BMC Genomics

BioMed Central

Research article

Open Access

Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals

Sven Nelander[1], Erik Larsson[1], Erik Kristiansson[2], Robert Månsson[3], Olle Nerman[2], Mikael Sigvardsson[3], Petter Mostad[2] and Per Lindahl*[1]

Address: [1]Sahlgrenska Academy, Department of medical and physiological biochemistry Box 440, SE-405 30 Göteborg, Sweden, [2]Chalmers Technical University, Department of mathematical statistics, Eklandagatan 76, SE-412 96 Göteborg, Sweden and [3]Lund Strategic Research Center for Stem Cell Biology and Cell Therapy, BMC B10, Klinikgatan 26, SE-221 48 Lund, Sweden

Email: Sven Nelander - sven.nelander@wlab.gu.se; Erik Larsson - erik.larsson@wlab.gu.se; Erik Kristiansson - erikkr@math.chalmers.se; Robert Månsson - Robert.Mansson@stemcell.lu.se; Olle Nerman - nerman@math.chalmers.se; Mikael Sigvardsson - mikael.sigvarsson@stemcell.lu.se; Petter Mostad - mostad@math.chalmers.se; Per Lindahl* - per.lindahl@wlab.gu.se

* Corresponding author

Published: 09 May 2005

BMC Genomics 2005, **6**:68 doi:10.1186/1471-2164-6-68

Received: 27 July 2004

Accepted: 09 May 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/68

Abstract

Background: The expression of *gene batteries*, genomic units of functionally linked genes which are activated by similar sets of cis- and trans-acting regulators, has been proposed as a major determinant of cell specialization in metazoans. We developed a predictive procedure to screen the mouse and human genomes and transcriptomes for cases of gene-battery-like regulation.

Results: In a screen that covered ~40 per cent of all annotated protein-coding genes, we identified 21 co-expressed gene clusters with statistically supported sharing of cis-regulatory sequence elements. 66 predicted cases of over-represented transcription factor binding motifs were validated against the literature and fell into three categories: (i) previously described cases of gene battery-like regulation, (ii) previously unreported cases of gene battery-like regulation with some support in a limited number of genes, and (iii) predicted cases that currently lack experimental support. The novel predictions include for example Sox 17 and RFX transcription factor binding sites that were detected in ~10% of all testis specific genes, and HNF-1 and 4 binding sites that were detected in ~30% of all kidney specific genes respectively. The results are publicly available at http://www.wlab.gu.se/lindahl/genebatteries.

Conclusion: 21 co-expressed gene clusters were enriched for a total of 66 shared cis-regulatory sequence elements. A majority of these predictions represent novel cases of potential co-regulation of functionally coupled proteins. Critical technical parameters were evaluated, and the results and the methods provide a valuable resource for future experimental design.

Background

To understand how gene expression is coordinated to produce hundreds of cell phenotypes from an identical complement of genes is a principal challenge in mammalian genome research. A commonly suggested model for terminal differentiation in metazoans is that the core features of the cellular phenotype are mediated by a set of genes that is regulated as a *gene battery*, i.e. a set of functionally coupled genes that are activated by similar cis- and trans-acting regulators [1-3]. Although the gene battery is an

idealized concept, concrete examples of gene battery-like regulation have been found in for example muscle subtypes [4-7], megakaryocytes [8], the epidermis [9] and lymphocytes [10,11].

A key step in the elucidation of gene battery-like regulation is to detect and functionally test the *cis* regulatory elements that mediate the co-regulation. A number of computation-based methods have been proposed to do this. In micro-organisms, computational methods have proven useful to detect modules of co-regulated genes [12,13]. In mammals, predictive models based on assumed co-regulation at the *cis* level have been constructed for liver- and skeletal muscle-selective gene regulation [14,15], and general tools have been developed for the regulatory analysis of co-expressed genes [16-18].

The aim of this work is to screen the mouse and human genomes and transcriptomes for instances where sharing of *cis*-regulatory sequences is statistically coupled to conserved co-expression of genes, i.e. cases that fall within or near the idealized gene battery concept. Another aim is to critically investigate technical parameters in order to maximize the sensitivity by which co-regulation of co-expressed genes can be detected. In a screen that covered ~40 per cent of all protein-coding genes according to the latest Ensembl annotation, we identified 21 co-expressed gene clusters with 66 cases of statistically supported sharing of cis-regulatory sequence motifs. The predictive value of the assignment of transcription factor binding sites was experimentally evaluated on EBF binding sites in a set of B-cell expressed gene clusters. The predicted cases of co-regulation include several previously known prototype examples of tissue specific regulation, but also novel predictions. All data are made available to the research community in the form of an internet resource that may serve as a starting point for further analysis.

Results

The analysis was based on the assumption that homologous genes in mouse and human are equivalent in most aspects of regulation and function. In particular, we assumed that the transcriptional regulation is conserved for orthologous genes. For example, the mouse gene *Myh1* and the human gene *MYH1* are assumed to share expression pattern and to share important cis-regulatory sequences. Below, the term 'ortholog pair' will be applied as a two-species equivalent of 'gene', for which expression and sequence data where retrieved for both mouse and human. Ensembl gene annotations [19,20] were applied thoughout the analysis.

Co-expressed gene sets were defined from a compendium of mouse and human expression data and tested for functional coupling

Previous results by our group and others have shown that statistical analysis of gene expression profiles in a large compendium of expression data can predict targets of differentiation processes, and identify functionally coupled genes [12,21,22]. In the following analysis, we specifically focused on a compendium derived from the recently completed Novartis expression atlas (SymAtlas) [23]. These data contain transcription profiles for 140 mouse and human tissues generated by hybridization on customized Affymetrix chips, and cover a large fraction of the mouse and human protein-coding genes. Sequence annotation of the Novartis probes linked the mouse and human data to approximately 17.000 unique Ensembl gene identities in each species (Table 1). Between the two datasets, 13282 non-redundant ortholog pairs could be identified by linkage of reciprocal Ensembl homology assignments (Table 1). In later steps, we excluded genes for which regulatory sequence could not be extracted (Methods), leaving 9561 ortholog pairs for clustering (Table 1). The final dataset included ~40% of the mouse and human Ensembl annotated genes.

Clustering

We clustered the mouse/human ortholog pairs based on their expression profiles across the 140 mouse and human tissues (Methods). We computed clusters at cut-off levels ranging from Pearson's correlation coefficient (hereafter termed PCC) 0.61 to 0.99. At the lowest applied cut-off, 57% of all ortholog pairs in the data were members of a cluster. The cluster sizes were distributed in a skewed manner, with a predominant formation of small clusters (Figure 1A). All analyses hereafter were performed at a PCC = 0.75 cut-off, which generated 160 clusters with 2407 ortholog pairs. This relatively stringent cut-off was chosen to reduce the number of non-relevant genes in the clusters. A higher cut-off did not seem reasonable given the noise-level of the microarray experiments (as judged by the Pearson correlation between replicated samples and between probes that are annotated for the same gene, data not shown).

Assessment of functional linkage

According to the definition, a gene battery should encode functionally linked proteins. We used Gene Ontology (GO) terms, protein-protein interactions (from the BIND database [24]), and manual curation to assess functional linkage within our expression clusters. First, we investigated the relationship between expression profile similarity of two ortholog pairs and their relative probability to share a functional annotation term or to encode interacting proteins (Figure 1B). There was a consistent correlation between co-expression and the relative probability

Table 1: Gene coverage of the analysis

	MOUSE	HUMAN	BOTH
Sequence data:			
Ensembl genes total:	23954	21961	
Ensembl transcripts total:	34076	35685	
Ortholog pairs of Ensembl genes:			20188
Ortholog pairs with upstream sequence extracted :			13272
Ortholog pairs with upstream sequence extracted (redundancies removed):			**12239** *
Expression data:			
Ensembl genes matching SymAtlas probes:	17552	16929	
ortholog pairs with expression data in both species:			**13282** **
Integrated dataset:			
Two-species expression data AND regulatory sequence:			**9561**

Numbers in the 'MOUSE' and 'HUMAN' columns signify the number of unique Ensembl identifiers in each respective species. Numbers in column 'BOTH' signify ortholog pairs of Ensembl entries. The overlap between the nonredundant sequence database (*) and the nonredundant expression dataset (**) was 9561 ortholog pairs.

for two genes to share a GO-term or to encode interacting proteins (Figure 1B).

Second, we studied the statistical over-representation of GO terms and interacting proteins inside the clusters (Methods). 30/32 clusters with ten or more ortholog pairs contained at least one over-represented GO term. The proportion of small clusters with over-represented GO terms was lower, which reflects a lack of statistical power in small clusters. Genes encoding interacting protein pairs were also over-represented inside clusters. 35 cases of protein-protein interaction between two genes *in the same cluster* were found. In contrast, 1000 simulations on permuted data revealed a median of 9 interactions (observations ranging between 5 and 21). The BIND data contained only 600 interactions that could be mapped to the dataset, which explains the seemingly low number of 35 interactions.

Finally, the clusters were annotated by manual curation. Six examples are shown in Figures 2, 3, 4, 5, 6, 7, and clusters with predicted regulators are listed in Table 2. For a full overview, see the web supplement http://www.wlab.gu.se/lindahl/genebatteries. Several clusters clearly represent gene sets that mediate *specialized features of different cell types*, including smooth muscle specific genes (cluster 40, Figure 2), B lymphocyte-specific genes (cluster 16, Figure 3), and genes selectively expressed in the testis (cluster 5, part shown in Figure 4). Further, we detected clusters that were related to *cellular processes or organelles*, including endoplasmatic reticulum (cluster 13, Figure 5), protein synthesis (cluster 1, Figure 6), and modulators of transcriptional regulation (cluster 65, Figure 7). A grand majority of the clusters were defined by peaks at different amplitudes in several tissues. As an example, the endoplasmatic reticulum cluster (Figure 5) was defined by a highly variant profile with strong expression in, for example, exocrine glands. Generally, the cluster profiles were conserved between species, in the sense that clusters were defined by peaks in the same organs. This effect was more pronounced for clusters with expression in a single organ, such as the testis (data not shown).

In combination, the GO term enrichment, the protein-protein interaction, and the manual curation convincingly show that clustered genes are functionally linked.

Regulatory DNA and descriptions of transcription factor binding sites were extracted and pre-processed

In the next part of the analysis, individual ortholog pairs were scored for transcription factor binding sites. Binding motifs were represented in the form of Position Frequency Matrices (hereafter denoted PFMs). Based on a fixed amount of upstream DNA sequence in each ortholog pair, a statistical score was computed to predict the potential for a site in the sequence to bind the factor corresponding to a PFM (described in detail in Methods). We extracted upstream + intronic sequence from the Ensembl database, in amounts of 2, 6 or 15 kb per gene (see Methods for details on boundaries). A filter was applied that removed ortholog pairs for which the transcription start differed between the two species (>1000 bp difference, see Methods). Filtering was successful for 12239 unique ortholog pairs (Table 1). In a subsequent filter, DNA sequence that was not conserved between mouse and human was removed, so-called phylogenetic footprinting.

