ARTICLE

doi:10.1038/nature11245

Architecture of the human regulatory network derived from ENCODE data

Mark B. Gerstein[1,2,3]*, Anshul Kundaje[4]*, Manoj Hariharan[5]*, Stephen G. Landt[5]*, Koon-Kiu Yan[1,2]*, Chao Cheng[1,2]*, Xinmeng Jasmine Mu[1]*, Ekta Khurana[1,2]*, Joel Rozowsky[2]*, Roger Alexander[1,2]*, Renqiang Min[1,2,6]*, Pedro Alves[1]*, Alexej Abyzov[1,2], Nick Addleman[5], Nitin Bhardwaj[1,2], Alan P. Boyle[5], Philip Cayting[5], Alexandra Charos[7], David Z. Chen[3], Yong Cheng[5], Declan Clarke[8], Catharine Eastman[5], Ghia Euskirchen[5], Seth Frietze[9], Yao Fu[1], Jason Gertz[10], Fabian Grubert[5], Arif Harmanci[1,2], Preti Jain[10], Maya Kasowski[5], Phil Lacroute[5], Jing Leng[1], Jin Lian[11], Hannah Monahan[7], Henriette O'Geen[12], Zhengqing Ouyang[5], E. Christopher Partridge[10], Dorrelyn Patacsil[5], Florencia Pauli[10], Debasish Raha[7], Lucia Ramirez[5], Timothy E. Reddy[10]†, Brian Reed[7], Minyi Shi[5], Teri Slifer[5], Jing Wang[1], Linfeng Wu[5], Xinqiong Yang[5], Kevin Y. Yip[1,2,13], Gili Zilberman-Schapira[1], Serafim Batzoglou[4], Arend Sidow[14], Peggy J. Farnham[9], Richard M. Myers[10], Sherman M. Weissman[11] & Michael Snyder[5]

Transcription factors bind in a combinatorial fashion to specify the on-and-off states of genes; the ensemble of these binding events forms a regulatory network, constituting the wiring diagram for a cell. To examine the principles of the human transcriptional regulatory network, we determined the genomic binding information of 119 transcription-related factors in over 450 distinct experiments. We found the combinatorial, co-association of transcription factors to be highly context specific: distinct combinations of factors bind at specific genomic locations. In particular, there are significant differences in the binding proximal and distal to genes. We organized all the transcription factor binding into a hierarchy and integrated it with other genomic information (for example, microRNA regulation), forming a dense meta-network. Factors at different levels have different properties; for instance, top-level transcription factors more strongly influence expression and middle-level ones co-regulate targets to mitigate information-flow bottlenecks. Moreover, these co-regulations give rise to many enriched network motifs (for example, noise-buffering feed-forward loops). Finally, more connected network components are under stronger selection and exhibit a greater degree of allele-specific activity (that is, differential binding to the two parental alleles). The regulatory information obtained in this study will be crucial for interpreting personal genome sequences and understanding basic principles of human biology and disease.

A central goal in biology is to understand how a limited cohort of transcription factors is able to organize the large diversity of gene-expression patterns in different cell types and conditions. Over the past decade, system-wide analyses of transcription-factor-binding patterns have been performed in unicellular model organisms, such as *Escherichia coli* and yeast, and have revealed a great deal of information about the organization of regulatory information[1–8]. These studies have provided insights into such features as network hubs[1], connectivity correlations[9], hierarchical organization[10,11] and network motifs[12,13]. Moreover, more complex networks that integrate disparate forms of genomic and proteomic data, such as protein–protein interactions and phosphorylation, have related gene regulation to other biological processes[14–16]. However, for humans, systems-level analyses have been a challenge due to the size of the transcription factor repertoire and genome, and only specific regulatory subnetworks with a handful of factors have been reported thus far[17–19]. The large-scale data from the ENCODE project now begins to enable such analyses[20]. Moreover, with the vast amount of human polymorphism data and genome sequences of many mammals[21,22], it is possible to obtain an unprecedented view of how selection relates to networks.

Here we present an analysis of the genome-wide binding profiles of 119 transcription-related factors, including sequence-specific, general and chromatin-acting factors. (For simplicity, we refer to all of these as transcription factors, and we use TFSS to denote canonical sequence-specific factors.) We first used the transcription-factor-binding data to analyse the co-association patterns between different factors, as well as their differential patterns in promoter-proximal and distal regulatory regions. We then organized the binding patterns into a stratified hierarchy representing the overall systems-level regulatory wiring. To this, we added other forms of network information, including non-coding RNA (ncRNA) regulation (especially microRNAs

[1]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [2]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [3]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, Connecticut 06511, USA. [4]Department of Computer Science, Stanford University, 318 Campus Drive, Stanford, California 94305, USA. [5]Department of Genetics, Stanford University, 300 Pasteur Drive, M-344 Stanford, California 94305, USA. [6]Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA. [7]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. [8]Department of Chemistry, Yale University, 225 Prospect Street, New Haven, Connecticut 06520, USA. [9]Department of Biochemistry and Molecular Biology, University of Southern California, Norris Comprehensive Cancer Center, 1450 Biggy Street, NRT 6503, Los Angeles, California 90089, USA. [10]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [11]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [12]Genome Center, University of California-Davis, 451 Health Sciences Drive, Davis, California 95616, USA. [13]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. [14]Department of Pathology, Stanford University, SUMC L235 (Edwards Bldg), 300 Pasteur Drive, Stanford, California 94305, USA. †Present address: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina 27710, USA.

*These authors contributed equally to this work.

(miRNAs))[23,24], protein–protein interactions[25,26], and protein phosphorylation[27]. We analysed this 'meta-network' for properties that differ based on hierarchical level and connectivity (for example, hubs versus non-hubs) and also searched for enriched network motifs. Finally, we surveyed the pattern of sequence variation over the network, examining selective pressure and allelic effects (preferential binding to the maternal or paternal allele). Several of our key findings are summarized below.

- Human transcription factors co-associate in a combinatorial and context-specific fashion; different combinations of factors bind near different targets, and the binding of one factor often affects the preferred binding partners of others. Moreover, transcription factors often show different co-association patterns in gene-proximal and distal regions.
- Different parts of the hierarchical transcription factor network exhibit distinct properties. For instance, the middle level has the most information-flow bottlenecks and, offsetting this, tends to have the most regulatory collaboration between transcription factors. Conversely, higher-level transcription factors have the greatest connectivity with other networks (for example, the phosphorylome).
- The occurrence of the feed-forward loops is strongly enriched in the transcription factor network, as are a number of motifs in which two genes co-regulated by a factor are bridged by a protein–protein interaction or regulating miRNA.
- Highly connected network elements (both transcription factors and targets) are under strong evolutionary selection and exhibit stronger allele-specific activity (this is particularly apparent when multiple factors are involved). Surprisingly, however, elements with allelic activity are under weaker selection than non-allelic ones.

Overview of data and processing

The ENCODE project has generated chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) data sets for 119 distinct transcription factors over five main cell lines (Supplementary Information, section B.1, and Supplementary Tables 1 and 2a). Each data set contains at least two biological replicates. In addition, for a select set of factors (Supplementary Fig. 1c), short interfering RNA (siRNA) experiments were performed, where the transcription factor was depleted and expression changes were quantified by RNA-seq (Supplementary Information, section B.2). Most of the factors (88, 74%) are TFSSs that can be subcategorized on the basis of their DNA-binding domain sequences (Supplementary Table 2a)[28]. A small subset (16, 13%) comprises POL2 and general transcriptional machinery; a final subset (15, 13%) consists of chromatin-modifying and remodelling factors.

