Published online 30 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D267–D273***
doi:10.1093/nar/gkn782

SysZNF: the C2H2 zinc finger gene database

Guohui Ding[1], Peter Lorenz[2], Michael Kreutzer[2], Yixue Li[1,3] and Hans-Juergen Thiesen[2,*]

[1]Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, P. R. China, [2]Institute of Immunology, Schillingallee 70, University of Rostock, D-18055 Rostock, Germany and [3]Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, P. R. China

Received August 12, 2008; Revised September 19, 2008; Accepted October 9, 2008

ABSTRACT

C2H2 zinc finger (C2H2-ZNF) genes are one of the largest and most complex gene super-families in metazoan genomes, with hundreds of members in the human and mouse genome. The ongoing investigation of this huge gene family requires computational support to catalog genotype phenotype comparisons of C2H2-ZNF genes between related species and finally to extend the worldwide knowledge on the evolution of C2H2-ZNF genes in general. Here, we systematically collected all the C2H2-ZNF genes in the human and mouse genome and constructed a database named SysZNF to deposit available datasets related to these genes. In the database, each C2H2-ZNF gene entry consists of physical location, gene model (including different transcript forms), Affymetrix gene expression probes, protein domain structures, homologs (and synteny between human and mouse), PubMed references as well as links to relevant public databases. The clustered organization of the C2H2-ZNF genes is highlighted. The database can be searched using text strings or sequence information. The data are also available for batch download from the web site. Moreover, the graphical gene model/protein view system, sequence retrieval system and some other tools embedded in SysZNF facilitate the research on the C2H2 type ZNF genes under an integrative view. The database can be accessed from the URL http://epgd.biosino.org/SysZNF.

INTRODUCTION

The zinc finger (ZNF) motif of $CX_{2\text{-}4}CX_3FX_5LX_2HX_{3\text{-}4}HTGEKPYX$ (X is any amino acid) forms an independent mini-domain folded around a central zinc ion and 'gripping' the DNA/RNA with the adjacent ZNFs (1). The two cysteines (C) and two histidines (H) in the motif are conserved. To distinguish this ZNF motif from other ZNF motifs, this motif was designated as 'Cys2His2-ZNF' or 'C2H2-ZNF'. Besides the C2H2-ZNF motifs, C2H2-ZNF proteins embody different effector domains, such as KRAB (Krüppel-associated box), SCAN, BTB and SET (2). These effector domains are usually found in the amino-terminal region of the proteins and often serve different roles in transcriptional regulation through their ability to interact with other cellular molecules (2). The KRAB domain initiates transcriptional repression by recruiting co-repressor complexes (3); the SCAN domain mediates homo- and hetero-oligomerization (4); the BTB domain confers the ability for homomeric and in some instances heteromeric dimerization (5); and the SET domains are considered protein–protein interaction domains (6). Even though the C2H2-ZNF motif is ancient and can be detected in Archaea and Eukaryota according to the Pfam database (7), most of the effector domains emerged more recently and many C2H2 genes containing them show lineage-specific expansion (2). One prominent example is the KRAB domain that is found in the largest subfamily of C2H2-ZNF proteins, initially identified as heptad repeat of leucines in KOX1/ZNF10 (8). This domain can be traced back to the base of the deuterostomes (9) but dramatically expanded only in mammals (3).

C2H2-ZNF proteins are assumed to influence the expression of downstream genes by mediating interactions, predominantly between DNA sequences and regulatory proteins. They have been implicated in processes and pathways such as development, cell proliferation and complex phenotypes including disease [e.g. mouse C2H2-ZNF gene *zfp90* as one of the key drivers of the omental fat pad mass trait (10)]. Defined C2H2-ZNF motifs have been applied to engineer custom zinc-finger nucleases (ZFN) for gene therapy (11). Designed ZFN have been used to inactivate genes in zebrafish (12,13) and to disrupt the CCR5 gene in $CD4^+$ T cells to acquire HIV-1 resistance (14). Thus, in-depth characterization of

*To whom correspondence should be addressed. Tel: +49 381 494 5870; Fax: +49 381 494 5882; Email: hans-juergen.thiesen@med.uni-rostock.de
Correspondence may also be addressed to Yixue Li. Tel: +86 21 54920089; Fax: +86 21 54920143; Email: yxli@sibs.ac.cn

C2H2-ZNF genes will broaden our understanding of the regulation of transcriptional networks as well as extend our repertoire of tools applicable in genetic engineering projects.

