SCIENTIFIC REPORTS

OPEN

# mTFkb: a knowledgebase for fundamental annotation of mouse transcription factors

Received: 1 December 2016
Accepted: 12 April 2017
Published online: 08 June 2017

Kun Sun[1,2], Huating Wang[1,3] & Hao Sun[1,2]

**Transcription factors (TFs) are well-known important regulators in cell biology and tissue development. However, in mouse, one of the most widely-used model species, currently the vast majority of the known TFs have not been functionally studied due to the lack of sufficient annotations. To this end, we collected and analyzed the whole transcriptome sequencing data from more than 30 major mouse tissues and used the expression profiles to annotate the TFs. We found that the expression patterns of the TFs are highly correlated with the histology of the tissue types thus can be used to infer the potential functions of the TFs. Furthermore, we found that as many as 30% TFs display tissue-specific expression pattern, and these tissue-specific TFs are among the key TFs in their corresponding tissues. We also observed signals of divergent transcription associated with many TFs with unique expression pattern. Lastly, we have integrated all the data, our analysis results as well as various annotation resources to build a web-based database named mTFkb freely accessible at http://www.myogenesisdb.org/mTFkb/. We believe that mTFkb could serve as a useful and valuable resource for TF studies in mouse.**

Transcription factors (TFs) are a family of proteins that could bind to specific DNA sequences, usually in enhancer or promoter regions, to regulate the expression of target genes, either positively (as an activator) or negatively (as a repressor)[1–3]. In human, around 8% of the total genes encode TFs[4]. TFs are found to be highly conserved among most of the organisms. For instance, the numbers of annotated TFs in human (Homo Sapiens) and mouse (Mus Musculus) are similar[5] and most of them are conserved between these two species. This highly conserved characteristic suggests that TFs are among the fundamental proteins for normal cellular functions[6]. Therefore, there is ongoing interest in the functional investigation of TFs. They are known essential regulators in normal cell function and tissue development. For instance, MyoD (Myogenic Differentiation 1) and Myf5 (Myogenic factor 5) play key roles in the development of limb and skeletal muscle[7,8]. Furthermore, TFs that are key to guide cell differentiation and tissue development are discovered to interact with regulatory DNA elements such as enhancers and promoters[3,9]. Recent studies also showed that key TFs could establish super-enhancers, clusters of enhancers with high activity, which are essential in controlling cell identity and disease[10,11]. In addition, more and more studies demonstrated the successful reprogramming of somatic cells using a "cocktail" containing key TFs of the target cell type[12]. Very interestingly, emerging reports demonstrated the biological phenomenon of divergent transcription from the promoters of TFs[13,14], which could be helpful in deciphering its significance and functional mechanism[14,15]. For instance, our group has recently discovered a novel long noncoding RNA, Linc-Yy1, which is transcribed from ~2 kb upstream of the Yy1 (Yin Yang 1) gene and serves as an important regulator of mouse skeletal myoblast differentiation through interaction with the Yy1 transcription factor[14]. Collectively, the existing studies reinforced that the TFs are among the most important regulators affecting the identity of cell/tissue type through diversified mechanisms of actions; it is thus imperative to identify the key TFs that are critical for the development of certain tissues.

Knowing their functional significance, however, most of the known TFs have yet to be characterized[16]. Existing studies in human found that the TFs are expressed in a tissue-dependent manner hence the expression pattern of the TFs across various tissues is closely correlated with their functions and could be used to mine the key TFs for the tissues[16–19]. Similar study however is still lacking in mouse, warranting the creation of a public

[1]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China. [2]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China. [3]Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong SAR, China. Correspondence and requests for materials should be addressed to H.W. (email: huating.wang@cuhk.edu.hk) or H.S. (email: haosun@cuhk.edu.hk)

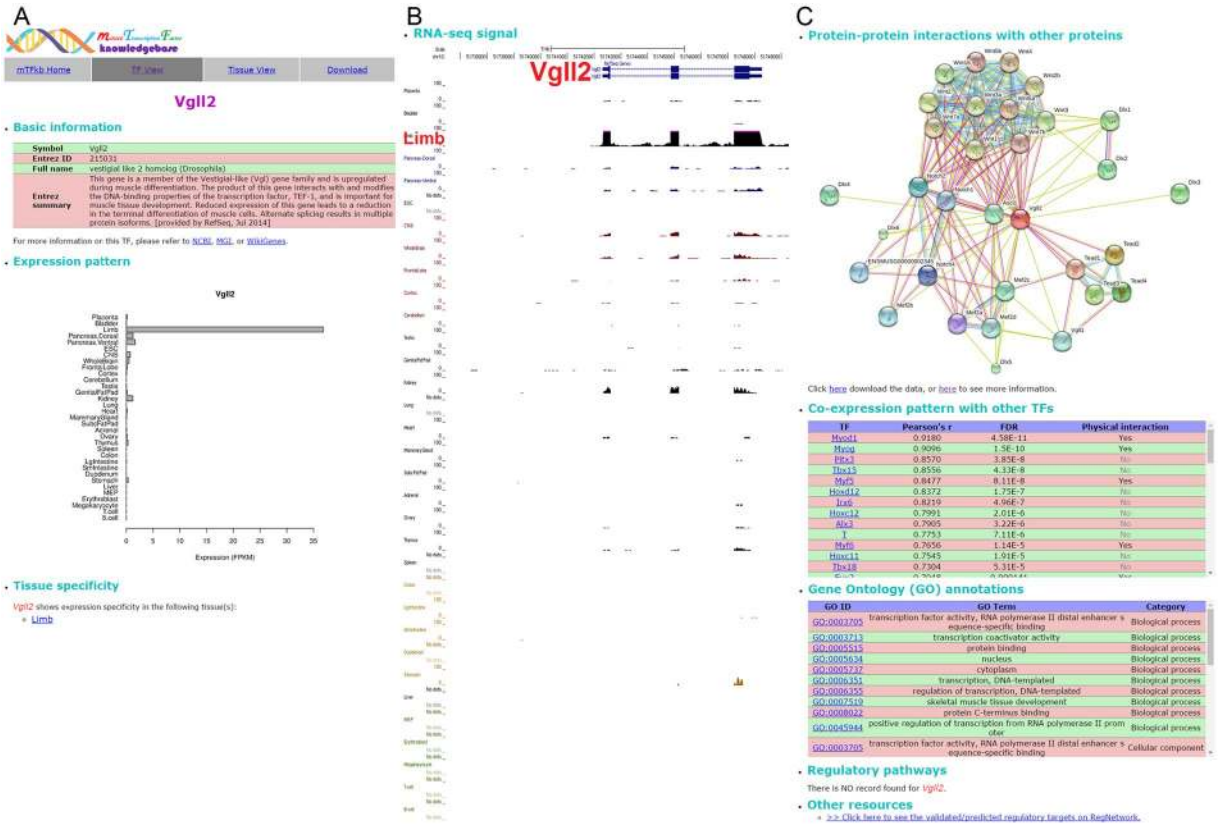


**Figure 1.** Snapshot of mTFkb webserver. Illustrated as an example is the transcription factor (TF) Vgll2. (**A**) Basic annotation, expression pattern and tissue-specificity identification; (**B**) Normalized RNA-seq signal across the mouse tissues analyzed; (**C**) Functional annotations of the TF including protein-protein interactions, co-expression pattern, Gene Ontology, regulatory pathways and targets.

