HIGH-THROUGHPUT TOOLS FOR FUNCTIONAL GENOMICS

by

CATHERINE L. GUAY

A dissertation submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational and Integrative Biology

Written under the direction of

Dr. Jongmin Nam

And approved by

_____

Dr. Jongmin Nam

_____

Dr. Eric Klein

_____

Dr. Jean-Camille Birget

_____

Dr. Sudhir Kumar

_____

Dr. Grace Brannigan, Program Director

Camden, New Jersey

May 2018

ABSTRACT OF THE THESIS

High-throughput tools for functional genomics

By: CATHERINE L. GUAY

Dissertation Director:

Dr. Jongmin Nam

Genetic information is stored in DNA sequences, referred to as the genome. Decoding the meaning of these sequences is a critical challenge in biology. The genome contains functional components that modulate information in the genome into cell type specific functions. Of these components, *Cis*-regulatory modules (CRMs) integrate transcription factor inputs to causally affect gene expression. The genomic locations and sequences of CRMs are useful for resolving interactions within gene regulatory networks, and are therefore useful to address fundamental questions pertaining to development, evolution, cancer, and genetic disease.

My work addressed major methodological limitations in CRM analysis. Previously, there was no functional method to identify CRMs at genome-scale. Thus, I developed a Genome-scale Reporter Assay Method for CRMs, or GRAMc, that utilizes random 25bp DNA barcodes (N25s) as reporters to quantitatively and reproducibly identify CRMs across an entire genome. The method was applied in cultured human liver cells and in sea urchin embryos. Due to the limited delivery rate of reporter constructs into embryos, it is often advantageous to test GRAMc libraries containing long (>1kb) genomic inserts.

Previously, there was no efficient way to characterize genome-scale reporter libraries containing long inserts. To overcome this limitation, I developed a bi-directional mate-pair library approach to characterize long insert containing libraries of genome-scale magnitude. Although genome-scale identification of CRMs is critical to increase *cis*-regulatory analysis, in embryos it is also necessary to characterize the spatial activity of CRMs. Available methods to identify the spatial activity of CRMs rely on image analysis. Due to the limited number of optically distinct reporter genes, image-based methods are limited to testing only a few CRMs at a time. To address this challenge, I developed a method for Multiplex and Mosaic Observation of Spatial information encoded In CRMs, or MMOSAIC, that increases throughput of spatial *cis*-regulatory analysis in embryos by several orders of magnitude over traditional imaging based approaches. Together, these tools for *cis*-regulatory analysis overcome several major limitations for the study of functional genomics.

DEDICATION

To my beloved children, Joelle, Toby, and Sophie

AKNOWLEDGEMENTS

I am extremely thankful for my advisor and mentor, Jongmin Nam. It is difficult to describe in brief the time that he dedicated and the energy he invested in helping me to grow scientifically. Jongmin saw potential in me when I did not. His ideas about the breadth of information that can be gleaned from the regulatory genome excited me about the research and motivated me to persevere when challenges arose. Without his efforts and ideas, this work truly would not be possible.

I am thankful for the advice and suggestions from my committee members: Dr. Eric Klein, Dr. Jean-Camille Birget, and Dr. Sudhir Kumar. I thank Dr. Angelica Gonzalez and Dr. Nir Yakoby for comments on the work.

I am grateful to my “laboratory half-brother,” Sean McQuade for his contribution to MMOSAIC.

I am abundantly appreciative of Dr. Dibyendu Kumar of the Waksman Institute Genomics Core Facility for his generous donation of next generation sequencing reagents. Dr. Kumar’s suggestions were also enormously helpful for maximizing our sequencing library throughput for the Illumina platform, which was essential for completion of the work.

I am grateful to the sea urchin community, in particular Greg Wray, for recognizing the potential of the high-throughput spatial method. The sea urchin community has also been a source of personal encouragement as well as critical commentary on my presentations of the work. I especially found motivation and encouragement through discussions with Jia Song.

I am grateful for the support of my teaching mentor Mary Craig, who encouraged me to pursue my Doctorate. Her wisdom and concern has helped me to overcome difficult aspects of graduate life.

I am grateful to my pseudo-lab-mates Nicole, Rob, and Vikrant, currently and formerly of the Yakoby lab, for scientific advice and discussions.

Lastly, I am eternally grateful to my loved ones. I thank my parents, my brother David, and the rest of my family for their love, encouragement, and support. My children have lived this work with me for the past five years, enduring visits to the lab and evidence of my frustrations. I also appreciate the steadfast strength and support that Paul has brought to my life as I strove to complete this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY MATERIAL

CHAPTER 1: INTRODUCTION

1.1 *Cis*-Regulatory Modules

Cis-regulatory modules (CRMs) are essential functional components of the genome that modulate information in the genome into cell type specific functions. CRMs provide the genomic platform for transcription factor (TF) binding, which leads to the regulation of target genes (Maniatis et al., 1987). The interaction of the bound CRM complex with the general transcriptional machinery affects the stability of RNA polymerase II at the transcription start site (TSS) and subsequently transcription (Fig.1) (Hardison and Taylor, 2012; Suryamohan and Halfon, 2015). The accessibility of TF binding sites in CRMs in combination with the presence of necessary TFs leads to temporal and tissue specific gene expression (Hardison and Taylor, 2012; Maniatis et al., 1987).

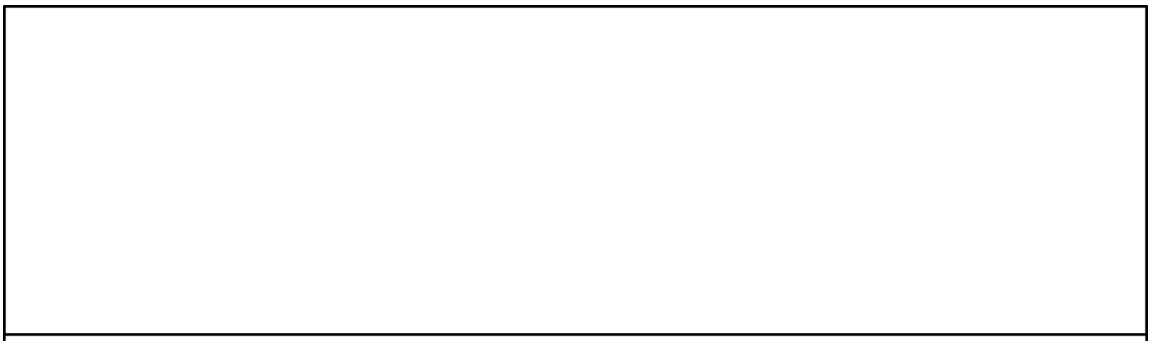


Figure 1: Cis-regulatory modules control gene expression. Cis-regulatory modules (CRMs) located throughout the genome provide a platform for transcription factor (TF) binding. The bound CRM complex affects the stability of the general transcription factors (GTFs) and RNA polymerase II (Pol II) at the transcription start site (TSS) of target genes, thereby controlling gene expression. Genes are regulated by multiple CRMs that are located either proximally or distally to the target gene. (reproduced from (Suryamohan and Halfon, 2015) with permission from the publisher)

For this reason, extensive efforts have been dedicated to studying CRMs to address fundamental questions pertaining to development, evolution, cancer, and genetic disease (Feigin et al., 2017; Ferrara et al., 2015; Lettice et al., 2002; Parker et al., 2014; reviewed in Sakabe et al., 2012). From such work, we know

that CRMs are modular, function combinatorially, respond to different transcription factor (TF) inputs, and can operate both proximally and distally with respect to their target genes (Davidson, 2006; Levine et al., 2014).

An important aspect of CRMs is their modularity, which enables them to function in combination to specify gene expression (Davidson, 2006). A typical gene may be embedded among tens of enhancers that can function in combination to precisely control the timing, location, and level of gene expression (Levine et al., 2014). This combinatorial complexity, however, confounds discovery of individual modules and ensuing elucidation of their modes of operation (Levine et al., 2014). An added challenge for CRM discovery is that while CRMs are often located near target genes, they have been shown to function as distally as 1 Megabase away from the target gene (Lettice et al., 2002; Shlyueva et al., 2014).

1.2 CRMs in disease

Gene expression changes due to variations in regulatory sequences are implicated in human disease (Mathelier et al., 2015; Pickrell, 2014). Approximately 0.1% of the human genome contains polymorphisms (The ENCODE Project, 2004), and >90% of implicated human disease associated variants occur within non-coding regions of the genome (Hindorff et al., 2009). Genome-wide Association Studies (GWAS) have identified single nucleotide polymorphisms (SNPs) that are enriched in non-coding regions and are associated with disease (The 1000 Genomes Project, 2010). Figure 2 illustrates one example of how polymorphisms in regulatory sequences can lead to disease

associated gene expression changes. Changes in CRMs can alter TF binding, which can result in a change in transcript abundance with potential phenotypic consequences. This example is similar to the hypothesized model of dysregulation of the SHH gene associated with preaxial polydactyly (PDD) (Wieczorek et al., 2010). Diseases such as α and β –thalessaemia, lung adenocarcinoma, T-ALL, and Burkitt lymphoma have all been associated with duplication, deletion, or hijacking of CRMs or alterations of TF binding sites in CRMs (Lower et al., 2009; Mansour et al., 2014; Taub et al., 1982; Van der Ploeg et al., 1980; Zhang et al., 2016).

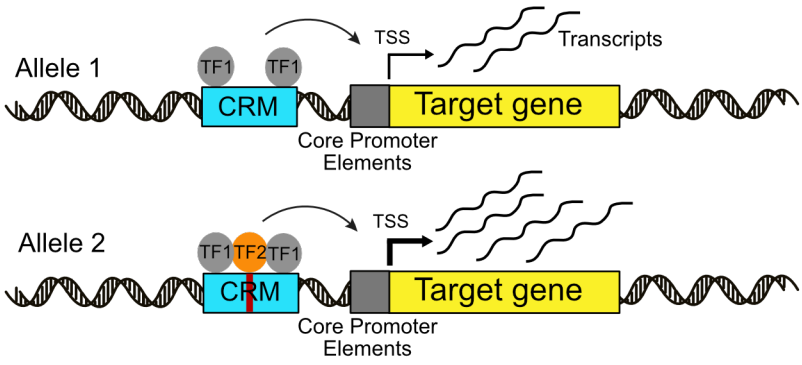


Figure 2: Sequence polymorphisms in CRMs and dysregulation of target genes. This illustration depicts a hypothetical example of how polymorphisms within regulatory sequences can alter transcription factor binding, resulting in changes in the level of transcript abundance that may lead to phenotypic changes associated with genetic diseases.

Causal defects in CRMs have been identified for a few diseases such as T_4 binding globulin deficiency, colorectal cancer, breast cancer, Gilbert's syndrome, and Schizophrenia (Ferrara et al., 2015; Fortini et al., 2014; Garritano et al., 2015; Iolascon et al., 1999; Roussos et al., 2014), but there are many gene expression-related defects for which the genetic cause remains unknown (Kwasnieski et al., 2014; Manolio et al., 2009; The 1000 Genomes Project, 2010). Nevertheless, in a few cases, therapies are being developed that correct

known causal variations in CRMs, mitigating the disease phenotype (Rincon et al., 2015). This underscores the long-term clinical potential of identifying and studying variations in CRMs as well as a need to identify regulatory regions of the genome to guide prioritization of functional testing for disease associated variations.

1.3 CRMs in development and evolution

Gene regulation through CRMs plays a fundamental role in development. Studies in comparative genomics suggest that the developmental complexity of an organism is not related to the total number of genes, but rather the manner in which these genes are expressed (Levine et al., 2014). Different gene expression programs lead to distinct cellular activities such as differentiation, proliferation, apoptosis, and motility, which determine cellular fate (Arthur, 2011), and thus differences in timing and level of gene expression during development can lead to phenotypic differences (Ludwig et al., 2011).

There is increasing evidence that sequence changes within CRMs provide the major means of evolution (reviewed in Wray, 2007). Several distinct CRMs that respond to different combinations of TF inputs can regulate the same target gene in different times and territories during embryonic development. Regulatory states can be decoupled, allowing evolution to act on each feature independently (Crocker and Stern, 2017). As a result, changes in individual CRMs result in gene expression changes during limited developmental windows and embryonic domains. It is this combinatorial complexity that enables the evolution of complex traits with reduced detriment to the organism (Davidson, 2006).

1.4 The utility of CRMs for resolving gene regulatory networks

CRMs provide a unique means to study gene regulation. Due to their pivotal role in conveying TF inputs into expression of target genes, CRMs are the crux of gene regulatory networks (GRNs). Identification of CRMs facilitates two-directional resolution of GRNs: their sequences enable motif enrichment analysis to identify potential TF inputs, and their locations may provide insights into potential effector gene outputs. Thus CRMs enable multifaceted analysis of regulatory information. It has been demonstrated that small reporter constructs harboring putative regulatory sequences are conducive to simultaneously testing CRM and endogenous gene responses to perturbation of effector genes, thereby validating gene regulatory interactions (Nam and Davidson, 2012). Despite the importance of CRMs for elucidating GRNs in the context of human cells or developmental model systems, the vast majority of these modules remain undiscovered. CRM predictions have been made for close to half a million regions of the human genome (The ENCODE Project, 2004, 2012). Yet to date, this number of potential CRMs far exceeds the number of validated CRMs (Suryamohan and Halfon, 2015; Visel et al., 2007; White, 2015).

1.5 Traditional methods of *cis*-regulatory analysis

The standard functional test for CRM activity is the reporter assay. To test a genomic sequence for CRM activity, the sequence is cloned upstream of a basal promoter, which allows for RNA polymerase II binding, and the open reading frame for a gene that produces a visually identifiable product. Some examples of reporters include proteins that fluoresce upon excitation at specific

wavelengths, such as Green or Red Fluorescent Protein (GFP or RFP), or enzymes for colorimetric determination of gene product abundance upon cleavage of a substrate, such as β-galactosidase and luciferase (reviewed in Arnone et al., 2004; Hall et al., 1983). The CRM::reporter construct is then introduced into the model system of interest by gene transfer using transfection, injection, or infection (McMahon et al., 1984; reviewed in Recillas-Targa, 2006). For traditional reporter assays, detection via imaging or spectrophotometry is required for analysis of reporter expression. The choice of reporter for a given application depends on a balance of detection sensitivity, quantitative range, and ease of use. Regardless of the application, the limited number of spectrally distinct visualizable reporters negates the potential for high-throughput multiplex CRM analysis.

1.6 Computational methods for *cis*-regulatory analysis

Computational alternatives to the traditional reporter assay were developed to increase CRMs discovery. These methods involve predictions based on interspecies comparisons, co-regulation analysis, analysis of conserved transcription factor binding site motifs, or correlation with DNase I sensitivity or histone marks to denote a genomic location a “CRM” (Madrigal and Krajewski, 2012; Pai et al., 2015; Pennacchio and Rubin, 2001; Shelton et al., 1997; Stranger et al., 2007; Thurman et al., 2012). These predictions, however, require validation through direct functional assays (Plank and Dean, 2014). Massively parallel reporter assays reveal that only 26% of tested ENCODE predicted enhancers function in their respective cell types, and predicted

repressive elements show no repressor function (Kwasnieski et al., 2014), indicating low accuracy and utility of such predictions. Authorities in the field of *cis*-regulatory analysis have noted that taken alone, predictions are inadequate, and large-scale functional testing in conjunction with genomic predictions is essential (Kellis et al., 2014). Arguably, since large-scale functional testing remains a requirement, a genome-scale functional assay for CRMs would alleviate the need for pre-requisite predictions.

1.7 High-throughput functional testing for CRMs

Within the past decade, several groups have developed high-throughput functional assays to identify and study CRMs *en masse* (summarized in Table 1). Fluorescence activated cell sorting-based methods (FACS-based) such as enhancer FACS-seq (eFs) and FIREWACh rely on isolation of CRM driven GFP positive cells followed by amplification of CRMs for next generation sequencing (NGS) (Dickel et al., 2014; Gisselbrecht et al., 2013; Murtha et al., 2014). These methods represent the lowest throughput of all of the high-throughput methods since they require stable integration and test only one CRM per cell.

Another widely used approach utilizes DNA barcodes as reporters (Akhtar et al., 2014; Arnold et al., 2013; Kheradpour et al., 2013; Kwasnieski et al., 2014; Melnikov et al., 2014; Nam et al., 2010; Nam and Davidson, 2012; Patwardhan et al., 2009; Savic et al., 2015a). Of these methods, Massively Parallel Reporter Assays (MPRAs) offer the greatest flexibility for large-scale variant testing within limited genomic regions (Ernst et al., 2016; Melnikov et al., 2012). The core of this approach is the use of small array synthesized CRM::reporter constructs

containing genomic test sequences of 145bp or less. Due to this size limitation, however, the MPRA approach and other approaches that rely on pre-designed barcode sets (Nam and Davidson, 2012) are not scalable to identify CRMs at genome-scale (Arnold et al., 2013).

Table 1: Comparison of high-throughput functional assays for CRMs

	GRAMc	MPRAs	STARR-seq	SuRE	FACS-based
Insert size/ Number of fragments in library	~800bp/ ~20M	≤145bp <3K	~1.25kb/ NR	~0.2-2kb/ ~270M	~154bp/ ~84K
Genome-wide	yes	No	yes	yes	No
Enables control of coverage	yes	no	no	no	no
Barcode:Insert ratio	10:1	13:1, 100:1	1:1	NR *	NA
Number of cfu/ ng of library transformed/ volume of competent cells	200M/ 30ng/ 50uL Electromax	27K or ~54K/ NR	NR/ **/ 625uL Electromax	~270M/ NR/ 80uL Clonecatcher	NR
Reproducibility/number of cells transfected	r = 0.947/ 200M HepG2	r = 0.69/ ~10M HepG2	NR/ 800M HeLa	NR/ 100M K562	NA
Limitations	NR	Limited to select regions; not genome-scale	Limited to enhancers; reported "considerable background"; no reported quantified reproducibility	Limited to promoters; highly variable insert size can skew library representation	Limited to select regions; 1 region tested/cell = limited throughput
Number of reporter libraries required to test every cell type in a given species	1	~400 libraries required for 1X coverage at current synthesis throughput ***	1	1	As many as the number of cell-types
References	In preparation	(Melnikov et al., 2012), (Kheradpour et al., 2013), (Patwardhan et al., 2009),	(Muerdter et al., 2018)	(van Arensbergen et al., 2017)	(Murtha et al., 2014), (Dickel et al., 2014; Gisselbrecht et al., 2013)

NR = not reported; NA = not applicable

* ~270M unique genomic fragments were reported for the library, but no description is provided for the total number of library constructs. Only ~26M barcodes reported for each test of the library.

** ~7.5ug of vector plus inserts was used for the final HD Fusion reaction, ethanol precipitated and then all was used to transform 625uL of Electromax competent cells.

*** Current throughput is up to 55K on Agilent platform.

To overcome this limitation, a method that uses self-transcribing active regulatory regions as reporters, or STARR-seq, was developed (Arnold et al., 2013). In this method, genomic inserts are cloned downstream of a basal promoter and active enhancers are self-transcribing. Enhancer activity is measured similarly to RNA-seq. Through this approach, the number of reporters is not limited to a predefined set because the enhancer is the reporter (Arnold et al., 2013). But to adapt this approach for use in the large human genome, the scale of assay and cost of sequencing is prohibitive. While the method has been applied to the mouse and human genomes at a reduced scale (Barakat TS, 2017; Vanhille et al., 2015) and the human genome at full scale (Muerdter et al., 2018), no data on reproducibility or evaluation of success criteria have been reported, suggesting that the method may have limited accuracy when adapted to the large human genome. This is not surprising in that each enhancer is tested with a single reporter and may further suffer the confounding effect of cryptic unstable transcripts (Berretta et al., 2008).

STARR-seq is also not suitable to resolve differences in CRM activity in a heterogeneous population of cells because each CRM is only detected using a single reporter. Thus, the observed activity reflects an average, not absolute, activity across the population of cells. By design, STARR-seq specifically detects enhancers, and is not sufficient to study other regulatory elements. Similarly, another method called Survey of Regulatory Elements (SuRE) is designed to only recover promoters (van Arensbergen et al., 2017). Details of these methods are included in Table 1.

1.8 The necessity of novel high-throughput tools for *cis*-regulatory analysis

Despite the many advances in high-throughput methods to identify and study CRMs, to date, there is still no truly quantitative functional assay that identifies both enhancers and promoters at genome-scale in large mammalian genomes. An ideal method would possess the highly quantitative capabilities of MPRAs that test each insert with multiple barcodes, achieve unbiased full genome-scale coverage as in STARR-seq, and identify both promoters and enhancers. To that end, a major focus of my work was to develop a Genome-scale Reporter Assay Method for CRMs, or GRAMc, which allows for precise control of genomic coverage in the reporter library, can accommodate larger genomic test fragments, and is truly quantitative and unbiased. Many high-throughput human genomic libraries contain regions of interest that are selected based on biochemical signals, requiring separate libraries for each cell type of interest. A GRAMc library, however, contains nearly all of potential CRMs, and therefore a single library can identify CRMs in any transfectable cell type. Results of the method are presented in Chapter 2, and a subsequent discussion is contained in Chapter 5.

The specific design of a genome-scale reporter library depends on the major biological question that it is intended to address. One of the most important parameters in this regard is the choice of genomic test fragment length. Libraries containing smaller genomic test fragments (~500-800bp) are conducive to finer resolution mapping of CRMs and motif enrichment analysis (Grant et al., 2011). Larger genomic test fragments may be desirable when considering orientation

specific CRM activity to distinguish enhancers and promoters. Since it has been shown that minimal enhancer activity may be affected by or coupled to nearby genomic sequences (Crocker and Stern, 2017; Telorac et al., 2016), it may also be advantageous to consider CRM activity in a larger context. While there are tools available for cloning larger genomic inserts up to 10kb (Gibson et al., 2009), there is currently no available method to efficiently characterize a genome-scale library that harbors test fragments larger than ~800bp. For this reason, another goal of my work was to develop a method to enable efficient characterization of large insert genome-scale libraries. Preliminary Results of the method are presented in Chapter 3 followed by a discussion of Future Directions in Chapter 5.

While multiplex reporter screening to identify CRM activity in a single cell-type or at a specific stage in development are attainable either by current methods or the newly established genome-scale approach, previously there were severe limits on multiplex spatial *cis*-regulatory analysis *in vivo.* A complete catalogue of CRMs, while abundantly useful, fails to address one of the most fundamental questions of regulation: how the activity of CRMs in different embryonic domains reflects different gene regulatory programs. Capitalizing upon a previously established multiplex method for cis-regulatory analysis *in vivo* (Nam and Davidson, 2012) and technologies developed for the GRAMc, another focus of my work was development of a radically different *in vivo* multiplex method that increases throughput of spatial *cis*-regulatory analysis by several orders of magnitude over traditional imaging based approaches. The method is

the Multiplex and Mosaic Observation of Spatial Information encoded In CRMs, or MMOSAIC. The published work is presented as Chapter 4, and future applications of the method are discussed in Chapter 5.

1.9 The HepG2 human hepatocyte model

Human cells that are grown in the laboratory provide the primary model to study human CRMs. While cultured cells are not considered a physiological model, they provide a snapshot of CRM activity in a single cell type. Transient assays of episomal reporter constructs are well-established in this model system (Gorman et al., 1982; Lee and Taichman, 1989). Reporter assays using cultured cells are also conducive to perturbation analysis. Culture conditions may be modified to simulate various physiological conditions. Cells can also be subjected to drug treatments as well as ectopic expression or knock-down of transcription factors. CRM sequences in reporter constructs can be altered and changes in reporter activity observed to identify causal regulatory sites.

The human liver carcinoma cell line, HepG2 is an established model of human hepatocytes. HepG2 cells have been used to study liver-like gene regulation and liver cancer (Costantini et al., 2013). Because the liver is the major site of drug metabolism, HepG2 is also widely used as a tool for toxicology through gene expression profiling (Gerets et al., 2012; Jeong et al., 2005) and screening of stably transfected toxicity responsive reporter lines. Importantly, there is an abundance of RNA-seq and global protein binding data available for this cell line. High-throughput reporter assays have been successfully applied to this cell line (Savic et al., 2015a; Savic et al., 2015b). For these reasons, HepG2

is the ideal cell-type to establish a new genome-scale functional assay for human cells.

1.10 Sea urchin: a model system for early embryogenesis

The sea urchin model system allows for rapid assay of thousands of synchronously developing, optically transparent embryos. Based on these characteristics, sea urchins have been established as an experimentally tractable model system to study early embryogenesis of Deuterostomes. Early studies on the molecular basis of development were performed using sea urchin (Flytzanis et al., 1987; Nocente-McGrath et al., 1989). Studies have demonstrated that core regulatory programs are highly conserved (Burke et al., 2014; Peter and Davidson, 2011) and that some human homologues of sea urchin CRMs not only function in sea urchin but also display similar spatial activities despite an evolutionary distance of >500MY (Royo et al., 2011). Due to the availability of a fully sequenced genome (Sodergren et al., 2006) and well established molecular tools to perturb effector genes (Lin and Su, 2016; Mao et al., 1996; reviewed in Materna, 2017), the sea urchin, *Strongylocentrotus purpuatus*, provides an ideal model system for formulating and testing hypotheses on GRNs (reviewed in Martik et al., 2016). Indeed, many effector and TF genes have been identified and studied in this system (Barsi et al., 2014; Howard-Ashby et al., 2006a, b; Howard-Ashby et al., 2006c; Tu et al., 2014). Thus, sea urchin is the ideal model system to establish both GRAMc *in vivo* as well as MMOSAIC.

CHAPTER 2: GRAMc to identify and study CRMs in cultured human cells

2.1 Rational design of a genome-scale functional test for CRMs

To address the need for a genome-scale method for *cis*-regulatory analysis, I developed a quantitative Genome-scale Reporter Assay Method for CRMs, or GRAMc, that utilizes random 25-base DNA barcodes as reporters. Several considerations were accounted for in the design of the method. First, because the GRAMc utilizes the traditional reporter construct configuration, it recovers both enhancers and promoters while avoiding the effects of cryptic unstable transcripts that may confound the reliability of STARR-seq (Berretta and Morillon, 2009). Second, GRAMc enables testing of larger inserts than MPRAs can accommodate, reducing the likelihood of CRM truncation. Third, the method utilizes a novel qPCR approach to precisely control genomic coverage during cloning of genomic fragments, and this coverage is maintained with a sufficiently large library. Fourth, the GRAMc protocol allows for precise control of barcode to insert ratio. Control of these last two parameters help to maximize reproducibility without unnecessary increase in library size and cost.

2.2 Overview of the GRAMc and construction of the human GRAMc library

I prepared a GRAMc library of ~200 million unique reporter constructs containing ~20 million unique ~800bp long fragments from the NG16408 human genome, designated Hs800_GRAMc. The NG16408 genome is of relatively normal karyotype and originates from an apparently healthy male. Genomic DNA originates from the whole blood compartment of a 12-year-old boy, rather than from Epstein-barr virus transformed lymphoblastoid cells or sets of unidentified

pooled donors. Thus, the NG16408 genome represents the most pristine source of publically available human genomic DNA.

To prepare genomic inserts for cloning into the GRAMc vector (Supp. 1), randomly fragmented, size-selected genomic DNA was repaired and ligated to a double adapter (Fig. 3a). The double adapter reduces the potential for insert concatenates or ligation of two of the same adapter to a single insert. The double adapter contains two bases of RNA for subsequent linearization using RNase HII.

Prior to cloning, coverage of the genomic inserts is optimized (Fig. 3b). Linearized adapter ligated inserts are serially diluted to obtain a desired genomic coverage. The coverage of each serial dilution is determined using qPCR to screen for the presence or absence of 11 randomly selected single copy genomic targets. For a dilution that contains ~4 M randomly sampled genomic DNA fragments of 800bp-in length (1x genomic coverage), the expected proportion of targets present is 0.6. Multiple replicates are amplified from the dilution that represents 1X coverage, and these replicates mixed to obtain the desired coverage. In the case of the Hs800_GRAMc library, 5 replicates of 1X coverage were prepared and mixed to obtain a library of 5X coverage, since a library of this depth is expected to represent close to 99% of the genome.

Inserts of the desired coverage are cloned using a 3-piece Gibson Assembly (Gibson et al., 2009) into the GRAMc vector upstream of a synthetic minimal super-core promoter, SCP1 (Juven-Gershon et al., 2006) and a GFP open reading frame (ORF) (Fig. 3c). SCP1 has been validated in both human

cells and sea urchin embryos (Guay et al., 2017). Although barcodes are used as reporters, the GFP ORF is included for subsequent qPCR-based quality control and downstream library applications. It was experimentally determined that pre-barcode amplification from a linearly assembled library resulted in less loss of genomic coverage compared to amplification from a circular library. This result was consistent with expectation (Laghi et al., 2004). Following assembly, the library is amplified and random 25-base DNA barcodes (N25) are added using a single cycle of PCR with a biotinylated primer. Barcoded constructs are isolated using streptavidin beads, PCR amplified, self-ligated and exonuclease treated for transformation into bacteria. Serial dilutions of a sample of the transformants are plated for colony estimation of the size of the library. The bacterially grown and prepared reporter library is characterized and then assayed in the desired system.

To characterize the Hs800_GRAMc library, I prepared Illumina paired-end sequencing libraries (Fig. 3d). Sequencing libraries were generated by removing the intervening SCP-GFP or vector backbone sequences from between N25 barcodes and their respective genomic inserts. Barcodes from paired-end reads were associated with inserts that were mapped onto the hg38 reference genome. Mapped fragments covered ~80% of the genome, and this coverage is expected to increase as the reference genome is updated.

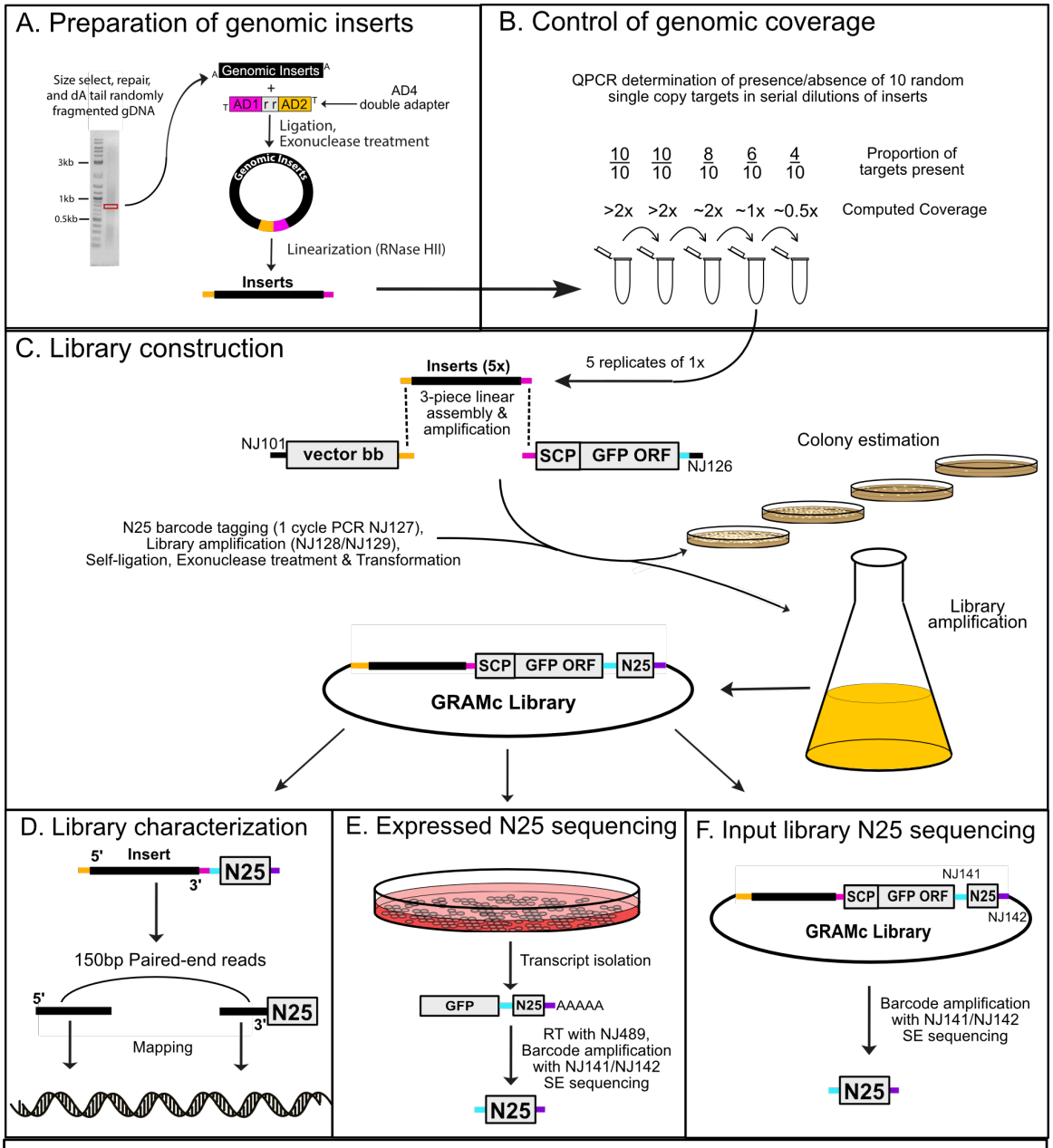


Figure 3: Overview of the GRAMc. A. Randomly sheered and size selected genomic inserts are TA ligated with the AD4 double adapter, containing sequences for insert amplification and two bases of RNA. The circularized ligates are linearized with RNAse HII. **B**. The coverage of serially diluted adapter ligated inserts is determined by comparison of the observed proportion of random genomic targets present per sample by qPCR with an expectation based on a Poisson probability model. A proportion of ~0.6 targets present represents 1X genomic coverage. **C**. A 5X mix of amplified inserts are cloned into the GRAMc vector in a 3-piece Gibson Assembly, followed by amplification and the addition of random 25-base (N25-mer) DNA barcodes via a single cycle of PCR. The barcoded library is amplified, self-ligated, and transformed into bacteria. A minute fraction of transformants are serially diluted and plated for colony counting and estimation of library size, while the remaining transformants are cultured for amplification of the library. **D**.For library characterization, intervening sequences are removed from the plasmid library such that inserts and N25 barcodes are separate only by short adapter sequences. Libraries containing inserts up to ~800bp in length are sequenced

Figure 3: cont. by 150bp Illumina paired-end sequencing. **E.** The GRAMc plasmid library is transfected into cultured cells. Cells are harvested after 24h of transfection and transcripts are isolated and expressed N25 barcodes amplicons are prepared for Illumina single-end sequencing. **F.** N25 barcode amplicons are prepared from the GRAMc plasmid input library for Illumina single-end sequencing. Expressed N25 barcodes are normalized to input library N25 barcodes to account for variations in barcode representation in the library.

2.3 GRAMc reproducibly identifies CRM activity in HepG2 cells

I tested the Hs800_GRAMc library by transient transfection in 2 batches of ~200 million HepG2 cells. Illumina NextSeq single-end sequencing of out-of-phase sequencing libraries was used to obtain ~250M reads/batch of expressed N25 barcodes and ~500M reads for input library N25 barcodes. We normalized read counts for respective N25s from cDNA batches to readcounts for the input library (Fig4). Activity is defined by the fold-change of normalized barcode expression over the average of the middle 30% of normalized barcode expressions for each batch. We included fragments as active if they were represented by greater than 10 reads in the input library and had a normalized barcode expression of ≥5-fold over background. We mapped activities to the hg38 reference genome. Using this stringent activity criteria, we identified ~45,000 unique active genomic fragments, referred to as CRMs or the set “G5”. Another ~110,000 fragments consistently displayed activity of ≥3-fold over background, and are collectively referred to as the set “G3”. Fragments that had activity of ≤1-fold of background are categorized as the set “L1.” This set represent a large set of functionally validated inactive fragments that are useful control for subsequent analysis.

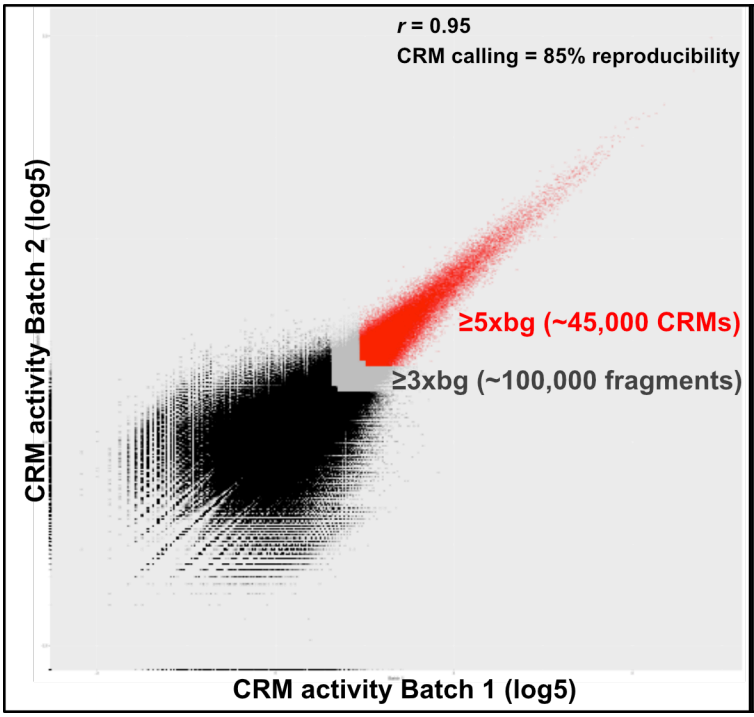


Figure 4: GRAMc reproducibly identifies CRMs in HepG2 cells. The Hs800_GRAMc library was tested in 2 independent batches of ~200M HepG2 cells. Background (BG) is defined as the average of the middle 30% of input nomalized N25 barcode expression. We denote a genomic fragment a CRM if it consistently displayed activity of ≥5-fold of BG in both replicates. We identified ~45,000 CRMs. Another ~100,000 genomic fragments consistently displayed activity of ≥3-5-fold of BG.

We evaluated the reproducibility of the replicates by computing the Pearson correlation coefficient (0.947) between the two batches (Fig. 4). To independently confirm consistency of the data, I cloned and tested a set of GRAMc tested fragments using an individual reporter assay (Fig. 5). I cloned 11 fragments from G5, 5 fragments from G3, and 5 fragments from L1 individually into the GRAMc vector upstream of SCP and independent of their original GRAMc library N25 barcodes. Quadruplicates of individual reporter constructs were co-transfected in HepG2 with an enhancer-less SCP-EGFP as an internal control, and assayed for CRM activity using QPCR. Reporter expressions for test constructs (GFP) and the basal control (EGFP) were normalized to their respective plasmid copies for each sample, and the fold-change of reporter expression over the average expression of the set L1 was determined for each fragment. Activities determined in the individual reporter assay corresponded

with GRAMc defined activity (R^2 = 0.83), suggesting that the majority of the reporter activity in each assay can be explained by other assay. Taken together with the high correlation between batches of GRAMc we conclude that the GRAMc reliably identifies CRM activity.

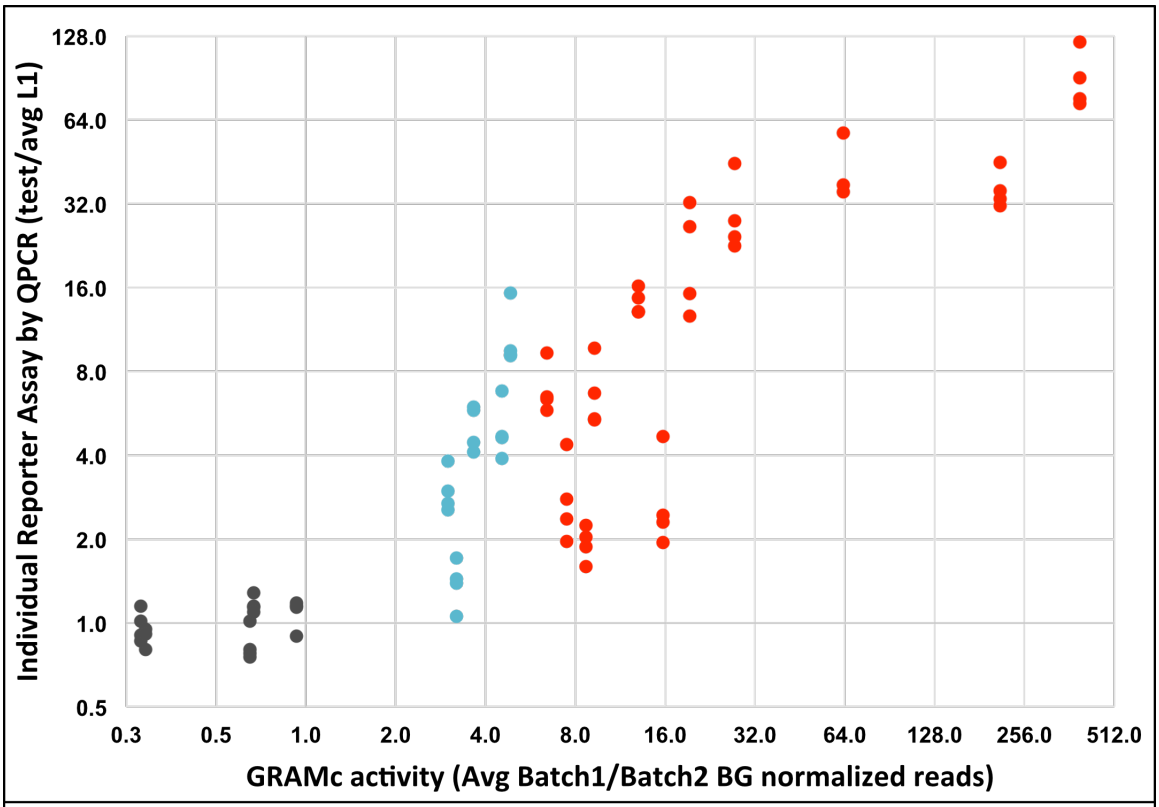


Figure 5: GRAMc identified CRMs are confirmed in an independent reporter assay. Select fragments from the GRAMc library (G5 in red, G3 in cyan, and L1 in gray) were tested in an independent individual reporter assay. Distinct assay platforms reported similar activities for the fragments tested (R^2 = 0.83).

