AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors

Hui Hu [1,2,†], **Ya-Ru Miao**[1,2,†], **Long-Hao Jia**[1], **Qing-Yang Yu**[1], **Qiong Zhang**[1] **and An-Yuan Guo**[1,2,*]

[1]Department of Bioinformatics and Systems Biology, Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, PR China and [2]Huazhong University of Science and Technology Ezhou Industrial Technology Research Institute, Ezhou, Hubei 436044, PR China

ABSTRACT

The Animal Transcription Factor DataBase (AnimalTFDB) is a resource aimed to provide the most comprehensive and accurate information for animal transcription factors (TFs) and cofactors. The AnimalTFDB has been maintained and updated for seven years and we will continue to improve it. Recently, we updated the AnimalTFDB to version 3.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/) with more data and functions to improve it. AnimalTFDB contains 125,135 TF genes and 80,060 transcription cofactor genes from 97 animal genomes. Besides the expansion in data quantity, some new features and functions have been added. These new features are: (i) more accurate TF family assignment rules; (ii) classification of transcription cofactors; (iii) TF binding sites information; (iv) the GWAS phenotype related information of human TFs; (v) TF expressions in 22 animal species; (vi) a TF binding site prediction tool to identify potential binding TFs for nucleotide sequences; (vii) a separate human TF database web interface (HumanTFDB) was designed for better utilizing the human TFs. The new version of AnimalTFDB provides a comprehensive annotation and classification of TFs and cofactors, and will be a useful resource for studies of TF and transcription regulation.

INTRODUCTION

Transcription factors (TFs) are special proteins with sequence specific DNA-binding domains (DBDs) that bind target DNA to promote or suppress gene transcription (1) and play key roles in all kinds of biological processes (2). Accurate identification of TFs is the basis for studying the function of TFs. There are several databases for TFs, for example, the current most comprehensive plant TFs were well-defined and established by the PlantTFDB databases (3,4). For animal TF databases, although there are some databases such as The Human Transcription Factors database (5) and REGULATOR (6), which focus on single genome and 77 metazoan species, respectively. Our AnimalTFDB is the first and most comprehensive animal TF database including classification and annotation of genome-wide TFs and cofactors. The AnimalTFDB was firstly built in 2011 (7) and in 2015 it was updated to AnimalTFDB v2.0 (8) with more species and annotations. It has been accessed by millions, cited by hundreds and widely used for the functional studies of animal TFs and TF prediction.

As one of the major regulator types in biological processes or diseases, TFs have been well studied in many aspects, such as functions and regulatory mechanism (9), evolutionary analysis (10), drug targets analysis (11,12), disease or phenotype of TFs (13–15), TF regulatory networks (16), TF target prediction (17), and TF-related single nucleotide polymorphisms (SNPs) (18). The regulatory networks and functional interactions between TFs and target binding sites play key roles in cancers and other diseases (19,20). The DNA binding sites of hundreds of vertebrate TFs have been determined and collected by several databases. HOCOMOCO contains transcription factor binding site (TFBS) models of several hundreds of human and mouse TFs (21). Besides, TRANSFAC and JASPAR (22) embrace TFBS of several animal species, and Cis-BP database (23) contains 6559 TFBS of 340 species. These resources laid the foundation of regulatory research for TFs. Since the TFBS is a short DNA sequence, genomic variants about SNPs and mutations will affect the TF binding and

---

*To whom correspondence should be addressed. Tel: +86 27 8779 3177; Fax: +86 27 8779 3177; Email: guoay@hust.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

regulation. Genome-wide association studies (GWAS) [24] identified many phenotype related variants genome widely which may be a useful resource to explore TF related variants and phenotypes [25,26].

In the past 4 years, the number of species in Ensembl database has increased by doubled. To meet the urgent demand of data-driven research, we upgraded AnimalTFDB to version 3.0, which covers more species, more TFs and cofactors with the latest annotation and new functions. In addition, TF related GWAS phenotype and TFBS information were integrated, as well as a TFBS prediction tool was provided. The new AnimalTFDB3.0 will be a useful resource for transcriptional regulation and comparative genomic research.

