CardioTF, a database of deconstructing transcriptional circuits in the heart system

Yisong Zhen

Background. Information on cardiovascular gene transcription is fragmented and far behind the present requirement at the systems biology level. To facilitate deep learning of genomic data, the CardioTF database was constructed for collecting cardiovascular information on heart transcription factors (TFs), position weight matrices, and enhancer sequences discovered using the ChIP-seq method. **Methods.** The Naïve-Bayes algorithm was used to classify literature and identify all PubMed abstracts on cardiovascular development. The nature language learning tool GNAT was then used to identify corresponding gene names embedded within these abstracts. Local Perl scripts were used to integrate and dump data from public databases into the MariaDB management system (MySQL). In-house R scripts were written to analyze and visualize the results. **Results.** Cardiovascular TFs from fly, Ciona, zebrafish, frog, chicken, mouse and human were collected on the basis of their human counterparts. Mouse homologs were singled out for defining the core TFs by intersecting four independent data sources. Position weight matrices from Jaspar, hPDI, and UniPROBE databases were deposited in the database and can be retrieved by their corresponding TF names. Gene enhancer regions from various sources of ChIP-seq data were deposited into the database and can be visualized by graphical output. **Discussion.** The database can be used as a portal to construct transcriptional network in cardiac development. **Availability and Implementation.** Database URL: http://www.cardiosignal.org/database/cardiotf.html

CardioTF, a database of deconstructing transcriptional circuits in the heart system

Yisong Zhen[1]

1. State Key Laboratory of Cardiovascular Disease,Fuwai Hospital, National Center for Cardiovascular Diseases Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100037, People's Republic of China

Corresponding Author:
Yisong Zhen
Beilishilu, 167, Beijing, 100037 P.R.China
Email address: zhenyisong@fuwaihospital.org

Abstract

Background. Information on cardiovascular gene transcription is fragmented and far behind the present requirement at the systems biology level. To facilitate deep learning of genomic data, the CardioTF database was constructed for collecting cardiovascular information on heart transcription factors (TFs), position weight matrices, and enhancer sequences discovered using the ChIP-seq method.

Methods. The Naïve-Bayes algorithm was used to classify literature and identify all PubMed abstracts on cardiovascular development. The nature language learning tool GNAT was then used to identify corresponding gene names embedded within these abstracts. Local Perl scripts were used to integrate and dump data from public databases into the MariaDB management system (MySQL). In-house R scripts were written to analyze and visualize the results.

Results. Cardiovascular TFs from fly, Ciona, zebrafish, frog, chicken, mouse and human were collected on the basis of their human counterparts. Mouse homologs were singled out for defining the core TFs by intersecting four independent data sources. Position weight matrices from Jaspar, hPDI, and UniPROBE databases were deposited in the database and can be retrieved by their corresponding TF names. Gene enhancer regions from various sources of ChIP-seq data were deposited into the database and can be visualized by graphical output.

Discussion. The database can be used as a portal to construct transcriptional network in cardiac development.

Availability and Implementation. Database URL:
http://www.cardiosignal.org/database/cardiotf.html

Introduction

Heart disease is a leading cause of morbidity and mortality in both infants and adults [1,2]. Insight into the cause of congenital heart diseases (CHDs) has led to the identification of mutations in essential cardiac transcription factors (TFs) [3]. At the opposite end of the temporal spectrum, adult cardiac disease can be traced to variants of gene regulatory sequences [4]. Thus, knowledge of TFs, their downstream targets, and regulatory genomic sequences in the heart transcriptional regulatory network will enhance our understanding of the heart disease process.

Although the torrent of data generated by high-through-put technologies is archived in databases such as ArrayExpress or GEO of NCBI [5,6], they do not contain cohesive knowledge and lack expert annotation. In addition, cardiac development has experienced accelerated growth that can be attributed to the use of various animal models. However, to our present understanding, few efforts have been committed to build a database which collects transcriptional information across species, thereby limiting the benefits from evolutionary perspectives.

To data, a limited number of databases have been dedicated to collecting similar data aspects. For example, BloodChIP [7] is a database of comparative genome-wide TF binding profiles in human blood cells. CistromeMap [8] is a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mice and humans. However, there is no unique database committed to cardiovascular development. This prompted us to combine information about TFs, position weight matrces (PWMs), and ChIP-seq results and release a one-stop site for all such information.

CardioTF was therefore constructed to capture transcriptional information regarding cardiovascular development. It documents TFs across species, including fly, *Ciona*, fish, frog,

chicken, mouse and human. To facilitate downstream analysis, CardioTF also collects PWM files from public databases. Moreover, to visualize the enhancers of the genes of interest, it provides a simple graphical output for revealing the enhancers discovered in various ChIP-seq experiments.

