

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, 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>.

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

3

4 Yisong Zhen¹

5

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

9

10 Corresponding Author:

11 Yisong Zhen

12 Beilishilu, 167, Beijing, 100037 P.R.China

13 Email address: zhenyisong@fuwaihospital.org

14 **Abstract**

15 **Background.** Information on cardiovascular gene transcription is fragmented and far behind the
16 present requirements of the systems biology field. To create a comprehensive source of data for
17 cardiovascular gene regulation and to facilitate a deeper understanding of genomic data, the
18 CardioTF database was constructed. The purpose of this database is to collate information on
19 cardiovascular transcription factors (TFs), position weight matrices, and enhancer sequences
20 discovered using the ChIP-seq method.

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

26 **Results.** Known cardiovascular TFs from humans and human homologs from fly, *Ciona*,
27 zebrafish, frog, chicken, and mouse were identified and deposited in the database. Position
28 weight matrices from Jaspar, hPDI, and UniPROBE databases were deposited in the database
29 and can be retrieved using their corresponding TF names. Gene enhancer regions from various
30 sources of ChIP-seq data were deposited into the database and were able to be visualized by
31 graphical output. Besides biocuration, mouse homologs of the 81 core cardiac TFs were selected
32 using a Naïve-Bayes approach and then by intersecting four independent data sources: RNA
33 profiling, expert annotation, PubMed abstracts and phenotype.

34 **Discussion.** The CardioTF database can be used as a portal to construct transcriptional network
35 of cardiac development.

36 **Availability and Implementation.** Database URL:
37 <http://www.cardiosignal.org/database/cardiotf.html>

38 **Introduction**

39 Heart disease is a leading cause of morbidity and mortality in both infants and adults [1,2].
40 Insights into the cause of congenital heart diseases (CHDs) have led to the identification of
41 mutations in essential cardiac transcription factors (TFs) [3]. At the opposite end of the temporal
42 spectrum, some cases of adult cardiac disease have been traced to variation in gene regulatory
43 sequences [4]. Thus, knowledge of TFs, their downstream targets, and the regulatory genomic
44 sequences involved in the heart development will enhance our understanding of heart disease.

45 Although the vast amounts of data generated by high-throughput technologies are archived in
46 databases such as ArrayExpress or GEO of NCBI [5,6], they do not contain cohesive knowledge
47 and lack expert annotation. In addition, the field of cardiac development has experienced
48 accelerated growth that can be attributed to the use of various animal models. However, to our
49 present understanding, few efforts have been made to create a database which collects cardiac
50 transcriptional information across species, thereby limiting the benefits from an evolutionary
51 perspective to study heart development.

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

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

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

71 **Materials & Methods**

72 **The project's code and data for reproducible research**

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

79 **Comprehensive collection and annotation of cardiac TFs**

80 Cardiac TFs were previously defined as regulators of cardiac gene expression, which can impact
81 the process of heart development, particularly the initiation and maintenance of the myocardium

82 [16]. In the CardioSignal database, efforts were also made to collate the cardiac specific
83 enhancers which drive gene expression in cardiomyocytes. At that time, cardiac specific
84 transcriptional factors were defined as genes that regulated expression of genes in the
85 myocardium. During development, the heart consists of three layers: the myocardium,
86 epicardium and endocardium [17,18]. Additionally, at least four heart-specific cell lineages
87 have been characterized, including cardiomyocytes, endothelial cells, epicardial cells, and
88 fibroblasts, the latter is derived mainly from epicardial cells through the epithelial to
89 mesenchymal transition (EMT) [19,20]. By definition, cardiac TFs themselves should be
90 involved in the steps of specification, determination, patterning, and differentiation that will
91 result in a heart fate. In our CardioTF database we collated, cardiac transcription factors which
92 are expressed in all layers of heart. The goal of the CardioSignal database was to use a machine
93 learning approach to find cardiac enhancers at the genome scale. In contrast, the CardioTF
94 database is constructed to study systems biology of transcription regulation. The cross-talk
95 between different layers will also be explored using this platform. Additionally, *Mesp1* is an
96 indicator of pre-cardiac mesoderm, which, to our knowledge, is the earliest marker that labels
97 all cardiac lineages [21]. In *Ciona*, the *Mesp* enhancer specifies heart precursor cells [22]. In
98 this sense, cardiac TFs should be all derived from *Mesp1* progenitor cells. In the initial screen,
99 we identified human TFs from previously published annotations [23], hence it was named
100 Wingender's annotation set. This dataset is comprehensive in annotating human TFs. We used
101 these human TFs, excluding human-specific TFs, as a reference to search for their homologs in
102 other species, including fly, *Ciona*, zebrafish, chicken, frog and mouse [24]. Human-specific
103 TFs are defined as genes which have no homologs in the mouse genome. The NCBI
104 HomoloGene database [25] was used as a reference to assess homologs between human and

105 mouse/zebrafish. Human homologs from other were retrieved from their central databases,
106 namely, FlyBase (Fly), BirdBase (Chicken), Aniseed (*Ciona*) and Xenbase (Frog) [26-29]. We
107 also documented the expression status for mouse TFs from four independent sources which
108 included annotation from the Cardiovascular Gene Ontology Annotation Initiative [30], Mouse
109 Genome Database (MGI) genes with cardiovascular phenotypes [31], PubMed abstract parsing
110 results and RNA expression profiling results.

