Bioinformatics, 31(8), 2015, 1258–1266
doi: 10.1093/bioinformatics/btu795
Advance Access Publication Date: 29 November 2014
Original Paper

OXFORD

Data and text mining

Computer-assisted curation of a human regulatory core network from the biological literature

Philippe Thomas[1], Pawel Durek[2,3], Illés Solt[1,4], Bertram Klinger[2,5], Franziska Witzel[2,5], Pascal Schulthess[2,5], Yvonne Mayer[1], Domonkos Tikk[4], Nils Blüthgen[2,5,*] and Ulf Leser[1,*]

[1]Humboldt-Universität zu Berlin, Institute for Computer Science, Knowledge Management in Bioinformatics, 10099 Berlin, Germany, [2]Institute of Pathology, Charité–Universitätsmedizin Berlin, [3]Deutsches Rheuma Forschungszentrum, Charitéplatz 1, 10117 Berlin, Germany, [4]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary and [5]Integrative Research Institute for the Life Sciences, Humboldt Universität zu Berlin, Philippstr. 13 Haus 18, 10115 Berlin, Germany

*To whom correspondence should be addressed.
Associate Editor: Igor Jurisica

Received on July 9, 2014; revised on November 13, 2014; accepted on November 26, 2014

Abstract

Motivation: A highly interlinked network of transcription factors (TFs) orchestrates the context-dependent expression of human genes. ChIP-chip experiments that interrogate the binding of particular TFs to genomic regions are used to reconstruct gene regulatory networks at genome-scale, but are plagued by high false-positive rates. Meanwhile, a large body of knowledge on high-quality regulatory interactions remains largely unexplored, as it is available only in natural language descriptions scattered over millions of scientific publications. Such data are hard to extract and regulatory data currently contain together only 503 regulatory relations between human TFs.

Results: We developed a text-mining-assisted workflow to systematically extract knowledge about regulatory interactions between human TFs from the biological literature. We applied this workflow to the entire Medline, which helped us to identify more than 45 000 sentences potentially describing such relationships. We ranked these sentences by a machine-learning approach. The top-2500 sentences contained ∼900 sentences that encompass relations already known in databases. By manually curating the remaining 1625 top-ranking sentences, we obtained more than 300 validated regulatory relationships that were not present in a regulatory database before. Full-text curation allowed us to obtain detailed information on the strength of experimental evidences supporting a relationship.

Conclusions: We were able to increase curated information about the human core transcriptional network by >60% compared with the current content of regulatory databases. We observed improved performance when using the network for disease gene prioritization compared with the state-of-the-art.

Availability and implementation: Web-service is freely accessible at http://fastforward.sys-bio.net/.

Contact: leser@informatik.hu-berlin.de or nils.bluethgen@charite.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcription factors (TFs) influence the rates by which their target genes are transcribed by binding to regulatory DNA-segments, like promoter or enhancer regions (Vaquerizas *et al.*, 2009). In vertebrates, the relationship between TFs and genes is complex: on one hand, regulation of a specific gene often involves a variety of TFs, acting in an independent, cooperative or competitive manner (Lemon and Tjian, 2000). On the other hand, specific TFs often are involved in the co-regulation of a multitude of target genes (Niehrs and Pollet, 1999). Furthermore, TFs often also regulate other TFs. These TF-TF relationships can be considered as the core of the full human gene regulatory network (GRN) that orchestrates many cellular processes by inducing or repressing genes which function is specifically required for a given environment, for a certain cell type, or at a certain point-in-time during development and cell differentiation. Well-studied parts of this GRN are involved in wound healing (Pratt *et al.*, 2008) or in development (Davidson and Erwin, 2006), and its dys-regulation is associated with many diseases (Vaquerizas *et al.*, 2009). It also has a particularly important role in cancer (Dang, 2012; Jürchott *et al.*, 2010), where a highly interconnected regulatory core network mediates different aspects of the disease (Stelniec-Klotz *et al.*, 2012).

Biological research over the last decades has identified thousands of individual regulatory interactions using specific, time-consuming and laborious experiments. Proving a direct regulatory relationship between a TF X and a gene Y typically comprises three individual evidences: (E1) binding of X to a genomic location related to Y, (E2) change in expression of Y upon activation of X, and (E3) abrogation of regulation of Y upon removal or alteration of the binding site. To-date, only some of these evidence types can be addressed in a high-throughput manner. In particular, binding of TFs to genomic DNA can be assessed on a genome-wide manner by chromatin immuno-precipitation (ChIP) followed by sequencing, as for example has been done on a large scale by the ENCODE project (Consortium, 2012). However, binding of TFs alone does not necessarily imply that downstream genes are regulated by the TF, and genome-wide measurements tend to be rather noisy (Waldminghaus and Skarstad, 2010). Consequently, classical low throughput, mechanistic studies are still considered the most reliable way of identifying regulatory interactions (Furey, 2012), and our knowledge on the topology of the regulatory networks still remains rather sketchy (Röttger *et al.*, 2012).

A central problem in compiling regulatory networks from high-confidence low-throughput mechanistic studies is that these are scattered over the large body of scientific literature. Accessing these data in a systematic manner is difficult, as it requires finding articles discussing such relationships, correctly identifying the involved genes, and checking for each of the required evidences described earlier. There are attempts to compile knowledge about GRNs in databases, including the recently established TF Encyclopedia that is a community-curated repository of information about different aspects of TFs (Yusuf *et al.*, 2012). TRANSFAC (Wingender, 2008), TRRD (Kolchanov *et al.*, 2002) and ORegAnno (Griffith *et al.*, 2008) are more established databases specifically focusing on regulatory relationships. However, these databases do not attempt to comprehensively cover the core GRN, but rather focus on particular TFs or on specific binding sites to compile binding site motifs. Notably, these three databases (henceforth abbreviated as RegDBs) together contain only 503 regulatory relationships between two human TFs for an estimated number of at least 2000 TFs in the genome (Vaquerizas *et al.*, 2009). This situation is in stark contrast to e.g. *Escherichia coli*, for which the RegulonDB (Collado-Vides *et al.*, 2009; Gama-Castro *et al.*, 2008) contains 369 regulatory relationships between the estimated 300 TFs in the genome (Vaquerizas *et al.*, 2009).