knowledgebase for mouse TFs, which provides fundamental annotations. In addition, despite the existence of several TF databases such as TFdb[20], TFCat[21] and DBD[22] that provide catalogs of TFs, functional characterizations are mostly lacking in these databases. RegNetwork[23], YY1TargetDB[24] and TFBSshape[25] integrated only the regulatory targets information of the TFs. Another widely-used TF database, AnimalTFDB[5], on the other hand, integrates annotations including Gene Ontology and regulatory pathway. In its 2.0 version[26], it also incorporated tissue expression data but a limited number of mouse tissues were included with no further analyses provided. We reason that a database integrating expression analyses as well as functional annotations is needed to facilitate the studies on mouse TFs. To this end, in this study, we employed the transcriptome data from more than 30 major mouse tissues to annotate the TFs. Our analysis identified *bona fide* key TFs in many mouse tissues and shed novel insights of their tissue-specific functionality. In addition, divergent transcription associated with the promoter/enhancer regions of many TFs was observed and also showed unique tissue-specific expression pattern. Furthermore, we integrated functional annotations from various resources including protein-protein interactions, Gene Ontology (GO) and regulatory pathways to develop a web-based database named mTFkb (mouse transcription factor knowledgebase) freely accessible to the academic community, which we believe will become a valuable resource for studying TFs in mouse.

## Results

**The mouse transcription factor knowledgebase.** We collected the whole transcriptome shotgun sequencing (a.k.a. RNA-seq) data for 33 major mouse tissues from the literature and profiled the expression pattern of 1,603 known mouse TFs (see Methods). Based on this data, we built a web-based database named mTFkb, which integrated all the expression data, our functional analysis results as well as functional annotations from various resources, freely available at http://sunlab.cpy.cuhk.edu.hk/mTFkb/. The database allows the users to inspect the expression profile and the analysis results for each mouse TF (the "TF View" page) or tissue (the "Tissue View" page) via a user-friendly interface. One snapshot was shown in Fig. 1 using Vgll2 (vestigial like 2 homolog (Drosophila)) as an example. Its basic information, expression pattern and tissue specificity can be fetched through the query function in the "TF View" page (Fig. 1A). Snapshot of RNA-seq signal tracks from each tissue is also included (Fig. 1B). Furthermore, we provide functional annotations including protein-protein interactions[27], Gene Ontology (GO)[28], and regulatory pathways[29] as well as other possible information (e.g. regulatory targets[23–25] if available) by integrating annotations from various resources (Fig. 1C). The detailed descriptions are provided in the following sections.

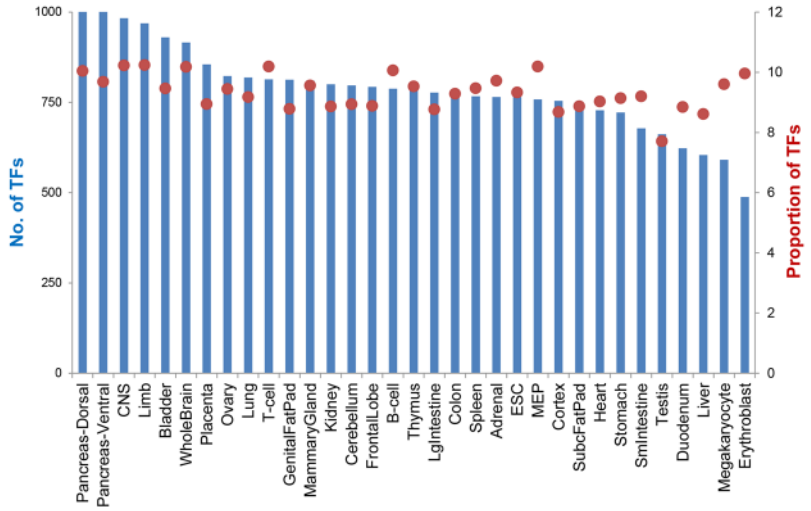


**Figure 2.** Numbers of transcription factors (TFs) expressed in each tissue (blue bars) and the proportion of the expressed TFs versus all expressed genes (red points, numbers are given as a percentage).

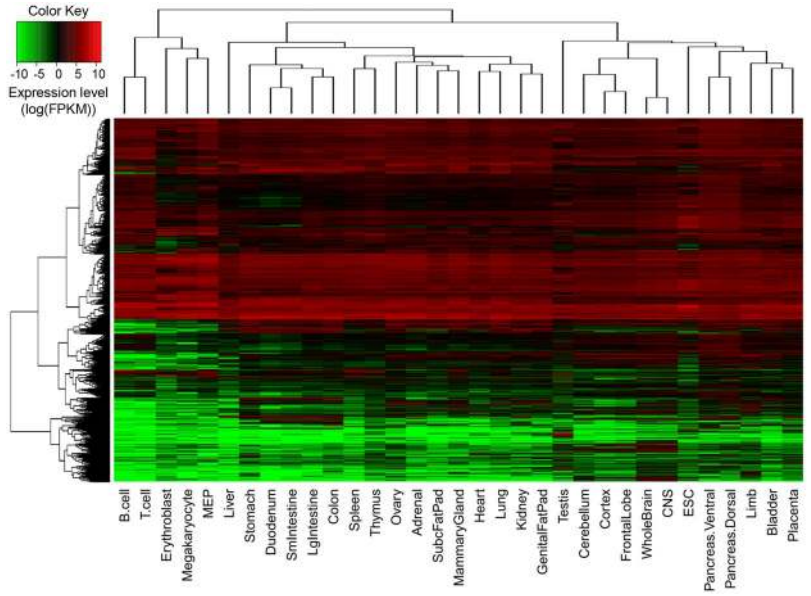


**Figure 3.** Hierarchical clustering of the mouse tissues using the expression values of the transcription factors.

**Expression pattern of the mouse TFs.** After profiling the expression values of the TFs using the RNA-seq data, we further investigated the expression pattern of the TFs across various tissues, which was also included in the "TF View" page. We found that the number of expressed TFs varies significantly among different tissues (Fig. 2, and the "Tissue View" page). For example, there were more than 1000 TFs expressed in pancreas tissues, while as a contrast, the number of expressed TFs in erythroblasts was only round 500. Still, when compared to the total number of genes expressed in each tissue, the proportions of the TFs were relatively stable (Fig. 2), which was consistent with previous findings in human[16].