To allow effective integrative analysis of these diverse data sets, we developed a uniform processing pipeline and quality-control measures (Supplementary Information, section B.1, and Supplementary Figs 1a, b and 2a; data at http://www.encodeproject.org). In total, we identified 7,424,765 peaks; 2,948,387 (~40%) were proximal (within ±2.5 kilobases) to annotated gene transcription start sites (TSSs).

Context-specific transcription factor co-association

We first examined the genome-wide co-association of all pairs of transcription factors by analysing the overlap between peaks of all pairs of factors[20]. Although many general trends can be identified, this approach does not take into account the context-specificity of transcription factor binding (that is, the observation that factors bind together in distinct combinations at different genomic locations, and that the co-binding of one pair of transcription factors is often affected by the binding of another transcription factor; Supplementary Information, section C.1). Therefore, we developed a framework focusing on the specific genomic regions bound by a particular transcription factor (the focus factor) and examined the co-association of all other factors (partner factors) within this context (Supplementary Fig. 2a). For each ~350-base-pair region in the focus-factor context, we extracted normalized binding signals of overlapping peaks of all transcription factors, generating a co-binding map. Figure 1a shows such a map for the GATA1 context. Here, factors that consistently co-associate with each other and a substantial proportion of GATA1 peaks are termed 'primary partners' (for example, group 6 transcription factors such as GATA2 and TAL1 in Fig. 1a). In addition to these factors, there are also groups of 'local partners' that co-associate with each other in the presence of GATA1, but only at specific subsets of GATA1-binding peaks (for example, JUN in group 7 and MAX in group 3; Fig. 1a and Supplementary Fig. 2c-1). These 'biclusters', typically containing two to five transcription factors, can be mutually exclusive or partially overlapping.

To identify systematically all primary and local partners for each focus-factor context, we used a machine-learning approach. We derived nonlinear, combinatorial models of each focus-factor's co-binding map relative to randomized control maps (Supplementary Information, section C.2, and Supplementary Fig. 2a, b). Analysis of multivariate rules in these models, in turn, identified pairs and higher-order clusters of significantly co-associated transcription factors. Moreover, these co-associations are robust to peak overlap and calling thresholds (Supplementary Information, section C.4).

The first statistic derived from the models is a relative importance (RI) score (Supplementary Information, section C.2.4.2), which gives the overall importance of each transcription factor in the model. It reflects the 'size' of the biclusters to which a particular transcription factor belongs, and it is related to the number of co-binding factors and the fraction of peak locations involved. For the GATA1 context (Fig. 1b and Supplementary Fig. 2c-2), primary partners TAL1, GATA2 and POL2, as well as local partners MAX and JUN, have high RI scores. To reveal further the partnering in the focus-factor context, we computed co-association scores between all pairs and higher-order sets of transcription factors (Supplementary Information, section C.2.4). These scores measure the impact of the co-dependency implicit in a particular pair on the model as a whole, and they more directly probe the co-occupancy of transcription factors in the focus-factor context than does the RI score. For the GATA1 context, the co-association scores revealed both expected and novel pairings (for example, MYC–MAX–E2F6 and CCNT2–HMGN3, respectively; Fig. 1b, Supplementary Fig. 2c-2 and Supplementary Information, section C.3.1.4). Furthermore, GATA1 is usually associated with enhancer activity. However, the co-association score shows that it is connected to both repressive (for example, NRSF (also called REST) and HDAC2) and activating factors (for example, P300). This discordant behaviour has been observed previously[29]; here, it is borne out by expression studies and knockdowns (Supplementary Information, section C.3.1.4). In particular, after GATA1 knockdown, we found that 94 targets of GATA1 were significantly upregulated, and only 54 were downregulated (Supplementary Fig. 2e-4). Finally, we analysed the functions of genes that lie near clusters of co-associated factors, and found that many are enriched for specific biological functions (Supplementary Fig. 2e-2). For example, one bicluster involving E2F6 (E2F6–GATA1–GATA2–TAL1) was enriched for genes related to myeloid differentiation, whereas another (E2F6–SP1–SP2–FOS–IRF1) was involved in DNA damage response (Supplementary Information, section C.3.3). Thus, distinct combinations of factors regulate specific types of genes.

Comparing co-association across contexts

Aggregate RIM and PPM

After establishing the co-binding structure in each transcription factor context, we compared our co-association statistics across contexts. In particular, we combined the RI scores for each transcription factor into a single matrix (RIM, Supplementary Fig. 2a). Clustering reveals nine functionally distinct classes of transcription factor contexts that fall into four broad groups: proximal, distal, repressive

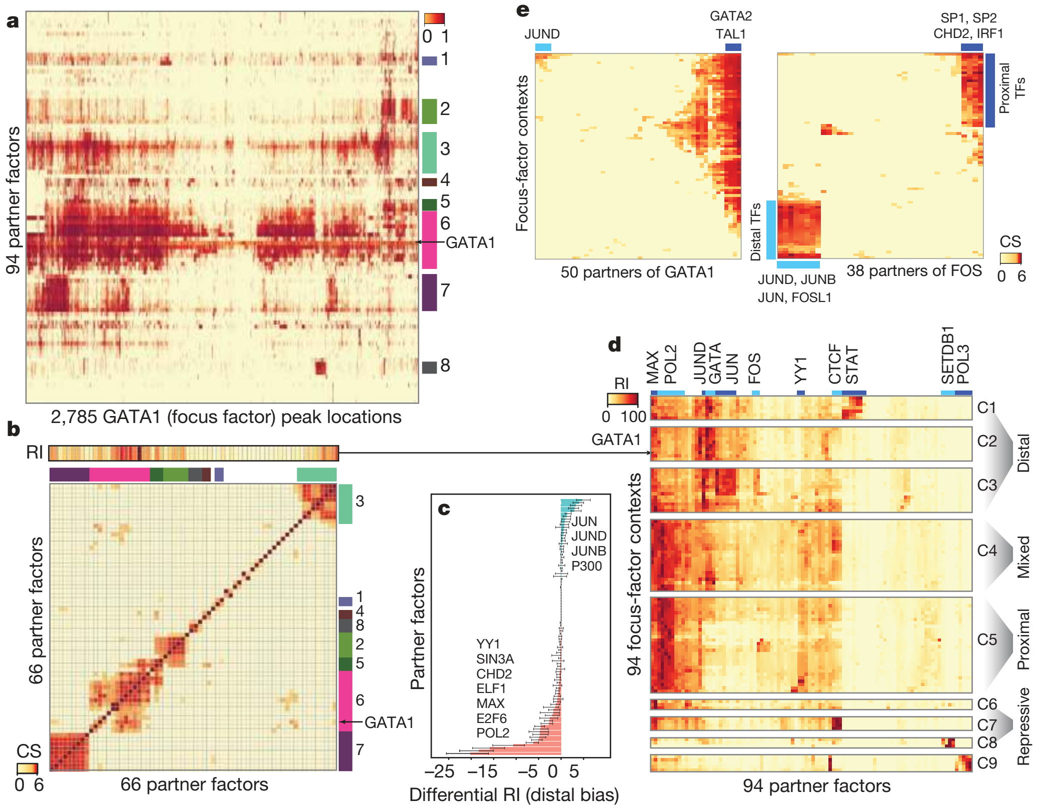

Figure 1 | Transcription factor co-association. a, The co-binding map for the GATA1 focus-factor context in K562 cells shows the binding intensity of peaks of all transcription factors (TFs) in K562 (rows) that overlap each GATA1 peak (columns). The coloured rectangles represent eight key clusters consisting of different combinations of co-associating partner-factors. **b**, The GATA1 context-specific relative importance (RI) scores of all partner factors (top) and the matrix of co-association scores (CS) between all pairs of factors (bottom). Primary and local partners of GATA have high RI scores. The co-association score matrix captures the eight clusters observed in **a**. **c**, Different partner factors are preferentially enriched at gene-distal (positive differential RI) and proximal (negative differential RI) GATA1 peaks. **d**, The aggregate factor importance matrix (RIM), obtained by stacking the RI scores of all partner factors (columns) from all focus-factor contexts (rows) in K562 cells, shows nine functionally distinct clusters (C1 to C9) of contexts that can be broadly grouped as distal, proximal, mixed and repressive. The blue rectangles highlight representative partner factors with high RI scores in the clusters. The arrow from **b** to **d** indicates that the GATA1 context-specific RI scores form one row in this matrix. **e**, Co-association variability map of partners (columns) of GATA1 (left panel) and FOS (right panel) over all K562 focus-factor contexts (rows). TAL1 and GATA2 show consistently high co-association scores with GATA1 over most focus-factor contexts, but JUND shows context-specific co-association. FOS shows marked changes in co-association score of partner factors over different contexts (for example, FOS–JUND in distal contexts and FOS–SP2 in proximal ones). (More details are available in Supplementary Fig. 2c, d, f-1, l-2.)

