Supplementary Material

Introduction

In this document we provide additional information about the different sources that we checked to obtain the input list of candidate transcription factors (TFs) for the TFcheckpoint database. In addition, we provide some details about the text curation effort that we undertook to establish the existence of experimental evidence that would support a DNA binding transcription factor (DbTF) classification of the putative TFs.

A) Overview of TF-data sources

We built a cumulative list of putative TFs by collecting all entries marked as transcription factor from sources with mammalian TFs (see Supplementary Table 1). This provided a list of 3462 unique putative TFs that we next checked for functional evidence in literature.

Supplementary Table 1: Transcription factor sources.

Sources	Entries	Species	URL	PubMed / Version / Date
TFCat (Fulton, et al., 2009)	1052	human, mouse, rat	http://www.tfcat.ca/	PMID: 19284633/ Release 1.0 / March 12, 2009
JASPAR (Sandelin, et al., 2004)	115	human, mouse, rat	http://jaspar.cgb.ki.se/	PMID: 18006571 / October 12, 2009
DBD (Kummerfeld and Teichmann, 2006)	1395	human, mouse, rat	http://www.transcriptionfactor.org/index.cgi?Home	PMID: 16381970 / Release 2.0
ORFeome (Messina, et al., 2004)	1770	human		PMID: 15489324 / October, 2004
AnimalTFDB (Zhang, et al., 2011)	1682	human, mouse, rat	http://115.156.249.50/TFDB/index.php	PMID: 22080564 / November 12, 2011
Vaquerizas *et al* (Vaquerizas, et al., 2009)	1909	human		PMID: 19274049 / April, 2009
Ravasi *et al* (Ravasi, et al., 2010)	1967	human, mouse		PMID: 20211142 / March 05, 2010
TcoF-DB (Schaefer, et al., 2011)	1864	human	http://cbrc.kaust.edu.sa/tcof/index.php	PMID: 20965969 / October 2010
GOC (Harris, et al., 2004)	1121	human, mouse, rat	http://amigo.geneontology.org	PMID: 10802651 / February 16, 2013

Data from 9 sources (column 1) were downloaded and used to assemble a comprehensive list of proposed TFs. The table shows the identifier(s) of the source; the number of unique entries obtained from that source, the species, the URL if the source is a database, the PubMed ID of the appropriate reference and the time of download.

Supplementary Table 2: Overlap between the TF sources included in the TFcheckpoint.

	AnimalTFDB	Vaquerizas *et al*	ORFeome	Ravasi *et al*	TFCat	JASPAR	GOC	DBD	TcoF-DB
AnimalTFDB	1682 (48.6)								
Vaquerizas *et al*	1435 (41.4)	1909 (55.1)							
ORFeome	1162 (33.6)	1225 (35.4)	1770 (51.1)						
Ravasi *et al*	1157 (33.4)	1256 (36.3)	1355 (39.1)	1967 (56.8)					
TFCat	683 (19.7)	685 (19.8)	785 (22.7)	852 (24.6)	1052 (30.4)				
JASPAR	111 (3.2)	111 (3.2)	109 (3.2)	114 (3.3)	108 (3.1)	115 (3.3)			
GOC	435 (12.6)	576 (16.6)	583 (16.8)	662 (19.1)	392 (11.3)	47 (1.4)	1121 (32.4)		
DBD	1307 (37.8)	1325 (38.3)	1093 (31.6)	1080 (31.2)	601 (17.4)	98 (2.8)	398 (11.5)	1395 (40.3)	
TcoF-DB	1278 (36.9)	1362 (39.3)	1332 (38.5)	1464 (42.3)	845 (24.4)	115 (3.3)	578 (16.7)	1198 (34.6)	1864 (53.8)

The overlap between resources is indicated in two ways: 1) The first number on the intersection of columns and rows shows the number of identical TFs between the two resources; the second number (between parentheses) indicate the percentage-wise overlap between different TF sources (taking the cumulative total (3462) as 100%).

Supplementary Table 3: Overview of the TF-sources with their strengths and weaknesses.