Meanwhile, we performed the hierarchical clustering of the tissues using the expression values of the TFs. As expected, the result in Fig. 3 showed that histologically related tissues were clustered together. For instance, the tissues from hematopoietic system (B-cells, T-cells, Erythroblasts, Megakaryocytes and MEP (Megakaryocyte-Erythroid Progenitor cell)), digestive system (Stomach, Duodenum, Small intestines, Large intestines, and Colon), and nervous system (Cerebellum, Cortex, Frontal lobe, Whole brain, and CNS (Central Nervous System)) were clustered together, separately. This result indicated that the expression values of the TFs are highly correlated with the histology and function of the corresponding tissue. In addition, we also found that some TFs expressed ubiquitously in most tissues while others expressed in only a small proportion of the tissues. To strengthen the notion, for each TF, we counted the number of tissues in which it is expressed. As shown in Fig. 4, we found that TFs expressed in a "U-shape" across the tissues, i.e., the majority of the TFs tend to express

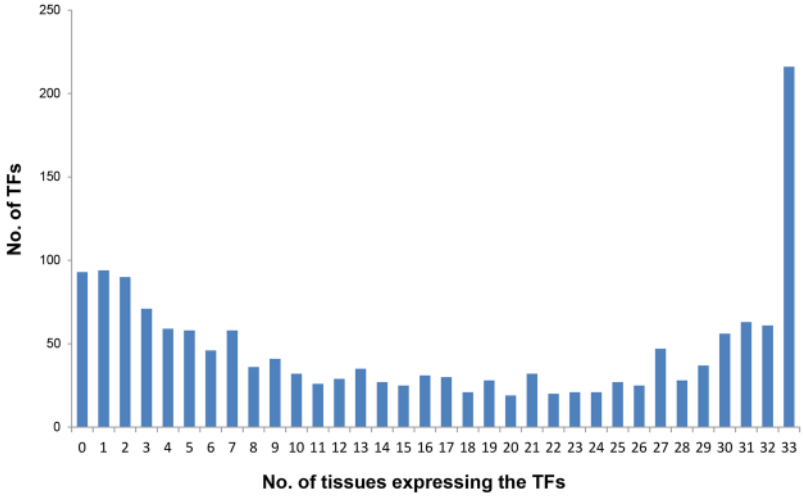


**Figure 4.** Distribution of transcription factors (TFs) based on the number of tissues in which they express.

in either most of the tissues or in a very limited number of tissues, suggesting diversified functional scenarios: some TFs are "housekeeping" while others may be highly tissue-specific. The latter are more likely to be key TFs defining and maintaining the cell/tissue identity[3,10,17], thus deserved a more intensive exploration.

**Exploration of key TFs in various mouse tissues.** As shown in Fig. 2, for all the tissue types, hundreds of TFs are expressed, while usually only a small proportion of them are potential key TFs which play important roles that are tightly related to the function and identity of the tissue type. As shown in Fig. 3 by the hierarchical clustering, we found that the dynamics of the TF expression are highly correlated with the tissue histology, the expression patterns was thus used to identify key TFs for the tissues[17]. To this end, we searched for tissue-specifically expressed TFs as the candidates key TFs (see Methods) (The "Tissue View" page). As a result, we found that around 30% (489 out of 1603) TFs showed tissue-specificity (Fig. 5A). On the other hand, the number of TFs that show specificity in each tissue type varies significantly (Fig. 5B). For instance, more than 70 TFs are specifically expressed in ESC while less than 20 in liver (Fig. 5B). The variation in the number of specifically-expressed TFs across the tissues might be correlated with the functional complexity of the tissues.

To investigate the performance of our approach, especially the ability to identify potential key TFs, we first examined several well-known master TFs for certain tissues. As shown in Fig. 6, Sox2 (SRY-box containing gene 2) and Pou5f1 are known master TFs in ESC[30,31], and indeed we found them specifically expressed in ESC (Fig. 6A and B); similarly, Myod1 and Myog (myogenin) are known master TFs in skeletal muscle development[32] and was found to display specificity in limb tissue (Fig. 6C and D); Foxi1 (forkhead box I1) and Foxn1 (foxhead box N1) were identified to be expressed in kidney and thymus, respectively, which is consistent with previous knowledge that they are key regulators of kidney[33] and thymus[34], respectively (Fig. 6E and F). In addition, we compared our result with the key TFs identified in human by D'Alessio *et al.*[17]. Interestingly, we found that in many tissues, the mouse orthologs of top-ranked key TFs in human were also identified to be specific in the homologous tissue in mouse. A comparison for the pancreas tissue ("pancreatic islet cells" in D'Alessio *et al.* versus "pancreas ventral" in mTFkb) was shown in Table 1 as an illustrating example and more results from other tissues could be found in Suppl. Table S1.

The tissue-specificity identified by mTFkb might suggest uncharacterized functions of the TFs in their corresponding tissues and this information could be especially valuable for the TFs that have not been comprehensively investigated. For instance, 1700003F12Rik and B930041F14Rik are two TFs coded by RIKEN cDNA 1700003F12 and B930041F14 genes, respectively, and their functions remain completely uncharacterized. Our data revealed their unique expression in testis (Fig. 7A) and adrenal glands (Fig. 7B), respectively, which will be helpful in guiding the functional studies in the future. On the other hand, Hoxa11 (homeo box A11) is known to be involved in repressing MyoD during limb muscle development[35]. Interestingly, in addition to the high expression in limb, we found that it is also enriched in bladder and colon, which suggested potentially uncharacterized functions (Fig. 7C). To this end, the antisense gene of human HOXA11 (i.e., HOXA11-AS) was demonstrated to be a biomarker for urothelial carcinoma[36] which also correlates with tumor size and metastasis in colorectal cancer[37], supporting that Hoxa11 may play some roles in the bladder and colon tissues. Similarly, Fig. 7D shows that Stat4 (signal transducer and activator of transcription 4) is specifically expressed in the lymphocytes and testis. It is known to be essential for mediating responses to IL12 in lymphocytes and regulating the differentiation of T helper cells[38], while its potential functions in the testis remain to be investigated. Our expression analysis thus provided valuable information for future functional and mechanistic studies.

