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

Yisong Zhen Corresp. 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, China

Corresponding Author: Yisong Zhen
Email address: zhenyisong@fuwaihospital.org

Background. Information on cardiovascular gene transcription is fragmented and far behind the present requirements of the systems biology field. To create a comprehensive source of data for cardiovascular gene regulation and to facilitate a deeper understanding of genomic data, the CardioTF database was constructed. The purpose of this database is to collate information on cardiovascular 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 natural 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.** Known cardiovascular TFs from humans and human homologs from fly, *Ciona*, zebrafish, frog, chicken, and mouse were identified and deposited in the database. Position weight matrices from Jaspar, hPDI, and UniPROBE databases were deposited in the database and can be retrieved using their corresponding TF names. Gene enhancer regions from various sources of ChIP-seq data were deposited into the database and were able to be visualized by graphical output. Besides biocuration, mouse homologs of the 81 core cardiac TFs were selected using a Naïve-Bayes approach and then by intersecting four independent data sources: RNA profiling, expert annotation, PubMed abstracts and phenotype. **Discussion.** The CardioTF database can be used as a portal to construct transcriptional network of 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 requirements of the systems biology field. To create a comprehensive source of data for cardiovascular gene regulation and to facilitate a deeper understanding of genomic data, the CardioTF database was constructed. The purpose of this database is to collate information on cardiovascular 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 natural 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. Known cardiovascular TFs from humans and human homologs from fly, *Ciona*, zebrafish, frog, chicken, and mouse were identified and deposited in the database. Position weight matrices from Jaspar, hPDI, and UniPROBE databases were deposited in the database and can be retrieved using their corresponding TF names. Gene enhancer regions from various sources of ChIP-seq data were deposited into the database and were able to be visualized by graphical output. Besides biocuration, mouse homologs of the 81 core cardiac TFs were selected using a Naïve-Bayes approach and then by intersecting four independent data sources: RNA profiling, expert annotation, PubMed abstracts and phenotype.

Discussion. The CardioTF database can be used as a portal to construct transcriptional network of 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]. Insights into the cause of congenital heart diseases (CHDs) have led to the identification of mutations in essential cardiac transcription factors (TFs) [3]. At the opposite end of the temporal spectrum, some cases of adult cardiac disease have been traced to variation in gene regulatory sequences [4]. Thus, knowledge of TFs, their downstream targets, and the regulatory genomic sequences involved in the heart development will enhance our understanding of heart disease.

Although the vast amounts of data generated by high-throughput technologies are archived in databases such as ArrayExpress or GEO of NCBI [5,6], they do not contain cohesive knowledge and lack expert annotation. In addition, the field of 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 made to create a database which collects cardiac transcriptional information across species, thereby limiting the benefits from an evolutionary perspective to study heart development.

At present, two branching efforts have been made to archive and analyze the data. One is to construct small scale databases, like BloodChIP or CistromeMap, which are dedicated to collecting specific types of data [7, 8]. The other approach is to establish a number of consortia, like ENCODE, modENCODE, and Epigenomics Roadmap, which are created to generate huge amounts of raw data and archive them [9-11]. In addition to these projects, analysis and visualization software are valuable resources that lead to deeper understanding of the data, and facilitate the generation of novel hypotheses. Central databases, like Ensembl and UCSC also have search functions which allow browsing of the results generated by the consortia mentioned

above [12, 13]. However, there are currently few databases committed exclusively to cardiovascular development [14]. This prompted us to combine information about TFs, position weight matrices (PWMs), and ChIP-seq results and create a one-stop site for information on cardiovascular development, thus facilitating systems biology studies in transcriptional network regulation [15].

CardioTF was therefore constructed to capture all transcriptional information relating to cardiovascular development. As a biocuration project, it documents TFs, PWM files and enhancers across species, including fly, *Ciona*, fish, frog, chicken, mouse and human. It also implements a search engine to query this information on the fly. In addition to the data-mining effort, core cardiac TFs are identified using Naïve-Bayes approach, which can be used as a roadmap alongside with further annotation for enhancers to generate gene regulatory network of heart development

Materials & Methods

The project's code and data for reproducible research

All the Perl scripts and R codes were uploaded to GitHub (https://github.com/zhenyisong/). The raw data including Weinstein meeting abstracts (positive_test_data plus positive_training_data), negative dataset (negative_test_data plus negative_training_data) (in zip format), intermediary files, which contain cross-validation results as well as other public data were uploaded to the CardioTF database server (http://www.cardiosignal.org/download/download.html). These raw data and source codes can be used to verify the findings.

Comprehensive collection and annotation of cardiac TFs

Cardiac TFs were previously defined as regulators of cardiac gene expression, which can impact the process of heart development, particularly the initiation and maintenance of the myocardium [16]. In the CardioSignal database, efforts were also made to collate the cardiac specific enhancers which drive gene expression in cardiomyocytes. At that time, cardiac specific transcriptional factors were defined as genes that regulated expression of genes in the myocardium. During development, the heart consists of three layers: the myocardium, epicardium and endocardium [17,18]. Additionally, at least four heart-specific cell lineages have been characterized, including cardiomyocytes, endothelial cells, epicardial cells, and fibroblasts, the latter is derived mainly from epicardial cells through the epithelial to mesenchymal transition (EMT) [19,20]. By definition, cardiac TFs themselves should be involved in the steps of specification, determination, patterning, and differentiation that will result in a heart fate. In our CardioTF database we collated, cardiac transcription factors which are expressed in all layers of heart. The goal of the CardioSignal database was to use a machine learning approach to find cardiac enhancers at the genome scale. In contrast, the CardioTF database is constructed to study systems biology of transcription regulation. The cross-talk between different layers will also be explored using this platform. In the initial screen, we identified human TFs from previously published annotations [21], hence it was named Wingender's annotation set. This dataset is comprehensive in annotating human TFs. We used these human TFs, excluding human-specific TFs, as a reference to search for their homologs in other species, including fly, *Ciona*, zebrafish, chicken, frog and mouse [22]. Human-specific TFs are defined as genes which have no homologs in the mouse genome. The NCBI HomoloGene database [23] was used as a reference to assess homologs between human and mouse/zebrafish. Human homologs from other were retrieved from their central databases,

