S1 | A census of human transcription factors: function, expression and evolution

Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, Nicholas M. Luscombe

Nature Reviews Genetics (2009)

Supplementary material

1. Identification of TF repertoire

We identified TFs by looking for proteins that bind DNA in a sequence-specific manner. To do so, we assembled a list of DNA-binding domains and families from the InterPro database (release 17)[1]. For each entry we examined the description and associated literature to assess their sequence-specific DNA-binding capabilities, which resulted in an accurate list of 347 domains and families (see Supplementary Table 1 [.txt file S2]). We then extracted 4,610 proteins from the International Protein Index (IPI; release 3.48) database[2] that showed a significant match with these selected DNA-binding domains. DNA-binding domain searches were performed using InterproScan[3] with default parameters and are available in the IPI database. This group of proteins mapped to 1,960 human genomic loci in the Ensembl database (release 51)[4].

Next, we manually inspected each locus using the information available in GeneCards[5], NCBI's Entrez Gene[6] and Uniprot[7], and grouped them according to our confidence in their functionality as sequence-specific DNA-binding TFs (see Supplementary Table 2 [.txt file S3]). Genes classified as 'a' or 'b' are probable TFs for which there is experimental evidence of regulatory function in any mammalian organism ('a') or have an equivalent protein domain arrangement ('b'); those classified as 'c' are possible TFs that contain non-promiscuous InterPro DNA-binding domains (i.e. InterPro domains that are only ever found in TFs), but for which we do not have further functional evidence. Finally, we removed from the list unlikely TFs (classified as 'x') that comprise predicted genes, contain promiscuous InterPro DNA-binding domains (i.e. DNA-binding domains that are also found in non-TFs) or have an established molecular function outside transcription (such as nucleoporins, threonine phosphatases or splicing factors). Finally, we also included 27 curated probable TFs from other sources ('other'), such as Gene Ontology (GO)[8] or TRANSFAC[9] containing undefined DNA-binding domains, and were therefore missed using the above procedure.

The InterPro domains are provided in Supplementary Table 1. The list of all loci, along with their quality assessment, is provided in Supplementary Table 2.

For the Analysis we focused on 1,391 TFs that were classified as 'a', 'b' or 'other'.

2. Coverage of the TF repertoire

We measured the coverage of our TF repertoire by evaluating the percentage of TFs recovered by our own method compared with a gold-standard set of loci annotated with the Gene Ontology terms "transcription factor activity" and "DNA-binding". There are 62 loci annotated with these terms derived from experimental evidence (as of Dec 2008). Of these, we recover 58 genes, approximating to a coverage of 94%.

To eliminate possible bias in GO annotations for the human genome, we performed the same analysis using the mouse genome. We extracted one-to-one orthologues for our human TF dataset from the Ensembl Compara database (version 51), and compared them with a gold standard set of loci annotated with the same GO terms as above. We recovered 175 out of 207 experimentally annotated genes, giving in an estimated coverage of 85%.

3. Microarray data analysis for TF expression

The SymAtlas dataset measures gene expression levels across 79 human tissue samples and cell lines using Affymetric GeneChip HG-133U arrays[10]. Each sample is represented by two biological replicates. For the Analysis, we focused on data for 32 major human tissue types.

We obtained the original .CEL files from the Genome Novartis Foundation. First, we processed the raw data using the three-step GCRMA algorithm[11] implemented in the BioConductor project[12]. This resulted in $\log_2$ values representing expression levels for each probeset.

Next, for each array we define minimum $\log_2$ values at which we define a probeset as being "present" in the biological sample. We use the PANP method[13], which estimates a false-positive rate at a given $\log_2$ value threshold using control probesets on the array. We also estimate a true-positive rate by checking for the presence of ESTs in the Unigene database[14] for libraries taken from equivalent tissue types (see Supplementary Table 4).

We compare our method with MAS.50 and arbitrary thresholds commonly used for this dataset (e.g. $\log_2(100)$, $\log_2(200)$). In Supplementary Figure 1 we show ROC curves displaying the sensitivity and specificity for these different methods. At 5% false positive rate, the PANP approach improves the sensitivity of detection by an average of 30% over MAS5.0.