Materials & Methods

Comprehensive collection and annotation of cardiac TFs

Cardiac TFs were previously defined to regulate the expression of cardiac genes, which impact the process of heart development. TFs themselves should be involved in the steps of specification, determination, patterning, and differentiation that will result in a myocardial fate. In principle, we recruited cardiac TFs involved in the development of the epicardium and endocardium, both of which, along with myocardium, should be from the Mesp1-expressing cell lineage [9]. In the initial screening, we collected the repertoire of cardiac TFs from previously published annotations [10]. We also used this human set, excluding human-specific TFs as the reference set to search for their homologs in other species, including fly, *Ciona*, zebrafish, chicken and mouse [11]. The human-specific TF is defined as the gene which has no homologs in the mouse genome. We also documented the mouse TF expression status from four independent sources at different developmental stages. They included annotation from the Cardiovascular Gene Ontology Annotation Initiative [12], Mouse Genome Database (MGI) genes with cardiovascular phenotype [13], PubMed abstract parsing results and RNA expression profiling results.

TFs from PubMed abstract parsing

The Weinstein Cardiovascular Conference provides a platform for talks and posters on all aspects of heart development and congenital heart disease. We used the Weinstein meeting

abstracts from year 2010 to 2013 as the positive group, thus including 954 abstracts. In this regards, we assumed that Weinstein-like abstracts deposited in PubMed are all from the cardiovascular community and focus on cardiovascular development. Abstracts from the negative group were from non-heart related journals, which were manually selected from PMC Open Access Subset at NCBI. The negative group includes 57080 abstracts. We split the whole data (positive and negative groups) into training (80%) and test (20%) set. The Naïve-Bayes module from The Comprehensive Perl Archive Network (CPAN) was used by the local Perl script to classify Weinstein-like abstracts. We used training set and adopted the 5×2 cross-validation proposed by Dietterich [14] to train and validate the data. A wrapper function was implemented to parse the abstract and calculate the word frequency. In doing so, the word frequency was forwarded to the algorithm as the only feature. A separated test set using optimized parameter was used to assess the algorithm performance. The selected abstracts were then processed by GNAT [15] using its default script (test100.sh) to recognize the gene name in mouse.

RNA expression profiling data procession

Affymetrix data (GSE1479) were processed by R using the MAS5 algorithm which provides a present call for each gene. The gene expression status was defined as “on” if the gene in any array of those data at certain developmental stage had a present call. RNA-seq data (GSE47950 and GSE29184) were re-analyzed using the recommended protocol [16]. Any gene with an FPKM value greater than 1 was defined as expressed.

Depositing PWM files

The gene symbol was used as the unique identifier to link the original database ID to our local database primary key. A local Perl script was written to change their format to TRANSFAC style,

which was used by our in-house CardioSignalScan program. PWM files were collected from Jaspar, UniPROBE and hPDI databases [17-19]. Users can retrieve their annotations by directing them to the respective database.

Orthologs of TFs from model systems

NCBI has its own collection of all ortholog gene collections using its unpublished algorithm. TFs from mouse, human and zebrafish are annotated by NCBI [20]. Frog, chicken and *Ciona* TF homolog annotations were downloaded from their center database including Xenbase, BirdBase and ANISEED [21-23]. Fly TFs, which have their counterparts in the human proteome, were annotated by the Inparanoid system [24]. Each TF collected in the database was assigned one treeID on the basis of its human counterpart. The treeID is equivalent to a TF family by the recommendation of TFClass.

Enhancer curation: TF-ChIP and Histone-ChIP data processing

Enhancer regions were defined as the ChIP-seq signals. Peak calling was performed using the recommended pipeline [25]. In brief, sequencing reads were aligned to the mm10/hg19 reference genome using Bowtie/Bowtie2.

Bowtie call

bowtie -m 2 -S -q -p 8

Peak calling was performed using the MACS peak calling algorithm [26].

MACS call

macs14 -t ERR231646.bam -c ERR231653.bam -g mm -n sham_Anti_H3k9ac

Detailed annotation of the peaks was performed by HOMER using annotatePeaks.pl package [27].

125 # Results

126 ## The database schema

127 Our database uses MariaDB, a drop-in replacement for MySQL, as a database management
128 system (DBMS). For answering questions about what information will be stored and how the
129 elements in it will be related to one another, we used unified modeling language (UML) to
130 describe the high-level database model. UML was originally developed as a graphical notation
131 for describing software designs in an object-oriented style. It has been extended, with some
132 modifications, to be a popular notation for describing database designs. Here we used UML
133 instead of entity/relationship diagram to design the relational database schema following
134 modeling principles, such as faithfulness, avoiding redundancy, and simplicity counts (Fig. 1). If
135 possible, we used composition distinguished by a line between two classes that ends in a solid
136 diamond at one end. The diamond implies that the label at the end must be 1:1. For example,
137 there is a composition from CardioTFmatrix to CardioTFCenter, which means that every matrix
138 annotation row (PWM related information) belongs to exactly one row in CardioTFCenter (one
139 type of TF may have more PWM records in a CardioTFmatrix table). A 1:1 label at the
140 CardioTFCenter end is implied by a solid black diamond.

141 ## Web Interface and search engine

142 CardioTF is a Perl website implemented using only Perl language to dynamically display the
143 graphical output while querying the database in the backend (Fig. 2). To aid cardiovascular
144 biologists, a search engine was created to allow users to tackle the following problems: (1)
145 identify homology information regarding the queried TF across six species and link to the
146 corresponding central databases outside CardioTF; (2) identify PWM file union of three public
147 databases regarding the queried TF; and (3) identify the enhancer regions revealed by ChIP-seq

148 data of the queried gene. Thus, the key aspects of the database functionality intend to construct
149 essential information required for transcriptional network in heart development.