Genes encoding C2H2-ZNF containing proteins make up ~40% of all human transcription factor genes (15) and constitute the second largest paralog gene family in human (16). Notably, numerous C2H2-ZNF genes, in particular those with KRAB and/or SCAN domain, are arranged in clusters in the human genome, indicating possible tandem *in situ* duplications of these genes (17). The expansion and ongoing evolution of these subfamilies of C2H2-ZNF genes occurred not only in human but also across mammals and in a lineage-specific fashion (2). The comparative analysis of all C2H2-ZNF genes in mammals gives a model for the interlinked evolution of SCAN, SCAN–KRAB and KRAB subfamilies (2).

Although, a complete catalogue of C2H2-ZNF genes can give a bird's-eye view of these genes and promote the research of these genes on the omics level, there is no special database existing to deposit these genes with related annotation information to date. A well characterized database related to human KRAB C2H2-ZNF genes online is the Human KZNF Gene Catalog, which was designed to derive a complete catalog of the KRAB–ZNF gene family (17). However, the Human KZNF Gene Catalog focuses on KRAB ZNF genes only and provides a compact web interface. Another web interface, based on the systematic characterization of ZNF proteins of all types (not just C2H2) in the mouse transcriptome and initially accessible at http://cassandra.visac.uq.edu.au/zf (18), appears to be off-line. There is thus a need for a comprehensive catalog of C2H2-ZNF genes with functional annotations and graphical interface.

Here, we provide an online resource, SysZNF, to catalog the C2H2-ZNF protein coding genes in human and mouse. Our goal was to develop a comprehensive database with a web interface that (i) arranges C2H2 gene information based on physical localization in the respective genome, (ii) provides downloadable and visual information on gene organization and protein domain composition, (iii) serves as starting point to search and examine ortholog and paralog relationships in different species and (iv) contains direct links from a particular data set to other resources of the scientific community.

In SysZNF, each C2H2-ZNF gene is represented as a card to integrate the related information from public databases. A graphical gene model system, a protein display system and a gene cluster display system were developed for these genes. The SysZNF database is maintained jointly by the Proteome Center at the University of Rostock in Germany and the Key Laboratory of Systems Biology at the Shanghai Institute for Biological Sciences in China from 2006. In the current version (Release 3.0), 740 fully annotated human and 780 mouse C2H2-ZNF genes are cataloged (Table 1). The resource can be freely accessed at http://epgd.biosino.org/SysZNF. The old versions except the first version developed in 2006 which was an in-house database, can also be visited.

METHODS

We regarded any gene encoding at least one C2H2-ZNF motif as a C2H2-ZNF gene. In our database, the HMM profile from the Pfam database (accession number: PF00096) (7) was used to represent the C2H2-ZNF motif. Then, the whole proteomes of human and mouse were scanned with this HMM profile by hmmpfam in the HMMER package (19). Lastly, a full annotation of these C2H2-ZNF genes was done. A brief description of these steps is listed below.

(i) Protein sequences and gene model information of human and mouse were downloaded from AceView (20).

(ii) All putative C2H2-ZNF motifs were identified in the protein sequences using the hmmpfam in HMMER package (19). We used a conservative value of 1E-5 to determine 765 C2H2-ZNF genes in human and 806 C2H2-ZNF genes in mouse. The population size of human C2H2-ZNF gene is similar to the size determined by other researchers which was 718 (2). For mouse C2H2-ZNF genes, 506 were reported previously (18). This smaller size may result from the incomplete dataset of the mouse transcriptome then.

(iii) The gene model information was compiled for the genes chosen in step (ii). The coordinates of 25 human C2H2-ZNF genes and 26 mice C2H2-ZNF in genomes were incomplete. These genes can not be mapped to any chromosome.

(iv) Additional domains (e.g. effector domain) were documented in the C2H2-ZNF protein sequences identified in step (ii) with hmmpfam (19) against the Pfam (7) database (Table 1).

(v) Information from other public databases was integrated, such as gene model information from AceView (20) and UCSC genome Browser (21), human and mouse synteny information from Ensembl (22), literature information from NCBI Entrez Gene (23) and iHop (24), cross references to Swissprot (25), Treefam (26), InterPro (27), dbPTM (28), etc. Three different criteria were used with decreasing impact: the first criteria is that two items from different databases have a similar physical region in the genome with an overlap of more than 80%; the second is that two items have similar sequences [similarity >90% using BLAST search (29)]; the third is that two items share authoritative IDs (e.g. NCBI Entrez Gene ID).