Finally, probesets were matched with gene loci using the mapping provided by the Ensembl database (version 51). A gene was defined as expressed if any of its probesets were present in both biological replicates at a false positive rate threshold of 5%. Furthermore, expression values were calculated for each probeset as the mean value of the two replicates per tissue available in the SymAtlas dataset. The re-analysed expression dataset is available at ArrayExpress[15] under accession number E-TABM-145.

4. Defining tissue-specific expression

Using the processed microarray data, we defined sets of "present" TFs that display tissue-specific expression. For this we calculated a propensity score, which normalises a probeset's expression relative to all other probesets using:

$$propensity = \frac{x_{ij} \sum_i \sum_j x_{ij}}{\sum_i x_{ij} \sum_j x_{ij}}$$

where P_{ij} is the propensity that a given probeset i is expressed at level x in tissue j relative to its expression in all other conditions. Therefore in this study, each probeset has 32 propensities corresponding to each tissue. For comparison, we also calculated propensities for a randomised gene expression dataset, in which expression values are shuffled among all tissues. Tissue-specific TFs were identified as those in which any of their probesets had propensity values higher than the top 5% of all randomised propensities. TFs classified as "present" with no tissue-specific probesets were classified as non-specific.

To test the results, we examined the GO annotations for groups of tissue-specific genes (including non-TFs) using the g:Profiler software[16]. Specific GO terms are enriched in different tissues; for example "immune system process" in whole blood, and "transmission of nerve impulse" in brain, confirming that appropriate genes were assigned to each tissue.

5. Detection of clusters of TF loci in the genome

Clusters were identified by comparing the number of TF and non-TF genes in a 500kb sliding window along each chromosome (100kb step size) using Ensembl (v51) coordinates. A p-value was calculated for each window with the Fisher exact test, using the expected numbers of TF and non-TFs for each chromosome independently. The p-values were adjusted for multiple-testing within each chromosome using FDR, and windows with $p < 0.05$ were defined as significantly enriched for TFs. Finally, we defined 23 clusters by merging overlapping windows with significant p-values. The list of clusters and their locations are presented in Supplementary Table 5.

6. Identification of TF homologues

For all TFs, we identified one-to-one, one-to-many and many-to-many orthologues using the GeneTree assignments from Ensembl Compara (version 51)[17]. 24 eukaryotic organisms were chosen based on the quality of their genome sequences and annotations within the Ensembl database (at least 2x coverage; non-beta annotations). These organisms were grouped into primates (human, chimpanzee, macaque, orangutan), mammals (mouse, rat, pig, cow, dog, horse), vertebrate (opossum, platypus, fugu, tetraodon, stickleback, zebrafish, frog, chicken), metazoa (ciona, mosquito, fly, worm), and all eukaryotes (baker's yeast).

TFs were hierarchically clustered based on the presence or absence of the previously identified orthologues across genomes, and the results were shown as heatmaps. TFs were classified into one of five groups (primate-specific, mammalian-specific, vertebrate-specific, metazoa-specific, and eukaryotic) by the presence of one-to-one orthologues. First, we checked if a TF has an orthologue in the yeast genome, if so it was classified as a eukaryotic TF. If not, next we checked if a TF has orthologues in at least two metazoan genomes, and if so it was classified as a metazoa-specific TF. If not, next we check for orthologues in at least two vertebrate genomes, then mammals and finally primates.

In order to examine the evolution of different TF families, for each taxonomic class we calculated the proportion of TFs belonging to each of the three largest families (C_2H_2-ZNF, homeodomain and helix-loop-helix; Figure 5B). The proportion is the number of TFs in each taxonomic group for each family divided by the total number of TFs classified in each taxonomic group.

21 TFs did not have orthologues in any other genome and therefore appeared to be unique to the human genome. We checked whether these were true human-specific by running BLAT sequence searches against the chimpanzee genome using the UCSC Genome Browser website[18]. Each TF returned a significant hit - in most cases with 100% sequence identity - along the entire length of its gene against unique regions of the chimpanzee genome. This suggested that these TFs were not truly human-specific, but appeared so because the gene was not yet annotated in the chimpanzee genome.

