Developmental Biology 422 (2017) 92–104

Contents lists available at ScienceDirect

Developmental Biology

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

Single embryo-resolution quantitative analysis of reporters permits multiplex spatial *cis*-regulatory analysis

Catherine L. Guay[a,1], Sean T. McQuade[a,1], Jongmin Nam[a,b,*]

[a] Center for Computational & Integrative Biology, Rutgers, The State University of New Jersey, Camden, NJ 08102, USA
[b] Department of Biology, Rutgers, The State University of New Jersey, Camden, NJ 08102, USA

ARTICLE INFO

Keywords:
Cis-regulatory
Spatial
High-throughput
Mosaic
Embryo
Gene regulatory network
Sea urchin
Reporter assay

ABSTRACT

Cis-regulatory modules (CRMs) control spatiotemporal gene expression patterns in embryos. While measurement of quantitative CRM activities has become efficient, measurement of spatial CRM activities still relies on slow, one-by-one methods. To overcome this bottleneck, we have developed a high-throughput method that can simultaneously measure quantitative and spatial CRM activities. The new method builds profiles of quantitative CRM activities measured at single-embryo resolution in many mosaic embryos and uses these profiles as a 'fingerprint' of spatial patterns. We show in sea urchin embryos that the new method, Multiplex and Mosaic Observation of Spatial Information encoded in *Cis*-regulatory modules (MMOSAIC), can efficiently predict spatial activities of new CRMs and can detect spatial responses of CRMs to gene perturbations. We anticipate that MMOSAIC will facilitate systems-wide functional analyses of CRMs in embryos.

1. Introduction

Systems-level understanding of gene regulation requires detailed characterization of many regulatory genes, their target genes, and *cis*-regulatory modules (CRMs) that mediate the regulatory interactions of the former two (Davidson, 2006). Of the three key components of gene regulatory networks, CRMs possess two unique features that can improve the speed and accuracy of solving complex gene regulatory network problems. First, because CRMs contain binding sites for transcription factors, their sequences provide unique opportunities to predict and validate the transcription factors that regulate them. Second, because CRMs often control nearby genes in the genome their genomic location facilitates identification of their target genes. Thus, CRMs serve as critical information hubs for understanding how gene expression is controlled (Buecker and Wysocka, 2012). In addition, since development of the entire animal body from a single fertilized cell is the product of genetically encoded regulatory programs, detailed examination of many CRMs will also reveal the genetic mechanisms of animal development.

Cis-regulatory analysis measures temporal and spatial activities of CRMs in normal and perturbed embryos. The conventional approach for a *cis*-regulatory analysis is to ***i***) build a reporter construct that contains a wild-type or mutated CRM, a core promoter that can bind RNA polymerase II, a reporter gene such as green fluorescent protein (GFP), and a core poly-(A)denylation signal, ***ii***) deliver the reporter construct into cells or embryos, ***iii***) examine temporal and spatial expression of the reporter gene, and ***iv***) compare reporter expression with gene expression patterns to build a *cis*-regulatory model for gene expression control (e.g., Yuh et al., 1998). While this approach has been instrumental for our current understanding of *cis*-regulatory mechanisms, it has become increasingly difficult to keep up with recent progress in genomics for measuring genomic signatures and gene expression patterns. In addition, as many eukaryotic genes are controlled by multiple CRMs, unaccounted CRMs due to less than comprehensive analysis would lead to misguided models and make experimental data difficult to interpret. Therefore, comprehensive measurement of the activities of these individual CRMs in isolation is essential for modeling and validating how they function together in the genomic context. The most critical bottleneck of these methods is the limited number of fluorescent reporters to distinguish activities of many CRMs during imaging analysis.

