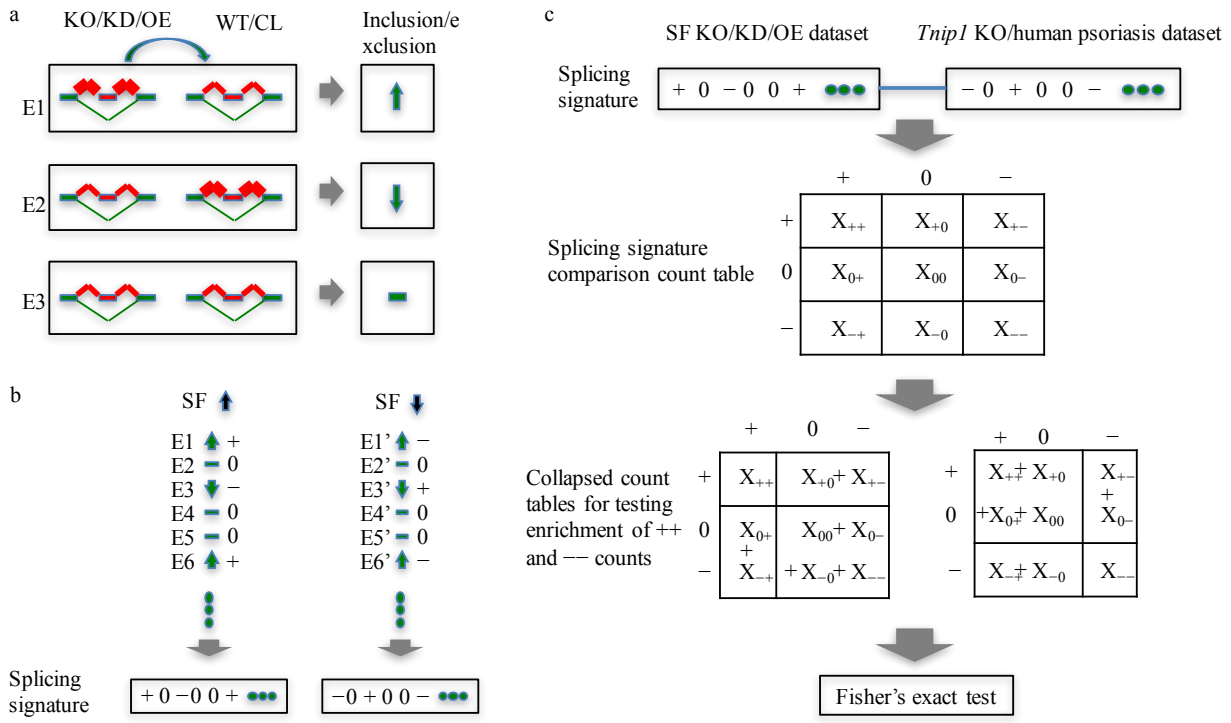


Figure 2. 1 Splicing signature comparison workflow for the discovery of candidate SFs that regulate alternative splicing in psoriasis.

(a) The splicing events direction in the perturbed group. (b) The regulation direction of DAS event. (c) Splicing signatures comparison workflow.

with a signature from the psoriasis data, a 3×3 contingency table was tabulated with rows and columns named +/−/0 and was used to see whether the two signatures match each other. To further check for the direction of the consistency, we collapsed the 3×3 table into two 2×2 tables so that the enrichment of + + events and − − events, respectively, could be tested using Fisher's exact test (Figure 2.1c). The SFs with significantly enriched + + events and − − events are the candidate SFs that regulate the splicing in psoriasis.

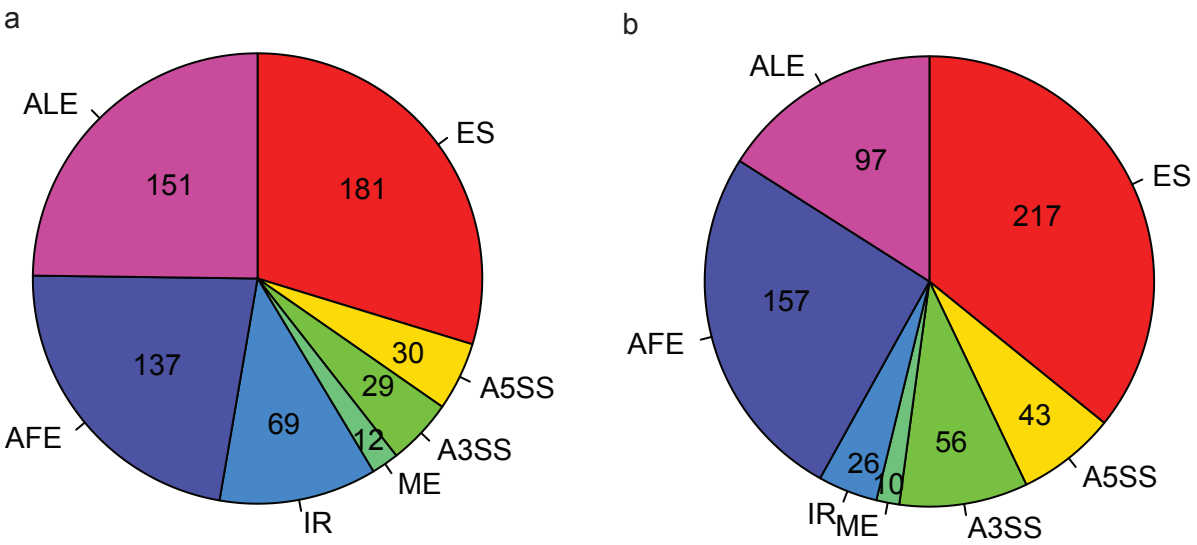


Figure 2. 2 Number of DAS events for the seven splicing event types.

DAS analyses were performed for the mouse and human datasets, involving seven splicing event types: ES, A5SS, A3SS, ME, IR, AFE, and ALE. Under $|\Delta\Psi| > 0.05$ and $q < 0.05$, the pie charts depict the number of DAS events for the seven splicing event types. (a) DAS analysis revealed 609 DAS events in the *Tnip1* KO mouse model dataset. (b) DAS analysis revealed 606 DAS events in the human psoriasis dataset.

2.3 Results

2.3.1 Revealing large-scale changes in alternative splicing by analyzing RNA-Seq data from psoriasis mouse model and human skin

To investigate the role of the splicing process in psoriasis, a psoriasis mouse model was studied first to detect splicing changes. In this mouse model, the gene *Tnip1* was knocked out[23]. Notably, *TNIP1* (the homologous gene of *Tnip1*) in humans is found in a psoriasis susceptibility locus[38]. It has been shown that *Tnip1* knockout (KO) mice exhibit macroscopical psoriasis-like

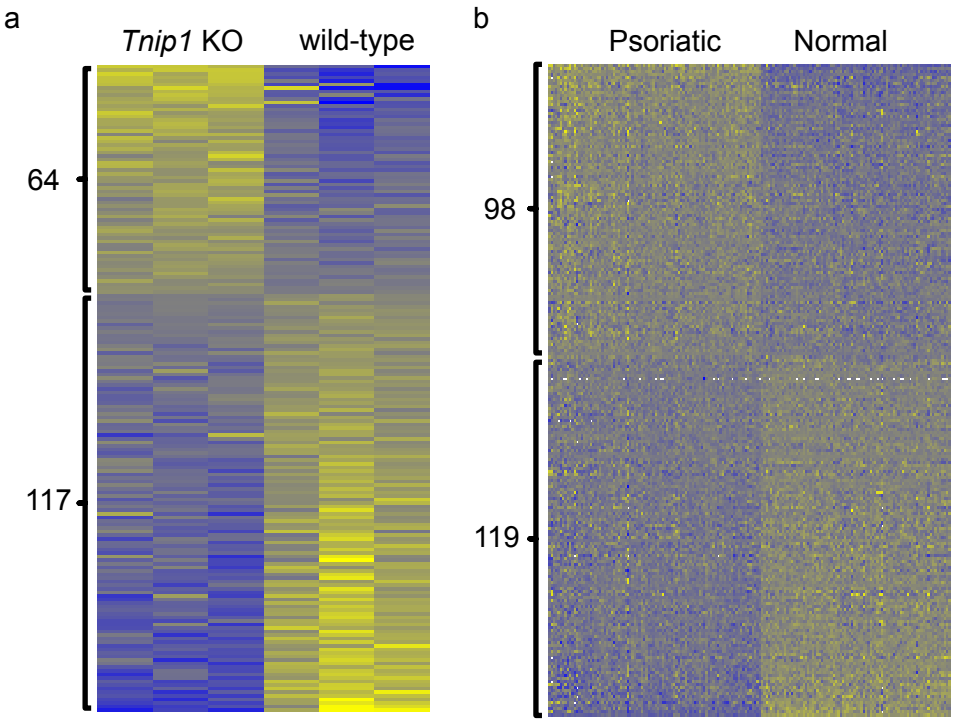


Figure 2. 3 Heat map of PSI values for alternative ES events in the *Tnip1* KO mouse model dataset and the human psoriasis dataset.

Yellow: high PSI. Blue: low PSI. (a) The heat map of the PSI values between three KO samples and three wild-type samples in mice. 64 of 181 ES events have more inclusion of variable exons in psoriasis, and 117 ES events have less inclusion of variable exons in psoriasis. (b) Heat map of the PSI values between 92 lesional psoriatic skins and 82 normal control skins in humans. 98 of 217 ES events have more inclusion of variable exons in psoriasis, and 119 ES events have less inclusion of variable exons in psoriasis.

phenotypes, such as redness and scaling, and microscopical psoriasis-like phenotypes, such as epidermal thickening, elongated rete-like ridges, papillomatosis, retention of nuclei within corneocytes, and infiltrations with different immune cell types[23]. To reveal splicing changes, the Dirichlet Multinomial (DMN) regression[27] was used to analyze the dataset from the *Tnip1* KO

mouse model. Benjamini-Hochberg–adjusted[29] *p*-value and the percent sliced in (PSI, Ψ) were estimated for seven types of splicing events. Under $|\Delta\Psi| > 0.05$ and $q < 0.05$, a total of 609 differential alternative splicing (DAS) events were identified. Figure 2.2a shows the number of DAS events for seven splicing types in the mouse model. To verify that the *Tnip1* KO mouse model recapitulated the main splicing features in human psoriasis, we performed a DAS analysis using RNA-Seq data from psoriasis patients and controls[18]. This DAS analysis identified 606 DAS events ($|\Delta\Psi| > 0.05$ and $q < 0.05$). Figure 2.2b shows the number of DAS events for seven splicing types in the human psoriasis dataset.

Our DAS results revealed many significant splicing events in the psoriasis mouse model. Figure 2.3 shows the heat map of PSI values for ES events in the *Tnip1* KO mouse model and in the human psoriasis dataset. In the *Tnip1* KO mouse model dataset, 64 splicing events have more inclusion of the variable exons in psoriasis, while 117 splicing events have less inclusion of the variable exons in psoriasis. In the human psoriasis dataset, 98 splicing events have more inclusion of the variable exons in psoriasis, while 119 splicing events have less inclusion of the variable exons in psoriasis. To reveal the biological functions of the genes with DAS events, gene ontology (GO) analysis was applied to detect the enriched GO terms for genes with DAS events in both the *Tnip1* KO mice and the human psoriasis dataset. Specifically, the GO term "regulation of wound healing, spreading of epidermal cells" was enriched in both mice and humans. The wound healing process is accelerated in psoriasis, suggesting the potential role of splicing changes in psoriasis[39]. In addition, the actin-filament–related GO terms "negative regulation of actin filament depolymerization" and "actin filament reorganization" were enriched in mice and humans, respectively. Dysregulation of actin filament is observed in psoriatic skins, indicating that splicing changes may contribute to the formation of psoriasis[40]. Therefore, our DAS analysis

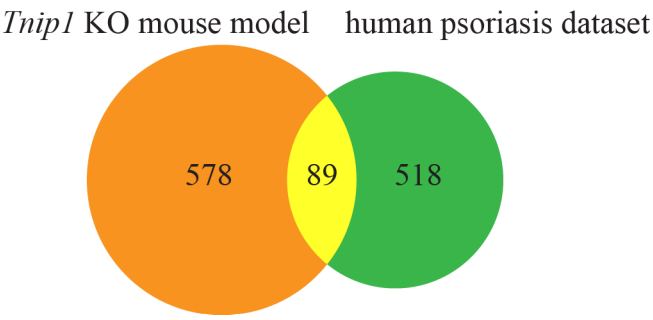


Figure 2. 4 Venn diagram of the genes with DAS events in the *Tnip1* KO mouse model dataset and the human psoriasis dataset.

To investigate the genes with DAS events, we ended up with 667 genes in the *Tnip1* KO mouse model dataset and 607 genes in the human psoriasis dataset. Mapping the human gene symbols to mouse homologous genes using HomoloGene resulted in 89 common homologous genes with DAS events in both species. Alternatively, 578 genes have DAS events in mice but not humans. On the other hand, 518 genes have DAS events in humans but not mice. Taking 12,233 homologous genes expressed in both species as the background genes, the Fisher's exact test showed significant enrichment of the common homologous genes with $p = 1.7\times10^{-32}$.

discovered large-scale splicing changes in psoriasis, providing feasible and promising new features to study the role of splicing in the pathogenesis of psoriasis.

2.3.2 Revealing conserved splicing events in both mice and humans by splicing conservation analysis

To identify the most critical splicing changes in psoriasis, we conducted a splicing conservation analysis to reveal the splicing changes common to both the *Tnip1* KO mouse model dataset and the human psoriasis dataset. By mapping mouse and human gene symbols using HomoloGene[32], we detected 89 homologous genes with DAS events in both mice and humans

(Figure 2.4). The Fisher's exact test showed significant enrichment of the common homologous genes with $p = 1.7\times10^{-32}$. This supports the conclusion that there is commonality in splicing underlying psoriasis in both mice and humans.

To further characterize the conservation of splicing in mice and humans, we compared the isoform sequences between them. By the splicing conservation analysis at the isoform level, we ended up with 24 homologous genes with conserved isoform sequences for the common splicing events in human and mouse gene annotation. The high proportion of conserved isoform sequences for the common splicing events (24 of 33) suggested feasible and promising conservation of splicing changes between the *Tnip1* KO mouse model dataset and the human psoriasis dataset.

To identify the splicing features in psoriasis, we further evaluated whether the common splicing events were conserved in the same isoform between the *Tnip1* KO mouse model dataset and the human psoriasis dataset. Specifically, we checked whether the splicing events shared the same inclusion pattern of variable exons in mouse and human. We ended up with 18 alternative splicing events conserved in the same isoform, which means that the splicing events have more or less inclusion of variable exons in the same way between the two species (Table 2.1). The corresponding 18 homologous genes with conserved alternative splicing events include *ABI1, ARHGAP12, ATP5C1, CTTN, DNM1L, EXOC1, FBLN2, FNBP1, GOLGA2, GOLGA4, MYH11, MYL6, MYO1B, PAM, SEC31A, SLK, SPAG9,* and *ZMYND11*. Of the 18 conserved splicing events, eight were largely spliced in both species, with over 10% PSI differences (Table 2.1). Our conservation analysis identified the 18 conserved splicing events, suggesting that the splicing features in the psoriasis mouse model dataset can be recapitulated in the human psoriasis dataset, and further, the 18 conserved splicing events can be promising targets to follow to study the splicing mechanism in psoriasis.

Table 2. 1 Identification of the conserved splicing events between the *Tnip1* KO mouse model dataset and the human psoriasis dataset.

Gene in human	Gene in mouse	ΔΨ in human	ΔΨ in mouse	Isoform conservation	PSI consistency
ABI1	Abi1	–0.133	–0.095	both	Y
ARHGAP12	Arhgap12	–0.069	–0.309	both	Y
ATP5C1	Atp5c1	0.104	0.259	both	Y
CTTN	Cttn	–0.054	–0.159	both	Y
DNM1L	Dnm1l	–0.096	–0.221	both	Y
EXOC1	Exoc1	0.186	0.198	both	Y
FBLN2	Fbln2	–0.172	–0.173	both	Y
FNBP1	Fnbp1	–0.137	–0.269	both	Y
GOLGA2	Golga2	–0.099	–0.144	both	Y
GOLGA4	Golga4	0.058	0.086	both	Y
MYH11	Myh11	–0.064	–0.225	both	Y
MYL6	Myl6	–0.101	–0.294	both	Y
MYO1B	Myo1b	–0.109	–0.206	both	Y
PAM	Pam	–0.096	–0.327	both	Y
SEC31A	Sec31a	–0.105	–0.115	both	Y
SLK	Slk	0.107	0.217	both	Y
SPAG9	Spag9	–0.100	–0.226	both	Y
ZMYND11	Zmynd11	0.061	0.365	both	Y
AXL	Axl	–0.060	0.091	both	N

Table 2.1 Continued.

Gene in human	Gene in mouse	ΔΨ in human	ΔΨ in mouse	Isoform conservation	PSI consistency
DMKN	Dmkn	–0.180	–0.078	hs_inc=mm_excl [a]	N
MLX	Mlx	0.089	–0.132	both	N
MPRIP	Mprip	–0.127	0.187	both	N
NDRG2	Ndrg2	0.126	–0.159	both	N
POSTN	Postn	–0.072	–0.195	hs_inc=mm_excl [b]	N

The column "PSI consistency" marks 'Y' for the 18 splicing events conserved in both mice and humans.

[a] The ΔΨ s are of the same negative signs, meaning psoriatic samples have more exclusion for the splicing events in DMKN/Dmkn of both species. However, the isoform with more inclusion of the variable exon in the human (hs_inc) is conserved with the isoform with more exclusion of variable exon in the mouse (mm_excl). Therefore, the event is not conserved.

[b] The ΔΨ s are of the same negative signs, meaning psoriatic samples have more exclusion for the splicing events in POSTN/Postn of both species. However, the isoform with more inclusion of the variable exon in the human (hs_inc) is conserved with the isoform with more exclusion of variable exon in the mouse (mm_excl). Therefore, the event is not conserved.

2.3.3 Revealing candidate splicing factors regulating splicing in psoriasis by splicing signature analysis in mouse

To further elucidate the splicing mechanism in psoriasis, we conducted SF screening to discover the candidate SFs that may regulate large-scale splicing events in psoriasis. Because a great number of splicing events are discovered in mouse psoriasis datasets, we hypothesize that SFs may play critical roles in the regulation of these events. To screen for the candidate SFs, we manually curated a list of RNA-Seq datasets with gain- or loss-of-function of mouse SFs[10]. Using the datasets in SFMetaDB, we systematically compared the splicing changes in the psoriasis mouse model dataset with the effects of SF perturbation using a splicing signature comparison workflow (Figure 2.1). Our splicing signature comparison approach screened the SF perturbation datasets related to a total 31 SFs for splicing regulators in the mouse psoriasis dataset, where nine SFs showed significant overlapping splicing changes in psoriasis, including NOVA1, PTBP1, PRMT5, RBFOX2, SRRM4, MBNL1, MBNL2, U2AF1, and DDX5, which are potential regulators responsible for splicing changes in psoriasis.

2.3.4 Confirming the key splicing regulators in humans

To confirm the importance of these nine SFs in mice, we performed a similar splicing signature comparison analysis in humans. Using the human homologous symbols of these nine mouse SFs, we curated on GEO[41] the human RNA-Seq datasets with these genes perturbed. Our curation resulted in four datasets for three of nine human SFs—GSE59884 [42] and GSE69656 [43] for PTBP1, GSE66553 for U2AF1, and GSE76487 [44] for MBNL1. The splicing signature comparison analysis (Figure 2.1) using splicing signatures from the human psoriasis dataset and the four datasets of the three human SF perturbation showed significantly overlapped splicing changes between the human psoriasis dataset and the SF perturbation datasets of three human

SFs— i.e., PTBP1, U2AF1, and MBNL1. These results suggest the important role of these three SFs in potentially regulating splicing in psoriasis.