To enlarge the body of experimentally asserted information about the human core GRN, we set out to develop, apply and evaluate a computer-assisted workflow for systematically finding and extracting experimental evidence for direct regulatory relationships between human TFs from the biological literature. This workflow comprises a state-of-the-art software to identify and normalize gene names in text; a machine-learning based classifier to judge whether a sentence in which a pair of genes co-occur describes a regulatory relationship between these two genes; and an extensive phase of human curation to check the truthfulness of the classifier's output on the sentence level and to provide an assessment of the strength of supporting evidences described in the containing article. We applied our workflow to all abstracts in PubMed. Altogether, we identified more than 18 million pairs of genes co-occurring in the same sentence. We automatically classified each of these sentences using a classifier trained on a manually annotated gold standard corpus of sentences describing regulatory relationships and inspected in detail the top-2500 sentences mentioning a pair of human TFs. 35% of those 2500 sentences report transcriptional interactions that were already covered by RegDBs. By manual curation, we found that 660 of the remaining 1625 sentences contained interesting information about gene regulatory relations, and further 322 sentences described co-operation or competition in transcription. Domain experts then studied all 459 full-text publications covering the 660 sentences to assess the trustfulness of the relationship with respect to the three lines of independent evidence mentioned earlier. This lead to the identification of 128 relationships supported by all three evidences, compared with only 35 described in the RegDBs. 310 relationships not previously covered by RegDBs were identified that are supported by at least one of the three evidences, compared with 503 described in the RegDBs. We performed an initial characterization of the expanded network and found it to be considerably larger, better connected and functionally different. It also led to improved performance when used for disease gene prioritization in four different RNA-Seq datasets.

2 Results and discussion

2.1 A workflow to extract the core regulatory network between human TFs

Mammalian cells harbor a complex regulatory core network of TFs regulating each other. We were interested in compiling this core network in an as-complete-as-possible manner using two sources of knowledge: The scientific literature and existing curated regulatory databases. We were particularly interested in comparing the respective coverage of both approaches and the quality of the network obtained by merging both sources.

To extract high-quality regulatory interactions from the literature, we first compiled a list of TFs by extending a hierarchical TF classification (Wingender *et al.*, 2013), and mapped the proteins to their respective genes (see Supplementary File S1). We then applied the workflow depicted in Figure 1. Mentions of TFs are identified in all PubMed abstracts using the state-of-the-art gene name recognition tool GNAT (Hakenberg *et al.*, 2011). GNAT was evaluated in several critical evaluations (Lu *et al.*, 2011; Morgan *et al.*, 2008) and achieves, according to these assessments, a precision of 82 % and a recall of 82 % for abstracts and precision/recall values of 54/47% for full-text articles. We identified 76 596 sentences

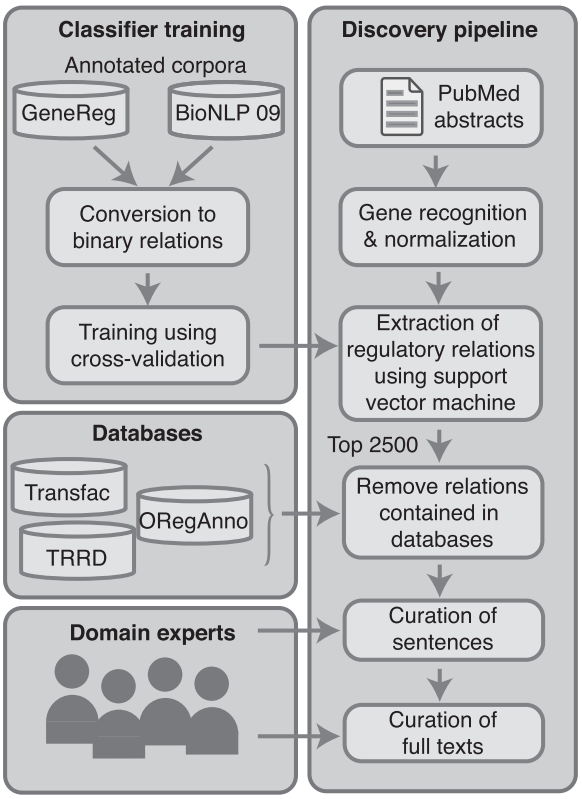

Fig. 1. Workflow of our approach. A classifier is trained on two gold standard corpora and applied to all abstracts in PubMed and the Open Access portion of PubMed Central. Positively classified sentences are partly curated manually and partly evaluated against RegulonDB

containing at least two human TFs in PubMed abstracts. Manual inspection of a sample of these sentences unveiled that large fractions consists of false positives, i.e. they do not describe regulatory interactions. Consequently, extracted sentences must be subjected to subsequent manual curation. As doing so for 76 596 sentences is not feasible, we followed a machine-learning based approach to prioritize sentences for manual curation. To this end, we trained a state-of-the-art classification algorithm on the union GeneReg and BioNLP'09, two freely available collections of manually annotated sentences on gene regulations. Example annotations for sentences containing regulatory events are shown in Figure 2A.

As for the second source of knowledge, we compiled a regulatory dataset from three well-established gene regulatory databases, namely TRANSFAC (Wingender, 2008), TRRD (Kolchanov *et al.*, 2002) and ORegAnno (Griffith *et al.*, 2008). Of these, TRANSFAC contained the largest amount of relationships between human TFs (373), TRRD contained 183 and ORegAnno contained 22. Surprisingly, we found that these databases had very little overlap, with only one relation being in all three databases (see Fig. 2B). In total, we could extract 503 unique regulatory interactions from these RegDBs, which we consider as a surprisingly low number, given that the human genome contains more than 2000 sequence-specific TFs.