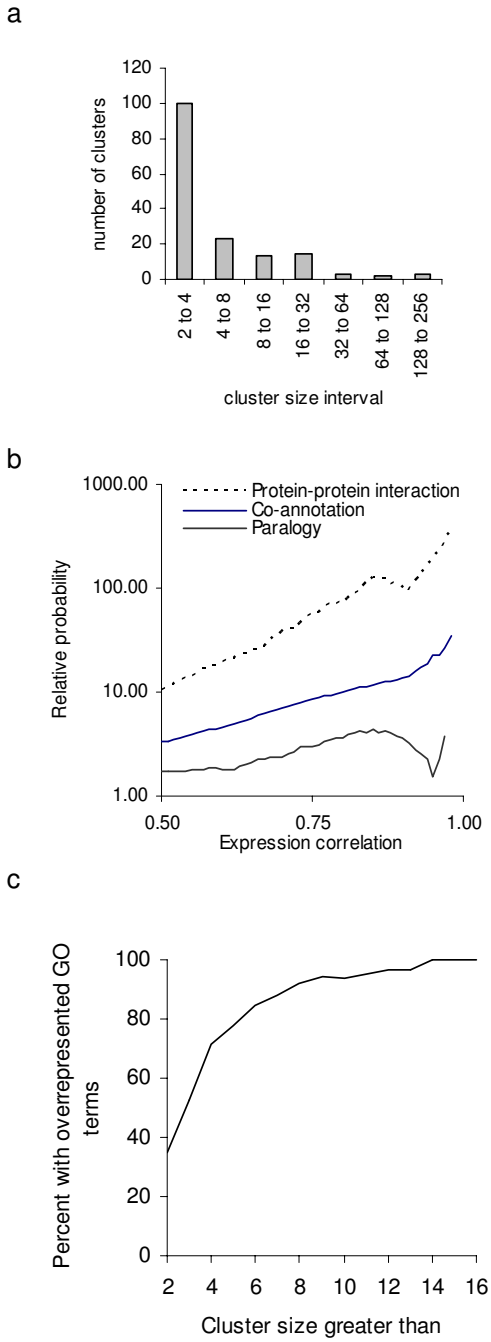


Figure 1

Cluster statistics A: Histogram showing the log number of clusters as a function of log cluster size, based on the clustering at Pearson correlation coefficient 0.75 cut-off. Numbers on the x axis denote cluster size intervals (2), (3–4), (5–8), (9–16),... B: Co-expression as a predictor for shared function, protein interaction and paralogy. We identified all gene pairs that correlated above or below a threshold T (X-axis). We measured the fraction of such pairs for which there was (i) a BIND database protein-protein interaction recorded in human, (ii) at least one shared gene ontology term, and (iii) evidence of paralogy. We then computed the relative probability for genes above T with this feature, compared to gene pairs below T. At expression correlation 0.80, co-expression was associated with a 100-fold relative probability for genes to encode protein interactors, a 10-fold probability for genes to share functional annotation, but only a 3-fold probability for genes to be paralogs. C: Fraction of clusters with at least one over-represented GO term (Y axis), as a function of cluster size (X axis). GO term over-representations were computed at a 10% false discovery rate.

Table 2: over-represented motifs detected at <10% false discovery rate

Cluster number	FDR	PFM Number	PFM Annotation
1: Protein synthesis	<2.5%	190	M00025:Elk-1, M00007:Elk-1
	<2.5%	110	M00050:E2F, MA0024:E2F
	<2.5%	57	M00108:NRF-2, MA0028:Elk-1, MA0062:NRF-2
	<2.5%	181	MA0076:SAP-1
	<10%	18	M00074:c-Ets-1(p54)
	<10%	78	M00262:Staf
2: Oocyte / fertilized egg	<2.5%	71	M00024:E2F
	<2.5%	190	M00025:Elk-1, M00007:Elk-1
	<2.5%	9	M00032:c-Ets-1(p54)
	<2.5%	110	M00050:E2F, MA0024:E2F
	<2.5%	57	M00108:NRF-2, MA0028:Elk-1, MA0062:NRF-2
	<2.5%	181	MA0076:SAP-1
	<10%	238	MA0088:Staf, M00264:Staf
3: Neural tissues	<2.5%	99	M00189:AP-2
	<2.5%	115	M00196:Sp1
	<2.5%	141	M00256:NRSF
	<10%	75	M00243:Egr-1
4: Lymphocytes	<2.5%	143	MA0050:Irf-1, M00062:IRF-1, M00063:IRF-2
	<10%	74	M00054:NF-kappaB, MA0061:NF-kappaB
	<10%	28	M00258:ISRE
5: Testis / spermatogenesis	<2.5%	109	M00281:RFX1
	<2.5%	142	MA0078:SOX17
	<10%	108	M00036:v-Jun
	<10%	248	M00041:CRE-BP1/c-Jun
	<10%	65	M00100:CdxA
6: Liver	<2.5%	16	M00134:HNF-4
	<2.5%	212	M00158:COUP-TF / HNF-4, MA0017:COUP-TF
	<2.5%	33	M00206:HNF-1
	<2.5%	203	MA0046:HNF-1, M00132:HNF-1
	<2.5%	234	MA0047:HNF-3beta, M00131:HNF-3beta
	<2.5%	113	MA0065:PPARgamma-RXRal
	<10%	46	M00155:ARP-1
	<10%	212	M00158:COUP-TF / HNF-4, MA0017:COUP-TF
	<10%	146	MA0071:RORalfa-1, M00156:RORalpha1
8: ECM	<10%	215	M00378:Pax-4
9: Cardiac muscle	<2.5%	223	M00026:RSRFC4
	<2.5%	144	M00152:SRF
	<2.5%	59	M00231:MEF-2
	<2.5%	222	M00232:MEF-2
	<2.5%	161	M00252:TATA
	<2.5%	259	M00418:TGIF, M00419:MEIS1
	<2.5%	160	MA0052:MEF2
	<10%	60	M00006:MEF-2
12: Skeletal muscle	<2.5%	201	M00184:MyoD, M00001:MyoD
	<10%	17	M00002:E47
	<10%	59	M00231:MEF-2
13: Endoplasmatic reticulum	<10%	190	M00025:Elk-1, M00007:Elk-1
	<10%	57	M00108:NRF-2, MA0028:Elk-1, MA0062:NRF-2
	<10%	181	MA0076:SAP-1

Table 2: over-represented motifs detected at <10% false discovery rate *(Continued)*

15: Erythrocyte	<10%	209	M00128:GATA-1, M00127:GATA-1
	<10%	122	M00203:GATA-X
	<10%	198	M00413:AREB6
16: B lymphocyte	<2.5%	133	MA0081:SPI-B
17: Kidney	<2.5%	33	M00206:HNF-1
	<2.5%	188	M00411:HNF-4alpha1
22: Cell cycle genes	<10%	110	M00050:E2F, MA0024:E2F
24: Pancreas	<10%	121	M00071:E47
	<10%	193	M00080:Evi-1, M00082:Evi-1
30: Small intestine	<10%	31	M00346:GATA-1, M00347:GATA-1, M00348:GATA-2
40: Smooth muscle	<2.5%	144	M00152:SRF
	<2.5%	245	M00186:SRF, M00215:SRF
	<2.5%	88	MA0083:SRF
44: Retina	<2.5%	196	M00087:Ik-2
45: Testis (mouse signal only)	<2.5%	164	M00253:cap
49: Lung/endothelium (mouse signal only)	<10%	66	M00199:AP-1, M00037:NF-E2
65: NfkappaB signalling	<2.5%	235	M00051:NF-kappaB (p50), MA0105:p50

Over-represented motifs arranged by cluster number. FDR column: False Discovery Rate (estimated probability for the over-representation to be a spurious detection). Motifs are are shown both by their numerical identifiers (PFM number) and by their annotation (PFM annotation). In cases where a PFM is a composite based on more than one source, the components are given separated by commas. The data where generated from the PCC = 0.75 clustering, 2 kb sequence database, at 90% phylogenetic conservation.

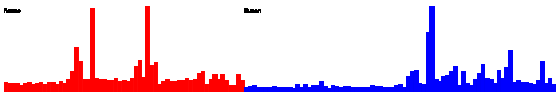
Phylogenetic footprinting was applied at different stringency, to allow the following optimization of the protocol (below). Furthermore, all exon sequence was removed from the analysis (Methods). Finally, the sequences were matched with the expression data based on annotation, the overlap being 9561 ortholog pairs (Table 1).

322 vertebrate PFMs were downloaded from the TRANSFAC and JASPAR databases [25,26]. Since the databases appeared to contain redundant or equivalent entries, highly similar PFMs were grouped and merged using single linkage hierarchical clustering and a PFM distance measure defined in [27], which reduced the number of PFMs from 322 to 266 (Methods). This step reduced the redundancy, but did not merge all identically annotated PFMs (see for example the redundant serum response factor (SRF) PFMs in Table 2, cluster 40).

Design of a predictive scoring system for transcription factor binding

After the retrieval and preprocessing of both sequences and PFMs, all individual sequences were tested for PFM matches using the MAST software [28], a software for identifying single or multiple motifs in sequences. MAST was set to compute one p-value for each PFM with respect to each sequence (Methods). Based on the p-values obtained from the MAST software, a *composite score* was defined as the product of the p-value in the mouse and human sequences of an ortholog pair (Methods). A *composite* score close to 0 indicates that both the mouse and human promoter sequence in the ortholog pair contains sequence elements that are in very good agreement with a certain PFM.

To address the biological validity of the MAST composite scores within the context of a set of co-expressed genes, we screened 48 ortholog pairs present in B-cell expressed clusters for individual EBF sites. In all, 24 individual EBF



Bonferroni P-value	Cluster coverage	Dataset coverage	GO term
1.00E+00	43% (3/7)	(56/9561)	motor activity
1.00E+00	29% (2/7)	(10/9561)	actin filament
1.00E+00	29% (2/7)	(23/9561)	muscle development
1.00E+00	29% (2/7)	(24/9561)	structural constituent of cytoskeleton

#	PFM	Cluster coverage	Dataset coverage	FDR
1	MA0083:SRF	86% (6/7)	1% (89/9561)	<2.5%
2	M00186:SRF|M00215:SRF	86% (6/7)	1% (114/9561)	<2.5%
3	M00152:SRF	71% (5/7)	1% (135/9561)	<2.5%
4	M00216:TATA	43% (3/7)	2% (220/9561)	<20%
5	M00059:YY1	43% (3/7)	3% (284/9561)	<20%
6	MA0090:TEF-1	43% (3/7)	3% (296/9561)	<20%

1	2	3	4	5	6	gene (MM)	gene (HS)	Annotation (mouse)
■	■	■				Actg2	ACTG2	ACTIN, GAMMA 2, SMOOTH MUSCLE, ENTERIC
■	■	■	■	■		Q8C3J0	MYH11	MYOSIN HEAVY CHAIN 11
■	■	■		■		Lpp	LPP	LIPOMA PREFERRED PARTNER
■	■	■	■			Tagln	TAGLN	TRANSGELIN (SMOOTH MUSCLE PROTEIN 22-ALPHA)
■	■	■		■	■	Acta2	ACTA2	ACTIN, AORTIC SMOOTH MUSCLE (ALPHA-ACTIN 2)
■	■				■	Myl9	MYL9	MYL9 PROTEIN (FRAGMENT)
			■		■	Lmod1	LMOD1	LEIOMODIN 1 (SMOOTH MUSCLE)