150 **Cardiovascular TFs in the database**

151 Wingender's annotation set was used as a benchmark to recruit TFs across species. Human-
152 specific TFs, defined as those with no orthologs in the mouse genome, were discarded because
153 no model system could be used to verify their function *in vivo*. Therefore, 1200 mouse TFs were
154 collected. Other established animal models for cardiovascular development include fly, *Ciona*,
155 zebrafish, frog, and chicken. TFs from these species were collected if they were homologs to the
156 above mouse TFs. The distribution of TFs from different species is shown in Fig. 3. The
157 expression status of mouse TFs was documented in four independent resources, namely RNA-
158 seq data re-analysis, phenotype annotation from MGI database annotation, expert
159 recommendation from the UK Cardiovascular Gene Annotation Initiative, and PubMed
160 relevance from classification of Weinstein-like abstracts (see the subsequent section and Table
161 S1).

162 **Weinstein TFs from PubMed analysis**

163 We determined the journals that favored Weinstein-style papers, which may contain gene
164 information regarding cardiovascular development. As expected, after using a machine learning
165 method, the journals were identified with the journal titles, among which are the 30 journals
166 most relevant to developmental biology. Two of the journals (*Circ. Res*. and *J Mol Cell Cardiol.*)
167 obviously are specialized in heart theme (Table S1: CardioJournalDistribution). We then used
168 GNAT, a tool that recognizes gene names in the literature, to recover all TFs mentioned in
169 Weinstein-style abstracts because we assumed that the TFs are studied by researchers in the
170 cardiovascular community (Fig. 4, Fig. S1 and Table S1).

PWM files collected in database

Public databases for PWM files include UniPROBE, Jaspar, and hPDI, and they provide PWM files for TFs. Jaspar PWM files are curated from the published literatures whereas the other two databases generate PWM files from experiments. Our database integrates these three sources, and the TF PWM can be queried on the basis of the TF name. Search results directly link to the original database through the PWM raw database key. The CardioTFmatrix class now contains 904 records, and these PWM files can be recognized by our local CardioSignalScan program to search for the motifs in genomics regions.

Core cardiovascular TFs

The 1200 mouse TFs were included in the cardiac TF dataset as the entry point to initiate deep annotation. To define the core dataset of cardiac TFs, we intersected four independent sources of cardiovascular TF collections. These 81 TFs are evidenced by their expression pattern, phenotype annotation, expert recommendation and PubMed relevance (Table S1). We also performed DAVID functional analysis, and found that TFs are particularly enriched in cardiac muscle differentiation (Fig. S2). Is this set of gene the minimum requirement to identify the cardiovascular system? Indeed, these four lines of evidence indicate that these TF genes play key roles in heart development. Do these TFs display specific expression patterns in heart developmental stages? The heatmap generated using seven RNA-seq data of various development stages did not reveal any specific pattern (Fig. S3). In adult tissues, these TFs did not exhibit enriched expression in the adult heart. In case of TFs which are never expressed at any stage of heart development, no specific expression pattern was revealed by the boxplot assay (Fig. S4-S5).

Cardiovascular enhancers collected in this database version

Few enhancers have already been verified by traditional biological experiments, for example, by using transgenic expression of isolated DNA fragment *in vivo* to analyze temporal-spatial patterns. Therefore, the ChIP-seq method provides a high-through-put approach to delineate enhancer regions at the genome scale. A standard protocol was used to identify genome-wide locations of transcription/chromatin factor binding sites or histone modification sites from ChIP-seq data(Fig. S6). The present database houses 511,893 enhancer records, covering different developmental stages of the heart.

Discussion

We identified 5442 TFs from six species, and integrate 904 PWM files from three PWM databases. We also collected 511,893 peak fragments for further analysis. Together, user can use the database to query homology information of various TFs across those species, PWM information corresponding to TFs and enhancers from the high-throughput ChIP-seq data.

Heart development in all vertebrates from fish to humans follows the same genre: fusion of myocardium and endocardium in the ventral midline to form a simple tubular heart; onset of function, and looping to the right side; chamber specification and formation; and finally, development of specialized conduction tissue, coronary circulation, innervation, and mature valves. Thus CardioTF database aims to integrate TF information across species for the entire cardiovascular research community. Our initial effort was to create a central portal to host transcriptional information from different species [28].

We defined a core set of TFs evidenced by four independent sources to support their roles in cardiovascular development. This may provide a roadmap for systems biology to construct transcriptional network in heart development. Current approaches by the Sperling group or the Pu group only report three to four TFs on ChIP-seq data [29,30]. This is much less than our

knowledge of these core cardiovascular TFs, which have multiple sources supporting their role in cardiovascular development.

Are all 1200 mouse TFs expressed at heart developmental stages? In contrast, our preliminary analysis indicates that approximately 200 TFs have no evidence of their expression pattern, phenotype, expert recommendation, and PubMed abstracts. Whether these TF genes are expressed or play roles in heart disease requires further data analysis.