111 **TFs from PubMed abstract parsing**

112 The Weinstein Cardiovascular Conference provides a platform for talks and posters on all
113 aspects of heart development and congenital heart disease. The Weinstein meeting abstracts
114 were extracted from the meeting abstract book from 2010 to 2013 by hand. As the positive
115 group, this data set included 954 abstracts. We assumed that Weinstein-like abstracts deposited
116 in PubMed are all from the cardiovascular community and focus on cardiovascular development.
117 Abstracts from the negative group were from non-heart related journals, which were manually
118 selected from the PMC Open Access Subset at NCBI. To choose the negative control journals,
119 the following criteria were set. First, well-known cardiovascular journals were excluded, such as
120 ‘Circulation’ and ‘Circulation Research’. Second, journals without key words ‘heart’ or ‘cardiac’
121 or ‘cardiovascular’ in their title were selected. Third, journals which are dedicated to the study
122 of other organs or diseases, for example, ‘Neuron’ or ‘Cancer’ were selected. Fourth, other
123 journal which are unlikely to publish articles about cardiovascular development and related
124 topics, such as journals about plants or viruses, were selected. Journals in the negative group
125 contained research from across kingdoms and topics obviously in other fields, such as
126 ‘Sleep_Disord’ or ‘Toxicology’. All journal names in the negative group were saved in a file
127 and uploaded onto the cardioTF server (negative_set.journal.txt).The negative group includes

128 57080 abstracts. We split the data (positive and negative groups) into a training (80%) and test
129 (20%) set. The Naïve-Bayes module from The Comprehensive Perl Archive Network (CPAN)
130 was used with a local Perl script to classify Weinstein-like abstracts. We used the training set
131 and adopted the 5×2 cross-validation proposed by Dietterich [32] to train and validate the data.
132 The parameter (the cutoff to decide whether an abstract is a true Weinstein-like) was selected
133 based on average predictive performance which led to an accuracy (ACC) of 0.99. A wrapper
134 function was implemented to parse the abstracts and calculate the word frequencies. This
135 function called two Perl modules (Lingua::EN::Splitter and Lingua::EN::StopWords) to extract
136 words and perform text analysis. The word frequency alone was forwarded to the algorithm.
137 The withheld test set using the optimized parameter was then used to assess the algorithm's
138 final performance. All publication abstracts from 2008 to 2013 were downloaded to the local
139 environment and analyzed by the algorithm. We targeted journals which had at least 6
140 publications classified as Weinstein-like abstracts in the six-year period (annual publication rate
141 is ≥ 1). Then all abstracts from the targeted journal were downloaded. This process was
142 repeated for all journals that met the criteria. The selected abstracts were then processed by
143 GNAT [33] using its default script (test100.sh) to recognize the mouse gene name. The PMID
144 was recorded when the gene name matched the name in the curated mouse TF set.

145 **RNA expression profiling data procession**

146 Affymetrix data (GSE1479) were processed by R using the MAS5 algorithm which provides a
147 present call for each gene (see the script ExtractAffy.R at Github) Gene expression status was
148 defined as “on” if the gene was expressed in any microarray at selected developmental stages
149 and had a present call. RNA-seq data (GSE47950 and GSE29184) were re-analyzed using the
150 recommended protocol [34]. Briefly, pre-processing software (FastQC) was used to estimate the

151 read length of raw data. If read length is above 50bp, Bowtie2 was used. Otherwise, Bowtie was
152 used. mm10 and hg10 are the genome builds used by UCSC. Index and annotation files for
153 Bowtie2/Bowtie were downloaded from Illumina iGenomes project. Genome sequences from
154 UCSC are repeat-masked with lower-case characters. Any gene with an FPKM value greater
155 than 1 was defined as expressed and this threshold was empirically set although justifiable

156 **Depositing PWM files**

157 The gene symbol was used as the unique identifier to link the original database ID to our local
158 database primary key. A local Perl script was written to change the format to the TRANSFAC
159 style, which was used by our in-house CardioSignalScan program [16]. PWM files were
160 collected from Jaspar, UniPROBE and hPDI databases [35-37]. Users can retrieve their
161 annotations by directing them to the respective database. All the PWMs from those three
162 sources for each TF were deposited in the database. We currently do not use the program
163 implemented by the Zhang lab [38] to check the similarity of PWMs and reduce the redundancy
164 in the collection of PWM files.

165 **Orthologs of TFs from model systems**

166 NCBI has its own gene orthologs that were identified using unpublished algorithm [39]. TFs
167 from mouse, human and zebrafish are annotated by NCBI Homologene [25]. Frog, chicken and
168 *Ciona* TF homolog annotations were downloaded from their central databases including
169 Xenbase, BirdBase and ANISEED. Fly TFs, which have counterparts in the human proteome,
170 were annotated by the Inparanoid system [40]. Each TF collected in the database was assigned
171 one treeID on the basis of its human counterpart. The treeID is equivalent to a TF family by the
172 recommendation of TFClass [23].

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

174 Enhancer regions were defined by ChIP-seq signals. We assume that pan-enhancer markers, like
175 H3K4me1 or H3K27ac [41], or lineage specific markers, like GATA4 or MEF2C [42] will
176 delineate true enhancer regions, although these collections will produce some false positive
177 records. Peak calling was performed using the recommended pipeline [43]. In brief, sequencing
178 reads were aligned to the mm10/hg19 reference genome using Bowtie/Bowtie2. mm9/hg9
179 represents the genome build assigned by UCSC. Index files for mm9/hg19 were downloaded
180 from the iGenome project. MACS1.4.2 was used to process all the ChIP-seq data. The default
181 cutoff for the p-value was 1e-05. This default value was used in all ChIP-seq analysis. This
182 protocol was adapted from published literature [44].

183 Bowtie call

```
184 bowtie -m 2 -S -q -p 8
```

185 Peak calling was performed using the MACS peak calling algorithm.

186 MACS call linux command

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

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

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

191 When possible, the control files were merged:

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

193 After MASC analysis was completed, annotatePeaks.pl in HOMER [45] was used to parse the
194 bed file from the MACS output. Then the parsed results were dumped into the MySQL table.

