

## RESEARCH ARTICLE

# ZokorDB: tissue specific regulatory network annotation for non-coding elements of plateau zokor

Jingxue Xin<sup>1,6,7,†</sup>, Junjun Hao<sup>2,†</sup>, Lang Chen<sup>1</sup>, Tao Zhang<sup>3</sup>, Lei Li<sup>1,5,7</sup>, Luonan Chen<sup>3,5</sup>, Wenmin Zhao<sup>4</sup>, Xuemei Lu<sup>2,5</sup>, Peng Shi<sup>2,5,\*</sup>, Yong Wang<sup>1,5,7,\*</sup>

<sup>1</sup> CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

<sup>2</sup> State Key Laboratory of Genetic Resources and Evolution, Laboratory of Evolutionary and Functional Genomics, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

<sup>3</sup> Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

<sup>4</sup> Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

<sup>5</sup> Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

<sup>6</sup> Department of Statistics, Department of Biomedical Data Science, Bio-X Program, Stanford University, Stanford, CA 94305, USA

<sup>7</sup> University of Chinese Academy of Sciences, Beijing 100049, China

\* Correspondence: ywang@amss.ac.cn, ship@mail.kiz.ac.cn

Received September 4, 2019; Revised December 16, 2019; Accepted December 23, 2019

**Background:** Plateau zokor inhabits in sealed burrows from 2,000 to 4,200 meters at Qinghai-Tibet Plateau. This extreme living environment makes it a great model to study animal adaptation to hypoxia, low temperature, and high carbon dioxide concentration.

**Methods:** We provide an integrated resource, ZokorDB, for tissue specific regulatory network annotation for zokor. ZokorDB is based on a high-quality draft genome of a plateau zokor at 3,300 m and its transcriptional profiles in brain, heart, liver, kidney, and lung. The conserved non-coding elements of zokor are annotated by their nearest genes and upstream transcriptional factor motif binding sites.

**Results:** ZokorDB provides a general draft gene regulatory network (GRN), *i.e.*, potential transcription factor (TF) binds to non-coding regulatory elements and regulates the expression of target genes (TG). Furthermore, we refined the GRN by incorporating matched RNA-seq and DNase-seq data from mouse ENCODE project and reconstructed five tissue-specific regulatory networks.

**Conclusions:** A web-based, open-access database is developed for easily searching, visualizing, and downloading the annotation and data. The pipeline of non-coding region annotation for zokor will be useful for other non-model species. ZokorDB is free available at the website ([bigd.big.ac.cn/zokordb/](http://bigd.big.ac.cn/zokordb/)).

**Keywords:** tissue specific regulatory network; non-coding element; plateau zokor; non-model species

**Author summary:** ZokorDB is a non-coding regulatory element annotation database providing abundant source of materials for studying DNA sequence, gene expression and regulation, as well as tissue-specific gene regulatory network for zokor, which is a non-model organism famous for adaptation to extreme environment and evolution study. The integrated pipeline to systematically annotate genome with non-coding region and gene regulatory network for zokor are helpful for other non-model species.

<sup>†</sup> These authors contributed equally to this work.

## INTRODUCTION

Plateau zokors (*Eospalax baileyi*) live in sealed burrows at elevations of 2,000–4,200 m at Qinghai-Tibet Plateau [1]. In addition to the extreme high altitude, they also survive in dark subterranean caves with high carbon dioxide (CO<sub>2</sub>) concentration. Plateau zokors live in a depth of 70–250 cm burrows about 5 months from November to March in order to deal with low temperature [2]. This environment aggravates the low oxygen and high CO<sub>2</sub> environment since the content of oxygen is 17.04%–18.43% and carbon dioxide is 0.22%–1.46% respectively in 18 cm depth burrows [3]. Surprisingly, the body temperature of plateau zokor is on average 34.72 °C and varies slightly when the ambient temperature ranges from 0 to 27 °C [3]. Their basal metabolic rates (BMR) are significantly higher than other fossorial rodents [4]. The large energy consumption is also a physiological characteristic for adaptation to cold conditions. Taken together, these facts exhibit that plateau zokor have morphological, physiological, and behavioral adaptations for living in the high altitude underground burrows [5,6]. It is a successful example for study of adaptation to extreme environment.

