

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

RNA as an Epigenetic Molecule During Cardiac Lineage Commitment

**Permalink**

<https://escholarship.org/uc/item/5q5455b5>

**Author**

George, Matthew Ryan

**Publication Date**

2018

Peer reviewed|Thesis/dissertation

RNA as an Epigenetic Molecule  
During Cardiac Lineage Commitment

by

Matthew Ryan George

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Developmental and Stem Cell Biology

in the

GRADUATE DIVISION



This dissertation is dedicated to my wife Paige, whom has granted me the honor of sharing a life and family; and to my daughter Marlo, born by 'the big red bridge' and my best reminder for why any of the work we do matters.

Special thanks to my graduate mentor Dr. Benoit Bruneau for creating an environment so driven in pursuit of knowledge and for empowering his trainees to embrace the cutting edge, while maintaining roots in the fundamental principles of basic science.

I would also like to express my gratitude to the other member of my thesis committee Dr. Deepak Srivastava and Dr. Brian Black for their guidance during my time at UCSF. The discoveries Deepak, Brian, and Benoit have illuminated throughout their scientific careers will be sources of inspiration for the rest of mine.

All experiments were designed and performed in the laboratory of Dr. Benoit Bruneau at the Gladstone Institute of Cardiovascular Disease, UCSF.

## RNA as an Epigenetic Molecule During Cardiac Lineage Commitment

Matthew Ryan George

The molecular machinery underlying heart development out of primordial anterior mesoderm remains incompletely understood. This is in part due to a restricted research emphasis on canonical genomic elements and the inability to precisely model the timing of cardiac lineage commitment. The genome is pervasively transcribed far beyond the coding exome. Therefore, we hypothesized that the transcriptome as a whole, including all noncoding and splice isoforms would most precisely define the identity of the cell as it differentiated into a cardiomyocyte. Using an enhancer fragment of the heart-critical *Smarcd3* Brg1/Brm associated factor (BAF) subunit, we developed an *in vitro* reporter differentiation system that precisely delineated the first heart specification from murine embryonic stem cells (mESCs). From this, *de novo* total RNA isoform coexpression networks were generated to reveal a staged progression by which hundreds of thousands of gene isoforms were organized into hundreds of subnetwork modules that dynamically programmed nascent mesoderm to become restricted to a cardiovascular fate.

Many transcripts within the nucleus bind and influence the genomic regulation of chromatin modifying complexes. Therefore, we next aimed to elucidate the BRG1/BAF RNA interactome during its key role at cardiac fate commitment. Using targeting immunoprecipitations coupled to molecular techniques to isolate and identify protein:RNA adducts, we found at least 7 subunits engaged RNA molecules via

previously unrealized discrete domains. Furthermore, these subunits interfaced RNA through tens of thousands of binding events that frequently coincided the defining transcript isoform transitions of this developmental window.

These interrogations into the transcriptional basis for cardiac lineage differentiation also identified 6 annotated long intergenic noncoding RNAs (lincRNA) with discrete gene structure, epigenetic regulation, and cardiac progenitor specificity *in vivo*. We ablated these lincRNAs in the embryo via Cas9 editing, which revealed regulatory roles within their local genomic environments, including between *Bmp4* and *Rubie*. While none of the 6 transcripts were required for proper heart morphogenesis, compound heterozygous *Bmp4<sup>+/-</sup>; Rubie<sup>+/-</sup>* offspring did reveal a genetic interaction between these genes in formation of the right ventricular outflow tract. These experiments provide new insight into the complex role RNA plays during cardiac lineage commitment and congenital heart disease.

## Table of Contents

|   |     |
|---|-----|
| <b>Chapter 1:</b> <i>De novo</i> transcriptome construction throughout cardiomyocyte differentiation reveals highly complex gene isoform expression network dynamics at the onset of cardiac lineage commitment ..... | 1   |
| References .....  | 30  |
| <br>  |     |
| <b>Chapter 2:</b> The BAF complex binds thousands of RNA transcripts within the nucleus via discrete domains of multiple subunits at the onset of cardiac lineage commitment .....                                    | 35  |
| References .....  | 96  |
| <br>  |     |
| <b>Chapter 3:</b> Ablation of cardiac lincRNAs in vivo reveals genetic interaction between Rubie and Bmp4 .....   | 112 |
| References .....  | 149 |

## List of Figures

### Chapter 1

|   |    |
|---|----|
| 1.1: The <i>Smarcd3</i> -F6-eGFP mESC reporter system .....                                     | 22 |
| 1.2: Transcriptional complexity during cardiac differentiation .....                            | 23 |
| 1.3: Differential ‘gene’ isoform expression at each phase of<br>cardiac lineage commitment..... | 25 |
| 1.4: Differential expression of network module hub gene isoforms.....                           | 27 |
| 1.5: Multigene co-processing and intragenic reverse transcription .....                         | 29 |

### Chapter 2

|   |    |
|---|----|
| 2.1: Identification of BAF subunit RNA binding domains.....                                   | 81 |
| 2.2: 3° BAF subunit structure modeling near RNA binding domains.....                          | 82 |
| 2.3: Subunit-specific RNA interactome detection using native RIP and eCLIP .....              | 84 |
| 2.4: Characteristics of Brg1/BAF RNA binding .....  | 85 |
| 2.5: BAF protein-protein binding interfaces predicted by eCLIP .....                          | 86 |
| 2.6: RNA motif recognition of BAF subunits.....   | 87 |
| 2.7: Targeting of BRG1/BAF DNA binding sites to genes containing RNA-bound<br>subunits .....  | 89 |
| 2.8: BAF complex RNA binding at genes of differential RNA isoform<br>splice transitions ..... | 90 |
| 2.9: SMARCE1 and SMARCA4 RNA co-binding .....   | 92 |
| 2.10: Non-BAF co-immunoprecipitated RNA binding proteins .....                                | 93 |



## Chapter 3

|   |     |
|---|-----|
| 3.1: Epigenetically regulated cardiac lincRNAs and genomic characterization<br>of lincRNA Rubie .....   | 139 |
| 3.2: Genomic characterization of <i>Handlr</i> and <i>Atcayos</i> lincRNAs .....  | 140 |
| 3.3: Genomic characterization of <i>HrtLincR4</i> and <i>HrtLincR5</i> lincRNAs.....  | 141 |
| 3.4: Genomic characterization of <i>HrtLincRX</i> , <i>5033406O09Rik</i> , <i>9630002D21Rik</i> ,<br>and <i>2810410L24Rik</i> lincRNAs.....                                 | 142 |
| 3.5: Molecular characterization of lincRNA cohort.....  | 143 |
| 3.6: LincRNA expression patterns <i>in vivo</i> .....   | 144 |
| 3.7: Cas9 ablation of cardiac lincRNAs <i>in vivo</i> and effects on local<br>gene expression .....   | 145 |
| 3.8: Requirements for lincRNA cohort for viable development .....   | 146 |
| 3.9: TAC hypertrophy models in <i>Handlr</i> and <i>Atcayos</i> null mice .....   | 147 |
| 3.10: Effect of lincRNA ablation, <i>Hand2</i> / <i>Handlr</i> compound heterozygosity,<br>and <i>Rubie</i> / <i>Bmp4</i> compound heterozygosity on heart development..... | 148 |

## Chapter 1

*De novo* transcriptome construction throughout cardiomyocyte differentiation reveals highly complex gene isoform expression network dynamics at the onset of cardiac lineage commitment

### Introduction

Clonal analyses of the early vertebrate embryo indicate that the heart derives from anterior lateral plate mesoderm<sup>1</sup>. Waves of gene transcription model newly gastrulated mesoderm to become cardiogenic, including overlapping progressions through lineages expressing the transcription factors (TFs) BRACHYURY<sup>2</sup>, EOMES<sup>3</sup>, and MESP1<sup>4</sup>, respectively. During this time, even before any morphogenesis becomes apparent, downstream lineage fates are being specified. This includes a restriction barrier that rapidly forms within MESP1<sup>+</sup> mesoderm, which divides cells that will eventually populate the left ventricle and atria (first heart field, FHF)<sup>5</sup> from those that will primarily form the right ventricle and outflow tract (second heart field, SHF)<sup>6</sup>. This process is significantly due to important transcription factor activity, including TBX5 in FHF<sup>7</sup> and ISL1 in SHF populations<sup>8</sup>, along with the dual heart field TF NKX2-5<sup>9</sup>. However, the molecular mechanisms responsible for early commitment of the heart lineage out of Mesp1<sup>+</sup> mesoderm likely occur before these factors are expressed and have remained unknown. While the first cardiac progenitors arise within Mesp1-expressing primordium, these cells are not heart-specific. Our laboratory has found that the first mesodermal expression of the Swi/Snf Brg1/Brm associated factor (BAF) chromatin remodeling complex subunit *Smarcd3* (*Baf60c*) labels the earliest pan-cardiac progenitors *in vivo*

prior to expression of *Isl1*, *Tbx5*, and *Nkx2-5*. Moreover, we have identified a particular enhancer region, named 'F6', which directs the earliest appearance of *Smarcd3* *in vivo*<sup>10</sup>. This unique reporter allows the isolation and detailed interrogation of this commitment step for the first time. Therefore, it presents an invaluable tool to model the transition from pre-cardiac to heart-specified mesoderm progenitors.

All developmental processes arise in some capacity from transcriptional events. Therefore, to better understand the nature of cardiac development, one must better decipher its underlying gene expression. We hypothesized that the *Smarcd3*-F6 reporter, in combination with directed embryonic stem cell differentiation into cardiomyocytes *in vitro* would allow the temporal specificity necessary to elucidate transcriptional archetypes that drive cardiac lineage commitment. We also predicted that the complexity of gene expression at the transcript isoform level would characterize previously unappreciated dynamics of this process. In doing so, we discovered a discrete transition window separating nascent mesoderm and cardiac progenitors. This lineage commitment phase was resolved and defined by a 3-step reorganization of hundreds of novel RNA isoforms from the concerted activity of distinct transcriptional subnetworks.

## Materials and Methods

### *Smarcd3-F6eGFP mESC Cell Line Engineering*

The Smarca4FLAG knock-in ES cell line<sup>11</sup> was used for targeting of the *Smarcd3-F6-Hsp68-nlsEGFP* construct to the *Hipp11* locus. Briefly, a modified shuttle vector containing a polylinker including *PacI*, *XhoI*, *SacII*, and flanking *Ascl* sites was purchased from IDT. A pGKNeo selection cassette was subcloned from the pL451 plasmid using *XhoI* and *SacII* into the modified shuttle vector. A *PacI* fragment including flanking H19 insulator sequences, the *Smarcd3-F6* enhancer, an Hsp68 minimal promoter, nlsEGFP coding sequence, WPRE mRNA stabilization sequence, and EF1alpha polyA sequence was subcloned into the modified shuttle vector. The entire reporter-selection construct was cloned into the *Hipp11* targeting vector using *Ascl*. The targeting vector was linearized using *Apal* and electroporated into ES cells. Following G418 selection, correctly targeted clones were screened by PCR and Southern blotting. For culturing, ES cells were maintained in 2i + LIF (ESGRO, EMD) media.

### *mESC differentiation into cardiomyocytes and cell sorting*

Directed cardiomyocyte differentiations were performed as previously described by Wamstad et al<sup>12</sup> using the *Smarcd3-F6nlsEGFP* mESC line with minor modifications to improve differentiation efficiency. Briefly, three days before differentiation induction (day -3), mESCs were split into 2i + LIF media on gelatin. The following day (day -2), 2i + LIF was replaced with FBS-LIF medium (15% FBS (HiClone) in DMEM + 1X non-essential amino acids + 1X sodium pyruvate + 1X GlutaMAX +  $\beta$ mercaptoethanol + 1X

penicillin/streptomycin + 1000U/ml LIF). The following day (day -1), cells were fed again with the same 15% FBS-LIF media to complete their conversion to epiblast-like stem cells. One day later (day 0), cardiac differentiation was initiated as per Wamstad et al. On day 4, 18 hours after plating and cardiac induction with VEGF, FGF10, and FGF2 (day 4.75), supernatant was collected and 0.22  $\mu$ m filtered. Cells were then washed with D-PBS (w/o Ca<sup>2+</sup>/Mg<sup>2+</sup>), dissociated from plates using TrypLE (Gibco), resuspended in filtered supernatant, and placed on ice. GFP<sup>+</sup> and GFP<sup>-</sup> populations were subsequently sorted into RNeasy Protect Cell Reagent (Qiagen) using a BD FACSAria II flow cytometer. Smarcd3-eGFP<sup>+</sup> and Smarcd3-eGFP<sup>-</sup> RNA was then purified from each population using the RNeasy Mini kit (Qiagen).

#### *RNA-seq data sets, mapping, and normalization*

Stored RNA from mESC differentiation into cardiomyocytes was re-sequenced in the Boyer Laboratory at MIT using the Illumina TruSeq Stranded Total library kit to generate 100nt paired end reads. This provided the raw reads for epiblast (EPI), mesoderm (MES), cardiac progenitor (CP), and cardiomyocyte (CM) groups (2 biological replicates each). 2i+LIF-maintained mESCs, day 4.75 Smarcd3-eGFP<sup>+</sup> and day 4.75 Smarcd3-eGFP<sup>-</sup> stranded RNA-seq libraries were prepared in the Gladstone Genomics Core using the Ovation Mouse FFPE RNA-Seq Multiplex System (NuGEN) and sequenced on an Illumina HiSeq 2000 with 150nt paired end reads. This generated the raw reads for inner cell mass (ICM), pre-cardiac mesoderm (pcMES), and cardiac mesoderm (cMES) groups (3 biological replicates each). All raw reads were mapped to the mouse genome (mm10) using STAR<sup>13</sup>. PCR duplicates were estimated with STAR and

removed. De novo transcriptomes were generated with StringTie<sup>14</sup> using Ensembl annotated gene isoforms as a reference scaffold. Gene isoform count data was tabulated with Cuffquant and Cuffnorm<sup>15</sup> without specialized scaling to maintain inherent variation between samples. Between sample variation was then reduced with the RUVg command of the 'RUVseq' R package<sup>16</sup> using the 30% of gene isoforms least-differentially transcribed between groups (calculated with the 'limma' R package<sup>17</sup>) as empirical controls and  $k = 1$  factors of variation. Variance was then stabilized using varianceStabilizationTransformation of the 'DESeq2' R package. 2D multidimensional scaling (MDS) on this isoform count matrix was calculated in R.

### *Gene Isoform Network Construction and Analysis*

Gene isoform networks were generated from a variance-stabilized matrix of isoform counts at each differentiation time point using the WGCNA<sup>18</sup> R package. Due to the number of isoforms analyzed, network construction was run with the blockwiseModules command with a soft threshold power  $sft = 8$ , networkType = signed, and topological overlap matrix TOM = signed. Dendrograms, heatmaps, module membership plots of gene isoforms and module eigenvectors were generated using WGCNA commands. P values for differential transcript levels between MES-pcMES and cMES-CP stages were calculated by linear models generated with the 'limma' R package, and the distribution of these p values was input into the 'qval'<sup>19</sup> R package to estimate the proportion of false positives called significant. Q values less than 0.05 were deemed significant. For pcMES-cMES comparisons, the minimum of 'limma' p values and Student's t-test (two tailed) p values were used as input for q value calculation. Network modules with

enriched representation in differentially expressed transcript subsets at MES-pcMES, pcMES-cMES, and cMES-CP transitions were calculated by hypergeometric tests. Gene ontology (GO) for genes associated with differentially expressed transcripts was calculated with Panther<sup>20</sup>.

### *RT-qPCR*

RNA was harvested at 6 hour intervals between day 4.0 and day 6.0 using Qiagen RNeasy Minelute columns. 300ng for each sample was reverse transcribed with Applied Biosystems High Capacity cDNA RT Kit. 10ng cDNA per sample was then input into qPCR reactions with TacMan (Applied Biosystems) master mix and primer/probes against *Mesp1*, *eGFP*, *Smarcd3*, *Nkx2-5*, *Tbx5*, *Tnnt2*, and *Actb*. Reactions were run and analyzed on a 7900HT (Thermo Fisher) cycler with absolute quantification. dCt values were generated between gene-specific primer Ct values and *Actb* internal control primer Ct values.

### *Immunocytochemistry*

At 6-hour intervals between day 4.0 and day 6.0 of cardiac differentiation, cells were resuspended in D-PBS and quickly imaged for *Smarcd3*-eGFP reporter expression. For immunocytochemistry at day 6.0, cells were washed with D-PBS, fixed in 4% paraformaldehyde in D-PBS at room temperature (RT) for 15 minutes, blocked and permeabilized for 30 min in 0.1% TritonX-100 + 10% FBS in D-PBS at RT, and incubated with primary antibodies in 2% FBS + 0.5% saponin in D-PBS (WASH) at 4°C overnight. 1° antibodies included anti-GFP (Abcam #ab13970; 1:5000 dilution) and anti-

TNNT2 (Thermo Scientific #MS295; 1:100 dilution). The following day, cells were washed twice with WASH and incubated for 2 hours at RT with Alexa Fluor (Invitrogen) 2° antibodies. Finally, cells were washed three more times with WASH, including DAPI (1:5000) in the third wash, before imaging.

### *Flow Cytometry*

At 6-hour intervals between day 4.0 and day 6.0, live cells were dissociated with TrypLE (Invitrogen), quenched with 10% FBS in DMEM + LIVE/DEAD stain (Invitrogen), 40µm cell-strained into flow cytometry tubes, and resuspended in D-PBS (Invitrogen) on ice. Cells were then sorted on a MACSQuant VYB cytometer to detect Smarcd3-F6 reporter eGFP expression. At day 12 of differentiation, cells were dissociated with TrypLE (Invitrogen), quenched with 10% FBS in DMEM + LIVE/DEAD stain, 40µm cell-strained into flow cytometry tubes, and fixed in 4% paraformaldehyde in D-PBS. Fixed samples were permeablized and blocked in 10% FBS + 0.2% saponin in D-PBS and stained with anti-TNNT2 (Thermo Scientific #MS295) 1° in 2% FBS + 0.2% saponin (WASH) for 30 min at RT. Cells were subsequently washed twice with WASH and incubated for 2hrs with Alexa Fluor-594 2° antibody before analyzing on a MACSQuant VYB cytometer.



## Results

### ***An upstream enhancer of Smarcd3 labels the onset of heart lineage commitment in vitro before Tbx5, Nkx2-5, and Tnnt2.***

The *Smarcd3*-F6 enhancer labeled the first specific pan heart-committed progenitors *in vivo*. Therefore, we utilized this novel tool to develop a system that could allow maximal temporal resolution of the transition out of primordial mesoderm into nascent cardiovascular tissue. Toward this we engineered a stable transgenic mouse embryonic stem cell (mESC) line containing the *Smarcd3*-F6 enhancer fragment driving nuclear-localized enhanced green fluorescent protein (nlseGFP, Fig1A). Subsequently, we modified the protocol developed in the Keller laboratory by Kattman et al<sup>21</sup> to efficiently differentiate this live reporter line (*Smarcd3*-nlseGFP) into beating cardiomyocytes through mesoderm and cardiac progenitors. We maintained mESCs in the naïve inner-cell mass-like state with GSK3 $\beta$  and Mek1/2 inhibitors along with LIF (2i + LIF), transitioned them through the primed epiblast-like state, and finally induced differentiation via *in vivo*-relevant growth factors of the Keller protocol.

Upon closely monitoring eGFP reporter expression and the transcription of key cardiac genes during differentiation, we discovered a precise time window 18 hours after cardiac induction at day 4.0 with VEGF, FGF10, and FGF2 supplementation (day 4.75) whereby the *Smarcd3*-F6 transgene reached maximal activity. This overlapped the rapid decline of *Mesp1* expression and preceded the activation of *Tbx5*, *Nkx2-5*, and *Tnnt2* that defined the cardiac progenitor state at day 5.3 in culture (Fig1B).

Consistently, 50-70% of the cell population was concomitantly eGFP<sup>+</sup> at a given time

during this interval, which translated into end-point TNNT2<sup>+</sup> purities after day 10 of 75% to >90% (Fig1C). Furthermore, by day 6 of differentiation, we found remaining *Smarcd3*-F6 reporter eGFP protein overlapped the appearance of nascent TNNT2 protein expression. This was observed even when the differentiation efficiency was drastically reduced with suboptimal BMP4 concentrations (Fig1D). Together, these findings along with *in vivo* data, suggested activity of the *Smarcd3*-F6 enhancer directly and specifically labeled the earliest known time point of heart lineage commitment. We subsequently performed fluorescence assisted cell sorting (FACS) on day 4.75 cells to isolate F6-eGFP<sup>+</sup> and F6-eGFP<sup>-</sup> populations. Since these differentiations resulted in high end-point cardiomyocyte purities, we could thus assign the eGFP<sup>-</sup> population as lagging in developmental time to the eGFP<sup>+</sup> cardiac-committed mesoderm. This system, with precise temporal resolution surrounding cardiac lineage commitment, was thus a model of cardiac development established *in vitro* for which analogs of the inner cell mass (ICM, day -2.0), epiblast (EPI, day 0), gastrulating mesoderm (MES, day 4.0), pre-cardiac mesoderm (pcMES, day 4.75: eGFP<sup>-</sup>), cardiac mesoderm (cMES, day 4.75: eGFP<sup>+</sup>), cardiac progenitor (CP, day 5.3), and beating cardiomyocyte (CM, day 10+) stages could be discretely interrogated and compared.

***Thousands of novel gene isoforms are expressed during cardiac differentiation, and their complex dynamics resolve the transition from mesodermal to cardiac progenitors.***

We hypothesized that in order to more completely understand cardiac lineage commitment, we needed to generate a comprehensive transcriptome of each stage in

this process. Therefore, in collaboration with the Boyer lab at MIT, we performed stranded total RNA-seq on samples that had been previously harvested during the differentiation of mESCs from the EPI stage to CM stage via MES and CP intermediates<sup>21</sup>. This provided greater read depth, read length, and strand information than previous efforts and allowed us to better dissect the nuance of transcription during these time points. Furthermore, we performed new Keller protocol differentiations of *Smarcd3*-F6-eGFP mESCs from the ICM stage to day 4.75, whereby eGFP<sup>+</sup> and eGFP<sup>-</sup> populations were isolated by FACS. After performing total stranded RNA-seq on these samples as well, we were able to generate an integrated *de novo* transcriptome encompassing the ICM, EPI, MES, pcMES, cMES, CP, and CM stages together using STAR read mapping<sup>13</sup> and StringTie splice variant assembly<sup>14</sup>. In doing so, we identified 124,953 discrete transcript splice forms from 44,478 unique genetic elements that were reliably expressed during the differentiation timecourse (Figure 2A). We referred to these units as ‘genes’, given their varied combination of protein coding isoforms, antisense transcripts to protein coding genes, noncoding RNAs, and previously unannotated loci. Multidimensional scaling of the *de novo* data set revealed clear transitions during MES to CP progression through pcMES and cMES intermediates, while these could not be clearly represented at a gross gene level (Fig2B). Furthermore, less than 40% of all expressed transcript isoforms and less than 1% of all detected ‘genes’ were fully Ensembl annotated (Fig2C). This suggested that the specific identity of the cell was best defined not just by the genes that were expressed but by their previously uncharacterized underlying transcript diversity at the isoform level.

Splice form signatures delineated closely related, yet discrete developmental stages of cardiac development. To understand how this complexity was organized, we performed signed and weighted correlation network analysis<sup>17</sup> on the transcriptome. In doing so, we could segregate these 124,953 expressed RNA molecules to 579 assigned module eigenvectors, revealing surprising paradigms of coordinated activity. 95% of the 13,406 alternatively spliced 'genes' could subdivide their respective isoforms into separate network clusters (Fig2C, D). Furthermore, these distinct modules displayed widely varying degrees of inter-eigenvector correlation. Thus, individual 'gene' isoforms were commonly segregated into distantly related subnetworks (Fig2D). These data pointed to the conclusion that an individual isoform often held greater intergenic expression connectivity than to the other transcripts within its own 'gene'. Finally, the prevailing expression patterns of these 579 modules encompassed diverse configurations, but parallel patterns over developmental time were shared across distinct transcript subnetworks (Fig 2E). This indicated that waves of convergent and divergent transcript network activities drove the gene expression transitions of cardiac differentiation.

***Differential expression and splicing of key transcript network module hubs underlie the specification of cardiac progenitors.***

Next, we analyzed differential RNA expression between the step wise commitment of early cardiac precursors that our *Smarcd3*-F6 model system was designed to elucidate. The total transcribed number of variably spliced 'genes' and isoforms per 'gene' remained constant throughout all stages of cardiac differentiation (Fig3A, B). However, despite the relatively short elapsed time between commitment stages, we discovered

hundreds of instances of significant differential splice form prevalence between each leg of the three-step process: 1. MES to pcMES; 2. pcMES to cMES; and 3. cMES to CP. These encompassed 1575 separate isoforms dispersed over 1368 'genes', for which approximately 17% of these 'genes' were alternatively spliced at more than one stage of the commitment process. In 90% of these cases, though, this was the result of a single transcript's dynamic expression pattern. These data indicated the primary change of a particular 'gene' during these developmental steps most often occurred through evolution of one of its individual RNA species (Fig3C). At the first, second, and third conversion stages, 616, 421, and 759 differentially spliced and/or expressed 'gene' transcripts were detected, respectively. These arose from 196, often distant, co-expression modules. 13 of these were significantly overrepresented (hypergeometric  $p < 0.005$ ) at one or more conversion steps (Fig 3D, E, F), indicating a specific role for their network dynamics in the MES to CP differentiation window. The concerted activity of even these 13 stage specific subnetworks required distant eigenvectors to join in collective transcript up- or down-regulation, thereby generating the defining transcriptional landscape of this developmental progression (Fig3G). These archetypes indicated again that divergent transcriptional subnetworks generated convergent mutual patterning effects at specific time points during lineage commitment.

Three of these eigenvector subnetworks were overrepresented at multiple stages of cardiac conversion (Fig.3D, E, F). Of these, 'Module 6', whose connected transcripts derived from many contractile protein genes, contained a significant number of upregulated transcript isoforms at both MES to pcMES (step 1) and cMES to CP (step

3) conversions (Fig3D, F). However, gene ontology (GO) analysis of loci containing step 1-upregulated protein coding isoforms revealed their enrichment in calcium ion homeostasis and cell junction assembly biological processes, but not contractile proteins. This was in contrast to step 3, where upregulated protein coding isoforms were overrepresented for cardiac muscle contractility and actomyosin structure functions (Fig3G). This dichotomy showed that subsets even within these connected subnetworks could enact differential effects at alternative stages of development.

We focused on the 13 significantly enriched network modules to establish how their differentially expressed 'gene' isoforms connected to other cluster members. By calculating intra-modular membership of each constituent, we found the transcripts that significantly changed at each step were often among the most connected subnetwork hubs (Fig. 4A-E). For example, the lone-expressed transcript of the critical mesoderm transcription factor *Mesp1*, which was downregulated from MES to pcMES stages, was the most connected transcript of 345 RNA isoforms in its module (Fig4A). Furthermore, despite numerous expressed isoforms, novel splice forms of *Cdk17* and the Myo6-interacting gene *Dock7<sup>22</sup>* were not only the lone significantly regulated RNA molecules of their respective genes, but they were also of the most connected central hubs of their co-expression subnetworks (Fig4B, C). Additional examples of differentially expressed module hubs included transcript isoforms of the cardiovascular patterning gene *Amot<sup>23</sup>*; *Prkab2*, important for muscle energy homeostasis<sup>24</sup>; *Cox4i2*; the actin cytoskeletal coregulator *Mprip<sup>25</sup>*; *Zmym4*; the ribonuclear splice regulator essential for heart development *Hnrnpul1<sup>26</sup>*; and another actin filament binding protein important for

cardiomyocyte proliferation *Sorbs2*<sup>27</sup> (Fig4D). These data suggested the RNA variants most dynamically expressed during mesoderm procession into cardiac progenitors were centrally connected to larger genomic coregulation networks.

***Differential isoform expression involves multigene co-processing and widespread intragenic reverse transcription during cardiac development.***

In studying the transcriptional properties of dynamically regulated loci in our model system, we established additional multifaceted transcriptional characteristics. Over the course of ICM to CM differentiation, we detected 503 loci where two or more separate annotated protein coding genes were co-processed and spliced as single genetic units. 80 of these loci involved the processing of transcripts that were differentially expressed during the 3-step cardiac commitment phase of MES to CP cells (Fig5A). For example, at the *Tubgcp6 / Hdac10* locus, a single transcript spanned both genes most abundantly at the EPI stage of differentiation. However, of the twenty expressed splice forms, a single processed transcript of histone deacetylase *Hdac10* was differentially enriched during pcMES to cMES progression (Fig5B). Additionally, we observed complex co-processing at the critical *Myh6 / Myh7* locus. From MES to pcMES stages, two novel transcripts were significantly upregulated, including an RNA molecule spliced from *Myh7* into *Myh6*, as well as an *Myh6*-specific transcript containing a retained intron. Subsequently, after conversion from cMES into CP stage cells, these isoforms were upregulated again, along with four additional splice forms. These new splice forms included two additional *Myh6*-to-*Myh7* spliced RNA species, as well as added *Myh6*- and *Myh7*-specific transcripts (Fig5C). Furthermore, evidence for *Myh6 / Myh7*

compound transcripts was also corroborated by GenBank expressed sequence tags<sup>28</sup>. These results added another layer of complexity to the transcriptional evolutions that took place during cardiac development.

