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 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, mouse and human 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 machine learning method and then by intersecting four independent data sources: RNA profiling, expert annotation, PubMed abstract 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, mouse and human 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 machine learning method and then by intersecting four independent data sources: RNA profiling, expert annotation, PubMed abstract 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) 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 involved in the heart transcriptional regulatory network will enhance our understanding of heart disease.

Although the vast amounts of data generated by high-throughput 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, 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

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 a machine learning approach, which could be used as a roadmap to generate gene regulatory network of heart development.

Materials & Methods

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. By definition, cardiac TFs themselves should be involved in the steps of specification, determination, patterning, and differentiation that will result in a heart fate. We recruited cardiac TFs involved in the development of the epicardium and endocardium, both of which, along with the myocardium, should be from the Mesp1-expressing cell lineage [15]. Mesp1 is an indicator of pre-cardiac mesoderm, which, to our knowledge, is the earliest marker that labels all cardiac lineages. In *Ciona*, the Mesp enhancer specifies heart precursor cells [16]. In the initial screen, we identified human TFs from previously published annotations [17]. 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 [18]. Human-specific TFs are defined as genes which have no homologs in the mouse genome. The NCBI HomoloGene database [19] was used as a reference to assess homologs between human and mouse/zebrafish. Other specie homologs of human were retrieved from their central databases, namely, FlyBase (Fly), BirdBase (Chicken), Aniseed (*Ciona*) and Xenbase (Frog) [20-23]. We also documented the expression status for mouse TFs from four independent sources which included annotation from the Cardiovascular Gene Ontology Annotation Initiative [24], Mouse Genome Database (MGI) genes with cardiovascular phenotype [25], 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 Weinstein meeting abstracts from 2010 to 2013 (954 abstracts) as the positive group. 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

105 selected. Journals in the negative group contained research from across kingdoms and topics
106 obviously in other fields, such as 'Sleep_Disord' or 'Toxicology'. All journal names in the
107 negative group were saved in a file and uploaded onto the cardioTF server
108 (negative_set.journal.txt).The negative group includes 57080 abstracts. We split the data
109 (positive and negative groups) into a training (80%) and test (20%) set. The Naïve-Bayes
110 module from The Comprehensive Perl Archive Network (CPAN) was used with a local Perl
111 script to classify Weinstein-like abstracts. We used the training set and adopted the 5×2 cross-
112 validation proposed by Dietterich [26] to train and validate the data. The parameter (the cutoff
113 to decide whether an abstract is a true Weinstein-like) was selected based on average predictive
114 performance which led to an accuracy (ACC) of 0.99. A wrapper function was implemented to
115 parse the abstract and calculate the word frequencies. This function called two Perl modules
116 (Lingua::EN::Splitter and Lingua::EN::StopWords) to extract words and perform text analysis.
117 In doing so, the word frequency was forwarded to the algorithm as the only feature. A test set
118 using optimized parameter was used to assess the algorithm's final performance. All publication
119 abstracts from 2008 to 2013 were downloaded to the local environment and analyzed by the
120 algorithm. We targeted journals which had at least 6 publications classified as Weinstein-like
121 abstracts in the six-year period (annual publication rate is >= 1). Then all abstracts from the
122 targeted journal were downloaded. All abstracts were downloaded from journals that met the
123 criteria. The selected abstracts were then processed by GNAT [27] using its default script
124 (test100.sh) to recognize the gene name in mouse. PMID was recorded when the gene name
125 matched with the name in the curated mouse TF set.

126 **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 the 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 (GSE47950 and GSE29184) were re-analyzed using the recommended protocol [28]. 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 [29-31]. 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 [32]. Frog, chicken and *Ciona* TF homolog annotations were downloaded from their central databases including Xenbase, BirdBase and ANISEED. Fly TFs, which have their counterparts in the human proteome, were annotated by the Inparanoid system [33]. 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 by ChIP-seq signals. Peak calling was performed using the recommended pipeline [34]. In brief, sequencing reads were aligned to the mm10/hg19

reference genome using Bowtie/Bowtie2. mm9/hg9 represent the genome build assigned by UCSC. Index files for mm9/hg19 were downloaded from the iGenome project.

Bowtie call

bowtie -m 2 -S -q -p 8

Peak calling was performed using the MACS peak calling algorithm [35].

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 [36].

