The transcription factor encyclopedia

Dimas, Y; Camenisch, G; Wenger, R H; Wasserman, Wyeth; et al

Abstract: Here we present the Transcription Factor Encyclopedia (TFe), a new web-based compendium of mini review articles on transcription factors (TFs) that is founded on the principles of open access and collaboration. Our consortium of over 100 researchers has collectively contributed over 130 mini review articles on pertinent human, mouse and rat TFs. Notable features of the TFe website include a high-quality PDF generator and web API for programmatic data retrieval. TFe aims to rapidly educate scientists about the TFs they encounter through the delivery of succinct summaries written and vetted by experts in the field. TFe is available at http://www.cisreg.ca/tfe.

DOI: https://doi.org/10.1186/gb-2012-13-3-r24

Originally published at:
Dimas, Y; Camenisch, G; Wenger, R H; Wasserman, Wyeth; et al (2012). The transcription factor encyclopedia. Genome Biology, 13(3):R24.
DOI: https://doi.org/10.1186/gb-2012-13-3-r24
The Transcription Factor Encyclopedia

Abstract

Here we present the Transcription Factor Encyclopedia (TFe), a new web-based compendium of mini review articles on transcription factors (TFs) that is founded on the principles of open access and collaboration. Our consortium of over 100 researchers has collectively contributed over 130 mini review articles on pertinent human, mouse and rat TFs. Notable features of the TFe website include a high-quality PDF generator and web API for programmatic data retrieval. TFe aims to rapidly educate scientists about the TFs they encounter through the delivery of succinct summaries written and vetted by experts in the field. TFe is available at http://www.cisreg.ca/tfe.
Background
As modulators of gene expression, transcription factors (TFs) act on all eukaryotic biochemical systems, driving ‘networks’ or ‘regulatory programs’ that define the developmental stages of life and maintain cells in dynamically changing microenvironments. From regulating muscle differentiation in embryonic development (MYOD) [1] to helping the kidneys reclaim water at times of dehydration (NR3C2) [2] and even instigate oncogenesis (MYC) [3], the pervasive roles of TFs are becoming increasingly appreciated and experimentally characterized. TFs are amongst the most highly studied class of proteins. Even though TFs comprise fewer than 5% of human protein-encoding genes [4,5], over 16% of gene-related papers address members of this critical class (Figure 1).

Increasingly, TFs are the focus of research aimed at deciphering the complex regulatory programs that allow a single genome to specify hundreds of phenotypically distinct cell types. The study of stem cell differentiation is dominated by efforts to understand how the activation of individual TFs can direct the progression to specific lineages. Perhaps the most important of these advances in recent years is the realization that, by introducing specific ‘sets’ of TFs into terminally differentiated cells, one can induce these cells to return to a pluripotent capacity [6,7]. A complete understanding of TFs and the processes that alter their activity is a fundamental goal of modern life science research.

Rapidly advancing knowledge in TFs is nearly impossible to track, with over 8,000 TF-related papers published in 2009 alone (Figure 1). In this light, the authors of this work believe that non-TF researchers are sometimes confront with the need to understand the properties of certain TFs that they come across within their research, as a potential participant in some differentiation, signaling or regulatory pathway they are studying. In this scenario, an accessible, high quality synopsis of the TF can catalyze rapid progress in the study, allowing researchers to chart an efficient approach. Such synopses have traditionally

Figure 1 New journal articles associated with human or mouse TFs. Over the past five years, 216,421 journal articles associated with human or mouse genes have been published and indexed in NCBI PubMed. This amount represents 5.59% of all articles published and indexed during the same time frame (3,871,190 articles). Out of the 216,421 articles associated with human or mouse genes, at least 34,943 are associated with human or mouse TFs, or 16.15%. This is astounding when considering that known TFs represent only 5% of the genome. The proportion of journal articles associated with TFs has also been rising steadily over the past five years - from 15.47% in 2005 to 16.81% in 2009. These figures were determined with a conservative set of approximately 3,200 human and mouse TF genes derived from the works of Fulton et al. [4] and Vaquerizas et al. [5] and the publicly available ‘gene2pubmed’ annotation from NCBI.
been obtained from published review articles, but the need for timely information about the growing pool of actively studied genes has increasingly led researchers to online information sources.

In the Internet Age, gene-specific resources have emerged that present information gathered from highly specialized biomedical databases. Examples of such resources include Entrez Gene [8] and GeneCards [9]. While automated content can be useful, many researchers seek summary descriptions of the proteins. The classic UniProt/Swiss-Prot [10,11] model for curated content is often viewed as a gold standard, while automated systems have emerged to extract key sentences from the research literature, such as iHOP [12] and WikiGenes [13]. The community participation model for maintaining current information exemplified by Wikipedia has arguably not been proven successful for small communities with specialized interests and need for peer-reviewed content, perhaps reflecting the limited time available from the small cadre of qualified experts. The Gene Wiki project within Wikipedia has been the most advanced effort, providing automated stub articles for many genes within the confines of Wikipedia [14]. However, the absence of a rigorous and enforced peer review process and the lack of oversight in monitoring contributor qualifications makes the model less than ideal for scientists who seek *bona fide* information in the digital realm.

TFs are proteins with special abilities and attributes not found in other classes of proteins. For example, they often work in pairs or networks to modulate specific regulatory pathways. They directly or indirectly bind to DNA. Some also interact with ligands or hormones. In short, the unique properties of TFs place special demands on - and presents opportunities for innovation with regards to - the kind of information TF-specific biomedical resources can offer, and how this information can be displayed to users such that it is intuitive, sensible, and helpful. There are many different kinds of TF-specific useful data that can be captured. Sequence-specific DNA binding TFs act on target genes, interact with other TFs to achieve specificity in action, and have structural characteristics that are predictive of DNA interaction mechanisms. A well-characterized TF will be represented by a binding profile that defines the target sequences to which it can bind. These class-specific properties have spurred the development of key databases, such as JASPAR [15], PAZAR [16] and TRANSFAC® [17]. These efforts, however, are constrained by limited capacity to identify and curate data from the scientific literature.

Based on the importance of TFs, the rapid accumulation of research advances in the scientific literature, and the need to provide class-specific information, we have created a new web-based platform called the Transcription Factor Encyclopedia (TFe). TFe’s mission is to facilitate the curation, evaluation, and dissemination of TF data. TFe espouses the principles of open access and promotes collaboration within the TF research community. It rewards scientists for contributing their data, and aims to optimize content quality ensuring expert editorship and multiple levels of peer review, both internal and external. TFe is curated and managed by the TFe consortium, a collaboration of over 100 TF researchers from throughout the world (see Figure 2 for the list of completed mini review articles that they contributed, and Figure 3 for their distribution by country). The objective of the TFe consortium is to produce concise mini review articles on pertinent human and mouse TFs, and to accelerate the curation of TF-specific data.

To date, the TFe consortium has prepared over 800 TF mini review articles, 136 of which are sufficiently complete to be presented here in the inaugural paper. Overall, the TFe database contains 184 original tables and diagrams, 221 TF binding site profiles, 3,083 non-redundant binding sequences, 2,334 genomic targets, 212 three-dimensional structural predictions, 6,308 protein-protein and protein-ligand interactions, 42,500 TF-to-disease predictions based on Medical Subject Headings (MeSH), and more.

The long-term goal of TFe is to create an online encyclopedic collection about well-studied TFs, combining a mixture of both expert-curated and automatically populated content.

**Resource content**

In this paper we present a collection of 136 mini review articles about human and mouse TFs. These articles are available on the TFe website [18]. Two versions of every article are available. A definitive version can be viewed online, while an abridged version can be downloaded in Portable Document Format (PDF) from the website. A sample PDF article is enclosed in Additional file 1, while Additional file 2 contains the raw data files.

The completed articles represent 15% of all TF articles that have been pre-populated with automated content in TFe. As for the remaining articles, most are awaiting an expert volunteer author or remain at a preliminary state of development. An ongoing effort aims to recruit appropriate authors to curate these ‘orphaned’ articles. In total, TFe currently hosts 803 TF articles, 216 of which are human, 585 of which are mouse, and 2 of which are rat. While TFs that bind directly to DNA are considered for inclusion in TFe at this time, a few contributed articles have addressed other TFs. Recent research has suggested that there are well over 1,300 TFs in the human genome [4,5]. With the increasing availability of data, our goal is to eventually characterize all TFs in the human and mouse genomes. See Additional file 3 for an inventory of all TF articles currently available in TFe alongside their classification, which is discussed in further detail below.
To ensure uniformity, all TF articles in TFe are written in a standardized format that was established in response to input and feedback from consortium members. The style emphasizes relatively short articles accompanied by a few figures and up to 75 references. These articles are concise, informative, and cater to a broad audience of life science researchers.
The article page (shown in Figure 4b) is the cornerstone of the TFe website, as it is where articles are accessed. Articles in TFe are organized into ten tabbed sections titled ‘Summary’, ‘Structure’, ‘TFBS’ (TF binding site), ‘Targets’, ‘Protein’, ‘Interactions’, ‘Genetics’, ‘Expression’, ‘Ontologies’, and ‘Papers’ (that is, references) (Figure 5). Above the tabs lies a standard header that displays pertinent information regarding the TF, including the TF symbol, species, classification, the date of the most recent revision, and an article completion score bar (Figure 5). Sections generally contain a mixture of author-curated and automatically populated content, typically in the form of an expert-written overview text - the author-curated portion - followed by several additional headings filled with a mixture of author-curated and automatically populated content. See Figure 6 for a comprehensive list of all automatically populated and manually curated content available in the article page. The automatically populated content represents data that we have incorporated into TFe from second and third party resources, including: BioGRID [19], Ensembl [20], Entrez Gene [8], Gene Ontology [21], MeSH [22], the Mouse Genome Database [23], Online Mendelian Inheritance in Man (OMIM) [24], PAZAR [25], RCSBP Protein Data Bank [26], the UCSC Genome Browser [27] and the Allen Brain Atlas [28]. More details on the software tools and data repositories utilized in the generation of automatically populated content found in each tab are presented in Table 1.