2.2 The workflow extracts functional regulatory interactions from abstracts with high precision

Out of 23 140 530 Medline sentences, 3 449 157 contained at least two proteins and 76 596 contained at least two human TFs. When we applied our classifier to the latter set of sentences, it labeled 48 901 as positive (see Fig. 3A). We sorted the corresponding pairs of TFs by classifier's confidence on supporting sentences and further curated abstracts and full articles (see next section) of the top-2500 pairs. We removed all pairs which were already contained in any of the RegDBs. We then asked domain experts to manually curate the remaining 1625 sentences (for detailed statistics see Supplementary Table S1 and Supplementary File S2). Curators found 660 (40.6%) sentences clearly indicating that the respective article describes a regulatory interaction between human TFs. We also asked the curators to assess the types of experiment evidence provided in the respective article (see Fig. 3C). Interestingly, much more publications describe regulation of expression (E2) and mutational analysis (E3) than binding of a TF (E1) (see Fig. 3B). Further 322 (19.8%) sentences were found to describe cooperativity or competition between the two TFs on a common target gene, thus hinting towards a functional relationship.

Of all 1625 manually inspected sentences, 643 (39.6%) contained no evidence that the article contains any information about a regulatory relationship. Thus, the precision of our text mining pipeline on the top-2500 pairs can be estimated at 74.3% also counting those pairs as positive that were already contained in one of the RegDBs, or 60.4% when only considering TF-pairs not in current databases. We also investigated how the truly interesting sentences are distributed among all ranked sentences. Figure 3D shows the precision, recall and F1-measure at increasing rank of the curated sentence. It is reassuring for our approach that high ranks show considerably higher precision than low ranking predictions. As the recall is not saturating, curating further sentences will most likely unveil much more regulatory relations (work ongoing). To exclude a bias incurred by differences between the human experts which performed the manual curation, we also compared the proportions of their different evaluation results. Overall, reviewers obtained fairly similar percentages of the different evaluation outcomes (see Supplementary Table S2).

2.3 Manual curation of full texts establishes a high-confidence regulatory network of human cells

In the first round of manual evaluation, sentences were only assessed by the question whether or not they indicate that the article they are contained in provide experimental evidence for a regulatory relationship between the two given TFs. In a second round of curation, we studied in detail the full text of all 459 publications containing at least one of the 660 relevant sentences to collect regulatory relationship that have sufficient experimental evidence in the article. Each evidence was manually classified according to the type of experiments that were reported on in the article. This led to the identification of 310 distinct TF-TF relationships supported by at least one of the three evidences, including 128 relationships supported by all three evidences, and 82 supported by exactly two. It is worth noting that some of the reviewed full-text publications contained additional annotations which had not been found in the abstracts, leading to additional support of already known pairs.

In contrast, the RegDBs to-date contain 503 TF-TF relationships supported by at least one evidence, but only 37 relationships supported by all three evidences (Fig. 3E, yellow bars). Furthermore, the vast majority of regulations in the databases, 352 out of 503, are solely supported by experiments showing a binding of a TF to the promoter region of another TF (E1). The reliability of these interactions remains unclear (Waldminghaus and Skarstad, 2010), rendering those relationships less valuable from a biological point-of-view. In contrast, altogether 189 TF-TF relationships found by

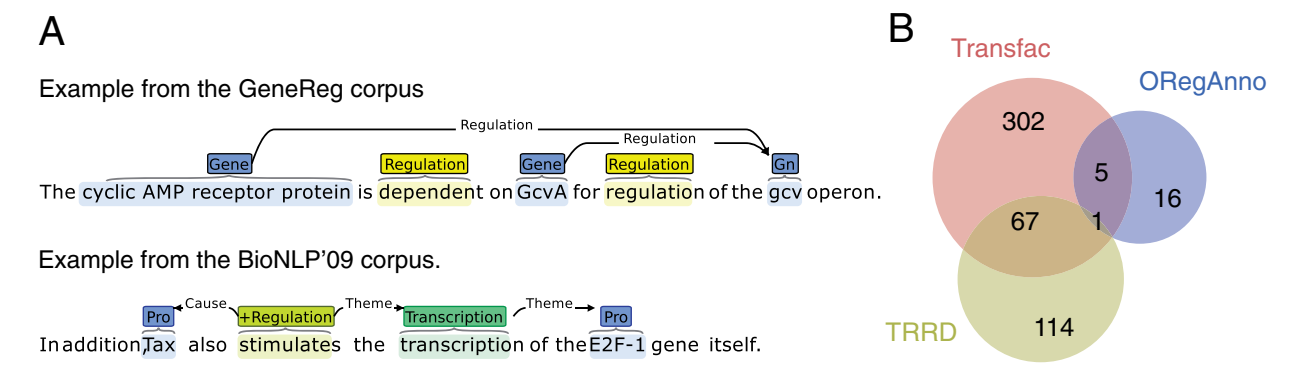

Fig. 2. Data sources. (**A**) Example sentences from the training corpora with regulatory relationship annotations visualized using (Stenetorp *et al.*, 2011). **(B)** Venn-diagram of the regulatory relations between two TFs in the databases TRANSFAC, TRRD and ORegAnno