Rapid development of high-throughput sequencing technologies sheds light on genome assembling of diverse species. For plateau zokor, a high-quality draft genome of one individual at 3,300 m was assembled [7]. They also identified hundreds of positively selected genes by comparative genomic study with other species. Among the positive selection regions, both coding genes and non-coding sequences are expected to be functional. Since nonsynonymous mutations change sequences of amino acid, their mechanism is straightforward. However, the function annotation of conserved non-coding sequences is more complex. They have potential to be regulatory elements [8]. But for the regulatory elements, the cell types they have function, their regulating genes, related signaling pathways, and biological mechanisms remain unknown. Recently, human ENCODE [9] and mouse ENCODE project [10] have identified a variety of functional regions for human and mouse, which provides new perceptions of chromatin organization and gene regulation [9]. Leveraging the public epigenomic data acquired from human and mouse to study regulatory elements for a variety of non-model animals is potentially helpful for understanding evolution.

As far as we know, there is no relevant database available for zokor up to now. This situation prompted us to construct a web-based, open-access database for easily searching, visualizing, and downloading data. Our database contains the basic zokor genome and annotation information, as well as conserved non-coding elements

compared with mouse, a close relative of zokor in the phylogenetic tree. For each conserved element, transcriptional factor motif binding sites (MFBS) and its nearest target gene are listed based on mouse genome knowledge. Furthermore, a refined GRN is reconstructed through incorporating matched RNA-seq and DNase-seq data from mouse ENCODE project. We also construct five tissue-specific regulatory networks by taking advantage of gene expression data from brain, heart, liver, kidney, and lung. All data mentioned above are collected in ZokorDB. Above valuable data sources will facilitate researchers gain insight into context-specific mechanism of adaptation to extreme environment.

## DATABASE CONSTRUCTION METHODS

### Data sources

We integrated several data resources in our database. Zokor genome and gene annotation data is based on a high-quality draft genome assembled from one individual at 3,300 m [7]. Gene expression data of five tissues (brain, heart, liver, kidney, and lung) are collected from a zokor living in Qinglin, China at 2,800 m. We used Cufflinks [11] to quantify gene expression by FPKM value, then we chose FPKM > 0.05 as a cutoff for genes expressed in that tissue. Table 1 demonstrates the number of tissue-specific genes for each tissue.

**Table 1** The number of tissue-specific genes in five tissues

Tissue	Number of tissue-specific genes
Brain	14,764
Heart	13,392
Kidney	13,849
Liver	13,045
Lung	14,410

Sample matched RNA-seq and DNase-seq data of 25 tissues/cell types of mouse from mouse ENCODE project are obtained from [12]. DNase-seq signals are quantified on promoters and enhancers of mouse. Motifs of 557 TFs collected by [12] are also downloaded. This data facilitates constructing a refined GRN.

### Genome assembly and gene annotation

To assemble the plateau zokor genome, we first used Allpaths-LG [13] for *de novo* genome assembly. Then we applied BAUM [14] to improve the draft assembly based on the result of Allpaths-LG. This step is to reduce the uncertainty caused by repetitive elements.

In detail, we assembled the genome of the plateau

zokor using BAUM [14] and Allpaths-LG [13]. BAUM is a pipeline that can perform reference-assisted genome assembly and improve assemblies from other assemblers iteratively. In the assembly of the plateau zokor genome, we first used Allpaths-LG, a popular assembly software based on *de Bruijn* graph, for *de novo* genome assembly. Although the *de Bruijn* graph frameworks are recognized efficient nowadays, it could hardly resolve the uncertainty caused by the widespread repetitive elements. Therefore, we applied BAUM to improve the draft assembly result of Allpaths-LG. BAUM used unique mapping reads for scaffolding and contig-extension, which ensured its sustainability during iterations. As for read mapping, we used SEME to control the sensitivity and specificity and to obtain unique mapping reads. After scaffolding and contig-extension, we used a robust method to decide whether two adjacent overlapped contigs should be merged.

