Cell

Resource

Circuitry and Dynamics of Human Transcription Factor Regulatory Networks

Shane Neph,[1,5] Andrew B. Stergachis,[1,5] Alex Reynolds,[1] Richard Sandstrom,[1] Elhanan Borenstein,[1,2,4,*] and John A. Stamatoyannopoulos[1,3,*]

[1]Department of Genome Sciences
[2]Department of Computer Science and Engineering
[3]Department of Medicine
University of Washington, Seattle, WA 98195, USA
[4]Santa Fe Institute, Santa Fe, NM 87501, USA
[5]These authors contributed equally to this work
*Correspondence: elbo@uw.edu (E.B.), jstam@uw.edu (J.A.S.)
http://dx.doi.org/10.1016/j.cell.2012.04.040

SUMMARY

The combinatorial cross-regulation of hundreds of sequence-specific transcription factors (TFs) defines a regulatory network that underlies cellular identity and function. Here we use genome-wide maps of in vivo DNaseI footprints to assemble an extensive core human regulatory network comprising connections among 475 sequence-specific TFs and to analyze the dynamics of these connections across 41 diverse cell and tissue types. We find that human TF networks are highly cell selective and are driven by cohorts of factors that include regulators with previously unrecognized roles in control of cellular identity. Moreover, we identify many widely expressed factors that impact transcriptional regulatory networks in a cell-selective manner. Strikingly, in spite of their inherent diversity, all cell-type regulatory networks independently converge on a common architecture that closely resembles the topology of living neuronal networks. Together, our results provide an extensive description of the circuitry, dynamics, and organizing principles of the human TF regulatory network.

INTRODUCTION

Sequence-specific transcription factors (TFs) are the key effectors of eukaryotic gene control. Human TFs regulate hundreds to thousands of downstream genes (Johnson et al., 2007). Of particular interest are interactions in which a given TF regulates other TFs, or itself. Such mutual cross-regulation among groups of TFs defines regulatory subnetworks that underlie major features of cellular identity and complex functions such as pluripotency (Boyer et al., 2005; Kim et al., 2008), development (Davidson et al., 2002a), and differentiation (Yun and Wold, 1996). On a broader level, cross-regulatory interactions among the entire complement of TFs expressed in a given cell type form a core transcriptional regulatory network, endowing the cell with systems-level properties that facilitate the integration of complex cellular signals, while conferring additional nimbleness and robustness (Alon, 2006). However, despite their central biological roles, both the structure of core human regulatory networks and their component subnetworks are largely undefined.

One of the main bottlenecks limiting generation of TF regulatory networks for complex biological systems has been that information is traditionally collected from individual experiments targeting one cell type and one TF at a time (Davidson et al., 2002a; Yuh et al., 1994; Kim et al., 2008; Roy et al., 2010; Gerstein et al., 2010). For example, the sea urchin endomesoderm regulatory network was constructed by individually perturbing the expression and activity of several dozen TFs and analyzing the effect of these perturbations on the expression of TF genes containing putative *cis*-regulatory binding elements for these factors (Davidson et al., 2002b; Yuh et al., 1994). More recently, genome-wide analysis combining chromatin immunoprecipitation of individual TFs with high-throughput sequencing (ChIP-seq) has been used to derive subnetworks of small numbers of TFs, such as those involved in pluripotency (Kim et al., 2008), or larger-scale networks combining several dozen TFs (Roy et al., 2010; Gerstein et al., 2010). However, such approaches are limited by three major factors: (1) the availability of suitable affinity reagents; (2) the difficulty of interrogating the activities of multiple TFs within the same cellular environment; and, perhaps most critically, (3) the sizable number of TFs and cellular states that need to be studied. De novo network construction methods based on gene expression correlations partly overcome the limitation of studying one TF at a time but lack directness and typically require several hundred independent gene expression perturbation studies to build a network for one cell type (Basso et al., 2005; Carro et al., 2010). Similarly, yeast one-hybrid assays offer a high-throughput approach for identifying *cis*-regulatory element binding partners (Walhout,

2006; Reece-Hoyes et al., 2011). However, such assays lack native cellular context, limiting their direct utility for building cell-type-specific networks. Given these experimental limitations, only a handful of well-described multicellular transcriptional regulatory networks have been defined, and those that do exist are often incomplete despite the numerous experiments and extended time (typically years) needed to construct them (Davidson et al., 2002a; Basso et al., 2005; Boyer et al., 2005; Kim et al., 2008; Roy et al., 2010; Gerstein et al., 2010).