(vi) Adjacent C2H2-ZNF genes with an interval distance <500 kb in the chromosome were considered to belong to a physical 'cluster' (2). Note that 500 kb is only the default setting in SysZNF, the user can input any distance to infer physical clusters of adjacent C2H2-ZNF genes in the web site for further analysis.

(vii) All-to-all BLAST search for the human and mouse protein sequences were applied to get the putative ortholog and paralog sequences. The synteny regions of these two species were downloaded from Ensembl (http://www.ensembl.org) directly (22).

Table 1. Summary of the C2H2-ZNF genes in human and mouse

	Human	Mouse
C2H2-ZNF genes	740	780
Additional domains[a]	17	19
Physical clusters in the genome[b]	90	95
Synteny regions[c]	156	

[a]Conserved additional domains (e.g. effector domains) comprised in C2H2-ZNF protein sequences that can be used to define new subfamilies. The domains counted here should be present in more than three genes.
[b]Clusters are defined by complying with two conditions: (i) they have at least two C2H2-ZNF genes and (ii) the intergenic distance between the included adjacent ZNF genes is within a physical interval <500 kb. Note that the number of the clusters will vary depending on the intergenic distance chosen. The SysZNF presents a tool to infer physical distance-based clusters with any interval C2H2-ZNF gene distance setting.
[c]The information on syntenic regions between human and mouse was downloaded from Ensembl (22).

(viii) A graphical gene model and protein viewer system was implemented. We developed and embedded a genomic sequence retrieval system in the database to access the sequences in the human or mouse genome easily.
(ix) A text search and sequence search system was developed and installed.

CONTENT IN THE DATABASE

The gene models of the C2H2-ZNF genes are at the center of the database. The region of each gene model is defined as the region between the minimum left position of all of its transcripts and the maximum right position. A total of 1520 C2H2-ZNF genes (740 and 780 genes in human or mouse, respectively) has been mapped to the genomes. Taking the protein sequences encoded by these genes of both species, the top 10 frequent effector domains encode KRAB domains (PF01352, 649 genes), BTB domains (PF00651, 97 genes), SCAN domains (PF02023, 89 genes), SET domains (PF00856, 15 genes), PHD domains (PF00628, 12 genes), Homeobox domains (PF00046, 10 genes), Zfx_Zfy_act domains (PF04704, 8 genes), bZIP_1 domains (PF00170, 6 genes), bZIP_2 domains (PF07716, 6 genes) and ELM2 domains (PF01448, 5 genes). The PHD, bZIP and ELM2 domains in the C2H2-ZNF proteins indicate new C2H2-ZNF subfamilies. Besides these effector domains, another frequently occurring additional domain is DUF1610 (PF07754, 140 genes), which is likely to bind zinc via its four well-conserved cysteine residues according to the Pfam database (7).

The C2H2-ZNF genes in SysZNF have been allocated to physical clusters whose C2H2-ZNF genes share physical proximity on the chromosome. The definition of proximity is based on the distribution of the interval distances between genes, and 200 kb (17) and 500 kb (our default) (2) have been used in the context of KRAB or C2H2-ZNF genes. Our cluster allocation has been implemented to point at an initial list of ZNF genes for further detailed analysis of evolutionary/synteny relationships. Based on the default definition of the gene cluster in the SysZNF, there are 90 C2H2-ZNF clusters detected in human chromosomes and 95 clusters in mouse chromosomes, which cover 69.2% human C2H2-ZNF genes and 71.7% mouse C2H2-ZNF genes. The coverage estimated here is a little different from the value reported in human previously (72%) (2), possibly due to different versions of the genomic coordination system. The top five biggest clusters in human accounting for 27% of human C2H2-ZNF genes are hosted on chromosome 19. For mouse, the top five biggest clusters covering 21% of all C2H2-ZNF genes are located on chromosomes 2, 13, 17, 4 and 10.

SysZNF integrated the synteny information from Ensembl (22). One hundred and fifty-six out of 349 synteny regions (46%) contain the C2H2-ZNF genes. SysZNF also presents direct cross references to public databases, such as AceView (20), iHop (24), Allen Brain Atlas (30), Human Protein Atlas (31), Swissprot (25), Treefam (26), InterPro (27), dbPTM (28), NCBI Map viewer (http://www.ncbi.nlm.nih.gov/projects/mapview/) (23) and UCSC genome Browser (http://genome.ucsc.edu/) (21). Thus, a plethora of information can be easily accessed to help in the formulation of hypotheses for further studies. In order to propose DNA binding signatures of particular C2H2 motifs, an embedded program forwards any user-defined ZNF array (limited to maximal four consecutive ZNFs) to a server at the Hebrew University of Jerusalem that predicts possible DNA bindings sites (32).