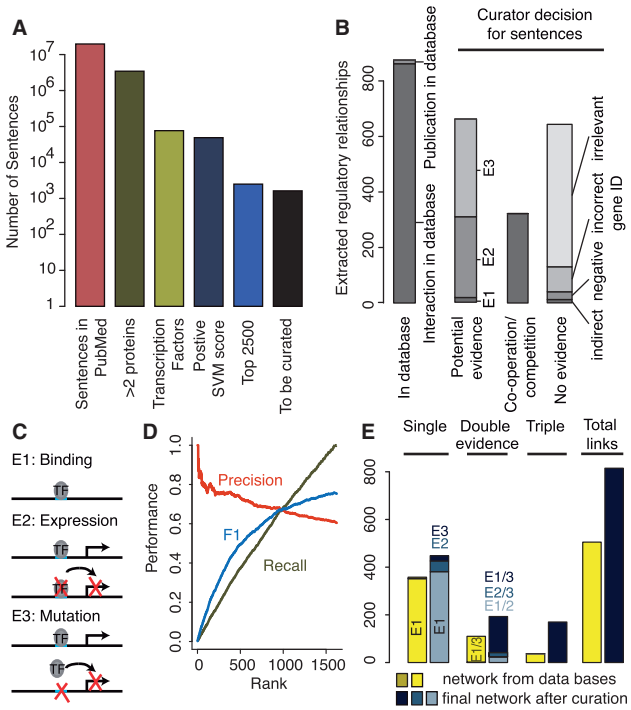

Fig. 3. Curation of regulatory interactions **(A)** Number of sentences that were considered in each of the steps of the pipeline. **(B)** Curator decisions for the 1625 sentences with highest rank. **(C)** The three evidence codes used for curation. **(D)** Precision, recall and F1-measure for manually curated sentences ranked by their confidence score. **(E)** Frequency of the different evidence levels for the existing relations in databases (yellow), or after full-text curation (blue)

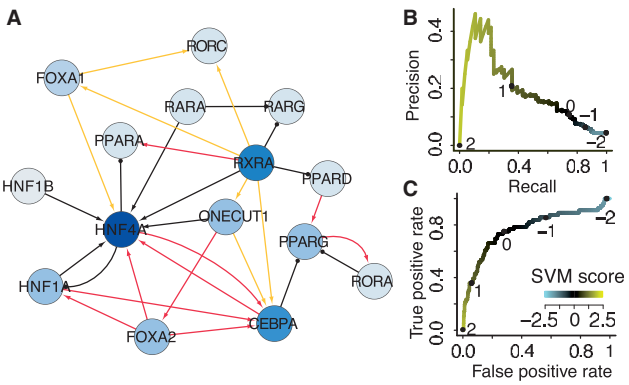

Fig. 4. Benchmarking of classifier capacity by complete manual curation. **(A)** Regulation network for liver specific TFs [list of TFs from (Tomaru *et al.*, 2009)]. Edges are categorized by respective source. *Black*, regulations contained in existing RegDBs; Red, regulations added by manual curation of the top scoring 2500 sentences; *Orange*, regulations found by manual curation of all 1435 sentences with co-occuring liver specific TFs. Arrow shapes indicate activation (arrow) and unclear regulation (circle). **(B)** Precision-recall plot on the fully curated subnetwork of 1435 sentences. **(C)** ROC curve to asses classifier performance on the fully curated subnetwork of 1435 sentences. Color coding reflects classifier score (see legend)

our curation workflow are supported by evidence E1 and also by E2 or E3 or both. Taken together, our approach increased the amount of known and experimentally asserted regulatory interactions in the human core regulatory network by 38% when compared with the RegDBs as previous state-of-the-art (compare blue and yellow bars in Fig. 3E; and Supplementary Table S3).

2.4 Comparison with a comprehensively hand-curated subnetwork

To systematically assess whether the precision of our approach could also be achieved using simpler methods, in particular simple co-occurrence of TFs in the same sentence, we decided to manually investigate all sentences with co-occurrence of two TFs of a specific subnetwork. We focused on a list of 19 liver-enriched TFs (Tomaru *et al.*, 2009), for which we obtained 1435 sentences from 781 publications mentioning at least two of these 19 factors. We then manually curated all these sentences, without using a classifier for filtering. Interestingly, only 61 (4.3%) of these 1435 sentences actually contained evidence for a regulatory relationship. This is in stark contrast to results from the classification, where 660 pairs out of 1625 (39.4%) were manually evaluated as relevant. These data suggest that our approach using top-ranked pairs increases the precision roughly 10-fold compared with pure co-occurrence and thus drastically reduces the amount of time needed to find and curate regulations.

For this particular 19-node subnetwork, only 12 interactions were found in the three databases, and the curation of the 2500 sentences led to the discovery of 10 additional interactions (see Fig. 4A). The full analysis of all sentences with co-occurring TF pairs yielded only seven additional connections, but required curation of 1435 sentences—a hard-earned improvement. By focusing on this set of 1435 sentences, we also systematically assessed classifier performance in

Table 1. Characteristics of the human regulatory network

	Only DBs	Only Curated	Combined
TFs	277	215	359
Regulatory relationships	503	332	807
Max degree	53	40	67
Average degree	3.58	2.97	4.38
Connected components	10	11	9
Diameter	10	10	9

(i) As represented in current databases; (ii) Data obtained by the workflow described in this article; (iii) Combined dataset.

terms of precision, recall and receiver operating characteristic (ROC) curve. Sentences with a high classifier score show a precision of about 0.4, and sentences with a positive score have a recall of 0.75 at a precision above 0.1 (see Fig. 4B). Similarly, the ROC-curve shows a very low false-positive rate at classifier scores above 1, and a false-positive rate of 0.35 at score 0 (see Fig. 4C).

2.5 Initial characterization of the human core regulatory network

As our approach has reconstructed the largest available GRN for human cells curated from low-throughput experiments so far, we were highly interested in the topological properties of the network. We found that the data obtained through our curation pipeline does not only increase the number of regulatory relationships in the human core network, but it also considerably changes the scope and structure of the network.

