Developmental Biology 413 (2016) 128–144

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/locate/developmentalbiology

Genomes and Developmental Control

The analysis of novel distal *Cebpa* enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification

Eric Bertolino [a,*], John Reinitz [a,b,c], Manu [c,d,*,1]

[a] Department of Molecular Genetics and Cell Biology, The University of Chicago, Chicago, IL 60637, USA
[b] Department of Statistics, The University of Chicago, Chicago, IL 60637, USA
[c] Department of Ecology and Evolution and Institute of Genomics and Systems Biology, The University of Chicago, Chicago, IL 60637, USA
[d] Department of Biology, University of North Dakota, 10 Cornell Street, Stop 9019, Grand Forks, ND 58202-9019, USA

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form
13 January 2016
Accepted 15 February 2016
Available online 3 March 2016

Keywords:
Cell fate
Gene regulation
Hematopoiesis
Silencers
Transcriptional modeling

ABSTRACT

C/EBPα plays an instructive role in the macrophage-neutrophil cell-fate decision and its expression is necessary for neutrophil development. How *Cebpa* itself is regulated in the myeloid lineage is not known. We decoded the *cis*-regulatory logic of *Cebpa*, and two other myeloid transcription factors, *Egr1 and Egr2*, using a combined experimental-computational approach. With a reporter design capable of detecting both distal enhancers and silencers, we analyzed 46 putative *cis*-regulatory modules (CRMs) in cells representing myeloid progenitors, and derived early macrophages or neutrophils. In addition to novel enhancers, this analysis revealed a surprisingly large number of silencers. We determined the regulatory roles of 15 potential transcriptional regulators by testing 32,768 alternative sequence-based transcriptional models against CRM activity data. This comprehensive analysis allowed us to infer the *cis*-regulatory logic for most of the CRMs. Silencer-mediated repression of *Cebpa* was found to be effected mainly by TFs expressed in non-myeloid lineages, highlighting a previously unappreciated contribution of long-distance silencing to hematopoietic lineage resolution. The repression of *Cebpa* by multiple factors expressed in alternative lineages suggests that hematopoietic genes are organized into densely interconnected repressive networks instead of hierarchies of mutually repressive pairs of pivotal TFs. More generally, our results demonstrate that *de novo cis*-regulatory dissection is feasible on a large scale with the aid of transcriptional modeling.

[1]Current address: Department of Biology, University of North Dakota, 10 Cornell Street, Stop 9019, Grand Forks, ND 58202-9019, USA.

1. Introduction

The spatiotemporal expression of genes is encoded in the genome by *cis*-regulatory sequences, which may be located tens to hundred of kilobases from the transcription start site (Carey et al., 2008; Spitz and Furlong, 2012). It is usually possible to empirically define *cis*-regulatory modules (CRMs) as sequences that act as enhancers (Banerji et al., 1983; Banerji et al., 1981) or silencers (Brand et al., 1985; Ogbourne and Antalis, 1998) of the activity of the core promoter in reporter assays. The activity of CRMs results from sequence-specific transcription factors that bind to their recognition sites and recruit cofactors which interact with the RNA polymerase II holoenzyme complex or remodel chromatin (Spitz and Furlong, 2012). Careful analysis of the CRMs of a few well-characterized genes (Fromental et al., 1988; Göttgens et al., 2002; Ondek et al., 1988; Schirm et al., 1987; Small et al., 1992; Wilson et al., 2011; Yuh et al., 1998) has revealed how the internal composition and structure of CRMs—the arrangement of transcription factor binding sites (TFBS), the TFs binding to them, and interactions between bound TFs—encodes the pattern of gene expression. For the vast majority of genes however, both the identities of CRMs as well as their *cis*-regulatory logic remain unknown.

Determining the *cis*-regulatory logic of individual genes is an important goal of functional genomics (Nam et al., 2010). First and foremost, determining the *cis*-regulatory logic of individual genes is a prerequisite for constructing high-quality gene regulatory networks (GRNs) (Levine and Davidson, 2005; Singh et al., 2014) and modeling them predictively. Second, even though the putative rules

* Corresponding authors.
E-mail addresses: eric.bertolino@gmail.com (E. Bertolino), manu.manu@und.edu (Manu).
[1] Current address: Department of Biology, University of North Dakota, 10 Cornell Street, Stop 9019, Grand Forks, ND 58202-9019, USA.

http://dx.doi.org/10.1016/j.ydbio.2016.02.030
0012-1606/

of *cis* regulation have been inferred by the analysis of a few genes (Cantor and Orkin, 2002; Göttgens et al., 2002; Kim et al., 2013; Small et al., 1993; Wilson et al., 2011), checking their generality requires that we repeat such analyses on a much larger scale.

Transcriptional regulation is an input-output problem. The key to unscrambling *cis*-regulatory logic is to map inputs (TF concentrations) to output (rate of transcription), conditioned by regulatory sequence. A necessary requirement for successfully decoding regulatory logic therefore is to include all three: TF concentrations, DNA sequence, and transcriptional output. Mainstream genomic approaches, such as Chromatin Immunoprecipitation followed by Sequencing (ChIP-Seq), RNA-seq, and massively parallel reporter assays (Arnold et al., 2013; Levo and Segal, 2014; Melnikov et al., 2012; Nam et al., 2010; Sharon et al., 2012), assay either input or output but not both. This fact necessitates the development of the means to include all three components in *cis*-regulatory decoding.

