博士論文(要約)

Variation of Transcription Factor Profiles and Their Roles in Mammalian Evolution

(転写因子のプロファイル変動と哺乳類進化における役割)

趙 新威
（Zhao Xin-Wei）

Variation of transcription factor profiles and their roles in mammalian evolution

By

Xin-Wei Zhao

A thesis submitted for the degree of Doctoral of Philosophy
in the Department of Agricultural and Environmental Biology of
graduate school of Agricultural and Life Sciences

The University of Tokyo

Supervisor: Prof. Hirohisa Kishino

Abstract:

Recently unexpectedly large variations were found in regulatory proteins. These variations may significantly affect regulatory network evolution by changing the expressions, molecular interactions, and post-translational modifications of the regulator. These events, however, have been assumed to be occasional or rare in the context of trans-regulatory research for a long time. Trans-species conserved nature was confirmed by comparing orthologous TFs from different species. On the contrary, macro evolution of TF families and its potential effect on species specificity are less studied. It may be acquisition of new TFs and loss of existing TFs that mainly make species unique. In this thesis, I had attempted to answer to this question by analysing the evolution of the mammalian TF families.

Multiple isolated transcription factors act as switches and contribute to species uniqueness

In some mammals, although there are good genome sequencing data, there are few studies on the identification of transcription factors in these species. By using the standard hidden Markov model (HMM), I constructed a database of TFs from the entire genomes of 96 species of mammals. I further annotate our database with orthologous groups information by OrthoDB. Now, there exist several animal TF databases, such as humanTFs, animalTFDB3, Riken mouse TFdb,FlyTF, TFCat, TFCONES, and ITFP. My database and the others are based on the similar pipeline by DNA binding domain (DBD) and HMMER. AnimalTFDB3 contains 125,135 TFs from 97 genomes, which ranges from Caenorhabditis elegans to mammal species like human. My database contains 140,821 TFs, focusing on 96 mammal species. This may provide a better solution in mammal's TF evolution history. Numbers of TFs varied largely between species, with minimum of 1,113 in platypus and maximum of 1,905 in chimpanzee.

To interpret the membership variation of TF families in evolutionary context, I constructed the phylogenetic trees of TF families and estimated the events of gene duplication and gene loss by reconciling the gene trees with the species tree. It was found that, in mammalian evolution, the TF families had increased their members by 37.8% and lost their members or their DBDs by 15.0%. As a result, each species has its unique set of TFs.

Largely because existing TF databases had insufficient coverage, previously constructed gene regulatory networks (GRNs) for mammals were mostly limited to TF-to-TF relationships and

cover only a small subset of orthologues. They revealed the conservation between orthologues, with all other TFs ignored. To account for the contribution of birth and death of TFs, I constructed more comprehensive GRNs (1,200+ TFs and 8000+ nodes and containing TF-to-TF and TF-to-TG edges) for mouse and rat generated from whole species gene interaction networks using STRING. Because mouse and rat have the same common ancestor not long ago, any differences between their GRNs can be assumed to be caused by recent changes.

By comparing the human TF-to-TF interaction network with that of mouse, I confirmed that two-thirds of TFs in the largest connected component were conserved. On the other hand, the other about 500 isolated TFs, which had been mostly treated as out of target in the preceding studies, were variable and prone to be lost from the genome. By comparing the number of isolated and primary connected TFs, I found that a large amount of human isolated TFs were enriched in the Cys2His2-like (zf-C2H2) TF family. The difference in the members of C2H2 zinc finger proteins, the largest TF family, was associated with lower expression of their interaction genes in human compared with mouse. Knock-out mice that lacked these interacting genes had abnormal phenotypes, such as a short tail or hairless skin, which are observed in humans.

As the contents of this chapter (page) are anticipated to be published in a paper in a scholarly journal, they cannot be published online. The paper is scheduled to be published within 5 years.

The effect of TF loss

To quantify the effect of TF loss on the target genes (TGs), I examined TG evolutionary rates and found the deceleration effect of TF loss on TGs over the long term. Because the rate of molecular evolution of a gene is negatively correlated with the strength of a functional constraint, the loss of a TF is expected to have a role in the adaptive evolution of the regulatory system. The molecular evolutionary rate of a gene is negatively correlated with its expression level. Consistently, in human and mice, TGs lacking TFs had higher expression levels. Functional annotation of TGs revealed functions mainly related to the cell cycle, cell migration, signal transduction, and inflammation. By comparing GRNs of mouse and rat, I found that loss of TF genes lost all its edges and DBD changes resulted in an average loss of 50.7% of the

edges, much larger than 30.1% edge loss due to all other factors besides DBD. So, TF and its DBD are the main factors affecting its subnetworks in GRN.

I used gene ontology (GO), a standard functional annotation method for comparative research among species, to reveal relationships between TFs and their functions. I found that 92.7% of TFs (1430 out of 1543 TFs) own unique set of GO items when comparing with any other TFs in mouse. Considering the low resolution of some GO items, this means each TFs can be taken as unique and have their own functions. In the list of GO items that are governed by single TFs, the regulating TFs vary among human, mouse and rat. Through the enrichment analysis of this small GO item list, I found that majority of these GO items are phenotype relate functions. These functions were found to be regulated by different TFs among species, suggesting changes in the regulatory pattern.

As the contents of this chapter (page) are anticipated to be published in a paper in a scholarly journal, they cannot be published online. The paper is scheduled to be published within 5 years.

Table of Contents

Chapter1 Introduction

Gene expression patterns vary among species—even among closely related species that share highly similar genomic sequences. These differences in gene expression and regulation are believed to be the major sources of species phenotypic variation and important factors in evolution.

For many years, mutations in TFs have been thought to be the least likely source of variation, mainly because they can be responsible for negative pleiotropic effects. When a mutation arises in protein-coding regions of a transcriptional regulator, multiple target genes of the regulator are simultaneously affected, potentially causing large-scale detrimental effects. Genetic perturbations of 304 human/mouse TF orthologs in mouse associate with phenotypes and many individual TF loci have strong GWAS signals for multiple diseases. HOX TF genes play a key role in proper body pattern formation [1], while SRY, a TF gene, is important for sex determination. In particular, C2H2 zinc finger proteins were found to diversify rapidly and to represent most of the rapidly evolving human TFs.

During the past decade, an ever-increasing number of hidden Markov models of DNA binding domains (DBDs) and the growing sensitivity of TF detection procedures based on these models have contributed to the expansion of TF databases. Several animal TF databases have been established, such as animalTFDB3 [2], Riken mouse TFdb [3], FlyTF [4], TFCat[5], TFCONES [6], ITFP [7], and humanTFs [8]. These databases collectively contain variable numbers of TFs from different species. Scanning of these databases suggests that the number of non-orthologous TFs is significant. Recent research on C2H2 TF families has also revealed the variability of TFs, but the relative frequency and consequences of global variation remain largely unexplored.

Although the systematic mapping of protein–protein interaction (PPI) is far from complete, it enables to understand the developmental and disease mechanisms at the system level by associating the global topology and dynamic characteristics of the interactome network with known biological characteristics. Orthologous human and mouse TFs show preserved TF–TF interactions in a TF-to-TF network. In contrast, information regarding the effects of non-orthologous TFs on gene regulatory networks is still limited. TFs with only non-TF interactions are usually ignored in TF-to-TF networks because they lack TF–TF interactions

and are considered non-conservative. Since orthologous TFs are shared by both species, they are expected to be the core elements of the regulatory networks. Species specificity may be generated by microevolution of these orthologous TFs or their downstream target genes in each species lineage or it may be generated by rewiring the transcription networks by acquisition of new TFs and loss of existing TFs in each lineage. Because the second scenario has been largely neglected, I attempted to characterize it in this paper. Some TF/protein interactions are less well documented; however, their conservation tends to be low and mutated TFs are likely to be lethal, so they are more likely to achieve lineage-specific adaptation (reviewed in [9]). What then, are the TFs with rare TF interactions in TF-to-TF networks? How do these TFs work to enable different numbers of TFs between species? Based on the above findings, I identified such transcription factors, and they conformed to the speculated characteristics described in previous studies. I further investigated the origins, consequences, and underlying regulatory logic of TF evolution for this set of isolated TFs.

Simplification and complication are both critical aspects of macroevolution. Simplification, that is, the reduction of biological complexity to varying degrees, has received less scientific attention than complexity. Examples of simplification-driven diversification across the tree of life include simplification events in the early history of metazoans, convergent losses of complexity in fungi, and simplification during early eukaryotic evolution (reviewed by [10]. Nonadaptive simplification, such as drift, can lead to the accumulation of slightly deleterious mutations in bacteria [11]. Adaptive genome reduction may also explain some important stages of eukaryotic evolution, such as the simplification of animal metabolism [12]. The ‘less-is-more principle’ suggests that loss of gene function is a common evolutionary response of populations undergoing an environmental shift and, consequently, a change in the pattern of selective pressures [13]. In this regard, TF patterns may naturally evolve along with macroevolution.

The important components of metazoan and embryonic-plant TF kits were present even earlier in their respective single-cell ancestors [14]. Given that the origin and expansion of TFs occurred long before the big bang of speciation, macroevolution has likely been driven by a more direct factor, possibly TF loss. TF loss leading to major diversification has occurred in eukaryotes–for instance, the convergent simplification of adaptin complexes in flagellar apparatus diversification [10]. As another example, the unexpectedly complex list of Wnt family signaling factors evolved in early multicellular animals about 650 million years ago

(Mya) [15]. Functional and phenotypic diversification of the mouth was caused by the loss of Wnt family signaling factors during animal evolution [16].

How TFs work when mammalian species quickly adapt to changing environments via macroevolution is poorly understood. Advances in comparative genomics have clearly shown that the exclusive use of genes as evolutionary units is an oversimplification of actual evolutionary relationships [17, 18]. The related concept of orthologous groups refers to a set of homologous genes that evolved from a single ancestral gene after a given speciation event. Given the close connection between orthologous groups and evolutionary events, I used orthologous groups, rather than genes or gene families, as the basic unit in this study to detect gain and loss events of global TFs in mammalian evolutionary history. Because of the lack of phenotypic data related to orthologous groups and the high resolution of orthologous genes in loss events, I used orthologous TFs to identify the association between TF loss and traits. Here, I show the pattern of TF loss enrichment in the macroevolutionary process. The role of TFs in macroevolutionary processes is further discussed by describing the correlation between TF loss, target gene (TG) expression and molecular evolutionary rate, which also between TF loss and species traits. This analysis may provide new insights into the role of TF loss under various macroevolutionary models as well as its contribution to the rapid adaptation of species to different environments.

Chapter2 Materials and Methods

2.1 Construction of a mammalian transcription factor (TF) database

2.1.1 Existing databases

The identification of transcription factors is the basis of transcription factor research. In the past 20 years, researchers have constructed some mammalian transcription factor databases. But today, a large number of databases stopped updating long ago. In the recently updated database, the more famous and representative databases are HumanTFs and AnimalTFDB.

In 2018, Lambert et al. published HumanTFs, a database of human transcription factors, in cell magazine [8]. This database focuses on humans and has published a list of 1,639 human transcription factors. 69 putative TFs with unknown DBD family are included in this list. Another database is AnimalTFDB3 [2]. This is version 3 of AnimalTFDB, which is the most recently updated version in 2019. AnimalTFDB3 contains 125,135 transcription factors from 97 species, 72 of which are mammals.

My research focuses on mammals. Because of the need to consider the evolutionary history of transcription factors and the number of species changes, more mammals are required to provide more dense transcription factor data to improve the accuracy of the final results. Because of this need, if possible, my database needs more mammals than 72 species. On the other hand, since AnimalTFDB3 has not been released when this research started, and AnimalTFDB2 contains only 41 mammalian species, it is necessary to establish a new mammalian transcription factor database.

2.1.2 Mammalian genome data

In order to obtain a suitable number of mammalian species, I chose the NCBI database and obtained all the protein sequences in 96 mammalian genomes. Since sequencing depth and quality may affect the quality of transcription factor identification, I confirmed the 96 mammalian genome data in the NCBI database. According to Table 2.1. Except for one species lacking sequencing depth information, there are 8 species from 5x to 10x, 55 from 10x to 100x, and 32 from 100x to 500x. Because these species have good sequencing depth and quality, it is reliable to use these genome sequences for further transcription factor identification.

Table 2.1 Genome information of 96 mammalian species

Lineage	SpeciesName	protein count	GC%	Coverage
Afrotheria	*Chrysochloris asiatica*	25227	41.8	66x
Afrotheria	*Elephantulus edwardii*	25209	41.5	62x
Afrotheria	*Echinops telfairi*	22926	43.6	78x
Afrotheria	*Loxodonta africana*	29784	40.9	7x
Afrotheria	*Orycteropus afer*	25544	42.1	44x
Afrotheria	*Trichechus manatus latirostris*	26315	41.6	150x
Carnivora	*Acinonyx jubatus*	27284	41.4	75.0x
Carnivora	*Ailuropoda melanoleuca*	36506	41.7	60x
Carnivora	*Canis lupus familiaris*	58776	41.1	90.0x
Carnivora	*Felis catus*	53033	43.67	20x
Carnivora	*Leptonychotes weddellii*	25718	43.8	82x
Carnivora	*Mustela putorius furo*	48107	41.8	162x
Carnivora	*Odobenus rosmarus divergens*	31370	41.7	200.0x
Carnivora	*Panthera tigris altaica*	29473	41.5	99x
Carnivora	*Ursus maritimus*	28887	41.7	101x
Cetartiodactyla	*Balaenoptera acutorostrata*	34821	41.4	92x
Cetartiodactyla	*Bison bison bison*	35554	42.2	60.0x
Cetartiodactyla	*Bubalus bubalis*	41499	42.2	70.0x
Cetartiodactyla	*Bos mutus*	28881	42	130x
Cetartiodactyla	*Camelus bactrianus*	28601	40.45	79.2x
Cetartiodactyla	*Camelus dromedarius*	26729	41.5	65x
Cetartiodactyla	*Camelus ferus*	31796	41.3	30x
Cetartiodactyla	*Capra hircus*	42687	42.69	50.0x
Cetartiodactyla	*Lipotes vexillifer*	26901	41.4	115x
Cetartiodactyla	*Ovis aries*	48308	42.4	166.0x
Cetartiodactyla	*Orcinus orca*	27870	41.7	200.0x
Cetartiodactyla	*Physeter catodon*	31522	41.3	75x
Cetartiodactyla	*Pantholops hodgsonii*	32279	42.4	67.0x
Cetartiodactyla	*Sus scrofa*	63577	41.5	65.0x
Cetartiodactyla	*Tursiops truncatus*	38849	41.5	114.5x
Cetartiodactyla	*Vicugna pacos*	33208	41.6	72.5x
Chiroptera	*Eptesicus fuscus*	24147	43.5	84x
Chiroptera	*Myotis brandtii*	40808	42.9	120x
Chiroptera	*Myotis davidii*	33106	43.1	110x
Chiroptera	*Myotis lucifugus*	41184	42.7	7x
Chiroptera	*Miniopterus natalensis*	29787	42.4	77.0x
Chiroptera	*Pteropus alecto*	33106	39.9	110x
Chiroptera	*Pteropus vampyrus*	33311	40.5	188.0x

Chiroptera	*Rousettus aegyptiacus*	48803	40	169.2x
Dasyuromorphia	*Sarcophilus harrisii*	24821	37.15	85x
Dermoptera	*Galeopterus variegatus*	32104	41.2	55x
Didelphimorphia	*Monodelphis domestica*	49112	38.14	NA
Insectivora	*Condylura cristata*	29166	41.9	113.1x
Insectivora	*Erinaceus europaeus*	29382	42	79x
Insectivora	*Sorex araneus*	23427	43.4	120x
Lagomorpha	*Oryctolagus cuniculus*	38463	44.05	7.48X
Lagomorpha	*Ochotona princeps*	25691	45.2	103x
Monotremata	*Ornithorhynchus anatinus*	24786	45.66	6x
Perissodactyla	*Ceratotherium simum simum*	33629	41.2	91x
Perissodactyla	*Equus asinus*	42247	41.4	42.4x
Perissodactyla	*Equus caballus*	36064	41.48	90.57x
Perissodactyla	*Equus przewalskii*	38416	41.3	85.63x
Pholidota	*Manis javanica*	41843	41.5	60.0x
Primates	*Aotus nancymaae*	47568	41.1	132.4x
Primates	*Colobus angolensis palliatus*	38656	41.6	86.8x
Primates	*Cercocebus atys*	65920	41.1	192.0x
Primates	*Cebus capucinus imitator*	53175	41	81x
Primates	*Callithrix jacchus*	45251	40.8	60x
Primates	*Chlorocebus sabaeus*	61803	40.93	95x
Primates	*Carlito syrichta*	33081	41	48x
Primates	*Gorilla gorilla*	46533	40.98	80x
Primates	*Homo sapiens*	79257	40.9	165.0x
Primates	*Macaca fascicularis*	40079	41.34	68x
Primates	*Mandrillus leucophaeus*	38336	41.6	117.2x
Primates	*Macaca mulatta*	36584	41.2	50x
Primates	*Microcebus murinus*	59023	41.34	221.6x
Primates	*Macaca nemestrina*	62876	41.3	113.1x
Primates	*Nomascus leucogenys*	38654	41.4	5.6x
Primates	*Otolemur garnettii*	26925	41.5	137x
Primates	*Pongo abelii*	37509	41.59	6x
Primates	*Papio anubis*	66659	41.05	104.0x
Primates	*Propithecus coquereli*	28194	43.2	104.7x
Primates	*Pan paniscus*	47451	42.32	26x
Primates	*Pan troglodytes*	79956	40.88	55x
Primates	*Rhinopithecus bieti*	49595	41.5	76.6x
Primates	*Rhinopithecus roxellana*	37291	41	53.7x
Primates	*Saimiri boliviensis*	36241	41.1	80x
Rodentia	*Cricetulus griseus*	32843	41.6	130x
Rodentia	*Chinchilla lanigera*	45449	41.4	87x
Rodentia	*Cavia porcellus*	37720	40.1	6.8x
Rodentia	*Dipodomys ordii*	29351	42.6	181.0x

Rodentia	*Fukomys damarensis*	43907	40.5	140x
Rodentia	*Heterocephalus glaber*	41292	40.95	90x
Rodentia	*Jaculus jaculus*	25473	42.7	78x
Rodentia	*Ictidomys tridecemlineatus*	38592	40.1	495.1x
Rodentia	*Mesocricetus auratus*	36852	43.2	115x
Rodentia	*Marmota marmota marmota*	31750	40.2	30x
Rodentia	*Mus musculus*	61940	42.49	60.0x
Rodentia	*Microtus ochrogaster*	30752	42.83	94x
Rodentia	*Nannospalax galili*	38552	41.6	86x
Rodentia	*Neotoma lepida*	24320	36.2	48.0x
Rodentia	*Octodon degus*	27336	42.5	80x
Rodentia	*Peromyscus maniculatus bairdii*	45588	42.7	110.0x
Rodentia	*Rattus norvegicus*	42093	42.34	10.0x
Scandentia	*Tupaia chinensis*	36148	42	80x
Xenarthra	*Dasypus novemcinctus*	38202	41.5	6x

These reference genomes information were obtained from NCBI.

2.1.3 Pfam and Hidden Markov Models

The Pfam[19] database integrates structural information and representative sequence information to confirm and classify protein domains or DNA binding domains (DBD) (Figure 2.1). Many of domain information are based on PDB [20] and Interpro database [21]. By categorizing different DBDs or protein domains, pfam defines a family. Therefore, the classification of transcription factor families in this study completely complies with the definition and classification of transcription factor families in the Pfam database. Through the seed sequences in the family, the Pfam database provides the hidden Markov model (HMM) of the DBD of this family by HMMER and all the parameter of HMM.

HMMER can be used for multiple sequence alignments[22-24] and sequence homologs search by HMM models [25-29]. HMMER is further improved on the basis of profile HMM architecture to make the appraisal result more reliable (Figure 2.2). The HMMER model has more insert states than the Profile HMM, which enables HMMER to read the entire sequence at one time and detect multiple repeated DBDs in a single sequence. On the other hand, HMMER's model allows from any match state to end state, which enables HMMER to detect incomplete DBD, which is closer to reality and further improves the sensitivity of detection. Therefore, the use of HMMER and HMM to identify transcription factors is widely used.

Through HMM, I collected sequences containing this DBD from all protein sequences of the 96 species. And this step of searching is realized by HMMER software.

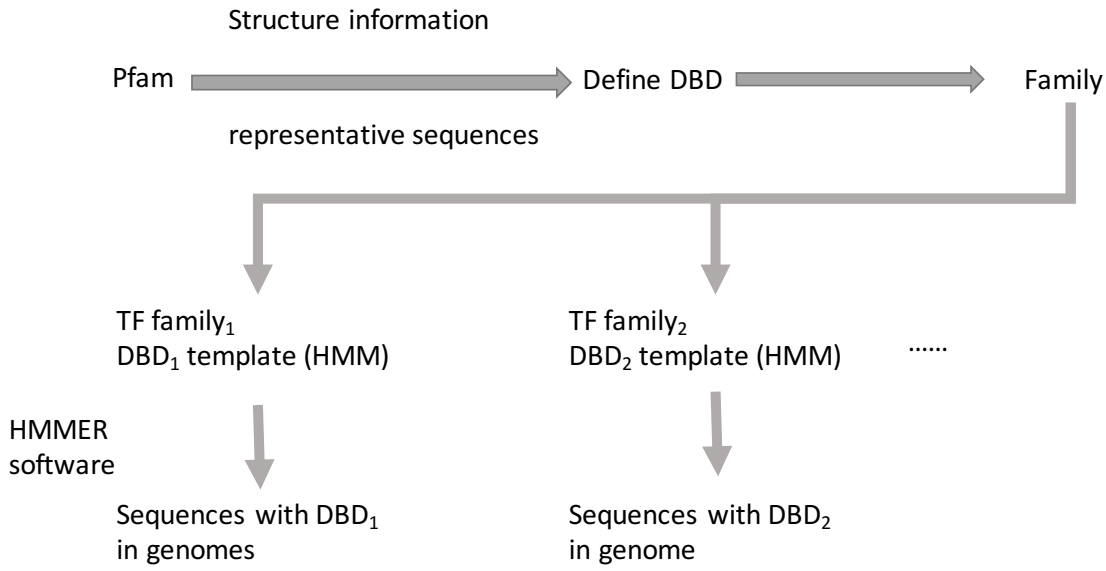

Figure 2.1 TF family and DBD defined by Pfam database.