2.3.5 Revealing potential candidate SFs that regulate splicing in psoriasis using conserved splicing events in SF perturbation datasets

To identify the potential SFs that regulate the conserved splicing events in psoriasis, we investigated the consistency of regulation direction of splicing events in the mouse/human dataset and the SF perturbation datasets. The 18 conserved ES events were significantly conserved in the *Tnip1* KO mouse model dataset and the human psoriasis dataset, indicating the key spliced genes in psoriasis. Upon checking whether the splicing events were positively/negatively regulated by the SF in the same way in the SF perturbed datasets and the psoriasis datasets (Figure 2.1b), we ended up with 12 SFs (CELF1, CELF2, DDX5, MBNL1, MBNL2, NOVA1, PRMT5, PTBP1, RBFOX2, SF3A1, SRRM4, and U2AF1) potentially regulating 13 splicing events (*Abi1, Arhgap12, Atp5c1, Cttn, Exoc1, Fbln2, Golga2, Golga4, Myl6, Pam, Sec31a, Spag9*, and *Zmynd11*) in the *Tnip1* KO mouse model dataset and three SFs (PTBP1, U2AF1, and MBNL1) potentially regulating five splicing events (*ABI1, CTTN, GOLGA2, MYL6*, and *PAM*) in the human psoriasis dataset. These results show the potential SFs that may regulate splicing events in psoriasis.

3. IDENTIFYING KEY FACTORS IN EPIDERMAL DEVELOPMENT AND COLD-INDUCED THERMOGENESIS*

A large volume of biological data is being generated for studying mechanisms of various biological processes. These precious data enable large-scale computational analyses to gain biological insights. However, it remains a challenge to mine the data efficiently for knowledge discovery. The heterogeneity of these data makes it difficult to consistently integrate them, slowing down the process of biological discovery. We introduce a data processing paradigm to identify key factors in biological processes via systematic collection of gene expression datasets, primary analysis of data, and evaluation of consistent signals [8]. To demonstrate its effectiveness, our paradigm was applied to epidermal development and identified many genes that play a potential role in this process. Besides the known epidermal development genes, a substantial proportion of the identified genes are still not supported by gain- or loss-of-function studies, yielding many novel genes for future studies. Among them, we selected a top gene for loss-of-function experimental validation and confirmed its function in epidermal differentiation, proving the ability of this paradigm to identify new factors in biological processes. In addition, this paradigm revealed many key genes in cold-induced thermogenesis using data from cold-challenged tissues, demonstrating its generalizability. This paradigm can lead to fruitful results for studying molecular mechanisms in an era of explosive accumulation of publicly available biological data.

---

* Reprinted with permission from "A data mining paradigm for identifying key factors in biological processes using gene expression data" by Jin Li, Le Zheng, Akihiko Uchiyama, Lianghua Bin, Theodora M. Mauro, Peter M. Elias, Tadeusz Pawelczyk, Monika Sakowicz-Burkiewicz, Magdalena Trzeciak, Donald Y. M. Leung, Maria I. Morasso, and Peng Yu., 2018. *Scientific Reports*, 8, 9083, Copyright 2018 by authors

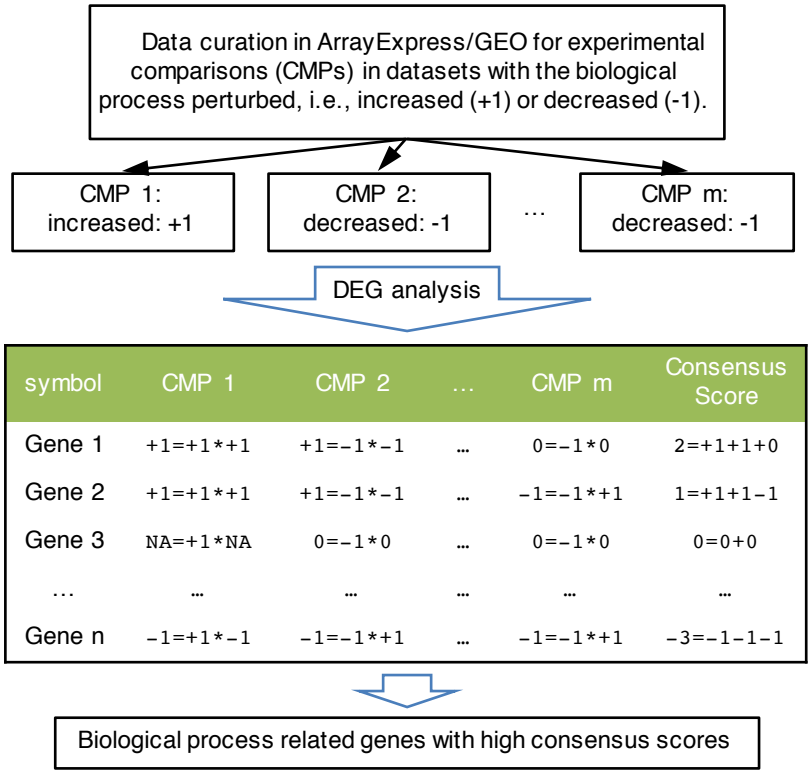


Figure 3. 1 Data processing paradigm flowchart.

Data curation was performed to identify the gene expression datasets with the given biological process perturbed (e.g., the process is increased in CMP 1 with $+1$ and is decreased in CMP 2 or CMP m with direction -1). $+1/-1/0$ represents the up-regulated, down-regulated, or unchanged genes, respectively. An affinity score of $+1/-1/0$ was calculated first by comparing the gene expression change and the regulation of the biological process, where $+1$ indicates that the gene (e.g., Gene 1 in CMP 1 and CMP 2) is positively related to the biological process, -1 indicates that the gene (e.g., Gene 2 in CMP m and Gene n in CMP 1) is negatively related to the biological process, and 0 indicates no relation of the gene to the biological process. No measurement (notated as NA, e.g., Gene 3 in CMP 1) indicates an unknown affinity of the gene in the dataset. By summing the affinity scores, a consensus score was calculated for genes in the perturbed datasets.

3.1 Introduction

The huge amount of data generated from previous biological studies provides a precious resource for mining new biological knowledge. A significant portion of the data is freely available in public repositories such as ArrayExpress [34] and Gene Expression Omnibus (GEO) [41]. For example, around one million series studies are publicly available in GEO. Due to the unstructured nature of the metadata associated with public data, manual curation is required [5, 6, 10, 45, 46], a step that is essential for collecting large-scale gene expression data.

Gene expression data facilitate the application of the network reconstruction approach for identifying key factors in biological processes. For example, Bhaduri et al. [47] applied the gene network reconstruction approach to explore epidermal differentiation regulators. Using network analysis, the *MPZL3* gene was identified as a highly connected hub required for epidermal differentiation. In addition, the *MPZL3* gene indirectly regulates epidermis genes, including *ZNF750*, *TP63*, *KLF4*, and *RCOR1*, through the *FDXR* gene and reactive oxygen species.

Complementing data analyses with more relevant data improves the identification of key factors in biological processes. Even though massive expression data can provide essential insights in revealing genetic interactions, there are confounding factors or "noise" introduced by technical variations, such as batch effects [48]. To obviate the "noise" and generate a consistent result, one solution is integrative analysis by comparing large-scale datasets [49]. In this section, we introduced a paradigm to integrate data collection and data analysis for mining key factors in specific biological processes (Figure 3.1). To demonstrate the power of our data processing paradigm, we evaluated key factors of two applications in skin biology and energy homeostasis.

The epidermis of skin mediates various functions that protect against the environment, such as microbial pathogen challenges, oxidant stress, ultraviolet light, chemicals, and mechanical

insults [50]. Therefore, it is critical to understand mechanisms of epidermal development to develop new treatment for human skin diseases [13]. Our paradigm predicts key factors in epidermal development by collecting related datasets and integrating the information. A fraction of genes are annotated in Gene Ontology (GO) or have strong functional validation based on gain-/loss-of-function studies [51]. The remaining genes are novel; their functionality has not been experimentally validated. We picked a top hit, suprabasin (*SBSN*), and performed loss-of-function experiments for the mouse homolog of gene *Sbsn* using RNA-Seq. The analysis validates that *Sbsn* knockdown in mouse keratinocyte cultures down-regulates cornified envelope genes, suggesting an essential role of *SBSN* in epidermal differentiation. These results demonstrate the effectiveness of our paradigm in discovering key factors of epidermal development.

As another application, cold-induced thermogenesis (CIT) can reduce body weight by increasing resting energy expenditure in mammals [52]. Genes involved in CIT can be promising therapeutic targets for treating obesity and diabetes. Thus, it is important to understand the underlying mechanism of CIT. Our paradigm detected potential CIT-related genes, including known CIT genes and novel ones, showing that the paradigm can be generalized easily to other biological processes. It is a promising integrative analysis approach to identify key factors in biological processes.

3.2 Methods

3.2.1 Curating gene expression data related to epidermal development

We collected gene expression datasets related to epidermal development by manual curation according to the following procedure. First, we searched ArrayExpress using the keyword (“epidermis+development” OR "epidermal+development”) AND organism: “homo sapiens”, retrieving only five studies, none of which could be reused to study the epidermal development

process because of no change in epidermal development in the datasets. Therefore, we started from known epidermal development genes to curate datasets with the process perturbed. Specifically, genes from the GO [31] epidermis development (accession GO:0008544) term were extracted first for humans. Then, the official symbol of each gene was queried on ArrayExpress for human microarray datasets. Each retrieved dataset was manually examined to retain only the datasets with at least one epidermis development gene being perturbed (i.e., knocked out, knocked down, or overexpressed). To ensure proper downstream statistical analysis, any dataset with no replicates was discarded.

3.2.2 Data processing paradigm of the perturbed expression data

To identify the genes related to a biological process, our data processing paradigm was performed on the gene expression data to capture the affinities between specific genes and the biological process. An affinity score of $+1$ or -1 means that the gene is positively or negatively related to the biological process. Specifically, if the expression of a gene is increased or decreased in a biological process that is increased, the gene has an affinity score of $+1$ or -1 for the biological process. Alternatively, if the biological process is decreased, these genes have an affinity score of -1 or $+1$. The affinity score was 0 or NA for the genes not differentially expressed or unmeasured. The detailed workflow of the paradigm is shown in Figure 3.1. For a biological process, systematic data curation is performed to collect gene expression datasets with the process perturbed (increased or decreased). Using DEG analysis [53-55], affinity scores are calculated for each gene in each comparison in each dataset. Finally, a consensus score is calculated by summing these affinity scores among the comparisons for each gene. High consensus scores suggest that the corresponding genes are potentially critical to the biological process. Thus, our paradigm is a general framework that can be used to identify the key factors in a biological process.

3.2.3 DEG analysis using Sbsn knockdown RNA-Seq data in mouse differentiating primary keratinocyte cultures

To identify the differentially expressed genes in mouse differentiating primary keratinocyte cultures in which *Sbsn* had been knocked down with siRNA, the following analysis was performed. The raw RNA-Seq reads were aligned to the mouse (mm10) genome using STAR (version 2.5.1b) [24] with default settings. The uniquely aligned reads were retained to calculate the read counts for each gene against the UCSC KnownGene annotation (mm10), and a count table was constructed by counting the number of reads aligned uniquely to each of the genes for each sample. DEG analysis was performed by DESeq2 [56]. To adjust the batch effect, a generalized linear model with a batch factor was used to model the read counts for all samples, and the Wald test was used to test the significance of differences in gene expression between *Sbsn* knockdown samples and controls. FDR adjusted *q*-values were then calculated from the *p*-values in the Wald test using the Benjamini-Hochberg procedure [29]. The log2-fold changes between *Sbsn* knockdown samples and controls were also calculated for each gene. The differentially expressed genes were identified under |log2-fold-change| > 0.5 and $q < 0.05$.

3.2.4. Comparisons on the curated datasets with respect to epidermal development

To assess the ability of the paradigm, differentially expressed genes using individual comparisons were compared to top identified genes. For individual comparisons, the genes were ordered by contrasts for the increased process or by negative contrasts for the decreased process. Because there were 295 genes in the epidermis development GO term (accession GO:0008544), the same number of genes was selected as top genes in each individual comparison. Further, 295 top genes were selected for the combined comparisons. Some genes tied at the rank of 295, so 24 instances of random sampling (to generate the same number of observations as individual

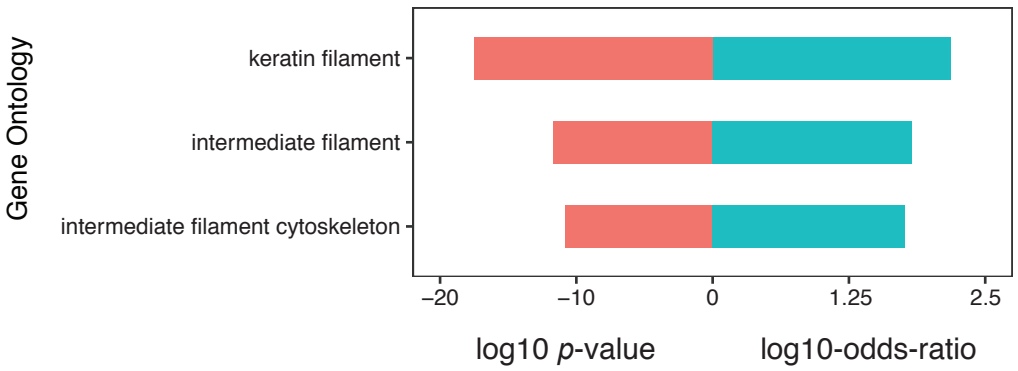


Figure 3. 2 GO terms enriched in the co-evolved genes of *SBSN*.

Three GO terms, keratin filament, intermediate filament, and intermediate filament cytoskeleton, were significantly enriched in the 59 co-evolved genes of *SBSN* in humans.

comparisons) were performed within these tied genes to keep only a total of 295 genes. Within each of these sets of 295 genes, the number of the genes in the epidermis development GO term was recorded. To test the difference of epidermal development genes between individual comparisons and top identified genes, one-sided Wilcoxon test was applied over the recorded number of epidermal development genes.

3.2.5. Phylogenetics-based GO analysis

Because the function of *SBSN* has not been elucidated, it is important to derive an unbiased indication regarding its biological function. For this purpose, a GO analysis based on a gene set derived by a phylogenetic approach was performed using the following procedure. Co-evolved genes of human *SBSN* were first detected using the human-centric binary phylogenetic matrix from Clustering by Inferred Models of Evolution (CLIME) [57]. The human-centric phylogenetic matrix in CLIME was built by searching the protein sequence of each gene in humans against the protein sequences in the rest of 138 fully sequenced eukaryotic organisms [58] and in a

"prokaryote" outgroup of 502 prokaryotic species using BLASTP [59]. In this matrix, rows are human genes, and columns are the 138 eukaryotic organisms together with the "prokaryote" outgroup. Each element in the matrix is binary, which takes 1 if the human protein sequence of the gene in the row is similar to the sequence of a protein in the species of the column; otherwise it takes 0. Then, the Fisher's exact test was applied to evaluate the significance that each gene was co-evolved with *SBSN* among 138 eukaryotic organisms and the "prokaryote" outgroup. A total of 59 genes were co-evolved with *SBSN* under $p < 1.0\times10^{-6}$. These 59 co-evolved genes were used to screen for the enriched GO terms using Fisher's exact test (with the null hypothesis H_0: log-odds-ratio < 2) with the genes appearing in all human GO terms as background [31].

The GO analysis resulted in three significantly enriched GO terms related to epidermal development: keratin filament, intermediate filament, and intermediate filament cytoskeleton (Figure 3.2). For example, keratin filament has shown to be critical in the formation of skin disorders [60]. These enriched GO terms identified by the co-evolved genes of *SBSN* indicate a potentially critical role of *SBSN* in epidermal development.

3.2.6. Expression increase of SBSN upon epidermal differentiation

To evaluate the gene expression changes of *SBSN* upon epidermal differentiation, a microarray dataset (GSE52651) measured in a 7-day time-course keratinocyte differentiation experiment was analyzed. Human progenitor keratinocytes were seeded onto devitalized dermis to enable keratinocyte differentiation into fully stratified epithelium, which captured dynamic changes in tissue regeneration [61]. With log2 transformation and quantile normalization of raw probe expression values, Figure 3.3 shows increased expression of *SBSN* upon epidermal differentiation starting from day 1. The early increase of its expression values upon the induction of differentiation indicates a potentially critical role of *SBSN* in epidermal differentiation.

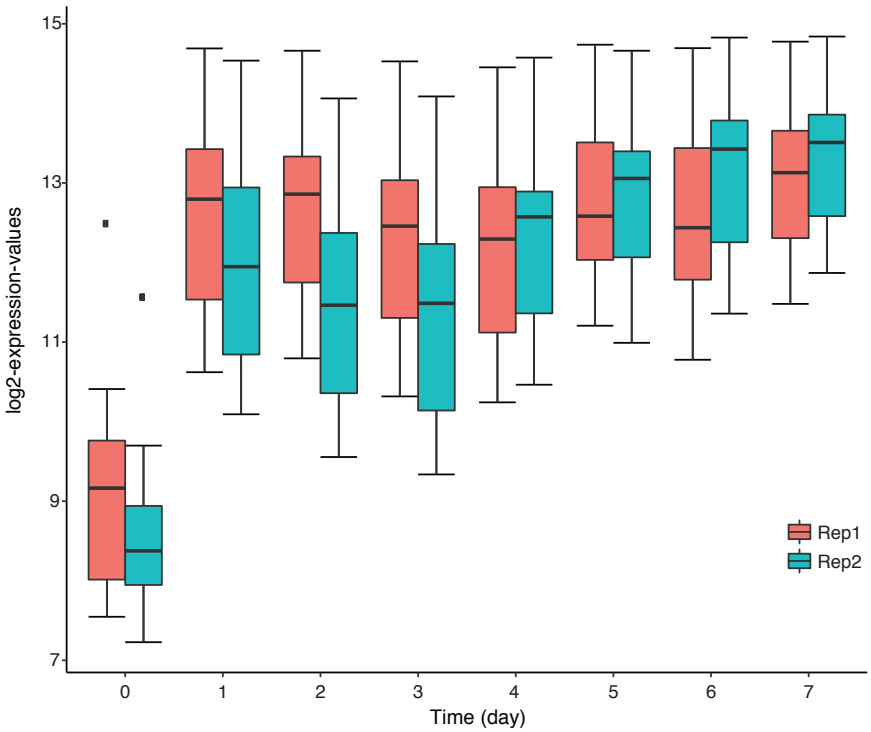


Figure 3. 3 Expression changes of *SBSN* in human keratinocytes upon epidermal differentiation.

To investigate the gene expression changes of *SBSN* upon epidermal differentiation, a time-course microarray dataset was used to measure the expression values of *SBSN*. Human keratinocytes were treated to induce differentiation for discrete time points of seven days. The boxplot shows the normalized log2-expression values of the 11 probes in the microarray mapped to *SBSN* with two biological replicates measured per day. The expression of *SBSN* was significantly up-regulated in days 1 to 7 compared with day 0 (*t*-test using linear contrast *p*-value $< 2.2\times10^{-16}$).

3.2.7. Clustering analysis using the affinity distance metric based on Fisher's exact test

To investigate the relationship of the 24 experimental comparisons in the curated datasets, clustering analysis was performed using an affinity distance metric. The affinity distance metric

was derived from an affinity score matrix calculated in the paradigm (Figure 3.1). An affinity score (annotated as $+1/-1/0$ or NA) of a gene in an experimental comparison examines the relatedness of the gene to a biological process. To evaluate the similarity of the results of two experimental comparisons, a 3×3 contingency table, labeled as $+1/0/-1$, was tabulated by counting the number of genes from the two columns in the affinity score matrix; the table then was collapsed into two 2×2 tables such that the enrichment of the genes having $+1$s or -1s in both experimental comparisons could be tested using Fisher's exact test. The geometric mean of the two *p*-values calculated from the two 2×2 tables corresponding to $+1$s and -1s was considered the affinity distance between the two experimental comparisons. A smaller affinity distance indicates a closer relationship between the two experimental comparisons. To examine the relationships among the 24 experimental comparisons in our curated datasets, the affinity distances were calculated for all pairs of 24 comparisons and were saved in an affinity distance matrix. Then, hierarchical clustering with complete linkage was applied to this matrix.

3.2.8. Empirical distribution of consensus score

To determine the cutoff of consensus scores, simulations were performed to generate the empirical distribution. Specifically, the consensus scores for all genes and 24 comparisons in epidermal development were used to construct an original score matrix, with rows as genes and columns as comparisons. To perform the simulation, the affinity scores for each comparison (each column) were permutated. After all columns were permutated, the consensus scores were calculated for each row. A total of 10,000 iterations of simulation were executed to generate the empirical distribution of consensus scores.