The database still lacks cell lineage-based expression profiling data, which quantify the expression level of various TFs and thus construct 4-D dynamic expression pattern *in vivo*. This information may be combined with cell lineage-based ChIP-seq data and create a super-resolution of enhancer tomography.

Conclusions

Modern translational medicine rests upon the progressive study of pathways and principles from model organisms such as yeast, fly, fish, and mouse to clinical studies in humans. Therefore, we recruited TFs from six model animals which are established models for research on cardiovascular development. These well-annotated homologs in different animals enable investigators to interrogate more fundamental problems in heart development.

We hope that in the near future, single-cell sequencing data may provide comprehensive gene expression information with detailed temporal-spatial resolution, thereby providing insight into the enhancer network that choreograph the gene specific expression patterns. CardioTF is the initial step into transcriptional network reconstruction.

Supplemental Information

Additional file 1: Four independent sources of cardiac TF lists and preferred journals for Weinstein-like papers. Format: XLSX. Sheet 1 (CardioJournalDistribution) includes the top 100

journals that favor Weinstein-like papers. Sheet 2 (NotHeartTFs) includes all TF names that is never appeared in the four lines of evidence. Sheet 3 (MGI_TFs) is the TF names from MGI database. Sheet 4 (PubMed_TFs) lists TF names from Weinstein-like PubMed abstracts. Sheet 5 (Cardio_GO_UK_TFs) includes TF names from UK Cardio-GO project. Sheet 5 (Cardio_lineage_TFs) includes TF names from RNA-seq or microarray analysis. Sheet 6 (core_TFs) defines the core 81 TFs regarding heart development.

Additional file 2: Format: PDF. Figure S1. Confusion matrix and precision-recall curve. By convention, the class label of the minority class is positive (Weinstein abstracts), while the class label of the majority class is negative (non-Weinstein-like abstracts). (A) The confusion matrix for a two-class problem. The first column shows the actual class label of the examples, and the first row presents their predicted class label. In the matrix, TP shows the true positive samples, FP shows the false positive samples, TN shows the true negative samples, and FN shows the false negative samples. (B) The precision-recall curve.

Additional file 3: Format: PDF. Figure S2. DAVID analysis of 81 core TFs. The results indicate that these TFs are truly associated with cardiac function by GO term enrichment analysis.

Additional file 4: Format: PDF. Figure S3. Heatmap of the 81 core cardiac transcriptional factors at the different stages of heart development.

Additional file 5: Format: PDF. Figure S4. RNA-seq expression pattern of the 81 core TFs across 13 adult tissues by the boxplot assay.

Additional file 6: Format: PDF. Figure S5. The expression profile across 13 adult tissues of TFs which are never expressed in the heart, as determined by the present four lines of evidence.

Additional file 7: Format: PDF. Figure S6. The work flow of parsing ChIP-seq data and dumping

262 into the MySQL database.

263 **Availability and requirements**

264 CardioTF database is freely available on the web at
265 http://www.cardiosignal.org/database/cardiotf.html.

266 **List of abbreviations**

267 TF: transcriptional factors; PWM: position weight marix; UML: unified modeling language;

268 **Competing interests**

269 The author declares that he has no competing interests.

270 **Authors' contributions**

271 YZ conceived of the project, designed and implemented the database, and wrote the manuscript.
272 The author read and approved the final manuscript.

273 **Acknowledgements**

274 This work was supported by National Natural Science Foundation of China [Grant number
275 31000644 to Y.Z.].The author is grateful to Prof. Rutai Hui, Prof. Weinian Shou, Dr. Tingting Li
276 and Dr. Jianxin Chen for their helpful comments. Special thanks to Prof. Paul Krieg (University
277 of Arizona), Matija Brozovic (ANISEED), Prof. Carl J. Schmidt (BirdBase) and Zichao Sang
278 (CardioSignal) for their comments or suggestions on data curation.

279 **References:**

280 1. van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ, et al.:
281 **Birth prevalence of congenital heart disease worldwide: a systematic review and meta-**
282 **analysis.** *J Am Coll Cardiol*, 2011, **58**(21):2241-7.

283 2. Celermajer DS, Chow CK, Marijon E, Anstey NM, Woo KS: **Cardiovascular disease in the**
284 **developing world: prevalences, patterns, and the potential of early disease detection**. *J Am*
285 *Coll Cardiol*, 2012, **60**(14):1207-16.
286 3. McCulley DJ, Black BL: **Transcription factor pathways and congenital heart disease.**
287 *Curr Top Dev Biol*, 2012, **100**:253-77.
288 4. Smith JG, Newton-Cheh C: **Genome-wide association studies of late-onset cardiovascular**
289 **disease**. *J Mol Cell Cardiol,* 2015, **83**:131-41.
290 5. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al.:
291 **ArrayExpress update--an archive of microarray and high-throughput sequencing-based**
292 **functional genomics experiments.** *Nucleic Acids Res*, 2011,**39**(Database issue):D1002-4.
293 6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al.: **NCBI GEO:**
294 **archive for functional genomics data sets--update.** *Nucleic Acids Res*, 2013, **41** (Database
295 issue):D991-5.
296 7. Chacon D, Beck D, Perera D, Wong JW, Pimanda JE: **BloodChIP: a database of**
297 **comparative genome-wide transcription factor binding profiles in human blood cells**.
298 *Nucleic Acids Res*, 2014, **42**(Database issue):D172-7.
299 8. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, et al.: **CistromeMap: a knowledgebase and**
300 **web server for ChIP-Seq and DNase-Seq studies in mouse and human.** *Bioinformatics*, 2012,
301 **28** (10):1411-12.
302 9. Bondue A, Blanpain C: **Mesp1: a key regulator of cardiovascular lineage commitment.**
303 *Circ Res*, 2010, **107**(12):1414-27.
304 10. Wingender E, Schoeps T, Dönitz J: **TFClass: an expandable hierarchical classification of**
305 **human transcription factors.** *Nucleic Acids Res*, 2013, **41** (Database issue):D165-70.