DATA SOURCE AND SUMMARY

All protein sequences of 97 animal genomes were downloaded from the Ensembl database (version 92) [27]. In AnimalTFDB3.0, we identified 125,135 TFs and 80,060 transcription cofactors in 97 animal species (Table 1) by using the improved prediction pipeline as described in next section. There are 1665 TFs (7.34% in protein-coding genes) and 1025 cofactors (4.52%) in human. The numbers of TFs and cofactors in 97 species were shown in Supplementary Table S1. Statistical data shows that TFs account for 5–8.5% of protein-coding genes in vertebrates, while this data reduced to 2.9% ~4% in other eukaryotic organisms (Supplementary Table S1). The 'Species' page was shown in Figure 1A. We collected a large amount of annotations from the NCBI Entrez Gene and Ensembl databases, including basic information, gene phenotypes, homologous genes, and Gene Ontology (GO). We acquired protein-protein interaction (PPI) data from BioGRID [28] and HPRD [29]. The protein functional domains were predicted by the PfamScan for all protein domain models in Pfam database, while the signaling pathway information was obtained from BioCarta (https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) and KEGG databases.

Next, 4257 gene-SNP pairs (2469 for TFs and 1796 for transcription cofactors) with the corresponding GWAS phenotypes were gathered from the latest GWAS Catalog [25] and dbSNP (release 144) [30]. Furthermore, TFBS for 18,952 TFs of 51 species were integrated from HOCOMOCO [21], TRANSFAC, JASPAR [22] and CIS-BP databases. In addition, we collected TF expression from TCGA [31], EMBL-EBI Expression Atlas [32], RNA-seq data published by Li *et al.* [33] and bgee database [34] of 22 animal species as well as the human protein expressions from Human Protein Map [35]. In AnimalTFDB3.0, the data amount and types are more comprehensive compared with the previous two versions (Table 1).

IMPROVED CONTENT AND NEW FEATURES

Animal TF family and assignment rules

TFs are typically characterized and classified into specific families by their conserved DBDs. We adjusted the TF families based on the AnimalTFDB2.0 by extracting several new families from the 'Others' groups or merge some families after systematically literature review. The five new TF families

extracted from the 'Others' group of previous version were zf-CCCH, LRRFIP, DACH, GCFC and CSRNP. In addition, we moved the CEP-1 family into the 'Others' group because it has only one TF and also merged the C/EBP and TF_bZIP 2 families into TF_bZIP family because of them with the same DBD [36]. Finally, we obtained 73 TF families in AnimalTFDB3.0 including an 'Others' group contained orphan TFs.

We set up three rules to classify a TF into its correct family. First, if a superfamily has several families, we classified the TFs based on the family specific domain. For example, the zf-C2H2 superfamily includes two families: zf-C2H2 and ZBTB. Proteins containing both zf-C2H2 and ZBTB domains were assigned into the ZBTB family, while proteins with only zf_C2H2 domain were classified into the zf-C2H2 family. The second rule is that if a TF has multiple unrelated DBDs, we will categorize it into the family with the smallest E value in DBD prediction. The third rule is that some proteins were predicted by some DBDs but they were annotated as enzymes based on their functional domains and functions, we removed them by their enzyme related domains. For example, we found some acetyltransferases were also predicted a zf-C2HC domain, so they were removed by the prediction result of acetyltransferase domain MOZ_SAS.