Table 3. 1 Result of dataset curation on GEO by the epidermis development GO term genes

GSE No.	Perturbed Gene	Perturbation	Experiment Tissue	Abbreviation	Tissue Type
GSE37637	EXOSC9	Overexpressed	Primary human keratinocytes	PHK	Epidermal tissue
GSE71017	GRHL2	Knockdown	Ovarian cancer cell line OVCA429	OVCA429	Cancer cell
GSE37049	GRHL3	Knockdown	Primary human normal neonatal keratinocytes	NHEK	Epidermal tissue
GSE32685	KLF4 ZNF750	Knockdown	Primary neonatal keratinocytes	HEKn	Epidermal tissue
GSE1676	RELA	Knockdown	HEK 293	HEK 293	Organotypic tissue
GSE62454	RUNX1	Knockdown	LNCaP cell line	LNCaP	Cancer cell
GSE24778	RUNX1	Knockdown	K562 cells	K562	Cancer cell
GSE8640	TFAP2A TFAP2C	Knockdown	MCF7	MCF7	Cancer cell
GSE28448	SMAD4 TIF1	Knockdown	HMEC-TR	HMEC-TR	Epithelial cell
GSE33495	TP63 TP63 TP63	Knockdown	Primary neonatal keratinocytes	HEKn	Epidermal tissue
GSE38039	ZNF750	Knockdown	HaCaT cells	HaCaT	Cancer cell
E-MTAB-1833	CUX1	Knockdown	Loucy cells	Loucy	Organotypic tissue
GSE27275	PITX2	Knockdown	Trabecular meshwork (TM) tissue from eye	TM	Trabecular meshwork cell
E-MTAB-900	RELA	Knockdown	HEK293T	HEK293T	Organotypic tissue
GSE70940	SMAD4	Overexpressed (8hr) Overexpressed (24hr) Overexpressed (48hr)	Pancreatic ductal adenocarcinoma (PDAC) cell line BxPC3	PDAC	Cancer cell
GSE28558	SNAI1	Knockdown	A549	A549	Cancer cell
GSE44203	TFAP2C	Knockdown	MCF7	MCF7	Cancer cell

3.2.9. DEG analysis using microarray data

For each of the curated human microarray datasets in Table 3.1, DEG analysis was performed as described below. To map microarray probes to gene symbols, the probe sequences were aligned to the transcript sequences of the GENCODE human annotation (release 25) [62]

using Bowtie (version 1.1.2) [63] with an exact match. The probes aligned to multiple genes were discarded. The raw microarray probe data were then rank-normalized by transforming the raw probe values to ranks scaled to [0, 1] by dividing the total number of probes in each platform. These scaled ranks were transformed by the variance-stabilizing transformation (VST) [64]. The resulted VST values were fit using linear models with adapted FDR in contrasts [29, 65]. The DEGs were identified as FDR ≤ 0.05.

3.2.10. DEG analysis using RNA-Seq data

For DEG analysis using RNA-Seq data, raw full-length of the single-end or the first end of the paired-end reads were first aligned to the transcriptome sequences annotated in GENCODE (mouse release M12) [66] eliminating pseudogenes using STAR (version 2.5.3a) [24] ignoring multiple alignment reads. A count table was tabulated of the number of reads aligned to each gene, discarding those reads aligned to multiple genes. Genes with low counts were filtered out from the count table. Normalization and DEG were conducted using DESeq2 [56]. FDR-adjusted q-values were computed using the Benjamini-Hochberg procedure [29]. The DEGs were identified as |log2-fold-change| > 0.5 and $q < 0.05$.

3.3 Results

3.3.1 Identification of candidate epidermal development genes

To identify key gene expression datasets that are likely to be related to epidermal development, data curation was performed. A total of 295 epidermis development genes (according to GO) were searched on ArrayExpress to query microarray datasets, and over 300 datasets were retrieved. Due to the limitation of the search function in ArrayExpress, many retrieved datasets did not have any perturbation of these epidermis development genes, even

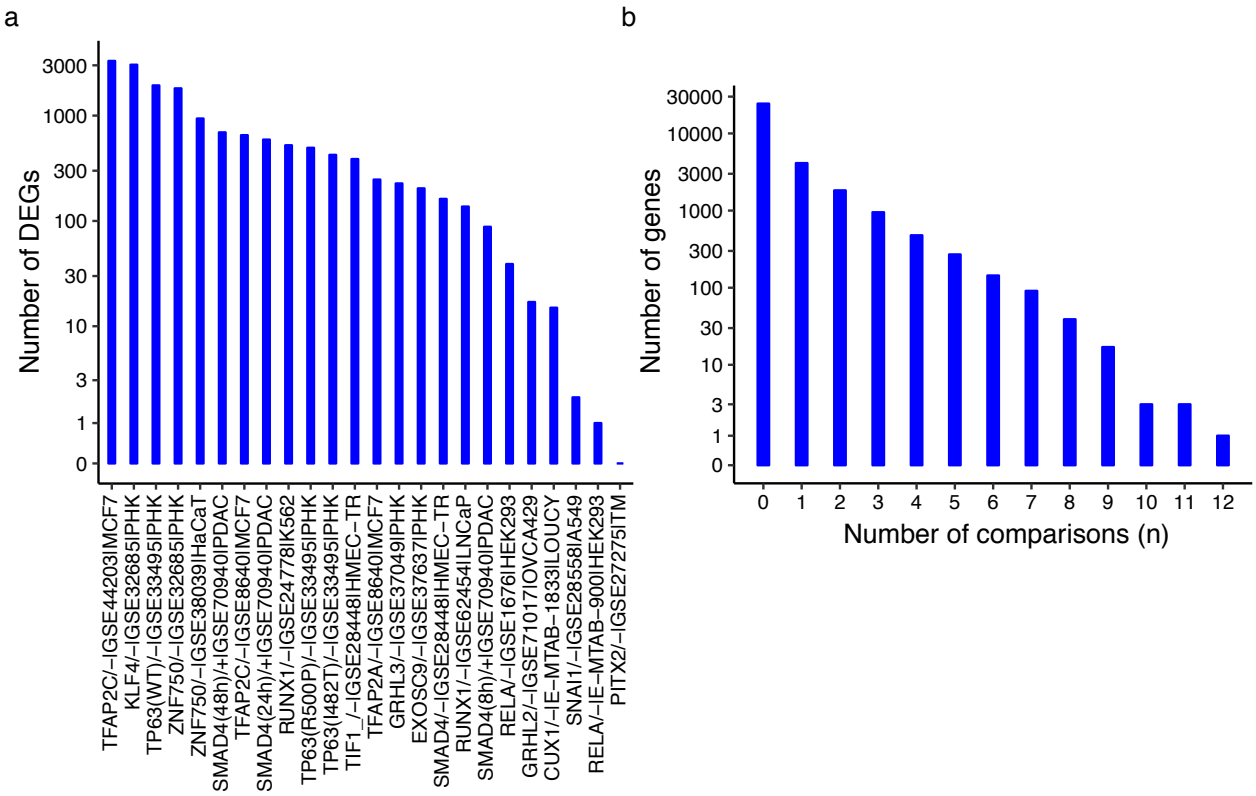


Figure 3. 4 DEG results of the curated microarray datasets.

To identify the differentially expressed genes of the 24 experimental comparisons in curated microarray datasets, DEG analysis was performed as mentioned in supplemental materials. DEGs were identified under $q \leq 0.05$. (a) The bar plot depicts the number of DEGs identified in each of the 24 experimental comparisons. (b) The figure depicts the number of genes differentially expressed in n comparisons out of all the 24 comparisons. A large number of DEGs were identified in the curated datasets. A small group of genes was differentially expressed in multiple datasets.

though the gene symbols were mentioned in the datasets. To overcome this problem, manual curation was performed on each retrieved dataset to retain relevant ones, and the manual curation

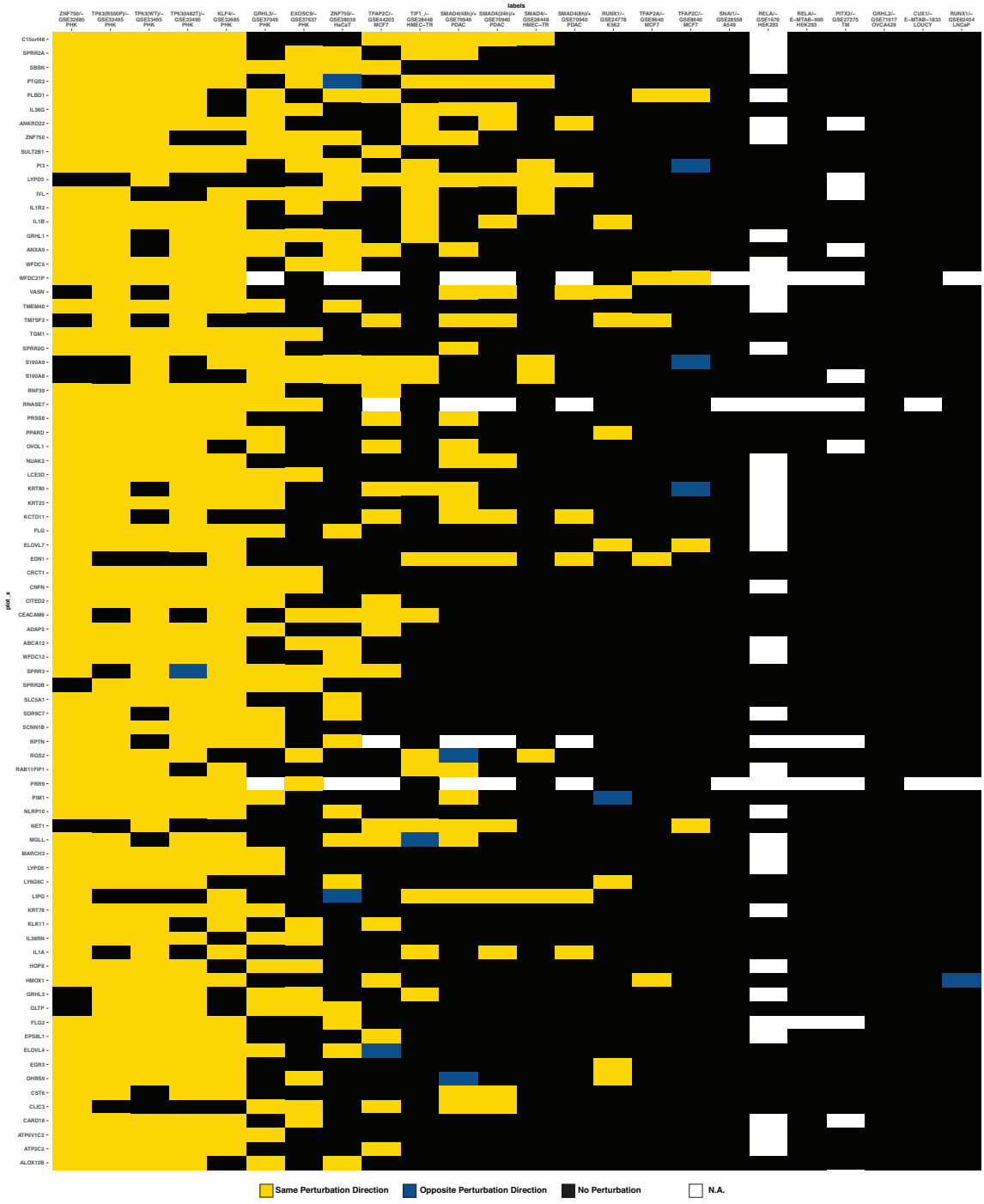


Figure 3. 5 Heatmap of the top genes (consensus score ≥ 6) in epidermal development derived from 24 experimental comparisons of the curated datasets.

Figure 3.5 Continued. Heatmap of the top genes (consensus score ≥ 6) in epidermal development derived from 24 experimental comparisons of the curated datasets.

To identify the candidate genes that are potentially important in epidermal development, the paradigm was applied to the curated datasets. A total of 81 top genes (consensus score ≥ 6) revealed a set of candidate genes involved in epidermal development. Each column in the heatmap represents one of the 24 experimental comparisons in the curated datasets. For example, "ZNF750/-; GSE32685; PHK" represents the dataset (GSE32685) in which *ZNF750* was knocked down in primary human keratinocytes. Each row corresponds to a gene that was examined in those experimental comparisons. The colors yellow/blue/black/white correspond to the affinity scores $+1/-1/0/$ NA, respectively. These 81 top genes showed an affinity in epidermal tissues, demonstrating potential roles of these genes in epidermal development.

resulted in 24 experimental comparisons from 17 datasets with gain or loss function of 14 epidermis development genes (Table 3.1).

To determine the candidate genes potentially involved in epidermal development, differential gene expression (DEG) analysis was performed on the 24 experimental comparisons of the curated microarray datasets. Differentially expressed genes were identified under $q \leq 0.05$. The large-scale gene expression changes derived from our curated datasets provided a list of candidate genes that may be potentially involved in epidermal development (Figure 3.4).

To identify genes that are potentially critical in epidermal development, consensus gene scores were summarized for each gene from affinities on the 24 experimental comparisons. Eighty-one genes were identified as key genes related to epidermal development with a consensus score

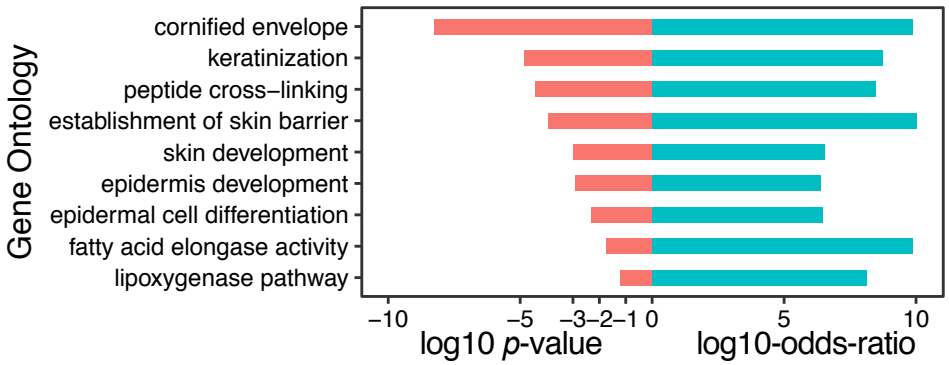


Figure 3. 6 Biological process and literature study of genes with consensus score ≥ 6.

To identify the biological process that the 81 top genes (consensus score ≥ 6) were involved in, a GO enrichment analysis was performed. The enriched GO pathways were plotted with a log10 *p*-value, along with their log10 odds ratios in the enrichment analysis.

≥ 6. The heatmap (Figure 3.5) shows a majority of these genes with a +1 affinity score in skin-related cell types. This information suggests that these top genes may play a role in epidermal development. To infer the biological processes involved, GO analysis was performed on these top genes using Fisher's exact test (the null hypothesis is log-odds-ratio < 2) with all the genes annotated in GO as the background. Several epidermis-related GO terms were enriched in these genes (Figure 3.6). For example, the essential GO terms in the epidermis were enriched, such as keratinocyte differentiation, epidermal cell differentiation, epidermis development, skin development, cornified envelope, and keratinization. In addition, the GO terms involved in skin barrier formation were also enriched, such as fatty acid elongase activity, lipoxygenase pathway, and establishment of skin barrier. These enriched GO terms suggest that the top identified genes are critical in epidermal development.

Because GO annotation is not complete for gene functions [67], we manually curated functional annotations for the top identified genes. Of these genes, besides the 18 genes annotated

in the GO term "epidermis development," only three genes have loss-of-function experiments supporting their role in epidermal development. However, the majority of these identified genes have no functional experimental validation on epidermal development. Of the three genes with literature evidence, *EDN1* (consensus score = 7) mediates the homeostasis of melanocyte (located at the bottom of epidermis) *in vivo* upon ultraviolet irradiation [68]. The loss function of *ELOVL4* (consensus score = 6) represses the generation of very-long-chain fatty acids, which is critical for the epidermal barrier function, showing the important role of *ELOVL4* in epidermis development [69]. The *in vitro* loss-of-function experiment of *HOPX* (consensus score = 6) leads to increased expression of cell differentiation markers in human keratinocytes, demonstrating its involvement in epidermal development [70].

To evaluate how well the roles of the identified genes are understood in epidermal development, we queried the PubMed literature database and examined the results. For each gene, the keyword used in the PubMed search was constructed as "<symbol>[tiab] AND (epidermis OR skin)". The search results showed that a large proportion of identified genes ($\sim$42% = 34/81) have no publications related to skin. Therefore, these understudied novel genes revealed potential candidate genes for new studies on epidermal development. In addition, the majority ($>$ 70%) of identified genes were not in the epidermis development GO term (Figure 3.7). These novel genes demonstrate the ability of the paradigm to discover unknown factors in epidermal development.

To demonstrate the effectiveness of the paradigm computationally, top-ranked genes using collective comparisons were compared to genes using individual comparisons. Figure 3.8 shows

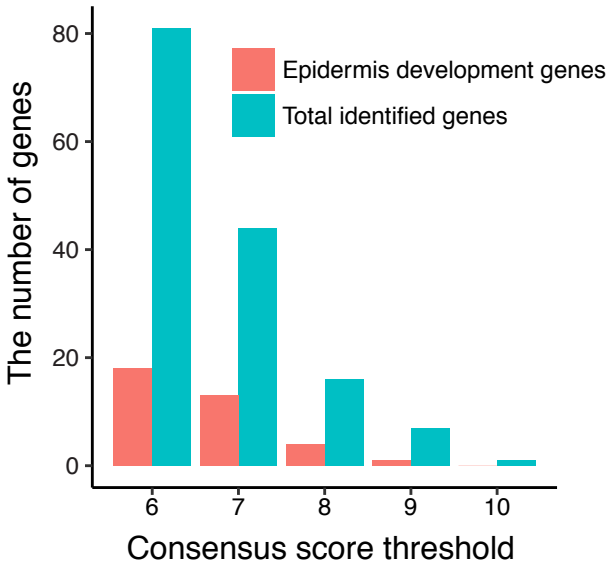


Figure 3. 7 The majority of identified genes were not annotated in the epidermis development GO term.

To evaluate the effectiveness of the paradigm in identifying new factors in epidermal development, the top identified genes were overlapped with the genes in the epidermis development GO term (GO:0008544). These identified genes were extracted using consensus score thresholds from ≥ 6 to ≥ 10. The green and red bars depict the number of total identified genes given the threshold and the number of the genes in the epidermis development GO term, respectively. The majority of identified genes (consensus score ≥ 6) were not in the epidermis development GO term.

the significantly (p-value = 3.6×10^{-9}) increased epidermal development genes identified by the paradigm compared to differentially expressed genes derived from individual comparisons.

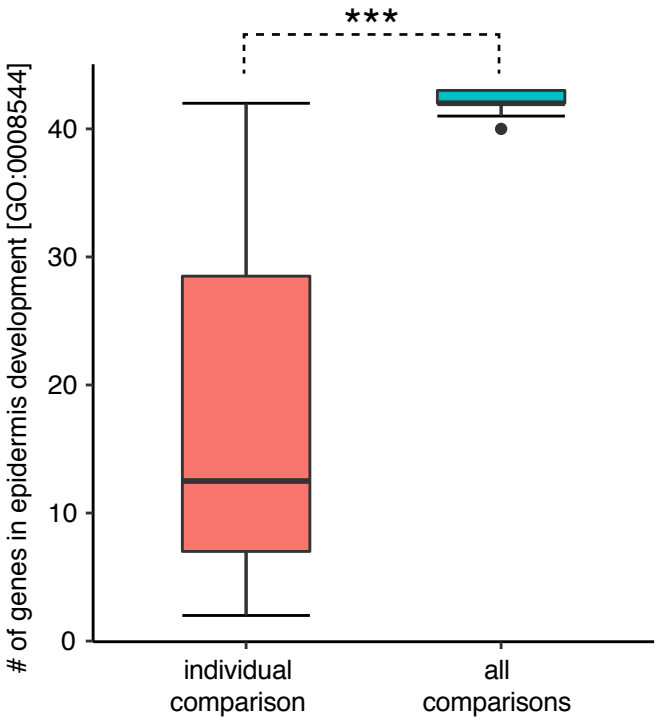


Figure 3. 8 The paradigm revealed an increased number of epidermal development genes. To demonstrate the power of the paradigm, differentially expressed genes derived from individual comparisons were compared to the top ranked genes using all the comparisons. One-sided Wilcoxon test was used to test the significance of the difference between the number of epidermal development genes from two approaches. (***: p-value < 0.001)

3.3.2 Validation of Sbsn role in epidermal differentiation by loss-of-function and other experiments

Among the identified genes, a top gene (*SBSN*) (with a high consensus score of 9) was selected to validate its role in epidermal development. A phylogenetics-based GO analysis revealed enriched GO terms related to epidermal development using co-evolved genes of *SBSN* (Figure 3.2). In addition, a time-course microarray dataset showed an increased expression of

SBSN upon epidermal differentiation (Figure 3.3). These results suggest a potentially critical role of *SBSN* in epidermal development. To determine the cellular component that *Sbsn* is involved with, we performed a study of the differentially expressed genes in differentiating mouse primary keratinocyte cultures from mice with *Sbsn* knockdown. In *Sbsn* knockdown mouse cultures, 326 genes were up-regulated, and 161 genes were down-regulated (Figure 3.9a). To investigate the functional roles of *Sbsn*, these differentially expressed genes were used to search for enriched GO terms [31] using Fisher's exact test (null hypothesized log-odds-ratio < 2) with the genes expressed in the *Sbsn* knockdown mouse culture and the controls as background. Specifically, the cornified envelope GO term was found enriched in the genes down-regulated upon *Sbsn* knockdown (p-value < 0.05), and eight cornified envelope genes were down-regulated (Table 3.2). These results suggest the role that *Sbsn* may play in epidermal differentiation and cornified envelope formation.