While pervasive antisense transcription at canonical genes had been indicated previously<sup>29</sup>, current assemblies did not annotate stable intragenic antisense transcription in a majority of the protein-coding genome. However, our stranded RNA-seq data set and *de novo* transcriptome assembly detected thousands more instances of this phenomenon than was acknowledged by the Ensembl91 database. In fact, the predominance of protein coding loci contained detectable antisense transcripts within their gene bodies (Fig5D). This pattern also carried over to dynamically expressed loci that underlied cardiac lineage commitment. For example, the *Agrn* gene, encoding a basal lamina glycoprotein required for proper cardiac contractility<sup>30</sup>, contained multiple significantly regulated isoforms during the cardiac phase transition. However, in addition to this, numerous lowly transcribed and previously unannotated antisense transcripts spanned its gene body (Fig5E). This pattern was reproduced throughout the transcribed genome at nearly every locus interrogated. Therefore, we hypothesized that antisense transcripts played a fundamental regulatory role in their local gene environments.



## Discussion

The *Smarcd3*-F6-eGFP reporter presented a unique tool to interrogate the onset of cardiac lineage commitment. Therefore, we incorporated this into the directed differentiation of mESCs into cardiomyocytes to precisely delineate the stages of mesoderm to cardiac progenitor developmental commitment. By constructing a *de novo* transcriptome spanning this differentiation, we uncovered new layers of gene expression dynamics. These included nearly tripling the number of known RNA isoforms transcribed during heart development and establishing hundreds of co-expression subnetworks that connected this transcriptome together. By doing this, we established the concept that cell identity during the dynamic cardiac commitment window was best defined by its profile of transcript splice forms more so than at the gross gene level. In addition, we showed that this process could be segregated into at least 3 discrete steps, defined by hundreds of transitions of these individual gene isoforms from the convergent activity of nearly 200 often disparate transcript network expression modules. Further, differentially expressed transcript variants happened to be highly connected subnetwork hubs of the most enriched network modules during these commitment steps. In addition, these subnetworks even contained intramodular compartments of RNA transcripts that differentially contributed to each step in the cardiac commitment process. Finally, our *de novo* transcriptome assembly revealed the pervasive nature of antisense transcription within canonical gene bodies, as well as dynamically regulated inter-gene transcript processing during the fate specification of nascent mesoderm.

These new insights into the transcriptional complexity of cardiac lineage commitment generated additional questions. What intermediate factors might integrate the observed transcript network organization that drove the commitment of gastrulating mesoderm into cardiac tissue? Also, what function might this widespread and previously unannotated transcription have during cardiac differentiation, including pervasive antisense transcription throughout canonical protein-coding genes? To address these questions, we next aimed to identify molecular components within the nucleus that might unify the multifaceted and complex nature of RNA expression and splicing over the course of mesoderm commitment toward the heart lineage.

## Description of Figures

**Figure 1.1.** The Smarcd3-F6-eGFP mESC reporter system.

A.) Top: Schematic representation of Smarcd3-F6 enhancer and mESC reporter cloning strategy. Bottom: Southern blot of stable single copy transgene insertion at HIP11 locus. B.) Top: Graphic representation of major developmental stages during mESC differentiation into cardiomyocytes. Bottom: RT-qPCR timecourse during mESC to cardiomyocyte differentiation. Error bars have been omitted for clarity. N = 4. C.) Left: Representative live cell flow cytometry of Smarcd3-F6-EGFP from day 4.0 to 4.75. Right: Representative flow cytometry for TNNT2 after 12 days of differentiation. N > 3. D.) Immunocytochemistry staining for Smarcd3-F6-eGFP reporter expression and TNNT2 at day 6 of differentiation. Left: optimized concentration of BMP4. Right: suboptimal BMP4 concentration

**Figure 1.2.** Transcriptional complexity during cardiac differentiation.

A.) Dendrogram of 124,493 transcript isoforms with 10 RNA-seq counts detected in  $\geq 2$  samples. B.) Left: 2D MDS plot for transcript isoform expression by stage of differentiation; window of cardiac commitment highlighted. ICM, inner cell mass-like (day -2.0); EPI, epiblast-like (day 0); MES, mesoderm (day 4.0); pcMES, pre-cardiac mesoderm (day 4.75, Smarcd3-F6-eGFP<sup>-</sup>); cMES, cardiac mesoderm (day 4.75, Smarcd3-F6-eGFP<sup>+</sup>); CP, cardiac progenitor (day 5.3); CM, cardiomyocyte (Day 10). Right: Adapted gene level PCA plot from Zhang Y, et al. Cell Stem Cell. 2016. C.) Top: Novel vs Ensembl-annotated transcript isoforms expressed during in vitro cardiac

differentiation. Middle: Fully Ensembl-annotated *de novo* gene elements vs gene elements containing novel transcripts. Bottom: Variably spliced gene elements with all transcript isoforms assigned to single network expression module vs multiple expression modules. D.) Inter-module eigenvector correlation between 479 clusters; Assignment of two Myo7a isoforms to uncorrelated network modules depicted. E.) Module eigenvector correlation to cardiac differentiation expression pattern models of 128 possible binary (ON/OFF) expression patterns.

**Figure 1.3.** Differential 'gene' isoform expression at each phase of cardiac lineage commitment.

A.) Total number of spliced 'genes' for each developmental stage. B.) Distribution of isoforms expressed per spliced 'gene' for each developmental stage; box, 1<sup>st</sup> and 3<sup>rd</sup> quartiles; line, median value; whiskers, min and max; points, outliers. C.) Summary of overall differentially expressed 'gene' and 'gene' isoforms from all three depicted transitions. D-F.) Differentially expressed 'gene' isoforms between MES and pcMES stages, pcMES and cMES stages, and cMES and CP stages, respectively, sorted by network module, FDR-adjusted p value < 0.05. Color represents row z-score of geometric means. Modules with significant overrepresentation explicitly labeled with example genes of interest; hypergeometric test p < 0.005, green, UP regulated; red, DOWN regulated. Annotated neighboring genes with shared splices separated by '/'. G.) Dendrograms of all 579 transcript network eigenvectors with overrepresented modules for each transition highlighted; Gene ontology enrichment displayed for differentially

expressed Module 6 transcripts at 1<sup>st</sup> and 3<sup>rd</sup> transitions; GO, gene ontology green, UP regulated; red, DOWN regulated. MES, mesoderm (day 4.0); pcMES, pre-cardiac mesoderm (day 4.75, Smarcd3-F6-eGFP<sup>-</sup>); cMES, cardiac mesoderm (day 4.75, Smarcd3-F6-eGFP<sup>+</sup>); CP, cardiac progenitor (day 5.3); Mod, network module.

**Figure 1.4.** Differential expression of network module hub gene isoforms.

A-C.) Left: Ranked module isoforms by module membership with differentially expressed module hub genes highlighted; Right: row z-score of geometric means for all expressed isoforms with assigned module number and intron/exon transcript schematic for Ensembl91 protein coding and *de novo* assemblies, respectively; differentially expressed isoforms are starred and boxed along with heat map of corresponding stage transition. D.) Additional ranked module isoforms by module membership with differentially expressed module hub genes highlighted.

**Figure 1.5.** Multigene co-processing and intragenic reverse transcription.

A.) Summary of total loci with Ensembl-annotated neighboring genes spliced into each other for all expressed 'genes' and 'genes' differentially expressed between MES and CP stages, respectively. B.) Example of multigene co-processing at the Tubgcp6/Hdac10 locus; Left: row z-score of geometric means for all expressed isoforms with assigned module number and intron/exon transcript schematic for Ensembl91 protein coding and *de novo* assemblies, respectively; box and star, differential expression at stage transition. C.) Example of multigene co-processing at the

Myh6/Myh7 locus; Left: row z-score of geometric means for all expressed isoforms with assigned module number and intron/exon transcript schematic for Ensembl91 and *de novo* assemblies, respectively; box and star, differential expression at stage transition.

D.) Summary of all Ensembl91 protein coding genes with annotated intragenic antisense transcribed elements vs Ensembl91 protein coding genes with detected antisense transcribed elements by *de novo* transcriptome construction. E.) Example of novel antisense transcript arising from *Agrn* locus. Intron/exon transcript schematic for Ensembl91 and *de novo* assemblies, respectively. red, (+) strand; blue, (-) strand.

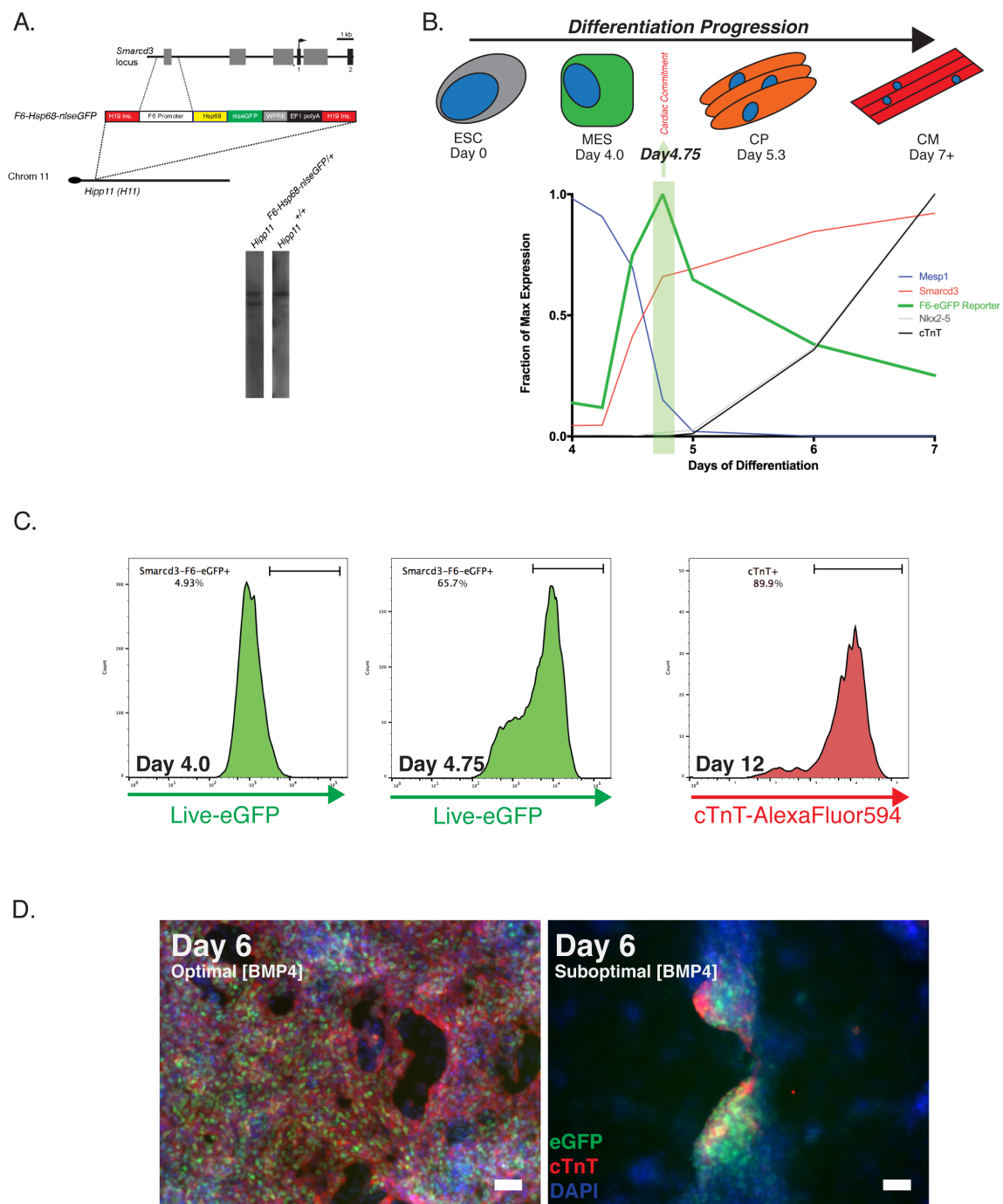


Figure 1.1. The Smarcd3-F6-eGFP mESC reporter system

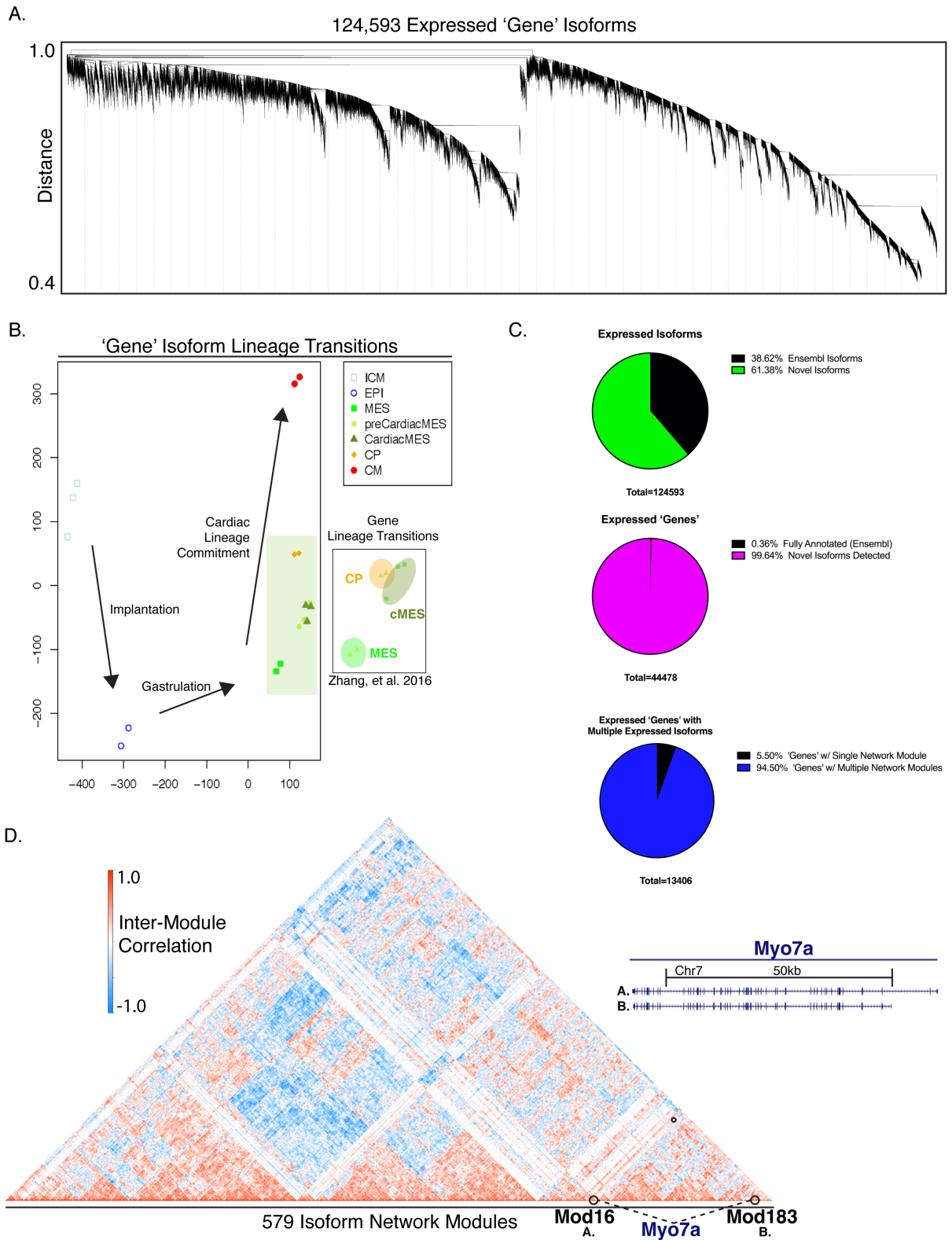


Figure 1.2. Transcriptional complexity during cardiac differentiation



E.

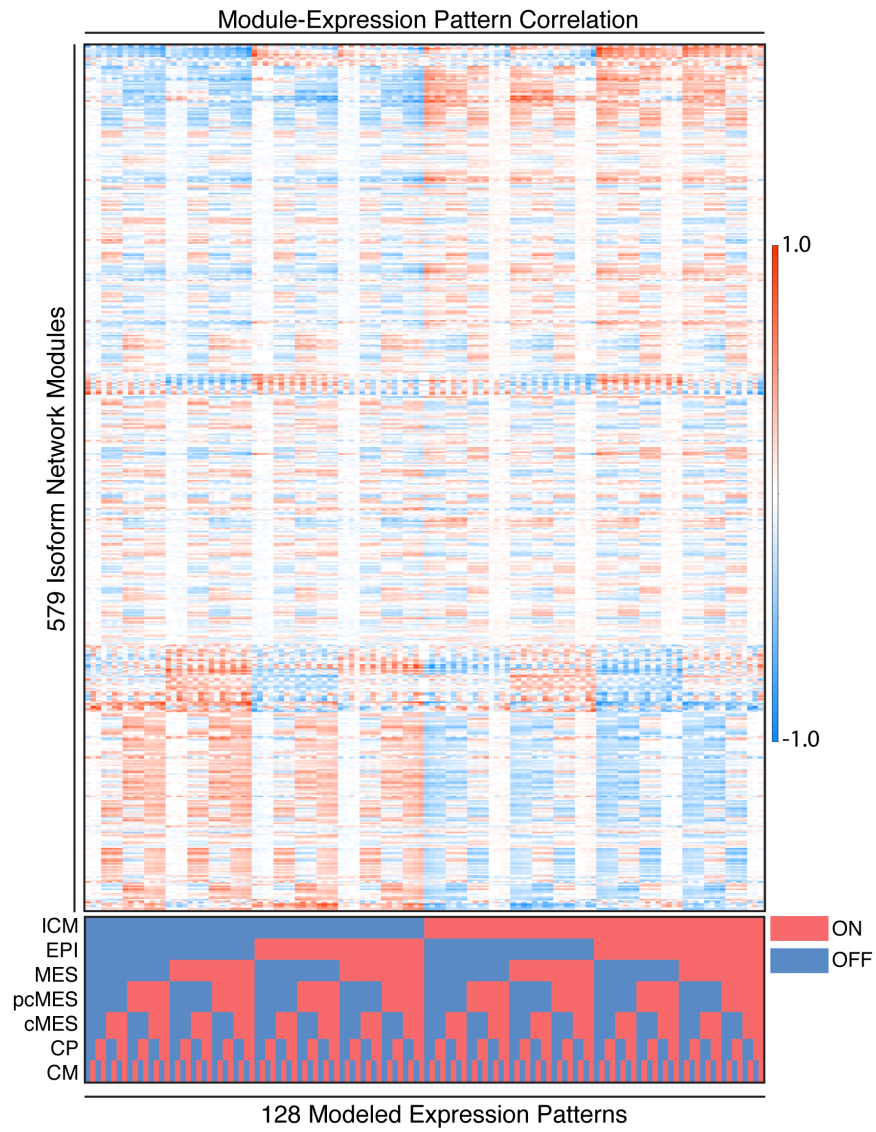


Figure 1.2. Transcriptional complexity during cardiac differentiation

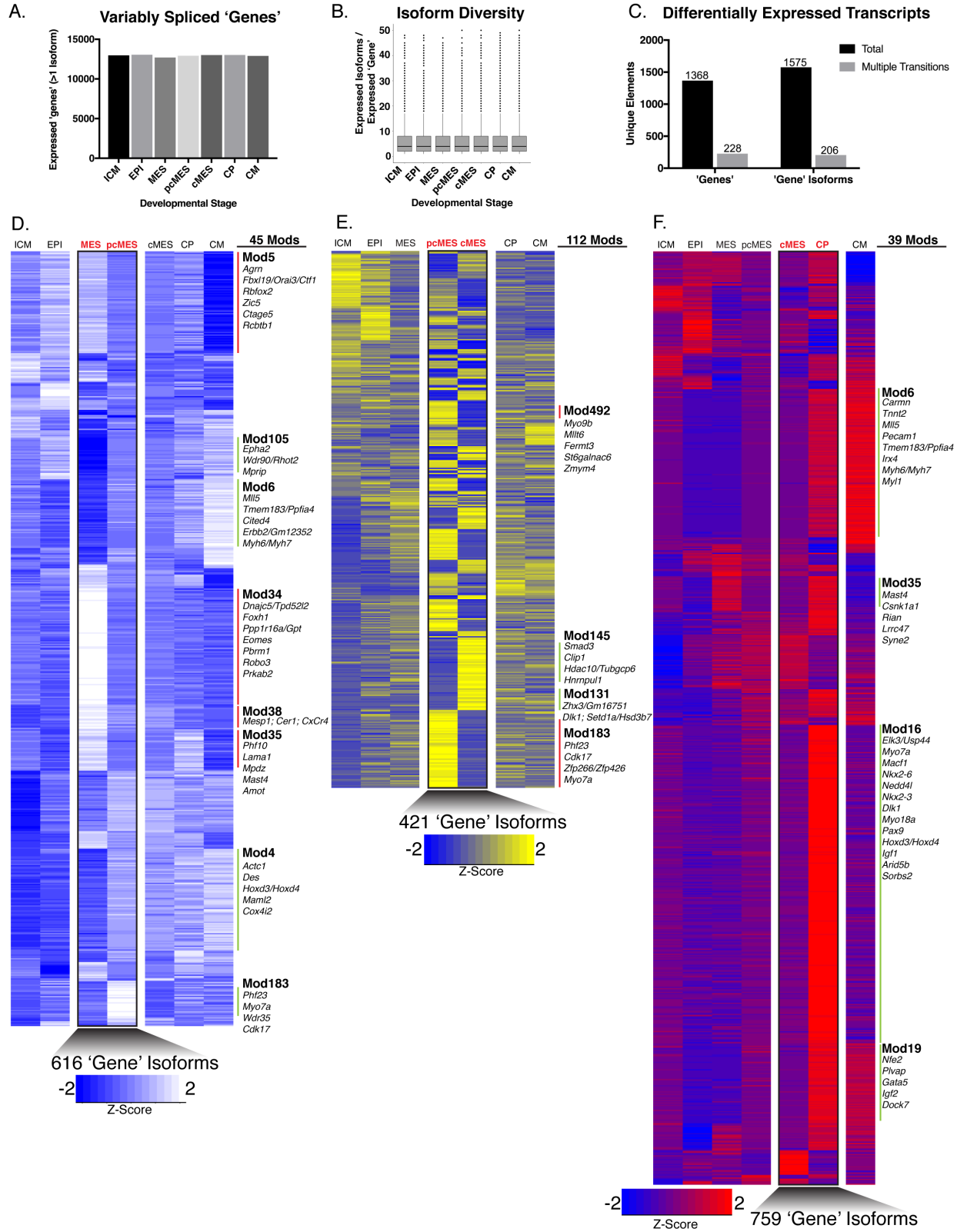


Figure 1.3. Differential 'gene' isoform expression at each phase of cardiac lineage commitment

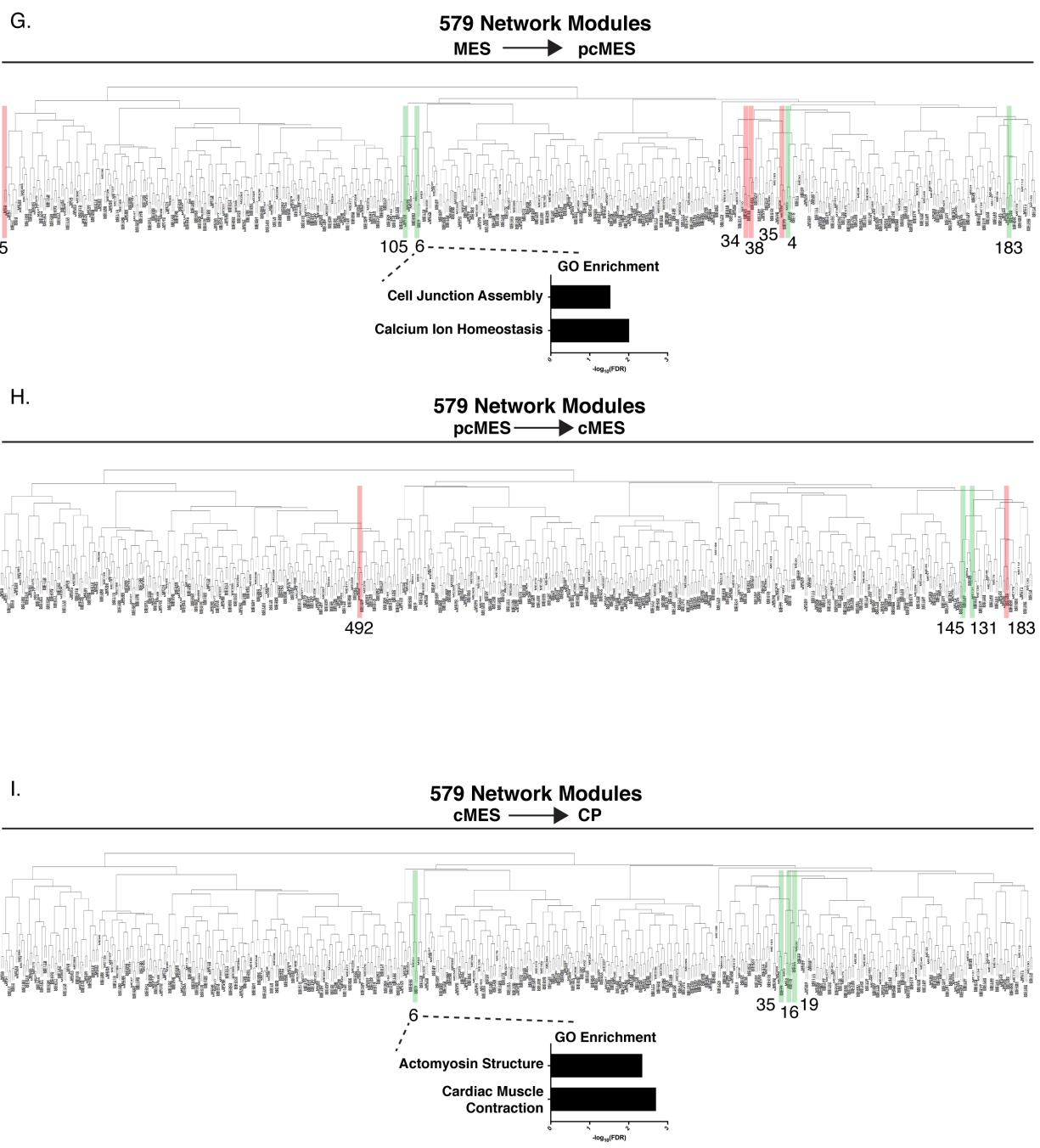


Figure 1.3. Differential 'gene' isoform expression at each phase of cardiac lineage commitment

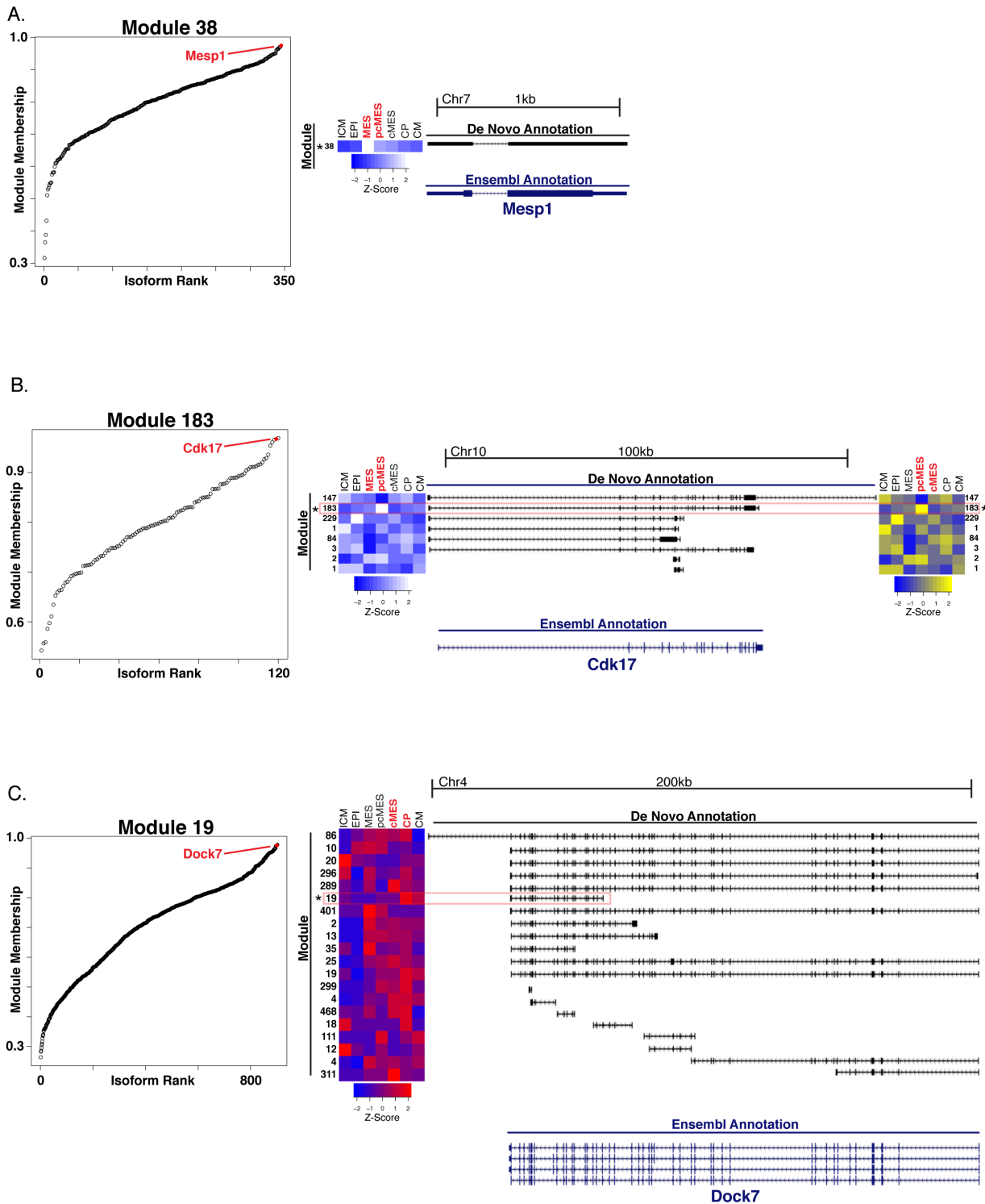


Figure 1.4. Differential expression of network module hub gene isoforms

D.

