Genomics 106 (2015) 151–158

Contents lists available at ScienceDirect

Genomics

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

High throughput technologies for the functional discovery of mammalian enhancers: New approaches for understanding transcriptional regulatory network dynamics

Lisa Dailey

NYU School of Medicine, Department of Microbiology, Kimmel Center for Stem Cell Biology, 550 First Avenue, MSB 252, New York, NY 10016, United States

ARTICLE INFO

Article history:
Received 1 March 2015
Accepted 8 June 2015
Available online 10 June 2015

Keywords:
Transcription
Chromatin
Enhancers
Transcription factors
Development
Genome

ABSTRACT

Completion of the human and mouse genomes has inspired new initiatives to obtain a global understanding of the functional regulatory networks governing gene expression. Enhancers are primary regulatory DNA elements determining precise spatio- and temporal gene expression patterns, but the observation that they can function at any distance from the gene(s) they regulate has made their genome-wide characterization challenging. Since traditional, single reporter approaches would be unable to accomplish this enormous task, high throughput technologies for mapping chromatin features associated with enhancers have emerged as an effective surrogate for enhancer discovery. However, the last few years have witnessed the development of several new innovative approaches that can effectively screen for and discover enhancers based on their functional activation of transcription using massively parallel reporter systems. In addition to their application for genome annotation, these new high throughput functional approaches open new and exciting avenues for modeling gene regulatory networks.

1. Introduction

The biological processes required for the development and survival of all organisms rely on mechanisms for precisely regulating the expression of genetic information. During embryogenesis the orchestrated generation of a multitude of specialized cell types from a single cell zygote depends on the establishment of regulated patterns of gene expression at each stage of development. Furthermore, whole genome sequencing has revealed the surprising observation that the trillions of cells of the adult human are specified by 21,000–25,000 genes, the same number of genes encoded within genome of mice or the simple nematode *Candida elegans* [1,2], suggesting that sophisticated and carefully regulated mechanisms controlling gene expression (rather than differences in gene number) are also likely to underlie differences in organismal complexity. Regulated gene expression is largely mediated by the concerted activities of DNA elements such as promoters, enhancers, silencers, and insulators, which together determine the level, timing, and cell-type specificity of gene transcription [3]. Of these, enhancers have been shown to be particularly instrumental for the temporally- and spatially-controlled activation of gene subsets in many developmental processes [4–7], and sequence variations at enhancers have been shown to cause or be implicated in human disease [8–10]. Thus identifying these regulatory DNAs within mammalian genomes, and how, where and when they are active, is fundamental for understanding embryonic development, cellular responses to environmental cues, and how sequence alterations within regulatory DNAs cause disease.

This review will summarize the initial discovery and basic features of enhancer function in mammalian cells and how these properties are being applied in current efforts to map enhancers genome-wide and to model transcriptional regulatory networks. The importance of integrating functional approaches into this effort will be discussed, and new high throughput technologies for the direct discovery of mammalian enhancers will be compared. I will also summarize how functionally-defined enhancers discovered using the technology FIREWACh can be used to identify key TFs of cell identity, and consider ways in which function-based high throughput approaches can be applied for refining chromatin modification interpretation, understanding the grammar of regulatory DNA elements, and dissecting the transcriptional networks and sub-networks driving cell fate decisions. For more detailed discussion of some of the approaches for genome-wide enhancer prediction and motif analyses that will be touched upon here, the reader is referred to [11–16].

2. Regulatory DNA elements directing transcription by RNA Polymerase II

Gene transcription by RNA Polymerase II (RNA PolII) relies on the concerted activities of two major classes of regulatory elements: promoters and enhancers. Promoters are classically considered as the

E-mail address: lisa.dailey@nyumc.org.

http://dx.doi.org/10.1016/j.ygeno.2015.06.004
0888-7543/

combined activities of a core promoter where transcription initiation takes place, and one or more adjacent *cis*-regulatory modules (CRMs) that bind sequence-specific TFs. The core promoter encompasses the transcription start site (TSS), and surrounding DNA sequences which together bind general transcription factors and facilitate RNA PolII complex binding [17–19]. Although they are generally considered as a generic entity, core promoters comprise structurally and functionally diverse subclasses that are defined by the combinations of sequence motifs they possess, such as the TATA box, TFIIB recognition element (BRE), initiator (Inr), and the downstream promoter element (DPE) [19,20]. These motifs lie within −40 to +40 nucleotides of the TSS, and are likely to have important roles in differentially regulated gene expression that have yet to be fully explored. In addition to core promoter sequences, "cis-regulatory modules" (CRMs) containing clusters of recognition sites for DNA sequence specific TFs are typically located within 200 bp or so of the TSS. Bound TFs either make direct contact with basal factors or recruit cofactors or mediator proteins to facilitate transcription initiation complex assembly and binding at the core promoter [19], and may also impact promoter clearance and elongation, or interactions with distal enhancers [21]. Thus promoter-proximal CRMs integrate activating or repressive signals that can contribute to the specificity of transcriptional initiation at the core promoter.