11. Hutson MR, Kirby ML: **Model systems for the study of heart development and disease.** *Semin Cell Dev Biol*, 2007, **18**(1):1-2.

12. Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S, Talmud PJ, et al.: **The representation of heart development in the gene ontology.** *Dev Biol*, 2011, **354**(1):9-17.

13. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE: Mouse Genome Database Group. **The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse.** *Nucleic Acids Res*, 2014, **42** (Database issue):D810-17.

14. Dietterich TG: **Approximate statistical tests for comparing supervised classification learning algorithms**. *Neural Computation*, 1998, **10**:1895-923.

15. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics*, 2008, **24**(16):126-32.

16. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al.: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc*, 2012, **7**(3):562-78.

17. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al.: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.** *Nucleic Acids Res*, 2014, **42** (Database issue):D142-7.

18. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML: **UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions.** *Nucleic Acids Res*, 2015, **43**(Database issue):D117-22.

19. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: **hPDI: a database of experimental human protein-DNA interactions.** *Bioinformatics*, 2010, **26**(2):287-9.

20. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS Comput Biol*, 2009, **5**(1): e1000262.

21. Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, et al.: **Xenbase, the Xenopus model organism database; new virtualized system, data types and genomes.** *Nucleic Acids Res*, 2015, **43** (Database issue):D756-63.

22. Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, et al.: **Gallus GBrowse: a unified genomic database for the chicken.** *Nucleic Acids Res*, 2008, **36** (Database issue):D719-23.

23. Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, et al.: **The ANISEED database: digital representation, formalization, and elucidation of a chordate developmental program.** *Genome Res*, 2010, **20**(10):1459-68.

24. Sonnhammer EL, Östlund G: **InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic.** *Nucleic Acids Res*, 2015, **43** (Database issue):D234-39.

25. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al.: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol*, 2013, **9**(11):e1003326.

26. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.** *Nat Protoc*,2012, **7**(9):1728-40.

27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C,Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell*,2010, **38**(4):576-89.

28. Xavier-Neto J, Davidson B, Simoes-Costa MS, Castro RA, Castillo HA, Sampaio AC, Azambuja AP. Evolutionary Origins of Hearts. In: Rosenthal, N, Harvey, R, editors. *Heart Development and Regeneration*. New York: Academic Press;2010. Vol I, p. 3-45.

29. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, Lange M,Tönjes M, Dunkel I, Sperling SR: **The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs.** *PLoS Genet,* 2011,**7**(2):e1001313.

30. He A, Kong SW, Ma Q, Pu WT: **Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart.** *Proc Natl Acad Sci U S A*, 2011,**108**(14):5632-37.

Figure 1(on next page)

Unified modeling language diagram for the Cardio-TF database design.

The six graphical notation boxes represent of six major classes, namely CardioTFmatrix, CardioTFCenter, CardioTree, CardioGeneNames, CardioEnhancer, and ChIPExpAssay. These classes are analogous to entity/relationship sets. Each class has only two sections, one for the class name and one for the attributes. The attribute of each class is associated with the type used in MariaDB. The “#” in front of an attribute indicates that it’s visibility is “protected”, thereby making it a primary key. These classes faithfully represent the real world.