2.4 GRAMc identified active regions display characteristic features of CRMs

We considered several known features of CRMs to confirm that GRAMc identified active regions are consistent with what is expected for CRMs. First, CRMs are located throughout the genome, but are enriched near expressed genes (Davidson, 2006; Levine et al., 2014; Shlyueva et al., 2014). To determine

if GRAMc identified active regions met this expectation for CRMs, we considered the genomic distribution of active fragments compared to expressed genes (Fig. 6). While N25 sequencing reads for the input library covered the genome with a relatively even distribution, reads for expressed N25s displayed a similar distribution to that of expressed genes, as shown for chromosome 1.

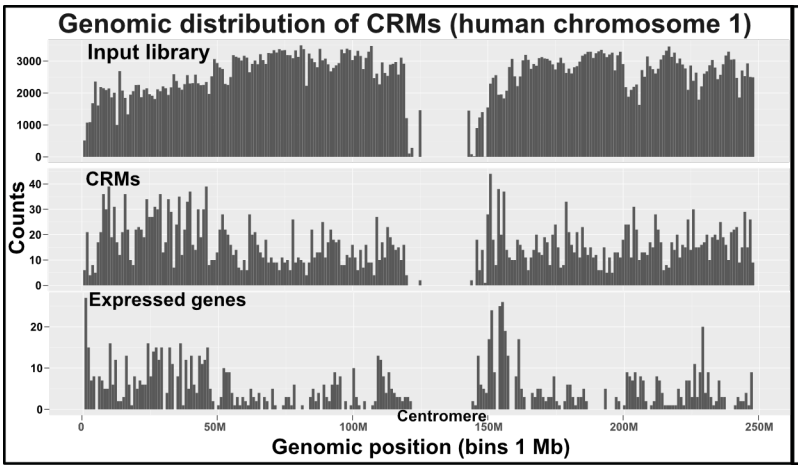


Figure 6: Distribution of GRAMc HepG2-CRMs across chromosome 1. The distribution of CRMs as defined by GRAMc (≥5-fold over background) differs from the distribution of genomic fragments tested in the input library. The distribution roughly overlaps with the distribution of expressed genes reported in (Mortazavi et al., 2013)

Second, CRMs can be located anywhere in the genome but are more abundant in the 5' proximal region of genes (Davidson, 2006; Levine et al., 2014; Shlyueva et al., 2014). We considered the enrichment of active fragments surrounding genes by computing the proportion of active fragments within 2kb bins up to 100kb upstream of the start of genes and up to 100kb downstream of the end of genes (Fig. 7). We observed that the proportion of active fragments surrounding genes is higher than the average proportion of active fragments across the genome. Furthermore, the proportion of active fragments is highest within the 2kb window immediately 5' proximal to genes, in keeping with our expectation for CRMs. Though CRMs are enriched in the 5' proximal region of genes as in expected for promoters and proximal enhancers, the method also

detects CRMs in other regions of the genomic, including intergenic regions and introns as expected for unbiased identification of enhancers.

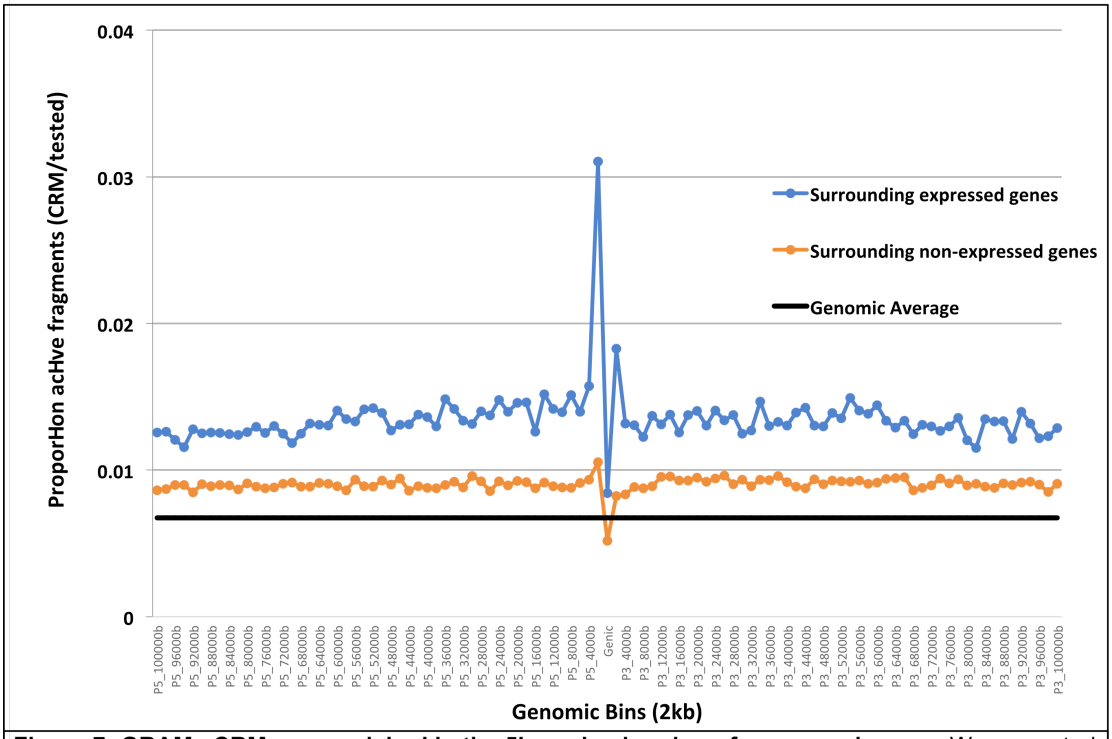


Figure 7: GRAMc CRMs are enriched in the 5' proximal region of expressed genes. We computed the proportion of active GRAMc test fragments within 2kb ins in the 100kb region surrounding genes. CRMs were enriched in the 5' proximal region of genes, and this enrichment was more pronounced for expressed genes versus non-expressed genes (FPKM ≥1) (Mortazavi et al., 2013).

We further compared the enrichment of active fragments surrounding genes with respect to expressed versus non-expressed genes (Mortazavi et al., 2013). We observed that the proportion of active fragments within the 2kb region upstream of expressed genes is almost 3 times higher (~4.6 times the genomic average) than the proportion of active fragments within the 2kb region upstream of unexpressed genes (~1.56 times the genomic average)(Fig. 7).

Third, It is estimated that more than two third of the human genome is comprised of repetitive transposable elements (TEs) (de Koning et al., 2011). It is believed that certain repetitive elements significantly contribute to the function

and evolution of the regulatory genome (Liang et al., 2015; Lynch et al., 2015). One type of TE, SINE/Alu, have been described as a means evolution in the primate lineage and a source of human diversity (Carroll et al., 2001). In particular, it has been shown that SINE/Alu overlap with tissue specific H3KMe1 enhancer marks and have been functionally validated to act as enhancers in large-scale reporter assays (Su et al., 2014). Similarly, another class of TEs has been shown to function as promoters (Bannert and Kurth, 2004; Cohen et al., 2009; Dunn et al., 2006). When our unbiased CRM data was cross-referenced with annotated repetitive elements (repeatmasker.org), the high-copy families, SINE/Alu and LTR/ERV1, were enriched 2.9-fold and 2.6-fold, respectively, in CRMs compared to random expectations, respectively. These repeat elements were among the most enriched in our dataset, aligning with the previous observation that GRAMc CRMs represent both promoters and enhancers.

Fourth, CRMs are the major component that integrates transcription factor inputs for the regulation of genes, and thus motifs for relevant regulatory inputs are enriched within CRMs (Davidson, 2006). We identified 120 out of 601 transcription factor binding site motifs from the HOCOMOCOv10 database that were enriched ≥3-fold in GRAMc identified HepG2-CRMs (Fig. 8). Taken together, overlap of GRAMc CRMs with these established features of CRMs bolsters confidence in their ability to recover regulatory information.

2.5 Motif enrichment in GRAMc HepG2-CRMs reveals novel regulatory interactions

Activators of CRMs can be predicted by computationally mining enriched putative transcription factor binding site (TFBS) motifs (e.g., Enuameh et al., 2013; Halfon et al., 2011; Mariani et al., 2017; Markstein et al., 2004; Xie et al., 2005). Traditionally, enrichment is computed against the null expectation that random or shuffled sequences of CRMs should contain no meaningful regulatory information. GRAMc, however, provides not only a large set of functionally validated CRMs, but also an even larger set of functionally validated inactive fragments. This copious regulatory information enables a more rigorous approach to calculate motif enrichment.

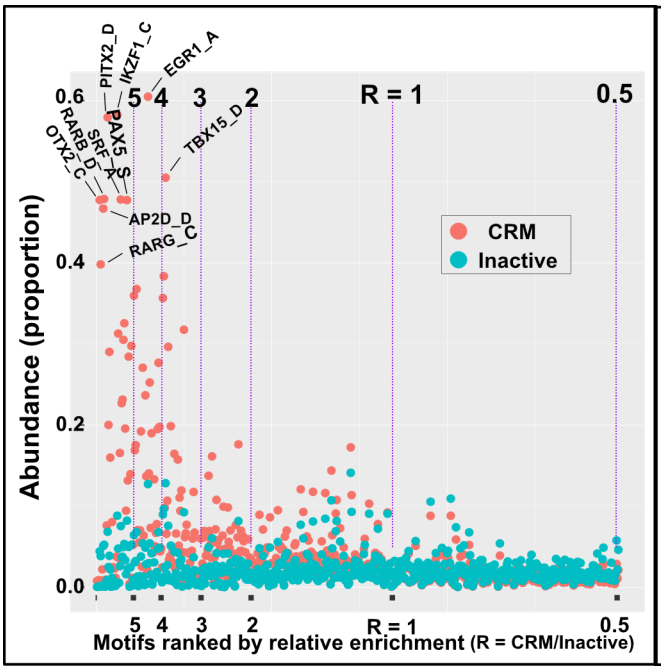


Figure 8: TFBS motifs are enriched in GRAMc CRMs. Using FIMO (*E*-value cutoff of 10E-5), we comuputed and plotted the enrichment of putative TFBS motifs for 601 human TFs in GRAMc CRMs (G5, active) relative to the enrichment of the motifs in GRAMc functionally validated inactive fragments (L1). Of the 601 motifs, 120 were enriched in CRMs. We further considered the abundance of each motif in CRMs by computing the proportion of fragments containing each motif. We identified a set of 12 TF for which motifs are both highly enriched (≥3-fold relative enrichment) and highly abundant (p ≥ 0.4) in GRAMc identified HepG2-CRMs.

To demonstrate the utility of our experimental data set, we further examined 12 TFBS motifs that are abundant (≥40% of CRMs) and enriched (≥3-fold vs. inactive fragments) in GRAMc-discovered CRMs (Fig.8). We first asked whether cognate transcription factors for the 12 motifs are expressed in HepG2 cells. Interestingly, only 5 transcription factors showed detectable expression

(FPKM ≥ 1) (Mortazavi et al., 2013). Consistent with their abundance, these transcription factors are global transcriptional regulators. To check whether their potential target CRMs are shared among expressed or non-expressed transcription factors, we clustered the 12 motifs based on motif co-occurrences within CRMs. Three distinct clusters (marked with red dots in Fig. 9) were detected by UPGMA clustering (Nei and Kumar, 2000). Interestingly, one cluster was exclusively shared among motifs for non-expressed transcription factors. This unexpected observation led to an interesting question of why these motifs are highly enriched in HepG2-CRMs. Because these transcription factors are expressed in other cell types, we hypothesized two alternative models for the role of the non-expressed transcription factors of enriched motifs: work as positive regulators or negative regulators of HepG2-CRMs in other cell types.

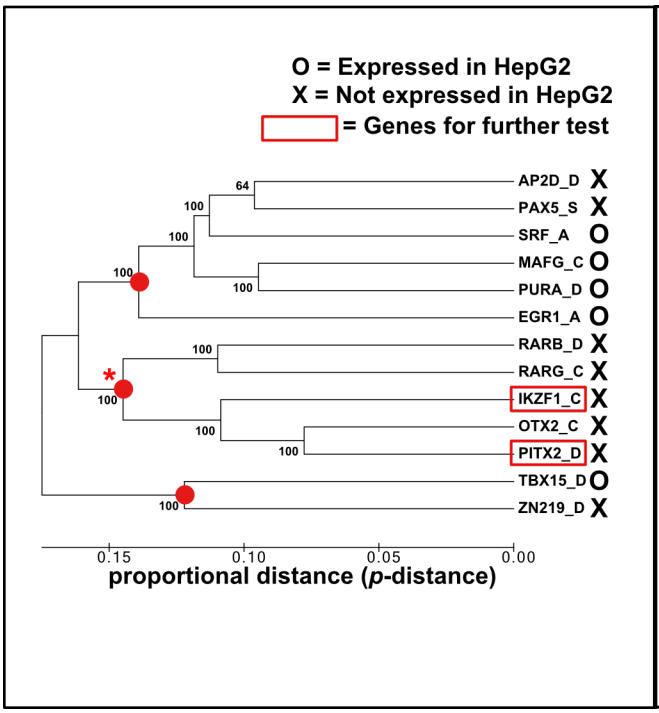


Figure 9: TFs share potential GRAMc CRM targets. UPGMA visualization of enriched and abundant motifs shows three major clusters of TFs that share potential CRM targets. TFBSs were clustered based on the proportion of shared CRMs. Proportional distance (*p*-distance) between two TFBSs is: 1 - proportion of shared CRMs. The three distinct clusters are marked with red dots. Several non-expressed TFs share many putative target CRMs. One cluster is shared exclusively by non-expressed TFs (marked by *).

2.6 High-throughput perturbation analysis distinguishes regulatory interactions

To test our hypotheses, we examined the effect of ectopic expression of pitx2 (a homeobox gene) and ikzf1 (an ikaros homolog) on CRMs in HepG2 cells. We co-transfected plasmids that can constitutively express mRNAs of *pitx2* (CMV::pitx2) or *ikzf1* (CMV::ikzf1) with a set of randomly selected ~80,000 GRAMc reporter constructs from the full GRAMc library. As a control experiment, we co-transfected plasmids that can constitutively express GFP mRNAs (CMV::gfp) with the same set of reporter constructs. We observed that ectopic expression of *pitx2* in HepG2 down-regulated the majority of CRMs by ≥2-fold and this down-regulation was more pronounced in Pitx2 motif-positive CRMs (Fig.10A). In the case of the ectopic expression of *ikzf1*, only 9 CRMs were down-regulated by ≥2-fold and 6 of the 9 down-regulated CRMs were positive for the IKZF1 motif (Fig.10B).

While this experiment does not provide in depth regulatory network analysis, we demonstrate an approach to rapidly screen putative regulatory inputs uncovered through the analysis of CRMs that were identified at genome-scale. Further discussion of the implications of the work are included in Chapter 5.

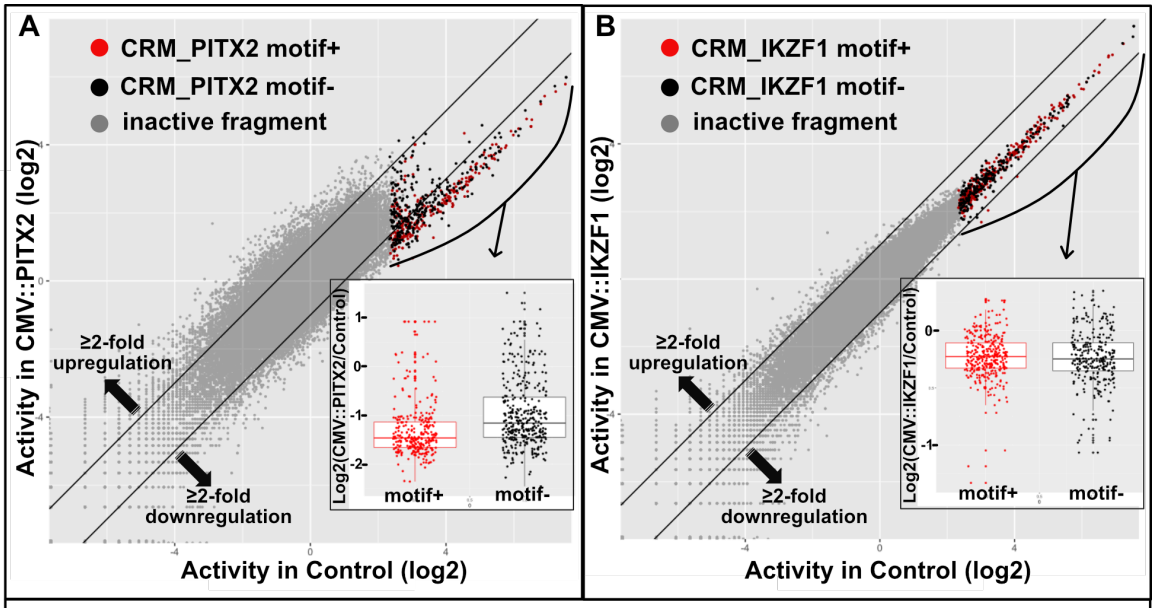


Figure 10: High-throughput perturbation analysis distinguishes potential regulatory interactions. Plasmids that constitutively express *pitx2* or *ikzf1* cDNA were co-transfected with a random subset of ~80,000 reporters from the Hs800_GRAMc library. As a control, we co-transfected a plasmid that constitutively expresses GFP cDNA with the library. Activities are shown in Log2 scale. CRMs were grouped into motif-positive (+, red) or motif-negative (-, black) groups. **A.** Ectopic expression of *pitx2* down-regulated the majority HepG2-CRMs (≥2-fold), and the effect was more pronounced for CRMs containing Pitx2 motifs, **B.** whereas ectopic expression of *ikzf1* led to down-regulation (≥2-fold) of only 9 HepG2-CRMs, 6 of which contained IKZF1 motifs. The distributions of Log2 ratio of CRM activities between perturbed and control cells are shown in box plots with superimposing frequency-jittered stripcharts.

CHAPTER 3: GRAMc in sea urchin, *S. Purpuratus*: overcoming the challenges of long insert genome-scale libraries

3.1 Previous CRM analysis in *S. Purpuratus*

How different CRMs contribute to precisely controlled gene expression within developing embryos is a fundamental question in biology. Previous work in *S. purpuratus* demonstrated that high-throughput multiplex testing of reporter expressions in combination with endogenous gene responses to perturbations *in vivo* can resolve CRM responses to different regulatory inputs (Nam and Davidson, 2012). This work revealed the potential advantage of high-throughput *in vivo* CRM analysis for interrogating gene regulatory logic. The application of GRAMc in *S. purpuratus* has the potential to uncover tens of thousands of new CRMs, enabling extension of this previous work. To that end, several GRAMc libraries were constructed from the genome of *S. purpuratus* (Table 3).

3.2 Construction of sea urchin dual orientation GRAMc libraries

Previously, the majority of temporally characterized *S. purpuratus* CRMs were ~3kb in length (Nam and Davidson, 2012). For comparative purposes, I generated 2 GRAMc libraries from the *S. purpuratus* genome with inserts of ~3kb in length. To enable distinction of promoters from enhancers, each library contains the same set of inserts cloned in opposite orientations: the forward orientation library, Sp3KF_GRAMc and the reverse orientation library, Sp3KR_GRAMc (Table 2 and see Supp.1 for vector information).

3.3 Characterization challenges for large insert genome-scale libraries

Characterization of millions of long inserts presents a major challenge. To date, the most cost effective high-throughput platform for characterizing GRAMc libraries is Illumina NextSeq, but this platform can only accommodate paired-end sequencing for inserts up to ~800bp in length. For this reason, Illumina paired-end sequencing is not feasible for sequencing genome-scale libraries such as the Sp3K_GRAMc libraries. Although the PacBio RS II platform can accommodate sequencing of longer inserts, this platform lacks the throughput required to fully characterize a GRAMc library. Therefore, development of an alternative approach was necessary to characterize the Sp3K_GRAMc libraries.

3.4 Method development: bi-directional mate-pair library characterization of long insert-GRAMc libraries

To overcome sequencing limitations for the characterization of long insert GRAMc libraries, I developed a new method that utilizes bi-directional mate-pair libraries: one mate-pair library in which the 5' ends of the genomic inserts and adapter sequences are abutted or "mated" to their respective barcodes, and one mate-pair library in which the 3' ends of the genomic inserts and adapter sequences are abutted or "mated" to their respective barcodes (Fig. 11). The 5' and 3' ends of genomic inserts are mapped to the reference genome (echinobase.org, spur4.2) using the Burrows-Wheeler Alignment (BWA) program (Li and Durbin, 2009). The 5' and 3' insert ends are combined using their shared N25 barcode as a key. Mapping is guided by a size restriction that reflects the target insert size. It is important to note that the same insert is represented in reads from multiple barcodes. For some barcodes, only one end of the genomic

insert may be identified in sequencing reads. For these, the other end of the insert can be identified based on clustering and mapping of the same inserts paired with other barcodes. This approach reduces the depth of bi-directional mate-pair sequencing required to map genomic inserts. Using this approach, the short reads required to characterize each library can be obtained on any high-throughput platform**.**

A full protocol for generating bi-directional mate-pair libraries is presented under Materials and Methods. In brief, the 5' ends of genomic inserts and their adjacent adapter sequences were mated to N25-barcodes and by removal of the vector backbone using inverse PCR ("5N25" sequencing library). Similarly, 3' ends of genomic inserts and their adjacent adapter sequences were mated to N25-barcodes by removal of the GFP ORF using inverse PCR ("3N25" sequencing library). A second round of PCR amplification yielded pairs of sequencing libraries, each containing only genomic inserts, intervening adapter sequences, N25 barcodes, and primer sites. The adapter sequences were later used to locate N25-barcodes and genomic insert ends within sequencing reads. During this second amplification, a molecule of biotin was introduced at the N25-barcode end of the sequencing libraries. Mate-pair libraries were sonicated and size selected to be ~125bp in length, resulting in ~50bp of genomic insert ends for subsequent mapping with BWA.

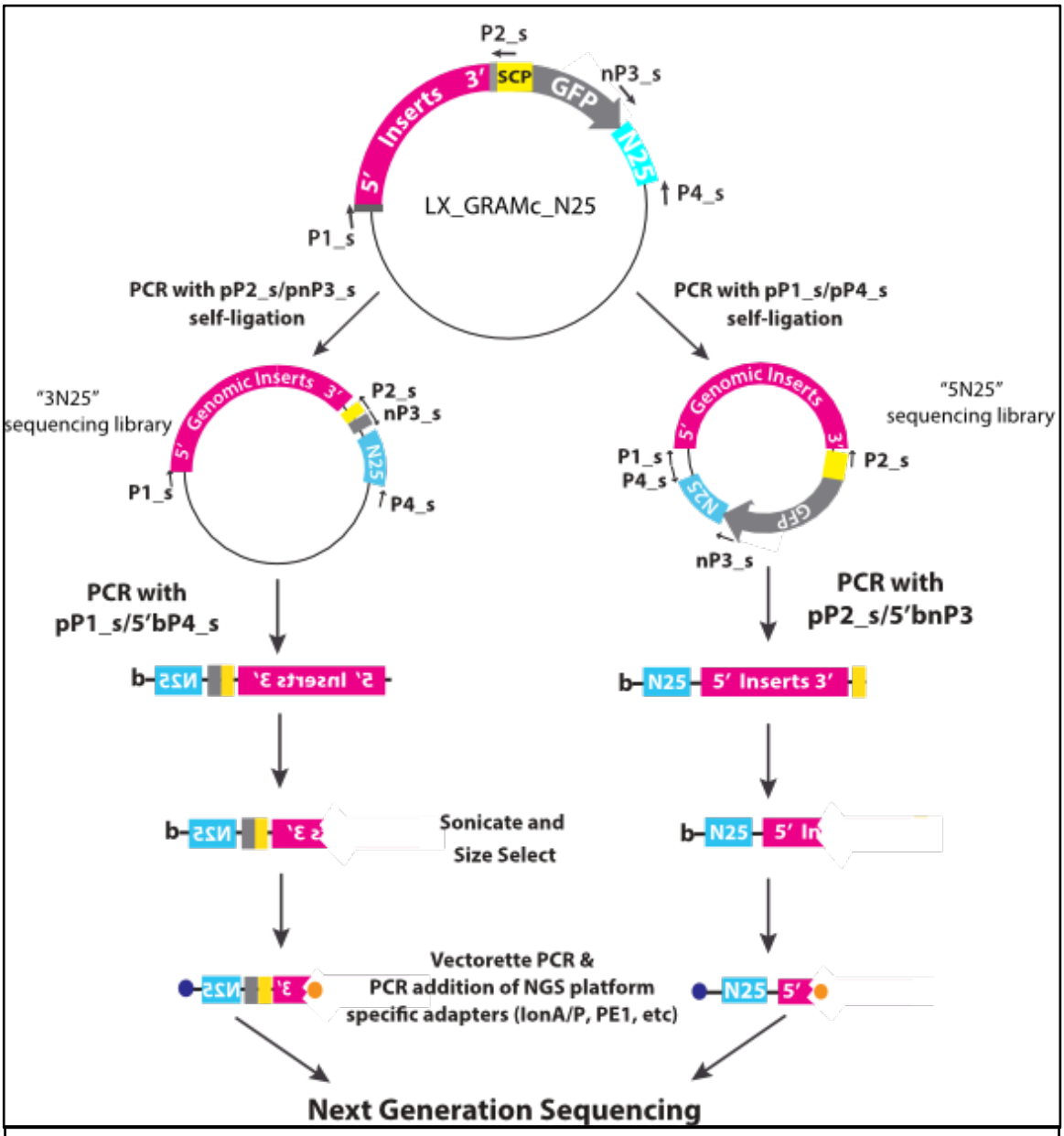


Fig. 11. Generation of a bi-directional mate-pair sequencing library. The plasmid reporter library is amplified through inverse PCR using two different primer pairs. The 3' ends of genomic inserts are brought within sequencing distance of the 5' end of N25 barcodes (5N25) through the removal of GFP. The 5' ends of genomic inserts are brought within sequencing distance of the 3' ends of N25 barcodes (3N25) by removal of the vector backbone. Inserts and barcodes are amplified from the self-ligated 3N25 and 5N25 library constructs by switching the primer pairs used for amplification. A molecule of biotin is included on the primer for the site adjacent to the barcode. Amplicons are sonicated and size selected to ~125bp. Fragments containing a barcode and ~50bp of a 3' or 5' end of an insert are isolated using streptavidin beads. Products are adapter ligated for Vectorette PCR and amplified. The 3N25 and 5N25 libraries are amplified to include Ion A and Ion P sites for sequencing on the Ion Torrent Proton platform (LifeTech).

Fragments containing N25-barcodes and their paired insert ends were

recovered using streptavidin beads. Vectorette PCR (Arnold and Hodgson, 1991) was used to prepare the bi-directional mate-pair libraries for Ion Torrent Proton sequencing. Additional Ion Xpress barcodes were added to the mate-pair libraries for simultaneous sequencing and rapid identification of 5N25 or 3N25 sequencing reads. 5N25 and 3N25 reads containing the same barcode were paired and subsequently mapped to the reference genome in accordance with a size restriction.

3.5 Bi-directional mate-pair characterization of sea urchin GRAMc libraries

Preliminary sequencing of the Sp3K_GRAMc libraries confirmed the success of the bi-directional mate-pair sequencing library approach. Mapped Ion Torrent Proton reads indicated a genomic coverage of ~58%, which is expected to increase as the *S.purpuratus* genome is further resolved. Shallow sequencing preliminarily revealed ~11M unique N25 barcodes for the Sp3KF_GRAMc library and ~18M unique N25 barcodes for the Sp3KR_GRAMc library.

3.6 Preliminary screening of the sea urchin forward orientation GRAMc library in embryos

We tested the Sp3KF_GRAMc library in 3 batches of ~3,000 (total 9,000) embryos that were sampled at the onset of gastrulation (24 hours post fertilization). On average, ~ 3M barcodes were tested in each batch of embryos. Though this portion of the work is preliminary, we can conclude from the number of tested barcodes that injection of >3,000 embryos per batch is require for true genome-scale analysis. While many previously unreported CRMs were identified in this preliminary screen, unbiased genome-scale analysis will require injection

of a larger number of embryos (discussed in Chapter 5). Despite the limitations of the current data, two newly identified fragments, Gpc1246_INT and FoxJ_5P, consistently showed high levels of activity (≥3-fold of background) across all 3 replicates, and were therefore selected for subsequent spatial characterization using MMOSAIC (see Chapter 4).

Because Gpc1246_INT and FoxJ_5P were highly active in all three sets of embryos, even when genome-scale testing was not achieved, we reasoned that they are likely to be broadly active. Upon transgenesis in sea urchin, small reporter constructs form concatenates with co-injected randomly fragmented genomic DNA, and are incorporated into a single cell in early stage embryos in a mosaic fashion. For this reason, some cells that overlap with the activity domain of the tested CRM will not harbor the transgenic reporter, resulting in no observed expression. When a CRM is broadly active, the likelihood of testing its associate incorporated reporter in cells that overlap with the CRM's activity domain is higher. We expect this to be reflected in a higher observed level of reporter expression for reporters paired with broadly active CRMs.

To further explore this hypothesis, we analyzed these CRMs using MMOSAIC (described in more detail in Chapter 4). According to the analysis, these CRMs show the highest overlap with previously characterized reference CRMs that mark the oral ectoderm. While this does not exclude the potential that these CRMs are more broadly active than in the oral ectoderm alone, it does indicate that their activity is unlikely to be truly ubiquitous. A discussion of the limits of MMOSAIC are included in Chapter 4. Further analysis to establish

prioritization of CRMs identified using GRAMc for subsequent spatial characterization is required once the library is tested in a larger number of embryos.

CHAPTER 4. Single embryo-resolution quantitative analysis of reporters permits multiplex spatial *cis*-regulatory analysis (Guay et al., 2017)

Catherine L. Guay, Sean T. McQuade, and Jongmin Nam (CLG and STM are co-first authors)

There were two major challenges to actualizing the spatial classification of CRMs. First, there was previously no sufficient method to identify multiple types of CRMs across the genome. Second, there was no efficient way to identify where in embryos the CRMs are active. Previously, spatial analysis of CRMs activity relied on imaging analysis of up to four optically distinct reporters, presenting a major bottleneck. The method for spatial *cis*-regulatory analysis presented in the paper, Multiplex and Mosaic Observation of Spatial Information encoded In CRMs, or MMOSAIC, is a radically different method that exploits the high-throughput multiplex potential of DNA barcode reporters and next generation sequencing to overcome the limitations of traditional imaging-based analysis. The new method can increase the throughput of spatial CRM analysis by several orders of magnitude.

I contributed the experimental methods that made the work possible. I also acquired experimental data and helped draft the manuscript.

The paper is reprinted here with permission from the publisher.

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.

Clusters of sequence reads from the same N25 barcodes were identified by single-linkage clustering allowing up to 2 mismatches (due to sequencing errors) in a Burrow-Wheeler Aligner search (Li and Durbin, 2009) against all identified N25 sequences. Each cluster identified in this analysis represents a unique barcode. The number of reads respectively from expressed barcodes and incorporated barcodes in each cluster are used to compute relative CRM activities with Eq. (1) and Eq. (2) in the Results section.

2.4. Generation of rank ordered profiles of reporter expressions

Rank ordered profile (ROP) of a CRM is a quantitative profile of ordered expression levels of 1000 barcoded reporters measured at single-embryo resolution. ROP generation is comprised of two steps: ***i***) random but controlled selection of 1000 barcodes and ***ii***) normalization of reporter expressions by the mean of relative expression levels of the sampled barcodes. In the first step, we randomly sampled 1000 barcodes that fit the following criteria: ***i***) barcodes that were incorporated up to the 116-cell stage, and thus are present in ~4 or more cells in an embryo at 18 hpf or 24 hpf and ***ii***) barcodes that are within 2–98 ranked percentiles of the barcodes that fulfilled the first criterion. The first criterion is to exclude barcodes that are carried in transgenic patches that are too small (< 4 cells) and thus require much more than 1000 barcodes to comprehensively cover, and the second criterion is to minimize the effect of outliers in reporter expressions. To find the scaling factor between the number of sequence reads and the number of cells that carry a unique barcode, we took the ratio of the number of reads for the most prevalent barcode to the number of total cells in an embryo at a given stage. For example, when the most prevalent barcode has ~200 reads from incorporated DNAs and the number of cells in an embryo is ~400, the scaling factor is 2. The scaling factor is then applied to estimate the number of cells from the number of sequence reads. For example, if a unique barcode has 2 reads from incorporated DNA, the number of cells that carry this barcode is estimated to be 4 (2×2 reads). In all experiments presented in this study, the number of the most prevalent barcode in each sample ranged from ~400 to ~540, thus the scaling factor was ~1. Therefore, the cutoff number of reads for selecting barcodes that were incorporated up to the 116-cell stage was 4. Note that the cutoff value for the number of reads from incorporated DNA can vary depending on sequencing depth and desired resolution of individual studies (e.g., 1/32th of an embryo instead of 1/116th).

In the second step, relative expression levels of individual barcodes were computed using Eq. (1) in the Results section. To neutralize the effect of expression levels and to only compare shapes of profiles, we normalized expression levels of individual barcodes to the mean levels of the 1000 sampled barcodes. Note that measurement of absolute level is not necessary, because it will be cancelled out during normalization to the mean level. Relative levels of normalized barcode expressions are ordered based on their ranks and resulting ROPs are visualized by using the ggplot2 package (Wickham, 2009).

Pair-wise distances (***D***) of ROPs were computed using Eq. (3) in the Results section. The statistical significance of an observed ***D*** value of a CRM between a control experiment and a perturbation experiment for a given CRM was computed by a computer simulation. The null hypothesis is that the observed ***D*** value between the two experiments occurred by chance alone. In this simulation, we bootstrap sampled (Efron and Tibshirani, 1986) two sets of 1000 barcodes in each experiment and computed a ***D*** value. We repeated this process 1000 times to generate a null distribution of ***D*** values for each experiment. This process generates two sets of null distributions respectively from a control experiment and a perturbation experiment. An averaged ***P***-value of the observed ***D*** value in the two null distributions is reported.

2.5. Imaging-based reporter assay

Co-injection of equimolar amount of Nodal_5P-RFP and CyIIIa_5P-GFP constructs were conducted as described in Nam et al. (2007). Three independent batches of injected embryos were observed for reporter expressions at 26 hpf. Fluorescent images were pseudo-colored and overlayed in the Fiji platform (Schindelin et al., 2012).

2.6. Ontogenic simulation

We developed a simulation strategy to mimick random mosaic incorporation and subsequent clonal amplification of an incorproated DNA based on the lineage map (Morrill and Marcus, 2004). The probability of a DNA incorporation event (***P***) is determined by our empirical distribution of DNA incorporation shown in Fig. 3**B** and by assuming all cells at the same time point have the same probability of incorporating reporter constructs,

$$P = P_i * 1/N_i$$

where P_i is the empirically determined probability of DNA incorporation at stage ***i***, and N_i the number of cells in an embryo at stage ***i***. After incorporation, the number of an incoporated reporter construct will increase following the cell lineage map. For any patch of a predefined set of active cells for a given CRM, expression level was set to 1. Reporter expression in an inactive cell was set to 0. Note that any non-zero value for an active cell will generate an identical profile after normalizing each value with the mean value of 10,000 simulations. Ontogenic simulation was conducted independently for each CRM. ROPs were plotted by using the ggplot2 package (Wickham, 2009). The MATLAB code for ontogenic simulation is available on GitHub (https://seanmcquade.github.io/Ontogenic-Simulation-of-Mosaic-Reporter-Assay/). Pair-wise distances of ROPs were computed using Eq. (3). To build the UPGMA tree shown in Fig. 8**D** we used the computer program MEGA5 (Tamura et al., 2011) with a matrix of pairwise distances of ROPs.

3. Results

MMOSAIC has two components: ***i***) a high-throughput single-embryo resolution reporter assay and ***ii***) data deconvolution to predict spatial patterns of CRMs based on a new theory (Fig. 1). The single-embryo resolution reporter assay measures expression levels of individual barcoded reporters in every single embryo from a large number of embryos. We then build quantitative profiles of normalized levels of reporter expressions driven by individual CRMs and use these profiles to predict spatial patterns. In the following, we will first describe the new single-embryo resolution reporter assay method and then will describe the underlying theory for the data deconvolution.

3.1. High-throughput cis-regulatory analysis at single-embryo resolution

The new quantitative method for single-embryo resolution *cis*-regulatory analysis is achieved by delivering each uniquely barcoded reporter construct into only one embryo during the entire experiment. Thus, although we pool many embryos with mosaic DNA incorporation (hereafter called mosaic embryos), transcripts of a unique barcode originate from one embryo. Note that one embryo can harbor many different reporter constructs, as long as each barcode is unique. At the heart of this method is our new 'extreme' barcoding technology, with which we can add many billions of unique random 25-mers (N25) to established constructs that already contain unique identifier (ID) sequences (Fig. 2**A**). By delivering only a small fraction of the reporter constructs into embryos, we can probabilistically introduce each unique barcode into only one embryo. For example, when we deliver 1000 molecules of constructs from a pool of 1,000,000 uniquely

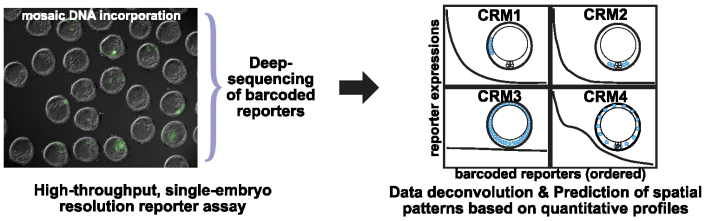


Fig. 1. Overview of MMOSAIC. MMOSAIC combines a novel high-throughput single-embryo resolution reporter assay with a new theory for data deconvolution to predict spatial activities of CRMs. Quantitative profiles of barcoded reporter expressions measured at single-embryo resolution in many embryos with mosaic DNA incorporation are used as fingerprints for spatial patterns. Expression levels of barcoded reporters are measured by a next-generation sequencing, and expression of a green fluorescent reporter gene is only for visualization. Spatial patterns of new CRMs can be predicted by matching their profiles to the profiles of CRMs with known spatial patterns. Hypothetical spatial patterns in embryos are marked cyan.

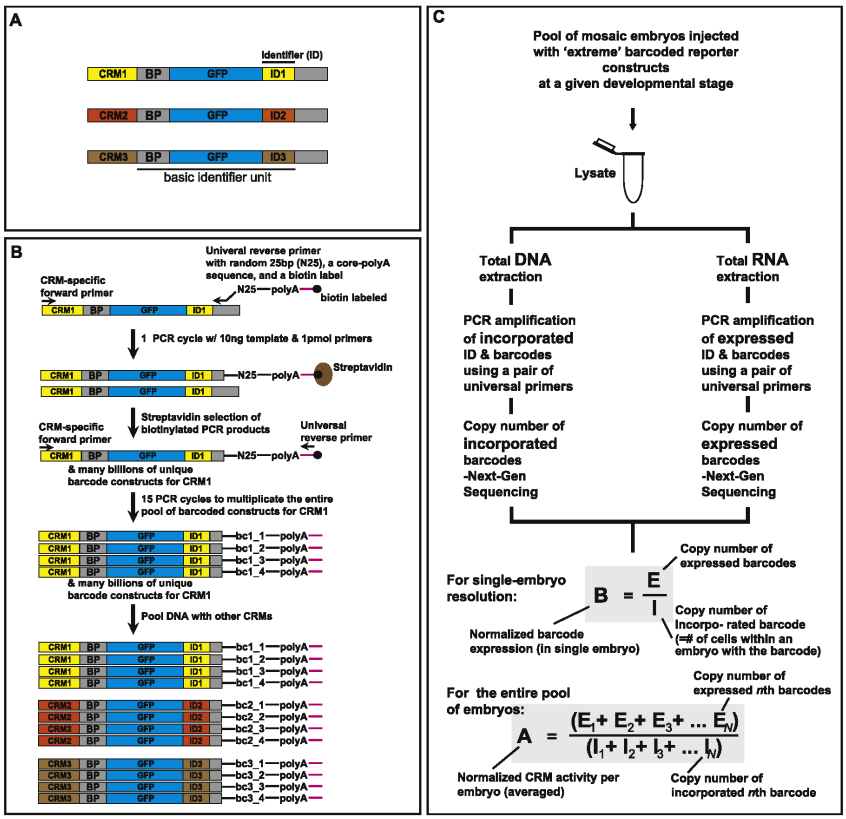


Fig. 2. Overview of barcoding technologies for quantitative *cis*-regulatory analysis. (A) Generation of reporter constructs in which unique identifier (ID) sequences are driven by a CRM (CRM::ID construct). The IDs sequences are used to identify CRMs. BP, basal or core promoter. (B) The "extreme" barcoding technology. Extreme barcodes are random 25 bp (N25) sequences, and we can add many billions of unique N25 barcodes to already established CRM::ID constructs. Extreme barcoded constructs for multiple CRMs are pooled together for injection. Next-generation sequencing is used to measure the copy numbers of individual barcodes from total RNA and total DNA. Because the number of unique barcodes tested in embryos is much smaller than the entire pool of barcodes, each barcode is tested in only one embryo in the entire experiment enabling single-embryo resolution *cis*-regulatory analysis. Extreme barcodes that are driven by individual CRMs are identified by the identifier sequences in the same sequencing reads. (C) Analytical procedure for single-embryo resolution *cis*-regulatory analysis. The entire pool of embryos carrying 'extreme' barcoded constructs at any given stage is harvested, and total genomic DNA and total RNA is extracted simultaneously. Barcodes in each sample are selectively PCR amplified by a pair of universal primers and will be sequenced. Relative expression level of each barcode (*B*) measured in this way is at single-embryo resolution, and the normalized activity of a given CRM (*A*) is computed by averaging the relative expression levels of all barcodes driven by the CRM.

barcoded reporter constructs, the probability of introducing the same barcode twice or more is extremely low (~5×10^{-7}). As a result, each unique barcode will be incorporated into only one initial cell, because there is only one sampled molecule for a unique barcode. (A single molecule of DNA cannot be incorporated more than once.) Fig. 2**B** outlines the experimental strategy of the new barcoding technology: ***i***) For each CRM driving a unique ID sequence in a reporter construct (CRM::ID) as shown in Fig. 2**A**, many billions of N25 are added by using a CRM-specific forward primer and a universal reverse primer that contains N25, a core poly-(A)denylation signal, and a biotin label at it's 5′-end. One cycle of PCR with these primers adds the unique barcodes to each molecule of CRM::ID constructs; ***ii***) Magnetic beads conjugated with Streptavidin capture only barcoded molecules; ***iii***) Ten to twenty PCR cycles with another pair of primers amplifies the entire pool of barcoded reporter constructs to make enough DNA for reliable measurement of concentration; ***iv***) Equimolar amounts of barcoded reporter constructs for different CRMs are pooled together and injected into embryos. In data deconvolution from sequence reads, different CRMs are distinguished by ID sequences, i.e., ID1 and ID2, and single-embryo resolution CRM activities are measured by using N25, i.e., bc1_1 and bc2_1.