References

1. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res* **35**, D224-8(2008).
2. Kersey, P.J. et al. The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985-8(2004).
3. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116-20(2005).
4. Hubbard, T.J. et al. Ensembl 2007. *Nucleic Acids Res* **35**, D610-7(2007).
5. Safran, M. et al. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18**, 1542-3(2002).
6. Maglott, D. et al. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **35**, D26-31(2007).
7. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**, D169-74(2009).
8. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9(2000).
9. Wingender, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-9(2000).
10. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7(2004).
11. Wu, Z. et al. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Associatio* **99**, 909-917(9)(2004).
12. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80(2004).
13. Warren, P. et al. PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering,* 108-115(2007).
14. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35**, D5-12(2007).
15. Parkinson, H. et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* **35**, D747-50(2007).
16. Reimand, J. et al. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**, W193-200(2007).
17. Vilella, A.J. et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-35(2009).
18. Karolchik, D. et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**, D773-9(2008).

Supplementary Tables

Supplementary Table 1 (available electronically as .txt file S2). List of Interpro DNA-binding domains and families used to characterise the human TFs repertoire.

Supplementary Table 2 (available electronically as .txt file S3). List of TF-encoding loci classified as 'a', 'b', 'c', 'x', or selected from other databases (classified as 'other'; see Supplementary Material). For TFs included in the repertoire, the list also contains accompanying information including: Ensembl gene IDs (release 51), HGNC identifiers, IPI IDs, associated DNA-binding Interpro domains and families, and tissue specificity if any.

Supplementary Table 3 (available electronically as .txt file S4). List of TF families and corresponding InterPro domains. InterPro domains were grouped on the basis of the InterPro parent-child relationships. TF that belong only to families with less than five TFs were classified as 'other'.

Supplementary Table 4. List of Unigene libraries used to estimate the sensitivity of the "present"/"absent" call method.

Unigene library ID	Tissue
6989	liver
252	liver
5551	pancreas
6760	pancreas
13019	pituitary gland
2587	placenta
6835	placenta
13001	placenta
13000	placenta
250	placenta
10196	ovary
253	ovary
6763	prostate
14129	prostate
14131	prostate
13710	testis
6833	kidney
6759	heart
3718	lymph node
2710	lymph node
2709	lymph node
3720	lymph node
3719	lymph node
2711	lymph node
6761	muscle
530	muscle
45	muscle
14414	muscle
6834	lung
249	lung
14590	spinal cord

Supplementary Table 5. List of human TF cluster locations (chromosomal positions according to Ensembl v51).

Chromosome	Start (bp)	End (bp)
1	245100000	245600000
2	176200000	177100000
3	44200000	45000000
5	177900000	178700000
6	28000000	28800000
7	26700000	27600000
7	63300000	64200000
7	148100000	149000000
10	38000000	38600000
12	52200000	53100000
12	131800000	132300000
16	2900000	3700000
17	43600000	44500000
19	9100000	9900000
19	11400000	12600000
19	19600000	20300000
19	20600000	21400000
19	21600000	22300000
19	23600000	24100000
19	41300000	43100000
19	48800000	49900000
19	57000000	58500000
19	61300000	63800000

Supplementary Figure 1. Receiver Operator Characteristic (ROC) curves measuring the sensitivity and specificity of microarray data for the lymph node. The solid black curve represents the quality of gene expression data at different detection thresholds at expression values between $\log_2(1)$ and $\log_2(200)$. Coloured dots represent different cut-offs: PANP at 1% and 5% error rates (blue), MAS5.0 (red) and commonly used arbitrary thresholds at $\log_2(100)$, $\log_2(150)$ and $\log_2(200)$ (orange). The best balance between sensitivity and specificity is achieved using the PANP 5% cut-off, which occurs at the shoulder of the ROC curve. Similar results were obtained for the other evaluated tissues (see Supplementary Table 4; data not shown).