namely, FlyBase (Fly), BirdBase (Chicken), Aniseed (*Ciona*) and Xenbase (Frog) [24-27]. We also documented the expression status for mouse TFs from four independent sources which included annotation from the Cardiovascular Gene Ontology Annotation Initiative [28], Mouse Genome Database (MGI) genes with cardiovascular phenotypes [29], 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. The Weinstein meeting abstracts were extracted from the meeting abstract book from 2010 to 2013 by hand. As the positive group, this data set included 954 abstracts. 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 the PMC Open Access Subset at NCBI. To choose the negative control journals, the following criteria were set. First, well-known cardiovascular journals were excluded, such as 'Circulation' and 'Circulation Research'. Second, journals without key words 'heart' or 'cardiac' or 'cardiovascular' in their title were selected. Third, journals which are dedicated to the study of other organs or diseases, for example, 'Neuron' or 'Cancer' were selected. Fourth, other journal which are unlikely to publish articles about cardiovascular development and related topics, such as journals about plants or viruses, were selected. Journals in the negative group contained research from across kingdoms and topics obviously in other fields, such as 'Sleep_Disord' or 'Toxicology'. All journal names in the negative group were saved in a file and uploaded onto the cardioTF server (negative_set.journal.txt).The negative group includes 57080 abstracts. We split the data (positive and negative groups) into a training (80%) and test

(20%) set. The Naïve-Bayes module from The Comprehensive Perl Archive Network (CPAN) was used with a local Perl script to classify Weinstein-like abstracts. We used the training set and adopted the 5×2 cross-validation proposed by Dietterich [30] to train and validate the data. The parameter (the cutoff to decide whether an abstract is a true Weinstein-like abstract) was selected based on average predictive performance which resulted in a classification accuracy (ACC) of 0.99. A wrapper function was implemented to parse the abstracts and calculate the word frequencies. This function called two Perl modules (Lingua::EN::Splitter and Lingua::EN::StopWords) to extract words and perform text analysis. The word frequency alone was forwarded to the algorithm. The withheld test set using the optimized parameter was then used to assess the algorithm's final performance. All publication abstracts from 2008 to 2013 were downloaded to the local environment and analyzed by the algorithm. We targeted journals which had at least 6 publications classified as Weinstein-like abstracts in the six-year period (annual publication rate is >= 1). Then all abstracts from the targeted journal were downloaded. This process was repeated for all journals that met the criteria. The selected abstracts were then processed by GNAT [31] using its default script (test100.sh) to recognize the mouse gene name. The PMID was recorded when the gene name matched the name in the curated mouse TF set.

RNA expression profiling data procession

Affymetrix data (GSE1479) were processed by R using the MAS5 algorithm which provides a present call for each gene (see the script ExtractAffy.R at Github) Gene expression status was defined as "on" if the gene was expressed in any microarray at selected developmental stages and had a present call. RNA-seq data were re-analyzed using the recommended protocol (all raw data identifiers can be retrieved from supplemental Table S2) [32]. Briefly, pre-processing software (FastQC) was used to estimate the read length of raw data. If read length is above 50bp,

Bowtie2 was used. Otherwise, Bowtie was used. mm10 and hg19 are the genome builds used by UCSC. Index and annotation files for Bowtie2/Bowtie were downloaded from Illumina iGenomes project. Genome sequences from UCSC are repeat-masked with lower-case characters. Any gene with an FPKM value greater than 1 was defined as expressed and this threshold was empirically set although justifiable

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 the format to the TRANSFAC style, which was used by our in-house CardioSignalScan program [16]. PWM files were collected from Jaspar, UniPROBE and hPDI databases [33-35]. Users can retrieve their annotations by directing them to the respective database. All the PWMs from those three sources for each TF were deposited in the database. We currently do not use the program implemented by the Zhang lab [36] to check the similarity of PWMs and reduce the redundancy in the collection of PWM files.

Orthologs of TFs from model systems

NCBI has its own gene orthologs that were identified using unpublished algorithm [37]. TFs from mouse, human and zebrafish are annotated by NCBI Homologene [23]. Frog, chicken and *Ciona* TF homolog annotations were downloaded from their central databases including Xenbase, BirdBase and ANISEED. Fly TFs, which have counterparts in the human proteome, were annotated by the Inparanoid system [38]. 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 [21].

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

Raw ChIP-seq data were recruited based upon two criteria: first, whether the source of tissue or cells is from heart or heart progenitor derived cells; second, the DNA-binding protein for the ChIP assay should be pan-enhancer markers or heart lineage specific TFs. In the latter case, the core heart TFs were proposed in our screening procedure. Enhancer regions were defined by ChIP-seq signals. We assume that pan-enhancer markers, like H3K4me1 or H3K27ac [39], or lineage specific markers, like GATA4 or MEF2C [40] will delineate true enhancer regions, although these collections will produce some false positive records. Peak calling was performed using the recommended pipeline [41]. In brief, sequencing reads were aligned to the mm10/hg19 reference genome using Bowtie/Bowtie2. Mm10/hg19 represents the genome build assigned by UCSC. Index files for mm10/hg19 were downloaded from the iGenome project. MACS1.4.2 was used to process all the ChIP-seq data. The default cutoff for the p-value was 1e-05. This default value was used in all ChIP-seq analysis. This protocol was adapted from published literature [42].

Bowtie call

bowtie -m 2 -S -q -p 8

Peak calling was performed using the MACS peak calling algorithm.

MACS call linux command

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

A Torque job script was written to submit the job to the supercomputer. After that, the format transformation was performed:

samtools view -bS -o tbx20_positive.bam positive_tbx20.sam

When possible, the control files were merged:

samtools merge out.bam in_1.bam in_2.bam in_3.bam

After MASC analysis was completed, the annotatePeaks.pl was run in HOMER [43] to parse

the bed file from the MACS output. Then the parsed results were dumped into the MySQL table. Public identifiers for the raw data can be retrieved from supplemental Table S2 and ChIP-seq experimental information has been recorded in the MySQL table 'ChIPExpAssay'.