and mixed (Fig. 1d, Supplementary Fig. 2f-1 and Supplementary Information, section C.3.4.1). Next, combining the co-association scores from all focus factors across different contexts provides an overall view of all the primary partners of each transcription factor in the form of a primary-partner matrix (PPM; Supplementary Fig. 2f-4). The RIM reflects the overall similarities in the binding context of focus factors, whereas the PPM highlights the specific factors that tend to co-bind with each other (mutual primary partners). To some degree, one can see the PPM as a subset of the relationships implicit in the RIM. That is, two factors can have similar binding contexts without explicit co-association—for example, two factors that tend both to bind promoters but near different sets of genes. Overall, the PPM shows well known sets of co-associated transcription factors, such as FOS–JUN (the AP1 complex[30,31]) and CTCF–RAD21–SMC3 (the cohesion complex[32,33]), as well as many novel co-associations, such as CHD2–ZBTB33, EGR1–ZBTB7A and CTCF–ZNF143–SIX5 (Supplementary Information, section C3.6.2). We confirmed one novel co-association (CEBPB–TAL1) using co-immunoprecipitation and mass spectrometry (Supplementary Table 3a).

Variability map

The variability map shows the degree of variability in the partners of a given transcription factor over contexts (as determined by the co-association score) (Supplementary Information, section C.2.5.5). For instance, Fig. 1e shows that GATA1 has mostly the same partners in many contexts (for example, TAL1 and GATA2 are partners over almost all contexts). However, a few partners (for example, JUND) are present in only some contexts. An extreme example is FOS, which completely changes its partners in different contexts (Fig. 1e, Supplementary Fig. 2l-2 and Supplementary Information, section C.3.6.1).

Cell-type differences

We analysed transcription factor co-association in the five main ENCODE cell types (Supplementary Information, section C.3.4). The GM12878 and K562 cell lines have the most common (31) transcription factor data sets (Supplementary Information, section C.3.5). Comparative analysis showed that over 80% of the transcription factor pairs had no significant change in co-association between K562 and GM12878 cell lines. However, there were a few marked examples of cell-line differences. For instance, FOS and JUND co-associate in K562 but not in GM12878 cells (Supplementary Information, section 3.5.1), despite the fact that most of the other partners of FOS are maintained in both cell lines.

Gene context: proximal versus distal

Overall, we found distinct partner preferences at proximal and distal sites. These results were robust to the choice of the distance used to define proximal and distal regions (Supplementary Fig. 2c-3). In particular, for the GATA1 context, we found that RI scores change markedly between proximal or distal sites (Fig. 1c and Supplementary Fig. 2c-3): typical core promoter transcription factors (for example, POL2, E2F6, MAX and ELF1) have a significant proximal promoter bias, whereas JUND, JUNB, JUN and P300 show preferential co-association with distal sites. Another way of analysing differences between proximal and distal sites is in the framework of the variability map, in which one can observe the changing partners of a transcription factor in different contexts. For instance, FOS has completely different partners with which it co-associates proximally and distally (Fig. 1e, Supplementary Fig. 2l-2 and Supplementary Information, section C.3.6.1).

Assembling pairwise interactions into hierarchies

Analysis of co-associations specifies the relationships between the DNA-binding profiles of multiple regulators. To obtain a systems-level perspective, we recast transcription factor associations as a network (Supplementary Fig. 4a), wherein the nodes are regulators or their targets, and the edges designate regulatory relationships. Here, we focussed on the global wiring pattern across all cell types. We expected different subnetworks within this framework to be active to different degrees in different cells.

Using our binding-site list, we identified an initial set of regulatory targets from genes having promoter-proximal binding sites. The resulting raw network consists of 500,542 promoter-associated interactions between transcription factors and all their putative targets, of which 4,809 are between pairs of factors (networks at http://encodenets.gersteinlab.org). We filtered this to identify the most confident interactions using a probabilistic model, giving 26,070 total interactions, with only 338 between transcription factors[34] (Supplementary Information, section D.1). We validated the performance of the filtering using the siRNA experiments; for each case, the targets identified by our model were more differentially expressed in siRNA-treated cells than were those identified by a simple peak-based method (Supplementary Fig. 1c–e).

We next computed common connectivity statistics for individual transcription factors, namely, out-degree (O), in-degree (I) and betweenness, which were then used to identify hubs and information-flow bottlenecks (Supplementary Information, section K). Of particular interest is the difference between out- and in degree ($O - I$), which measures the direction of information flow (Supplementary Fig. 3a). A positive value suggests that a transcription factor is located 'upstream' in the network, whereas a negative value indicates that it is 'downstream'. We further defined a normalized version of this 'hierarchy height' metric, $h = (O - I)/(O + I)$. We found that this can be approximated by three levels (Supplementary Fig. 3c), with top-level, 'executive' transcription factors regulating many other factors ($h \approx 1$), and bottom-level 'foreman' factors more regulated than regulating ($h \approx -1$). For purposes of visualization, we used a simulated-annealing procedure to optimally and robustly arrange the 119 transcription factors into three discrete levels (with the number of downward-pointing edges maximized) (Fig. 2a and Supplementary Information, section D.2).

Layering on distal, ncRNA and protein interactions

The filtered transcription factor hierarchy consists of the strongest promoter-associated interactions. Building upon this skeleton, we added additional types of connections.

Interactions involving distal regulatory elements (for example, enhancers) are more difficult to identify than those involving proximal elements. Here, we used a statistical model[35]. This identifies distal sites with potentially many binding transcription factors using chromatin features. These regions were associated with a gene if their changing pattern of chromatin marks across cell lines correlates with the expression of that gene (Supplementary Information, section E.1). Overall, the model identified 19,258 distal edges (Fig. 2a).

The regulatory interactions between transcription factors and ncRNAs constitute an additional layer of information to add to the meta-network. We used transcription factor peaks proximal to ncRNAs to identify transcription-factor-to-ncRNA regulation. Next, we incorporated miRNA-to-transcription-factor regulatory interactions from TargetScan[36] (Supplementary Information, section E.2). Finally, we incorporated physical protein–protein interactions[26], as well as predicted phosphorylations (Supplementary Information, section F.3, and Supplementary Fig. 7a). Overall, these different interactions form a dense meta-network that we analysed further for interesting biological properties.

Relating network connectivity and genomic properties

We next correlated measures for the connectivity and hierarchical position of each transcription factor with a wide variety of genomic and proteomic properties (Fig. 2c, Table 1 and Supplementary Table 4, P values in the latter).