More than mapping an input to an output, transcriptional regulation is a problem of mapping multiple inputs to a single output, since CRMs are regulated by multiple interacting TFs. For example, the CRM driving the expression of the second stripe of the *even-skipped* gene of *Drosophila* is bound by 7 TFs at about 20 binding sites (Arnosti et al., 1996b; Janssens et al., 2006; Small et al., 1992). The binding of CRMs by multiple TFs is widespread. Studies in multiple cellular contexts, including the hematopoietic system (Heinz et al., 2010; Wilson et al., 2010a), have detected combinatorial binding of lineage-specifying TFs. More generally, the ENCODE and modENCODE projects (ENCODE Project Consortium et al., 2012; Gerstein et al., 2012, 2010) have identified Highly Occupied Targets (HOTs)—DNA sequences occupied by multiple TFs—which occur at a frequency higher than one expected by chance (Nègre et al., 2011). TFs interact in complex manners to control the spatiotemporal program of gene expression. Many activators are known to promote gene expression synergistically, TFs can bind cooperatively, and repressors interfere with the activator function (Arnosti et al., 1996a; Cantor and Orkin, 2001, 2002; He et al., 2012a; Heinz et al., 2010; Kulkarni and Arnosti, 2005; Small et al., 1993, 1996). Multiple interacting inputs make large-scale *cis*-regulatory inference challenging since there is not a straightforward correspondence between TF binding and gene expression (Calero-Nieto et al., 2014). Addressing this complexity of *cis*-regulation requires that we devise a computational attack on the problem.

Here, we present a new approach for reverse engineering the *cis*-regulatory logic of a target gene. Our approach overcomes the challenge of regulatory complexity by integrating multiple datasets—evolutionarily conserved non-coding DNA sequence, genome-wide gene expression data, TF binding preferences, and reporter activity data—using a transcriptional model that explicitly simulates mechanisms of TF interaction. Our premise is that datasets assaying multiple aspects of gene regulation, in combination with the rules of gene regulatory interaction encapsulated in the model, will constrain the number of *cis*-regulatory schemes consistent with activity data. Our transcriptional model is of the so-called "thermodynamic" type. Thermodynamic models have been used to quantitatively predict CRM activity during development (He et al., 2012b; Janssens et al., 2006; Kazemian et al., 2010; Kim et al., 2013; Reinitz et al., 2003; Segal et al., 2008; Zinzen et al., 2006). In contrast to the previous applications of such models

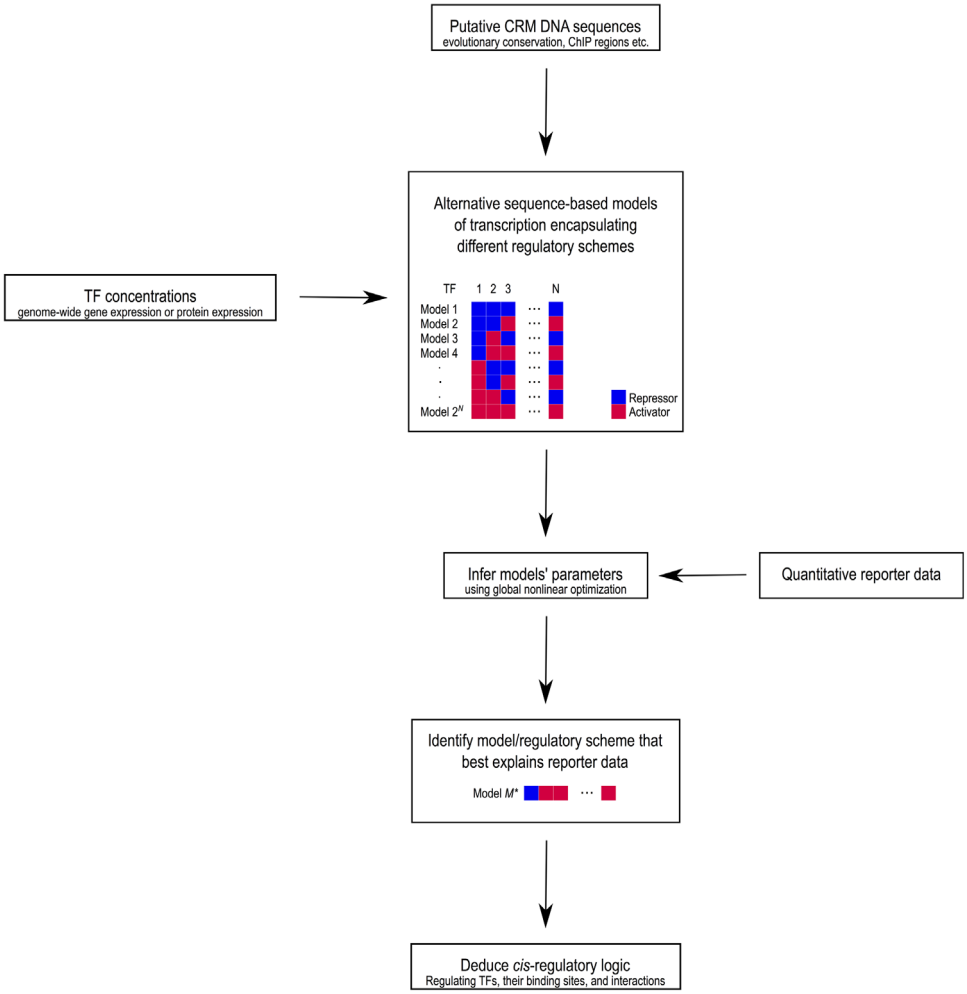

Fig. 1. A schematic illustration of the methodology for reverse engineering *cis*-regulatory logic.