To address these limitations, several high-throughput reporter assay methods for CRMs have been developed (Nam et al., 2010; Nam and Davidson, 2012; Melnikov et al., 2012; Patwardhan et al., 2012; Smith et al., 2013a; White et al., 2013; Arnold et al., 2013). These tools take advantage of the virtually unlimited diversity of DNA oligomers to barcode and track many CRMs in parallel. However, application of these methods has been limited to quantitative measure-

* Corresponding author at: Department of Biology, Rutgers, The State University of New Jersey, Camden, NJ 08102, USA.
E-mail address: jn322@camden.rutgers.edu (J. Nam).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ydbio.2017.01.010
Received 13 October 2016; Received in revised form 31 December 2016; Accepted 15 January 2017
Available online 16 January 2017

ment of CRM activities in cells or embryos. More recently, a method that combines barcoded reporters with fluorescent activated cell sorting (FACS) enabled simultaneous isolation of many CRMs that are active in the same cells as marker CRMs for two or three predefined cell types (Gisselbrecht et al., 2013). Although these new methods have considerably increased throughput, their application in multicellular systems such as embryos has been hampered due to lack of or limitation in spatial information.

Here, using sea urchin embryos, we develop a radically different and highly scalable method, Multiplex and Mosaic Observation of Spatial Information encoded in *Cis*-regulatory modules (MMOSAIC), to simultaneously measure both quantitative and spatial activities of CRMs. MMOSAIC is based on two well-known observations in a variety of model systems: ***i***) stochastic and mosaic incorporation of linear reporter constructs into only one cell in an early embryo and ***ii***) unequal clonal replication of the incorporated reporter constructs depending on cell lineages during embryogenesis (Flytzanis et al., 1985, 1987; McMahon et al., 1985). The level of reporter expression in a mosaic embryo is determined by the combination of intrinsic activity of a given CRM and cells that harbor the construct at the time of measurement. Since a large sample size neutralizes the effect of random mosaic DNA incorporation, we hypothesize that the quantitative profiles of single-embryo resolution reporter expressions measured in a sufficiently large number of embryos is determined by spatial activity of a given CRM. Using our new single-embryo resolution reporter assay method, we show that the quantitative profile of single-embryo resolution reporter expressions measured in a large number of mosaic embryos can be used for spatial *cis*-regulatory analysis without or minimally relying on imaging tools.

2. Materials and methods

2.1. Making of reporter constructs and extreme barcoding

The size of CRMs used in this study ranges from 351 bp to 2716 bp, and the majority of the CRMs are ~2 kb-long (Supporting File 1). Reporter constructs were generated in two steps following the procedure outlined in the Results section (Fig. 2**A** and **B**). The first step is to fuse individual CRMs with unique identifier (ID) sequences (hereafter called CRM::ID constructs). Two different strategies to build CRM::ID constructs were employed as the project progressed. To generate constructs shown in Figs. 5 and 7, we used pre-established reporter constructs from Nam and Davidson (Nam and Davidson, 2012). These constructs contain a modified version of sea urchin *nodal* basal promoter in which a functional SMAD site was removed (Nam et al., 2010). To generate a large number of CRM::ID constructs shown in Fig. 6, we used a pre-barcoded library of empty reporter vectors (freely available on request) that already contain a pan-bilaterian Super Core Promoter 1 (SCP1) (Juven-Gershon and Kadonaga, 2010), a GFP ORF, and ~100 million random barcodes (25 bp-long). The pool of pre-barcoded vectors was linearized by inverse PCR using primer pairs targeting 5′-upstream of SCP1 (primers Lin1 and Lin2 in Supporting File 2). Each of the 17 CRMs was amplified by PCR using primer pairs with ≥15 bp flanking sequences that overlap with the arms of linearized vectors. The names and sequences of CRM-specific primers are also available in Supporting File 2. Amplified individual CRMs were ligated to the vector by Gibson Assembly (Gibson, 2011) and were transformed into DH10B by electroporation. At a minimum of 2 colonies per CRM were selected for Sanger DNA sequencing to identify ID sequences. A total of 58 unique IDs for the 17 CRMs were identified. This approach using pre-barcoded vectors is our most up-to-date protocol to generate up to thousands of CRM::ID constructs.