Each section - with the exceptions of the ‘Ontologies’ and ‘Papers’ sections - begins with a brief, expert-written summary statement from the authors followed by relevant figures, lists, and tables. For instance, the ‘Summary’ section is designed to begin with a 500-word (maximum) overview followed by one or two captioned figures. The ‘Targets’ section contains a 200-word overview focusing on the TF’s regulatory role, followed by a table of genomic targets populated by the author and additional data automatically extracted from PAZAR. The expert-written summaries in TFe are meant to provide the reader with some perspective, highlight key points, and reveal tacit knowledge. For a complete list of features available in each section, please see Additional file 4.

Here we discuss each of the ten tabbed sections - ‘Summary’, ‘Structure’, ‘TFBS’, ‘Targets’, ‘Protein’, ‘Interactions’, ‘Genetics’, ‘Expression’, ‘Ontologies’, and ‘Papers’ - in greater detail.

**Summary tab**

The ‘Summary’ tab presents insightful overview text written by expert authors, one or more figures as supplied by them, and a list of relevant references. Authors also have the option to post noteworthy links - for instance, to a Wikipedia entry for the TF.

Like every other tab, the ‘Summary’ tab user interface is a content viewer and editor combined into one. When expert authors wish to implement changes to their articles, they may ‘sign in’ to TFe using their personalized user accounts. After this is done, they are able to see the normally hidden editing interface that allows them to upload text, figures, figure captions, references, external links, and data, depending on the tab. The
The editing interface supports the widely used wiki syntax to allow basic text formatting, such as bolding, italicizing, underlining, and the creation of bulleted and numbered lists. All text entered in wiki syntax is converted to HTML by a local installation of the MediaWiki software. Authors also have the option to add PubMed references.
anywhere in their text by using special tags that look like `(pmid:16371163)` - without the quotes. These tags are automatically converted to a proper citation (Van-couver style) by the TFe software. Figures can be uploaded in many different image formats, while figure captions are submitted as text. PubMed citations are also supported in figure caption text.

**Structure tab**

The ‘Structure’ tab contains author-provided overview text regarding the structural properties of the TF, followed by - if available - the predicted three-dimensional structure of the TF’s DNA binding domain. These ‘structural predictions’, which were created by the consortium using a custom-made pipeline, are available for download as both high-resolution Portable Network Graphics (PNG) images and Protein Data Bank (PDB) formatted files. The materials and methods used in their construction are discussed in the Materials and methods section of this paper.

**TFBS tab**

A key property of TFs is the DNA sequences to which they bind. In the world of TF research, such DNA sequences are often called ‘transcription factor binding sites’, or ‘TFBS’ for short. Knowledge of TFBS patterns is key to identifying putative binding sites in genomic sequences and to the identification of sets of genes regulated by the TF in promoter analysis.

In light of this, disseminating TFBS data is a crucial part of TFe’s mission. The ‘TFBS’ tab contains a summary of the DNA binding characteristics of the TF, alongside one or more DNA binding target site data, when sufficient data are available. A graphical depiction of the target site pattern is displayed in the form of a sequence logo, along with a brief summary text from the author. This information is extracted from the PAZAR regulatory sequence database.

TFe authors are able to create new binding models by inputting a list of binding sites, experimental evidence, and references in the ‘TFBS’ tab through a submission interface that is visible to authors only. It is possible for authors to submit target sequences that exist in a genome, or artificial sites, such as those generated in a SELEX (Systematic Evolution of Ligands by Exponential Enrichment) experiment. When we receive a submission through this TFBS form system, we forward the supplied...
information to a team of curators who review the information for errors and, if appropriate, deposit the annotation into the PAZAR database. Because PAZAR and TFe are programmatically linked, the annotation deposited in PAZAR will also appear in TFe.

**Targets tab**

Related to the ‘TFBS’ tab, the ‘Targets’ tab presents users with an introductory text followed by a list of genes directly regulated by the subject TF sourced from the PAZAR database. At a minimum, the ‘Targets’ list recapitulates the information in the ‘TFBS’ tab, but oftentimes, expert authors provide additional genes known to be regulated by the TF but for which the specific DNA target sequence is unknown. Authors can add additional targets by using a specialized editing interface that is accessible upon sign in.

**Protein tab**

The ‘Protein’ tab presents information about the functional consequences of protein modifications or distinctions between protein isoforms. Authors summarize
such information in free text entries. As a late addition to the system identified as a need during the beta-testing process, the section has yet to be populated for many entries.

**Interactions tab**

Interactions between TFs and ligands or proteins are reported in this tab. While automated content from the BioGRID database is included, authors may also provide information about additional interactions not reported in the external system through a specialized submission interface. Authors have a limited set of interaction types (Table 2) from which to pick labels. If the gene encoding the TF is subject to transcriptional regulation in a selective manner, the regulating TFs are reported in this section.

**Genetics tab**

TFs perform powerful genetic roles in the development and physiology of organisms. Therefore, the genetic properties of TFs can have powerful consequences upon the phenotype of an organism. The ‘Genetics’ tab presents two sets of data linking TFs to phenotype, in addition to the prerequisite expert-written summary. The first is a ‘cloud’ of TF-to-disease associations composed with MeSH terms. The second set of data linking TFs to phenotype is a list of Mouse Genome Database mammalian phenotype terms associated with the mouse homolog of the TF protein.

**Expression tab**

The ‘Expression’ tab reports expression data from the GNF Expression Atlas, sourced from the UCSC Genome Browser, and observed regional expression in the brain according to the Allen Brain Atlas. Authors are encouraged to provide a text description of known expression properties of the TF gene.

**Ontologies tab**

Annotated characteristics of the TF are reported in the ‘Ontologies’ tab. Gene Ontology terms linked to the gene are extracted from Entrez Gene for display. The
Table 2 List of predefined interaction types

| Interaction type                                           | Gene | Ligand |
|------------------------------------------------------------|------|--------|
| Acts on upstream signaling pathway                         |      |        |
| Competitive inhibition                                     |      |        |
| Genetic                                                    |      |        |
| Indirect                                                   |      |        |
| Multimerization                                            |      |        |
| Not specified                                              |      |        |
| Physical: deacetylation                                    |      |        |
| Physical: dephosphorylation                                |      |        |
| Physical: desumoylation                                    |      |        |
| Physical: deubiquitination                                 |      |        |
| Physical: enzyme modification: acetylation                 |      |        |
| Physical: enzyme modification: methylation                 |      |        |
| Physical: enzyme modification: phosphorylation             |      |        |
| Physical: enzyme modification: protein cleavage             |      |        |
| Physical: enzyme modification: sumoylation                 |      |        |
| Physical: indirect altering posttranslational modifications |      |        |
| Physical: sequestering                                     |      |        |
| Physical: translocation                                    |      |        |
| Physical: ubiquitination                                   |      |        |
| Physical: undefined direct interaction                     |      |        |
| Physical: with another TF                                  |      |        |
| Physical: with another TF: complex binds DNA               |      |        |
| Physical: with co-activator affecting recruitment          |      |        |
| Physical: with co-repressor affecting recruitment          |      |        |
| Regulatory: decreases expression of this TF                |      |        |
| Regulatory: increases expression of this TF                |      |        |
| Unknown                                                    |      |        |

When a protein-to-protein or protein-to-ligand is added to a TFe article either by the author or automatically from BioGRID, the author of that article has the option to define the interaction type if this information is known. To promote a standardized vocabulary for describing interaction types, a list of possible interaction types between proteins and ligands is provided to the author. This list continues to be adjusted and expanded based on need and author feedback.

automatically populated TF-to-MeSH associations for all MeSH terms outside of diseases are reported, following the same procedure as introduced in MeSHOP.