Table 3. 2 Eight enriched cornified envelope genes in *Sbsn* knockdown mouse differentiating keratinocyte cultures.

Cnfn
Lce1g
Lce1h
Lce3c
Lce3d
Lce3e
Sprr2d
Sprr2e

Atopic dermatitis (AD) is the most common chronic inflammatory skin disease [71]. IL-4, a type 2 cytokine, contributes to the development of AD. Because broad defects of cornified

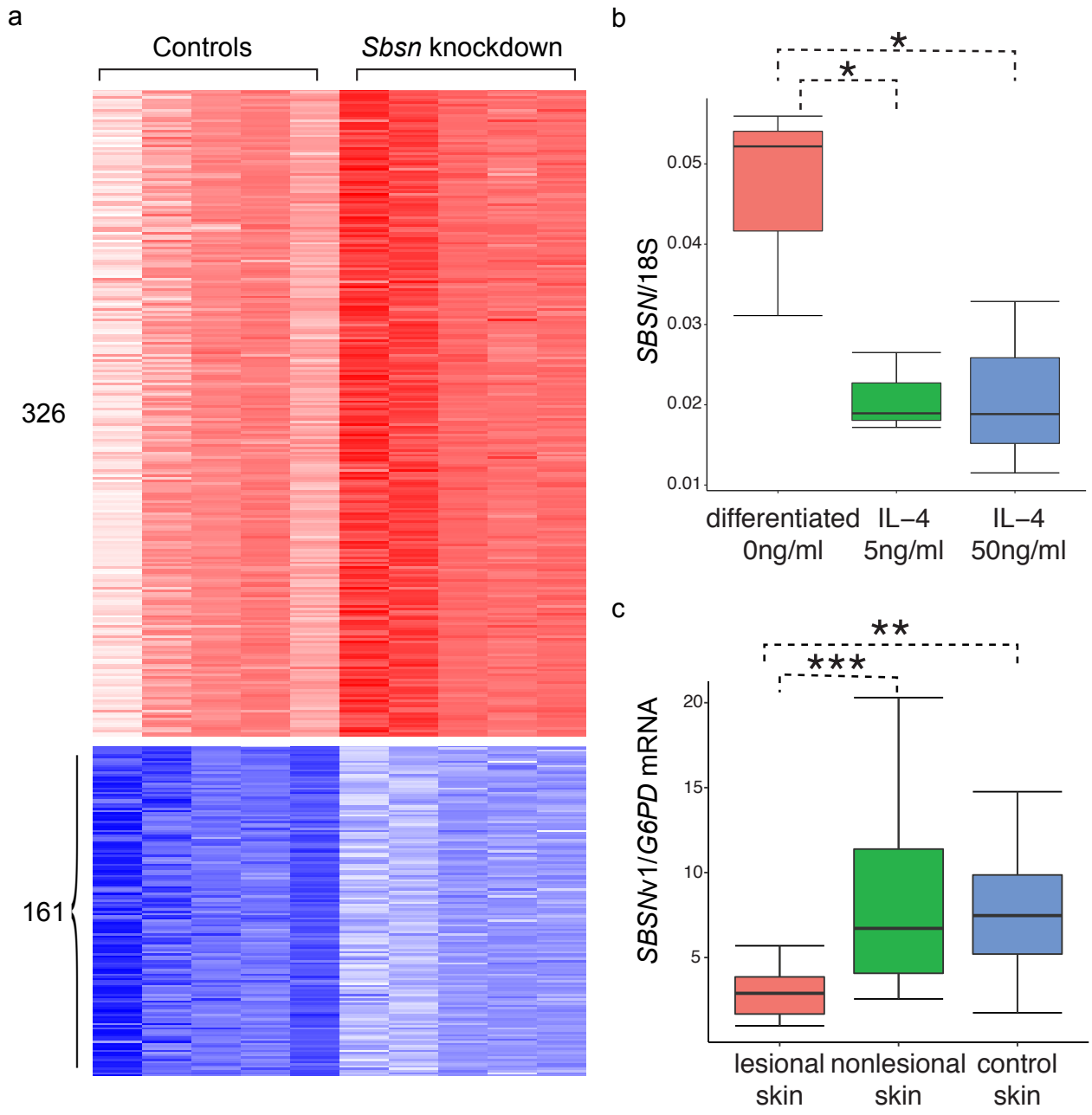


Figure 3. 9 Validations of *SBSN* in epidermal differentiation.

Figure 3.9 Continued. Validations of *SBSN* in epidermal differentiation. (a) Heatmap of the expression levels between *Sbsn* knockdown mice and controls. Expression levels are shown for genes differentially expressed (under |log2-fold-change| > 0.5 and q-value < 0.05) upon *Sbsn* knockdown. Red and white colors indicate high and low expression levels (arc-sine hyperbolic transformed normalized counts by DESeq and scaled by standard deviations) for 326 up-regulated genes, respectively. Blue and white colors indicate high and low expression levels for 161 down-regulated genes, respectively. (b) Expression values of *SBSN* normalized by 18S rRNA in differentiated keratinocytes upon IL-4 treatment. To evaluate the gene expression changes of *SBSN* during keratinocyte differentiation upon IL-4 treatment, an RT-PCR experiment was performed with nine differentiated cells with and without IL-4 treatments (three replicates per condition). The expression values of *SBSN* were normalized by the expression levels of 18S rRNA. The boxplot shows a significant decrease of *SBSN* expression at two IL doses (5 ng/ml and 50 ng/ml) (*: p-value < 0.05). (c) Expression values of full-length *SBSN* transcript (v1) in AD skins. To evaluate the expression changes of *SBSN* in AD skins, expression values were measured in AD skins for the *SBSN* transcripts via RT-PCR. The expression levels were normalized by the expression levels of *G6PD*. The full-length *SBSN* transcript showed significantly decreased expression levels in AD lesional skins compared to AD nonlesional and control skins (***: p-value < 0.001, **: p-value < 0.01, *: p-value < 0.05).

envelope have been identified in AD [72], *SBSN* may play a critical role in AD via defective cornification. To investigate the putative role of *SBSN* in AD, differentiated primary normal human

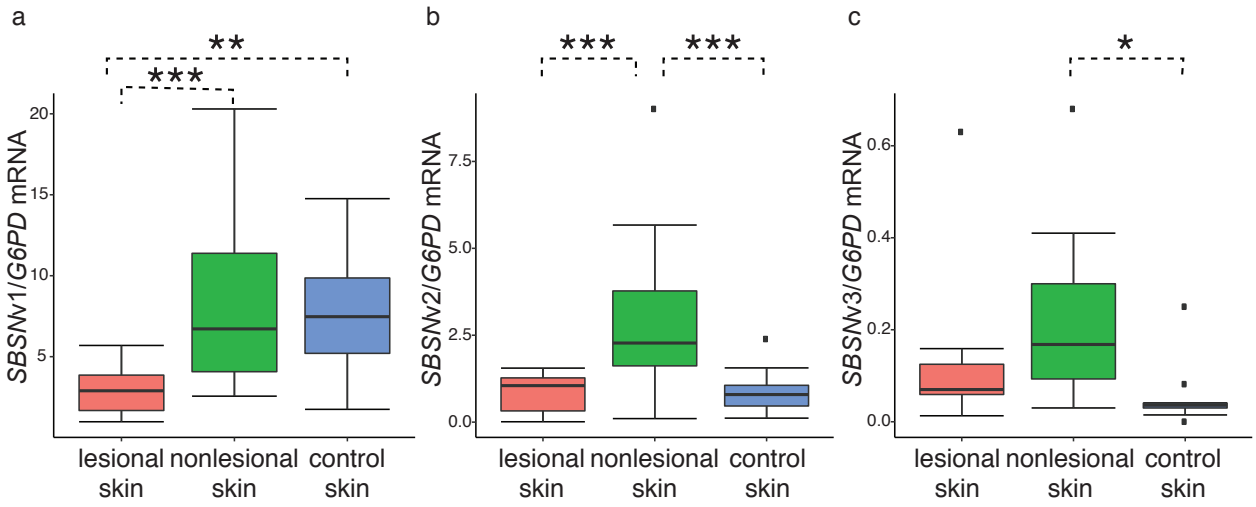


Figure 3. 10 Expression values of *SBSN* transcript v2 and v3 in AD skins.

To evaluate the expression changes of *SBSN* in AD skins, expression values were measured in AD skins for the *SBSN* transcripts via RT-PCR. The expression levels were normalized by the expression levels of *G6PD*. (a) As shown in Figure 3.9c, the *SBSN* transcript v1 showed significantly decreased expression levels in AD lesional skins compared to AD nonlesional and control skins. (b) The *SBSN* transcript v2 showed significantly different expression changes between AD lesional versus nonlesional skins and nonlesional versus control skins. (c) The *SBSN* transcript v3 showed significantly different expression changes between AD nonlesional versus control skins (***: p-value < 0.001, **: p-value < 0.01, *: p-value < 0.05).

epidermal keratinocytes (NHEKs) were cultured to examine the expression levels of *SBSN* upon IL-4 treatments via RT-PCR. In the presence of IL-4 (at doses of 5 ng/ml and 50 ng/ml), *SBSN* mRNA levels in the differentiated cells were significantly decreased as compared to differentiated cells without cytokine treatment (Figure 3.9b). These decreased expression levels of *SBSN* upon IL-4 treatment suggest a critical precursor role of *SBSN* in the development of AD via disruption of cornification—and further indicate an important role of *SBSN* in epidermal differentiation.

To investigate the role of *SBSN* in AD, expression levels of three *SBSN* transcripts were measured in AD lesional/nonlesional and control skins via RT-PCR. A total of 49 skin biopsies were measured, consisting of 16 AD lesional skin biopsies, 16 AD nonlesional skin biopsies, and 17 healthy controls. The expression levels of *SBSN* transcripts were normalized to *G6PD*. *SBSN* transcript v1 (NM_001166034.1) showed a significantly decreased level in AD lesional skin compared to AD nonlesional skin and controls (Figure 3.9c). The decreased expression levels of the full-length transcript of *SBSN* suggests an important role of this *SBSN* isoform in AD.

The expression of the full-length transcript of *SBSN* (v1) was significantly decreased in AD lesional skin compared to nonlesional skin and healthy controls (Figure 3.9c). The transcript v2 (NM_198538.3) of *SBSN* showed significantly decreased levels in AD lesional skin compared to nonlesional skin, but not controls. However, the transcript v3 (NM_001166035.1) showed no significant expression changes in AD lesional skin compared to nonlesional skin and controls, even though nonlesional skin showed an increased expression compared to controls (Figure 3.10). The v2 and v3 *SBSN* transcript variants had lower expression compared to the full-length transcript (v1) ($\sim$10% and $<$ 1% of v1 in healthy controls). Because the v2 and v3 *SBSN* transcript variants were much less abundant compared to the full-length transcript (v1), the full-length transcript of *SBSN* may be the *SBSN* isoform critical in AD.

3.3.3 Generalization of the paradigm as demonstrated by its application on CIT

To investigate the generalizability of our integrative analysis approach, we applied the paradigm to reveal thermogenesis genes in tissues upon cold exposure. We collected ten gene expression datasets from GEO (Table 3.3). These gene expression data were collected from tissues of mice treated with cold temperature to induce thermogenesis. Both microarray and RNA-Seq data were collected. Because thermogenesis is always activated upon cold exposure, the direction

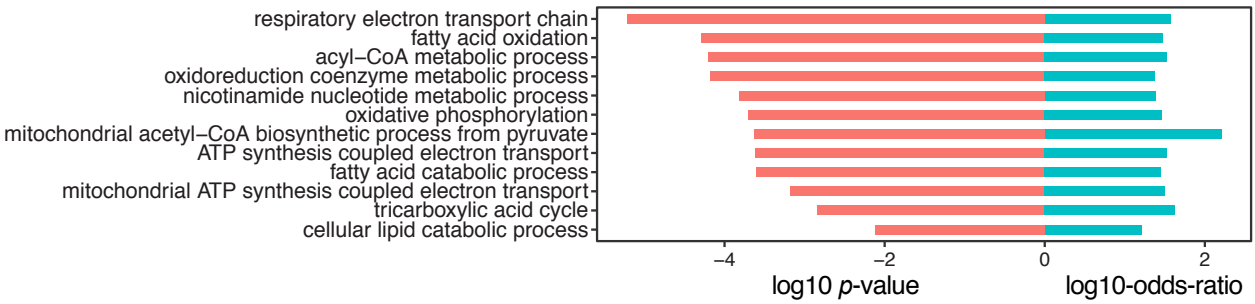


Figure 3. 11 Enriched GO terms of identified genes in CIT.

The identified genes (consensus score ≥ 6) were used to screen GO terms, and the figure depicts the enriched GO terms. The *p*-values and odds ratio from Fisher's exact test were recorded for each GO term.

of thermogenesis is thus increased in all the 24 comparisons within the ten collected datasets. Using DEG analysis, the paradigm calculated the consensus scores for measured genes from 24 comparisons and identified 153 genes with a consensus score ≥ 6. These 153 identified genes were then used to perform GO analysis. Enriched GO terms are related to energy homeostasis (Figure 3.11). Literature curation confirmed the functional evidence in CIT of some identified genes. For example, elongation of very-long-chain fatty acids (*Elovl3*, consensus score = 13) in ablated mice showed a proliferated metabolic rate in a cold environment, indicating a higher capacity for brown fat-mediated nonshivering thermogenesis. Thus, *Elovl3* is a key regulator for CIT in adipose tissue upon cold exposure [73]. As another example, carnitine palmitoyltransferase 2 (*Cpt2*, consensus score = 11) depletion mediates the fatty acid oxidation in adipose tissue, which is required for CITs, suggesting the critical role of *Cpt2* in CIT [67, 74]. This second application of our paradigm in CIT suggests that the paradigm can be generalized to other biological processes. Our paradigm is a simple but important integrative data processing approach for gene expression data.

Table 3. 3 Ten gene expression datasets of adipose tissue upon cold exposure.

GEO Accession	Experimental Tissue	Platform
GSE13432	White adipose tissue	Affymetrix Mouse Genome 430 2.0 Array
GSE40486	Brown adipose tissue; Skeletal muscle	Illumina mouseRef-8 v1.1 expression beadchip
GSE44138	Brown adipose tissue; White adipose tissue; Liver	Illumina Mouse Ref-6 V1
GSE51080	Brown adipose tissue; Mesenteric white adipose tissue; Posterior subcutaneous white adipose tissue	Affymetrix Mouse Genome 430 2.0 Array
GSE63031	Interscapular brown adipose tissue; Inguinal white adipose tissue; Epididymal white adipose tissue	Illumina Genome Analyzer II
GSE64909	Brown adipose tissue	Illumina MouseWG-6 v2.0 R2 expression beadchip
GSE70437	Interscapular brown adipose tissue	Illumina HiSeq 1500
GSE74062	Epididymal white adipose tissue	Affymetrix Mouse Exon 1.0 ST Array
GSE74899	Inguinal white adipose tissue	Affymetrix Mouse Transcriptome Array 1.0
GSE77534	Brown adipose tissue	Illumina HiSeq 2000

4. CITGENEDB: GENES ENHANCING OR SUPPRESSING COLD-INDUCED THERMOGENESIS*

Cold-induced thermogenesis increases energy expenditure and can reduce body weight in mammals, so the genes involved in it are thought to be potential therapeutic targets for treating obesity and diabetes. In the quest for more effective therapies, a great deal of research has been conducted to elucidate the regulatory mechanism of cold-induced thermogenesis. Over the last decade, a large number of genes that can enhance or suppress cold-induced thermogenesis have been discovered, but a comprehensive list of these genes is lacking. To fill this gap, we examined all of the annotated human and mouse genes and curated those demonstrated to enhance or suppress cold-induced thermogenesis by *in vivo* or *ex vivo* experiments in mice. The results of this highly accurate and comprehensive annotation are hosted on a database called CITGeneDB, which includes a searchable web interface to facilitate broad public use [67]. The database will be updated as new genes are found to enhance or suppress cold-induced thermogenesis. It is expected that CITGeneDB will be a valuable resource in future explorations of the molecular mechanism of cold-induced thermogenesis, helping pave the way for new obesity and diabetes treatments.

4.1 Introduction

Cold-induced thermogenesis (CIT) is a process by which mammals increase their resting energy expenditure in cold temperatures. CIT can be activated in two types of adipose tissues: white adipose tissue (WAT), which mainly stores fat, and brown adipose tissue (BAT), which mainly releases stored energy[52]. Direct activation of BAT contributes to heat generation. In

---

* Reprinted with permission from "CITGeneDB: A comprehensive database of human and mouse genes enhancing or suppressing cold-induced thermogenesis validated by perturbation experiments in mice" by Jin Li, Su-Ping Deng, Gang Wei, and Peng Yu, 2018. *Database (Oxford)*. bay012-bay012, Copyright 2018 by authors

addition, induction of brown-adipocyte-like cells (beige or "brite") in WAT depots in the process, called "browning," also promotes heat production[75, 76]. Since the activation of CIT can significantly contribute to increasing resting metabolic rate in humans[77], which in turn reduces body weight, genes affecting CIT can be potential targets for antiobesity therapies.

Many genes that can affect heat production during cold exposure have been discovered. For example, Okada *et al.* found that knockout of *Acot11* increased oxygen consumption rates in both primary brown adipocytes and isolated BAT from the mutant mice, and up-regulated BAT thermogenic genes after exposure to a 4°C environment for 96 hours[78], indicating a suppressive role of *Acot11* in BAT thermogenesis. As another example, knockout of *Zfp423* decreased oxygen consumption of mouse subcutaneous WAT and down-regulated the expression of a number of WAT browning marker genes, such as *Cidea* and *Elovl3*, after exposing $Zfp423^{-/-}$ mice to progressively colder temperatures, suggesting that gene *Zfp423* can promote the browning of WAT under cold-exposure conditions[79]. Besides these single-gene examples, some genes may function synergistically to control CIT. For example, the knockout of both *Nova1* and *Nova2* increases thermogenesis in adipose tissue upon cold exposure, but the single knockout of *Nova1* does not show a significant effect[80]. These results suggest that the discovered genes may significantly contribute to uncovering the regulatory machinery of CIT.

Despite the rapid progress experienced in this field, a complete list of the genes involved in CIT is still missing. For example, the Gene Ontology (GO) Consortium does not have the term "cold-induced thermogenesis," and the most closely related term, "adaptive thermogenesis," (GO:1990845) has only 16 annotated mouse genes, among which the majority (13 of 16) are actually annotated to a child term, "diet induced thermogenesis." Of the four genes (*Ucp1*, *Ucp2*, *Ucp3*, and *Pm20d1*) directly annotated to "adaptive thermogenesis," only two (*Ucp1* and *Pm20d1*)

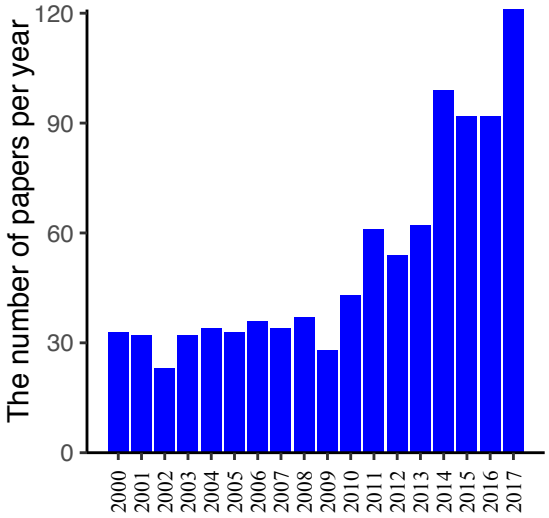


Figure 4. 1 Amount of papers about "cold-induced thermogenesis" published per year since 2000.