195 **Recognition of transcription factor binding sites (TFBSs) in enhancer**

196 CardioSignalScan was previously implemented to identify transcription factor binding sites [16].

197 However, this local program (see cardiophylo.pl in GitHub) is brute-force solution which

198 consumes computational time with linear complexity ($O(mn)$). In the Big O notation, m is the
199 matrix length and n is the DNA string length. Therefore, it is unrealistic to scan sequences
200 longer than 3000bp with this local program. This prompted us to choose MOODS [46] instead,
201 which reduces the computational time proportionally to position weight matrix length ($O(m)$). A
202 wrapper module was written to calculate the threshold that gauges the match. The cutoff was
203 empirically defined to be 0.75 (range from 0 to 1 and 1 is most conserved score).

204
$$\text{threshold} = \text{min_log_score} + (\text{max_log_score} - \text{min_log_score}) * \text{cutoff}$$

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

206 **Gene ontology analysis**

207 DAVID analysis (version 6.7) was performed using the 81 TFs as the input gene list, official
208 gene symbols as the identifiers and the entire mouse gene set as the background. The functional
209 annotation clusters results were used. The classification stringency was set to the default
210 (medium).

211 **Results**

212 **The database schema**

213 Our database uses MariaDB, a drop-in replacement for MySQL, as the database management
214 system (DBMS). To address how information will be stored and how the elements will be
215 related to one another, we used the unified modeling language (UML) to describe the high-level
216 database model [47]. UML was originally developed as a graphical notation for describing
217 software designs in an object-oriented style. It has been extended, and modified and is now a
218 popular notation for describing database designs. Here we used UML instead of an
219 entity/relationship diagram to design the relational database schema following modeling
220 principles, such as faithfulness, avoiding redundancy, and simplicity counts [47] (Fig. 1). Where

221 possible, we used a composition distinguished by a line between two classes that ends in a solid
222 diamond at one end. The diamond implies that the label at the end must be 1:1. For example,
223 there is a composition from CardioTFmatrix to CardioTFCenter, which means that every matrix
224 annotation row (PWM related information) belongs to exactly one row in CardioTFCenter (one
225 type of TF may have more PWM records in a CardioTFmatrix table). A 1:1 label at the
226 CardioTFCenter end is implied by a solid black diamond.

227 **Web Interface and search engine**

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

236 Wingender's annotation set [23] was used as a benchmark to recruit TFs across species. The
237 frozen version of this dataset contains 1564 human TFs. Among them, only 1513 TFs have
238 corresponding Entrez gene records. Human-specific TFs, defined as those with no orthologs in
239 the mouse genome, were discarded because no model system could be used to verify their
240 function *in vivo*. This step excluded a further 313 human TFs which have no counterpart in
241 mouse from the homolog annotation. Therefore, 1200 mouse TFs were collected. Other
242 established animal models for cardiovascular development include fly, *Ciona*, zebrafish, frog,
243 and chicken. TFs from these species were collected if they were homologs to the above mouse

244 TFs. The distribution of TFs from different species is shown in Fig. 3. The expression status of
245 mouse TFs was verified by four independent resources, namely RNA-seq data re-analysis [41],
246 phenotype annotations from the MGI database [31], expert recommendation from the UK
247 Cardiovascular Gene Annotation Initiative [30], and PubMed relevance from classification of
248 Weinstein-like abstracts (see the subsequent section and Table S1).

249 **Weinstein TFs from PubMed analysis**

250 We identified the journals that favored Weinstein-style papers, which were likely contain
251 information on genes involved in cardiovascular development. As expected, after using a Naïve-
252 Bayes method, the journals we identified were among the 30 journals most relevant to
253 developmental biology. Two of the journals (*Circ. Res.* and *J Mol Cell Cardiol.*) obviously
254 publish research specifically in the area of heart system. (Table S1: CardioJournalDistribution).
255 If normalized and ranked by publication rate, the above conclusion still holds true although two
256 different heart journals (*Eur J Echocardiogr*, *Heart Rhythm*) are in the top 30 list (Table S1:
257 CardioJournalDistribution_norm) in this case. We then used GNAT, a tool that recognizes gene
258 names in the literature, to recover all TFs mentioned in Weinstein-style abstracts because we
259 assumed that these TFs are studied by researchers in the cardiovascular community (Fig. 4, Fig.
260 S1 and Table S1).

261 **PWM files collected in database**

262 Public databases for PWM files include UniPROBE, Jaspar, and hPDI, and they provide PWM
263 files for TFs. Jaspar PWM files are curated from the published literatures whereas the other two
264 databases generate PWM files from experiments. Our database integrates these three sources,
265 and the TF PWM can be queried on the basis of the TF name. Search results directly link to the
266 original database through the PWM raw database key. The CardioTFmatrix class contains 904

267 records, and these PWM files can be recognized by our local CardioSignalScan program to
268 search for the motifs in genomic regions.

269 **Core cardiovascular TFs**

270 The 1200 mouse TFs were included in the cardiac TF dataset as the entry point to initiate deep
271 annotation. To define a core set of cardiac TFs, we intersected four independent sources of
272 cardiovascular TF collections. Inclusion of the resulting 81 TFs is supported by their expression
273 status, phenotype annotation, expert recommendation and PubMed relevance (Table S1). We
274 also performed DAVID functional analysis, and found that these TFs are particularly enriched
275 in cardiac muscle differentiation (Fig. S2). 10-20% of these TFs which are enriched in the
276 Annotation Cluster 7 including Gata4 ,Gata6, Smad7, Nkx2-5, Tbx2, Tbx5, Foxc1, Foxp1,
277 Prox1, Rara ,Rarb, Rxra, Rxrb, Zfp207. These 14 TFs, are annotated in the DAVID as being
278 involved in cardiac muscle formation. As we know, the heart system includes the endocardium
279 which is a specialized layer derived from endothelial cells. Cluster 8 from our DAVID analysis
280 includes genes expressed in endothelial cells such as Smad5, Smad7, Meis1, Nkx2-5, Tbx20,
281 Epas1, Foxc1, Foxo1, Hey1, Hand2, Hif1a, Mef2c, Prrx1, Prox1, Srf, Nr2f2, Tcf21, Vezf1,
282 Zfp207. Is this set of genes the minimum requirement for cardiac development? Indeed, these
283 four sources of supporting evidence indicate that these TFs genes play a key role in heart
284 development. We wanted to determine if these TFs display specific expression patterns in heart
285 development. A heatmap was generated using seven RNA-seq data sets, including samples from
286 embryonic cells, mesoderm cells, cardiac progenitors, nascent cardiomyocytes and adult heart
287 tissue. This heatmap did not reveal any specific patterns (Fig. S3). In adult tissues, these TFs did
288 not exhibit enriched expression in the adult heart. In the case of TFs which are never expressed

289 at any stage of heart development, no specific expression pattern was revealed by the boxplot
290 assay (Fig. S4-S5).