Recognition of transcription factor binding sites (TFBSs) in enhancer

CardioSignalScan was previously implemented to identify transcription factor binding sites [16]. However, this local program (see cardiophylo.pl in GitHub) is brute-force solution which consumes computational time with linear complexity (O(mn)). In the Big O notation, m is the column length of the matrix and n is the length of the input DNA string. Therefore, it is unrealistic to scan sequences longer than 3000bp with this local program. This prompted us to choose MOODS [44] instead, which reduces the computational time proportionally to position weight matrix length (O(m)). A wrapper module was written to calculate the threshold that gauges the match. The cutoff was empirically defined to be 0.75 (range from 0 to 1 and 1 is most conserved score).

threshold = min_log_score + (max_log_score – min_log_score) * cutoff

This step avoids using *P*-values to assess the significance of TFBS.

Gene ontology analysis

DAVID analysis (version 6.7) was performed using the 81 TFs as the input gene list, official gene symbols as the identifiers and the entire mouse gene set as the background. The functional annotation clusters generated by DAVID were identified by TFs (Figure S2). The classification stringency was set to the default (medium).

Results

The database schema

Our database uses the MariaDB, a drop-in replacement for MySQL, as the database management system (DBMS). To address how information will be stored and how the elements will be related to one another, we used the unified modeling language (UML) to describe the high-level database model [45]. UML was originally developed as a graphical notation for describing software designs in an object-oriented style. It has been extended, and modified and is now a popular notation for describing database designs. Here we used UML instead of an entity/relationship diagram to design the relational database schema following modeling principles, such as faithfulness, avoiding redundancy, and simplicity counts [45] (Fig. 1). Where possible, we used a composition distinguished by a line between two classes that ends in a solid diamond at one end. The diamond implies that the label at the end must be 1:1. For example, there is a composition from CardioTFmatrix to CardioTFCenter, which means that every matrix annotation row (PWM related information) belongs to exactly one row in CardioTFCenter (one type of TF may have more PWM records in a CardioTFmatrix table). A 1:1 label at the CardioTFCenter end is implied by a solid black diamond.

Web Interface and search engine

CardioTF is a Perl website implemented using only Perl language to dynamically display the graphical output while querying the database in the backend (Fig. 2). To aid cardiovascular biologists, a search engine was created to allow users to: (1) identify homology information for the queried TF across six species and link to the corresponding central databases outside CardioTF; (2) identify PWM file union of three public databases regarding the queried TF; and (3) identify the enhancer regions revealed by ChIP-seq data of the queried gene. Thus, the database is able to perform the key functions required to construct a transcriptional network of heart development.

Cardiovascular TFs in the database

Wingender's annotation set [21] was used as a benchmark to recruit TFs across species. The frozen version of this dataset contains 1564 human TFs. Among them, only 1513 TFs have corresponding Entrez gene records. Human-specific TFs, defined as those with no orthologs in the mouse genome, were discarded because no model system could be used to verify their function *in vivo*. This step excluded a further 313 human TFs which have no counterpart in mouse from the homolog annotation. Therefore, 1200 mouse TFs were collected. Other established animal models for cardiovascular development include fly, *Ciona*, zebrafish, frog, and chicken. TFs from these species were collected if they were homologs to the above mouse TFs. The distribution of TFs from different species is shown in Fig. 3. The expression status of mouse TFs was verified by four independent resources, namely RNA-seq data re-analysis [39], phenotype annotations from the MGI database [29], expert recommendation from the UK Cardiovascular Gene Annotation Initiative [28], and PubMed relevance from classification of Weinstein-like abstracts (see the subsequent section and Table S1).

Weinstein TFs from PubMed analysis

We identified the journals that favored Weinstein-style papers, which were likely contain information on genes involved in cardiovascular development. As expected, after using a Naïve-Bayes method, the journals we identified were among the 30 journals most relevant to developmental biology. Two of the journals (*Circ. Res*. and *J Mol Cell Cardiol.*) obviously publish research specifically in the area of heart system. (Table S1: CardioJournalDistribution). If normalized and ranked by publication rate, the above conclusion still holds true although two different heart journals (*Eur J Echocardiogr, Heart Rhythm*) are in the top 30 list (Table S1: CardioJournalDistribution_norm) in this case. We then used GNAT, a tool that recognizes gene

names in the literature, to recover all TFs mentioned in Weinstein-style abstracts because we assumed that these TFs are studied by researchers in the 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 contains 904 records, and these PWM files can be recognized by our local CardioSignalScan program to search for the motifs in genomic 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 a core set of cardiac TFs, we intersected four independent sources of cardiovascular TF collections. Inclusion of the resulting 81 TFs is supported by their expression status, phenotype annotation, expert recommendation and PubMed relevance (Table S1). We also performed DAVID functional analysis, and found that these TFs are particularly enriched in cardiac muscle differentiation (Fig. S2). 10-20% of these TFs which are enriched in the Annotation Cluster 7 including Gata4 ,Gata6, Smad7, Nkx2-5, Tbx2, Tbx5, Foxc1, Foxp1, Prox1, Rara ,Rarb, Rxra, Rxrb, Zfpm2. These 14 TFs, are annotated in the DAVID as being involved in cardiac muscle formation. As we know, the heart system includes the endocardium which is a specialized layer derived from endothelial cells. Cluster 8 from our DAVID analysis includes genes expressed in endothelial cells such as Smad5, Smad7, Meis1, Nkx2-5, Tbx20,

Epas1, Foxc1, Foxo1, Hey1, Hand2, Hif1a, Mef2c, Prrx1, Prox1, Srf, Nr2f2, Tcf21, Vezf1, Zfpm2. Is this set of genes the minimum requirement for cardiac development? Indeed, these four sources of supporting evidence indicate that these TFs genes play a key role in heart development. We wanted to determine if these TFs display specific expression patterns in heart development. A heatmap was generated using seven RNA-seq data sets, including samples from embryonic cells, mesoderm cells, cardiac progenitors, nascent cardiomyocytes and adult heart tissue. This heatmap did not reveal any specific patterns (Fig. S3). In adult tissues, these TFs did not exhibit enriched expression in the adult heart. In the 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 fragments *in vivo* to analyze temporal-spatial patterns. Therefore, the ChIP-seq method provides a high-throughput 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 stages of heart development. Searching for a single enhancer also provides a user interface to scan the TFBSs. The binding matrices provided in the list come from the core cardiac TFs. The flat text output file contains the matrix key for the TFs in the database and sorts these hits in the increasing order.