Correlations with distal edges

Distal edges have a different degree distribution than do proximal ones (Fig. 2a and Supplementary Fig. 5). Inspection reveals that many point upward in the transcription factor hierarchy, opposite to most proximal edges. Furthermore, we found many transcription factors with low in-degree values in the proximal network but high in-degree values in the distal one, indicating that they are heavily regulated through enhancers (Supplementary Fig. 5a). Some of these are well known condition- and tissue-specific regulators (for example, IRF4 and GATA1)[37].

Correlations within the proximal network

Upper-level transcription factors tend to have more targets than lower-level ones, both overall and when considering only other transcription factors as targets. As measured by betweenness in proximal regulation, middle-level transcription factors form information-flow bottlenecks (Fig. 2c). Moreover, betweenness in the proximal network is correlated with more distal regulation. This tends to increase the information flow through mid-level bottlenecks even more. (See Supplementary Information section F.3.6 for clarification of the implications.)

Correlation with protein interactions and the phosphorylome

We found that top-level transcription factors tend to have more partners in the protein–interaction network than do lower-level ones (Fig. 2c and Table 1). We further studied how transcription factors in different levels are regulated by kinases. Although there is no significant difference in terms of the number of kinases regulating transcription factors at different levels, we found that if the phosphorylome is arranged into a hierarchy using the same approach used for organizing the transcription factor network, kinases at the bottom tend not to phosphorylate transcription factors, but they tend to be regulated by them (particularly by top-level factors; Supplementary Fig. 7).

Correlation with ncRNAs

We found that top- and middle-level transcription factors have the highest total number of ncRNA targets (Fig. 2c, Table 1 and Supplementary Fig. 6a), consistent with our findings for protein-coding targets. We then developed a score indicating the fraction of a transcription factor's total regulation devoted to ncRNAs, relative to protein-coding genes (Supplementary Information, section E.2); this identified several factors that preferentially target ncRNAs, such as BDP1 and BRF2 (Supplementary Fig. 6b, c).

Matching the pattern for ncRNAs in general, most of the transcription factors involved in miRNA regulation tend to be top- or middle-level ones (Fig. 2c). Moreover, highly connected transcription factors tend to regulate more miRNAs and to be more regulated by them

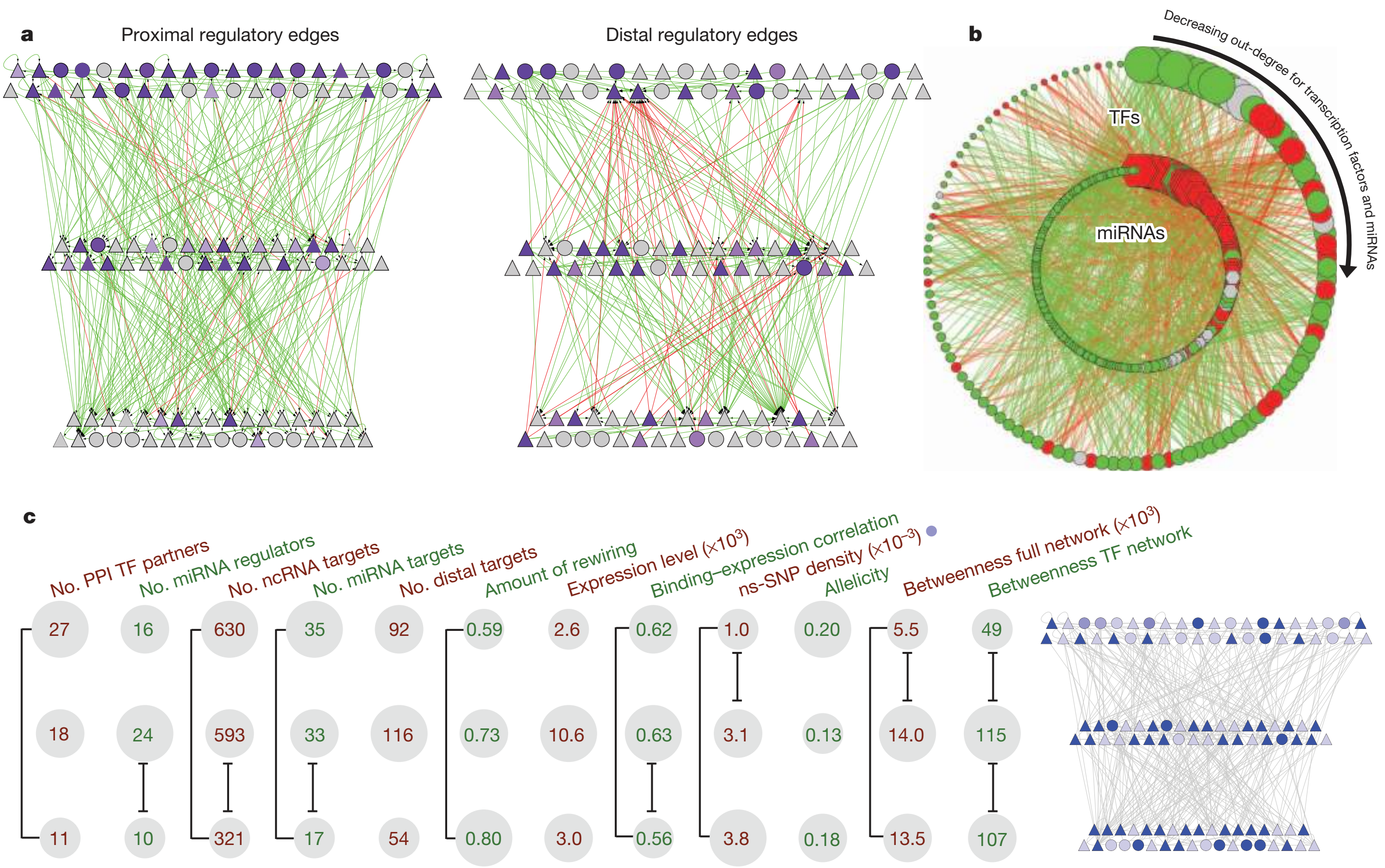

Figure 2 | Overall network. a, Close-up representation of the transcription factor hierarchy. Nodes depict transcription factors. TFSSs are triangles, and non-TFSSs are circles. Left: proximal-edge hierarchy with downward pointing edges coloured in green and upward pointing ones coloured in red. The nodes are shaded according to their out-degree in the full network (as described in Table 1). Right: factors placed in the same proximal hierarchy but now with edges corresponding to distal regulation coloured green and red, and nodes re-coloured according to out-degree in the distal network. The distal edges do not follow the proximal-edge hierarchy. **b**, Close-up view of transcription-factor–miRNA regulation. The outer circle contains the 119 transcription factor, whereas the inner circle contains miRNAs. Red edges correspond to miRNAs regulating transcription factors; green edges show transcription factors regulating miRNAs. Transcription factors and miRNAs each are arranged by their out-degree, beginning at the top (12:00) and decreasing in order clockwise. Node sizes are proportional to out-degree. For transcription factors, the out-degree is as described in Table 1; for miRNAs, it is according to the out-degree in this network. Red nodes are enriched for miRNA–transcription factor edges and green nodes are enriched for transcription factor–miRNA edges. Grey nodes have a balanced number of edges (within ±1). **c**, Average values of various properties (topological, dynamic, expression-related and selection-related—ordered consistently with Table 1) for each level are shown for the proximal-edge hierarchy. The top, middle and bottom rows correspond to the top, middle and bottom of the hierarchy, respectively. The sizing of the grey circles indicates the relative ordering of the values for the three levels. Significantly different values ($P < 0.05$) using the Wilcoxon rank-sum test are indicated by black brackets. The proximal-edge hierarchy depicted on the right shows non-synonymous SNP (ns-SNP) density, where the shading corresponds to the density for the associated factor. (See Supplementary Fig. 4 for more details.)