After final iteration of building scaffolds with contig-extension and contig merging, the genome size of our assembly achieved 2.63 Gb, with contig N50 and scaffold N50 being 311 Kb and 2.78 Mb respectively. The GC content of a genome is indispensable for genome analysis. In our assembly, the mean and standard deviation of GC content was 41.2% and 4.2%. The continuity of an assembly is also critical for the follow-up genome analysis. We calculated the gap number in each scaffold and the size of each gap. The average number of gaps in a scaffold is 27.5. The mean and standard deviation of the gap sizes inside our assembly are 853.9 and 1219.6 respectively. Comparing with the previous *de novo* assembly identified 208,451 non-redundant transcripts with a contig N50 only 2,433 base pairs [7], our genome assembly achieved significant larger contigs.

With the assembled genome of the plateau zokor, we applied three gene prediction software, Genscan [15], Augustus [16] and GlimmerHMM [16]. Then we used the cleaned reads from RNA-seq data of 5 tissues to refine the transcriptome.

### Cis-regulatory element annotation

The pipeline of cis-regulatory element annotation procedures is shown in Fig. 1A. First we aligned zokor genome, including 38,364 scaffolds, to mouse (mm9) with GMAP software [17]. 22,839 fragments of zokor genome conserved in mouse are identified. These conserved regions are defined as cis-regulatory elements (RE). For identification of high-quality REs, we implemented open chromatin regions, derived from 25 DNase-seq samples from ENCODE, as candidate RE set. Among 22,839 conserved fragments, 4,183 ones overlap with above defined candidate REs of mouse.

### TF-RE-TG triplet model

Our aim is to model how a TF will regulate a TG via conserved regulatory element. Given one TG, we start by enumerating all the nearby elements  $e_0, e_1, e_2, \dots, e_m$  around TG to serve as cis-regulatory elements. Those elements are grouped into two classes by their distance to the TG. Those include the promoter of TG (denoted by  $e_0$ ) and  $m$  potential enhancers ( $e_1, e_2, \dots, e_m$ ). We assume that TF may regulate this TG's expression by physical binding to those cis-regulatory elements. We then use the following information to extract the TF-RE-TG triplet and infer if TG is likely to be regulated by this TF via REs. The evidence we considered in our model are (Fig.1B):

1. TF binds to TG's cis-elements. We scan TG's cis-regulatory elements for all positions with substantial similarity to TF's sequence motif or position weight matrix (PWM). Considering the proximity of those binding sites (we use the proximity of binding sites to its center for union peak), for cis-element  $e_i$ , we derive a binding strength  $\{B_i, i = 0, 1, \dots, m\}$ , here  $B_i$  denotes the strength that TF regulates TG via cis-element  $e_i$ . We assume that the association strength of TF-TG pair  $r$  via cis-element  $e_i$  is a weighted sum of intensities of all the motif binding sites:

$$B_i = \sum_{l=1}^{k_i} PWM\ Score_{il} e^{-\frac{d_{il}}{d_{i0}}},$$

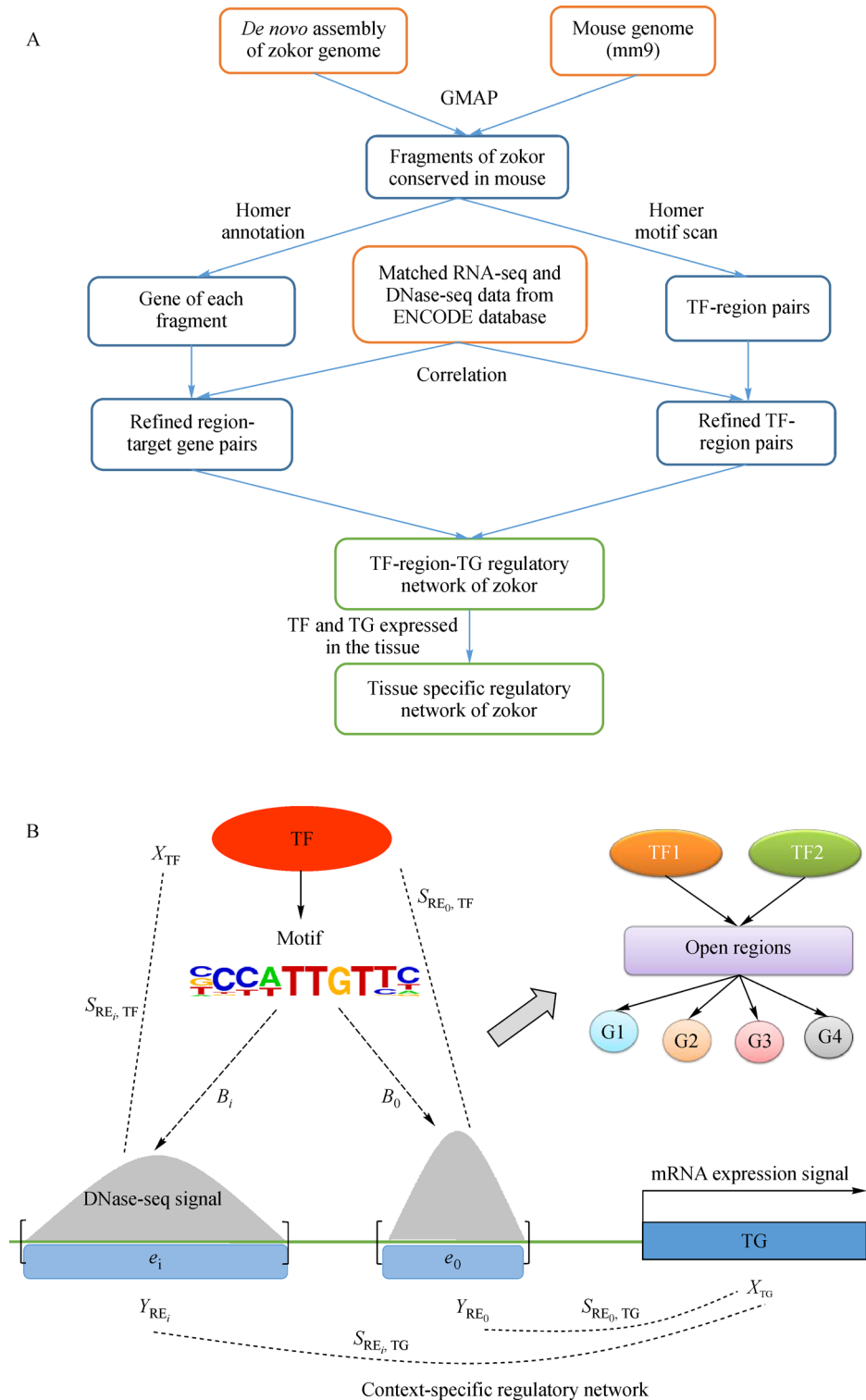
where  $PWM\ Score_{il}$  is the strength of the  $l$ -th binding site in RE  $e_i$  for TF-TG pair  $r$ . For enhancer,  $d_{il}$  is the distance between the center of RE and the  $l$ -th binding site.  $d_{i0}$  is a constant. We set  $d_{i0} = 500$  bp,  $i \neq 0$  for union peaks in our implement.

2. The expression of TF and TG. The gene expression levels of TF and TG, i.e.,  $X_{TF}, X_{TG}$ , in this condition or cell type can be measured by RNA-seq data. We used Cufflinks [11] to quantify gene expression by FPKM value. Selecting TFs and target genes expressed in a cell type with the criteria  $X_{TF} > 0$ , and  $X_{TG} > 0$ .

Following this procedure, we will predict a set of TF-RE-TG triplet. Pooling all the triplets together, we will have a TF-RE-TG network, where TF, RE, TG are nodes and their regulations are derived from triplet.

### Construction of gene regulatory network

For the 22,839 cis-regulatory elements, we identified the TFs binding to them and their regulatory target genes. The whole GRN construction method is demonstrated in Fig. 1A, B. First, we utilized Homer [18] to scan motif binding sites on each conserved RE. Totally 47,514,939 motif binding sites are obtained. Second, we assigned the nearest gene to each RE based on genomic positions.