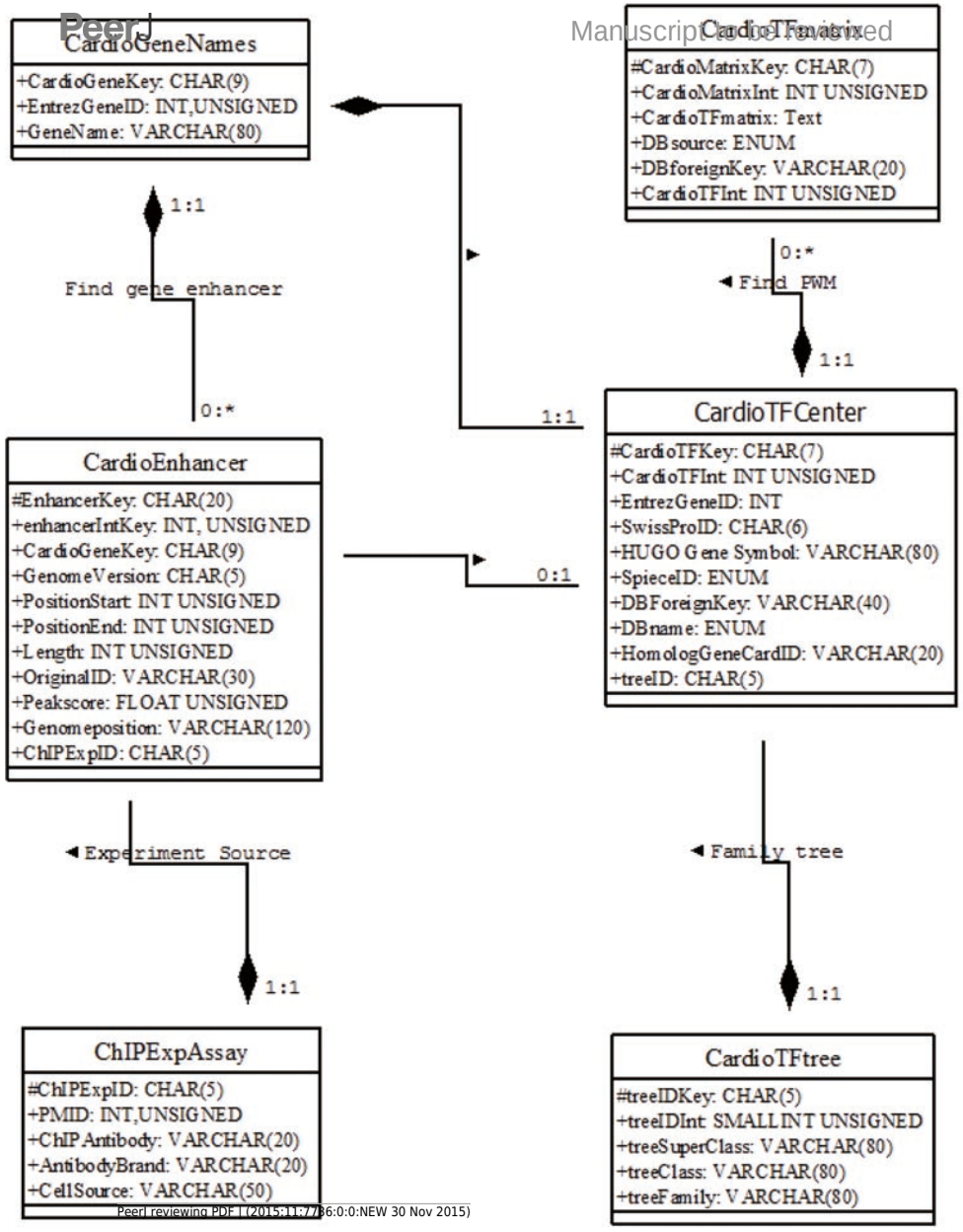


Figure 2(on next page)

The search engine and the web interface of the database.

(A) The search engine was implemented to perform three functions: querying TFs, their PWMs and gene enhancers (B) Web graphical output of Gata4 enhancers in mouse. Black lines indicate the enhancer region found by the ChIP-seq scanning program. (C) Query results for the GATA4 TF across species. TFs are listed and indexed according to their database identifiers.

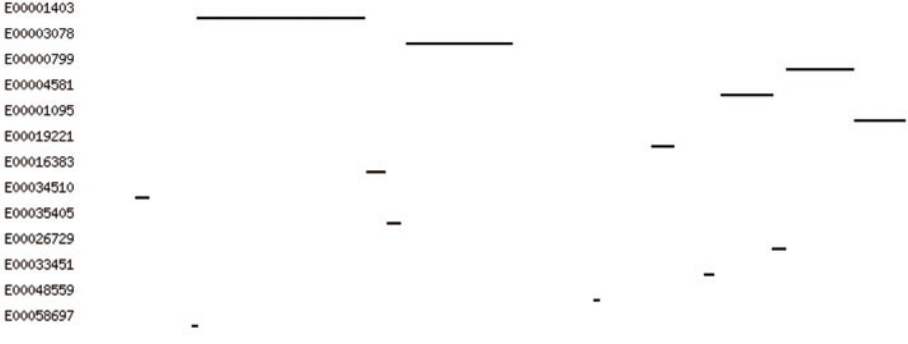
A

CardioTF Module: Cardiac Transcriptional Factor (TF)