Discussion

We identified 5442 TFs from six species, and integrated 904 PWM files from three PWM databases. We also collected 511,893 peak fragments for further analysis. The on-the-fly search tool was implemented to match the core cardiac TFBSs in the specified enhancer sequence. Our database provides a framework where users can query homology information for various TFs across species and PWM information corresponding to TFs and enhancers from high-throughput ChIP-seq data. The curation for cardiac enhancers and TFs facilitate future efforts to construct transcription network.

The database now contains the six species which are model organisms for studying heart development (Figure 3A). *Ciona* is an ideal model used to study the early specification step in the *Mesp*-lineage. Zebrafish is well suited to perform imaging analysis and for performing quantitative study. The chicken is well suited for lineage tracing *ex vivo*. The mouse provides a well-studied mammalian model and is most similar to humans. The fly is a valuable model for testing new concepts and is easy to study at the genomic level. New models may be added in the future if they have unique advantage over the other models.

Previously, we constructed the CardioSignal database which collates cardiac factors driving genes expressed in a tissue-specific or quantitative manner. Most enhancers archived in the database are expressed specifically in the myocardium. CardioTF is a complementary database that accumulates cardiac TFs expressed in the epicardium, endocardium and myocardium. This information including PWMs can be used not only to find the features of enhancers in a machine learning approach (such the left-right patterning of the heart) but also to reconstruct regulatory network in systems biology.

We defined a core set of TFs using four independent sources, namely, RNA profiling, expert curation, PubMed abstract parsing and phenotype annotation to support their roles in

cardiovascular development. In RNA-seq analysis, after pre-processing and post-processing of the adult heart data, the count table contained 48995 genes. After filtering genes with FPKM > 1, there were 20863 genes remaining. Roughly half of these annotated genes were abandoned because their gene expression level was too low. To analyze PubMed abstracts, a Naïve-Bayes approach was used to classify Weinstein-style abstracts and then pick out the TFs embedded in those abstracts. Most journals containing the key words "heart" or "cardiac" or "cardiovascular" focus on heart pathology or physiology instead of development biology. For example, journals such as "Cardiovascular research", "Circulation", and "Circulation research", accept papers related to the adult cardiovascular system. Most papers published in these journals take a more translational approach which is oriented to bench-side work. Our analysis identified two heart related journals (which is obvious from their name) suggesting that the algorithm was successful in finding the wording pattern present in Weinstein abstracts. The results indicate that most developmental biology manuscripts relating to cardiovascular system are sent to specialized journals that focus on development. Top-tier cardiovascular journals are more likely to publish papers describing the adult cardiovascular system. Our text analysis, whether normalized by publication number or not, had a tendency of identifying journals favored by heart developmental biologists and journals that specialized in developmental biology (Table S1).

Traditional definitions of heart specific TFs can often be ambiguous and should not include TFs that only regulate cardiac muscle. In the present definition, heart specific TFs must be detected to be expressed in heart tissue which includes the endocardium and epicardium layers. Both DAVID Annotation Cluster 7 and Cluster 8 contain TFs involved in heart muscle or vascular formation. The DAVID analysis result did not reveal any other clusters with genes involved in kidney, liver or the formation of other organs. Even the expression of *GATA4* and *MEF2C*,

which are *de facto* cardiac TFs, is not restricted to cardiomyocytes. These TFs are expressed in cardiac progenitors at a certain stage of development. The present approach is empirical and proposes a method which uses four independent data sources to identify true cardiac TFs based on their expression, phenotype, community opinion and PubMed abstracts. These 81 core TFs could be used further to support simulation study to infer the significance in future.

In general, the set of core cardiac TFs identified by these sources provide a roadmap for systems biology to construct a transcriptional network of heart development. Current approaches by the Sperling group or the Pu group only report three to four TFs based on ChIP-seq data [40,46]. Similar approaches by other genome biologists who tried to find cardiac enhancers on a genome scale have been reviewed elsewhere [47, 48]. However, the information generated from these studies is well below our knowledge of these core cardiovascular TFs, which have multiple sources supporting their role in cardiovascular development.

We archived 1200 mouse TFs and wanted to determine at what stage of heart development they were expressed. 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 analysis.

The database still lacks cell lineage-based expression profiling data, which will quantify the expression level of various TFs and thus construct a 4-D dynamic expression pattern *in vivo*. This information could be combined with cell lineage-based ChIP-seq data to 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 mice to clinical studies in humans. Therefore, we

recruited TFs from six model organisms which are established models for research on cardiovascular development. The identification and collation of these well-annotated homologs from different organisms will enable investigators to better address the complexities of transcriptional network on heart development [48].

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 transcription networks that contribute to heart development or heart diseases.. CardioTF database try to collate these *cis* and *trans* information and take the initial steps in the construction of a comprehensive transcriptional network.

Supplemental Information

Additional file 0: 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 have never appeared in the four sources of supporting 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 known to be involved in heart development. Sheet 7 (CardioJournalDistribution_norm) includes journals ranked by publication rate (normalized version)

Additional file 1: ChIP-seq and RNA-seq raw data collection. Format: XLSX. Sheet 1 (ChIP-seq data) includes all raw ChIP-seq data information, which are also deposited in MySQL table 'ChIPExpAssay'. The MACS processed data are dumped into MySQL table 'CardioEnhancers'.

Sheet 2 (RNA-seq data) includes all raw RAN-seq information. These data were used to generate the Supplemental Figure S3.

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 patterns for the 81 core TFs across 13 adult tissues presented by the boxplot.

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 our four sources supporting evidence.