291 **Cardiovascular enhancers collected in this database version**

292 Few enhancers have already been verified by traditional biological experiments, for example, by
293 using transgenic expression of isolated DNA fragments *in vivo* to analyze temporal-spatial
294 patterns. Therefore, the ChIP-seq method provides a high-throughput approach to delineate
295 enhancer regions at the genome scale. A standard protocol was used to identify genome-wide
296 locations of transcription/chromatin factor binding sites or histone modification sites from
297 ChIP-seq data (Fig. S6). The present database houses 511,893 enhancer records, covering
298 different stages of heart development. Searching for a single enhancer also provides a user
299 interface to scan the TFBSs. The binding matrices provided in the list come from the core
300 cardiac TFs. The flat text output file contains the matrix key for the TFs in the database and
301 sorts these hits in the increasing order.

302 **Discussion**

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

309 The database now contains the six species which are model organisms for studying heart
310 development (Figure 3A). *Ciona* is an ideal model used to study the early specification step in
311 the *Mesp*-lineage. Zebrafish is well suited to perform imaging analysis and for performing

312 quantitative study. The chicken is well suited for lineage tracing *ex vivo*. The mouse provides a
313 well-studied mammalian model and is most similar to humans. The fly is a valuable model for
314 testing new concepts and is easy to study at the genomic level. New models may be added in the
315 future if they have unique advantage over the other models.

316 Previously, we constructed the CardioSignal database which collates cardiac factors driving
317 genes expressed in a tissue-specific or quantitative manner. Most enhancers archived in the
318 database are expressed specifically in the myocardium. CardioTF is a complementary database
319 that accumulates cardiac TFs expressed in the *Mesp-1* lineage, thus including TFs expressed in
320 the epicardium, endocardium and myocardium. This information including PWMs can be used
321 not only to find the features of enhancers in a machine learning approach (such the left-right
322 patterning of the heart) but also to reconstruct regulatory network in systems biology.

323 We defined a core set of TFs using four independent sources, namely, RNA profiling, expert
324 curation, PubMed abstract parsing and phenotype annotation to support their roles in
325 cardiovascular development. In RNA-seq analysis, after pre-processing and post-processing of
326 the adult heart data, the count table contained 48995 genes. After filtering genes with FPKM > 1 ,
327 there were 20863 genes remaining. Roughly half of these annotated genes were abandoned
328 because their gene expression level was too low. To analyze PubMed abstracts, a Naïve-Bayes
329 approach was used to classify Weinstein-style abstracts and then pick out the TFs embedded in
330 those abstracts. Most journals containing the key words “heart” or “cardiac” or “cardiovascular”
331 focus on heart pathology or physiology instead of development biology. For example, journals
332 such as “Cardiovascular research”, “Circulation”, and “Circulation research”, accept papers
333 related to the adult cardiovascular system. Most papers published in these journals take a more
334 translational approach which is oriented to bench-side work. Our analysis identified two heart

335 related journals (which is obvious from their name) suggesting that the algorithm was successful
336 in finding the wording pattern present in Weinstein abstracts. The results indicate that most
337 developmental biology manuscripts relating to cardiovascular system are sent to specialized
338 journals that focus on development. Top-tier cardiovascular journals are more likely to publish
339 papers describing the adult cardiovascular system. Our text analysis, whether normalized by
340 publication number or not, had a tendency of identifying journals favored by heart
341 developmental biologists and journals that specialized in developmental biology (Table S1).

342 Traditional definitions of heart specific TFs can often be ambiguous and should not include TFs
343 that only regulate cardiac muscle. In the present definition, heart specific TFs must play a
344 specific role in establishing the *Mesp-1* lineage, this includes regulators of both the
345 endocardium and epicardium layers. Both DAVID Annotation Cluster 7 and Cluster 8 contain
346 TFs involved in heart muscle or vascular formation. The DAVID analysis result did not reveal
347 any other clusters with genes involved in kidney, liver or the formation of other organs. Even
348 the expression of *GATA4* and *MEF2C*, which are *de facto* cardiac TFs, is not restricted to
349 cardiomyocytes. These TFs are expressed in cardiac progenitors at a certain stage of
350 development. The present approach is empirical and proposes a method which uses four
351 independent data sources to identify true cardiac TFs based on their expression, phenotype,
352 community opinion and PubMed abstracts. These 81 core TFs could be used further to support
353 simulation study to infer the significance in future.

354 In general, the set of core cardiac TFs identified by these sources provide a roadmap for systems
355 biology to construct a transcriptional network of heart development. Current approaches by the
356 Sperling group or the Pu group only report three to four TFs based on ChIP-seq data [42,48].
357 Similar approaches by other genome biologists who tried to find cardiac enhancers on a genome

358 scale have been reviewed elsewhere [49, 50]. However, the information generated from these
359 studies is well below our knowledge of these core cardiovascular TFs, which have multiple
360 sources supporting their role in cardiovascular development.