Papers tab

The 'Papers' tab provides a set of recommended articles pertinent to the TF. Authors indicate the most useful introductory readings and other key papers with a two circle rating system. Two full circles indicate an excellent paper in the author’s opinion, while no circles still indicate a very good and noteworthy paper.

System features

In this section of the paper, we discuss the important features of our platform. These features include: (1) our system of classifying TFs; (2) our concept of 'content inheritance', or how articles of very closely related TFs may derive content from each other when biologically appropriate; (3) our structural prediction system; (4) our data on TF binding sites; (5) our TF-to-disease association predictions; (6) our PDF rendering system; (7) TFe's data export capabilities; and (8) the article completion score.

Classification of transcription factors

The classification of TFs into 'groups', 'families' and 'subfamilies' is a very important feature of TFe. Over the past few years, there have been efforts to identify and classify all TFs within the human and mouse genomes [4,5]. While there are potentially several different strategies for classifying TFs, one promising approach is to group them based on DNA binding domain structures. Building upon the work of Fulton et al. in the Transcription Factor Catalog (TFCat) project [4], we have organized all TFs in TFe into various groups, families and subfamilies as previously mentioned (Additional file 3).

Content inheritance

When comparing orthologous TFs, or recently evolved paralogs within a species, it is commonly observed that homologous TFs are well conserved structurally and functionally [29]. Indeed, some homologous TFs are so well conserved that there is often no information that distinguishes the homologs. However, in TFe we have opted to create separate articles for all TFs, including homologous TFs. For instance, we have several articles for the TF NFE2L2 - one each for human, mouse, and rat. In doing so, we aim to provide maximum flexibility to our authors who may wish to discuss key subtle differences between closely related proteins.

The drawback to this approach is that, in some cases, we end up with multiple articles for what could be considered as functionally synonymous TFs. These TFs, due to their extreme likeness, would inevitably share common attributes such as binding site profiles, interactors, and target genes. In this situation, it becomes important to keep all shared attributes current and synchronized across the different articles. To assist with this information management process, we implemented a content inheritance system that enables authors to define small clusters of homologous TFs for which certain data may be automatically shared. Under this system, the article that is more annotated - the 'parent article' - donates text, figures, and data as appropriate to the article that is less annotated, the 'child article'. However, authors and editors are able to override the automatic sharing of data when it is not reflective of the underlying biology.

Structural predictions

We have developed a custom computational pipeline for predicting the three-dimensional protein structures of the DNA binding domains of TFs. The final output of our pipeline is a PDB formatted file of the predicted structure, alongside a short segment of double-stranded
DNA for positional reference. The DNA molecules are stylistic and do not represent particular sequences such as the consensus sequence for the TF. We have generated standardized PNG image renderings of these PDB files for web and print purposes. Figure 7 contains a representative sample of the structural predictions, one from each family of TFs featured in our first release. To date, 212 structural predictions have been generated, with the emphasis of effort focused on TFs with articles that are nearing completion. All structural predictions are available for download in PDB format under the 'Structure' section in the articles of their respective TFs. A brief summary of the materials and methods used in our protocol can be found in the Materials and methods section below.

Transcription factor binding site data

One of the goals of TFe is to encourage experts to assist in the curation of TFBS sequences and generation of binding profiles. Working in partnership with PAZAR [25], an open source and open access TF and regulatory sequence annotation database, our consortium gains access to a powerful curation platform with which it can store, annotate, and manage data, as well as retrieve additional data from other projects in PAZAR. Our initial collection of 100 reviews collectively contain 3,083 unique binding site sequences from the PAZAR database, of which a total of 452 sequences have been donated to PAZAR by the consortium. From this set of binding site sequences, we have generated 221 binding models and extracted 1,436 genomic targets for 199 different TFs. In addition, 898 genomic targets have been entered manually by our authors to supplement this genomic target dataset. See Additional file 2 to see the binding data of released articles and Additional file 5 for key binding profiles that have been generated in the TFe project.
Disease associations

Many TFs are implicated in disease. Out of a growing list of 1,321 human TFs we compiled from the work of Vaquerizas et al. [5] and Fulton et al. [4], 197 are currently linked to one or more diseases in the OMIM Morbid Map [30]. In light of the strong connection between TFs and disease, we have predicted 42,500 TF-to-disease associations. This was done by using the ‘Entrez Gene to PubMed’ (‘gene2pubmed’) and MeSH datasets that are available at the National Center for Biotechnology Information (NCBI). With mainly these two datasets, with additional datasets such as OMIM and GeneRIF to further strengthen our predictions, we developed a protocol that makes the connection between TF-encoding genes, papers that discuss these genes, and the MeSH terms that are tagged to the papers. By indirectly mapping disease-oriented MeSH terms to TF-encoding gene identifiers, we are able generate a list of MeSH terms that are associated with each TF. Statistical analysis is applied to the raw connections to determine their strength - mainly by reflecting the frequency of TF-term co-occurrence in light of the number of papers that refer to either the TF or the term. This information can be viewed as a table or as a ‘cloud’ under the ‘Genetics’ tab.

PDF rendering

We have built a PDF rendering engine in TFe that transforms articles into condensed, four-page PDF ‘mini summaries’ available for printing (Figure 8). These summaries can be downloaded by clicking on the ‘Download article (PDF)’ link that is prominently displayed on all article pages on the TFe website. We have included a sample in Additional file 1.

While the articles as they appear on the TFe website permit great flexibility in terms of length and variety of content, the PDF format is more structured and compact. Thus, the PDF version of the articles can be described as the ‘abridged’ form of the article. When necessary, we are keen to remind users that there is additional content on the TFe website that cannot be incorporated into the abbreviated PDF article.

In our effort to encourage authors to write more balanced articles that fulfill the prescribed style, we ration the available space for each section. For instance, one third of the last page is strictly allocated to the Genetics and Expressions paragraphs. If an author chooses not to comment on those sections, that space will remain blank - to motivate authors to do something about it. Conversely, if the author provides more text than allowed, the surplus text will simply be trimmed to the nearest sentence.

The PDF feature was created to produce an article format that more closely resembles a ‘journal paper’, with pleasant typesetting and pagination. Indeed, many open source journals that publish exclusively online still invest significant resources to generate definitive PDF copies for all of their articles, even when HTML versions are adequate for practical purposes. We envision that for some users, once a TF has significantly piqued their interest for further perusal, they would be inclined to review the web version to access the most complete and up-to-date information.

Behind the scenes, our PDF rendering engine is based on in-house code and the dompdf 0.5.1 open source module. It uses fuzzy logic to handle the modifications necessary to determine the best solution of text, images, captions, and data tables to make the page layouts as aesthetically pleasing as ‘machinely’ possible. These modifications include changing the sizes of the figures, truncating excess text, reformatting the references, and calculating trade-offs between having larger figures and more data in data tables at the expense of less text, or keeping more text at the expense of having fewer figures and sparser data tables.

Data export

One of the goals of TFe is to make TF data easily accessible to all. To support this goal, we built a web-based application programming interface (API) to facilitate a straightforward approach for extracting data from the TFe website. In addition to the TFe API, we have built a spreadsheet generator that allows visitors to download Excel (.xls) formatted files containing all of the information that is available through the web API, as a service for users who are not inclined to use the programmer-oriented web API. In short, virtually all forms of data available in TFe, including binding sequences, genomic targets, interactors, key papers, and even ontology terms, can be downloaded through the API, the spreadsheet generator, or PDF renderer. The TFe web-based API and its accompanying documentation can be found on the TFe website.

The presence of a machine-friendly API is what sets TFe apart from most other biomedical wikis. For easy parsing, the API sends data in tab-delimited plain text format. Since the API is web-based and communicates through the ubiquitous HTTP protocol, it is compatible with all common scripting and programming languages, including PHP Hypertext Preprocessor (PHP), Perl, and Python. See Figure 9 for an illustration of how the data retrieval process works when using the TFe web API.

Article completion score

An article completion score (ACS) is automatically computed for every article in TFe. The ACS can range from 0% to 100%. Its purpose is to reflect the depth of annotation present in the article - the article’s level of
completeness. To illustrate, nearly complete articles typically have an ACS of 90% or more. The ACS is prominently displayed in the header of all articles on the TFe website in the form of a ‘progress bar’ that changes from orange to green as the score approaches 100% (Figure 5). The ACS system is designed to help authors determine whether their articles are sufficiently complete, and more importantly, identify article sections that are in need of more attention. By clicking on the ‘see what’s missing’ link to the right of the progress bar, authors can view a list of suggestions that they can undertake to increase the score of their article, such as ‘please provide more information in the Overview section of the Summary tab.’

The ACS evaluates the completeness of TFe articles based on several factors, which include, among others, the amount of text, figures, references, and data contributed for each article. Overall, 19 attributes (Table 3) are taken into account in the computation of the ACS.