To measure reporter expression and to correct for DNA copy number at a desired stage, we simultaneously extract total DNA and total RNA from the same lysate of embryos (Fig. 2**C**). PCR with a pair of universal primers selectively amplifies the entire set of barcodes from total genomic DNA and total cDNA, and next-generation sequencing is used for counting the relative copy numbers of barcodes. (Note that any potential amplification biases specific to each barcode should be common in both genomic DNA and cDNA templates, so the effect of such bias to final results should be minimal.) Because the sequence reads also contain IDs specific to individual CRMs, each N25 barcode can be computationally assigned to a CRM. Relative expression level of a unique barcode in a single embryo (***B***) is computed by the following equation.

$$B = E/I \tag{1}$$

where ***E*** is the number of sequence reads originated from expressed unique barcode and ***I*** the number of sequence reads from incorporated unique barcode. Note that ***I*** is proportional to the number of cells that carry the barcode within a single embryo, as each barcode is present in only one embryo. One can also estimate the traditional and averaged relative CRM activity (***A***) measured by reporter expressions using the following equation.

$$A = \frac{(E_1 + E_2 + E_3 + \ldots E_N)}{(I_1 + I_2 + I_3 + \ldots I_N)} \tag{2}$$

where E_N and I_N respectively are the numbers of sequence reads from expressed and incorporated N^{th} unique barcode, and ***N*** the total number of unique barcodes driven by a given CRM. Because the scales of ***B*** and ***A*** also depend on the number of sequence reads, we further normalize the data to the background leaky activity measured by using a co-injected inactive DNA fragment. The advantages of background normalization are *i*) reduction of variations among samples and *ii*) elimination of the need for measuring absolute levels (Nam et al., 2010; Nam and Davidson, 2012). We used Nodal_3P fragment to measure background activity, because no detectable positive or negative activity was observed with the fragment (Nam et al., 2007).

Using the new technology, we conducted a proof-of-concept experiment for single-embryo resolution *cis*-regulatory analysis. Four reporter constructs that respectively contain three known sea urchin CRMs, Nodal_5P, Nodal_INT, and Delta_5P, and one inactive DNA fragment, Nodal_3P (Nam and Davidson, 2012; Revilla-i-Domingo et al., 2004; Nam et al., 2007; Smith and Davidson, 2008), were "extreme" barcoded by the procedure outlined in Fig. 2**A and B** , and were co-injected into fertilized sea urchin eggs. We sampled 300 embryos at 18 h post fertilization (hpf), which is a routine number of embryos for traditional imaging-based *cis*-regulatory analysis. Quantitative reporter assay was conducted as outlined in Fig. 2**C**.

In this experiment, our QPCR analysis detected an average of 25 molecules of reporter constructs per cell. Therefore, in a successful single-embryo resolution analysis, if each of barcoded constructs was incorporated at 1-cell stage and thus all cells in an embryo carry the same 25 barcoded constructs, the expected number of unique barcodes in the 300 embryos is 7,400 (25 barcodes×300 embryos). If the total number of unique barcodes in 300 embryos were smaller than 7400, it indicates that some barcodes are present in more than one embryo. In an opposite scenario, if all barcoded constructs were incorporated at 18 hpf, every single cell in the 300 embryos should carry different barcodes. Therefore, the expected number of total barcodes in the 300 embryos is ~3,000,000 (25 molecules×~400 cells in an embryo at 18 hpf×300 embryos). Subsequent analysis of ~1.7 million sequence reads detected a total of 95,148 unique barcodes, which is far greater than the minimal number of 7400 barcodes for single-embryo resolution analysis. In addition, under the assumption that DNA incorporation occurred only once in each embryo, an earlier experiment estimated that stable incorporation of injected DNA into a replicating nuclear form occurs most often in a single cell at the 8 or 16 cell stages (Hough-Evans et al., 1988). In this case, one would expect 60,000 (25×8×300) or 120,000 (25×16×300) unique barcodes, and the number of unique barcodes that we detected, 95,148, is within this range. These results suggest that the barcodes that we detected are indeed from single embryos.

If barcodes are incorporated at 8-cell or 16-cell stages as estimated by Hough-Evans et al. (1988), one will be measuring reporter expressions in 8 or 16 different partitions of an embryo. Because there is no reason to assume only one incorporation event per embryo, in the following, we further investigated the temporal profile of DNA incorporation, which will inform us of the distribution of the sizes of transgenic patches that carry individual reporter constructs.

3.2. Temporal profile of DNA incorporation

To infer the temporal profile of DNA incorporation events, we numerically examined the relative numbers of sequence reads for the entire set of unique barcodes incorporated in the 300 embryos. In single-embryo resolution experiments, the copy number of a barcode cannot exceed the total number of cells in an embryo at a given developmental stage. Therefore, the copy numbers of individual barcodes should range between ~400 to 1 in embryos at 18 hpf: barcodes that are incorporated at 1-cell stage should be present in all ~400 cells, and unincorporated or most recently incorporated barcodes should be present in 1 cell. According to a sea urchin cell lineage map (ref), the number of cells increases by a factor of approximately 2, i.e., 1, 2, 4, 8, 16, 32, 60, ~116, ~216, and ~400, depending on the stage of development. Therefore, the average copy number of a barcode in an 18 h embryo is approximately 400 divided by the number of cells in an embryo at the time of incorporation (Fig. 3**A**). Since the numbers of sequence reads are proportional to the copy numbers of barcodes, the number of reads for individual barcodes are scaled by the ratio of 400 and the number of reads from the most prevalent barcode. The number of reads are then binned to different stages of incorporation: the low and high boundaries of a bin are defined by the geometric means of two adjacent bins. For example, boundaries for the barcodes that are incorporated at 8-cell stage, bin_8, are defined as $\sqrt{400/16\times400/8} <$ bin_8$\leq \sqrt{400/8\times400/4}$. Note that cell division in sea urchin embryos becomes less synchronous after the 5th cleavage (32-cell stage), thus our analysis is less accurate for incorporations in later stages.

When we examined ~1.7 million sequencing reads from incorporated barcodes, the number of reads varied from 1 to 540 for each of the 95,148 unique barcodes. Because the number of reads from the most prevalent barcode was already close to 400, for simplicity in data analysis, we assigned each barcode to the abovementioned bins with

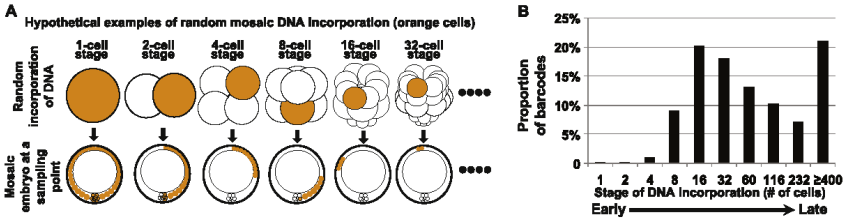


Fig. 3. Mosaic incorporation of injected DNAs in sea urchin embryos. (A) The timing of DNA incorporation and hypothetical examples of the pattern of mosaicism. Injected DNAs at one cell stage will be incorporated into nucleus for stable replication in one cell at any stage (cells marked in orange), and the ontogenic division of the transgenic cell will determine the degree and pattern of mosaicism at a later stage embryo. For example, incorporation at one cell stage will result in an embryo that carries DNA in all cells, and incorporation in one cell at 32-cell stage will generate a later embryo that carries DNA in approximately 1/32th of the cells that it contains. Note that these examples are only one example for each developmental stage, and the location of transgenic patch is determined by the initial cell that incorporates the injected DNA. After the 6th cleavage, divisions of cells in an embryo become much less synchronous, thus the degree of mosaicism depends heavily on the ontogeny of the initial cell where DNA incorporation occurred. (B) The relative frequency of the timing of DNA incorporation in early embryos up to 18 h. The stage of DNA incorporation for each barcode was estimated based on the relative copy numbers of barcodes measured from the total genomic DNA extracted from a pool of 300 mosaic embryos. DNA incorporation peaks at 16-cell stage and gradually decreases. The abnormally high peak at the last stage is for barcodes that are present as single copy in the pool, and we expect that the majority of these barcodes are not incorporated. Note however that a significant number of barcodes in this category are also expressed, illuminating the possibility of even single-cell resolution *cis*-regulatory analysis.

the scaling factor of 1. Fig. 3**B** shows that DNA incorporation peaks at 16–32 cell stages, and incorporation at even later stages is common. The discrepancy between the Hough-Evans and colleagues' estimate (1988) and our own suggests that multiple events of incorporation per embryo are prevalent. Note that the discrepancy is not due to the use of scaling factor 1, because use of scaling factor of 0.74 (400/540) pushes incorporations to even slightly later stages. Most strikingly, we also detected that ~20% of barcodes are present in only one cell, and we suspect that the majority of single copy barcodes are unincorporated. Nevertheless, some of single copy barcodes should have been incorporated as we detected expression of 7.7% of these barcodes. These results suggest that the new method can measure reporter expressions in many small partitions primarily in 1/8th–1/64th of an embryo.

In the following, we will describe the underlying theory for MMOSAIC that enables application of the single-embryo resolution reporter assay to a multiplex spatial *cis*-regulatory analysis.

3.3. Theory for MMOSAIC

Although mosaic DNA incorporation is a notorious problem that hinders spatial analysis of reporter expression, somewhat counter-intuitively, MMOSAIC turns this problem into an advantage for high-throughput spatial analysis of CRMs. The key is to measure reporter expression in every single mosaic embryo and build a profile or 'fingerprint' of quantitative reporter expressions from a sufficiently large number of embryos. By comparing the profile of an unknown CRM to reference profiles of known CRMs, one can predict spatial activity of a given CRM.

To further explain this strategy, let us first consider cases in which incorporation of injected reporter constructs occurs in one cell at the 32-cell stage (Fig. 4**A**). (In real experiments, a reporter construct can be incorporated at any stage.) Depending on the ontogenic division of the initial cell, a predefined set of cells will carry the reporter construct in later stages. Therefore, to test a given CRM in all possible embryonic territories, we need all 32 different types of incorporation. To sample all possible 32 types with a 99% confidence interval, assuming the uniform probability of DNA incorporation among the 32 cells, because this process is sampling with replacement, we need ~150 different barcodes. When we extend this computation to cover almost all types of random incorporation up to the 116-cell stage, we would need ~1000 barcodes per CRM. When the number of barcodes per CRM becomes very large, the overall distributions of mosaic incorporations between different CRMs become virtually identical.

Next, let us consider two hypothetical CRMs with different spatial activities, CRM1 for oral ectoderm and CRM2 for non-skeletogenic mesodermal cells (Fig. 4**B**). For the convenience of explanation, we will show reporter expressions in the same set of mosaic embryos shown in Fig. 4**A**. In each mosaic embryo, where cells active for each CRM are colored in cyan and cells carrying the reporter construct are colored in orange, reporters will be expressed only in the cells colored in both cyan and orange. For example, two copies of bc1_1 for CRM1 are present in an embryo and both are in inactive cells, thus the expression level of bc1_1 is 0/2. On the other hand, two copies of bc1_3 are present in an embryo and both are in active cells, thus the expression level of bc1_3 is 2/2. When the normalized levels of barcode expressions measured from six embryos are rank ordered, each CRM displayed a distinctive distribution of reporter expressions or rank ordered profile (ROP).

Let us also consider two additional patterns as shown in Fig. 4**C**, ubiquitous pattern driven by CRM3 and an evenly distributed spotty pattern by CRM4. These two patterns are interesting, because normalized expression levels of barcodes that are incorporated at early stages (e.g., 4-cell stage) will be uniform for both CRMs, and thus will be ineffective in distinguishing the two spatial patterns. However, as described in Fig. 4**C**, smaller transgenic patches of late stage-incorporated barcodes will differentially overlap with the two patterns, and thus will be able to distinguish the two patterns. Although the hypothetical patterns are compared side-by-side in the same embryo, in real experiments identifying such pairs of barcodes is not necessary, when large numbers of barcodes per CRM are compared.

The aforementioned hypothetical examples illustrate the principle that transgenic patches that differentially overlap with different spatial patterns are the key determinants of the ROP-based spatial analysis. To examine diverse spatial patterns, because different spatial patterns can be resolved by different sizes of transgenic patches, one needs to collectively analyze a large number of barcodes that are incorporated at different stages. For this reason, the spatial resolution of the new method is determined by the size of the smallest transgenic patches examined. We aim to resolve spatial differences up to 1/116th of an embryo. This is because use of 1000 barcodes that are incorporated up to the 116-cell stage will be sufficient to include almost all possible transgenic patches that are incorporated at 1~116-cell stages and is small enough for efficient high-throughput analysis. To achieve this level of resolution, because ~70% of barcodes are incorporated between 1-cell and 116-cell stages (see Fig. 3**B**), one needs to inject ~1500 barcodes per CRM. To comprehensively examine barcodes in embryos that harbor ~1500 barcodes per CRM, the number of sequence reads required is ≥100,000 per CRM.

Note that an inherent limitation of the new method in relation to the complexity of cell lineage is investigated by computer simulation below.

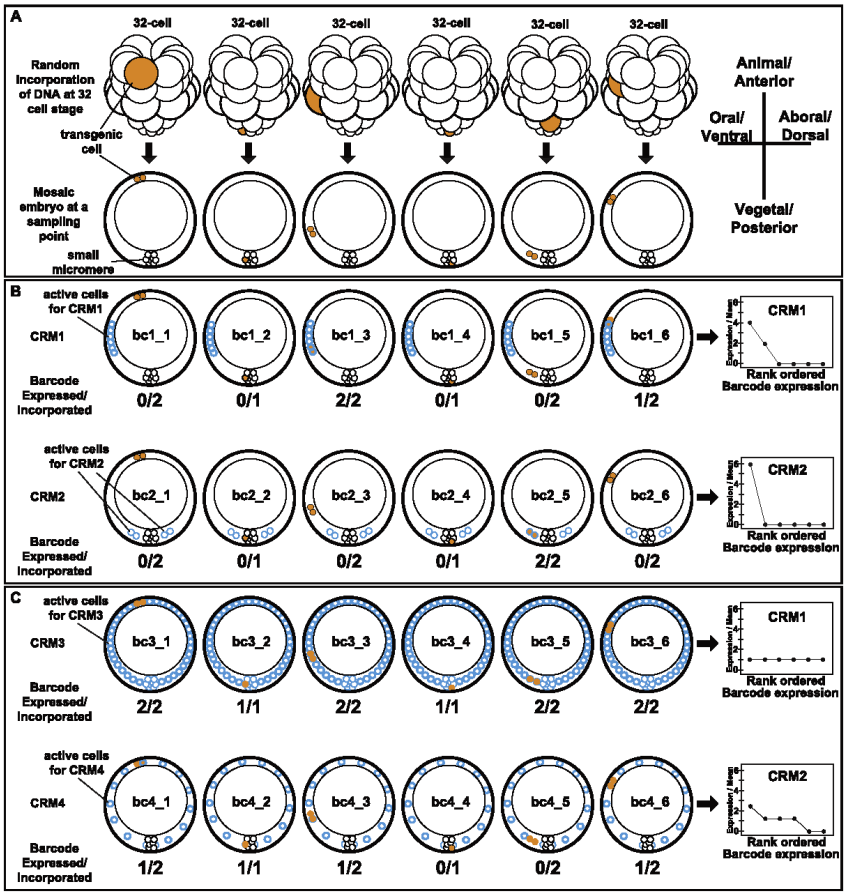


Fig. 4. Hypothetical examples of mosaic DNA incorporation and single-embryo resolution reporter expressions. With several hypothetical examples, we illustrate our working hypothesis that the profile of single-embryo resolution reporter expressions measured from many mosaic embryos can be used to infer spatial activity of a CRM. (A) Six hypothetical examples of DNA incorporation at 32-cell stage and resulting pattern of mosaicism at a later stage. Orange cells are where injected DNAs are incorporated into nucleus and stably replicated during development. The number of transgenic cells varies depending on the ontogeny of the initial cell for DNA incorporation. For example, small micromere cells stop dividing after reaching eight cells. Two axes of embryos are shown on the right side of the embryos. (B) & (C) Construction of the profile of single-embryo resolution reporter expressions driven by a given CRM. Reporter expressions in single embryos driven by two CRMs with different spatial activities are shown for six mosaic embryos from A. Active cells for each CRM are marked with cyan and cells that carry reporter constructs are marked with orange, and each barcode reporter (bc1s for CRM1, bc2s for CRM2, bc3s for CRM3, and bc4s for CRM4) is expressed in cells where the two colors overlap. The copy number of each barcode expressed and that of incorporated are listed below each embryo. The expression levels of barcodes per embryo after correcting for DNA copy number are used to build a profile for each CRM. For the same set of six mosaic embryos, the two CRMs of different spatial activities are distinguishable. In real profiles, individual CRMs will have ≥1000 barcodes, and these profiles are "fingerprints" of individual CRMs.

3.4. Rank ordered profiles (ROPs) can distinguish similar and dissimilar spatial patterns

For the practical application of the theory, we examined single-embryo resolution data generated by using four control constructs mentioned above: Two CRMs, Nodal_5P and Nodal_INT, recapitulate expression pattern of *nodal* in ~50 cells in the presumptive oral ectoderm throughout early embryogenesis (Li et al., 2014); Delta_5P recapitulates the expression pattern of the sea urchin *delta* gene in ~20–30 cells in non-skeletogenic mesodermal cells at 18 h embryos (Materna and Davidson, 2012); Nodal_3P is an inactive DNA fragment and was used to measure background opportunistic expression of reporters (Nam et al., 2007). A total of ~1.7 million reads and ~1.5 million reads from incorporated barcodes and expressed barcodes were obtained, respectively. Following the procedure outlined in Fig. 4**B**, barcodes for each CRM were rank ordered based on the number of sequence reads from expressed barcodes normalized to those from incorporated barcodes (Fig. 5). We randomly sampled a set of 1000 barcodes that are incorporated up to the 116-cell stage for each CRM.

The quantitative profiles shown in Fig. 5**A** are barcode expressions normalized to DNA copy number of the same barcode (or number of cells that harbor the barcode) at 18 hpf. The average expression levels of barcodes per CRM are shown within parentheses. As noted earlier (Nam et al., 2010; Nam and Davidson, 2012), even the negative

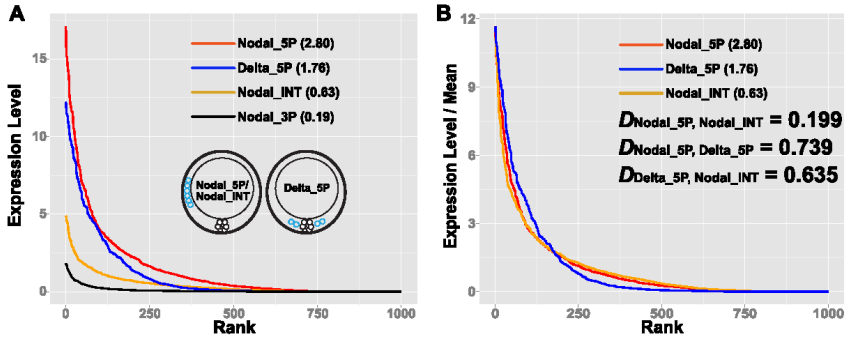


Fig. 5. ROPs and spatial activities. We empirically visualize our theory by using four controls: two CRMs from the *nodal* gene, Nodal 5P and Nodal INT, display significantly overlapping spatial activities in the oral ectoderm; one CRM from the *delta* gene, Delta R6, is active in the non-skeletogenic mesoderm; and Nodal 3P is an inactive fragment and used as negative control. (A) ROPs of four CRMs at 18 hpf. Of many tens of thousands of barcodes tested for each CRM, we randomly sampled 1000 barcodes that fit the following criteria: *i*) barcodes that are incorporated up to the 116-cell stage and *ii*) barcodes that are within 2–98% interval of the original ranks. These criteria were respectively applied to avoid under-sampling of barcodes for small transgenic patches and to minimize the effect of outliers. Relative activities of individual CRMs were computed using Eq. (2). Spatial patterns of the three CRMs are shown with cyan circles. (B) Comparing shapes of ROPs. The levels of barcode expressions are inversely scaled to the averaged level of each CRM, which is given within parentheses. Some of the high ranks of Nodal 3P are too high to show in this figure, thus are omitted for the convenience of presentation. The relative shapes and pairwise distances (***D***) of the profiles are consistent with known activities of the drivers.

control, Nodal_3P, displayed residual activity. Our result shows that the leaky background expression is attributed by a small number of randomly incorporated constructs, while the majority of constructs remained silent. All three positive control constructs displayed expression levels more than 3-fold higher than that of Nodal_3P. In addition, the number of expressed barcodes is far greater for active CRMs than for Nodal_3P, which can be an effective quantitative criterion for discovering weak CRMs.

To test whether the shape of ROPs can be used as "fingerprints" of spatial *cis*-regulatory patterns, expression levels of individual barcodes were inversely scaled to the averaged expression level of the 1000 sampled barcodes for each CRM (Fig. 5**B**). After initial visual examination of the profiles of the three active CRMs, as expected from their known spatial activities, profiles of Nodal 5P and Nodal_INT were distinct from that of Delta_5P.

Next, we quantitatively measured pairwise differences (***D***) of the profiles using a simple nonparametric method with the equation shown below.

$$\mathbf{D}_{\mathrm{CRM1,CRM2}} = \sum_{n=1}^{N} [|B_{\mathrm{CRM1,n}} - B_{\mathrm{CRM2,n}}|/\langle B_{\mathrm{CRM1,n}} + B_{\mathrm{CRM2,n}}\rangle/2]/N \quad (3)$$

where $\boldsymbol{B}_{\mathbf{CRM1,n}}$ and $\boldsymbol{B}_{\mathbf{CRM2,n}}$ respectively are ***n***th ranked barcode expression for two CRMs compared, and $[|\boldsymbol{B}_{\mathbf{CRM1,n}}-\boldsymbol{B}_{\mathbf{CRM2,n}}|/(\boldsymbol{B}_{\mathbf{CRM1,n}}+\boldsymbol{B}_{\mathbf{CRM2,n}})/2]$ the absolute difference of expressions for ***n***th ranked barcodes normalized to their mean. When the mean is 0, it was replaced by 1. Consistent with their known spatial activities, $\boldsymbol{D}_{\mathbf{Nodal\_5P,\ Nodal\_INT}}$ was very small (0.199), while $\boldsymbol{D}_{\mathbf{Nodal\_5P,\ Delta\_5P}}$ (0.739) and $\boldsymbol{D}_{\mathbf{Delta\_5P,\ Nodal\_INT}}$ (0.635) were considerably higher.

The above example demonstrates that the ROP-based spatial analysis has the potential to distinguish similar patterns from different patterns. In the following, we expand MMOSAIC to other embryonic territories and to predict spatial patterns of new CRMs.

3.5. ROP-based prediction of CRMs with similar spatial patterns

To predict CRMs with similar spatial patterns, we first measured the expected variation of ROPs of 17 selected CRMs originated from 14 genes with territory-specific gene expression patterns. (The names and sequences of the CRMs are provided as Supporting File 1.) When there is no experimental variation with infinite number of barcodes, the expected ***D*** value for two identical spatial patterns should be 0. However, in real experiments, ***D*** values can be greater than 0 even between CRMs of identical spatial patterns. To measure variation of ***D*** values, we barcoded the 17 CRMs with a total of 58 different IDs, which were then extreme barcoded (Fig. 6**A**). Different IDs for the same CRM are expected to have identical profiles. ROPs of the 58 reporter constructs were measured in a single batch of 300 embryos at 24 hpf as described in Fig. 2**C**. We chose 24 hpf due to the availability of most spatial patterns. A total of ~3.6 million reads for expressed barcodes and ~5.1 million reads for incorporated barcodes were analyzed. The total number of unique barcodes detected was 433,332, and the number of unique barcodes per CRM: ID construct was ≥1950. As described above, the ROPs for individual CRM::ID constructs were generated from randomly selected 1000 barcodes that were incorporated up to the 116-cell stage. When the ROPs of the 58 different IDs were compared, ***D*** values between pairs of IDs for the same CRM showed a tight distribution with the maximum value of 0.115 ($\boldsymbol{D}_{\mathbf{CRM\_same}}$ in Fig. 6**B**). In addition, the median of $\boldsymbol{D}_{\mathbf{CRM\_same}}$ values were ~0.035 (red line in Fig. 6**B**). On the contrary, ***D*** values between different CRMs showed much wider distribution with the maximum value of 0.425 ($\boldsymbol{D}_{\mathbf{CRM\_diff}}$ in Fig. 6**B**). On the basis of these results, two different CRMs with $\boldsymbol{D}_{\mathbf{CRM\_diff}} \leq 0.035$ (equivalent to ~50 percentile of $\boldsymbol{D}_{\mathbf{CRM\_same}}$) can be considered to have similar spatial activities in 24 hpf embryos. When there are more than one CRM with low ***D*** value for a CRM of interest, we used a CRM with the lowest ***D*** value for single-linkage clustering. In this way, we classified the 17 CRMs into five groups (Fig. 6**C**). The list of ***D*** values used for clustering are provided as Supporting File 3.

Of the five predicted groups, four groups contained at least one reference CRM that has previously been examined in blastula stage embryos: Group 1 contains two oral ectoderm-specific CRMs, Nodal_5P and Nodal_INT (Nam et al., 2007; Range et al., 2007); Group 2 contains three endodermal CRMs with ectodermal ectopic activities, Endo16_All, Eve_5P, and Wnt8_A (Kirchhamer et al., 1996; Smith et al., 2008; Minokawa et al., 2005); Group 3 contains a skeletogenic mesoderm-specific CRM, Alx_J (Damle and Davidson, 2011); Group 4 contains an endoderm-specific CRM with very low ectopic activity, Gatae_m10 (Lee et al., 2007). Because there is no reference CRM in Group 5, we can only predict that spatial activity of

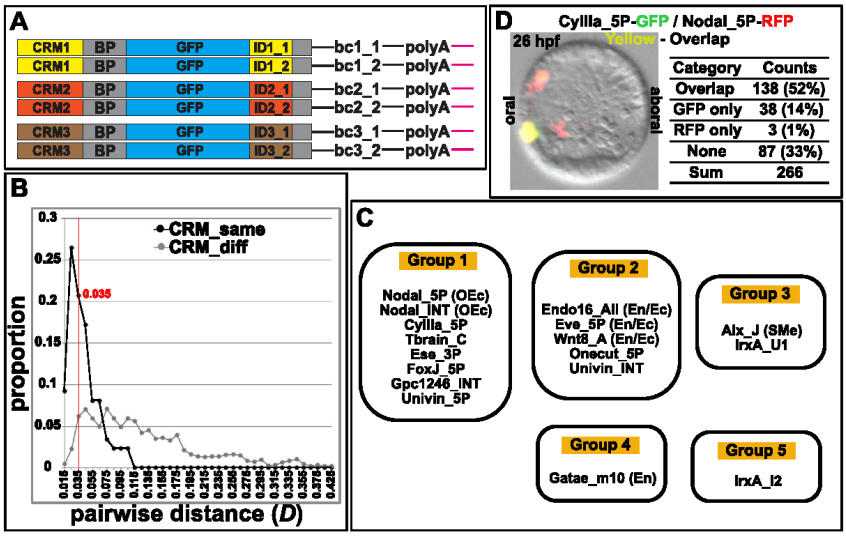


Fig. 6. Empirical validation of the ROP-based prediction of spatial activities. Seventeen CRMs were selected for ROP-based spatial analysis at 24 hpf. Seven CRMs have spatial information available at 24 hpf. (A) Building of reporter constructs. The 17 CRMs were represented by a total of 58 different IDs with a minimum of 2 IDs per CRM. CRM::ID constructs were extreme barcoded by PCR as described in Fig. 2A. (B) Comparison of reporter constructs driven by the same CRMs and those driven by different CRMs at 24 hpf. ROPs between the same CRMs with different IDs (black dots, CRM same) have smaller D values (Eq. (3)) than D values between different CRMs (gray dots, CRM diff). We empirically determined that the D values for the same CRMs were ≤0.115 at 24 hpf, with the median of $D_{\mathrm{CRM\_same}}$ values ≤0.035 (red vertical line). (C) Classification of the 17 CRMs based on the ROPs. Known spatial activities of 7 CRMs are consistent with the classification of the ROP-based groups. Known spatial activities of CRMs are shown within parentheses: **Ec**, Ectoderm; **OEc**, Oral Ectoderm; **En**, Endoderm; **SMe**, Skeletogenic Mesoderm. (D) Experimental validation of a prediction. Our method predicted that Nodal 5P module and CyIIIa 5P module overlap at 24 hpf. The former is marker for oral ectoderm in the blastula stage embryo and the latter for aboral ectoderm in the pluteus larval stage. Spatial activity of CyIIIa 5P module at the blastula stage embryo has not been reported. When Nodal 5P (RFP) and CyIIIa 5P (GFP) were examined in coinjected embryos, the two modules showed overlapping activities (yellow) in the blastula stage embryos. Summary statistics of scoring reporter expressions in three independent batches of embryos is shown. This result confirms our prediction for Group 1.

IrxA_I2 module is distinct from other groups. Given the aboral ectoderm expression of *irxa* in blastula stage embryos (Su et al., 2009), it is possible that IrxA_I2 is active in aboral ectoderm. Overall, the ROP-based prediction classified CRMs with similar spatial activities together and separated CRMs with distinct spatial activities.

Of the 10 CRMs with predicted spatial activities, CyIIIa_5P module (Hough-Evans et al., 1988) was clustered with two oral ectoderm-specific CRMs, Nodal_5P and Nodal_INT (D ≤0.03). This is interesting, because CyIIIa_5P module was assumed to be aboral ectoderm-specific in blastula stage embryos based on its activities in gastrula stage embryos (Kirchhamer et al., 1996). To validate our prediction, we coinjected CyIIIa_5P::GFP and Nodal_5P::RFP in one-cell stage embryos and observed their activities at 26 hpf in three batches of embryos. Nodal_5P module was used as marker for oral ectoderm as in Nam et al. (2007). Consistent with our prediction, CyIIIa_5P displayed overlapping activity with Nodal_5P in 138 out of 266 (52%) blastula stage embryos observed (Fig. 6**D**). Thirty-eight (14%) embryos showed only GFP expression and three (1%) embryo showed only RFP expression. Eighty-seven (33%) embryos showed neither GFP nor RFP expression. This result suggests that spatial activity of CyIIIa_5P is slightly broader than that of Nodal_5P. Note that over-injection of DNA can cause overlapping expression of GFP and RFP in nearly all embryos due to predominant ectopic reporter expressions. Although comparison with endogenous gene expression patterns (e.g., *nodal* expression in oral ectoderm) would be the most definitive validation of our prediction, distinctive expression of GFP in the 14% of the embryos cannot be explained by overinjection.

In summary, our results showed that the new method has a sufficient resolution to assign CRMs into different germ layers. However, we note that more comprehensive future studies than the present study are necessary to fully understand the resolution of the new method in various developmental stages.

3.6. Simultaneous measurement of quantitative and spatial cis-regulatory responses to gene perturbations for gene regulatory network analysis

Since CRMs causally mediate gene networks, the most immediate application of MMOSAIC is to simultaneously measure quantitative and spatial CRM responses to gene perturbations. Quantitative response of a CRM to gene perturbation can be computed by using Eq. (2). In the case of spatial response, Fig. 7**A** shows expected changes in profiles depending on how a CRM responds to perturbations: ***i***) when spatial activity of a CRM does not change, regardless of level, the normalized profile to its mean activity will be identical or very similar to the control profile; ***ii***) when CRM activity expands to a larger number of cells in an embryo, the quantitative profile will show fewer zeros and a lower level of high-ranked values in comparison to the control profile; ***iii***) when CRM activity contracts to a fewer number of cells in an embryo, the proportion of the zero-valued tail will increase.

To test the theory, we combined our single-embryo resolution *cis*-regulatory analysis with a gene perturbation experiment. Responses of a new set of 16 CRMs to *nodal* mRNA overexpression (MOE) were measured at 18 hpf (Fig. 7). Extreme barcoded reporter constructs of the 16 CRMs were coinjected with *nodal* mRNA in a perturbation experiment and were injected alone in a control experiment. We sampled a single batch of 300 embryos for each experiment. We identified ~1.3 million barcodes and ~1.6 million barcodes by analyz-

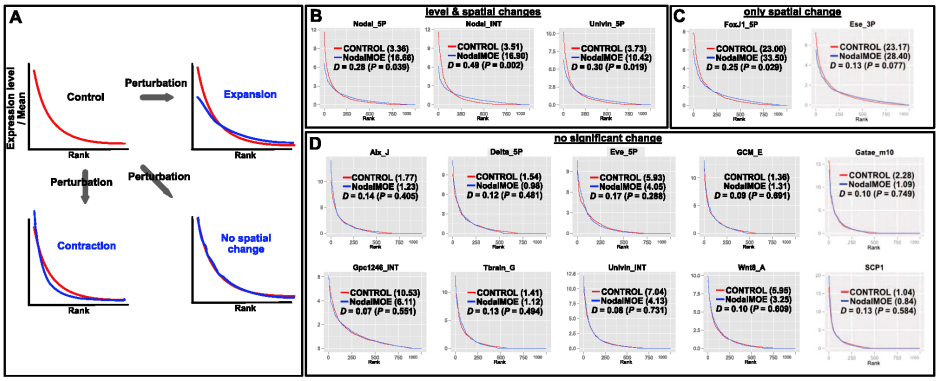


Fig. 7. Measuring CRM responses to gene perturbations. Changes in ROPs of 14 CRMs were measured in *nodal* mRNA over-expressed (MOE) embryos at 18 hpf. CRM activities normalized to the background activity are shown within parentheses and pairwise distance (***D***) between two profiles and ***P***-values are also shown for each comparison. To minimize sample-to-sample variations, CRM activities were normalized to the background leaky activity of Nodal_3P fragment in each condition. (A) Expected changes in profiles when CRM activity expands or contracts. (B) CRMs that displayed both level and spatial changes. All three CRMs were up-regulated and expanded by *nodal* MOE. These CRM responses have been confirmed by independent experiments. (C) CRMs that showed significant spatial changes but not level changes. These CRM responses have not been confirmed by independent experiments. (D) CRMs that showed no significant change.

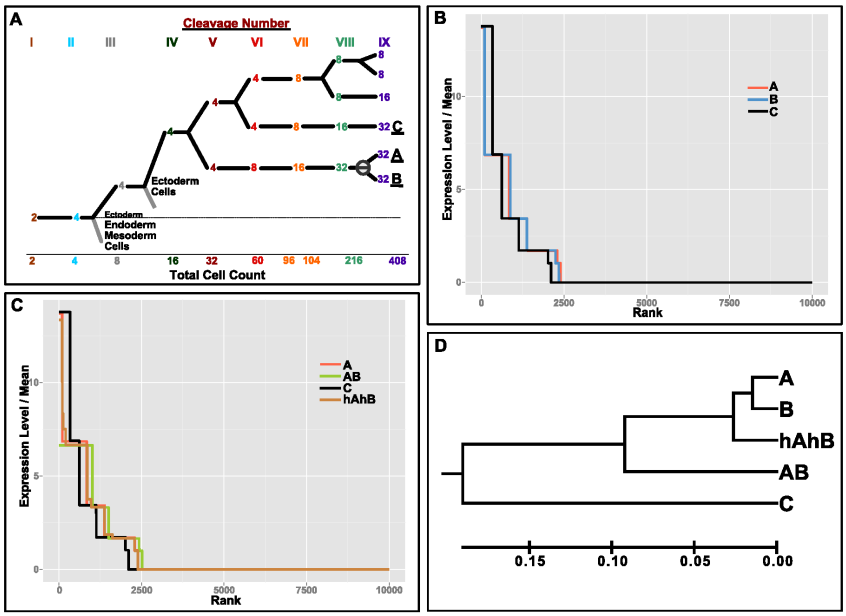


Fig. 8. Complexity of cell lineage and spatial resolution. Ontogenic simulation was used to computationally generate ROPs of three distinct spatial patterns of identical cell numbers, **A**, **B** and **C**. Reporter constructs were randomly incorporated following the frequency distribution of DNA incorporation shown in Fig. 3B, and propagation of an incorporated construct followed sea urchin cell lineage map (Morrill and Marcus, 2004). (A) Location of three patterns in the cell lineage map up to 9th cleavage (116-cell stage). Note that, due to latitudinal orientation of the 9th cleavage that separates **A** and **B**, any reporter constructs that are incorporated 9th or earlier cleavage events will be equally shared by both **A** and **B**. The numbers of cells after each cleavage are shown in nodes, and this map shows only animal half of the sea urchin embryo that leads to ectoderm. (B) Computationally generated ROPs of the three spatial patterns. In the simulation, we assumed homogeneous reporter expressions among active cells with no noise in measurement. (C) ROPs of two additional patterns. (D) A tree based on the pairwise distances (*D*) of ROPs. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method (Nei and Kumar, 2000) was used.

ing ~65 million reads from the control sample and ~56 million reads from the perturbed sample, respectively. The number of barcodes per CRM varied from ~48,000 to ~164,000. Note that the large number of barcodes tested in these experiments is an outcome of a conventional microinjection condition. As in the case of ROP-based prediction of spatial patterns, we randomly sampled 1000 barcodes that were

incorporated up to the 116-cell stage. To minimize variations among samples, CRM activities were normalized to the background leaky activity of Nodal_3P. Pair-wise distances (***D***) of ROPs were computed by using Eq. (3). The statistical significance of ***D*** values were computed by a computer simulation (see the Methods section).

Sea urchin *nodal* is a key initiator of oral ectoderm gene regulatory network and overexpression of *nodal* mRNA oralized embryos (Duboc et al., 2004). Earlier studies showed that *nodal* MOE up-regulated activities of Nodal_5P, Nodal_INT (Nam et al., 2007; Range et al., 2007), and Univin_5P (Nam and Davidson, 2012). Consistent with the earlier results, activities of the three CRMs were significantly up-regulated (≥3-fold) and expanded ($P < 0.05$) in *nodal* MOE embryos (Fig. 7**B**). We also found two CRMs, FoxJ1_5P and Ese_3P, that only displayed expansion of spatial patterns without a significant level change (Fig. 7**C**). Endogenous gene expression patterns of *foxj1* and *ese* are ectodermal and broad at 18 h embryos, respectively (Rizzo et al., 2006; Tu et al., 2006). It is noteworthy that all *nodal*-responsive CRMs are members of Group 1 (Fig. 6**C**), which most likely represent ectodermal CRMs. The remainder did not show any significant changes (Fig. 7**D**).

These results demonstrate that our method can simultaneously detect changes in both quantitative and spatial activities of multiple CRMs in response to perturbations without relying on imaging tools. In addition, the excess number of barcodes tested in the 300 embryos in this experiment demonstrated that MMOSAIC is easily scalable to ≥850 CRMs (≥1.3 million/1500 barcodes per CRM) in just 300 embryos.

3.7. Effect of complexities in cell lineage on spatial resolution

When two CRMs are active in the same number of different cells, it is not easy to intuitively predict whether ROPs can distinguish them. To examine how complexities in cell lineages affect the spatial resolution of the new method, we computationally simulated the ROPs of CRMs that are active in the same number of cells that belong to three different cell lineages (cell groups **A**, **B**, and **C** in Fig. 8**A**). Groups **A** and **B** have symmetrical histories, while that of group **C** is asymmetrical to those of **A** and **B**. Each group contains 32 cells that share the same lineage history up to the 9th cleavage. Reporter constructs respectively active in only cell groups **A**, **B**, and **C** were randomly incorporated 10,000 times and their expression level after the 9th cleavage for each incorporation event were recorded mimicking the experimental analysis.

Fig. 8**B** shows simulated quantitative profiles of three CRMs that are exclusively active in cell groups **A**, **B**, and **C**, respectively. Cell groups **A** and **B** that have nearly identical profiles due to their symmetrical cell lineage histories, and the small difference is the result of random sampling. On the other hand, the profile of cell group **C** is distinct from those of **A** and **B** due to their asymmetric cell lineages. To further examine how asymmetric variations in cell lineages can change ROPs, we tested two spatial patterns within cell groups **A** and **B**: a spatial pattern that encompasses the entire set of 64 cells in **A** and **B** (**AB**) and another pattern that includes 16 cells from **A** and 16 cells from **B** (**hAhB**). The ROPs of **AB** and **hAhB** were distinct from those of **A**, **B**, and **C**, suggesting that spatial patterns that are of asymmetric cell lineages can be distinguished by the ROP (Fig. 8**C**). A cladogram constructed based on the pair-wise differences of the profiles using Eq. (3) further supports this conclusion (Fig. 8**D**).

Although current version of ontogenic simulation does not consider experimental noises, and thus is limited in accurately simulating ROPs, we could illustrate an inherent inability the ROP-based method to resolve two patterns with symmetrtical cell lineages.

4. Discussion

In this paper, by using experimental and computational approaches, we demonstrated the feasibility of a new method for multiplex spatial *cis*-regulatory analysis. We showed that the ROPs of reporter expressions measured at single-embryo resolution from many mosaic embryos can be used to find similar/different spatial CRM activities in normal and perturbed embryos. In the following, we will discuss potential pitfalls and prospects of the new method.