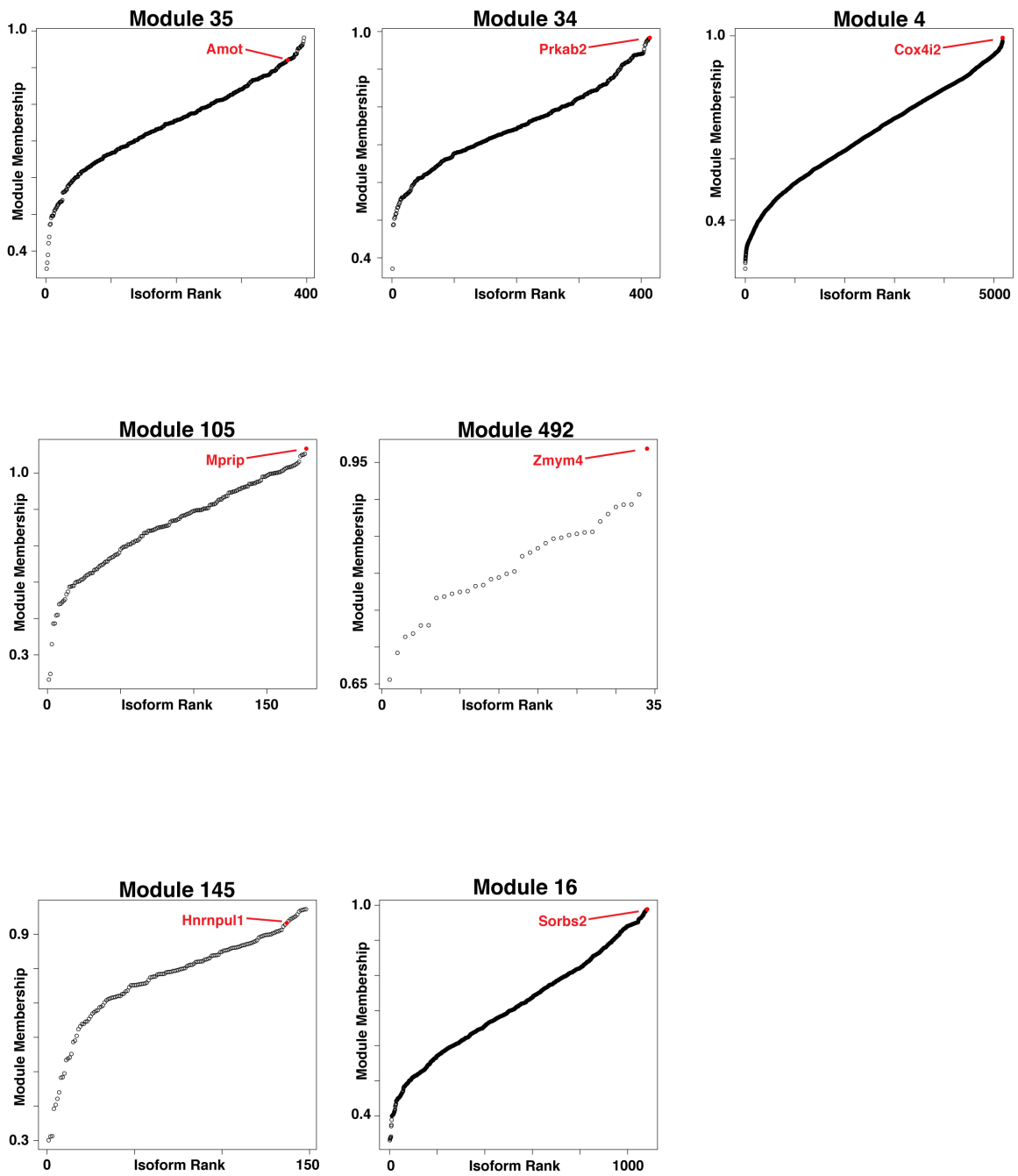


Figure 1.4. Differential expression of network module hub gene isoforms

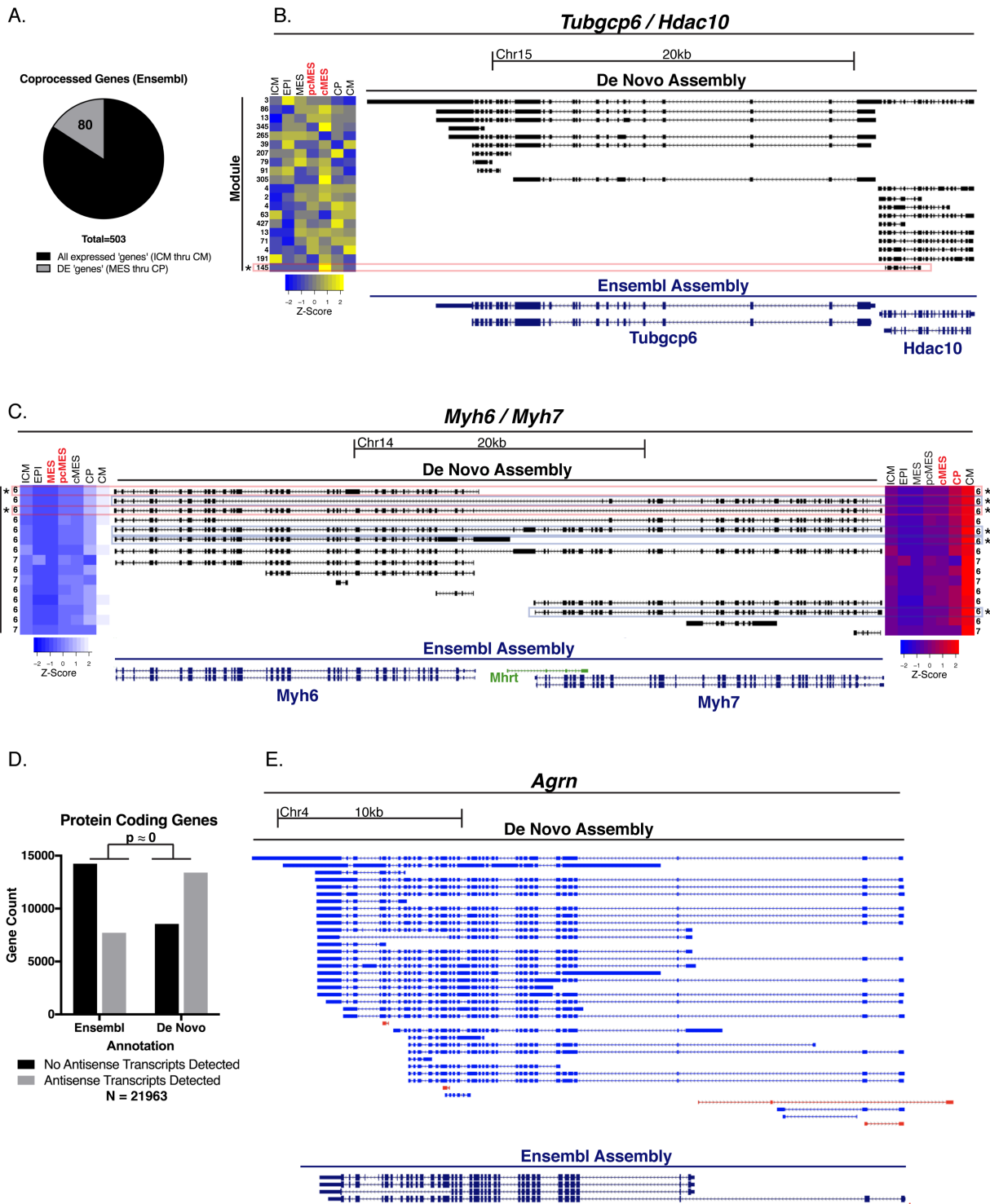


Figure 1.5. Multigene co-processing and intragenic reverse transcription

## Chapter 1 References

1. Meilhac, S. M., F. Lescroart, C. Blanpain and M. E. Buckingham (2014). "Cardiac Cell Lineages that Form the Heart." *Cold Spring Harbor Perspectives in Medicine* 4(9): a013888.
2. Perea-Gomez, A., A. Camus, A. Moreau, K. Grieve, G. Moneron, A. Dubois, C. Cibert and J. Collignon (2004). "Initiation of gastrulation in the mouse embryo is preceded by an apparent shift in the orientation of the anterior-posterior axis." *Curr Biol* 14(3): 197-207.
3. Costello, I., I. M. Pimeisl, S. Drager, E. K. Bikoff, E. J. Robertson and S. J. Arnold (2011). "The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation." *Nat Cell Biol* 13(9): 1084-1091.
4. Saga, Y., S. Miyagawa-Tomita, A. Takagi, S. Kitajima, J. Miyazaki and T. Inoue (1999). "MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube." *Development* 126(15): 3437-3447.
5. Kelly, R. G., M. E. Buckingham and A. F. Moorman (2014). "Heart Fields and Cardiac Morphogenesis." *Cold Spring Harbor Perspectives in Medicine* 4(10): a015750.
6. Kelly, R. G. (2012). "The second heart field." *Curr Top Dev Biol* 100: 33-65.

7. Bruneau, B. G., M. Logan, N. Davis, T. Levi, C. J. Tabin, J. G. Seidman and C. E. Seidman (1999). "Chamber-specific cardiac expression of Tbx5 and heart defects in Holt-Oram syndrome." Dev Biol **211**(1): 100-108.
8. Engleka, K. A., L. J. Manderfield, R. D. Brust, L. Li, A. Cohen, S. M. Dymecki and J. A. Epstein (2012). "Islet1 derivatives in the heart are of both neural crest and second heart field origin." Circ Res **110**(7): 922-926.
9. Schwartz, R. J. and E. N. Olson (1999). "Building the heart piece by piece: modularity of cis-elements regulating Nkx2-5 transcription." Development **126**(19): 4187-4192.
10. Devine, W. P., J. D. Wythe, M. George, K. Koshiba-Takeuchi and B. G. Bruneau (2014). "Early patterning and specification of cardiac progenitors in gastrulating mesoderm." Elife **3**.
11. Attanasio, C., A. S. Nord, Y. Zhu, M. J. Blow, S. C. Biddie, E. M. Mendenhall, J. Dixon, C. Wright, R. Hosseini, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, V. Afzal, B. Ren, B. E. Bernstein, E. M. Rubin, A. Visel and L. A. Pennacchio (2014). "Tissue-specific SMARCA4 binding at active and repressed regulatory elements during embryogenesis." Genome Res **24**(6): 920-929.
12. Wamstad, J. A., J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer and B. G. Bruneau (2012). "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage." Cell **151**(1): 206-220.



13. Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.
14. Pertea, M., D. Kim, G. M. Pertea, J. T. Leek and S. L. Salzberg (2016). "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown." Nat Protoc **11**(9): 1650-1667.
15. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat Protoc **7**(3): 562-578.
16. Risso, D., J. Ngai, T. P. Speed and S. Dudoit (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." Nature Biotechnology **32**: 896.
17. Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research **43**(7): e47-e47.
18. Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**: 559.
19. Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A **100**(16): 9440-9445.

20. Mi, H., A. Muruganujan, J. T. Casagrande and P. D. Thomas (2013). "Large-scale gene function analysis with the PANTHER classification system." Nat Protoc **8**(8): 1551-1566.
21. Kattman, S. J., A. D. Witty, M. Gagliardi, N. C. Dubois, M. Niapour, A. Hotta, J. Ellis and G. Keller (2011). "Stage-specific optimization of activin/nodal and BMP signaling promotes cardiac differentiation of mouse and human pluripotent stem cell lines." Cell Stem Cell **8**(2): 228-240.
22. O'Loughlin, T., T. A. Masters and F. Buss (2018). "The MYO6 interactome reveals adaptor complexes coordinating early endosome and cytoskeletal dynamics." EMBO Rep **19**(4).
23. Aase, K., M. Ernkvist, L. Ebarasi, L. Jakobsson, A. Majumdar, C. Yi, O. Birot, Y. Ming, A. Kvanta, D. Edholm, P. Aspenstrom, J. Kissil, L. Claesson-Welsh, A. Shimono and L. Holmgren (2007). "Angiomotin regulates endothelial cell migration during embryonic angiogenesis." Genes Dev **21**(16): 2055-2068.
24. The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Research **45**(D1): D158-D169.
25. Mulder, J., A. Ariaens, D. van den Boomen and W. H. Moolenaar (2004). "p116Rip targets myosin phosphatase to the actin cytoskeleton and is essential for RhoA/ROCK-regulated neuriteogenesis." Mol Biol Cell **15**(12): 5516-5527.

26. Ye, J., N. Beetz, S. O'Keeffe, J. C. Tapia, L. Macpherson, W. V. Chen, R. Bassel-Duby, E. N. Olson and T. Maniatis (2015). "hnRNP U protein is required for normal pre-mRNA splicing and postnatal heart development and function." Proc Natl Acad Sci U S A **112**(23): E3020-3029.
27. Molck, M. C., M. Simioni, T. Paiva Vieira, I. C. Sgardiolli, F. Paoli Monteiro, J. Souza, A. C. Fett-Conte, T. M. Félix, I. Lopes Monlléo and V. L. Gil-da-Silva-Lopes (2017). "Genomic imbalances in syndromic congenital heart disease." Jornal de Pediatria **93**(5): 497-507.
28. Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2013). "GenBank." Nucleic Acids Res **41**(Database issue): D36-42.
29. Katayama, S., Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M. Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard and C. Wahlestedt (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-1566.
30. Hilgenberg, L. G., B. Pham, M. Ortega, S. Walid, T. Kemmerly, D. K. O'Dowd and M. A. Smith (2009). "Agrin regulation of alpha3 sodium-potassium ATPase activity modulates cardiac myocyte contraction." J Biol Chem **284**(25): 16956-16965.

## Chapter 2

The BAF complex binds thousands of RNA transcripts within the nucleus via discrete domains of multiple subunits at the onset of cardiac lineage commitment

### Introduction

Interrogation into the transcriptional dynamics of cardiac lineage commitment revealed a strikingly diverse, yet highly connected, network of gene transcript splicing and expression. Therefore, we aimed to gain a better understanding of the molecular components that might influenced and integrate this complexity.

In order to fit nearly two meters of genomic DNA (gDNA) into the nucleus of a cell, it must be compacted several thousand-fold. This is in part accomplished by wrapping DNA around nucleosomes to form chromatin<sup>1</sup>. As a result, these nucleosomes must be dynamically arranged and moved throughout the genome to regulate chromatin structure and, thus, developmental and cell type-specific gene expression. ATP-dependent chromatin remodelers facilitate this movement, including Swi/Snf BRG1-associated factor (BAF) complexes<sup>2</sup>, which can induce both gene stimulation and repression. In mammals BAF complexes are polymorphic and comprised of distinct subunit arrangements, which result in varying protein and chromatin interactions in differentiating lineages. During heart development, BRG1 (SMARCA4), a core ATPase of the BAF complex, and several associated subunits are absolutely required for proper organogenesis<sup>3</sup>. As depicted in Chapter 1 of this manuscript, its subunit BAF60C (SMARCD3), one of three *Baf60* isoforms<sup>4</sup>, also conveys cardiac lineage-specific

function<sup>5</sup>. This includes expression in the first restricted heart-destined population, where it enhances molecular interactions between progenitor transcription factors and initiates the cardiac gene program<sup>6</sup>. Of interest, the most critical BRG1 function during cardiomyocyte differentiation also occurs at this early stage<sup>7</sup>. These factors implicate BRG1/BAF as a critical gene regulator at the onset of cardiogenesis. However, alternative lineages share similar subunit compositions. This suggests that additional layers of regulation determine BAF function during heart development.

RNA has been known to critically regulate the epigenomic state of the developing embryo for over twenty years<sup>8</sup>. However, it has only recently become apparent how pervasive this function is<sup>9</sup>. In fact, RNA interactions with chromatin modifying complexes may be as fundamental as those between proteins. Techniques meant to enrich and increase sensitivity to transcription continuously expand the scope of widespread RNA expression that permeates and connects the genome, even between distal loci<sup>10</sup>. This includes the expression of mRNAs, thousands of putative long noncoding RNAs (lncRNAs), nascent transcripts, as well as other RNA species. Chromatin-modifying complexes bind hundreds to thousands of RNA species through discrete subunit domains<sup>11,12</sup>. Additionally, epigenome-shaping complexes require RNA interactions for the maintenance of higher-order chromatin conformation<sup>13</sup>, gene expression<sup>14</sup>, and subsequent cell identity<sup>15</sup>. Recently, multiple groups have implicated specific interactions between BRG1/BAF and individual RNA transcripts that modulate and/or recruit its gene activating or repressing behavior<sup>16,17</sup>. However, to date, it is unknown how/if the BAF complex interfaces the transcriptome as a whole during its critical

function early throughout mesoderm commitment to the heart lineage.

We hypothesized that RNA binding was a fundamental component of the BAF complex's critical function during cardiac lineage specification. To test this, we developed and modified technologies to study BAF-RNA interactions from both protein- and RNA-centric perspectives. In doing so, we found that at least 7 subunits of the complex combined to stably bind tens of thousands of diverse RNA species within the nucleus via discrete, previously unappreciated protein domains. RNA interactions coincided with nearly every detectable stable DNA binding event of BRG1/BAF, as well as a majority of the loci that were differentially spliced and/or expressed during the transitions of cardiac lineage commitment. Finally, at this developmental window, we found numerous additional co-localized RNA binding proteins, including splice regulators, Wnt signaling effectors, and histone binding proteins.

## Materials and Methods

### *Smarce1 N- and C-terminal 3xFLAG tagging,*

We could not efficiently immunoprecipitate (IP) SMARCE1 via C-terminal 3x FLAG tagging. Therefore, we engineered pCDNA3.1(+) (Addgene) to contain both N- and C-terminal 3xFLAG tags, each with 8Gly polylinker sequences, flanking an AflIII restriction site. Smarce1 coding sequence was cloned from E9.5 mouse cDNA, ligated into AflIII-linearized vector with ColdFusion cloning kit (Systems Biosciences) and amplified with manufacturer-supplied e.coli. Cloning was verified by Sanger sequencing and restriction digest.

### *HEK293FT transient transfection*

HEK293FT cells were grown in HEK medium (DMEM + 10% FBS + 1X NEAA + 1x sodium pyruvate + 1X penicillin/streptomycin) to 80% confluence on gelatin coated 10cm tissue culture plates. One day before transfection, cells were split with TrypLE (Invitrogen) to allow 75% confluence the following day. For each 10cm dish, 9 $\mu$ g pCDNA-Smarce1-N-C-3xFLAG in 300 $\mu$ L DMEM (Gibco) were added to 27 $\mu$ g polyethyleneimine (PEI, 20kD linear, Sigma) in a separate aliquot of 300 $\mu$ L DMEM, vortexed immediately for 15 x 1s pulses, and allowed to complex at RT for 20 min. After this, 3.6ml DMEM + 2% FBS (HiClone) were added to complexes (4.2mL final volume) and pipetted gently onto D-PBS-washed HEK293FT cells. Transfection was allowed to proceed for 2-3 hours before replacing media with HEK medium. 48 hours later, cells were harvested in cold D-PBS by cell scraping. Nuclei were extracted using EZ Nuclei

Isolation Buffer (Sigma-Aldrich), pelleted at 500g for 5 min, and resuspended in 300 $\mu$ L LYSE buffer (20mM Tris-HCl pH7.4 + 120mM NaCl + 1% NP-40 + 0.2% sodium deoxycholate + 1X HALT (Fisher Scientific) protease inhibitors) per 10cm plate harvested. For gentle lysis, nuclei were swelled on ice in LYSE buffer and frozen overnight at -80°C.

### *Smarce1 CLIP and infrared RNA labeling*

Frozen Smarce1-N-C-3xFLAG-transfected HEK293FT nuclei were thawed quickly for 3 min at 37°C and placed on ice. Next, 0.5% (final) saponin + 0.1% (final) Triton X-100 + 20U TurboDNase were added, and nuclei were incubating at 37°C and 1200rpm in a thermomixer for 15 min. Subsequently, 10ul (4U/ $\mu$ L stock) RNase I (LifeTech) per 20M cells were added and nuclei were incubated again at 37°C and 1200rpm in a thermomixer for exactly 5 min before returning to ice. To extract, samples were rotated at 4°C for 30 minutes and centrifuged at 20,000g for 10 min. Supernatant was transferred to new 1.5mL tube and pipetted to mix. This extract was then evenly distributed to 4 new tubes, and NaCl (5M stock) was added to 150mM x 2, 450mM, and 750mM final concentration, respectively. After rotating for 30min at 4°C, extracts were transferred to 24 well ultra-low attachment plates (Corning) and UV crosslinked with 400mJ/cm<sup>2</sup>. The contents of each well were then added to 1.2mL IP buffer (50mM Tris-HCl + 120mM NaCl + 0.1% NP-40) containing 20ul packed Protein G magnetic beads (Pierce) complexed with 2 $\mu$ g FLAG M2 antibody (Sigma-Aldrich) to immunoprecipitate by rotating overnight at 4°C. The next day, beads were washed in 1.0mL IP buffer



followed by 1.0mL WASH buffer (50mM Tris-HCl + 1M NaCl + 1% Triton X-100 + 0.1% NP-40). Beads were transferred to a new tube for T4 PNK (NEB) reactions in PNK buffer (70mM Tris-HCl pH 6.5 + 10mM MgCl<sub>2</sub> + 1.0μL PNK + 4U murine RNase Inhibitor + 1U TurboDNase) at 37°C for 20 min. Then, beads were rinsed again with WASH and 3' end-tailed with ddATP-N<sub>3</sub> using yPAP (MCLab) in RNA TAILING buffer (20mM Tris pH 7.4 + 0.6mM MnCl<sub>2</sub> + 20μM EDTA + 100μg/mL BSA + 10% glycerol + 200U yPAP + 0.25mM ddATP-N<sub>3</sub> + 10U murine RNase Inhibitor) for 20 min at 37°C with shaking. Beads were sequentially washed with both IP buffer and WASH buffer and transferred to a new tube. Azide moieties were then infrared labeled in 50μL reactions containing 10μM ir800-DBCO (Licor) + 10U murine RNase Inhibitor in D-PBS for 40 minutes at 37°C with shaking. Labeled samples were again sequentially washed with both IP buffer and WASH buffer and transferred to a new tube. 2x 40μL elutions were then performed with 150μg/mL 3x FLAG peptide in IP buffer for 1 hour at 4°C with shaking, and supernatants from each elution were combined in a new tube. Volume was reduced to 15μL in a SpeedVac without heating and samples were run on a NuPage (ThermoFisher) 4-12% Bis-Tris gel in MOPS buffer at 150V for 90 minutes. Fluorescently labeled RNA was imaged in-gel on a LiCor Odyssey Fc imager. Total eluted SMARCE1 for each condition was then verified by SYPRO Ruby (ThermoFisher) protein stain and UV imaging.

#### *mESC culture and differentiation into cardiac mesoderm*

All experimental groups utilized the *Smarcd3*-F6nlEGFP; *Brg1*-3xFLAG mESC line

described in Chapter 1. Control (without 3xFLAG tag; NoFLAG) mESCs were generated using the same Smarcd3-F6nlseGFP reporter vector nucleofected into E14 mESCs. For culturing, ES cells were maintained in 2i + LIF media. Directed cardiomyocyte differentiations were performed as previously described by Wamstad et al<sup>18</sup> using the modifications described in Chapter 1. All subsequent experiments into BAF complex RNA binding were performed at day 4.75 of the protocol in samples containing at least 60% F6-eGFP<sup>+</sup> cells.

#### *Mass spectrometry and RNPxl pipeline*

Day 4.75 cardiac mesoderm was washed in cold D-PBS, crosslinked on ice with 800mJ/cm<sup>2</sup>, and 20-80M nuclei were isolated as before with EZ Nuclei Isolation Buffer (Sigma-Aldrich). Samples were swelled in LYSE buffer (+ 100U/mL RNaseOUT (ThermoFisher) for 5min on ice and frozen overnight at -80°C to disrupt membranes (300µL per 20M cells). The following day, nuclei were thawed quickly for 3 min at 37°C and placed on ice. 40U Turbo DNase per 20M cells were added and samples were placed on thermomixer for 5 min at 37°C and 1200rpm. Lysed nuclei were then centrifuged at 20,000g at 4°C for 20 min to clear. Supernatant was added to 1.5mL IP buffer (50mM Tris-HCl pH 7.4 + 120mM NaCl + 0.1% NP-40 + 1mM EDTA + 200U/mL RNaseOUT + 1X HALT protease inhibitors) containing 25µL packed bead volume of M2 FLAG magnetic beads. BAF complexes were immunoprecipitated (IP) by rotating overnight at 4°C. The next day, beads were washed 4 x with cold WASH buffer (50mM Tris-HCl pH 7.4 + 120mM NaCl + 0.1% NP-40) then transferred to new 1.5mL tube.

Bound material was eluted 2x with 60 $\mu$ L WASH buffer + 150 $\mu$ g/mL 3x FLAG peptide (Sigma-Aldrich) by rotating at 4 $^{\circ}$ C for 30min. Subsequently, samples were processed as described by Kramer et al<sup>19</sup>, with numerous modifications. 80U RNaseOUT, 1.5 $\mu$ g MS-grade Trypsin (Sigma-Aldrich), and 0.1% SDS (final concentration) was added to digest overnight at 37 $^{\circ}$ C. Next, size exclusion chromatography was performed using two passes through 7kD molecular weight cutoff polyacrylamide columns (Pierce). Samples were then diluted to 200 $\mu$ L to 90mM Tris-HCl pH 8.0, 3mM MgCl<sub>2</sub>, and 2M urea. 25U benzonase were added to digest all nucleic acid for 30 min at 37 $^{\circ}$ C, followed by addition of 1 $\mu$ g RNaseA (ThermoFisher) and 1U RNaseI<sub>f</sub> (NEB) 1hr at 52 $^{\circ}$ C in thermomixer at 800rpm. Finally, 1 $\mu$ g MS-grade Trypsin was again added to digest overnight at 37 $^{\circ}$ C. Prepped samples were diluted 1:1 with 2.5% trifluoro acetic acid (TFA) and desalted on graphite columns (Pierce). Samples were eluted off columns with 4 x 100 $\mu$ L 50% acetonitrile + 0.1 $\mu$ L% formic acid and dried in SpeedVac (ThermoFisher) without heating. For MS, each sample was resuspended in 10 $\mu$ L 0.1% formic acid. Samples were run in the Gladstone Mass Spectrometry Core on an Orbitrap Fusion mass spectrometer. The HPLC ran a gradient from 3%B to 36%B over 97 minutes, followed by a 10 min wash at 95%B. The total run time was 108 min. The mass spec collected a full scan in the Orbitrap at 120,000 resolution followed by MS/MS of the most abundant peaks with HCD fragmentation. The HCD normalized collision energy was 27%, and the MS/MS spectra were detected in the Orbitrap at 15,000 resolution. The mass spec triggered as many MS/MS spectra as it could with a duty cycle of 3 seconds. Between samples, trypsin-digested BSA standard was run as a blank to reduce carryover. Raw

mass spec reads were processed using the RNPxl<sup>19</sup> pipeline using a custom database of all Uniprot entries referenced to the 31 BAF complex subunits expressed during mESC through CM differentiation<sup>20</sup>. Centroided peaks were filtered using the cRAPome database<sup>21</sup>, negative controls without UV irradiation, and non BAF-specific IP'd material from nuclei without the endogenous 3xFLAG tag, respectively. Only top quartile-scoring RNA-peptide adducts were kept for analysis. Pipeline was performed in biological triplicate, and protein domains with multiple high-scoring adducts ( $-\log_{10} > 2.0$ ) and representation in all three replicates were deemed positive for RNA binding. For visualization, protein:RNA adducts were mapped to the UCSC genome browser (mm10) using iPIG software<sup>22</sup>. For non-BAF subunit RNA adduct identification, searches were performed against all mus musculus Uniprot entries, Uniprot entries associated with histone modifications, or Uniprot entries associated with signaling, respectively.