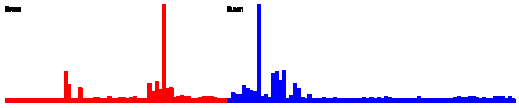
Figure 2
A smooth muscle differentiation battery: The bar chart (left) illustrates the average expression level of cluster members (Y axis) across arbitrarily ordered tissues (X axis) in two species (red = mouse and blue = human). Three tables list over-represented functional terms (upper small table), over represented motifs (PFMs) (middle table), and cluster members (lower table).

sites in 15 different ortholog pairs were detected (supplementary data, additional file 1). To test the functionality (in terms of EBF binding) of these sites, we examined the potential of 22 basepair duplex oligo-nucleotides spanning the presumed sites to compete for protein binding in Electrophoretic Mobility Shift Assays (EMSA:s). EBF binding capacity was assayed using nuclear extracts from the pre-B cell line 40-EI, a labelled mouse mb-1 promoter high affinity EBF site [29], and competitor oligo-nucleotides covering the new potential binding sites. In the absence of competitor oligo-nucleotide, a prominent DNA/protein complex (mb-1/EBF) could be detected whereas this complex was undetectable after the addition of a 300 fold molar excess of the unlabeled binding site (mb-1). The identity of the protein and the specificity of the binding were verified by competition with a point mutated mb-1/EBF site and by the inclusion of an EBF specific antibody into the reaction mixture (Figure 8B). The point mutated EBF binding site was unable to abolish complex formation even in a 1000-fold molar excess (Figure 8B), indicating that we detect specific protein DNA interactions with this experimental set up. 18 out of the 24 new binding sites competed for complex formation when added in a 300- or 1000-fold molar excess, and thus have the ability to bind EBF in vitro (Figure 8A). We conclude that the large majority of binding sites were able to functionally interact with the predicted protein, and that the composite score in principle detected factor binding. It should however be emphasized that the quality of predictions is dependent on the quality of the binding site descriptions, and the result does not necessarily imply that other predicted factors bind.

A statistical procedure was used to detect enriched motifs in the clusters

To test whether the identified clusters represent potential gene batteries, i.e. contain shared *cis-regulatory elements*, we designed a procedure to detect significant over-representation of orthologs that match a PFM inside a cluster. The procedure is based on a modification of Fisher's exact test, which tests for dependency between two events (in this case cluster membership vs detection of a motif) [30]. We introduced a procedure to optimize the composite score thresholds for individual PFMs. In brief, we selected the threshold that gave the lowest Fisher test p-value in any one cluster. This was based on the assumption that non-random distribution of detections over clusters

#	PFM	Cluster coverage	Dataset coverage	FDR
1	MA0081:SPI-B	25% (6/24)	3% (241/9561)	<2.5%
2	M00414:AREB6	29% (7/24)	6% (549/9561)	<20%
3	M00210:OCT-x|M00161:Oct-1	8% (2/24)	0% (16/9561)	<20%



P-value	Cluster coverage	Dataset coverage	GO term
7.70E-01	54% (13/24)	(1233/9561)	integral to membrane
1.00E+00	13% (3/24)	(22/9561)	external side of plasma membrane
1.00E+00	8% (2/24)	0% (4/9561)	C-X-C chemokine receptor activity
1.00E+00	8% (2/24)	0% (5/9561)	B-cell activation
1.00E+00	13% (3/24)	(37/9561)	heterophilic cell adhesion
1.00E+00	25% (6/24)	(333/9561)	receptor activity
1.00E+00	8% (2/24)	0% (8/9561)	lymph gland development
1.00E+00	8% (2/24)	0% (8/9561)	protein coupled
1.00E+00	17% (4/24)	(134/9561)	rhodopsin-like receptor activity
1.00E+00	13% (3/24)	(65/9561)	sugar binding

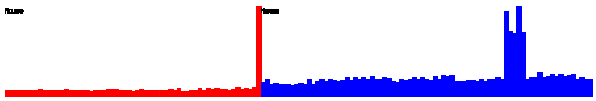
1	2	3	gene (MM)	gene (HS)	Annotation (mouse)
■	■	■	Vpreb3	VPREB3	PRE-B LYMPHOCYTE GENE 3
	■		Cd79a	CD79A	MB-1
			Cd79b	CD79B	B-CELL-SPECIFIC GLYCOPROTEIN B29
■			BB219290	NM_032738	FC RECEPTOR HOMOLOG EXPRESSED IN B CELLS
■	■		Cd19	CD19	B-LYMPHOCYTE ANTIGEN CD19 PRECURSOR
			Cd22	CD22	B-LYMPHOCYTE CELL ADHESION MOLECULE
			Ms4a1	MS4A1	B-CELL SURFACE PROTEIN CD20 HOMOLOG
			Pou2af1	POU2AF1	POU DOMAIN CLASS 2, ASSOCIATING FACTOR 1
			Swap70	NM_015055	SWAP COMPLEX PROTEIN
			A530094C12Rik	BANK1	A530094C12Rik WEAKLY SIMILAR TO BANK (FRAGMENT)
			A230020G22Rik	NM_052938	A230020G22Rik FCRH1-PENDING.
			Blr1	BLR1	C-X-C CHEMOKINE RECEPTOR TYPE 5 (CXC-R5)
			Tnfrsf13c	TNFRSF13C	TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY MEMBER 13C
			Cd72	CD72	B-CELL DIFFERENTIATION ANTIGEN LYB-2 (CD72).
			Fcer2a	FCER2	LOW AFFINITY IMMUNOGLOBULIN EPSILON FC RECEPTOR
■	■		AI586015	NM_012108	BCR DOWNSTREAM SIGNALING 1
	■		Ltb	LTB	LYMPHOTOXIN-BETA (LT-BETA) (TUMOR NECROSIS FACTOR C
■	■		1810037B05Rik	NM_005449	1810037B05Rik
			Gpr18	GPR18	G PROTEIN-COUPLED RECEPTOR 18
			P2ry10	P2RY10	PURINERGIC RECEPTOR P2Y, G-PROTEIN COUPLED 10
	■		Cxcr4	CXCR4	PRE-B-CELL-DERIVED CHEMOKINE RECEPTOR (PB-CKR)
■			Cd83	CD83	Cd83 CD83 ANTIGEN
			Plcg2	PLCG2	PHOSPHOLIPASE C, GAMMA 2
		■	Ly78	LY64	CD180 ANTIGEN PRECURSOR (LYMPHOCYTE ANTIGEN 78)

Figure 3
B-lymphocyte differentiation battery: Tables and charts are organized as in figure 2.

reflects biological function, as has been proposed in [31]. The Fisher test p-values with the optimized thresholds are hereafter termed *p-scores*.

The tests of multiple detection thresholds, multiple clusters and 266 PFMs led to a need to compensate for mass testing. This was done by estimating false discovery rates (FDR) based on simulations on randomized data. In brief,

Bonferroni P-value	Cluster coverage	Dataset coverage	GO term
3.30E-10	8% (15/185)	(42/9561)	spermatogenesis
9.40E-01	2% (4/185)	0% (6/9561)	flagellum (sensu Eukarya)
1.00E+00	2% (3/185)	0% (6/9561)	biogenesis
1.00E+00	2% (3/185)	0% (7/9561)	sperm motility
1.00E+00	2% (3/185)	0% (9/9561)	activity
1.00E+00	2% (3/185)	0% (9/9561)	microtubule
1.00E+00	2% (4/185)	(23/9561)	nucleosome
1.00E+00	2% (3/185)	(11/9561)	chromosome
1.00E+00	2% (4/185)	(27/9561)	actin cytoskeleton
1.00E+00	1% (2/185)	0% (4/9561)	cytoplasmic dynein complex



#	PFM	Cluster coverage	Dataset coverage	FDR
1	MA0078:SOX17	8% (15/185)	2% (165/9561)	<2.5%
2	M00281:RFX1	11% (21/185)	4% (398/9561)	<2.5%
3	M00036:v-Jun	10% (19/185)	4% (361/9561)	<10%
4	M00100:CdxA	2% (4/185)	0% (16/9561)	<10%
5	M00041:CRE-BP1/c-Jun	8% (15/185)	3% (271/9561)	<10%
6	M00280:RFX1	6% (12/185)	2% (212/9561)	<20%

1	2	3	4	5	6	gene (MM)	gene (HS)	Annotation (mouse)
■						Stk22c	STK22C	TESTIS-SPECIFIC KINASE 3
		■				1700095G12Rik	NM_199339	1700095G12Rik
	■				■	Txnrd3	TXNRD3	THIOREDOXIN REDUCTASE 3
						Klhl10	KLHL10	KELCH-LIKE 10
						Spa17	SPA17	SPERM SURFACE PROTEIN SP17
					■	4931407G18Rik	NM_033122	SIMILAR TO TESTIS DEVELOPMENT PROTEIN NYD-SP26
					■	Pdcl2	PDCL2	PHOSDUCIN-LIKE PROTEIN 2
						Tnp1	TNP1	SPERMATID NUCLEAR TRANSITION PROTEIN 1
						4930532L20Rik	ANKRD7	4930532L20Rik WEAKLY SIMILAR TO ANKYRIN MOTIF
	■			■		Akap3	AKAP3	A-KINASE ANCHOR PROTEIN 3
						Prm2	PRM2	SPERM HISTONE P2 PRECURSOR (PROTAMINE MP2)
■		■				Mcsp	MCSP	SPERM MITOCHONDRIAL ASSOCIATED CYSTEINE-RICH PROTEIN
						1700001N01Rik	TSR2_HUMAN	TESTIS SPERMATOCYTE APOPTOSIS-RELATED GENE 2 PROTEIN
		■				Cappa3	CAPZA3	F-ACTIN CAPPING PROTEIN ALPHA-3 SUBUNIT
						Tenr	NM_139243	TESTIS NUCLEAR RNA BINDING PROTEIN
				■		Akap4	AKAP4	A-KINASE ANCHOR PROTEIN 4 PRECURSOR
						Plcz1	PLCZ1	PHOSPHOLIPASE C, ZETA 1; PHOSPHOLIPASE C ZETA.
						Ddx4	DDX4	DEAD-BOX PROTEIN 4
					■	Gk-rs2	GK2	GLUCOKINASE ACTIVITY, RELATED SEQUENCE 2
	■					1700013F07Rik		1700013F07Rik
						Cct6b	CCT6B	T-COMPLEX PROTEIN 1, ZETA-2 SUBUNIT
								(REMAINING GENES SHOWN IN SUPPLEMENTARY)

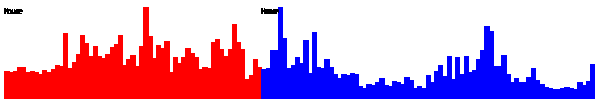
Figure 4
Testis selective battery: Over representation of RFX and SOX17 motifs indicates new roles for these factors as coordinators of testis selective gene expression. Tables and charts are organized as in figure 2.

we compared the outcome when using permuted and when using observed data at different p-scores, and defined the FDR as the ratio between the two (Figure 9A). This procedure allowed us to choose a significance threshold with a controlled expected number of spurious detections. Simulations were repeated 100 times. Since optimization of detection thresholds were repeated in each simulation round, no bias in disfavour of the control case was introduced.