Table 1 | Correlating properties with centrality and hierarchy height

Category	Property	Correlation with: Degree centrality‡ Full	TF–TF	Betweenness centrality Full	TF–TF	(O − I)/(O + I) TF–TF
Topology	Number of TF partners in PPI	0.28†	0.27†	0.25*	0.33†	0.08
Topology	Number of miRNA regulators	0.24*	0.33†	−0.02	0.00	0.29†
Topology	Number of ncRNA targets	0.65†	0.49†	0.34†	0.35†	0.22*
Topology	Number of miRNA targets	0.62†	0.50†	0.33†	0.34†	0.19*
Topology	Number of distal targets	0.32†	0.24*	0.19*	0.23*	0.07
Dynamics	Amount of rewiring	−0.14	−0.12	0.44*	0.35	−0.42*
Expression	Expression level	0.14	0.12	0.23*	0.27*	−0.04
Expression	Binding–expression correlation	0.41†	0.31†	0.30†	0.36†	0.19*
Selection properties for factors	ns-SNP density	−0.19*	−0.27*	−0.01	−0.03	−0.22
Selection properties for factors	Allelicity	0.20	0.28*	−0.10	−0.16	0.18
Selection properties for targets	ns-SNP density	−0.05†	–	–	–	–
Selection properties for targets	dN/dS	−0.05†	–	–	–	–

Spearman correlation values of various properties (topological, dynamic, expression-related and selection-related) with centrality measures and hierarchy height. Only properties that are significantly correlated with centrality or hierarchy height are listed. For a full set of properties, *P* values and explanations, see Supplementary Tables 4 and 6. dN/dS, non-synonymous to synonymous mutation ratio.

* Spearman correlation $P < 0.05$.

† Spearman correlation $P < 0.01$.

‡ Degree centrality refers to out-degree, except for selection properties on targets, in which case it refers to in-degree. In particular, out-degree in the full transcription factor target network refers to the 'Targets' column in Supplementary Table 4a, and the same quantity is used throughout Fig. 2.

(Table 1 and Fig. 2b). However, when we analyse transcription-factor–miRNA regulation in detail we find that the factors most involved in miRNA regulation tend to either largely regulate or be regulated by miRNAs (Fig. 2b and Supplementary Fig. 4d). That is, there are few high-degree transcription factors with 'balanced regulation' (similar numbers of incoming and outgoing edges, relative to a control; Supplementary Fig. 3m). The same pattern can be seen for miRNAs (Supplementary Fig. 3l).

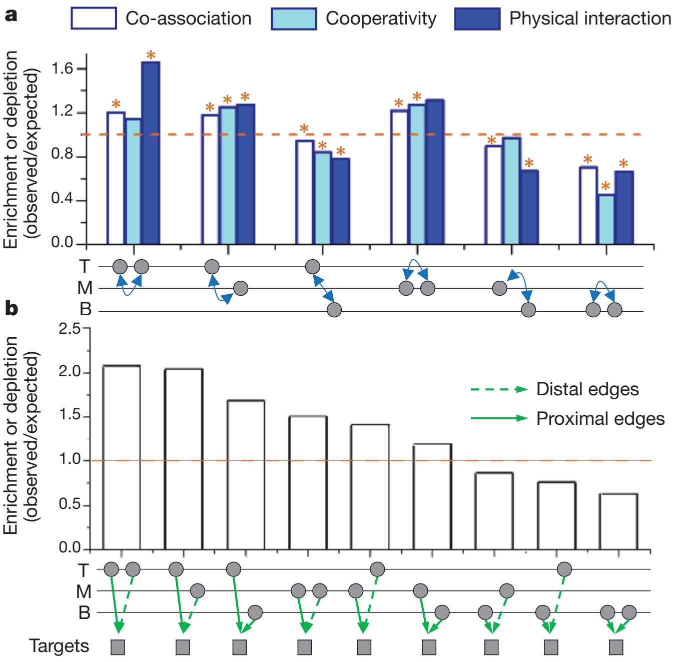

Figure 3 | Collaboration between levels. a, Enrichment of collaborating transcription factor pairs from different levels (top (T), middle (M) and bottom (B)). The factors are represented by two nodes below each bar graph. The dashed orange line indicates the expected level of collaboration. Significant enrichment above or depletion below that level is marked by asterisks ($P < 0.05$). (See Supplementary Information section G.1.2 for more details.) **b**, Enrichment of proximal and distal co-regulatory pairs in the network hierarchy. Co-regulatory pairs from different levels are shown by the two nodes below each bar.

Correlation with families and functional categories

Chromatin-related factors are enriched at the top of the hierarchy, whereas TFSSs are enriched in the middle (Supplementary Table 5a and Supplementary Information, section F.1). Also, TFSSs show a greater degree of tissue specificity and are more highly regulated by miRNAs than are general and chromatin-related factors (Supplementary Information, section F.4), indicating that they may be more finely tuned in their expression. Examining functional enrichment, we found that transcription factors at the top of the hierarchy tend to have more general functions, and those at the bottom tend to have more specific functions (Supplementary Table 5c and Supplementary Information, section F.1).

Correlation with network dynamics

We studied how transcription factors change their binding patterns among different cell types, principally between the K562 and GM12878 cell lines. We quantified the amount of 'rewiring' as the fraction of unshared targets, normalized by the union of two target sets (Supplementary Information, section 3.5). We found that this 'rewiring score' is negatively correlated with hierarchy height (Fig. 2c and Table 1). This means that the targets of lower-level transcription factors tend to change more between cell types, consistent with their role in more specialized processes.

Correlation with gene expression

We calculated the average expression levels of transcription factors across 34 tissues[26]; highly connected factors tend to be highly expressed. We further examined the relationship between connectivity and expression by calculating, for each transcription factor, the correlation between its binding signal around its targets and the level of target expression (Supplementary Information, section F.3.4). This binding–expression correlation is positively correlated with factor connectivity. Moreover, transcription factors at the top and middle levels show a greater correlation. Thus, more 'influential' transcription factors tend to be better connected and higher in the hierarchy. (This degree of 'influence" becomes even clearer when one considers weighting the correlation by the number of transcription factor targets, given that higher-level factors tend to have more targets.) However, somewhat surprisingly, a model integrating the binding–expression relationships of all the highly connected transcription factors has about the same predictive power for expression as a model integrating all the less connected ones, indicating that the weak binding–expression relationships of the less influential factors are collectively quite influential (Supplementary Information, section F.3.4)[38].

Collaboration between hierarchy levels

We explored how transcription factors in the top, middle and bottom (T, M and B, respectively) levels of the hierarchy collaborate, in terms of both inter-level (TM, MB, TB) and intra-level (TT, MM, BB) relationships (Fig. 3a). We examined three kinds of collaboration: co-association (as described earlier), physical interactions, and target-expression cooperativity. We defined two transcription factors as being cooperative if their shared targets are significantly different in expression from their unshared targets (Supplementary Information, section G.2). Overall, we found that collaborations involving the middle level (and to a lesser extent, the top one) tended to be enriched. In particular, TM and MM transcription factor pairs influenced gene expression cooperatively. Next, all co-associations involving top- and middle-level factors are enriched, whereas those involving the bottom level are depleted. A similar pattern was observed for protein–protein interactions, with TT and TM co-regulation more likely to occur between physically interacting transcription factors (Fig. 3a and Supplementary Information, section G.1).

Finally, we analysed how proximal and distal sites 'collaborate'. We identified pairs of transcription factors that bind to the promoter and distal regulatory regions of the same target gene (Supplementary Information, section G.3) and studied their respective locations in the factor hierarchy. We found an asymmetry between proximal and distal regulation, with transcription factors associated through promoter regulation more likely to reside in upper levels (Fig. 3b).