Given that the human genome encodes > 1,000 TFs (Vaquerizas et al., 2009) and that human cellular diversity spans hundreds of different cell types and an even greater number of cellular states, we sought to develop an accurate and scalable approach to analyze transcriptional regulatory networks suitable for the application to any cellular or organismal state. The discovery of DNaseI footprinting over 30 years ago (Galas and Schmitz, 1978) revolutionized the analysis of regulatory sequences in diverse organisms and directly enabled the discovery of the first human sequence-specific TFs (Dynan and Tjian, 1983). In the context of living nuclear chromatin, DNaseI treatment preferentially cleaves the genome within highly accessible active regulatory DNA regions, creating DNaseI-hypersensitive sites (DHSs) (Wu et al., 1979; Kuo et al., 1979; Wu, 1980; Stalder et al., 1980). Within DHSs, DNaseI cleavage is not uniform but is rather punctuated by sequence-specific regulatory factors that occlude bound DNA, leaving "footprints" that demarcate TF occupancy at nucleotide resolution (Hesselberth et al., 2009; Pfeifer and Riggs, 1991). DNaseI footprinting is a well-established method for identifying direct regulatory interactions and provides a powerful generic approach for assaying the occupancy of specific sequence elements with *cis*-regulatory functions (Karin et al., 1984; Kadonaga et al., 1987).

DNaseI footprinting has been applied widely to study regulatory interactions between TFs and to identify cell- and lineage-selective transcriptional regulators (Dynan and Tjian, 1983; Karin et al., 1984; Tsai et al., 1989). In the context of the ENCODE Project, we applied digital genomic footprinting (Hesselberth et al., 2009) to delineate millions of human DNaseI footprints genome-wide in 41 diverse cell types. Combining DNaseI footprints with defined TF recognition sequences accurately and quantitatively recapitulates ChIP-seq data for individual TFs, while simultaneously interrogating the genomic occupancy of potentially all expressed DNA-binding factors in a single experiment (Neph et al., 2012a).

By performing systematic analysis of TF footprints in the proximal regulatory regions of each TF gene, we develop a foundational experimental paradigm for comprehensive, unbiased mapping of the complex network of regulatory interactions between human TFs. In such networks, TFs comprise the network "nodes," and the cross-regulation of one TF by another the interactions or network "edges." Furthermore, iterating this paradigm across diverse cell types provides a powerful system for analysis of TF network dynamics in a complex organism. Here, we use genome-wide maps of in vivo DNaseI footprints to assemble an extensive core human regulatory network comprising connections among 475 sequence-specific TFs and analyze the dynamics of these connections across 41 diverse cell and tissue types.

RESULTS

Comprehensive Mapping of TF Networks in Diverse Human Cell Types

To generate TF regulatory networks in human cells, we analyzed genomic DNaseI footprinting data from 41 diverse cell and tissue types (Neph et al., 2012a). Each of these 41 samples was treated with DNaseI, and sites of DNaseI cleavage along the genome were analyzed with high-throughput sequencing. At an average sampling depth of ~500 million DNaseI cleavages per cell type (of which ~273 million mapped to unique genomic positions), we identified an average of ~1.1 million high-confidence DNaseI footprints per cell type (range 434,000 to 2.3 million at a false discovery rate of 1% [FDR 1%]; Neph et al., 2012a). Collectively, we detected 45,096,726 footprints, representing cell-selective binding to ~8.4 million distinct 6–40 bp genomic sequence elements. We used well-annotated databases of TF-binding motifs to infer the identities of factors occupying DNaseI footprints (Wingender et al., 1996; Bryne et al., 2008; Newburger and Bulyk, 2009) (Experimental Procedures) and confirmed that these identifications matched closely and quantitatively with ENCODE ChIP-seq data for the same cognate factors (Neph et al., 2012a).

To generate a TF regulatory network for each cell type, we analyzed actively bound DNA elements within the proximal regulatory regions (i.e., all DNaseI hypersensitive sites within a 10 kb interval centered on the transcriptional start site [TSS]) of 475 TF genes with well-annotated recognition motifs (Wingender et al., 1996; Bryne et al., 2008; Newburger and Bulyk, 2009) (Figure 1A). Repeating this process for every cell type disclosed a total of 38,393 unique, directed (i.e., TF-to-TF) regulatory interactions (edges) among the 475 analyzed TFs, with an average of 11,193 TF-to-TF edges per cell type (Data S1). Given the functional redundancy of a minority of DNA-binding motifs (Berger et al., 2008), in certain cases multiple factors could be designated as occupying a single DNaseI footprint. However, most commonly, mappings represented associations between single TFs and a specific DNA element. Because DNaseI hypersensitivity at proximal regulatory sequences closely parallels gene expression (The ENCODE Project Consortium, 2012), the annotation process we utilized naturally focuses on the expressed TF complement of each cell type, enabling the construction of a comprehensive transcription regulatory network for a given cell type with a single experiment.

De Novo-Derived Networks Accurately Recapitulate Known TF-to-TF Circuitry

To assess the accuracy of cellular TF regulatory networks derived from DNaseI footprints, we analyzed several well-annotated mammalian cell-type-specific transcriptional regulatory subnetworks (Figures 1B and 1C). The muscle-specific factors MyoD, Myogenin (MYOG), MEF2A, and MYF6 form a network that was uncovered using a combination of genetic and physical studies, including DNaseI footprinting, and is vital for specification of skeletal muscle fate and control of myogenic development and differentiation (Naidu et al., 1995; Yun and Wold, 1996; Ramachandran et al., 2008). Figure 1B juxtaposes the known regulatory interactions between these factors determined