The detection of over-represented PFMs was affected by DNA amount and masking, but not affected by gene paralogy

Using the described algorithm, we analyzed the number of times a motif was over-represented when using different amounts of DNA sequence per gene (2, 6 or 15 kb) and different stringency in the phylogenetic footprinting (>0% (keep all sequence), >60%, >70%, >80% or >90% identity). The analysis was performed at 10% false discovery rate and with the clustering obtained at PCC = 0.75 clustering cut-off. At all DNA amounts, higher stringency

Bonferroni P-value	Cluster coverage	Dataset coverage	GO term
6.70E-08	38% (11/29)	(142/9561)	endoplasmic reticulum
1.00E+00	21% (6/29)	(152/9561)	intracellular protein transport
1.00E+00	14% (4/29)	(43/9561)	isomerase activity
1.00E+00	17% (5/29)	(126/9561)	Golgi apparatus
1.00E+00	17% (5/29)	(130/9561)	protein transporter activity
1.00E+00	10% (3/29)	(72/9561)	electron transporter activity
1.00E+00	7% (2/29)	(26/9561)	GTPase activity
1.00E+00	31% (9/29)	(1233/9561)	integral to membrane
1.00E+00	10% (3/29)	(162/9561)	GTP binding
1.00E+00	7% (2/29)	(67/9561)	protein folding



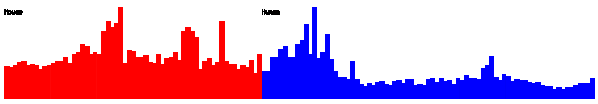
#	PFM	Cluster coverage	Dataset coverage	FDR
1	M00025:Elk-1|M00007:Elk-1	38% (11/29)	(911/9561)	<10%
2	1|MA0062:NRF-2	34% (10/29)	8% (785/9561)	<10%
3	MA0076:SAP-1	31% (9/29)	7% (711/9561)	<10%
4	M00251:XBP-1	10% (3/29)	0% (46/9561)	<20%
5	M00287:NF-Y	34% (10/29)	(999/9561)	<20%
6	M00032:c-Ets-1(p54)	31% (9/29)	9% (870/9561)	<20%

1	2	3	4	5	6	gene (MM)	gene (HS)	Annotation (mouse)
	■					Ssr4	SSR4	TRANSLOCON-ASSOCIATED PROTEIN, DELTA SUBUNIT PRECURSOR
■	■	■		■	■	P4hb	P4HB	PROTEIN DISULFIDE ISOMERASE PRECURSOR (PDI)
■		■			■	Sec61a1	SEC61A1	PROTEIN TRANSPORT PROTEIN SEC61 ALPHA SUBUNIT ISOFORM 1
						Ppib	PPIB	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE B PRECURSOR
						Il25	C19orf10	LYMPHOCYTE ANTIGEN 6 COMPLEX, LOCUS E LIGAND
■		■		■	■	Tra1	TRA1	ENDOPLASMIC RETICULUM PROTEIN 99
				■		Hspa5	HSPA5	78 KDA GLUCOSE-REGULATED PROTEIN PRECURSOR (GRP 78)
				■		Grp58	GRP58	PROTEIN DISULFIDE ISOMERASE A3 PRECURSOR
	■			■		1110032D12Rik	P24_HUMAN	COP-COATED VESICLE MEMBRANE PROTEIN P24 PRECURSOR (P24A)
						2310008M10Rik	NM_021227	2310008M10Rik
							SEC61G	
■	■	■		■	■	Ddost	DDOST	GLYCOSYLTRANSFERASE 48 KDA SUBUNIT PRECURSOR
■	■	■	■		■	Copb2	COPB2	COATOMER BETA' SUBUNIT
						Ssr1	SSR1	SIGNAL SEQUENCE RECEPTOR, ALPHA
■	■	■			■	Ssr2	SSR2	SIGNAL SEQUENCE RECEPTOR, BETA
						02-sep	NEDD5	SEPTIN 2 (NEDD5 PROTEIN)
						Armet	ARMET	ARMET PROTEIN PRECURSOR
			■	■		Mgat2	MGAT2	MANNOSIDE ACETYLGLUCOSAMINYLTRANSFERASE 2
	■					Alg5	ALG5	DOLICHYL-PHOSPHATE BETA-GLUCOSYLTRANSFERASE
				■		Cai	PDA4_HUMAN	PROTEIN DISULFIDE ISOMERASE A4 PRECURSOR
■	■	■			■	Arf4	ARF4	ADP-RIBOSYLATION FACTOR 4
						Tmem4	TMEM4	TRANSMEMBRANE PROTEIN 4; PUTATIVE SECRETED PROTEIN ZSIG9
■	■	■	■		■	1810073M23Rik	T9S3_HUMAN	TRANSMEMBRANE 9 SUPERFAMILY PROTEIN MEMBER 3 PRECURSOR
■	■	■			■	Cope	COPE	COATOMER EPSILON SUBUNIT
						JTB_MOUSE	JTB	JTB PROTEIN PRECURSOR
■				■		Srpr	SRPR	SIGNAL RECOGNITION PARTICLE RECEPTOR ('DOCKING PROTEIN')
■						0610006K04Rik	NM_032344	0610006K04Rik
				■		Sec23b	SEC23B	PROTEIN TRANSPORT PROTEIN SEC23B
						H47	SELS_HUMAN	SELENOPROTEIN S (MINOR HISTOCOMPATIBILITY ANTIGEN H47)

Figure 5
Endoplasmatic reticulum associated genes: Over representation of XBP-1, NRF and RTS motifs suggest novel functions for NRF and ETS family factors in the regulation of ER-related genes. Tables and charts are organized as in figure 2.

phylogenetic footprinting appeared to be beneficial, and between DNA amounts, 2 kb and 4 kb compared favourably over 15 kb (Figure 9C). We conclude that optimal results are obtained when using a limited amount of sequence per gene (see Discussion).

A potential confounding factor in the analysis is that similarity in upstream sequence may be attributable to factors other than shared cis-regulatory elements. The most important such factor is likely to be gene paralogy, since over-represented motifs might simply represent matches to non-functional (not yet diverged) sequence in a co-

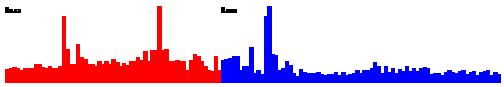


Bonferroni P-value	Cluster coverage	Dataset coverage	GO term
7.40E-46	78% (32/41)	(152/9561)	protein biosynthesis
2.60E-31	54% (22/41)	(78/9561)	ribosome
1.70E-26	63% (26/41)	(247/9561)	intracellular
2.10E-06	68% (28/41)	(104/9561)	structural constituent of ribosome
1.70E-05	12% (5/41)	0% (8/9561)	Eukarya)
1.10E-04	24% (10/41)	(148/9561)	RNA binding
1.60E-02	10% (4/41)	(10/9561)	small ribosomal subunit
8.00E-02	7% (3/41)	0% (4/9561)	Eukarya)
1.10E-01	10% (4/41)	(15/9561)	translational elongation
1.00E+00	7% (3/41)	(15/9561)	translation elongation factor activity

#	PFM	Cluster coverage	Dataset coverage	FDR
1	1|MA0062:NRF-2	34% (14/41)	8% (785/9561)	<2.5%
2	M00025:Elk-1|M00007:Elk-1	37% (15/41)	10% (911/9561)	<2.5%
3	MA0076:SAP-1	22% (9/41)	3% (331/9561)	<2.5%
4	M00032:c-Ets-1(p54)	24% (10/41)	6% (567/9561)	<10%
5	M00141:Lyf-1	32% (13/41)	11% (1065/9561)	<20%
6	MYC|MA0093:USF	5% (2/41)	0% (9/9561)	<20%

1	2	3	4	5	6	gene (MM)	gene (HS)	Annotation (mouse)
						Rpl13a	RPL13A	60S RIBOSOMAL PROTEIN L13A
				■			RPL14	
				■		Rpl35	RPL35	RIBOSOMAL PROTEIN L35
■				■		Rpl8	RPL8	60S RIBOSOMAL PROTEIN L8
						Rplp1	RPLP1	60S ACIDIC RIBOSOMAL PROTEIN P1
	■					Rpl28	RPL28	60S RIBOSOMAL PROTEIN L28
■				■		2210402A09Rik	RPS10	40S RIBOSOMAL PROTEIN S10
■	■	■	■			Rps18	RPS18	40S RIBOSOMAL PROTEIN S18
						2410030A14Rik	RPS21	40S RIBOSOMAL PROTEIN S21
■	■	■	■	■		Rpl19	RPL19	60S RIBOSOMAL PROTEIN L19
						Rpl39	RPL39	60S RIBOSOMAL PROTEIN L39
■	■					Rps3a	RPS3A	40S RIBOSOMAL PROTEIN S3A
						Rps4x	RPS4Y	40S RIBOSOMAL PROTEIN S4
■						Rps16	RPS16	40S RIBOSOMAL PROTEIN S16
■	■	■	■	■		Rpl6	RPL6	60S RIBOSOMAL PROTEIN L6
	■					Rps3	RPS3	40S RIBOSOMAL PROTEIN S3
	■			■		2010004J23Rik	RPL4	60S RIBOSOMAL PROTEIN L4 (L1)
■	■	■	■	■		Rps7	RPS7	40S RIBOSOMAL PROTEIN S7 (S8)
						Rpl12	RPL12	60S RIBOSOMAL PROTEIN L12
						Rpl27	RPL27	RIBOSOMAL PROTEIN L27
						Fau	FAU	UBIQUITIN-LIKE PROTEIN FUBI
	■					Naca	NACA	NASCENT POLYPEPTIDE-ASSOCIATED COMPLEX ALPHA POLYPEPTIDE
■	■	■	■			Eef1b2	EEF1B2	ELONGATION FACTOR 1-BETA (EF-1-BETA)
				■		Rps8	RPS8	40S RIBOSOMAL PROTEIN S8
■	■	■	■			Gnb2	GNB2L1	GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT-LIKE PROTEIN
■	■		■			Rps14	RPS14	40S RIBOSOMAL PROTEIN S14 (PRO2640)
						Rpl23	RPL23	60S RIBOSOMAL PROTEIN L23 (L17)
						Eef1a1	EEF1A1	ELONGATION FACTOR 1-ALPHA 1 (EF-1-ALPHA-1)
■	■	■	■				RPL30	
					■	Fbl	FBL	FIBRILLARIN (NUCLEOLAR PROTEIN 1)
				■		Rpl9	RPL9	60S RIBOSOMAL PROTEIN L9
				■		Rps23	RPS23	40S RIBOSOMAL PROTEIN S23
					■	Eif3s5	EIF3S5	EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SUBUNIT 5
■	■	■	■			Eif3s3	EIF3S3	EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SUBUNIT 3
				■		0610025G13Rik	RPL38	60S RIBOSOMAL PROTEIN L38
						Rpl36a	RPL36A	60S RIBOSOMAL PROTEIN L44 (L36A)
						Eif3s6ip	EIF3S6IP	EUKARYOTIC TRANSLATION INITIATION FACTOR 3, SUBUNIT 6 INT PROT
						Eif3s7	EIF3S7	EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SUBUNIT 7
						Gtf3a	GTF3A	GENERAL TRANSCRIPTION FACTOR III A
■	■	■	■			Rps5	RPS5	40S RIBOSOMAL PROTEIN S5
				■		Eef2	EEF2	ELONGATION FACTOR 2 (EF-2)

Figure 6
Ribosomal genes: Tables and charts are organized as in figure 2.