ACCESS AND WEB FEATURES

The SysZNF was implemented as a MySQL relational database. An interactive web interface was created using Java Server Pages technology, hosted in an Apache Tomcat web server on a Linux machine. The schema of the database design is available on the web site. The graphical gene model/protein view system, sequence retrieval system and some other tools were developed in Java.

Browsing C2H2-ZNF genes

The C2H2-ZNF genes in the SysZNF can be browsed through chromosomes or by effector/additional domains (Figure 1A). In the chromosome browsing view, the deposited genes are marked as red line adjacent to the chromosome. A chromosome is linked (mouse click) to a list with all the C2H2-ZNF genes residing in this chromosome. In the protein domain browsing view, each effector/additional domain can be chosen by clicking and the user will be forwarded to a page with all C2H2-ZNF genes with this domain. For both views, the user can download all the protein sequences or nucleic acid sequences in a list. If the list has been generated through the domain view, the domain regions of the protein sequences of this list can be downloaded.

Searching SysZNF

SysZNF supports three types of searching modes (Figure 1B). The first type is text search. The user can enter any keyword using 'QUICK SEARCH'. The gene

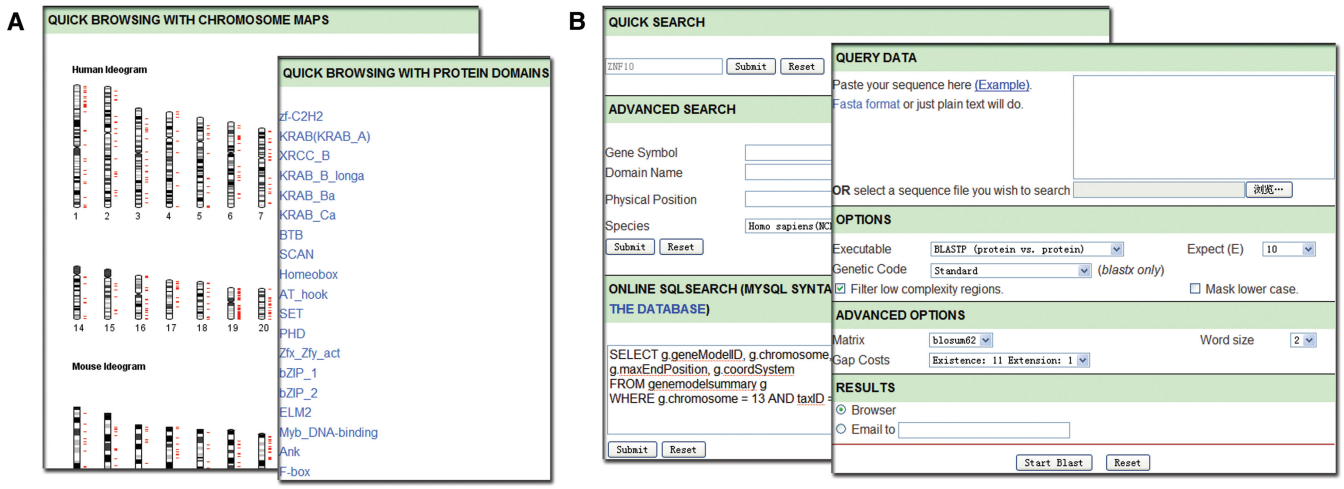

Figure 1. Browsing and searching in SysZNF. (**A**) Browsing SysZNF through chromosomes or by domains. (**B**) Text strings, SQL and bio-sequences searching. The gene symbol, protein domain name and physical location can be used as search fields in the 'ADVANCED SEARCH'. Only the 'search' statement could be used in the 'ONLINE SQL SEARCH'. Both protein and nucleic acids sequences can serve as input in the BLAST search page. The user can also access the result of BLAST by email.

symbol, protein domain name and physical position can be used in the 'ADVANCED SEARCH'. The second type is the online SQL search. This utility is designed for the advanced user who is familiar with the schema of the database. For security reason, only the 'select' statement is supported. The third one is the sequence search. The NCBI BLAST package (29) was used as the sequence search engine. Both protein sequences and DNA/RNA sequences can be used as input sequences. The result can be browsed directly or accessed later by email. An ID will be assigned for each result. The user can retrieve his/her result by this ID. Note that the BLAST result will be deleted one day later. In addition, some example queries can be found in the help page of the database (http://epgd.biosino.org/SysZNF/help.jsp#example).