### *3° protein modeling*

Primary protein sequence flanking RNA binding domains of was input into the RaptorX<sup>23</sup> structure prediction server. Canonical protein domains were assigned using Uniprot KB<sup>24</sup>. Resulting structure predictions were visualized with Pymol software<sup>25</sup>.

### *Native RIP-seq*

At day 4.75 of differentiation, nuclei from both experimental (*Brg1*-3xFLAG tagged cells) and NoFLAG controls were isolated using EZ Nuclei Isolation Buffer (Sigma-Aldrich) and swelled in 300 $\mu$ L LYSE buffer (20mM Tris-HCl pH7.4 + 120mM NaCl + 1% NP-40 + 0.2% Sodium deoxycholate + 1X HALT (Fisher Scientific) protease inhibitors + 20U

murine Rnase Inhibitor (NEB)) per 10cm plate for 5 minutes. Experimental groups were processed in biological triplicate, while NoFLAG controls were collected in duplicate (20-40M nuclei per biological replicate). Nuclei were frozen overnight at -80°C, thawed quickly the next day at 37°C for 3 min, and placed on ice. 8U TurboDNase were added per 20M cells, and extraction was performed by incubating at 37°C with 1200rpm shaking in thermomixer for 10 minutes. Samples were pelleted at 20,000g and 4°C for 20 min, and 2% of nuclear supernatant was stored at -80°C as input control. The remaining supernatant was added to 30µL packed FLAG M2 magnetic beads (Sigma-Aldrich) in 2.0mL total IP buffer (50mM Tris-HCl pH 7.4 + 120mM NaCl + 0.1% NP-40 + 1mM EDTA + 100U murine RNase inhibitor + 1X HALT protease inhibitors). IP was performed by rotating at 4°C for 4 hours. Beads were washed 6 x with 0.5mL IP buffer and transferred to a new tube. Immunoprecipitated material was eluted with 3 x 250µL 150µg/mL 3x FLAG peptide (Sigma-Aldrich) in IP buffer. 0.2% SDS and 100µg/mL proteinase K were added, and samples were incubated at 37°C for 15 minutes to release RNA from protein. RNA was isolated with phenol:chloroform + isoamyl alcohol (125:24:1, pH 4.5, ThermoFisher) and further concentrated and cleaned with Zymo Spin Columns (Zymo Research). RIP RNA length was then normalized by incubating at 94°C in FAST AP buffer (10mM Tris HCl pH 7.4 + 5mM MgCl<sub>2</sub> + 100mM KCl + 0.02% Triton X-100) for 5 minutes. Libraries were amplified as per the eCLIP protocol described by Von Nostrand et al<sup>26</sup> using RiL19 adapter and unique molecular identifier (UMI) ligations. Samples were indexed with NEBNext (NEB) primers and sequenced on an Illumina Nextseq obtaining PE75 reads with PhiX controls. Raw read images were

sorted by index and converted to fastq files with assigned UMIs using bcl2fastq, which were trimmed using fastq-mcf<sup>27</sup>. These raw reads were mapped to the mouse genome (mm10) with STARv2.5.2b<sup>28</sup>. *de novo* transcriptome annotations were then generated with Stringtie<sup>29</sup> using the custom transcriptome created in Chapter 1 as a template. RIP transcript counts were measured using Cuffquant and Cuffnorm<sup>30</sup>, and variation was controlled using RUVs of the RUV-seq R package<sup>31</sup>. Transcripts were only deemed BAF complex RIP interactors if they were enriched  $\geq 3$ -fold over input and NoFLAG controls with  $q < 0.05$  for both comparisons (calculated with Limma R package<sup>32</sup>).

#### *eCLIP-seq*

At day 4.75, cells were washed in cold D-PBS, crosslinked on ice at 350mJ/cm<sup>2</sup>, and nuclei were isolated, swelled, and lysed as described above for native RIP-seq (stored at -80°C). Experimental samples were generated in biological duplicate by pooling irradiated nuclei from 5-7 separate differentiations (400-500M nuclei, each; 900M nuclei total). NoFLAG control samples were also generated in biological duplicate of 50M pooled nuclei per sample. Nuclear extractions were performed as for native RIP-seq with 1 modification. After TurboDNase treatment, 40U of RNase I (ThermoFisher) were added, and samples were again incubated in a thermomixer for exactly 5 additional minutes. Subsequently, nuclear extractions, 2% input control storage, M2 FLAG IPs, bead washes, and elution was performed as described for native RIP-seq on both experimental and NoFLAG control groups. NoFLAG eluate was then stored at -80°C. Eluted BAF complexes and accompanying crosslinked RNA were then input into a second IP with antibodies targeting ARID1A (Santa Cruz), PBRM1 (, SMARCA4,

SMARCC1, SMARCC2, SMARCD3, or SMARCE1, respectively, rotating overnight at 4°C. A secondary IP was also performed using anti-NEUROD4 non-BAF subunit antibody for reference. Samples were then processed as per the eCLIP<sup>26</sup> protocol to generate RNA-seq libraries for each subunit-specific IP. This protocol required samples to be immobilized 3 days on beads for all processing steps. To enrich protein-crosslinked RNA, BAF-specific samples, input controls, and NoFLAG controls were electrophoresed by SDS-PAGE in NuPage (ThermoFisher) 4-12% Bis-Tris gels with MOPS buffer at 150V. To visualize end products from each individual IP, 20% of each sample was run in parallel and stained in-gel with SYPRO Ruby. Only the most abundant BAF subunits (i.e. SMARCC1) could be detected after the 3-day process. Therefore, to validate proper antibody targeting, we performed another dual IP for each antibody (BRG1-3xFLAG followed by subunit-specific IP) with equivalent buffers and washes performed over a single day. To maintain size-matched consistency, we extracted 50kD to 300kD regions for all samples and controls. Libraries were quantified by RT-qPCR to determine the minimum required cycles for sufficient amplification. All subunit-associated libraries could be adequately amplified with less than 24 cycles, while Smarce1 libraries could be amplified with 14-16 cycles, similar to size-matched input control libraries. Indexed eCLIP libraries were sequenced on an Illumina Nextseq obtaining PE75 reads with PhiX controls. Raw read images were sorted by index and converted to fastq files with assigned UMIs using bcl2fastq, which were trimmed using fastq-mcf. These raw reads were mapped to the mouse genome (mm10) with STARv2.5.2b using the following options:

```
--genomeLoad LoadAndKeep --outSAMtype BAM SortedByCoordinate --  
readFilesCommand zcat --outSAMattrIHstart 0 --outFilterType BySJout --  
outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin  
20 --chimSegmentMin 20
```

PCR artifacts were removed using UMI-tools<sup>33</sup>. Read 2 (oriented to proper RNA strand) from each paired bam file was extracted with samtools and input to the GEM/GPS<sup>34</sup> peak caller using supplied CLIP read distribution file and the following options:

```
--strand_type 1 --nrf --nf --smooth 5 --nd 2 --outBED --local_control --f SAM --k_min 4 --  
k_max 14 --poisson_control --k_neg_dinu_shuffle
```

Only significant GPS peaks with enrichment greater than 10-fold over size-matched input libraries (q value < 0.01) were kept. Of these, peaks were only assigned to a particular subunit if both biological replicates overlapped within a 7nt resolution around the called binding event. Furthermore, any binding region that overlapped a GPS called peak within a 20nt window of NoFLAG controls was removed. For binding regions assigned to multiple BAF subunits (due to incomplete complex dissociation), the peak was assigned to the IP that produced the greatest enrichment over size-matched input controls.

### *eCLIP Analysis*

The distribution of eCLIP peaks for each subunit were calculated against Ensembl<sup>91</sup><sup>35</sup> annotations, the de novo transcriptome described in Chapter 1, enhancer regions called



by Wamstad et al<sup>18</sup>, and PSYCHIC<sup>36</sup> enhancer domains, respectively. All charts and graphs were generated in R and Prism 7. Wiggle tracks were generated using Eseq v1.05<sup>37</sup> and uploaded to the UCSC Genome Browser for visualization. Statistical enrichment of eCLIP peaks within differentially spliced genomic regions was computed by hypergeometric tests. Gene ontology (GO) enrichment was calculated using Gorilla<sup>38</sup> using GPS called peaks on size-matched Input controls as background.

#### *RNA sequence motif enrichment*

For all eCLIP binding sites, stranded sequence was expanded to approximately 12 nucleotides and motif enrichment analysis was performed with HOMER<sup>39</sup> using the -rna, -len 6,8,10 and -noweight options. Sequences from eCLIP size-matched Input controls were used for background comparison.

#### *RNA 2° structure prediction*

For each subunit, the top 27-30 enriched RNA binding sites that also could be mapped to detected transcript exons were selected for secondary structure prediction. From each binding region (approximately 7 nucleotides), stranded sequence was expanded 40nt upstream and downstream to provide local folding environment. These sequences were then input into TurboFold v6.0.1<sup>40</sup>. Resulting 'ct' files were then visualized with Assemble2.3<sup>41</sup>. Random 86nt sequences extracted from all annotated protein exons were also processed for comparison.

### *Live-cell permeabilization and immunocytochemistry*

At day 4.0 of differentiation, 3x FLAG cells were plated into gelatin-coated 8-well  $\mu$ -Slides (Ibidi) at 300K cells per well. 18 hours later (day 4.75), wells were washed with D-PBS<sup>+/+</sup> (with Mg<sup>2+</sup> and Ca<sup>2+</sup>). To extract soluble nuclear components from native nuclei (before crosslinking), cells were incubated for 3 min in cold CSK buffer (10 mM PIPES-KOH pH 7.0 + 100 mM NaCl + 300 mM sucrose + 3 mM MgCl<sub>2</sub> + 1X HALT protease inhibitors) with 0.5% Triton-X100. Then wells were gently washed 2x with D-PBS<sup>+/+</sup> and fixed in 3% paraformaldehyde in D-PBS<sup>+/+</sup>. For BRG1-3xFLAG immunocytochemistry, wells were blocked and permeabilized with D-PBS containing 0.1% Triton X-100 and 10% goat serum (Hyclone) for 30 min at room temperature (RT). Then, cells were stained overnight at 4°C with 1:100 dilution of mouse anti-FLAG M2 antibody (Sigma) in WASH buffer (D-PBS + 1-% goat serum + 0.1% Tween-20). The following day, wells were gently washed 3 x 5 min with WASH buffer. Subsequently, goat anti-mouse AlexaFluor-594 (1:1000, ThermoFisher) in WASH buffer was added for 2 hours at RT. Wells were again washed 3 x 5 min with WASH buffer, where DAPI (1:300 of 1mg/mL stock) was included in final wash. D-PBS was then added for imaging on a Keyence epifluorescence microscope with 100X objective using Zeiss Immersol 518F oil.

### *Chromatin Immunoprecipitation and sequencing*

Chromatin immunoprecipitation (as adapted by Hota et al, in review<sup>20</sup>) of day 4.75 cardiac mesoderm was performed according to O'Geen et al<sup>42</sup> with modifications in

duplicate from pooled material of 2-4 differentiations (30M cells per sample). Briefly, cells were harvested and crosslinked with 2mM disuccinimidyl glutarate (DSG) for 45 mins, washed twice with PBS, followed by crosslinking with 1% formaldehyde for 15 mins. Cells were quenched with 0.125M glycine for 5 mins, washed in D-PBS thrice and stored at -80°C. Frozen pellets of dual crosslinked cells were thawed, washed in PBS and lysed in D-PBS containing 1% Triton X-100 for 3 mins on ice followed by 25 cycles of dounce with a tight pestle. Nuclei were collected at 350g for 5min and washed once with MNASE buffer (50mM Tris-HCl, pH7.6, 1mM CaCl<sub>2</sub>, 0.2% Triton X, 5mM sodium butyrate with 1x HALT protease inhibitor and 0.5mM PMSF added just before use). Nuclei were digested with 400U of micrococcal nuclease (ThermoFisher) for exactly 5 mins at 37°C. MNase digestion was stopped by adding STOP buffer (10mM EDT, 10mM EGTA and 0.1% SDS). Chromatin were sonicated for short time (2 cycles, 30s ON, 1 min OFF at output 4 in a VirSonic sonicator), centrifuged at 10,000g for 10mins and supernatant stored at -80°C. Extent of MNase digestion was measured by agarose gel electrophoresis. Chromatin (40ug) was diluted to 5-fold in IP dilution buffer (50mM Tris.Cl, pH7.4, 150mM NaCl, 1% NP-40, 0.25% sodium deoxycholate, 1mM EDTA) without EDTA and pre-cleared with 25ul of M-280 goat anti-rabbit IgG dyna beads (ThermoFisher) for 2 hrs followed by addition with 1ug of anti-Brg1 antibody (Abcam, 110641) for 12-16hrs. 5% of samples were set aside as input before antibody addition. Antibody bound BRG1-DNA complexes were immunoprecipitated using 25ul of M-280 goat anti-rabbit IgG dyna beads for 2hrs, washed twice with IP dilution buffer, five times with IP wash buffer (100 mM Tris HCl, pH 9.0, 500mM Lithium chloride, 1% NP-40 and

1% Sodium deoxycholate) and thrice with IP wash buffer containing 150mM NaCl. DNA was eluted with 200ul of elution buffer (10mM Tris.Cl, pH 7.5, 1mM EDTA and 1%SDS), NaCl was added to both ChIP and input samples (0.52M), and crosslinking was reversed for 6 – 12 hrs at 65°C. Samples were digested with RNase A, and proteinase K in presence of 3ug of glycogen, and DNA purified with Ampure beads (Beckman Coulter) and eluted in 50ul of TE. To prepare libraries for ChIP-Sequencing, DNA was end repaired, A-tailed followed by adapter ligation (Illumina TruSeq) and 14 cycles of PCR amplification. PCR amplified DNA was size selected (200 -400bp) followed by gel extraction (Qiagen) and eluted in 20ul TE. The concentration and size of eluted libraries was measured (Qubit and Bioanalyzer) before sequencing in a NEBNextSeq sequencer. Reads (single end 75bp) were trimmed using fastq-mcf and aligned to mouse genome mm10 assembly using Bowtie2<sup>43</sup>. Minimum mapping quality score was set to 30. Peaks were called with MACS2 and default settings using Galaxy. Heatmaps of aligned reads were generated with Easeq v1.05 software. BRG1 bound regions within 5kb of gene were directly assigned to said gene. Intergenic bound regions were assigned to target genes using PSYCHIC enhancer-promoter database adapted from mm9 to mm10 with UCSC liftOver.

## Results

### ***At the onset of heart lineage commitment, at least 7 Brg1-associated BAF complex subunits bind RNA via discrete domains.***

The BAF complex is critical for establishing the gene regulatory environment at early stages of cardiac mesoderm development. Since this complex could interact with individual noncoding transcripts, we hypothesized that RNA binding was a fundamental molecular component of its function, including at this important developmental time period. Moreover, we predicted this functionality would result in pervasive interactions with numerous RNA transcripts. To first test this, we designed an immunoprecipitation assay using HEK293FT nuclei that were transiently transfected with a Smarce1 dual 3x FLAG-tagged vector. In this experiment, sub-nanometer nuclear RNA:protein binding events stable in increasing concentrations of NaCl were fixed with 254nm ultraviolet (UV) radiation. SMARCE1 was immunoprecipitated to pull down BAF complex components along with their fixed RNA partners, and these accompanying transcripts were fluorescently labeled at stoichiometric ratios (1:1) (Fig1A). Thus, RNA adduct-induced electrophoretic mobility shifts could be visualized. We observed UV-dependent shifts at molecular weights synonymous with core BAF complex subunits that could not be interrupted by up to 750mM NaCl prior to crosslinking (Fig1B). This gave the indication that RNA binding could be occurring throughout the complex via multiple subunits. We next looked to identify which complex members might engage in RNA interactions during cardiac lineage commitment. To do this, we employed a modified approach to the RNPxl pipeline<sup>19</sup> on isolated nuclei from UV-crosslinked day 4.75

cardiac mesoderm. After immunoprecipitating SMARCA4 along with its subunit constituents using an endogenous 3xFLAG tag, we digested proteins into tryptic peptides and used size exclusion chromatography (SEC) to isolate peptides bound to larger RNA molecules. Next, we digested RNA into short (1-4nt) peptide-bound fragments, thereby creating pseudo-phosphopeptides that could be purified on graphite columns. Finally, these peptide-RNA species were analyzed by mass spectroscopy with RNPxl algorithms to detect and score the confidence of BAF complex fragments covalently bound to RNA (Fig1C). Due to the numerous possible nucleotide-peptide permutations of RNA composition, RNA length, and chemical modifications incurred during library preparation, a single RNA-bound peptide region could produce numerous mass spectra. Among top quartile RNA adduct-scoring peptides, we looked for those containing multiple high confidence binding scores and reproducible (both within and between biological replicates, n = 3) intra-protein RNA adducts (Fig1D). As a result, seven of the 31 known BAF complex subunits expressed during cardiac differentiation<sup>20</sup> fulfilled these criteria, and these RNA crosslinks were UV-dependent and specific to the BRG1 immunoprecipitation. In agreement with Han et al<sup>44</sup> and Cajigas et al<sup>45</sup>, we observed discrete RNA binding domains to exist within SMARCA4 (BRG1), SMARCC1 (BAF155), and SMARCC2 (BAF170). Additionally, our analyses were able to add ARID1A (BAF250A), PBRM1 (BAF180), SMARCD3 (BAF60C), and SMARCE1 (BAF57) to this list. Importantly, as SMARCA4 was the sole bait used for immunoprecipitation, all these identified subunits were in complex with at least SMARCA4. These data established RNA binding through discrete domains of numerous BAF complex subunits

at the onset of cardiac lineage commitment.

***Structural modeling of BAF subunits reveals spatial orientation of functional domains to RNA-binding domains.***

Mass spectrometry detected consistent UV-induced RNA adducts within distinct regions of BAF proteins of between 40 and 140 amino acids (Fig2A). We next chose to investigate the spatial relationship of these RNA binding domains (RBD) to other canonical functional groups. To do this, we used RaptorX protein structure prediction<sup>23</sup>, which employed artificial deep neural networks to solve folding conformations, even without widely annotated sequence or structural homology for comparison. ARID1A, a subunit required for heart development and BAF complex DNA occupancy<sup>46</sup>, displayed an orientation in which the N- and C- termini folded within the protein to position four LxxLL motifs in proximity to its ARID A-T rich interacting domain. Furthermore, our data indicated its RBD comprised a region lacking clear 2° structure that folded along an interface with this assembly (Fig2B). This posited its RNA transcript interactions into physical proximity to the protein's ARID-dependant DNA binding<sup>47</sup> and LxxLL-mediated transcription factor coactivation<sup>48</sup>.

Our model of PBRM1, a defining constituent of the pBAF complex required for heart chamber development<sup>49</sup>, aligned bromodomain (BRD) 3 thru 6 parallel to each other in around a central axis. Within this region, it bound RNA via an unfolded linker between BRD 4 and 5, as well as along one of the four  $\alpha$ -helices of BRD 5 (Fig2C). These points of binding inserted PBRM1 RNA interactions into a position to influence bromodomain

recognition of target histones<sup>50</sup>.

Previous groups' attempts to determine the RBDs of SMARCA4 through *in vitro* experiments claimed it to be within either the DExxc N-terminal portion of its helicase domain<sup>44</sup> or HSA and BRK domains of the N-terminal half of the protein<sup>45</sup>, respectively. However, our crosslinked snapshot of actual interactions taking place within the nucleus pointed in the opposite direction. Our structural models predicted the SMARCA4 BRD folded back against its C-terminal helicase domain. The predominance of RNA binding subsequently took place within the tether between these groups, which flanked SnAC and AT-hook domains (Fig2D). These two regions were previously shown to be critical for nucleosome remodeling<sup>51</sup> and DNA binding<sup>52</sup>, respectively. Therefore, RNA binding within BRG1 could physical impact numerous protein functions, including domain conformation, DNA and histone recognition, and catalytic activity.

Despite sharing approximately 70% amino acid homology as well as common core domains, the BAF scaffold subunits SMARCC1 and SMARCC2 contained RBDs in distinct regions of each respective protein. Each protein model positioned SANT histone binding<sup>53</sup> and SWIRM BAF assembly domains<sup>54</sup>, respectively, held in place within a structural coiled coil scaffold (Fig2E, F). Our calculated models inferred that SMARCC1'S RNA interface lied within an unstructured loop connecting its coiled C-terminal end with the SANT and SWIRM domains (Fig2E). On the other hand, SMARCC2'S RBD contained an  $\alpha$ -helix tightly folded near its SWIRM domain (Fig2F). Therefore, SMARCC2's RNA interface was expected to be more likely to directly impact canonical functions of its core domains such as complex assembly and histone



interactions. In contrast, SMARCC1's RBD was modeled to engage its RNA targets distal to these integral structures. However, this region contained residues capable of being phosphorylated and forming isopeptide crosslinks with SUMO2<sup>55</sup>. Therefore, SMARCC1 RNA binding could play a role in subunit post-translational modification.

SMARCD3, a required subunit for proper heart gene expression programs<sup>56</sup>, contained its RBD at the N-terminus. However, this region did not exhibit any predicted structural connection with the predominant SWIB domain-containing core of the protein (Fig2G). Therefore, RNA binding was unlikely to play a direct role in its cofactor recruitment to chromatin<sup>57</sup> but was instead expected to participate in intermolecular contacts with other complex constituents. In contrast, the core subunit SMARCE1 bound RNA directly between a tight assembly of a structural coiled coil region and high mobility group (HMG) box DNA-binding domain (Fig2H). Therefore, its RBD was positioned in direct proximity to its HMG box chromatin interface<sup>58</sup>.

These models provided context into the spatial organization of BAF subunit RBDs in relation to canonical functional domains. Of note, each subunit RBD contained substantial stretches of amino acids without clear secondary structure. Other groups previously found that RNA recognition domains frequently existed in unstructured free states and only became ordered after undergoing RNA interactions<sup>59</sup>. We therefore suggested that these BAF RBDs could have strong allosteric impacts throughout their respective proteins upon the conformational changes that coincided RNA detection.

***Multifaceted transcription within the nucleus engages the BAF complex via tens of thousands of discrete binding events***

We wanted to better understand the identity of RNA molecules that were interacting with complex subunits during cardiac lineage commitment. To test this, we first performed RNA immunoprecipitation with high throughput sequencing (RIP-seq) under physiological salt conditions (120mM NaCl) at day 4.75 of differentiation. In doing so, we isolated nuclei and used the endogenous 3xFLAG tag to pull down BRG1 with any stably bound subunits and/or RNA molecules. Over 2500 individual transcript isoforms were significantly associated with the complex at this critical transition, and greater than 50% of those were unannotated in the Ensembl 91 database. These included new splice isoforms as well as antisense transcripts throughout canonical gene bodies. Additionally, hundreds of BAF-enriched transcripts were undetected even in the *de novo* transcriptome we previously generated throughout mESC to cardiomyocyte differentiation in Chapter 1 (Fig3A). These results suggested novel and often lowly expressed transcripts were the RNA species interfacing BRG1/BAF.

To improve the sensitivity and specificity in detecting each subunit's RNA interactome, we UV-crosslinked the same cardiac mesoderm at day 4.75 and normalized transcript lengths with short RNase treatment. Again, we first pulled down BRG1/BAF from isolated nuclei in 120mM NaCl. Subsequently, eluted complexes from this step were input into a second IP with antibodies targeting ARID1A, PBRM1, SMARCA4, SMARCC1, SMARCC2, SMARCD3, and SMARCE1, respectively, under high salt and reducing conditions. Libraries were then prepared using the eCLIP protocol<sup>26</sup>, which

was designed to allow cDNA synthesis up to protein:RNA adducts induced by UV irradiation. This two-step IP and subsequent library preparation required immobilization of protein targets for three total days with numerous high stringency washes. Only after this extended period did we notice obvious purification of individual target subunits from other BAF proteins (Fig3B). RNA-sequencing of subunit-specific libraries successfully generated between 1.3M (ARID1A) and 21.0M (SMARCE1) mapped reads (mm10) after PCR duplicate removal. In doing so, we were able to dissect subunit-specific binding at gene loci that revealed diverse transcript binding characteristics (Fig3C). We found greater than 33,000 instances of discrete, reproducible, and specific RNA binding events with BAF subunits. Furthermore, SMARCE1 and SMARCA4 were responsible for over 27,000 and 3700 of these, respectively, while PBRM1 and ARID1A engaged only 236 and 254, respectively (Fig4A). For each locus, including the 9896 protein coding genes containing eCLIP binding peaks, subunits bound RNA in single events. However, SMARCE1 most often engaged an individual gene's transcripts in 1-3, and upwards of 6, events (Fig4B). Moreover, at protein coding loci, approximately half of all bound RNA transcripts were antisense to the protein mRNA strand (Fig4C). This was true regardless of whether a particular gene had detectable antisense transcript annotations either by Ensembl 91 or even the *de novo* transcriptome generated in Chapter 1. Therefore, we could conclude that intragenic antisense transcription within canonical genes was often responsible for engaging the BRG1/BAF complex. For all RNA-binding subunits, the majority of eCLIP peaks lied within 2kb of protein coding genes (min = 64%, ARID1A; max: 87%, SMARCA4). Also, between 90%-98% of each

subunit's eCLIP binding events could be assigned to a detected transcribed element (Ensembl 91- or *de novo*-annotated coding and noncoding loci) or enhancers.

Our experiments showed that BAF subunits maintained association with each other despite stringent IP conditions. Thus, even with the enrichment for each 2° IP target after 3 days of the eCLIP protocol, we postulated that proteins with the strongest interactions (i.e. direct contacts) to each other would also overlap in their identified RNA binding profiles. We in fact did observe this, as 4646 of the eCLIP events were shared by more than one subunit-specific IP, albeit to varying levels of enrichment.

Furthermore, these inter-subunit agreements were not random. In fact, three pairs of BAFs were most-correlated in their exact agreement for enriched eCLIP binding peaks. Pulldowns targeting SMARCA4 and SMARCC1, PBRM1 and SMARCC2, as well as SMARCE1 and SMARCD3, respectively, displayed a high degree of identically-assigned RNA binding regions (Fig5A). We next modeled inter-subunit amino acid binding probabilities using RaptorX complex contact prediction\*. In doing so, we were able to predict the likelihood of interfacial contacts flanking each pair's respective RBDs. While the primary functional domains of SMARCC1 showed strong interacting probabilities with SMARCA4's helicase and bromodomain, our model also indicated that SMARCA4 interacted with RNA along its interface with the SMARCC1 SWIRM domain. However, we could not infer any clear interaction between SMARCC1's RBD and the C-terminus (amino acid 1071-1570) of BRG1 (Fig5B). Thus, we concluded that SMARCA4 was most likely to bind RNA at a convergent, multi-subunit chromatin interface, while SMARCC1's RBD lied outside this structural region.

We previously modeled the RNA binding domains of PBRM1 and SMARCC2 to flank their BRD and SWIRM chromatin-interacting regions, respectively. Furthermore, these canonical domains were predicted to contain high probability intermolecular amino acid contacts. In addition, the PBRM1 and SMARCC2 RBDs were also calculated to complex at a common interface. Therefore, RNA binding and chromatin binding by these subunits was modeled to take place within a shared functional space.