Recognition of transcription factor binding sites (TFBSs) in enhancer

CardioSignalScan was previously implemented to identify the transcription factor binding sites. However, this local program is brute-force solution which consumes computational time exponentially making unrealistic to scan sequences longer than 3000bp. This prompted us to choose MOODS [37] instead, which reduces the computational time proportionally to position weight matrix length. 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).

threshold = min_log_score + (max_log_score – min_log_score) * cutoff

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

Results

The database schema

Our database uses MariaDB, a drop-in replacement for MySQL, as the database management system (DBMS). To address how information will be stored nad how the lements will related to one another, we used the unified modeling language (UML) to describe the high-level database

model [38]. 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 [38] (Fig. 1). When 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 tackle the following problems: (1) identify homology information for the queried TF across six species and linking 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 fucntions required to construct a transcriptional network of heart development. **Cardiovascular TFs in the database**

Wingender's annotation set [16] was used as a benchmark to recruit TFs across species. 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*. Therefore, 1200 mouse TFs were

196 collected. Other established animal models for cardiovascular development include fly, *Ciona*,
197 zebrafish, frog, and chicken. TFs from these species were collected if they were homologs to the
198 above mouse TFs. The distribution of TFs from different species is shown in Fig. 3. The
199 expression status of mouse TFs was verified by four independent resources, namely RNA-seq
200 data re-analysis [39], phenotype annotations from the MGI database [40], expert
201 recommendation from the UK Cardiovascular Gene Annotation Initiative [24], and PubMed
202 relevance from classification of Weinstein-like abstracts (see the subsequent section and Table
203 S1).

204 **Weinstein TFs from PubMed analysis**

205 We identified the journals that favored Weinstein-style papers, which were likely contain gene
206 information one genes involved in cardiovascular development. As expected, after using a
207 machine learning method, the journals we identified were among the 30 journals most relevant
208 to developmental biology. Two of the journals (*Circ. Res*. and *J Mol Cell Cardiol.*) obviously
209 publish research specifically in the area of heart system. (Table S1: CardioJournalDistribution).
210 If normalized and ranked by publication rate, the above conclusion still holds true although two
211 different heart journals (*Eur J Echocardiogr, Heart Rhythm*) are in the top 30 list (Table S1:
212 CardioJournalDistribution_norm) in this case.We then used GNAT, a tool that recognizes gene
213 names in the literature, to recover all TFs mentioned in Weinstein-style abstracts because we
214 assumed that these TFs are studied by researchers in the cardiovascular community (Fig. 4, Fig.
215 S1 and Table S1).

216 **PWM files collected in database**

217 Public databases for PWM files include UniPROBE, Jaspar, and hPDI, and they provide PWM
218 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 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. Inclusion of these 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 TFs are particularly enriched in cardiac muscle differentiation (Fig. S2). Is this set of gene 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 of various development stages. 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 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-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 developmental stages of the heart. The searching result of a single enhancer also provide user interface to scan the TFBSs. The binding matrices provided in the list are came from the core cardiac TFs. The output in the flat text format contains the matrix key of the TFs in the database and list of hits in the order.

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. The on-the-fly searching 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 of various TFs across those species and PWM information corresponding to TFs and enhancers from high-throughput ChIP-seq data. 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 Mesp-1 lineage, thus including 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 by four independent sources, namely, RNA profiling, expert curation, PubMed abstract parsing and phenotype annotation to support their roles in

cardiovascular development. In analyze PubMed abstracts, a machine learning approach was used to classify the Weinstein-style abstracts and then pick out the TFs embedded in those abstracts. 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). The overlap of core cardiac TFs identified by these sources provide a roadmap for systems biology to construct 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 [41,42]. Similar approach by other genome biologists who tried to find cardiac enhancers on a genome scale have been reviewed elsewhere [43]. 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 data 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 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 contribute to the gene specific expression patterns. CardioTF is the initial step into the construction of a comprehensive 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 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 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 like to thank the two anonymous reviewers and Dr. Rackham for their valuable comments and suggestions to improve the quality of the paper. 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..: **ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments.** *Nucleic Acids Res*, 2011,**39**(Database issue):D1002-4.