To examine the popularity of studies about thermogenesis in recent years, the number of papers was retrieved from the PubMed database by querying "cold" and "thermogenesis" in titles and abstracts on Dec. 11, 2017. The bars in the figure depict the number of papers relevant to CIT from 2000 to 2017. The number of papers published per year had been steadily relatively low before 2011, and the number increased between 2011 and 2017.

are annotated via experimental evidence (Inferred from Mutant Phenotype (IMP)), whereas the other two are annotated via phylogenetic transfer (Inferred from Biological Ancestor (IBA)). Since *Ucp1* is involved in CIT according to the paper cited in GO for *Ucp1*, it should be annotated to "cold-induced thermogenesis" if GO had this term. These pieces of evidence confirm the incompleteness of CIT gene annotation in GO.

This lack of completeness may be explained by the fact that research in the field of CIT has been booming since 2011 (Figure 4.1), two years after BAT was identified in human adults by positron-emission tomography/computed tomography (PET/CT)[81], and it may be difficult for

GO to keep up with the pace of progress in this specific area. Therefore, it is crucial to construct a complete list of genes enhancing/suppressing CIT, for the lack of completeness may hinder the progress of fully elucidating the mechanisms of CIT.

To fill this gap, we curated papers from PubMed and Google in a semiautomatic fashion. The papers show that CIT is affected when given genes are perturbed (by knockout, knockdown, overexpression, etc.). In addition, we structured the genes, the PubMed identifiers (PMIDs) of the corresponding papers, and other important data into a database called CITGeneDB.

In this section, we introduce our effort to curate CIT-related human and mouse genes based on retrieved publications from PubMed and Google. We worked to construct and describe CIT-related data, after which we built the database CITGeneDB to host the important information of the curated CIT-related genes. In addition, we created a web interface for CITGeneDB to facilitate access to the metadata of these genes while also sharing them with various research communities.

4.2 Methods

4.2.1 CIT-related papers retrieval of all the human and mouse genes validated in mice experiments

To construct a complete list of the potential papers for CIT-enhancing/suppressive human and mouse genes, all the human and mouse gene symbols were first retrieved from the HUGO Gene Nomenclature Committee (HGNC) and Mouse Genome Informatics (MGI). Each gene was searched against the PubMed database via PubMed API ESearch in Entrez Programming Utilities (E-utilities) using the query "<gene_symbol>[tiab] AND cold[tiab] AND (thermogenic[tiab] OR thermogenesis[tiab]) NOT Review[pt] NOT Comment[pt] NOT Editorial[pt] NOT News[pt] NOT Published Erratum[pt] AND eng[la]". This search returned over 1,500 gene-paper pairs. Since a single paper may mention multiple genes and a gene may appear in multiple papers, we ended up

with ~200 human and mouse genes in over 1,000 papers. These results were retained for further curation for the CIT-enhancing/suppressive genes.

Some CIT-enhancing/suppressive genes may still be missing in the above search results because the titles, abstracts, and corresponding medical subject heading (MeSH) annotations used by the PubMed search for the papers describing these genes may not contain all of the keywords. To capture these missing genes, a second PubMed search for all the human and mouse genes was performed using the keywords but ignoring "cold," which was used in the first search. This search returned ~5,000 additional gene-paper pairs, with ~3,000 additional papers for ~400 additional human and mouse genes. Since these additional genes may not be related to CIT, we kept only the genes found by querying "<gene_symbol> cold thermogenesis" on Google. For each of these genes, Google usually does well to rank a relevant paper (if there is one) as the first hit because it uses click-through rate[82], a very effective metric for ranking webpages. Here, we checked the top three webpages for each gene kept to further ensure a high recall.

4.2.2 Curation of CIT-enhancing/suppressive human and mouse genes

Our curation criterion for inclusion of a gene was that at least one thermogenesis phenotype, such as body temperature, energy expenditure, or oxygen consumption, must be significantly changed *in vivo* or *ex vivo* by the perturbation of the gene in an animal model or using tissues from an animal model in a cold-exposure condition. All animal models, such as knockout (including conditional knockout), overexpression, and drug/antibody inhibition, were considered, as long as the gene was perturbed. In other words, after the mice with a perturbed gene had been exposed to cold, some thermogenesis phenotypes were measured *in vivo* or were measured in harvesting tissues from the mice *ex vivo*. If any such phenotype was significantly changed, the perturbed gene was included. For example, the body temperature of *Hdac3* conditional knockout

mice was significantly decreased in the cold condition[83], indicating that the gene can enhance heat production upon cold exposure. As another positive example, triglyceride storage of BAT harvested from *Fabp4/5* double knockout mice shrank after a 4-hour exposure to a cold environment (4°C), demonstrating that *Fabp4/5* together can promote CIT. To be stringent, genes that were tested only *in vitro* or without cold exposure were not considered. For instance, Bai *et al*. only studied *Celf1 in vitro* and not in a cold-exposure condition[84]; thus, *Celf1* was not considered according to our curation criterion.

For consistency, all official gene symbols were recorded on CITGeneDB. For mice, MGI (http://www.informatics.jax.org/marker) was used to look up official symbols. Although genes hosted on CITGeneDB were mostly from mice, there are studies with human genes introduced in mice for experimentation. For these human genes, HGNC (https://www.genenames.org) was used to look up the official symbols.

To deal with special cases, we used the following approaches. When a large number of papers (e.g., >100) were returned for a gene by PubMed, the gene symbol usually was a common English word (e.g., JUN or NOV). In this case, it was not efficient to manually examine all the papers returned by PubMed, as mostly these symbols took their common English meaning in the returned papers. To overcome this limitation of PubMed, we instead searched the query term "gene <gene symbol> cold thermogenesis" on Google. Some papers could not be found by the above methods due to a lack of related keywords, but they still described genes enhancing/suppressing thermogenesis. In these cases, we manually added them to CITGeneDB.

4.3 Results

4.3.1 Statistics of enhancing/suppressive human and mouse genes in CITGeneDB

CITGeneDB is a comprehensive resource of CIT-enhancing and -suppressive human and mouse genes. Only genes confirmed in perturbation experiments using mouse models are recorded. Some information about the experiments is included in the database, such as the official symbols of the perturbed genes and perturbation type. In addition, the PMIDs of the corresponding references for each gene are stored in the database.

CITGeneDB currently has 95 CIT-enhancing genes and 47 CIT-suppressive genes. The perturbation type is knockout for most of these genes, with the exception of overexpression using adenovirus (e.g., *HOXC10,* PMID:28186086), point mutation (e.g., *Tshr,* PMID:18559984), deletion mutation (e.g., *Kdm6b*, PMID: 26625958), antibody neutralization (e.g., *Acvr2b*, PMID:22586266), and conditional transgenic overexpression (e.g., *Wnt10b*, PMID:15190075). Most genes can individually function in CIT, but some genes need to work synergistically to affect CIT. Mostly, they have redundant or similar functions, such as *Fabp4/5*, *Nova1/2*, and *Adrb1/2/3*. It should be noted that although our current study focuses only on CIT, the genes that were curated also may be involved in other types of thermogenesis. For example, *Ucp1* is involved in diet-induced thermogenesis. Moreover, CITGeneDB will be periodically updated when more human or mouse genes are found in new literature. Continuously updating this database will likely maintain its impact on obesity and diabetes research.

4.3.2 Web interface for CITGeneDB

To facilitate database access, we developed a web interface that allows users to browse and search. Users can select the number of entries (label 1) to show on a page (Figure 4.2). For each entry, the main information (label 2) includes “Official symbol” (from MGI or HGNC), “PubMed

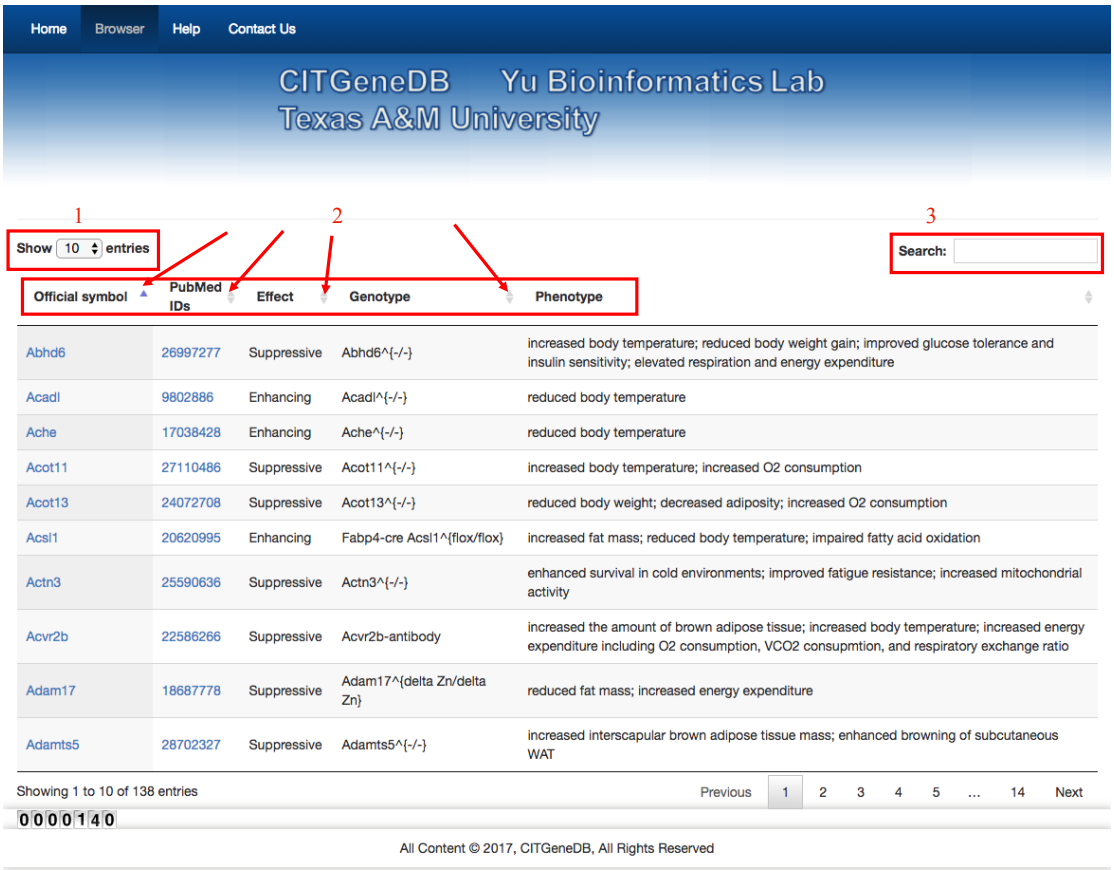


Figure 4. 2 Web interface of CITGeneDB.

To share the CIT-enhancing/suppressive genes, the CITGeneDB web interface was created. In the figure, label 1 is for setting the maximum number of entries on one page. Label 2 represents the main information of CIT genes including official symbols (the official gene symbol from MGI or HGNC), PMIDs of the papers supporting the thermogenesis role of the genes, Effect (whether the gene enhances or suppresses thermogenesis), Genotype (what genes were perturbed and how the genes were perturbed), and Phenotype (affected phenotypes supported by experiments). Label 3 provides the search box for the inquiry about CIT-enhancing/suppressive genes.

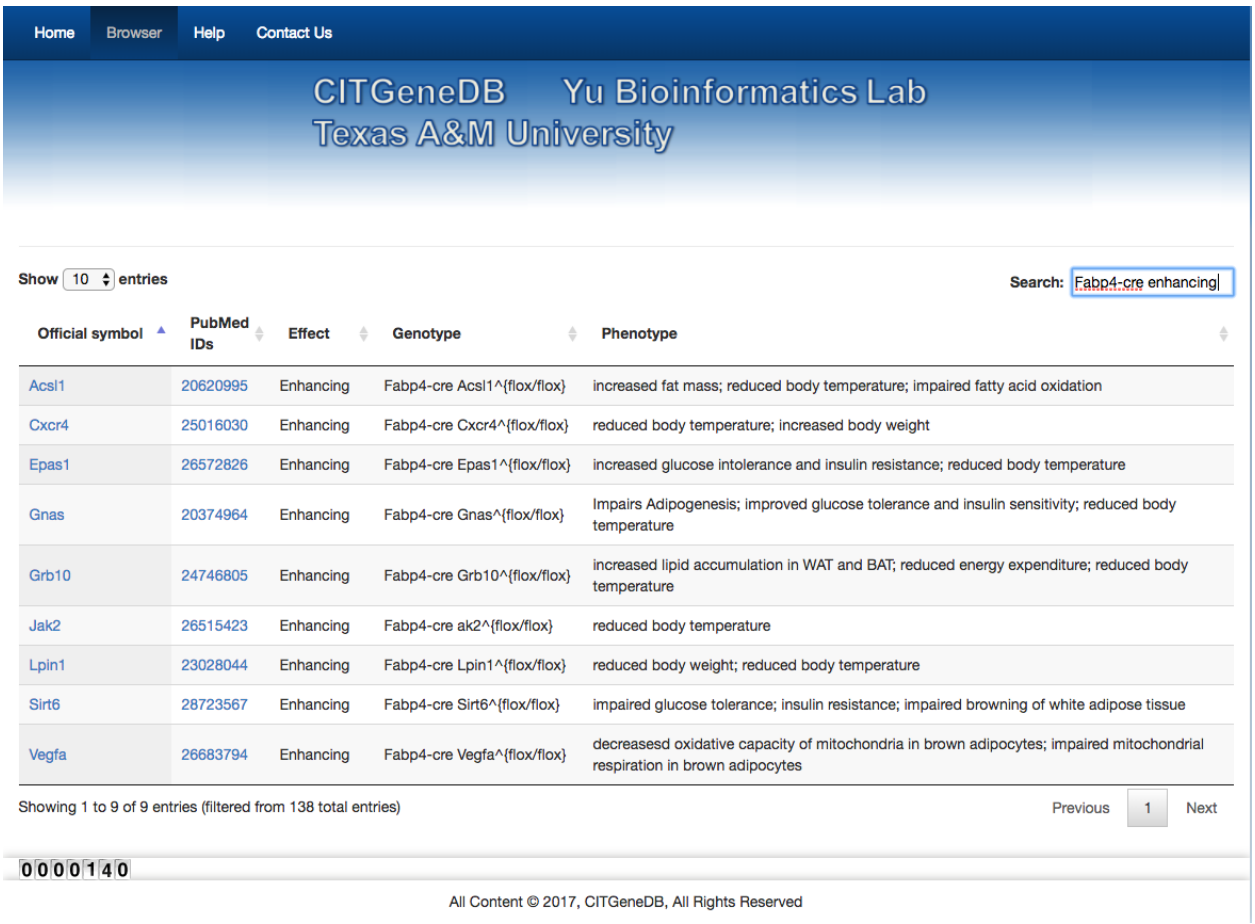


Figure 4. 3 Search result example.

When the keyword "Fabp4-cre enhancing" was searched, nine entries were returned. *Acsl1*, *Cxcr4*, *Epas1*, *Gnas*, *Grb10*, *Jak2*, *Lpin1*, *Sirt6*, and *Vegfa* were all demonstrated to enhance thermogenesis in cold conditions using the fabp4-cre-based conditional knockout mouse models.

IDs" (of the papers supporting the thermogenesis role of the genes), "Effect" (whether the gene enhances or suppresses thermogenesis), "Genotype" (what genes were perturbed and how the genes were perturbed), and Phenotype (affected phenotypes supported by experiments). Each column can be sorted alphabetically by clicking the corresponding information bars, and keywords

can be searched in all the columns of the table via the search box (label 3) to obtain the corresponding entries. For example, the result entry is shown in Figure 4.3 for the search "Fabp4-cre enhancing." Nine entries were returned that have been experimentally tested as enhancing roles in thermogenesis upon cold exposure using *Fabp4*-cre-based conditional knockout mice.

5. SFMETADB: PUBLIC RNA-SEQ DATASETS WITH PERTURBED SPLICING FACTORS*

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 [10]. 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.

5.1 Introduction

Due to the lack of fully structured metadata, the wide use of the valuable RNA-Seq datasets in public repositories such as ArrayExpress and Gene Expression Omnibus (GEO) may be restricted, despite structured metadata having been used elsewhere for raw data usability [85]. For

---

* Reprinted with permission from "SFMetaDB: a comprehensive annotation of mouse RNA splicing factor RNA-Seq datasets" by Jin Li, Ching-San Tseng, Antonio Federico, Franjo Ivankovic, Yi-Shuian Huang, Alfredo Ciccodicola, Maurice S. Swanson, and Peng Yu, 2017. *Database (Oxford)* bax071-bax071, Copyright 2018 by authors

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 [86]. To fill this gap, manual curation has been devoted to developing and maintaining metadata databases [45]. For example, microarray and RNA-Seq datasets have been curated for the downstream analyses in Expression Atlas [87]. We previously launched the RNASeqMetaDB [5] database to facilitate the access of the metadata of public available mouse RNA-Seq datasets. In this section, 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, 19903 human protein-coding genes encode 54896 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 [88]. The second isoform of *CD44*, CD44V8-10, is an important marker for human gastric cancer and increases tumor initiation in gastric cancer cells [89]. 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 [90]. 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 ~10x 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 [91]. For another example, FANCE is a part of the Fanconi anemia (FA) 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 [92]. 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 [93]. 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 [94]. The membrane-bound receptor isoform promotes apoptosis while the soluble isoform inhibits apoptosis.

Alternative splicing is commonly mediated by RNA splicing factors [95]. 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 [96]. Moreover, the splicing factor NOVA2 uniquely mediates the alternative splicing of many axon guidance related genes during cortical development [97]. As another example, the splicing factor PTBP1 suppresses *Pbx1* exon 7 and the neuronal PBX1A isoform in embryonic stem cells (ESCs) during neuronal development [98].

In this section, 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.

5.2 Methods

5.2.1 RNA-Seq dataset curation and SFMetaDB web server deployment

We extracted 353 RNA splicing factors annotated in Gene Ontology (GO) (accession GO:0008380) [31] and Kyoto Encyclopedia of Genes and Genomes (KEGG) (entry mmu03040) [99] for mice. Then, we queried ArrayExpress [86] and GEO [41] 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). Because some splicing factors were studied in more than one dataset, a total of 56 splicing factors were found.

To facilitate the access to these datasets, we launched the database SFMetaDB (http://sfmetadb.yubiolab.org/). 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.

5.2.2 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 [31] and KEGG pathway (entry mmu03040). UniProt annotates the conservative Pfam domain families for the canonical sequences of the splicing factors [100]. From these domain annotations, we calculate the numbers of the splicing factors in Pfam domain families. Figure 5.1 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.