TF prediction pipeline

Based on the TF family and classification rules, we built the TF prediction pipeline. The Hidden Markov Model (HMM) profiles for DBDs of 58 TF families were downloaded from the latest Pfam database (version 31.0) [37] and 14 TF families were reconstructed based on the DBD sequences from classical species (human, mouse, zebrafish and fly) by ourselves with HMMER (v3.1b2) [38]. The self-build HMM files of 14 TF families are downloadable in the 'Download' or 'Document' page. Next, we ran the hmmsearch program in HMMER package to search all the protein sequences against all DBD HMM profiles to predict TFs in each species. To improve the accuracy of prediction result, we set different E-value thresholds for different families (Supplementary Table S2 and online document page) based on our manual curation rather than using a fixed cutoff. For instance, E-value 1e–3 for zf-C2H2 domain while 1e–20 for zf-CCCH. In addition, orphan TFs with only one member in their families and reported as TFs by literature were categorized into the 'Others' group.

Identification of transcription cofactors and their family rules

Here, we defined transcription cofactors are proteins that can modify chromatin status or interact with TFs to activate or repress the transcription of genes. In AnimalTFDB3.0, the chromatin remodeling factors were merged into transcription cofactors. Same as the version 2.0, we collected the human transcription cofactors from Tcof-DB v2 database [39] and GO database according to the related GO terms. Finally, we obtained 1,025 transcription cofactors in human after manual curation and removing redundant genes. Cofactors in the other 96 species were identified by performing mutually best-hit BLAST between each of them and human with *E*-value ≤1e–4, coverage ≥50% and identity ≥30%.
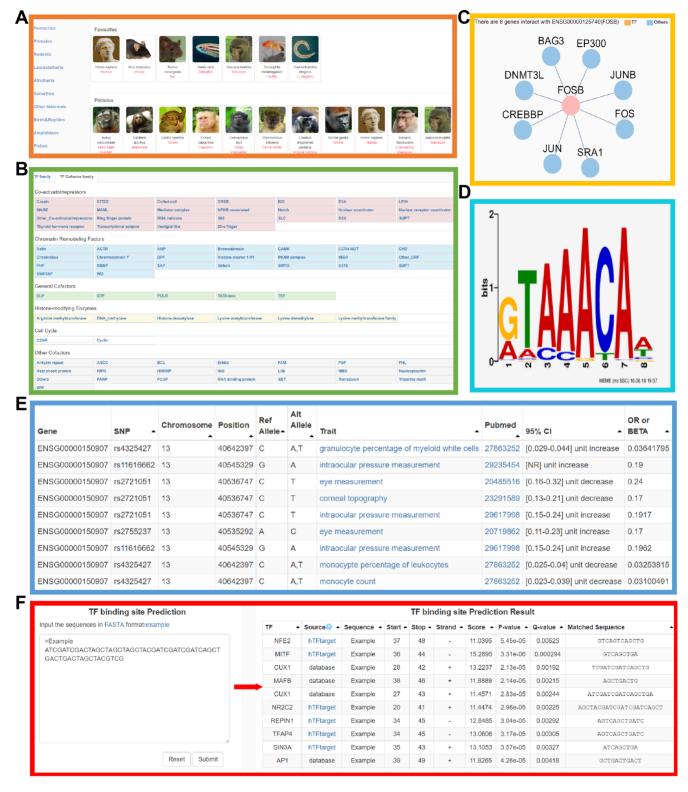
Figure 1. New features of AnimalTFDB3.0. (**A**) Part of the 'Species' page. (**B**) The families and categories of transcription cofactors. (**C**) An example for PPI network. (**D**) An example of TFBS information. (**E**) The GWAS phenotype related information of human TFs. (**F**) The TFBS prediction server and the example of prediction result.