**Divergent transcription associated with TFs.** To explore whether divergent transcription associated with TFs is a prevalent phenomenon in mouse tissues, we further examined the normalized RNA-seq signals that could be obtained by querying the TF through the "TF View" page (Fig. 1B). For many TFs, we could observe a certain level of RNA-seq signal at the promoter/enhancer regions, indicating the potential existence of divergent

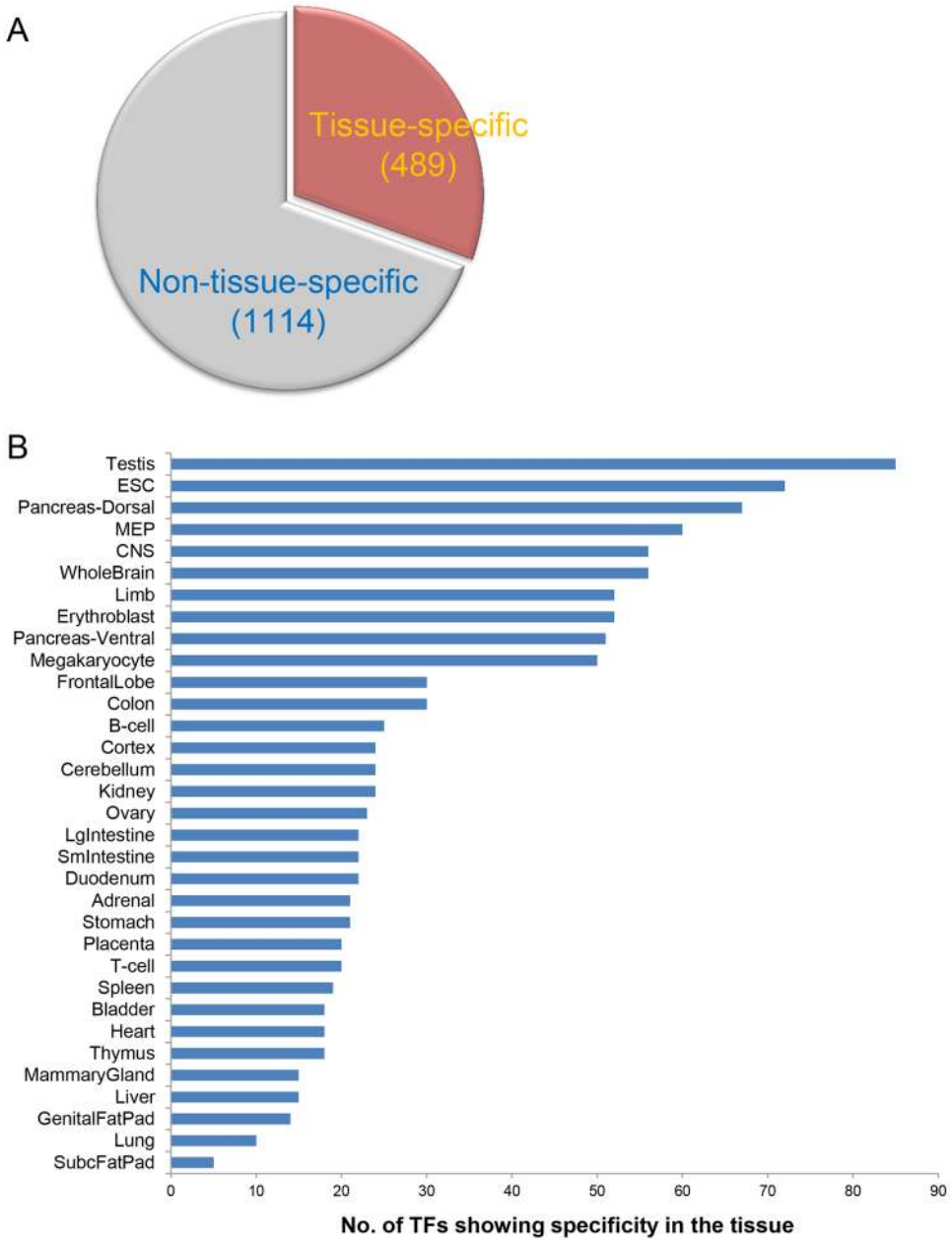


**Figure 5.** (**A**) 489 out of 1603 (30.5%) transcription factors (TFs) show tissue specific expression pattern while 1114 are non-tissue specific. (**B**) Number of tissue specific TFs in each mouse tissue.

transcription associated with the TFs. Yy1 and Myod1 were plotted in Fig. 8 as examples. Consistent with our recent report, there is strong RNA-seq signal for the divergent transcript of Yy1 gene (i.e., Linc-Yy1)[14]. However, despite the fact that Yy1 is ubiquitously expressed among most mouse tissues[24,39], the signal of Linc-Yy1 could only be observed in limb and the nervous system (Fig. 8A). The function of Yy1 and its interplay with Linc-Yy1 has been characterized in the muscle development[40]; it however remains to be determined whether Linc-Yy1 interacts with Yy1 during the development of nervous system considering that Yy1 is a known important regulator during the nervous system development[41]. Similarly, RNA-seq signal was also observed in the promoter of Myod1 in limb tissue where MyoD is highly expressed (Fig. 8B), which warrants further investigation in the future. Collectively, these findings suggested that divergent transcripts display unique tissue-specific expression pattern independent of the associated TFs.

**Functional annotations of the TFs.** The above expression analysis, key TF annotation and divergent transcription have, to some degree, provided information on the functional aspects of each TF. To further strengthen the functional annotations, we analyzed various features of a TF, including protein-protein interactions[27], co-expression pattern, Gene Ontology (GO)[28], regulatory pathways[29] as well as other annotation resources (e.g. regulatory targets[23,24] and DNA binding pattern[25]) by integrating existing information into mTFkb (Fig. 1C). As