361 We archived 1200 mouse TFs and wanted to determine at what stage of heart development they
362 were expressed. Our preliminary analysis indicates that approximately 200 TFs have no
363 evidence of their expression pattern, phenotype, expert recommendation, and PubMed abstracts.
364 Whether these TF genes are expressed or play roles in heart disease requires further analysis.

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

369 **Conclusions**

370 Modern translational medicine rests upon the progressive study of pathways and principles from
371 model organisms such as yeast, fly, fish, and mice to clinical studies in humans. Therefore, we
372 recruited TFs from six model organisms which are established models for research on
373 cardiovascular development. The identification and collation of these well-annotated homologs
374 from different organisms will enable investigators to better address the complexities of heart
375 development [50].

376 We hope that in the near future, single-cell sequencing data may provide comprehensive gene
377 expression information with detailed temporal-spatial resolution, thereby providing insight into
378 the enhancer networks that contribute to gene specific expression patterns. CardioTF is the
379 initial step in the construction of a comprehensive transcriptional network.

380 **Supplemental Information**

381 Additional file 1: Four independent sources of cardiac TF lists and preferred journals for
382 Weinstein-like papers. Format: XLSX. Sheet 1 (CardioJournalDistribution) includes the top 100
383 journals that favor Weinstein-like papers. Sheet 2 (NotHeartTFs) includes all TF names that
384 have never appeared in the four sources of supporting evidence. Sheet 3 (MGI_TFs) is the TF
385 names from MGI database. Sheet 4 (PubMed_TFs) lists TF names from Weinstein-like PubMed
386 abstracts. Sheet 5 (Cardio_GO_UK_TFs) includes TF names from UK Cardio-GO project.
387 Sheet 5 (Cardio_lineage_TFs) includes TF names from RNA-seq or microarray analysis. Sheet
388 6 (core_TFs) defines the core 81 TFs known to be involved in heart development. Sheet 7
389 (CardioJournalDistribution_norm) includes journals ranked by publication rate (normalized
390 version)

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

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

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

402 Additional file 5: Format: PDF. Figure S4. RNA-seq expression patterns for the 81 core TFs
403 across 13 adult tissues presented by the boxplot.

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

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

408 **Availability and requirements**

409 CardioTF database is freely available on the web at
410 <http://www.cardiosignal.org/database/cardiotsf.html>.

411 **List of abbreviations**

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

413 **Acknowledgements**

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

419 **References:**

- 420 1. van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ,
421 Roos-Hesselink JW...: **Birth prevalence of congenital heart disease worldwide: a systematic**
422 **review and meta-analysis.** *J Am Coll Cardiol*, 2011, **58**(21):2241-7.
- 423 2. Celermajer DS, Chow CK, Marijon E, Anstey NM, Woo KS: **Cardiovascular disease in the**
424 **developing world: prevalences, patterns, and the potential of early disease detection.** *J Am*
425 *Coll Cardiol*, 2012, **60**(14):1207-16.

- 426 3. McCulley DJ, Black BL: **Transcription factor pathways and congenital heart disease.**
427 *Curr Top Dev Biol*, 2012, **100**:253-77.
- 428 4. Smith JG, Newton-Cheh C: **Genome-wide association studies of late-onset cardiovascular**
429 **disease.** *J Mol Cell Cardiol*, 2015, **83**:131-41.
- 430 5. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I,
431 Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E,
432 Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A.:
433 **ArrayExpress update--an archive of microarray and high-throughput sequencing-based**
434 **functional genomics experiments.** *Nucleic Acids Res*, 2011,**39**(Database issue):D1002-4.
- 435 6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA,
436 Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N,
437 Davis S, Soboleva A.: **NCBI GEO: archive for functional genomics data sets--update.**
438 *Nucleic Acids Res*, 2013, **41** (Database issue):D991-5.
- 439 7. Chacon D, Beck D, Perera D, Wong JW, Pimanda JE: **BloodChIP: a database of**
440 **comparative genome-wide transcription factor binding profiles in human blood cells.**
441 *Nucleic Acids Res*, 2014, **42**(Database issue):D172-7.
- 442 8. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, Wang S, Chen J, Shen L, Duan X, Hu S, Li
443 W, Long H, Zhang Y, Liu XS.: **CistromeMap: a knowledgebase and web server for ChIP-**
444 **Seq and DNase-Seq studies in mouse and human.** *Bioinformatics*, 2012, **28** (10):1411-12.
- 445 9. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL,
446 Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee
447 G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A,
448 Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I,

- 449 van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R,
450 Hubbard TJ. **GENCODE: the reference human genome annotation for The ENCODE**
451 *Project. Genome Res.*, 2012, 22(9):1760-74.
- 452 10. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai
453 EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston
454 RH; modENCODE Consortium. **Unlocking the secrets of the genome. Nature.**, 2009,
455 459(7249):927-30
- 456 11. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. **Epigenomics:**
457 **Roadmap for regulation. Nature.** 2015, 518(7539):314-6.
- 458 12. Mangan ME, Williams JM, Kuhn RM, Lathe WC 3rd. The UCSC Genome Browser: **What**
459 **Every Molecular Biologist Should Know. Curr Protoc Mol Biol.**, 2014, 1;107:19.9.1-19.9.36.
- 460 13. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P,
461 Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S,
462 Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T,
463 McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS,
464 Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S,
465 White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J,
466 Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM.
467 **Ensembl 2013. Nucleic Acids Res.**, 2013, 41(Database issue):D48-55.
- 468 14. Djordjevic D, Yang A, Zadoorian A, Rungrueecharoen K, Ho JW. **How difficult is**
469 **inference of mammalian causal gene regulatory networks? PLoS One.**, 2014, 9(11):e111661.
- 470 15. Blais A, Dynlacht BD. **Constructing transcriptional regulatory networks. Genes Dev.** ,
471 2005, 19(13):1499-1511.