While the N-terminus of SMARCD3 was not predicted to engage in intramolecular folding, this region was modeled to contact the HMG box, RBD, and coiled coil domains of SMARCE1. These regions of SMARCE1 in turn could broadly interact throughout SMARCD3, including with the SWIB cofactor-recruiting domain. Therefore, this model suggested that SMARCD3 and SMARCE1 could bind RNA at a common interface, and any intramolecular contacts by the SMARCD3 N-terminal RBD would likely require SMARCE1 intermediates.

***SMARCE1 and SMARCA4 RNA-binding domains favor GC-rich transcript regions, while ARID1A, PBRM1, SMARCC1, SMARCC2, and SMARCD3 consistently interact with RNA 2° structure.***

The presence of discrete RNA binding domains within BAF subunits and the tight resolution of immunoprecipitated transcript regions led us to predict that these proteins recognized either sequence or 2° structural motifs. We performed primary motif enrichment analysis with HOMER<sup>39</sup> algorithms comparing 12 nucleotide windows flanking eCLIP binding sites for ARID1A, PBRM1, SMARCA4, SMARCC1, SMARCC2,

SMARCD3, and SMARCE1, respectively. We were unable to identify notable enrichment for any 1° sequence. However, the two most promiscuous RNA binding subunits, SMARCA4 and SMARCE1, each showed strong preference for GC-rich regions (Fig6A). In agreement with this, we often observed their associated eCLIP peaks to reside near transcriptional start sites (TSS), which frequently are known to be enriched for CpGs<sup>60</sup>. Next, we asked if these seven subunits bound their RNA targets via recognized secondary structure. For each protein, we selected the most enriched RNA crosslink sites that mapped to detectable transcript exons. From these exons, we extracted approximately 85nt sequences centered around the eCLIP peaks. We then used TurboFold<sup>40</sup> software to test for homologous secondary structures aligned at RNA binding sites between these highly enriched RNA molecules. Regular features were not consistently aligned in randomly selected exon centers (n=30, data not shown). Nor could we find reproducible 2° motifs near eCLIP crosslinks for SMARCA4 or SMARCE1. However, we could detect consistent structural features for the other 5 subunits. ARID1A, SMARCC1, and SMARCD3 bound RNA at apical loops modeled in 89% (n=27), 85% (n=27), and 73% (n=30) of aligned sequences, respectively (Fig6B, D, F). SMARCC2's modeled secondary structure preference also aligned to stem loops in 70% (n=30) of regions, but these often contained extended helical base pairing (Fig6E). In contrast to these subunits, PBRM1 was enriched along stem helices containing interspersed mismatched nucleotides in 80% of aligned models (n=30, Fig6C). Furthermore, these homologous structural alignments could detect instances of multiple subunit binding events at secondary structures along the same RNA molecule, including

along the 7SL RNA signal recognition particle component<sup>61</sup> (SRP, Fig6G). These results suggested that wide-spread RNA binding via Smarce1 and Smarca4 were GC content dependent, while RNA secondary structures produced within gene loci were responsible for engaging other RBD-containing BAF subunits.

***Stable BRG1 association with the genome correlates with BAF subunit RNA binding at target loci.***

We wanted to understand where RNA interactions were taking place in the nucleus. Toward this, we first looked to gain perspective on the relative amount of insoluble DNA-bound BRG1 (SMARCA4) during cardiac lineage commitment versus free BRG1/BAF throughout the nucleoplasm. On day 4.75 of differentiation, we permeabilized live nuclei *in situ* to allow soluble BAF complexes to wash out. After depleting the unbound fraction and subsequently performing immunocytochemistry on remaining nuclear components, we observed distinct puncta of remaining insoluble SMARCA4. These putative regulatory centers were preferentially associated with genomic regions outside of dense heterochromatin. In addition, these associations were cell cycle dependent, whereby BRG1 did not maintain stable genomic contact with replicated and condensed DNA during metaphase (Fig5A). These results showed that only a subset of total expressed Brg1 protein was engaged in stable chromatin interactions within the nucleus during heart lineage commitment, and these contacts needed to be restored after each cell division.

We next asked if these stable DNA-binding events were accompanied by the pervasive

RNA interactions we detected by eCLIP-seq. At day 4.75 of differentiation, we performed BRG1 chromatin immunoprecipitation with sequencing (ChIP-seq). These experiments found 6430 stable BRG1 binding sites throughout the genome (Fig5B). Thousands of these ChIP peaks were greater than 5kb away from a target gene, and only approximately 15% of regulatory DNA regions interact with their nearest gene neighbor. Therefore, we implemented PSYCHIC enhancer-promoter contact predictions based on Hi-C chromatin capture data generated in numerous cell types<sup>36</sup>. These DNA:DNA contact annotations were applied to BRG1 ChIP sites to establish high confidence genomic targets of occupied regions. Between direct (within 5kb of gene body) and PSYCHIC-identified DNA targets, we could assign 6014 of these DNA binding sites (94%) to annotated genes. Furthermore, we detected eCLIP RNA binding peaks within target gene transcripts (+/- 2kb) of nearly 89% of these DNA binding events (Fig5C). For example, BRG1/BAF directly targeted the *Mesp1* locus, which coincided with SMARCA4 and SMARCE1 RNA binding events and significant gene expression dynamics during lineage conversion (maximum  $\Delta Z\text{-score}_{\text{MES-CP}} = -4.4$ ,  $q < 0.05$ ). In contrast, RNA interactions could not be detected at nearby *Mesp2*, which was not as dynamically expressed during this time window (maximum  $\Delta Z\text{-score}_{\text{MES-CP}} = -1.1$ , not significant; Fig5d). Intergenic BRG1/BAF DNA binding sites could also be targeted to distal genes containing RNA binding events. This included a stable DNA interaction approximately 50kb upstream of critical mesendodermal transcription factor *Sox17*<sup>62</sup> in a region with high-confidence DNA-DNA interactions with *Mrpl15* and *Lypla1*, as well as *Sox17*. SMARCC1 AND SMARCE1 RNA binding events were detected within *Sox17*,



while SMARCE1 stably interfaced *Lyp1a1* transcripts. However, eCLIP peaks were not detected at *Mrp15*. Of note, both *Sox17* and *Lyp1a1* were differentially spliced during cardiac commitment phase transitions, while *Mrp15* was not (Fig5E). Finally, we discovered that SMARCA4 physically targeted and stably interfaced RNA within at least 15 of its core subunit genes (Fig5F), indicating a great degree of feedback regulation on BAF protein expression and potential composition.

We could assign only 54% of all RNA binding events to DNA-bound BRG1/BAF (Fig.5C). This was most likely the result of three possible factors. First, many of the observed RNA binding events may have taken place within the soluble BRG1/BAF fraction. However, the proportion of each subunit's eCLIP peaks detected in target loci of DNA-bound BRG1 was similar for all constituents (data not shown). Therefore, we could not establish RNA binding to specific subunits that favored DNA-independent interactions or sequestration of the complex away from the genome. Second, BRG1 was previously shown to be particularly difficult to target in CHIP experiments<sup>63</sup>. Thus, many BRG1/BAF DNA interactions could have gone undetected. Third, PSYCHIC assignment of enhancer-promoter interactions did not include cardiac progenitor-specific chromatin conformations or previously unannotated genomic regions in its classifications. Therefore, future experiments must be performed to fully assign transcript binding proximal and distal to the genome. Nonetheless, these data indicated that the vast majority of stable BRG1/BAF interactions within the genome were accompanied by RNA binding within transcripts of target loci.

***The differentially expressed and spliced genes that define the stage transitions of cardiac lineage commitment are enriched for BRG1/BAF RNA binding.***

We found that the BRG1/BAF complex engaged thousands of transcripts throughout the genome at the transitional stages of cardiac lineage commitment. We thus wanted to better understand how this widespread phenomenon might coincide with the dynamic transcript isoform transitions that defined the step-wise progression of mesoderm into committed cardiac progenitors. Of the 39,177 expressed genomic elements ('genes') from mesoderm to cardiac progenitor stages (day 4 to day 5.3 of differentiation), approximately 30% contained high confidence eCLIP RNA binding events. However, we found that BAF subunits bound transcripts within 55% ( $p < 6.5 \times 10^{-48}$ ) of 'genes' that had differentially expressed and/or spliced transcript isoforms during one or more of the lineage transitions of this developmental window (Fig8A). Additionally, 10 of 10 differentially expressed hubs of overrepresented transcript subnetworks during lineage commitment (identified in Chapter 1) had BRG1/BAF RNA binding events within their respective 'gene'. Furthermore, we could also attribute stable BRG1/BAF DNA binding to at least 7 of these important hub-containing genes (Fig8B). These data implicated BAF RNA binding as an integral component of gene regulation during the onset of heart development.

The association of RNA interactions within differentially spliced gene targets of DNA-bound BRG1/BAF took place at many notable loci. For example, our transcriptome analyses from Chapter 1 discovered that 1 of 5 expressed RNA isoforms of the *Carmn* cardiac super enhancer-associated lincRNA<sup>64</sup> was significantly upregulated from the

cMES to CP transition phase. eCLIP experiments subsequently established SMARCE1 RNA binding events within the third exon of this novel *Carmn* isoform, as well as throughout the *Bvhrt / Carmn* bidirectionally-transcribed enhancer region. In addition, we found multiple BRG1/BAF ChIP binding events within this enhancer, which could physically interact with *Arhgef37* and the casein kinase negative Wnt regulator *Csnk1a1*<sup>65</sup>. These target genes also contained ARID1A and SMARCE1 RNA binding, respectively, for which 1 of 10 expressed *Csnk1a1* isoforms (ENSMUST00000170862) was dramatically upregulated ( $\log_2$  fold change = 6.7,  $q < 2.5 \times 10^{-4}$ ) between cMES to CP stages (Fig8C). Although these significant isoforms of *Carmn* and *Csnk1a1* shared common expression at this discrete period, they each were expressed within disparate transcriptome subnetworks (Chapter 1: module 6 vs 35, respectively). Therefore, BRG1/BAF chromatin occupancy and RNA binding coincided with the convergence of these transcripts' expression dynamics during this time window.

BRG1/BAF also occupied the maternally imprinted enhancer region nearby *Meg3*<sup>66</sup> that could physically target both the lincRNA *Rian*, as well as the noncanonical NOTCH ligand *Dlk1*. *Rian* and *Dlk1* each contained SMARCE1 RNA binding events with their respective locus transcripts. Moreover, two DLK1 isoforms, ENSMUST00000109843 and ENSMUST00000124293, were differentially expressed within distinct transcript subnetworks (Chapter 1: modules 131 and 16, respectively) during the MES through CP gene isoform phase transitions. ENSMUST00000109843 was upregulated at the pcMES to cMES transition, while both a novel *Rian* isoform and ENSMUST00000124293 were significantly enriched after conversion to the CP stage.

Thus, BRG1/BAF enhancer occupation, along with SMARCE1 RNA interactions corresponded to divergent splice form expression of the important Notch pathway regulator DLK1.

Our transcriptome reconstruction from Chapter 1 discovered differential splicing at 80 co-processed ‘genes’, whereby multiple neighboring protein coding genes were transcribed and processed as single genomic elements. 78% (63/80) of these loci contained transcript binding events to BAF subunits (Fig8E). For example, *Tmem183a* and *Ppfia4*, a liprin family member implicated in neuromuscular junction formation<sup>67</sup>, were co-transcribed and spliced during cardiac differentiation. However, at both MES to pcMES and cMES to CP conversions, respectively, this ‘gene’ was differentially spliced to enrich the presence of a novel *Ppfia4* RNA isoform. At the same time, BRG1/BAF stably engaged the genome directly at this splice border. Accompanying this, SMARCE1, SMARCC1, and SMARCA4 bound RNA transcripts of this co-processed ‘gene’, both at the TSS and within *Ppfia4* (Fig8F). In Chapter 1, we described differential splicing of co-processed *Myh6/Myh7* during cardiac lineage development, and Han et al had previously described BRG1’s involvement in the stoichiometry of these gene products upon cardiomyocyte maturation<sup>16</sup>. However, at this early lineage commitment window, we did not observe RNA binding attributed to SMARCA4 or any other BAF subunit at the *Myh6/Myh7* locus. Nonetheless, our experiments suggested that additional layers of complexity were involved in BRG1/BAF modulation of MYH6:MYH7 ratios.

Of note, all subunits were proportionally associated with loci that underwent differential

isoform splicing over the course of mesoderm specification into cardiac progenitors. In addition, we could not ascribe particular subunit RNA interactions to either transcript up- or down-regulation (data not shown). Therefore, we concluded that, while BAF-RNA binding events were integral to the regulation of transcript expression dynamics, additional mechanistic factors contributed to determining the functional consequence of gene behavior at these loci.

***RNA interactions tether SMARCE1 to transcripts associated with multiple developmental programs, while co-binding with SMARCA4 targets the complex to genes specific to progression into the cardiac lineage.***

Exact eCLIP binding agreement between separate BAF subunit-targeted immunoprecipitations predicted physical protein-protein interactions between SMARCA4-SMARCC1, SMARCD3-SMARCE1, and SMARCC2-PBRM1, respectively. However, we next analyzed which BRG1-associated subunits were binding RNA concomitantly at the same locus. Despite the predicted RBD colocalization of SMARCD3 to SMARCE1 and SMARCC2 to PBRM1, respectively, we did not observe a strong correlation of RNA co-binding between these subunits. However, we did find a high degree of correlation of 500nt windows that contained both SMARCA4 and SMARCE1 eCLIP peaks (Fig9a). These frequently took place at gene promoters, whereby each respective subunit could bind sense or antisense orientations to the resident gene (Fig9B). For example, both SMARCA4 and SMARCE1 co-bound transcripts on opposite strands at the promoter region of the *Prkab2*, a subunit of the AMPK complex important for muscle energy homeostasis<sup>68</sup>. Furthermore, this gene

contained an important network module hub noncoding transcript. Of note, RNA binding at this locus coincided with significant reduction of this noncoding RNA during the first phase of mesoderm conversion into the cardiac lineage (Fig9B).

We previously did not observe overt subunit-specific RNA-binding correlation to significant gene isoform changes during cardiac lineage commitment. Therefore, we next asked if particular BAF subunits might preferentially interact with transcripts of genes associated with certain biological functions. Therefore, we performed gene ontology (GO) enrichment analysis on each cohort of loci containing eCLIP peaks. In doing so, we found that SMARCE1's vast array of RNA binding was enriched at genes responsible for multiple developmental pathways, especially anterior/posterior embryo patterning (Fig9C). However, SMARCA4 RNA interactions were enriched for ontologies specific to heart lineage development, such as Wnt and Bmp4 signaling, development of neuromuscular processes, and establishment of planar cell polarity (Fig9D). Therefore, we concluded that subunit-specific RNA binding did coincide with gene function-specific interactions.

***At the onset of cardiac lineage commitment, BRG1/BAF colocalizes in the nucleus with multiple RNA-binding proteins, including those critical to transcript splice regulation, histone modification, and Wnt signaling.***

The immunoprecipitations we performed to detect BAF subunit RNA adducts in day 4.75 isolated nuclei were carried out under physiologically relevant (native) salt conditions. Therefore, we asked whether additional co-localized RNA-binding proteins

might also have been isolated along with the complex. We found at least 16 RBD-containing proteins that reproducibly IP'd out of nuclear isolates along with the endogenous SMARCA4-3xFLAG bait (Fig10A). Notably, ESRP2, a critical splice factor involved in Wnt signaling<sup>69</sup>, FGFR2 regulation, and epithelial to mesenchymal transition<sup>70</sup>, bound RNA via its 3<sup>rd</sup> RNA recognition motif (RRM) and co-localized with BRG1/BAF. Additionally, we found an RNA binding domain within the 3<sup>rd</sup> malignant brain tumor domain of SFMBT1, a histone recognition protein known to interact with LSD1 and polycomb transcriptional repressor complexes<sup>71</sup>. MED12, a subunit of Mediator complexes implicated in Wnt and Shh signaling, was previously shown to bind activating ncRNAs to facilitate Mediator gene targeting, kinase activity, and chromatin conformation<sup>72</sup>. We also discovered BRG1 co-localized with RNA-bound MED12 in cardiac mesoderm. However, our analysis determined its RBD to lie at the N-terminus of the protein, instead of the central core as suggested by Lai, et al<sup>73</sup>. RTF1, a known RNA-binding member of the Paf1 transcriptional activation and elongation complex<sup>74</sup> was previously shown to be required for normal Wnt and Shh signaling and heart development<sup>75</sup>. Recently, Fischl et al also showed the Paf1 complex as a key regulator of differential transcript fate and nuclear export<sup>76</sup>. We detected RTF1 reliably associated with BRG1/BAF in the nucleus during cardiac lineage commitment, where its RBD was modeled in an unfolded linker between structural coiled coil and PLUS3 DNA-binding domains<sup>77</sup> (Fig10B).

We were also surprised to find co-localized and previously unannotated RNA binding in numerous proteins canonically associated with the nuclear envelope, endoplasmic (ER)

/ sarcoplasmic reticulum (SR), and Golgi apparatus. These included RYR2, required for SR calcium release, cardiac contractility and early heart tube formation<sup>78</sup>, which contained an RBD near its calmodulin-binding region. Interestingly, calcium/calmodulin was also shown to be required for BAF complex chromatin remodeling<sup>79</sup>. Additionally, we were able to pull down ARHGAP21, a nuclear and golgi-associated rho GTPase activator important for regulation of actin dynamics<sup>80</sup> that bound RNA proximal to its alpha-catenin (CTNNA1) interface. Another Golgi protein GOLPH3, a known actin cytoskeleton-binding protein important for cell migration<sup>81</sup>, contained a previously unannotated RBD and co-localized with BRG1/BAF. We could also co-IP RNA-bound NCLN and SLC33A1, ER-associated Nodal<sup>82</sup> and Bmp4<sup>83</sup> signaling antagonists, respectively, along with the nuclear envelope/ ER-associated protein RNF180. RNF180 contained an RBD near its RING zinc finger ubiquitin ligase domain and ZIC2 binding domain (Fig10C). Of note, ZIC2 was recently shown critical for Nodal regulation, as well as lateral patterning of the heart<sup>84</sup>. Therefore, we could associate ribonuclear components to the association of these ER/Golgi proteins with BRG1/BAF in the nucleus. These unexpected findings were particularly interesting given eCLIP detection of SMARCA4, SMARCC1, and SMARCC2 docking along 7SL RNA of the SRP, which begins assembly in the nucleus and targets its receptor within the ER<sup>85</sup>.

Six additional BRG1/BAF associated RNA binding proteins were isolated from cardiac mesoderm nuclei. These included CSNK1G3, a casein kinase involved in Wnt signal transduction<sup>86</sup>; NEUROD4, a transcription factor important for Notch signaling<sup>87</sup>; and Tab3, a cardioprotective TGF $\beta$ -activated protein kinase<sup>88</sup>. In addition, discrete RBDs



and co-IP with BRG1/BAF were also found for ADCY9, an adenylyl cyclase important for cardiovascular function but not canonically associated with the nucleus<sup>89</sup>; the glycogen debranching enzyme AGL, important for cardiomyocyte energy metabolism<sup>90</sup>; and ADAMTS19, a protein with metalloprotease and thrombospondin functions eventually expressed in the atrioventricular canal and outflow tract<sup>91</sup> (Fig10D). These results showed that the BAF complex, itself engaged in thousands of RNA contacts, co-localized with proteins of diverse enzymatic and regulatory functionalities while they too engaged RNA transcripts within a shared local environment.

## Discussion

These experiments were designed to test the hypothesis that RNA binding was an intrinsic component of BRG1/BAF complex function at the onset of cardiac lineage commitment. In doing so, we aimed to identify protein-RNA interactions that were actually taking place *in vivo* within the nuclei of developmentally relevant cells. We first discovered that at least 7 BRG1-bound BAF subunits were docked with RNA in the nucleus at this developmental stage via discrete protein RBDs. Furthermore, these RBDs were oriented within each protein to be capable of impacting the DNA, histone, and protein binding activities of their functional domains. We next discovered that these subunits respectively engaged RNA via hundreds- to tens of thousands- of binding events, predominantly within protein coding genes and enhancers, which were correlated to nearly all stable BRG1 interactions with the genome. These wide spread RNA transcript contacts connected BRG1/BAF (mostly via SMARCE1 recognition of GC-rich RNA domains) to the transcriptional state of thousands of genes, often over large linear distances. Also, BAF RNA docking was significantly correlated to the differential gene transcript splicing and expression that defined the transitions of mesoderm-to-heart specification. At protein coding loci, BAF subunits equally engaged antisense, as well as sense, RNA molecules. Therefore, we found little distinction between underlying transcript dynamics regulating coding vs noncoding genomic elements. Furthermore, our experiments identified the BAF complex in contact with numerous additional proteins, which also engaged RNA transcripts, thereby suggesting an integration of RNA binding to splice regulation, gene coactivation, and signal

transduction functions at BRG1/BAF nuclear targets during early heart development.

These experiments provide the first insight into the pervasive nature of RNA binding taking place within the nucleus between BRG1/BAF and the widespread transcription that underlies the transition of nascent mesoderm into the cardiac lineage. We hypothesize that RNA scaffolds provide a dynamic interface to gene regulatory machinery that drives gene expression and cell identity. These experiments introduce numerous correlative discoveries that open the door for mechanistic interrogation. Future work must aim to specifically interrupt these RNA interactions in order to establish their requirement for proper BAF localization and functionality, gene expression, and heart organogenesis.

## Description of Figures

**Figure 2.1.** Identification of BAF subunit RNA binding domains. A.) Schematic for unbiased labeling and detection of BAF complex-bound RNA in transiently transfected HEK293FT nuclei. B.) Left, Infrared fluorescence of labeled RNA in gel after SDS-PAGE; SYPRO Ruby protein stain in same gel to detect SMARCE1 bait; Right, measured fluorescence intensity across UV<sup>-</sup> and UV<sup>+</sup> samples and increasing salt concentration. C.) Schematic for purification of BAF complex RNA-bound peptides and detection by mass-spectrometry. D.) UCSC Genome Browser tracks of identified BAF subunit RNA binding domains detected by RNPxl and mapped with iPiG<sup>93</sup>.

**Figure 2.2.** 3° BAF subunit structure modeling near RNA binding domains. A.) Summary of subunits with high confidence RNA binding domains; RBD, RNA binding domain (with amino acid positions). B.) 3° modeled structure of ARID1A N- and C-terminal LxxLL domains and core ARID domain. C.) 3° modeled structure of amino acids 411-990 of PBRM1 containing bromodomains 3-6. D.) 3° modeled structure of SMARCA4 amino acids 1071-1570 containing C-terminal helicase domain and bromodomain. E.) 3° modeled structure of SMARCC1 amino acids 201-950 containing SANT and SWIRM domains. F.) 3° modeled structure of SMARCC2 amino acids 201-950 containing SANT and SWIRM domains. G.) 3° modeled structure of full length SMARCD3 amino acids 1-483 SWIB domain. H.) 3° modeled structure of full length SMARCE1 amino acids 1-411 containing COILED COIL and HMG-BOX domains. RNA, site of RNA crosslinking; amino acids without clear structure or spatial relevance to RBD

omitted.

**Figure 2.3.** Subunit-specific RNA interactome detection using native RIP and eCLIP. A.) Left: Summary of total and novel significantly enriched RNA isoforms after native immunoprecipitation without crosslinking of BRG1-3xFLAG tagged bait; Middle: Cumulative annotation of 2505 significantly enriched RNA isoforms, Right: UCSC Genome Browser tracks of representative native RIP-enriched antisense transcript from the *Bcl2l11* locus. B.) Top: Schematic for sequential immunoprecipitations; UV-induced RNA crosslinks depicted by purple bolts and red curves, respectively; Bottom Right: In-gel Sypro Ruby stain of immunoprecipitation products after 1 day and 3 days on beads, respectively; kD, kilodaltons; BAF, whole BAF complex after IP#1; arrow, remaining SMARCC1 after 3 days on beads; Bottom Left: Bioanalyzer gel-like image of resulting amplified libraries for each subunit IP; dashed line, combined molecular weight of library primers. C.) UCSC Genome Browser eCLIP tracks at *Bcl2l11* locus for SMARCC1- and SMARCE1-specific secondary immunoprecipitations; green highlight, RIP-seq enriched region; asterisks, significantly enriched eCLIP peaks. IP, immunoprecipitation.

**Figure 2.4.** Characteristics of BRG1/BAF RNA binding. A.) Total discrete eCLIP binding events detected for each BAF subunit. B.) Number of eCLIP binding events per subunit per protein coding gene containing at least 1 subunit binding event, respectively. C.) Fraction of subunit eCLIP binding events that were antisense to resident protein coding gene. D.) Genome-wide distribution of eCLIP binding events for each subunit.

**Figure 2.5.** BAF protein-protein binding interfaces predicted by eCLIP. A.) Binary

frequency dendrogram of subunit eCLIP peaks with exact agreement to 7nt resolution; Extra-BAF associated RNA-binding transcription factor NEUROD4 included for reference. B.) RaptorX contact heatmap of predicted inter-protein amino acid contacts between SMARCC1 and SMARCA4. C.) RaptorX contact heatmap of predicted inter-protein amino acid contacts between SMARCC2 and PBRM1. D.) RaptorX contact heatmap of predicted inter-protein amino acid contacts between SMARCD3 and SMARCE1. Star, RNA binding domain; increasing probability of amino acid contact shaded from white to black within respective heatmap; N, N-terminus of protein; C, C-terminus of protein.

**Figure 2.6.** RNA motif recognition of BAF subunits. A.) Primary sequence enrichment for 12-14nt windows surrounding SMARCA4 and SMARCE1 eCLIP binding sites, respectively; percentage of sites containing enrichment and p-value explicitly stated. B.) Representative 2° structure motifs for ARID1A eCLIP binding sites. C.) Representative 2° structure motifs for PBRM1 eCLIP binding sites. D.) Representative 2° structure motifs for SMARCC1 eCLIP binding sites. E.) Representative 2° structure motifs for SMARCC2 eCLIP binding sites. F.) Representative 2° structure motifs for SMARCD3 eCLIP binding sites. G.) Representative examples of co-bound RNA species with corresponding 2° structures at subunit interaction sites. Black curves, 10nt binding interface centered around eCLIP binding sites.

**Figure 2.7.** Targeting of Brg1/BAF DNA binding sites to genes containing RNA-bound subunits. A.) Top: Schematic of live *in situ* nuclei permeabilization and washout of soluble components; Bottom: immunocytochemistry against BRG-3xFLAG and DAPI

nuclear stain, respectively with pixel fluorescence intensity quantified at right; NoPerm, cells fixed before immunostaining; dashed lines, nuclear perimeter; scale bar, 10 $\mu$ m; A.U., arbitrary units. B.) Heatmap of Brg1-ChIP read density centered around 6430 binding sites determined by analysis with MACS2. Scale represents approximately 4-fold dynamic range. C.) Left: Ratio of BRG1 ChIP binding sites with assigned gene targets that contained eCLIP BAF subunit-RNA binding events; Right: Ratio of all eCLIP binding events that could be directly assigned to BRG1 ChIP binding sites. D.) Representative UCSC Genome Browser tracks of BRG1 ChIP directly targeted to *Mesp1* but not *Mesp2* and subunit-specific eCLIP binding events of *Mesp1* transcripts. E.) Representative UCSC Genome Browser tracks of intergenic BRG1 ChIP binding site assigned to multiple distal loci with PSYCHIC enhancer predictions and subunit-specific eCLIP binding events of *Sox17* and *Lyp1a1* transcripts. F.) BAF subunit genes targeted by DNA-bound BRG1 and subunit-specific eCLIP sites. Asterisks, significantly enriched eCLIP peaks, red gene labels, significant differential splicing and/or expression during mesoderm to cardiac progenitor transitions.

**Figure 2.8.** BAF complex RNA binding at sites of differential RNA isoform splice transitions. A.) Summary of BAF subunit eCLIP binding events at all expressed ‘genes’ (MES, pcMES, cMES, and CP stages) vs at ‘genes’ with differential transcript isoforms at MES thru CP transitions. B.) Summary of BRG1 ChIP and eCLIP binding at genes with differentially expressed subnetwork hub transcript isoforms as described in Chapter 1. C.) Above: UCSC Genome Browser tracks depicting BAF eCLIP and ChIP binding within *Bvhrt* and *Carmn* enhancer associated lincRNAs and eCLIP peaks within

transcripts of target genes *Arhgef37* and *Csnk1a1*, respectively; asterisks, significantly enriched RNA binding sites; Below: Upregulation of *Rian* noncoding RNA transcript and *Csnk1a1* coding RNA transcript isoforms, respectively; box, isoform depicted in expression schematics; black asterisks, significant expression change between cMES and CP stages; white asterisks, location of SMARCE1 RNA binding. D.) Above: UCSC Genome Browser tracks depicting ChIP binding near *Meg3* locus and eCLIP peaks within transcripts of target *Rian* lincRNA and *Dlk1* locus, respectively; asterisks, significantly enriched eCLIP peaks. Below: Differential expression of *Dlk1* protein-coding isoforms and novel *Rian* transcript over pcMES-cMES-CP transitions; asterisks, significant differential expression and/or splicing. E.) Ratio of differentially spliced (MES-pcMES-cMES-CP) multi-processed 'genes' with greater than one Ensembl-annotated gene containing BAF subunit eCLIP peaks. F.) Above: Representative UCSC Genome Browser tracks depicting BRG1 ChIP and SMARCA4, SMARCC1, and SMARCE1 eCLIP peaks, respectively, within the *Ppfia4/Tmem183a* co-processed locus. Below: Upregulation of novel *Ppfia* RNA transcript isoform; box, isoform depicted in expression schematic; asterisk, significant expression change between MES-pcMES and cMES-CP stages, respectively.