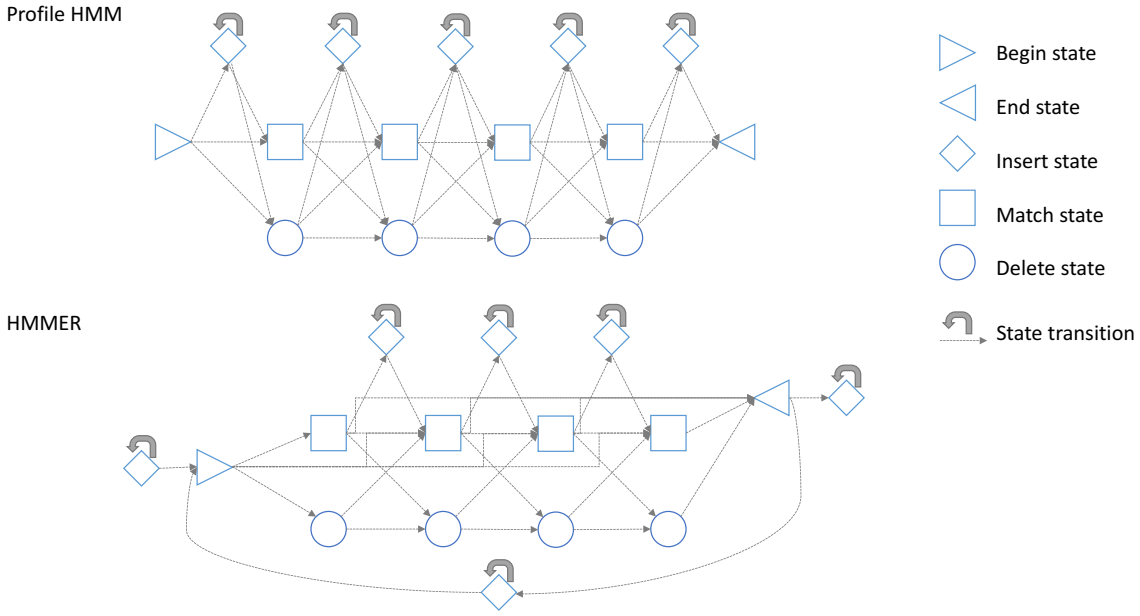

Figure 2.2 Model architecture of profile HMM and HMMER. Diamond is insert state; square is match state; circle is delete state; triangle is begin or end state. Match states and insert states of proteins each have emission distribution of 20 possible amino acid symbols.

2.1.4 Detect TFs from genomes

I detected TFs by HMMER from all protein sequences of the 96 species based on the Hidden Markov models of the 66 transcription factor families (Figure 2.3). Standardizing these protein sequences names without deleting or changing the sequences themselves, enables them to meet the requirements of the HMMER software, which is conducive to further processing in the future.

With a batch program, I obtained the transcription factor family information. I adopted an E value of 0.0001 to ensure high reliability. For each single run, I got all the candidate TFs with one type of DBD from one species. By 6,336 ($96 * 66$) runs, I got these 6336 datasets which covered 96 mammal species and 66 TF families.

To remove redundancies, protein names were annotated, and only protein isoforms with the highest scores were retained; in addition, alternative splicing types were filtered out after TF detection for each DBD. The duplications were cleared after identification. This step decreased the false negative of transcription factors, which might cause by the alternative splicing and sequencing depth, to make the results more accurate.

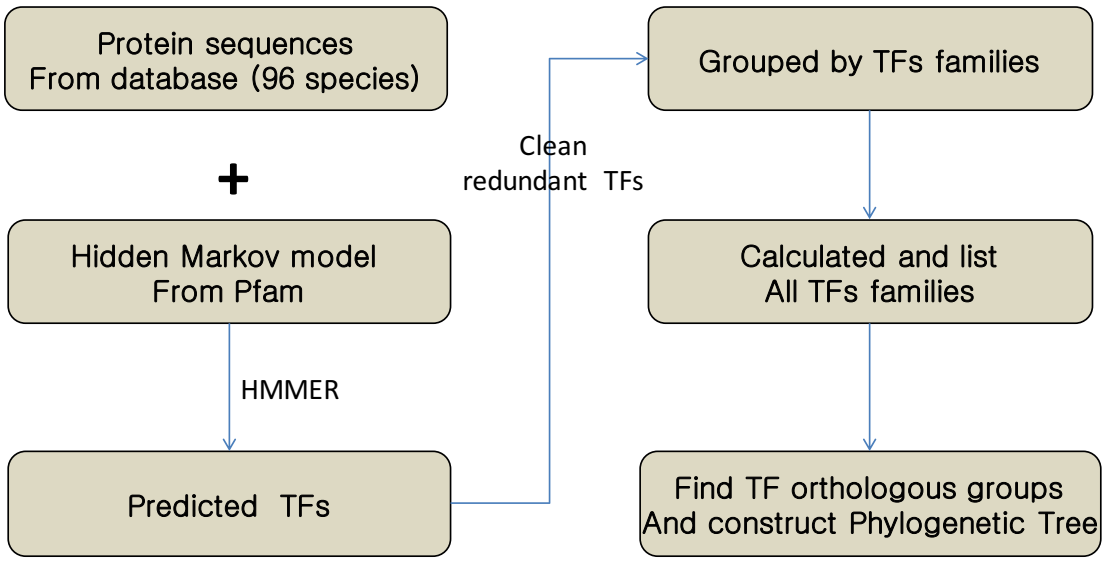

Figure 2.3 Pipeline of TF database construction.

2.1.5 Detect number variation among each TF families and species

Through the previous steps, I got 6,336 datasets of transcription factors from 66 families and 96 species (Table 2.2). Note that a transcription factor, which has multiple DBDs, is included in different transcription factor families. I divided the number of the same transcription factor family in different species by its average value to get the relative size of the same transcription factor family in different species. Using the same method to standardize all 66 transcription factor families, I obtained the relative sizes of different transcription factor families in different species, and used this result to draw a heat map.

In order to study the relationship between transcription factor families, I used the number of transcription factor families in different species to compare the transcription factor families in pairs to obtain the correlations between transcription factor families, and use these correlations to draw heat maps.

In order to obtain the total number of transcription factors in different species, I added the number of transcription factors of different families of the same species to get the sum of the number of transcription factors of this species, and then arranged them according to the order of the species on the species tree.

Table 2.2 TF family size among 96 mammalian species

	AF.4	AP2	ARID	AT_ho	BTB	BTD	bZIP_1	bZIP_2	bZIP_M	CBF	CBFB_	CBFD_	CEP1.	CG.1	CP2	CSD	CUT	DM	E2F_T	EBP	Ets	Forkhe
Chrysochloris_asiatica	4	0	14	1	129	2	42	43	28	3	1	33	0	2	6	8	8	7	10	2	29	43
Elephantulus_edwardii	4	0	13	0	126	2	39	40	30	3	0	29	0	2	6	8	7	5	10	2	29	42
Echinops_telfairi	4	0	12	1	127	2	41	39	27	2	1	20	0	2	6	8	6	3	9	2	29	34
Loxodonta_africana	4	0	12	1	129	2	40	37	26	3	1	39	0	2	6	9	5	5	10	3	28	41
Orycteropus_afer	4	0	15	1	132	2	39	41	27	4	1	35	0	2	6	9	6	7	11	2	28	42
Trichechus_manatus_latirostris	4	0	15	1	127	2	43	44	30	3	1	38	0	2	6	9	7	7	9	2	28	40
Acinonyx_jubatus	6	0	14	0	112	2	35	34	21	2	1	22	0	1	7	7	5	3	10	2	27	25
Ailuropoda_melanoleuca	6	0	14	1	128	2	39	39	27	3	1	31	0	2	6	10	5	5	10	2	28	42
Canis_lupus_familiaris	4	0	14	1	127	2	37	40	26	3	1	40	0	2	7	9	7	8	10	2	29	40
Felis_catus	8	0	14	0	127	2	39	39	26	3	1	33	0	2	6	9	6	6	10	2	29	40
Leptonychotes_weddellii	7	0	13	1	128	2	44	42	25	3	1	25	0	1	7	9	8	6	9	2	28	41
Mustela_putorius_furo	4	0	14	1	131	2	39	43	23	3	1	22	0	2	7	10	5	6	10	2	28	41
Odobenus_rosmarus_divergens	4	0	15	1	133	2	41	43	29	3	1	39	0	2	6	9	6	8	10	2	28	45
Panthera_tigris_altaica	6	0	14	0	117	2	36	36	19	3	1	24	0	2	6	7	6	3	9	2	26	23
Ursus_maritimus	4	0	15	0	119	2	37	37	22	3	1	39	0	2	7	10	6	3	10	2	28	32
Balaenoptera_acutorostrata	5	0	13	1	125	1	36	38	25	3	1	35	0	2	7	10	6	6	10	2	29	38
Bison_bison_bison	5	0	13	0	128	2	35	35	25	3	1	40	0	2	6	8	5	7	10	2	25	39
Bubalus_bubalis	7	0	14	1	135	2	41	41	30	3	1	37	0	2	6	10	7	7	11	2	28	42
Bos_mutus	6	0	14	1	128	2	38	41	30	3	1	34	0	2	7	9	7	6	9	2	27	38
Camelus_bactrianus	5	0	12	0	129	2	42	42	28	3	1	36	0	2	6	9	6	4	9	2	28	37
Camelus_dromedarius	5	0	13	0	129	2	40	40	28	3	1	35	0	2	6	8	6	6	9	2	29	42
Camelus_ferus	6	0	12	0	120	2	33	36	23	3	1	31	0	1	6	8	5	3	9	2	29	35
Capra_hircus	6	0	14	1	132	2	40	43	28	3	1	36	0	2	6	11	7	8	10	2	27	45
Lipotes_vexillifer	7	0	13	1	132	2	37	39	27	3	1	32	0	2	7	8	7	7	10	2	29	43
Ovis_aries	5	0	13	0	125	2	38	40	26	3	1	31	0	2	7	10	8	8	10	3	26	40
Orcinus_orca	5	0	14	1	135	2	40	42	28	3	1	36	0	2	6	9	7	7	10	2	28	44
Physeter_catodon	6	0	13	0	130	1	37	40	25	3	1	34	0	2	8	8	6	7	9	2	29	34
Pantholops_hodgsonii	6	0	14	0	125	2	38	40	26	3	1	32	0	2	7	9	7	4	10	3	28	33
Sus_scrofa	7	0	13	1	121	1	34	37	23	3	1	34	0	1	5	8	10	7	9	2	29	39
Tursiops_truncatus	6	0	9	0	124	2	37	37	25	2	1	24	0	1	6	9	5	8	8	2	27	33
Vicugna_pacos	5	0	12	0	129	2	39	41	26	3	1	31	0	2	7	8	6	4	9	2	28	36

Table 2.2 TF family size among 96 mammalian species (continued)

	GAGA	GATA	GCM	GTF2I	HLH	HMG_	Homec	Hormo	HPD	HSF_D	HTH_p	IRF	LAG1.	MBD	MH1	Myb_D	NCU.G	NDT80	Nrf1_D	P53	PAX	PC4
Chrysochloris_asiatica	4	26	2	3	102	49	209	50	2	5	2	9	2	10	11	29	1	2	1	3	9	2
Elephantulus_edwardii	4	17	2	3	102	52	207	48	2	5	2	9	2	7	12	28	1	2	1	4	9	2
Echinops_telfairi	2	16	2	3	92	45	189	50	2	5	2	10	2	9	10	29	1	2	1	4	9	1
Loxodonta_africana	4	16	2	4	95	46	208	48	2	6	2	10	2	9	12	27	1	1	1	3	9	1
Orycteropus_afer	4	18	2	3	103	48	208	49	2	5	2	10	2	10	12	33	1	2	1	3	9	1
Trichechus_manatus_latirostris	7	16	2	3	99	46	212	50	2	5	2	10	2	8	12	31	1	2	1	3	9	1
Acinonyx_jubatus	7	17	2	4	69	34	159	48	2	4	2	10	2	7	11	30	1	2	1	3	8	2
Ailuropoda_melanoleuca	3	16	2	3	96	45	202	49	2	4	2	10	2	9	11	32	1	1	1	3	9	2
Canis_lupus_familiaris	4	16	2	3	96	46	210	49	2	6	1	8	2	9	12	29	1	2	1	3	8	2
Felis_catus	5	16	2	3	91	43	210	48	2	4	2	11	2	8	10	32	1	2	1	3	9	2
Leptonychotes_weddellii	3	17	2	2	96	47	208	51	2	3	2	8	2	8	12	29	1	2	1	4	9	1
Mustela_putorius_furo	6	16	2	3	93	44	204	49	2	4	1	10	2	8	12	30	1	2	1	3	9	2
Odobenus_rosmarus_divergens	4	18	3	2	103	49	217	49	2	7	2	11	2	10	12	31	1	2	1	3	9	1
Panthera_tigris_altaica	3	15	2	3	73	39	168	46	2	5	1	10	2	8	9	29	1	2	1	3	9	2
Ursus_maritimus	3	15	2	3	73	41	175	48	2	5	2	9	2	9	11	30	1	2	1	3	9	2
Balaenoptera_acutorostrata	5	16	2	3	89	45	203	47	2	3	2	9	1	8	12	27	1	2	1	3	9	1
Bison_bison_bison	2	17	2	3	89	52	196	48	2	8	0	10	2	8	12	29	1	2	1	3	9	1
Bubalus_bubalis	3	17	2	3	102	48	215	48	2	7	2	10	2	9	12	32	1	2	1	3	9	2
Bos_mutus	4	17	2	3	89	44	203	49	2	7	2	9	2	8	11	32	1	2	1	3	9	2
Camelus_bactrianus	0	15	2	3	89	44	199	49	2	5	2	10	2	9	12	31	1	2	1	3	9	1
Camelus_dromedarius	2	16	2	3	90	47	205	49	2	7	2	10	2	9	10	31	1	2	1	3	9	1
Camelus_ferus	1	14	2	3	78	39	173	49	2	8	2	9	2	8	10	31	1	2	1	4	9	1
Capra_hircus	4	16	2	3	105	50	216	49	2	8	2	10	1	9	12	31	1	2	1	3	9	1
Lipotes_vexillifer	3	18	2	3	104	45	202	47	2	5	2	10	2	9	12	31	2	2	1	3	9	1
Ovis_aries	4	14	2	4	94	47	198	47	2	8	2	11	1	8	9	29	1	2	1	3	9	2
Orcinus_orca	4	18	2	3	103	51	207	49	2	5	2	10	2	9	12	31	1	2	1	3	9	1
Physeter_catodon	2	15	2	3	85	38	199	47	2	4	2	10	1	8	10	30	1	2	1	3	9	1
Pantholops_hodgsonii	2	17	2	3	79	42	193	48	2	6	2	10	2	9	10	30	1	2	1	3	9	2
Sus_scrofa	3	15	1	4	97	49	201	48	2	5	2	10	2	10	12	28	1	2	3	3	11	1
Tursiops_truncatus	5	16	2	3	93	41	185	46	2	4	2	9	2	7	10	25	1	2	1	3	7	1
Vicugna_pacos	2	16	2	4	87	42	198	50	2	5	2	10	2	9	10	32	1	2	1	3	8	1

Table 2.2 TF family size among 96 mammalian species (continued)

	Pou	RFX_D	RHD_[	Runt	SAND	SRF.T	STAT_	T.box	TEA	TF_AP	TF_Ot	THAP	TIG	TSC22	Tub	zf.BED	zf.C2H	zf.C2H	zf.C4	zf.LITA	zf.MIZ	zf.NF.X
Chrysochloris_asiatica	14	9	11	3	6	5	7	17	4	5	3	10	24	3	5	19	482	6	47	2	6	2
Elephantulus_edwardii	15	8	10	3	7	5	7	16	4	5	3	6	23	4	5	19	433	6	46	2	6	2
Echinops_telfairi	12	8	10	3	5	5	7	18	4	5	3	8	23	4	5	17	461	6	45	3	6	2
Loxodonta_africana	13	9	10	3	7	4	7	17	4	5	3	9	21	4	5	19	532	6	46	2	6	2
Orycteropus_afer	13	10	10	3	7	5	7	18	3	5	3	9	23	4	5	31	538	6	46	3	6	2
Trichechus_manatus_latirostris	12	9	10	3	6	5	7	17	4	5	3	10	23	4	5	24	555	5	45	2	6	2
Acinonyx_jubatus	12	8	10	3	6	5	7	16	4	5	4	10	22	4	5	9	476	7	44	2	5	2
Ailuropoda_melanoleuca	16	9	10	3	7	5	7	17	3	5	3	12	24	4	5	12	542	7	47	3	6	2
Canis_lupus_familiaris	13	9	10	4	8	5	7	17	4	5	3	12	24	4	5	11	521	6	46	3	7	2
Felis_catus	15	8	10	3	7	5	7	17	4	5	3	12	24	5	5	11	531	7	46	2	5	2
Leptonychotes_weddellii	15	6	11	3	8	5	8	18	3	5	3	10	21	5	5	12	492	7	44	2	6	2
Mustela_putorius_furo	14	8	10	3	7	5	7	17	4	5	3	11	24	5	5	11	537	6	46	4	7	2
Odobenus_rosmarus_divergens	15	8	10	3	7	5	7	17	4	5	3	12	24	4	5	11	522	7	47	3	7	2
Panthera_tigris_altaica	12	8	9	3	6	4	7	17	4	5	2	10	21	5	5	11	471	7	43	3	5	2
Ursus_maritimus	14	9	10	3	7	5	8	17	4	5	3	9	21	4	5	10	509	7	44	3	5	2
Balaenoptera_acutorostrata	13	8	10	3	7	5	6	17	4	5	3	12	22	4	5	10	495	6	44	2	6	2
Bison_bison_bison	13	8	10	3	7	5	7	17	3	5	3	9	24	4	5	10	509	6	44	2	6	2
Bubalus_bubalis	14	9	10	3	8	5	7	17	4	5	3	11	24	3	5	10	480	6	45	2	6	2
Bos_mutus	12	8	10	3	7	4	7	17	4	5	3	11	24	3	5	10	496	6	44	3	6	2
Camelus_bactrianus	17	9	10	3	6	5	7	17	4	5	3	9	23	4	5	11	526	7	45	2	6	2
Camelus_dromedarius	15	9	10	3	7	5	8	17	4	5	3	10	23	4	5	13	508	7	46	2	6	2
Camelus_ferus	12	10	10	3	6	4	7	18	4	5	3	9	24	4	5	14	504	6	45	3	6	2
Capra_hircus	14	8	9	3	9	5	7	17	4	5	3	12	24	4	5	11	535	6	46	3	7	2
Lipotes_vexillifer	16	9	9	3	7	5	8	17	5	5	3	12	23	5	5	13	487	6	46	2	6	2
Ovis_aries	14	9	10	3	8	4	7	16	4	5	4	11	23	4	5	12	508	6	44	3	6	2
Orcinus_orca	13	8	10	3	8	5	7	17	4	6	3	13	24	4	5	12	523	6	46	2	6	2
Physeter_catodon	13	7	10	3	6	5	7	17	4	5	3	11	23	4	4	8	475	5	45	2	6	2
Pantholops_hodgsonii	13	8	10	3	7	4	7	16	5	5	3	7	24	3	5	7	448	6	43	2	6	2
Sus_scrofa	13	9	11	5	5	7	8	16	4	6	3	10	23	5	4	12	494	5	42	3	8	3
Tursiops_truncatus	11	8	10	3	6	5	6	16	4	5	3	7	22	4	3	12	450	6	39	2	6	1
Vicugna_pacos	15	9	10	3	7	5	7	17	4	5	3	10	23	5	5	10	507	6	46	2	6	2
Eptesicus_fuscus	15	8	10	3	7	5	7	17	4	5	3	9	24	5	5	12	475	6	43	2	7	2

Table 2.2 TF family size among 96 mammalian species (continued)

	AF.4	AP2	ARID	AT_ho	BTB	BTD	bZIP_1	bZIP_2	bZIP_M	CBF	CBFB_	CBFD_	CEP1.	CG.1	CP2	CSD	CUT	DM	E2F_T	EBP	Ets	Forkhe
Myotis_brandtii	5	0	15	0	124	2	37	38	27	3	1	23	0	2	7	7	5	3	11	2	28	37
Myotis_davidii	6	0	13	0	125	2	37	36	23	3	1	21	0	2	7	9	5	3	10	2	28	33
Myotis_lucifugus	5	0	13	0	127	2	34	36	25	3	1	29	0	2	8	10	6	4	11	2	26	29
Miniopterus_natalensis	5	0	13	0	129	2	39	42	30	3	1	30	0	2	6	7	6	3	10	2	28	34
Pteropus_alecto	5	0	15	1	129	2	39	39	26	3	1	33	0	2	7	8	5	7	10	2	27	43
Pteropus_vampyrus	6	0	14	1	132	2	39	39	24	3	1	27	0	2	7	8	6	7	10	2	29	42
Rousettus_aegyptiacus	4	0	16	1	131	2	41	40	26	3	1	32	0	2	7	9	7	8	10	2	28	47
Sarcophilus_harrisii	4	0	15	0	121	2	37	36	23	3	1	22	0	2	8	8	6	3	9	2	26	35
Galeopterus_variegatus	6	0	15	1	130	2	38	38	25	3	1	38	0	2	9	8	7	4	11	2	29	36
Monodelphis_domestica	4	0	14	1	129	2	36	35	25	3	1	23	0	2	5	8	7	7	11	2	26	42
Condylura_cristata	5	0	12	1	123	2	38	39	27	3	1	28	0	2	7	9	6	6	8	2	29	35
Erinaceus_europaeus	5	0	13	0	126	2	41	39	26	3	1	26	0	1	6	9	6	6	10	2	29	40
Sorex_araneus	4	0	12	0	127	2	42	39	27	3	1	29	0	2	6	8	7	4	9	2	27	32
Oryctolagus_cuniculus	5	0	13	1	122	2	35	40	24	3	1	31	0	2	6	9	6	8	11	2	28	37
Ochotona_princeps	4	0	14	0	132	2	42	39	27	3	1	27	0	2	6	7	6	6	10	2	29	43
Ornithorhynchus_anatinus	4	0	13	0	109	2	28	28	18	3	1	24	0	1	6	7	6	4	9	2	23	28
Ceratotherium_simum_simum	5	0	14	1	133	2	40	41	28	3	1	38	0	1	7	8	5	8	9	2	29	40
Equus_asinus	5	0	14	0	127	2	39	39	26	3	1	41	0	1	6	8	6	6	10	2	29	34
Equus_caballus	4	0	14	1	123	2	36	36	26	3	1	37	0	1	6	9	6	4	10	2	29	36
Equus_przewalskii	7	0	13	0	126	2	36	38	26	3	1	34	0	2	6	9	6	5	9	3	30	32
Manis_javanica	5	0	14	1	127	2	39	42	29	3	1	24	0	2	6	8	5	8	9	2	27	41
Aotus_nancymaae	4	0	13	1	127	2	37	39	27	3	1	36	0	2	6	11	7	7	10	2	28	43
Colobus_angolensis_palliatus	4	0	13	1	122	2	38	38	25	3	1	34	0	2	6	8	6	6	11	2	29	42
Cercocebus_atys	4	0	14	1	128	2	38	41	28	3	1	33	0	2	6	8	8	7	11	2	28	45
Cebus_capucinus_imitator	5	0	15	1	131	2	39	41	26	3	1	32	0	2	6	9	7	7	11	2	28	46
Callithrix_jacchus	4	0	15	1	126	2	37	41	26	3	1	33	0	2	6	11	7	7	12	2	29	42
Chlorocebus_sabaeus	4	0	14	1	131	2	36	41	27	3	1	34	0	2	6	8	6	7	11	2	29	45
Carlito_syrichta	6	0	12	0	124	2	37	37	27	3	1	41	0	2	6	8	6	4	10	2	25	33
Gorilla_gorilla	4	0	14	1	128	2	38	43	27	3	1	36	0	2	7	8	8	7	11	3	27	41
Homo_sapiens	4	0	15	1	133	2	39	40	25	3	1	26	0	2	6	9	7	7	11	2	28	50
Macaca_fascicularis	4	0	13	1	130	2	38	40	26	3	1	32	0	2	6	8	7	7	10	2	31	44
Mandrillus_leucophaeus	5	0	13	0	123	2	36	36	26	3	1	36	0	2	7	8	6	5	10	2	29	33