- 472 16. Zhen Y, Wang Y, Zhang W, Zhou C, Hui R. **CardioSignal: a database of transcriptional**
473 **regulation in cardiac development and hypertrophy.** *Int J Cardiol.*, 2007, 116(3):338-347.
- 474 17. Moorman AF, Christoffels VM. **Cardiac chamber formation: development, genes, and**
475 **evolution.** *Physiol Rev.*, 2003, 83(4):1223-1267.
- 476 18. Fishman MC, Chien KR. **Fashioning the vertebrate heart: earliest embryonic decisions.**
477 *Development.* , 1997, 124(11):2099-2117.
- 478 19. Evans SM, Yelon D, Conlon FL, Kirby ML. **Myocardial lineage development.** *Circ Res.*,
479 2010,107(12):1428-1444.
- 480 20. Moore-Morris T, Cattaneo P, Puceat M, Evans SM. **Origins of cardiac fibroblasts.** *J Mol*
481 *Cell Cardiol.*, 2016, 91:1-5.
- 482 21. Bondue A, Blanpain C: **Mesp1: a key regulator of cardiovascular lineage commitment.**
483 *Circ Res*, 2010, **107**(12):1414-27.
- 484 22. Satou Y, Imai KS, Satoh N. **The ascidian Mesp gene specifies heart precursor cells.**
485 *Development.* 2004, 131(11):2533-41.
- 486 23. Wingender E, Schoeps T, Dönitz J: **TFClass: an expandable hierarchical classification of**
487 **human transcription factors.** *Nucleic Acids Res*, 2013, **41** (Database issue):D165-70.
- 488 24. Hutson MR, Kirby ML: **Model systems for the study of heart development and disease.**
489 *Semin Cell Dev Biol*, 2007, **18**(1):1-2.
- 490 25. NCBI Resource Coordinators. **Database resources of the National Center for**
491 **Biotechnology Information.** *Nucleic Acids Res.* 2015, 43(Database issue):D6-17.
- 492 26. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ; FlyBase
493 consortium. **FlyBase: establishing a Gene Group resource for Drosophila melanogaster.**
494 *Nucleic Acids Res.* 2016, 44(D1):D786-92.

- 495 27. Karpinka JB, Fortriede JD, Burns KA, James-Zorn C, Ponferrada VG, Lee J, Karimi K,
496 Zorn AM, Vize PD.: **Xenbase, the Xenopus model organism database; new virtualized**
497 **system, data types and genomes.** *Nucleic Acids Res*, 2015, **43** (Database issue):D756-63.
- 498 28. Schmidt CJ, Romanov M, Ryder O, Magrini V, Hickenbotham M, Glasscock J, McGrath
499 S, Mardis E, Stein LD.: **Gallus GBrowse: a unified genomic database for the chicken.**
500 *Nucleic Acids Res*, 2008, **36** (Database issue):D719-23.
- 501 29. Tassy O, Dauga D, Daian F, Sobral D, Robin F, Khoueiry P, Salgado D, Fox V, Caillol D,
502 Schiappa R, Laporte B, Rios A, Luxardi G, Kusakabe T, Joly JS, Darras S, Christiaen L,
503 Contensin M, Auger H, Lamy C, Hudson C, Rothbacher U, Gilchrist MJ, Makabe KW, Hotta K,
504 Fujiwara S, Satoh N, Satou Y, Lemaire P.: **The ANISEED database: digital representation,**
505 **formalization, and elucidation of a chordate developmental program.** *Genome Res*, 2010,
506 **20**(10):1459-68.
- 507 30. Khodiyar VK, Hill DP, Howe D, Berardini TZ, Tweedie S, Talmud PJ, Breckenridge R,
508 Bhattarcharya S, Riley P, Scambler P, Lovering RC.: **The representation of heart**
509 **development in the gene ontology.** *Dev Biol*, 2011, **354**(1):9-17.
- 510 31. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE: Mouse Genome Database Group.
511 **The Mouse Genome Database: integration of and access to knowledge about the**
512 **laboratory mouse.** *Nucleic Acids Res*, 2014, **42** (Database issue):D810-17.
- 513 32. Dietterich TG: **Approximate statistical tests for comparing supervised classification**
514 **learning algorithms.** *Neural Computation*, 1998, **10**:1895-923.
- 515 33. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G: **Inter-species normalization**
516 **of gene mentions with GNAT.** *Bioinformatics*, 2008, **24**(16):126-32.

- 517 34. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn
518 JL, Pachter L.: **Differential gene and transcript expression analysis of RNA-seq**
519 **experiments with TopHat and Cufflinks.** *Nat Protoc*, 2012, **7**(3):562-78.
- 520 35. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S,
521 Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A,
522 Wasserman WW.: **JASPAR 2014: an extensively expanded and updated open-access**
523 **database of transcription factor binding profiles.** *Nucleic Acids Res*, 2014, **42** (Database
524 issue):D142-7.
- 525 36. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML: **UniPROBE, update 2015: new tools**
526 **and content for the online database of protein-binding microarray data on protein-DNA**
527 **interactions.** *Nucleic Acids Res*, 2015, **43**(Database issue):D117-22.
- 528 37. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: **hPDI: a database of experimental human**
529 **protein-DNA interactions.** *Bioinformatics*, 2010, **26**(2):287-9.
- 530 38. Schones DE, Sumazin P, Zhang MQ. **Similarity of position frequency matrices for**
531 **transcription factor binding sites.** *Bioinformatics.*, 2005, **21**(3):307-13.
- 532 39. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs**
533 **inference projects and methods.** *PLoS Comput Biol*, 2009, **5**(1): e1000262.
- 534 40. Sonnhammer EL, Östlund G: **InParanoid 8: orthology analysis between 273 proteomes,**
535 **mostly eukaryotic.** *Nucleic Acids Res*, 2015, **43** (Database issue):D234-39.
- 536 41. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L,
537 Lobanenkov VV, Ren B. **A map of the cis-regulatory sequences in the mouse genome.**
538 *Nature.*, 2012, **488**(7409):116-20.