4.1. Properties of MMOSAIC

Because the new method's data form is unconventional and does not directly present spatial pattern, there will be some degree of uncertainties until many independent applications of the method become available. We therefore will discuss potential pitfalls first. Two factors are critical for the spatial resolution of the new method: ***i***) the diversity of random mosaic patches and ***ii***) the degree of unequal rate of division among lineages. First, the number of cells in an embryo at the time of DNA incorporation determines the size and diversity of random mosaic patches at the time of sampling, which in turn affect spatial resolution of our new method. For example, DNA incorporation in the one cell-stage embryo will have homogeneously incorporated DNAs in every cell in a later stage embryo. This, however, will have no diversity in the size of patches, thus uninformative in our method. Similarly, reporter constructs that were incorporated just before sampling are present in only one cell, and are therefore equally as uninformative as a whole embryo incorporation. Our empirical survey of DNA mosaicism show that the majority of random events of DNA incorporations occurred at 16, 32, and 60-cell stages, resulting in a complex set of transgenic patches that can distinguish more than 60 distinct spatial patterns. Second, when all cells in an embryo divide at an equal rate and are of the same size and spatial patterns of CRM activities are exclusively bound to cell lineages (monophyletic), spatially distinct CRMs that are active in the same number of cells in an embryo would be indistinguishable in our method as shown in computer simulation (Fig. 8). Alternatively, when active cells for a given CRM have multiple origins and do not share a common ancestor (polyphyletic), the number of spatial patterns that can be distinguished by quantitative profile will be much larger. In addition, cells in different lineages often have different level of transcriptional activities due to variations in size, which can further increase the resolution of the new method. As embryogenesis progresses, these variations will only increase further enhancing the resolution of our method. Note that a set of 1000 barcodes that we use to generate a ROP is large enough to cover almost all random patches orignated from incorporation events up to the 116-cell stage, and this will set the resolution limit of the new method, i.e., differences in few cells cannot be distinguished with 1000 barcodes per CRM.

Another challenge in the reporter assay is the choice of basal promoters. Because it is impractical to use endogenous basal promoters for each CRM in large scale experiments, we used a synthetic, pan-bilaterian Super Core Promoter 1 (SCP1) (Juven-Gershon et al., 2006) as a heterologous basal promoter in the reporter constructs. It is possible that some CRMs may not be compatible with SCP1 or may have ectopic activities in combination with SCP1. Although we have not seen clear cases of such an artifact, it recommended to use multiple different basal promoters or use endogenous basal promoters when possible.

4.2. Accuracy of MMOSAIC

MMOSAIC has successfully classified CRMs in a manner that is consistent with known activities of reference CRMs in different germ layers. MMOSAIC could also distinguish endodermal CRMs with different degrees of ectopic activities (Groups 2 and 4 in Fig. 6**C**). In addition, all *nodal*-responsive CRMs detected are grouped with oral ectoderm-specific CRMs (Group 1 in Fig. 6**C**) where *nodal* is a key regulator in sea urchin embryos (Duboc et al., 2004). Because we validated only one out of 10 predictions, our assessment of the accuracy of MMOSAIC is still preliminary.

Additional questions that need to be addressed are: ***i***) what is the resolution of MMOSAIC? ; ***ii***) is there any difference in the resolution of MMOSAIC between different embryonic territories? ; ***iii***) do developmental stages have an impact on the accuracy and resolution of MMOSAIC? Comprehensive examination of the accuracy of MMOSAIC would require systematic examination of many CRMs with varying degrees of spatial differences in different embryonic territories and in different developmental stages. Given the challenges to even obtain such a set of CRMs, iterative approaches that combine large-scale prediction of spatial CRM activities and focused examination of CRMs of interest would provide a better understanding on the accuracy of the new method. These newly characterized CRMs will in turn serve as references to better predict spatial patterns of other CRMs. It is also possible that an improved version of our simulation-based approach with the consideration of realistic parameters (e.g., experimental and biological noises) and an accurate map for cell lineages (Villoutreix et al., 2016) would help overcome some of practical limitations.

In our understanding, MMOSAIC is best suited for large-scale analyses of CRMs which can then be aided by more targeted, high-resolution analysis by traditional imaging-based methods.

4.3. Scalability of MMOSAIC

The strength of MMOSAIC over traditional imaging-based approach is in its scalibility. Three factors are critical for the scalability of the new method: ***i***) construction of reporter constructs with unique identifier sequences, ***ii***) the number of sequencing reads, and ***iii***) delivery of DNA into embryos. First, it is not pragmatic to individually build a large number (≥100) of CRM::ID constructs with fusion PCR that we used to construct four test constructs. However, use of the library of pre-barcoded vectors that contain ~100 million unique barcodes can signifiantly accelerate cloning of several hundred CRMs. Second, the number of sequencing reads necessary for each CRM and the number of CRMs determine the minimal number of sequencing reads necessary for an experiment. In our experience with sea urchin embryos, ≥100,000 reads are necessary per CRM per condition to collect enough data for an ROP. Given the throughput of major sequencing equipment, a single sequencing run would generate enough sequencing reads to analyze several hundreds of CRMs. Third, in the case of the scale of DNA delivery, DNA microinjection into sea urchin embryos is already efficient enough to analyze hundreds of CRMs in parallel (e.g., Nam and Davidson, 2012 and present study). In addition, we did not see any noticeable developmental defects in embryos that carry ~1000 molecules of reporter constructs per cell (Nam et al., 2010). Therefore, less than 1000 embryos should be sufficient to analyze ~1000 CRMs. Similar transgenesis methods that causes mosaic DNA incorporation are already proven to work in other systems such as zebrafish embryos (e.g., Smith et al., 2013b) and tunicate embryos (e.g., Stolfi and Christiaen, 2012). In addition, the new method can also be applied to other systems such as frog embryos (Chesneau et al., 2008) and chicken embryos (Betancur et al., 2010) that also result in DNA mosaicism upon transgenesis.

4.4. Prospect for predicting of spatial activities of new CRMs by profile matching

A large number of cell-type specific CRMs can be a valuable resource to computationally examine the extent of known regulatory inputs and also to predict new regulatory logics. As in the case of sequence homology search, our new method can be applied to massively predict spatial patterns of new CRMs by matching their ROPs to those of known spatial patterns. The spatial resolution of this approach will in part depend on the number of reference profiles. Currently, there are ≤30 well-characterized sea urchin CRMs that represent different embryonic territories covering all three germ layers of the embryo. Because cell lineages leading to three primary germ layers are asymmetric to each other, we anticipate that MMOSAIC can be used for rapidly identifying CRMs that are specific to endoderm, mesoderm, and ectoderm cells, even with these relatively small number of reference CRMs. Once ROP of a CRM has been established, these profiles can be re-analyzed as the number of reference profiles increases, further enhancing its resolution without additional experiments.

4.5. Application to gene regulatory network analysis

The most significant advantage of the new method over earlier multiplex approaches and imaging-based approaches is that one can now simultaneously detect both quantitative and spatial responses of many CRMs to gene perturbations. This aspect is very useful, because spatial activity of a CRM can change without affecting overall quantitative activity, and these changes will remain undetected in earlier multiplex methods. Because a typical eukaryotic gene is controlled by multiple CRMs, directly measuring the causal effect of gene perturbations on individual CRMs will enhance the resolution of gene regulatory network models. This advantage will be most obvious for the cases where two or more CRMs function as "OR" logic and, thus causality is difficult to detect by measuring gene expression alone. Identification of causal regulatory inputs onto individual CRMs will lead to more comprehensive models of gene networks, beyond the resolution of approaches that rely on gene expression changes. As it has been demonstrated in previous studies (e.g., Bothma et al., 2015; Delpretti et al., 2013), comprehensive information on all relevant individual CRMs for a gene of interest will ultimately lead to the understanding of CRM function in the genomic context.

Author contributions

CLG, Acquisition of experimental data, Drafting of the manuscript; STM, Acquisition of simulated data, Drafting of the manuscript; JN, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting of the manuscript.

Acknowledgements

The authors appreciate Kevin Abbey for maintaining cluster server; Pat Leahy at Caltech's Kerckhoff Marine Lab for collecting sea urchins; Rugers Wakman Genomics Core Facility for sequencing; Kyle Kendzierski and Brianna Polk for maintaining sea urchin aquarium; and Kwangwon Lee, Eric Klein, and Darius Balciunas for their comments on an earlier version of the manuscript. We also thank two anonymous reviewers for their suggestions on an earlier version of the manuscript. This work was supported by Rutgers Start-up to JN.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ydbio.2017.01.010.

References

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., Stark, A., 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074–1077.

Arnone, M.I., Dmochowski, I.J., Gache, C., 2004. Using reporter genes to study cis-regulatory elements. In: Ettensohn, C.A., Wessel, G.M., Wray, G.A. (Eds.), Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental approaches 74. Elservier Academic Press, San Diego, 621–652.

Betancur, P., Bronner-Fraser, M., Sauka-Spengler, T., 2010. Genomic code for Sox10 activation reveals a key regulatory enhancer for cranial neural crest. Proc. Natl. Acad. Sci. USA 107, 3570–3575.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bothma, J.P., Garcia, H.G., Ng, S., Perry, M.W., Gregor, T., Levine, M., 2015. Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. eLife 4.

Buecker, C., Wysocka, J., 2012. Enhancers as information integration hubs in development: lessons from genomics. Trends Genet. 28, 276–284.
Chesneau, A., et al., 2008. Transgenesis procedures in Xenopus. Biol. Cell. 100, 503–521.
Damle, S., Davidson, E.H., 2011. Precise cis-regulatory control of spatial and temporal expression of the alx-1 gene in the skeletogenic lineage of s. purpuratus. Dev. Biol. 357, 505–517.
Davidson, E.H., 2006. The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Academic Press/Elservier, San Diego, CA.
Delpretti, S., Montavon, T., Leleu, M., Joye, E., Tzika, A., Milinkovitch, M., Duboule, D., 2013. Multiple enhancers regulate Hoxd genes and the Hotdog LncRNA during cecum budding. Cell Rep. 5, 137–150.
Duboc, V., Rottinger, E., Besnardeau, L., Lepage, T., 2004. Nodal and BMP2/4 signaling organizes the oral-aboral axis of the Sea Urchin embryo. Dev. Cell. 6, 397–410.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat. Sci. 1, 54–75.
Flytzanis, C.N., Britten, R.J., Davidson, E.H., 1987. Ontogenic activation of a fusion gene introduced into sea urchin eggs. Proc. Natl. Acad. Sci. USA 84, 151–155.
Flytzanis, C.N., McMahon, A.P., Hough-Evans, B.R., Katula, K.S., Britten, R.J., Davidson, E.H., 1985. Persistence and integration of cloned DNA in postembryonic sea urchins. Dev. Biol. 108, 431–442.
Gibson, D.G., 2011. Enzymatic assembly of overlapping DNA fragments. Methods Enzymol. 498, 349–361.
Gisselbrecht, S.S., 2013. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nat. Methods.
Hough-Evans, B.R., Britten, R.J., Davidson, E.H., 1988. Mosaic incorporation and regulated expression of an exogenous gene in the sea urchin embryo. Dev. Biol. 129, 198–208.
Juven-Gershon, T., Kadonaga, J.T., 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev. Biol. 339, 225–229.
Juven-Gershon, T., Cheng, S., Kadonaga, J.T., 2006. Rational design of a super core promoter that enhances gene expression. Nat. Methods 3, 917–922.
Kirchhamer, C.V., Yuh, C.H., Davidson, E.H., 1996. Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc. Natl. Acad. Sci. USA 93, 9322–9328.
Lee, P.Y., Nam, J., Davidson, E.H., 2007. Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. Dev. Biol. 307, 434–445.
Li, E., Cui, M., Peter, I.S., Davidson, E.H., 2014. Encoding regulatory state boundaries in the pregastrular oral ectoderm of the sea urchin embryo. Proc. Natl. Acad. Sci. USA 111, E906–E913.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Materna, S.C., Davidson, E.H., 2012. A comprehensive analysis of Delta signaling in pre-gastrular sea urchin embryos. Dev. Biol. 364, 77–87.
McMahon, A.P., Flytzanis, C.N., Hough-Evans, B.R., Katula, K.S., Britten, R.J., Davidson, E.H., 1985. Introduction of cloned DNA into sea urchin egg cytoplasm: replication and persistence during embryogenesis. Dev. Biol. 108, 420–430.
Melnikov, A., et al., 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. 30, 271–277.
Minokawa, T., Wikramanayake, A., Davidson, E.H., 2005. cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. Dev. Biol. 288, 545–558.
Morrill, J.B., Marcus, L., 2004. Cell lieage chart through tenth cleavage in the sea urchin Lytechinus variegatus. In: Ettensohn, C.A., Wessel, G.M., Wray, G.A. (Eds.), Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental approaches 74. Elservier, 840.
Nag, A., Narsinh, K., Kazerouninia, A., Martinson, H.G., 2006. The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. RNA 12, 1534–1544.
Nam, J., Davidson, E.H., 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. PLoS One 7, e35934.
Nam, J., Dong, P., Tarpine, R., Istrail, S., Davidson, E.H., 2010. Functional cis-regulatory genomics for systems biology. Proc. Natl. Acad. Sci. USA 107, 3930–3935.
Nam, J., Su, Y.H., Lee, P.Y., Robertson, A.J., Coffman, J.A., Davidson, E.H., 2007. Cis-regulatory control of the nodal gene, initiator of the sea urchin oral ectoderm gene network. Dev. Biol. 306, 860–869.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford Press, New York, New York.
Patwardhan, R.P., et al., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. 30, 265–270.
Range, R., Lapraz, F., Quirin, M., Marro, S., Besnardeau, L., Lepage, T., 2007. Cis-regulatory analysis of nodal and maternal control of dorsal-ventral axis formation by Univin, a TGF-beta related to Vg1. Development 134, 3649–3664.
Revilla-i-Domingo, R., Minokawa, T., Davidson, E.H., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. Dev. Biol. 274, 438–451.
Rizzo, F., Fernandez-Serra, M., Squarzoni, P., Archimandritis, A., Arnone, M.I., 2006. Identification and developmental expression of the ets gene family in the sea urchin (Strongylocentrotus purpuratus). Dev. Biol. 300, 35–48.
Schindelin, J., et al., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682.
Smith, J., Davidson, E.H., 2008. Gene regulatory network subcircuit controlling a dynamic spatial pattern of signaling in the sea urchin embryo. Proc. Natl. Acad. Sci. USA 105, 20089–20094.
Smith, J., Kraemer, E., Liu, H., Theodoris, C., Davidson, E.H., 2008. A spatially dynamic cohort of regulatory genes in the endomesodermal gene network of the sea urchin embryo. Dev. Biol. 313, 863–875.
Smith, R.P., et al., 2013b. A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. Genome Biol. 14, R72.
Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., Ahituv, N., 2013a. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet..
Stolfi, A., Christiaen, L., 2012. Genetic and genomic toolbox of the chordate Ciona intestinalis. Genetics 192, 55–66.
Su, Y.H., Li, E., Geiss, G.K., Longabaugh, W.J., Kramer, A., Davidson, E.H., 2009. A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. Dev. Biol. 329, 410–421.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Tu, Q., Brown, C.T., Davidson, E.H., Oliveri, P., 2006. Sea urchin Forkhead gene family: phylogeny and embryonic expression. Dev. Biol. 300, 49–62.
Villoutreix, P., Delile, J., Rizzi, B., Duloquin, L., Savy, T., Bourgine, P., Doursat, R., Peyriéras, N., 2016. An integrated modelling framework from cells to organism based on a cohort of digital embryos. Sci. Rep. 6, 37438.
White, M.A., Myers, C.A., Corbo, J.C., Cohen, B.A., 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. USA 110, 11952–11957.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York.
Yuh, C.H., Bolouri, H., Davidson, E.H., 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896–1902.

CHAPTER 5. DISCUSSION and FUTURE DIRECTIONS

I generated several GRAMc libraries from both human and *S. purpuratus* genomes. I also prepared small-scale subsets of the GRAMc libraries for a variety of applications. Information about the libraries, their degree of characterization and their current applications are listed in Table 2. The applications have been previously described in previous Chapters describing the results or will be discussed in this chapter.

Table 2: List of GRAMc and small-scale libraries and their applications

Library	Number of constructs / Approximate coverage	Insert to N25 ratio	Characterization status	Application
Hs800_GRAMc	~200M/ 5X	1:10	Completed (Illumina 150bp PE)	Genome-scale identification of HepG2-CRMs; Prediction of regulatory interactions; tested in a single batch of sea urchin embryos
Hs800_300K	~300K/ NA	NA (expected 1:1)	Same as Hs800	NA
Hs800_180K	~180K/ NA	NA (expected 1:1)	Same as Hs800	Tested in sea urchin embryos
Hs800_80K	~80K/ NA	NA (expected 1:1)	Same as Hs800	TF perturbation transfection in HepG2 cells; Double barcoded (SE_Hs800)
Hs800_50K	~50K/ NA	NA (expected 1:1)	Same as Hs800	Double barcoded (SE_Hs800)
SE_Hs800	~?	NA *	Completed	Preliminary MMOSAIC testing of human fragments in sea urchin
Sp3KF_GRAMc	~24M/ 5X	1:17	Partially characterized (Ion Torrent Proton)	Tested in 3 batches of embryos at early gastrula stage (24hpf); Novel CRMs used for subsequent spatial CRM analysis
Sp3KR_GRAMc	~30M/ 5X	1:20	Partially characterized (Ion Torrent Proton)	NA
Sp800_GRAMc	~44M/ 5X	1:9	Not characterized	
SE_Sp800	?	NA *	Completed	Preliminary MMOSAIC testing in sea urchin

* Libraries were extreme barcoded for single-embryo resolution analysis

5.1 GRAMc provides reliable and abundant regulatory information

I have developed a method that reliably identifies both enhancers and promoters at genome-scale when tested in human cells. We have also shown that the method is adaptable for use *in vivo*, as demonstrated in sea urchin embryos. Although the design of GRAMc reporter constructs follows the traditional format and the methodology behind GRAMc appears simple, refinement of the GRAMc protocol required rigorous optimization. All of the optimization addressed various aspects of the same problem: high-throughput methods of analysis are inherently noisy, reducing the reliability of and the confidence in the data obtained. By process mapping and deconstruction of the overall GRAMc protocol, I identified and controlled for every optimizable step. A series of quality control measures were implemented to ensure success of the method. Based on the reproducibility of GRAMc data, the high correlation between GRAMc data and an independent reporter assay, and the consistency of GRAMc identified CRMs with several reported features of CRMs, we have a high degree of confidence in the reliability of GRAMc data.

We have demonstrated that functionally validated identification of both active and inactive genomic fragments at genome-scale provide new insights into novel potential mechanisms of regulation. Previous high-throughput analyses have excluded non-expressed factors under the logic that they play no apparent role in the cell-type of interest. Our analysis, however, underscores the potential boon from analytical inclusion of non-expressed factors. Indeed it was because of our confidence in the genome-scale data that led us to conclude that these

factors were not merely artifacts, but rather must have some biologically relevant role in the regulation of HepG2-CRMs.

The observed changes HepG2-CRMs in response to the ectopic expression of the non-expressed transcription factors PitX2 and IKZF1 is consistent with several studies on the regulatory role of these factors. In mice, the gene *pitx2* is expressed in and is required for hematopoietic function of the fetal liver, and shut down of both *pitx2* and hematopoietic function of the fetal liver is essential for the differentiation of the adult liver from the fetal liver (Kieusseian et al., 2006). Similarly, *ikzf1* is a key regulator of hematopoietic development (Davis, 2011), but its expression in the fetal liver is unknown. It is possible that *pitx2* (and/or *ikzf1* to a minor degree) keeps HepG2-CRMs repressed in the fetal liver, and clearance of *pitx2* is critical for the activation of HepG2-CRMs and gene expression in the adult liver. Our results highlight that motif enrichment analysis with large numbers of functionally validated CRMs and inactive "control" DNAs defined in one cell type can guide gene regulatory network analysis not only within that cell type, but also between different cell types and conditions, potentially providing an effective avenue to construct inter-cell type gene regulatory networks.

5.2 Future direction for analyzing the role of *pitx2* and *ikzf1* in regulating HepG2-CRMs

Our analysis indicates that Pitx2 and IKZF1 potentially share a majority of CRM targets. Given that both factors are also reported to play a role in hematopoiesis, and may potentially regulate CRMs in the same cell type, a future

direction of the work would be to test their combined regulatory effect on CRM activity when both factors are simultaneously ectopically expressed in HepG2 cells. As a further consideration, gene expression profiling by RNA-seq of these same samples may elucidate whether the co-expression of these factors can drive HepG2 cells towards a more hematopoietic-like, liver regenerative-like, or fetal liver-like state. This could be done using distance-based clustering of the gene-expression profile in perturbed HepG2 cells with gene-expression profiles contained in Gene Expression Omnibus (GEO) or LifeMap® Discovery database. Similar to the analysis reported in (Nam and Davidson, 2012), overlaying CRM and nearby gene expression responses to TF perturbations in HepG2 cells may provide an effective strategy for interrogating GRN.

5.3 Future application of GRAMc identified HepG2-CRMs: toxicity associated biomarker discovery

Changes in gene expressions in response to drug treatments are causally mediated by cis-regulatory modules (CRMs) such as enhancers and promoters (Luizon and Ahituv, 2015). There are a several well-known CRMs such as the promoters of cytochrome P450 genes that serve as biomarkers for drugs' cellular toxicity. Whereas subtle toxicological effects may not be apparent at level of gene expression changes, the use of a catalogue of drug responsive liver CRMs may enable detection of subtle indications of toxicity. Therefore a future application the GRAMc in human cells is to test the ability of a subset of the Hs800_GRAMc library to identify CRMs that respond to known liver-toxic drugs. To accomplish this, we have enriched the Hs800_GRAMc library for CRMs using

an RNA probe set against ~60,000 fragments with CRM activity of ≥3-fold over background and ~1,000 inactive fragments as an internal control. We will test the enriched library in batches of HepG2 cells using drug treatments with known Cholestatic toxicity (cyclosporin A and chlorpromazine) and known Hepatocellular toxicity (acetaminophen and tetracycline) along with non-toxic controls (adefovir and clonidine). We will identify sets of CRMs that respond differentially to the distinct types of drug toxicity as well as a set of CRMs that respond to both types of drug toxicity.

5.4 Future direction for GRAMc in *S.purpuratus*

To further capitalize upon the CRMs contained in the Sp3K_GRAMc libraries, an increased depth in characterization is required. In addition, more finely mapped CRM activity in sea urchin may be achieved through more thorough characterization and testing of the Sp800_GRAMc library in sea urchin embryos. Currently, only a fraction of the Sp800_GRAMc library has been characterized following subsequent double barcoding of reporter constructs (the original barcode serves as the ID) for the purpose of further analysis using MMOSAIC.

Preliminarily, we tested the Sp3KF_GRAMc library in *S. purpuratus* embryos at 24hpf. In this organism, 24hpf is the onset of gastrulation. For Deuterstomes such as sea urchins, this stage marks the formation of the trilaminar embryo, containing the endoderm, mesoderm, and ectoderm, from totipotent cells. At this stage, embryonic gene regulatory programs are well-established, and analysis of gene expression is not confounded by the

persistence of maternal factors. With the ensuing ingression of cells and epithelial to mesenchymal cell transitions, the onset of gastrulation marks the beginning of a dramatic shift in regulatory programs. Following the transition from totipotency, cells begin to acquire increasingly more specified regulatory states. For this reason, stage-specific genome-scale identification of CRMs at mid- and late- gastrulation would enrich the temporal analysis of CRM activity.

Preliminarily, we tested the SE_Sp800 library at the onset of gastrulation (24hpf) and in late gastrula stage embryos (48hpf). For earlier stage embryos, 4078 out of 38375 tested fragments that passed read filtering had an activity of ≥3-fold of BG, but for later stage embryos, 2885 out of 39738 fragments had an activity of ≥3-fold of BG. This decrease in the proportion of active fragments between 24hpf embryos and 48hpf embryos, may indicate that a stage specific genome-scale profile of CRM activity may facilitate the prioritization of library fragments for subsequent spatial analysis.

5.5 Future direction and application of MMOSAIC

We tested the potential to perform trans-species spatial characterization of CRMs by MMOSAIC in ~2000 early gastrula stage sea urchin embryos (24hpf) using ~1000 sea urchin fragments randomly selected from the Sp800_GRAMc library and ~1000 human fragments randomly selected from the Hs800_GRAMc library. Of the fragments tested, 35 urchin fragments and 25 human fragments had an activity of >3-fold over background. Both urchin and human fragments displayed spatially restricted activities (Supp. 5). This preliminary test addressed

the potential to perform high-throughput CRM analysis for other species using sea urchin as a test bed.

Reciprocally, in collaboration with Darius Balciunas's lab, we preliminarily tested the applicability of MMOSAIC in other organisms using the single embryo library of *S. purpuratus* reference CRMs tested in the zebrafish, *Danio rerio*. Embryos sampled at 30% Epiboly, 70% Epiboly, and 24hpf showed an increasing degree of spatial restriction of sea urchin CRMs. As a model of vertebrate development that has been utilized for functional genomics studies pertaining to human genetic neurological disorders, trans-species analysis of human, zebrafish, and sea urchin CRMs tested in both *D. rerio* and *S. purupatus* embryos could provide a functional means to study the evolution of complex genetic disease. For this reason, both GRAMc and subsequent MMOSAIC analysis both in sea urchin and zebrafish are attractive future directions for the work.

5.5 Concluding statements

I have developed three major tools to aid studies in functional genomics: a genome-scale reporter method for use in cultured cells and in embryos, a method to characterize genome-scale libraries that harbor large inserts, and a multiplex method for high-throughput spatial *cis*-regulatory analysis. These methods provide a vast resource to address questions pertaining to development, evolution, and human health.

CHAPTER 6. MATERIALS and METHODS

6.1 General molecular methods

All PCR reactions were done using Q5 polymerase in a standard 20uL reaction as per the manufacturer's specifications (NEB M0491) unless otherwise noted.

All gel purifications were done using the ZymoClean Gel DNA recovery kit. All column purifications of DNA were done using the Zymo-6 column (which is the same as the Zymo IC column) unless otherwise noted. All elutions were done using nuclease free water.

All primers were purchased from Integrated DNA Technologies and their sequences are contained in Supp. 2. Primers for the published work (Chapter 4) are included in Supp. 6.

6.2 QPCR

All qPCR was performed using the Quant Studio 6 Flex system (ThermoFisher) using the 384-well plate format. Primers for qPCR were designed using Primer3 software (Koressaar and Remm, 2007; Untergasser et al., 2012) with an amplicon size parameter of ~120bp and an optimal Tm of 59C. All QPCR was performed using either Power Sybr Green 2X Master Mix () or HOTfire Evagreen Reagent () in a final reaction volume of 10uL.

6.3 Reverse transcription

All cDNA was generated from total RNA using the High Capacity cDNA Synthesis Kit (ThermoFisher)

6.3 AD4 adapter preparation

An AD4 double-adapter was prepared as follows: p-AD4_F and p-AD4_R oligos were prepared at a concentration of 4pmol/uL in 1x T4 ligation buffer and placed on a block heated to 95C. The block was allowed to cool to 25C over the course of an hour. Annealed adapters were either used fresh, or aliquoted and stored at -80C, and thawed on ice immediately prior to ligation.

6.4 GRAMc vector preparation

The GRAMc vector was constructed by cloning the Super Core Promoter (Juven-Gershon et al.) upstream of the GFP ORF in the pGEMT easy vector by inverse PCR. The vector was linearized by AflII/HindIII overnight digestion and amplified using 10 cycles of PCR as two separate pieces each from 20ng of vector template for subsequent three-piece linear Gibson assembly as follows: the SCP-GFP cassette was amplified using primers NJ-95 and NJ-145, and the vector backbone was amplified with NJ-146 and NJ-96 using an annealing temperature of 62C with a 2 minute extension. A sequence of six phosphorothioated bases are included at the 5' end of the NJ-145 and NJ-146 primers to prevent loss of primer sites due to nuclease activity during Gibson Assembly.

6.5 Construction of the Hs800_GRAMc library

6.5.1 Preparation of human genomic inserts for cloning into the GRAMc vector

NG16408 genomic DNA was obtained from Coriell Institute. This DNA has a normal karyotype and originates from an apparently healthy male. For generation of genomic inserts, 20ug of NG16408 DNA was sonicated in 200uL of

water using a XXXX sonicator at 20% Amperage with 3 cycles of 15s pulses and 10s off. DNA was column cleaned using a Zymo-25 column, and ~800bp fragments were size selected using gel electrophoresis on a 1.2% agarose gel, purified using the Zymo-5 gel purification kit, and an aliquot was size confirmed on a 2% Agarose E-gel (ThermoFisher G501802). The remaining purified fragments were repaired in a 25uL PreCR reaction (NEB M0309) containing 1X Thermopol Buffer, 100uM dNTPs, 1X NAD+, and 0.5uL of PreCR enzyme for 30 minutes at 37C. PreCR treated fragments were column purified using a Zymo-6 column and subjected to a half reaction of the End Repair/dA Tailing Module (NEB) followed by a half reaction of the TA Ligation Module (NEB ?) using a 10:1 adapter to insert molar ratio of the annealed AD4 double adapter. Unligated adapters and genomic inserts were removed with 20U each of Exonuclease I (NEB M0293) and Exonuclease III (NEB M0206) in a 50uL reaction that was supplemented to 1X with Cutsmart buffer. Ligates were column cleaned (Zymo-6) and then linearized with 15U of RNase HII (NEB M0288) in a 30uL reaction in 1X Thermopol buffer for 90 minutes at 37C. Linearized inserts were purified using 20uL of Axygen AxyMag beads (cat?), supplemented to a final concentration of 17% PEG and 10mM $MgCl_2$ followed by 3 washes with 70% ethanol and elution in 30uL of water.

6.5.2 Coverage estimation

To determine the amount of template inserts that represent 1X genomic coverage, dilutions of 0.5 ng/ul, 0.25 ng/ul, 0.1 ng/ul, 0.05 ng/ul, and 0.025ng/uL of inserts were made. The optimum cycle number for PCR amplification of 0.5ng

of inserts using primers NJ-213 and NJ-214 was determined to be 16 cycles by a cycle test, with an annealing temperature of 61C and an extension time of 1 minute. Cycle numbers were increased accordingly for each serial dilution. Amplifications of each dilution were bead cleaned as described previously, and 8ng/well of each were loaded in duplicate along with 8ng/well of NG16408 stock DNA for qPCR against the following single copy targets: ACTA1, ADM, ADAM12, AXL-2, CFB, DLX5, Kiss1, NCOA6, Notch2, RPP30, and TOP1. Increase in cycle number for each target compared to the unprocessed genomic DNA was computed, and targets that displayed a cycle number increase of greater than 5 cycles were counted as absent.

The proportion of targets present by QPCR were compared to the Poisson probability model of randomly selecting a genomic insert at random from a pool of genomic inserts of a given coverage, ***X***. This probability can be calculated using the equation:

$$\boldsymbol{P} = 1 - (1 - \boldsymbol{p})^{\boldsymbol{XN}}$$

in which,

$$\boldsymbol{p} = \frac{insert\ size}{genome\ size}$$

and ***N*** is the number of partitions of the genome for the given insert size. Based on this model, *P* = 0.63. Of the dilutions, the 0.1ng contained 6 of the 11 targets or a proportion of 0.545, representing between 0.5X and 1X coverage. In this manner, 0.2ng of inserts were determined to represent ~1X genomic coverage. Equimolar amounts of independently amplified replicates were mixed to obtain a

pool of inserts at 5X genomic coverage.

6.5.3 Gibson assembly of the Hs800_GRAMc library

A 30ng portion of the 5X genomic insert mix was combined in a 1:1:1 molar ratio with the prepared GRAMc vector backbone (100ng) and the prepared SCP-GFP cassette (30ng) (see section 6.2 for details on vector preparation). The DNA was mixed 1:1 with NEBuilder HiFi Assembly MasterMix (NEB E2621) to a final volume of 16uL and was incubated at 50C for 20 minutes. The assembled DNA was column cleaned and eluted in 20uL of water.

To prepare the assembled library for barcoding, 8ng of the purified assembly was amplified in 9 cycles of PCR, as determined by a cycle test, with primers NJ-101 and NJ-126 using an annealing temperature of 62C and a 5 minute extension time. The amplified assembly was column cleaned. In this way, the low cycle number minimizes the potential for the PCR-based mutations while generating multiple copies of each assembled fragment for testing each fragment with multiple barcodes.

6.5.4 Barcoding and isolation of the Hs800_GRAMc library

To add barcodes to the amplified assembled library, 150ng of DNA was subjected to 1 cycle of PCR with a biotinylated, random 25-base barcoding primer, NJ-127, in a 50uL Q5 reaction with an annealing temperature of 60C for 40 seconds and an extension time of 15 minutes. A second primer, NJ-126, is included in the reaction, but merely serves as a competitor to reduce the potential for template switching that may result in erroneous and confounding pairing of barcodes. Primers were removed using 50uL of Axygen AxyMag

beads, supplemented to a final concentration of 17%PEG 8000 and 10mM $MgCl_2$ followed by 3 washes with 70% ethanol and elution in 20uL of water.

To isolate the barcoded library, the AxyMag cleaned DNA was mixed with an equal volume of 2X washing/binding buffer (2M NaCl, 10mM Tris, 1mM EDTA) and added to 20uL of Dynabead MyOne Streptavidin C1 beads (ThermoFisher 65002) that were prepared by 3 washes with 100uL of 2X washing/binding buffer. Binding was performed by incubation for 20 minutes at room temperature with tapping to mix in 2 minute intervals. The bound library was washed with 3 times with 100uL of 1X washing/ binding buffer (1M NaCl, 5mM Tris, 0.5mM EDTA). Beads were quickly washed once with 20uL of water to remove traces of EDTA and then resuspended in 50uL of nuclease free water and quantified using Qubit dsDNA dye.

To determine the minimum cycles of PCR required for library amplification, 1ul of the resuspended C1 beads were cycle tested with primers NJ-128 and NJ-129 with an annealing temperature of 61C and an extension time of 5 minutes. Since heat elution was not used to release the library from the C1 beads prior to PCR amplification, 1uL is the maximum volume that can be used in a 20uL PCR reaction. Above this volume, C1 beads inhibit PCR. To prepare the barcoded library, 24 replicate Q5 reactions containing 1uL of template were amplified in 9-cycles with primers NJ-128 and NJ-129 with an annealing temperature of 61C and an extension time of 5 minutes. Replicates were combined, AxyMag bead cleaned, gel isolated from a 1% agarose gel using a zymo column, and further

bead cleaned with AxyMag since guanidinium salts used during gel extraction can inhibit the subsequent ligation.

6.5.5 Low concentration self-ligation and final preparation of the Hs800_GRAMc library

To reduce the potential formation of concatenates that could lead to mis-pairing or ambiguities during barcode to insert association, the library is self-ligated at an extremely low concentration. To achieve this, 125ng of the final purified library is prepared for ligation in a 600uL ligation reaction using 1X standard T4 ligase buffer (NEB B0202) and 7ul of high concentration T4 ligase (NEB M0202T). The enzyme is added last, after the DNA is well dispersed, and the ligation is incubated for 4hours at room temperature. Following ligations, the reaction was supplemented with 57uL of Lambda exonuclease buffer and 30U each of Exonuclease I and Lambda exonuclease (NEB M0262) are added directly to the reaction for 1hour at 37C. At the end of the digestion, the reaction mixture is spiked with 1uL of ProteinaseK (ThermoFisher ?) for 15 minutes at 37C. The circularized library is AxyMag bead cleaned using 25uL of beads supplemented to a final concentration of 15% PEG 8000 and 10mM $MgCl_2$, followed by 4 washes with 70% ethanol and elution in 6.5uL of water. Qubit BR dye was used to quantify the amount of library 0.5uL eluate.

6.5.6 Growing and estimating the size of the Hs800_GRAMc library

Duplicate electroporations of 25uL of ElectroMAX DH10B competent cells (ThermoFisher 18290015) were each done using 30ng of library ligates (12ng/uL). Each replicated was resuspended into 1ml of SOC media immediately

following electroporation and then combined. A 1uL aliquot from the 2ml of combined library was removed and used to make serial dilutions of the transformed library. A portion of these dilutions were plated, representing 1/200,000, 1/2,000,000. 1/20,000,000, and 1/200,000,000 of the total library, respectively. Based on colony counts for the respective plates, the Hs800_GRAMc library was estimated to contain ~200 million unique constructs.

The remaining transformed library was immediately used to inoculate 180ml of LB containing 100ug/ml Ampicillin without recovery. The culture was grown using an orbital shaking incubator set at 37C and 250 RPM for 16 hours. A 30ml portion of the bacterially amplified library was mixed with 20ml a 50% Glycerol solution and frozen as aliquots. The remaining 150ml of culture were processed using the ZymoPure Plasmid Maxiprep Kit (Zymo ?).

To expand the library further, 5ml of bacterial glycerol stock was used to inoculate 1L of LB containing 100ug/ml Ampicillin. Library plasmids were purified using the ZymoPure Plasmid Gigaprep Kit. This preparation of the library was not used for the current analysis, and it should be noted that the coverage of the input library may vary from that of the original maxiprepped Hs800_GRAMc library.

6.5.7 Characterization of the Hs800_GRAMc library by paired-end sequencing

To prepare the Hs800_GRAMc library for characterization, 5ng of the plasmid library were amplified by inverse PCR with the phosphorylated primers NJ-209 and NJ-141 to remove the GFP ORF and abut the 3’ end of genomic

inserts with the 5’ end of N25 barcode sequences (denoted as “Hs800_23”). A separate inverse PCR amplification was done from 5ng of the plasmid library with the phosphorylated primers NJ-208 and NJ-142 to remove the vector backbone and abut the 5’ end of genomic inserts with the 3’ end of N25 barcode sequences (denoted as “Hs800_14”). A total of 20 replicates for each of the two primer pairs were prepared using an annealing temperature of 60C with an extension time of 3.5minutes for a total of 10 cycles. Replicates for each of the two amplifications were combined, column concentrated, gel isolated and AxyMag bead cleaned as described previously.

Each of the two purified amplifications was self-ligated at a concentration of 75ng in a total ligation volume of 350uL, containing 1x T4 ligase buffer and 3uL of NEB high concentration ligase, overnight at room temperature. To remove unligated DNA, ligations were supplemented with 20U of both Exonuclease I and Exonuclease III at 37C for 1hr followed by incubation with Proteinase K for 10 minutes at 37C. AxyMag bead cleaning was performed as described and ligates were eluted in 30uL of water.

To amplify insert::N25 cassettes from the circularized first round PCR products, the 4 replicates containing 2ng of Hs800_14 ligates were amplified using NJ-209 and NJ141 (now denoted as Hs800_1423), and 4 replicates containing 2ng Hs800_23 ligates were amplified using NJ-208 and NJ142 (now denoted as Hs800_2314) with an annealing temperature of 60C and an extension time of 90 seconds for a total of 8 cycles. Products were column

cleaned, gel isolated, and bead-cleaned for subsequent PCR amplification to add PE adapter sequences for Illumina sequencing.

The Illumina sequencing platform is not suitable for sequencing low diversity libraries that contain excess adapter sequences such as the Hs800_GRAMc library. To increase diversity of the Hs800_1423 and Hs800_2314 sequencing libraries for sequencing on the Illumina NextSeq platform. Each library was amplified using 7 different, out of phase PE1 containing primers. For the Hs800_1423 library, 2ng of template were used per each separate reaction with the PE2 containing primer NJ-401 and each of the following partial PE1 containing primers: NJ-400, NJ-504, NJ-505, NJ-506, NJ-507, NJ-508, and NJ-509 with an annealing temperature of 60C and an extension time of 90 seconds for a total of 7 cycles. For the Hs800_2314 library, 2ng of template were used per each separate reaction with the PE2 containing primer NJ-403 and each of the following partial PE1 containing primers: NJ-402, NJ-498, NJ-499, NJ-500, NJ-501, NJ-502, and NJ-503 with an annealing temperature of 60C and an extension time of 90 seconds for a total of 7 cycles. Individual amplifications were column cleaned, gel isolated, and AxyMag bead cleaned. Each of the 7 out-of-phase Hs800_1423 libraries was amplified using NJ-497 and NJ-401 to complete the PE1 adapter sequence. Each of the 7 out-of-phase Hs800_2314 libraries were amplified using NJ-497 and NJ-403 to complete the PE1 adapter sequence. For each amplification, 2ng of respective library templates were amplified in 6 cycles of PCR with an annealing temperature of 60C and an extension time of 90 seconds. Libraries were again

purified, gel isolated, and AxyMag bead cleaned. Equimolar amounts of the 14 out-of-phase libraries were combined and submitted to the Waksman Institutes Genomics Core Facility.

6.6 Preparation of random sub-sets of the Hs800_GRAMc library

To obtain small-scale subsets of the GRAMc library, an ~50uL chunk of frozen glycerol stock was diluted into 2ml of LB media, recovered with shaking at 250 RPM in an orbital incubator at 37C for 20 minutes. A series of 2-fold dilutions were made, 10uL of which was used for 2 10-fold dilutions for plating, and the remaining 1ml of which was used to seed 150ml LB-Amp cultures for overnight growth as described previously. Plated colonies for each serial dilution were used to estimate the number of constructs in each random sub-set of the Hs800_GRAMc library. Cultures that were estimated to contain ~300K colonies, ~180K colonies, ~80K colonies, and ~35K colonies were processed using the ZymoPure Plasmid Maxiprep Kit.

6.7 Construction of the Sp3KF and Sp3KR_GRAMc libraries

The AflII/HindIII digested GRAMc vector (described previously) was prepare for both forward and reverse cloning of *S. purpuratus* genomic inserts. For forward cloning, the vector backbone and SCP-GFP cassettes were amplified as described for the Hs800_GRAMc library. For preparation of the reverse vector, NJ-97 and NJ-145 were substituted to amplify the SCP-GFP cassette, and NJ98 and NJ146 were substituted to amplify the vector backbone. Annealing and extension conditions remained the same.