Table 2.2 TF family size among 96 mammalian species (continued)

	GAGA	GATA	GCM	GTF2I	HLH	HMG_	Homec	Hormo	HPD	HSF_D	HTH_p	IRF	LAG1.	MBD	MH1	Myb_D	NCU.G	NDT80	Nrf1_D	P53	PAX	PC4
Myotis_brandtii	3	19	2	3	81	48	188	47	2	5	2	10	2	8	10	29	1	2	1	3	9	1
Myotis_davidii	2	16	2	3	77	43	172	47	2	5	2	9	2	8	11	26	1	2	1	3	8	1
Myotis_lucifugus	5	25	2	3	82	42	190	46	1	6	2	8	2	11	11	26	1	2	2	2	8	1
Miniopterus_natalensis	1	17	2	3	84	47	184	46	2	6	2	10	2	8	11	28	1	2	1	4	9	1
Pteropus_alecto	4	15	2	3	92	47	204	47	2	6	2	9	2	9	12	29	1	2	1	3	9	1
Pteropus_vampyrus	3	16	2	3	92	41	196	48	2	3	2	10	2	9	13	29	2	2	1	3	9	1
Rousettus_aegyptiacus	4	16	2	4	103	49	210	49	2	6	2	10	2	9	12	29	1	2	1	3	9	2
Sarcophilus_harrisii	2	63	2	4	90	45	197	50	2	4	2	9	2	8	12	30	1	1	1	4	9	1
Galeopterus_variegatus	4	17	2	4	89	48	218	49	2	6	2	10	3	8	12	33	1	2	1	3	9	2
Monodelphis_domestica	2	40	2	3	102	47	194	46	2	5	2	9	2	9	12	28	1	1	1	3	9	1
Condylura_cristata	1	11	2	3	87	39	200	49	2	3	2	10	2	8	12	30	1	2	1	3	9	1
Erinaceus_europaeus	7	14	2	3	92	50	192	48	2	4	2	9	2	9	10	32	1	2	1	3	10	2
Sorex_araneus	2	15	2	3	86	46	190	48	2	4	2	10	2	8	13	30	1	2	1	3	9	1
Oryctolagus_cuniculus	2	33	2	3	88	43	202	46	2	5	2	9	2	9	11	29	1	2	1	3	9	2
Ochotona_princeps	3	15	2	4	96	50	203	49	2	4	2	9	2	9	12	28	1	2	1	3	8	1
Ornithorhynchus_anatinus	3	10	2	2	75	35	134	41	2	5	1	6	2	7	11	29	1	2	1	3	8	1
Ceratotherium_simum_simum	5	18	2	3	99	48	205	49	2	5	2	10	2	9	12	29	1	2	1	3	9	1
Equus_asinus	4	15	2	3	93	45	201	49	2	8	2	10	2	8	12	29	1	2	1	3	9	1
Equus_caballus	3	14	2	3	80	41	194	48	2	8	2	10	2	9	10	32	1	2	1	3	9	2
Equus_przewalskii	3	13	2	3	88	43	190	49	2	5	2	9	2	9	9	28	1	2	1	3	9	2
Manis_javanica	2	18	2	3	89	44	190	50	2	3	1	10	2	9	11	31	1	2	1	4	8	1
Aotus_nancymaae	5	19	2	4	97	46	209	49	2	5	2	7	2	10	12	30	1	2	1	3	9	1
Colobus_angolensis_palliatus	6	18	2	3	88	46	205	47	2	5	2	9	2	9	11	29	1	1	1	3	9	1
Cercocebus_atys	7	17	2	7	103	53	213	49	2	7	2	9	2	10	12	30	1	2	1	3	9	1
Cebus_capucinus_imitator	5	18	2	3	104	54	215	48	2	6	2	9	2	10	11	30	1	2	1	3	9	1
Callithrix_jacchus	5	16	4	5	97	46	212	48	2	6	2	9	2	10	12	30	1	2	1	3	9	1
Chlorocebus_sabaeus	8	17	2	4	103	57	216	48	2	6	2	9	2	10	12	30	1	2	1	3	9	1
Carlito_syrichta	2	16	2	3	77	44	188	47	2	4	2	8	2	8	10	29	1	2	1	3	9	1
Gorilla_gorilla	6	17	2	4	93	54	215	50	2	5	2	9	2	9	11	30	1	2	1	3	9	1
Homo_sapiens	7	18	2	4	103	49	215	48	2	7	2	9	2	12	12	30	1	2	1	3	9	1
Macaca_fascicularis	8	18	2	3	101	51	216	48	2	6	2	9	2	10	12	31	1	2	1	3	9	1
Mandrillus_leucophaeus	7	17	2	4	85	43	209	49	2	4	2	9	2	8	10	30	1	1	1	3	9	1

Table 2.2 TF family size among 96 mammalian species (continued)

	Pou	RFX_D	RHD_[	Runt	SAND	SRF.T	STAT_	T.box	TEA	TF_AP	TF_Ot	THAP	TIG	TSC22	Tub	zf.BED	zf.C2H	zf.C2H	zf.C4	zf.LITA	zf.MIZ	zf.NF.X
Myotis_brandtii	15	8	10	3	8	4	8	17	4	6	3	8	23	4	5	9	497	6	45	2	7	2
Myotis_davidii	15	8	10	3	8	4	7	16	4	5	3	8	25	4	5	12	484	6	44	2	6	2
Myotis_lucifugus	15	8	10	3	7	5	7	19	2	5	2	9	23	4	4	10	502	6	43	3	7	2
Miniopterus_natalensis	12	8	10	3	8	5	7	17	4	5	3	9	24	5	5	12	459	6	42	2	7	2
Pteropus_alecto	15	9	10	3	7	5	7	16	4	5	3	12	24	4	5	11	484	6	45	3	7	2
Pteropus_vampyrus	14	9	11	3	8	5	7	17	3	5	3	12	26	4	4	11	505	6	47	2	7	2
Rousettus_aegyptiacus	14	9	11	3	9	5	7	17	4	5	3	12	26	4	5	12	501	6	47	3	7	2
Sarcophilus_harrisii	14	7	10	2	4	4	7	17	2	5	2	7	23	3	5	10	371	7	46	2	6	2
Galeopterus_variegatus	15	8	10	3	7	5	7	17	4	5	3	12	25	5	5	15	613	7	46	2	7	2
Monodelphis_domestica	13	9	10	3	5	5	6	19	2	5	3	8	23	5	5	16	389	7	45	2	7	2
Condylura_cristata	15	9	9	3	4	5	7	16	4	5	3	11	23	5	5	8	419	6	45	2	6	2
Erinaceus_europaeus	14	7	10	3	3	5	7	16	4	5	3	9	23	4	5	9	395	7	46	2	6	2
Sorex_araneus	12	8	10	2	6	5	7	16	4	5	3	9	23	4	5	15	415	6	46	3	6	2
Oryctolagus_cuniculus	15	11	11	3	3	4	7	16	4	5	3	10	24	4	5	14	486	6	42	3	6	2
Ochotona_princeps	14	8	10	3	4	5	8	17	4	5	3	10	23	4	5	9	414	6	46	3	7	2
Ornithorhynchus_anatinus	13	6	8	2	5	4	7	18	4	5	1	7	22	4	5	10	238	6	37	3	8	3
Ceratotherium_simum_simum	15	9	10	3	8	5	7	17	4	5	3	10	24	4	4	12	561	6	45	3	6	2
Equus_asinus	13	9	10	3	9	5	7	17	4	5	3	8	23	4	5	13	548	6	46	3	6	2
Equus_caballus	13	8	10	3	8	5	7	17	4	5	2	9	24	4	5	10	534	7	45	2	6	2
Equus_przewalskii	13	9	9	3	7	5	7	17	4	5	2	10	21	6	5	13	555	7	44	2	6	2
Manis_javanica	13	9	10	3	9	5	7	17	4	5	3	11	24	4	5	12	511	5	45	3	6	2
Aotus_nancymaae	15	9	10	3	8	5	8	17	5	5	3	11	24	4	5	14	635	6	44	2	6	2
Colobus_angolensis_palliatus	15	9	9	3	8	4	8	17	4	5	3	10	23	4	5	14	660	6	46	2	6	2
Cercocebus_atys	15	9	10	3	8	5	7	17	4	5	3	12	23	3	5	12	673	6	46	2	6	2
Cebus_capucinus_imitator	16	8	10	3	9	5	7	17	4	5	3	12	23	4	5	12	636	6	45	2	6	2
Callithrix_jacchus	13	8	10	3	8	6	7	18	4	5	3	11	23	4	5	13	649	6	45	2	6	2
Chlorocebus_sabaeus	15	8	10	3	9	5	7	18	4	5	3	12	23	4	5	14	680	6	46	3	6	2
Carlito_syrichta	15	8	9	3	7	5	9	17	4	5	3	11	22	4	7	10	594	6	44	2	6	2
Gorilla_gorilla	16	9	10	4	8	4	7	18	3	5	3	10	22	4	5	13	677	6	47	2	8	2
Homo_sapiens	16	9	10	3	8	6	7	17	4	5	3	12	31	4	5	15	685	6	46	5	6	2
Macaca_fascicularis	14	9	10	3	9	5	7	17	4	5	3	12	24	5	4	11	683	6	46	3	6	2
Mandrillus_leucophaeus	14	9	10	3	8	4	7	17	4	5	3	10	23	5	5	15	651	6	45	3	6	2

Table 2.2 TF family size among 96 mammalian species (continued)

	AF.4	AP2	ARID	AT_ho	BTB	BTD	bZIP_1	bZIP_2	bZIP_M	CBF	CBFB_	CBFD_	CEP1.	CG.1	CP2	CSD	CUT	DM	E2F_T	EBP	Ets	Forkhe
Macaca_mulatta	5	0	14	1	132	2	38	40	27	3	1	37	0	2	6	8	7	7	11	2	30	45
Microcebus_murinus	6	0	14	1	132	2	38	38	27	3	1	35	0	2	6	8	7	8	10	2	28	45
Macaca_nemestrina	4	0	14	1	129	2	36	38	27	3	1	35	0	2	6	8	7	7	10	2	29	45
Nomascus_leucogenys	5	0	13	0	127	2	39	40	27	3	1	37	0	2	6	9	5	6	11	2	29	36
Otolemur_garnettii	4	0	13	1	130	2	41	40	28	3	1	33	0	2	6	8	7	7	10	2	27	40
Pongo_abelii	5	0	14	1	132	2	36	38	25	3	1	43	0	1	6	9	5	7	11	2	29	39
Papio_anubis	5	0	13	1	126	2	38	40	26	3	1	31	0	2	6	8	6	4	11	2	30	41
Propithecus_coquereli	4	0	14	1	131	2	40	40	28	3	1	40	0	2	6	9	8	8	10	2	30	44
Pan_paniscus	4	0	14	0	129	2	38	41	25	3	1	34	0	2	6	9	6	5	10	2	29	37
Pan_troglodytes	4	0	16	1	132	2	37	39	25	3	1	29	0	2	6	10	7	7	11	2	28	48
Rhinopithecus_bieti	4	0	12	1	131	2	39	41	27	3	1	36	0	2	7	8	7	7	11	2	28	42
Rhinopithecus_roxellana	5	0	13	1	130	2	39	41	29	3	1	38	0	2	7	9	7	8	11	2	31	45
Saimiri_boliviensis	4	0	13	0	127	2	38	39	25	3	1	33	0	2	7	9	6	4	11	2	29	37
Cricetulus_griseus	4	0	13	0	124	2	35	39	24	3	1	23	0	2	6	8	5	4	9	2	26	35
Chinchilla_lanigera	4	0	14	0	129	2	38	41	28	3	1	22	0	2	6	8	7	5	10	2	27	38
Cavia_porcellus	5	0	14	0	130	2	38	42	26	3	1	26	0	2	6	8	6	7	10	2	27	33
Dipodomys_ordii	5	0	14	0	125	2	39	40	29	3	1	34	0	2	6	8	8	5	8	3	26	42
Fukomys_damarensis	4	0	15	0	131	2	37	39	24	3	1	27	0	2	6	9	6	3	9	2	26	35
Heterocephalus_glaber	4	0	14	1	134	2	40	43	29	3	1	37	0	2	6	10	6	7	10	2	27	43
Jaculus_jaculus	5	0	13	0	128	2	41	40	29	3	1	18	0	2	6	8	6	7	10	2	27	42
Ictidomys_tridecemlineatus	4	0	14	1	131	2	40	41	27	3	1	33	0	2	6	10	7	7	10	2	28	42
Mesocricetus_auratus	4	0	14	1	135	1	38	40	29	3	1	26	0	2	6	7	7	6	10	2	26	43
Marmota_marmota_marmota	6	0	13	0	124	2	41	41	27	3	1	29	0	1	7	10	6	8	8	2	28	36
Mus_musculus	4	0	15	1	141	2	38	38	26	3	1	31	0	2	6	9	7	7	10	2	27	44
Microtus_ochrogaster	6	0	14	0	131	3	38	40	29	4	1	30	0	1	6	8	7	7	10	2	27	45
Nannospalax_galili	6	0	13	0	134	2	40	40	27	3	1	31	0	2	6	8	7	7	9	2	26	45
Neotoma_lepida	7	0	13	0	131	2	33	37	24	3	1	21	0	2	7	9	9	5	11	1	27	43
Octodon_degus	5	0	12	1	131	3	39	38	27	3	1	34	0	2	7	7	7	7	9	2	28	40
Peromyscus_maniculatus_bairdii	5	0	14	1	132	2	40	41	28	3	1	26	0	2	6	8	6	7	10	2	26	44
Rattus_norvegicus	5	0	15	0	143	3	41	43	29	3	1	31	0	2	6	9	6	7	10	2	27	44
Tupaia_chinensis	5	0	15	0	131	2	38	38	23	3	1	32	0	2	6	10	5	5	10	2	29	38
Dasypus_novemcinctus	5	0	14	0	125	2	36	38	24	3	1	32	0	1	7	8	6	7	10	2	28	42

Table 2.2 TF family size among 96 mammalian species (continued)

	GAGA	GATA	GCM	GTF2I	HLH	HMG_	Homec	Hormo	HPD	HSF_D	HTH_p	IRF	LAG1.	MBD	MH1	Myb_D	NCU.G	NDT80	Nrf1_D	P53	PAX	PC4
Macaca_mulatta	7	17	2	4	103	51	218	48	2	6	2	9	2	10	12	31	1	2	1	3	9	1
Microcebus_murinus	3	17	2	3	103	49	221	48	2	5	2	10	2	10	12	30	1	2	1	3	9	1
Macaca_nemestrina	7	17	2	4	102	55	218	48	2	6	2	9	2	10	12	31	1	2	1	4	9	1
Nomascus_leucogenys	7	16	2	5	86	52	206	48	2	6	2	9	2	10	10	31	1	2	1	3	9	1
Otolemur_garnettii	3	18	2	4	100	50	206	48	2	3	2	10	2	9	12	32	1	2	1	3	9	1
Pongo_abelii	8	17	3	5	99	52	217	49	2	5	1	10	2	11	12	31	1	2	1	3	10	1
Papio_anubis	8	18	2	5	99	49	217	46	2	5	2	9	2	10	12	31	1	2	1	3	9	1
Propithecus_coquereli	3	17	1	3	103	46	205	48	2	4	2	10	2	8	11	31	1	2	1	3	9	2
Pan_paniscus	5	16	2	4	94	49	212	48	2	6	2	9	2	9	11	30	1	2	1	3	9	1
Pan_troglodytes	9	18	2	4	105	57	225	48	2	6	2	9	2	11	12	31	1	2	1	3	9	1
Rhinopithecus_bieti	6	17	2	5	104	47	217	49	2	6	2	9	2	9	11	29	1	2	1	3	9	1
Rhinopithecus_roxellana	9	17	2	4	103	54	221	48	2	7	2	9	2	9	11	30	1	2	1	3	9	1
Saimiri_boliviensis	5	17	2	3	94	43	207	49	2	5	2	8	2	9	11	31	1	2	1	3	9	1
Cricetulus_griseus	3	23	2	3	81	42	172	48	2	6	2	9	2	10	11	31	1	2	1	3	9	2
Chinchilla_lanigera	3	17	2	4	97	48	205	49	2	6	2	8	2	10	12	31	1	2	1	3	9	1
Cavia_porcellus	2	24	2	3	92	47	205	48	2	5	2	9	2	10	10	32	1	2	1	3	9	1
Dipodomys_ordii	4	22	2	3	91	47	201	48	2	5	2	9	2	11	12	30	1	2	1	3	9	2
Fukomys_damarensis	2	14	2	3	77	44	186	48	2	5	2	9	2	9	11	30	1	2	1	3	9	1
Heterocephalus_glaber	2	19	2	4	98	50	200	52	2	5	2	9	2	10	13	33	1	2	1	3	9	1
Jaculus_jaculus	3	14	1	4	99	49	205	49	2	3	2	10	2	9	12	30	1	2	1	3	9	1
Ictidomys_tridecemlineatus	2	18	3	3	102	44	211	49	2	6	2	10	2	10	12	32	1	2	1	3	9	2
Mesocricetus_auratus	3	21	2	5	96	49	207	44	2	5	1	9	1	9	12	28	1	2	1	3	9	1
Marmota_marmota_marmota	1	20	3	3	93	41	201	48	2	7	2	10	2	10	11	33	1	2	1	3	8	2
Mus_musculus	3	35	2	3	100	50	248	49	2	6	2	9	2	11	12	31	1	2	1	3	9	1
Microtus_ochrogaster	2	16	2	3	104	49	209	48	2	6	2	9	3	10	12	31	1	2	1	3	9	1
Nannospalax_galili	3	16	2	3	102	49	211	49	2	6	2	10	2	10	12	31	1	2	1	3	9	1
Neotoma_lepida	1	13	2	3	89	54	206	49	2	5	2	9	2	6	10	27	0	2	1	4	9	2
Octodon_degus	4	17	2	3	102	50	207	49	2	5	2	9	3	10	12	32	1	2	1	3	9	2
Peromyscus_maniculatus_bairdii	2	28	2	3	105	50	220	50	2	5	2	9	2	10	12	32	1	2	1	4	9	1
Rattus_norvegicus	2	23	2	3	102	53	229	51	2	6	2	9	3	10	12	31	1	2	1	3	9	1
Tupaia_chinensis	4	18	2	3	94	49	215	48	2	6	2	10	2	9	12	31	1	2	1	3	9	3
Dasypus_novemcinctus	8	16	2	4	97	50	195	48	2	9	2	11	2	10	12	33	1	2	1	3	9	2

Table 2.2 TF family size among 96 mammalian species (continued)