As shown in Table 1, the number of interactions increases by ~60% and the number of TFs contained in the network increase by ~30% compared with those previously described in a RegDB. The density of the network raises considerably; the average degree of nodes increases from 3.58 to 4.38 (both in-degree—the number of TFs regulating a TF—and out-degree—the number of TFs that regulate other TFs—increases, see Fig. 5B). The expanded network is better connected (from 10 to 9 connected components), and the diameter shrinks from 10 to 9. Thus, our workflow both increases the number of TFs captured by the network and the amount of knowledge on each TF within the network.

Figure 5A shows the full network, combining RegDBs and the novel curated data. Nodes whose betweenness centrality score is the highest, i.e. hubs in the network, are colored. Such genes have frequently been associated to the onset of genetic diseases (Ideker and Sharan, 2008) and, in particular, cancer (Li *et al.*, 2012). The lists of the top-10 genes ranked by betweenness-centrality for either network are shown in Table 2. Although several known regulatory hubs like SP1, FOS, MYC and P53 show high betweenness in both networks, a number of important cancer genes, like BRCA1, MYB and ESR1, rank highly only in the new combined network. A particular interesting case is HOXD13, which ranks highly in the combined network, but is not even contained in the RegDB network. These cases point to a bias in the selection of TFs and regulatory relationships that are included into a RegDB. To investigate publication bias in our network due to occurrence in current literature, we counted the occurrence of each TFs in PubMed, and plotted it against the degree of the nodes. Clearly, degree and occurrence in PubMed correlate (see Fig. 5C), indicating that the degree of TFs in the network is largely determined by how intensely a TF is investigated in the research community. However, such publication bias is evidently inevitable for any literature-based approach, including all literature-curated databases.

Analysis of the *E.coli* GRN has unveiled that certain wiring patterns, so-called network motifs, are recurrent (Shen-Orr *et al.*, 2002). One of the most important over-represented pattern is feed-forward loops by which a TF regulates a target both directly and indirectly through a second TF. This motif, for instance, has been implicated in sign-sensitive delays in signal processing or response acceleration (Mangan and Alon, 2003). We applied motif analysis (Wernicke and Rasche, 2006) to test if specific three-node network motifs are over-represented also in the human core network. Similarly to *E.coli*, we found that feed-forward loop patterns, in their different flavors, are the only 3-node network motifs that are strongly over-represented (Fig. 5D, $Z > 3$, $P < 0.01$). Interestingly, some of these feed-forward loops contain two-node feedbacks, by which two TFs that are within the feed-forward loop motif mutually regulate each other.

2.6 Effects on gene prioritization in four cancer types

To test the effect of the expanded network on a typical experimental data analysis procedure, we obtained RNA-Seq datasets from four types of human cancer (lung, prostate, liver and lung) together with corresponding healthy tissue samples from The Cancer Genome Atlas Research Network (2013). For each cancer type, we performed a network biology analysis which has proven more robust for obtaining cancer-associated genes than conventional gene-based differential analysis in several studies (Fuller *et al.*, 2007; Ideker and Krogan, 2012; Ortutay and Vihinen, 2009; Taylor *et al.*, 2009; Winter *et al.*, 2012). Therein, correlation values between genes in each sample first are mapped to a regulatory network build from background knowledge. Next, genes in each network, i.e. healthy and cancerous, are ranked according to their graph centrality. The final ranking of genes is obtained by assessing the change in the centrality rank and compared with sets of genes known to be very likely associated to the respective cancer. Results are considerably better in three out of the four cases and on-a-par in the forth, when using the expanded network as background compared with using the RegDB network (see Fig. 6). To test whether these improvements could be artefacts of the increased network size, we also created randomized networks of the same size and performed the same analysis. In three cases, results of the curated network are significantly better than expected by chance; in the forth case, results are still better but not significantly (*P*-value cutoff 0.05).

2.7 A database of human TF-TF regulatory relationships

The data assembled from several databases (e.g. TRANSFAC, TRRD, ORegAnno) and our manual curation efforts was aggregated into a database, which is accessible by our web-service *FastForward* available at http://fastforward.sys-bio.net/. The web-service allows users to search for proteins as regulators or as targets of an arbitrary regulator. Result of an example search for the TF *c-Fos* is shown in Supplementary Figure S2(a). On the left-hand side of the result page, a list of TFs and TF complexes containing *c-Fos* is shown; the right-hand side shows target genes regulated by another TF. A detailed view of all genes regulated by *c-Fos* is shown in Supplementary Figure S2(b), including the type of regulatory relationship (i.e. activation, inhibition or unknown) and the presence/absence of the three individual evidences. Hovering over a specific evidence provides links to the respective publications. A particular feature of FastForward is that users can also search for TF families. For instance, a search for *FOS* returns all TFs associated with the FOS TF family (e.g. Fra-1, Fra-2, JDP-2,. . .). The database is also provided as