OPTIONS: Cardiac Gene Name

QUERY: Gata4 submit

B



C

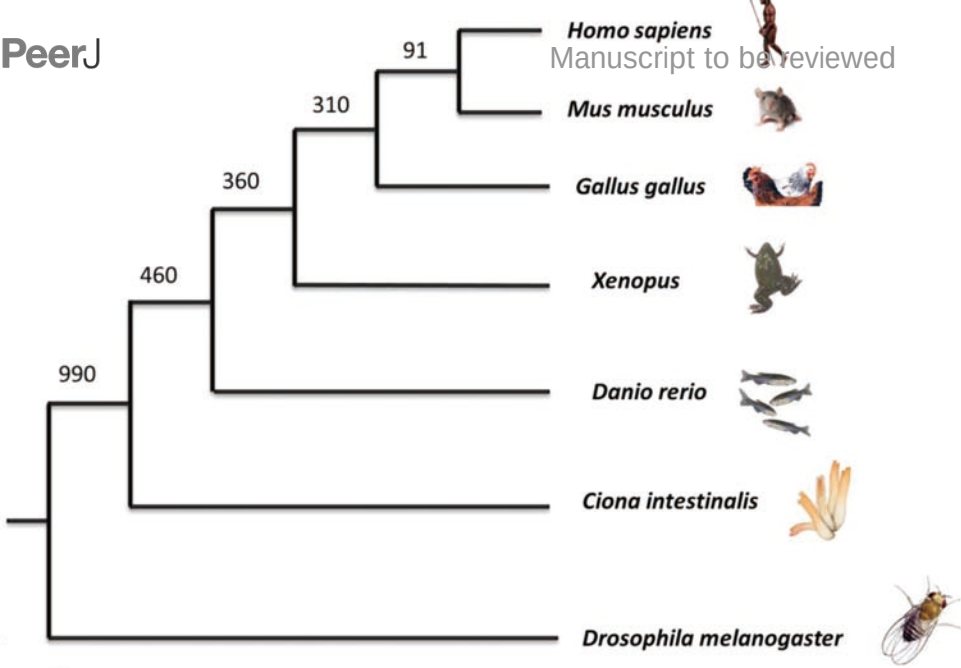
CardioTF Identifier	Gene Symbol	Species Name	EntrezGeneID	SwissProID	DB_link
TF02887	gata4	*Danio rerio*	30483	B8JKU1	ZFIN
TF01887	GATA4	*Homo sapiens*	2626	P43694	NCBI
TF04086	gata4	*Xenopus (Silurana) tropicalis*	549703	Unknown	XenBase
TF04714	GATA4	*Gallus gallus*	396392	Unknown	BirdBase
TF05400	PNR	*Drosophila melanogaster*	44849	P52168	FlyBase
TF01888	Gata4	*Mus musculus*	14463	Q08369	MGI

Figure 3(on next page)

TF distribution across species in the database.

(A) Phylogenetic tree showing the main animal models commonly used in heart development research and their evolutionary relationship. The divergence times in millions of years ago (Mya) are shown on the basis of multigene and multiprotein studies. Branch lengths are not proportional to time (B) Distribution TFs across six species. All TFs have homologs in their human counterparts. The unit of Y-axis is TF number.