The ACS was implemented during TFe’s beta testing period, when we observed that authors need guidance as to the expected level of article content. Prior to the implementation of the ACS, a large majority of ‘completed articles’ were deemed substantially incomplete. The ACS metric has established a standard for author contribution and helps authors attain this standard by highlighting sections of deficient articles that require further attention, and notifies authors how the deficiency can be remedied. While developed de novo, subsequent feedback indicates that the progress tracking scores are reminiscent of content tracking scores utilized in the LinkedIn social networking system. The response to the ACS has been positive. Within six months of implementation, the completion scores of all articles increased from 40.6% to 60.2% (Figure 10). The scoring metrics for computing the ACS are presented in Table 3.

User authentication
The user authentication system of TFe, which handles the ‘sign in’ and ‘sign out’ functions, is built upon the Perl CGI::Session module. All account passwords stored in the TFe database are encrypted to safeguard the privacy and security of TFe users.

Software
In this section, we discuss in technical detail the software that runs the TFe website, mainly its user interface and system architecture.
Figure 9 Using the TFe web API. Adventures in bioinformatics often involve large amounts of data retrieval and computation not amenable to manual labor. Thus, in place of humans, software is written to automate the grunt work, which may include computing vast quantities of data or obtaining large amounts of information from resources in the cloud, such as NCBI. To give researchers the option to retrieve data from TFe in an automated fashion, we have implemented a simple yet powerful web API. This figure provides a summary of what a data transaction may look like when using the TFe web API. In this case, the goal of the data retrieval exercise is to obtain all MeSH disease terms associated with the transcription factor ‘ATF3’.

1. Bioinformatician writes a small "script" program to automatically retrieve the desired information from the TFe Web API.

2. The program is set to run on its own.

Get all TFe TFIDs
For each TFe TFID,
Get the TFID’s symbol
If symbol equals "ATF3",
Get MeSH disease term

Get all TFe TFIDs
http://www.cisreg.ca/cgi-bin/tfe/api.pl?code=all-tfids

For each TFe TFID,
Get symbol
http://www.cisreg.ca/cgi-bin/tfe/api.pl?code=symbol&tfid=118
http://www.cisreg.ca/cgi-bin/tfe/api.pl?code=symbol&tfid=119
http://www.cisreg.ca/cgi-bin/tfe/api.pl?code=symbol&tfid=120

... NFE2L2
ATF2
ATF3

If symbol equals "ATF3",
Get MeSH disease term
http://www.cisreg.ca/cgi-bin/tfe/api.pl?code=mesh-disease&tfid=120

Adenocarcinoma 0.17760114421424 467 2 63 224213 20...
Amino Acid Metabolism, Inborn Errors 0.0614416365875486 46...
Anoxia 0.0113406899391781 467 2 63 46934 2010-06-01
Avitaminosis 0.156084229322004 467 1 63 49774 2010-06...
Blood Protein Disorders 0.160228839379989 467 1 63 5121...
The TFe software is a database-driven website application that runs the TFe website. For end users, the TFe website is an information resource where researchers can read peer-reviewed, expert-written summaries on pertinent TFs as well as obtain a wide variety of TF-related data, including binding sequences, genomic targets, and TF-to-disease associations. See Figure 6 for a complete list of all types of information available on TFe.

Table 3 Computing the TFe Article Completion Score

| Tab                  | Scoring element                  | Target                                      | Points | Weight  |
|----------------------|----------------------------------|----------------------------------------------|--------|---------|
| Summary              | Overview text                     | 500 words                                    | 10 points | 8.33%  |
| Summary              | References in overview text       | 3 references                                 | 5 points | 4.17%  |
| Summary              | Figures                           | 1 figure                                     | 10 points | 8.33%  |
| Structure            | Overview text                     | 200 words                                    | 5 points | 4.17%  |
| Structure            | TFBS Binding site profiles        | 1 binding site profile                       | 10 points | 8.33%  |
| Targets              | Overview text                     | 200 words                                    | 5 points | 4.17%  |
| Targets              | Targets                           | 10 targets in total (both author and auto)   | 10 points | 8.33%  |
| Protein              | Isoforms text                     | 200 words                                    | 5 points | 4.17%  |
| Protein              | Covalent modifications text       | 200 words                                    | 5 points | 4.17%  |
| Interactions         | Overview text                     | 200 words                                    | 5 points | 4.17%  |
| Interactions         | Ligands                           | 1 ligand                                     | 1 point  | 0.83%  |
| Interactions         | Interactions                      | 10 interactors in total (both author and auto) | 10 points | 8.33%  |
| Interactions         | Interactions                      | All ‘nature of interaction’ fields annotated | 10 points | 8.33%  |
| Genetics             | Overview text                     | 250 words                                    | 5 points | 4.17%  |
| Expression           | Overview text                     | 200 words                                    | 5 points | 4.17%  |
| Papers               | Papers                            | 15 papers                                    | 10 points | 8.33%  |
| Papers               | Papers                            | 3 papers marked as ‘recommended’            | 3 points | 2.50%  |
| (all)                | Links                             | 1 link                                       | 1 point  | 0.83%  |

120 points 100%

The TFe Article Completion Score (ACS) is based on the 19 components listed here. Authors earn points for completing each component, up to the prescribed maximum for that component (‘Maximum points’). So, for example, an author would be granted one point for adding one link, but no additional points would be granted if the author adds a second or third link. This prevents the author from adding 24 links to boost their ACS score by 20%. A ‘fully complete’ article would net 120 points, which gives a score of 100%.

Overview

The TFe software is a database-driven website application that runs the TFe website. For end users, the TFe website is an information resource where researchers can read peer-reviewed, expert-written summaries on pertinent TFs as well as obtain a wide variety of TF-related data, including binding sequences, genomic targets, and TF-to-disease associations. See Figure 6 for a complete list of all types of information available on TFe.

As previously discussed, the TFe website also features a password-protected user interface that allows expert authors to create and edit TF articles, upload data, report technical problems (that is, bugs), and submit anonymous peer reviews of other articles. It also features a built-in Customer Relationship Management (CRM)-like tool to help the administrators recruit new authors, as well as manage and communicate with the rest of the consortium. In short, the TFe website is a specialized and integrated software platform that has

Figure 10 The completion scores of authored articles in TFe. The y-axis of this graph is the article completion score (ACS), while the bars on the x-axis represent the 176 authored TF articles in TFe (some of which are still works-in-progress), ordered such that higher scoring articles are positioned on the right (for clarity). In this graph, the completion scores of the 176 articles from three different periods - Q2 2009, Q4 2009, and Q2 2010 - are superimposed to demonstrate that the scores have been increasing over time. Within six months of the implementation of the ACS system in Q2 2009, the completion scores of authored TFe articles have increased from 40.6% to 60.2%, thus attesting to the effectiveness of this feedback mechanism (see Q2 2009 versus Q4 2009).
been custom-built to facilitate a community-curated TF wiki project.

User interface
The TFe website, which can be accessed at [18], features a familiar and streamlined graphical user interface that is written in Extensible Hypertext Markup Language (XHTML) 1.0 Transitional, Cascading Style Sheets (CSS), and JavaScript.

On the homepage, a large ‘universal’ search box dominates the center of the screen (Figure 4a). This search box allows users to quickly access TFe’s built-in search engine, which accepts 18 different types of queries, including gene symbols, fragments of binding sequences, and the names of researchers who are associated with particular TFs through their publication records. Alternatively, visitors can click on the ‘go to a random article’ link to view a random article on the article page.

Displayed in Figure 4b and as previously discussed, the article page is the centerpiece of the TFe-user interaction as it is where the bulk of TFe content lies. It features a compact yet informative and graphically rich header with key pieces of information about the TF, followed by the described ACS indicating the TF article’s level of completeness or ‘depth’. Below the ACS, the contents of the article are divided into ten tabs labeled ‘Summary’, ‘Structure’, ‘TFBS’, ‘Targets’, ‘Protein’, ‘Interactions’, ‘Genetics’, ‘Expression’, ‘Ontologies’, and ‘Papers’. A row of navigation links is placed unobtrusively on the left side of the page. Other noteworthy pages on the TFe website include the classification page (Figure 4c) and the browse page. The classification page presents an organized hierarchy of TFs based on the TFCat [4] and the extended TF classification system of Vaquerizas et al. [5]. The browse page allows users to browse for TF articles based on various attributes such as name, classification, and level of completeness.

System architecture
The TFe website software is written almost entirely in the Perl programming language, using the ‘LAMP’ (Linux, Apache, MySQL, Perl/PHP) paradigm for developing web-based applications. The Perl programming language was chosen for its robust text-manipulation capabilities and widespread support within the bioinformatics research community. In developing the website software, we have incorporated Perl and PHP modules and software packages to handle specialized tasks - such as reading the database, generating PDF files, and resizing images. See Table 4 for a list of Perl and PHP modules and software packages incorporated into the TFe software.