	Pou	RFX_D	RHD_[	Runt	SAND	SRF.T	STAT_	T.box	TEA	TF_AP	TF_Ot	THAP	TIG	TSC22	Tub	zf.BED	zf.C2H	zf.C2H	zf.C4	zf.LITA	zf.MIZ	zf.NF.X
Macaca_mulatta	15	9	10	3	9	5	7	18	4	5	3	13	25	4	5	15	709	6	46	3	6	2
Microcebus_murinus	15	9	10	3	9	5	7	17	5	5	3	12	25	4	5	14	590	6	46	3	6	2
Macaca_nemestrina	15	8	10	3	7	5	7	17	4	5	3	11	23	4	5	10	680	6	45	3	6	2
Nomascus_leucogenys	12	8	10	3	7	4	8	17	4	5	2	12	23	4	5	14	680	6	46	2	6	2
Otolemur_garnettii	14	8	10	3	5	5	7	17	4	5	3	11	24	4	5	11	562	7	44	3	7	2
Pongo_abelii	16	10	9	3	6	5	7	17	4	5	3	12	24	5	4	13	720	5	45	2	6	2
Papio_anubis	16	9	10	3	8	5	7	17	4	5	3	12	23	5	5	14	674	6	45	4	6	2
Propithecus_coquereli	14	8	10	3	7	5	7	17	4	5	3	12	24	4	5	11	576	6	47	2	7	2
Pan_paniscus	16	8	10	3	8	5	7	17	4	5	3	10	23	4	5	12	672	6	47	2	6	2
Pan_troglodytes	16	9	10	3	9	5	7	18	5	5	3	12	24	4	5	14	723	6	47	2	6	2
Rhinopithecus_bieti	16	9	10	3	7	5	9	17	4	5	3	10	24	4	5	12	693	6	42	3	6	2
Rhinopithecus_roxellana	15	9	10	3	8	5	7	17	4	5	3	11	24	4	5	15	688	6	46	2	6	2
Saimiri_boliviensis	13	9	10	3	7	5	7	17	4	6	3	9	24	4	5	14	603	6	45	2	6	2
Cricetulus_griseus	9	9	10	3	9	5	8	17	4	5	3	8	26	4	5	12	444	8	41	2	6	2
Chinchilla_lanigera	13	9	10	3	8	5	7	17	4	5	3	11	23	7	5	11	493	5	46	5	7	2
Cavia_porcellus	14	9	10	3	7	4	7	16	4	5	3	9	23	6	5	16	478	5	45	2	8	2
Dipodomys_ordii	15	9	10	3	7	5	8	17	4	5	3	7	22	4	5	13	419	5	46	2	7	2
Fukomys_damarensis	15	8	10	3	6	4	6	17	4	5	3	11	23	4	5	10	470	5	46	3	6	2
Heterocephalus_glaber	15	9	10	3	6	5	6	17	4	5	3	11	23	6	4	18	503	6	50	3	7	2
Jaculus_jaculus	14	8	10	3	6	6	7	17	4	5	3	8	24	4	5	12	440	6	47	3	7	2
Ictidomys_tridecemlineatus	13	9	10	3	7	5	7	17	4	5	3	11	24	4	5	22	516	6	45	3	7	2
Mesocricetus_auratus	15	9	9	3	8	5	7	17	4	5	3	7	22	4	5	12	439	7	46	2	6	2
Marmota_marmota_marmota	13	8	9	3	8	5	7	17	4	5	3	12	23	4	5	16	531	5	44	3	5	3
Mus_musculus	15	9	10	3	10	5	7	17	4	5	3	7	32	4	6	14	586	6	47	3	6	2
Microtus_ochrogaster	15	9	10	3	8	5	7	17	4	5	3	8	24	5	5	9	450	5	45	2	6	2
Nannospalax_galili	15	9	10	3	9	5	7	17	4	5	3	9	24	6	5	13	531	6	46	3	6	2
Neotoma_lepida	17	6	9	3	5	5	7	17	3	5	2	8	18	3	5	4	283	5	39	4	4	1
Octodon_degus	16	8	9	3	7	5	7	17	4	5	3	11	24	6	5	12	476	5	45	2	6	2
Peromyscus_maniculatus_bairdii	13	8	10	3	8	5	7	17	4	5	3	9	24	4	6	12	471	6	46	3	6	2
Rattus_norvegicus	14	8	10	3	5	5	7	17	4	5	3	8	28	4	5	11	517	6	48	2	6	2
Tupaia_chinensis	15	9	10	3	7	5	7	17	4	5	3	11	24	4	5	10	503	6	45	3	6	2
Dasypus_novemcinctus	13	9	10	3	5	5	7	17	4	5	3	11	23	4	5	11	542	6	43	2	6	2

2.2 Inference of the species tree, the gene tree, the TF family trees and TF orthologous group trees

In the first project, species tree downloaded from TIMETREE database [30]. Unmatched species among 96 mammals were fitted by their closely related species to provide consistence on divergent time with other clades. This species tree provided both tree topology of mammalian species and divergence times. In the second project, the species tree was estimated by integrating gene trees by multi-species coalescent model, following Wu et al. [31]. In total, 823 single-copy genes were used to infer a species tree based on the coalescent-based Njst method that takes account of lineage sorting due to ancestral polymorphism [32]. The topology of the inferred species tree was consistent with that of Tarver et al. (2016), who placed treeshrew (Tupaia chinensis) as the root lineage of Glires [33]. The phylogenetic position of treeshrew is not yet resolved, however, and several researchers consider treeshrew to be the root lineage of Euarchonta, not Glires [34]. As an alternative species tree in this study, I consequently used a tree in which the position of treeshrew was fixed at the root of Euarchonta rather than Glires. Following the same method as Wu et al. 2017, I got the divergence times of 96 mammals based on the inferred branch effect (the product of genomic rate and time) and fossil calibrations [31]. I used the Atlantogenata topology in main research, while Exafroplacentalia and Epitheria topology as complementary analysis (Figure 2.4-2.6).

I estimated the maximum likelihood tree for each gene using IQ-TREE software [35], which automatically performed model selection and determined the best data partitions. The best evolutionary model for each gene was independently selected based on the Bayesian information criterion and used for inference of the nucleotide tree. All gene trees were calculated using 1,000 bootstrap replicates.

Sequences of each TF family from the 96 mammalian species were pooled together and aligned using MAFFT7 [36] and MUSCLE3.8 [37]. The aligned datasets were imported into DAMBE5 [38], converted to MEGA format, and used to construct phylogenetic trees of mammalian TF families in MEGA6 [39]. Among them, only 48 neighbor-joining trees of TF families had small member size and could be constructed.

As the contents of this chapter (page) are anticipated to be published in a paper in a scholarly journal, they cannot be published online. The paper is scheduled to be published within 5 years.

2.4 Protein-protein interaction network and TF-to-TF network

2.4.1 network construction

Mice (Mus musculus) and rats (Rattus norvegicus) are closely related species that diverged 20.9 million years ago [30]. All differences between mouse and rat networks can be assumed to have arisen recently. I therefore used mouse and rat data to detect factors that affect network evolution. Humans and mice are more genetically and phenotypically diverged, and much research has been conducted on these two species. I thus looked for human and mouse phenotype and expression differences caused by network changes. Whole protein network data of humans, mice, and rats (from STRING [40]) were used to construct PPI networks for these species. Within each network, all interactions based on 5 main sources (Figure 2.9) and had confidence scores ≥ 0.4 (medium + high confidence) from STRING. Global PPI networks for mice (19,505 nodes and 847,065 edges), rats (19,920 nodes and 1,099,355 edges), and humans (18,720 nodes and 782,253 edges) were then constructed (Figure 2.10).

In human, 1555 TFs had TF interactions or non-TF interactions. To detect isolated TFs (TF with only non-TF interactions or disconnected from the main TF group), these 1555 human TF nodes and TF-to-TF interactions were used to construct the TF-to-TF network. STRING collects protein-protein interactions based on multiple types of evidence: co-expression, high-throughput laboratory experiments, previous knowledge in databases, genomic context predictions and automated text-mining. For network construction, I adopted interactions when there was any evidence regarding the type of interaction. If there is noise in the database, the networks may include false positive interactions but the chance of false negatives is minimized to give reliable information on isolated TFs. I also constructed the TF-to-TF networks of human, mice and rat by STRING.

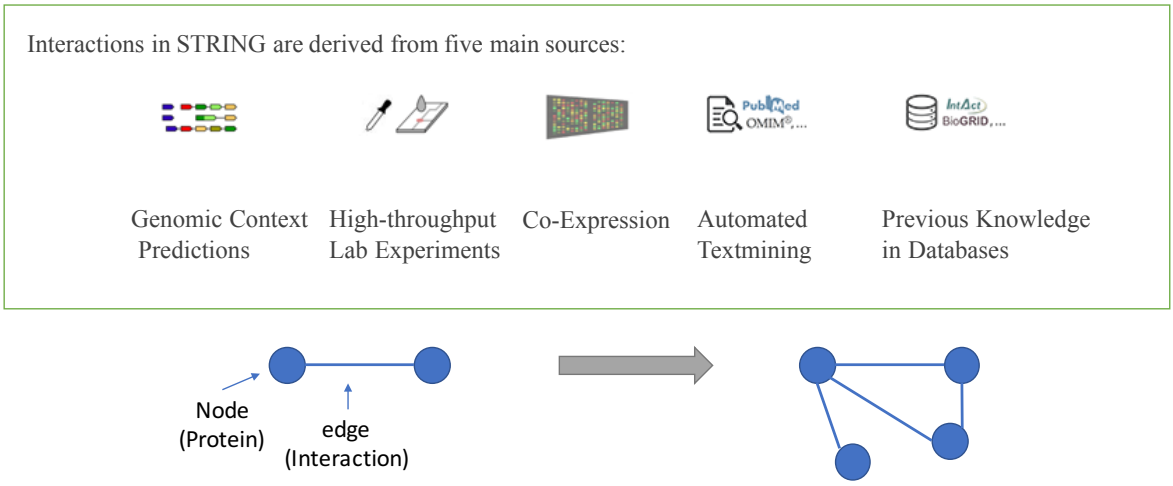

Figure 2.9 Protein-protein interactions (PPIs) in STRING database. Five main sources of PPIs from STRING database listed in green box. Below is an example of how interactions are built into a network. The blue circles are nodes and the lines are interactions.

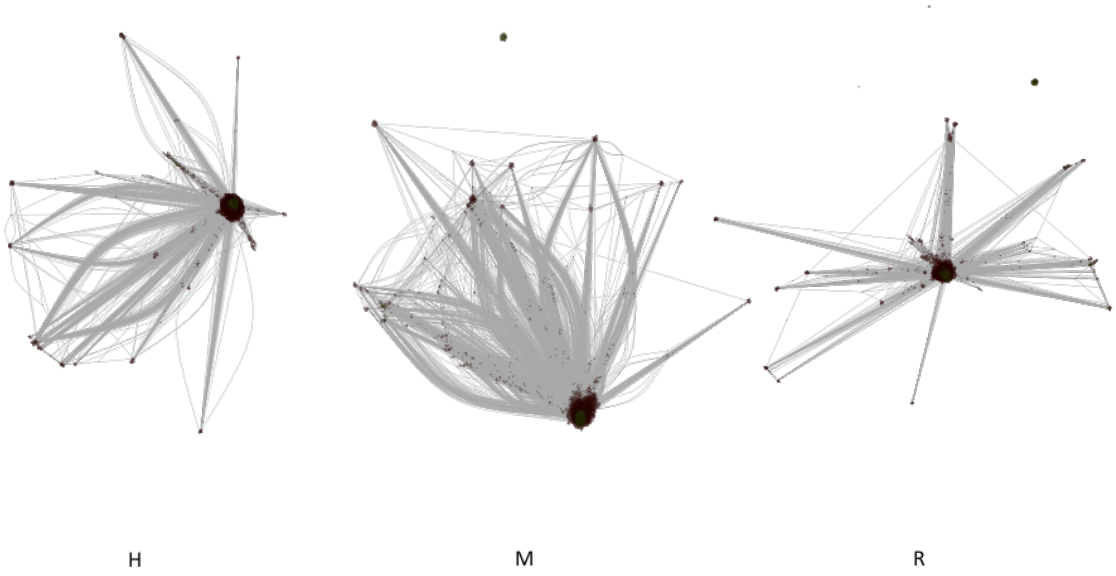

Figure 2.10 PPI network of human, mouse and rate by R package. H:human , M:mouse, R:rat. Black points are nodes; Gray lines are interactions. These shows the overall shape of PPI network.

2.4.2 Edge change ratios of transcription factors in PPI network

Calculate the number of edges in rats and mice for each node by cytoscape3.5.1 (Figure 2.11). Through the TF database and blast data, the TF list of rats and mice was manually compared, and then transcription factors in the two species were divided into 3 categories. TF loss: TF gene is lost in one of the species. DBD loss: TF gene exists in both species, but DBD is lost in one species. Others (no loss): TF gene exists in both species and its DBD is not lost. The number of edges of transcription factor i in mice is M_i, and the number of edges in rats is R_i. The edge change ratio C_i of transcription factor i between rat and mouse, was calculated as:

$$C_i = \frac{|R_i - M_i|}{|R_i + M_i|}$$

Then count the edge change ratios of transcription factors in each group (Figure 2.12).

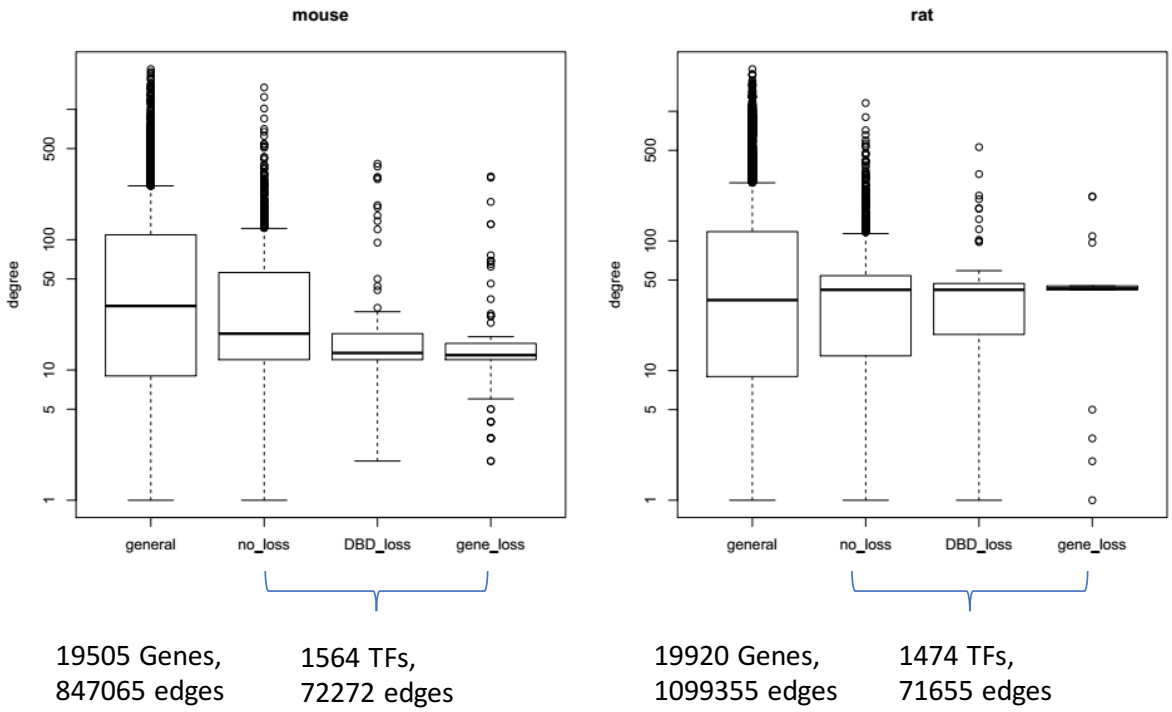

Figure 2.11 Edge number of each nodes in mouse and rat. Gene_loss: TF gene is lost in one of the species. DBD_loss: TF gene exists in both species, but DBD is lost in one species. Others (no_loss): TF gene exists in both species and its DBD is not lost. General: All genes.

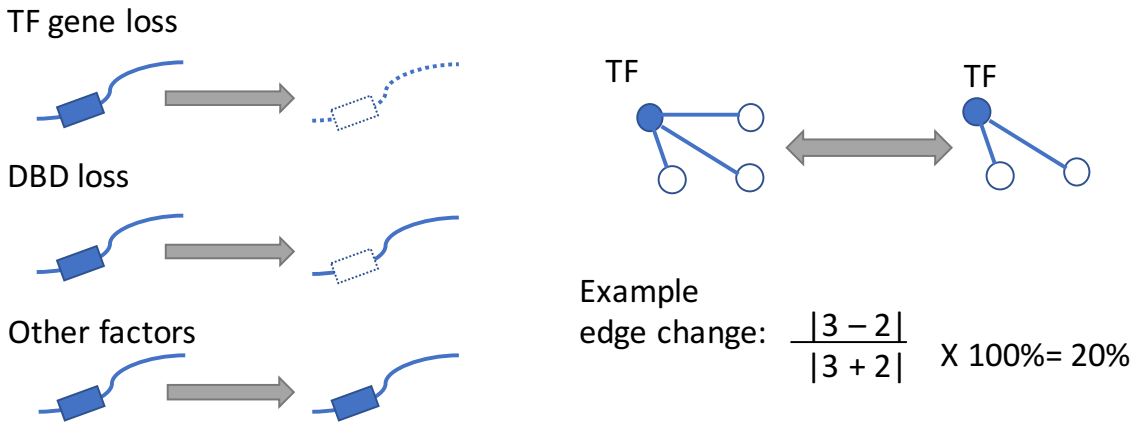

Figure 2.12 Example of edge change on Loss of TFs, loss of DBDs and the loss of interactions. On the left, the line is the sequence; the dashed line is the sequence loss; the blue rectangle is the DBD; and the dashed box is the DBD loss. On the right, the blue ball is TF; the line is edge; and the blue ring is the advanced nodes of TF.

2.4.3 Functional Cartography of the Human PPI Network

Using a previously published functional cartography protocol [41], I characterized each gene in the human PPI network according to its within-module degree z-score (z) and participation coefficient (p). The within-module degree z-score of node i, z_i, was calculated as:

$$z_i = \frac{k_i - \overline{k_{s_i}}}{\sigma_{k_{s_i}}}$$

where k_i is the number of links between node i and other nodes in its module, k_{s_i} is k averaged over all nodes in s_i , and $\sigma_{k_{s_i}}$ is the standard deviation of k in s_i. The participation coefficient of node i, p_i , was calculated as:

$$p_i = 1 - \sum_{s=1}^{N_M} (\frac{k_{is}}{k_i})^2$$

where k_{is} is the number of links between node i and other nodes in module s, and k_i is the total degree of node i.

Genes were classified into eight groups: (1) those with no experimental interactions, (2) ultra-peripheral nodes ($z < 2.5$ and $p < 0.05$), (3) peripheral nodes ($z < 2.5$ and $0.05 \leq p < 0.625$), (4) non-hub connector nodes ($z < 2.5$ and $0.625 \leq p < 0.8$), (5) non-hub kinless nodes ($z < 2.5$ and $p \geq 0.8$), (6) provincial hubs ($z \geq 2.5$ and $p < 0.3$), (7) connector hubs ($z \geq 2.5$ and $0.3 \leq p < 0.75$), and (8) kinless hubs ($p \geq 0.75$).

2.5 Gene expression

2.5.1 Negative binomial regression analysis of the effect of TF membership variation on gene expression

Gene expression data of 15,796 orthologous human and mouse genes in five organs (cerebellum, heart, kidney, liver, and testis) were retrieved [42, 43]. After standardization of these data as transcripts per kilobase per million reads, similar average expression levels were observed in each organ between humans and mice. To analyze the effect of the variation in the membership of TF families on the expression of their interacting genes, all orthologous genes were separated into five types: (1) genes without TF interactions, (2) genes with orthologous TF interactions, (3) genes with interactions with human- and mouse-specific TFs, (4) genes with interactions with human-specific TFs absent in mice, and (5) genes with interactions with mouse-specific TFs absent in humans. For each species and organ, I estimated gene expression profiles by negative binomial regression:

$$\log E[\text{expression}|\text{gene type} = C_k] = \alpha + \beta_k$$

using glm.nb in the R package MASS [44, 45]. In this equation, the coefficient β_k is the log mean expression of other groups relative to the reference group (genes without TF interactions).

2.5.2 The effect of TF loss on TG expression profiles: human–mouse comparison

Gene expression data of 15,796 orthologous human and mouse genes in five organs (cerebellum, heart, kidney, liver and testis) were retrieved [42]. The average levels of these expression data, which were standardized as transcripts per million kilobases, were similar between humans and mice in these organs. Regulatory information on TFs and their TGs was obtained from TRRUST v.2 [46]. I chose up-regulatory interactions to analyse the effect of TF losses on the expressions of their TGs. According to the TF list in the TRRUST database and the human and mouse TF lists in my database, the TFs in TRRUST that can up-regulate the expression of TGs were divided into two types that exist in the species and those that are lost. According to the data of these interactions, I got the corresponding TG. Through the changes in the expression of TG in these two groups, the overall impact of the loss of transcription factors on the expression of target genes were detected.

As the contents of this chapter (page) are anticipated to be published in a paper in a scholarly journal, they cannot be published online. The paper is scheduled to be published within 5 years.

2.9 TF-GO bipartite graphs for humans and mice

Gene ontology (GO) data of human and mouse TFs were retrieved [47, 48]. The intersection of every TF associated with a GO term was checked between humans and mice, and the proportion of intersecting TFs relative to the average number of TFs was obtained by local polynomial regression using loess in R [45, 49].

2.10 Data Availability

Protein sequences: NCBI [50]; http://www.ncbi.nlm.nih.gov

DNA binding domain (DBD) models: Pfam [19]; https://pfam.xfam.org

Orthologous groups: OrthoDB; www.orthodb.org

Species tree: TIMETREE [30]; http://www.timetree.org

Protein interaction data: STRING [40]; https://string-db.org

Gene ontology data: (1) Gene Ontology Consortium [47], http://www.geneontology.org and (2) g:Profiler [48], https://biit.cs.ut.ee/gprofiler

Gene name data: DAVID [51]; https://david.ncifcrf.gov

Phenotypic data: MGI [52]; http://www.informatics.jax.org

Life history traits: ADW; http://animaldiversity.org

Chapter 3 Multiple isolated transcription factors act as switches and contribute to species uniqueness

3.1 TF database compared with the existing databases

To detect the accuracy of TF annotation, I compared our 1625 human TFs list (TF list extracted from Table S3.1 that collects 140,821TFs in 96 mammalian genomes) with 2 well-known TF databases, AnimalTFDB3 and HumanTFs. In humanTFs, there are 1639 TFs listed. And in AnimalTFDB3, 1665 human TF listed (Figure 3.1). 1402 TFs listed in all 3 databases, 82 TFs only in HumanTFs, 123 TFs only in AnimalTFDB3, 140 TFs only in our human TF list. My database and the others are based on the similar pipeline by DBD and HMMER. AnimalTFDB3 and HumanTFs use human genome in ensemble and I use NCBI's human genome. The DBD list (even the number of DBD) is different among these three databases. These may lead to the differences among human TF numbers and lists.

AnimalTFDB3 contains 125,135 TFs from 97 genomes, which ranges from Caenorhabditis elegans to mammal species like human. My database contains 140,821 TFs, focusing on 96 mammal species. This may provide a better solution in mammal's TF evolution history. To avoid the limitation of mammalian species, I further annotated our database with orthologous groups information by OrthoDB. By this way, it is possible to trace the TFs in mammalian species back to those of bacterial species.

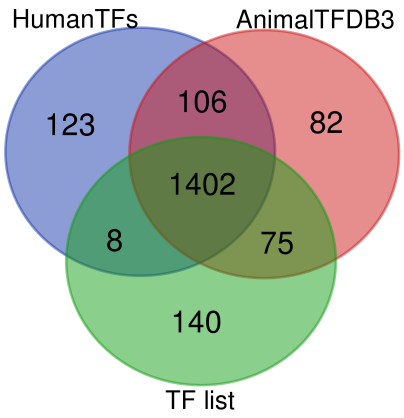

Figure 3.1 Venn diagram of human TF lists in HumanTFs, AnimalTFDB3, and our database. The numbers in the red, green, and blue circles are the number of human transcription factors in AnimalTFDB3, our database and HumanTFs. The number of overlapping parts of the circles is the number of overlapping parts of transcription factors in different databases.