C2H2-ZNF gene entry

The C2H2-ZNF gene entry consists of physical location, gene model, Affymetrix probe sets, protein domain, homologs, related literature and cross-references to public databases (Figure 2). A tool bar was designed to facilitate the browsing of this page (Figure 2A). The user can download the genomic sequences from an embedded sequence retrieval system (Figure 3A), UCSC, Ensembl or NCBI. We developed a graphical gene model/protein viewing system (Figure 2A). The elements in the gene model were rendered as special logo. For example, a red arrowhead denotes the start codon and a black one denotes the stop codon of an isoform. The isoforms are aligned according to their physical locations in the gene model figure. The protein domains were displayed as rectangle, whose colors are corresponding to the *E*-value of this domain predicted by HMMER. The indicators of gene model elements and protein domains in the images can be clicked to retrieve the related sequences.

If a gene belongs to a C2H2-ZNF gene cluster, a page with the cluster information can be linked (Figure 2B). All C2H2-ZNF genes in this cluster will be displayed in a figure and listed below. The genes listed can be linked out to the gene entry page.

We compared the human C2H2-ZNF genes with the mouse C2H2-ZNF genes with protein sequences and derived the putative orthologs and paralogs (Figure 2C). The homolog pairs can be aligned by bl2seq in the NCBI BLAST package. If the gene is in a syntenic region of human and mouse, the user can access it by clicking a cross link (Figure 2D). In the synteny page, the physical coordinates of the region in both species are displayed. These regions can be accessed by Ensembl ContigView (22). The C2H2-ZNF genes in this region are also listed.

The bottom region in the gene entry page is the literature list of this gene and includes some cross-references to other databases (Figure 2E). Figure 2E displays only part of the whole page (to access the whole page, the user can browse the database at http://epgd.biosino.org/SysZNF).

Some utilities

A genomic sequence retrieval system was developed in Java and embedded in SysZNF (Figure 3A). The genomic sequences can be retrieved by physical location, gene symbol and Affymetrix probeset name one by one or in batch style. When using gene symbols or Affymetrix probeset name, only the gene symbols deposited in the SysZNF are supported.

When the user wants to combine different C2H2-ZNF motifs in a protein to a ZNF array, a compact tool named FunnyFingerSelector can be used (Figure 3B). From each gene entry or designed C2H2-ZNF array, three to four individual motifs can be arbitrarily selected to construct a novel arrangement of motifs. Then, this C2H2 array can be used to predict putative DNA binding sites (32). The tool is helpful for comparing theoretical binding site predictions with experimentally obtained specificities on the way (i) to engineered transcription factors with desired functions and (ii) to the description of transcriptional networks mastered by C2H2-ZNF genes.