The TFe website software is designed to run quickly and efficiently, yet remain relatively simple for programmers and system administrators to maintain. One challenge we had to overcome during the development process was keeping the software fast and responsive despite its size and complexity. One solution was to purposefully fragment the TFe website software into over 40 independent components. Each component serves a single unique purpose - for instance, to generate the homepage, or to search the database, or to display articles. Each component can be summoned separately and without disturbing the other components. This fragmentation allows us to improve the speed and responsiveness of the TFe website, as at any given time only a fraction of the entire TFe software is being executed by the server.

To reduce code repetition, we placed shared functions such as those that generate the page header or navigation links - in a shared module that can be summoned by any component as needed. We call this module the ‘TFe core module’ because it forms the nucleus of the TFe website software. To further increase speed, we divided this TFe core module into three separate components: (1) a component that contains the vast majority of shared functions called ‘tfe.pm’; (2) a component that contains only those functions involved with database reads and writes called ‘db.pm’; and (3) a component that deals with maintenance and update functions called ‘update.pm’. See Figure 11 for a schematic representation of the TFe website software.

With regards to hardware architecture, the TFe website software is currently implemented in Linux-based (CentOS) environment using a dedicated virtual server. The TFe software stores data in both the UNIX file system (that is, for images and PDF files) and a MySQL database, both of which are physically located in a

| Table 4 Perl and PHP modules used in TFe |
|-----------------------------------------|
| Language | Module | Purpose |
|----------|--------|---------|
| Perl     | CGI    | Web browser interface |
| Perl     | CGI::Session | User login |
| Perl     | Crypt::Blowfish | Data encryption and random string generation |
| Perl     | DBI     | MySQL database interface |
| Perl     | GD::Image | Creation of TF binding site diagrams |
| Perl     | HTML::Detoxifier | User input filtering |
| Perl     | Image::Resize | Image resizing and formatting |
| Perl     | LWP::Simple | Interface between TFe and web-based APIs |
| Perl     | pazar   | Data retrieval from PAZAR |
| Perl     | pazar::gene | Data retrieval from PAZAR |
| Perl     | pazar::reg_seq | Data retrieval from PAZAR |
| Perl     | TFBS::PatternGen:MEME | Creation of TF position weight matrices |
| Perl     | Image::Resize | Image resizing and formatting |
| PHP      | dompdf 0.5.1 | PDF generation |

Listed on this table are the second and third party modules incorporated into TFe, with their respective programming languages, usage in TFe, and current web addresses at the time of publication.
proximal storage area network. To share computational load and optimize service responsiveness, a dedicated database server executes all complex database queries. This database server is connected to both the primary TFe web server and storage area network via fiber optic.

**Development process**

The TFe system was developed over a period of three years. Early prototypes in 2007 were subjected to intense testing and continuous refinement by the programming team during what we refer to as the ‘pre-alpha stage’. In the 2008 ‘alpha stage’, external quality control testing was initiated by inviting ten authors to provide feedback on the software’s design, features, and usability. By 2009, the software had evolved to a more stable and mature form. At this ‘beta stage’, we invited over 100 TF experts from around the world to contribute articles.

---

**Figure 11 Software architecture** This schematic demonstrates the conceptual structure of the TFe software. Written mainly in the Perl programming language, the software is essentially a collection of Perl ‘scripts’ that runs on an Apache web server, in a UNIX-compatible environment. The software relies on MySQL for data storage, and a number of third party modules. Over 40 ‘front line’ scripts (shown as the red rectangle) generate individual pages such as the home page and article page. These front line scripts are backed by a cluster of three TFe Perl modules (shown as the green circles): (1) the ‘database updater’, which is summoned pro re nata whenever the TFe database needs to be maintained or updated with new content from external sources such as NCBI; (2) the ‘main module’, which contains shared subroutines such as those that generate page headers; and (3) the ‘database handler’, which forms the gateway between all components of the TFe software and the TFe database. The database (shown as the yellow cylinder) is stored on a separate database server and communicates with the rest of the TFe software via fiber optic. It contains cached copies of third party resources so that the TFe software does not have to constantly retrieve data from the ‘cloud’. This optimizes performance. The web API (shown as the purple rectangle) is directly connected to the ultra small and efficient database handler module. In bypassing activation of the large main module and database updater, the web API is able to run faster than the web-based interface. GO, Gene Ontology; MGI, Mouse Genome Informatics.

---

Yusuf et al. Genome Biology 2012, 13:R24
http://genomebiology.com/2012/13/3/R24
Page 17 of 25
Over the next six months, TF experts responded to our invitations and began producing articles. To cope with the influx of feedback, we implemented an online feedback form. We upgraded our bug tracking process by adopting MantisBT, a web-based system that is available at [31]. All feedback was reviewed and prioritized for system modification if justified. Small changes were addressed immediately.

A rigorous backup regimen occurs on a daily and weekly basis to help us quickly and fully recover in the event of catastrophic system failure.

**Discussion**

Three prominent systems have been introduced that rely more heavily on the community-contributed content wiki model. These are: (1) WikiProteins [32]; (2) WikiGenes [13]; and (3) Gene Wiki [14]. WikiProteins uses automated procedures to extract information from multiple resources, a text-based procedure to summarize these data, and a wiki-based format to collect user-supplied information. Similarly, WikiGenes uses a text-based procedure based on the iHOP service to present automated content organized under categorized subjects, and users are encouraged to provide content and corrections to the system, with their identities displayed to acknowledge contributions. Gene Wiki, the product of which resides within Wikipedia, automates the creation and maintenance of ‘stub’ articles on genes, thus creating a systematic framework for gene content. Despite the quality of these systems, examples of deep community commitment to contribute content are rare. By visual inspection, most entries in these systems still contain mostly automated content.

A striking divergence from the classic model is GeneTests [33], in which expert authors are recruited for each subject gene, taking intellectual ownership of an article of substantial importance to the clinical genetics community. When contrasting GeneTests to the aforementioned wiki-based systems, two qualities contribute prominently to the success of the former. First, GeneTests addresses a niche, allowing content to be tailored to the needs of a target audience. Second, the scientists who write articles on GeneTests are strongly acknowledged, allowing them to receive recognition for their intellectual contributions. While lasting participation in - and the continuing evolution of - GeneTests may ultimately derive from the intense commitment of the project’s directors, it stands out as one of the rare cases in which prominent genetics researchers contribute original content to a community resource.

TFe represents a new direction in scientific communication of gene-specific information. Combining automated data presentation with expert-user reviews, the wiki-based system provides succinct reports about TFs, one of the most highly studied classes of proteins. The highly engaged efforts by researchers worldwide demonstrate that a wiki-based system can attract active participation and meet high quality standards of scholarly content. With over 100 mini reviews presented in the initial release, TFe represents one of the largest community participations in a gene-focused wiki project.

While the term wiki has become loosely applied over the years, in reality the term refers to a specific class of software that allows shared development of a document. However, in its most basic sense the term is commonly used to reflect the philosophy that information is best made accessible and editable by anyone pro bono. The wiki model has caught the attention of some scientists, who see it as a powerful tool that can hasten the pace of scientific communication. In the wake of Wikipedia’s success, there emerged a high profile rallying call to create a gene-function wiki for scientists [34], and several groups have heeded this call by creating various scientific wikis, some built from the ground up [13] and some derived from existing general purpose wiki engines [32,35].

Unfortunately, as evidenced by WikiProteins [32], WikiGenes [13], and to a lesser extent Gene Wiki [14], scientific wikis have generally struggled to attract the level of community involvement envisioned by their founders. There are several contributing influences for the observed low rate of participation. The success of Wikipedia is in part attributable to the enthusiasm of a tiny fraction of the large global community of Internet users who are willing to contribute content. The scientific community with expertise on a specific topic, on the other hand, is small. Thus, even if the participation rate among these scientists remains comparable to the participation rate of the global community of Internet users who contribute content to Wikipedia, there would still be far fewer scientists contributing. To make matters worse for proponents of scientific wikis, scientists seem generally less willing to participate in these sorts of endeavors than the average user, reflecting perhaps the enormous demands on their time or the relative age of the experts. For many, their limited time is dedicated to rewarding tasks, such as performing experiments and reporting on the results in peer-reviewed journals. Earning new publications appears to be a strong motivator for many scientists. Few are willing to spend the same amount of time and effort to expand a wiki article that resides in the public domain and from which they would not receive any substantial credit.

Recognizing these constraints, a critical component of the success of TFe is the provision for authorship credit. Furthermore, we strive to actively identify and recruit authors, as opposed to waiting for contributors to contact us. Without addressing these two aspects, we doubt
that we would be able to attain the same level of community involvement. Ultimately, the support of a journal willing to publish the resulting mini reviews in the form of this article (subject to passing a peer-review process) was a key motivator for many authors to participate in the project.

The retention of peer-review within the wiki-based article development process is scientifically critical. Readers of the system must hold high confidence in the quality of the reports. To meet this standard, all participating authors were encouraged to provide anonymous peer review reports for a set of articles. Approximately 40% of TFe authors participated in this voluntary peer review program as peer reviewers of other TFe articles.