3.2 TF families vary greatly in scale among mammalian species

TF families vary in scale because of gene duplication and loss, as well as the loss of DBDs. To examine variation in the membership of TF families, I de novo detected 140,821putative TFs belonging to 66 TF families in 96 mammalian genomes (Supplementary Materials, Table S3.1). The total number of TFs varied substantially among species (Figure 3.2). For example, Neotoma lepida had 1337 TFs, whereas a closely related species, Peromyscus maniculatus bairdii, had 1628. Using a standardized number of each TF family as a control, I observed that variation in membership was also very widespread among these TF families (Figure 3.3a). I examined the correlation between TF families and found that 97.9% of TF family pairs (1973 out of 2016) were not strongly correlated ($r < 0.5$). This result indicates that number variations in each TF family tend to be independent of other families. In the TF-family correlation matrix and heatmap shown in Figure 3.3b, only three small clusters have members that are strongly correlated with one another: (1) bZIP_1, bZIP_2, and bZIP_Maf; (2) BTD and LAG1_DNAbind; and (3) HMG_box, BTB, Homeobox, Forkhead, and HLH. bZIP_1, bZIP_2, and bZIP_Maf are all present in 14 mouse TF genes, while BTD and LAG1_DNAbind are both located in two mouse TF genes. Two members of cluster 3, HMG_box and HLH, are both found in the gene encoding protein S9YBX2. In other words, these strong correlations mostly result from genes sharing multiple DBDs rather than the co-occurrence of gene duplications or losses.

Figure 3.2 Total number of TFs in 96 mammalian species. The black bar is the total number of TFs in a species. The species trees are time trees from TimeTree. The adjacent strips of different colors on the species tree are different lineages.

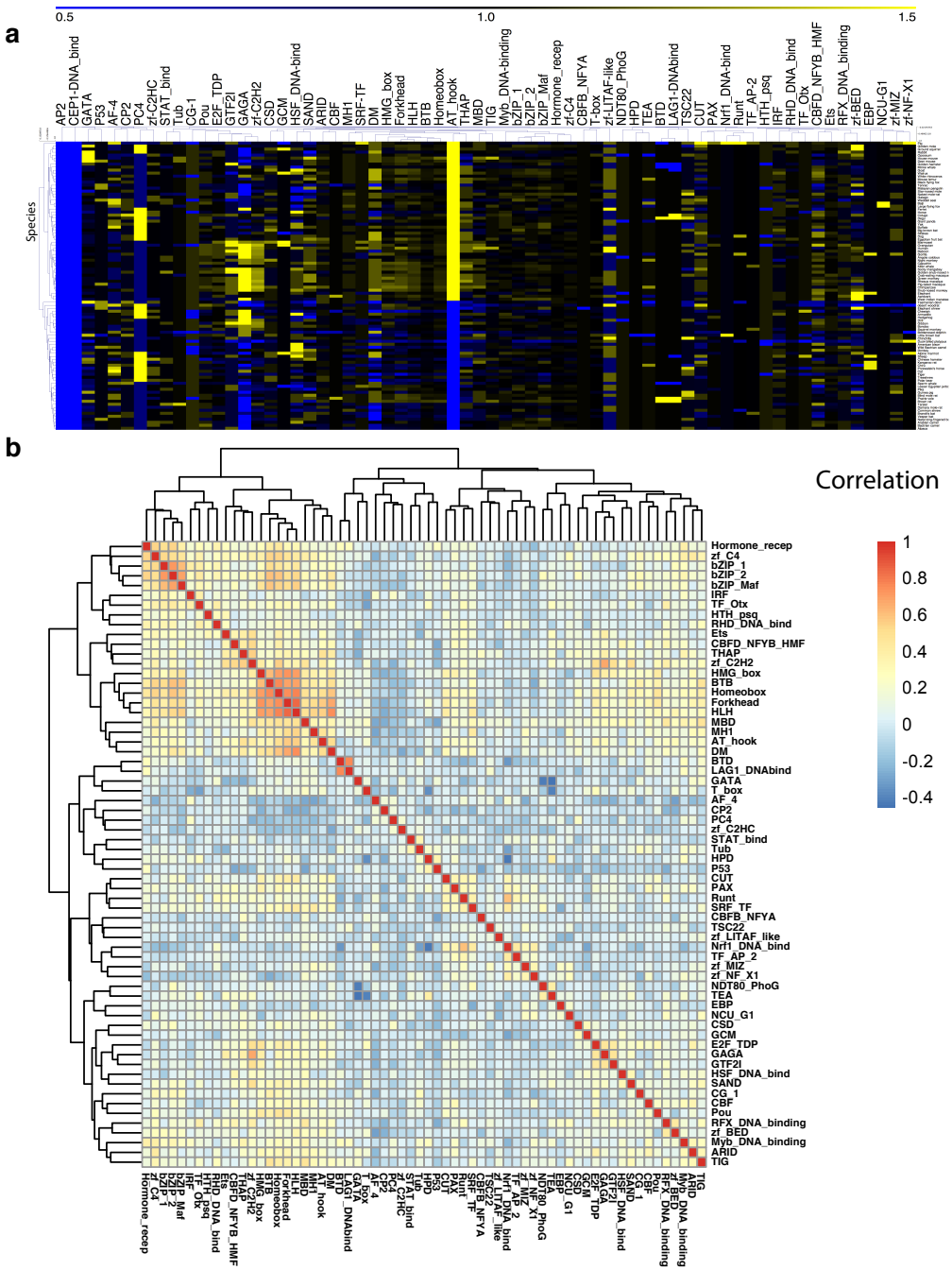

Figure 3.3 Variation in the number of transcription factors (TFs) within and among TF families. The dendrograms show the hierarchical relationships. (a) Variation in the number of TFs within each TF family. X-axis: TF family; y-axis: mammalian species. The average number of TFs in each TF family was standardized to 1 (black). The colors on the heat map represent the degree of TF number variation, where blue is low and yellow is high. (b) Correlation of TF number variation among TF families. The colors on the heat map represent the degree of correlation (blue, low; red, high).

Large TF families, such as C2H2, have been found to rapidly diversify. Families with limited members are usually thought to be more conserved and are less researched. To reveal the detailed history of variation in TF family membership, phylogenetic trees of 48 small size TF families were reconciled with the mammalian species tree. The membership of different TF families was found to have changed along nearly all branches of the mammalian species tree (Figure 3.4-3.6). Compared with the common mammalian ancestor, an average of 37.8% of the TFs of a mammalian species arose during its evolution, whereas 15.0% disappeared. This high level of turnover, more than half of the TFs of a species, indicates that TF families have generally undergone substantial alteration through isolated TFs. Unlike TF orthologs [53], these TF families as a whole are not as conserved as previously thought. TF formation and loss have occurred even more extensively along recent branches. These TF formation and loss events have shaped the unique TF profile of each species. Among 48 TF families (Supplementary Materials, Table S3.2), abundant gains and losses have taken place in families such as GATA and Forkhead. Members of the GATA TF family, which include more than 15% of all gained TFs, are inducers of the pluripotency reprogramming and may serve as important mediators of cell fate conversion (Figure 3.4). The Forkhead TF family, which includes 14.5% of all lost TFs, regulates cell growth, proliferation, differentiation, and longevity (Figure 3.5). The functional importance of TFs is therefore not dependent of evolutionary conservation. TF gains and losses have been prevalent during mammalian evolution (Figure 3.6). Since the software Notung only provides event numbers, I could not check the proteins that experienced the events in detail. To obtain a clear picture on the effect of TF losses, I focused on human and mice and conducted quantitative analysis. I will try to find better ways to apply quantitative analysis to whole mammal species in future research.

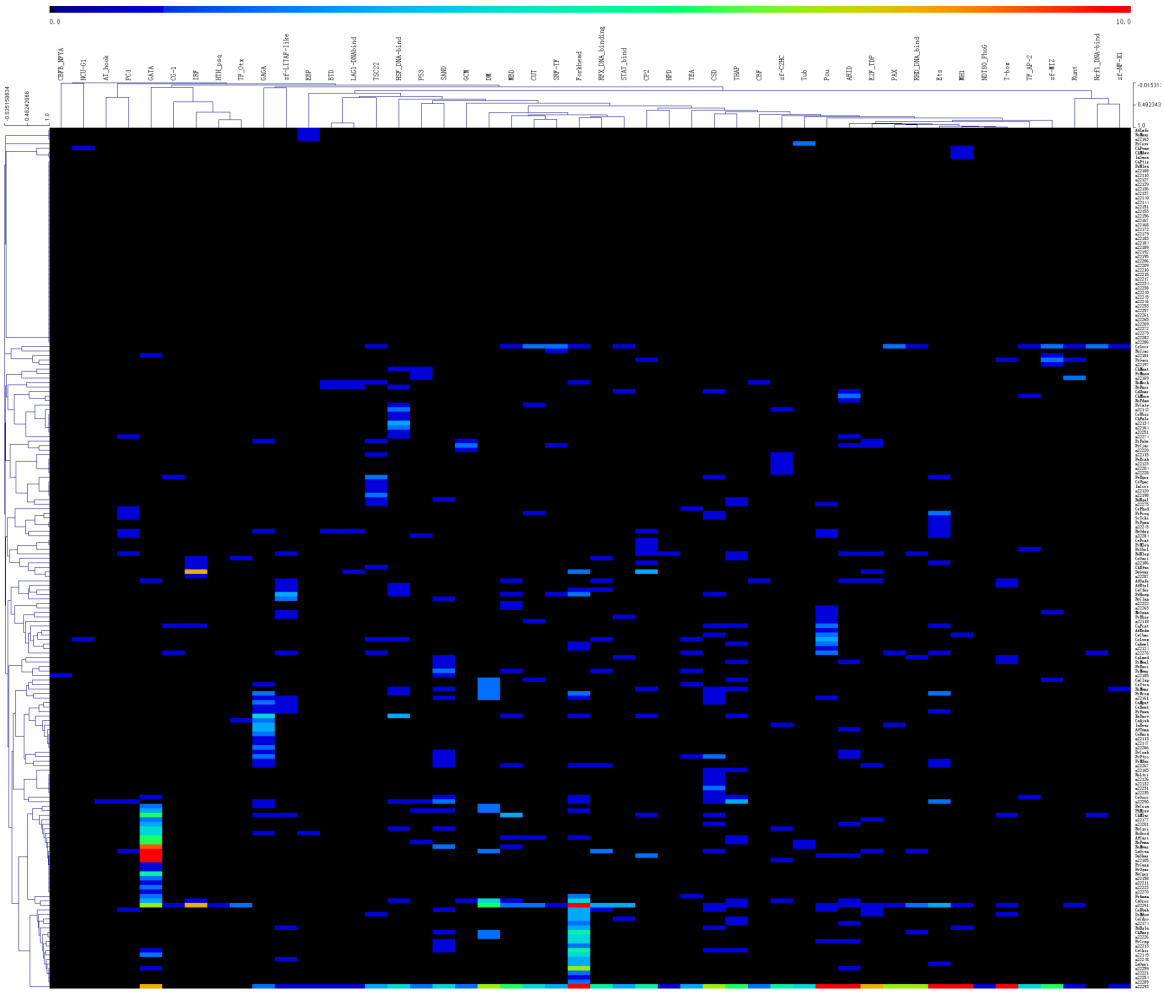

Figure 3.4 TF gains among 48 TF families. X-axis: TF family; y-axis: mammalian species. Tree: hierarchical clustering. From blue to red, the number of events increase.

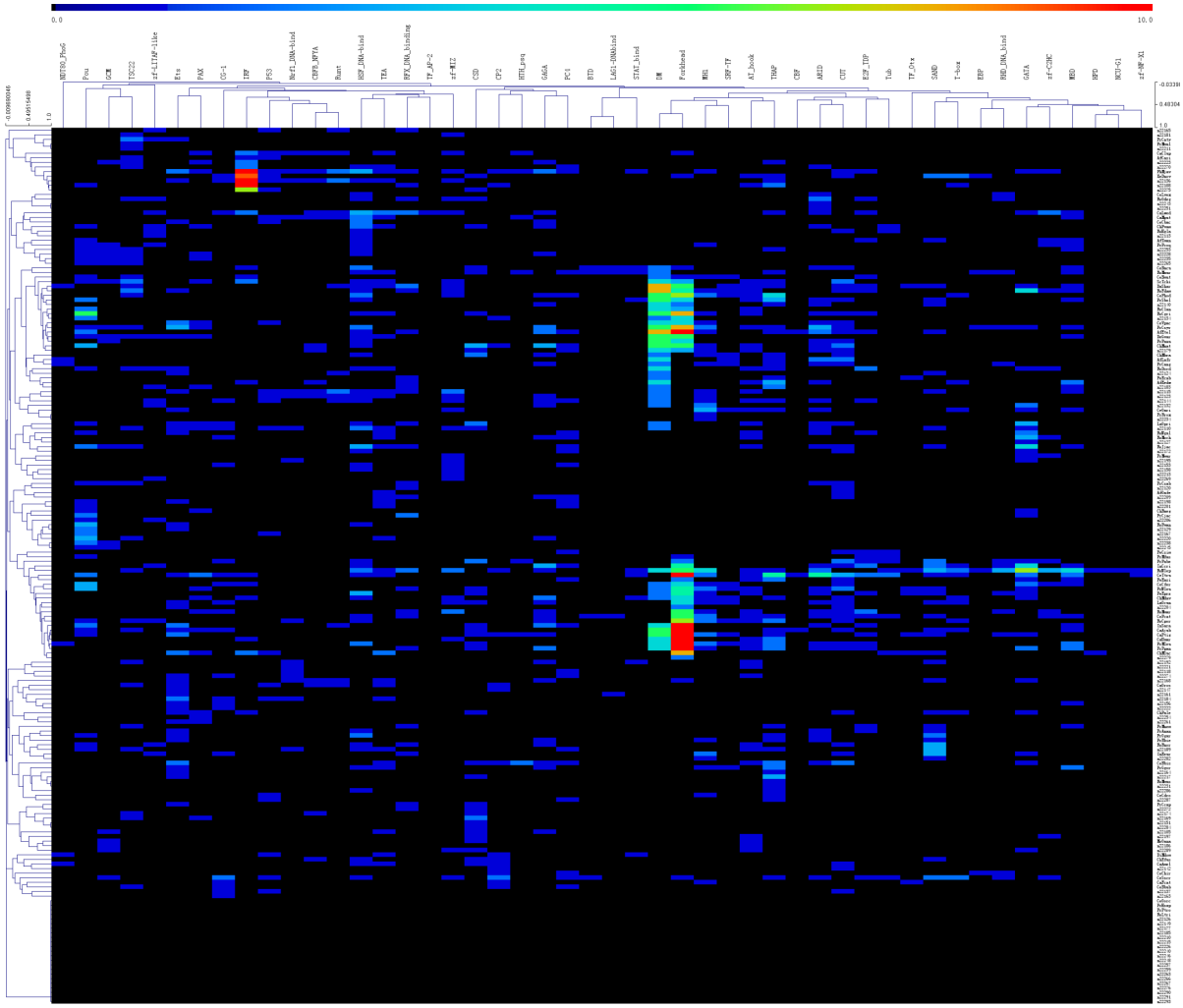

Figure 3.5 TF loss among 48 TF families. X-axis: TF family; y-axis: mammalian species. Tree: hierarchical clustering. From blue to red, the number of events increase.

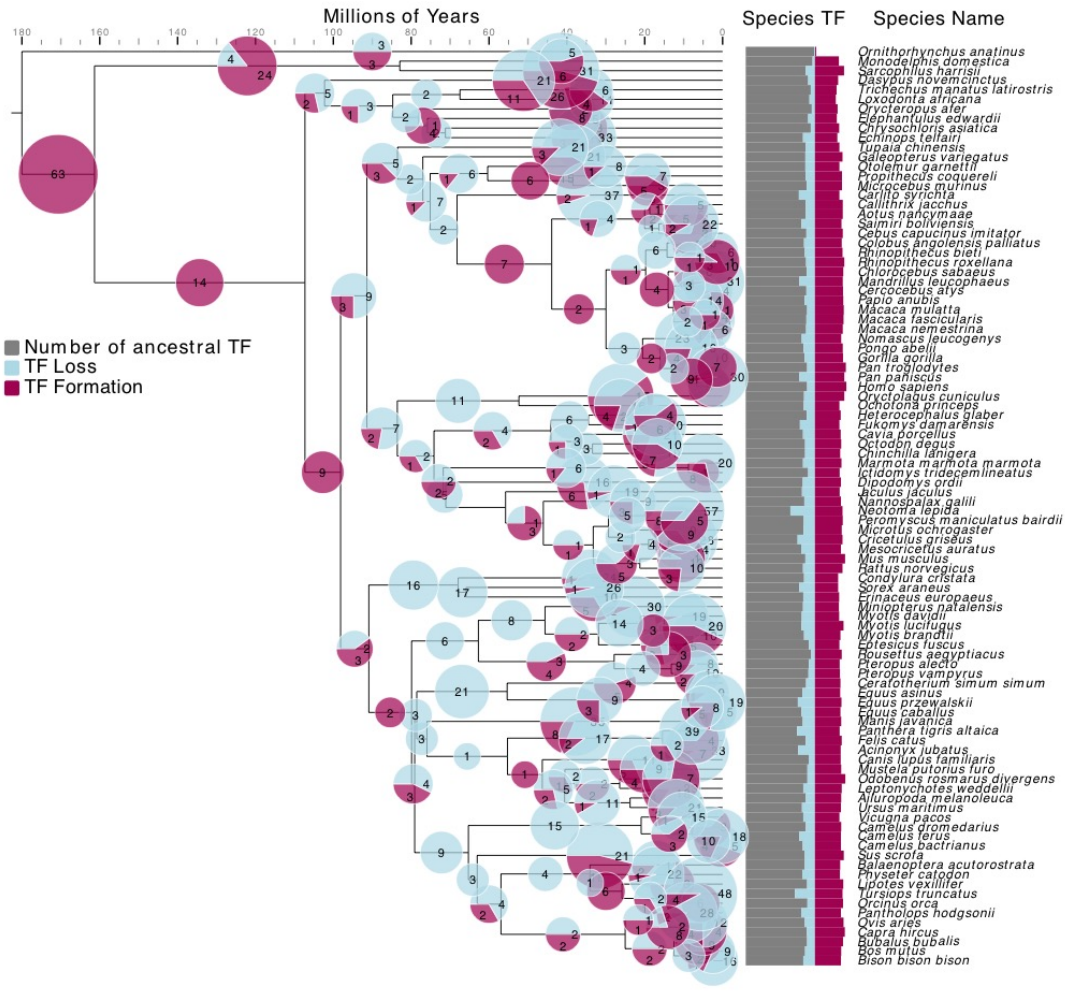

Figure 3.6 Atlas of formation and loss events in 48 transcription factor (TF) families from 96 mammalian species over 177 million years. The size of each pie chart is proportional to the number of TF gains and losses on each branch; light blue indicates TF loss events, and red indicates TF formation events. The bar chart displays the total number of TF gains (red) and losses (blue) in each species over 177 million years. The gray bars indicate ancestral TFs. Species tree is from TimeTree.

3.3 Isolated TFs in a human TF-to-TF network often have no orthologs in mouse

I constructed a PPI network of nearly all human genes and a TF-to-TF network based on the detected TF list from the whole gene network (see Materials and Methods). Interactions were found between 1555 of the 1625 human TFs in the PPI network. This means these 1555 TFs have been previously investigated and that interactions have been determined with other TFs or non-TFs. One-third (515) were isolated from the other 1040 TFs (no conserved co-expression, high-throughput laboratory experiments, previous knowledge in databases, genomic context predictions or automated text-mining interaction), but were connected with non-TF genes (Figure 3.7, Table 3.1, and Table S3.1). Out of 1040 TFs in the large connected component of the network, 507 (48.8%) were lethal when mutated, and only 40 (3.8%) were not found in mice. In contrast, 26 (5.0%) of the 515 isolated TFs were lethal when mutated, and 189 (36.7%) were absent in mice. The average degree (number of connections) of the 515 isolated TFs in the human gene TF-to-TF network was 10.5 ± 8.8 (mean ± standard deviation), whereas TFs in the large connected component had an average degree of 77.9 ± 127.8. TFs having fewer documented interactions are less conserved and less lethal and are therefore more likely to enable lineage-specific adaptation (reviewed in [9]). Isolated TFs are consistent with the characteristics of this type of TF. Overall, TFs that are isolated in the TF-to-TF network generated TF number variation, and the human TFs absent in mice are more dispensable for TF–TF interactions. I additionally conducted a functional cartographic analysis of all TF and non-TF genes in the human PPI network. TFs were not at the core of the human PPI network, but were on the periphery, even compared with non-TF genes. This observation is consistent with the variable TF profile uncovered when non-orthologous TFs are also considered. However, TFs in the large connected component, which is enriched in orthologous TFs, are evolutionarily conserved. In human TF profile, 229 TFs are different when compared with mice. Among the 229 TFs, 189 belong to isolated TFs. The isolated TFs are largely human-specific; they contribute most to TF profile differentiation, at least among human and mice.