**Figure 1. The pipeline of zokorDB construction for tissue specific regulatory network annotation for non-coding element of plateau zokor.** (A) The whole procedures include genome assembly and gene annotation, cis-regulatory element annotation, and construction of gene regulatory network. (B) The schematic diagram of constructing context specific gene regulatory network. TF-RE-TG triplet is modeled by first calculating the strength of TF binding to TG's cis-elements. Then TF-RE, and RE-TG relations are refined by computing the SCC across tissues/cell types from ENCODE database.

Thus, after two steps calculation we obtained a draft GRN of potential TFs binding to conserved REs and regulate the expression of target genes, *i.e.*, TF-RE-TG triplets. However, this draft network may include many false positive nodes and edges since motif occurrence does not necessarily indicate TF binding and conserved regions may not regulate the nearest genes. Therefore, we incorporated 25 matched gene expression (RNA-seq) and chromatin accessibility data (DNase-seq) collected by [12]. Among the 22,839 conserved REs, 4,183 ones overlap with promoters/enhancers of mouse. Then we calculated spearman correlation coefficients (SCC) between chromatin accessibility of REs and expression level (FPKM) of corresponding genes across 25 matched samples. Let  $X_{TF,c}$  and  $X_{TG,c}$  be the gene expression levels of TF and TG in cell-type  $c$ .  $Y_{RE,c}$  represents the openness score of the regulatory element (RE) in celltype  $c$ . Openness score is quantified the openness score for the RE by a simple fold change score, which computes the enrichment of read counts in peak by comparing with a large background region. Then we convert  $X_{TG,c}$  and  $Y_{RE,c}$  to ranks  $rgX_{TG,c}$  and  $rgY_{RE,c}$ , and define the SCC of one RE-TG pair across all cell-types as follows.

$$S_{RE,TG} = \frac{\sum_{i=1}^C (rgX_{TG,i} - \overline{rgX_{TG}}) - (rgY_{RE,i} - \overline{rgY_{RE}})}{\sqrt{\sum_{i=1}^C (rgX_{TG,i} - \overline{rgX_{TG}})^2 (rgY_{RE,i} - \overline{rgY_{RE}})^2}},$$

where  $\overline{rgX_{TG}}$  and  $\overline{rgY_{RE}}$  are the average of expression level of TG and openness score of RE.  $C$  is the total number of cell types with matched RNA-seq and DNase-seq samples. Chosen  $S_{RE,TG} > 0$ , which indicates cis-regulatory elements and target genes have positive correlations, as a loose threshold, we obtained 2,170 candidate elements. Next, we selected candidate TFs binding to certain cis-regulatory elements through motif binding occurrence. Again, we computed SCC between openness score of candidate REs and expression level of corresponding TFs identified by MFBS as following formulation shows.

$$S_{RE,TF} = \frac{\sum_{i=1}^C (rgX_{TF,i} - \overline{rgX_{TF}}) (rgY_{RE,i} - \overline{rgY_{RE}})}{\sqrt{\sum_{i=1}^C (rgX_{TF,i} - \overline{rgX_{TF}})^2 (rgY_{RE,i} - \overline{rgY_{RE}})^2}}.$$

RE and TF pairs with SCC  $p$ -value  $< 0.05$  are selected.

We obtained a refined GRN with 225,747 TF-RE-TG triplet relationships in total. Finally, according to tissue-specific genes (Refer to “Data sources”), we selected TF-RE-TG triplets for each tissue if the TF and TG are expressed in that tissue. Table 2 shows the number of triplets identified in every tissue.

We downloaded the TF-TG regulation pairs from RegNetwork database [19] and used it as approximate gold standard positive set. After two steps computing SCC of RE-TG and RE-TF, we could identify 93,786 TF-TG pairs in total. Among them, 718 are overlapped in the database, *i.e.*, 0.77% is true. Then, we calculated the positive rate in the background distribution which was randomly selected 93,786 TF-TG pairs from TF-TG relations only predicted by motif occurrence in the conserved regions nearest a gene, *i.e.*, candidate TF-TG pairs before filter by SCC. The average positive rate is 0.43%, which is lower ( $\sim 1.8$  fold) than the rate predicted by our model.