Author identification was a challenge. We initially sought participation from existing collaborators and subsequently from peer referrals. During this early part of the project, we were able to recruit a core team of about ten authors who also became our de facto ‘alpha testers’, thus allowing us to incorporate user feedback during the application development process. These authors - and eventually other authors as well - had significant input into the TFe system throughout its formation.

Given the large number of characterized TFs, we ultimately needed a larger-scale approach. To this end, we identified researchers who frequently appeared as the senior author in publications that discuss a specific human or mouse TF (using an automated analysis of articles in PubMed). Overall, 251 authors were individually contacted via email. About 59% (149 authors) agreed to participate, in addition to 10 authors who were directly invited at the outset of the project, and 2 authors who expressed interest and joined without invitation. About 65% of the participants developed articles sufficiently for inclusion in this report.

Moving forward, TFe can be expanded, advancing the effort to the ultimate goal of a high-quality article for every human TF. For the future, we plan to adopt a more targeted approach by working with communities of authors who represent specific structural groups of TFs (for example, nuclear receptors) or TFs that function within a specific biological context (for example, diabetes). Such efforts can be partnered with sponsoring journals that agree to reward the community efforts with a citable publication.

Citing the resource
To cite TFe as a concept or software tool, cite this paper. To cite specific mini review articles found on the TFe website, please use the following format when possible:

Author(s) last name followed by initials: <TF symbol in bold and proper capitalization>. In Yusuf D et al.: The Transcription Factor Encyclopedia. Genome Biology 2012, 13:R24.

Example:

Bolotin E, Schnabl JM, Sladek FM: HNF4a. In Yusuf D et al.: The Transcription Factor Encyclopedia. Genome Biology 2012, 13:R24.

Conclusions
TFe is a new web-based platform for facilitating the collection, evaluation, and dissemination of TF data. It is organized and curated by a consortium of TF experts from around the world whose goal is to develop concise mini review articles on pertinent human and mouse TFs. TFe contains a wealth of TF information consisting of both automatically populated and manually curated content. Over 100 released articles are currently available, with more to come. By offering multiple data export options that include the web API, the PDF generator, and spreadsheet generator, TFe strives to be a convenient and accessible resource. The TFe is available at http://www.cisreg.ca/tfe.

Materials and methods
TFe is an amalgamation of several different and highly involved projects. For the sake of brevity, here we present only the most important key points regarding the materials and methods we employed in creating TFe. Thus, we selectively describe the materials and methods used in creating: (1) our TF classification system; (2) our TF binding profiles; (3) our TF protein structure predictions; and (4) our TF-to-disease associations. We describe the latter two in greater detail.

Transcription factor classification system
With few exceptions, all TF genes and classification information in TFe were sourced from TFCat, a large collection of predicted and confirmed mouse TF genes [4]. This collection is based on Entrez Gene identifiers. However, not all TF genes described in TFCat were added to TFe, as TFe is focused on those TFs that bind directly to DNA in a sequence-specific fashion. Thus, with few exceptions, only TFs tagged with the function-based taxonomy of ‘DNA-Binding: sequence specific’ in TFCat were added. Ultimately, out of about 1,764 mouse TF genes catalogued in TFCat, 585 were suitable enough to be imported from TFCat to TFe.

TFs in TFe are organized into ‘groups’ and ‘families’ based on their DNA binding evidence and transcriptional activation functions. This method of TF classification is inherited from TFCat. ‘Groups’ of TFs represent the highest level of organization in this classification system. Within each group exist different ‘families’ of TFs. For nuclear receptors, this classification system is further extended with a ‘subfamily’ category. Placement
of nuclear receptors within the subfamily category is guided by recommendations from the Nuclear Receptors Nomenclature Committee [36]. For a comprehensive list of the groups, families, and subfamilies that are represented in TFe, refer to Additional file 3.

**Transcription factor binding profiles**

Most of the profiles in TFe are generated through manual curation. Binding site data from our authors are submitted via a web-based form. Submissions were processed by the curatorial staff of the PAZAR database who confirm the quality of the submitted information and enter the data into the TFe division of the PAZAR database. Authors may submit either genomic coordinates or TF binding motifs, such as those generated in selection and amplification experiments.

**Protein structure predictions**

In summary, DNA binding transcription factors have been extensively studied and can be grouped according to a structural classification system [4]. For each of the small set of structural domains known to facilitate sequence-specific protein-DNA interactions, solved protein structures have been reported. Thus, it is feasible to produce homology-based models for many DNA-binding domains of proteins represented in TFe by using these solved protein structures as templates.

We generated a set of 202 predicted protein structures—homology-based predictions of the DNA binding domains of TFs. To do this, we developed a custom pipeline, written in Python, that incorporates two tools well-known in the realm of protein studies: HMMER [37] and Modeller [38]. Our protocol is based on the work of Morozov and Siggia [39], in which templates are selected to optimize similarity of DNA-binding residues. This method has been shown to increase modeling accuracy at the DNA-binding interface.

There are three main steps in generating the structural predictions: (1) building the template library; (2) finding a suitable template for each unsolved structure we would like to model; and (3) creating the structural prediction using the template as a guide.

**Building the template library**

We downloaded the entire RCSB PDB database [40] and the Protein Families (Pfam) Pfam-A HMMs database [41]. Using a custom Python script, we identified and extracted records from the PDB database that appear to contain a DNA binding domain and depict a protein-DNA binding interface (see Additional file 6 for a list of PDB records extracted). Each record is fragmented into one or more files, such that each file contains only one chain and the DNA residue. Using HMMER and the Pfam-A HMMs database, we analyzed each fragmented PDB record to catalogue all Pfam domains contained in the protein sequence. The result of this exercise is a list of relationships between Pfam domains and PDB records (Additional file 7). This constitutes our template library.

**Finding a suitable template for each unsolved structure**

For each unsolved TF protein structure, we looked for Pfam domains in the protein sequence by reviewing protein domain annotations provided by Entrez Gene. Since we are focused on modeling just the DNA binding domain of the TF protein, we removed the rest of the protein sequence. We then looked for templates in our PDB set that contain the same Pfam binding domains. We take these matching templates and compare each individually with our unsolved protein structure until the most suitable template is found. Our comparisons, which are done by an alignment tool, are scored based on similarity of the DNA-binding domain residues. For TFs known to form homodimers, a homodimeric template is selected.

**Creating the structural prediction**

After the most appropriate template is found, we input the unsolved protein sequence and the chosen template to Modeller 9v2, which constructs the predicted structure. After the structure is complete, we transfer the DNA residue from the template to the model by superimposing the two protein structures in three-dimensional space to find the most optimal superimposition, copying the DNA residue from the template to the model, and transposing the DNA residue per the superimposition coordinates. As mentioned earlier, the DNA molecules are stylistic and do not represent particular sequences - for example, the consensus sequence for the TF. The final predicted structure is rendered using iMol for presentation on the website.

**Transcription factor to disease associations**

TFs are a class of proteins that are highly implicated in disease. Thus, we have made disease annotations an important component of TFe. Under the ‘Genetics’ section of all TFe articles, we have implemented a ‘cloud’ report of associated MeSH disease terms, along with their respective P-values. These annotations were generated in-house using a novel pipeline. Conceptually, the pipeline works as follows.

In PubMed, most - if not all - articles are tagged with a list of MeSH terms by NCBI curators. Some of these terms refer to diseases such as ‘Diabetes Mellitus, Type 2’ or ‘Aniridia’. In addition, articles are often tagged with the identifiers of genes that are featured prominently in the report. Ultimately, an article is tagged with a list of MeSH terms, and also a list of genes. In TFe, we have leveraged these annotations to infer associations between TFs and certain diseases. For instance, mutations in the TF PAX6 are causal for the genetic disorder
aniridia [42]. An automated analysis of all PAX6-referring ring articles identifies the term ‘Aniridia’ as appearing far more often than expected by chance (Fisher’s exact P-value $3.2 \times 10^{-84}$).

In the end, we generated 58,807 predicted TF-to-disease associations for the TFs in TFe (mean of 74.2 associations per TF) with a scoring threshold of 0.05. These associations can be viewed under the ‘Genetics’ tab on the TFe website. An overview of our approach is as follows.

Creating the associations

We derived these associations utilizing data from PubMed and Entrez Gene. In the PubMed database, the publications indexed in PubMed are associated with MeSH terms. For instance, in PubMed, a publication about the well-characterized gene TP53 may be associated with the MeSH terms ‘Cell Line, Tumor’, ‘Oncogene Proteins, Fusion’, ‘Tumor Suppressor Protein p53’, and the like. We refer to this set of data as ‘mesh2pubmed’ as it links MeSH terms to PubMed references.