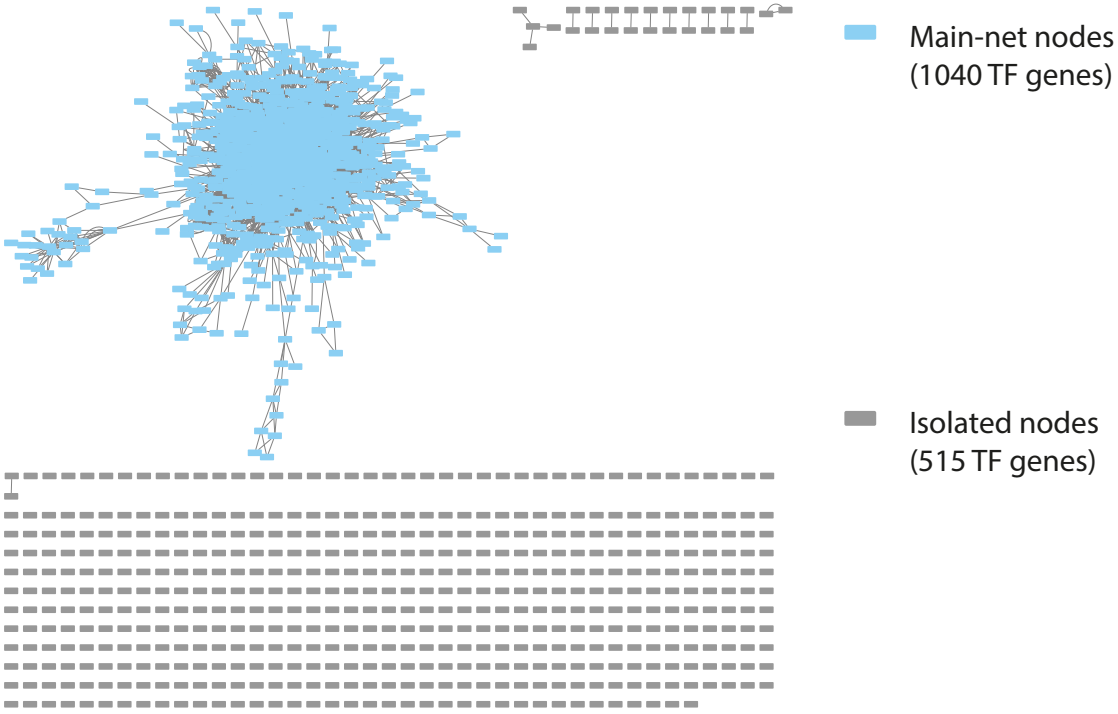

Figure3.7 Human TF-to-TF network that shows interactions between transcription factors. Gray blocks are Isolated TFs; Blue Blocks are Main-net TFs; Lines are TF-to-TF interactions.

Table 3.1. Different features of large-component and isolated transcription factors (TFs).

TF Type	TF Number	TFs with Lethal Phenotype	TFs Absent in Mouse	Degrees
large component TFs	1040	507 (48.8%)	40 (3.8%)	77.9 ± 127.8
Isolated TFs	515	26 (5.0%)	189 (36.7%)	10.6 ± 8.8

Values were acquired by network analysis and TF annotation. Large-component TFs refer to the largest connected component in a TF-to-TF network. Isolated TFs comprise one four-TF component, 12 two-TF components, and other TFs with no TF–TF interactions. Degree indicates the average number of degrees of TFs in a human gene interaction network. The "lethal" phenotype was assigned to genes identified from a search using the keyword "lethal".

3.4 Genes interacting with TFs in humans and mice have similar expression profiles but are more highly expressed in mice

Variation in the membership of TF families influences the PPI network. The formation of TFs adds new edges, while the loss of TF genes removes them. To determine the effect of DBD loss, I compared the global PPI networks of two closely related species, mice and rats. The mouse network contained 19,505 nodes and 847,065 edges, while the rat network consisted of 19,920 nodes and 1,099,355 edges. Within these networks, I focused on the TF subnetworks (1440 and 1288 TF genes in mice and rats, respectively) and their interacting genes. Without considering DBD loss, roughly the same numbers of orthologous TFs were found to interact, with a relative difference of 30.1 ± 22.3% (Supplementary Materials, Table S3.3). When DBD loss was considered, the relative difference in the number of interacting genes increased to 50.7 ± 27.9%. In general, a change in a DBD doubled the variation in the number of interacting genes. By functional cartography of the human PPI network (Figure3.8, Table 3.2), I found transcription factors are more in the periphery of PPI network.

Variation in TF-interacting genes among species may affect their expression profiles. Figure 3.9 shows the expression profiles of orthologous genes in humans and mice (Supplementary Materials, Table S3.4) relative to the expression of non-TF-interacting genes. Generally, the relative expression of TF interacting genes compared to non-TF interaction genes is higher in mice than in humans, although the difference is small in the testis. In the cerebellum and testis, genes interacting with human- and mouse-specific TFs have higher expression levels, especially in humans. In the heart, genes interacting with human-specific TFs have the highest expression, especially in mice. In the liver, genes interacting with orthologous TFs have the highest expression in both humans and mice. Variation in expression profiles is small in the kidney.

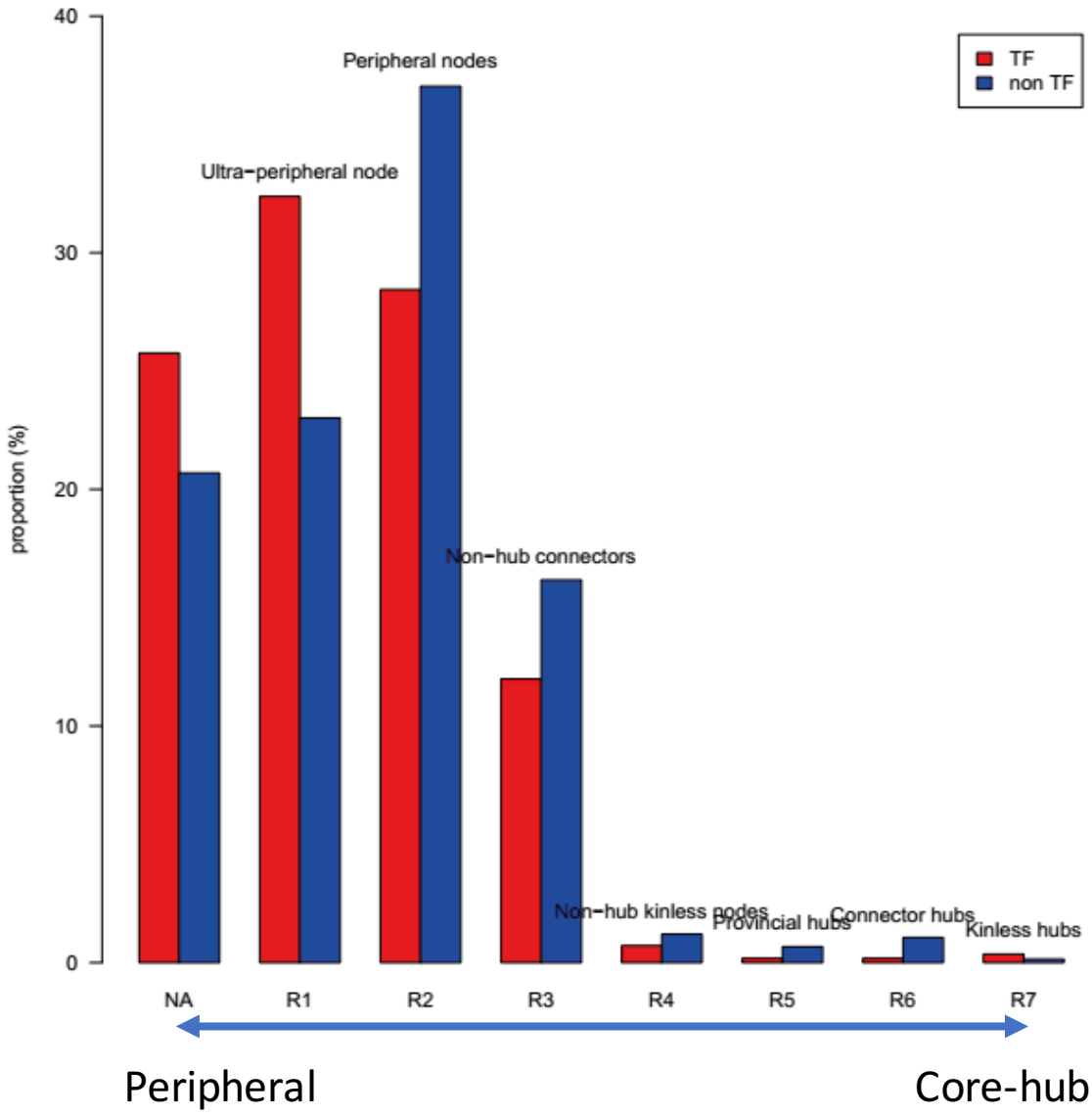

Figure3.8 Functional cartography of experimentally determined gene interactions. I characterized each gene in the human PPI network according to its within-module degree *z*-score (*z*) and participation coefficient (*p*). Genes were classified into eight groups: (NA) those with no experimental interactions, (R1) ultra-peripheral nodes ($z < 2.5$ and $p < 0.05$), (R2) peripheral nodes ($z < 2.5$ and $0.05 \leq p < 0.625$), (R3) non-hub connector nodes ($z < 2.5$ and $0.625 \leq p < 0.8$), (R4) non-hub kinless nodes ($z < 2.5$ and $p \geq 0.8$), (R5) provincial hubs ($z \geq 2.5$ and $p < 0.3$), (R6) connector hubs ($z \geq 2.5$ and $0.3 \leq p < 0.75$), and (R7) kinless hubs ($p \geq 0.75$).

Table 3.2 Gene in core hubs

	within_module_degree	participation_coefficient	role	Type
FLNA	2.638795333	0.779336735	R7: Kinless hubs	NonTF
PPP1CC	2.812761616	0.815193572	R7: Kinless hubs	NonTF
CSNK1E	3.776699562	0.803506209	R7: Kinless hubs	NonTF
IGKV3D-7	3.753569747	0.771626947	R7: Kinless hubs	NonTF
ENSG00000223931	6.798587327	0.817452513	R7: Kinless hubs	NonTF
IGLL5	3.193238407	0.802092952	R7: Kinless hubs	NonTF
BTRC	2.582959388	0.85467128	R7: Kinless hubs	NonTF
BABAM1	3.279226392	0.80625	R7: Kinless hubs	NonTF
PIK3R3	2.903835891	0.76953125	R7: Kinless hubs	NonTF
USP11	2.666990665	0.853741497	R7: Kinless hubs	NonTF
CTNNB1	4.14314548	0.811791383	R7: Kinless hubs	NonTF
BECN1	2.535077737	0.8384	R7: Kinless hubs	NonTF
KEAP1	2.80624304	0.793950851	R7: Kinless hubs	TF
PBX2	2.666666667	0.824196597	R7: Kinless hubs	TF
DAZAP1	3.051858105	0.312	R6: Connector hubs	NonTF
LDHB	3.718067748	0.451325462	R6: Connector hubs	NonTF
LDHA	3.622894942	0.435502959	R6: Connector hubs	NonTF
USP25	4.005862286	0.580246914	R6: Connector hubs	NonTF
ILF3	3.731411259	0.696548397	R6: Connector hubs	NonTF
ECH1	2.751629402	0.324260355	R6: Connector hubs	NonTF
CHCHD2	3.08088606	0.441275437	R6: Connector hubs	NonTF
PRPF8	4.629316502	0.337775927	R6: Connector hubs	NonTF
NDUFS3	4.858872012	0.445555556	R6: Connector hubs	NonTF
OTUD4	3.328201177	0.6304	R6: Connector hubs	NonTF
HLA-DRB1	2.547649317	0.617346939	R6: Connector hubs	NonTF
HLA-DRA	2.771748099	0.692901235	R6: Connector hubs	NonTF
COQ9	6.702709295	0.54254907	R6: Connector hubs	NonTF
RMND5A	2.753220395	0.545	R6: Connector hubs	NonTF
DUT	3.718067748	0.367643107	R6: Connector hubs	NonTF
CFL1	4.479450197	0.317906574	R6: Connector hubs	NonTF
SLC25A3	3.005990158	0.49621417	R6: Connector hubs	NonTF
ACTR2	3.885698622	0.642509465	R6: Connector hubs	NonTF
PCBP1	3.718067748	0.548575603	R6: Connector hubs	NonTF
C15ORF48	5.122277338	0.403045102	R6: Connector hubs	NonTF
U2AF2	4.081334152	0.439670139	R6: Connector hubs	NonTF
MRPL23	3.335727022	0.318772137	R6: Connector hubs	NonTF
STX12	2.677600654	0.550925926	R6: Connector hubs	NonTF
STX6	2.90207017	0.411242604	R6: Connector hubs	NonTF
HSD17B10	3.622894942	0.47761194	R6: Connector hubs	NonTF

MSN	2.671166881	0.315061728	R6: Connector hubs	NonTF
MRPL22	2.610305921	0.380485632	R6: Connector hubs	NonTF
IGBP1	5.716696677	0.441207076	R6: Connector hubs	NonTF
HNRNPM	3.601849595	0.636824197	R6: Connector hubs	NonTF
CDKN1A	2.54415177	0.453703704	R6: Connector hubs	NonTF
NHP2L1	3.259360626	0.667854756	R6: Connector hubs	NonTF
PPP2R1A	2.504565696	0.726963546	R6: Connector hubs	NonTF
PPP2CA	2.82304537	0.576594346	R6: Connector hubs	NonTF
TRAF2	3.500180183	0.676020408	R6: Connector hubs	NonTF
HNRNPU	4.012836358	0.693530246	R6: Connector hubs	NonTF
STIP1	5.050487034	0.335638388	R6: Connector hubs	NonTF
ALDOC	2.956685299	0.324031653	R6: Connector hubs	NonTF
NDUFA4	5.714939322	0.470100309	R6: Connector hubs	NonTF
LATS2	3.670062883	0.67638484	R6: Connector hubs	NonTF
HSPB1	3.813240554	0.5603125	R6: Connector hubs	NonTF
TPI1	3.147030911	0.497222222	R6: Connector hubs	NonTF
PRDX6	2.575994075	0.316115702	R6: Connector hubs	NonTF
HNRNPK	2.985369451	0.537854671	R6: Connector hubs	NonTF
SNRPA1	3.464854008	0.357659435	R6: Connector hubs	NonTF
POLR2E	2.545609321	0.478737997	R6: Connector hubs	NonTF
ABHD16A	3.015113446	0.73015873	R6: Connector hubs	NonTF
HNRNPH1	2.711378276	0.573129252	R6: Connector hubs	NonTF
ICT1	3.137884904	0.315631451	R6: Connector hubs	NonTF
COPS5	3.099323019	0.466942149	R6: Connector hubs	NonTF
AHCYL1	6.192560707	0.4394	R6: Connector hubs	NonTF
SRRM2	2.916871657	0.421682099	R6: Connector hubs	NonTF
HNRNPR	3.32785842	0.654979304	R6: Connector hubs	NonTF
DDX5	2.642880482	0.488040123	R6: Connector hubs	NonTF
UQCRC2	2.544358548	0.708189546	R6: Connector hubs	NonTF
SNRNP200	3.396356214	0.317384953	R6: Connector hubs	NonTF
CDK2	3.5897596	0.562860438	R6: Connector hubs	NonTF
NUP107	3.988175128	0.585848075	R6: Connector hubs	NonTF
NUP153	2.529086667	0.674718867	R6: Connector hubs	NonTF
COPS6	4.582345448	0.477828541	R6: Connector hubs	NonTF
VDAC3	2.808148039	0.4896875	R6: Connector hubs	NonTF
HSPA8	2.575994075	0.668628809	R6: Connector hubs	NonTF
SF3B1	3.670347389	0.401107266	R6: Connector hubs	NonTF
MYO5C	4.093515837	0.7024	R6: Connector hubs	NonTF
DSTN	4.669795809	0.409972299	R6: Connector hubs	NonTF
XPO1	3.242203717	0.517510708	R6: Connector hubs	NonTF
SF3A1	4.560818708	0.303757356	R6: Connector hubs	NonTF
U2AF1	3.396356214	0.399024539	R6: Connector hubs	NonTF

SRSF1	3.053867245	0.394445487	R6: Connector hubs	NonTF
DHX15	3.190862833	0.63239645	R6: Connector hubs	NonTF
SNRPD1	2.848373864	0.391020408	R6: Connector hubs	NonTF
SF3B3	4.423823121	0.399551066	R6: Connector hubs	NonTF
SF3A3	2.711378276	0.309573361	R6: Connector hubs	NonTF
EFTUD2	2.77987607	0.309901738	R6: Connector hubs	NonTF
TXN	3.051858105	0.365384615	R6: Connector hubs	NonTF
EIF3H	2.526976695	0.709572742	R6: Connector hubs	NonTF
RYBP	2.541924884	0.577777778	R6: Connector hubs	NonTF
DOK1	2.840642326	0.680555556	R6: Connector hubs	NonTF
MRPL53	3.665463886	0.358842618	R6: Connector hubs	NonTF
AMOT	3.325994488	0.640095181	R6: Connector hubs	NonTF
UIMC1	2.950415073	0.701388889	R6: Connector hubs	NonTF
EXOSC8	2.636595074	0.380165289	R6: Connector hubs	NonTF
PIK3R1	2.678090412	0.682222222	R6: Connector hubs	NonTF
CLTB	3.266340979	0.4453125	R6: Connector hubs	NonTF
C3	2.666666667	0.642857143	R6: Connector hubs	NonTF
BIRC2	2.560443091	0.6304	R6: Connector hubs	NonTF
MARK2	3.390604559	0.656	R6: Connector hubs	NonTF
ATM	2.985826233	0.620498615	R6: Connector hubs	NonTF
B2M	3.553690788	0.41322314	R6: Connector hubs	NonTF
CDC5L	3.32785842	0.562019013	R6: Connector hubs	TF

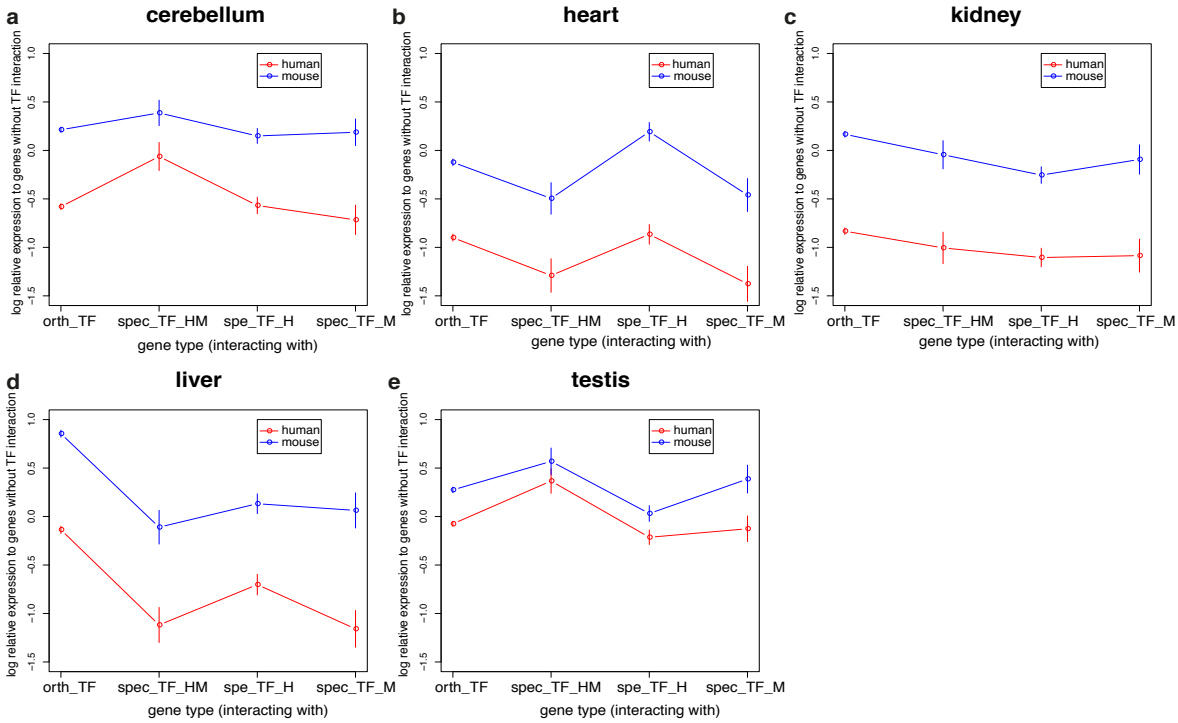

Figure 3.9 Log expression levels of transcription factor (TF)-interacting and non-interacting human and mouse orthologous genes. (**a**)Expression of genes in cerebellum. (**b**)Expression of genes in heart. mouse orthologous genes. (**c**)Expression of genes in kidney. (**d**)Expression of genes in liver. (**e**)Expression of genes in testis. Standard error bars are attached to the means. Orthologous genes were divided into four groups according to their interactions with TFs, namely, those that interact with orthologous TFs (orth_TF), human- and mouse-specific TFs (spec_TF_HM), human- but not mouse-specific TFs (spec_TF_H), and mouse- but not human-specific TFs (spec_TF_M).

3.5 Loss of human TFs in mice reveals knockout-phenotypes of their targets in humans

Human-isolated TFs are enriched in the Cys2His2-zinc finger (C2H2-zf) TF family. To compare the effect of reducing their expression levels in humans and mice, I focused on the C2H2-zf-containing Krüppel-associated box (C2H2-KRAB) family, the largest individual genome-encoded transcriptional repressor family of higher organisms. I surveyed the knock-out phenotypes of 672 C2H2-KRAB-interacting genes. A total of 9827 mammalian phenotype terms were recorded (Supplementary Materials, Table S3.5). I then collected information on mouse C2H2-KRAB-interacting genes that do not participate in this interaction and that are known to be responsible for specific mammalian phenotypes (Table 3.3, Table 3.4). I looked at the function of genes whose expression is regulated by human specific TFs and these human specific TFs also belong to C2H2-KRAB family. Because transcription factors can up-regulate or down-regulate the expression of target genes, and the C2H2-KRAB family is mainly down-regulated, thus the target genes of C2H2-KRAB are selected as extra criteria to keep the direction of regulation as consistent as possible. According to the knockout data of mouse genes, the phenotypic changes that may be caused by the down-regulation of target gene expression caused by the presence of transcription factors in humans are simulated. As demonstrated by C2H2-KRAB knock-out phenotypes in mice, morphological differences in the corresponding phenotype between humans and mice are due, at least partially, to the reduced expression levels of these interacting genes in humans relative to mice (Table 3.3). For example, the “short tail” (vs. “long tail”) phenotype in knock-out mice is consistent with the absence of tails in humans, while “delayed tooth eruption” (vs. “continually growing teeth”) in knock-out mice is comparable to permanent teeth in humans. Other examples include hair and skin phenotypes. Although information about target genes of species-specific TFs is lacking, I also found a similar trend in TF to target-gene regulation. The human-specific TF, SHOX, activates the expression of its target gene, FGFR3. The absence of SHOX in mice may contribute to the lower expression of mouse FGFR3. In mice, a humanized FGFR3 gene leads to “short tail” phenotypes, whereas knock-out of FGFR3 causes “long-tail” phenotypes (Table 3.4). These findings indicate that species-specific TFs are responsible for diverse target-gene expression because of altered regulatory interactions and that divergent expression of genes shapes species-specific phenotypes (Supplementary Materials, Figure 3.10).