**Figure 2.9.** Smarce1 and Smarca4 RNA co-binding. A.) Binary frequency dendrogram of co-bound subunit eCLIP peaks within 500nt window of each other; Extra-BAF associated RNA-binding transcription factor NEUROD4 included for reference. B.) Above: Representative UCSC Genome Browser tracks depicting Smarca4 and Smarce1 binding within *Prkab2* locus, respectively; asterisks, significantly enriched



eCLIP peaks; Below: Downregulation of *Prkab2* noncoding RNA transcript isoform; box, isoform depicted in expression schematic; asterisk, significant expression change between MES and pcMES stages. C.) Enriched gene ontology (GO) biological process terms for loci containing Smarce1 eCLIP peaks. D.) Enriched GO biological process terms for loci containing Smarca4 eCLIP peaks; FDR, false discovery rate.

**Figure 2.10.** Non-BAF co-immunoprecipitated RNA binding proteins. A.) Summary of Smarca4 co-localized non-BAF proteins containing high confidence RNA binding domains; RBD, RNA binding domain (with amino acid positions). B.) RaptorX 3° structure models of nuclear splice regulators and histone modifying proteins; RRM, RNA recognition motif; MBT, malignant brain tumor domain. C.) RaptorX 3° structure models of proteins associated with golgi apparatus, nuclear envelope, and/or endoplasmic/sarcoplasmic reticulum; RYR, ryanodine receptor domain; CALM, calmodulin; CTNNA1, catenin $\alpha$ 1; RHO GAP, Rho GTPase activating protein domain; ZIC2, zinc finger protein 2; RING ZF, ring zinc finger domain; PTDINS4P; phosphatidylinositol-4-phosphate. D.) RaptorX 3° structure models of additional signaling, metabolic, and transcription factors; ATP, adenosine triphosphate; RANBP2 ZF, ranbp2-type zinc finger domain; TSP1, thrombospondin type-1 domain; bHLH, basic helix-loop-helix domain. Amino acid numbers explicitly labeled; RNA, site of RNA crosslinking; amino acids without clear structure or relevance to RBD omitted.

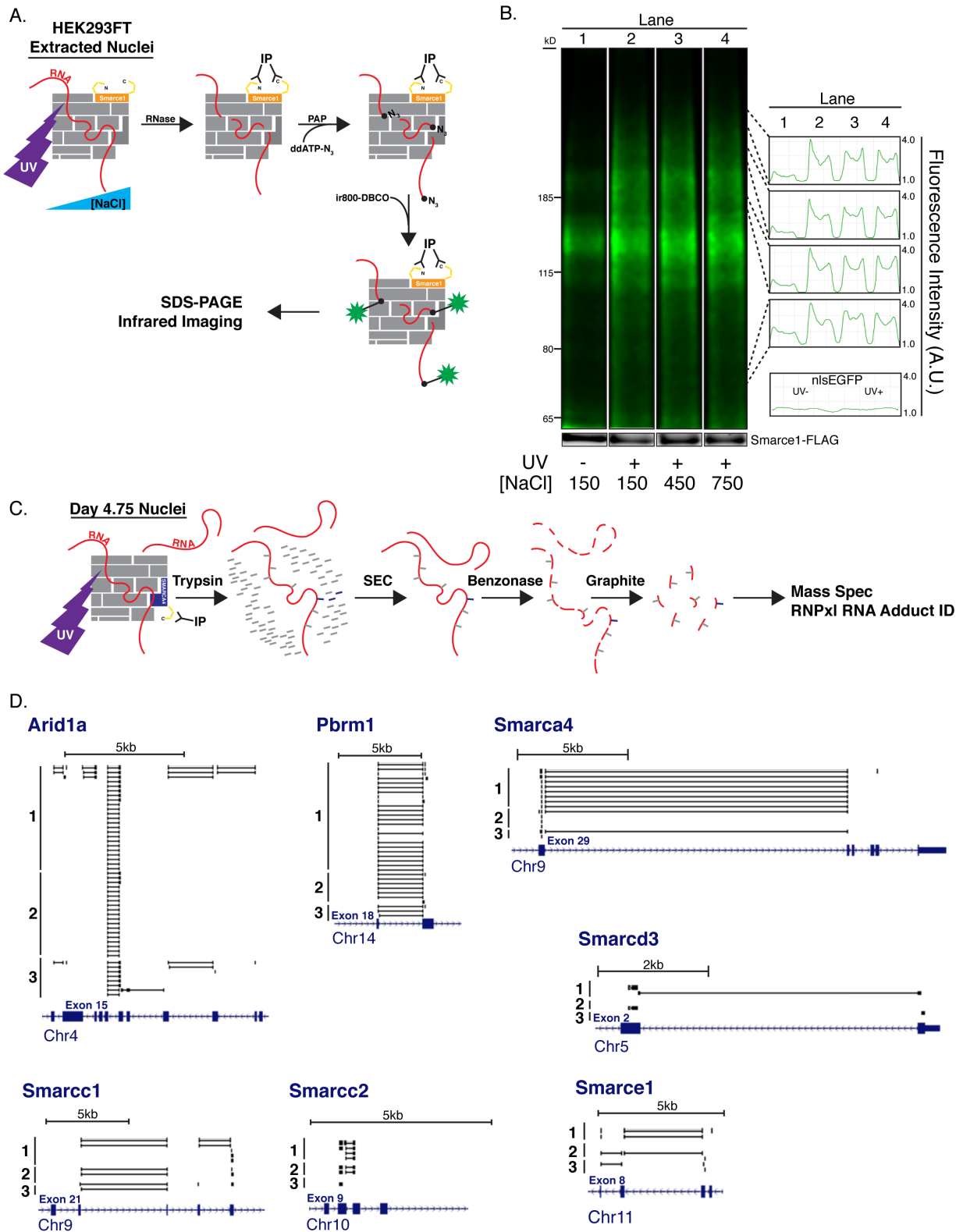


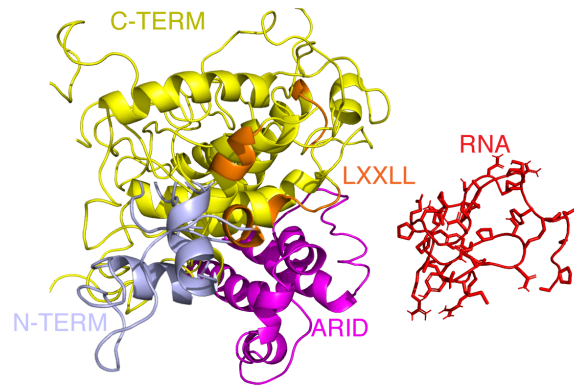
Figure 2.1. Identification of BAF subunit RNA binding domains

A.

| <i>Subunit</i> | <i>RBD</i> |
|----------------|------------|
| ARID1A         | 1202-1257  |
| PBRM1 (1&2)    | 621-769    |
| SMARCA4 (1)    | 1381-1399  |
| SMARCC1        | 772-832    |
| SMARCC2        | 241-264    |
| SMARCD3        | 10-69      |
| SMARCE1        | 148-188    |

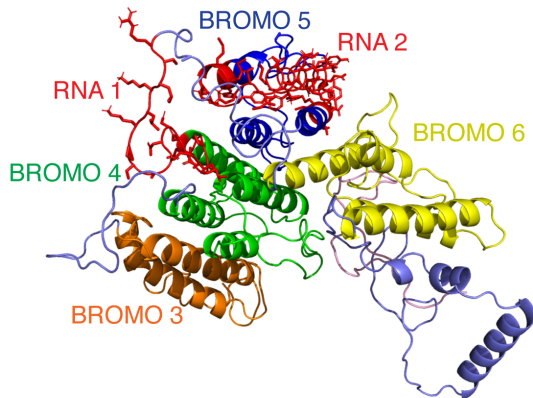
B.

**ARID1A (N-term, Arid, C-term)**



C.

**PBRM1 (411-990)**



D.

**SMARCA4 (1071-1570)**

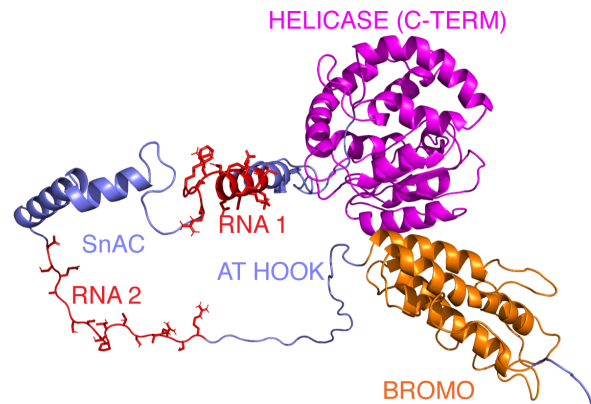


Figure 2.2. 3° BAF subunit structure modeling near RNA binding domains

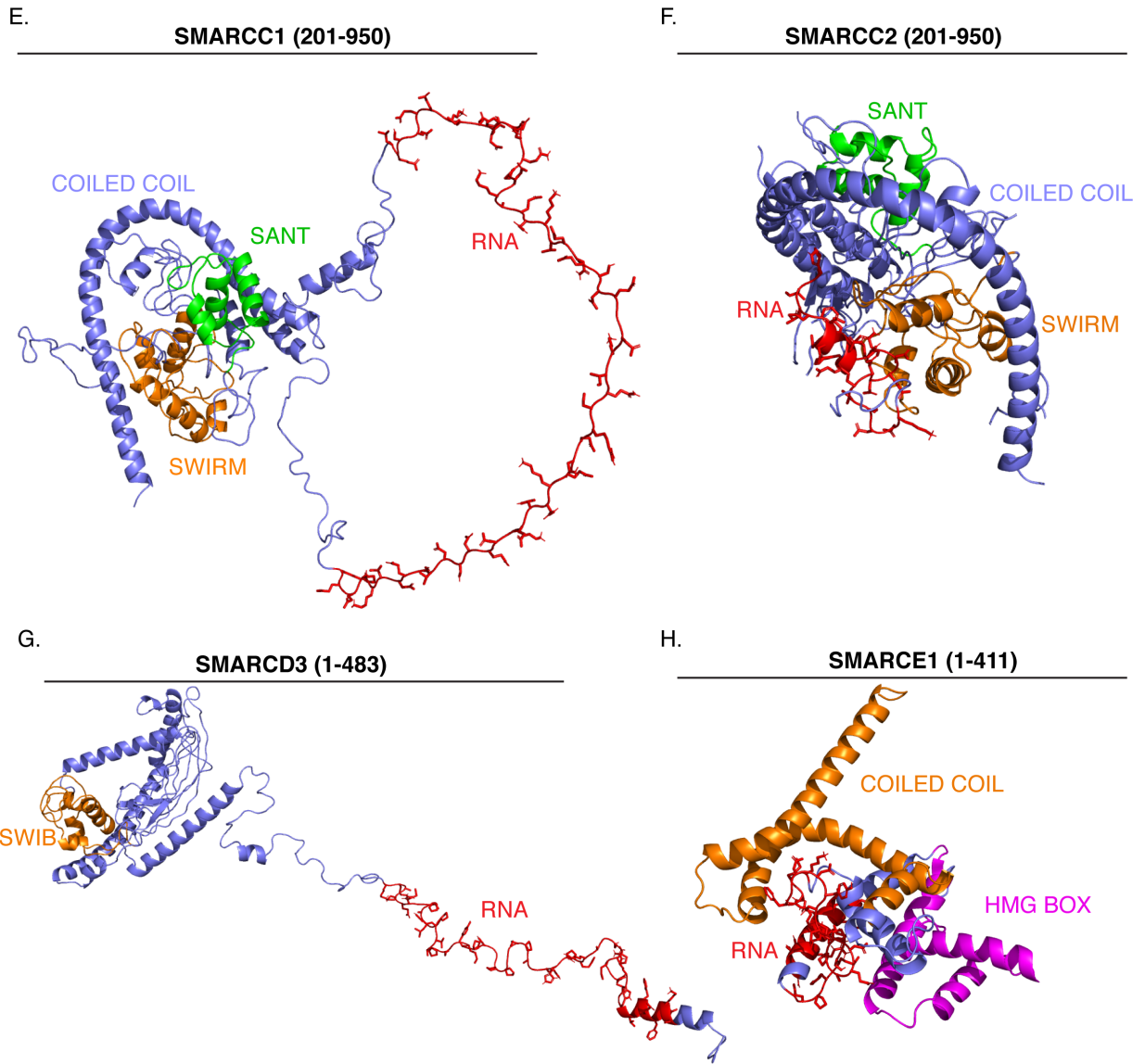


Figure 2.2. 3° BAF subunit structure modeling near RNA binding domains

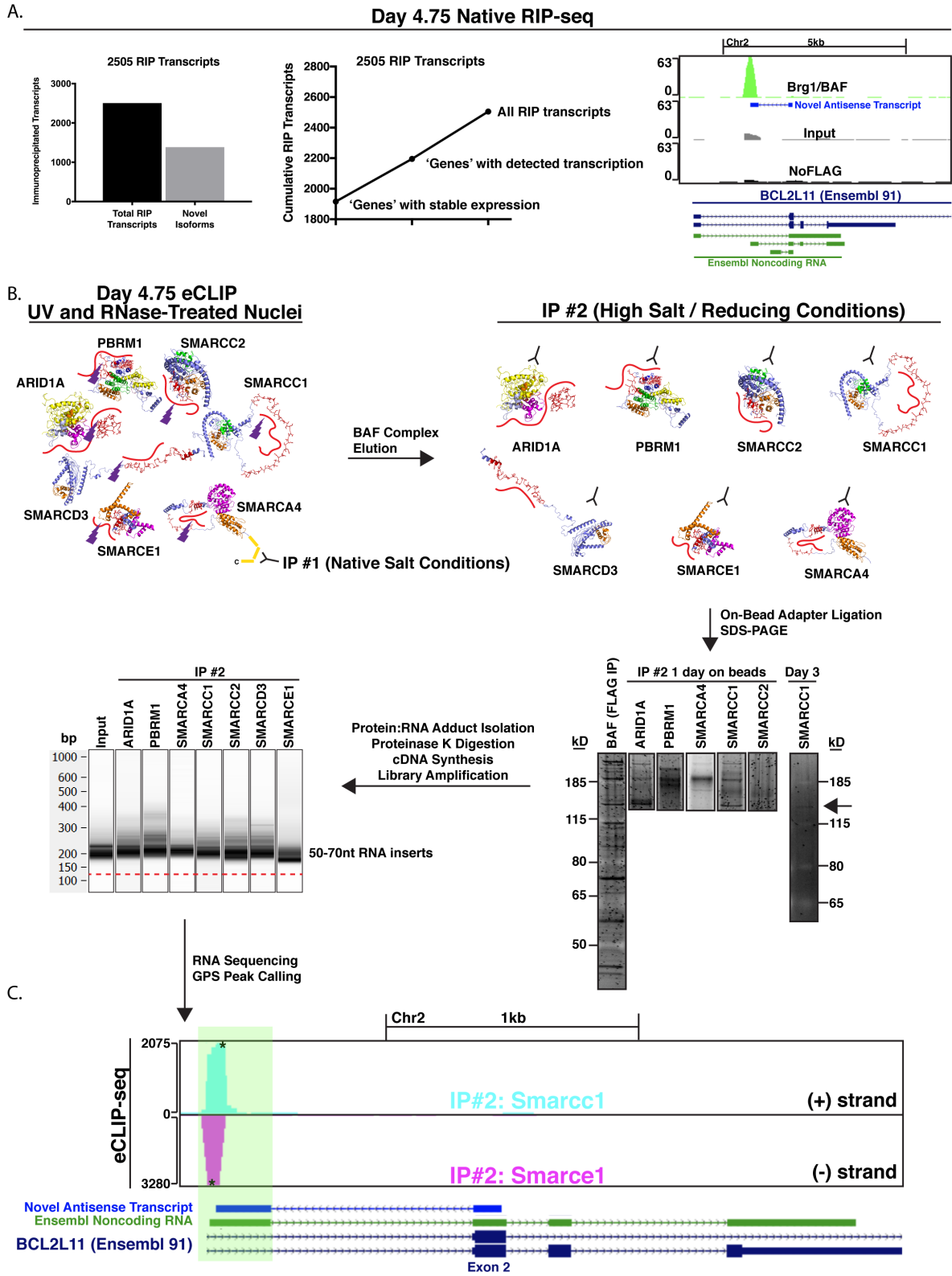


Figure 2.3. Subunit-specific RNA interactome detection using native RIP and eCLIP

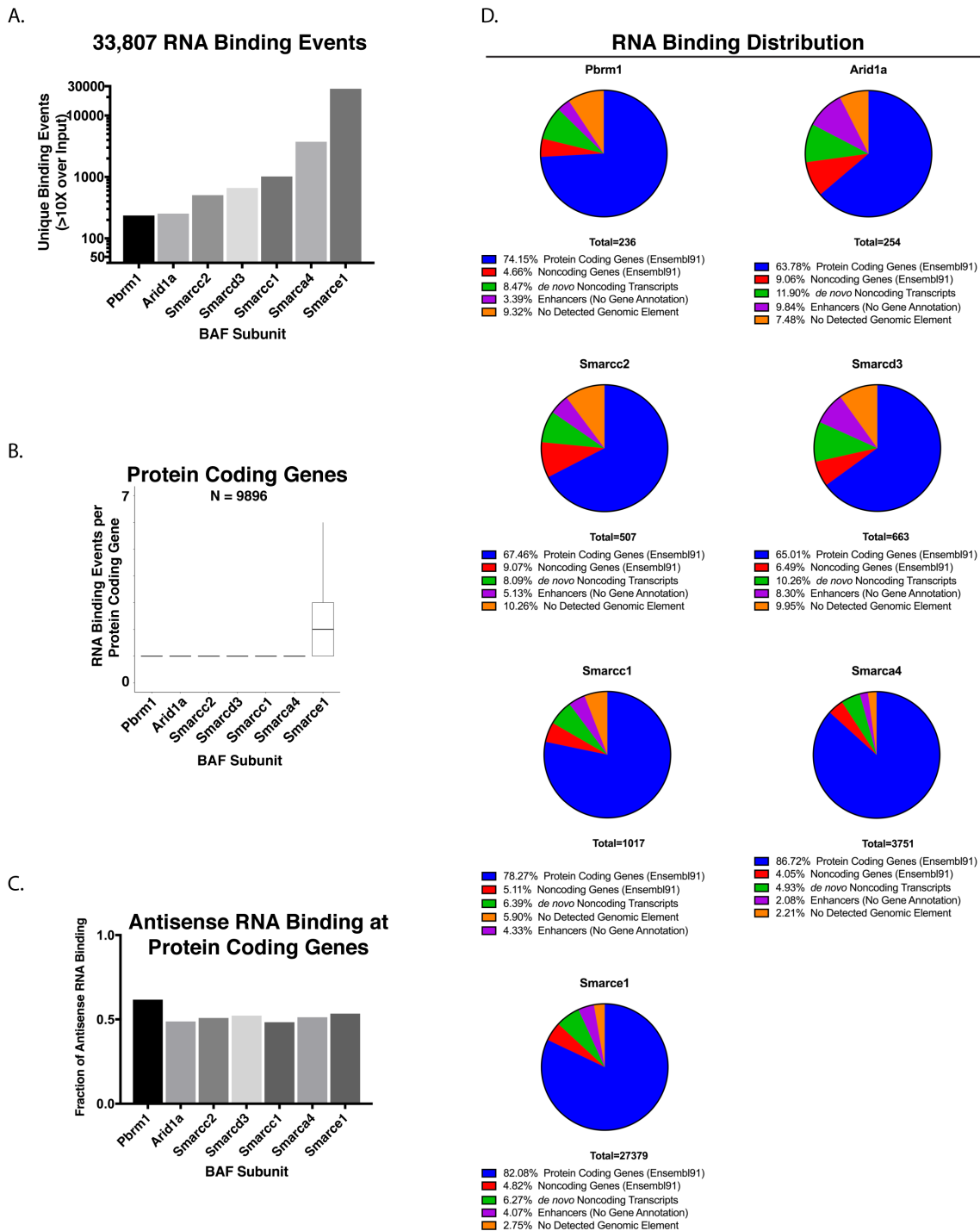


Figure 2.4. Characteristics of Brg1/BAF RNA binding

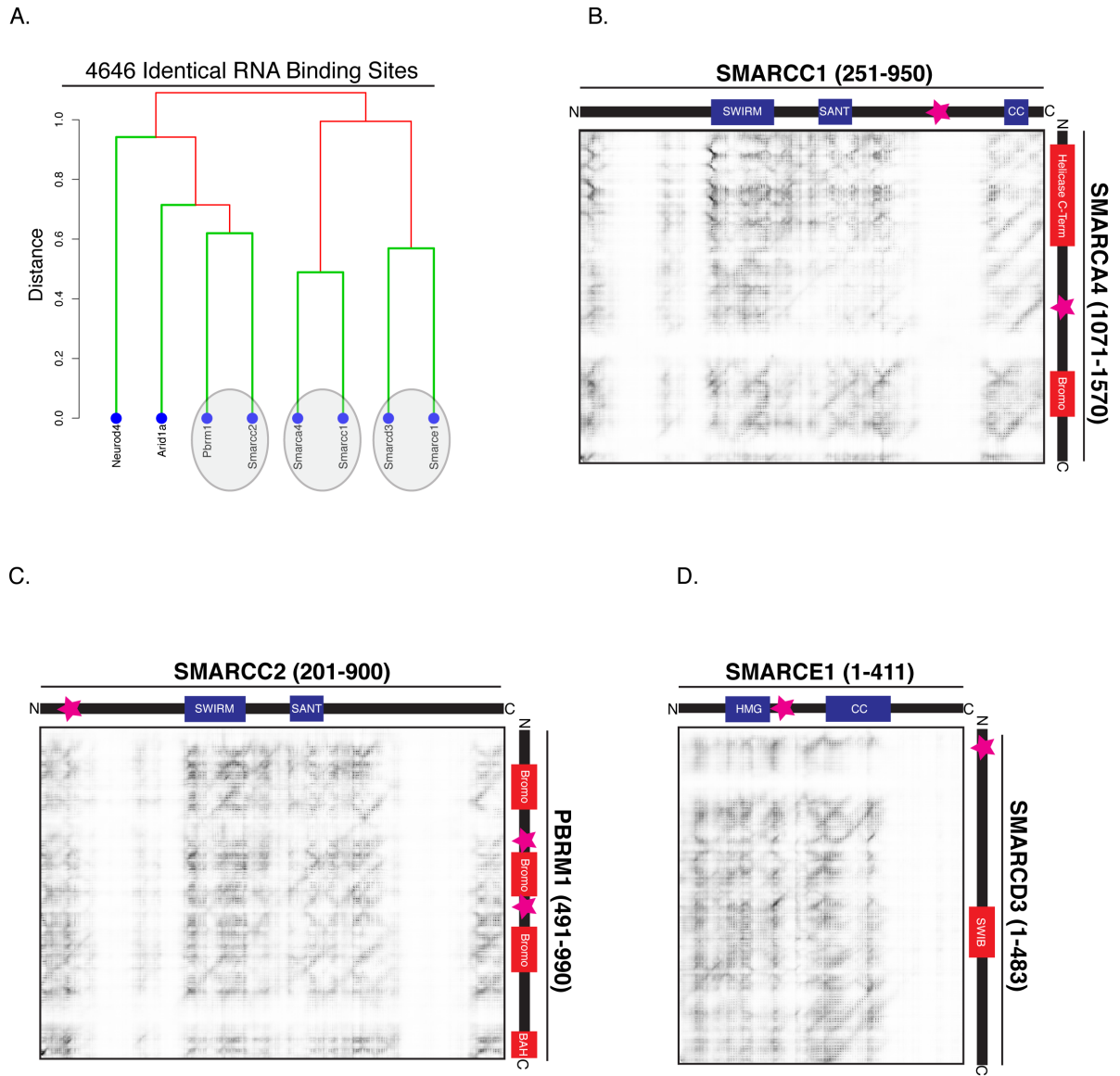


Figure 2.5. BAF protein-protein binding interfaces predicted by eCLIP

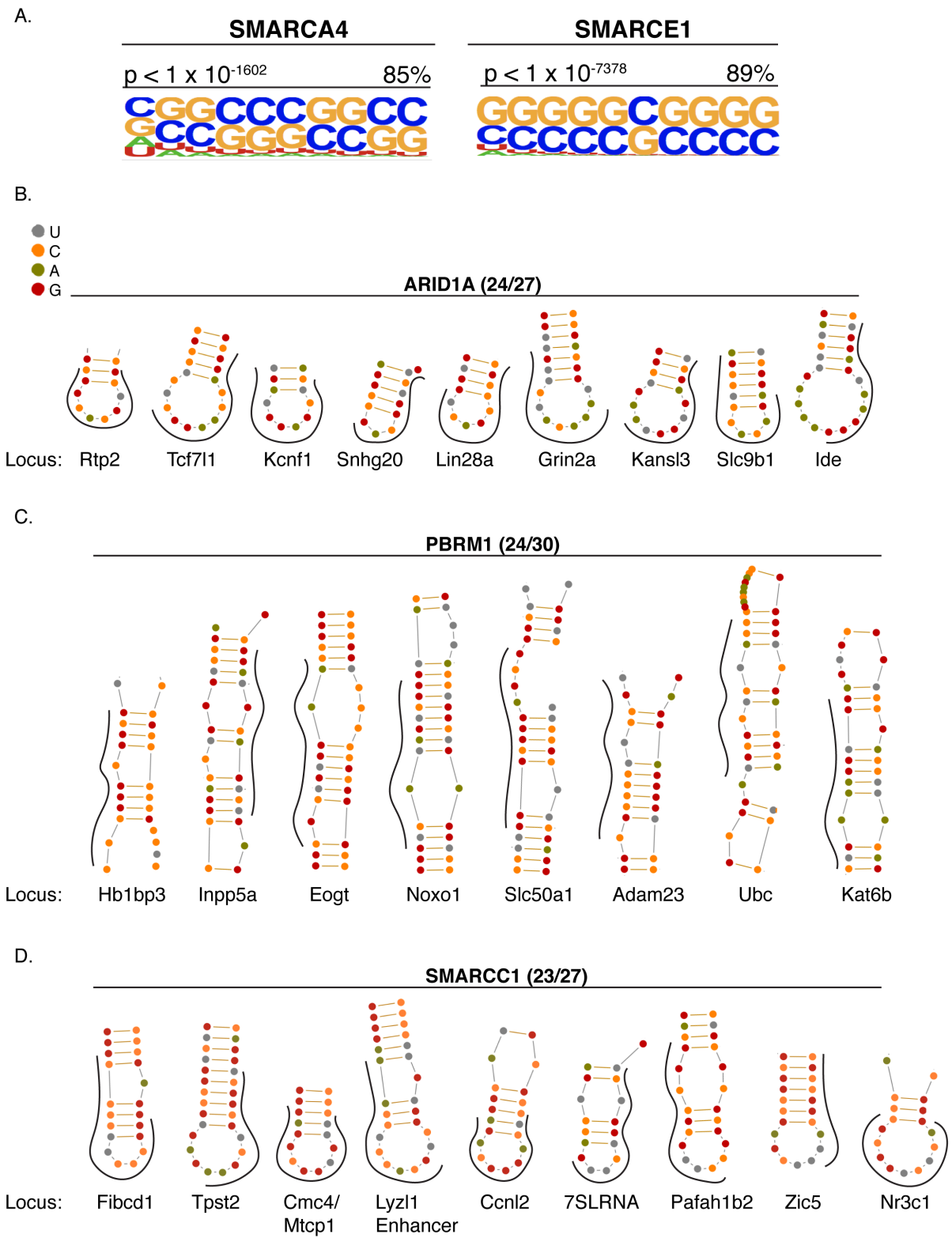


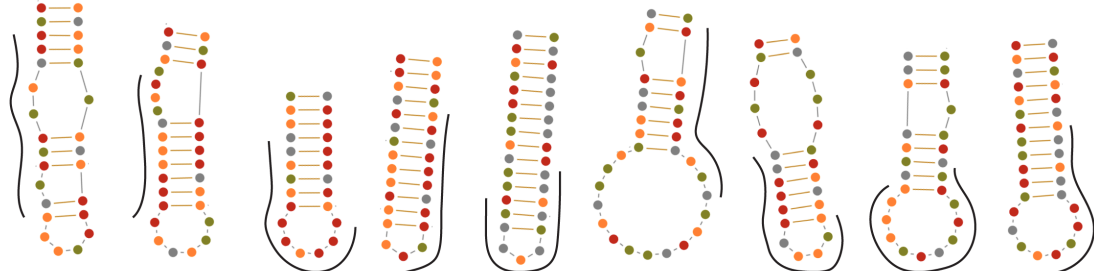
Figure 2.6. RNA motif recognition of BAF subunits



E.