In the Entrez Gene database, there similarly exists the ‘gene2pubmed’ dataset that associates PubMed references with genes. Given these resources, it is possible to create a link between genes and MeSH terms through PubMed references. The end result is a set of ‘many-to-many’ associations between MeSH terms and genes, such that each MeSH term is associated with numerous genes, and vice versa. As MeSH is a hierarchical controlled vocabulary maintained by curators and only the most specific relevant terms are ultimately associated with each PubMed article, each MeSH term is expanded to include all of its more generic parent terms. For instance, the MeSH term ‘Diabetes Mellitus, Type 2’ would be expanded to include ‘Diabetes Mellitus’, ‘Glucose Metabolism Disorders’, ‘Metabolic Diseases’, and ‘Nutritional and Metabolic Diseases’ (the latter being the broadest and most generic term).

Following this exercise we are left with millions of gene to MeSH associations, and - in particular - 662,163 associations between TF-encoding genes and MeSH terms. Yet, not all associations are informative. For instance, the MeSH term ‘Humans’ is associated with many genes and - in practice - offers little annotation value. On the other hand, the association - or multiple associations - of a relatively rare term such as ‘Leukemia, Erythroblastic, Acute’ with a TF-encoding gene may offer greater insight into the function of that gene.

To evaluate the quality of these associations, we computed Fisher’s exact test P-value scores (Equation 1) for each TF to MeSH term association. In this equation, $n$ is the number of articles associated with the gene via ‘gene2pubmed’; $k$ is the number of $n$ articles associated with the gene annotated with the MeSH term; $N$ is the number of articles in PubMed; and $m$ is the number of articles in PubMed annotated with the MeSH term. For the background set, we compute the average rate of occurrence for all possible gene-to-MeSH term associations, taking into account not only whether an association exists, but also how often the same association occurs in each gene. In short, relatively rare MeSH terms that are associated multiple times with the same gene will yield a low (significant) P-value, while relatively common MeSH terms that are associated a few times with the same gene will yield a high (insignificant) P-value. We have not applied corrections for multiple hypothesis testing in our P-values, although we plan to implement this option in the future.

$$Pr(K \leq k) = \sum_{i=0}^{k} \binom{N}{n} \binom{N-m}{n-i} / \binom{N}{n}$$

(1)

On the TFe website, these data are displayed to the user through the use of MeSH term clouds, where clicking on a term in the cloud launches a PubMed search displaying the relevant articles.

This procedure resulted in 662,163 TF to MeSH term associations, 333,909 of which return a P-value of ≤ 0.05. About 58,807 of these associations are of TFs in the TFe database. Overall, TFs are significantly linked to 2,121 out of over 4,400 disease terms described in the MeSH vocabulary.

Additional material

Additional file 1: PDF of released article. The PDF version of the human FOXL2 article. PDF versions of released TFe articles can be accessed on the TFe website.

Additional file 2: Data of released articles. Additional data related to the released mini review articles to supplement the four-page PDF versions, arranged in alphabetical order by TF name.

Additional file 3: Classification of transcription factors in TFe. There are 803 human, mouse, and rat articles in TFe, most of which are organized into groups, families and subfamilies of TFs. The classification scheme utilized in TFe is derived from the work of Fulton et al. [5] in TFCat. There are 8 & large groups, which are further subclassified into 34 families. Several TFs, namely nuclear receptors, are even further subclassified into subfamilies.

Additional file 4: The TFe article structure. Articles in TFe are organized into ten tabs labeled ‘Summary’, ‘Structure’, ‘TFBS’, ‘Targets’, ‘Protein’, ‘Interactions’, ‘Genetics’, ‘Expression’, ‘Ontologies’, and ‘Papers’. Each tab, with the exception of the Ontologies and Papers tabs, typically begins with a brief overview written by the authors, followed by a mixture of tables and figures that features data from both the authors and second (that is, PAZAR) or third party (that is, BioGRID) sources.

Additional file 5: Binding models produced in the TFe project. Images of the binding models produced in TFe that are sufficiently characterized to be used in a study.

Additional file 6: PDB records depicting protein-DNA binding interface. A list of PDB records depict a protein-DNA binding interface.
**Abbreviations**

ACs: article completion score; APl: application programming interface; CIHR: Canadian Institutes of Health Research; MeSH: Medical Subject Headings; MSFHR: Michael Smith Foundation for Health Research; NCBI: National Center for Biotechnology Information; OMM: Online Mendelian Inheritance in Man; PDB: Protein Data Bank; PDF: Portable Document Format; PHP: PHP Hypertext Preprocessor; RSCB: Research Collaboratory for Structural Bioinformatics; TF: transcription factor; TRBS: transcription factor binding site; TFCat: Transcription Factor Catalog; TFe: Transcription Factor Encyclopedia.

**Acknowledgements**

We thank all authors who have made this project possible, especially our alpha and beta testers who have worked closely with us to refine the contents and user interface of this resource. We thank David J Arenillas, and Miroslav Hatas for providing programming assistance and systems administration support. We highlight author Frances M Sladek (University of California, Riverside) for extensive suggestions about the system interface and contents. We thank Frederick Pio (Simon Fraser University) for early assistance with TF structural modeling. TFw was supported by the Pleiades Promoter project, which is funded by Genome Canada, Genome British Columbia, GlaxoSmithKline R&D Limited, the British Columbia Mental Health and Addiction Services, Child and Family Research Institute, the University of British Columbia Institute of Mental Health, and the University of British Columbia Office of the Vice President Research. It is now supported by funding from the National Institutes of Health (grant no. R01GM084875 to WVV). Computer hardware resources utilized in this project were supported by the Gene Regulation Bioinformatics Laboratory funded by the Canada Foundation for Innovation (CFI). We sincerely thank the multitude of funding agencies throughout the world that have turned this project into reality. Here are our funding acknowledgments. FY is supported by funding from the University of British Columbia Work Study Program as well as the Faculty of Medicine of the University of British Columbia. EB is funded by T32 training grant. WAC is funded by the National Sciences and Engineering Research Council of Canada, the Michael Smith Foundation for Health Research (MSFHR) and the Canadian Institutes of Health Research/MSFHR Strategic Program in Bioinformatics. DLF is supported by a Canadian Institutes of Health Research (CIHR) Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Award and MSFHR Graduate Doctoral Scholarship award. OHR is funded by CHAARM in Europe. MCV is funded by INSERVA, EFSD, and ANR. Ovid is funded by a CIHR CCS-M Graduate Award and a Junior Graduate Scholarship EM. EMSL is funded by Canadian Institutes of Health Research, Sharon Stewart Trust, and Genome British Columbia. GUR’s work is supported by funding from Deutsche Forschungsgemeinschaft. EFW’S work is supported by Cancer Research UK and Breast Cancer Campaign. RK was funded by a research fellowship of the Faculty of Medical Sciences, Newcastle University. MSW is funded by a CASE Studentship from the BBSRC in partnership with Novartis. grants number HL091219 and R24 EY017540. AHB is funded by Leukaemia and Lymphoma Research and Cancer Research UK. SPR is funded by the National Institutes of Health (NIH) grants ES11863, ES018998 and HL66109. RLC is funded by CIHR and the Heart and Stroke Foundation of Canada. ML is funded by the National Institutes of Health (NIH) grant K08 CA123049. MV is funded by the National Institutes of Health (NIH) grants ES11893 and HL 66109. AR is funded by the National Institute for General Medical Sciences, Nebraska Department of Health and the National Cancer Institute. MZ is funded by the Italian Association for Cancer Research (AIRC) and the Italian Ministry of Education, University and Research (MIUR-PRIN). SF is funded by the National Institutes of Health (grants: CA45250, 1U54HG004558, and GM07373). P.J. is funded by the National Institutes of Health (grants CA45250, 1U54HG004558, and U01 ES017154). PJB is funded by Leukaemia and Lymphoma Research. KLP is funded by the National Institutes of Health (NIH) grant F32HD068113, and SRJ is funded by NIH grant HD42024. LDP is funded by the Spanish Ministry of Science and Innovation, as well as the 7th Framework research Program of the European Union. RHM is funded by the EU FP7 and the Swiss National Science Foundation (grant 3100A3_129962/1). MV is supported by the Ministry of Science and Higher Education (grant IP 2010 026770). MR was supported by studentships from the Faculté des Études Supérieures at Université de Montréal and from the Montreal Center for Experimental Therapeutics in Cancer. S/M is funded by the Canadian Institutes for Health Research MOP 13-147) and the Anthony and Gloria Carbone and Lila and Nathan Rabinowitz Breast Cancer Research Chair at Université de Montréal. JO is supported by the Ministry of Science and Higher Education (grant N 401 071439). SRJ is funded by the National Institutes of Health grant R01 HD42024. MB is funded by the Research Institute for Children, Children’s Hospital New Orleans, LA. MBB performed this work in collaboration with the Diana Hels University Medical Research Foundation. MSL is funded by the Research Institute at Children’s Hospital, New Orleans, and the National Institutes of Health (NIH) grant R01DK064136. KNN and JMD are funded by Canadian Institutes of Health Research (grant MOP-84320). MV is funded by Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie. JH is funded by a CIHR MOP-89808. PN is funded by the Department of Medical and Dental Sciences, University of Gothenburg. CFW is funded by the National Institutes of Health (grant 044215) and the Oklahoma Center for Adult Stem Cell Research. JMD is supported by funding from the CIHR (grant MOP-84320) and the Natural Sciences and Engineering Research Council of Canada (NSERC grant RG5123/00-2010). SW is funded by the National Institutes of Health grants CA75123, CA95026, and CA146033. DJP is funded by Australian Research Council and the National Health and Medical Research Council of Australia PSI is funded by the Welcome Trust and Breast Cancer Campaign. JIK is funded by the Dutch Organization for Medical Research (ZonMW). BB is funded by the Netherlands organization for health research and development ZonMW (grant VIDI917-66-310), the Landsteiner foundation (grant 0608), and the National Institutes of Health (grant NIH/ NIAID R01-A1080564-01). RHG is supported by funding from the National Heart, Lung and Blood Institute (HL108679) at the National Institutes of Health (NIH), and New York State Stem Cell Science (NYSTEM). MWW is a recipient of a Doctoral Award from the Canadian Institutes of Health Research. MRH is funded by the Canadian Institutes of Health Research and the Cancer Research Society. EC is funded by the Dina Gordon-Malkin Ontario Graduate Scholarship in Science and Technology. GNPE have been supported by funding from the Medical Research Council (MRC) in the United Kingdom, the British Heart Foundation, the Wellcome Trust, and Action Medical Research. MKB is funded by the Medical Research Council (MRC) in the United Kingdom, grant G0800202. OH is funded by the National Institutes of Health (NIH) grant R01CA28888. VR is funded by the NIH grant SR01 CA097226-03. DPL is funded by the Canadian Institutes of Health Research, Cancer Research Society, and the Leukemia and Lymphoma Society of Canada. SB is funded by the National Institutes of Health grants GM072939 and HL081205, National Heart, Lung, and Blood Institute Specialized Centers of Clinically Oriented Research grant P50HL084945, Center for Childhood Asthma in the Urban Environment grant P50ES015903, National Institute on Environmental Health Sciences Center grant P30 ES050381, and a Clinical Investigator Award from the Flight Attendant Medical Research Institute. C.A is funded by NIHES grant ES07141. RFH is funded by the National Institutes of Health (grant R01 MH058669). MCB was funded by National Institutes of Health (NIH) training grant T32GM007183 and also by the NIH grant R01 NS050386 (through KJM). KJM is funded by the National Institutes of Health (grants R01NS050386, R01NS044262). KD is funded by the Swedish Medical
References