Table 1. Data summary in AnimalTFDB3.0 database

AnimalTFDB	Version 1.0	Version 2.0	Version 3.0
Species	50	65	97
TF families	72	70	73
TF genes	52,722	72,336	125,135
Cofactor genes	9066	21,053	80,060
CRFs genes	3476	6502	Merged into cofactors
Cofactor families	0	0	83
Species with expression data	0	9	22
Phenotype	No	Yes	Yes
DBDs WebLogo	No	Yes	Yes
TF prediction server	No	Yes	Yes
BLAST search	No	Yes	Yes
PPI network	No	No	Yes
GWAS	No	No	Yes
TFBS	No	No	Yes
TFBS prediction server	No	No	Yes

Transcription cofactors were divided into 83 families and the following five major categories according to their protein families and functions (Figure 1B). Genes in the 'Coactivator/repressors' category with the annotation of coactivator or corepressor; 'Histone-modifying Enzymes' category contains genes encoding histone modification enzymes; 'Chromatin Remodeling Factors' genes were collected according to the description of GO annotations related to chromatin remodeling but excluding the histone modification enzymes; Genes in 'General Cofactors' category are transcription cofactors involving in initiation or elongation process of transcription; 'Cell Cycle' genes are cell cycle associated transcription cofactors; cofactors did not belong to the above categories were classified as 'Other Cofactors'.

Gene expression

In AnimalTFDB3.0, we provided gene expression information of TFs and transcription cofactors of 22 species, which contains normal tissues, cell lines and cancers in human as well as normal tissues and cells in other species. These expression data showed the ratio of expressed TFs varied from 37% to 99% and cofactors from 41% to 100% in 22 species (Supplementary Table S3). The human TF and cofactor expression in 16 normal tissues collected from EBI Expression Atlas (http://www.ebi.ac.uk/gxa/download.html) illustrates that totally 52.62% TFs and 88.15% cofactors expressed in the 16 tissues. For TFs, there are 6% of them expressed in only one tissue, and for cofactors, the data reduced to 1%.

PPI network

TFs act as important regulators in the transcription process, and a large number of proteins interact with them directly or indirectly to affect the transcription. We got PPI data for 19 species from BioGRID (28) and human PPI data from HPRD (29). In order to illustrate the interaction explicitly, we visualized the PPI networks by Cytoscape.js (http://js.cytoscape.org/) (Figure 1C). The two colors of the network node represent a TF or other gene, and the edges represent the interaction of these proteins with the selected TF.

TF related GWAS phenotypes

We collected the latest human GWAS data and SNP annotation data from GWAS Catalog (25) and dbSNP (release 144) (30) respectively. By mapping the GWAS identified phenotype associated SNPs to the genomic locations of TF and transcription cofactor genes respectively, we obtained a list of SNPs located in TFs and cofactors along with 2469 TF-SNP pairs (680 TFs) and 1796 cofactor-SNP pairs (538 cofactors) with the corresponding GWAS phenotypes. The data indicates that 40.84% TFs and 52.49% cofactors relate to disease phenotypes. For each GWAS SNP locates in TFs or cofactors, the position, disease and reference literature were shown on the page (Figure 1E).

TFBS and its prediction server

TFs regulate gene transcription by binding to specific DNA sequences on target genes. We extracted TFBS of vertebrate TFs by integrating data from HOCOMOCO (v11) (21), JASPAR (22), TRANSFAC (version 2017) and CIS-BP (23) databases, which including TFBS for 18,952 TFs (1335 human TFs, 886 mouse TFs and TFs of other 49 species). The MEME Suite (40) was used to draw the logo of each TFBS (Figure 1D).

Identify the TF targets is a key step for understanding the TF functions. To help users identify TF binding sites on their nucleotide sequences, a TFBS prediction server (http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/tfbs_predict) was built in current version. The TF motif matrices of human for prediction was gathered from TRANSFAC, JASPAR, HOCOMOCO and CIS-BP databases. We also collected TF motifs from hTFtarget (http://bioinfo.life.hust.edu.cn/hTFtarget), which were predicted by peaks calling of ChIP-Seq data using MACS2 (41). The TFBS prediction server will scan these TFBS matrices on user input sequences to predict the TFBS by the motif detection function of the FIMO tool (42) in MEME Suite. In the prediction result, TFBS sequence, score, P-value, Q-value, and detailed alignment information will be shown (Figure 1F).