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

Availability and requirements

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

List of abbreviations

TF: transcription factors; PWM: position weight matrix; UML: unified modeling language;

Acknowledgements

The author is grateful to Prof. Rutai Hui, Prof. Weinian Shou, Dr. Tingting Li and Dr. Jianxin Chen for their helpful comments. Special thanks to Prof. Paul Krieg (University of Arizona), Matija Brozovic (ANISEED), Prof. Carl J. Schmidt (BirdBase) and Zichao Sang (CardioSignal) for their comments or suggestions on data curation. I would also like to thank Dr. Rosannah C. Cameron at the Albert Einstein College of Medicine for her assistance editing the manuscript.

References:

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

2. Celermajer DS, Chow CK, Marijon E, Anstey NM, Woo KS: **Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection**. *J Am Coll Cardiol*, 2012, **60**(14):1207-16.

3. McCulley DJ, Black BL: **Transcription factor pathways and congenital heart disease.** *Curr Top Dev Biol*, 2012, **100**:253-77.

4. Smith JG, Newton-Cheh C: **Genome-wide association studies of late-onset cardiovascular disease**. *J Mol Cell Cardiol,* 2015, **83**:131-41.

5. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A..:

446 **ArrayExpress update--an archive of microarray and high-throughput sequencing-based**
447 **functional genomics experiments.** *Nucleic Acids Res*, 2011,**39**(Database issue):D1002-4.
448 6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA,
449 Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N,Robertson CL, Serova N,
450 Davis S, Soboleva A..: **NCBI GEO: archive for functional genomics data sets--update.**
451 *Nucleic Acids Res*, 2013, **41** (Database issue):D991-5.
452 7. Chacon D, Beck D, Perera D, Wong JW, Pimanda JE: **BloodChIP: a database of**
453 **comparative genome-wide transcription factor binding profiles in human blood cells**.
454 *Nucleic Acids Res*, 2014, **42**(Database issue):D172-7.
455 8. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, Wang S, Chen J, Shen L, Duan X, Hu S, Li
456 W, Long H, Zhang Y, Liu XS..: **CistromeMap: a knowledgebase and web server for ChIP-**
457 **Seq and DNase-Seq studies in mouse and human.** *Bioinformatics*, 2012, **28** (10):1411-12.
458 9. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL,
459 Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee
460 G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A,
461 Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I,
462 van Baren J, Brent M, Haussler D, Kellis M,Valencia A, Reymond A, Gerstein M, Guigó R,
463 Hubbard TJ. **GENCODE: the reference human genome annotation for The ENCODE**
464 Project. *Genome Res*., 2012, 22(9):1760-74.
465 10. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai
466 EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston
467 RH; modENCODE Consortium. **Unlocking the secrets of the genome.** *Nature.*, 2009,
468 459(7249):927-30

11. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. **Epigenomics: Roadmap for regulation**. *Nature.* 2015, 518(7539):314-6.

12. Mangan ME, Williams JM, Kuhn RM, Lathe WC 3rd. The UCSC Genome Browser: **What Every Molecular Biologist Should Know**. *Curr Protoc Mol Biol.*, 2014, 1;107:19.9.1-19.9.36.

13. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. **Ensembl 2013**. *Nucleic Acids Res.,* 2013, 41(Database issue):D48-55.

14. Djordjevic D, Yang A, Zadoorian A, Rungrugeecharoen K, Ho JW. **How difficult is inference of mammalian causal gene regulatory networks?** *PLoS One.*, 2014, 9(11):e111661.

15. Blais A, Dynlacht BD. **Constructing transcriptional regulatory networks.** *Genes Dev. ,* 2005, 19(13):1499-1511.

16. Zhen Y, Wang Y, Zhang W, Zhou C, Hui R. **CardioSignal: a database of transcriptional regulation in cardiac development and hypertrophy.** *Int J Cardiol.*, 2007, 116(3):338-347.

17. Moorman AF, Christoffels VM. **Cardiac chamber formation: development, genes, and evolution.** *Physiol Rev*., 2003, 83(4):1223-1267.

18. Fishman MC, Chien KR. **Fashioning the vertebrate heart: earliest embryonic decisions.** *Development.* , 1997, 124(11):2099-2117.

491 19. Evans SM, Yelon D, Conlon FL, Kirby ML. **Myocardial lineage development.** *Circ Res,*.
492 2010,107(12):1428-1444.
493 20. Moore-Morris T, Cattaneo P, Puceat M, Evans SM. **Origins of cardiac fibroblasts.** *J Mol*
494 *Cell Cardiol.,* 2016, 91:1-5.
495 21. Wingender E, Schoeps T, Dönitz J: **TFClass: an expandable hierarchical classification of**
496 **human transcription factors.** *Nucleic Acids Res*, 2013, **41** (Database issue):D165-70.
497 22. Hutson MR, Kirby ML: **Model systems for the study of heart development and disease.**
498 *Semin Cell Dev Biol*, 2007, **18**(1):1-2.
499 23. NCBI Resource Coordinators. **Database resources of the National Center for**
500 **Biotechnology Information**. *Nucleic Acids Res.* 2015, 43(Database issue):D6-17.
501 24. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ; FlyBase
502 consortium. **FlyBase: establishing a Gene Group resource for Drosophila melanogaster**.
503 *Nucleic Acids Res.* 2016, 44(D1):D786-92.
504 25. Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, Karimi K,
505 Zorn AM, Vize PD.: **Xenbase, the Xenopus model organism database; new virtualized**
506 **system, data types and genomes.** *Nucleic Acids Res*, 2015, **43** (Database issue):D756-63.
507 26. Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, McGrath
508 S, Mardis E, Stein LD.: **Gallus GBrowse: a unified genomic database for the chicken.**
509 *Nucleic Acids Res*, 2008, **36** (Database issue):D719-23.
510 27. Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V,Caillol D,
511 Schiappa R, Laporte B, Rios A, Luxardi G, Kusakabe T, Joly JS, Darras S, Christiaen L,
512 Contensin M, Auger H, Lamy C, Hudson C, Rothb?cher U, Gilchrist MJ, Makabe KW, Hotta K,
513 Fujiwara S, Satoh N, Satou Y, Lemaire P..: **The ANISEED database: digital representation,**

formalization, and elucidation of a chordate developmental program. *Genome Res*, 2010, **20**(10):1459-68.

28. Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S, Talmud PJ, Breckenridge R, Bhattarcharya S, Riley P, Scambler P, Lovering RC.: **The representation of heart development in the gene ontology.** *Dev Biol*, 2011, **354**(1):9-17.

29. 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.

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

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

32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H,Salzberg SL, Rinn JL, Pachter L.: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc*, 2012, **7**(3):562-78.

33. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B,Sandelin A, Wasserman WW..: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.** *Nucleic Acids Res*, 2014, **42** (Database issue):D142-7.

34. 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.

537 35. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: **hPDI: a database of experimental human**
538 **protein-DNA interactions.** *Bioinformatics*, 2010, **26**(2):287-9.
539 36. Schones DE, Sumazin P, Zhang MQ. **Similarity of position frequency matrices for**
540 **transcription factor binding sites.** *Bioinformatics.*, 2005, 21(3):307-13.
541 37. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs**
542 **inference projects and methods.** *PLoS Comput Biol*, 2009, **5**(1): e1000262.
543 38. Sonnhammer EL, Östlund G: **InParanoid 8: orthology analysis between 273 proteomes,**
544 **mostly eukaryotic.** *Nucleic Acids Res*, 2015, **43** (Database issue):D234-39.
545 39. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L,
546 Lobanenkov VV, Ren B. **A map of the cis-regulatory sequences in the mouse genome.**
547 *Nature*., 2012, 488(7409):116-20.
548 40. He A, Kong SW, Ma Q, Pu WT: **Co-occupancy by multiple cardiac transcription factors**
549 **identifies transcriptional enhancers active in heart.** *Proc Natl Acad Sci U S A*,
550 2011,**108**(14):5632-37.
551 41. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang
552 J.: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol*,
553 2013, **9**(11):e1003326.
554 42. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.**
555 *Nat Protoc*,2012, **7**(9):1728-40.
556 43. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C,Singh H,
557 Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-**
558 **regulatory elements required for macrophage and B cell identities.** *Mol Cell*,2010,
559 **38**(4):576-89.

560 44. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. **MOODS: fast search for**
561 **position weight matrix matches in DNA sequences**. *Bioinformatics.,* 2009, 25(23):3181-2.
562 45. Ullman JD, Widom J. A first course in database systems, 3rd. Beijing. Pearson Education
563 Asia Limited & China Machine Press; 2008. p.140-147.
564 46. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, Lange M,Tönjes M,
565 Dunkel I, Sperling SR: **The cardiac transcription network modulated by Gata4, Mef2a,**
566 **Nkx2.5, Srf, histone modifications, and microRNAs.** *PLoS Genet,* 2011,**7**(2):e1001313.
567 47. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN,
568 Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS,
569 Holloway AK, Boyer LA, Bruneau BG. **Dynamic and coordinated epigenetic regulation of**
570 **developmental transitions in the cardiac lineage**. *Cell.* 2012, 151(1):206-20.
571 48. Wamstad JA, Wang X, Demuren OO, Boyer LA. **Distal enhancers: new insights into heart**
572 **development and disease.** *Trends Cell Biol. ,*2014, 24(5):294-302.

Figure 1(on next page)

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

The six boxes represent the six major classes, namely CardioTFmatrix, CardioTFCenter, CardioTree, CardioGeneNames, CardioEnhancer, and ChIPExpAssay. These classes are analogous to entity/relationship sets. Each class has 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.

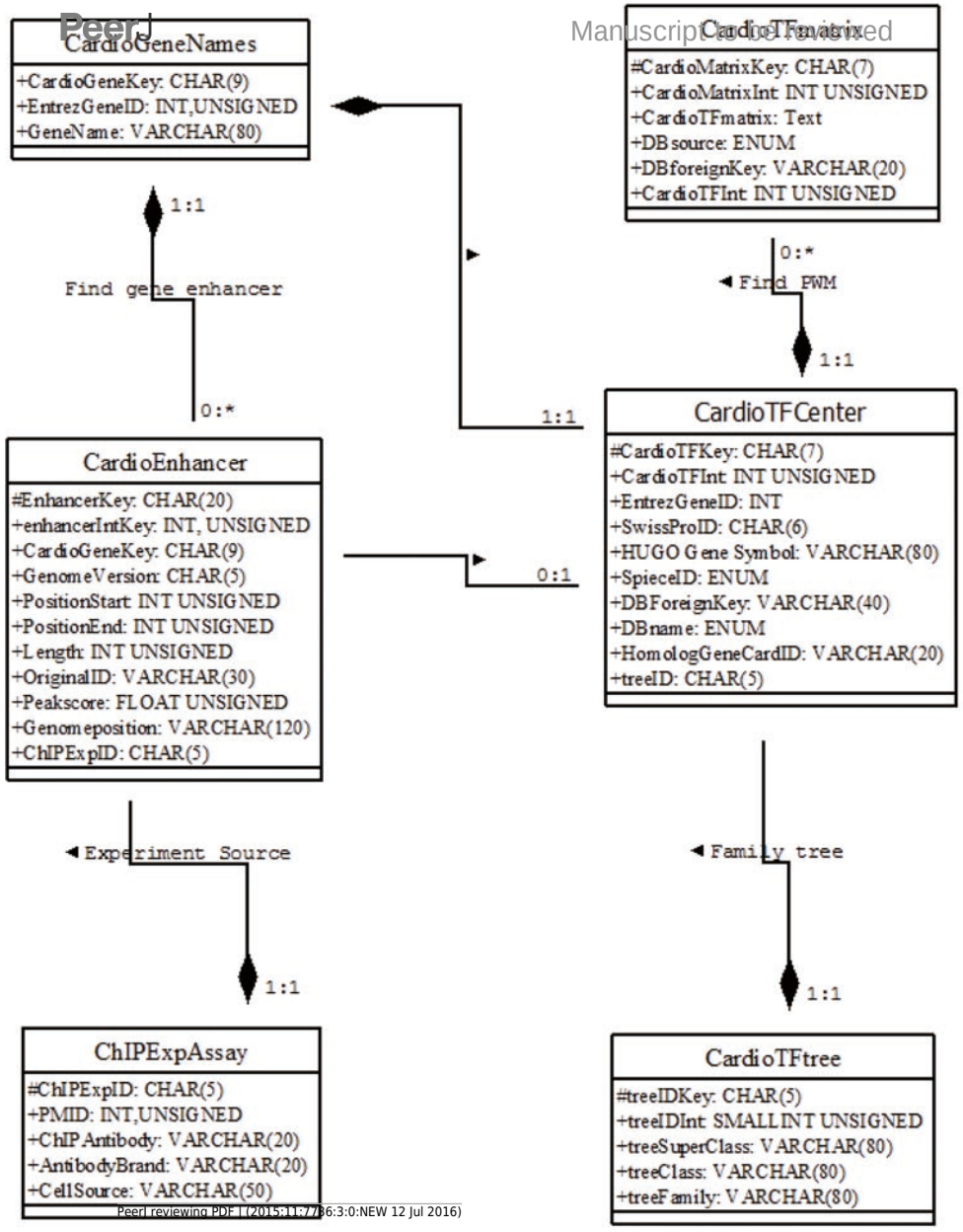
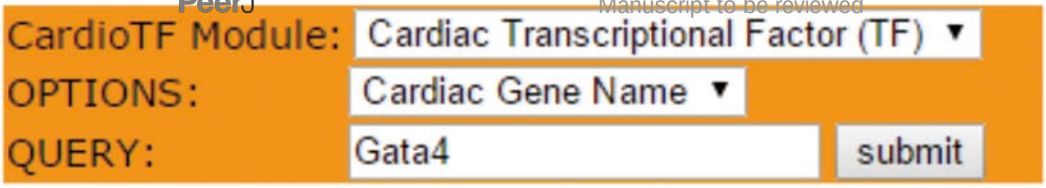