P-value	Cluster coverage	Dataset coverage	GO term
1.00E+00	67% (2/3)	(358/9561)	transcription factor activity
1.00E+00	67% (2/3)	(677/9561)	regulation of transcription, DNA-dependent
1.00E+00	67% (2/3)	(1300/9561)	nucleus

#	PFM	Cluster coverage	Dataset coverage	FDR
1	(p50)|MA0105:p50	100% (3/3)	2% (215/9561)	<2.5%
2	M00053:c-Rel|MA0101:c-REL	67% (2/3)	0% (46/9561)	<10%
3	M00208:NF-kappaB	100% (3/3)	4% (399/9561)	<10%
4	(p65)|MA0107:p65	67% (2/3)	2% (144/9561)	<20%

1	2	3	4	gene (MM)	gene (HS)	Annotation (mouse)
				Relb	RELB	TRANSCRIPTION FACTOR RELB
				Nfkbie	NFKBIE	NF-KAPPAB INHIBITOR EPSILON (NF-KAPPA-BIE)
				Nfkb2	NFKB2	CELLS 2

Figure 7
NF-kappaB pathway: Over representation of REL and NFkappaB motifs indicates feed back signalling. Tables and charts are organized as in figure 2.

expressed gene family. In total, 22343 cases of pairwise paralogy were detected in the dataset of 9561 ortholog pairs using BLAST. All genes with transcripts matches at a BLAST E-value less than 1.0 were defined as potential paralogs. Of the 2407 ortholog pairs that clustered, 219 genes had a paralog inside the same cluster. This was clearly an over-representation since filling clusters with random genes from the dataset produced an average of 122 genes with a paralog inside a cluster (values between 104 to 131 observed in 10 randomization rounds). In a side-by-side comparison, we therefore analyzed the effect of removing paralogs from a cluster before testing for over-represented motifs (see Methods), as opposed to not removing paralogs but the same number of randomly selected genes. The results showed no increase in the number of over-represented motifs when allowing paralogs in the same cluster (Figure 9B).

66 over-represented motifs were observed in 21 clusters

The final analysis was performed with 2 kb sequence and 80% phylogenetic footprinting (results generated with variations of these parameters are found in the web supplement). Again, the clustering generated at PCC = 0.75 was chosen for analysis. The over-representation algorithm was run in the above cases, using 100 simulations to estimate false discovery rate thresholds. Over-represented motifs with FDR:s less than 2.5% and 10% were recorded. Key features of these results are presented in Tables 2 and 3, and the complete results are available by web browser http://www.wlab.gu.se/lindahl/genebatteries.

A predictive screen performed with these settings associated 66 motifs to co-expression in a total of 21 clusters at FDR thresholds 2.5% and 10% (Table 2). There was a clear tendency for motifs to be detected in most large clusters, and for smaller clusters to lack over-represented motifs. This can be directly explained by the fact that even a high degree of motif presence in a small cluster can be attributed to spurious detections, and that the Fisher test accounts for that. The over-represented motifs were validated against the literature, and fell into three main categories: [1)] previously described cases of gene-battery-like regulation, [2)]previously unreported cases of gene battery-like regulation with some support in a limited number of genes, and [3)] hypothetical cases of gene battery regulation (Table 3, Discussion).

In the typical case, an over-represented motif did not cover all cluster members but rather a fraction. The average coverage was 15%, with observations ranging from 1% to 100%. The limited coverage can be exemplified by the relatively large (104 ortholog pairs) cluster of liver-selective genes, which contained 8 over-represented motifs at FDR<10%, where the PFM annotation implied HNF1, HNF4, ARP1, PPAR-γ, and COUP-TF as the binding factors (Table 2). These different motifs were detected in between 8 and 19 ortholog pairs, indicating a cluster coverage of less than 20%. An example of very high coverage was observed in the cluster that contained smooth muscle differentiation markers, in which 6/7 (86%) ortholog pairs were positive for a Serum Response Factor (SRF) motif. We compared our results for HNF-1 to

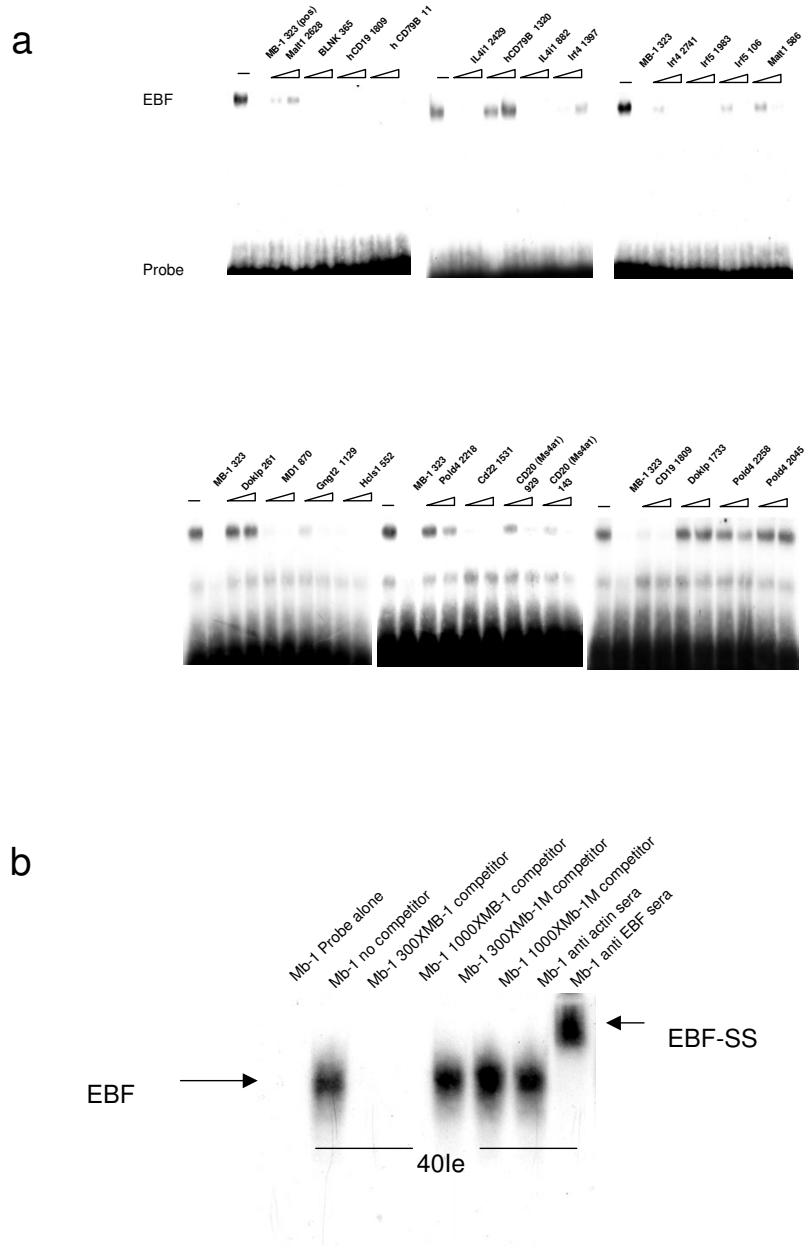


Figure 8
EMSA validation of EBF binding sites: A: The figure displays EMSAs in which binding of EBF to a mb-1 promoter EBF site is competed for by the inclusion of 300 or 1000-fold molar excess of unlabelled oligonucleotides that correspond to the predicted motifs. The name of the gene and the position of the motif is given in the figure. (m) indicates mouse and (h) human. "EBF" shows the position of the DNA/protein complex, and "Probe" indicates the position of free DNA. See supplementary information for a detailed description of the sites. B: The mb-1 promoter EBF site interacts specifically with EBF protein in a pre-B cell nuclear extract. The figure displays an autoradiogram in which a labelled EBF binding site from the mb-1 promoter has been incubated with nuclear extracts from 40EI pre-B cells and competitors or antibodies as indicated. EBF denotes the bound EBF protein and EBF-SS the super-shifted complex obtained by the addition of the EBF reactive antibody to the reaction mixture.