HumanTFDB web interface

Human is the most concerned and most studied species. In order to facilitate people to directly browse or search human

TFs, we separately built a webpage for Human Transcription Factor Database (HumanTFDB, http://bioinfo.life. hust.edu.cn/HumanTFDB/). In the HumanTFDB, users can browse, search, and download human TFs and cofactors. It also retains the web servers 'Predict TF', 'Predict TFBS' and 'Blast' tool.

DISCUSSION

As the increasing of sequenced and well annotated animal genomes, we updated AnimalTFDB to version 3.0 and several new features were added. AnimalTFDB3.0 provided TFs and cofactors in 97 animal genomes. Most importantly, the accuracy of TF prediction result was improved by adjusting the TF family assignment rules and prediction cutoffs. We have compared the human TFs in AnimalTFDB3.0 with TFs in a recent paper (5) and TRANSFAC data. Among the 1639 TFs in Lambert's paper, 1566 (95.55%) of them are in our AnimalTFDB3.0. The remaining 73 genes (4.45%) were commented as 'Likely to be sequence specific TF' in their website or without literature evidence. However, most of the 157 unique TFs in AnimalTFDB3.0 were explicit TFs, such as transcriptional repressors (LR-RFIP1, LRRFIP2, MIER1, MIER1, ID1/2/3/4 etc.) and activators (SMAD2, SMAD6, SMAD7, UBTF, TCF19, TCF25 etc.). Among the 736 human TFs in TRANS-FAC, 598 (81.00%) of them were TFs or cofactors in AnimalTFDB3.0. Most of the remaining 138 genes are not TFs, such as, nuclear ribonucleoproteins (HNRNPA1, HNRN-PDL and HNRNPL), transporters (ABCG2, SLC22A1, SLC22A3 and SLC6A2), and enzymes (ADAR, HNRN-PAB and HSD17B4). These comparisons highlight the accuracy of our TF prediction results.

The GWAS phenotype related information of human TF and TFBS information will provide useful resources for researchers to further exploration of TF function and regulation. The TFBS prediction server and PPI network will be helpful for user to analyze TF target and its regulatory network. The HumanTFDB web interface is very convenient for researchers to study human TFs. Overall, we believe these improvements will make AnimalTFDB more comprehensive and more useful. There is no doubt that the genomic data of various species will continue to grow. We will continue to update the AnimalTFDB database regularly to make it as a core resource for TF regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank colleagues in data production and database construction in groups of Ensembl, dbSNP, TCGA, TRANSFAC, JASPAR, HOCOMOCO and hTF-target. We are also grateful to our users and all members in our lab for their valuable suggestions and comments.