Enriched network motifs

Apart from its global structure, we further studied the network from the perspective of its constituent building blocks; that is, network motifs, which are small connectivity patterns that carry out canonical functions[39]. We systematically searched for motifs, first in the promoter-regulation hierarchy and then in the meta-network including distal, miRNA and protein–protein interactions. Our procedure was to instantiate all possible motifs for broad template patterns and then determine which of these were significantly over- or under-represented relative to a random control[40] (Supplementary Information, section H). For instance, starting with all possible three-transcription-factor motifs in the proximal network (Fig. 4a), we found the most enriched motif to be the well-studied feed-forward loop (FFL)[39]. In agreement with the observed collaborations within the hierarchy, many FFLs involve the middle level (Supplementary Fig. 9a). Moreover, by analysing the expression levels of the constituent genes of the FFLs over many tissues, we found that many were positively correlated, highlighting the tight regulation implicit in the motif (Fig. 4a and Supplementary Information, section H.1).

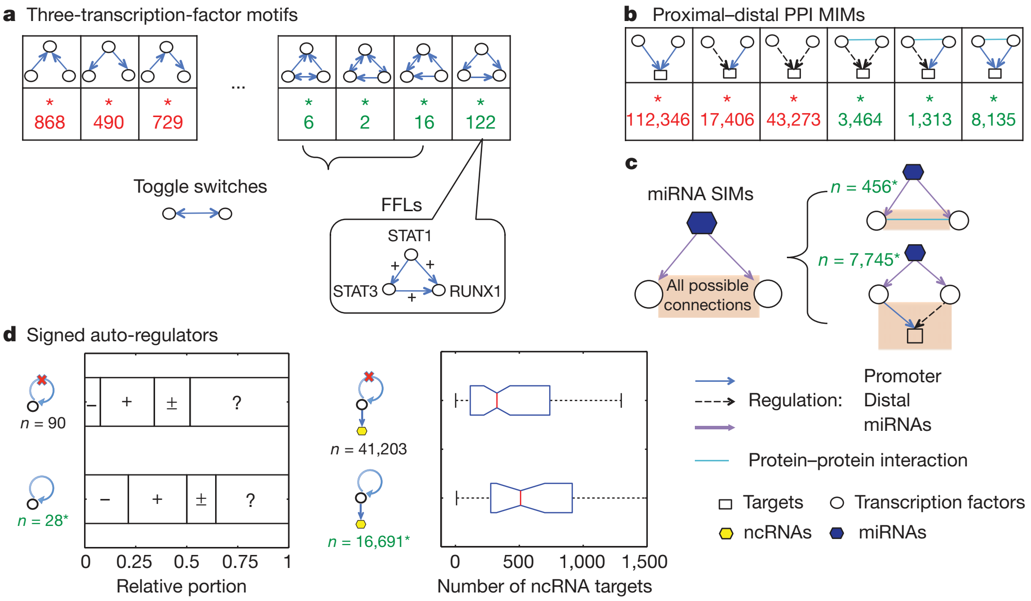

Figure 4 | Motif analysis. Motifs are accompanied by the number of occurrences, *n*. Enriched motifs are highlighted in green; depleted ones in red. An asterisk means that the corresponding enrichment/depletion is statistically significant ($P = 1 \times 10^{-5}$). The motifs are sorted such that those at the ends have more significant *P* values. (See Supplementary Fig. 9h for more details.) **a**, Systematic search of three-transcription-factor motifs. The most enriched motif is the FFL. A particular example formed by STAT1, STAT3 and RUNX1 is highlighted. Here, the '+' symbol on an edge indicates that the correlation between the gene expression of the source and the target across tissues is positive. Other motifs containing a toggle-switch regulation on top of the basic FFL design are also indicated. **b**, Proximal–distal PPI MIMs. Here we searched all motifs involving the co-regulation of two transcription factors (which could be either proximal or distal) with (or without) a protein–protein interaction between them. Motifs containing the protein–protein interaction tended to be enriched. **c**, miRNA SIMs. The two enriched motifs resulting from enumerating all motifs in which a miRNA targets two transcription factors that are connected in various ways are shown. These two motifs contain a protein complex of two transcription factors and a cooperative pair of promoter and distal regulatory transcription factors. **d**, The auto-regulator motif is enriched in the transcription factor–transcription factor network: 28 of all factors are auto-regulators. Moreover, auto-regulators are more likely to be repressors (−) relative to non-auto regulators, and they tend to have more ncRNAs as their targets. In the box plots, the red line indicates the median, the blue box shows the interquartile range (IQR), and whiskers extend out to 1.5 IQR.

Finally, we found further enriched three-transcription-factor motifs containing an additional regulation on top of that in a FFL. This creates a mutual regulation between a pair of transcription factors, instantiating a toggle-switch, which has been shown to have an essential role in the determination of cell fate[41].

Next, we analysed another template: all possible multiple-input modules (MIMs, defined in Supplementary Information, section K) involving promoter and distal regulation and a protein–protein interaction (proximal–distal PPI MIMs, Fig. 4b). We found that co-regulating transcription factors are likely to interact physically, indicating that they work together as a complex. Moreover, the motif ranking second in enrichment consists of a distal regulatory relationship, a promoter regulatory relationship, and a protein–protein interaction. This is suggestive of a common picture of DNA looping, with an interacting complex of transcription factors binding to the promoter and enhancer simultaneously.

The connection between co-regulated entities extends to miRNA regulation. We surveyed all possible instances of a miRNA regulating two transcription factors (miRNA SIM, Fig. 4c) and found that the miRNAs are more likely to regulate a pair of physically interacting factors. This enrichment indicates that, to avoid unwanted cross-talk, a miRNA tends to shut down an entire functional unit (that is, transcription factor complex) rather than just a single component. Similarly, we found that miRNAs tend to target a pair of transcription factors binding both proximally and distally (Fig. 4c). This suggests that miRNA represses the expression of both promoter and distal regulators to shut down a target completely. Apart from miRNAs, we also studied motifs involving other kinds of ncRNAs. Among motifs involving a transcription factor regulating two ncRNAs, there is great enrichment for both ncRNAs to be long intergenic noncoding RNAs (lincRNAs) (Supplementary Information, section H.2).

Finally, we found the network to be enriched for auto-regulators (28 out of 119 transcription factors), a simple but important motif, which are commonly found in networks exhibiting multistability[42]. Moreover, we found that the auto-regulators tend to be repressors, representing a well known design principle for maintaining steady state[39] (Fig. 4d).

Allelic behaviour in a network framework

We examined the relationship between sequence variation and transcription factor regulation. In particular, we investigated the coordination between allele-specific binding and allele-specific expression[43,44]. We used the sequenced data sets for the GM12878 cell line, which has a deeply sequenced diploid genome (Supplementary Information, section I.1). We extended pairwise analysis of allele-specific behaviour[20] to study higher-order coordination of multiple factors regulating a common target. We first generated the unfiltered, promoter-regulation network for GM12878 cells and then identified a subnetwork within it representing the difference between maternal- and paternal-specific networks (Supplementary Information, section I.2). This subnetwork is shown in Fig. 5a, with 4,798 transcription-factor-target edges coloured red or blue to represent predominantly maternally or paternally regulated targets; the targets are similarly coloured to indicate predominantly maternal or paternal expression. We found that of the 4,798 allele-specific binding cases of a single factor regulating its associated target, 57% showed coordinated allelic binding and expression. We then found that for the cases in which two transcription factors regulate a common target, 63% were consistent (that is, both factors bind to the same allele that is expressed). For those cases in which triplets of transcription factors regulate a common target, the consistency increased to 65%. This trend continues, demonstrating that, as one increases the degree of combinatorial regulation, there is a progressively stronger relationship between expressed and regulated alleles.