5.3 Results

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 [101], thus we use MBNL1 to demonstrate the usage of SFMetaDB, which confirms the advantage of SFMetaDB

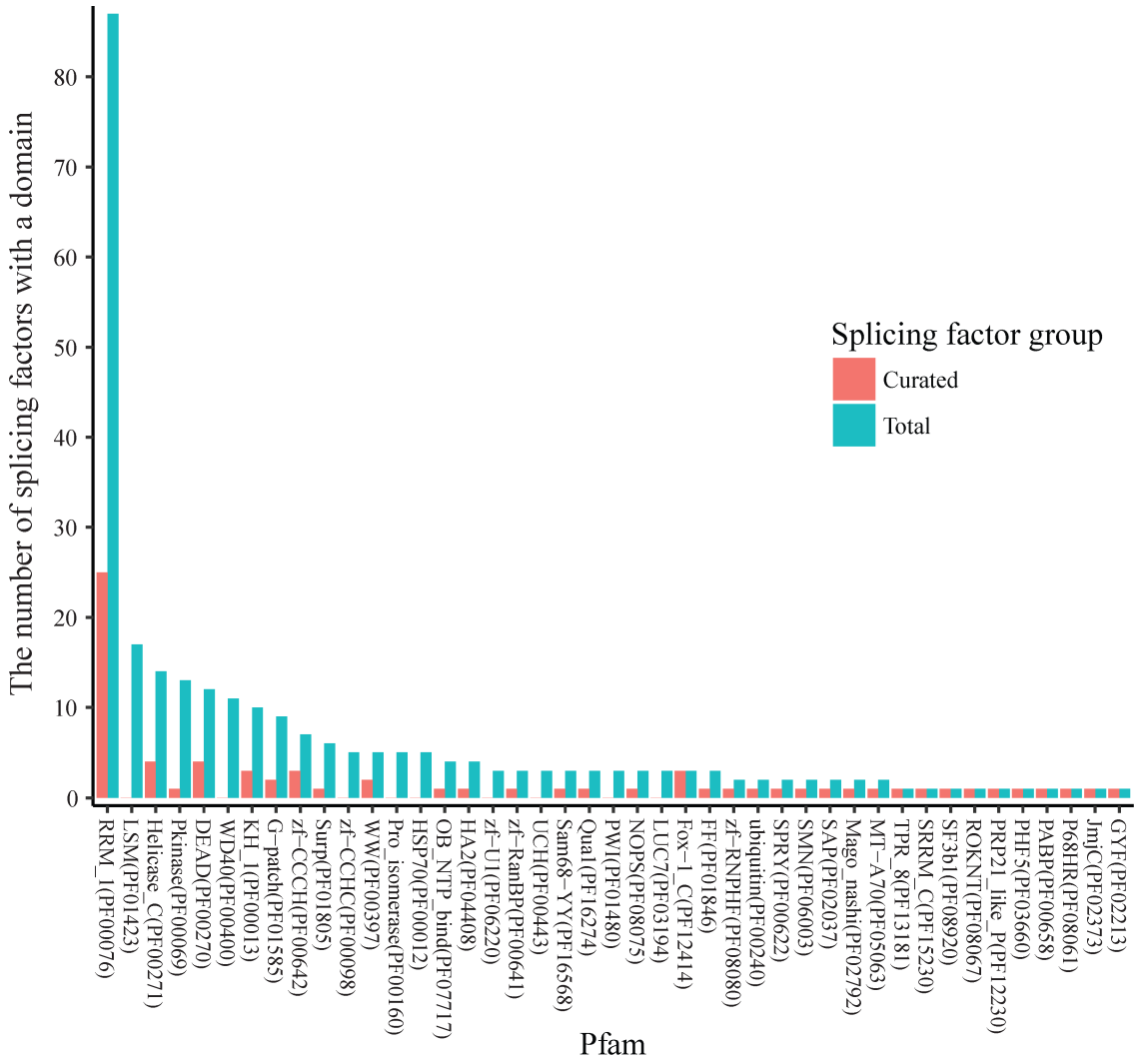


Figure 5. 1 The occurrence of Pfam domain families in splicing factors

over ArrayExpress [10]. A query of MBNL1 on SFMetaDB returns the accurate datasets related with *Mbnl1* loss- or gain-of-function experiments. 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 5.1 Continued. 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 ≥ 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.

However, ArrayExpress returned a total of 13 mouse RNA-Seq datasets with the query Mbnl1, and eight 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. 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 [100]. 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 5.1, 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) have been studied yet [102], 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) [103] that can be used for splicing analysis. In fact, twenty-five unstudied splicing factors 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.

6. RBPMETADB: PUBLIC RNA-SEQ DATASETS WITH PERTURBED RNA-BINDING PROTEINS*

RNA-binding proteins may play a critical role in gene regulation in various diseases or biological processes by controlling post-transcriptional events such as polyadenylation, splicing, and mRNA stabilization via binding activities to RNA molecules. Due to the importance of RNA-binding proteins in gene regulation, a great number of studies have been conducted, resulting in a large amount of RNA-Seq datasets. However, these datasets usually do not have structured organization of metadata, which limits their potentially wide use. To bridge this gap, the metadata of a comprehensive set of publicly available mouse RNA-Seq datasets with perturbed RNA-binding proteins were collected and integrated into a database called RBPMetaDB [6]. This database contains 278 mouse RNA-Seq datasets for a comprehensive list of 163 RNA-binding proteins. These RNA-binding proteins account for only ~10% of all known RNA-binding proteins annotated in Gene Ontology, indicating that most are still unexplored using high-throughput sequencing. This negative information provides a great pool of candidate RNA-binding proteins for biologists to conduct future experimental studies. In addition, we found that DNA-binding activities are significantly enriched among RNA-binding proteins in RBPMetaDB, suggesting that prior studies of these DNA- and RNA-binding factors focus more on DNA-binding activities instead of RNA-binding activities. This result reveals the opportunity to efficiently reuse these data for investigation of the roles of their RNA-binding activities. A web application has also been implemented to enable easy access and wide use of RBPMetaDB. It is expected that RBPMetaDB

* Reprinted with permission from "RBPMetaDB: A comprehensive annotation of mouse RNA-Seq datasets with perturbations of RNA-binding proteins" by Jin Li, Su-Ping Deng, Jacob Vieira, James Thomas, Valerio Costa, Ching-San Tseng, Franjo Ivankovic, Alfredo Ciccodicola, Peng Yu, 2018. *Database (Oxford)* bay054, Copyright 2018 by authors

will be a great resource for improving understanding of the biological roles of RNA-binding proteins.

6.1 Introduction

A lack of fully structured metadata limits the wide use of valuable RNA-Seq datasets in public repositories such as Gene Expression Omnibus (GEO) [3] and ArrayExpress [86]. To fill this gap, manual curation has been shown to be an effective way to collect data resources [67] and has been applied to develop and maintain metadata databases[45]. For example, microarray and RNA-Seq datasets have been curated for the downstream analyses in Expression Atlas [87] and in epidermal development. We previously launched two databases, RNASeqMetaDB [5] and SFMetaDB [10], to facilitate access to the metadata of publicly available mouse RNA-Seq datasets with perturbed disease-related genes and splicing factors, respectively. In this section, we present a new database, RBPMetaDB, for the metadata of RNA-Seq datasets with perturbed RNA-binding proteins (RBPs).

RBPs play a critical role in multiple cellular processes in eukaryotes. RBPs bind to double- or single-stranded RNA molecules and are potential key factors in biological processes, such as pre-mRNA splicing, RNA methylation, and protein translation [104]. Besides influencing each of these processes, RBPs also provide a link between them [105]. The perturbation of these intricate networks can destroy the coordination of complex post-transcriptional events and lead to disease.

According to recent genomic data and evidence derived from animal models, RBPs play a crucial role in the pathogenesis of many complex human diseases, including neurological disorders [106], Mendelian diseases[107], and cancer [108]. These diseases have been demonstrated to have strong associations with aberrant functions or expression of RBPs, which can impact many different genes and pathways. Some diseases can be caused by loss of function of RBPs, such as

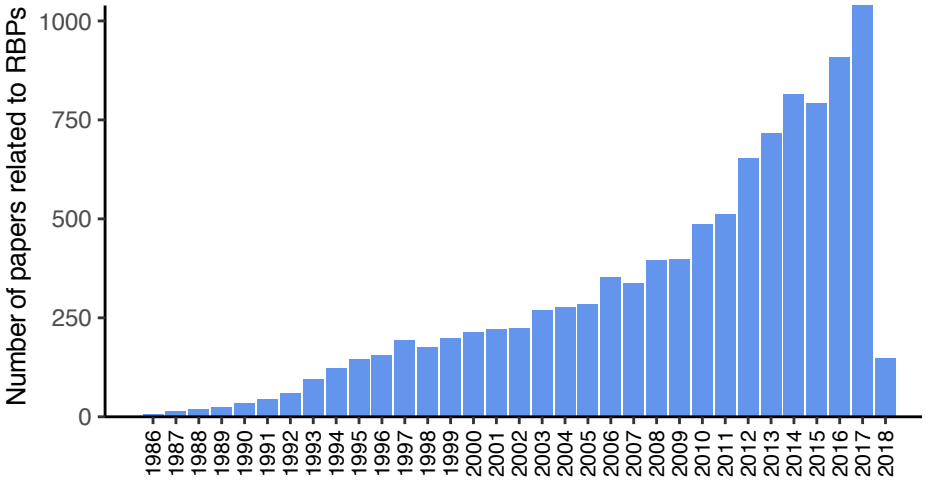


Figure 6. 1 The rapid growth of papers related to RPBs in PubMed.

Approximately 10,000 papers related to RPBs are indexed on PubMed according to the query of "RNA binding protein"[tiab] OR "RNA binding proteins"[tiab] at the time of writing. Since 2012, the number of papers published per year has been increasing more rapidly than ever before. In 2017 alone, over 1,000 papers were published.

Fragile X syndrome, paraneoplastic neurologic syndromes, and spinal muscular atrophy [104]. For example, Fragile X syndrome can be caused by the deficiency of gene fragile X mental retardation (*FMR1*)[109]. Alternatively, some diseases can be caused by gain of function of RBPs, including myotonic dystrophy, Fragile X tremor ataxia syndrome, and oculopharyngeal muscular dystrophy (OPMD) [104]. For instance, OPMD is generated by the accumulation of aggregates in the nuclei of skeletal muscle fibers caused by mutants in the protein PABPN1 [110]. And a deficiency of PABPN1 can induce progressive muscle weakness in muscular dystrophy[111].

To investigate the functions of RBPs in biological processes or diseases such as the ones mentioned above, a large number of studies have been conducted, resulting in exponential growth of RBP-related papers in recent years (Figure 6.1). For example, more than 1,000 papers were

published in 2017 alone. Among the studies on RBPs, a large number of RNA-Seq datasets have been generated in loss- or gain-of-function experiments and are publicly available from online repositories like GEO [3]. However, because GEO does not have a stringent requirement for metadata of the submitted datasets, the metadata are non-uniformly maintained across different datasets, resulting in inconsistent dataset annotation and sometimes ambiguity. Such a deficiency makes it difficult to identify useful datasets with high precision and recall, which limits the wide use of the datasets.

To address this challenge, we worked to curate RNA-Seq datasets from GEO and ArrayExpress with one or more RBPs being perturbed, e.g., by knock-out, knock-down, or overexpression. Important dataset annotations such as genotypes and PubMed references were manually curated to ensure high accuracy. Curated datasets can be used in gene expression analysis [55, 112] and alternative splicing analysis [27, 113] for biological hypothesis generation [13] via a signature comparison approach [11]. To facilitate the use of our curated datasets, the metadata information of these datasets was imported into a database called RBPMetaDB. It should be mentioned that our database differs greatly from Expression Atlas in the sense that the latter is not mainly about datasets where specific genes are perturbed and so are not guaranteed to be complete in this aspect.

In this section, we describe our main curation methods used in constructing RBPMetaDB and the statistics of the database. To demonstrate the use of RBPMetaDB, a number of promising candidate RBPs have been identified by comparing RBPs with RNA-Seq datasets and all the RBPs annotated in Gene Ontology (GO). In addition, a web application has been developed to host the database to broaden the use of curated metadata and the original raw datasets among biomedical communities.

6.2 Methods

6.2.1 Metadata curation of GEO/ArrayExpress RNA-Seq datasets and RBPMetaDB web application deployment

To collect RNA-Seq datasets for RBPs from GEO comprehensively, we first extracted 1,587 mouse RBPs annotated in GO (accession GO:0003723) [31]. Each of these RBPs was queried against GEO for mouse RNA-Seq data using the query (<official_symbol>[Title] OR <official_symbol>[Description]) NOT SuperSeries[Description] AND gse[Entry Type] AND "Mus musculus"[porgn:__txid10090] AND ("expression profiling by high throughput sequencing"[DataSet Type] OR "non coding rna profiling by high throughput sequencing"[DataSet Type])") and against ArrayExpress using the query (<official_symbol> AND organism:"Mus musculus" AND exptype:"sequencing assay" AND exptype:"rna assay"). These queries resulted in 1,194 unique datasets in mice. Due to the limitations of the search functions of GEO and ArrayExpress, many of these datasets do not have perturbed RBPs despite the official symbols of some RBPs being mentioned in the titles or descriptions of the datasets. To retain the datasets with perturbations of RBPs, we manually curated each dataset [46] and retained datasets with biological replications per comparison condition, with at least one RBP being knocked-out, knocked-down, or overexpressed (along with the corresponding wild-type or control samples) in mice. For the datasets that do not have associated PubMed IDs on GEO and ArrayExpress, we manually added the PubMed IDs.

To facilitate access to these datasets, we launched a database called RBPMetaDB (http://rbpmetadb.yubiolab.org). RBPMetaDB is implemented using Flask (http://flask.pocoo.org), a microframework for web development in Python. The MySQL database is used for data storage. The website of RBPMetaDB is freely available, and it presents the

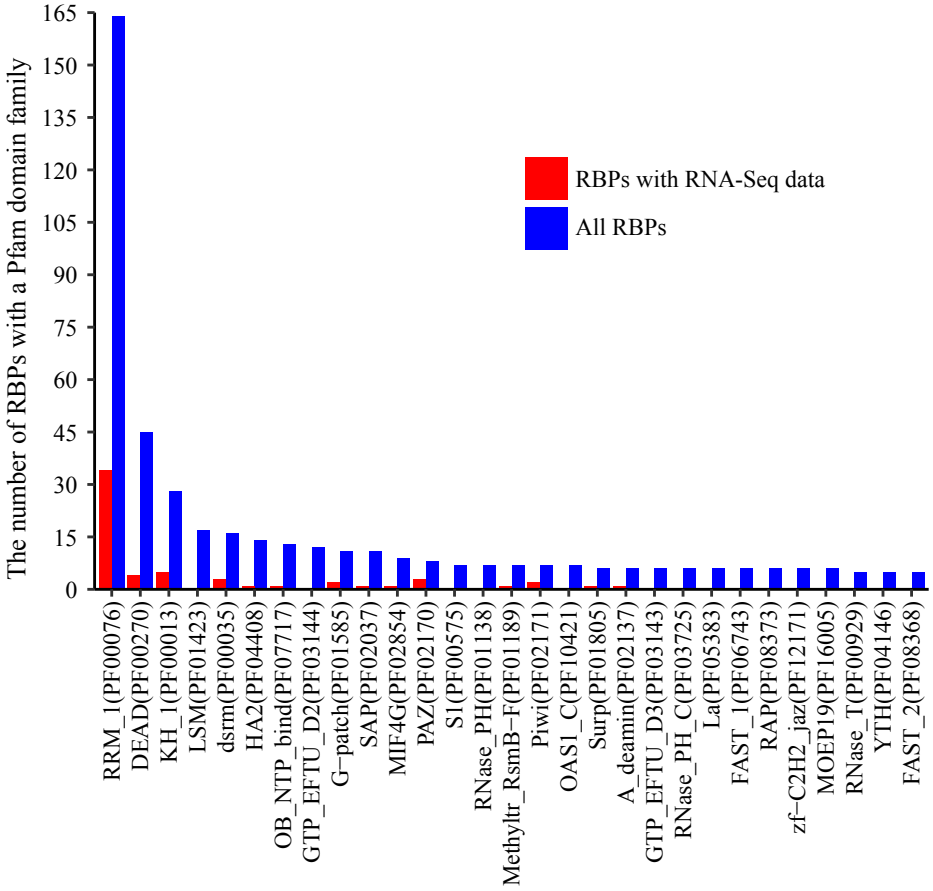


Figure 6. 2 The number of RBPs containing a domain from a Pfam family with RNA-binding activity.

Blue bars indicate the number of RBPs containing a domain from a family among all RBPs, and red bars indicate the numbers of RBPs containing a domain from a family among the RBPs with associated RNA-Seq datasets. Only families with a blue bar with ≥ 5 RBPs are shown.

GEO/ArrayExpress accession numbers, descriptions, number of samples, associated curated RBPs, perturbation, and PubMed references for each RNA-Seq dataset.

6.2.2 Domain structure analysis of RBPs

Protein domain structure analysis of RBPs was performed to identify critical RBPs for future studies. First, all RBPs annotated to the "RNA binding" GO term (GO: 0003723) were retrieved using the R package GO.db [114]. Using the UniProt annotation of the Pfam families assigned to the RBP protein domains [100], the number of RBPs with specific Pfam families was calculated using RBPs with the curated RNA-Seq datasets and using the total RBPs, respectively. To investigate the RNA binding effect, the number of RBPs of Pfam families with RNA binding activity was calculated, where RNA-related Pfam families were searched using the RESTful interface in the Pfam database. Figure 6.2 plots the number of RBPs with Pfam families specific to RNA binding for the RBPs with RNA-Seq data and all the RBPs. By comparing the domain families of the RBPs with RNA-Seq datasets to those of all the RBPs, the RBPs in relatively less-studied domain families can be promising candidates for future RBP studies.

6.3 Results

6.3.1 Data statistics

RBPMetaDB has 292 RNA-Seq datasets with 187 perturbed RBPs, which account for only ~10% of all annotated RBPs. Among these 187 RBPs, over 30% of them have more than one corresponding RNA-Seq dataset. Approximately 90% of datasets in RBPMetaDB have only one perturbed RBP, meaning that most studies are small-scale and well-focused. Also, RBPs with RNA-Seq data tend to have DNA-binding activity. To systematically examine the DNA-binding activity of RBPs, the GO term "DNA binding" (GO:0003677) was used to extract the genes with DNA-binding activity. By overlapping with DNA-binding proteins, 66 RBPs with RNA-Seq datasets and 207 RBPs without RNA-Seq datasets were shown to have DNA-binding activity. Taking the total 1,587 RBPs as background, Fisher's exact test showed an enrichment of DNA-

binding activity in RBPs with datasets compared to RBPs without datasets (p-value$< 5.8\times10^{-14}$). Specifically, for RBPs in RBPMetaDB, the proportion between RBPs with and without DNA-binding activity is 0.68 (66 over 97). On the contrary, the proportion of RBPs that do not have RNA-Seq datasets is only 0.17 (207 over 1,219). This large difference suggests that many datasets in RBPMetaDB were collected for their DNA-binding activity instead of RNA-binding activity,

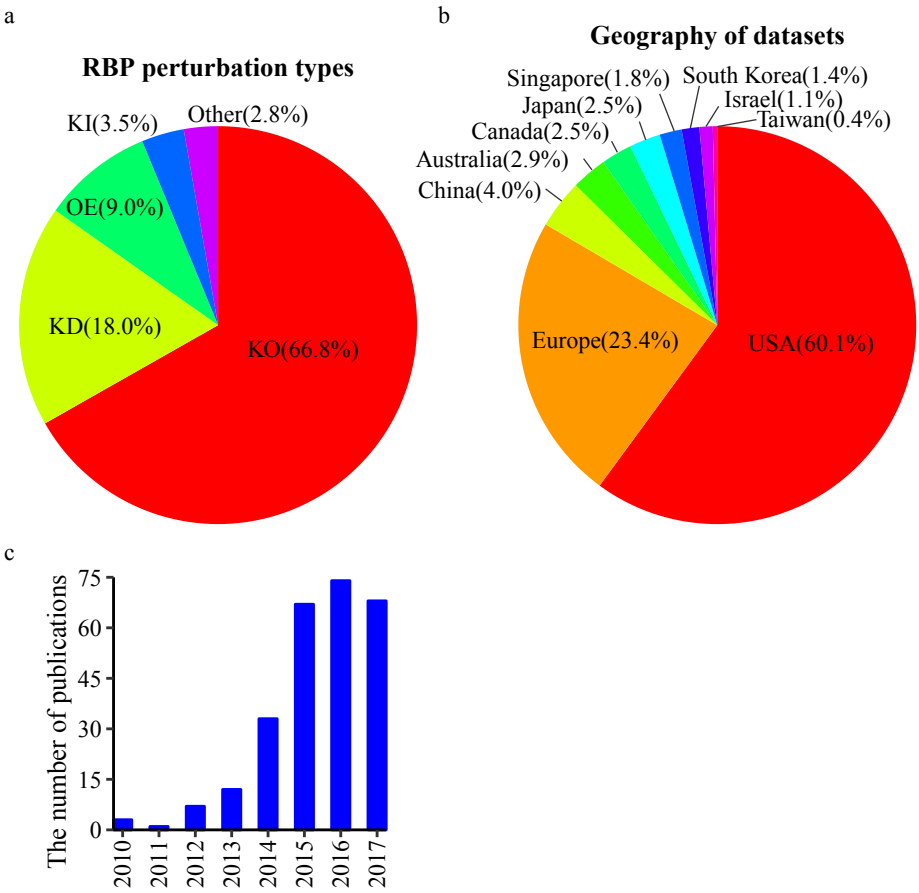


Figure 6. 3 Statistics of curated RNA-Seq datasets for RBPs.