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. Smith,N.C. and Matthews,J.M. (2016) Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors. *Curr. Opin. Struct. Biol.*, **38**, 68–74.
2. Bella,L., Zona,S., Nestal de Moraes,G. and Lam,E.W.-F. (2014) FOXM1: a key oncofoetal transcription factor in health and disease. *Semin. Cancer Biol.*, **29**, 32–39.
3. Guo,A.-Y., Chen,X., Gao,G., Zhang,H., Zhu,Q.-H., Liu,X.-C., Zhong,Y.-F., Gu,X., He,K. and Luo,J. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, **36**, D966–D969.
4. Jin,J., Tian,F., Yang,D.-C., Meng,Y.-Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.
5. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
6. Wang,K. and Nishida,H. (2015) REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC Bioinformatics*, **16**, 114.
7. Zhang,H.-M., Chen,H., Liu,W., Liu,H., Gong,J., Wang,H. and Guo,A.-Y. (2012) AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–D149.
8. Zhang,H.-M., Liu,T., Liu,C.-J., Song,S., Zhang,X., Liu,W., Jia,H., Xue,Y. and Guo,A.-Y. (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76–D81.
9. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
10. Villar,D., Flicek,P. and Odom,D.T. (2014) Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat. Rev. Genet.*, **15**, 221–233.
11. Butt,T.R. and Karathanasis,S.K. (1995) Transcription factors as drug targets: opportunities for therapeutic selectivity. *Gene Expr.*, **4**, 319–336.
12. Papavassiliou,K.A. and Papavassiliou,A.G. (2016) Transcription factor drug targets. *J. Cell. Biochem.*, **117**, 2693–2696.
13. Tshori,S. and Nechushtan,H. (2012) Mast cell transcription factors–regulators of cell fate and phenotype. *Biochim. Biophys. Acta*, **1822**, 42–48.
14. Ang,Y.-S., Rivas,R.N., Ribeiro,A.J.S., Srivas,R., Rivera,J., Stone,N.R., Pratt,K., Mohamed,T.M.A., Fu,J.-D., Spencer,C.I. *et al.* (2016) Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell*, **167**, 1734–1749.
15. Liu,C.-J., Hu,F.-F., Xia,M.-X., Han,L., Zhang,Q. and Guo,A.-Y. (2018) GSCALite: a web server for gene set cancer analysis. *Bioinformatics*, doi:10.1093/bioinformatics/bty411.
16. Fuxman Bass,J.I., Sahni,N., Shrestha,S., Garcia-Gonzalez,A., Mori,A., Bhat,N., Yi,S., Hill,D.E., Vidal,M. and Walhout,A.J.M. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, **161**, 661–673.
17. Kilpatrick,A.M., Ward,B. and Aitken,S. (2014) Stochastic EM-based TFBS motif discovery with MITSU. *Bioinforma. Oxf. Engl.*, **30**, i310–i318.
18. Ponomarenko,J.V., Orlova,G.V., Merkulova,T.I., Gorshkova,E.V., Fokin,O.N., Vasiliev,G.V., Frolov,A.S. and Ponomarenko,M.P. (2002) rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. *Hum. Mutat.*, **20**, 239–248.
19. Zhang,H.-M., Kuang,S., Xiong,X., Gao,T., Liu,C. and Guo,A.-Y. (2015) Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief. Bioinform.*, **16**, 45–58.
20. Walhout,A.J.M. (2006) Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.*, **16**, 1445–1454.

21. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.

22. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

23. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

24. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.

25. MacArthur,J., Bowler,E., Cerezo,M., Gil,L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.

26. Li,M.J., Wang,P., Liu,X., Lim,E.L., Wang,Z., Yeager,M., Wong,M.P., Sham,P.C., Chanock,S.J. and Wang,J. (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.

27. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.

28. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.

29. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

30. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

31. Tomczak,K., Czerwińska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Poznan Pol.*, **19**, A68–A77.

32. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Füllgrabe,A., Fuentes,A.M.-P., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.

33. Li,J.J., Huang,H., Bickel,P.J. and Brenner,S.E. (2014) Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.*, **24**, 1086–1101.

34. Bastian,F., Parmentier,G., Roux,J., Moretti,S., Laudet,V. and Robinson-Rechavi,M. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In: Bairoch,A, Cohen-Boulakia,S and Froidevaux,C (eds). *Data Integration in the Life Sciences*. Springer, Berlin, Heidelberg, Vol. **5109**, pp. 124–131.

35. Kim,M.-S., Pinto,S.M., Getnet,D., Nirujogi,R.S., Manda,S.S., Chaerkady,R., Madugundu,A.K., Kelkar,D.S., Isserlin,R., Jain,S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.

36. Vinson,C., Myakishev,M., Acharya,A., Mir,A.A., Moll,J.R. and Bonovich,M. (2002) Classification of human B-ZIP proteins based on dimerization properties. *Mol. Cell. Biol.*, **22**, 6321–6335.

37. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

38. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. and Finn,R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.

39. Schmeier,S., Alam,T., Essack,M. and Bajic,V.B. (2017) TcoF-DB v2: update of the database of human and mouse transcription co-factors and transcription factor interactions. *Nucleic Acids Res.*, **45**, D145–D150.

40. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.

41. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

42. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.*, **27**, 1017–1018.