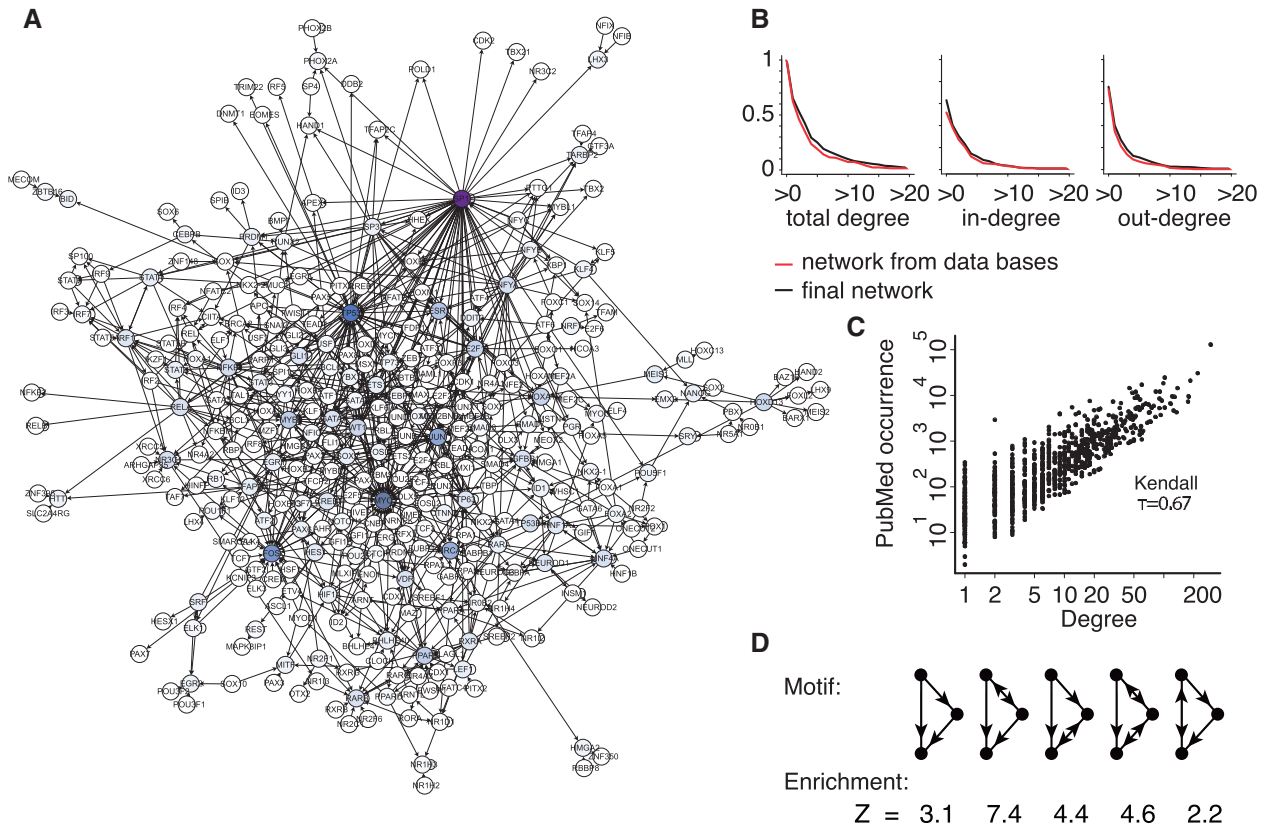

Fig. 5. Topology of the final, manually curated network. **(A)** Core GRN of human cells, including all evidence levels. Gray shade indicates centrality. **(B)** Overall degree, in-degree and out-degree increase by expanding the network. Degree distributions of the network obtained from databases (black/solid line) and the final network including the curated full texts. **(C)** The degree of each node strongly correlates with occurrence of the gene in PubMed. **(D)** The 3-node motifs that are significantly over-represented in the network (compared with a randomized network with the same degree distributions) contain feed-forward loops

Table 2. List of the 10 highest ranked betweenness centrality genes and the corresponding score in the three networks

Rank	Only DBs		Only curated		Combined	
1	SP1	0.493	MYC	0.496	SP1	0.319
2	FOS	0.177	TP53	0.177	MYC	0.251
3	TP53	0.177	SP1	0.135	TP53	0.168
4	MYC	0.097	HOXA10	0.13	FOS	0.118
5	JUN	0.093	PPARG	0.126	JUN	0.095
6	HNF4A	0.083	GATA1	0.121	BRCA1	0.083
7	WT1	0.08	MEIS1	0.111	ESR1	0.056
8	IGFBP1	0.063	ESR1	0.105	MYB	0.048
9	NR3C1	0.062	MYB	0.096	PPARG	0.047
10	BRCA1	0.057	CEBPD	0.094	E2F1	0.04

tab separated file, allowing for simple import into other databases or analysis pipelines.

3 Conclusions

Consistently verified knowledge on human regulatory relationships is still scarce and only achievable through costly low throughput experiments. Nevertheless, such knowledge is of utmost importance for further advancing research in human regulatory processes; this is especially true when serving as background knowledge in the analysis of high-throughput datasets. Here, we presented a text-mining assisted pipeline for targeted curation of a human core

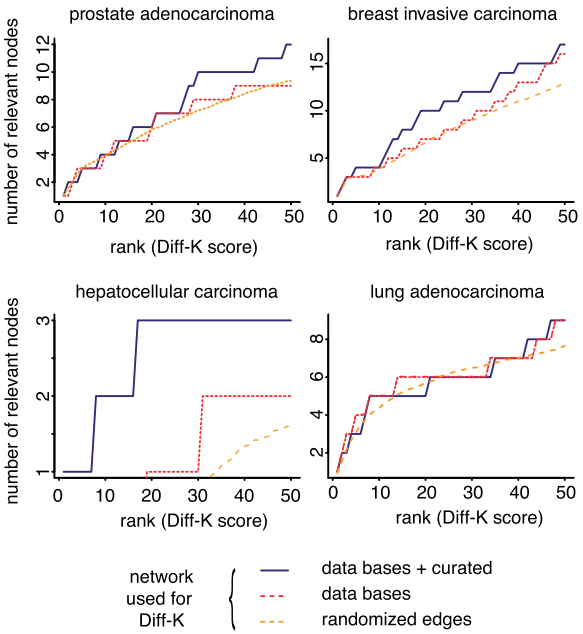

Fig. 6. Recovery rate of cancer-type associated genes using curated network versus the existing network from databases or randomized networks. Genes were ranked as described in the text and compared with gold standard gene sets. When using the expanded network, recovery of gold standards is much improved in three out of the four cases

transcriptional network. We showed that using this pipeline provides a much less costly (in terms of human labor hours) approach to the curation of regulatory relationships. Furthermore, we applied our pipeline to abstracts from Medline and could thus, after an extensive manual post-processing step, generate the (to the best of our knowledge) largest human TF network available today. Initial results of studying the properties of this network and using it in systems biology evaluations show promising results. Curation of the literature following our pipeline is ongoing and should lead to further growth of our datasets in the near future.