Table 3.3 Mammalian phenotypes of representative genes that interact with KRAB-C2H2 and have low expression in humans.

Organ	Mouse Normal Phenotype	Gene with KRAB-C2H2 Interactions	Mammalian Phenotype	Mouse Knock-Out Phenotype	Human (Monkey) Normal Phenotype
Tail	Horizontal tail	CACNA1B	MP:0003382	Straub tail	(Vertical tail)
	Long tail	RIPK4	MP:0000592	Short tail	Without tail
Tooth	Continually growing teeth	OTUD7A	MP:0003053	Delayed tooth eruption	Permanent tooth
Hair	Fur-covered	CTSL2	MP:0000414	Alopecia	Hairless
		CTSL2	MP:0000417	Short hair	
Skin	Epidermis < 25 μm	CTSL2	MP:0001219	Thick epidermis	Epidermis > 50 μm
	Normal dermis	CTSL2	MP:0001245	Thick dermal layer	Thicker dermis than mouse

Organs exhibiting obvious phenotypic divergence between humans and mice are listed. Mammalian and mouse knock-out phenotypes were obtained from MGI. The phenotypes for all analyzed genes are listed in Supplementary Materials, Table S5.

Table 3.4 The effect of loss of SHOX in mouse inferred from the tail phenotypes associated with mutations in mouse FGFR3, the orthologous target gene of human SHOX.

Symbol	Mutation allele/type of mouse	Phenotype of mutation allele
$Fgfr3^{tm1Led}$	Targeted (Null/knockout)	long tail
$Fgfr3^{tm2Schl}$	Targeted (Null/knockout)	long tail
$Fgfr3^{tm3.1Cxd}$	Targeted (Humanized sequence)	short tail
		domed cranium
$Fgfr3^{tm1Llm}$	Targeted (Humanized sequence)	small snout
		short tail
		domed cranium

Knock-out phenotypes were obtained from MGI.

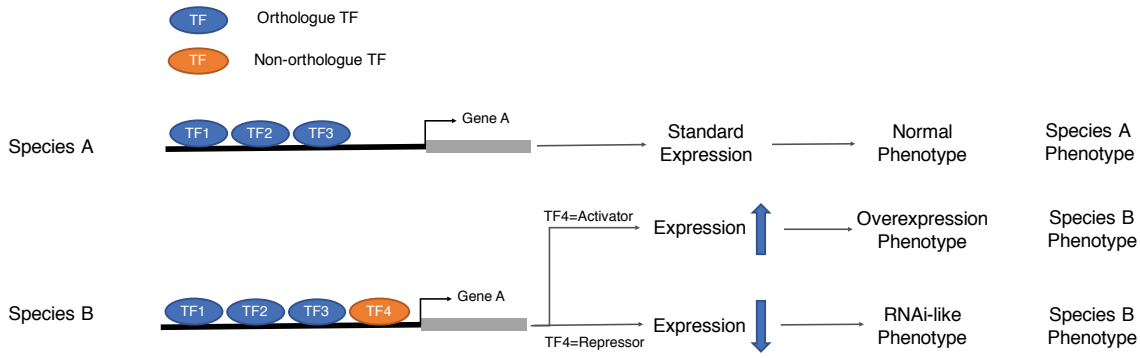

Figure 3.10 Shaping of species-specific phenotypes by species–specific TFs. The blue ellipse is orthologue TF and the orange ellipse is non-orthologue TF.

The same logic can also be applied to pathways. Small animals, such as mice, have a high metabolic rate. The glycolytic pathway is the basic pathway that supports the metabolic demands of different organisms. The isolated TFs can modulate five connected genes (TPI1, NLK, ALDOA, PFKL, and PFKM) in the glycolytic pathway, possibly making these genes more plastically regulated (Figure 3.11a). Only two TFs (ZNF224 and ZNF256) interacting with ALDOA in humans are absent in mice. ZNF224 represses transcription of the ALDOA gene, and ZNF256 is a transcriptional repressor. Consequently, ALDOA has relatively lower expression in humans than in mice. In all five organs, expression of the ALDOA gene is nearly double in mice compared with that in humans (Figure 3.11b). This result indicates that the pathway can also be affected by the evolution of isolated TFs.

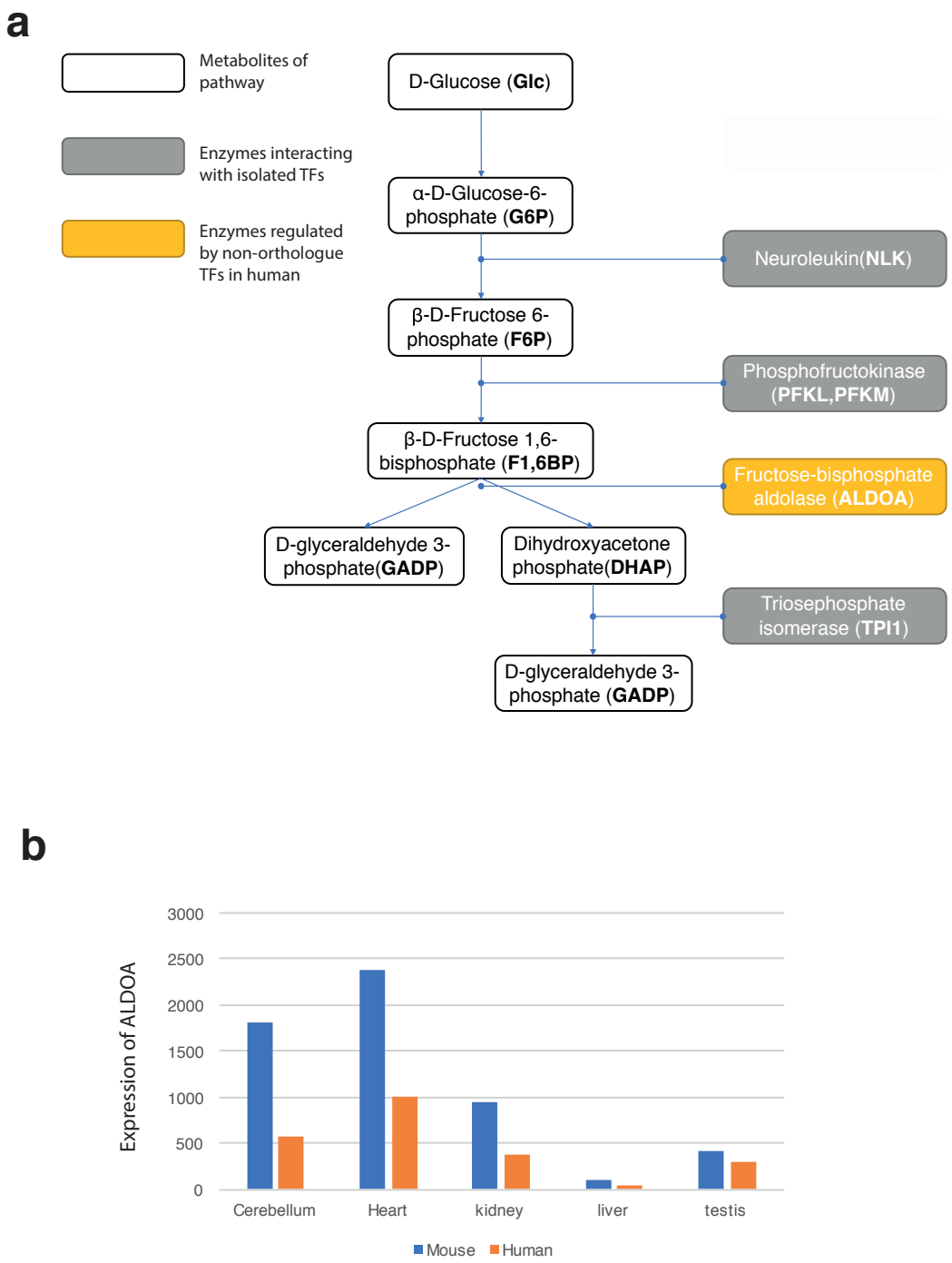

Figure 3.11 Glycolytic pathway component ALDOA and its expression levels in humans and mice. (a) Initial steps of glycolytic pathway. White block: metabolite of pathway. Gray block: enzyme interacting with isolated TF. Orange block: enzyme interacting with non-orthologous TFs in human. (b) Expression of ALDOA. The blue bar is the expression of ALDOA in mouse. The orange bar is the expression of ALDOA in human.

3.6 Human and mouse biological functions are regulated by similar numbers of TFs but different TF family members

Human and mouse biological functions were found to be regulated by similar numbers of TFs (Figure 3.12a) but by different members of TF families (Figure 3.12b). The number of GO items in human, mouse and rat are similar (Figure 3.13). GO terms associated with a small number of TFs are mostly regulated by orthologous TFs. However, for GO terms regulated by many TFs (as many as 400, i.e., $\sim e^6$), the proportion of orthologous TFs is as small as 50%. I conducted GO and pathway enrichment analyses on these two TF groups and their interacting genes (Figure 3.14). Even though the numbers of isolated TFs and their interacting genes were much smaller than those of the other set of genes, their functional profiles were very similar regarding GO terms and pathways. This outcome indicates that the isolated TFs are not null-function, though their interaction with those functions may be weaker. Although the amount of functional change caused by the formation or loss of isolated TFs is small, the related phenotype is still affected. These TFs, especially the isolated ones, thus function through their formation or loss like multiple switches that open or close to generate a unique phenotype or a divergent function during speciation.

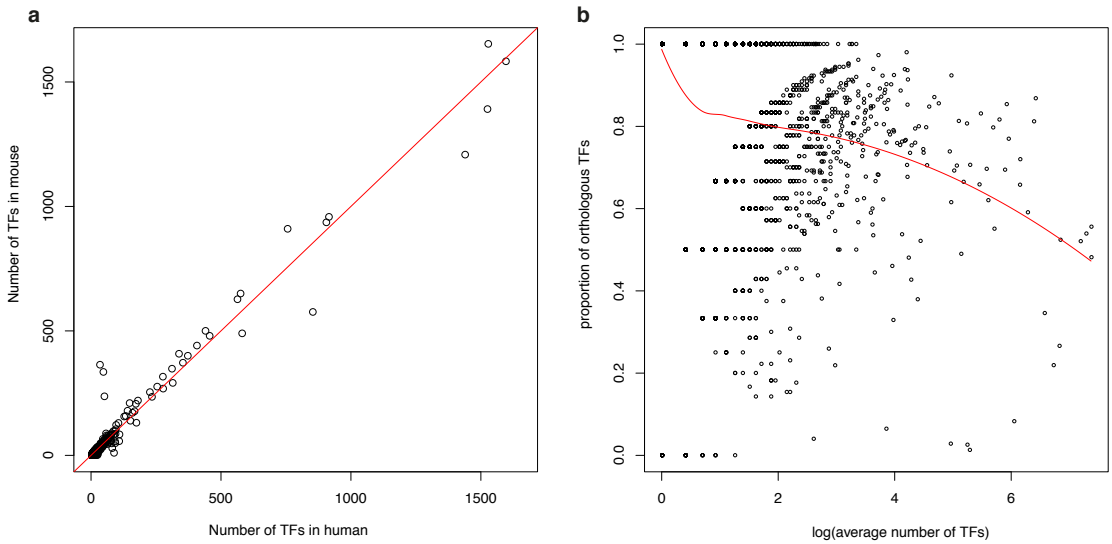

Figure 3.12 Shared and specific transcription factors (TFs) that regulate gene ontology (GO) terms in humans and mice. (**a**) Comparison of the number of TFs regulating GO terms in humans and mice. (**b**) Proportion of orthologous TFs relative to the average number of TFs. The red line in (**a**) represents the average number of TFs regulating GO terms in humans and mice. The smooth red curve in (**b**) represents the predicted proportion of orthologous TFs regulating GO terms.

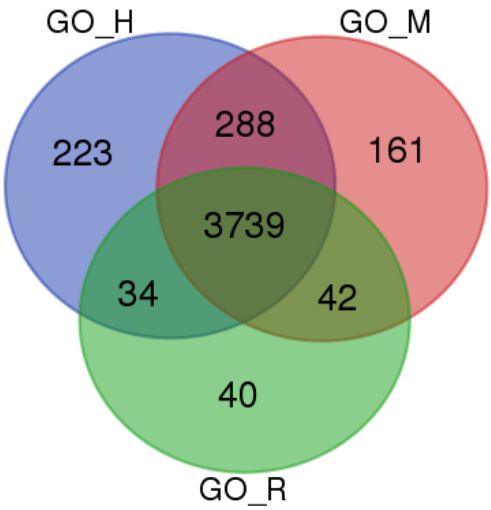

Figure 3.13 The number of GO items in human, mouse and rat.

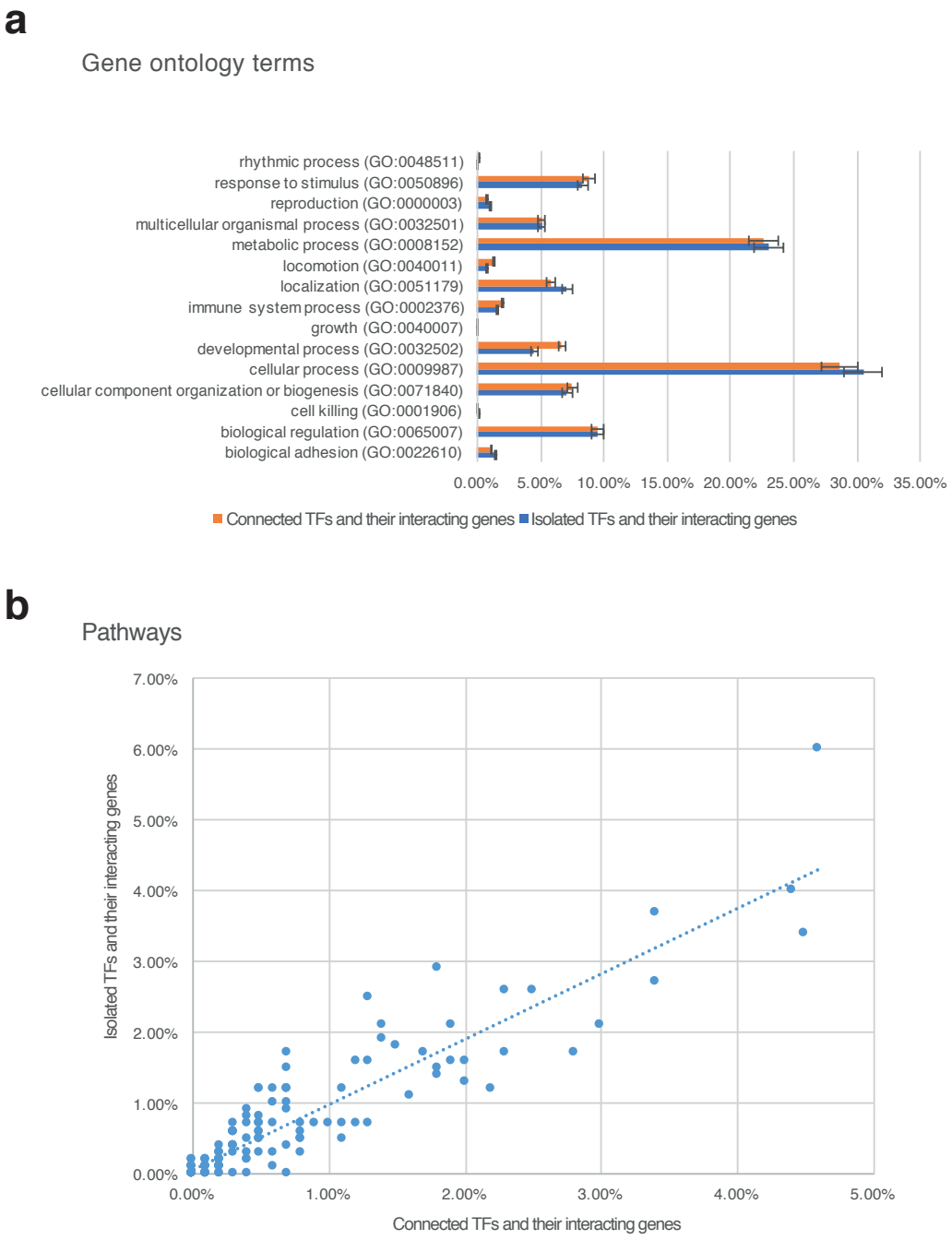

Figure 3.14 Enrichment analysis of gene ontology terms and pathways. (a) Enrichment analysis of gene ontology (GO) terms. X-axis: GO terms; Y-axis: percentage of genes in GO term. Orange bar: connected TFs and their interaction genes. Blue bar: isolated TFs and their interaction genes. (b) Enrichment analysis of pathways. X-axis: percentage of connected TFs and their interaction genes. Y-axis: percentage of isolated TFs and their interaction genes. Blue dot: pathway.

3.7 discussion

Our TF-to-TF network is based on the STRING database, which collects protein–protein interactions based on several types of evidence (see Materials and Methods). Interactions with genes for which there is little information may be under-represented in the list. However, because of the large amount of human RNA-seq data, the co-expression data coverage is comprehensive. TFs can regulate gene expression, so if such regulation exists, it is likely to be detected by "conserved co-expression" in STRING. Evidence of co-expression and from high-throughput laboratory experiments may include unbiased information on the TF-with-protein interactions. I adopted the interactions when there was any evidence regarding the type of interaction; therefore, the isolation of TFs is likely to be real.

The TF database was constructed by collecting sequences with DBD. Some proteins own DBD bud does not have regulatory function and some proteins have regulatory function but do not include sequences that are similar to known DBD domain. The number of functional annotations and DBDs are growing but these are still incomplete for now. The quality of the annotation of regulatory function varies among species. Therefore, our analysis of acquisition and loss of transcription factors may be affected by the variation of the quality of functional annotation. The analysis will become more solid as many well-annotated genomes across whole mammal species become available.

In recent years, studies of the C2H2 TF family and several other TF genes have revealed the evolution of TFs. A relationship between TF sequence evolution and changes in DNA binding properties has also been found. Reports showing that TFs are evolutionarily conserved were based primarily on TFs with known DNA-binding sequence specificities, whereas reports showing that TFs are evolutionarily variable always considered entire TF families. I therefore hypothesized that there is another type of TF that, along with well-studied TFs, contribute to overall TF evolution. Three factors have been proposed to explain how TF evolution has circumvented the problem of negative pleiotropy: (1) alternative splicing, (2) short linear motifs, and (3) simple sequence repeats. Until now, however, the regulatory logic behind overall TF evolution remains unknown.

It was found that one-third of TFs constitute a new TF type that is isolated in the human TF-to-TF network and that tends to be peripheral in the network of PPIs. These TFs have rarely

been reported in previous human TF-to-TF network studies. The characteristics of isolated TFs are consistent with the protein characteristics related to lineage-specific phenotypes. Mutations of these isolated TFs are far less lethal than those of other TFs, indicating the high tolerance of the regulatory network to the evolution of these genes. The less strongly interacting genes encoding these isolated TFs contribute to less pleiotropic regulation. The other two-thirds of TFs make up a large connected TF component of the human TF-to-TF network containing nearly all TFs with known DNA-binding specificities.

The comparative study of mammalian TFs presents an overview of TF member variation and demonstrates that TF evolution in mammals is ubiquitous—with changes observed in closely related species, not just between humans and mice. Starting from the same TFs in the shared common ancestor, the turnover of TFs during mammalian evolution and species–specific formation and loss events have gradually led to unique sets of TFs. In our human-mouse model, the overall force of TF formation and loss tends to be unilateral, with the overall expression level of interacting genes in a species being either relatively higher or lower. Changing the expression level of functional genes will consequently change phenotypes and pathway efficiency, an idea that is confirmed by the evidence in this study.

An isolated TF has a GO functional term overlay similar to that of connected TFs, which means that isolated TFs can also adjust a wide range of functions that are mainly regulated by connected TFs. Each GO term was found to be regulated in humans and mice by a similar number of TFs, which are largely non-orthologous.

The gain and loss of TFs, mainly the isolated ones, may not be a useless process, even though these changes are prevalent and tolerable to organisms. These changes will largely affect the properties of an interacting gene, such as its interaction and expression. When interacting TFs are absent or newly emerging, the same interacting genes will have different expression levels. As TF evolution has been frequent and widespread throughout mammalian history, large-scale phenotypes and pathway efficiencies have been shaped among species. These observations improve our understanding of the consequences of TF evolution.

I therefore hypothesized that these connected TFs follow the common TF regulatory pattern, with their conserved members possibly forming the backbone structure of the regulatory network. In contrast, the variable isolated TFs tune the flow of the regulatory network and give

rise to species uniqueness by acting as on/off switches. This scenario explains how TFs can evolve while tolerating negative pleiotropic effects and identifies a major source of TF evolution and why TF numbers vary among species.

This situation may be best visualized by regarding the members of TF families as regulatory switches. During evolution, species may have modified the flow of the regulatory network by selecting different on/off states. Isolated TFs are an ideal tool for accomplishing this task: the relatively less lethal phenotypes of isolated TFs make them more tolerant to changes during speciation. In addition, emerging TFs in different species can diversify the expression profiles of their target genes, resulting in an adaptive phenotype for each species. Consequently, phenotypes have evolved by turning multiple switches on and off—in other words, through the formation and loss of isolated TFs.

Supplementary information

includes large tables in link:

https://drive.google.com/drive/folders/1gifcglSa5X5BoMeOzsfaKlGNy5Kc3j9Q?usp=sharing

Table S3.1: Isolated TF list and connected TF list in human.

Table S3.2: Formation and loss events in 48 TF families. The number of gain events on branches and the number of loss events on branches.

Table S3.3: Number of edges between different types of TFs in mouse and rat gene interaction networks. (a) TFs with DBD loss among mouse and rat. (b) TFs with gene loss among mouse and rat. (c) TFs without loss among mouse and rat.

Table S3.4: Human and mouse gene expression data. RNA-seq data of 15,796 orthologous genes in cerebellum, heart, kidney, liver and testis.

Table S3.5: Mammalian phenotypes of genes that interact with KRAB-C2H2 and have low expression in humans. Mammalian phenotypes of genes were obtained from MGI.

As the contents of this chapter (page) are anticipated to be published in a paper in a scholarly journal, they cannot be published online. The paper is scheduled to be published within 5 years.