6.7.1 Preparation of *S. purpuratus* genomic inserts for cloning into the GRAMc vector

For generation of genomic inserts, 30ug of *S. purpuratus* genomic DNA in 200uL of water was sonicated with a single 25s pulse at 20% Amperage. Genomic fragments of ~3kb were size selected by gel isolation on a 1% agarose gel with subsequent purification as described previously. DNA was PreCR treated, dA Tailed, adapter ligated, linearized and purified as described previously.

6.7.2 Coverage estimation

Dilutions of linearized adapter ligated genomic inserts were made and amplified as described for the Hs800_GRAMc library. The amount of adapter ligated inserts representing slightly less than 1X coverage of the *S. purpuratus* genome was determined to be 0.1ng by qPCR for the presence or absence of the following single copy targets: BMP2/4, Delta, FoxA, FoxQ2, Lefty, Lim1, Nodal, Tgif, Ubq, and Univin. Genomic DNA and amplified dilutions of inserts were loaded as duplicates of 2ng/well for each qPCR reaction. Because the 0.1ng/uL dilution of inserts represents slightly less than 1X coverage, equimolar amounts of 7 replicates amplified in 16 cycles of PCR with conditions described previously were mixed to obtain ~5X coverage.

6.7.3 Gibson assembly of the Sp3K_GRAMc libraries

Gibson assembly was performed as described previously, but using 200ng of inserts with 200ng of either the forward or the reverse orientation vector backbone and 60ng of the respective forward or reverse orientation SCP-GFP

cassette. To prepare the assembled library for barcoding, 10ng of the purified assembly was amplified in 12 cycles of PCR, as determined by a cycle test, with primers NJ-101 and NJ-126 using an annealing temperature of 62C and an 8 minute extension time. The amplified assembly was column cleaned.

6.7.4 Barcoding and isolation of the Sp3K_GRAMc libraries

To add barcodes to the amplified assembled library, 150ng of DNA was subjected to 1 cycle of PCR with a biotinylated, random 25-base barcoding primer, NJ-127, in a 50uL Q5 reaction with an annealing temperature of 60C for 40 seconds and an extension time of 15 minutes. A second primer, NJ-126, is included in the reaction, but merely serves as a competitor to reduce the potential for template switching that may result in erroneous and confounding pairing of barcodes. Primers were removed using 50uL of Axygen AxyMag beads, supplemented to a final concentration of 17%PEG 8000 and 10mM $MgCl_2$ followed by 3 washes with 70% ethanol and elution in 20uL of water.

To isolate the barcoded library, the AxyMag cleaned DNA was mixed with an equal volume of 2X washing/binding buffer (2M NaCl, 10mM Tris, 1mM EDTA) and added to 20uL of Dynabead MyOne Streptavidin C1 beads (ThermoFisher 65002) that were prepared by 3 washes with 100uL of 2X washing/binding buffer. Binding was performed by incubation for 20 minutes at room temperature with tapping to mix in 2 minute intervals. The bound library was washed with 3 times with 100uL of 1X washing/ binding buffer (1M NaCl, 5mM Tris, 0.5mM EDTA). Beads were quickly washed once with 20uL of water to

remove traces of EDTA and then resuspended in 50uL of nuclease free water and quantified using Qubit dsDNA dye.

To determine the minimum cycles of PCR required for library amplification, 1ul of the resuspended C1 beads were cycle tested with primers NJ-128 and NJ-129 with an annealing temperature of 61C and an extension time of 8 minutes. For PCR amplification, 16 replicates with 1.5ul of resuspended C1 beads as templates were prepared using primers NJ-128 and NJ-129 with an annealing temperature of 61C and an extension time of 8 minutes for a total of 7 rounds of amplification. Barcoded libraries were column concentrated, gel purified, and AxyMag bead cleaned.

6.7.5 Low concentration self-ligation of the Sp3K_GRAMc libraries

Sp3KF_GRAMc and Sp3KF_GRAMc libraries were each self-ligated at a concentration of 600ng in 500uL of 1X ligase buffer supplemented to 7.5%PEG 8000 and 25uL of NEB high concentration ligase for 4 hours at room temperature. Ligations were supplemented with 57uL of Lambda exonuclease buffer, 30U each of Exonuclease I and Lambda Exonuclease for 1 hour at 37C before adding 3uL of ProteinaseK solution for 15 minutes at 37C. The libraries were AxyMag bead cleaned as described previously.

6.7.6 Growing and estimating the size of the Sp3K_GRAMc libraries

Between 70-90ng of each library was used to transform 100ul of ElectroMAX DH10B competent cells. Dilution plating and seeding of the libraries was performed as described. The Sp3KF_GRAMc library was estimated to contain ~24 million colonies and the Sp3KR_GRAMc library was estimated to

contain ~30 million colonies. Cultures were grown overnight and plasmids were prepared as described for the Hs800_GRAMc library. Glycerol stocks of each library were also prepared.

6.7.7 Characterization of the Sp3K_GRAMc libraries

To generate mate-pair sequencing libraries for the Sp3K_GRAMc libraries, 1ug of each library was cut with Cas9 (NEB) that was pre-incubated for 15 minutes at room temperature with 10uL of sgRNAs designed against either the vector backbone or the GFP ORF. CRISPR linearization of the library was carried out for 90 minutes at 37C in a total reaction volume of 200uL. The reaction was column cleaned and 10ng of the linearized templates were used 4 replicate inverse PCR amplification as follows: Sp3KF_back bone cut was amplified with NJ-142 and NJ-208 to remove the vector backbone (Sp3KF_14); Sp3KF_GFP cut was amplified with NJ-209 and NJ-141 to remove the GFP ORF (Sp3KF_23); Sp3KR_back bone cut was amplified with NJ-142 and NJ-209 to remove the vector backbone (Sp3KR_24); Sp3KR_GFP cut was amplified with NJ-208 and NJ-141 to remove the GFP ORF (Sp3KR_13). An annealing temperature of 58C and an extension time of 8 minutes were used in a total of 5 cycles. The replicates column cleaned and gel isolated.

Purified libraries were self-ligated at a concentration of 100ng/125uL to mate inserts with N25 barcodes as described for the Hs800_GRAMc library. Unligated DNA was removed by digestion with ExoI and ExoIII as described for the Hs800_GRAMc library. Ligates were amplified in 5 replicates Q5 reactions each from 4ng of template in a total of 14 cycles as determined by test PCR.

Second round amplifications were performed substituting biotinylated versions of the primers flanking the N25 barcodes as follows: Sp3KF_14 was amplified with NJ-209 and NJ-143 (Sp3KF_1423); Sp3KF_23 was amplified with NJ-144 and NJ-208 (Sp3KF_2314); Sp3KR_24 was amplified with NJ-143 and NJ-208 (Sp3KR_2413); Sp3KR_13 was amplified with NJ-209 and NJ-144 (Sp3KR_1324). Sp3KF_1423 and Sp3KR_2413 were used to generate “5N25” mate-pair libraries and Sp3KF_2314 and Sp3KR_1324 were used to generate “3N25” mate-pair libraries.

To prepare mate-pair libraries for Vectorette PCR (Arnold and Hodgson, 1991), 1ug of each library was sonicated in 200uL of water for 13.5 minutes at 30% amplitude with a pulse of 20 seconds on and 10 seconds off. The fragmented mate-pair libraries were purified using 50 uL of AxyMag beads supplemented to a final concentration of 17% PEG and 10mM MgCl2. DNA was eluted in water and repaired using a 25uL PreCR reaction as described previously. DNA was size-selected for ~200-250bp by isolation from a 2% agarose gel. Purified DNA was Endrepaired/dAtailed using a half reaction as described for genome-scale library building. End repaired was followed with a half reaction of TA ligation (NEB) with 45pmol (1uL) of annealed Vect53/Vect57 adapters (NJ-149/NJ-150, annealed as described for the AD4 double adapter). Fragments containing N25 barcodes were isolated using 20uL of Dynabeads MyOne Streptavidin C1 beads followed by heat elution in 20uL of water.

To amplify the vectorette ligated 3N25 and 5N25 libraries for sequencing, primers NJ-300 and NJ-141 were used for the Sp3KF_1423 library and the

Sp3KR_2413 library (5N25 libraries), and primers NJ-300 and NJ-142 were used for the Sp3KF_2314 library and the Sp3KR_1324 library (3N25 libraries). Each entire library was amplified in a 50uL Q5 reaction, with one round of amplification using an annealing temperature of 65C with a 20 second extension, followed by 9 cycles of amplification with an annealing temperature of 60C and an extension time of 20 seconds. Libraries were column purified, eluted in 20uL, and ~5ng (0.5-1uL) of each library was amplified with 6 cycles of PCR with Ion A and Ion P containing primers using an annealing temperature of 62C and an extension time of 20s. The primers used were as follows: for Sp3KF_2314 NJ-200 and NJ-205; for Sp3KF_1423, NJ-201 and NJ-205; for Sp3KR_1324, NJ-130 and NJ-205; for Sp3KR_2413, NJ-131 and NJ-205.

The Sp3KF bi-directional mate-pair libraries (5N25 and 3N25) sequenced on 3 Ion Torrent Proton 200v3 chips using 2uL, 4uL and 6.5uL of mixed 100pM library in 100uL of water for templating on One Touch 2. The Sp3KR bi-directional mate-pair libraries were sequenced on 1 Torrent Proton 200v3 chips using 6.5uL of mixed 100pM library in 100uL of water for templating on One Touch 2. Additionally, the Sp3KF and Sp3KR libraries were each sequenced on a Proton 200v3 chip using HiQ reagents.

6.7.8 Testing the Sp3KF_GRAMc library in *S. purpuratus*

The Sp3KF_GRAMc and Sp3KR_GRAMc libraries were linearized by cutting the vector backbone using CRISPR as described for library characterization, and 4 replicates of 10ng of each were amplified with 8 cycles of PCR using the primers NJ-208 and NJ-78 for the Sp3KF library and primers NJ-

209 and NJ-78 for the Sp3KR library with an annealing temperature of 60C and a 4.5 minute extension. The amplified libraries were column concentrated, gel isolated, and further purified for injection

Library DNA was prepared for injected in embryos as described previously (Nam et al., 2010). Injected embryos were grown at 15C and sampled at 24 hours post fertilization (hpf) and divided into three cohorts (~how many embryos each?). Total RNA and genomic DNA were extracted from the samples using the AllPrep DNA/RNA Micro kit (Qiagen) following the protocol described in (Nam et al., 2010). Total RNA was used to make cDNA using the High Capacity cDNA Reverse Transcription Kit following the manufacturers protocol with the addition of 5pmole of a GRAMc library specific RT oligo, NJ-489. Copy numbers of incorporated GFP template was measured by QPCR as described in (Nam et al., 2010; Revilla-i-Domingo et al., 2004). $1/40^{th}$ 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 in 12 cycles of PCR using a pair of universal primers, NJ-142 and NJ-264. Amplicons were prepared for Ion Torrent sequencing with the following primer pairs: cDNA 1: NJ-197/NJ-199, cDNA 2: NJ-198/NJ-199, cDNA 3: NJ-130/NJ-199, gDNA 1: NJ-132/NJ-199, gDNA 2: NJ-133/NJ-199, gDNA 3: NJ-134/NJ-199. The library was templated for sequencing of 2 Ion Torrent Proton 200v3 chips using 4ul and 7ul of 100pM mixed libraries according to the manufacturer's protocol.

6.8 Construction of the Sp800_GRAMc library

Genomic inserts for the Sp800_GRAMc library were prepared from S. purpuratus genomic DNA using the conditions described for the Hs800_GRAMc library. Coverage was determined as described for the Sp3K _GRAMc libraries, and equimolar amounts of 5 replicates amplified from 0.045ng of template were mixed to obtain 5X coverage. Inserts were assembled, amplified, barcoded, isolated, and amplified as described previously. The final self-ligation was done at a concentration of 200ng/600uL and processed as described previously. Separate electroporations of 2x25ul of Electromax DH10B were done using 5ng and 4ng of purified ligates, resulting in ~22M and ~20M colonies respectively as per colony estimation. The libraries were grown separately, but combined following plasmid maxi prep. Glycerol stocks of library cultures were stored separately.

The Sp800_GRAMc library was prepared for Illumina sequencing as described for the Hs800_GRAMc library. The prepared sequencing library is in storage at the Waksman Institute Genomics Core Facility.

6.9 Cell culture

HepG2 cells (ATCC HB-8065) were grown under supplier recommended conditions of EMEM supplemented with 10% fetal bovine serum without antibiotics. HepG2 cells were used within no more than 16 passages from receipt for all experiments. All experiments were performed in cells that underwent a minimum of 5 passages from thawing.

6.10 Testing the Hs800_GRAMc library in HepG2

6.10.1 Genome-scale transfection

For each genome-scale transfection batch, 10^7 cells were seeded in 30ml media in each of 10x150mm culture dish and allowed to attach for 30 hours. Cells were transfected with 100ug of the Hs800_GRAMc library using 100uL of DNA-IN for HepG2 reagent (MTI-Globalstem) in 4ml of Optimem (ThermoFisher) prepared according to the manufacturer's protocol.

6.10.2 Lysate collection, RNA preparation, and cDNA synthesis

Cells were washed with 1X PBS 26hours after transfection and were collected by scraping in 2.4ml RNA-STAT-60 (AMSBIO) per plate. Lysates were combined and prepared according to the manufacturer's protocol with the addition of a second ethanol wash.

Isolated total RNA was resuspended in 1.7ml of nuclease free water digested for a minimum of 4 hours at 37C in a 2ml reaction containing 1X DNase I Buffer, 100U of DNase I(NEB M0303), and 900U each of ExoI and ExoIII. The progress of DNA digestion was monitored by QPCR against GFP (NJ-443 and NJ-444). To accomplish this, a diluted sample of RNA was heat inactivated at 80C for 20 minutes and loaded at an equivalent volume of ~1000cell/well. As needed, DNase digestion was allowed to proceed overnight. Following digestion, nucleases were removed by extraction with Phenol:Chloroform:Isoamyl alcohol (25:24:1) and ethanol precipiatated overnight at -20C followed by two washes with 75% ethanol. RNA was resuspended in 1-ml of water, and a sample was removed and diluted for quality control. An equivalent volume of the total RNA containing ~4000cells (~1ug) was used to make cDNA using the High Capacity

cDNA Reverse Transcription Kit following the manufacturers protocol with the addition of 5pmole of a GRAMc library specific RT oligo, NJ-489 and used as the standard for maximum cDNA synthesis from transcripts. The remaining RNA was used as a Non-RT control for subsequent qPCR.

To process expressed barcodes, the remaining total RNA was diluted to 1.420ml, and 2000pmol of GRAMc_RT_oligo (NJ-489) was added. The RNA and primer mixture was incubated at 65C for 1 minute and then chilled on ice to reduce secondary structures in the RNA and to allow for annealing of the RT oligo. To this, 200uL of High Capacity buffer, 80uL of dNTP and 100uL of Multiscribe were added, mixed thoroughly, incubated for 10minutes at room temperature, and then incubated for 4 hours at 37C before using the equivalent volume of 100cells to monitor the progression of cDNA synthesis via qPCR against GFP in comparison to the small scale 4000C QC control. The reaction was spiked with M-MuL-V (NEB M0253)) and additional dNTPs as required and allowed to proceed overnight as required.

Upon completion of the RT reaction, the samples were ethanol precipitated to reduce the volume. RNA/cDNA was resuspended and prepared for digestion with 1000U of RNase If (NEB M0243) in a 500uL reaction with 1X NEB3 buffer at 37C overnight. For removal of excess protein, 1uL of Proteinase K solution was added to the reaction and incubated at 37C for 15 minutes. cDNA was ethanol precipitated overnight at -20C, washed 3x with 80% ethanol. cDNA pellets were resuspended in 200uL of water and heated to 95C for 10 minutes to

destroy residual Proteinase K. A sample of the cDNA library was subjected to QC by QPCR.

6.10.3 Preparation of expressed N25s for NGS

The entire pool of expressed N25s were amplified using primers NJ-141 and NJ-142 in 8 replicates of a 50ul Q5 PCR reaction using an annealing temperature of 62C and an extension time of 1minute for a total of 8 cycles. Replicates were combined for each batch. A 50uL aliquot was processed from each batch as follows. Unwanted long DNAs were bound using a 0.5X volume of AxyMag beads for 20 minutes at room temperature. The desired short amplicons (65bp) from the supernatant were further purified each batch using duplicate Zymo column and eluted each in 20uL of water.

To prepare amplicons for sequencing, 2ng of 1st round amplified and cleaned N25 barcodes were subjected to another 9 cycles of amplification with NJ-141 and NJ-142. N25 barcodes from 2ng of the input library were also amplified in 9 cycles of PCR from a mixture of uncut/CRISPR backbone cut/CRISPR GFP cut using the NJ-141 and NJ-142 primers. Amplicons were prepared both for IonTorrent sequencing (Batch 1: NJ197 and NJ-523; Batch2: NJ-198 and NJ-523) and Illumina NextSeq sequencing (Input: 14 out-of-phase libraries using NJ-400/NJ-504/NJ-505/NJ-506/NJ-507/NJ-508/NJ-509 with NJ364 or NJ-402/NJ-498/NJ-499/NJ-500/NJ-501/NJ-502/NJ-503 with NJ-399). For all of these amplifications, an annealing temperature of 65C and an extension time of 20seconds were used for a total of 6 cycles. Batch 1 and Batch 2 were also

prepared for out-of-phase Illumina NexSeq sequencing as described for the input library.

6.11 Perturbation of the Hs800_80K library

6.11.1 Small-scale transfection

Cells were seeded in duplicates of ~2M cells per 10 cm^2 plate for transfection with each of 3 co-transfections: 80K library + CMV::pitx2 (Genscript OHu17480D), 80K library + CMV::IKZF1 (Genscript OHu28016D), and 80K library + CMV::EGFP (Clontech pEGFP-C1). Cells were cultured for ~24h prior to transfection. Cells were co-transfected with 9ug of the 80K library and 3ug of the respective expression vector using 36uL of DNA-IN for HepG2 reagent (MTI-Globalstem) and 1.2ml of Optimem (ThermoFisher) prepared according to the manufacturer's protocol.

6.11.2 Lysate collection, RNA preparation, and cDNA synthesis

Cells were collected by trypsinization and washing with 1X DPBS 24h after transfection. A portion of $1/10^{th}$ of the cells was saved for Western Blot analysis to confirm expression of Pitx2 and IKZF1. The remaining cells were lysed and processed using the Zymo-Duet kit with the IIICG column for both DNA and RNA, without on-column DNaseI treatment. DNA was eluted in 100uL and RNA was eluted in 80uL and treated with DNase I (8U)/ExoI(100U)/ExoIII(100U) (NEB) for a minimum of 4hr at 37C in a total reaction volume of 100uL in 1X DNaseI buffer. Assuming ~10M cells per sample, an ~10,000 cell equivalent of gDNA and an ~5000 cell equivalent of nuclease treated RNA were tested with QPCR with GFP as target to confirm the quality of transfection and completion of

DNase digestion, respectively. Reactions were spiked with another 2U of DNaseI as needed. RNA was column cleaned using a Zymo-IIIC column and eluted in 50uL of water. An equivalent of ~4000cells was used as a measure of quality control in a standard RT reaction as described in the genome-scale protocol. The remaining RNA was incubated with 80pmole of GRAMc_RT_oligo (NJ-489) used for cDNA synthesis in an 80uL 1X High-Capacity cDNA synthesis reaction using 8uL of Multiscribe and 3.2uL of dNTP, but without the use of random primers, for 4hrs to overnight at 37C as per a quality control QPCR following 2hrs of RT. Upon completion of DNA digestion, 4uL of NEB3 and 2uL of RNase If were added to the reaction for 2hr at 37C, then spiked with Proteinase K for 15min at 37C, and heat inactivated for 10min at 95C followed by overnight ethanol precipitation and resuspension in 30uL of water.

6.11.3 Preparation of expressed N25s for NGS

N25 barcodes were preliminarily amplified as described previously, but using 6cycles of a single 50uL Q5 reaction, and IX barcoded for Ion Torrent sequencing using the following primer pairs: For control-1: NJ-197/NJ523, for control-2: NJ-198/NJ523, for Pitx2-1: NJ-199/NJ523, For Pitx2-2: NJ-132/NJ523, for IKZF1-1: NJ-133/NJ523, for IKZF1-2: NJ-134/NJ523.

6.12 Individual reporter cloning

A CRM-less N25-barcode library of the SCP-GRAMc vector was previously prepared by linearization of the GRAMc vector with AflII/HindIII followed by barcoding with NJ-127 with NJ126 as competitor, followed by addition of a polyadenylation sequence by amplification with NJ-128 and NJ-129

as described previously. Colony estimation and plasmid library preparation was done as described for the Hs800_GRAMc library, to yield an SCP1-GRAMc barcoded library of ~100M contructs.

The empty vector library was amplified using NJ-95 and NJ-96. Individual genomic regions were amplified using primers listed in Supp. 3 and Supp. 4 and cloned into the P1 and P2 sites of the vector upstream of SCP. IGV extracted genomic sequences are contained in Supp. 4.

6.13 Processing of sequencing reads and N25 normalization

The sequences of N25 barcodes from individual sequence reads were identified by trimming flanking adapter sequences: for sequences for which the nP3 site was adjacent to the PE1 or IonA primer site, GRAMcP3s-SE adapter was used; for sequences for which the P4 site was adjacent to the PE1 or IonA primer site, the GRAMcP4s-SE adapter was used. The sequences of adapters are provided as Table 4. 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.

Clusters of sequence reads from the same N25 barcodes were identified by single-linkage clustering allowing up to 2 mismatches (due to sequencing errors) in a Burrow-Wheeler Aligner search (Li and Durbin, 2009) against all identified N25 sequences. Each cluster identified in this analysis represents a unique barcode.

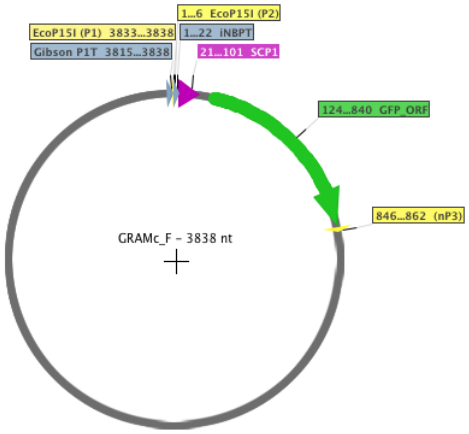
Expressed N25 barcode counts were input normalized to N25 barcode counts from the input library. Input normalized barcode read counts were further normalized to the average of the middle 30% of input normalized barcode read counts.

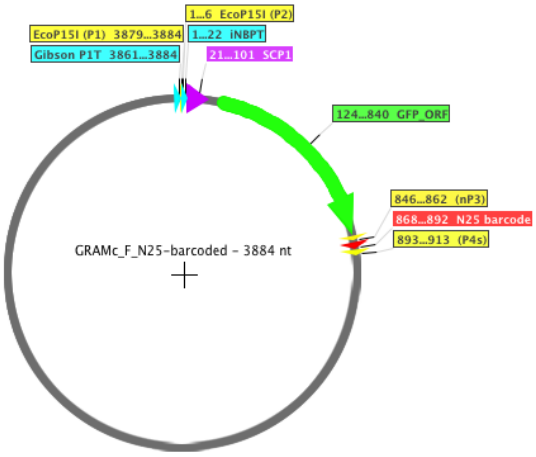
6.14 Motif enrichment analysis

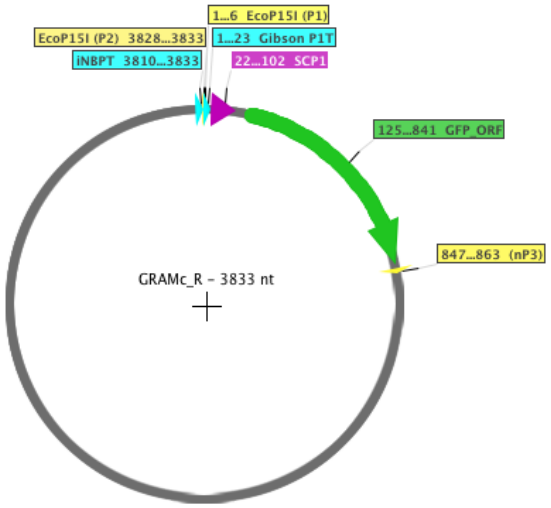
Motif enrichment analysis was performed using FIMO within the MEME suite (Grant et al., 2011).

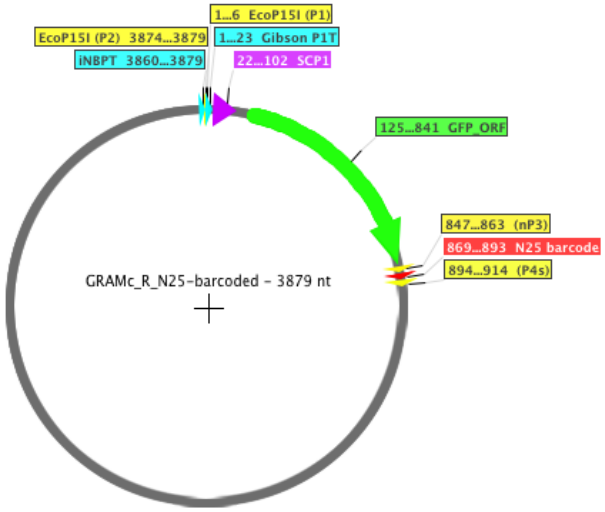
CHAPTER 7:

Supp. 1: GRAMc forward and reverse vectors






1...6 EcoP15I (P1)
EcoP15I (P2) 3828...3833
1...23 Gibson P1T
iNBPT 3810...3833
22...102 SCP1
125...841 GFP ORF
847...863 (nP3)
GRAMc_R - 3833 nt


1...6 EcoP15I (P1)
EcoP15I (P2) 3874...3879
1...23 Gibson P1T
iNBPT 3860...3879
22...102 SCP1
125...841 GFP ORF
847...863 (nP3)
869...893 N25 barcode
894...914 (P4s)
GRAMc_R_N25-barcoded - 3879 nt

Supp. 2: Primers

Primer Name	Description	Sequence
p-AD4_F	Anneal for adapter ligation	/p/CTGCTGAATCACTAGTGAATTATTACCC rUrU CAAGACACTACTCTCCAGCAGT
p-AD4_R	Anneal for adapter ligation	/p/CTGCTGGAGAGTAGTGTCTTG rArA GGGTAATAATTCACTAGTGATTCAGCAGT
NJ-78	end_core_polyA	CAC AAA CCA CAA CTA GAA TGC A
NJ-95	Amplification of SCP-GFP region of GRAMc vector or 3-piece Gibson Assembly	CTGCTGGAGAGTAGTGTCTTGtACTTATATAAGGGGGTG GG
NJ-96	Amplification of GRAMc vector backbone for 3-piece Gibson Assembly	CTGCTGAATCACTAGTGAATTCGCGG
NJ-97	P1r_SCP1_lin	CTGCTG AATCACTAGTGAATT GtACTTATATAAGGGGGTGGG
NJ-98	EcoP15I_P2r_lin	CTGCTG GAGAGTAGTGT CTT CCGCCTGCAGGTCGAC
NJ-101	Pre-barcoding library amplification	GGCGCGCCGCTGAGGGAGT
NJ-112	SpDelta_QF	ACG GAG CTA CAT GCC TGA AC
NJ-113	SpDelta_QR	TCA CAA TGG ACC GAA TCA GA
NJ-116	SpLefty_QF	CGT AGT CGC CAC ATC AGA GA
NJ-117	SpLefty_QR	CAG ATA CAT CAT GGG CAA CG
NJ-120	SpLim_QF	GTA TCC GAT CCG TTG ACG AC
NJ-121	SpLim_QR	TAG CCT TGC ATT CAC AGC AC
NJ-122	SpTgif_QF	GCT CTA CCT ATC TCG CTT GGC
NJ-123	SpTgif_QR	TGG TGA ACT TGT CAG GGT CT
NJ-124	SpFoxq2_QF	CAA GCA CCT TTT GCT CTG TGA CAT
NJ-125	SpFoxq2_QR	CTC CGC TAC GTC CAG CCT T
NJ-126	Pre-barcoding library amplification	AATTCGCCCTATAGTGAGTCGTA
NJ-127	Barcoding primer	/Biosg/TACAGTCCGACGATCCAGCAG (N:25252525)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N) (N)(N)(N)(N)(N)(N)(N)(N)(N)GGCGCGCCGCTGAGGGA GTC TAGAG
NJ-128	Post-barcode library amplification	/p/CACAAACCACAACTAGAATGCAGTGAAAAAAATGCTT TATTT GTTTACAGTCCGACGATCCAGCAG

NJ-129	Post-barcode library amplification	/p/AATTCGCCCTATAGTGAGTCGTA
NJ-130	GRAMc_Ion-A_IX5_P4s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CAG AAG GAA CGA TTA CAG TCC GAC GAT CCA GCA G
NJ-131	GRAMc_Ion-A_IX6_nP3s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTG CAA GTT CGA TTA GAC TCC CTC AGC GGC
NJ-132	GRAMc_Ion-A_IX7_P4s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTC GTG ATT CGA TTA CAG TCC GAC GAT CCA GCA G
NJ-133	GRAMc_Ion-A_IX8_P4s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TTC CGA TAA CGA TTA CAG TCC GAC GAT CCA GCA G
NJ-134	GRAMc_Ion-A_IX9_P4s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG TGA GCG GAA CGA TTA CAG TCC GAC GAT CCA GCA G
NJ-141	pGRAMc_nP3_short	/5Phos/TAGACTCCCTCAGCGGC
NJ-142	pGRAMc_P4_short	/5Phos/TACAGTCCGAcgatc CAGCAG
NJ-143	bGRAMc_nP3_s	/5BiosG/TAGACTCCCTCAGCGGC
NJ-144	bGRAMc_P4_s	/5BiosG/TACAGTCCGACGATCCAGCAG
NJ-145	Amplification of SCP-GFP region of GRAMc vector for 3-piece Gibson Assembly	G*G*C* G*C*G* CCGCTGAGGGAGT
NJ-146	Amplification of GRAMc vector backbone for 3-piece Gibson Assembly	A*A*T* T*C*G* CCCTATAGTGAGTCGTA
NJ-149	Vect53	CTCTCCCTTCTCGAATCGTAACCGTTCGTACGAGAATC GCTGTCCTCTCCTTC T
NJ-150	Vect67	GAAGGAGAGGACGCTGTCTGTCGAAGGTAAGGAACGG ACGAGAGAAGGGAGAG
NJ-153	SpUbq_QF	CACAGGCAAGACCATCACAC
NJ-154	SpUbq_QR	GAGAGAGTGCGACCATCCTC
NJ-155	SpFoxA_QF	CCAACCGACTCCGTATCATC
NJ-156	SpFoxA_QR	CGTAGCTGCTCATGCTGTGT
NJ-157	SpNodal_QF	gacaacccaagcaaccac
NJ-158	SpNodal_QR	CGCACTCCTGTACGATCATG
NJ-159	SpUnivin_QF	CAGCTCTACCATCAGCCACA
NJ-160	SpUnivin_QR	TCGAGCAATCTAAGGCGTTT
NJ-161	SpBMP2/4_QF	CCAGCAAGGTCGAAGAACTC
NJ-162	SpBMP2/4_QR	CTCTACCCGACGATGAT

NJ-179	CRSP_F_T7_backbone	TTAATACGACTCACTATAGGTCGTAGTTATCTACACGAC GGTTTTAGAGCTAGAAATAG
NJ-180	CRSP_F_T7_GFP	TTA ATA CGA CTC ACT ATA GGC GCG CTG AAG TCA AGT TCG AGT TTT AGA GCT AGA AAT AG
NJ-181	CRSP_F_T3_backbone	AAT TAA CCC TCA CTA AAG GTC GTA GTT ATC TAC ACG ACG GTT TTA GAG CTA GAA ATA G
NJ-182	CRSP_F_T3_GFP	AAT TAA CCC TCA CTA AAG GCG CGC TGA AGT CAA GTT CGA GTT TTA GAG CTA GAA ATA G
NJ-197	GRAMc_Ion-A_IX1_P4s	CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG CTA AGG TAA CGA TTA CAG TCC GAC GAT CCA GCA G
NJ-198	GRAMc_Ion-A_IX2_P4s	CCATCTCATCCCTGCGTGTCTCCGACTCAGTAAGGAGA ACGATTACAGTCCGACGATCCAGCAG
NJ-200	GRAMc_Ion-A_IX3_P4s	CCATCTCATCCCTGCGTGTCTCCGACTCAGAAGAGGAT TCGATTACAGTCCGACGATCCAGCAG
NJ-201	GRAMc_Ion-A_IX4_nP3s	CCATCTCATCCCTGCGTGTCTCCGACTCAGTACCAAGA TCGATTAGACTCCCTCAGCGGC
NJ-205	GRAMc_Ion-P_Vect53_3	CCT CTC TAT GGG CAG TCG GTG ATC GTT CGT ACG AGA ATC GCT G
NJ-208	pGRAMc_P1s_NoT	/5Phos/ATTCACTAGTGATTCAGCAG
NJ-209	pGRAMc_P2s_NoT	/5Phos/GACACTACTCTCCAGCAG
NJ-213	Amplification of linearized AD4 adapter ligated fragments	GCGAATTCACTAGTGATTCAGCAGT
NJ-214	Amplification of linearized AD4 adapter ligated fragments	CAAGACACTACTCTCCAGCAGT
NJ-268	Hs-Top1_QF	ACT TCG TGT GGA GCA CAT CA
NJ-264	GRAMc_P3	TTG TGA CCG CTG CTG GGA TCA C
NJ-269	Hs-Top1_QR	CGT TTC TCA ACA GGG ACC TT
NJ-270	Hs-ACTA1_QF	ATG GTC GGT ATG GGT CAG AA
NJ-271	Hs-ACTA1_QR	TCT CCA TGT CAT CCC AGT TG
NJ-276	Hs-AXL_QF2	CTG TCA GAC GAT GGG ATG G
NJ-277	Hs-AXL_QR2	TAA GGG GTG TGA GGA TGG AG
NJ-286	Hs-CFB_QF	CAA GCA GAC AAG CAA AGC AA
NJ-287	Hs-CFB_QR	GAT AAA GGG CAT CAG GCA GA
NJ-278	Hs-DLX5_QF	TAC ACA AGT GCA GCC AGC TC
NJ-279	Hs-DLX5_QR	GAG TAA GAG AGA GCA GCC CAT C

NJ-280	Hs-NOTCH2_QF	AAA TGC CTC ACA GGC TTC AC
NJ-281	Hs-NOTCH2_QR	CAC TGG CAC TGG TAG GAA CC
NJ-282	Hs-RPP30_QF	CTG CTT CCA GGA GAC CTG AC
NJ-283	Hs-RPP30_QR	TTT GTG GTG ATT TCC CCC TA
NJ-284	Hs-ADM_QF	GGT CGG ACT CTG GTG TCT TC
NJ-285	Hs-ADM_QR	CTT GCG CGA CTA TTC CTT GT
NJ-288	Hs-Kiss1_QF	ACC TGC CGA ACT ACA ACT GG
NJ-289	Hs-Kiss1_QR	TTT GGG GTC TGA AGT TCA CTG
NJ-292	Hs-NCOA6_QF	TGG CTT CTC AGC AGG ACA G
NJ-293	Hs-NCOA6_QR	TGC TGG ACA TTT TGA TTT GC
NJ-294	Hs-ADAM12_QF	CAG TTG CAG CAG GAA GGA CT
NJ-295	Hs-ADAM12_QR	TCC ACA AAT CTG TTC CCA CA
NJ-364	PE2_GRAMC_P4s	CAA GCA GAA GAC GGC ATA CGA GAT GTG ACT GGA GTT CAG ACGTGT GCT CTT CCG ATC TAC AGT CCG ACG ATC CAG CAG
NJ-365	PE1_GRAMC_P1s	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCT ACACGACGCTCTTCCGATCTTTCACTAGTGATTCAGCA G
NJ-399	PE2_GRAMC_P3s	CAA GCA GAA GAC GGC ATA CGA GAT GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA GAC TCC CTC AGC GGC
NJ-400	PE1_GRAMC_P3s	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TTA GAC TCC CTC AGC GGC
NJ-401	PE2_GRAMC_P2s	CAA GCA GAA GAC GGC ATA CGA GAT GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TAC ACT ACT CTC CAG CAG
NJ-402	PE1_GRAMC_P4s	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TTA CAG TCC GAC GAT CCA GCA G
NJ-403	PE2_GRAMC_P1s	CAA GCA GAA GAC GGC ATA CGA GAT GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TTT CAC TAG TGA TTC AGC AG
NJ-404	EGFPC1_QF1	AAG GGC ATC GAC TTC AAG GA
NJ-405	EGFPC1_QR1	GGC GGA TCT TGA AGT TCA CC
NJ-414	GRAMc_Ion-A_IX24_P4s	CCATCTCATCCCTGCGTGTCTCCGACTCAG AACCTCATTC GAT TACAGTCCGACGATCCAGCAG
NJ-415	GRAMc_Ion-A_IX25_P4s	CCATCTCATCCCTGCGTGTCTCCGACTCAG CCTGAGATAC GAT TACAGTCCGACGATCCAGCAG

NJ-443	GRAMc_GFP_QF2	GCCCTGTCTAAAGATCCCAA
NJ-444	GRAMc_GFP_QR2	CTTGTACAGCTCGTCCATGC
NJ-489	GRAMc_RT_oligo	TAC AGT CCG ACG ATC C
NJ-498	PE1s_GRAMc_2N_P4s	TACACGACGCTCTTCCGATCT NN TACAGTCCGACGATCCAGCAG
NJ-499	PE1s_GRAMc_4N_P4s	TACACGACGCTCTTCCGATCT NNNN TACAGTCCGACGATCCAGCAG
NJ-500	PE1s_GRAMc_6N_P4s	TACACGACGCTCTTCCGATCT NNNNNN TACAGTCCGACGATCCAGCAG
NJ-501	PE1s_GRAMc_8N_P4s	TACACGACGCTCTTCCGATCT NNNNNNNN TACAGTCCGACGATCCAGCAG
NJ-502	PE1s_GRAMc_10N_P4s	TACACGACGCTCTTCCGATCT NNNNNNNNNN TACAGTCCGACGATCCAGCAG
NJ-503	PE1s_GRAMc_12N_P4s	TACACGACGCTCTTCCGATCT NNNNNNNNNNNN TACAGTCCGACGATCCAGCAG
NJ-504	PE1s_GRAMc_2N_nP3s	TACACGACGCTCTTCCGATCT NN TAGACTCCCTCAGCGGC
NJ-505	PE1s_GRAMc_4N_nP3s	TACACGACGCTCTTCCGATCT NNNN TAGACTCCCTCAGCGGC
NJ-506	PE1s_GRAMc_6N_nP3s	TACACGACGCTCTTCCGATCT NNNNNN TAGACTCCCTCAGCGGC
NJ-507	PE1s_GRAMc_8N_nP3s	TACACGACGCTCTTCCGATCT NNNNNNNN TAGACTCCCTCAGCGGC
NJ-508	PE1s_GRAMc_10N_nP3s	TACACGACGCTCTTCCGATCT NNNNNNNNNN TAGACTCCCTCAGCGGC
NJ-509	PE1s_GRAMc_12N_nP3s	TACACGACGCTCTTCCGATCT NNNNNNNNNNNN TAGACTCCCTCAGCGGC
NJ-523	GRAMc_Ion-P_nP3s	CCTCTCTATGGGCAGTCGGTGAT tagactccctcagcggc
NJ-575	GRAMc_test1_F	TTCACTAGTGATTCAGCAGGAGTGCCATCATGATTCATAAATAG
NJ-576	GRAMc_test1_R	ACACTACTCTCCAGCAGGTACTTAATATTTGAGGTTACTCGTAG
NJ-577	GRAMc_test2_F	TTCACTAGTGATTCAGCAGCACCTGACCACTAGTGGG
NJ-578	GRAMc_test2_R	ACACTACTCTCCAGCAGCACTTTGGAATCCAAATTTCCAG
NJ-579	GRAMc_test3_F	TTCACTAGTGATTCAGCAGCAAGTACAGCATTGACTGAGC
NJ-580	GRAMc_test3_R	ACACTACTCTCCAGCAGAGACAGAGCTGACACACAC
NJ-589	GRAMc_test8_F	TTCACTAGTGATTCAGCAGTTATTTTGCTTACAGGGCCAG
NJ-590	GRAMc_test8_R	ACACTACTCTCCAGCAGGTGACACAGGAGCTTATATATATATAAGC
NJ-591	GRAMc_test9_F	TTCACTAGTGATTCAGCAGTACAATCCACCTACTTAAAGTGTG

NJ-592	GRAMc_test9_R	ACACTACTCTCCAGCAGTTAAATAGAGACGGGGTTTCA C
NJ-691	G5_1_F	TTC ACT AGT GAT TCA GCA GCC TTT CTA ACT TGG GTC ATT TCT G
NJ-692	G5_1_R	ACA CTA CTC TCC AGC AGC TTT CTT TAT CTA CAG CAA ACA GG
NJ-693	G5_2_F	TTC ACT AGT GAT TCA GCA GCA CAA GAT ACA TGT AGC TGA ATT TAG
NJ-694	G5_2_R	ACA CTA CTC TCC AGC AGT ATT TTT AGT AGA GAC GGG GTT TCA C
NJ-695	G5_3_F	TTC ACT AGT GAT TCA GCA GAA ACC CTC TAG GTC CTT TAA C
NJ-696	G5_3_R	ACA CTA CTC TCC AGC AGG GAT TAC AGG AAT GTG CCA C
NJ-697	G5_4_F	TTC ACT AGT GAT TCA GCA GAA AAC ACC ACG TAG TTT GGC
NJ-698	G5_4_R	ACA CTA CTC TCC AGC AGA GAC TTC TCA ATT CAT CTG TAT AC
NJ-699	G5_5_F	TTC ACT AGT GAT TCA GCA GAA GCC AGC GTT GCC CAT C
NJ-700	G5_5_R	ACA CTA CTC TCC AGC AGG CCT CAG CCT CCT GAG TAG
NJ-701	G5_6_F	TTC ACT AGT GAT TCA GCA GGT AAA TCC AAT CCC AGG TTG
NJ-702	G5_6_R	ACA CTA CTC TCC AGC AGG CCA CCA TGT TTG GCT ATT TTC
NJ-705	G3_1_F	TTC ACT AGT GAT TCA GCA GAG TTT TGG TAT TTT AAT ACT CTT G
NJ-706	G3_1_R	ACA CTA CTC TCC AGC AGC ATT GGT TAA GTG TAG CAA AC
NJ-707	G3_2_F	TTC ACT AGT GAT TCA GCA GAT CAT TTT TCT TTC CGA GAT GTT G
NJ-708	G3_2_R	ACA CTA CTC TCC AGC AGT ATT TTT TTT GAG ATG GAG TTT CGC
NJ-709	G3_3_F	TTC ACT AGT GAT TCA GCA GCC CGT TCC ACA AGG ATC TGT G
NJ-710	G3_3_R	ACA CTA CTC TCC AGC AGC TCC GGA ATA GCT GGG ATT AC
NJ-711	G3_4_F	TTC ACT AGT GAT TCA GCA GTC TCC TTA TAA ATA TCT TTC ACT TCC
NJ-712	G3_4_R	ACA CTA CTC TCC AGC AGA GAA TTA AGG GGG AAA AGT TG
NJ-713	G3_5_F	TTC ACT AGT GAT TCA GCA GGT GGA ATC TGG AGG CCA G
NJ-714	G3_5_R	ACA CTA CTC TCC AGC AGT TGT TGG CTC TGG TTT TTC TTT G
NJ-717	L1_1_F	TTC ACT AGT GAT TCA GCA GCT TCC TTC CTA CCT TCT TTT TC
NJ-718	L1_1_R	ACA CTA CTC TCC AGC AGA AAA CCT GGG AGT CCC AAA G