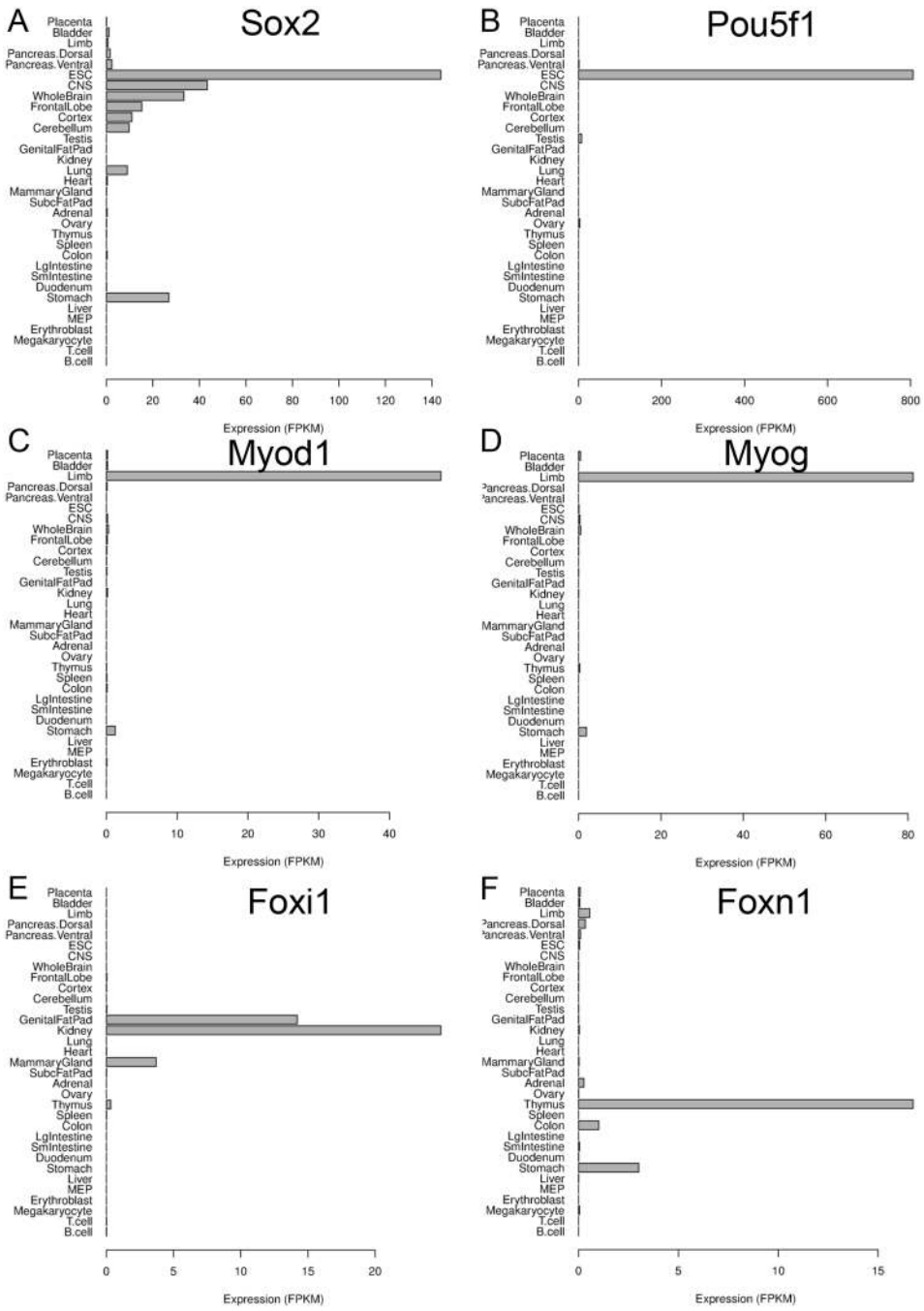


**Figure 6.** Expression pattern of selected known key transcription factors (TFs). (**A**) Sox2. (**B**) Pou5f1. (**C**) Myod1. (**D**) Myog. (**E**) Foxi1. (**F**) Foxn1.

shown in Fig. 1A and B, analysis of RNA-seq data showed that Vgll2 is specifically expressed in the limb tissue, suggesting that it may serve as a key TF in the muscle system which is consistent with the previous knowledge of its involvement in skeletal muscle differentiation[42,43]. To gain further understanding of its role, inspection of protein-protein interaction led to the discovery that Vgll2 interacts with a cluster of Wnt proteins, Mef2 (myocyte enhancer factor-2) family proteins (Mef2a, Mef2b, Mef2c and Mef2d), Notch1 and Notch2 as well as Tead (TEA domain) family proteins (Fig. 1C). Consistently, previous study had shown the interaction between Vgll2 and Mef2d in C2C12 cell line (a widely used mouse myoblast cell line)[44]; the interaction with the Wnt signaling pathway had also been discovered in Xenopus[45]. In addition, by co-expression analysis, we found that Vgll2 was associated with Myod1, Myog, Pitx3 (Paired Like Homeodomain 3), Tbx15 (T-box 15), Myf5, etc (Fig. 1C); among them physical interactions were also identified with Myod1, Myog and Myf5, which are well-known regulatory TFs of skeletal muscle development[32], suggesting its possible functional connection with mouse skeletal muscle development. Consistently, GO analysis revealed a GO term of "skeletal muscle tissue development" related with

| Rank in human pancreas | Transcription factor | Occurrence in mTFkb |
|---|---|---|
| 1 | RFX6 | Yes |
| 2 | INSM1 | No |
| 3 | PAX6 | Yes |
| 4 | ISL1 | Yes |
| 5 | NEUROD1 | Yes |
| 6 | GLIS3 | No orthologue in mouse |
| 7 | NR5A2 | Yes |
| 8 | ZNF165 | No orthologue in mouse |
| 9 | ARX | Yes |
| 10 | MNX1 | Yes |
| 14 | MAFB | No |
| 371 | PDX1 | Yes |
| 543 | PAX4 | No |
| 755 | NEUROG3 | Yes |

**Table 1.** Co-occurrence of key transcription factors in human and mouse pancreas tissues.

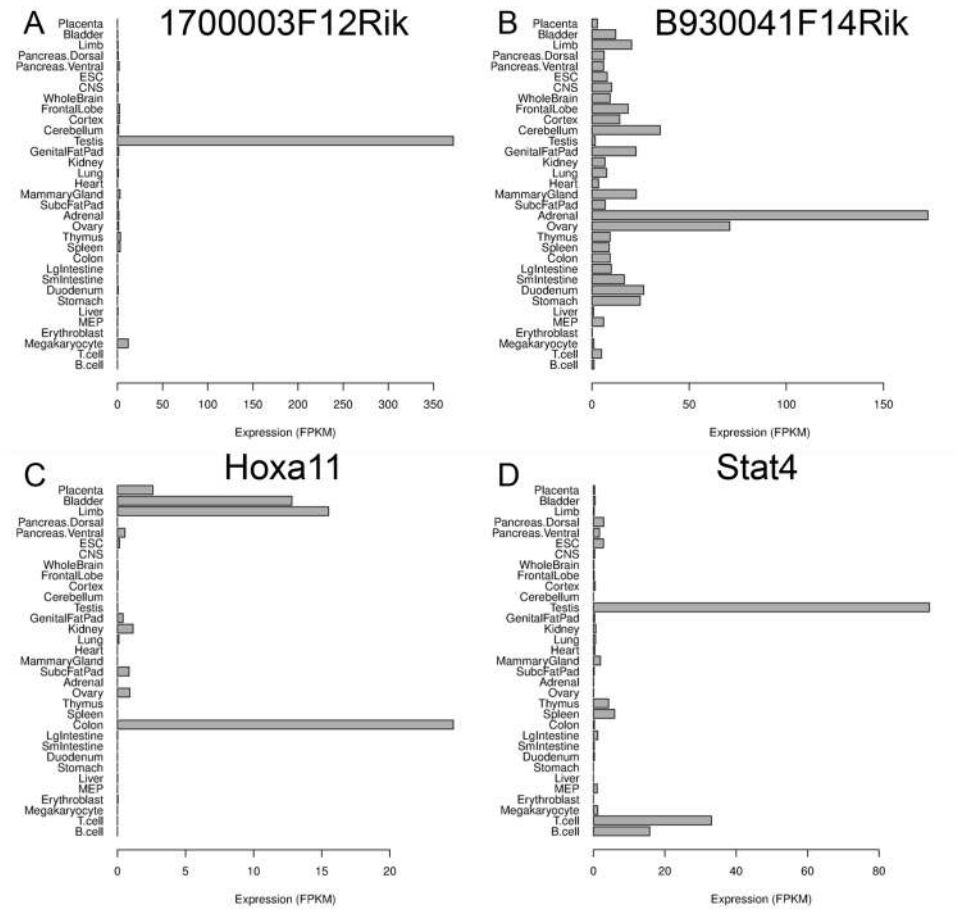


**Figure 7.** Expression pattern of selected transcription factors (TFs) with uncharacterized functions. (**A**) B930041F14Rik. (**B**) 1700003F12Rik. (**C**) Hoxa11. (**D**) Stat4.