355 6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA,
356 Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N,Robertson CL, Serova N,
357 Davis S, Soboleva A..: **NCBI GEO: archive for functional genomics data sets--update.**
358 *Nucleic Acids Res*, 2013, **41** (Database issue):D991-5.
359 7. Chacon D, Beck D, Perera D, Wong JW, Pimanda JE: **BloodChIP: a database of**
360 **comparative genome-wide transcription factor binding profiles in human blood cells**.
361 *Nucleic Acids Res*, 2014, **42**(Database issue):D172-7.
362 8. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, Wang S, Chen J, Shen L, Duan X, Hu S, Li
363 W, Long H, Zhang Y, Liu XS..: **CistromeMap: a knowledgebase and web server for ChIP-**
364 **Seq and DNase-Seq studies in mouse and human.** *Bioinformatics*, 2012, **28** (10):1411-12.
365 9. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL,
366 Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee
367 G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A,
368 Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I,
369 van Baren J, Brent M, Haussler D, Kellis M,Valencia A, Reymond A, Gerstein M, Guigó R,
370 Hubbard TJ. **GENCODE: the reference human genome annotation for The ENCODE**
371 Project. Genome Res., 2012, 22(9):1760-74.
372 10. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai
373 EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston
374 RH; modENCODE Consortium. Unlocking the secrets of the genome. Nature., 2009,
375 459(7249):927-30
376 11. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap
377 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. Bondue A, Blanpain C: **Mesp1: a key regulator of cardiovascular lineage commitment.** *Circ Res*, 2010, **107**(12):1414-27.

16. Satou Y, Imai KS, Satoh N. The ascidian Mesp gene specifies heart precursor cells. *Development.* 2004, 131(11):2533-41.

17. Wingender E, Schoeps T, Dönitz J: **TFClass: an expandable hierarchical classification of human transcription factors.** *Nucleic Acids Res*, 2013, **41** (Database issue):D165-70.

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

19. NCBI Resource Coordinators. **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2015, 43(Database issue):D6-17.

400 20. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ; FlyBase
401 consortium. **FlyBase: establishing a Gene Group resource for Drosophila melanogaster**.
402 *Nucleic Acids Res*. 2016, 44(D1):D786-92.
403 21. Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, Karimi K,
404 Zorn AM, Vize PD.: **Xenbase, the Xenopus model organism database; new virtualized**
405 **system, data types and genomes.** *Nucleic Acids Res*, 2015, **43** (Database issue):D756-63.
406 22. Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, McGrath
407 S, Mardis E, Stein LD.: **Gallus GBrowse: a unified genomic database for the chicken.**
408 *Nucleic Acids Res*, 2008, **36** (Database issue):D719-23.
409 23. Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V,Caillol D,
410 Schiappa R, Laporte B, Rios A, Luxardi G, Kusakabe T, Joly JS, Darras S, Christiaen L,
411 Contensin M, Auger H, Lamy C, Hudson C, Rothb?cher U, Gilchrist MJ, Makabe KW, Hotta K,
412 Fujiwara S, Satoh N, Satou Y, Lemaire P..: **The ANISEED database: digital representation,**
413 **formalization, and elucidation of a chordate developmental program.** *Genome Res*, 2010,
414 **20**(10):1459-68.
415 24. Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S, Talmud PJ, Breckenridge R,
416 Bhattarcharya S, Riley P, Scambler P, Lovering RC.: **The representation of heart**
417 **development in the gene ontology.** *Dev Biol*, 2011, **354**(1):9-17.
418 25. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE: Mouse Genome Database Group.
419 **The Mouse Genome Database: integration of and access to knowledge about the**
420 **laboratory mouse.** *Nucleic Acids Res*, 2014, **42** (Database issue):D810-17.
421 26. Dietterich TG: **Approximate statistical tests for comparing supervised classification**
422 **learning algorithms**. *Neural Computation*, 1998, **10**:1895-923.