(a) The distribution of perturbation types: knock-out (KO), knock-down (KD), overexpression (OE), knock-in (KI), and other (e.g., point mutations of RBPs or treatment with inhibitors of RBPs) among all the curated datasets. The percentages are shown between parentheses. Knock-out experiments are the most common. (b) The curated datasets are generated from research labs worldwide. The US is the dominant country with a contribution of 60.1% of all the datasets. (c) The number of associated publications for the datasets increased from 2010 to 2017. The slow-down of increase in 2016 and the drop in 2017 are likely due to the missing PMIDs annotation for a subset of the recently released datasets on GEO.

and these datasets are likely to be underanalyzed for RNA-binding activity, providing a cost-

effective opportunity to reanalyze these datasets to study their related RNA biology. For example, *Ezh2* is the most-studied gene, with 35 RNA-Seq datasets in RBPMetaDB. However, most studies of EZH2, as a catalytic subunit of Polycomb Repressive Complex 2 (PRC2), focus on its capacity for mono-, di-, and trimethylation of histone H3 on lysine K27 (H3K27me1/2/3) [115].

Figure 6.3a shows that the main RBP perturbation type of all the datasets in RBPMetaDB, is knock-out (~67%). The rest is knock-down (~18%), overexpression (~9%), knock-in (~3.5%), and other (~2.8%, e.g., treated with inhibitors or point mutation). Figure 6.3b shows that the US and Europe dominate the generation of RNA-Seq datasets for studying RBPs, with contributions of 60.1% and 23.4% of all the datasets, respectively. In addition, Figure 6.3c shows an increasing number of papers published about the RNA-Seq datasets in RBPMetaDB from 2010 to 2017. This increasing research interest worldwide will stimulate more investigation on RBPs.

6.3.2 Comparison of RBPs using protein domain analysis

Protein domains, as conserved protein structural units, typically characterize certain functional aspects of a protein, and proteins sharing similar domains tend to share similar functions. Since RBPs bind to RNAs, they should have RNA-binding domain. We therefore extracted the domain family information of all the RBPs according to Pfam domain family annotation[102]. Figure 6.2 shows the protein domain families ordered by the number of RBPs with a domain from a Pfam family, and only families with RNA-binding activities with ≥ 5 annotated RBPs are shown. The most dominant domain family is RRM_1 (RNA recognition motif, PF00076) and the RBPs with domains from this family are relatively well-studied (34 in RBPMetaDB over all 164 annotated). RBPs with domains from four additional families are fairly well-studied, including DEAD (DEAD/DEAH box helicase, PF00270), KH_1 (KH domain, PF00013), dsrm (double-stranded RNA-binding motif, PF00035), and HA2 (helicase-associated

domain, PF04408). However, none of the RBPs with domains from two highly dominant domain families, LSM (PF01423) and GTP_EFTU_D2 (PF03144), has related RNA-Seq datasets yet, and they may be good candidates for future high-throughput sequencing studies.

What's more, among the RBPs without related RNA-Seq datasets, 140 RBPs already have one or more mouse models on the International Mouse Strain Resource (IMSR) [103]. For example, the gene Cleavage Stimulation Factor Subunit 2 Tau Variant (*Cstf2t*) has been demonstrated to be an important stage-specific regulator of *Crem* mRNA processing that controls *Crem* polyadenylation in mouse testis. *Cstf2t* can lead to an overall decrease of the Crem mRNAs generated from internal promoters in $Cstf2t^{-/-}$ mice [116, 117]. Therefore, these 140 RBPs can be promising candidates for RNA-Seq studies in the future.

6.3.3 Web interface

To facilitate the use of RBPMetaDB, a user-friendly website has been launched. The website allows users to access all the key information related to the curated RNA-Seq datasets, including the GEO/ArrayExpress accession numbers, dataset titles, numbers of samples, associated RBPs, perturbation types, and PubMed IDs (Figure 6.4). The contents in these fields are linked to the corresponding entries in GEO/ArrayExpress, metadata information for each dataset, MGI gene symbol, and PubMed, respectively. In the table view of the website, the first 10 entries are shown by default, but the user can easily select the number of entries to be visualized from a pop-up menu on the left side (Label A). Each table has six columns about the metadata in RBPMetaDB (Label B), and all columns can be sorted in ascending or descending order by clicking column headers. The search boxes at the bottom of all the fields support field-specific search by regular expression (Label C). For example, to search for multiple gene symbols in the "RNA binding proteins" column, one can specify the gene symbols joined by "|". By searching a

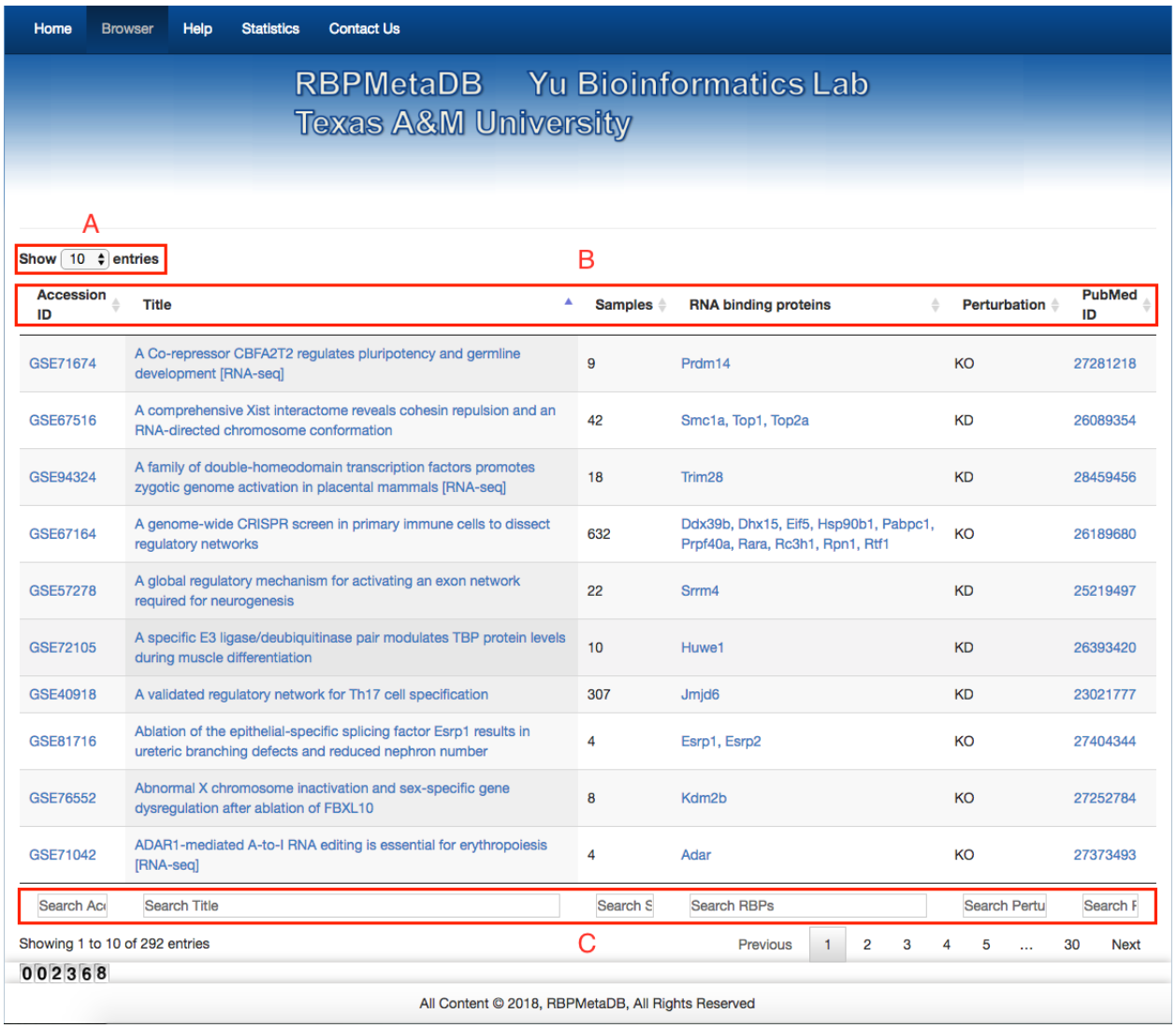
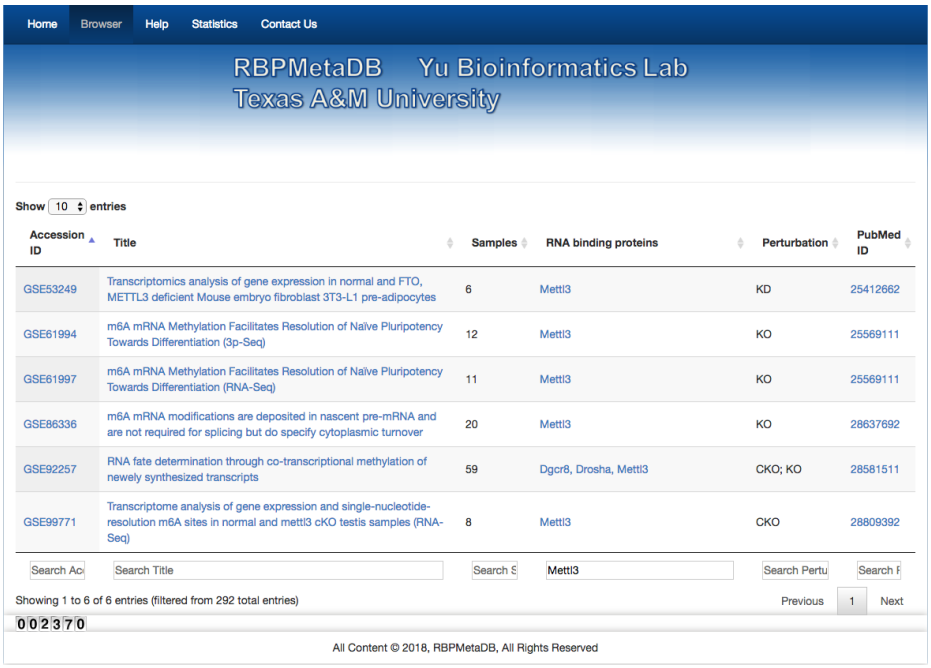


Figure 6. 4 Web interface of RBPMetaDB.

The RBPMetaDB website presents information about the mouse RNA-Seq datasets with perturbed RBPs. Label A refers to the maximum number of entries shown on a page. Label B is about the relevant information for each RNA-Seq dataset including GEO accession numbers, titles of the datasets in GEO, number of samples, official gene symbols from Mouse Genome Informatics (MGI), perturbation types of the RBPs associated with a dataset, and PMIDs of the related papers. Label C refers to the field specific search boxes.

gene of interest, users can find all RNA-Seq datasets with the gene perturbed. Take as an example

a



b

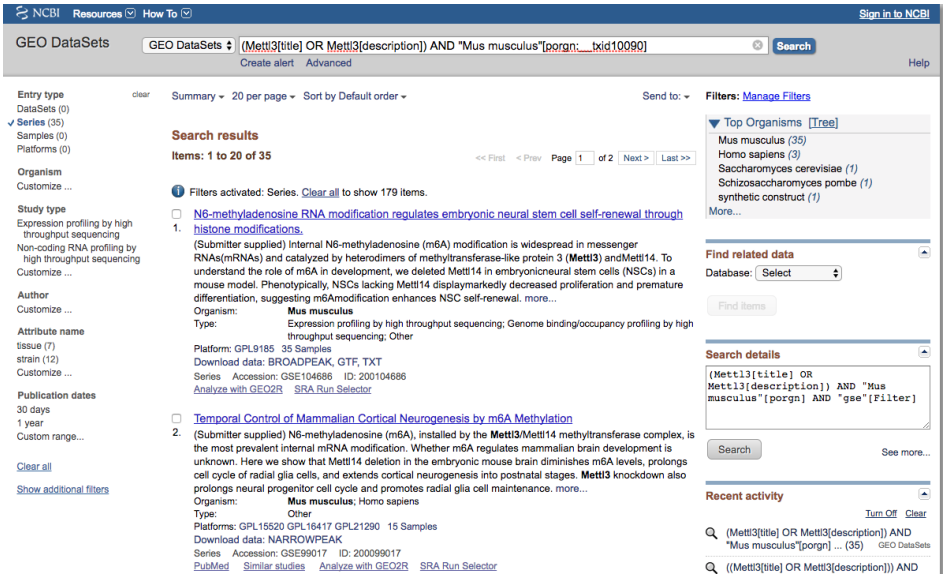


Figure 6. 5 A use case of RBPMetaDB for the mouse RPB METTL3.

(a) Here is a use case of RPB METTL3 to demonstrate the advantage of RBPMetaDB over GEO. By using the keyword "Mettl3," RBPMetaDB accurately returns six mouse RNA-Seq datasets with *Mettl3* perturbed. (b) However, GEO returns 35 mouse RNA-Seq datasets without identifying which datasets are from experiments with *Mettl3* perturbed.

METTL3, which is an important enzyme involved in the post-transcriptional methylation of internal adenosine residues in eukaryotic mRNA [118]— it can be demonstrated that RPBMetaDB greatly outperforms GEO in terms of search efficiency. When the keyword “Mettl3” is searched on RBPMetaDB, it returns six highly accurate mouse RNA-Seq datasets from *Mettl3* loss- or gain-of-function studies (Figure 6.5a). GEO returns 35 mouse RNA-Seq datasets with the query of “Mettl3” in dataset titles and descriptions (Figure 6.5b), but it is impossible to directly identify which RNA-Seq datasets are from loss- or gain-of-function experiments of *Mettl3*. On the contrary, RBPMetaDB does not return irrelevant datasets of a given RBP, and it returns more accurate results than GEO.

7. SUMMARY

This dissertation has described integrated analysis methods using public data. Using public RNA-Seq data, differential gene expression analysis and differential alternative splicing analysis have identified conservative expression and splicing changes in mouse psoriasis. The findings provide a better understanding of gene expression and alternative splicing mechanism underlying human skin disease psoriasis. A new data mining paradigm of pairing data collection and data analysis has revealed key genes in epidermal development and cold-induced thermogenesis. The experimental validations have demonstrated the power of the proposed paradigm in the epidermal development and cold-induced thermogenesis. The combining of systematic data collection and data analysis was shown to be effective approach in data analysis. As another contribution in this dissertation, several data resource have been constructed. Public RNA-Seq data for splicing factors and RNA-binding proteins have been systematically annotated and collected. The metadata databases of the public data provided a precious data resource for studying splicing factors and RNA-binding proteins. To collect a comprehensive list of genes that regulate cold-induced thermogenesis, many biologists have been collaborated to annotate genes that induce cold-induced thermogenesis phenotype changes supported by perturbed experiments. The comprehensive list of regulatory genes provides promising candidates to study the underlying mechanisms of cold-induced thermogenesis. In summary, this dissertation has developed several integrated analysis methods using public data, leading to better understanding of the mechanisms underlying psoriasis, epidermal development and cold-induced thermogenesis. This dissertation has also built data resources for splicing factors, RNA-binding proteins, and cold-induced thermogenesis.

REFERENCES

[1] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nat Rev Genet,* vol. 10, pp. 669-80, Oct 2009.

[2] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi*, et al.*, "HITS-CLIP yields genome-wide insights into brain alternative RNA processing," *Nature,* vol. 456, pp. 464-9, Nov 27 2008.

[3] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky*, et al.*, "NCBI GEO: archive for functional genomics data sets-update," *Nucleic Acids Research,* vol. 41, pp. D991-D995, Jan 2013.

[4] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet,* vol. 10, pp. 57-63, Jan 2009.

[5] Z. Guo, B. Tzvetkova, J. M. Bassik, T. Bodziak, B. M. Wojnar, W. Qiao*, et al.*, "RNASeqMetaDB: a database and web server for navigating metadata of publicly available mouse RNA-Seq datasets," *Bioinformatics,* vol. 31, pp. 4038-40, Dec 15 2015.

[6] J. Li, S. P. Deng, J. Vieira, J. Thomas, V. Costa, C. S. Tseng*, et al.*, "RBPMetaDB: a comprehensive annotation of mouse RNA-Seq datasets with perturbations of RNA-binding proteins," *Database (Oxford),* vol. 2018, Jan 1 2018.

[7] R. Agarwal and V. Dhar, "Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," *Information Systems Research,* vol. 25, pp. 443-448, Sep 2014.

[8] J. Li, L. Zheng, A. Uchiyama, L. Bin, T. M. Mauro, P. M. Elias*, et al.*, "A data mining paradigm for identifying key factors in biological processes using gene expression data," *Sci Rep,* vol. 8, p. 9083, Jun 13 2018.

[9] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat Genet,* vol. 40, pp. 1413-5, Dec 2008.

[10] J. Li, C.-S. Tseng, A. Federico, F. Ivankovic, Y.-S. Huang, A. Ciccodicola*, et al.*, "SFMetaDB: a comprehensive annotation of mouse RNA splicing factor RNA-Seq datasets," *Database,* vol. 2017, pp. bax071-bax071, 2017.

[11] Z. Li, J. Li, and P. Yu, "l1kdeconv: an R package for peak calling analysis with LINCS L1000 data," *BMC Bioinformatics,* vol. 18, p. 356, Jul 27 2017.

[12] M. Jangi and P. A. Sharp, "Building robust transcriptomes with master splicing factors," *Cell,* vol. 159, pp. 487-98, Oct 23 2014.

[13] J. Li and P. Yu, "Genome-wide transcriptome analysis identifies alternative splicing regulatory network and key splicing factors in mouse and human psoriasis," *Sci Rep,* vol. 8, p. 4124, Mar 7 2018.

[14] L. C. Tsoi, S. L. Spain, J. Knight, E. Ellinghaus, P. E. Stuart, F. Capon*, et al.*, "Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity," *Nat Genet,* vol. 44, pp. 1341-8, Dec 2012.

[15] D. Ellinghaus, L. Jostins, S. L. Spain, A. Cortes, J. Bethune, B. Han*, et al.*, "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci," *Nat Genet,* vol. 48, pp. 510-8, May 2016.

[16] F. O. Nestle, D. H. Kaplan, and J. Barker, "Psoriasis," *N Engl J Med,* vol. 361, pp. 496-509, Jul 30 2009.

[17] J. Kim, C. H. Oh, J. Jeon, Y. Baek, J. Ahn, D. J. Kim*, et al.*, "Molecular Phenotyping Small (Asian) versus Large (Western) Plaque Psoriasis Shows Common Activation of IL-

17 Pathway Genes but Different Regulatory Gene Sets," *J Invest Dermatol,* vol. 136, pp. 161-72, Jan 2016.

[18] B. Li, L. C. Tsoi, W. R. Swindell, J. E. Gudjonsson, T. Tejasvi, A. Johnston*, et al.*, "Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms," *J Invest Dermatol,* vol. 134, pp. 1828-38, Jul 2014.

[19] S. Koks, M. Keermann, E. Reimann, E. Prans, K. Abram, H. Silm*, et al.*, "Psoriasis-Specific RNA Isoforms Identified by RNA-Seq Analysis of 173,446 Transcripts," *Front Med (Lausanne),* vol. 3, p. 46, 2016.

[20] F. Xue, X. Li, X. Zhao, L. Wang, M. Liu, R. Shi*, et al.*, "SRSF1 facilitates cytosolic DNA-induced production of type I interferons recognized by RIG-I," *PLoS One,* vol. 10, p. e0115354, 2015.

[21] L. Chen, J. M. Tovar-Corona, and A. O. Urrutia, "Alternative splicing: a potential source of functional innovation in the eukaryotic genome," *Int J Evol Biol,* vol. 2012, p. 596274, 2012.

[22] J. M. Mudge, A. Frankish, J. Fernandez-Banet, T. Alioto, T. Derrien, C. Howald*, et al.*, "The origins, evolution, and functional potential of alternative splicing in vertebrates," *Mol Biol Evol,* vol. 28, pp. 2949-59, Oct 2011.

[23] S. K. Ippagunta, R. Gangwar, D. Finkelstein, P. Vogel, S. Pelletier, S. Gingras*, et al.*, "Keratinocytes contribute intrinsically to psoriasis upon loss of Tnip1 function," *Proc Natl Acad Sci U S A,* vol. 113, pp. E6162-E6171, Oct 11 2016.

[24] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha*, et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics,* vol. 29, pp. 15-21, Jan 1 2013.

[25] S. Anders, P. T. Pyl, and W. Huber, "HTSeq--a Python framework to work with high-throughput sequencing data," *Bioinformatics,* vol. 31, pp. 166-9, Jan 15 2015.

[26] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler, "The UCSC known genes," *Bioinformatics,* vol. 22, pp. 1036-1046, 2006.

[27] P. Yu and C. A. Shaw, "An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function," *Bioinformatics,* vol. 30, pp. 1547-54, Jun 1 2014.