1. Brand-Saberi B: Genetic and epigenetic control of skeletal muscle development. Ann Anot 2005, 187:199-207.

2. Balsamo A, Ciccaglione A, Gennari M, Sippell WG, Menabo S, Baronio F, Reipe FG: Functional characterization of naturally occurring NR3C2 gene mutations in Italian patients suffering from pseudohypoaldosteronism type 1. Eur J Endocrinol 2007, 156:249-256.

3. Field KJ, Spandidos DA: The role of ras and myc oncogenes in human solid tumours and their relevance in diagnosis and prognosis (review). Anticancer Res 1990, 10:1-22.

4. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladec K: TFCat: The Curated Catalog of Mouse and Human Transcription Factors. Genome Biol 2009, 10:129.

5. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: A census of human transcription factors: function, expression and evolution. Nat Rev Genet 2009, 10:252-263.

6. Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 2006, 126:663-676.

7. Takahashi K, Tanabe K, Ohmuki M, Nanta M, Ichisaka T, Tomoda K, Yamanaka S: Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007, 131:861-872.

8. Maglott D, Ottoy J, Prittt KD, Tatsuda T: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2007, 35:D26-31.

9. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Onderd T, Chafqa-Caspi V, Lancet D: GeneCards 2002: towards a complete, object-oriented, human gene compendium. Bioinformatics 2002, 18:1542-1543.

10. UniProt Consortium: The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 2010, 38:D142-146.

11. Boeckmann B, Barroch A, Appweller R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Pahn I, Pilbrow S, Schneider M: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003, 31:365-370.

12. Fernandez JM, Hoffmann R, Valencia A: IHOP web services. Nucleic Acids Res 2007, 35:W21-26.

13. Hoffmann R: A wiki for the life sciences where authorship matters. Nat Genet 2008, 40:1047-1051.

14. Huss JW, Lindenbaum P, Martone M, Roberts D, Pizarro A, Valafar F, Hogensieb JB, Su AI: The Gene Wiki: community intelligence applied to human gene annotation. Nucleic Acids Res 2010, 38:D633-639.

15. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Adato A, Ben-Dor U, Esterman N, Rosen N, Peter I, Onderd T, Chafqa-Caspi V, Lancet D: JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 2010, 38:D105-110.

16. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. Genome Biol 2007, 8:R207.

17. Matus V, Kell-Margoulis OV, Fricke E, Liebich I, Land S, Barbre-Dirie A, Reuter I, Chelkimey D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Petapov B, Savel H, Kel AE, Wingender E: TRANSFAC and its module TRANScmerge: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006, 34:D633-639.

18. Yiridoe E, Kyrkos J, Altman J, Velez E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 2010, 38:D105-110.

19. Portales-Casamar E, Kirov S, Lim J, Lithwick S, Swanson MI, Ticoll A, Snoddy J, Wasserman WW: PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. Genome Biol 2007, 8:R207.

20. Yiridoe E, Kyrkos J, Altman J, Velez E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 2010, 38:D105-110.
20. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragon E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahuari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, et al. *Ensembl* 2009. Nucleic Acids Res 2009, 37:D609-697.

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matesic J, Richardson JE, Ringwald M, Rubin GM, Sherlock G. *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.

22. Lipscomb CE. *Medical Subject Headings (MeSH)*. Bull Med Libr Assoc 2000, 88:265-266.

23. Bult CJ, Kadon JA, Richardson JE, Blake JA, Eppig JT. *The Mouse Genome Database: enhancements and updates*. Nucleic Acids Res 2010, 38:D586-592.

24. McKusick VA. *Mendelian Inheritance in Man and its online version, OMIM*. Am J Hum Genet 2007, 80:386-604.

25. Portales-Casamar E, Arellano D, Lim J, Saurin WJ. *The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences*. Nucleic Acids Res 2009, 37:D554-60.

26. Berman H, Henrick K, Nakamura H. *Announcing the worldwide Protein Data Bank*. Nat Struct Biol 2003, 10:980.

27. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Gardiner G, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. *The UCSC Genome Browser database: update 2010*. Nucleic Acids Res 2010, 38:D613-619.

28. Jones AR, Overly CC, Sunkin SM. *Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain*. J Mol Evol 2000, 50:103-115.

29. Hamosh A, Scott AF, Ambrose JS, Bochini CA, McKusick VA. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders*. Nucleic Acids Res 2005, 33:D514-517.

30. Mantis Bug Tracker. [http://www.mantisbt.org/].

31. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Telephone M, Herve A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wolfe J, Meijssen G, Moeller E, Roes PJ, Borner K, Barach O. *Calling on a million minds for community annotation in WikiProteins*. Genome Biol 2008, 9:R89.

32. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. *Mendelian Inheritance in Man and its online version, OMIM*. Nat Genet 2000, 25:25-29.

33. Pagon RA. *GeneTests: an online genetic information resource for health care providers*. J Med Libr Assoc 2005, 94:343-348.

34. Wang K. *Gene-function wiki would let biologists pool worldwide resources*. Nature 2006, 24:393-395.

35. Maier H, Dohr S, Grote K, Amrash M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Gardiner G, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. *The UCSC Genome Browser database: update 2010*. Nucleic Acids Res 2010, 38:D613-619.

36. Jones AR, Overly CC, Sunkin SM. *The Allen Brain Atlas: 5 years and beyond*. Nat Rev Neurosci 2009, 10:821-828.

37. Lowry JA, Atchley WR. *Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain*. J Mol Evol 2000, 50:103-115.

38. Hamosh A, Scott AF, Ambrose JS, Bochini CA, McKusick VA. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders*. Nucleic Acids Res 2005, 33:D514-517.

39. Mantis Bug Tracker. [http://www.mantisbt.org/].

40. Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Telephone M, Herve A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wolfe J, Meijssen G, Moeller E, Roes PJ, Borner K, Barach O. *Calling on a million minds for community annotation in WikiProteins*. Genome Biol 2008, 9:R89.

41. Eddy SR. *A probabilistic model of local sequence alignment that simplifies statistical significance estimation*. PLoS Comput Biol 2008, 4:e1000069.

42. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. *Comparative protein structure modeling using MODELLER*. Curr Protoc Protein Sci 2007, Chapter 2, Unit 2.9.

43. Released mini review articles in the TFe website. [http://www.cisreg.ca/cgi-bin/tfe/browselv.pl].

Cite this article as: Yusuf et al: *The Transcription Factor Encyclopedia*. Genome Biology 2012 13:R24.