- 539 42. He A, Kong SW, Ma Q, Pu WT: **Co-occupancy by multiple cardiac transcription factors**
540 **identifies transcriptional enhancers active in heart.** *Proc Natl Acad Sci U S A*,
541 2011,**108**(14):5632-37.
- 542 43. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang
543 J.: **Practical guidelines for the comprehensive analysis of ChIP-seq data.** *PLoS Comput Biol*,
544 2013, **9**(11):e1003326.
- 545 44. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.**
546 *Nat Protoc*,2012, **7**(9):1728-40.
- 547 45. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H,
548 Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-**
549 **regulatory elements required for macrophage and B cell identities.** *Mol Cell*,2010,
550 **38**(4):576-89.
- 551 46. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. **MOODS: fast search for**
552 **position weight matrix matches in DNA sequences.** *Bioinformatics.*, 2009, 25(23):3181-2.
- 553 47. Ullman JD, Widom J. A first course in database systems, 3rd. Beijing. Pearson Education
554 Asia Limited & China Machine Press; 2008. p.140-147.
- 555 48. Schlesinger J, Schueler M, Grunert M, Fischer JJ, Zhang Q, Krueger T, Lange M, Tönjes M,
556 Dunkel I, Sperling SR: **The cardiac transcription network modulated by Gata4, Mef2a,**
557 **Nkx2.5, Srf, histone modifications, and microRNAs.** *PLoS Genet*, 2011,**7**(2):e1001313.
- 558 49. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN,
559 Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS,
560 Holloway AK, Boyer LA, Bruneau BG. **Dynamic and coordinated epigenetic regulation of**
561 **developmental transitions in the cardiac lineage.** *Cell.* 2012, 151(1):206-20.

562 50. Wamstad JA, Wang X, Demuren OO, Boyer LA. **Distal enhancers: new insights into heart**
563 **development and disease.** *Trends Cell Biol.* ,2014, 24(5):294-302.

564

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.

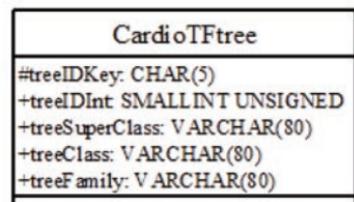
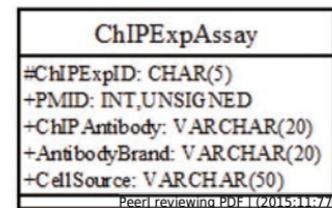
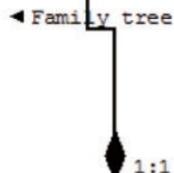
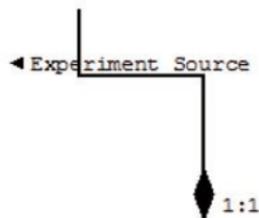
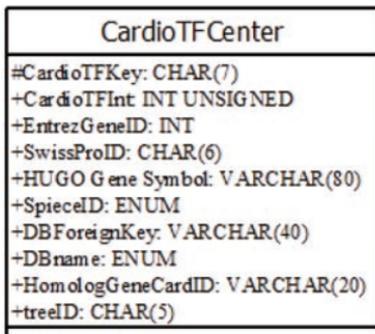
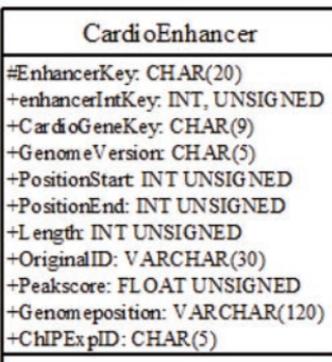
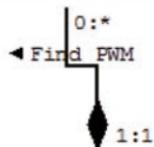
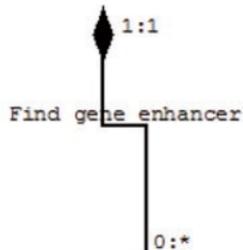
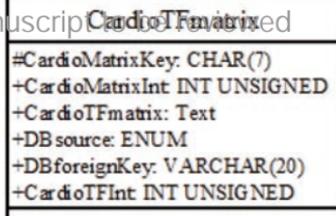
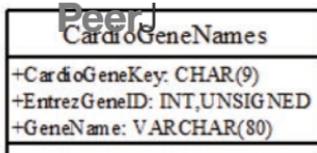


Figure 2(on next page)

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.