● U  
● C  
● A  
● G

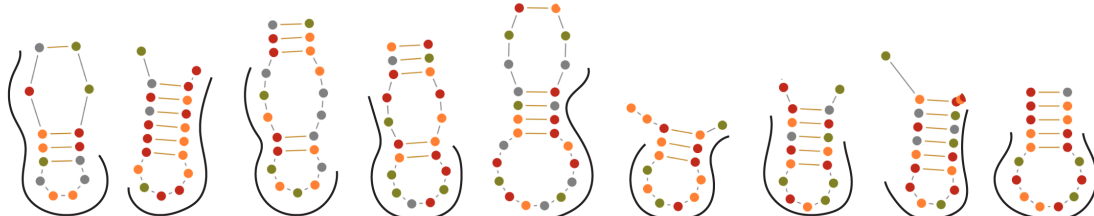
**SMARCC2 (21/30)**



Locus: Neat1 Gm16279 Ripor3 Gm14236 Gm22519 Zfp870 Eif3a Hnrnpc Golgb1

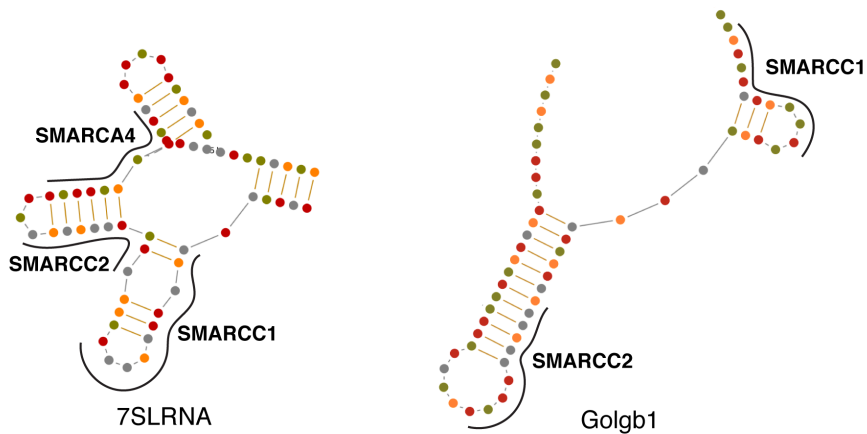
F.

**SMARCD3 (22/30)**



Locus: Elof1 Atp6v1b2 Adgr1 Nova2 P4ha1 Tjp2 Zxda Hp Ppp2r5e

G.



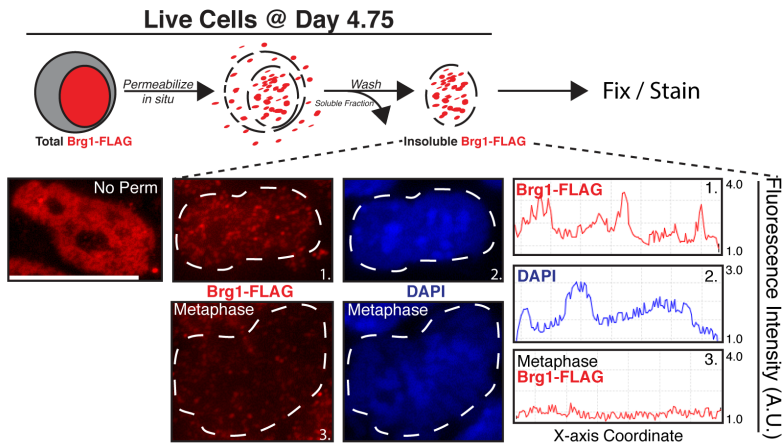
Locus:

7SLRNA

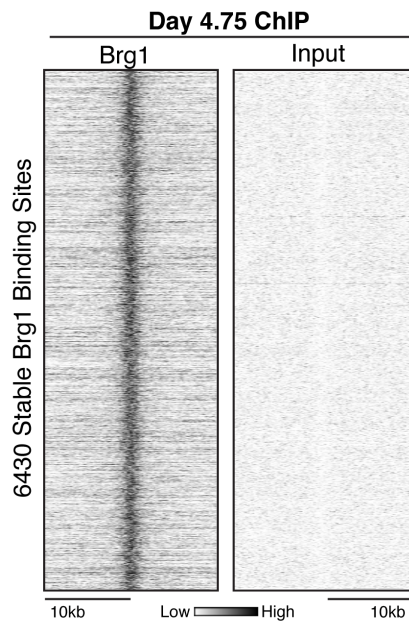
Golgb1

Figure 2.6. RNA motif recognition of BAF subunits

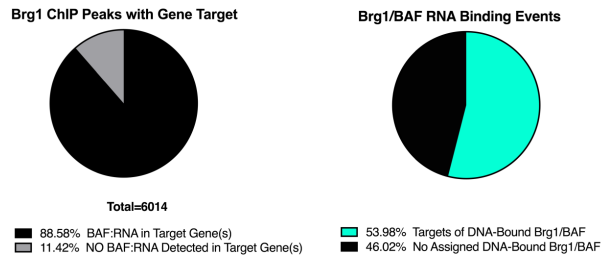
A.



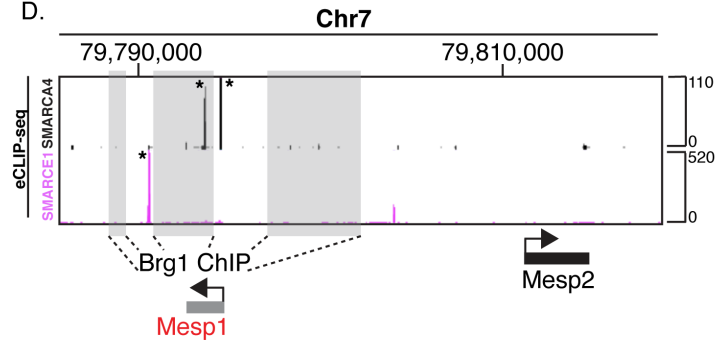
B.



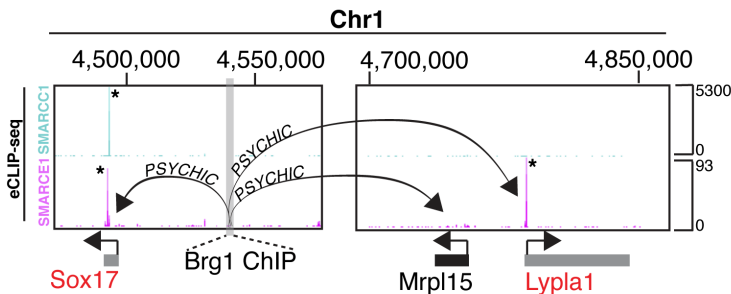
C.



D.



E.



F.

**Target BAF Genes**  
**Brg1 ChIP + BAF eCLIP**

|               |                |
|---------------|----------------|
| <i>Actl6a</i> | <i>Smarca2</i> |
| <i>Actl6b</i> | <i>Smarca4</i> |
| <i>Arid1a</i> | <i>Smarcc1</i> |
| <i>Arid1b</i> | <i>Smarcc2</i> |
| <i>Bcl7c</i>  | <i>Smarcd1</i> |
| <i>Dpf1</i>   | <i>Smarcd3</i> |
| <i>Dpf2</i>   | <i>Ss18l1</i>  |
| <i>Dpf3</i>   |                |

Figure 2.7. Targeting of BRG1/BAF DNA binding sites to genes containing RNA-bound subunits

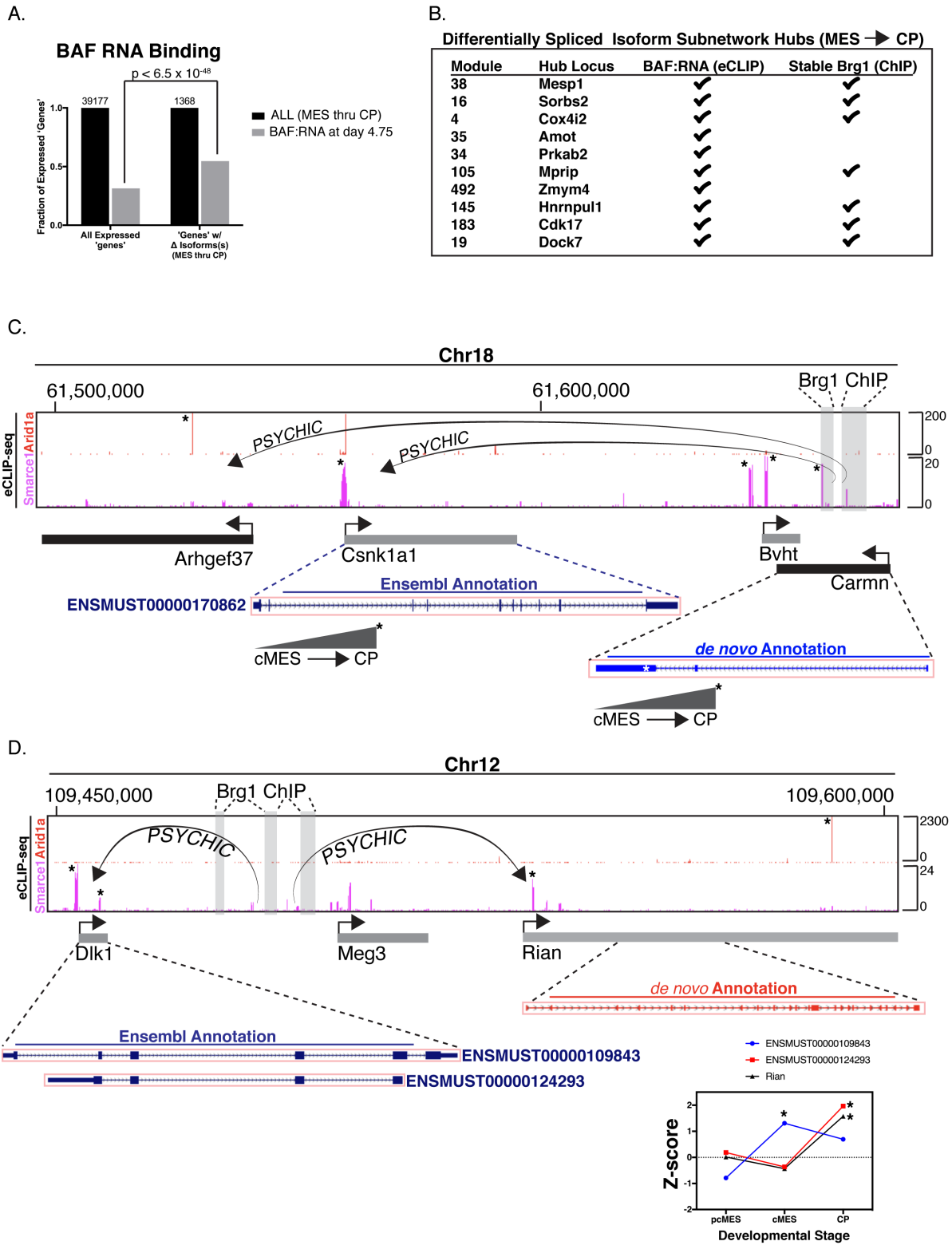
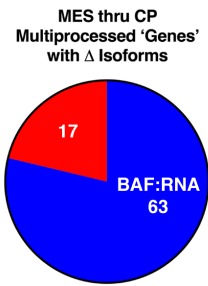


Figure 2.8. BAF complex RNA binding at genes of differential RNA isoform splice transitions

E.



F.

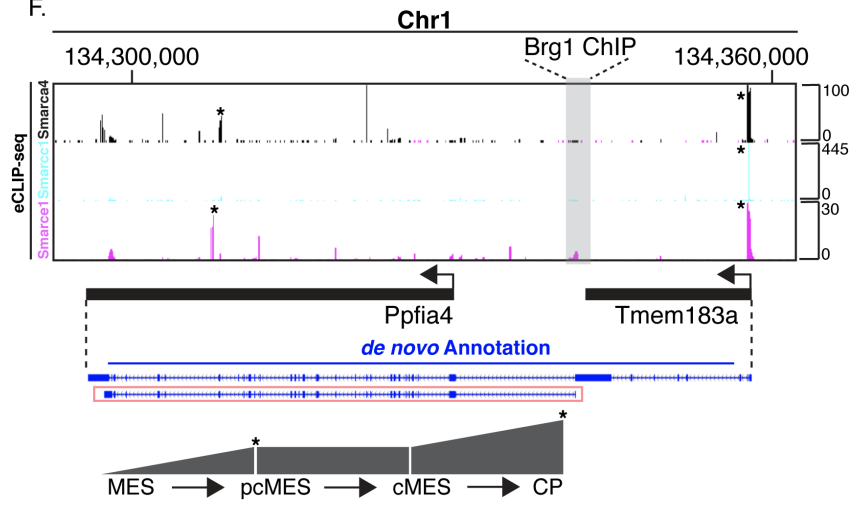


Figure 2.8. BAF complex RNA binding at genes of differential RNA isoform splice transitions

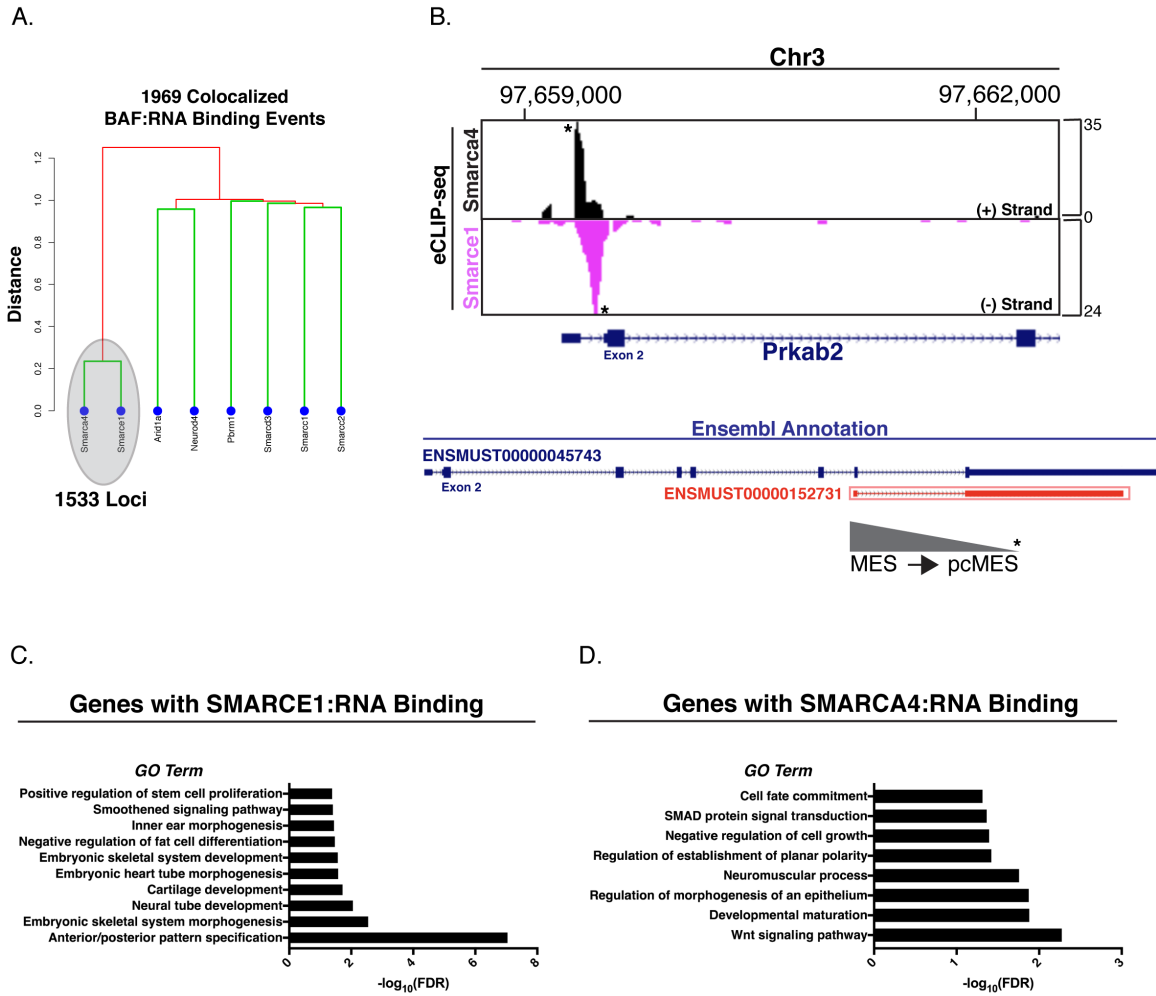


Figure 2.9. SMARCE1 and SMARCA4 RNA co-binding

A.

| <i>PROTEIN</i> | <i>RBD</i> | <i>PROTEIN</i> | <i>RBD</i> |
|----------------|------------|----------------|------------|
| ESRP2          | 532-550    | RNF180         | 539-557    |
| SFMBT1         | 303-318    | GOLPH3         | 28-52      |
| MED12          | 284-296    | AGL            | 1232-1268  |
| RTF1           | 308-328    | CSNK1G3        | 62-79      |
| RYR2           | 3564-3580  | ADCY9          | 576-595    |
| ARHGAP21       | 1653-1673  | TAB3           | 548-561    |
| NCLN           | 500-518    | ADAMTS19       | 974-993    |
| SLC33A1        | 16-39      | NEUROD4        | 10-23      |

B.

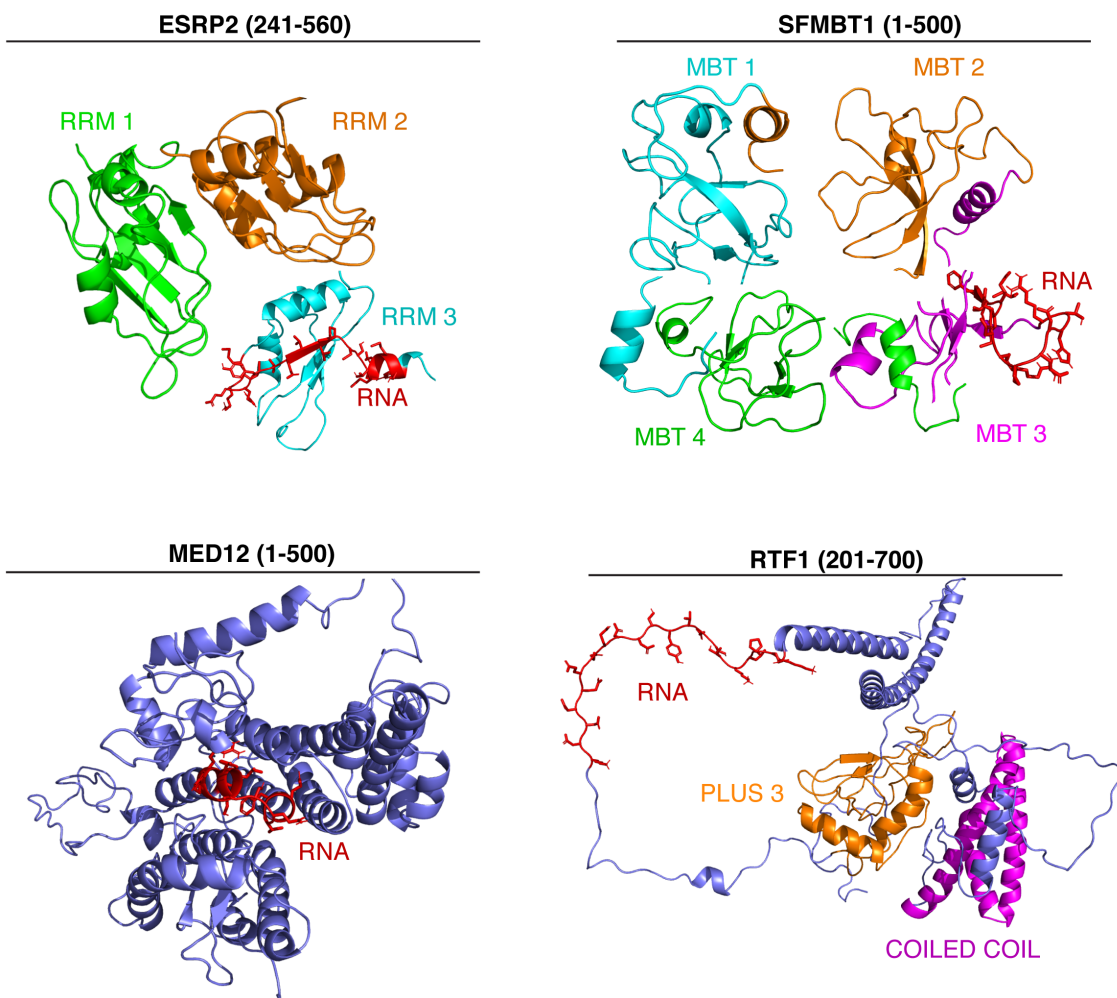


Figure 2.10. Non-BAF co-immunoprecipitated RNA binding proteins

C.

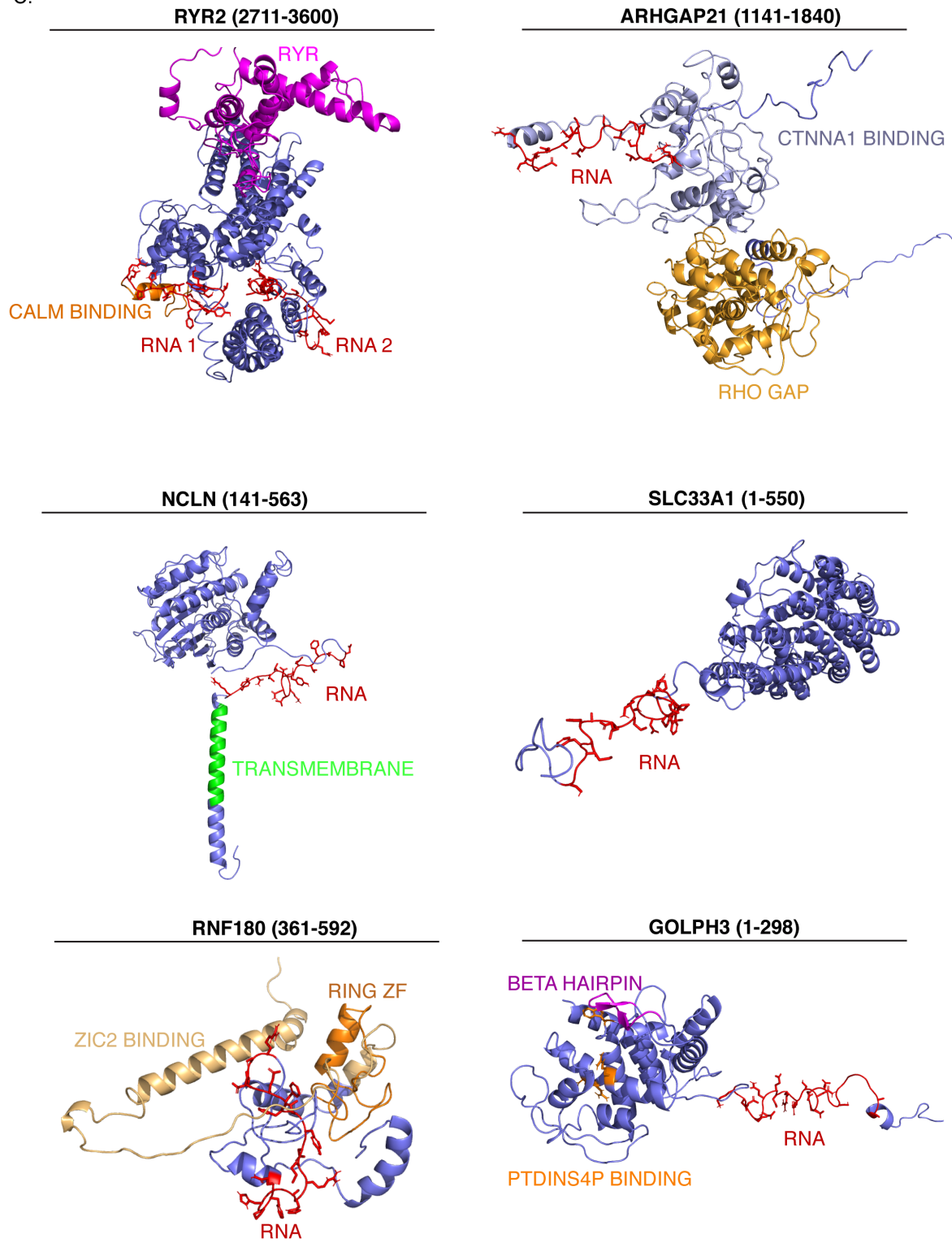
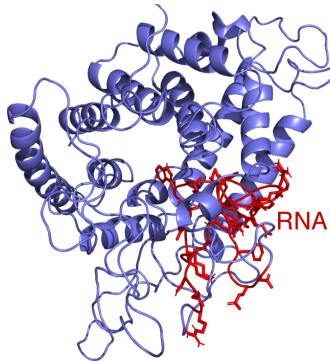


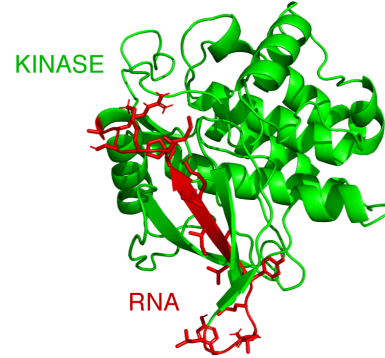
Figure 2.10. Non-BAF co-immunoprecipitated RNA binding proteins

E.

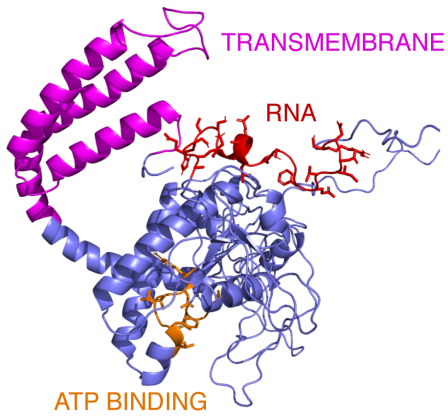
**AGL (1061-1532)**



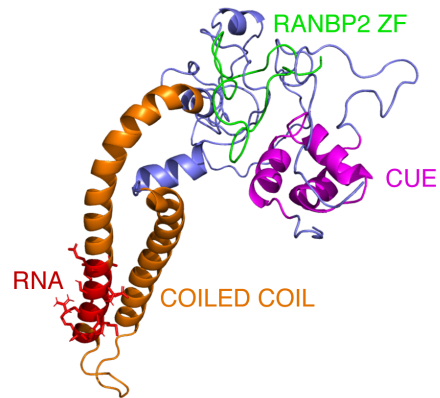
**CSNK1G3 (1-424)**



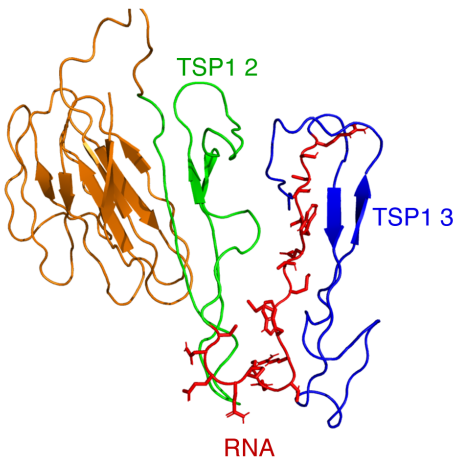
**ADCY9 (211-760)**



**TAB3 (1-716)**



**ADAMTS19 (791-1050)**



**NEUROD4 (1-150)**

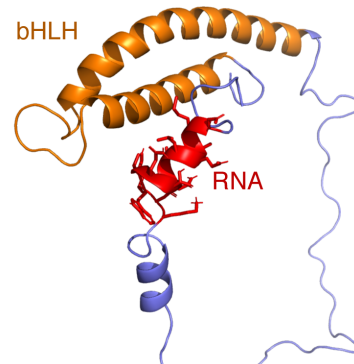


Figure 2.10. Non-BAF co-immunoprecipitated RNA binding proteins



## Chapter 2 References

1. Olins, A. L. and D. E. Olins (1974). "Spheroid chromatin units (v bodies)." *Science* 183(4122): 330-332.
2. Hota, S. K. and B. G. Bruneau (2016). "ATP-dependent chromatin remodeling during mammalian development." *Development (Cambridge, England)* 143(16): 2882-2897.
3. Hang, C. T., J. Yang, P. Han, H.-L. Cheng, C. Shang, E. Ashley, B. Zhou and C.-P. Chang (2010). "Chromatin regulation by Brg1 underlies heart muscle development and disease." *Nature* 466(7302): 62-67.
4. Puri, P. L. and M. Mercola (2012). "BAF60 A, B, and Cs of muscle determination and renewal." *Genes & Development* 26(24): 2673-2683.
5. Lickert, H., J. K. Takeuchi, I. von Both, J. R. Walls, F. McAuliffe, S. Lee Adamson, R. Mark Henkelman, J. L. Wrana, J. Rossant and B. G. Bruneau (2004). "Baf60c is essential for function of BAF chromatin remodelling complexes in heart development." *Nature* 432: 107.
6. Devine, W. P., J. D. Wythe, M. George, K. Koshiba-Takeuchi and B. G. Bruneau (2014). "Early patterning and specification of cardiac progenitors in gastrulating mesoderm." *Elife* 3.
7. Alexander, J. M., S. K. Hota, D. He, S. Thomas, L. Ho, L. A. Pennacchio and B. G. Bruneau (2015). "Brg1 modulates enhancer activation in mesoderm lineage

commitment." Development.