Vgll2 (Fig. 1C). Exploring RegNetwork[23], a data repository of regulatory relationships for human and mouse through "other resources", we found no known regulatory targets of Vgll2 probably due to the lack of sufficient functional studies of Vgll2 in mouse (Fig. 1C). Our functional annotations thus had revealed unknown aspects of Vgll2 involvement in limb development that could be tested experimentally in the future.

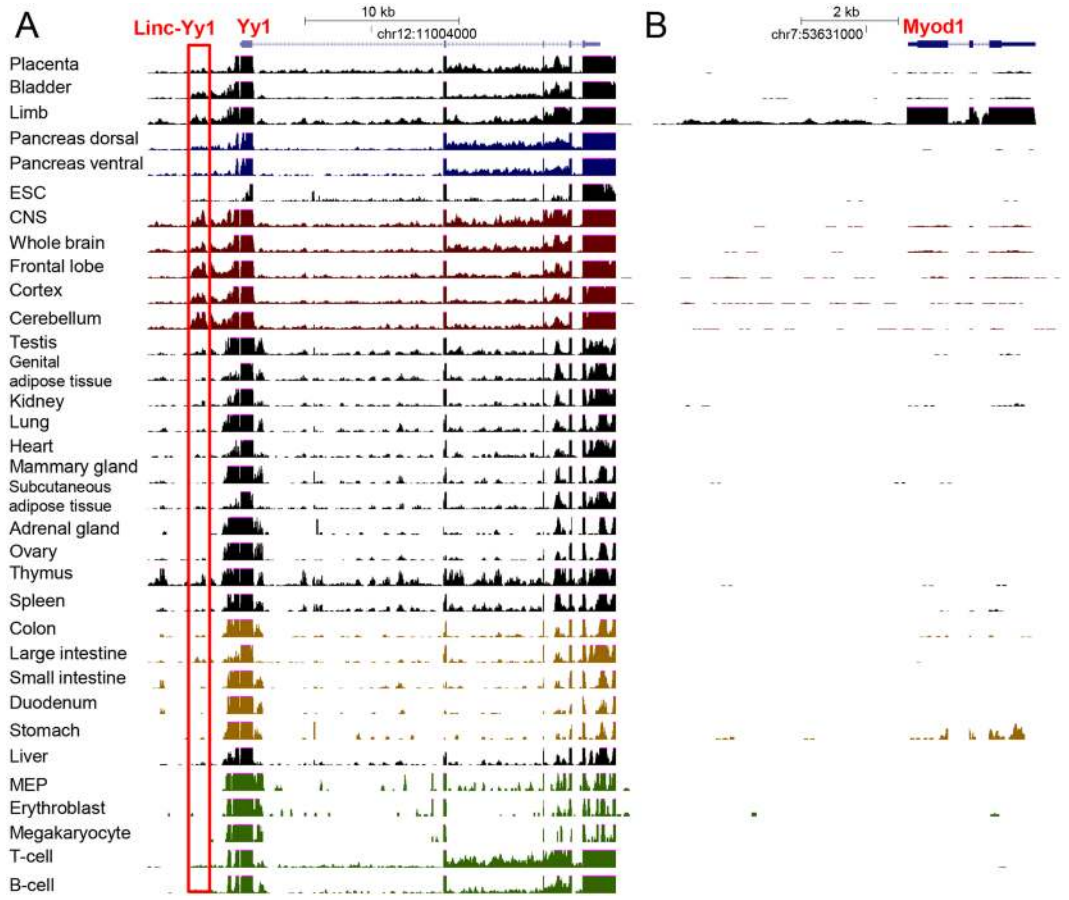


**Figure 8.** Normalized RNA-seq signal across the mouse tissues for (**A**) Yy1 and (**B**) Myod1. The red box in (**A**) indicates the genomic coordination of Linc-Yy1, the characterized divergent transcript of Yy1.

## Discussion

In this paper, we present mTFkb, a web-based database dedicated to the annotation of mouse TFs. mTFkb integrates the expression data from 33 major mouse tissues and provides novel insights into the expression pattern of the TFs. mTFkb is freely available thus allowing users to inspect the data for any TF and tissue via the web interface. Unlike most of other TF databases that only provide the catalog of the TFs or limited functional information, mTFkb provides the fundamental functional annotations including the tissue-specificity identification, key TF interference, RNA-seq signal profiling, divergent transcription screening, protein-protein interaction, co-expression analysis, GO annotation as well as regulatory pathway/targets. In addition, mTFkb covers the major tissues in mouse therefore serves as a comprehensive and valuable resource for fundamental functional annotation of TFs.

Among all the functional analyses that mTFkb provides, the identification of key TFs is the most valuable. Inferring the functional importance of known and unknown TFs in certain tissues will be valuable to guide the selection of the most important TFs in the tissue of interest for future mechanistic investigations. These key TFs likely represent the functional components in the "cocktail" of TFs used for cell reprogramming[17]. Not surprisingly, we found that the key TFs identified in mTFkb highly resembled those predicted in human since most TFs are highly conserved between human and mouse. For example, we found that most of the co-occurred TFs in Table 1 had been proved to play key roles in the Pancreas tissue. For instance, Rfx6 (regulatory factor X6) and Pax6 (paired box 6) are essential to maintain the functional identity of pancreatic beta-cells[46] and islet cells[47], respectively; Isl1 (ISL LIM homeobox 1) is a well-known key regulator for pancreatic islets and functions in the maturation, proliferation and survival of the endocrine pancreas[48]. Collectively our data suggest that the tissue-specific TFs included in mTFkb are of high confidence to be the *bona fide* key TFs of the corresponding tissues.

Besides the expression pattern analysis, we also profiled divergent transcription associated with TFs. Unfortunately, due to the lacking of a comprehensive catalog of the divergent transcripts, a quantitative analysis could not be performed. Nevertheless, our findings suggested the wide existence of divergent transcription and distinct tissue-specific expression pattern from the associated TFs. For instance, MyoD is a tissue-specific TF for limb and the divergent transcription signal was only observed in limb (Fig. 8B). In contrast, even though Yy1 did not show strong tissue-specificity, its divergent transcript (i.e., Linc-Yy1) showed tissue-specific expression pattern (Fig. 8A). In this regard, the RNA-seq signal mTFkb provides could serve as a resource for inspecting the presence and expression pattern of the potential divergent transcription for further studies of their functionality.

Lastly, by integrating functional annotations from various existing resources, further analysis of a TF's role through protein-protein interaction, co-expressed TFs, GO analysis and regulatory pathways/targets become possible. As demonstrated by the example of Vgll2 in Fig. 1, the integrated information can serve as the foundation for future functional exploration of a TF in certain tissue/cell.