NJ-719	L1_2_F	TTC ACT AGT GAT TCA GCA GAC CTT CTT ACT TCT TAA GGG GG
NJ-720	L1_2_R	ACA CTA CTC TCC AGC AGT CTG CGA GTC CTC CTC TTC TTT G
NJ-723	L1_4_F	TTC ACT AGT GAT TCA GCA GGC AAC CAG CTT GGA AAT TTC TC
NJ-724	L1_4_R	ACA CTA CTC TCC AGC AGA GAC TTC GAC TTC TTC GGA TG
NJ-725	L1_5_F	TTC ACT AGT GAT TCA GCA GGA TCA TAC CAT TGC ACT CAA G
NJ-726	L1_5_R	ACA CTA CTC TCC AGC AGA GAG TAG AAC GAG AGA TAC CAG
NJ-727	L1_6_F	TTC ACT AGT GAT TCA GCA GAA CTA ACA TGG CTG ATG CCT TG
NJ-728	L1_6_R	ACA CTA CTC TCC AGC AGT ATT TGG TTT GCT TAG AGT CCT CCT CTG
NJ-729	EGFP_5p_F	ATG GTG AGC AAG GGC GAG
NJ-730	EGFP_3p_R	TTA TCT AGA TCC GGT GGA TC
NJ-731	EGFP_GRAMc_gibson_F	GAT CCA CCG GAT CTA GAT AAG CCT CTA GAC TCC CTC AGC GGC GC
NJ-732	EGFP_GRAMc_gibson_R	CTC GCC CTT GCT CAC CAT TTG TGA TTC ACT TGT AAG ATG ACG

Supp. 3: Adapter sequences for read trimming

>GRAMcP1sr
CTGCTGAATCACTAGTGA

>GRAMcP1s
TCACTAGTGATTCAGCA

>GRAMcP2sr
CTGCTGGAGAGTAGTGT

>GRAMcP2s
ACACTACTCTCCAGCA

>GRAMcP3sr
GCGCGCCGCTGAGGGAGT

>GRAMcP3s
ACTCCCTCAGCGGCGCGC

>GRAMcP4sr
TGCTGGATCGTCGGAC

>GRAMcP4s
AGTCCGACGATCCAGCA

Supp. 4: Hg38 genomic sequences and primers for individual reporter cloning

>Test1 (universally active)
>chr11:41152726-41153475_plus
Forward: NJ-575
Reverse: NJ-576

ATCATGATTCATAAATAGCTAATACTGTACAGAATTAGAGAAGTGACCAAAGT
CCAAGGATAACTTTTTTTGAATGGACAGTCTTTGGAGGTCACTTTGTCTGTA
GAAGGTTGGTCTCCTTTCTGGGGAAATGGGAAATTCCTGATCTCAACAAGCT
CATACCACTAGCCAACTAATAAAAATAGGTAATAAAGTAAGATTCATCTTATA
GTCTTATATGTAGAATTCCTTCCTTACCTTGAAGAATAAGGTTTGCCATGAGC
TTCTCATGTCTATATGTTTGCCTTAGTCTCCTACCACCAACTCCATCTCATAC
TCTTACGTGAATCACGTCACTTGTAGCTTCCTAAGCCCTTAATTCTTAGCAAG
TCTGCATCTTTCTCATGCTGCTCCAACAGCCTGGGATTCTCTTCGCCTCCAA
GCTGCCTTATGAAGTCTCTACTCCCTTAAAGAATCTGATTAAAACTGTCATTT
CCATGAAAAAAATTATTAAATTACACCAAACTTTGTGCTGCCATAGCCTCAGT
TACTAGATATTGAGTACATGCATCAATCAACTGAGTTGAATACTTTTTGAAGG
GAATGATGGTATTTGTTGTTGTTTTTTCTCATCTTTGCATCCCTAGAACACAG
TACATGACTCACACTGTTTACTCAATAAATATGTACTGAATAAGTTTATGAAT
GATTCATTGAAAGCCCAAATCAGAGAACAAGGAGAGTCCCTCAATTAATTAA
AAAAGAAACTA

>Test2 (universally active)
>chr6:22827845-22828628_minus
Forward: NJ-577
Reverse: NJ-578

CAGCCCTCGGTTCTCAGGAAAATTTTATATCACAGAGGGGAATCAGGGTCA
GCGATTTCAGGGGCAGTCCCTTTCACTGCTTAATCCAACCTGGCTGACTTAA
TGATATAAATATAGATTTACTGCCTCTGCCCCCTGAGCACGGAGCCACCACC
TTCACTCTAAAGGTAACTCCACCTGGTAATTAACTGTGAAATATTGAGTGGC
CAGTGTCTGACCTGAGCGTTTGTGCCAAGCAAAAGTTTCACTTGTTTTGTTT
TTTTTTTTCTCTCTTTCACTGCCAGTTTGTGTGTACGTGTATTTTTAAGTATTA
ACTCTCATGGAAGCACATGCCTATTTAAAATTCCCAGACGATGAGTCAGACC
CGTTCAGTGAGGTTCCTGGAGAGTCACATGATGTGATTTGATGAAACTTCAC
CTTATGAAGGGTTGGAGGAAAAAATAAAAAAAAACACTTTCAGACTTATATAC
ATATACACATATACATACACATACATATATATTATTGTGGATATATACTTGTAT
ATAATGTCATATAATTTTGCATACATACACACACCCACACAGTAGAGAAACA
GACAGACACACATATACATGCAGTTAAAAAAAATCCCTTTGCTTTGTTTCTCT
GTTTCCCTCAATAGATATTTAAAAATCTGGTGAGCCAACAGCTCATGCCTCTT
TGAAACAAGAAGACAGCAGAAGCCCTGTAGCTACATGTGCTGCATATGAGT
TTTAGTTATGTAGCTACCCTCTGAAATCACAATAGTGAAACTGGAAATTT

>Test3 (universally active)
>chr8:630933-631871_minus
Forward: NJ-579
Reverse: NJ-580

AAAGGCCTGGGGCGGGGGTCGGGGGGCGGCACCTGCAAACCTTCACATC
ACCCCACCAGGTCATCCTAGAAAATTAGTCAGTGCCTGTAAACATGGCAGTT
CTCGGGAAGCTGAGAACGTTCCAGAGACGATGGTGTGGACGGCTGCACCA
GTGTGAAGCGCTCAGTGCCGCTGAACAGCACACTGAAAAAGGGCTCTGGT
GGTAAATTTTATGTTGTGTATATTTTACCACGATGTAAACAAACAAACAAAAA
ACTCTCGGCATTGCCCCCACTCCCTGGCAGTGTCTATTGTGGGAGGAGAGA
CCGAAATTCTCAGGACACACCCAGGCCTCAAGACTTCTCGCCCAATCCGTC
ACCACTTCCTGGCGCAGACATCGGACTGTTAAGGCCCCTCCACTTCCCGCT
CAGGTTACAGACCCCAGGGCACATCCCCCCATCCTCACCCGCCTGCATGAC
CAGGCTGCCCCCTGCCCCGCACACCTCTCTCTGAGTAGCCTCCTGTCTTCC
CTCTGGCAGCTGAGTCAGCTTCACCACCTCACTGGGTCTGGAACAGCCAAC
TCCTGACACTTTCACACTCACAGAGGTGGAGCAGGGGCACGGGGGCTGGG
CACCACCAGTGTGTGGGCAGCACCCAGGCATTAAACACAGCAGAGGATGG
CGCAGGCACCCCTGTTCTCCTCCCAGAGCCAAGCTTCAGGCCATGTCCAGC
GGGGGAGGCTGTGAGTCACCTCTGCCTCATGTGGGTGATCATAGGAGGGT
GTGAGTCAGCTCTGTCCACATGGTTGCTCATGGGAGGGTATGAGTCAGCTC
TGTCAATGTGGGTGGTGGGTGGTCACGGGAGGGTGTGAGTCAGCTCTGTC
CACGTGGTTGCTCATAGGAGGTTGTGAGTCAGCTCTGTCCATGTGGGGTGC
TCACAGGAGGGTGTGTGTCAGCTCT

>Test8
>chr9:78563843-78564651_minus 2
Forward: NJ-589
Reverse: NJ-590

CAGGGCCAGCTGCCCTTAACTTGATGTTCCTAAATTAGGTAAGACACTGAGT
TTCTTCAGCTGTTTACACTTTGCTGTGAATGTGTTTATTAGACCCTTTTCTAA
GAACAAGCAAAGCTAGGAAGAAAAAATATATGTGTGGTCCGTGGTTCTTGG
GGATGAAGACCAGGGCATATATTTTATGGTGAGACCAGTTGCTCCCTCTAA
GGGTATCAGATTTTTCCTTGGGTTATTCTACTAACATAGAGATGGAGACATTT
TTCAAACTGCTTCGGTCATATTTTAGTGGCCATTTAATGCATCAAAATCAAGG
GGGCAGACAATAGGATACTATACAGGATGGCCAACTTTCCTGGCTCTGCTG
GCCTCATACCCCACCCAGATAAACTCCCTTCTTTCTCCTGTATCCATCACAT
CTTCCTTACCCTTATAGGATTGCAGTTGCACAGGACAACAGGAGTTTGGATC
CATATACATATGTATAGATATAGATATAGCTGGCTGGGCACGGTGGCTCACA
CCTGTAATCCTAGCACTTTGGGAGGCCGAGGTGGGTGGATCACCTGAGGTC
AGGAGTTCAAGACCAGCCTGGACAACATGGTGAAACCCCATCTCTACAAAA
ATACAAAAATTAGCCGGGCATGATGGTGGGTGCCTGTAATCCCAGCTACTC
GGGAGGCTGAGGTGGGAGAACTGCTTGAACCCGGGAGGTAGAGGCTGCA
GTGAGACAAGATTGCACCATTGCACTCCAGCCTGGGCAACAGAGGGAGACT
CCATCTCAAGAATATATATAGATAAAGATATAGAT

>Test9
>chr12:123700613-123701394_minus 1
Forward: NJ-591
Reverse: NJ-592

ATACAATCCACCTACTTAAAGTGTGCAATTCGATGGTTTCTGGTATAGTCAC
AAAGCTGTGACTATCAGCAGATTCTATTTTAGAACATTTTATCACTCTGCGAA
GAAATCCTTTACCCACGGCAGTCTCTCCCCATTTCCTCCCAACTCCTCCACC
TCCGGCATCCACTAATCTGCTGTCTCTATACATTGCCCATTCTGGACATTGC
ATATAAATTGAGTCATACAATATGTGGTCTTTTGTGTCTGGCTTCTATCACAA
TATTTTCAATCACTTTATTTTCTTTTAAGTTAACAAAATTACCTTTATAAATTAC
AAAGTCAATTCTCAAATGCCTGTATTCCTGGCAATCAAAAACAGCAATGCAA
ACTGAAGTGCACTAACAATCTTCAAACCTTGCTTGGACTTTGTTTTAATGTAT
TTTATTTTATTTTGTTTTATTTTTAGAGTCAGGATCTTGCTCTGTCATCCAGGT
TGGAGTGTGATGGTGTAATCAGAGCTCACTGTAGCCTCAAACTCCTAGGCT
CCAGCGATCCTCTTTCCTCAGCATCCCAAGTAGCTGGGACTGCAGGGACAT
GCCACCACATTCGACTAATTTTTTAAATTTTTCATAGAGACAAGGTCTCACTA
TGTTGCCCAGGCTGGTCTTAAACTCCTGGCCTCGGCCTGGCACGGTGGCTA
ACGCCTATAATCCCAGCACTTTGGGAGGCCGAGGCAGGTGGATCACGGGG
TCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTAT
ttaaaaaatacaaaaattagccaggt

>G5_1
chr6 1137373 1138163 I_62.34_64.52_85 63.43 -
Forward: NJ-691
Reverse: NJ-692

CCTTTCTAACTTGGGTCATTTCTGACTCCTTCAAATAACAAGAATATTCTTTT
CAAATCTGGGGATGTATTCAGTCCACATGGTCATAGATCTTAAAGAAGTCAT
TTTATCCTTCACTTCCCCTTTTGCTGGGTGTTTCCAAAGTCTCCTAGAACTGA
GCTAAAACAGTACACTTGACTACTAGAGCTGAAGTTTACACACAATTCTGCA
GTGTAAAACTTTTGTTTCCAGTTCTGTATCCTTTAGACGAATTATTCTCCTGT
GCTACAAATTCTATAATCTCAAATTTCTGAATATACCGAGAACGAGTGATACC
AGGATTTAAAGCGGACCACTGGGGGGTTCATTAGAGTTGAAGATGCTGCGA
TTAGGTAATATCCCTCTGAAGGATTTTTGCAATTCACTTGATTATACACAGCC
ATAAGCTATTTTTTAAAATCCTGTTTGCTGGCCGGGCGCGGTGGCTCACGC
CTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGATGGATCACGAGGTCAG
GAGATCGAGACCATCCTGGCTAACACAGTGAAACCCGTGTCTACTAAAAATA
CAAAAATTAGCCGGGCGCGGTGGCGGGCACCTGTAGTCCCAGCTACTCGG
GAGGCTGAGGCAGGAAAATGGCTTGAACCCGGGAGGGGGAGCTTGCAGTG
AGCAGAGATCATGCCACTGCACTCCATCCTGGGTGACAGAGCAGGACTCCA
TCTCAAAATGAATGAATAAATAAATAAATAAATAAATAAATAAAATCCTGTTTG
CTGTAGATAAAG

>G5_2

chr7 7620832 7621599 I_18.41_20.31_76 19.36 + I/P
Forward: NJ-693
Reverse: NJ-694

AAGATACATGTAGCTGAATTTAGAATCTGTTGAGAACACAGAAGGCAGCAGC
AGTGTTTCTAAATGTCACTGACTTAATGATTTTAACTAATAAAATTTTTAAAGA
TTCAAATGCATCAACCCAAATCATCCATTTCTGAAATGTATTTTAGATAAATC
CCTTGAAGTGTTCTTTTGTTGCAAATGGGTGTCATTATGCATTAGTTTCTTTC
TCCAAATCCCTGCAATCCCCTAACCTAATTTCCAATAAGACAATCTATCAGAC
AATTTATCTCTAGTTCCTGTGTTTTAAAATTATGGAAGCTAAATGTTGTGTTC
CAGATTGTTTAGGGAGTCATACCTAGAAGCATTTTGTGCCTTATACTTCTTTC
TACTCTTTTGATTAAGAAAACATAAATGTGTATATAACTGGTATATTTTCCAAT
TATTTTAGGCGAACGTCGCAAATGTATTTTGGCCCTTTTGCGGGGGCCTATT
TTTCCTTTTAAAAAGTACCAACCATTCAACTTTGAATAACAGATTATACAGTG
GAGCATTATTTTTTTGAAAAATAAATTATCTTAGGTAGAATTTTAGAGTTTGTA
TATACTCATCATTTATTTTCCCTGTACTGCTTAAAATGTATTCTTAAATTTGTA
TGTTGGGTCAGGCGCGGTGGCTCACACCTGTAATCCCAGCACTTTGGGAG
GCCGAGGCAGATGGATCACAAGGTCAGGAGTTCGAGACCAGCCTGGCCAA
TATGTGAAACCCCGTCTCTACTAAAAATA

>G5_3
chr1 37890375 37891109 I_12.67_13.34_33 13.00 + I/P
Forward: NJ-693
Reverse: NJ-696

AACCCTCTAGGTCCTTTAACTTGATCATTCTATTCTGTACATCTATTCCACAG
GAATAATCCAAAATAAAAAGAAAGGCAATGAGAATTTATATTTTATTATTATTT
GCATTGTACTGTATTTATAGTCGTAACTAAAATACTGATACAAGTAAGTGTTT
TATCTTGACCAGACAGATTCTGGAGATATTCAGTAGGGGGCACTGGAGTAC
AGACTTCAGATATTTATATAATAACACACCTCAGGGGCGAGTCAGGAAAGAG
TAAATAAAACAGGAAAAAAAGTCACTTCAGAGTCAGGAAACACCTAAAAGAA
TGAAAGAGTGAAATCAAACGTTTCTAATCTACATTTTCAAGGGAGGGCTAGA
CTGCCAGTTCCAAATCCTAGATTGACCACTTGATAATTCTGGGACCTTGGGT
GAGTTATATAACTTCTCATTACCTGAGTTTCCCTAACTGTAAAACAGAAAAAA
GTACTTACCTCTTAGGCTAAAATAAGATAGTCCAAATAAGAGCAGGGTGCCT
GGCAGTAAGTGCAAATAAATGTCAGGTATTATTAAGAACATTTCCAGCCAGG
CACGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGAAGCCAAGATGGGAG
GATCAGATGAGGTCACAAGTTCGAGACCAGCCTGACCAACATGGAGAAAAC
CCATCTCTATTAAAAATACAAAATTAGCCAGGTGTGGTGGCACATTCCTGTA
ATCC

>G5_4
chr14 20288372 20289134 I_27.78_26.89_49 27.34 -
Forward: NJ-697
Reverse: NJ-698

AAAACACCACGTAGTTTGGCCTGCTTTATTCTCCCTAGGCCTGTATTCTTTGT
TTTAGGATCACTTCAATTTATTGGAGTAATACCCCCATTTTTTTCTTTATACAA
AGCAGTTTATTTCTCTCCCAGTTAAAAGAATTCCGACAATAGACAGCCCCAG
GCAGGAATGGCAACTCCACAAAGTTGCTAGAGACCAGCTTTCTTCCATTTTC
CCATTCTGTCATTCCCTAGGGGTGGTCTTCACTATGGTCCAAGATGACAGAG
CAATACTGTTCTAATAGTGTTACTAGGCTGAAGGACCAGAGAGAGTTAGATA
AATCCTCAGTTTCTTTAGAGTCTAGTGACATGTTTTATAACAGTTTATCTTAAT
ATCCTTGTGGTAAAATGGAGAAATGTGAGTGAGGATTAATAACCAATTGAAT
TGGCCTGGCACGGTGGCTCAGCCTGTAATCCCACCACTTTGGGAGTCCAAG
GCGGGTGGATCACGAGGTCAGGAGTTCAAGATCAGCCTGGCCAAGATGGT
GAAACTCCGTCTCTACTAAAAATACAAAAATTAGCTGGGCGTGGTGGCAGG
CACCTGTAATCACAGCTACTGGGGAGGCTGAGGCAGAGAATTGCTTGAACC
TGGGAAGTGGAGGTTGCAGTGAGCCGATATCATGCCACTGCACTCCAGCAT
GGGCGACAGAGCGAGACTCCGTCTCAAACAGCAACAACAACTGAATGGCT
GTAATACATTAATGTATACAGATGAATTGAGAAGTCT

>G5_5
chr15 74439985 74440809 l_205.78_218.16_14 211.97 +
Forward: NJ-699
Reverse: NJ-700

AAGCCAGCGTTGCCCATCTTGTTTAGCTCTGACACAGAGCTTTCTTCCCCAG
TGGAAGACTCCCAAACAGTGGTCCTAAATCTGCCCTCTGAGGCAACAAGGA
CATGTCTAATCTCTCTTCCTGGAATACGGCCCTTCAGTGATTTCACAGAGGT
CACATCCTTCTAAATCGGCTCTTTATTCAGGGTAAATAACTCCAGTTTCAAAT
AGATCCCTCATGGGGCTTTCCTTGAGAGGCTCTCTAATCCTGGTTGATTTCT
AAGACACGATCAGTCCGTCATTTAGCAAACTTTTACTGACCCCTCCTAGGTG
CTAAACAGTCCAGATTGTGTCTCTCTCTTTCAAAGATCAGGCCTCAGATCAG
AACACAACACCTAGTGTGGAGGCTGAGCAGTAAAGGGCTGAAAGACTAGCT
TCCTTCCTCCTTATTCTAGCACCTCTAAAGGCACTCTTAATGGGTTGAGACA
ACACGTTGTTAAATCCTCTTGATCTCAGCCCAGGGTTATCCCTGGGCCAATT
CTCAACCATCAGTAGTGGCTCTCTGGGATGCTCATTTGAGAAAGATTATTGA
ATCTGGGCCCAGGAGAAGAAATGAGACTCGATTCTGCATAAGAGCAATTAA
TTTTCTGGTCCCCCTTGGCCGGGCGCGGTGGCTCATGCCTGTAATCCCAGC
ACTTTGGGAGGCCGAGGTGGGCGGATCACGAGGTCAGGAGATCGAGACCA
TCCTGGCTAACACGGTGAAACCCCATCTCTACTGAAAATACAAAAAATTAGC
TGGGCGTGGTGGCGGGCACCTGTAGTCCCAGCTACTCAGGAGGCTGAGG

>G5_6
chr1 4465004 4465791 l_434.09_352.01_24 393.05 +
Forward: NJ-701
Reverse: NJ-702

GTAAATCCAATCCCAGGTTGATATGGTCTGGCTGTGTCCCCACCCAAATCTC
ATTTTGAATTGTAGCTCCCATAATTCCCCTGTGTTGTGGGAGGGACCTGGTG
GGAGATAATTGAATCATGGGGGTGGTTTCCCCCATACTGTTCTCATGGTAGT
GAATAAGTCTCATGAGATCTGATGGTTTTATAAAGGGTTTCCCCCTTTTGCTT
GGTTCTCATTTCTCTCTTGCCTGCCACCATGTAAGCTGTGCCTTTTGCCTTC
CACCATGATTGTGAGGCCTCCTCAGTCACGTGGAACCATGAGTCCATTAAA
CCTCTTTTCTTTATAAATTACCCAGTCTCGGGTATGTCTTTATCAGCAGCGTG
AAAATGGACTAATACACAGGCGTATACCCAAGAGAAGTGAAGACACATGTC
CCCATGGAAACCTGTCTGAGAATGTTCATAGCAGCCTGACTCATAGGCACC
AAAAGGTGAAAACAGCCCAAATATCCATCAACAAATGAGTGGATGAACAAAA
TGTGCCATGTCCAAATGGACTGTTATTTGACAGTGGAAAGGAATGAAAGACT
GACACTCAGTACAAGGAGGGTGAACCTTCAGGACACATCACGCTCAGTAAA
AGATACAGAAGCCGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCAC
TTTGGGAGGCCGAGGCAGGTGGATCACGAGGTCAGGAGTTTGAGACCAGC
CTGACCAACATGGTGAAACCCTGTCTCTACTAAAAATAAAAAAGAAAATAGC
CAAACATGGTGGC

>G3_1
chr11 84433280 84434177 I_3.24_2.73_32 2.99 + I/P
Forward: NJ-705
Reverse: NJ-706

AGTTTTGGTATTTTAATACTCTTGTATATTACTACGTAACCAGAAATTGCACTT
TCAATTTATCTTAAAGATTCTTCAACTGGCCTTTTAGAGTAAGACAATCTGGA
GATTGTCTAATTAAGAGATTTAAGTTGAAACGTGACAACAGGCATACAAGTA
AGATACTAACTTCTGAATTACTTACACAGGGTCAAAACAAAGAAGAAAGGAA
TAATTATTTGAATGGAGGTATTCTATCCCAGCCAGGCAATCTGGATTTCACAT
TATCACTTTTAAAATAATTTTATTCAGAGAATTTGTTGGTACAAATGAACTTTT
AAATGAAAGAATAACAAAGAATCTCTTATGTAGCTTATTTAGAAAAATGTCAA
TGTGAATCACCTAACATCTTTGACTTTTTTCAAAAAGAGAGCACATGTACAAT
TGCCAGGCATTTAGGTCCTTTTTGATTAAAGACATTTTTGGATTCTATAGGAA
TAAGCAATGAGTACAAATGCCAACGTTAAGGTTAAAATGATATGACAATTAAA
AGCTATTCAAATAGAATAATAGGTGTCCCAGCATGGTGGCTCATGCCTGTAA
TCCCAGCACTTTGGGAGGCCAAGGTGGGCAGATCACTTGAGCTCAGGAGTT
CGAGACCACTTTGGGGAACATGGGGAAACCTCATCTCTACAAAAAAATACAA
AAATTAGCCGGGCATATTGACACACACATGTGTTCCCAGCTACTCGGAAGG
CTGCGGTGGGAGGATCACCTGAGTCTGGAGAGGTTGAGACTGCAGTTTAA
GCAGTGATTGTGCCACTGTACTCCAGTCTGAGCTACAGTGAGACTCCATCT
CAAAAAAAAAAAAAAAAAGAATAATAGGTGACAAAACAAGTTTGCTACACTTA
ACCAATG

>G3_2
chr12 2927744 2928540 I_3.21_4.09_28 3.65 + I/P
Forward: NJ-707
Reverse: NJ-708

ATCATTTTTCTTTCCGAGATGTTGAATGTGCACAAAAAGTGGTTCAGAGAAA
AAAGCCACACTTGTTAATTCCATCTGTAAATCTATGATTCAAGGACCATTTCA
GTGAGGGGTAACTTCACCCCGCAGATACTAGACACAAGGGAATTTAGTTTA
GGTCTTGCTCTGCCTCTTTCTCTTGGAGAAATGATTCTTCCTGAGCCTGTAC
TCCCACATATAAATCTGAGGCTCGTCACTCTTTCATAACACACATTGATCACA
CTTTATTTTACATCTGCATTCCTCTCCAGATTATATACTCCATGAACAAGAAA
CAATAGAAAACAAAAAGACTATCAGTATCTCCTCCTCTAACTCAAATGTAAGA
ATAAAATGAAAAAATAGTTTTTCATCCAGCAGTTTTAAATTTCCTATAAAAGAT
GCTCCATAAATCCAAAGTACTAGTATTTCAAACATTCCTAGTTAGATTTAGTT
AACTAGACAGAGATTAAGAAAATGCCAGGGGGCCAGCCATGGTGGCTCAC
GCCTATAATCCCAGCACTTTGGGAGGCTTAGGCGGACAGATCACCTGAGGT
TGGGAGTTCGAGGCCAGCCTGACCAACATGGAGAAACCCCGTCTCTACTAA
AAATAAAAAATTAACCGGATGTGGTGGCACATGCCTGTAATCCCAGCTACTC
GAGAAGCTGAGGCAGGAGAATCGCTTGAACCTAGGAGGCGGAGGTTGCGG
TGAGCCCAGATTGCACCATTGCACTCCACCCTGGGCAACGAGAGCGAAACT
CCATCTCAAAAAAAATA

>G3_3
chr9 74920250 74921035 I_4.78_4.92_60 4.85 -
Forward: NJ-709
Reverse: NJ-710

CGTTCCACAAGGATCTGTGGTTACCCAGCCACTGTGGGGGTACATGGAGCA
GTGTTTGTTCTAGCACAAGACATTCCCTGCTTAAGTTTTGTCTGCCAGCTCC
AGGCCTCCATTCCCTGTCCTGATGTTCTGCCCTCCCTGGATAATCTTTTGGA
CCCGGAAGGAATTCCTTCCTCTCCTTAAATACCAGGTATATAGAGTTCTTTTC
TTTGGGACTTCTCTGCTCCCTGACACCAATTTCCCTGTCTGGGTGAGGACAA
CATCATCTTAGGTAGGACTGTTGGCATTAGGGTTACCTGCAGCTATTTTTTTT
TTTATTTTTTTATTGATAATTCTTGGGTGTTTCTCACAGAGGGGGACCTGCAG
CTATTACAAAAATTAAAGTATGAAGAGCGATATCTGATTCGACCATGTAAAAG
TAACAGAGGTTTCAGGCATTTACATTTTATTATCCTCTCTGTCTAATTACCTT
GTAACACTGTGAGTGTGGCTATCATTAATTGACTCTGATTCAGTTAAACATG
CACCACGATCATCTGTTCATGTGTGTTTATTATGCTATAACACTAGTCACCTT
ATAATGTTTGTAATTGAAAAAAAAAAACAGGCCATGCGCAGTGGCTCACGCC
TGTAATCCCAGCACTTTGGGAGGCCGAGGCGGCCAGATCACCTGAGGTCA
GGAGTTCAAGACCAGCCTGACCAACGTGGTGAAACTCTGTCTCTACTAAAA
ATACAAAATTAGCCGGGCTTGGTGGCGCAAGCCTGTAATCCCAGCTATTCC
GGAG

>G3_4
chr5 789596 790308 I_3.45_2.94_39 3.20 -
Forward: NJ-711
Reverse: NJ-712

TCTCCTTATAAATATCTTTCACTTCCTTGGTTAAATACATATCTAGGTAATTTT
TTGTAGCTATAATAAATGAGATTGCCTTCCTGATTTGGTTCTTGGCCAGATTG

TTATTGGTATATAAGAATGCTATTGATTTTTGTACATTAATTTTGTGTCCTTAA
ATGCTACAGAATTTATCAAATCTAAGAGTTTTTTGGTGGAGTCTTAATATGGT
TTGGCTGTGTCCCCATCCAAATCTCATCTTGAATTGTAACTCCCATAATTCCC
ATGTGTTGTGGGAGGGACTGGGTAGGAGACAATTGAATCATGGGGGCAGTT
TTCACCATCCTTTTCTCATGGGAGTGAATAAGTCTCACGAGACCTGACGGTT
TTATAAGGGGAAACCCCTTTCACTTGGCTTTCACCCTCTCTCTCGCCTGCTG
ATGTGTATGATCTGCCTTTTGCCGTCCACCATGACTGTGAGGCCTCCCCAG
CCATGTGGAACTGTGAGTCCATGAAACCTCTTTTTGTTTATAAATTACCCAGT
CTCAGGTATGTCTTTACCAGCAGTGATACAAGTCTTCAGGGTTTTCTATATA
GAATATCATTTCATCAGCAGAAAGGGATAGTTTGGGCTCCTTCTTCTCCAGT
GTGCATTCCCTTTATTTCTTTCTCTTGACTGATTGCTCTGGCCAGGACTTTCA
GTACTGTGTTGCATAAGAGTGGTGAAAGT

>G3_5
chr19 271467 272331 I_4.41_4.68_29 4.54 +
Forward: NJ-713
Reverse: NJ-714

GTGGAATCTGGAGGCCAGGCGCGGTGGCTCACGCCTATAATCCCAGCACT
TTGGGAGGCCGAGGCGGGTGGATCACGAGGTCAAGAGATCGAGACAATCC
TGACCAACATGGTGAAACCCTGTCTCTACTAAAAATACAAAAATTAGCTGAG
CGTGGTGACGCGTTCCTGTAATCCCAGCTACTTGGGAGGCTGAGGCAGGA
GAATCACTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCC
ACTGCACTCAAGCCTGGTGACAGAGTGAGACTCCGTCTCAAAACAAAAACA
AAACAAAAACAATTATTGAACCATTGTATCTCCACCTGGAATGGAGCAGTGA
CTTATGTAAGTGGCAGATGATGGGAAGCAGGTCTCTCACTGCTGAGGTGGG
AGGTTACAGATTAGCAAAGGGAGGCTACAATGATCCATGTGATAATAAACCA
GAGGGGCAACATCAGCATCAACTCATACTTAGCCTAATCCAGACACCGACA
GCTCCACATAGAAATCTCGATAATCAGGCGGTCCACATAGGTTGGAAACAC
ACAGGTGTTTCCTGGCACGTCAGCTGTAAGACCTAAAGGCATTGAGGCCGC
AGGGGCAATGAGTGCATTCAGCTCCTAAATATAGTTTTTCATGCGATTCTAC
AGTGAAAGGAACCAGGGCTCTTTGAAGAAATGGCTTCACTAAAACTAAGGC
AGTAAAAACATGAGCCTGGATGATCCTGAAGTATCAAAAAGTAAGGAACTGC
TCAGAACACAGACACACACACTCTCACCCCACACACACACACATACACACAA
GCACACACACACACAGATTATGACAAAGAAAAACCAGAGCCAACAA

>L1_1
chr1 1154867 1155707 I_0.99_0.86_21 0.93 +
Forward: NJ-717
Reverse: NJ-718

CTTCCTTCCTACCTTCTTTTTCTTTTTTTTTTGACAGAGTCTGGCTCTGTCACC
CAGGTTGGAGTGCAGCGGTGCAATCTCGGCTCACTGCAACCTCCGCCTCCT
GGGTCCAAGTGATTCTTCTGCCTCAGCCTCCTGAGAAGCTGGGATTACAGG
TGCCTGCCACCACACCTGGCTAATTTTTGCATTTTTAGTAGAGACAAGGTTT
CACCATGTTGGCCAGGATGGTCTCAAACTCCTGACCTCAGGTGATTCTCCC

GCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCATGAGCCAACGCGTCCA
GACCCCAGTTTTCTTTTCAGCAATTTTAAATCTTCTGTTAAGCCTATCCATTG
AGTTTTAAATTGTGGTTATATTTTCATTTCTAAAAGCTCATTGTCTTTTTAAAT
CTACTACATAATTTTTATAAGCTAAAGTTCTCTGCAAATATAAAACGTCTTTTT
TTGTTTTCAACACACTGATTTTAGTGTTTTTTGGTCTGTGTCTGATAATCCAAT
ATCTCAGGCCCCTGTGGGGCTTGCTGGTGTCTGATTCCATCCGTTCTTGCTT
ATATCATATCTTTGCGAGACTGGTTATCTTCGACGGCGTGCTGGTCATTGTA
CTTGAACAAGTATCTGTAGGACGATTTGATGCCTCGAGCAAAGGCACTGCC
CCATTCGGCTTTCCTCTGCCAGACACCTGGGACAATACTGCCTGGGCCTGA
GTGGCCTTTGGAAGACCACTGTGGAAGTGACTTGAACTGGGGCTGCGAGG
ATGGAGTCACTTCACTTCGTCCTCACTTTGGGAGTGGAGCCCTTTGGGACT
CCCAGGTTTT

>L1_2
chr8 342822 343535 I_0.69_0.60_25 0.65 +
Forward: NJ-719
Reverse: NJ-720

CCTTCTTACTTCTTAAGGGGGACAGGCCAAAAAGCTCCCTTGGCCTCCAATC
CCACTCATCGGGGCTCCATTCTCAGGGCGTAATCACCTCCCAAAGCCCCTG
GTCCTAGTATCAGCACCTCGGGGATTTGGATTTCAACAGATAAACTTGGGG
GACCCAGCATTCAGACTATGACAGCTAGAAACACAAACATACACAACACTCC
TGAGCTCAGCCTTGTCTCCAGAGTTCTGCTCTTTCCAAAACTAAAGTTAGGC
CCTCAGTGACTCTGTAATTCCACTCCTCTGTGGAAGTCATAGTGCCCAGATC
TATTTATGCCAATAGAAATGAGACAGACAGACCCCGTCACCCAGCCCATTAC
TCTCACTAAGCAAGATTATTCACACTGATACATATCTTATCATTGCAGGCCAT
TAGCAAGGTGCATGTAATGTTCTGCATGCCCTGGGGAGACTAGAAATAATG
ACAGTTGTCAACGCTCCCTTCCTTCCAGAGCCCAAGACGCTGTCTCCAATGT
CAGTCCCAATGACGGCTGAACAATGACATTGGCCGTTGTCCTGGCTGCCTT
CTCCGACGCCCTCTGCATGTGACCCCAGCAAATAAATCTCTGCACTCCTCTA
CCTCCTTAGAGAGTGGAAATCCTCATGGAGAAGCAGGCACAAGTGTGAACT
GCTTATGTGAGTGGTGTCAAAAGAAGAGGAGGACTCGCAGA

>L1_4
chr10 1331187 1332088 I_0.76_0.57_64 0.67 +
Forward: NJ-723
Reverse: NJ-724

CAACCAGCTTGGAAATTTCTCAAAATGTTAAATACATTTTAACTCATCAATCC
TACTGCCAAGTATATACCCAAGAGAAATAAAAACATATATCTACATAAAACCA
CATATTCAAGTGTTCACAGTGGCAATATTCATAATAGACAAAAAGTAGAAATG
ACCCAAATGTTGGATGGATCAGCTGACAAATGGATAAATGAGATTCTGGACA
TCCATACAGTGGAATATTACTCAGCTATAAAAAGAAATGAAGTTCTGACGTAT
GCCACACCCTAAAGACACCAAACTTAGAGAAGCCAGACACAAATGCTCACT
TATGTGTGATTCTATTTACCTGAAATGTCCAGAATAGGCAAATCATAGAGAC
AGAAAGATCAGTGGTTGCTAGGGGCCAGGGGAGGGGAAAATGCTAACAGG

TTTGTTGTGAGGTCAGAAAAATGTTCCAAAATTCCTATCCAAAGATAAGAATG
ATGGTTGCAGAACACTGTGAATATAGTAAAGACCTTTGAATTGTACTTTAAAG
GGTGATTTTTATGGTATCTGATTTATAGCCCAATTAAGTTGTTATAAAAATGC
TGAGATCAAAGTATCAGTAAAAGTTCTAGATTTTTTTCCTTTACCTCAGTGAA
TCATCTACATATCGCACACTTTGAAGACTTTAACACACTACACTAGTGAATTC
TGGACCCATGTATTTGGAATTATTCAGAGAGCTTTTAGGGTTTTTCAGAGATT
TCCTTCCCCATGGTATGGTGGTGGCTATTTGAACAGTTTGGGAAATTTTCAG
AAAAGATACTCCTGCAGTTGTCAGATTGGCTGCAACACTTGGCTTTGGTTGA
CAGGGCTCAGGGTGCCATCTGCCATTTTCTTCAGTCATATTCATCCGAAGAA
GTCGAAGTCT

>L1_5
chr16 89802861 89803690 I_0.38_0.17_18 0.28 -
Forward: NJ-725
Reverse: NJ-726

GATCATACCATTGCACTCAAGCCCAGGCGACAGAGCAAGCCTCCGTCTCAC
AAAAAAAAAAAAAAAGAAAAAAATCTGACTATAAACTTGTTCCCTTTCAAATA
GAAGTGCAAGAGGAAAAAATCTGGTTCATCTAAGGGTCCTGATTGGGTTTG
GTGAACGCTAGTGTAGCTGTGGTCAGGTGAGCAGTAACTTCTGCAGCGTTC
CTCAGGTTAGGAGCAGCTCACAGGGGTCCTGGTGGGCCCAAGCCATATGC
TCACCCGCCCCTCTGAAGTGGATGGTCTGTGCCATTCACAGTTGCTTGGAG
ACCATGATGTCCATAAATAACTTTGGTTTTATGGTTTTTCAGATTTTGTTCAAA
TGTTTGTTTTGAGGGGATTTCAGAAAAACTCAGATCTGAGAAGAACTGTGGA
GCCTGAAAAAATGCCGCAGGTAGGAGGAAAAGTCAAGTGAAGAGTTTAGCG
GAAGGAGCCGTCCTTATTCCAAGTGTTGAAATGTATTTACGTGTCTCTCTATT
GAAACGTGCTGTTTGTACCTATTTACGGGGTACATGTGATACTTGTTTTGTG
CATAGAGTGTAATGATCAAGTCAGGGTGTTTAGAGTCTCCGTCACCTCCAGT
ATGTATCATTTCTGTGTGTTGGCAACATTTCACATGTTGCAGTTACTTTGAAA
TATACAATCTGTTGTTAACTATAGCCGCCTCACTCTGCTGTGGAACATTAGA
ACTTACTTCTTCTGTCTAGCTGTATGTTTGTACCCATTAACCAACTCTCTTCG
TTCCCCCACTGCATATACCCTTCTCATCTCTGGTATCTCTCGTTCTACTC

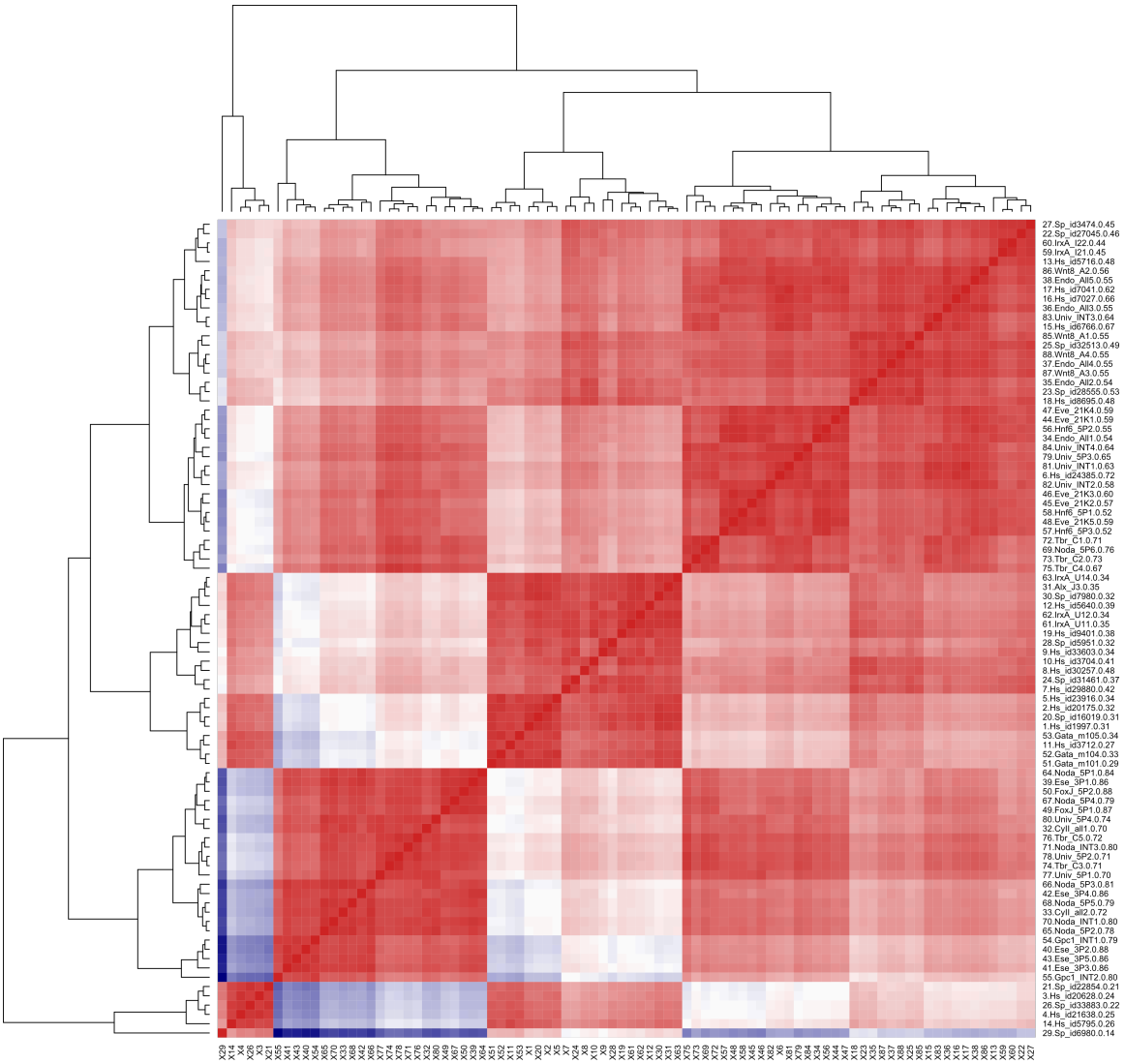
>L1_6
chr3:45837895-45838760 I_0.31_0.27_11 0.29 -
Forward: NJ-727
Reverse: NJ-728

AACTAACATGGCTGATGCCTTGCTCTTAGCTGACTGCTGACCTGCTCTGTTA
TTTGGGTGCTAGCCCTACCTCCATGGGCTTGCCTCTTAAATATTTGTCAAGC
TGTCATCTGGTTTAAAGTACAGACCAGTTAGTGGTTCTTTCCAAATCAAATAT
ACTTGTTAAAGCTGTATTTAATGGGGTAAGGGTTAAGCTGCTATAACAGAGT
TGCTGTTTTTTGCTGTAAGGTTAAGTGGCTTGAAGAATACAGAAGTTTATTTC
GCTCTCACTTAACAGCCATCTCACTGCTCATCTCAGTTCAGTGAGAGGGTCA
GTGCCTCTCCATGAGGTAGAGACGTGGGGCCCTTTCATCTTGCGGCTCCCC
CAGCCTCTAGGGCATTGATGTTGTCTGCATGATTCAGTCAAGTCGCCACCTC

TTCCAGGTTTCAGAGAAACAGGGAGGGGCAGAGGTACACCCCATATCTTAA
GACCAGGTCTTTGGAAGTAGACCACAACATTCATTGCTGAGAAGTAGTCAG
CAGTGACTTGATTGCAAGGAGGGCAAGGAAGTGTCATCTAGGTGGGCAGC
CATGCAGCCAGGAAGAAGGGAAGAATAGAATTTGATGGAGATCTAGCAGTC
TCGACCACAGGATATCTTATGCATTTTCATTGTTGGCTCTTTGATCAGATCTT
CAAAACTAACAATTAAATTACCTCTTTCTTTTTCAGGCAGAGTTGGGCCTAAA
TGAGCACCATCAAAATGAAGTTATTAATTATATGCGTTTTGCTCGTTCAAAGA
GAGGCTTGAGACTCAAAACTGTAGATTCCTGCTTCCAAGACCTCAAGGAGA
GCAGGTATCAGAGGAGGACTCTAAGCAAACCAAAT

Supp. 5: Heatmap showing clustering of 800bp human and sea urchin genomic fragments with reference sea urchin CRMs using MMOSAIC

Red indicates lower d-values (higher similarity of ROPs), and blue indicates higher d-values (lower similarity of ROPs).