423 27. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization**
424 **of gene mentions with GNAT.** *Bioinformatics*, 2008, **24**(16):126-32.
425 28. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H,Salzberg SL, Rinn
426 JL, Pachter L.: **Differential gene and transcript expression analysis of RNA-seq**
427 **experiments with TopHat and Cufflinks.** *Nat Protoc*, 2012, **7**(3):562-78.
428 29. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S,
429 Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B,Sandelin A,
430 Wasserman WW..: **JASPAR 2014: an extensively expanded and updated open-access**
431 **database of transcription factor binding profiles.** *Nucleic Acids Res*, 2014, **42** (Database
432 issue):D142-7.
433 30. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML: **UniPROBE, update 2015: new tools**
434 **and content for the online database of protein-binding microarray data on protein-DNA**
435 **interactions.** *Nucleic Acids Res*, 2015, **43**(Database issue):D117-22.
436 31. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: **hPDI: a database of experimental human**
437 **protein-DNA interactions.** *Bioinformatics*, 2010, **26**(2):287-9.
438 32. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs**
439 **inference projects and methods.** *PLoS Comput Biol*, 2009, **5**(1): e1000262.
440 34. Sonnhammer EL, Östlund G: **InParanoid 8: orthology analysis between 273 proteomes,**
441 **mostly eukaryotic.** *Nucleic Acids Res*, 2015, **43** (Database issue):D234-39.
442 34. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang
443 J.: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol*,
444 2013, **9**(11):e1003326.

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

36. 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.

37. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. **MOODS: fast search for position weight matrix matches in DNA sequences**. *Bioinformatics.,* 2009, 25(23):3181-2.

38. Ullman JD, Widom J. A first course in database systems, 3rd. Beijing. Pearson Education Asia Limited & China Machine Press; 2008. p.140-147.

39. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. **A map of the cis-regulatory sequences in the mouse genome.** *Nature*., 2012, 488(7409):116-20.

40. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG. **Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage**. *Cell.* 2012, 151(1):206-20.

41. 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.

42. 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.

43. Wamstad JA, Wang X, Demuren OO, Boyer LA. **Distal enhancers: new insights into heart 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 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.

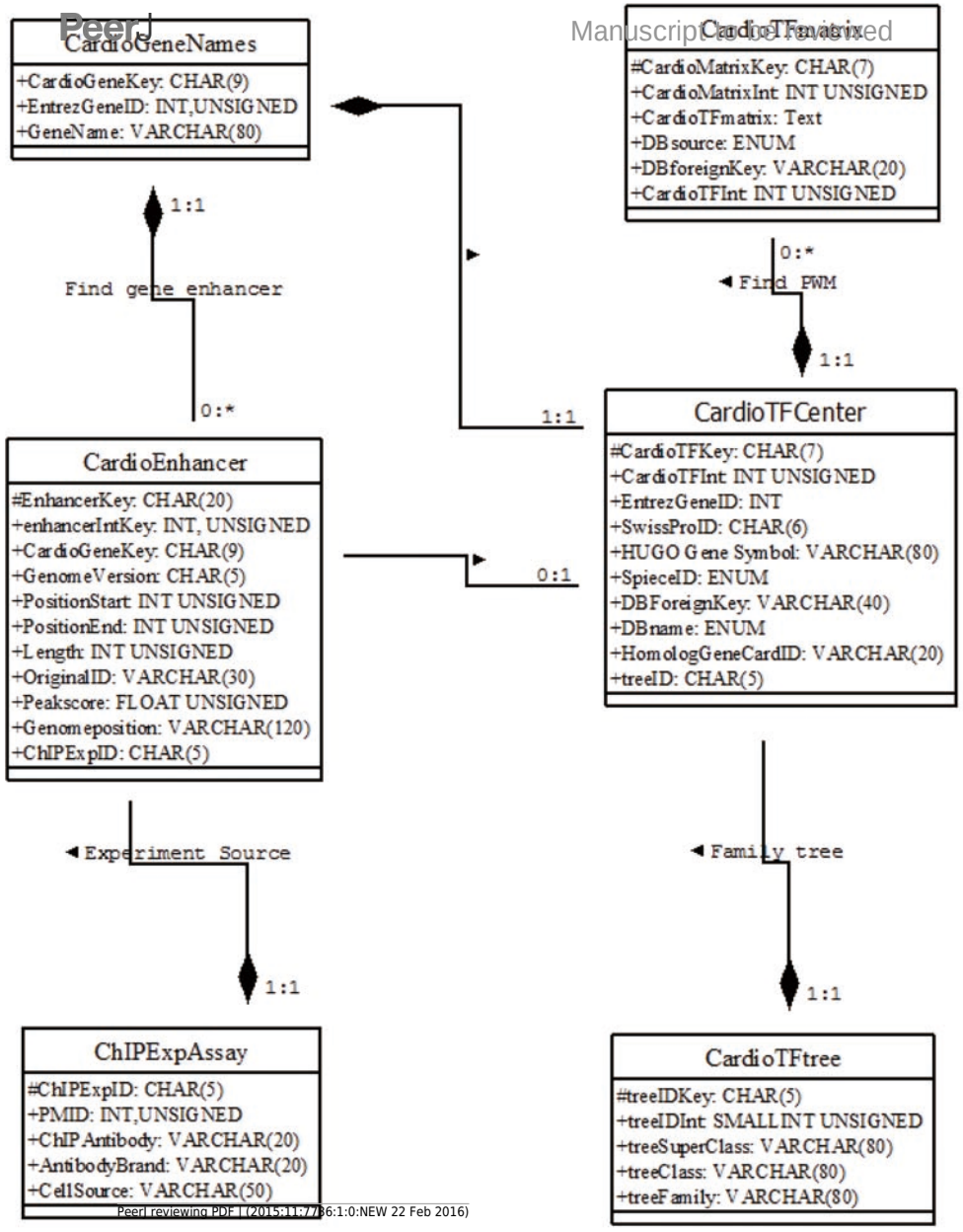
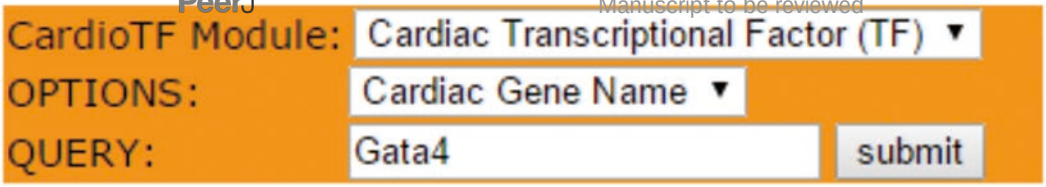