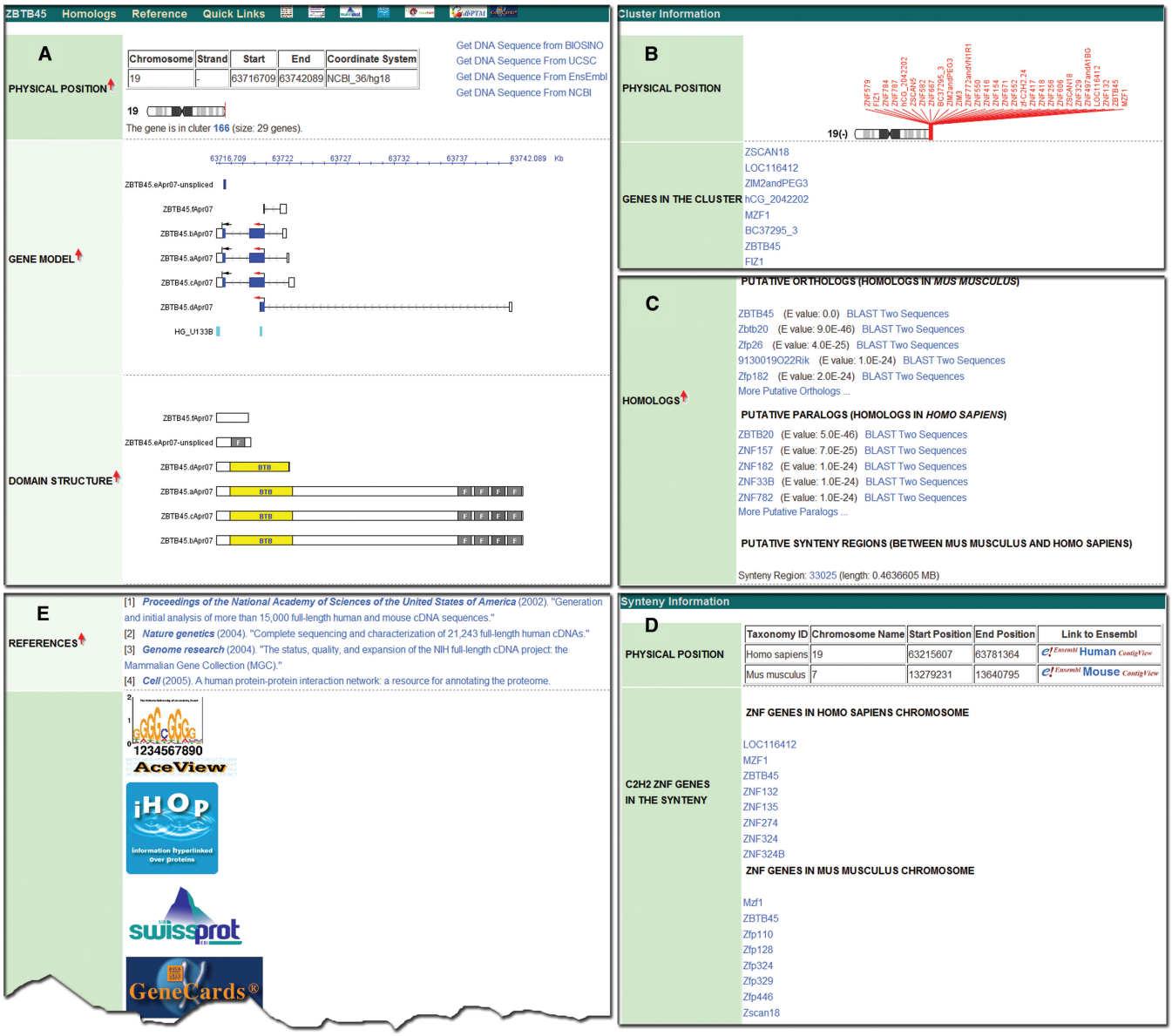

Figure 2. Screenshot of a C2H2-ZNF gene entry. (**A**) Physical coordinates, gene model and domain structures. (**B**) A C2H2-ZNF gene cluster. (**C**) Putative homologs and synteny regions. (**D**) A detailed synteny region between the human and mouse. (**E**) Literatures and cross-references related to the gene in this entry. The toothed margin of this screenshot denotes that it is only part of the whole gene entry page.

SysZNF also supports a tool to cluster the hosted C2H2-ZNF genes (Figure 3C; http://epgd.biosino.org/Utilities2007/funnyCluster/). The user can set any interval distance between two adjacent C2H2-ZNF genes in the chromosome. In the result page, this tool will give a list of all genes and corresponding cluster ID and a summary of each cluster.

Data Availability

All the data in SysZNF are freely available without password protection. The result pages for the query system provide the utilities to batch download the sequences. The nucleic acid and peptide sequences, probe set locations, gene model information and cluster information in the database can be downloaded in flat file on the 'DOWNLOAD' page. The source code for SysZNF is available from the corresponding authors under the GNU General Public License 2.0. Users are solicited to reference the usage of this database.

FUTURE DIRECTIONS

SysZNF was originally developed to systematically catalog the C2H2-ZNF genes of human and mouse to make them accessible for further studies. Once mammalian genomes are completed, C2H2-ZNF sets of these species will have to be integrated as well. In particular, a full understanding of ZNF biology does require to link