Supp 6: Supporting File 1 for Single embryo-resolution quantitative analysis of reporters permits multiplex spatial *cis*-regulatory analysis

Supporting File 1.

#CRM-specific primers are marked as Capital in Supporting File 2.

#Primers to linearize pre-barcoded empty vectors
>Lin1
CTGCTGGAGAGTAGTGTCTTGTACTTATATAAGGGGGTGGG
>Lin2
CTGCTGAATCACTAGTGAATTCGCGG

#Primers for making extreme barcoded constructs
>SE_N25 (5'-biotinylated N25 primer to add N25mers)
CACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTTTACAGTCC
GACGATGGNNNNNNNNNNNNNNNNNNNNNNNNCCGACGATCCAGCAG

>EndCore-PolyA (univeral reverse primer)
CACAAACCACAACTAGAATGCA

#Reporter-specific oligomer for Reverse transcription
>SE_RT_oligo
TACAGTCCGACGATGG

#Primers for QPCR measurement of GFP DNA
>GFP_QF (for QPCR)
AGGGCTATGTGCAGGAGAGA
>GFP_QR (for QPCR)
CTTGTGGCCGAGAATGTTTC

#Universal primers to amplify barcodes from cDNA and gDNA
>P3
TTGTGACCGCTGCTGGGATCAC
>P5
GTTTACAGTCCGACGATGG

#Primers for building sequencing libraries
>Ion-A_IX_P5 (forward primer for sequencing library, XXXXXXXXXX is for IonXpress barcodes)
CCATCTCATCCCTGCGTGTCTCCGACTCAGXXXXXXXXXXGATGTTTACAGT
CCGACGATGG
>Ion-P_P3 (reverse primer for sequencing library)
CCTCTCTATGGGCAGTCGGTGAttgtgaccgctgctgggatcac

#Adapters for sequence trimming
>AD_P5
CATCGTCGGACTGTAAAC
>AD_Int
CGACGATCCAGCA
>AD_P3
ACTCCCTCAGCGGCGCGC

Supp. 7: Supporting File 2 for Single embryo-resolution quantitative analysis of reporters permits multiplex spatial *cis*-regulatory analysis

Supporting File 2. The sequences of CRMs used in this study. Capitalized parts are primer sites.

>nodal_5p
TTTGATCAGATGATTATTGCATGCtcaaaacatgaataaatttcatatgtgcaaaaatcgtatatta
tcgtatttccgaaaattcccgatgaataattatgcaaatatattcacattatttttgtaacttcttattttcatttcttttctgttg
ttaaacgagtgtatgaagcccacattctgacgagactgttgtgctttccttagatcacaacacatgaagccaattgt
tttcaaagctgccaattaaagcagtctgaagctcagctaaacattattcaccgaattgtactcttcaaaatgaagta
tcgagaaagaagatagactaatgcgtaacacttagcaggactacccttttattgtcgtttggttacaactaattatag
tgtgcttagagtcctacaaaagagccgattcattgttaattaggaagacgggatgggtacatgtcgagcagccgg
gatgggttctatgcatagtggcccgagaaccgttgctttcattcaccgtggttcggctgaatggtttttagcatgaatg
cttgaattgcgaaacaatgcgcgcgcatgttgagctaggggtctttgaaaagAGTGTCTGATTAGTGT
GGCAAGGC

>nodal_int
TTTCAATTTTGCGGGCAAATGtatttcagttgagcgaaatcaaaattactttcttactatttacgggaa
gattagactatcttgcacatgtgctagtcgagtctttttcattgactttgtttttgagtctctccccccctcacctctcaaaca
cctgtacgaagtggctgtaaattctgccccaaattgttttctaattaactcgtctgagcctttccgttaccggataatct
atcaaggtgaaaagaggcattcatcatgaaaacgattaacccttcacacccgctctatcacaaagaagaaga
ataaaagggatgggacgagactcttttgttcgagggtccatccgaacgccgaagttcgagtcctctgtgtgggga
gcatggcgggggtagcgcttttcattcgagtggttgctttaatagcctggcttctgaagccttttgtgggtcgagactg
aatgagttaaactatgggctctcagaacaaagcccaaaattgatacgtcgcactagtacactgaaatggctggg
ttaataatattctcgtctggtatcgagctgggatgtagacataactattaataagattgatttgccatcggttcttcata
aaatgtaatgtcattaccttacttttgacttgtataaaaacaaatgtatcttgcttgtctgttgataatttaactaatcaaat
gttcttttctttctttttatttcagatCGTCAAATCTTCTTCCGACCTTGT

>nodal_3p
GGCCTTTGCCTTGAACTCTGAactcttacaatttgccaaactcttttgaacgcccacatggattgggt
gatcaagtagactcgatcagcaatcttaaagaaactgcaatcttggctctttttattatcgtatttttttttgttaacccgtc
gattgattttctttgcgatttttacttgttttatatttaccttatgaatgaatgtatttttcataagtttttgcagacttcgaaaca
attgaatacgagaggtacatggcacatgggggcgcggtaccacctaattttgtggggagatattacccgcacaa
gttagacgcatacaacaaaatctagaattgagagacggaaaatgcgttatcgaagtttgatgaacgttttattatc
aaaccattcatgatattttcaatttcatcgactttccaaggtaatacaaatagaaataatgtactcagtactatagtaa
ttatgttcttctacaaatttggatacataatgtatccgagcacatagccctagctatgcatttttaaaatgtacttttatttat
tgtatatttgtgtgtaaattatcaattcagagaatatgcatattattaaaaataaagcaattataaattcaaaaaccttt
gttatgattCTTGATTCATTTGTGACCTCTTTAA

>cyiiia_5p
TGTATTTCTAAATAAAGAGGGGATGtacatttataagattattaggtgggtttaatacccaaatatc
gacttcaaaatttgagtagggctgcagctttctcagaaccccgcatcgtatatacattcccataaacacacacac
acattcacacaaaatgtaacttttttaatacttgaaaacataattatcattctttgatgtacagcacatagttgacaca

aaggggatgttttttaatagttgaaaacaccgtgctttctgtgagagggaatgaccctgatccccaacccatac
acacatgtaaacatccaggacgtagataaggggcatgtgaatatgtgcacccctccccctccctccccaa
ccatcaaatgataagaaaatatacagaaaggtgaaagtggagaggaggaaatgtaacgtaaaatgaatatg
ctatcttccccacaccctcccattgtaacttcgcattttatccatcgggtaagattgatcccctccccacctcaccct
ccccccccacccccctcctactaaacctgagcccctcaatgcctggacccacgttacgccattttacgcactttccg
gccgtttttattagaccctatactcgtaatgtaaaagggttttgcagccatgttatttttcacggtatatctaagtgtgttttt
cggggaaaaacttgacaaagttactccctgtggataaatttcctaatttgcgggtcctagttttagttacttatgacaa
acaaattatgaaaagcaaaacattaaggtaaaatacaggaataacataaagaggtctcaaggttttgaacagtt
gaaatgttgttaaattacatctcacaaacaactgtataaaataatttaaaataagtagacacacggagagattgtg
ggacatgtcgaacactcctctgcagacccttgtgtcagatgcaaggattttatacgaaatctatcagaacagacg
caggccggttgtggtcatggttttaaaagtagaaagtttgggtcattttgtcggttgtccgacatttgctcctgctaaaa
ttgccttgcatcaatttcatactcacactttgttcttgacttcaatccttgacccaacaatatatctaaccgttacccttaa
cctaaccttaaagccaaatgtaactctaaccccgatccctcgtatgtgacttaataaagacagccgcaagggac
cattttcgtaatgtaaccagatttaatgcataagtctttgtgaagtgcctcaatacatattatatcattcgcctcataag
aataacaacgtttcaatagtcagaatgcaaagtacatgttgctaggtaggtcaagccatattgtttgtcctacctgg
caacaacttatgtatgccgttcggattttgtctattgaaatactattattcttatcaggtgaatgaaatttttattgaggc
acctcactattaaagacttatgcattaaatttggttacattacgaaactggcctattggaggaggagtacgcatctta
caaatcgtacggcggcgcacaaatgcggggttcgccctgatctgggcctctaatggcggcggaaagattaagt
atgtaagaaataagtaatgcatggatctcacatgaatacacaagcaaccggcacaagggaatggacttaaag
agctagcgagaatcattcaaccataatggaaactctgattggaccacggtgaatgggatatactgagctacttta
aaaggagtatttaattcaggttatgcacgctctggtggtatgaggtttcaataggccagacgttataggaaagattg
cgattgaagactcccaagaaaagggtaacaatggagcgataaaattgtttctccccttgagttaatgctttttgtca
tgcctaattatacctcggtgaggctcggtcacaccgcctaaacgattctatacccccctccaaactgagccccattt
aagcttaatcctattttggtcccgtactggccaacagtatgtagacgcctaagtttattattgcattgtcattcattgtcg
cgacatacttgtagttagctttttattcctccttactttctttgggtattagctgcgaagcttatctattggcctcgttgtcag
atttgtttttaaaaagaataaatgaagcgcaaacaaactttattaagcaaaaaagcaccgaatctcatttgcatat
ccttttcaatgcattccttatctgccctgaggcgtacgatgtgtctaaattgtctccttatttggtcaaaaaccctggac
ataactctcgcttgggggtctttgtccaagaaaggggtagtacattacttggtcccccacagtatcatttcactctcta
ccaagcaatcaagcaggtggtgtcatccagttctctttctcttctctctctaactcggtaagtttttttgcgatcattgaat
aagtgttgctaccttgtattagccttaagggcttggtggtaataTGTTTATGATATTGATACCACCAA
TT

>tbrain_c
ACATGACTTGATTGCAGCTTAGGCACtttaacaaaaaaagagtctttagaattctttgaacgttgt
acagatgggtccacctcatttatttttaacacatttcaatcaagaggtgttagaatataagccagactatcctattaat
tctcaatcctaaatctcacctcgaattgataatttgttttttgccttgcttgccatctggtatcgccaaatgaacaatatta
tagactaaatattttcaagtgggttcctgggacggccaagacaacgatgcgccttttccgcatctctttctctctccct
atctctctctctctctccctatttttctctctttatttctctttctctctcacaccctccctttccccctttttgttctctttctttttccc
actctttctctccctctcttcaaaatgaatcatgttcaatcaaattgtgtcctttactctatcttgaaatatctgtttgtatatt
ataaaagagatatcattctgcaagatcttgtttcttaccctcgctgtcattctattgaaatgtatcgcatttgaaatgata
tgttttttgttgtttgtttgttttttggttttttactttctcatatataccgaatatatatcacaccttccactgtcttcgcctctctct
ttttctctctctctctctctctctctctctttttctctctctttctcttcctctttctctctctttctctcactttcgcaccctgattctc
ctacgtaatttgacaatgttttaacctgtttcatcatcttattcttatgtacaaaaaaaacacaacatagagatagaca
ttatggataagcaactaaatttggtttctataaacgcattatacatgggaaattacatatttcatgactgtagcttcttct
ctggcataaaattcttatgttcttagttaaattgattcacattcactatatctgtattgttttgaagctttaattaagccaaa
acatattgattaaaagtagttttctaatctaacaagtaaattatgttgaaagctatgataaacacttttttaaatagaca

ttacgagaataatggaatatttatgaacaaataacatgattttcgaaaaaaaaaataaaaatatatatatatatatg
cgaaactctttacattatatttttattgtacagcgtacatttacatgtacttcgtacatattatatacttcgtacataattatc
gggcagcttcatgtaaataaaaatcgatcttttgtcggaatgttagtgtcggtatgtgacacatggcggcgtatcaa
gggatcattaattaatcccctgttgggtaccggtatttcggctcccacgaattatgtgtttttaatgtgtttattggtaaac
aaaaacatagcatcctgcgatgatgaacaaaggaaacgcgtcagaacagcgcgtgactatgtatttttacgac
cgtggacttgtttgTGTCCGCGCTCATTAATTTGCTTGT

>ese_3p
ATGCATACAAAGATTAACACCTGCTgtactagagaatccagttagttgacctaggccttagttact
gcctaaattgccgccaaacacatccattcagctagcacatgtctattacattctctgcctatacaactcaacagggt
catgagggtttagaatgaaagaaattcctgtcacattgtgacatatcgtatgtataataagatgaggatcctaatca
aatgattggttgagagcgtacacgtgaccaggcatttattttgccttacacgtaattttactaatgcccggtcatcagt
aaaatcatgtcattaaacataatttataaaacaaataagacatcatcttataaaacaaataataaacatcattttgtt
cgggcattagtaaaattacccacactcgatacctccgaaacagtccagacaccctcggctatcgcctggggcgt
gttagactgttctaaaggtaccttgtgtgggtaattcttactaatgacctcactgctgtgtagtatttgtatacagtcaaa
cctatataattattagagaccgccctagaaaaacaaaataatgtggttactatttaaatgcaggtttgtaagtattttg
gggatattgggggtaccactcaatttgggtttactgtctttttttgtttttccttcaaaatattcctcaattgtaactgttctg
aacaatacgttgatgtcagttgatatatgattattggtctatatatttatgtttaaatcttattttacatatctgttattattaca
cgacacggttgattagtatcaaacttgggggtacttgtattcaagctttgtacatggtattacatttaattaataatcatt
attgccaattttatattgaattgaggtgccgtgtagatttttacttaaaattgcacacttgtttattatatgtattttataata
atttagaatgtaacaaatttaataccaattgaatttgactggatatatatatttttacggataatgaattataccaccac
tcaatttgggtttactgtctctctctctttctttcaaaatctttctcaattttaattgttggacaatacgttgatgtcagttgata
tatgattattggtcctgtgaagtcctatgtacctttaattatgtttaaatgcgattgctcttgatttttttttcttcttctgttcatat
ctatttatttaaagtgtcaatttatacagttttattgtacaaggtatatatacatataactatattgtgaaccatttatatgc
ctgccttttttactgtatgaaatccaacttatggttgcaataaataaatgatttaatatcatattgatggggaggtgact
acaaagtgttcttgttttagtctccatttccttttcatcttattaccaagaatctccattctgctttccgtttcatttactttcat
gtttaggagttatgcaatgatgcgagatgtgacgtacatcagattaaagtattcaagtagaaatgcatcctgttaca
aagaaacgtaggcactacatctatagtccgctgccatataagggtaacccggcgcagtctgagacaggccctg
caagttttacgcatttacaatgcatagcattgaggaggtatcgaggatgtgcaggggccctcttgcatgactcaat
agacatgcgtatatatagacagactaattagggatctctttaacaaagaaccaagatgaaagcatagaaactcc
agtgtccttgggaagggcgtcgtcttttgacattatccgattacatgcattgtgacgaaacagtgacgtaattatcac
ccatgttgacatgaggaagtttgtagttatcgaaaatttaccttggtttaacatgttcaaatgcagcctagtatgatta
gcttattgatttctcattgtatagcctaccagtagtctgcaGCCACAGACCCTAATTTGACACTTT

>foxj1_5p
CGGATTCCGAAGTCCACTAGATAAtgtaggttaaaacctattctaagataagatatattatgtcaa
acgaaaggaatgttgaatcattgaatgtgacttctgtgtttgatatttctttataaagatgttcaaaatatacttcttaatt
gtcaaacgtctattcaggctcaacaaattattggctattttctatcattttgaagttgaccgaccctccccccccccccc
cgtttttctgatttttgttttaataaatttgttcttctgtccgtcatctcaatgttttttttaataaatagcctttctctttcctatttct
cgtcctctctatctgcccctcccaactcctcccatcccaaacatctttctacacccccctcccttctctccctcaccctc
tctctctctctctctctctctctctctctctctctctctccacccctctctctcgtatggcagaggtggatacagaccgat
cagtcaccgcacgtcgctcactgaaggcagaagactaaaaacaaatatgtatatggtgacaggagagtgtttc
gagagtgagagactcgaagaaaaagaaaagagacaaaagcaaagcaaatagacgcaattagaacacaa
gtattgcaaaaatagaagcaaacgtcaaatgaagcagcatgcatcacctcaataccattgctcctcagcacgta
tacactatcgacactataccgtttatacgtgctctgatgaaggctggagtgaggatttaaagataaagttgagttct
ggccatgcaattgcattgtgaaagaaacggtgtggaatcgatcagaaaaggttgttcatctcgcatataagtatc

ggtgtgtactgccaggttctggaatcatcaaaatgataagaaatctggcccagttttggtaaagttacaagcactgt
acacagaaatatgtatatgggactacactaacatgtggcataaaaatccgtcccgatcgttatgctgttacatgcc
atatttagaatattaagacgtttgaagagccattcttttcagggaacatttttttagctgatgttcttccacacttgcataat
acaatgataacgctcacaaattatttttcacctcttcccagaactcaacctaattctctttaagccatcctcatttaagc
catcctcatcagtttcacctatctccttttcagaagccgattactttgaccttaattatcattggttgactgaccgcatag
ttccgggtcaatgtcgggtattatcaatagcaatcgatgcgattgctattatcttgcttcaccagtgagtgcaattgttc
aataccttcattaggagttcatcttgattgattgggcgtgtagaggaagaaggtcaagttcacgcaggtattcccct
ggaatcactgggtcacgatcttgcgaattagatgccatgttgaatctgtatataagattttccaaagatcacgttttta
atttcctctttaatttttttccatattgcgaagaaagggaagaatactaacattcttcttttctctcatccacgccctttctttt
tgatactctctctctctctctctctccctcgctgatttcgtagaaatcggagttctcgatcacttccatttctcgtcctttc
actcattatgtgtgcgcccccccctctctctctctctctctctctctctctctccctctctgttttcatcccccctcttagctc
ccactatataaatatccccctttctacagtaataacatgacaatattagggcattcactcatatccttcttcaccatcctt
cttgatacgcaatgccatgtaatcaaagaaagatatattgataattgcatgtttcattaattgtttcggtctcaaattca
atatacgatgtttcgatttataaggagtgcttctcccccagtttgagttccctccaataccacttaccttgcaattactat
acaatttcggcctcctctgtaccttgcatcggttttgaaattgccctcactaagattgatcttttgtgcacagtgttttatttt
tctgattaaaaaaactatatctaaaaagaaaattatgttttcacgataatcaaattaatataacataatcaattgaca
ttatgtattcaaactcatacgtatgattataacaatcgtacattataccgaaatacattttaaattaaaattagcatacc
ggcataatcatgtgaatagtattacatgtaatggaggtttattgaacagtcgattgtgaattgataccatgtcaatctt
aatttaagtgtctggtgttttctaccgaattgtgattcaggtgaggaataagtgcccctattaacgtcaacgaacaat
aatccttaactgtattaatgtcttagggaaatcaatgacgtattttgatgaagactggcaggtaagtctcaaaacaa
taccttcagaacaattaatggagatcaagtgtctctgaaacgttgatagaatgacattatttgtaaagaagatcaa
aaNNNNNNNCACGCGGATTGTTATCGCAGAT

>gpc1246_int
AAACAGTTGGCTGATTCTAACTTTAAAGttacagttctggcctatttctaggtttttaagatccctta
attcaattaagttctgaatgaatgagaccgccttggggagttctaacgggctaattatgactaactgtcctagctaat
atgaggatcaaaataaggaggtagttatccacttgttgagcctctaaaaagacactgaaattaaaagagctgaa
gaatttgttgcttcctctcttagagtcttaatggtcattttggggaatgatacaccattaaatgaagtactttataagattt
caaccctcatttaatcatcatcaatgttggtcacttttcctctttcgtcattgttttcgttcatcctcctccttggacaccga
actgatgaactcgtggtctaattatcagctttaaggttgagcaatcgcttaagtccctttcattgtaggccgtcatcaa
gtgatcacattccacgctgaaaattagatctttttattcactcctgttatctcaccttctcagactctattgttgccaattta
attgtctctcgcactctcggaatacatagtgattcctttgaagcccggtagccctattgtgatcgaagatccttccga
agctccctaccacaattccaaatttcttcccaagccattaaggattgaaagaggagtgtcacaatcgccaatccct
gttcgactctttgatgaacccctctctctccctgcattcagactatcaggtttcatcgtatagctatacaagtatttgtgt
agtctgaacgttatatcaagaaacgttgtttacgagatgttgtattacaaaaactcctgtttaagcatcgttatgaac
gaaaactcaattaaaaatcattttcctgttctagtctgaatgtggtatcgtacgaatatccgttaaactgggtaggact
tagcatgaccacgcactgatgactggccggctgtttctgtatatggtttctctagtctgaatgctcctttcgctatatggt
tattctttaatacaacgtttctgtatacaagttttgtatgaatttcctgtagtgtgaacagaccatttgaatgacccttct
cccctgtcctcttctgtgttgatcaatccccttgaatagatgatgaatttacggagacacaaaacaacatctaaca
aaactaatcttctgtttttggacagctattgtgccccatctcaatcttagagctgtagtcttgttgtacagacatttggac
ctttatgatctttcgtgaaatttaggtaaagatgtactttctgtagatgttgatggaaagaactgtcttttaaaatagcca
tcggatagttaatattcatgcagcacattttcgtaaaagaagatttgatttattgaagaaagtgaaagaatgtattaa
tctgctgtaactttcaattgtcctatcttgaagaagcataaatatgtagttcttcaggttttggttcttaattcttaggatga
aagaggcacatgtaggacctagagatatttatttgaacatgactttatggactctggttaattgacgaagaacttctt
tggtgtactcagcattgcgggtctgaacactgaaactagtcaaaagggtaggatttgaagaaagagaagaata
acttatcttgtgaatgaattgcaaacagacaatataagcaggagggaaacagaacgattaatgtgtatgcatgc

ctgttgctgtgtttgtcatgacgacagccaaagtatggagaatagagggaagagagaaagaagggaagaatt
agcgtttggatgtatagagagtgagaaagaatacgtgtttatcgattgcttagagacagaagggcactgacttgat
gtaaagtcaggatttaagacccttctggctctcagagaagtgggtcgagaagagtattgagaaatgtgaggaga
aaaggaaacttaataataataactttgttctttttatagtgaatttcatgcctgttaacatttcaatgcgctttgcaatat
acatatataaactacaacaagatcatgaaaggaagatgaaacagaacaatgaagggaagaggaaggtag
aaagagacgaggaagaagttaaaaggtataggaatagactgcgaaagtgaagatgcatgcaattcctgcag
agacaatgagacatgcttcaacaagtattaatactagtagcagaggcggtatcattcaattagttcaaggcagtt
gggtgtattgtgctctcatgaataaagctatattctaggtcgttagttctcattcaaccttcactggggccctttatttgcc
ccctttttctattgactgcccccagaatggcacccctcccctttgttgtcctgctccttcaaagactttctcccccctacca
aagaatagtctctgtacctctcaccatttccccatagggttcgcatacccctcattttcaaatactaaaatcatccaat
tttgtctcaagggtagtgtgcatttgggtgttttggcgtacaaaagcagcaacaaattgtttgacgtacaaacctctta
atcttgataagctcattttataaatggtcggtttttatttaatactcaacatatctaagagttacagtcgatgtagactac
agacttcttatcagcatcttgtgtctcattggagaaagtttccCCAGGGTTAAACAAATGTGATCTT
A

>univin_5p
CTCTCTCTTTCTCGCTCTCTCTttccctctctcaaacacactcacccattctcttcactcctgttccact
cctctcactctttctcctttccttcctttcttcgtcacccagcctctctccccttctccccctttcagggagaacctttcatg
acacggaaaagcccgtgggcttatccatattatgatccagaaaagtatttcgtagcacactttcttctttcgaaagc
attttaacgttttggcatgagacggtgtgaggatctatcgtagccggaaaatatttattcgtcattttgattcaactatct
tcttgttaaaatatcctcaccaaatggaatatagcagcatctctcgagaggagcagagtttgttcaagggtatttca
cttccgttaatatttatggtcgctgtttggaatgtatccagcttaatgtattgtgtcatttaagaaaagggagaatgtttg
gataaagccgcttcattatggccacgtaaatccggttatcggaataatacggagctttgtataatgctaattgtgtaa
ttaatgagcgcagtgtggttccctcatgcttgcggacctcccgtcaatgcttgcaggccaattaggctggctagact
gctgagagctacatccaaactcccattctcaccagctctctgccaacaagaaggtctaccagcgcctcggaca
agctaccaacgcaaccaacaccgtgggaaatcaagaaagactccaaaatattactttgaacgacagaaggg
ttcaagtacgatcacatagctttcgttTGCCACCTTCCAACAGGAGAAGC

>endo16_5p
GTTACGCAGTTTTGTATATCGGatatcgtgacattaatttataatatatgatgacatttttgcttcatattt
tgcggtaccgggggatggtgattttaacatggggataaagatattgcatcaagatttgcacaagctcttgttctaat
atccaccttgccccccccccccccatctgctcccctctcactccctatttcttcttcttcgtcttcttcttctcgtcttctccttt
tccccttttgtacatatatcattttctttatcctctctctctctccctcgtttccatttacaccttcctctgtttctgtttctgtatctc
actctttttttcttacactgacccagttcgcactccctctttatttttccttcacccagactcgatctctctctctctcctgct
ctatatctctctgtatatctgtctctatgtgtgtgtgggtgtgtgtatgtgtgtgtgcgtgcgtgtgtgtgcgtgtgtgtgcgt
gtgtgtgtgtgtgtgtctctctctctctctctctttctctctctctcactctctaagtatatatctccttccccatttttctctttcc
cacttttctctttccccctctcaaatattgataaaaagaatacataatttgggttttctgttgtacgcagaaaacctaaa
tgtcgtattccttcacaaatattcgacttcgaacacattccttgcagaaatgtgtctctaatcacatcctcctaatacat
tatgatacaattttatttaaggaaaatgttgtcgtcaaatgtatggggctcccaacgcttcaaaggggctttaaagtt
atcatatgaatgtaacctaaaccttctgaattactgtcatgatattgggcactgctgggatgattttatcaatgaccaa
accgtaacttttgataaaatgtcattgcgcgtaaagtagacgaccgcccctcctcttccttcctttcgagtcgttgatc
ctccctccaaaaaaagtcttattatgacgaaataataagtatgaatagtattaggaacagatagtatctcgatacg
gagtcttgttgaatacaatgcttatacacgataatgtggacgcactttgcacactgttatctatcaaatttcttcaaag
gaacatgtctagccttgttttaccgacataaacagccagctttggttgctaagtgtagtgcagacggcggatcggg
ttaaagcatgaataagatatgttttaagtctttggttttctctaaactgtgtgtgataacagaagaaaattgatgttttag
ttacaggaacatattttggagtaggtaacaggtattggtatactggctcctggtacactaatgtacactcacactca

aaacgtggcttatggagaagggtttggatttccaattcggggttgtttttgtatatcaaataacaaatgaagagggc
acttctaattcagaaacgttatcatgaataaagactttaactttgttggtttttcaaattaatttcggatgtttttcagtcga
ttccgatgagataaaaccccgaatatacctggaactgaacgttctttgtttccacgaaagatttaatgcccgtaaac
acaaacatctgacaaatgcattaaacttcatcaaaccatgtaacaaaagcgtaattcctcttaaatcgcccttactt
ctaagaacgtcgtaaatcgcaacgtctcaaaaatattgaccaaaatgatcacgccattatcatcgcggaggatt
aagtgatcaagctaccaagagattacatcatttcaaagttatcgcatcccaggttaaactgtttgagttttggattgtg
ctatcaaagacaaaggggtgtaactttacccccCTCATCAAGAGCGGAGGGT

>eve_5p
CAAATACGTAGGCGCACTGATATGAttcccagacgtatttatattaagctagatagtgatgatcac
ttttcctttcacatttcacctaattaagatatattgaagtgagttgttaatggtatatatatcgctaagtaccttcctgtaat
gtataccacatttagtgttcataagctcaaaatgacgattctcatgatttttacggcgcctcaaatctgataaaacatt
caaatgagaaacctgtgttaataataataatactaataataattacatttatagggcgctttttacgggcgtttctata
agcgcacactttgtgttattacatgtcgaataaataaaatcatcttgcctcgtgtatttaagtacactgtcactgatatc
gaattctctcacgatcttgtcaccttaatgctaaacttttacttttcattatcgtgttatgtttgtatacagtgaaatagcat
gttagtacatgtaccaaaacctgtttgatatgtttttccacttgatatgatacatacacgatgtaccatacacctgagg
gcaacgaccccaagaagatgcctggcttttaacgtaaaacggagctcaatttcatctgcttttatcgataaatcca
attagccttatgatcagtgactaatcaagatatacccaaagaaatatcatcgaactcacagcgctacctttaatgtt
ccttaaataccttacaagaaactctctaagttgatcttcgaattaaactgaaaatcgaaactctctttttgttgtttatttttt
agtgtgcgtttgccatcccgacatctttgtgagtaaataatatggagcattgtgatgatgtagactagtgtaggagct
aagcttttctattagtcgaatgaaaggaagatttgggaagatacaacgccaagggctcgtaaaatggtacacag
acttcagtgaatgcatggatagccagaaagctgtaaagttaaagaaacaaagtatatggtattaacaggatattt
ggaggggagacaaaagtatgaatagtttcaatgacgactggctttcttttcgaatcttcgccctttatcacggcggc
gatcagaaagcgaaagctgtaagaaaggaataagacgacccaatttgtgtcatgaatagatatcagtgtctcat
ccccagcatctatatcaacgagccttttatattgttcataggggtcgagcgcggtgccactgtgtgctattgacgtg
taaatgagcgagccttttcaacttgactggttcgatagaacgcgcttttgttgtttttatttgcccgccgtgtccgcacg
ccattgtgcgcgctcgctattgatgactccgtgtaatggagctttgattattacatcgtttagagtcttcggcttggaag
cggctattagcataacaagaggggggaacgccccctcattggtgcgcaaagttcaaaagcactagccccttattg
ggtgttacctaatctggtaaggtcatgaattattttaactcaccgtatttatatcggaagaaagcagcattatcatcatt
ctacaatttcgtcacccaagagagaagttatcatgtcttcatgttttcaagccgccccgatgagaaagtgaactcgt
ctgcattgttaatgtgctactgatatcagttgtaacttgatacgcaagtggtgtttataatcacatcctttaaatatcgcc
gtcgaaaatatcacttggacattttctaactctgccggtttcgacaaccacttcgtcagtgagcctgactggagaag
cactttctgaccggtcgacatcaatttccgaaagtcaaagagcgccgacgataaattgatgtttagatatttcatcc
aatcgagagagagaaagttttaggactagtgcttctcgtcccgatttgatttggagttgatttaaagtcgaggattgg
tcgacagcaagcaaacattgatcgagtggtttctgtttaacatggacaagagtttctccagcaccatgctccccac
gactgagaaccatcatcaagattctaacaacaacgttaccaccacacccagagcatcacccggtggcgtagc
agccccgtctGAAGCTCGTCATCATACGCTCTTT

>wnt8_a
ACCGTGTTATACACCTACAAGTCTTaatattaagtctctcatatcgataagtttacgaatagcaca
gcacacgatgaacacaagtgtcaagcaggccttcatcaataaattcctcttacaatattcaaccgagcagagta
agatatcgacaaattacccttttacaaatggtaatttgtaaaggagtgttttcttatcggacaataaagcgtcccattt
gctgtgaaagaaacggaggtattagatacttagaaagctcttaatcgtacactccaagtcttttaactgcatacaa
ctctcgtcagaaagctctatccagagtttaatctttaatcttaaactattattagtgattgaatatctgtctaaatgcatg
ctcttacgcttcatcatctttttgttcgcttactccttccttttcaaaccagctaccgcctcgaggtcactttaggttccagt
ttgaaaaaaggaagtttaaagaaacgctacccaaactttgagaaagcctgaaatccccttttgttcctgcaaattg

aaagctcggtgggcgatctattcttctcttgctcacgagagcgcgtccttgcaaagagcagacgatgcgaacaa
aagtgcccccaatttagccgccaaaacaatagcgatgcgcctattgctttgatcatcccacggatctattcatgcg
aacatgagctgcattctcattggaccactcatgccttattatgctcgccgctccaatggcaagccgtcatccagctg
tcatgtcaaggtatcatttatgctctttttgatgaggagacaagtcaaaggtgtgtcgatgtaatttataagagctac
aactctgataaacttcattagcttttgagaagttcgaatcgtcaacaagtgtctggtcagctctcaacttttgttagcct
cacagcaggagatagtaattgatacacatcttctgtaaacaccatctcgcatcttatccttaacccttaTCGTGT
TTGAAATACATCACCAT

>onecut_5p
GTTTAACTGTGCTCAGTTATGTAaacatatcgaatttctccccacagtctgtctaatttatcgaacaa
acattttctaatctcccgcccgaaactaggacattcgtgctaagctccaaaagacgttgacctgtgtaggttaacg
cgtagtctgtttaaaactaaagcatatggcaggttaacaatacatggcaatatacccccctgttctgcacacacaga
attgttacaatgaagttgttactttgaacaggtagttagtgagacaattgcgcatgcgttgtggaacgtaacaaagg
ctttcatacatggttgagcgagctgagtcctgtaggggttttacgtgtaatgctattgtaatgtCATTGAGGGG
AGAGAGAGAAAGA

>univin_int
AAGAAGGGCCTCCTCCTtcctatttttcgaaaaaacaggacgctaaacatgtttttgtattaatttggcct
caaacaaaatgacctgacatgaagggctagtatttgtatactctccatctccctctcacgctggttgttcccatcatct
ctccccccccccccccctctctctctctctctctctctcagctacttattcacagtttatggatttcctaagtctcggacat
tttttactcgtctttcgtatctctttattaagttcacggaacgaatttttcattgtcaaaaacgggttatttagtaataagatt
tttaattcgccccgcccccctatttcctcgaaagagatgaaaagatggcgggtaacacttatcacctacccatttgtt
tattcactgacctggtctattgaaaacttcaataccttggtcttgttagttaaagacaagattggaaacgcgatgggt
tattttcaagaaagcggatcccagtgtaattaattacttattcaataactttcgatcaacgagcttattaattccttgtac
ctcgtttaatcctgaagtgggtcaaacaaagaggtttcGAAGCTGCGTGGCTCAGAAGC

>alx_j
CAACCCCCGGACACAAAAGGtcattttggcggtaacatttcctatgatgtttaagggtcatttcttttgag
aaacgcgaaagcacctgatttgcactcttgaccaatgaccgtgcccgaagcccagcggtgtataatagcacaa
aaaggatgccgctcgccctattcataacagatacacacgcacacacacagcgtcactcgtgaagtcctattgtat
tgtgtgtacggtgcactgtacctacccaaccgaggactgctcgcgtgcgagtggtgtgtggatgtgtgtacagaa
agaatactcgcacaatctgtataaaaactttggcacgcgcgGGTTTAGATACCACAACTTGTTTC
TTC

>irxa_u1
TCCATGTCGAGGACAAGCTATTaactttcagacggttaagatgagtagtgattgcgtaacagttaa
tgatgctaatttctggctgctaaattcgtgtaactaaaagcggaagaaaggccttgttaatggtgacttgattgttttat
ggcctagtaagtcacccgggagcggaattcccataataaatcgagacacctgagaaacctgtttttctttaaata
aaatgaaggtgacaaataattagctatcatgtaacgacaggcgctctgacatcccatcttttcaacaacaggata
caaataaatacatgtcagaatagattcagataaaagtagttatttttgtgtgtttatttcattttattgattgtctgtaatg
gaaaatgcgtaatgttatttttaacattaactgcagggtgaaacctggatcacttaaagacatataatgatattttca
taaagttttaaggcatcgacaatattgtttggcataaaagtgtaatactgttacaacacgaatgtaaccaatgaattt
ttcgaaaacgtttaccgaacgcctatatctttatgagcattacttcataatgatctacatttcactcgtttaacatctgct
aagggttggtgtacatatcggttaaacattaattactttaacttgaaggaaattagaaccatgcaaacccaccaaa
gatctcaaaagtatgactttgtataaatgaaagtgcactgagcatgaatatatatttcataaatacttgaaaaagca
tttatacatctattgaaattaagatcaataatattatatatctgaaaggcactttaaatacaaaatagtagagataga

gaaaaaaagatgatgcaaagatgcttacagacacatacaaataaacaagtgaatatgataggagagagaa
agagagagagagagagagagagagaaagaaaaaaagagagagaaagaaagagagagagagagag
agagagagagagagagagagaaatagcgacaatttctagacagaagcataaagttgggggtgagttgtaaa
tctcggatcatttttactctgcttcaattcgaacagagatggcgtaataacatccagcctagggagctgacggcga
gtgtatacacagcatcatgcatatagtcgttcatcgcatgcatgagcatatgttttacatcagtcccaggcatgatac
ttgttatgattatttcaatatataattcttttaggaacgctccctaataaataaacacatattcatcaaactcccacaaa
aatgcttgagtttgaaattaattatgagatttcacacccccaaaacacaggcttccataaaaagaattgatttctga
gcagaggcctatcaaaataaaataaaacgtttcagttttttaaattttaatttcttgaatgtgatgactattttctacttctt
ggaatgtatctatatttgaaattcattaaagcaattccaatatacatgataataataactatcactcgtagtttagtaa
agaaatgccagcgctaaattaaatttagatgattagactattttgacaataacttatcttggagcaacaaaaaaaa
caaaaagattcacgtgatttatttaaaacatgaatttctttaatacaaccagttttgatagaccaccacttatacttag
cgatgtgaacttggacagtggacgtgttgttgctattgccatgtcgcagagttcggcaaatacgcgagtctatactc
tgtccgcctcgctaggcgtaaagctatggcagccataaactgacaattttcaacaatctaataattcagcccgtct
cgcattgggcgatggtaaaataatttcacgaagagggaaaaaaactctcattcacttaggagactggttagcttc
gcacccgccaatcatgatgacgtagcacgataagtgtctgcttaatgggtaagggttgttgtgattggctgcttaaa
gtgattgagcggtacacgtcacagtagtctgggcggggcttgtgcatggctgagttgggtgactgagcgcccctg
tgtatggcgataacggggctactggtatggtgtagcatggtatgtgtattagggggtagcggtcatagacaaggg
ggaaacgacccatgggtctgaaaaggtctagctttagagacgcgcagggctgacacaaactcgtgaaaggc
gagcaaagcgagtagcgcacGGGCCTCTAACCATTTCAGATG