While regulatory sequences controlling most regulated gene transcription in bacteria and unicellular eukaryotes are primarily located near the TSS, a wealth of studies in vertebrates and *Drosophila* have underscored the dominant role played by distally-located enhancers in directing differentially regulated patterns of gene expression at promoters. Enhancers were first identified using functional assays where a 200 bp DNA segment derived from a region upstream of the early gene promoter within the SV40 DNA tumor virus genome was found to be capable of activating transcription of a linked chicken β-globin gene promoter several hundred-fold in transfected HeLa cells [22] (Fig. 1). Importantly, activation of reporter gene transcription occurred in an orientation- and distance-independent manner, and was observed even if the SV40 element was placed several kb upstream- or downstream of the promoter [23]. Accordingly, this activating entity was given the name 'enhancer' [22]. The first eukaryotic gene enhancer was subsequently discovered as a DNA segment from within an intron of the Ig heavy-chain gene that could activate transcription from the rabbit β-globin gene promoter or SV40 T Ag promoter even over distances of several kilobases [24]. However, unlike the SV40 enhancer, which functions in a wide variety of cell types, the Ig enhancer only activated the reporter gene in transfected B lymphocyte-derived cells, consistent with the cell-specific expression of the Ig gene. Functional studies further identified additional cellular enhancers, each of which were similarly found to activate transcription in a cell type- or developmental-specific manner [25].

3. Cell type-restricted enhancer activity is mediated by the combinatorial action of multiple TFs

These early functional studies supported the notion that enhancers play a fundamental role in directing complex and dynamic patterns of cell- or developmental-stage-specific transcription. Implicit in this observation is that only a subset of the many thousands of potential enhancers within a genome is active in a given cell type or context, and determining the basis for this selective enhancer activation is key to understanding how spatio-temporal gene expression patterns are established.

Enhancers generally span only a few hundred base pairs of DNA, and the functional units are one or more CRMs which contain clusters of binding sites for multiple transcription factors [26]. It was realized early on that the sequence-specific DNA binding properties of TFs to motifs within these enhancers could be used as a tool for identifying key TF mediators of cell-specific gene expression. An example of this is illustrated by studies of the Fgf4 enhancer that uncovered fundamental aspects of embryonic stem cell (ESC) biology. The murine fgf4 gene is expressed within the blastocyst inner cell mass in vivo and cultured ESCs or closely related F9 embryonal carcinoma (EC) cells, and depends on the activity of a 150 bp enhancer within the gene's 3′ UTR [27–30]. Functional analysis of enhancer sequence variants and in vitro binding assays showed that ESC-specific enhancer activity depends on the combinatorial activity of a cooperatively bound TF complex composed of OCT4 and SOX2, and a third TF that is most likely a KLF protein [28,31,32]. Importantly, these TFs act synergistically, and the enhancer is only active in cells that express all three TFs. Thus analysis of this single functionally characterized enhancer enabled the first report of SOX2 activity

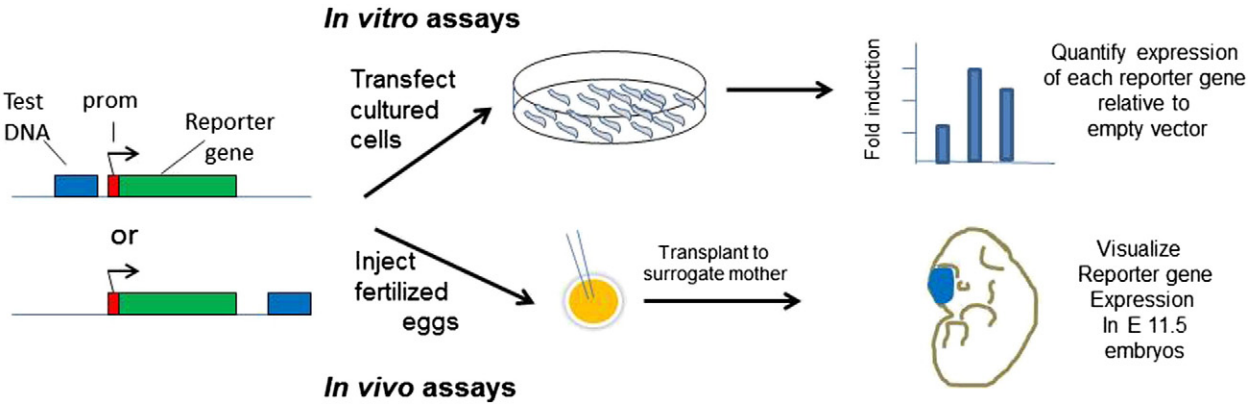

Fig. 1. Traditional reporter assays for enhancer discovery. Generally, a reporter plasmid would consist of a portion of a cellular- or viral promoter (i.e. minimal promoter, red box) that will provide sequences required for transcriptional initiation (arrow). The reporter gene (green box) encodes a gene product used for detection and quantification of transcriptional activation. These include fluorescence-based gene products such as GFP or YFP that permit direct microscopic visualization of expression, or enzymatic activities that require addition of a substrate to generate a measurable product that parallels transcriptional output from the reporter. Popular examples of this are luciferase or lacZ. Genomic regions to be tested for enhancer activity (blue box) are cloned anywhere upstream or downstream of the promoter-reporter gene cassette. *In vitro analysis*: the reporter plasmids lacking or containing the test enhancer DNA are independently introduced into cultured cells using transfection, and assayed 24–48 h later either by luciferase assays read by a luminometer (luciferase) or quantitative cytometry (GFP). The level of transcription detected in the absence of the test enhancer represents basal transcription and the activity of the plasmid containing the test enhancer is measured relative to the basal level to determine whether the test DNA harbors enhancer activity. *In vivo analysis*: Generally, plasmids used for these analyses harbor a lacZ reporter gene which, when stained with -galactosidase substrate, permits visualization of the tissues or cell types in which it is expressed. Each test plasmid is introduced into a fertilized mouse egg that is then transferred for implantation into a surrogate mother for development. Embryos or progeny are assayed for lacZ expression. The VISTA project is a large and ongoing study of this sort that originally sought to assess highly conserved genomic regions for enhancer function in 11.5 day embryos [81]. This analysis is now also applied for the validation of some of the regions identified as potential enhancers in high throughput studies summarized in Section 4, and thus far has identified 1154 mouse or human enhancers.