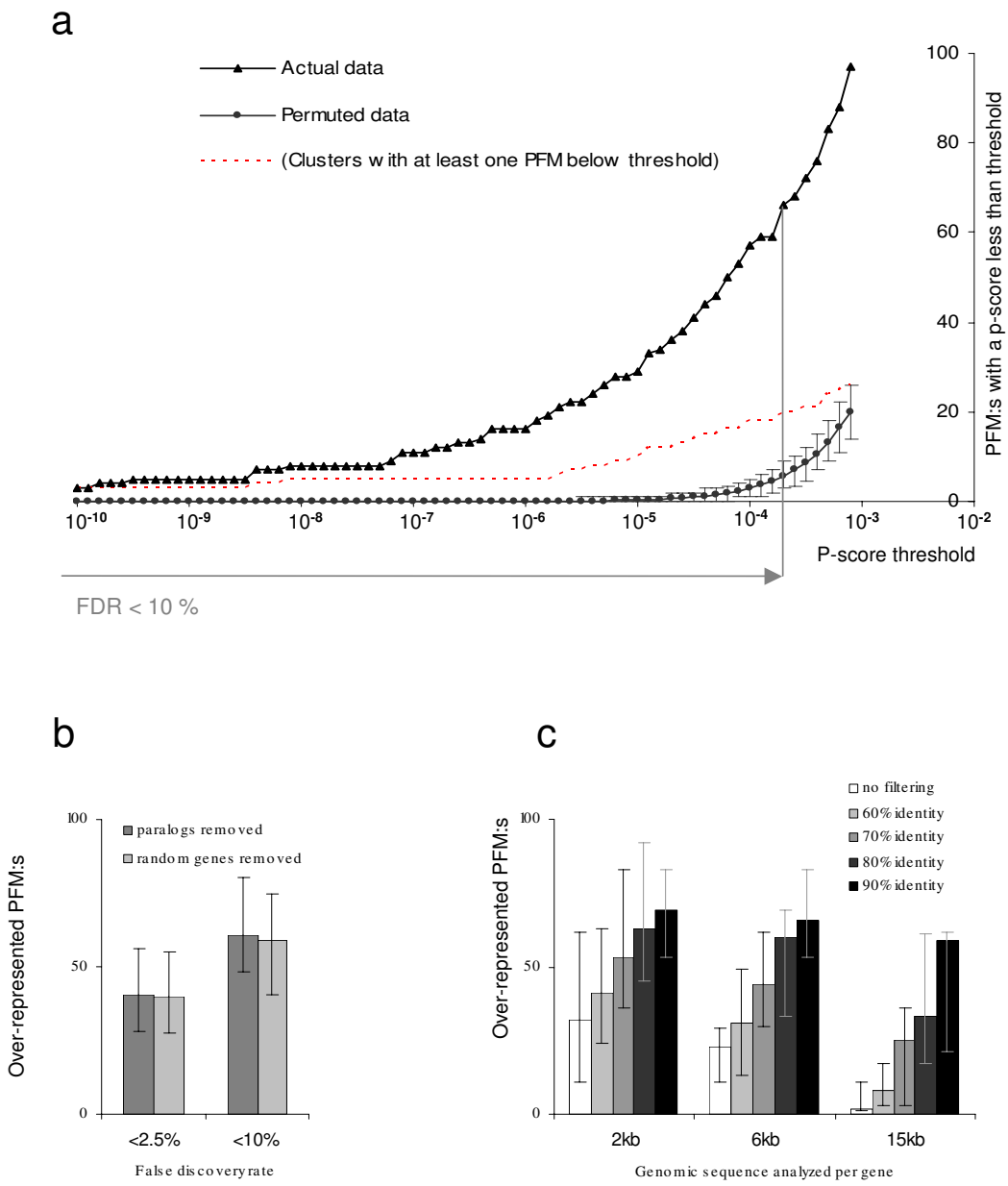


Figure 9

False discovery rate estimation. A: Illustration of how false discovery rates (FDRs) were estimated by use of simulations. Values on the y axis represent the number of motifs below a certain *p-score* (x axis). Triangles show results for observed data, filled circles show results for permuted data (error bars show the 90% confidence interval from 100 simulations). The FDR was calculated as the ratio between the simulated expectation and the observation. Red dotted line: The number of clusters with at least one enriched motif. Note that several motifs where over-represented in the same cluster. B: Removal of paralogous genes from each cluster did not affect the number of detected motifs. Consequently, co-expressed paralogs is not an important source of false positives. C: The amount of DNA used per gene, and the phylogenetic footprinting stringency has a strong effect on the number of detected over-represented motifs. The sensitivity is higher when the amount of DNA is reduced. Error bars in B and C were obtained by using the 5th and 95th percentiles in the simulation to define the FDR.

Table 3: Interpretation of over-represented motifs with respect to published evidence. (See footnote for definition of the categories.)

I: Expected cases		
Number	**Function / expression**	**Transcription factors**
4	lymphocytes	Irf1/Irf2/ISRE, NFKB [34-36]
6	liver	HNF-1 alpha and beta, HNF4 alpha [32, 59-61]
9,12	cardiac and skeletal muscle	MyoD/E47, MEF2 family factors, SRF [4, 6, 48]
15	erythroid cells	GATA-1 [62]
16	B lymphocytes	Spi-B, Oct-I* [37, 40, 63]
22	cell cycle	E2F family factors [49, 50]
40	smooth muscle	SRF [5, 64]
II: Extended		
3	neural tissue	NRSF [65]
5	testis	SOX17, RFX2 (RFX1-RFX3) [41-43]
6	liver	ARP-1/COUP-TF, PPARγ [66, 67]
13	ER	XBP-1 [68]
15,16	erythroid cells/ B cells	AREB6* [38]
17	kidney	HNF-4alpha, HNF-1-alpha/beta [69, 70]
III: Unexpected		
1	protein synthesis	NRF and ETS family factors
2	oocyte	E2F and ETS family factors
6	liver	ROR-alpha
8	ECM	Pax4
9	cardiac muscle	TALE family factors TGIF and MEIS1
13	ER	NRF and ETS family factors
24	pancreas	E47
30	small intestine	HNF4-alpha and GATA factors
44	retina/eye	Ik-2
45	testis	cap

Criteria for inclusion in the different groups: I. Existing evidence of co-regulation of cluster member by the factor, and/or a mutant mouse phenotype that affects differentiation of the cell type. II. Evidence of the factor acting on a limited number of genes in the tissue. Evidence of the transcription factor being expressed in the tissue. III. No or limited evidence in the literature. * = Detected at 20% FDR.

a whole-genome experimental screen for targets of this factor in hepatocytes [32]. 28% (29/104) of the ortholog pairs inside the liver cluster contained experimentally verified HNF-1 sites, which indicates a high but not complete coverage. Cross-comparison of experimentally and computationally identified HNF-1 targets showed a 69% – 71% agreement (results for the two different HNF-1 PFMs at 2 kb/gene masked for 90% phylogenetic conservation).

As a complement to using database motifs, we applied *de novo* motif elicitation to screen the regulatory sequence of each cluster for over-represented DNA motifs that were not present in databases. A two-step procedure based on the MEME algorithm [33] was developed (Methods). The use of a cluster to define a PFM will lead to a bias when testing for over-representation. A simulation that involved motif elicitation from randomized clusters was used to account for this bias (Methods). 28 de novo motifs were identified as over-represented in relation to the null simulation (Supplementary data, additional file 1). Of these 28 motifs, 13 matched a described database motif (based on the PFM similarity score described in Methods and in footnote of supplementary table). 9/13 of the matching motifs corresponded to motifs that were over-represented in Table 3. The remaining 15 motifs may represent novel regulatory elements. These motifs, and a description of the procedure, are available in the supplementary information.

Discussion

The gene battery theory predicts that the *core features* of differentiated cells are encoded by gene batteries, i.e. groups of functionally coupled, co-expressed genes that are regulated by similar sets of *cis-regulatory elements*. The primary aim of this article was to survey the mouse and human genomes for cases that fall within or close to the idealized gene battery concept.

Hierarchical clustering of an extensive compendium of micro-array expression data in mouse and human identified large numbers of co-expressed gene groups. A grand majority of the larger clusters were significantly enriched for genes that shared one or several GO-terms, indicating that co-expressed genes are functionally coupled. Moreover, interactions were significantly more common between proteins that are encoded by clustered pairs compared to randomly chosen pairs. 21 clusters, finally, were significantly enriched for genes that shared potential *cis-regulatory elements*

Confirmation and extension of several cases of gene battery-like regulation

The predicted over represented motifs (Table 2) fell into three principal categories: (i) predictions in agreement with described gene batteries, (ii) predictions supported by a limited number of observations, where the analysis predicted gene battery-like regulation, and (iii) novel predictions of hypothetical gene batteries. The findings are summarized with respect to these categories in Table 3.

For type (i) predictions the method is useful for proposing new target genes. As an example, the smooth muscle cluster contains 4 validated serum response factor (SRF) targets (*Acta2*, *Actg2*, *Myl11*, *Tagln*) [5]. The method identifies two more genes in this cluster, *Myl9* and *Lpp* as potential SRF targets in smooth muscle (Figure 2). The findings in two clusters that are relevant for lymphocytes (cluster 4 in the web supplement) and B-lymphocytes (cluster 16, Figure 3) are further examples of type (i) predictions. Cluster 4 contains genes that are expressed by various lymphocyte populations. Significantly enriched motifs were Irf1/Irf2, NF-kappaB, and ISRE. Irf1 and Ir2 appear to bind the same sites [34]. Irf1 knockout mice are impaired in their myelopoiesis [35]. NFkappaB is a well-known regulator of inflammation and immune functions [36]. Cluster 16 contains genes that are specifically expressed in B-lymphocytes. SpiB was identified as the most over-represented PFM followed by AREB6 and Oct-1. Knockout of SpiB caused specific defects in B-cell terminally differentiated functions [37]. AREB6 is a hematopoietic transcriptional repressor [38]. Whereas Oct-1 seems to be dispensable for B-cell development [39], it regulates B lymphocyte genes in combination with specific co-activators [40].

In type (ii) cases, a regulator with some degree of documentation is statistically supported to play a role in gene battery-like regulation. As an example, RFX family proteins have been shown to regulate a limited number of genes in spermatogenesis, and SOX17 has been shown to be expressed in the testis [41-43]. Here, we demonstrate that motifs corresponding to these factors are present in 8–11% of testis-selective genes (Figure 4). 29% of kidney-specific genes were covered by HNF-1 and HNF-4 sites, suggesting that these genes may be general regulators in the kidney.

Finally, type (iii) cases represent novel predictions which can be viewed as testable hypotheses of regulatory mechanisms. A complete list of PFMs in the different categories with references to the literature is presented in Table 3. Examples include a potential functional role for an Ikaros-like motif in retina-selective genes (cluster 44 in the web supplementary), and a Pax family factor acting on extracellular matrix genes with strong expression in arteries (cluster 8 in the web supplementary). Interestingly, a combination of Ets family factor motifs, Nuclear respiratory factor motifs, and E2F motifs was detected in several clusters of a housekeeping character (Table 3).

Incomplete coverage of cis-regulatory motifs indicates that other mechanisms than co-regulation may contribute to coordinated expression

The gene battery concept predicts that upstream transcription factors coordinate the expression of target genes through binding to similar *cis-regulatory elements* in the target genes. In a recent report, Alloco and co-workers [44] studied the relationship between co-expression in yeast (as determined by correlation in a set of 610 arrays) and the probability of sharing a transcription factor (as determined by experiments). In their experiments, an expression profile correlation of 0.85 implied a probability of transcription factor sharing of >0.5. The incomplete coverage in our analysis may similarly indicate that other mechanisms than co-regulation contribute to the coordinated expression. Importantly, it is technically difficult to accurately quantify the fraction of genes in a cluster that respond to a factor. The presence of false positive gene cluster members reduces the coverage. Uncertainty in the motif assignment to ortholog pairs may also reduce the coverage. In one case study, we could use data from an independent experimental screen to confirm that the liver-selective cluster had a limited coverage of HNF-1 responders (results). Further experiments are required to resolve to what extent the lack of coverage reflects alternative regulatory mechanisms or technical limitations.

Data limitations

The tissues in the dataset represent samples of limited morphological resolution at a fixed time point. As a

consequence, co-expressed gene groups that are active under specific developmental phases, under specific environmental stresses, or in small and localized anatomical structures are likely to escape detection. Moreover, co-expressed gene groups with a peak in one tissue only are unresolved for single cell types. Low coverage of motifs in such clusters may reflect contaminating genes that derive from another cell type.

The Affymetrix technology that were used to generate the expression measurements has been validated and match results obtained with tag sampling for sufficiently abundant genes [45,46]. Low intensity signals did not correlate well, indicating that such genes are less likely to form clusters (data not shown). Cross hybridization between related genes may, in theory, contribute to correlating expression levels. We can exclude that clusters primarily form as a result of cross-hybridization, since (1) members of well-characterized gene families appear in different (the expected) clusters (eg smooth muscle and cardiac actins) and since (2) co-expressed genes tended to share GO terms or to encode interacting proteins to a higher degree than they tended to be paralogs (supplementary data, additional file 1). We cannot however exclude subtle effects related to cross-hybridization, since results with cDNA microarrays indicate this possibility [47].

Transcription factors in the same family often have similar DNA binding properties and bind to the same sites on target genes. This leads to an ambiguity in the interpretation of PFM annotations. The MEF2 motif, is for example a receptor for the mammalian MEF2A, MEF2B, MEFC and MEFD transcription factors [48]. Similarly, the E2F motif is a receptor for a family of 9 E2F family proteins that form heterodimers with another family of proteins, the DP proteins, in a way that affects the binding affinity [49]. Different E2F proteins act at different stages in the cell cycle [50], and this is not captured by our method.

One additional data limitation is that a fraction of Ensembl transcripts may lack sequence in their 5' ends. The accuracy of our transcription starts is therefore dependent on the quality of the Ensembl transcript database.

Method considerations and perspectives

Our approach is technically related to the Toucan and ConFac tools [17,51]. Important differences and extensions are the introduction of a composite scoring system, a procedure to optimize PFM thresholds, and the use of simulation at the level of the whole clustering to measure significances.