Figure 2(on next page)

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 regions found by the ChIP-seq scanning program. These regions are from the same specie, but might be from different experiments. A user can check the experiment inforamtion by clicking the E0000XXX link which represents the primary key (ID) for this enhancer region, thus allowing the user to save it and retrieve the information later. (C) Query results for the GATA4 TF across species. TFs are listed and indexed according to their database identifiers.

A



B

C

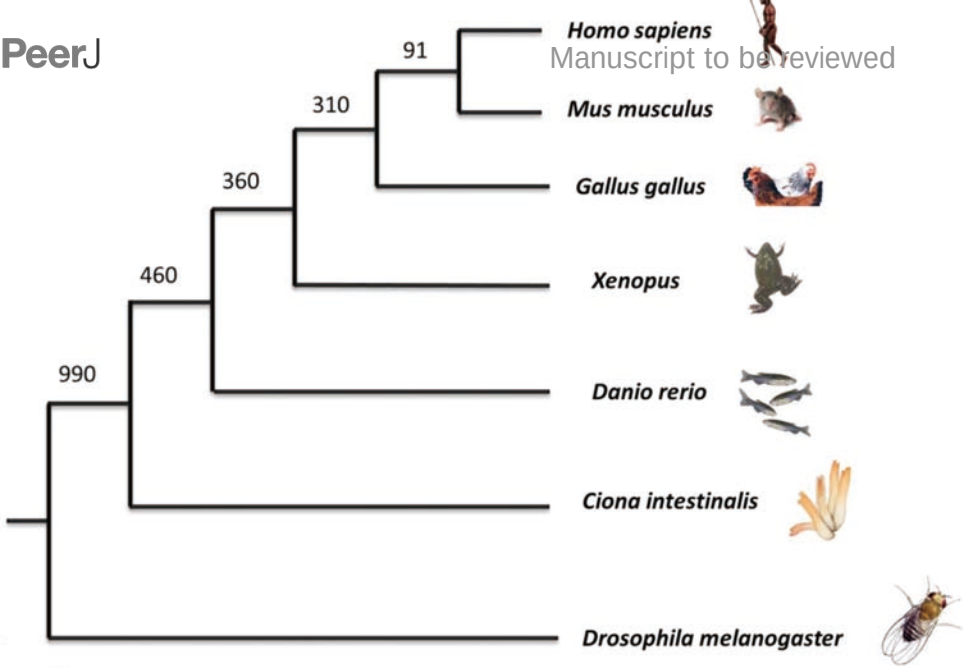
CardioTF Identifier	Gene Symbol	Species Name	EntrezGeneID	SwissProID	DB_link
TF02887	gata4	*Danio rerio*	30483	B8JKU1	ZFIN
TF01887	GATA4	*Homo sapiens*	2626	P43694	NCBI
TF04086	gata4	*Xenopus (Silurana) tropicalis*	549703	Unknown	XenBase
TF04714	GATA4	*Gallus gallus*	396392	Unknown	BirdBase
TF05400	PNR	*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 of TFs across six species. All TFs have homologs in humans. The unit of Y-axis is TF number.