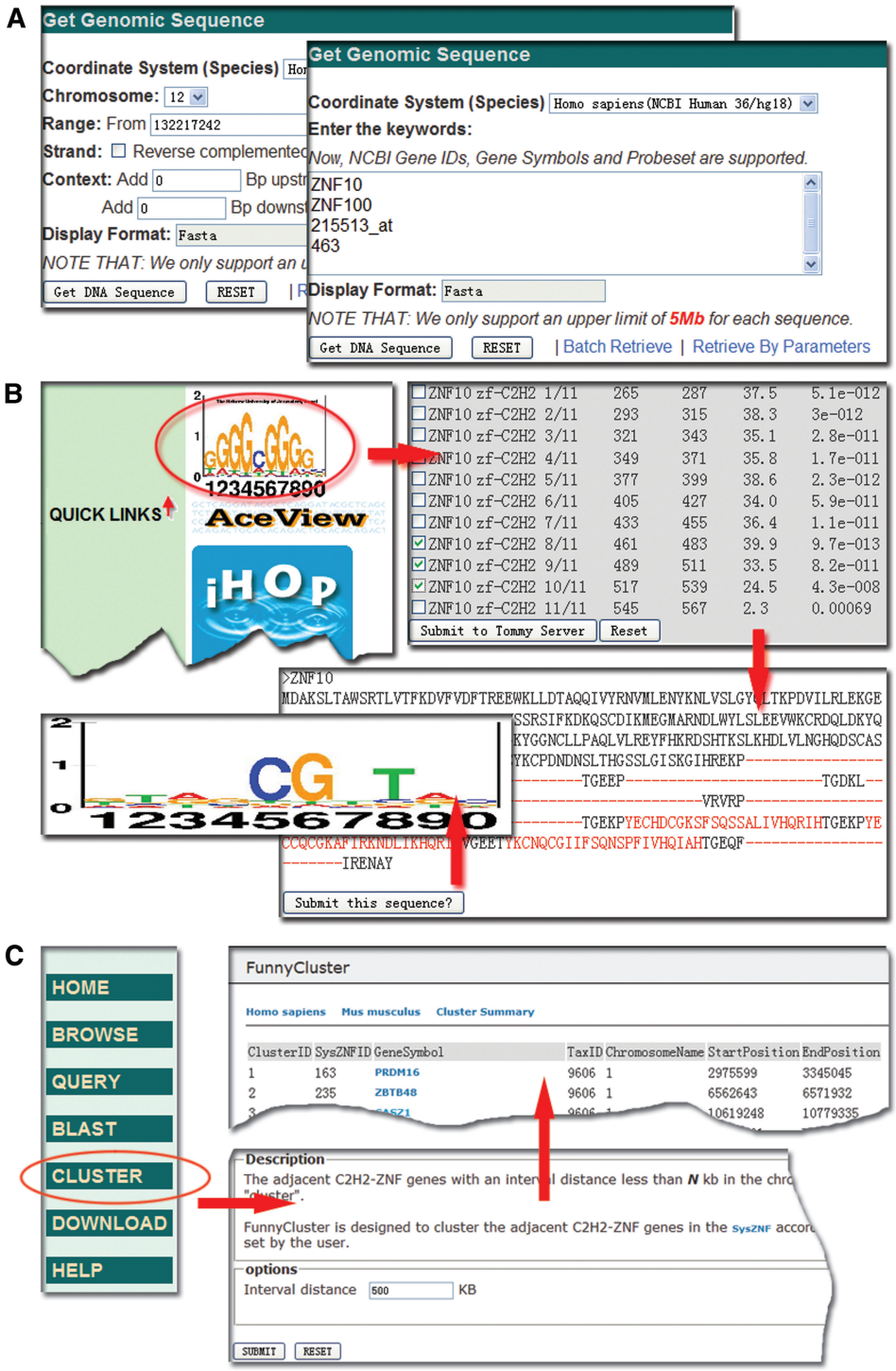

Figure 3. Some embedded tools. (**A**) Genome sequence retrieval system. (**B**) FunnyFingerSelector, a tool to combine individual fingers and predict the DNA binding sites of the resulting ZNF array. (**C**) FunnyCluster, a tool to infer physical distance-based clusters of C2H2-ZNF genes in SysZNF according to any user defined interval distance.

molecular data to topological data of genes' products, e.g. like the cell and tissue-specific expression patterns presented by the Allen Brain Atlas (30) or the Human Protein Atlas (31). In the future, we plan to add more species for comparative analysis of C2H2-ZNF gene families and to update SysZNF annually. Important expansions in the future will be to add more utilities for the analysis of C2H2-ZNF genes, in particular, to support systems biological approaches in modeling C2H2-ZNF gene functions in mammalian evolution.

ACKNOWLEDGEMENTS

We thank Chuan Wang, Yun Li, Zhen Wang, Hong Li, Guangyong Zheng for helpful comments and suggestions.