The performance of the method was not affected by the removal of paralogous genes from the clusters (Figure 9B), and clustering of such genes is clearly not a significant source of false positive motif prediction. Consequently, we decided to include the paralogs in the final analysis since we believe that co-expression of such genes may be a result of shared *cis-regulatory elements*.

We further evaluated the effects of using different DNA amounts in the analysis. In principal, it is motivated to include a substantial amount of DNA sequence per gene, since mammalian enhancers are frequently located far upstream in relation to the transcription start, or downstream in intronic DNA. However, the benefits of including as much sequence as possible must be balanced against the risk of introducing vast amounts of non-informative sequence into the analysis. The evaluation clearly favoured using limited amounts of DNA, and high stringency phylogenetic footprinting (Figure 9C). The optimal future alternative may be to assemble a small amount of DNA from a large area of the genome using algorithms that sort out regulatory DNA regions more efficient than phylogenetic footprinting.

The approach presented here needs to be extended to generate more complete models of gene battery regulation. Modelling of *cis-regulatory element* combinations and the relative position and spacing of *elements* are examples of such extensions. The use of PCC as a measure of co-expression may need to be reconsidered if extending the approach to very large datasets, since this measure may be sensitive to noninformative signals in a majority of samples. Further, the expression levels of potential trans-regulators and co-factors (based on protein interaction networks) can be introduced. Most likely, the method will benefit from a more accurate identification of regulatory regions (Crawford-04)

Conclusion

We screened the mouse and human genomes and transcriptomes for instances of gene battery-like regulation. Comparative clustering was highly predictive of gene function and protein interaction, which indicates that potential gene batteries could be identified this way. Based on a statistical composite score for motifs in ortholog pairs, and a simulation approach to determine significance levels, we found 21 instances of statistically supported gene battery-like regulation that were conserved between mouse and human. These included known cases of gene battery regulation in tissues such as muscle, lymphocytes, erythrocytes, and liver. A second category of predictions included regulators with some degree of documentation, e.g. in testis, kidney, and endoplasmatic reticulum. Finally, new candidate gene batteries with statistically enriched *cis-regulatory* motifs were listed.

The results of this investigation emphasizes the need to study differentiation in terms of larger transcriptional units, and extends the methodology for doing this.

Methods

Annotation and preprocessing of gene expression datasets

Target sequences for the Novartis Gene Atlas V2 mouse and human expression datasets [23] were matched against the Ensembl [20] collection of mouse and human transcripts using BLAST [52]. In cases where the E-value exceeded 10^{-20}, the BLAST search was re-done against Ensembl gene sequences. If there was no match below 10^{-20}, the probe set was excluded from further analysis. The resulting datasets covered 17552 mouse and 16929 human unique Ensembl genes (Table 1). Mouse/human orthologous gene pairs were formed using Ensembl homology maps. Redundant occurrences of the same gene in more than one ortholog pair were avoided according to the following procedure: the Ensembl human orthologs for each Ensembl mouse gene were identified. When more than one ortholog was assigned to a mouse gene, the one with the lowest positional disagreement d (defined below) was chosen. The procedure was repeated for all Ensembl human genes. Reciprocally matching ortholog pairs were identified, and others were excluded from further analysis. The mouse and human expression profiles of each ortholog pair were normalized with respect to mean and standard deviation and combined into a single larger profile, finally yielding an expression dataset with 13282 non-redundant ortholog pairs (Table 1).

Preparation of upstream DNA sequence

For each ortholog pair, mouse and human candidate regulatory sequence was extracted from Ensembl. The sequence extraction algorithm starts with an ortholog pair, localizes the 5' end of the transcript in the genome in each species, and computes a value that measures the positional disagreement between the transcript 5' ends in the two species. If the disagreement is too large, the ortholog pair is excluded from the analysis. Of the 12239 ortholog pairs in the expression dataset, 9561 satisfied this criterion (Table 1). A full description of the sequence extraction procedure is available in the online supplement.

Three different lengths of DNA were extracted: 2 kb, 6 kb and 15 kb. The 2 kb dataset contained nucleotide positions ranging from -2000 to -1 relative to the transcription start, the 6 kb dataset contained positions -4000 to +2000 and the 15 k dataset contained positions -10000 to +5000.

Phylogenetic footprinting

The sequence datasets were subjected to phylogenetic footprinting, i.e. removal of poorly conserved sequence. The mouse and human sequences of each ortholog pair were aligned by use of the LAGAN software [53] (standard settings). Similarity was defined as the number of identical nucleotides in a 20 bp window. Nucleotides in windows with similarity below the threshold were removed. In all, five different similarity threshold were applied: >0% (no footprinting – use all sequence), >60%, >70%, >80% and >90% identity.

Removal of exonic sequence

Each candidate regulatory sequence was aligned to the corresponding transcripts (pairwise BLAST, e-value threshold 0.01). Nucleotides aligning with one or more transcripts were removed.

Clustering

The set of 9561 orthologs pairs, for which both regulatory sequence and expression data could be assembled, were clustered with respect to expression pattern using hierarchical clustering [54] with average group linkage and Pearson's correlation coefficient as distance measure. In the average group linkage algorithm, cluster distances are defined as the distances between cluster means. We defined the cluster mean as the arithmetic mean of all cluster members. Correlation thresholds between 0.61 and 0.99 were applied in steps of 0.01.

Assembly of 266 non-redundant motif position weight matrices

Motifs represented as position frequency matrices (PFMs) were downloaded from the TRANSFAC [55] and JASPAR [25] databases. Non-vertebrate matrices were filtered out. Highly similar matrices were grouped and merged using single linkage hierarchical clustering, reducing the number of PFMs from 322 to 266. Distances between matrices were calculated using a probabilistic method [56]. Individual positions between matrices were compared using the chi square test, and p-values for all overlapping positions were combined using the geometric mean. Sense and antisense of motifs were compared for all possible frameshifts with at least 75% overlap. Clustered motifs (score > 0.5) were added together in overlapping positions, and flanking positions were discarded. PFMs were transformed into position weight matrices (PWM:s) to make them compatible with the MAST software (see below). The value of each matrix element was calculated according to the following formula:

$$\log_2 \frac{\frac{n+\sqrt{N}/4}{N+\sqrt{N}}}{p(b)}$$

where n is the raw count from the corresponding position in the PFM, N is the the number of observations (sum of each position/column in the PFM), $\sqrt{N}/4$ and $\sqrt{N}$ are

pseudocounts and $p(b)$ is the background frequency of the corresponding nucleotide.

Scoring of regulatory sequences for motif position weight matrices

Scoring of individual sequences

For each ortholog pair, both mouse and human regulatory sequences were scored for all motif PWM's using the MAST software [57]. MAST was set to compute whole-sequence p-values (*-seqp* setting), using a first order Markov chain background. The Markov chain background data where computed from unmasked genomic sequence -4000 upstream to +2000 downstream in all ortholog pairs.

MAST composite scores

The p-values reported by MAST for the mouse and human sequences of an ortholog pair were multiplied for each PWM. The result was treated as a composite score that reflects the overall "signal" for a certain binding site in the regulatory sequences of an ortholog pair. The composite score is a product of two p-values but should not be interpreted as a p-value, since the mouse and human sequences are highly dependent.

Algorithms to detect over-representation of GO terms and motifs

GO over-representation

Clusters were evaluated for over-representation of GO terms using Fisher's exact test [30]. Due to the large number of tests (the number of clusters times the number of GO terms), the resulting p-values were corrected using the Bonferroni method [30].

Motif over-representation

The test was applied to the 9561 ortholog pairs with both sequence and expression data. These constituted the *population*. For each PWM, genes with MAST composite scores below a threshold were defined as *labeled* and the others as *unlabeled*. Further, each cluster was considered as a *sample* from the population. The algorithm is briefly sketched here and is available in detail in the online supplement:

Step 1: Compute p-scores under the null hypothesis

First, ortholog pairs were permuted across the dataset, making each cluster a random selection of genes. Second, an optimal composite score threshold was found for each PWM. This was done by computing the Fischer test p-value for over-representation of labeled ortholog pairs for all clusters k at a range of detection thresholds ranging from 10^{-8} to 10^{-3} in stepwise increases by a factor of $10^{0.5}$. The threshold chosen for each PWM was the one that gave the best p-value for that PWM in any cluster. P-values for over-representation of all motifs in all clusters were finally calculated using the optimized thresholds. We refer to this statistic as the *p-score*. This whole procedure was repeated for 100 iterations, which resulted in empirical estimates of how many over-represented motifs we could expect below a certain p-score under the null hypothesis.

Step 2: Compute p-scores for the observed data

This step was identical to step 1 but without permutations and repetitions.

Step 3. Compute the false discovery rate

After the simulation, we defined the false discovery rate (FDR) at p-score p as the expected number of over-represented motifs in the null simulation, divided by the corresponding value in the observed data.

Electrophoretic mobility shift assay (EMSA)

40-EI and 230–238 cells were grown in RPMI medium supplemented with 10% FCS, 10 mM HEPES, 2 mM pyruvate, 50 μM 2-mercaptoethanol and 50 μg gentamicin per ml (complete RPMI media). STAT activation in 230–238 cells was achieved by 5 hours of incubation with 0.5 ng/ml recombinant mouse interferon gamma (Immunokontact, Germany). Nuclear extracts were prepared according to Schreiber et al. [58]. DNA probes were labelled with γ[^{32}P] ATP (Amersham Biosciences, UK) by incubation with T4 polynucleotide kinase (Roche Diagnostics, Mannheim, Germany), and purified on a mini Quick Spin Oligo Column (Roche Diagnostics, Sweden). Nuclear extracts were incubated with labelled probe (20,000 cpm, 3 fmol) for 30 min at room temperature in binding buffer (10 mM HEPES [pH 7.9], 70 mM KCl, 1 mM dithiothreitol, 1 mM EDTA, 2.5 mM MgCl2, 4% Glycerol) with 0.75 μg Poly(dI/dC) (Amersham Pharmacia Biotech, UK). When EBF-binding was investigated, 1 mM $ZnCl_2$ was supplemented to the mixture. The samples were separated on a 6% polyacrylamide TBE gel, which was dried and subjected to autoradiography. In the supershift experiments (Figure 8B), DNA competitors or antibodies (anti-EBF SC-15333, anti-Actin SC-1616, Santa Cruz Biotech) were added 10 min before the addition of the DNA probe. To visualize the super-shifted complex, the unbound probe was run out of the gel. The following Oligonucleotides were used for EMSA: *mb-1* sense: 5'-AGCCACCTCTCAGGGGAATTGTGG-3'; *mb-1* antisense: 5'-CCACAATTCCCCTGAGAGGTGGCT-3'; mutated *mb-1*sense: 5'-AGCCACCTCTCAGCCGTTTTGTGG-3'; mutated *mb-1* antisense: 5'-CCACAAAACGGCTGAGAG-GTGGCT-3';

List of abbreviations

EBF Early B cell factor

EMSA Electrophoretic Mobility Shift Assay

GO term Gene Ontology term

PCC Pearson's correlation coefficient

PFM Position frequency matrix

PWM Position weight matrix

Authors' contributions

The overall computational strategy applied was conceived by SN, EL and PM together with EK, PL, and ON. SN and EL performed the bulk of the analysis and the manuscript was drafted by SN, EL and PL with contributions from all authors. The experimental validation of EBF sites was conceived and performed by RM and MS. The *de novo* detection of motifs was conceived and performed by EK.