Figure 2(on next page)

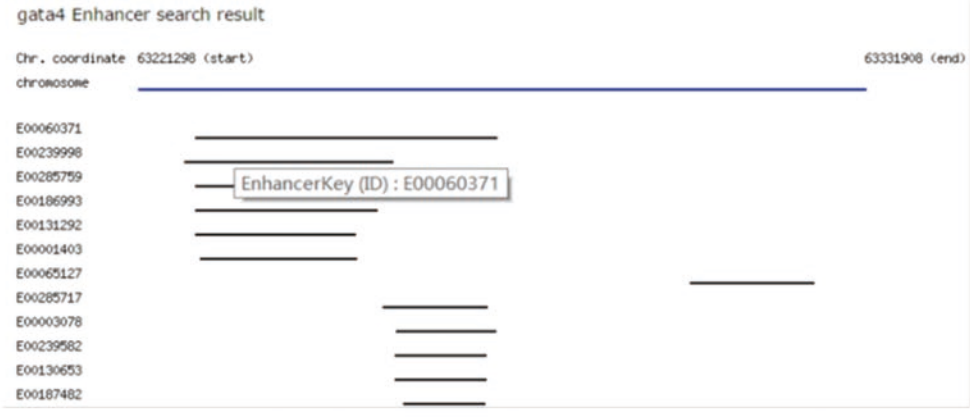
The search engine and the web interface of the database.