[28] G. Casella and R. L. Berger, *Statistical inference*, 2 ed.: Thomson Learning, 2001.

[29] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 57, pp. 289-300, 1995.

[30] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nat Methods,* vol. 7, pp. 1009-15, Dec 2010.

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry*, et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet,* vol. 25, pp. 25-9, May 2000.

[32] N. R. Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* vol. 44, pp. D7-19, Jan 4 2016.

[33] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res,* vol. 30, pp. 3059-66, Jul 15 2002.

[34] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam*, et al.*, "ArrayExpress update--trends in database growth and links to data analysis tools," *Nucleic Acids Res,* vol. 41, pp. D987-90, Jan 2013.

[35] S. Anders and W. Huber, "Differential expression of RNA-Seq data at the gene level–the DESeq package," *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL),* 2012.

[36] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel*, et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science,* vol. 313, pp. 1929-35, Sep 29 2006.

[37] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero*, et al.*, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci Transl Med,* vol. 3, p. 96ra77, Aug 17 2011.

[38] Y. A. Lee, F. Ruschendorf, C. Windemuth, M. Schmitt-Egenolf, A. Stadelmann, G. Nurnberg*, et al.*, "Genomewide scan in german families reveals evidence for a novel psoriasis-susceptibility locus on chromosome 19p13," *Am J Hum Genet,* vol. 67, pp. 1020-4, Oct 2000.

[39] V. B. Morhenn, T. E. Nelson, and D. L. Gruol, "The rate of wound healing is increased in psoriasis," *Journal of Dermatological Science,* vol. 72, pp. 87-92, Nov 2013.

[40] J. H. Choi, D. K. Choi, K. C. Sohn, S. S. Kwak, J. Suk, J. S. Lim*, et al.*, "Absence of a human DnaJ protein hTid-1S correlates with aberrant actin cytoskeleton organization in lesional psoriatic skin," *J Biol Chem,* vol. 287, pp. 25954-63, Jul 27 2012.

[41] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res,* vol. 30, pp. 207-10, Jan 01 2002.

[42] Z. Ge, B. L. Quek, K. L. Beemon, and J. R. Hogg, "Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway," *Elife,* vol. 5, Jan 08 2016.

[43] S. Gueroussov, T. Gonatopoulos-Pournatzis, M. Irimia, B. Raj, Z. Y. Lin, A. C. Gingras*, et al.*, "An alternative splicing event amplifies evolutionary differences between vertebrates," *Science,* vol. 349, pp. 868-73, Aug 21 2015.

[44] L. Fish, N. Pencheva, H. Goodarzi, H. Tran, M. Yoshida, and S. F. Tavazoie, "Muscleblind-like 1 suppresses breast cancer metastatic colonization and stabilizes metastasis suppressor transcripts," *Genes Dev,* vol. 30, pp. 386-98, Feb 15 2016.

[45] B. Qin, M. Zhou, Y. Ge, L. Taing, T. Liu, Q. Wang*, et al.*, "CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human," *Bioinformatics,* vol. 28, pp. 1411-2, May 15 2012.

[46] Z. Li, J. Li, and P. Yu, "GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata," *Database,* vol. 2018, pp. bay019-bay019, 2018.

[47] A. Bhaduri, A. Ungewickell, L. D. Boxer, V. Lopez-Pajares, B. J. Zarnegar, and P. A. Khavari, "Network Analysis Identifies Mitochondrial Regulation of Epidermal Differentiation by MPZL3 and FDXR," *Dev Cell,* vol. 35, pp. 444-57, Nov 23 2015.

[48] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nature genetics,* vol. 32, pp. 490-495, 2002.

[49] F. K. Kavvoura and J. P. Ioannidis, "Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls," *Human genetics,* vol. 123, pp. 1-14, 2008.

[50] P. M. Elias, "The skin barrier as an innate immune element," *Seminars in Immunopathology,* vol. 29, pp. 3-14, Apr 2007.

[51] P. D. Thomas, "The Gene Ontology and the Meaning of Biological Function," *Methods Mol Biol,* vol. 1446, pp. 15-24, 2017.

[52] J. M. Ellis, L. O. Li, P. C. Wu, T. R. Koves, O. Ilkayeva, R. D. Stevens*, et al.*, "Adipose acyl-CoA synthetase-1 directs fatty acids toward beta-oxidation and is required for cold thermogenesis," *Cell Metab,* vol. 12, pp. 53-64, Jul 07 2010.

[53] X. Qian, X. Li, T. O. Ilori, J. D. Klein, R. P. Hughey, C. J. Li*, et al.*, "RNA-seq analysis of glycosylation related gene expression in STZ-induced diabetic rat kidney inner medulla," *Front Physiol,* vol. 6, p. 274, 2015.

[54] Z. Guo, J. F. Gonzalez, J. N. Hernandez, T. N. McNeilly, Y. Corripio-Miyar, D. Frew*, et al.*, "Possible mechanisms of host resistance to Haemonchus contortus infection in sheep breeds native to the Canary Islands," *Sci Rep,* vol. 6, p. 26200, 2016.

[55] S. Osenberg, A. Karten, J. Sun, J. Li, S. Charkowick, C. A. Felice*, et al.*, "Activity-dependent aberrations in gene expression and alternative splicing in a mouse model of Rett syndrome," *Proc Natl Acad Sci U S A,* May 16 2018.

[56] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol,* vol. 15, p. 550, 2014.

[57] Y. Li, S. E. Calvo, R. Gutman, J. S. Liu, and V. K. Mootha, "Expansion of biological pathways based on evolutionary inference," *Cell,* vol. 158, pp. 213-25, Jul 03 2014.

[58] A. G. Bick, S. E. Calvo, and V. K. Mootha, "Evolutionary diversity of the mitochondrial calcium uniporter," *Science,* vol. 336, p. 886, May 18 2012.

[59] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J Mol Biol,* vol. 215, pp. 403-10, Oct 5 1990.

[60] E. Fuchs, "Keratins and the skin," *Annu Rev Cell Dev Biol,* vol. 11, pp. 123-53, 1995.

[61] V. Lopez-Pajares, K. Qu, J. Zhang, D. E. Webster, B. C. Barajas, Z. Siprashvili*, et al.*, "A LncRNA-MAF:MAFB transcription factor network regulates epidermal differentiation," *Dev Cell,* vol. 32, pp. 693-706, Mar 23 2015.

[62] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski*, et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Res,* vol. 22, pp. 1760-74, Sep 2012.

[63] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome biology,* vol. 10, p. R25, 2009.

[64] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics,* vol. 18, pp. S105-S110, 2002.

[65] L. ZHENG and P. YU. (2017). *brt: Biological Relevance Testing,* <https://cran.r-project.org/web/packages/brt/index.html>.

[66] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C. K. Chen, J. Chrast*, et al.*, "GENCODE: producing a reference annotation for ENCODE," *Genome Biol,* vol. 7 Suppl 1, pp. S4 1-9, 2006.

[67] J. Li, S.-P. Deng, G. Wei, and P. Yu, "CITGeneDB: a comprehensive database of human and mouse genes enhancing or suppressing cold-induced thermogenesis validated by perturbation experiments in mice," *Database,* vol. 2018, pp. bay012-bay012, 2018.

[68] S. Hyter, D. J. Coleman, G. Ganguli-Indra, G. F. Merrill, S. Ma, M. Yanagisawa*, et al.*, "Endothelin-1 is a transcriptional target of p53 in epidermal keratinocytes and regulates ultraviolet-induced melanocyte homeostasis," *Pigment Cell Melanoma Res,* vol. 26, pp. 247-58, Mar 2013.

[69] V. Vasireddy, Y. Uchida, N. Salem, S. Y. Kim, M. N. A. Mandal, G. B. Reddy*, et al.*, "Loss of functional ELOVL4 depletes very long-chain fatty acids (≥ C28) and the unique ω-O-acylceramides in skin leading to neonatal death," *Human molecular genetics,* vol. 16, pp. 471-482, 2007.

[70] J. M. Yang, S. M. Sim, H. Y. Kim, and G. T. Park, "Expression of the homeobox gene, HOPX, is modulated by cell differentiation in human keratinocytes and is involved in the expression of differentiation markers," *Eur J Cell Biol,* vol. 89, pp. 537-46, Jul 2010.

[71] P. M. Brunner, E. Guttman-Yassky, and D. Y. Leung, "The immunology of atopic dermatitis and its reversibility with broad-spectrum and targeted therapies," *J Allergy Clin Immunol,* vol. 139, pp. S65-S76, Apr 2017.

[72] E. Guttman-Yassky, M. Suarez-Farinas, A. Chiricozzi, K. E. Nograles, A. Shemer, J. Fuentes-Duculan*, et al.*, "Broad defects in epidermal cornification in atopic dermatitis identified through genomic analysis," *J Allergy Clin Immunol,* vol. 124, pp. 1235-1244 e58, Dec 2009.

[73] R. Westerberg, J. E. Mansson, V. Golozoubova, I. G. Shabalina, E. C. Backlund, P. Tvrdik*, et al.*, "ELOVL3 is an important component for early onset of lipid recruitment in brown adipose tissue," *J Biol Chem,* vol. 281, pp. 4958-68, Feb 24 2006.

[74] J. Lee, J. M. Ellis, and M. J. Wolfgang, "Adipose Fatty Acid Oxidation Is Required for Thermogenesis and Potentiates Oxidative Stress-Induced Inflammation," *Cell Reports,* vol. 10, pp. 266-279, Jan 13 2015.

[75] R. Berry and M. S. Rodeheffer, "Characterization of the adipocyte cellular lineage in vivo," *Nat Cell Biol,* vol. 15, pp. 302-8, Mar 2013.

[76] P. Seale, B. Bjork, W. Yang, S. Kajimura, S. Chin, S. Kuang*, et al.*, "PRDM16 controls a brown fat/skeletal muscle switch," *Nature,* vol. 454, pp. 961-7, Aug 21 2008.

[77] V. Ouellet, S. M. Labbe, D. P. Blondin, S. Phoenix, B. Guerin, F. Haman*, et al.*, "Brown adipose tissue oxidative metabolism contributes to energy expenditure during acute cold exposure in humans," *J Clin Invest,* vol. 122, pp. 545-52, Feb 2012.

[78] K. Okada, K. B. LeClair, Y. Zhang, Y. Li, C. Ozdemir, T. I. Krisko*, et al.*, "Thioesterase superfamily member 1 suppresses cold thermogenesis by limiting the oxidation of lipid droplet-derived fatty acids in brown adipose tissue," *Mol Metab,* vol. 5, pp. 340-51, May 2016.

[79] M. L. Shao, J. Ishibashi, C. M. Kusminski, Q. A. Wang, C. Hepler, L. Vishvanath*, et al.*, "Zfp423 Maintains White Adipocyte Identity through Suppression of the Beige Cell Thermogenic Gene Program," *Cell Metabolism,* vol. 23, pp. 1167-1184, Jun 14 2016.

[80] S. Vernia, Y. J. Edwards, M. S. Han, J. Cavanagh-Kyros, T. Barrett, J. K. Kim*, et al.*, "An alternative splicing program promotes adipose tissue thermogenesis," *Elife,* vol. 5, Sep 16 2016.

[81] A. M. Cypess, S. Lehman, G. Williams, I. Tal, D. Rodman, A. B. Goldfine*, et al.*, "Identification and importance of brown adipose tissue in adult humans," *N Engl J Med,* vol. 360, pp. 1509-17, Apr 09 2009.

[82] H. W. Zhao and Y. F. Huang, "An improved method for combination feature selection in web click-through data mining," *2012 International Symposium on Information Science and Engineering (Isise),* pp. 381-385, 2012.

[83] M. J. Emmett, H. W. Lim, J. Jager, H. J. Richter, M. Adlanmerini, L. C. Peed*, et al.*, "Histone deacetylase 3 prepares brown adipose tissue for acute thermogenic challenge," *Nature,* vol. 546, pp. 544-548, Jun 22 2017.

[84] Z. Bai, X. R. Chai, M. J. Yoon, H. J. Kim, K. A. Lo, Z. C. Zhang*, et al.*, "Dynamic transcriptome changes during adipose tissue energy expenditure reveal critical roles for long noncoding RNA regulators," *PLoS Biol,* vol. 15, p. e2002176, Aug 2017.

[85] A. Mitchell, F. Bucchini, G. Cochrane, H. Denise, P. ten Hoopen, M. Fraser*, et al.*, "EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data," *Nucleic Acids Res,* vol. 44, pp. D595-603, Jan 04 2016.

[86] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams*, et al.*, "ArrayExpress update-simplifying data submissions," *Nucleic Acids Research,* vol. 43, pp. D1113-D1116, Jan 28 2015.

[87] R. Petryszak, M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett*, et al.*, "Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants," *Nucleic Acids Res,* vol. 44, pp. D746-52, Jan 04 2016.

[88] M. Todaro, M. Gaggianesi, V. Catalano, A. Benfante, F. Iovino, M. Biffoni*, et al.*, "CD44v6 Is a Marker of Constitutive and Reprogrammed Cancer Stem Cells Driving Colon Cancer Metastasis," *Cell Stem Cell,* vol. 14, pp. 342-356, Mar 6 2014.

[89] W. M. Lau, E. Teng, H. S. Chong, K. A. P. Lopez, A. Y. L. Tay, M. Salto-Tellez*, et al.*, "CD44v8-10 Is a Cancer-Specific Marker for Gastric Cancer Stem Cells," *Cancer Research,* vol. 74, pp. 2630-2641, May 1 2014.

[90] Y. Lu, W. Xu, J. Ji, D. Feng, C. Sourbier, Y. Yang*, et al.*, "Alternative splicing of the cell fate determinant Numb in hepatocellular carcinoma," *Hepatology,* vol. 62, pp. 1122-31, Oct 2015.

[91] D. Vecellio Reane, F. Vallese, V. Checchetto, L. Acquasaliente, G. Butera, V. De Filippis*, et al.*, "A MICU1 Splice Variant Confers High Sensitivity to the Mitochondrial Ca2+ Uptake Machinery of Skeletal Muscle," *Mol Cell,* vol. 64, pp. 760-773, Nov 17 2016.

[92] F. Bouffard, K. Plourde, S. Belanger, G. Ouellette, Y. Labrie, and F. Durocher, "Analysis of a FANCE Splice Isoform in Regard to DNA Repair," *J Mol Biol,* vol. 427, pp. 3056-73, Sep 25 2015.

[93] R. P. Hulse, R. A. Drake, D. O. Bates, and L. F. Donaldson, "The control of alternative splicing by SRSF1 in myelinated afferents contributes to the development of neuropathic pain," *Neurobiol Dis,* vol. 96, pp. 186-200, Dec 2016.

[94] J. R. Tejedor, P. Papasaikas, and J. Valcarcel, "Genome-wide identification of Fas/CD95 alternative splicing regulators reveals links with iron homeostasis," *Mol Cell,* vol. 57, pp. 23-38, Jan 8 2015.

[95] B. Cieply and R. P. Carstens, "Functional roles of alternative splicing factors in human disease," *Wiley Interdiscip Rev RNA,* vol. 6, pp. 311-26, May-Jun 2015.

[96] O. Villate, J. V. Turatsinze, L. G. Mascali, F. A. Grieco, T. C. Nogueira, D. A. Cunha*, et al.*, "Nova1 is a master regulator of alternative splicing in pancreatic beta cells," *Nucleic Acids Res,* vol. 42, pp. 11818-30, Oct 2014.

[97] Y. Saito, S. Miranda-Rottmann, M. Ruggiu, C. Y. Park, J. J. Fak, R. Zhong*, et al.*, "NOVA2-mediated RNA regulation is required for axonal pathfinding during development," *Elife,* vol. 5, May 25 2016.

[98] A. J. Linares, C. H. Lin, A. Damianov, K. L. Adams, B. G. Novitch, and D. L. Black, "The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation," *Elife,* vol. 4, p. e09268, Dec 24 2015.

[99] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res,* vol. 28, pp. 27-30, Jan 01 2000.

[100] C. UniProt, "UniProt: a hub for protein information," *Nucleic Acids Res,* vol. 43, pp. D204-12, Jan 2015.

[101] P. Konieczny, E. Stepniak-Konieczna, K. Taylor, L. J. Sznajder, and K. Sobczak, "Autoregulation of MBNL1 function by exon 1 exclusion from MBNL1 transcript," *Nucleic Acids Res,* vol. 45, pp. 1760-1775, Feb 28 2017.

[102] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell*, et al.*, "The Pfam protein families database: towards a more sustainable future," *Nucleic Acids Res,* vol. 44, pp. D279-85, Jan 04 2016.

[103] J. T. Eppig, H. Motenko, J. E. Richardson, B. Richards-Smith, and C. L. Smith, "The International Mouse Strain Resource (IMSR): cataloging worldwide mouse and ES cell line resources," *Mamm Genome,* vol. 26, pp. 448-55, Oct 2015.

[104] K. E. Lukong, K. W. Chang, E. W. Khandjian, and S. Richard, "RNA-binding proteins in human genetic disease," *Trends Genet,* vol. 24, pp. 416-25, Aug 2008.

[105] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS Lett,* vol. 582, pp. 1977-86, Jun 18 2008.

[106] H. Zhou, M. Mangelsdorf, J. Liu, L. Zhu, and J. Y. Wu, "RNA-binding proteins in neurological diseases," *Sci China Life Sci,* vol. 57, pp. 432-44, Apr 2014.

[107] A. Castello, B. Fischer, M. W. Hentze, and T. Preiss, "RNA-binding proteins in Mendelian disease," *Trends Genet,* vol. 29, pp. 318-27, May 2013.

[108] B. Pereira, M. Billaud, and R. Almeida, "RNA-Binding Proteins in Cancer: Old Players and New Actors," *Trends Cancer,* vol. 3, pp. 506-528, Jul 2017.

[109] E. J. Mientjes, I. Nieuwenhuizen, L. Kirkpatrick, T. Zu, M. Hoogeveen-Westerveld, L. Severijnen*, et al.*, "The generation of a conditional Fmr1 knock out mouse model to study Fmrp function in vivo," *Neurobiol Dis,* vol. 21, pp. 549-55, Mar 2006.

[110] A. Abu-Baker and G. A. Rouleau, "Oculopharyngeal muscular dystrophy: recent advances in the understanding of the molecular pathogenic mechanisms and treatment strategies," *Biochim Biophys Acta,* vol. 1772, pp. 173-85, Feb 2007.

[111] P. Dion, V. Shanmugam, C. Gaspar, C. Messaed, I. Meijer, A. Toulouse*, et al.*, "Transgenic expression of an expanded (GCG)13 repeat PABPN1 leads to weakness and coordination defects in mice," *Neurobiol Dis,* vol. 18, pp. 528-36, Apr 2005.

[112] W. D. Cress, P. Yu, and J. Wu, "Expression and alternative splicing of the cyclin-dependent kinase inhibitor-3 gene in human cancer," *Int J Biochem Cell Biol,* vol. 91, pp. 98-101, Oct 2017.

[113] L. Dai, K. Chen, B. Youngren, J. Kulina, A. Yang, Z. Guo*, et al.*, "Cytoplasmic Drosha activity generated by alternative splicing," *Nucleic Acids Res,* vol. 44, pp. 10454-10466, Dec 01 2016.

[114] M. Carlson, "GO.db: A set of annotation maps describing the entire Gene Ontology," 2017.

[115] I. van Kruijsbergen, S. Hontelez, and G. J. Veenstra, "Recruiting polycomb to chromatin," *Int J Biochem Cell Biol,* vol. 67, pp. 177-87, Oct 2015.

[116] P. N. Grozdanov, A. Amatullah, J. H. Graber, and C. C. MacDonald, "TauCstF-64 Mediates Correct mRNA Polyadenylation and Splicing of Activator and Repressor Isoforms of the Cyclic AMP-Responsive Element Modulator (CREM) in Mouse Testis," *Biol Reprod,* vol. 94, p. 34, Feb 2016.

[117] P. N. Grozdanov, J. Li, P. Yu, W. Yan, and C. C. MacDonald, "Cstf2t Regulates expression of histones and histone-like proteins in male germ cells," *Andrology,* Apr 19 2018.

[118] J. M. Bujnicki, M. Feder, M. Radlinska, and R. M. Blumenthal, "Structure prediction and phylogenetic analysis of a functionally diverse family of proteins homologous to the MT-A70 subunit of the human mRNA:m(6)A methyltransferase," *J Mol Evol,* vol. 55, pp. 431-44, Oct 2002.