4 Materials and methods

4.1 TF classification

We assembled a list of human TFs as follows. We started with a list of 1690 human TFs and their respective isoforms from (Wingender *et al.*, 2013) in the version present on the following website: http://www.edgar-wingender.de/huTF_classification.html, version: June 19, 2011. We expanded this list by an additional 274 human TFs assembled from literature, TRANSFAC, TRRD and ORegAnno. HGNC gene names collected from literature are mapped to Entrez Gene identifiers using BioMart (Haider *et al.*, 2009). Mappings retrieved by BioMart are than manually evaluated for correctness. Our final list comprised 1056 unique Entrez Gene identifiers of human TFs.

4.2 Curated regulatory databases

We compiled the existing knowledge for interactions between human TFs from the following regulatory databases (RegDBs): TRANSFAC [(Wingender, 2008), Release 12.1], TRRD (Kolchanov *et al.*, 2002) and ORegAnno (Griffith *et al.*, 2008). We considered only relationships which were annotated with supporting evidence through at least one low-throughput experiment (e.g. no high density Chip–Chip) and at least one publication. Supporting experiments were classified into one or more of the three evidences categories (see Supplementary File S3).

4.3 Corpora for training the sentence classifier

We train predictive models for recognizing binary regulatory relationships in text using two existing corpora: First, we used the GeneReg corpus [version 1.0; (Buyko *et al.*, 2010)], which is a set of 314 manually annotated Medline abstracts about gene regulation in *E.coli*. We considered all interactions having a gene/protein as regulator, yielding 1164 positive pairs. The remaining 1616 pairs were used as negative examples. Second, we used the bio-molecular event corpus of the BioNLP'09 Shared Task (Kim *et al.*, 2009) consisting of 951 abstracts that were selected by MeSH terms 'Human', 'Blood cells' and 'Transcription Factor'. Of all annotated relationships, we considered all cases where the expression of a protein is regulated by another protein as positive, resulting in 295 positive and ~10 000 negative training examples. Classifiers often tend to keep the same positive to negative ratio seen in the training phase (Chawla *et al.*, 2004). We counteracted this problem by applying higher penalty costs for errors in the minority class (Veropoulos *et al.*, 1999).

4.4 Classifying TF-pairs in sentences

We applied the relation extraction library described in (Tikk *et al.*, 2013) to identify regulatory relationships between pairs of genes within a sentence. This library integrates sentence parsers (syntax and dependency), format conversion routines, experiment management and 13 algorithms for supervised relationship classification. Based on the results from (Tikk *et al.*, 2010), we used the two best performing methods in our experiments: First, the *shallow linguistic* [SL; (Giuliano *et al.*, 2006)] kernel builds high-dimensional context profiles based on words, stems and POS tags in near proximity of the mentions and in the sentence containing the pair. Second, the *all-paths graph* [APG; (Airola *et al.*, 2008)] kernel requires that sentences are first parsed to derive their dependency structure (de Marneffe and Manning, 2008). APG then uses all features from the SL classifier plus features derived from all paths connecting the mentions in the dependency graphs.

4.5 Curation

We tagged all abstracts from PubMed (as of June 2010) using GNAT (Hakenberg *et al.*, 2011). We removed all sentences which do not contain at least two human TFs and classified each remaining pair using our classifier described earlier. Positively classified pairs were ranked according to the classifiers confidence, and the top-2500 were selected as candidates for further evaluation. For manual curation, we filtered all candidates that were already present in TRANSFAC, TRRD or ORegAnno and also those candidates that were mentioned in publications already curated in one of these knowledge bases. The remaining 1625 candidate sentences were randomly split into five parts and manually evaluated by domain experts for evidence of regulatory relationships.

We adopted a two-phase curation, where in the first phase the experts had to judge if the sentence suggests that the article contains information about gene regulatory interactions, and in a second phase the experts read the full-text articles. The final network was then constructed using only those interactions where curators found experimental evidence for regulatory interactions in the full texts. To assess the benefit of our classification-based approach compared with simple co-occurrence, we also curated all 1435 sentences containing two different TFs from a list of 19 liver enriched TFs (Tomaru *et al.*, 2009), irrespectively of how these sentences were classified.

4.6 Network construction

For network analysis, we mapped TFs to vertices and regulatory relationships to edges in a graph. One obstacle in analyzing the extracted relationships in this manner is that TFs are often complexes of proteins. For example, the TF AP-1 describes a complex that contains a protein of the FOS family, and a protein of the JUN family. To provide the highest level of detail for different types of analysis, we retained the information if a TF was a single protein or a complex during curation (and in our database) by means of our hierarchical classification scheme (Wingender *et al.*, 2013). For network analysis, we decided to map the TFs to the genes that encode the proteins contained in the TFs for network analysis. Thereby, we generated a network between genes, and each link between a TF and a target gene may become multiple links if the TF is a complex, or can contain several members of a family of proteins.

4.7 Differential centrality analysis

RNA-Seq dataset for healthy and cancerous samples were obtained from the Cancer Gene Atlas (Accession ids: lung adenocarcinoma, LUAD; prostate adenocarcinoma, PRAD; liver hepatocellular carcinoma, LIHC; breast invasive carcinoma, BRCA). Spearman correlation of gene expression values within healthy and cancerous samples, respectively, where computed and added as edge weights to a background network of TFs. The centrality of TFs in the two networks per cancer was compared and TFs were finally ranked by the

Diff-K measure (Fuller *et al.*, 2007). We compared these ranked lists to cancer-specific gene lists obtained from MalaCards (Rappaport *et al.*, 2013) to assess the ability of the background network to recover known TF-disease associations. For evaluating the usefulness of our curation approach, we performed this analysis twice, using once the Reg-DB network as background and once the expanded network, and compared results.