## Materials and Methods

**Data collection.** 33 RNA-seq datasets, one for each mouse tissue, were collected from multiple sources including the ENCODE Project (Adrenal glands, Bladder, Cerebellum, CNS, Colon, Cortex, Duodenum, Frontal lobe, Genital adipose tissue, Heart, Kidney, Large intestine, Limb, Liver, Lung, Mammary Gland, Ovary, Placenta, Subcutaneous adipose tissue, Small intestine, Spleen, Stomach, Testis, Thymus, Whole brain, Erythroblast, Megakaryocyte and MEP)[49], Guttman *et al.* (ESC)[50], Kim *et al.* (T-cell and B-cell)[51] and Rodriguez-Seguel *et al.* (Pancreas ventral and Pancreas dorsal)[52]. The detailed information could be found in Suppl. Table S2. For consistency, all the RNA-seq datasets involved in this study were generated using Poly-A extraction protocol and sequenced on Illumina sequencer.

A list of 1,675 mouse TFs was obtained from the most updated version of TFdb (Riken TF Database)[20], which is a widely used database in the literatures. Among the 1,675 annotated TFs, 1,603 (95.5%) could be found in RefSeq[53] gene annotation and used in the analyses.

**RNA-seq data processing and expression profiling.** After downloading the raw sequencing reads from the RNA-seq datasets, a preprocessing procedure was first performed to remove 1) sequencing adaptors; 2) low-quality base-pairs; and 3) PCR duplications using in-house programs[54]. Then the filtered reads were aligned to the mouse reference genome (UCSC mm9/NCBI 37) using TopHat (version 2.0.9)[55] guided by the RefSeq[53] genes (the "-G" option of Tophat) with default parameters. Gene expression profiling was performed using Cufflinks (version 2.1.1)[56] against the RefSeq genes with default parameters. Cufflinks employs a built-in normalization scheme to improve the estimation of expression[57]. The expression level of the genes were quantified as FPKM (Fragments Per Kilobase of transcript per Million mapped reads)[56] values which had been demonstrated to be a reasonable measurement for expression quantifications[58]. A FPKM value of 5 was used as the threshold to call a gene/TF as "expressed" in each tissue type. A value of 1 was added to each raw FPKM value of the expression matrix before transforming to log2 scale (i.e., log-normalization) for the downstream data analyses[18,19]. The above log-normalization method has been demonstrated to be an appropriate normalization method for tissue-specificity analysis[18]. Indeed, after the log-normalization, we found that the expression distributions were similar across the samples and the hierarchical clustering result was based on the tissue histology rather than the laboratory of origin, confirming that the normalized expression profiles appeared consistent and comparable across the samples[18,57–59]. Hierarchical clustering of the tissues using expression values of the TFs was performed using R. For each tissue, the RNA-seq signal was extracted from the TopHat mapping result and normalized by the total number of aligned reads using in-house programs. For co-expression analysis, we calculated the Pearson's correlation for all the TF pairs using the expression values across all the tissues and the p-values were further adjusted using the Bonferroni correction method.

**Identification of key TFs.** We defined key TFs as those expressing in a tissue-specific manner and at a relatively high level in the corresponding tissues[17]. To identify tissue specifically expressed TFs, an algorithm adapted from Kadota *et al.*[60] was employed. This algorithm considers the task of tissue-specific gene identification as an "outliner identification" problem. The main advantage of this algorithm is that objective decisions could be made because the procedure is independent of a significance level[60]. Basically, for each TF, its expression values among various tissues were collected; the tissue specific expression in certain tissue was identified as "outliners" compared to the remaining tissues. Next, considering that the key TFs should express at a relatively high level in the corresponding tissue, we further filtered out the candidate TF-tissue pairs in which the expression of the TF is low (a FPKM value of 10 was used as the cutoff) in the candidate tissue. The implementation of this algorithm is freely accessible on our website.

## References


1. Latchman, D. S. Transcription factors: an overview. *Int J Biochem Cell Biol* **29**, 1305–1312, doi:10.1016/S1357-2725(97)00085-X (1997).
2. Lee, T. I. & Young, R. A. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**, 77–137, doi:10.1146/annurev.genet.34.1.77 (2000).
3. Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626, doi:10.1038/nrg3207 (2012).
4. Weirauch, M. T. & Hughes, T. R. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. *Subcell Biochem* **52**, 25–73, doi:10.1007/978-90-481-9069-0_3 (2011).
5. Zhang, H. M. *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40**, D144–149, doi:10.1093/nar/gkr965 (2012).
6. Hsia, C. C. & McGinnis, W. Evolution of transcription factor function. *Curr Opin Genet Dev* **13**, 199–206, doi:10.1016/S0959-437X(03)00017-0 (2003).
7. Kablar, B. *et al.* MyoD and Myf-5 differentially regulate the development of limb versus trunk skeletal muscle. *Development* **124**, 4729–4738 (1997).
8. Buckingham, M. *et al.* The formation of skeletal muscle: from somite to limb. *J Anat* **202**, 59–68, doi:10.1046/j.1469-7580.2003.00139.x (2003).
9. Smale, S. T. Pioneer factors in embryonic stem cells and differentiation. *Curr Opin Genet Dev* **20**, 519–526, doi:10.1016/j.gde.2010.06.010 (2010).
10. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319, doi:10.1016/j.cell.2013.03.035 (2013).
11. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947, doi:10.1016/j.cell.2013.09.053 (2013).