## Database implementation

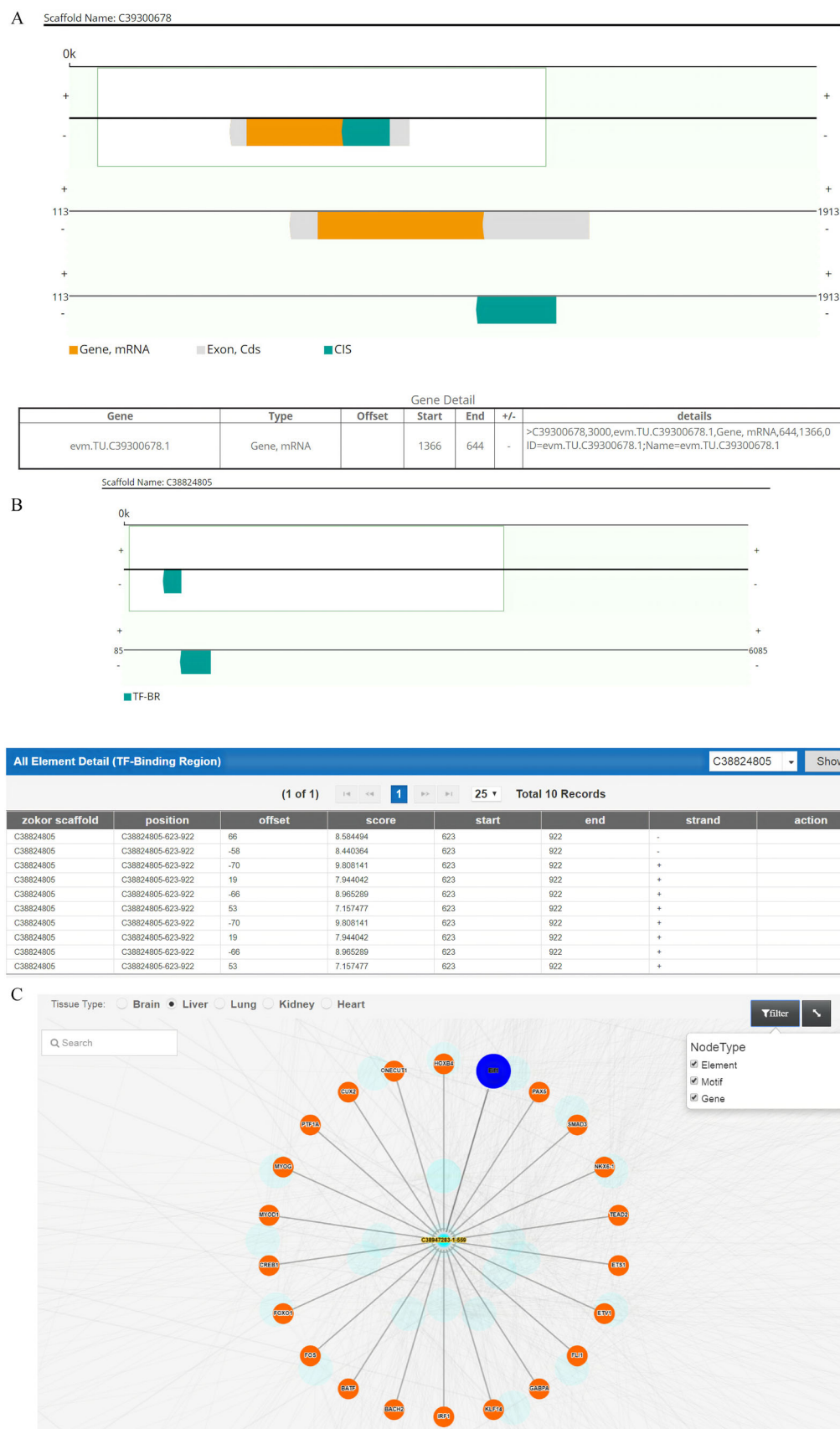
ZokorDB web system is implemented in JAVA/JSF framework. The zokor genome scaffolds, gene annotation information, REs, motif binding sites, and their associated TFs/TGs are listed in tab-delimited text files and stored into MySQL database. All data can be directly downloaded from our website. Also, the database supports searching and browsing zokor genome scaffolds, genes, and REs through genome browser. Cytoscape Web [20] is implemented when visualizing networks between TFs, REs, and TGs.