FUNDING

BMBF of Germany (CHN07/38); National High-Tech R&D Program (863) (2006AA02Z334, 2006AA020406); National Basic Research Program of China (2006CB910700, 2004CB720103, 2004CB518606, 2003CB715901); National Natural Science Foundation of China (30621091); Ministry of Science and Technology (2006CB943900, 2007CB947904, 2007CB947100, 2007CB948000); Key Research Program (CAS) (KSCX2-YW-R-112); Chinese Academy of Sciences (International Technology Collaboration Project, 2007DFA31040). Funding for open access charges: BMBF of Germany (CHN07/38).

Conflict of interest statement. None declared.

REFERENCES

1. Klug,A. and Schwabe,J.W. (1995) Protein motifs 5. Zinc fingers. *FASEB J.*, **9**, 597–604.
2. Tadepally,H.D., Burger,G. and Aubry,M. (2008) Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol. Biol.*, **8**, 176.
3. Margolin,J.F., Friedman,J.R., Meyer,W.K., Vissing,H., Thiesen,H.J. and Rauscher,F.J. 3rd (1994) Kruppel-associated boxes are potent transcriptional repression domains. *Proc. Natl Acad. Sci. USA*, **91**, 4509–4513.
4. Sander,T.L., Stringer,K.F., Maki,J.L., Szauter,P., Stone,J.R. and Collins,T. (2003) The SCAN domain defines a large family of zinc finger transcription factors. *Gene*, **310**, 29–38.
5. Huynh,K.D. and Bardwell,V.J. (1998) The BCL-6 POZ domain and other POZ domains interact with the co-repressors N-CoR and SMRT. *Oncogene*, **17**, 2473–2484.
6. Min,J., Zhang,X., Cheng,X., Grewal,S.I. and Xu,R.M. (2002) Structure of the SET domain histone lysine methyltransferase Clr4. *Nat. Struct. Biol.*, **9**, 828–832.
7. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
8. Thiesen,H.J. (1990) Multiple genes encoding zinc finger domains are expressed in human T cells. *New Biol.*, **2**, 363–374.
9. Birtle,Z. and Ponting,C.P. (2006) Meisetz and the birth of the KRAB motif. *Bioinformatics*, **22**, 2841–2845.
10. Schadt,E.E., Lamb,J., Yang,X., Zhu,J., Edwards,S., Guhathakurta,D., Sieberts,S.K., Monks,S., Reitman,M., Zhang,C. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.
11. Cathomen,T. and Joung,J.K. (2008) Zinc-finger nucleases: the next generation emerges. *Mol. Ther.*, **16**, 1200–1207.
12. Doyon,Y., McCammon,J.M., Miller,J.C., Faraji,F., Ngo,C., Katibah,G.E., Amora,R., Hocking,T.D., Zhang,L., Rebar,E.J. *et al.* (2008) Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 702–708.
13. Meng,X., Noyes,M.B., Zhu,L.J., Lawson,N.D. and Wolfe,S.A. (2008) Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 695–701.
14. Perez,E.E., Wang,J., Miller,J.C., Jouvenot,Y., Kim,K.A., Liu,O., Wang,N., Lee,G., Bartsevich,V.V., Lee,Y.L. *et al.* (2008) Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat. Biotechnol.*, **26**, 808–816.
15. Messina,D.N., Glasscock,J., Gish,W. and Lovett,M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.
16. Ding,G., Sun,Y., Li,H., Wang,Z., Fan,H., Wang,C., Yang,D. and Li,Y. (2008) EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogon information. *Nucleic Acids Res.*, **36**, D255–D262.
17. Huntley,S., Baggott,D.M., Hamilton,A.T., Tran-Gyamfi,M., Yang,S., Kim,J., Gordon,L., Branscomb,E. and Stubbs,L. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.*, **16**, 669–677.
18. Ravasi,T., Huber,T., Zavolan,M., Forrest,A., Gaasterland,T., Grimmond,S. and Hume,D.A. (2003) Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome Res.*, **13**, 1430–1442.
19. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
20. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl 1)**, S1211–S1214.
21. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E., Siepel,A. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
22. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
23. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
24. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
25. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
26. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 update. *Nucleic Acids Res.*, **36**, D735–D740.
27. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
28. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Lein,E.S., Hawrylycz,M.J., Ao,N., Ayres,M., Bensinger,A., Bernard,A., Boe,A.F., Boguski,M.S., Brockway,K.S., Byrnes,E.J. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.
31. Uhlen,M., Bjorling,E., Agaton,C., Szigyarto,C.A., Amini,B., Andersen,E., Andersson,A.C., Angelidou,P., Asplund,A., Asplund,C. *et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics*, **4**, 1920–1932.
32. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.