Additional material

Additional File 1

Supplementary.doc is a word file that contains all the supplementary information referred to in the text.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-6-68-S1.doc]

Acknowledgements

We thank Tore Samuelsson and Magnus Alm-Rosenblad for advice and for making the EVO high-performance computer available. We also thank Anne Uv, Klas Kullander and Christer Betsholtz for valuable comments on draft versions of the manuscript. The work was funded by the European Commission: The Sixth Framework Programme, The Swedish Cancer Foundation, and the Swedish Research Council for Medicine.

References

1. Davidson EH: **Genomic Regulatory Systems. Development and evolution.** Academic Press; 2001.
2. Morgan TH: **Embryology and Genetics.** New York , Columbia University Press; 1934.
3. Britten RJ, Davidson EH: **Gene regulation for higher cells: a theory.** *Science* 1969, **165(891):**349-357.
4. Dias P, Dilling M, Houghton P: **The molecular basis of skeletal muscle differentiation.** *Semin Diagn Pathol* 1994, **11(1):**3-14.
5. Owens GK, Kumar MS, Wamhoff BR: **Molecular regulation of vascular smooth muscle cell differentiation in development and disease.** *Physiol Rev* 2004, **84(3):**767-801.
6. Cripps RM, Olson EN: **Control of cardiac development by an evolutionarily conserved transcriptional network.** *Dev Biol* 2002, **246(1):**14-28.
7. Firulli AB, Olson EN: **Modular regulation of muscle gene transcription: a mechanism for muscle cell diversity.** *Trends Genet* 1997, **13(9):**364-369.
8. Wang X, Crispino JD, Letting DL, Nakazawa M, Poncz M, Blobel GA: **Control of megakaryocyte-specific gene expression by GATA-1 and FOG-1: role of Ets transcription factors.** *Embo J* 2002, **21(19):**5225-5234.
9. Ma S, Rao L, Freedberg IM, Blumenberg M: **Transcriptional control of K5, K6, K14, and K17 keratin genes by AP-1 and NF-kappaB family members.** *Gene Expr* 1997, **6(6):**361-370.
10. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124(10):**1851-1864.
11. Bartholdy B, Matthias P: **Transcriptional control of B cell development and function.** *Gene* 2004, **327(1):**1-23.
12. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34(2):**166-176.
13. Simonis N, van Helden J, Cohen GN, Wodak SJ: **Transcriptional regulation of protein complexes in yeast.** *Genome Biol* 2004, **5(5):**R33.
14. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278(1):**167-181.
15. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11(9):**1559-1566.
16. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26(2):**225-228.
17. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31(6):**1753-1764.
18. Frech K, Werner T: **Specific modelling of regulatory units in DNA sequences.** *Pac Symp Biocomput* 1997:151-162.
19. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T: **Ensembl 2004.** *Nucleic Acids Res* 2004, **32 Database issue:**D468-70.
20. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30(1):**38-41.
21. Nelander S, Mostad P, Lindahl P: **Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes.** *Genome Res* 2003, **13(8):**1838-1854.
22. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14(6):**1085-1094.
23. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101(16):**6062-6067.
24. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31(1):**248-250.
25. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32 (Database issue):**D91-4.
26. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28(1):**316-319.
27. Zhang JC, Kim S, Helmke BP, Yu WW, Du KL, Lu MM, Strobeck M, Yu Q, Parmacek MS: **Analysis of SM22alpha-deficient mice reveals unanticipated insights into smooth muscle cell differentiation and function.** *Mol Cell Biol* 2001, **21(4):**1336-1344.
28. Bailey TL, Gribskov M: **Combining evidence using p-values. application to sequence homology searches.** *Bioinformatics* 1998, **14:**48-54.
29. Hagman J, Belanger C, Travis A, Turck CW, Grosschedl R: **Cloning and functional characterization of early B-cell factor, a regulator of lymphocyte-specific gene expression.** *Genes Dev* 1993, **7(5):**760-773.
30. Rice JA: **Mathematical statistics and data analysis.** 2nd edition. Belmont, CA , Duxbury Press; 1995:xx, 602, A49 p..
31. Dieterich C, Rahmann S, Vingron M: **Functional inference from non-random distributions of conserved predicted transcription factor binding sites.** *Bioinformatics* 2004, **20 Suppl 1:**I109-I115.

32. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA: **Control of pancreas and liver gene expression by HNF transcription factors.** *Science* 2004, **303(5662):**1378-1381.
33. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
34. Kuo TC, Calame KL: **B lymphocyte-induced maturation protein (Blimp)-1, IFN regulatory factor (IRF)-1, and IRF-2 can bind to the same regulatory sites.** *J Immunol* 2004, **173(9):**5556-5563.
35. Testa U, Stellacci E, Pelosi E, Sestili P, Venditti M, Orsatti R, Fragale A, Petrucci E, Pasquini L, Belardelli F, Gabriele L, Battistini A: **Impaired myelopoiesis in mice devoid of interferon regulatory factor 1.** *Leukemia* 2004, **18(11):**1864-1871.
36. Aggarwal BB: **Nuclear factor-kappaB: the enemy within.** *Cancer Cell* 2004, **6(3):**203-208.
37. Bartel FO, Higuchi T, Spyropoulos DD: **Mouse models in the study of the Ets family of transcription factors.** *Oncogene* 2000, **19(55):**6443-6454.
38. Turner J, Crossley M: **Basic Kruppel-like factor functions within a network of interacting haematopoietic transcription factors.** *Int J Biochem Cell Biol* 1999, **31(10):**1169-1174.
39. Wang VE, Tantin D, Chen J, Sharp PA: **B cell development and immunoglobulin transcription in Oct-1-deficient mice.** *Proc Natl Acad Sci U S A* 2004, **101(7):**2005-2010.
40. Spiegelman BM, Heinrich R: **Biological control through regulated transcriptional coactivators.** *Cell* 2004, **119(2):**157-167.
41. Wang R, Cheng H, Xia L, Guo Y, Huang X, Zhou R: **Molecular cloning and expression of Sox17 in gonads during sex reversal in the rice field eel, a teleost fish with a characteristic of natural sex transformation.** *Biochem Biophys Res Commun* 2003, **303(2):**452-457.
42. Katoh M: **Molecular cloning and characterization of human SOX17.** *Int J Mol Med* 2002, **9(2):**153-157.
43. Horvath GC, Kistler WS, Kistler MK: **RFX2 is a potential transcriptional regulatory factor for histone H1t and other genes expressed during the meiotic phase of spermatogenesis.** *Biol Reprod* 2004, **71(5):**1551-1559.
44. Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** *BMC Bioinformatics* 2004, **5(1):**18.
45. Ishii M, Hashimoto S, Tsutsumi S, Wada Y, Matsushima K, Kodama T, Aburatani H: **Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis.** *Genomics* 2000, **68(2):**136-143.
46. Kim HL: **Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells.** *Exp Mol Med* 2003, **35(5):**460-466.
47. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18(3):**405-412.
48. Black BL, Olson EN: **Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins.** *Annu Rev Cell Dev Biol* 1998, **14:**167-196.
49. Tao Y, Kassatly RF, Cress WD, Horowitz JM: **Subunit composition determines E2F DNA-binding site specificity.** *Mol Cell Biol* 1997, **17(12):**6994-7007.
50. Attwooll C, Denchi EL, Helin K: **The E2F family: specific functions and overlapping interests.** *Embo J* 2004.
51. Karanam S, Moreno CS: **CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W475-84.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3):**403-410.
53. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13(4):**721-731.
54. Sokal RR, Michener CD: **A Statistical Method for Evaluating Systematic Relationships.** *Univ Kans Sci Bull* 1958, **38:**1409-1438.
55. Wingender E: **Recognition of regulatory regions in genomic sequences.** *J Biotechnol* 1994, **35(2-3):**273-280.
56. Schones DE, Sumazin P, Zhang MQ: **Similarity of position frequency matrices for transcription factor binding sites.** *Bioinformatics* 2004.
57. Bailey TL, Baker ME, Elkan CP: **An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases.** *J Steroid Biochem Mol Biol* 1997, **62(1):**29-44.
58. Schreiber E, Matthias P, Müller M, Schaffner W: **Rapid detection of octamer binding proteins with "mini-extracts", prepared from a small number of cells.** *Nucleic Acids Res* 1989, **17:**6419.
59. Li J, Ning G, Duncan SA: **Mammalian hepatocyte differentiation requires the transcription factor HNF-4alpha.** *Genes Dev* 2000, **14(4):**464-474.
60. Ishiyama T, Kano J, Minami Y, Iijima T, Morishita Y, Noguchi M: **Expression of HNFs and C/EBP alpha is correlated with immunocytochemical differentiation of cell lines derived from human hepatocellular carcinomas, hepatoblastomas and immortalized hepatocytes.** *Cancer Sci* 2003, **94(9):**757-763.
61. Arrese M, Karpen SJ: **HNF-1 alpha: have bile acid transport genes found their "master"?** *J Hepatol* 2002, **36(1):**142-145.
62. Izraeli S: **Leukaemia -- a developmental perspective.** *Br J Haematol* 2004, **126(1):**3-10.
63. Hu CJ, Rao S, Ramirez-Bergeron DL, Garrett-Sinha LA, Gerondakis S, Clark MR, Simon MC: **PU.1/Spi-B regulation of c-rel is essential for mature B cell survival.** *Immunity* 2001, **15(4):**545-555.
64. Parmacek MS: **Transcriptional programs regulating vascular smooth muscle cell development and differentiation.** *Curr Top Dev Biol* 2001, **51:**69-89.
65. Chen ZF, Paquette AJ, Anderson DJ: **NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis.** *Nat Genet* 1998, **20(2):**136-142.
66. Gavrilova O, Haluzik M, Matsusue K, Cutson JJ, Johnson L, Dietz KR, Nicol CJ, Vinson C, Gonzalez FJ, Reitman ML: **Liver peroxisome proliferator-activated receptor gamma contributes to hepatic steatosis, triglyceride clearance, and regulation of body fat mass.** *J Biol Chem* 2003, **278(36):**34268-34276.
67. Kang S, Spann NJ, Hui TY, Davis RA: **ARP-1/COUP-TF II determines hepatoma phenotype by acting as both a transcriptional repressor of microsomal triglyceride transfer protein and an inducer of CYP7A1.** *J Biol Chem* 2003, **278(33):**30478-30486.
68. Lee AH, Iwakoshi NN, Glimcher LH: **XBP-1 regulates a subset of endoplasmic reticulum resident chaperone genes in the unfolded protein response.** *Mol Cell Biol* 2003, **23(21):**7448-7459.
69. Ryffel GU: **Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF)1 and HNF4 families: functional and pathological consequences.** *J Mol Endocrinol* 2001, **27(1):**11-29.
70. Stoffel M, Duncan SA: **The maturity-onset diabetes of the young (MODY1) transcription factor HNF4alpha regulates expression of genes required for glucose transport and metabolism.** *Proc Natl Acad Sci U S A* 1997, **94(24):**13209-13214.