(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



B



C

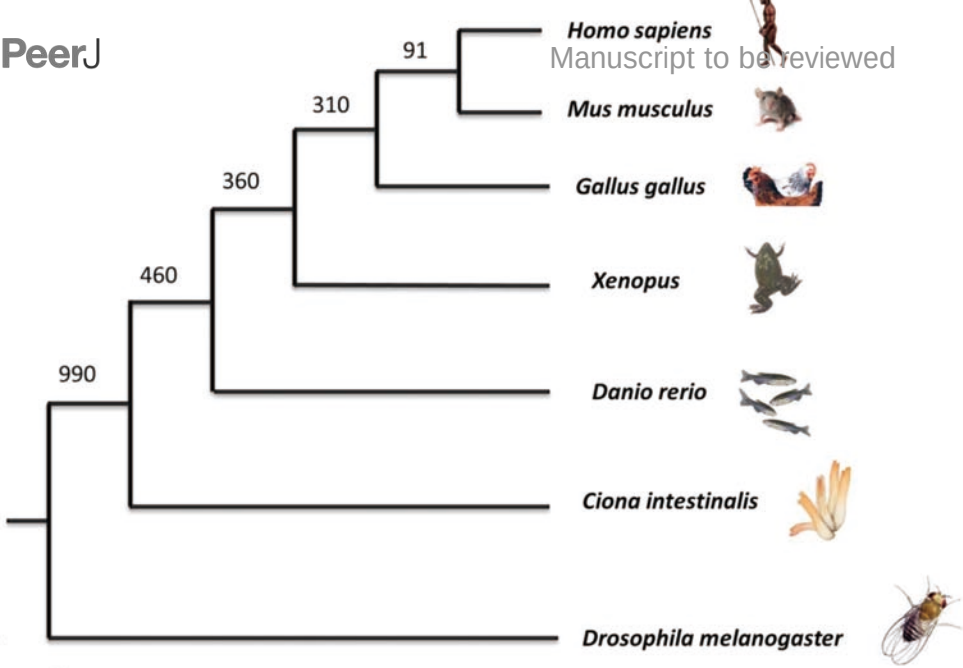
CardioTF Identifier	Gene Symbol	Species Name	EntrezGeneID	SwissProID	DB_link
TF02887	gata4	*Danio rerio*	30483	B8JKU1	ZFIN
TF01887	GATA4	*Homo sapiens*	2626	P43694	NCBI
TF04086	gata4	*Xenopus (Silurana) tropicalis*	549703	Unknown	XenBase
TF04714	GATA4	*Gallus gallus*	396392	Unknown	BirdBase
TF05400	PNR	*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