Acknowledgement

I want to express my special respect and thanks to my supervisor Prof. Hirohisa Kishino. When I first arrived japan, I was nervous about the new environment and new life here. At that time, I was almost a wet-lab student, familiar with skills like cell culture and RNAi, but my knowledge of bioinformatics or statistics was limited. Now I feel that I am really lucky to do research with Prof. Hirohisa Kishino. Not only because my dry-lab skill strengthened greatly here, but more importantly, his enthusiasm and concentration in scientific research have set me an excellent example. My experience at the University of Tokyo is precious and left me unforgettable memories.

I want to express my deep gratitude to Dr. Jiaqi Wu. In terms of programming and phylogenetic trees, she provided many useful programmes and valuable opinions for our research. These helps greatly promoted our research progress. When I first came to Japan, my Japanese was not very good. She helped me a lot in daily life, such as helping me fill out Japanese written documents and telling me a lot of customs in Japan. This helped me to adapt quickly to life in Japan. I also want to express my thanks to Dr. Takahiro Yonezawa, her husband. He provided valuable suggestions on mammalian species.

I want to thank all members in the laboratory of Biometry and Bioinformatics of the University of Tokyo. They are all very kind and provide me many helps. I really miss the experience of growing soybeans with many people in our laboratory.

I want to pay my special thanks to my Master supervisor, Prof. Ze Zhang and hope he can fully recover from cerebral infarction soon. I also hope to thanks Prof. Wei Sun. They taught me a lot in research and experimentation, and let me know more about model organisms and genomes. From these experiences, I became interested in transcription factors.

I hope to express my deep gratitude to my parents. They gave me life and brought me up. Because of them, I can do research without many concern. Without their support and love, I may fail my PhD. courses.

References

1. Krumlauf, R., *Hox Genes in Vertebrate Development.* Cell, 1994. **78**(2): p. 191-201.
2. Zhang, H.M., et al., *AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors.* Nucleic Acids Research, 2015. **43**(D1): p. D76-D81.
3. Kanamori, M., et al., *A genome-wide and nonredundant mouse transcription factor database.* Biochemical and Biophysical Research Communications, 2004. **322**(3): p. 787-793.
4. Pfreundt, U., et al., *FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database.* Nucleic Acids Research, 2010. **38**: p. D443-D447.
5. Fulton, D.L., et al., *TFCat: the curated catalog of mouse and human transcription factors.* Genome Biology, 2009. **10**(3).
6. Lee, A.P., et al., *TFCONES: A database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements.* BMC Genomics, 2007. **8**.
7. Zheng, G.Y., et al., *ITFP: an integrated platform of mammalian transcription factors.* Bioinformatics, 2008. **24**(20): p. 2416-2417.
8. Lambert, S.A., et al., *The human transcription factors.* Cell, 2018. **175**(2): p. 598-599.
9. Nowick, K. and L. Stubbs, *Lineage-specific transcription factors and the evolution of gene regulatory networks.* Brief Funct Genomics, 2010. **9**(1): p. 65-78.
10. O'Malley, M.A., J.G. Wideman, and I. Ruiz-Trillo, *Losing Complexity: The Role of Simplification in Macroevolution.* Trends Ecol Evol, 2016. **31**(8): p. 608-621.
11. Kuo, C.H. and H. Ochman, *Deletional bias across the three domains of life.* Genome Biol Evol, 2009. **1**: p. 145-52.
12. Zmasek, C.M. and A. Godzik, *Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires.* Genome Biol, 2011. **12**(1): p. R4.
13. Olson, M.V., *When less is more: gene loss as an engine of evolutionary change.* Am J Hum Genet, 1999. **64**(1): p. 18-23.
14. de Mendoza, A. and A. Sebe-Pedros, *Origin and evolution of eukaryotic transcription factors.* Curr Opin Genet Dev, 2019. **58-59**: p. 25-32.
15. Kusserow, A., et al., *Unexpected complexity of the Wnt gene family in a sea anemone.* Nature, 2005. **433**(7022): p. 156-60.
16. Somorjai, I.M.L., et al., *Wnt evolution and function shuffling in liberal and conservative chordate genomes.* Genome Biol, 2018. **19**(1): p. 98.
17. Gabaldon, T. and E.V. Koonin, *Functional and evolutionary implications of gene orthology.* Nat Rev Genet, 2013. **14**(5): p. 360-6.
18. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspective on protein families.* Science, 1997. **278**(5338): p. 631-7.
19. El-Gebali, S., et al., *The Pfam protein families database in 2019.* Nucleic Acids Research 2018.
20. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
21. Mitchell, A.L., et al., *InterPro in 2019: improving coverage, classification and access to protein sequence annotations.* Nucleic Acids Research, 2019. **47**(D1): p. D351-D360.
22. Eddy, S.R., *Multiple alignment using hidden Markov models.* Ismb, 1995. **Vol. 3**: p. 114-120.

23. Eddy, S.R., *Hidden Markov models.* Current Opinion in Structural Biology, 1996. **6**(3): p. 361-365.
24. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-763.
25. Wheeler, T.J. and S.R. Eddy, *nhmmer: DNA homology search with profile HMMs.* Bioinformatics, 2013. **29**(19): p. 2487-2489.
26. Eddy, S.R., *Accelerated Profile HMM Searches.* Plos Computational Biology, 2011. **7**(10).
27. Mistry, J., et al., *Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.* Nucleic Acids Research, 2013. **41**(12).
28. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Research, 2011. **39**: p. W29-W37.
29. Johnson, L.S., S.R. Eddy, and E. Portugaly, *Hidden Markov model speed heuristic and iterative HMM search procedure.* Bmc Bioinformatics, 2010. **11**.
30. Kumar, S., et al., *TimeTree: a resource for timelines, timetrees, and divergence times.* Molecular Biology and Evolution, 2017. **34**(7): p. 1812-1819.
31. Wu, J., T. Yonezawa, and H. Kishino, *Rates of Molecular Evolution Suggest Natural History of Life History Traits and a Post-K-Pg Nocturnal Bottleneck of Placentals.* Curr Biol, 2017. **27**(19): p. 3025–3033 e5.
32. Liu, L. and L. Yu, *Estimating species trees from unrooted gene trees.* Syst Biol, 2011. **60**(5): p. 661-7.
33. Tarver, J.E., et al., *The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference.* Genome Biol. Evol., 2016. **8**(2): p. 330–44.
34. Mason, V.C., et al., *Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates.* Science Advances, 2016. **2**(8): p. e1600633.
35. Nguyen, L.T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Mol Biol Evol, 2015. **32**(1): p. 268-74.
36. Yamada, K.D., K. Tomii, and K. Katoh, *Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees.* Bioinformatics, 2016. **32**(21): p. 3246-3251.
37. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
38. Xia, X.H., *DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution.* Journal of Heredity, 2017. **108**(4): p. 431-437.
39. Tamura, K., et al., *MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0.* Molecular Biology and Evolution, 2013. **30**(12): p. 2725-2729.
40. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Research, 2017. **45**(D1): p. D362-D368.
41. Guimera, R. and L.A. Nunes Amaral, *Functional cartography of complex metabolic networks.* Nature, 2005. **433**(7028): p. 895-900.
42. Brawand, D., et al., *The evolution of gene expression levels in mammalian organs.* Nature, 2011. **478**(7369): p. 343-8.
43. Petryszak, R., et al., *Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants.* Nucleic Acids Research, 2016. **44**(D1): p. D746-52.
44. Venables, W.N., B.D. Ripley, and W.N. Venables, *Modern applied statistics with S.* 4th ed. Statistics and computing. 2002, New York: Springer. xi, 495 p.
45. Team, R.C., *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2014, Vienna, Austria.

46. Han, H., et al., *TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions.* Nucleic Acids Research, 2018. **46**(D1): p. D380-D386.
47. Carbon, S., et al., *Expansion of the Gene Ontology knowledgebase and resources.* Nucleic Acids Research, 2017. **45**(D1): p. D331-D338.
48. Reimand, J., et al., *g:Profiler-a web server for functional interpretation of gene lists (2016 update).* Nucleic Acids Research, 2016. **44**(W1): p. W83-W89.
49. Cleveland, W.S., *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, 1979(74:368): p. 829-836.
50. Agarwala, R., et al., *Database resources of the national center for biotechnology information.* Nucleic Acids Research, 2016. **44**(D1): p. D7-D19.
51. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.
52. Smith, C.L., et al., *Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse.* Nucleic Acids Research, 2018. **46**(D1): p. D836-D842.
53. Stergachis, A.B., et al., *Conservation of trans-acting circuitry during mammalian regulatory evolution.* Nature, 2014. **515**(7527): p. 365.
54. Kimura, M., *Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.* Nature, 1977. **267**(5608): p. 275-6.
55. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly.* Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14338-43.
56. Nowick, K., M. Carneiro, and R. Faria, *A prominent role of KRAB-ZNF transcription factors in mammalian speciation?* Trends Genet, 2013. **29**(3): p. 130-9.
57. van der Vinne, V., et al., *Cold and hunger induce diurnality in a nocturnal mammal.* Proc Natl Acad Sci U S A, 2014. **111**(42): p. 15256-60.
58. Qu, S.L., et al., *Mipu1 overexpression protects macrophages from oxLDL-induced foam cell formation and cell apoptosis.* DNA Cell Biol, 2014. **33**(12): p. 839-46.
59. Beaney, K.E., et al., *Variant rs10911021 that associates with coronary heart disease in type 2 diabetes, is associated with lower concentrations of circulating HDL cholesterol and large HDL particles but not with amino acids.* Cardiovasc Diabetol, 2016. **15**(1): p. 115.
60. Meziane, H., et al., *The homeodomain factor Gbx1 is required for locomotion and cell specification in the dorsal spinal cord.* PeerJ, 2013. **1**: p. e142.
61. Moller, A.P., et al., *Immune defense and host sociality: a comparative study of swallows and martins.* Am Nat, 2001. **158**(2): p. 136-45.
62. Silk, J.B., *The adaptive value of sociality in mammalian groups.* Philos Trans R Soc Lond B Biol Sci, 2007. **362**(1480): p. 539-59.
63. Schoch, K., et al., *A Recurrent De Novo Variant in NACC1 Causes a Syndrome Characterized by Infantile Epilepsy, Cataracts, and Profound Developmental Delay.* Am J Hum Genet, 2017. **100**(2): p. 343-351.
64. Gilliam, F., *Social Cognition and Epilepsy: Understanding the Neurobiology of Empathy and Emotion.* Epilepsy Curr, 2015. **15**(3): p. 118-9.
65. Ganesh, S., et al., *Targeted disruption of the Epm2a gene causes formation of Lafora inclusion bodies, neurodegeneration, ataxia, myoclonus epilepsy and impaired behavioral response in mice.* Hum Mol Genet, 2002. **11**(11): p. 1251-62.
66. Clement, S., et al., *The lipid phosphatase SHIP2 controls insulin sensitivity.* Nature, 2001. **409**(6816): p. 92-7.
67. Isaacs, A.M., et al., *A mutation in Af4 is predicted to cause cerebellar ataxia and cataracts in the robotic mouse.* J Neurosci, 2003. **23**(5): p. 1631-7.

68. Lynch, V.J., et al., *Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals.* Nat Genet, 2011. **43**(11): p. 1154-9.
69. Melo, J.B., et al., *Cutis Aplasia as a clinical hallmark for the syndrome associated with 19q13.11 deletion: the possible role for UBA2 gene.* Mol Cytogenet, 2015. **8**: p. 21.
70. Simpson, J.C., et al., *Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway.* Nat Cell Biol, 2012. **14**(7): p. 764-74.
71. Stelzer, G., et al., *The GeneCards suite: from gene data mining to disease genome sequence analyses.* Current protocols in bioinformatics, 2016. **54**: p. 1-30.
72. Boggio, E.M., et al., *Visual impairment in FOXG1-mutated individuals and mice.* Neuroscience, 2016. **324**: p. 496-508.
73. Hurov, K.E., C. Cotta-Ramusino, and S.J. Elledge, *A genetic screen identifies the Triple T complex required for DNA damage signaling and ATM and ATR stability.* Genes Dev, 2010. **24**(17): p. 1939-50.
74. Majumdar, S., A. Singh, and D.C. Rio, *The human THAP9 gene encodes an active P-element DNA transposase.* Science, 2013. **339**(6118): p. 446-8.
75. Alders, M., et al., *Disruption of a novel imprinted zinc-finger gene, ZNF215, in Beckwith-Wiedemann syndrome.* Am J Hum Genet, 2000. **66**(5): p. 1473-84.
76. Yasukochi, Y., et al., *Six novel susceptibility loci for coronary artery disease and cerebral infarction identified by longitudinal exome-wide association studies in a Japanese population.* Biomed Rep, 2018. **9**(2): p. 123-134.
77. Smith, C.M., et al., *The mouse Gene Expression Database (GXD): 2019 update.* Nucleic Acids Research, 2018.
78. Liu, L., et al., *A Targeted, Next-Generation Genetic Sequencing Study on Tetralogy of Fallot, Combined With Cleft Lip and Palate.* J Craniofac Surg, 2017. **28**(4): p. e351-e355.
79. Patel, N., et al., *ZBTB42 mutation defines a novel lethal congenital contracture syndrome (LCCS6).* Hum Mol Genet, 2014. **23**(24): p. 6584-93.
80. Sharma, S., et al., *An siRNA screen for NFAT activation identifies septins as coordinators of store-operated Ca2+ entry.* Nature, 2013. **499**(7457): p. 238-42.
81. Sivan, G., et al., *Human genome-wide RNAi screen reveals a role for nuclear pore proteins in poxvirus morphogenesis.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. 3519-24.
82. Jeannot, E., et al., *Nuclear factor I X is a recurrent target for HPV16 insertions in anal carcinomas.* Genes Chromosomes Cancer, 2018. **57**(12): p. 638-644.
83. Raghuram, V., et al., *Assessment of mutations in KCNN2 and ZNF135 to patient neurological symptoms.* Neuroreport, 2017. **28**(7): p. 375-379.
84. Wu, H., et al., *TALE nickase-mediated SP110 knockin endows cattle with increased resistance to tuberculosis.* Proc Natl Acad Sci U S A, 2015. **112**(13): p. E1530-9.
85. Ren, M., et al., *Ponatinib suppresses the development of myeloid and lymphoid malignancies associated with FGFR1 abnormalities.* Leukemia, 2013. **27**(1): p. 32-40.
86. Grismayer, B., et al., *Rab31 expression levels modulate tumor-relevant characteristics of breast cancer cells.* Mol Cancer, 2012. **11**: p. 62.
87. Siatecka, M. and J.J. Bieker, *The multifunctional role of EKLF/KLF1 during erythropoiesis.* Blood, 2011. **118**(8): p. 2044-54.
88. Lomniczi, A., et al., *Epigenetic regulation of puberty via Zinc finger protein-mediated transcriptional repression.* Nat Commun, 2015. **6**: p. 10195.

89. Fleige, A., et al., *Serum response factor contributes selectively to lymphocyte development.* J Biol Chem, 2007. **282**(33): p. 24320-8.
90. Unoki, M., J. Okutsu, and Y. Nakamura, *Identification of a novel human gene, ZFP91, involved in acute myelogenous leukemia.* Int J Oncol, 2003. **22**(6): p. 1217-23.
91. Ooi, Y.S., et al., *Genome-wide RNAi screen identifies novel host proteins required for alphavirus entry.* PLoS Pathog, 2013. **9**(12): p. e1003835.
92. Passegue, E., et al., *Chronic myeloid leukemia with increased granulocyte progenitors in mice lacking junB expression in the myeloid lineage.* Cell, 2001. **104**(1): p. 21-32.
93. Minoretti, P., et al., *A W148R mutation in the human FOXD4 gene segregating with dilated cardiomyopathy, obsessive-compulsive disorder, and suicidality.* Int J Mol Med, 2007. **19**(3): p. 369-72.
94. Kronke, J., et al., *Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells.* Science, 2014. **343**(6168): p. 301-5.
95. Bastian, H. and P. Gruss, *A murine even-skipped homologue, Evx 1, is expressed during early embryogenesis and neurogenesis in a biphasic manner.* EMBO J, 1990. **9**(6): p. 1839-52.
96. Hammoud, S., et al., *Sequence alterations in the YBX2 gene are associated with male factor infertility.* Fertil Steril, 2009. **91**(4): p. 1090-5.
97. Lee, J., et al., *Functional polymorphism in H2BFWT-5'UTR is associated with susceptibility to male infertility.* J Cell Mol Med, 2009. **13**(8B): p. 1942-51.
98. Konig, R., et al., *Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication.* Cell, 2008. **135**(1): p. 49-60.
99. Madani, N., et al., *Implication of the lymphocyte-specific nuclear body protein Sp140 in an innate response to human immunodeficiency virus type 1.* J Virol, 2002. **76**(21): p. 11133-8.
100. Mesuraca, M., et al., *Expression profiling and functional implications of a set of zinc finger proteins, ZNF423, ZNF470, ZNF521, and ZNF780B, in primary osteoarthritic articular chondrocytes.* Mediators Inflamm, 2014. **2014**: p. 318793.
101. Metzger, J., et al., *Expression levels of LCORL are associated with body size in horses.* PLoS One, 2013. **8**(2): p. e56497.
102. Li, P., et al., *Increased ZNF84 expression in cervical cancer.* Arch Gynecol Obstet, 2018. **297**(6): p. 1525-1532.
103. Chittoor, G., et al., *Genetic variation underlying renal uric acid excretion in Hispanic children: the Viva La Familia Study.* BMC Med Genet, 2017. **18**(1): p. 6.
104. Yamaguchi-Kabata, Y., et al., *Integrated analysis of human genetic association study and mouse transcriptome suggests LBH and SHF genes as novel susceptible genes for amyloid-beta accumulation in Alzheimer's disease.* Hum Genet, 2018. **137**(6-7): p. 521-533.
105. Wang, Y., et al., *Cdx gene deficiency compromises embryonic hematopoiesis in the mouse.* Proc Natl Acad Sci U S A, 2008. **105**(22): p. 7756-61.
106. Mortlock, D.P. and J.W. Innis, *Mutation of HOXA13 in hand-foot-genital syndrome.* Nat Genet, 1997. **15**(2): p. 179-80.
107. Mackay, D.J., et al., *Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57.* Nat Genet, 2008. **40**(8): p. 949-51.
108. Zhang, E.E., et al., *A genome-wide RNAi screen for modifiers of the circadian clock in human cells.* Cell, 2009. **139**(1): p. 199-210.

109. Crujeiras, A.B., et al., *Identification of an episignature of human colorectal cancer associated with obesity by genome-wide DNA methylation analysis.* Int J Obes (Lond), 2019. **43**(1): p. 176-188.
110. Banerjee, P., et al., *TULP1 mutation in two extended Dominican kindreds with autosomal recessive retinitis pigmentosa.* Nat Genet, 1998. **18**(2): p. 177-9.
111. Broen, K., et al., *A polymorphism in the splice donor site of ZNF419 results in the novel renal cell carcinoma-associated minor histocompatibility antigen ZAPHIR.* PLoS One, 2011. **6**(6): p. e21699.
112. Aughey, R.J., C.A. Goodman, and M.J. McKenna, *Greater chance of high core temperatures with modified pacing strategy during team sport in the heat.* J Sci Med Sport, 2014. **17**(1): p. 113-8.
113. Edqvist, P.H., S.M. Myers, and F. Hallbook, *Early identification of retinal subtypes in the developing, pre-laminated chick retina using the transcription factors Prox1, Lim1, Ap2alpha, Pax6, Isl1, Isl2, Lim3 and Chx10.* Eur J Histochem, 2006. **50**(2): p. 147-54.
114. Miletich, I., et al., *Developmental stalling and organ-autonomous regulation of morphogenesis.* Proc Natl Acad Sci U S A, 2011. **108**(48): p. 19270-5.
115. Murthi, P., et al., *Homeobox gene DLX4 expression is increased in idiopathic human fetal growth restriction.* Mol Hum Reprod, 2006. **12**(12): p. 763-9.
116. Song, O.R., et al., *ArfGAP1 restricts Mycobacterium tuberculosis entry by controlling the actin cytoskeleton.* EMBO Rep, 2018. **19**(1): p. 29-42.
117. Warner, N., et al., *A genome-wide small interfering RNA (siRNA) screen reveals nuclear factor-kappaB (NF-kappaB)-independent regulators of NOD2-induced interleukin-8 (IL-8) secretion.* J Biol Chem, 2014. **289**(41): p. 28213-24.
118. Jackson, M., et al., *A murine specific expansion of the Rhox cluster involved in embryonic stem cell biology is under natural selection.* BMC Genomics, 2006. **7**: p. 212.
119. Passananti, C., et al., *The product of Zfp59 (Mfg2), a mouse gene expressed at the spermatid stage of spermatogenesis, accumulates in spermatozoa nuclei.* Cell Growth Differ, 1995. **6**(8): p. 1037-44.
120. Saga, Y., *Genetic rescue of segmentation defect in MesP2-deficient mice by MesP1 gene replacement.* Mech Dev, 1998. **75**(1-2): p. 53-66.
121. Li, Y. and R.R. Behringer, *Esx1 is an X-chromosome-imprinted regulator of placental development and fetal growth.* Nat Genet, 1998. **20**(3): p. 309-11.
122. Kranz, D. and M. Boutros, *A synthetic lethal screen identifies FAT1 as an antagonist of caspase-8 in extrinsic apoptosis.* EMBO J, 2014. **33**(3): p. 181-97.
123. Waddell, P.J., N. Okada, and M. Hasegawa, *Towards resolving the interordinal relationships of placental mammals.* Syst Biol, 1999. **48**(1): p. 1-5.
124. O'Leary, M.A., et al., *The placental mammal ancestor and the post-K-Pg radiation of placentals.* Science, 2013. **339**(6120): p. 662-7.
125. Albalat, R. and C. Canestro, *Evolution by gene loss.* Nat Rev Genet, 2016. **17**(7): p. 379-91.
126. Zhang, J. and J.R. Yang, *Determinants of the rate of protein sequence evolution.* Nat Rev Genet, 2015. **16**(7): p. 409-20.
127. Pal, C., B. Papp, and L.D. Hurst, *Highly expressed genes in yeast evolve slowly.* Genetics, 2001. **158**(2): p. 927-31.
128. Chuang, T.J. and T.W. Chiang, *Impacts of pretranscriptional DNA methylation, transcriptional transcription factor, and posttranscriptional microRNA regulations on protein evolutionary rate.* Genome Biol Evol, 2014. **6**(6): p. 1530-41.