## DATABASE USAGE AND ACCESS

The zokor database can be accessed through a web interface, where three main functions: genome browser, cis-regulatory elements, and networks are implemented. In genome browser, users can easily browse a certain gene and scaffold in the zokor genome (Fig. 2A). The genes and elements are shown in different colors in double strands. And the green box can be adjusted for the neighbor positions. Furthermore, cis-REs conserved between zokor and mouse are listed as a text format table, where one can simply find their associated motif

**Table 2 Information of high-quality tissue-specific regulatory network**

Tissue	Number of TFs	Number of REs	Number of TGs	Number of TF-RE-TG triplets
Brain	517	2,967	877	117,645
Heart	418	2,855	825	100,858
Kidney	446	2,948	858	111,869
Liver	387	2,831	814	93,530
Lung	452	2,964	863	225,747



**Figure 2. Database usage and accession.** (A) Genes and cis-regulatory elements, represented by different color bars, are shown in genome browser. When clicking the colored bars representing genes, exons, and CIS, the below table indicates detailed description of the selected region. The green box in the first track can be moved along genomic position, and its corresponding region is revealed in double strands by the below track. (B) Information of cis-RE. cis-REs conserved between zokor and mouse are listed as a text format table below, where one can simply find their associated motif binding sites on the RE. Also, the RE is visualized in the above genome browser. By clicking the green RE, users can easily obtain detailed information of this RE. (C) Visualization of subnetwork of certain target gene, RE, and TFs. For example, when searching Elf1 as a target gene in liver, the subnetwork shows Elf1's RE C3884283-1-559, and the predicted binding TFs, *i.e.*, Hoxb4, Pax4, and Smad3, etc.

binding sites on the RE. Also, the RE is visualized in the genome browser (Fig. 2B). The data incorporates chromosomes, zokor elements, mapped mouse elements and annotations, motif binding sites, TF motif matched relations, and genes, which can be easily searched, listed, and downloaded. Furthermore, five tissue-specific gene regulatory networks between transcription factors (motifs), elements, and target genes are displayed through a network format using Cytoscape Web plugin, where users could flexibly change and search a gene in the whole tissue-specific regulatory network. Users can easily search a certain gene in a specific tissue, for example, *Elf1* as a target gene in liver, then the network shows *Elf1*'s RE C38947283-1-559, and the predicted binding TFs, *i.e.*, *Hoxb4*, *Pax4*, and *Smad3*, etc. (Fig. 2C). Finally, all relevant text format data files can be freely downloaded from "Download" option. One can also view and search the element, conserved region, gene annotation, motif, scaffold, and gene data through "View Data" option.

## DISCUSSION AND CONCLUSIONS

Plateau zokor, a special species from Qinghai-Tibet Plateau, is an excellent example to study animals adapt to extreme environment and evolution. As the first database for plateau zokor, ZokorDB collects huge amount of data including genome, gene annotations, and conserved non-coding element annotations (motif binding sites and nearest target genes). Through borrowing conserved information from mouse, a close relative of zokor in the phylogenetic tree, ZokorDB provides abundant source of materials for studying DNA sequence, gene expression and regulation, as well as tissue-specific GRN for zokor, which may further facilitate genotype-phenotype understanding of adaptation and evolution. Furthermore, we plan to collect more samples from various altitude levels for SNPs and gene expression data in ZokorDB.

For constructing gene regulatory network, we do not have gold standard TF-TG relations for tissue specific context information in zokor. Thus, it is difficult to quantify the false positive rate. However, we found that conservation between zokor and mouse, element activity defined by DNase-seq in ENCODE database, correlation between RE and TG, and TF and RE, can reduce the false positive rate. The basic assumption is the biological conservation, including DNA sequence, regulatory elements (RE), transcription factors (TF), target genes (TG), and the regulatory relationships between TF, RE, and TG, within similar species, *i.e.*, zokor and mouse. By conservation between zokor and mouse, we obtained totally 22,839 conserved regions. After we used mouse DNase-seq peaks to filter, only 4,183 regions left, which

reduced one fifth of the total regions. Not only the biological knowledge can reduce candidate regulatory regions, it is also useful in RE-TG, and TF-RE prediction. If we only use sequence information based on motif scan at each element, 47,514,939 motif-element pairs are identified. Then we used spearman correlation between RE and TG, TF and RE, only 225,747 TF-RE-TG triplets are identified. As we can see, filter steps can efficiently reduce large amount of candidate regulatory relations.