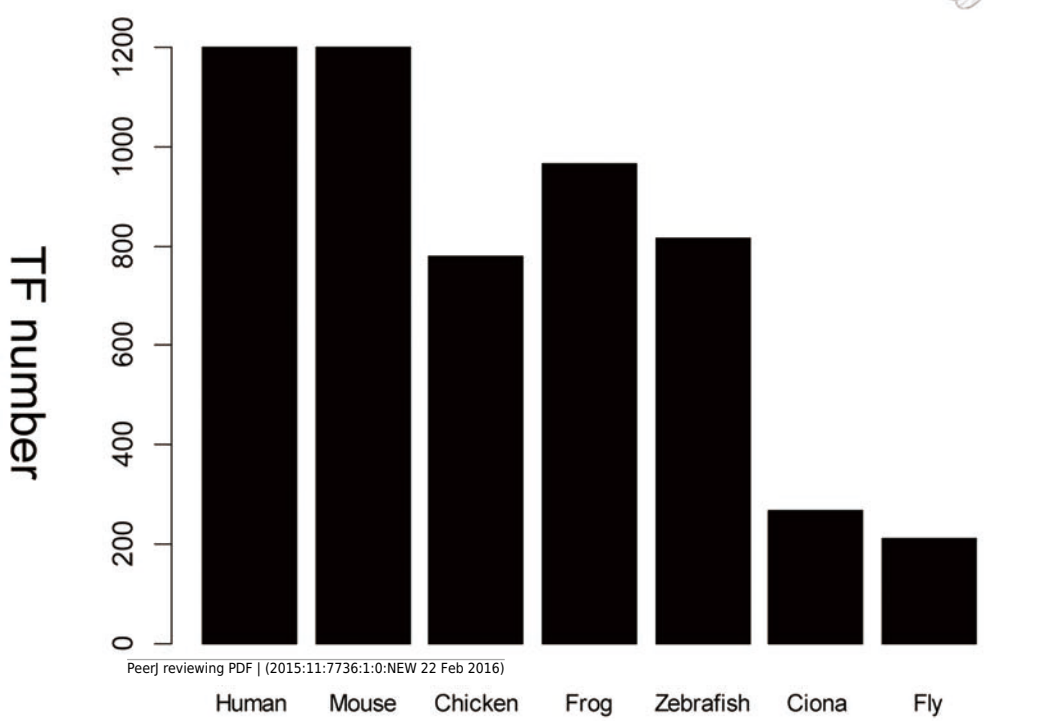
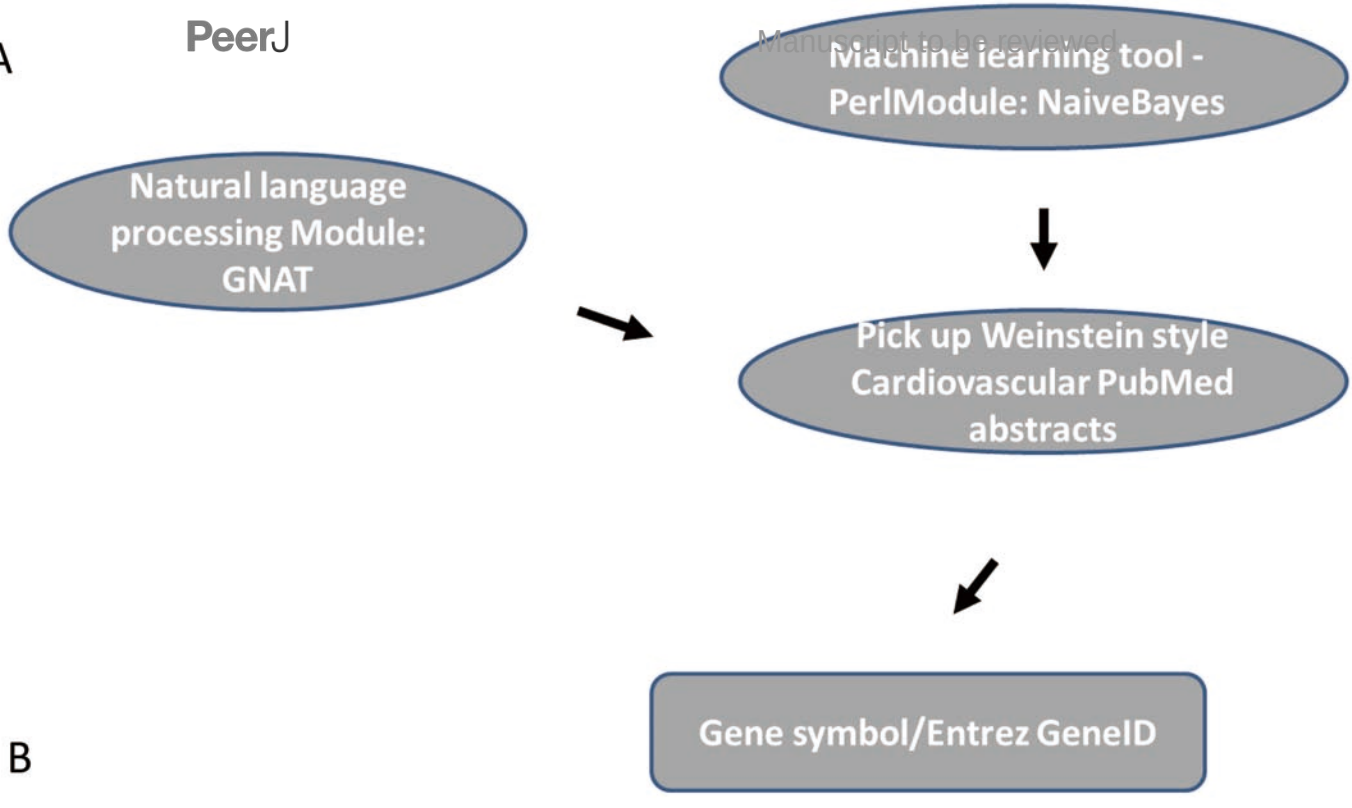


Figure 4(on next page)

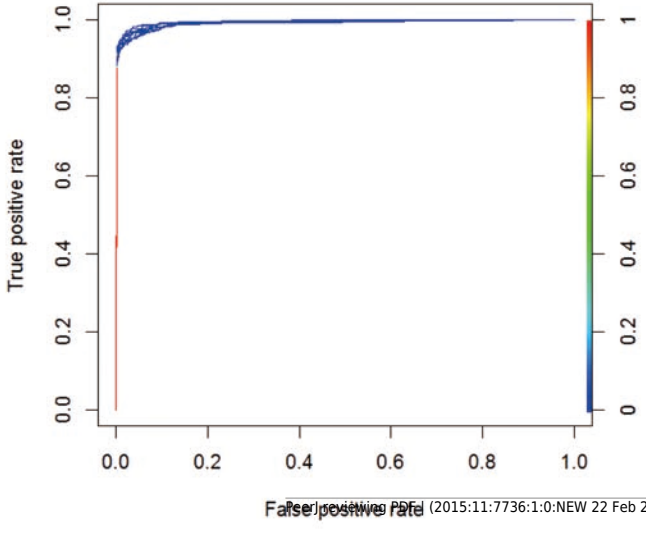
Machine learning protocol to select TFs described in Weinstein-like papers.

(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