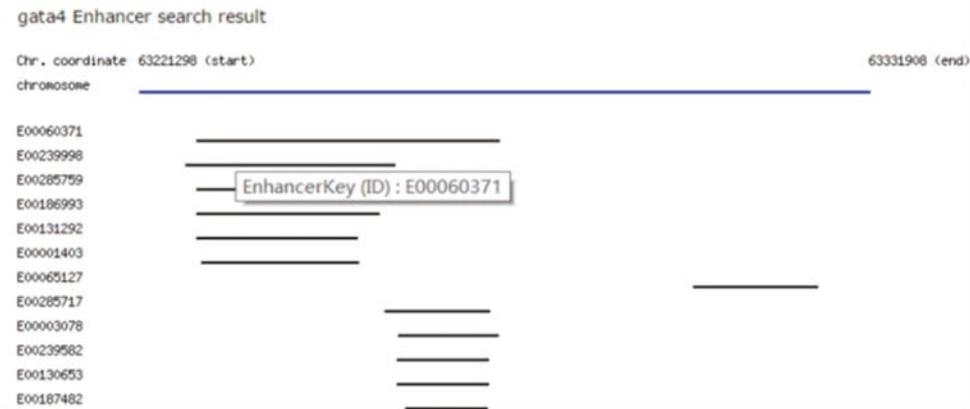
A

CardioTF Module: Cardiac Transcriptional Factor (TF) ▾

OPTIONS: Cardiac Gene Name ▾

QUERY: Gata4 submit

B



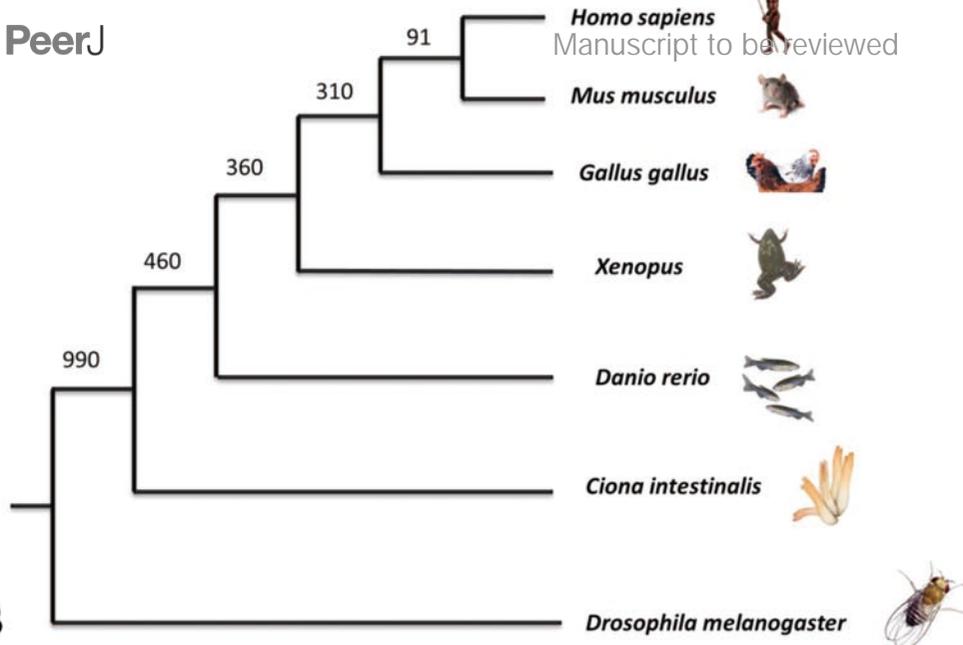
C

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

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.



B

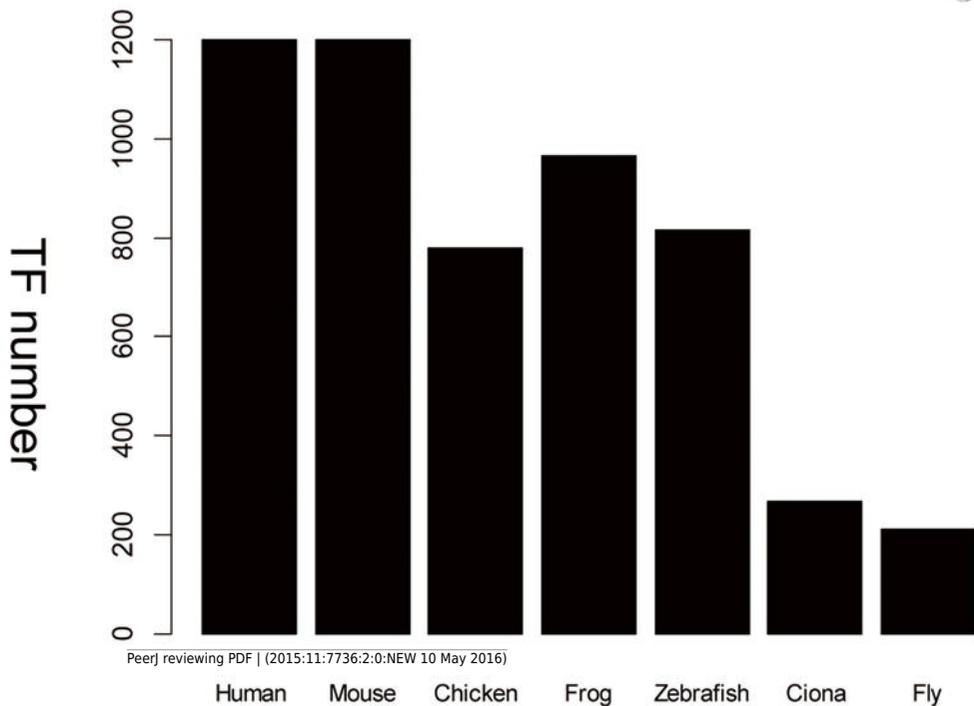


Figure 4(on next page)

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

Natural language
processing Module:
GNAT



Machine learning tool -
PerlModule: NaiveBayes

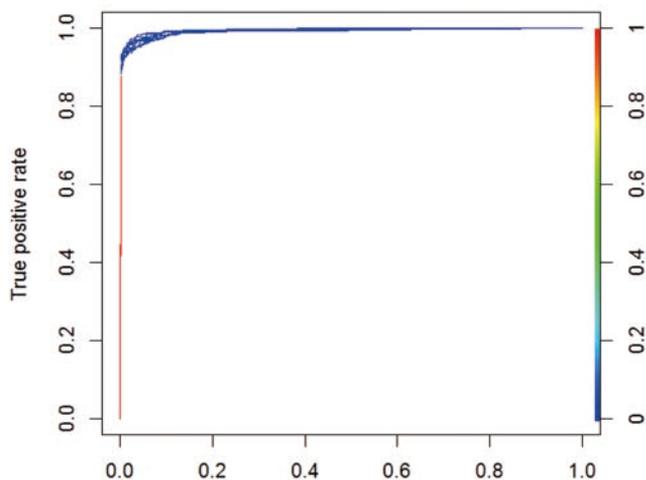


Pick up Weinstein style
Cardiovascular PubMed
abstracts



Gene symbol/Entrez GeneID

B



Sensitivity	0.9316
-------------	--------

Precision	0.9516
-----------	--------

F1 Score	0.9415
----------	--------