The second step is to 'extreme' barcode the CRM::ID constructs as summarized in Fig. 2**B**: ***i***) one cycle of PCR adds a set of random 25-mer (N25) barcodes to 10 ng of CRM::ID constructs by using a CRM-specific forward primer and a common biotinylated reverse primer (primer SE_N25 in Supporting File 2) that contained N25 and a core-polyA signal (Nag et al., 2006); ***ii***) extreme barcoded PCR products were isolated by using strepatavidin conjugated magnetic beads (Life Technologies, Carlsbad, CA) followed by 15 cycles of PCR using CRM-specific forward primers and a universal reverse primer (primer EndCore-PolyA in Supporting File 2). PCR products were purified with Zymo DNA Clean & Concentrator kit (Zymo Research, Irvine, CA). We used Q5 High Fidelity PCR kit (New England Biolabs, Ipswich, MA) following manufacturer's instruction in all PCR in this study.

2.2. Delivery of barcoded reporter constructs, isolation of RNA/DNA, and preparation of sequencing libraries

Equimolar amounts of pooled extreme barcoded constructs were injected as described in Nam et al. (2010) with reduced injection volume. The constituents of a 10 μl injection solution are ≤7.5 ng of pooled reporter constructs, 1.2 μl of 1 M KCl, and ≤130 ng of randomly sheared sea urchin genomic DNA as carrier (Arnone et al., 2004). The intended number of unique barcodes delivered per CRM was ≥1500 in the entire pool of injected embryos.

For the *nodal* mRNA overexpression (MOE) experiment, in vitro transcribed *nodal* mRNA was added to the injection solution to the final concentration of 80 pg/μl as described in Nam and Davidson (Nam and Davidson, 2012).

Injected embryos grown at 15 °C were sampled at 18 h post fertilization (hpf) or at 24 hpf and total RNAs and genomic DNAs from these samples were extracted using AllPrep DNA/RNA Micro kit (Qiagen, Valencia, CA) following the protocol described in Nam et al. (2010). Total RNA was used for cDNA synthesis using High Capacity cDNA Reverse Transcription Kit following manufacturer's instruction (Thermo Fisher Scientific, Grand Island, NY) and 5 pmole of reporter-specific oligomer (primer SE_RT_oligo in Supporting File 2). Copy numbers of incorporated GFP template was measured by QPCR as described in (Nam et al., 2010; Revilla-i-Domingo et al., 2004). 1/40th of an ethanol precipitated cDNA pool was used for QPCR to check reverse transcription using Power SYBR Green Master Mix (Thermo Fisher Scientific, Grand Island, NY). The remainder of genomic DNAs and cDNAs were used to PCR amplify incorporated and expressed barcodes using a pair of universal primers (primers P3 and P5 in Supporting File 2). Gel isolated amplicons were used to build sequencing libraries for IonProton sequencing (Thermo Fisher Scientific, Grand Island, NY) using a pair of custom-designed primers (primers Ion-P_P3 and Ion-A_IonXpress_P5 in Supporting File 2). Sequencing libraries from different samples (e.g., expressed barcodes vs. incorporated barcodes) were barcoded with different IonXpress sequences provided by the manufacturer of the IonProton platform. We used Q5 High Fidelity PCR kit (New England Biolabs, Ipswich, MA) following manufacturer's instruction.

Equimolar amount of sequencing libraries for expressed barcodes and incorporated barcodes for each sample were pooled and sequenced in the IonProton platform following the manufacturer's instruction. The intended number of sequence reads per CRM were ≥100,000 to sufficiently cover expressed and incorporated barcodes.

2.3. Processing of sequence reads

The sequences of IDs and N25 barcodes from individual sequence reads were identified by trimming flanking adapter sequences: adapters AD_P5 and AD_Int for IDs; adapters AD_P3 and AD_Int for N25 barcodes. The sequences of adapters are provided as Supporting File 2. To trim adapter sequences we used the computer program Trimmomatic (Bolger et al., 2014) allowing up to two mismatches. Sequence reads that did not contain all three adapters were excluded in further analysis. The sequences of IDs were used to link N25 reads to CRMs and the sequences of N25 barcodes were further analyzed as follows.