The degree of allele-specific behaviour of each transcription factor can be quantified by a statistic that we call 'allelicity'. The allelicity of a transcription factor is defined as the fraction of single nucleotide

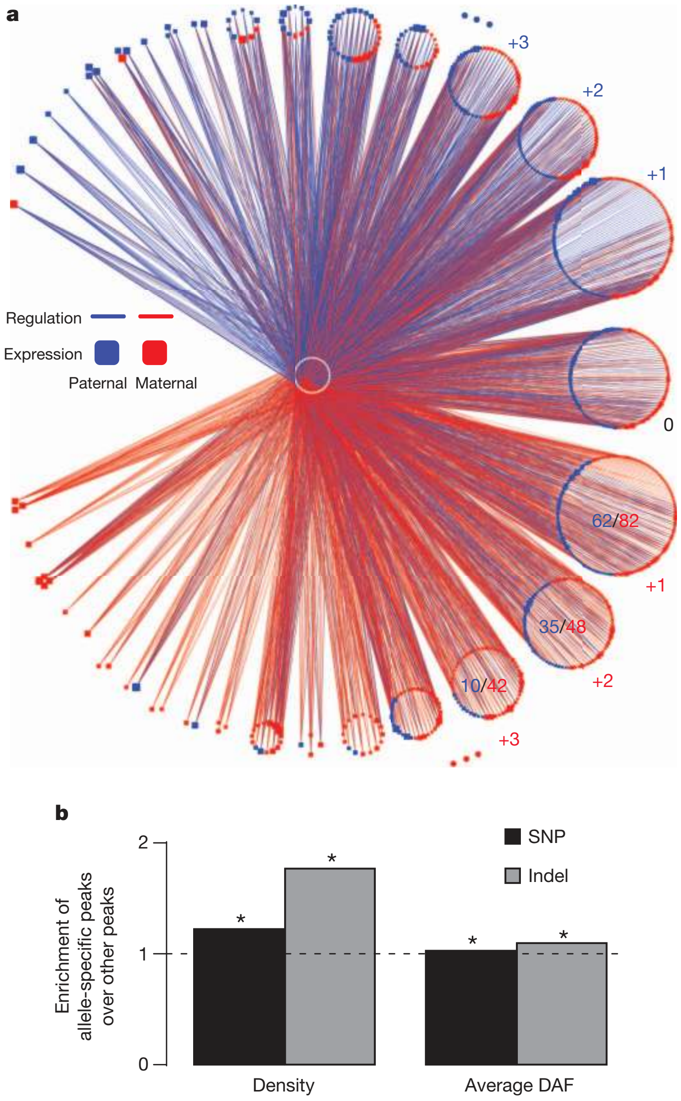

Figure 5 | Allelic effects. a, An 'allelic effects network' depicting the increasing coordination between allele-specific binding and allele-specific expression as the number of factors regulating a target increases. Central white nodes denote transcription factors, and peripheral nodes denote targets, which are blue (red) if they are expressed from the paternal (maternal) allele. Blue (red) edges denote allele-specific binding to the paternal (maternal) allele. This network represents the strongest differences between the paternal- and maternal-specific regulatory networks. As one goes around the larger circle anticlockwise (clockwise), each of the small circular clusters represents targets with progressively more paternal (maternal) regulation, indicated by the small blue (red) numbers to the side of the clusters. Moreover, within each of the clusters the fraction of predominantly paternally (maternally) expressed targets increases as one goes around the larger circle. As an illustration, this fraction is explicitly indicated by the ratios within three of the larger clusters at the bottom right. **b**, Relationship between transcription factor allelicity and selection. The bar height is the ratio of the degree of selection (as measured by SNP density or average DAF) in those binding peaks showing allelic behaviour to the degree of selection in all other binding peaks. Asterisks represent significant differences ($P < 0.05$, Wilcoxon rank-sum test). (See Supplementary Information section I.2 and Supplementary Fig. 10b, c for details.)

polymorphisms (SNPs) that exhibit allele-specific binding out of all the SNPs that may potentially exhibit it (Supplementary Information, section I.3). Thus, qualitatively, allelicity may be thought of as the sensitivity of a transcription factor's binding to maternal-versus-paternal variants. Using our network described here, we find that transcription factors with higher degrees of allelicity tend to have more target genes, indicating that these factors tend to vary more in their binding with sequence (Table 1). Finally, we found that small insertions and deletions (indels) tended to cause disproportionally more of these allelic events than did SNPs (Supplementary Table 6g).

Selection in a network context

Previous studies have examined the relationship between evolutionary selection and position in the human protein–protein interaction network[45]. However, the analogous relationship in the regulatory network has not yet been explored.

Selection

To address this, we first analysed the selective pressure on both transcription factors and their targets. We predominantly used non-synonymous SNP density from the 1000 Genomes Pilot[21] to determine selection among modern-day humans (Supplementary Information, section J). We also verified our results using other measures of selection (that is, derived allele frequency (DAF) and the ratio of non-synonymous to synonymous SNP rates (pN/pS statistic) (Supplementary Information, section J)). For selection over longer timescales, we calculated the ratio of non-synonymous to synonymous substitution rates in human–chimp orthologue alignments (dN/dS). We found significant negative correlation between the regulatory in-degree of target genes and both their non-synonymous SNP density and dN/dS values (Table 1 and Supplementary Table 6e). Thus, target genes regulated by more transcription factors are under stronger negative selection. Similarly, we found that there is a significant negative correlation between transcription factor regulatory out-degree and non-synonymous SNP density (Table 1 and Supplementary Table 6d). We observed a consistent result with transcription factor dN/dS values and other measures of selection, although these are not all as statistically significant (Supplementary Table 6d and Supplementary Information, section J). This shows that transcription factors regulating more targets tend to be under stronger negative selection. Moreover, within the transcription factor hierarchy, we found that factors at the top are under significantly stronger negative selection (Fig. 2c, Table 1 and Supplementary Table 6b).

Consistent with all of these results relating connectivity with constraint, we found that genes tolerant of loss-of-function mutations[46], which are under weaker negative selection, have a significantly lower total degree $(I + O)$ than other genes (Supplementary Information, section J).

Selection and allelic effects

Finally, we attempted to relate selection and allelic effects. We extracted transcription-factor-binding peaks in promoters and gene bodies showing allele-specific binding, and compared the selective pressure in these against a control (binding peaks within the same regions without allele-specific binding). We found that transcription-factor-binding peaks exhibiting allelic effects have higher SNP densities relative to the control (Fig. 5b). Moreover, binding peaks with no allelic effects show a skew in the DAF spectrum towards rarer SNPs, relative to allele-specific binding ones (Fig. 5b and Supplementary Fig. 10c). The same trend holds true for indels and structural variants (Fig. 5b and Supplementary Fig. 10b, c). Interestingly, these results indicate that allelic regulation seems to be under less selective constraint.

Discussion

This study provides the first detailed analysis of how human regulatory information is organized. A number of clear design principles emerge from it. Many of these are shared with model organisms (Supplementary Table 7), demonstrating that they are general features of transcription factor regulation. First, we found that the connectivity and hierarchical organization of regulatory factors is reflected in many genomic properties. For instance, top-level transcription factors have their binding more strongly correlated with the expression of their targets, perhaps indicating that they are more influential, as reported for model organisms[47]. Next, the middle-level contains information-flow bottlenecks and much connectivity with miRNA and distal regulation. Targeting these bottlenecks (for example, by drugs) is likely to most strongly affect the flow of information through regulatory circuits. To some degree, the cell mitigates the effect of bottlenecks by having pairs of middle-level transcription factors collaborate in regulation. (Co-regulation mitigates bottlenecks.) Third, the regulatory network seems to be built from repeated reuse of small, modular motifs. In particular, regulation between levels involves many feed-forward loops, which could be used to filter fluctuations in input stimuli. Again, these properties are shared with model organisms; the network motifs and cooperating middle-level have been observed in yeast[48].