Furthermore, we tested if the observed improvements are only an effect of the increased network size, but not due to specific novel TFs and relationships. To this end, we generated randomized networks as competitors for the expanded network as follows. We started with the Reg-DB network, as it is the common core contained in all networks considered here. We added as many TFs as the expanded network has more than the Reg-DB network, drawn randomly from our list of human TFs. In this process, the chance to draw a specific TF from the list equals its relative occurrence in PubMed. We then computed a random mapping between the additional TFs in the expanded network and the randomly chosen additional TFs in the randomized network and added as many random edges to each added TFs in the randomized network as its counterpart has in the expanded network. We generated 100 networks for each cancer type following this procedure and computed the distribution of recovery rates (see Fig. S1 in Supplementary Materials).

Funding

This work was funded by German Academic Exchange Service and the German Federal Ministry of Education and Research [0315417B, 0316184A,B, 01GQ1001C, 0315261].

Conflict of Interest: none declared.

References

Airola,A. *et al.* (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, **9(Suppl. 11)**, S2.

Buyko,E. *et al.* (2010). The GeneReg corpus for gene expression regulation events—an overview of the corpus and its in-domain and out-of-domain interoperability. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. ISBN: 2-9517408-6-7.

Chawla,N. *et al.* (2004). Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, **6**, 1–6.

Collado-Vides,J. *et al.* (2009). Bioinformatics resources for the study of gene regulation in bacteria. *J. Bacteriol.*, **191**, 23–31.

Consortium,T.E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Dang,C.V. (2012). Myc on the path to cancer. *Cell*, **149**, 22–35.

Davidson,E.H. and Erwin,D.H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, **311**, 796–800.

de Marneffe,M.-C. and Manning,C.D. (2008). The Stanford typed dependencies representation. In *Proceedings of the COLING'08 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8.

Fuller,T.F. *et al.* (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome*, **18**, 463–472.

Furey,T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.

Gama-Castro,S. *et al.* (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36(Suppl. 1)**, D120–D124.

Giuliano,C. *et al.* (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, The Association for Computer Linguistics, Trento, Italy, pp. 401–408.

Griffith,O.L. *et al.* (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.

Haider,S. *et al.* (2009). BioMart Central Portal–unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.

Hakenberg,J. *et al.* (2011). The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

Ideker,T. and Krogan,N.J. (2012). Differential network biology. *Mol. Syst. Biol.*, **8**, 565.

Ideker,T. and Sharan,R. (2008). Protein networks in disease. *Genome Res.*, **18**, 644–652.

Jürchott,K. *et al.* (2010). Identification of y-box binding protein 1 as a core regulator of mek/erk pathway-dependent gene signatures in colorectal cancer cells. *PLoS Genet.*, **6**, e1001231.

Kim,J.-D. *et al.* (2009). Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, Boulder, Colorado, pp. 1–9.

Kolchanov,N.A. *et al.* (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.*, **30**, 312–317.

Lemon,B. and Tjian,R. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.

Li,B.-Q. *et al.* (2012). Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. *PLoS One*, **7**, e33393.

Lu,Z. *et al.* (2011). The gene normalization task in biocreative iii. *BMC Bioinformatics*, **12(Suppl. 8)**, S2.

Mangan,S. and Alon,U. (2003). Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U S A*, **100**, 11980–11985.

Morgan,A. *et al.* (2008). Overview of biocreative ii gene normalization. *Genome Biol.*, **9(Suppl. 2)**, S3.

Niehrs,C. and Pollet,N. (1999). Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.

Ortutay,C. and Vihinen,M. (2009). Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.* **37**, 622–628.

Pratt,C. *et al.* (2008). Transcriptional regulatory network analysis during epithelial-mesenchymal transformation of retinal pigment epithelium. *Mol. Vis.*, **14**, 1414–1428.

Rappaport,N. *et al.* (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)*, **2013**, bat018.

Röttger,R. *et al.* (2012). How little do we actually know? On the size of gene regulatory networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 1293–1300.

Shen-Orr,S.S. *et al.* (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.

Stelniec-Klotz,I. *et al.* (2012). Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS. *Mol. Syst. Biol.*, **8**, 601.

Stenetorp,P. *et al.* (2011). BioNLP shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Oregeon, USA, pp. 112–120.

Taylor,I.W. *et al.* (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.

The Cancer Genome Atlas Research Network (2013). *The Cancer Genome Atlas*. http://cancergenome.nih.gov.

Tikk,D. *et al.* (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput. Biol.*, **6**, e1000837.

Tikk,D. *et al.* (2013). A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC Bioinformatics*, **14**, 12.

Tomaru,Y. *et al.* (2009). Identification of an inter-transcription factor regulatory network in human hepatoma cells by Matrix RNAi. *Nucleic Acids Res.*, **37**, 1049–1060.

Vaquerizas,J. *et al.* (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Veropoulos,K. *et al.* (1999). Controlling the sensitivity of support vector machines. In *Proceedings of Inertnational Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden, pp. 55–60.

Waldminghaus,T. and Skarstad,K. (2010). Chip on chip: surprising results are often artifacts. *BMC Genomics*, **11**, 414.

Wernicke,S. and Rasche,F. (2006). FANMOD: a tool for fast network motif detection. *Bioinformatics*, **22**, 1152–1153.

Wingender,E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.*, **9**, 326–332.

Wingender,E. *et al* (2013). TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, **41**, D165–D170.

Winter,C. *et al*. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.*, **8**, e1002511.

Yusuf,D. *et al*. (2012). The transcription factor encyclopedia. *Genome Biol.*, **13**, R24.