	TFCat	JASPAR	DBD	ORFeome	AnimalTFDB	Vaquerizas *et al*	Ravasi *et al*	TcoF-DB	GOC
Literature evidence	+	+	-	-	-	-	-	-	+
Comprehensiveness	+	-	+	+	+	+	+	+	-
Sequence/structure similarity based	+	+	+	+	+	+	+	-	+
Experimental evidence	-	-	-	-	-	-	-	-	+
Only DbTFs	-	+	-	-	-	-	-	-	+
TF curation guidelines	-	-	-	-	-	-	-	-	+

The strengths and weaknesses of these sources are highlighted based on presence '+' and absence '-' of the 6 different features, respectively.

B) DbTF annotation procedure

A DbTF by definition binds to specific DNA sequences in the promoter or enhancer region of a gene and regulates the transcription of the associated gene. Therefore, in our DbTF annotation procedure we considered the following two functional properties as the minimum criteria to qualify a protein as DbTF:

> i) there is evidence that the protein binds to specific DNA sequences and
> ii) the protein has been demonstrated to be involved in RNAPII dependent regulation of transcription.

Next, we compiled a list of experimental assays for protein-DNA interaction and transcription regulation (Supplementary Table 4) in order to identify the above evidence types for TFs in scientific publications.

Then we looked for specific scientific publications that would contain evidence to qualify TFs according to our DbTF annotation criteria. We started checking the already existing TF annotations by inspecting the literature that their annotations referred to. The majority of these existing annotations

came from GOC (174 DbTFs), JASPAR (112 DbTFs) and TFCat (231 DbTFs). Next, we searched the literature for experimental evidence supporting the remaining TF candidates, by performing searches for gene names in the following resources: UniProt (http://www.uniprot.org/), NCBI's Entrez Gene (Maglott, et al., 2007), iHOP (Hoffmann and Valencia, 2004), Gene Cards (Safran, et al., 2002), and NCBI's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). This yielded additional literature references for 466 DbTFs.

Supplementary Table 4: List of experimental assays for creating DbTF annotation.

Experimental assays
Specific DNA binding
Electrophoretic mobility shift assay (EMSA)
Electrophoretic mobility supershift assay (EMSA supershift)
DNA footprinting
DNase I footprinting (DNA footprint)
Methylation interference assay (MIC)
Ultraviolet (UV) footprinting (UV-footprint)
Dimethylsulphate footprinting (DMS-footprint)
Hydroxy radical footprinting (Hydroxy-footprint)
Potassium permanganate footprinting (KMnO4-footprint)
Affinity chromatography technology
DNA Pull-down assay
Southwestern blot assay (SW-blot)
In vitro evolution of nucleic acids (SELEX)
X-ray crystallography
RNAPII dependent transcription regulation*
Reporter gene assay
TG expression assay (Primer-specific PCR, Northern blot, Ribonuclease protection assay)

* To measure the transcription regulation, TF should have been identified by any of the following methods: wild-type TF overexpression, mutated TF overexpression, TF knock down (RNAi/antisense RNA).

References

Fulton, D.L., *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors, *Genome biology*, **10**, R29.
Harris, M.A., *et al.* (2004) The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Research*, **32**, D258-D261.
Hoffmann, R. and Valencia, A. (2004) A gene network for navigating the literature, *Nature genetics*, **36**, 664.
Kummerfeld, S.K. and Teichmann, S.A. (2006) DBD: a transcription factor prediction database, *Nucleic Acids Research*, **34**, D74-D81.
Maglott, D., *et al.* (2007) Entrez Gene: gene-centered information at NCBI, *Nucleic acids research*, **35**, D26-31.
Messina, D.N., *et al.* (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression, *Genome Res*, **14**, 2041-2047.
Ravasi, T., *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man, *Cell*, **140**, 744-752.
Safran, M., *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium, *Bioinformatics*, **18**, 1542-1543.
Sandelin, A., *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res*, **32**, D91-94.
Schaefer, U., Schmeier, S. and Bajic, V.B. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins, *Nucleic Acids Research*, **39**, D106-D110.
Vaquerizas, J.M., *et al.* (2009) A census of human transcription factors: function, expression and evolution, *Nature reviews. Genetics*, **10**, 252-263.
Zhang, H.-M., *et al.* (2011) AnimalTFDB: a comprehensive animal transcription factor database, *Nucleic Acids Research*.