A



B

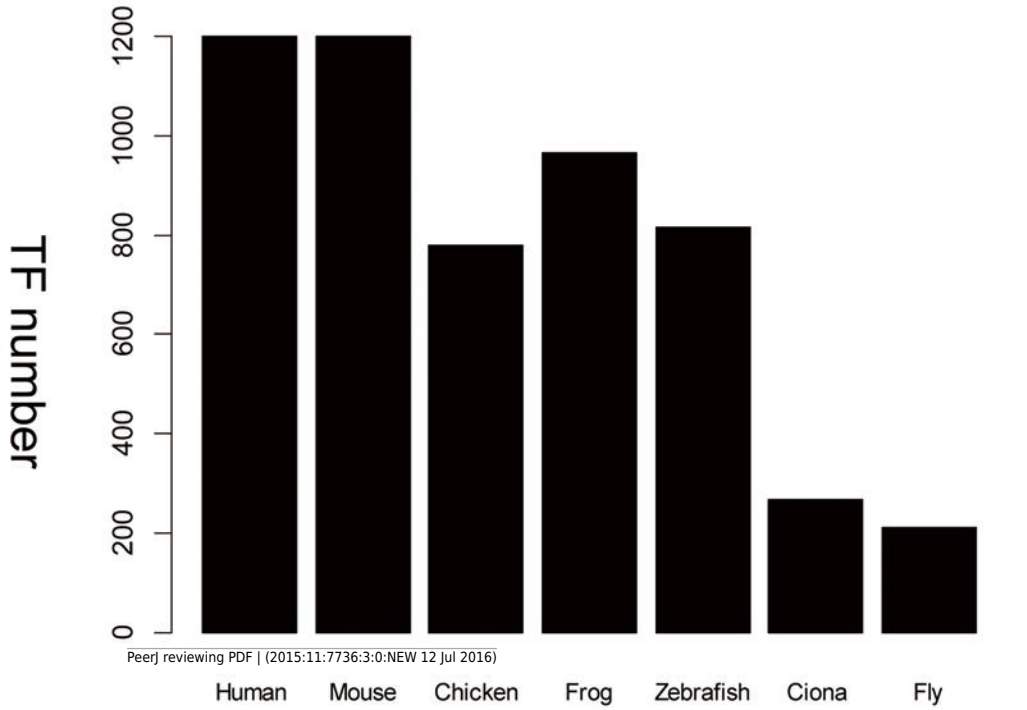
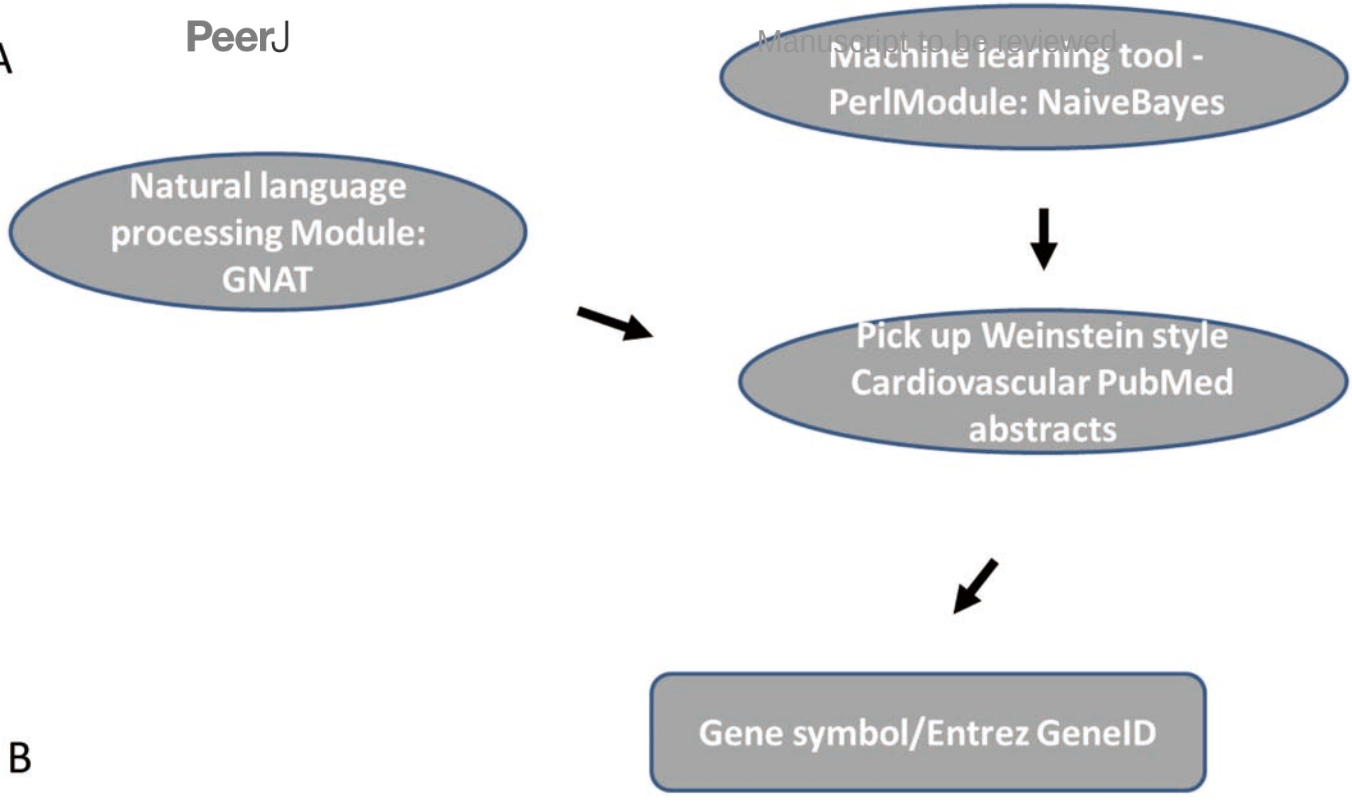


Figure 4(on next page)

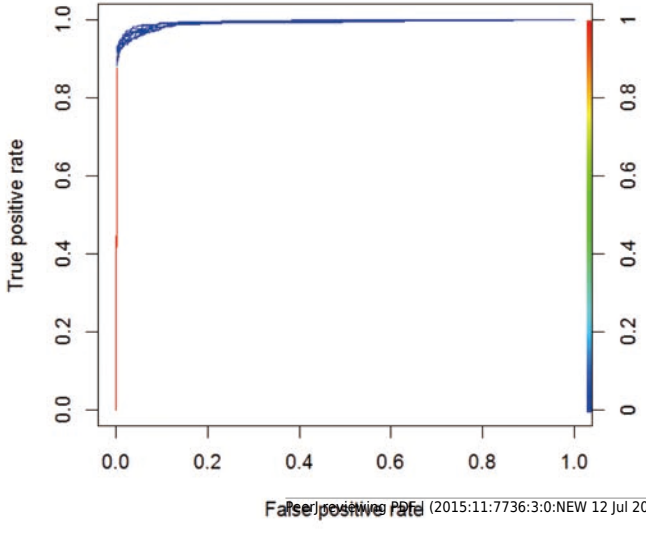
Machine learning protocol used to select TFs described in Weinstein-like papers.

(A) The pipeline used to select TF gene symbols from Weinstein PubMed abstracts. First, a 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