>gatae_m10
AGATTTGTATTCACCACTCGCagaagccattactttgataaattagttgtacatcactaaatcaatatt
caatattctggaactttgagcaaaatccgtttttgttttagcggcaaaccggcgggctaaagcacacgaactcaa
acataaatcttggcacgtgtacccgtagattggcggtaagctttgtagtcattctactggcaccatccttacatacac
tcattctcccacctccttttttctctttgtcgctaaatttgattttgaagttggtttgttcttctacccgtgtatcgttgttcatcg
agcaaatacgtatgttgtttgtcgtcgcacttgagccgcgaaataatcacagcactacaaagaaaacactccgg
cgcaaactgctgaaaacaaatcgtatgaatcatgtgaactgtgttttatgaatgccatcgcaagcagacaattca
ctttgctttgtaaagctttgtacttctatgaaagcgatctcgtacagttacaagggattagtaagagattaatggatga
acctggacaaatgctagaataaaagggtagcttgggatTACAACCTTTTTGATTAGCGCCGC

>irxa_i2
ACGACTTTGCGAATGGATTATGatattcataactcaccaatcttggccggagggaagatttgatcg
agaagaaagcttccctttctccatcaactctaagctcatcaattatctctcattcattgaaatgagaaatgatccaaa
tttggcacgaagaaaagagagaaataaattataccgcatcattgttatctcacacaataacgtcagaaggcgat
atggatagacagattttatcaaagaaattctggaatataatttttcttcctttctatctatcttcgtctctccatttctctctcc
gactctcttttatatttctatacttgctacaaagtttaacaattacgtttctaatagtcaggatatataatccctctaaagtt
cgctgtaaatggcgtgaaaaaaagagatcttcttataataatgtgcggttagcatataataatatgtggttaagata
tatgagatgttaaggtgtaaaagaaatatctggcagggaaaaggaaggataacgttggtcaaattaagaggatt
tcattgatgactttgtaaattcaccaggggatgataaaaggccacatcgcgtttgattaactagtcaataatattgt
taatctttcaacagaggtttatttttcaaaattttatggtgtttttgagaaatgttttgaccatgcaaactaagaaatcaa
ataaagtatatcaaatttcaaaataagacaagtcatttataatgatcatcagttctgtatcgaggtgttttagataaac
cagattaactataaactagtctttatgaaaagtcatgatgttgcattttgttgatttaattctaacaaaaaacaattaaa
cttggaatggattcgccttaatgtataatgactatctagatgtattggcaaacaatcacttgatatctttggaatgatct
aaatgtagttgaataagtagtagtaattttatgcaaacttggtgtggtaattgttagttgttattccgagcatatgatgta
aaaggagagatggtttggatggataaaaatgatatttgtgatcaattaggaatcaaagatgcgagcaaatggag
agtttgactttcatctggttatagcatcaattatttaggccatccccctttccatgactgcctccatctctctctccctatgt

ccctgtctctctctctcttcacctctcttttcttctctcctccgtccgagttcctcccatttcctctcttcacctctccctctccct
ctctctcatctcccctgtctctgtgtctctgtctgtctgtctcattgccccctctctgtttctctccccaccctctctctatatcc
ctctatatccctctctatccctctcgtctctttctctcccttcattttcttaatatttcgcgtgagttctttcatgaccaatgag
gggcttaatgggcgatcataagtatgatgccccgggcaatagtatatattgaatataccctctaaagacagtgat
atagtacaaaaatataccctctctatagagaccatgattaataacgatcattttcatctacactctcacatctcagtt
cgataaaccctcagcttaatagtttaatttcaaacagcatcgttgttatttttcatttaccaatcgagctgatcttataga
tatagttcggatatggttgtctttttatttgtttaaaaaccaatttgtgaaaactcgatacgaataaaaacctatggatat
aaacgaaagattgtggtttgcattaaacgaggaagataattacatgagcgcgtttatggacaagtttaacaataa
atatagaaataaacaattttgtagaatttggtgctttgttgttttagatattgataatatgcaaactaaaccttcacttga
caacctgggggtgaataatgattaataaagtacttgaaacgggcattgcatatttacattaagaagagcgtgtaa
aacctgcgtgcatggtgggtttaagccgcaacacacaatcattctacattctaatccagaattatcattatatcaga
agccatgcggtctgaatggagaaaaagcctgttacttgctcagcaacaatacaaaaacgaaacatgagattta
ggacggggacaccatccatatccaaaagttatcttaaagctcctcataaccgatgtaattttcgaggccaacattc
aacattgcattcatcgcaccaaacgacttaaccgtgcttttaagcagagccccttcttacttcttttcatttttttgccttgt
cgatcaagcatttaatggtagagacaagcgaagaattgtgtgccgagcgatttaatggcgaacaataccgatcc
gatgcaaaccgatcgatacaagaggcattgtccactacagatggatgtaaagcttaaggaatgaaaagcaag
agagttttgaatacaaatgatccacgtatatagataattcGAATTTGTATACGGCAAGTCCG

>tbr_g
TATAGGACCGTGTTATATACCTCtctaccccccccctcccccctccaaaaaaaaaaacccagca
acattctttataatttctaacattaatttctagccgatgaaatcattaataattgtgtccgagctgcttttttcttctgtctcttc
ttctccttcttcagtacattcttcgtcttttccttttcatctcacttttcttcttcttctctttcttctttttctttttgaatctcctcttttttctt
ctttttcttcttcttatgctacttcttcttcttcctcctctttccttgttctaagacttgttcttctcctttattctcattcccctcttttttta
tggtctctagcaagatatgttactatttttttaaataatgtaagttatatcatttacacaattaacaagtccatttattcgat
catcgatgtgtgaatgtagggctacattgcaaaaacaaaataattataaggcccctgtgcgtgctttacagtgata
acacaaagcggtcattgacctcttcctgtctggtgatcggtcaactaatcccttccggttggacgtgaacttcgacc
gctggtttactcatgatgtcgtcacatgagatttcatggaatttattccatataattatagatatatgacaccagtagca
gtattagtagagtacagtagtggtttatagtaggttagtagtagtagtagtagtacagtagtagtagtagtagtagta
gtagtagtagtagtagtagtagtagtagtagtagtatagtagtagtagtagtagtagtatagtagtagtagtagt
agcagtagtagttgttgttgttgttgttcttgttgttgtaatattgggtagaagtatagcagcaagagaaatagaagta
gccgtagcactatgtgtgcatgactttgctttaaaaacagatccagacataatcattatagacatagtttacatgctg
aagatgggtatccatctgaaaggatcgataaaatacagatttgactatcatatctggaaacttaagccctccctcc
tcacctccctcacccttggctccgctggttattcaatgctcaagtgagtaaaacgtgtaaggcttttcaacttttcacg
ataataaagacaataattaccatcttggcatcttgttaaagaaaccttttgtacattgcacttttttttttttacatggcaga
agtttcatgttatgcaatataattatatgtatgcttacgtctaagaggtttaccttaaactgctcttttcagttcacagttatt
tagatctaaatttcgtatttttaacccttttcctgaataaatattttatatccagatttgcctctattatacctatcaaaacac
aatgacatatggtggttccacaaatatatgttcatttttgttctagatggttattcttccagactatttttttcttcttcgtcgtc
gtctaaatgttatttcgagtcgcgtgcctccctggcggccaatcaacgcgcgactctccattgtctggtagccccac
gagatctgtctgaagtgcctcattaacaagtcccctaagtggttgacttactcatgacgtaatacaaagaaagtca
gccaagcgaagcgtgtgctcaatccatgcaaagttattgaacaaaacttgtgggcggtttcggtcacccttttttgtt
acagaagcattcgcagctttatttttgacagttccttgcaaatatgatggcagttcgtctcacgcctaaatgagatga
tcacatacaaatcgcaccgggggaatttcggaaaaagtgttaaaatcgcagtgagaatttcatcagcgttcgcg
ccttctcgcttctgtgtttatccatgtaatttgtgactgaatttTCGCACTCCGACTCTAACCCT

>delta_5p
CTGCCAGCCCCTATCTATTGgtatgtttagtaaagtgtggagaaatgaaaaactcaaaattgtcttttt
ggtggatatttatgactacaatgaccttgtaagttgtagtagtggcacgtgcttggaacagggctagcccgtatttgc
atatcattacaggaagaagtcattcacaaacgtacgggtttccggcttttatcgcgcgctctgtgaatacggttctct
gattggccgcagcaatcgtcaatgaatacactcgggtcaggggctctcttgcagaactgggaagaaaataata
gtaaaaccagacaaccgaacgggtgcattaagaagggaaaaagcaagacaaccaccctaacaacaatag
cagactaacaaacgaagcgcaggtgcaggcaagttgcttggacccagctgaccacacaatttgaatagattg
aaagaacccagtgatgaaaaaccttctcgtggacccattcaatctattgacagggggcacgcgtcccgcaca
atagcagcacacttgctgacgagggtagggcctcgttgaacaaaagaactttggagaagttgaggggggagcct
cagaacaggaaaaaaaagtgagggacaatgctaatataaagacctggccactgagatacgtgtgccgccgtt
atacctcctctttgttggagaataaccaactgaaatacatcagactgatggaatttgaacggtactgtaactattttc
cagaaagagccacaacttatgccaagttctacacagaaaattatgaagtttatcgaatgtgatgaaaaataatta
tacatcagctgactaaatataatataaattttaaaaataaaaaagggtaaaagatggaatctgacaggttcaata
ttcataaaacgatttctagagtccattagttacttgactgttagcttttagatgtttttataatttgtcccgtatgaatgaa
catgttgtattggtagtaaggtttaggcttcttgagtttgaaggcacgcgaactaatgtgttatcctttacaaactaga
gctaacacaaagtcctataaactgccaagccagcatgcaattagccataaaatgaaggttttagttaacaaaca
taaaagaagagctcagtgaggaggaaaaaacaggcccggcttgttttgtagaaggcacatcgataatgaaca
ttggaccgtggcaagagataatccggatttaaagcacacacaaaaaaatccagcaccacgcgaccacgggc
cataatcctatgcatatgcaaaacattggcacgagccgagggctggtatataaaggagagcgtcttgctcagtc
ggggttattaaacaaaaagcagtcctcacgaacagtcaaacttttgatatacttggaacttacgaagccagcattt
cgcaacaagtgatacgacattatacctggatttacaatctgacagtcggattatacacacgtgcatccttctcggc
gatgataaagtaattttcgcgagttttcatcaatttctacatttatatcgtatggatgagcaacattttagttagatatata
tttcgaatatatttttttatttgagttggagtatatacacagtactgtcaaatttgacattatacttgtgcaacatgagaact
tggcatcagtgacaaaagtagattgccgccctctctcgtgattatttgttcagtctatcctcgaaaccaaggcacgc
gtgtgggagctctgattggtcggctgatcggcttgtccacacttatgggcggggcgtagtaaacgcagagagag
agagcccgaaaatgacaaagagagcaaacggagcaagctaagttgatcaaaGCTCGCAGACTTG
AGAGGACTG

>gcm_e
AACCTTGCAGTACACACCTGCgaactcaactttatcttgaagtgacgtcatctttaatattaaaaaca
caacattctagtataagtactcaataaatcgagcaaaaggggatgggagagtcggttcattagtcttcccacgcc
aacattgtcaacagtataataggcctatatgtttactgggtccctttaaatgtgtttgcaaacaatcacctttgcgga
cttgtgctttcttcgtagatcggcaacattcaatgaactctttcggacttcattttctttgtgtaccgctccgatcctatattt
ttctcacgccttcttaacgaacacaaAGCCGCGCGTAACGGTACC

Supp. 8: Supporting File 3 for Single embryo-resolution quantitative analysis of reporters permits multiplex spatial *cis*-regulatory analysis

Supporting File 3**.** List of CRMs with smallest ***D*** value (cutoff = 0.035).

Group (Fig.5C)	CRM with ID	CRM with ID	*D* value	
1	CyIIIa_5P_ID1	Univin_5P_ID1	0.033514	
1	CyIIIa_5P_ID2	Nodal_5P_ID3	0.023593	
1	Tbrain_C_ID1	Nodal_5P_ID6	0.02325	
1	Tbrain_C_ID2	Nodal_5P_ID6	0.030461	
1	Tbrain_C_ID3	Univin_5P_ID2	0.01601	
1	Tbrain_C_ID4	Nodal_5P_ID6	0.028373	
1	Tbrain_C_ID5	Univin_5P_ID2	0.022333	
1	Univin_5P_ID4	Nodal_5P_ID2	0.029472	
1	Nodal_5P_ID1	Ese_3P_ID_ID1	0.019755	
1	Nodal_5P_ID2	Nodal_INT_ID1	0.021224	
1	Nodal_5P_ID4	FoxJ_5P_ID1	0.027504	
1	Nodal_5P_ID5	Nodal_INT_ID1	0.028867	
1	Nodal_INT_ID3	Tbrain_C_ID5	0.028758	
1	Ese_3P_ID1	FoxJ_5P_ID2	0.018464	
1	Ese_3P_ID3	Nodal_5P_ID3	0.030672	
1	Ese_3P_ID4	Nodal_5P_ID1	0.024707	
1	Ese_3P_ID5	Gpc1246_INT_ID1	0.030335	
2	Endo16_All_ID1	Onecut_5P_ID2	0.017012	
2	Endo16_All_ID2	Wnt8_A_ID1	0.026612	
2	Endo16_All_ID3	Wnt8_A_ID2	0.021914	
2	Endo16_All_ID4	Wnt8_A_ID4	0.023933	
2	Endo16_All_ID5	Wnt8_A_ID2	0.023023	
2	Onecut_5P_ID3	Endo16_All_ID1	0.028795	
2	Eve_5P_ID1	Endo16_All_ID1	0.025395	
2	Eve_5P_ID2	Onecut_5P_ID1	0.033852	
2	Eve_5P_ID3	Onecut_5P_ID1	0.024623	
2	Eve_5P_ID4	Univin_INT_ID2	0.025701	
2	Eve_5P_ID5	Onecut_5P_ID2	0.021139	
2	Onecut_5P_ID1	Endo16_All_ID1	0.020598	
2	Univin_INT_ID1	Eve_5P_ID4	0.029797	
3	Alx_J_ID3	IrxA_5P_ID4	0.022177	
3	IrxA_5P_ID1	Alx_J_ID3	0.02894	
3	IrxA_5P_ID2	Alx_J_ID3	0.023552	
4	Gatae_m10_ID1	IrxA_5P_ID4	0.044848	> 0.035
4	Gatae_m10_ID4	IrxA_5P_ID4	0.045082	> 0.035
4	Gatae_m10_ID5	IrxA_5P_ID4	0.059685	> 0.035
5	IrxA_I2_ID1	Eve_5P_ID1	0.056772	> 0.035
5	IrxA_I2_ID2	Endo16_All_ID1	0.056922	> 0.035

7. WORKS CITED

Akhtar, W., Pindyurin, A.V., de Jong, J., Pagie, L., Ten Hoeve, J., Berns, A., Wessels, L.F., van Steensel, B., van Lohuizen, M., 2014. Using TRIP for genome-wide position effect analysis in cultured cells. Nature protocols 9, 1255-1281.

Arnold, C., Hodgson, I.J., 1991. Vectorette PCR: a novel approach to genomic walking. PCR methods and applications 1, 39-42.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., Stark, A., 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339, 1074-1077.

Arnone, M.I., Dmochowski, I.J., Gache, C., 2004. Using reporter genes to study cis-regulatory elements. Methods Cell Biol 74, 621-652.

Arthur, W., 2011. Evolution : a developmental approach. Wiley-Blackwell, Chichester, West Sussex.

Bannert, N., Kurth, R., 2004. Retroelements and the human genome: new perspectives on an old relation. Proceedings of the National Academy of Sciences of the United States of America 101 Suppl 2, 14572-14579.

Barakat TS, H.F., Zhang M, Rendiero AF, Bock C, Chambers I, 2017. Functional dissection of the enhancer repertoire in human embryonic stem cells. bioRxiv 146696.

Barsi, J.C., Tu, Q., Davidson, E.H., 2014. General approach for in vivo recovery of cell type-specific effector gene sets. Genome research 24, 860-868.

Berretta, J., Morillon, A., 2009. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO reports 10, 973-982.

Berretta, J., Pinskaya, M., Morillon, A., 2008. A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in S. cerevisiae. Gene Dev 22, 615-626.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Burke, R.D., Moller, D.J., Krupke, O.A., Taylor, V.J., 2014. Sea urchin neural development and the metazoan paradigm of neurogenesis. Genesis 52, 208-221.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W.S., Henke, J., Makalowski, W., Jorde, L.B., Deininger, P.L., Batzer, M.A., 2001. Large-scale

analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol 311, 17-40.

Cohen, C.J., Lock, W.M., Mager, D.L., 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448, 105-114.

Costantini, S., Di Bernardo, G., Cammarota, M., Castello, G., Colonna, G., 2013. Gene expression signature of human HepG2 cell line. Gene 518, 335-345.

Crocker, J., Stern, D.L., 2017. Functional regulatory evolution outside of the minimal even-skipped stripe 2 enhancer. Development 144, 3095-3101.

Davidson, E.H., 2006. The regulatory genome : gene regulatory networks in development and evolution. Academic, Burlington, MA ; San Diego.

Davis, K.L., 2011. Ikaros: master of hematopoiesis, agent of leukemia. Ther Adv Hematol 2, 359-368.

de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics 7, e1002384.

Dickel, D.E., Zhu, Y., Nord, A.S., Wylie, J.N., Akiyama, J.A., Afzal, V., Plajzer-Frick, I., Kirkpatrick, A., Gottgens, B., Bruneau, B.G., Visel, A., Pennacchio, L.A., 2014. Function-based identification of mammalian enhancers using site-specific integration. Nature methods 11, 566-571.

Dunn, C.A., Romanish, M.T., Gutierrez, L.E., van de Lagemaat, L.N., Mager, D.L., 2006. Transcription of two human genes from a bidirectional endogenous retrovirus promoter. Gene 366, 335-342.

Enuameh, M.S., Asriyan, Y., Richards, A., Christensen, R.G., Hall, V.L., Kazemian, M., Zhu, C., Pham, H., Cheng, Q., Blatti, C., Brasefield, J.A., Basciotta, M.D., Ou, J., McNulty, J.C., Zhu, L.J., Celniker, S.E., Sinha, S., Stormo, G.D., Brodsky, M.H., Wolfe, S.A., 2013. Global analysis of Drosophila Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. Genome research 23, 928-940.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., Kellis, M., 2016. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nature biotechnology 34, 1180-1190.

Feigin, M.E., Garvin, T., Bailey, P., Waddell, N., Chang, D.K., Kelley, D.R., Shuai, S., Gallinger, S., McPherson, J.D., Grimmond, S.M., Khurana, E., Stein, L.D., Biankin, A.V., Schatz, M.C., Tuveson, D.A., 2017. Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. Nature genetics 49, 825-833.

Ferrara, A.M., Pappa, T., Fu, J., Brown, C.D., Peterson, A., Moeller, L.C., Wyne, K., White, K.P., Pluzhnikov, A., Trubetskoy, V., Nobrega, M., Weiss, R.E., Dumitrescu, A.M., Refetoff, S., 2015. A novel mechanism of inherited TBG deficiency: mutation in a liver-specific enhancer. The Journal of clinical endocrinology and metabolism 100, E173-181.

Flytzanis, C.N., Britten, R.J., Davidson, E.H., 1987. Ontogenic activation of a fusion gene introduced into sea urchin eggs. Proceedings of the National Academy of Sciences of the United States of America 84, 151-155.

Fortini, B.K., Tring, S., Plummer, S.J., Edlund, C.K., Moreno, V., Bresalier, R.S., Barry, E.L., Church, T.R., Figueiredo, J.C., Casey, G., 2014. Multiple functional risk variants in a SMAD7 enhancer implicate a colorectal cancer risk haplotype. Plos One 9, e111914.

Garritano, S., Romanel, A., Ciribilli, Y., Bisio, A., Gavoci, A., Inga, A., Demichelis, F., 2015. In-silico identification and functional validation of allele-dependent AR enhancers. Oncotarget.

Gerets, H.H., Tilmant, K., Gerin, B., Chanteux, H., Depelchin, B.O., Dhalluin, S., Atienzar, F.A., 2012. Characterization of primary human hepatocytes, HepG2 cells, and HepaRG cells at the mRNA level and CYP activity in response to inducers and their predictivity for the detection of human hepatotoxins. Cell Biol Toxicol 28, 69-87.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature methods 6, 343-U341.

Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., 3rd, Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., Gamble, C.E., Iagovitina, A., Singhania, A., Michelson, A.M., Bulyk, M.L., 2013. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nature methods 10, 774-780.

Gorman, C.M., Moffat, L.F., Howard, B.H., 1982. Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. Mol Cell Biol 2, 1044-1051.

Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017-1018.

Guay, C.L., McQuade, S.T., Nam, J., 2017. Single embryo-resolution quantitative analysis of reporters permits multiplex spatial cis-regulatory analysis. Developmental biology 422, 92-104.

Halfon, M.S., Zhu, Q., Brennan, E.R., Zhou, Y., 2011. Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. BMC Genomics 12, 578.

Hall, C.V., Jacob, P.E., Ringold, G.M., Lee, F., 1983. Expression and regulation of Escherichia coli lacZ gene fusions in mammalian cells. J Mol Appl Genet 2, 101-109.

Hardison, R.C., Taylor, J., 2012. Genomic approaches towards finding cis-regulatory modules in animals. Nature reviews. Genetics 13, 469-483.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362-9367.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006a. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Developmental biology 300, 90-107.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006b. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. Developmental biology 300, 74-89.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Tu, Q., Oliveri, P., Cameron, R.A., Davidson, E.H., 2006c. High regulatory gene use in sea urchin embryogenesis: Implications for bilaterian development and evolution. Developmental biology 300, 27-34.

Iolascon, A., Faienza, M.F., Centra, M., Storelli, S., Zelante, L., Savoia, A., 1999. (TA)8 allele in the UGT1A1 gene promoter of a Caucasian with Gilbert's syndrome. Haematologica 84, 106-109.

Jeong, W.S., Keum, Y.S., Chen, C., Jain, M.R., Shen, G., Kim, J.H., Li, W., Kong, A.N., 2005. Differential expression and stability of endogenous nuclear factor E2-related factor 2 (Nrf2) by natural chemopreventive compounds in HepG2 human hepatoma cells. J Biochem Mol Biol 38, 167-176.

Juven-Gershon, T., Cheng, S., Kadonaga, J.T., 2006. Rational design of a super core promoter that enhances gene expression. Nature methods 3, 917-922.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., Dunham, I., Elnitski, L.L., Farnham, P.J., Feingold, E.A., Gerstein, M., Giddings, M.C., Gilbert, D.M., Gingeras, T.R., Green, E.D., Guigo, R., Hubbard, T., Kent, J., Lieb, J.D., Myers, R.M., Pazin, M.J., Ren, B., Stamatoyannopoulos, J.A., Weng, Z., White, K.P.,

Hardison, R.C., 2014. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America 111, 6131-6138.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., Kellis, M., 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome research 23, 800-811.

Kieusseian, A., Chagraoui, J., Kerdudo, C., Mangeot, P.-E., Gage, P.J., Navarro, N., Izac, B., Uzan, G., Forget, B.G., Dubart-Kupperschmitt, A., 2006. Expression of Pitx2 in stromal cells is required for normal hematopoiesis. Blood 107, 492-500.

Koressaar, T., Remm, M., 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23, 1289-1291.

Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., Cohen, B.A., 2014. High-throughput functional testing of ENCODE segmentation predictions. Genome research 24, 1595-1602.

Laghi, L., Randolph, A.E., Malesci, A., Boland, C.R., 2004. Constraints imposed by supercoiling on in vitro amplification of polyomavirus DNA. The Journal of general virology 85, 3383-3388.

Lee, J.I., Taichman, L.B., 1989. Transient expression of a transfected gene in cultured epidermal keratinocytes: implications for future studies. J Invest Dermatol 92, 267-271.

Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S.W., Heutink, P., Hill, R.E., Noji, S., 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proceedings of the National Academy of Sciences of the United States of America 99, 7548-7553.

Levine, M., Cattoglio, C., Tjian, R., 2014. Looping back to leap forward: transcription enters a new era. Cell 157, 13-25.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Liang, K.C., Tseng, J.T., Tsai, S.J., Sun, H.S., 2015. Characterization and distribution of repetitive elements in association with genes in the human genome. Comput Biol Chem 57, 29-38.

Lin, C.Y., Su, Y.H., 2016. Genome editing in sea urchin embryos by using a CRISPR/Cas9 system. Developmental biology 409, 420-428.

Lower, K.M., Hughes, J.R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., Goriely, A., Ayyub, H., Sloane-Stanley, J., Vernimmen, D., Langford, C., Garrick, D., Gibbons, R.J., Higgs, D.R., 2009. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proceedings of the National Academy of Sciences of the United States of America 106, 21771-21776.

Ludwig, M.Z., Manu, Kittler, R., White, K.P., Kreitman, M., 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS genetics 7, e1002364.

Luizon, M.R., Ahituv, N., 2015. Uncovering drug-responsive regulatory elements. Pharmacogenomics 16, 1829-1841.

Lynch, V.J., Nnamani, M.C., Kapusta, A., Brayer, K., Plaza, S.L., Mazur, E.C., Emera, D., Sheikh, S.Z., Grutzner, F., Bauersachs, S., Graf, A., Young, S.L., Lieb, J.D., DeMayo, F.J., Feschotte, C., Wagner, G.P., 2015. Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. Cell reports.

Madrigal, P., Krajewski, P., 2012. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Frontiers in genetics 3, 230.

Maniatis, T., Goodbourn, S., Fischer, J.A., 1987. Regulation of inducible and tissue-specific gene expression. Science 236, 1237-1245.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747-753.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., Loh, M.L., Hunger, S.P., Sanda, T., Young, R.A., Look, A.T., 2014. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 346, 1373-1377.

Mao, C.A., Wikramanayake, A.H., Gan, L., Chuang, C.K., Summers, R.G., Klein, W.H., 1996. Altering cell fates in sea urchin embryos by overexpressing SpOtx, an orthodenticle-related protein. Development 122, 1489-1498.

Mariani, L., Weinand, K., Vedenko, A., Barrera, L.A., Bulyk, M.L., 2017. Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. Cell Syst 5, 187-201 e187.

Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., Levine, M., 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. Development 131, 2387-2394.

Martik, M.L., Lyons, D.C., McClay, D.R., 2016. Developmental gene regulatory networks in sea urchins and what we can learn from them. F1000Res 5.

Materna, S.C., 2017. Using Morpholinos to Probe Gene Networks in Sea Urchin. Methods Mol Biol 1565, 87-104.

Mathelier, A., Shi, W., Wasserman, W.W., 2015. Identification of altered cis-regulatory elements in human disease. Trends in genetics : TIG 31, 67-76.

McMahon, A.P., Novak, T.J., Britten, R.J., Davidson, E.H., 1984. Inducible expression of a cloned heat shock fusion gene in sea urchin embryos. Proceedings of the National Academy of Sciences of the United States of America 81, 7490-7494.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., Kellis, M., Lander, E.S., Mikkelsen, T.S., 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nature biotechnology 30, 271-277.

Melnikov, A., Zhang, X., Rogov, P., Wang, L., Mikkelsen, T.S., 2014. Massively parallel reporter assays in cultured mammalian cells. Journal of visualized experiments : JoVE.

Mortazavi, A., Pepke, S., Jansen, C., Marinov, G.K., Ernst, J., Kellis, M., Hardison, R.C., Myers, R.M., Wold, B.J., 2013. Integrating and mining the chromatin landscape of cell-type specificity using self-organizing maps. Genome research 23, 2136-2148.

Muerdter, F., Boryn, L.M., Woodfin, A.R., Neumayr, C., Rath, M., Zabidi, M.A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R.R., Schernhuber, K., Arnold, C.D., Stark, A., 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. Nature methods 15, 141-149.

Murtha, M., Tokcaer-Keskin, Z., Tang, Z., Strino, F., Chen, X., Wang, Y., Xi, X., Basilico, C., Brown, S., Bonneau, R., Kluger, Y., Dailey, L., 2014. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nature methods 11, 559-565.

Nam, J., Dong, P., Tarpine, R., Istrail, S., Davidson, E.H., 2010. Functional cis-regulatory genomics for systems biology. Proceedings of the National Academy of Sciences of the United States of America 107, 3930-3935.

Nam, J.M., Davidson, E.H., 2012. Barcoded DNA-Tag Reporters for Multiplex Cis-Regulatory Analysis. Plos One 7.

Nei, M., Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford Press, New York, New York.

Nocente-McGrath, C., Brenner, C.A., Ernst, S.G., 1989. Endo16, a lineage-specific protein of the sea urchin embryo, is first expressed just prior to gastrulation. Developmental biology 136, 264-272.

Pai, A.A., Pritchard, J.K., Gilad, Y., 2015. The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLoS genetics 11, e1004857.

Parker, H.J., Bronner, M.E., Krumlauf, R., 2014. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. Nature 514, 490-493.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., Shendure, J., 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nature biotechnology 27, 1173-1175.

Pennacchio, L.A., Rubin, E.M., 2001. Genomic strategies to identify mammalian regulatory sequences. Nature reviews. Genetics 2, 100-109.

Peter, I.S., Davidson, E.H., 2011. Evolution of gene regulatory networks controlling body plan development. Cell 144, 970-985.

Pickrell, J.K., 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet 94, 559-573.

Plank, J.L., Dean, A., 2014. Enhancer function: mechanistic and genome-wide insights come together. Molecular cell 55, 5-14.

Recillas-Targa, F., 2006. Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals. Mol Biotechnol 34, 337-354.

Rincon, M.Y., Sarcar, S., Danso-Abeam, D., Keyaerts, M., Matrai, J., Samara-Kuko, E., Acosta-Sanchez, A., Athanasopoulos, T., Dickson, G., Lahoutte, T., De Bleser, P., VandenDriessche, T., Chuah, M.K., 2015. Genome-wide Computational Analysis Reveals Cardiomyocyte-specific Transcriptional Cis-regulatory Motifs That Enable Efficient Cardiac Gene Therapy. Molecular therapy : the journal of the American Society of Gene Therapy 23, 43-52.

Roussos, P., Mitchell, A.C., Voloudakis, G., Fullard, J.F., Pothula, V.M., Tsang, J., Stahl, E.A., Georgakopoulos, A., Ruderfer, D.M., Charney, A., Okada, Y., Siminovitch, K.A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P.K., Plenge, R.M., Raychaudhuri, S., Fromer, M., Purcell, S.M., Brennand, K.J.,

Robakis, N.K., Schadt, E.E., Akbarian, S., Sklar, P., 2014. A role for noncoding variation in schizophrenia. Cell reports 9, 1417-1429.

Royo, J.L., Maeso, I., Irimia, M., Gao, F., Peter, I.S., Lopes, C.S., D'Aniello, S., Casares, F., Davidson, E.H., Garcia-Fernandez, J., Gomez-Skarmeta, J.L., 2011. Transphyletic conservation of developmental regulatory state in animal evolution. Proceedings of the National Academy of Sciences of the United States of America 108, 14186-14191.

Sakabe, N.J., Savic, D., Nobrega, M.A., 2012. Transcriptional enhancers in development and disease. Genome biology 13, 238.

Savic, D., Partridge, E.C., Newberry, K.M., Smith, S.B., Meadows, S.K., Roberts, B.S., Mackiewicz, M., Mendenhall, E.M., Myers, R.M., 2015a. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. Genome research 25, 1581-1589.

Savic, D., Roberts, B.S., Carleton, J.B., Partridge, E.C., White, M.A., Cohen, B.A., Cooper, G.M., Gertz, J., Myers, R.M., 2015b. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/ enhancer-binding protein beta binding sites. Genome research 25, 1791-1800.

Shelton, D.A., Stegman, L., Hardison, R., Miller, W., Bock, J.H., Slightom, J.L., Goodman, M., Gumucio, D.L., 1997. Phylogenetic footprinting of hypersensitive site 3 of the beta-globin locus control region. Blood 89, 3457-3469.

Shlyueva, D., Stampfel, G., Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15, 272-286.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J.A., Dean, M., Elphick, M.R., Ettensohn, C.A., Foltz, K.R., Hamdoun, A., Hynes, R.O., Klein, W.H., Marzluff, W., McClay, D.R., Morris, R.L., Mushegian, A., Rast, J.P., Smith, L.C., Thorndyke, M.C., Vacquier, V.D., Wessel, G.M., Wray, G., Zhang, L., Elsik, C.G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M.J., Mackey, A.J., Maglott, D., Panopoulou, G., Poustka, A.J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C.A., Worley, K., Durbin, K.J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A., Satija, R., Severson, T., Gonzalez-Garay, M.L., Jackson, A.R., Milosavljevic, A., Tong, M., Killian, C.E., Livingston, B.T., Wilt, F.H., Adams, N., Belle, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Geneviere, A.M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A.J., Goldstone, J.V., Cole, B., Epel, D., Gold, B., Hahn, M.E., Howard-Ashby, M., Scally, M., Stegeman, J.J., Allgood, E.L., Cool, J., Judkins, K.M., McCafferty, S.S., Musante, A.M., Obar, R.A., Rawson, A.P., Rossetti, B.J., Gibbons, I.R., Hoffman, M.P., Leone, A., Istrail, S., Materna, S.C., Samanta, M.P., Stolc, V., Tongprasit, W., Tu, Q., Bergeron, K.F., Brandhorst, B.P., Whittle, J.,

Berney, K., Bottjer, D.J., Calestani, C., Peterson, K., Chow, E., Yuan, Q.A., Elhaik, E., Graur, D., Reese, J.T., Bosdet, I., Heesun, S., Marra, M.A., Schein, J., Anderson, M.K., Brockton, V., Buckley, K.M., Cohen, A.H., Fugmann, S.D., Hibino, T., Loza-Coll, M., Majeske, A.J., Messier, C., Nair, S.V., Pancer, Z., Terwilliger, D.P., Agca, C., Arboleda, E., Chen, N., Churcher, A.M., Hallbook, F., Humphrey, G.W., Idris, M.M., Kiyama, T., Liang, S., Mellott, D., Mu, X., Murray, G., Olinski, R.P., Raible, F., Rowe, M., Taylor, J.S., Tessmar-Raible, K., Wang, D., Wilson, K.H., Yaguchi, S., Gaasterland, T., Galindo, B.E., Gunaratne, H.J., Juliano, C., Kinukawa, M., Moy, G.W., Neill, A.T., Nomura, M., Raisch, M., Reade, A., Roux, M.M., Song, J.L., Su, Y.H., Townley, I.K., Voronina, E., Wong, J.L., Amore, G., Branno, M., Brown, E.R., Cavalieri, V., Duboc, V., Duloquin, L., Flytzanis, C., Gache, C., Lapraz, F., Lepage, T., Locascio, A., Martinez, P., Matassi, G., Matranga, V., Range, R., Rizzo, F., Rottinger, E., Beane, W., Bradham, C., Byrum, C., Glenn, T., Hussain, S., Manning, G., Miranda, E., Thomason, R., Walton, K., Wikramanayke, A., Wu, S.Y., Xu, R., Brown, C.T., Chen, L., Gray, R.F., Lee, P.Y., Nam, J., Oliveri, P., Smith, J., Muzny, D., Bell, S., Chacko, J., Cree, A., Curry, S., Davis, C., Dinh, H., Dugan-Rocha, S., Fowler, J., Gill, R., Hamilton, C., Hernandez, J., Hines, S., Hume, J., Jackson, L., Jolivet, A., Kovar, C., Lee, S., Lewis, L., Miner, G., Morgan, M., Nazareth, L.V., Okwuonu, G., Parker, D., Pu, L.L., Thorn, R., Wright, R., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. Science 314, 941-952.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P., Dermitzakis, E.T., 2007. Population genomics of human gene expression. Nature genetics 39, 1217-1224.

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., Han, J.D., 2014. Evolution of Alu elements toward enhancers. Cell reports 7, 376-385.

Suryamohan, K., Halfon, M.S., 2015. Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol 4, 59-84.

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., Leder, P., 1982. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proceedings of the National Academy of Sciences of the United States of America 79, 7837-7841.

Telorac, J., Prykhozhij, S.V., Schone, S., Meierhofer, D., Sauer, S., Thomas-Chollier, M., Meijsing, S.H., 2016. Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements. Nucleic Acids Res 44, 6142-6156.

The 1000 Genomes Project, C., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

The ENCODE Project, C., 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636-640.

The ENCODE Project, C., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A., 2012. The accessible chromatin landscape of the human genome. Nature 489, 75-82.

Tu, Q., Cameron, R.A., Davidson, E.H., 2014. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. Developmental biology 385, 160-167.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S.G., 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res 40, e115.

van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J., van Steensel, B., 2017. Genome-wide mapping of autonomous promoter activity in human cells. Nature biotechnology 35, 145-153.

Van der Ploeg, L.H., Konings, A., Oort, M., Roos, D., Bernini, L., Flavell, R.A., 1980. gamma-beta-Thalassaemia studies showing that deletion of the gamma- and delta-genes influences beta-globin gene expression in man. Nature 283, 637-642.

Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T., Fernandez, N., Ballester, B., Andrau, J.C., Spicuglia, S., 2015. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun 6, 6905.

Visel, A., Minovitsky, S., Dubchak, I., Pennacchio, L.A., 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35, D88-92.

White, M.A., 2015. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics 106, 165-170.

Wieczorek, D., Pawlik, B., Li, Y., Akarsu, N.A., Caliebe, A., May, K.J., Schweiger, B., Vargas, F.R., Balci, S., Gillessen-Kaesbach, G., Wollnik, B., 2010. A specific mutation in the distant sonic hedgehog (SHH) cis-regulator (ZRS) causes Werner mesomelic syndrome (WMS) while complete ZRS duplications underlie Haas type polysyndactyly and preaxial polydactyly (PPD) with or without triphalangeal thumb. Hum Mutat 31, 81-89.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. Nature reviews. Genetics 8, 206-216.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M., 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434, 338-345.

Zhang, X., Choi, P.S., Francis, J.M., Imielinski, M., Watanabe, H., Cherniack, A.D., Meyerson, M., 2016. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nature genetics 48, 176-182.

CHAPTER 9: CURRICULUM VITAE

CATHERINE L. GUAY
6304 Harvey Avenue
Pennsauken, NJ 08109
(609)-351-6229
catherinelguay@gmail.com

Education

- 2018 PhD., Rutgers University, New Jersey
- 2002 B.S., Messiah College, Pennsylvania

Professional Experience

- 2013-2018: Teaching Assistant, Center for Computational and Integrative Biology, Chemistry Department, Rutgers University
- 2010-2013: Part-Time Lecturer, Chemistry Department, Rutgers University
- 2007-2008: Technician II, Yan Lab, Fox Chase Cancer Center
- 2002-2007: Technician I, Burch Lab, Fox Chase Cancer Center

Publications

- **Guay CL**, McQuade ST, and Nam J. Single embryo resolution quantitative analysis of reporters permits multiplex spatial cis-regulatory analysis. Developmental Biology. 2017 Feb 15;422(2): 92-104. doi: 10.1016/j.ydbio.2017.01.010

- Liao S, **Guay C**, Toczylowski T, and Yan H. Analysis of MRE11's function in the 5'→3' processing of DNA double-strand breaks. Nucleic Acids Research. 2012;40(The 1000 Genomes Project):4496-4506. doi:10.1093/nar/gks044

- Adamo RF*, **Guay CL***, Edwars AV, Wessels A, and Burch JBE. GATA-6 gene enhancer contains nested regulatory modules for primary myocardium and the embedded nascent atrioventricular conduction system. Anat. Rec. 2004; 280A: 1062–1071. doi:10.1002/ar.a.20105

  * Contributed equally to the work

Presentations

Oral presentations:

- **Developmental Biology of the Sea Urchin Meeting XXIV,** April 5-9, 2017. Woods Hole, MA; **Testing the utility of sea urchin embryos to discover human embryonic *cis*-regulatory modules** (selected talk)

- **Mid-Atlantic Society of Developmental Biology Regional Meeting 2016,** May 20-21, 2016. Washington, DC; **High-throughput spatial *Cis*-regulatory analysis by turning chaos into order** (selected talk)

- **Developmental Biology of the Sea Urchin Meeting XXIII,** October 7-11, 2015. Woods Hole, MA; **Elusive causal linkages revealed by combinatorial *cis*-regulatory perturbations of *univin.*** (selected talk)

Poster presentations:

- **Society of Developmental Biology 75th Annual Meeting/International Society of Differentiation 19th International Conference,** August 4-8, 2016. Boston, MA; **High-throughput spatial Cis-regulatory analysis by turning chaos into order** (poster)

Awards

- 2017: **CCIB Best Student Paper Award,** 3rd Place

- 2017-2018: **Teaching Assistant**, Supported by the Center for Computational and Integrative Biology (CCIB), Rutgers University

- 2016: **CCIB Retreat Best Poster Award**

- 2016-2017: **Teaching Assistant**, Supported by the Center for Computational and Integrative Biology (CCIB), Rutgers University

- 2016: **Graduate Student Travel Fund**, Rutgers University

- 2015-2016: **Teaching Assistant**, Supported by the Center for Computational and Integrative Biology (CCIB), Rutgers University

- 2015: **Graduate Student Travel Fund**, Rutgers University

- 2014-2015: **Teaching Assistant**, Supported by the Center for Computational and Integrative Biology (CCIB), Rutgers University

- 2013-2014: **Teaching Assistant**, Supported by the Center for Computational and Integrative Biology (CCIB), Rutgers University