By contrast, the differences in proximal and distal regulation seem to be a unique feature of human regulation. This finding is evident in the analysis of both transcription factor co-association and network structure. The proximal–distal differences reflect the much larger intergenic space in humans than model organisms and the commensurately larger amount of distal binding. Finally, analysis of conservation indicates that more highly connected parts of the network are under stronger selection, consistent with results from model organisms. However, one unique finding for humans is 'allelic' effects. More highly connected transcription factors are more likely to exhibit allele-specific binding. Interestingly, we found that the actual allele-specific binding sites tend to be under less selection. Unravelling this interaction between selection and regulatory networks will be crucial to interpreting variants in the many personal genome sequences expected in the future. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

Detailed methods associated with each section of the paper are in a similarly titled section of the Supplementary Information. In particular, an overview of our data processing pipeline is in Supplementary Information, section B.

Received 9 December 2011; accepted 22 May 2012.

1. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* **298,** 799–804 (2002).
2. Balazsi, G., Barabasi, A. L. & Oltvai, Z. N. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli. Proc. Natl Acad. Sci. USA* **102,** 7841–7846 (2005).
3. Yu, H. Y. & Gerstein, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl Acad. Sci. USA* **103,** 14724–14731 (2006).
4. Hu, Z. Z., Killion, P. J. & Iyer, V. R. Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genet.* **39,** 683–687 (2007).
5. Balaji, S., Babu, M. M. & Aravind, L. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli. J. Mol. Biol.* **372,** 1108–1122 (2007).
6. Jothi, R. *et al.* Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol. Syst. Biol.* **5,** 294 (2009).
7. Barabási, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nature Rev. Genet.* **5,** 101–113 (2004).
8. Kim, H. D., Shay, T., O'Shea, E. K. & Regev, A. Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science* **325,** 429–432 (2009).
9. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296,** 910–913 (2002).
10. Ma, H. W., Buer, J. & Zeng, A. P. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5,** 199 (2004).
11. Balaji, S., Iyer, L. M., Aravind, L. & Babu, M. M. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J. Mol. Biol.* **360,** 204–212 (2006).
12. Milo, R. *et al.* Network motifs: Simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
13. Cosentino Lagomarsino, M., Jona, P., Bassetti, B. & Isambert, H. Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proc. Natl Acad. Sci. USA* **104,** 5516–5520 (2007).
14. Ptacek, J. *et al.* Global analysis of protein phosphorylation in yeast. *Nature* **438,** 679–684 (2005).
15. Beyer, A., Bandyopadhyay, S. & Ideker, T. Integrating physical and genetic maps: from genomes to interaction networks. *Nature Rev. Genet.* **8,** 699–710 (2007).
16. Yu, H. Y., Xia, Y., Trifonov, V. & Gerstein, M. Design principles of molecular networks revealed by global comparisons and composite motifs. *Genome Biol.* **7,** R55 (2006).
17. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117 (2008).
18. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
19. Reed, B. D., Charos, A. E., Szekely, A. M., Weissman, S. M. & Snyder, M. Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional toles and combinatorial regulation of distinct classes of genes. *PLoS Genet.* **4,** e1000133 (2008).
20. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* http://dx.doi.org/10.1038/nature11247 (this issue).
21. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
22. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
23. Barski, A. *et al.* Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res.* **19,** 1742–1751 (2009).
24. Ozsolak, F. *et al.* Chromatin structure analyses identify miRNA promoters. *Genes Dev.* **22,** 3172–3183 (2008).
25. Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39,** D698–D704 (2011).
26. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140,** 744–752 (2010).
27. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144,** 296–309 (2011).
28. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10,** 252–263 (2009).
29. Kerenyi, M. A. & Orkin, S. H. Networking erythropoiesis. *J. Exp. Med.* **207,** 2537–2541 (2010).
30. Curran, T. & Franza, B. R. Fos and Jun: the AP-1 Connection. *Cell* **55,** 395–397 (1988).
31. Chinenov, Y. & Kerppola, T. K. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene* **20,** 2438–2452 (2001).
32. Rubio, E. D. *et al.* CTCF physically links cohesin to chromatin. *Proc. Natl Acad. Sci. USA* **105,** 8309–8314 (2008).
33. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132,** 422–433 (2008).
34. Cheng, C., Min, R. & Gerstein, M. TIP: A probabilistic method for identifying transcription factor target genes from ChIP-Seq binding profiles. *Bioinformatics* **27,** 3221–3227 (2011).
35. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
36. Friedman, R. C., Farh, K. K. H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).
37. Baron, M. H. & Farrington, S. M. Positive regulators of the lineage-specific transcription factor GATA-1 in differentiating erythroid cells. *Mol. Cell. Biol.* **14,** 3108–3114 (1994).
38. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* http://dx.doi.org/10.1101/gr.136838.111 (2012).
39. Alon, U. Network motifs: theory and experimental approaches. *Nature Rev. Genet.* **8,** 450–461 (2007).
40. Cheng, C. *et al.* Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.* **7,** e1002190 (2011).
41. Zhou, J. X. & Huang, S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends Genet.* **27,** 55–62 (2011).
42. Burda, Z., Krzywicki, A., Martin, O. C. & Zagorski, M. Motifs emerge from function in model gene regulatory networks. *Proc. Natl Acad. Sci. USA* **108,** 17263–17268 (2011).
43. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (2010).
44. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
45. Kim, P. M., Korbel, J. O. & Gerstein, M. B. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proc. Natl Acad. Sci. USA* **104,** 20274–20279 (2007).
46. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).
47. Bhardwaj, N., Kim, P. M. & Gerstein, M. B. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci. Signal.* **3,** ra79 (2010).
48. Bhardwaj, N., Yan, K.-K. & Gerstein, M. B. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proc. Natl Acad. Sci. USA* **107,** 6841–6846 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the ENCODE Project (National Institutes of Health/National Human Genome Research Institute (NIH/NHGRI)) for funding. We thank P. Bickel and B. Brown for conversations. Funding has also been provided by the NIH Predoctoral Training Program in Biophysics (D.C.; T32 GM008283-24) and the Sackler institute (G.Z.-S.). M.H. thanks G. Nair for designing database operations.

Author Contributions Work on the paper was divided between data production and analysis. The analysts were A.A., R.A., P.A., S.B., N.B., D.Z.C., C.C., D.C., Y.F., M.H., A.H., E.K., A.K., J.Le., R.M., X.J.M., J.R., A.S., J.W., K.-K.Y., K.Y.Y. and G.Z.-S. The data producers were N.A., A.P.B., P.C., A.C., Y.C., C.E., G.E., P.J.F., S.F., J.G., F.G., P.J., M.K., P.L., S.G.L., J.Li., H.M., R.M.M., H.O'G., Z.O., E.C.P., D.P., F.P., D.R., L.R., T.E.R., B.R., M.Sh., T.S., S.M.W., L.W. and X.Y. Larger efforts in analysis and data production are ascribed to the joint first authors. Author contributions to specific exhibits and files are shown in Supplementary Information, sections N and O. Overall project management was carried out by the two corresponding authors, M.B.G. and M.Sn.

Author Information Data sets described here can be obtained from the ENCODE project website at http://www.encodeproject.org and from http://encodenets.gersteinlab.org. More detail on data availability is in Supplementary Information, sections B and N. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.G. (mark.gerstein@yale.edu) or M.Sn. (mpsnyder@stanford.edu).