12. Buganim, Y., Faddah, D. A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat Rev Genet* **14**, 427–439, doi:10.1038/nrg3473 (2013).
13. Lepoivre, C. *et al.* Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics* **14**, 914, doi:10.1186/1471-2164-14-914 (2013).
14. Zhou, L. *et al.* Linc-YY1 promotes myogenic differentiation and muscle regeneration through an interaction with the transcription factor YY1. *Nat Commun* **6**, 10026, doi:10.1038/ncomms10026 (2015).
15. Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci USA* **110**, 2876–2881, doi:10.1073/pnas.1221904110 (2013).
16. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252–263, doi:10.1038/nrg2538 (2009).
17. D'Alessio, A. C. *et al.* A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. *Stem Cell Reports* **5**, 763–775, doi:10.1016/j.stemcr.2015.09.016 (2015).
18. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* (2016).
19. Zhu, J. *et al.* Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Sci Rep* **6**, 28400, doi:10.1038/srep28400 (2016).
20. Kanamori, M. *et al.* A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun* **322**, 787–793, doi:10.1016/j.bbrc.2004.07.179 (2004).
21. Fulton, D. L. *et al.* TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* **10**, R29, doi:10.1186/gb-2009-10-3-r29 (2009).
22. Wilson, D., Charoensawan, V., Kummerfeld, S. K. & Teichmann, S. A. DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* **36**, D88–92, doi:10.1093/nar/gkm964 (2008).
23. Liu, Z. P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* (*Oxford*) **2015**, 10.1093/database/bav095 (2015).
24. Guo, A. M., Sun, K., Su, X., Wang, H. & Sun, H. YY1TargetDB: an integral information resource for Yin Yang 1 target loci. *Database (Oxford)* **2013**, bat007, doi:10.1093/database/bat007 (2013).
25. Yang, L. *et al.* TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res* **42**, D148–155, doi:10.1093/nar/gkt1087 (2014).
26. Zhang, H. M. *et al.* AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76–81, doi:10.1093/nar/gku887 (2015).
27. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447–452, doi:10.1093/nar/gku1003 (2015).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29, doi:10.1038/75556 (2000).
29. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361, doi:10.1093/nar/gkw1092 (2017).
30. Zhang, S. & Cui, W. Sox2, a key factor in the regulation of pluripotency and neural differentiation. *World J Stem Cells* **6**, 305–311, doi:10.4252/wjsc.v6.i3.305 (2014).
31. Shi, G. & Jin, Y. Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res Ther* **1**, 39, doi:10.1186/scrt39 (2010).
32. Braun, T. & Gautel, M. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat Rev Mol Cell Biol* **12**, 349–361, doi:10.1038/nrm3118 (2011).
33. Vidarsson, H. *et al.* The forkhead transcription factor Foxi1 is a master regulator of vacuolar H-ATPase proton pump subunits in the inner ear, kidney and epididymis. *PLoS One* **4**, e4471, doi:10.1371/journal.pone.0004471 (2009).
34. Romano, R. *et al.* FOXN1: A Master Regulator Gene of Thymic Epithelial Development Program. *Front Immunol* **4**, 187, doi:10.3389/fimmu.2013.00187 (2013).
35. Yamamoto, M. & Kuroiwa, A. Hoxa-11 and Hoxa-13 are involved in repression of MyoD during limb muscle development. *Dev Growth Differ* **45**, 485–498, doi:10.1111/dgd.2003.45.issue-5-6 (2003).
36. Luo, H., Zhao, X., Wan, X., Huang, S. & Wu, D. Gene microarray analysis of the lncRNA expression profile in human urothelial carcinoma of the bladder. *Int J Clin Exp Med* **7**, 1244–1254 (2014).
37. Li, T. *et al.* Expression and clinicopathological significance of the lncRNA HOXA11-AS in colorectal cancer. *Oncol Lett* **12**, 4155–4160, doi:10.3892/ol.2016.5129 (2016).
38. Wurster, A. L., Tanaka, T. & Grusby, M. J. The biology of Stat4 and Stat6. *Oncogene* **19**, 2577–2584, doi:10.1038/sj.onc.1203485 (2000).
39. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25**, 1125–1142, doi:10.1038/sj.onc.1209080 (2006).
40. Lu, L. *et al.* Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis. *EMBO J* **32**, 2575–2588, doi:10.1038/emboj.2013.182 (2013).
41. He, Y. & Casaccia-Bonnefil, P. The Yin and Yang of YY1 in the nervous system. *J Neurochem* **106**, 1493–1502, doi:10.1111/jnc.2008.106.issue-4 (2008).
42. Mielcarek, M., Gunther, S., Kruger, M. & Braun, T. VITO-1, a novel vestigial related protein is predominantly expressed in the skeletal muscle lineage. *Gene Expr Patterns* **2**, 305–310, doi:10.1016/S0925-4773(02)00386-6 (2002).
43. Chen, H. H., Maeda, T., Mullett, S. J. & Stewart, A. F. Transcription cofactor Vgl-2 is required for skeletal muscle differentiation. *Genesis* **39**, 273–279, doi:10.1002/gene.20055 (2004).
44. Maeda, T., Chapman, D. L. & Stewart, A. F. Mammalian vestigial-like 2, a cofactor of TEF-1 and MEF2 transcription factors that promotes skeletal muscle differentiation. *J Biol Chem* **277**, 48889–48898, doi:10.1074/jbc.M206858200 (2002).
45. Faucheux, C. *et al.* Vestigial like gene family expression in Xenopus: common and divergent features with other vertebrates. *Int J Dev Biol* **54**, 1375–1382, doi:10.1387/ijdb.103080cf (2010).
46. Piccand, J. *et al.* Rfx6 maintains the functional identity of adult pancreatic beta cells. *Cell Rep* **9**, 2219–2232, doi:10.1016/j.celrep.2014.11.033 (2014).
47. Hart, A. W., Mella, S., Mendrychowski, J., van Heyningen, V. & Kleinjan, D. A. The developmental regulator Pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. *PLoS One* **8**, e54173, doi:10.1371/journal.pone.0054173 (2013).
48. May, C. L. The role of Islet-1 in the endocrine pancreas: Lessons from pancreas specific Islet-1 deficient mice. *Islets* **2**, 121–123, doi:10.4161/isl.2.2.10908 (2010).
49. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, doi:10.1038/nature11247 (2012).
50. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503–510, doi:10.1038/nbt.1633 (2010).
51. Kim, J., Sturgill, D., Tran, A. D., Sinclair, D. A. & Oberdoerffer, P. Controlled DNA double-strand break induction in mice reveals post-damage transcriptome stability. *Nucleic Acids Res* **44**, e64–e64, doi:10.1093/nar/gkv1482 (2016).

52. Rodriguez-Seguel, E. *et al*. Mutually exclusive signaling signatures define the hepatic and pancreatic progenitor cell lineage divergence. *Genes Dev* **27**, 1932–1946, doi:10.1101/gad.220244.113 (2013).
53. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, D130–135, doi:10.1093/nar/gkr1079 (2012).
54. Sun, K., Zhao, Y., Wang, H. & Sun, H. Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)–application in human skeletal muscle cells. *PLoS One* **9**, e84500, doi:10.1371/journal.pone.0084500 (2014).
55. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111, doi:10.1093/bioinformatics/btp120 (2009).
56. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515, doi:10.1038/nbt.1621 (2010).
57. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**, R22, doi:10.1186/gb-2011-12-3-r22 (2011).
58. Dillies, M. A. *et al*. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14**, 671–683, doi:10.1093/bib/bbs046 (2013).
59. Uhlen, M. *et al*. Transcriptomics resources of human tissues and organs. *Mol Syst Biol* **12**, 862–862, doi:10.15252/msb.20155865 (2016).
60. Kadota, K. *et al*. Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiol Genomics* **12**, 251–259, doi:10.1152/physiolgenomics.00153.2002 (2003).


## Acknowledgements

We thank Mr. Tao Liu for his technical support. The work described in this paper was substantially supported by the General Research Funds (GRF) from the Research Grants Council (RGC) (project codes: 14133016, 14100415, 14116014, 476113 to H.W. and 14102315, 14113514, 473713 to H.S.); Collaborative Research Fund (CRF) (project code: C6015-14G to H.W. and H.S.) from the Hong Kong Special Administrative Region, China; and one grant from the Ministry of Science and Technology of China (2014CB964700 to H.W.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Contributions

Conceived of the study: K.S. Analyzed the data: K.S. Contributed analysis tools: H.W. H.S. Wrote the manuscript: K.S. H.W. H.S.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-02404-w

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017