A



B

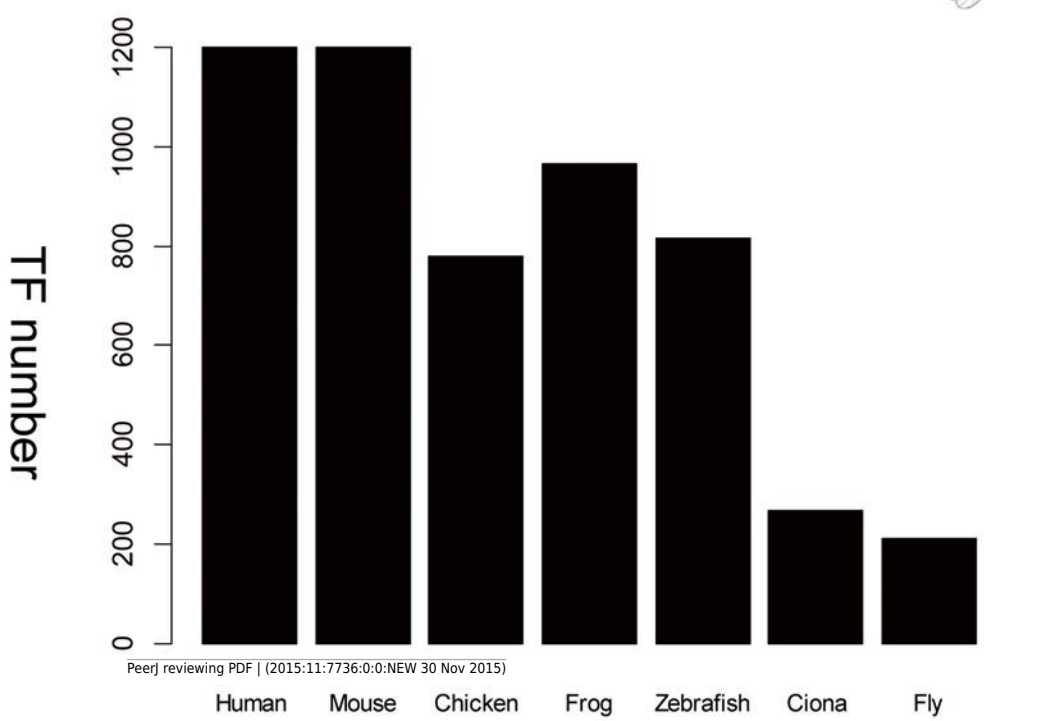
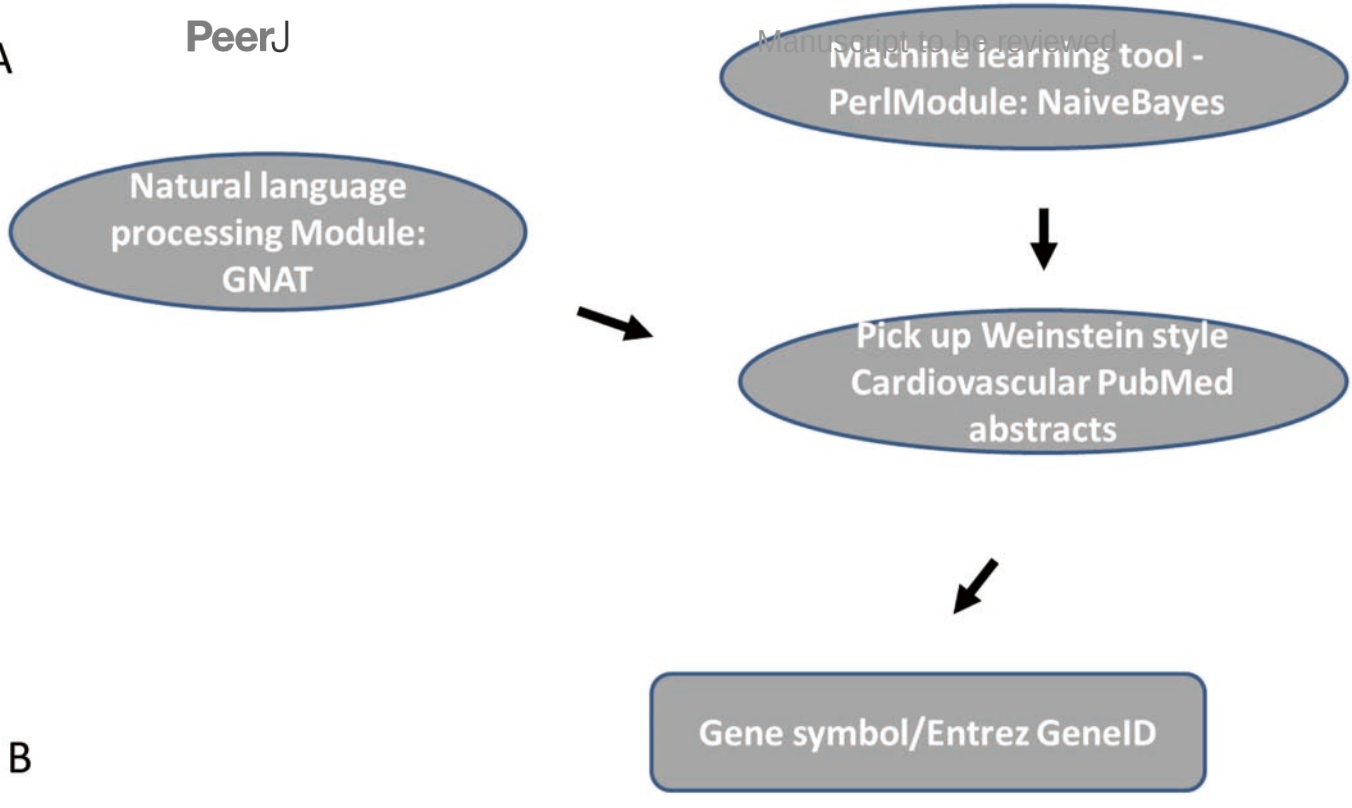


Figure 4(on next page)

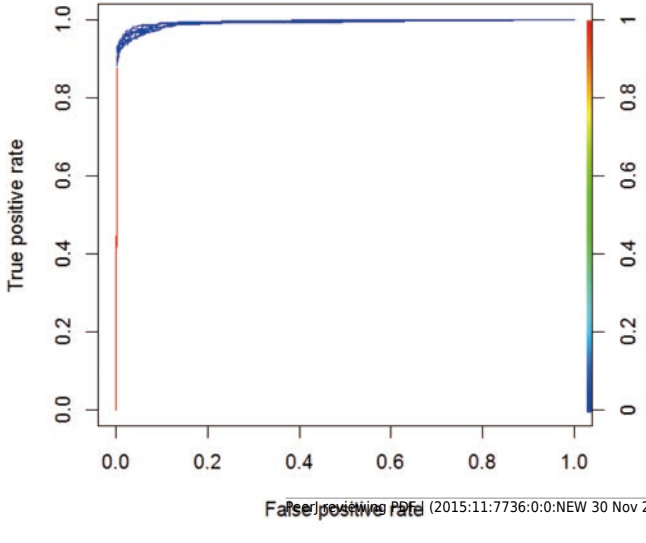
Machine learning protocol to select TFs described in Weinstein-like papers.

(A) The pipeline to select TF gene symbols from Weinstein PubMed abstracts. First, Naïve-Bayes module was used to select Weinstein-like papers from PubMed abstracts. Second, GNAT, a software that recognizes gene symbols, was used to identify all TF names from these Weinstein-like papers. (B) ROC curve and prediction performance judged by sensitivity, precision and F1 score.

A



B



Sensitivity	0.9316
Precision	0.9516
F1 Score	0.9415