One advantage of our database is to provide potential non-coding regulatory regions for a non-model animal zokor by referring annotation information of mouse. In fact, human ENCODE, mouse ENCODE, and mod ENCODE projects [21] have provided abundant information about genomic functional elements in the model organisms human, mouse, worm, and fly. Based on DNA sequence conservation, we can take advantages of the information to boost researches on other non-model species. Through GRN a more detailed picture of transcription factors regulating target genes by means of cis-regulatory elements may offer clear mechanisms of biological processes. Although in nature species are diverse, and regulatory mechanisms are different from various animals, genomic functional regions of model organisms could still help us gain new insight into part of genomic organization and gene regulation of a variety of species. Our procedures to integrate genomic, transcriptomic, and even epigenomic data and tissue specific regulatory network annotation methods to construct ZokorDB set up an example for other non-model species. We expect more database implementation of other species under the trend that more and more *de novo* assembled genomes are available.

## ACKNOWLEDGEMENTS

ZokorDB is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13000000). The authors are also supported by the National Natural Science Foundation of China (NSFC) (Nos. 11871463, 11871462, 61671444 and 61621003). We thank all the lab members for discussions on data collection, genome alignment, annotation, GRN reconstruction. We thank Dr. Yilei Wu and his group for help on database design and management.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Jingxue Xin, Junjun Hao, Lang Chen, Tao Zhang, Lei Li, Luonan Chen, Wenmin Zhao, Xuemei Lu, Peng Shi and Yong Wang declare that they have no conflicts of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors

## REFERENCES

1. Fan, N. and Shi, Y. (1982) A revision of the zokors of subgenus *Eospalax*. *Acta Theriol. Sin.*, 2, 183–199, In Chinese

2. Fan, N. and Gu, S. (1981) The structure of the tunnel system of the Chinese zokor. *Acta Theriol. Sin.*, 1, 67–71, In Chinese
3. Zeng, J., Wang, Z. and Shi, Z. (1984) Metabolic characteristics and some physiological parameters of the mole rat (*Myospalax baileyi*) in an alpine area. *Acta Biol. Plat. Sin.*, 3, 163–171
4. McNab, B. K. (1984) The metabolism of fossorial rodents: a study of convergence. *Ecology*, 47, 712–733
5. Reichman, O., Smith, S.C. (1990) Burrows and burrowing behavior by mammals. *Curr. Mammal.*, 2, 197–244
6. Zhang, Y. M. and Liu, J. (2003) Effects of plateau zokors (*Myospalax fontanierii*) on plant community and soil in an alpine meadow. *J. Mammal.*, 84, 644–651
7. Shao, Y., Li, J. X., Ge, R. L., Zhong, L., Irwin, D. M., Murphy, R. W. and Zhang, Y. P. (2015) Genetic adaptations of the plateau zokor in high-elevation burrows. *Sci. Rep.*, 5, 17262
8. Hardison, R. C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, 16, 369–372
9. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74
10. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515, 355–364
11. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511–515
12. Duren, Z., Chen, X., Jiang, R., Wang, Y. and Wong, W. H. (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA*, 114, E4914–E4923
13. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C. and Jaffe, D. B. (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.*, 18, 810–820
14. He, Y. X., Qi, X. B., Ouzhuluobu, , Liu, S., Li, J., Zhang, H., Baimakangzhuo, Bai, C., Zheng, W., Guo, Y., *et al.* (2018) Blunted nitric oxide regulation in Tibetans under high-altitude hypoxia. *Natl. Sci. Rev.*, 5, 516–529
15. Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268, 78–94
16. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.*, 32, W309–W312
17. Wu, T. D. and Watanabe, C. K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859–1875
18. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38, 576–589
19. Liu, Z. P., Wu, C., Miao, H., Wu, H. (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford)* 2015, bav095
20. Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q. and Bader, G. D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26, 2347–2348
21. The modENCODE Consortium. (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 33, 1787–1797