8. Brown, C. J., A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi and H. F. Willard (1991). "A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome." *Nature* 349(6304): 38-44.

9. Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M.

Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai and Y. Hayashizaki (2005). "The transcriptional landscape of the mammalian genome." *Science* 309(5740): 1559-1563.

10. Bell, J. C., D. Jukam, N. A. Teran, V. I. Risca, O. K. Smith, W. L. Johnson, J. M. Skotheim, W. J. Greenleaf and A. F. Straight (2018). "Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts." *Elife* 7.

11. Yang, Y. W., R. A. Flynn, Y. Chen, K. Qu, B. Wan, K. C. Wang, M. Lei and H. Y. Chang (2014). "Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency." *Elife* 3: e02046.

12. Cifuentes-Rojas, C., A. J. Hernandez, K. Sarma and J. T. Lee (2014). "Regulatory interactions between RNA and polycomb repressive complex 2." *Mol Cell*

55(2): 171-185.

13. Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel and R. Shiekhattar (2013). "Activating RNAs associate with Mediator to enhance chromatin architecture and transcription." *Nature* 494(7438): 497-501.
14. Guttman, M., J. Donaghey, B. W. Carey, M. Garber, J. K. Grenier, G. Munson, G. Young, A. B. Lucas, R. Ach, L. Bruhn, X. Yang, I. Amit, A. Meissner, A. Regev, J. L. Rinn, D. E. Root and E. S. Lander (2011). "lincRNAs act in the circuitry controlling pluripotency and differentiation." *Nature* 477(7364): 295-300.
15. Wang, P., Y. Xue, Y. Han, L. Lin, C. Wu, S. Xu, Z. Jiang, J. Xu, Q. Liu and X. Cao (2014). "The STAT3-binding long noncoding RNA Inc-DC controls human dendritic cell differentiation." *Science* 344(6181): 310-313.
16. Han, P., W. Li, C. H. Lin, J. Yang, C. Shang, S. T. Nuernberg, K. K. Jin, W. Xu, C. Y. Lin, C. J. Lin, Y. Xiong, H. Chien, B. Zhou, E. Ashley, D. Bernstein, P. S. Chen, H. V. Chen, T. Quertermous and C. P. Chang (2014). "A long noncoding RNA protects the heart from pathological hypertrophy." *Nature* 514(7520): 102-106.
17. Cajigas, I., D. E. Leib, J. Cochrane, H. Luo, K. R. Swyter, S. Chen, B. S. Clark, J. Thompson, J. R. Yates, 3rd, R. E. Kingston and J. D. Kohtz (2015). "Evf2 lncRNA/BRG1/DLX1 interactions reveal RNA-dependent inhibition of chromatin remodeling." *Development* 142(15): 2641-2652.
18. Wamstad, J. A., J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson,

H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer and B. G. Bruneau (2012). "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage." *Cell* 151(1): 206-220.

19. Kramer, K., T. Sachsenberg, B. M. Beckmann, S. Qamar, K. L. Boon, M. W. Hentze, O. Kohlbacher and H. Urlaub (2014). "Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins." *Nat Methods* 11(10): 1064-1070.

20. Hota, et al. BRG1/BRM-associated factor complex subunit diversity promotes temporally distinct gene expression programs in cardiogenesis. In review.

21. Mellacheruvu, D., Z. Wright, A. L. Couzens, J.-P. Lambert, N. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardi, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras and A. I. Nesvizhskii (2013). "The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data." *Nature methods* 10(8): 730-736.

22. Kuhring, M. and B. Y. Renard (2012). "iPiG: integrating peptide spectrum matches into genome browser visualizations." *PLoS One* 7(12): e50246.

23. Källberg, M., H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu and J. Xu (2012). "Template-based protein structure modeling using the RaptorX web server." *Nature Protocols* 7: 1511.
24. The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase." *Nucleic Acids Research* 45(D1): D158-D169.
25. The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.unpublished
26. Van Nostrand, E. L., G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman and G. W. Yeo (2016). "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)." *Nat Methods* 13(6): 508-514.
27. Aronesty, E (2011). ea-utils : "Command-line tools for processing biological sequencing data"; <https://github.com/ExpressionAnalysis/ea-utils>
28. Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29(1): 15-21.
29. Pertea, M., D. Kim, G. M. Pertea, J. T. Leek and S. L. Salzberg (2016). "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown." *Nat Protoc* 11(9): 1650-1667.

30. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." *Nat Protoc* 7(3): 562-578.
31. Risso, D., J. Ngai, T. P. Speed and S. Dudoit (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature Biotechnology* 32: 896.
32. Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research* 43(7): e47-e47.
33. Smith, T., A. Heger and I. Sudbery (2017). "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy." *Genome Research* 27(3): 491-499.
34. Guo, Y., S. Mahony and D. K. Gifford (2012). "High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints." *PLoS Computational Biology* 8(8): e1002638.
35. Zerbino, D. R., P. Achuthan, W. Akanni, M R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D.

- Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates and P. Flicek (2018). "Ensembl 2018." *Nucleic Acids Research* 46(D1): D754-D761.
36. Ron, G., Y. Globerson, D. Moran and T. Kaplan (2017). "Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains." *Nature Communications* 8: 2237.
37. Lerdrup, M., J. V. Johansen, S. Agrawal-Singh and K. Hansen (2016). "An interactive environment for agile analysis and visualization of ChIP-sequencing data." *Nature Structural & Molecular Biology* 23: 349.
38. Eden, E., R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini (2009). "GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." *BMC Bioinformatics* 10: 48.
39. Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." *Mol Cell* 38(4): 576-589.
40. Tan, Z., Y. Fu, G. Sharma and D. H. Mathews (2017). "TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs." *Nucleic Acids Res* 45(20): 11570-11581.



41. Jossinet, F., T. E. Ludwig and E. Westhof (2010). "Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels." *Bioinformatics* 26(16): 2057-2059.
42. O'Geen, H., S. Fietze and P. J. Farnham (2010). "Using ChIP-seq Technology to Identify Targets of Zinc Finger Transcription Factors." *Methods in molecular biology* (Clifton, N.J.) 649: 437-455.
43. Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9(4): 357-359.
44. Han, P., W. Li, C. H. Lin, J. Yang, C. Shang, S. T. Nuernberg, K. K. Jin, W. Xu, C. Y. Lin, C. J. Lin, Y. Xiong, H. Chien, B. Zhou, E. Ashley, D. Bernstein, P. S. Chen, H. V. Chen, T. Quertermous and C. P. Chang (2014). "A long noncoding RNA protects the heart from pathological hypertrophy." *Nature* 514(7520): 102-106.
45. Cajigas, I., D. E. Leib, J. Cochrane, H. Luo, K. R. Swyter, S. Chen, B. S. Clark, J. Thompson, J. R. Yates, 3rd, R. E. Kingston and J. D. Kohtz (2015). "Evf2 lncRNA/BRG1/DLX1 interactions reveal RNA-dependent inhibition of chromatin remodeling." *Development* 142(15): 2641-2652.
46. Chandler, R. L., J. Brennan, J. C. Schisler, D. Serber, C. Patterson and T. Magnuson (2013). "ARID1a-DNA interactions are required for promoter occupancy by SWI/SNF." *Mol Cell Biol* 33(2): 265-280.
47. Patsialou, A., D. Wilsker and E. Moran (2005). "DNA-binding properties of ARID

family proteins." *Nucleic Acids Research* 33(1): 66-80.

48. Savkur, R. S. and T. P. Burris (2004). "The coactivator LXXLL nuclear receptor recognition motif." *J Pept Res* 63(3): 207-212.

49. Wang, Z., W. Zhai, J. A. Richardson, E. N. Olson, J. J. Meneses, M. T. Firpo, C. Kang, W. C. Skarnes and R. Tjian (2004). "Polybromo protein BAF180 functions in mammalian cardiac chamber maturation." *Genes Dev* 18(24): 3106-3116.

50. Fujisawa, T. and P. Filippakopoulos (2017). "Functions of bromodomain-containing proteins and their roles in homeostasis and cancer." *Nature Reviews Molecular Cell Biology* 18: 246.

51. Sen, P., P. Vivas, M. L. Dechassa, A. M. Mooney, M. G. Poirier and B. Bartholomew (2013). "The SnAC domain of SWI/SNF is a histone anchor required for remodeling." *Mol Cell Biol* 33(2): 360-370.

52. Morrison, E. A., J. C. Sanchez, J. L. Ronan, D. P. Farrell, K. Varzavand, J. K. Johnson, B. X. Gu, G. R. Crabtree and C. A. Musselman (2017). "DNA binding drives the association of BRG1/hBRM bromodomains with nucleosomes." *Nat Commun* 8: 16080.

53. Boyer, L. A., R. R. Latek and C. L. Peterson (2004). "The SANT domain: a unique histone-tail-binding module?" *Nature Reviews Molecular Cell Biology* 5: 158.

54. Da, G., J. Lenkart, K. Zhao, R. Shiekhattar, B. R. Cairns and R. Marmorstein (2006). "Structure and function of the SWIRM domain, a conserved protein module

found in chromatin regulatory complexes." *Proc Natl Acad Sci U S A* 103(7): 2057-2062.

55. Li, M., X. Xu, C. W. Chang, L. Zheng, B. Shen and Y. Liu (2018). "SUMO2 conjugation of PCNA facilitates chromatin remodeling to resolve transcription-replication conflicts." *Nat Commun* 9(1): 2706.

56. Takeuchi, J. K. and B. G. Bruneau (2009). "Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors." *Nature* 459(7247): 708-711.

57. Bennett-Lovsey, R., S. E. Hart, H. Shirai and K. Mizuguchi (2002). "The SWIB and the MDM2 domains are homologous and share a common fold." *Bioinformatics* 18(4): 626-630.

58. Stros, M., D. Launholt and K. D. Grasser (2007). "The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins." *Cell Mol Life Sci* 64(19-20): 2590-2606.

59. Clery, A., S. Jayne, N. Benderska, C. Dominguez, S. Stamm and F. H. Allain (2011). "Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-beta1." *Nat Struct Mol Biol* 18(4): 443-450.

60. Illingworth, R. S. and A. P. Bird (2009). "CpG islands--'a rough guide'." *FEBS Lett* 583(11): 1713-1720.

61. Akopian, D., K. Shen, X. Zhang and S.-o. Shan (2013). "Signal Recognition Particle: An essential protein targeting machine." *Annual review of biochemistry* 82: 693-721.

62. Liu, Y., M. Asakura, H. Inoue, T. Nakamura, M. Sano, Z. Niu, M. Chen, R. J. Schwartz and M. D. Schneider (2007). "Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells." *Proc Natl Acad Sci U S A* 104(10): 3859-3864.
63. Internal Bruneau Lab data. Unpublished.
64. Ounzain, S., R. Micheletti, C. Arnan, I. Plaisance, D. Cecchi, B. Schroen, F. Reverter, M. Alexanian, C. Gonzales, S. Y. Ng, G. Bussotti, I. Pezzuto, C. Notredame, S. Heymans, R. Guigo, R. Johnson and T. Pedrazzini (2015). "CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis." *J Mol Cell Cardiol* 89(Pt A): 98-112.
65. Schittek, B. and T. Sinnberg (2014). "Biological functions of casein kinase 1 isoforms and putative roles in tumorigenesis." *Molecular Cancer* 13: 231.
66. Uchida, S. (2017). "Besides Imprinting: Meg3 Regulates Cardiac Remodeling in Cardiac Hypertrophy." *Circ Res* 121(5): 486-487.
67. Zürner, M. and S. Schoch (2009). "The mouse and human Liprin- $\alpha$  family of scaffolding proteins: Genomic organization, expression profiling and regulation by alternative splicing." *Genomics* 93(3): 243-253.
68. Thornton, C., M. A. Snowden and D. Carling (1998). "Identification of a novel AMP-activated protein kinase beta subunit isoform that is highly expressed in skeletal muscle." *J Biol Chem* 273(20): 12443-12450.

69. Bebee, T. W., J. W. Park, K. I. Sheridan, C. C. Warzecha, B. W. Cieply, A. M. Rohacek, Y. Xing and R. P. Carstens (2015). "The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development." *Elife* 4.
70. Warzecha, C. C., T. K. Sato, B. Nabet, J. B. Hogenesch and R. P. Carstens (2009). "ESRP1 and ESRP2 are epithelial cell type-specific regulators of FGFR2 splicing." *Molecular cell* 33(5): 591-601.
71. Zhang, J., R. Bonasio, F. Strino, Y. Kluger, J. K. Holloway, A. J. Modzelewski, P. E. Cohen and D. Reinberg (2013). "SFMBT1 functions with LSD1 to regulate expression of canonical histone genes and chromatin-related factors." *Genes Dev* 27(7): 749-766.
72. Rocha, P. P., M. Scholze, W. Bleiss and H. Schrewe (2010). "Med12 is essential for early mouse development and for canonical Wnt and Wnt/PCP signaling." *Development* 137(16): 2723-2731.
73. Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel and R. Shiekhattar (2013). "Activating RNAs associate with Mediator to enhance chromatin architecture and transcription." *Nature* 494(7438): 497-501.
74. Dermody, J. L. and S. Buratowski (2010). "Leo1 subunit of the yeast paf1 complex binds RNA and contributes to complex recruitment." *J Biol Chem* 285(44): 33671-33679.
75. Langenbacher, A. D., C. T. Nguyen, A. M. Cavanaugh, J. Huang, F. Lu and J.-N. Chen (2011). "The PAF1 complex differentially regulates cardiomyocyte specification."

Developmental biology 353(1): 19-28.

76. Fischl, H., F. S. Howe, A. Furger and J. Mellor (2017). "Paf1 Has Distinct Roles in Transcription Elongation and Differential Transcript Fate." *Molecular Cell* 65(4): 685-698.e688.

77. de Jong, R. N., V. Truffault, T. Diercks, E. Ab, M. A. Daniels, R. Kaptein and G. E. Folkers (2008). "Structure and DNA binding of the human Rtf1 Plus3 domain." *Structure* 16(1): 149-159.

78. Takeshima, H., S. Komazaki, K. Hirose, M. Nishi, T. Noda and M. Iino (1998). "Embryonic lethality and abnormal cardiac myocytes in mice lacking ryanodine receptor type 2." *Embo j* 17(12): 3309-3316.

79. Lai, D., M. Wan, J. Wu, P. Preston-Hurlburt, R. Kushwaha, T. Grundstrom, A. N. Imbalzano and T. Chi (2009). "Induction of TLR4-target genes entails calcium/calmodulin-dependent regulation of chromatin remodeling." *Proc Natl Acad Sci U S A* 106(4): 1169-1174.

80. Barcellos, K. S., C. L. Bigarella, M. V. Wagner, K. P. Vieira, M. Lazarini, P. R. Langford, J. A. Machado-Neto, S. G. Call, D. M. Staley, J. Y. Chung, M. D. Hansen and S. T. Saad (2013). "ARHGAP21 protein, a new partner of alpha-tubulin involved in cell-cell adhesion formation and essential for epithelial-mesenchymal transition." *J Biol Chem* 288(4): 2179-2189.

81. Xing, M., M. C. Peterman, R. L. Davis, K. Oegema, A. K. Shiau and S. J. Field

(2016). "GOLPH3 drives cell migration by promoting Golgi reorientation and directional trafficking to the leading edge." *Mol Biol Cell* 27(24): 3828-3840.

82. Haffner, C., U. Dettmer, T. Weiler and C. Haass (2007). "The Nicastrin-like protein Nicalin regulates assembly and stability of the Nicalin-nodal modulator (NOMO) membrane protein complex." *J Biol Chem* 282(14): 10632-10638.

83. Liu, P., B. Jiang, J. Ma, P. Lin, Y. Zhang, C. Shao, W. Sun and Y. Gong (2017). "S113R mutation in SLC33A1 leads to neurodegeneration and augmented BMP signaling in a mouse model." *Disease Models & Mechanisms* 10(1): 53-62.

84. Dykes, I. M., D. Szumska, L. Kuncheria, R. Puliyadi, C.-m. Chen, C. Papanayotou, H. Lockstone, C. Dubourg, V. David, J. E. Schneider, T. M. Keane, D. J. Adams, S. D. M. Brown, S. Mercier, S. Odent, J. Collignon and S. Bhattacharya (2018). "A Requirement for Zic2 in the Regulation of Nodal Expression Underlies the Establishment of Left-Sided Identity." *Scientific Reports* 8: 10439.

85. Jacobson, M. R. and T. Pederson (1998). "Localization of signal recognition particle RNA in the nucleolus of mammalian cells." *Proceedings of the National Academy of Sciences of the United States of America* 95(14): 7981-7986.

86. del Valle-Pérez, B., O. Arqués, M. Vinyoles, A. G. de Herreros and M. Duñach (2011). "Coordinated Action of CK1 Isoforms in Canonical Wnt Signaling." *Molecular and Cellular Biology* 31(14): 2877-2888.

87. Bae, Y.-K., T. Shimizu and M. Hibi (2005). "Patterning of proneuronal and inter-

proneuronal domains by hairy and enhancer of split-related genes in zebrafish neuroectoderm." *Development* 132(6): 1375-1385.

88. Sheng, Y., C. Xu and W. Zeng (2017). "TAB3 defect induces augmented cardioprotection loss from ischemic injury." *Cell Biol Int* 41(7): 787-797.

89. Tardif, J.-C., D. Rhainds, M. Brodeur, Y. Feroz Zada, R. Fouodjio, S. Provost, M. Boulé, S. Alem, J. C. Grégoire, P. L. L'Allier, R. Ibrahim, M.-C. Guertin, I. Mongrain, A. G. Olsson, G. G. Schwartz, E. Rhéaume and M.-P. Dubé (2016). "Genotype-Dependent Effects of Dalcatrapib on Cholesterol Efflux and Inflammation: Concordance With Clinical Outcomes." *Circulation. Cardiovascular Genetics* 9(4): 340-348.

90. Cheng, A., M. Zhang, M. Okubo, K. Omichi and A. R. Saltiel (2009). "Distinct mutations in the glycogen debranching enzyme found in glycogen storage disease type III lead to impairment in diverse cellular functions." *Hum Mol Genet* 18(11): 2045-2052.

91. Vrljicak, P., R. Cullum, E. Xu, A. C. Y. Chang, E. D. Wederell, M. Bilenky, S. J. M. Jones, M. A. Marra, A. Karsan and P. A. Hoodless (2012). "Twist1 Transcriptional Targets in the Developing Atrio-Ventricular Canal of the Mouse." *PLoS ONE* 7(7): e40815.



## Chapter 3

Ablation of cardiac lincRNAs *in vivo* reveals genetic interaction between

*Rubie* and *Bmp4*

### Introduction

The majority of the mammalian genome is transcribed throughout development, while only a small fraction of this yields functional protein<sup>1</sup>. The remaining noncoding RNA is arbitrarily classified into long (lincRNA) and short transcripts based upon length greater or less than 200nt. To date, fewer than 10 lincRNAs have been strongly implicated to be important for cardiac development *in vivo*<sup>2,3,4</sup>. However, these RNA molecules were often products of antisense transcription at canonical protein coding regions. Our studies into the transcriptional complexity at gene loci have brought us to the conclusion that pervasive bidirectional transcription takes place at essentially all genes. Therefore, we classify many of the annotated long ncRNAs that have been studied to be fundamental regulatory components of their respective protein coding genes instead of discrete genes themselves. However, thousands of putative intergenic lincRNAs (lincRNAs), with little protein coding potential, exist as stand-alone units<sup>5</sup>. They can exhibit characteristics indicative of active regulation, such as the histone modifications H3K4me3 and H3K27Ac at their promoters and H3K36me3 throughout their gene body, splicing, 5' m7G capping, and polyadenylation<sup>6,7,8</sup>. LincRNAs also can display considerable sequence conservation and are dynamically expressed in specific tissues at developmentally discrete times. The energy investment a cell puts toward the processing and maintenance of these transcripts indicates their putative importance to

the cell and organism. Therefore, considerable focus must be employed to systematically test the function and requirement of these noncoding genes for proper embryogenesis.

We were most interested in lincRNAs that might act to influence the early commitment of nascent mesoderm into the cardiac lineage. We hypothesized that as yet unstudied transcripts were important for this most fundamental stage of cardiac development. Therefore, we screened for the expression of candidates during mouse embryonic stem cell (mESC) *in vitro* differentiation into cardiomyocytes (CM) through nascent mesoderm (MES), cardiac mesoderm (cMES), and cardiac progenitor (CP) intermediates. Of more than 100,000 considered long noncoding RNA annotations, we identified a cohort of lincRNAs with epigenetic regulation, clear gene structure, and cardiac progenitor specificity, which we then validated *in vivo* in the early embryo. Furthermore, ablation of these noncoding genes revealed their local regulatory roles. In particular, for the first time, we were able to establish a genetic interaction between the noncoding transcript *Rubie* and *Bmp4* in the early developing heart.

## Materials and Methods

### *Informatic search for cardiac lincRNAs*

RNA-seq and ChIP-seq reads from mESC to CM differentiations described in Chapter 1 and 2 were mapped to the mouse genome and aligned to Noncode v4<sup>9</sup> annotated lincRNAs. The following criteria were used to generate a candidate list of lincRNAs. 1.) Less than 0.5 fragments per kilobase per million reads (FPKM) in mESCs. 2.) Greater than 1.0 FPKM at CP stage of differentiation (expression at other time points was not considered). 3.) Positive trimethylation of histone 3 lysine 4 (H3K4me3) ChIP-seq signal at TSS during CP stage. 4.) Positive acetylation of histone 3 lysine 27 (H3K27Ac) at TSS during CP stage. 5.) At least 1 exon splice in transcript. 6.) No splice events into neighboring protein coding genes 7.) TSS at least 1kb from nearest protein-coding gene TSS.

### *Whole mount in situ hybridization*

Primers were designed to amplify *in situ* probe templates between 440bp and 1.5kb for each candidate lincRNA off cDNA from CP stage of *in vitro* differentiation. Templates were electrophoresed in 1.0% agarose gel and purified using QIAquick gel extraction kit (Qiagen). These templates were then TOPO TA cloned into pCR4-TOPO using the TOPO TA cloning kit (Invitrogen) and Sanger sequenced to validate orientation in plasmid and proper composition. 2µg linearized vector for each lincRNA template were then input into digoxigenin (DIG) RNA synthesis kit reactions (Roche) in 40µL total volume using either T7 or T3 primers, depending on template orientation. Transcription

was carried out for 2 hours at 37°C. Afterward, 8U DNase I (NEB) were added to each reaction and incubated for 15 min at 37°C to degrade DNA. DNase reactions were quenched with 1.5µL EDTA, and DIG-RNA probes were cleaned and concentrated with RNeasy Mini Columns (Qiagen), EtOH precipitated and washed, and resuspended in 20µL H<sub>2</sub>O. DIG probes were then diluted to 100µg/mL in HYB buffer (50% formamide + 5X SSC pH 4.5 + 50µg/mL yeast tRNA + 75µg/mL heparin + 0.2% Tween-20 + 0.5% CHAPS + 5mM EDTA). E7.5 through E12.5, mouse embryos were liberated from the uterus and dissected from extraembryonic tissues and membranes. Embryos were washed with D-PBS and fixed overnight in 4% paraformaldehyde and then washed 3x in PBTw (PBS + 0.1% Tween-20) on ice. Embryos were dehydrated in MeOH series (25%, 50%, 75%, 2x 100%, 5 min each). Then, samples were rehydrated by reversing this series including 2 extra PBTw washes. Embryos were bleached in 6% H<sub>2</sub>O<sub>2</sub> in PBTw for 15 minutes at RT with rocking. Embryos were washed 3x 5 min in PBTw and treated with 10ug/mL proteinase K for 5 min (E7.5), 10 min (E8.5), 20 min (E9.5), or 30 min (E10.5) rocking at RT and then quenched 2x with 2mg/mL glycine in PBTw followed by 3x 5min washes in PBTw. Embryos were re-fixed in 4% paraformaldehyde + 0.2% glutaraldehyde for 20 min with rocking and washed an additional 5x 5min with PBTw. Embryos were then rinsed 2x in 65°C HYB buffer and incubated in HYB buffer for 3 hours at 65°C. Then, lincRNA-specific probes (in HYB) were added, respectively, to final concentration of 1µg/mL and hybridized overnight at 65°C. Embryos were rinsed 3x 5 min in 65°C WASH1 buffer (50% formamide + 5X SSC pH 4.5 + 1% SDS) and then incubated 2x 30 min again in 65°C WASH1 buffer. Next, embryos were washed 2x 30

min in 65°C WASH2 (50% formamide + 2X SSC pH 4.5 + 0.1% Tween-20), followed by 3x 5 min RT washes in TTBS (25mM Tris HCl pH 7.4 + 135mM NaCl + 2.5mM KCl + 0.1% Tween-20). Embryos were then blocked in TTBS containing 20% sheep serum for 3 hours at RT and stained overnight with alkaline phosphatase (AP) conjugated anti-DIG Fab fragments in TTBS + 1% sheep serum (1:5000, Roche). Embryos were then rinsed 3x 5min in RT TTBS, followed by 6x 1hr TTBS washes at RT. A final TTBS wash was then performed overnight at 4°C. Embryos were then washed 2x 30 min in AP buffer (100mM Tris pH 9.5 + 50mM MgCl<sub>2</sub> + 100mM NaCl + 0.1% Tween-20) at RT. Then, Boeringer Purple AP substrate was added to embryos to initiate staining reactions. Reactions were allowed to progress in the dark until suitable contrast was observed. AP reactions were quenched with 3x PBTw washes containing 1mM EDTA, followed by multiple PBTw pH5.5 washes. A final fixation was then performed overnight in 4% paraformaldehyde and 0.1% glutaraldehyde at 4°C. Finally, embryos were dehydrated again in methanol series and stored in 100% MeOH at -20°C. Embryos were imaged on an upright microscope, and images were white balanced with Adobe Photoshop.

### *Cas9 lincRNA knockout, mouse husbandry, and genotyping*

All mouse experiments were carried out in accordance with IACUC protocols and cared for by the UCSF LARC. For each lincRNA, two cut sites were targeted to induce a 2-3kb deletion flanking the TSS/promoter. Two sequence-specific truncated single guide RNA (tru-gRNA)<sup>10</sup> regions were separately cloned into pX330 (Addgene). After generating T7 promoter-containing sgRNA templates by PCR using Phusion TAC polymerase (NEB),

tru-sgRNAs were transcribed using the Hiscribe T7 High Yield RNA Synthesis Kit (NEB). Tru-sgRNA was extracted using Trizol reagent (Invitrogen) and dual chloroform purifications before immunoprecipitating with isopropanol. Each tru-gRNA pair was then resuspended in sterile 5mM Tris-HCl in sterile water before pronuclear injection by the Gladstone transgenic mouse core (UCSF). Injections were carried out as described by Yang et al<sup>11</sup>. To increase efficiency of obtaining deletions for each target site, all pairs were co-injected into each of 70 FVB/n pronuclei. All genotyping was performed on tail clips stored at -20°C. To extract gDNA, tail clips were suspended in 100µL 50mM NaOH in H<sub>2</sub>O and incubated at 95°C for 40 minutes. Tubes were agitated to break up tissue, and remaining solids were allowed to settle before use. pH was normalized by addition of 5-10µL of 1M Tris HCl pH 7.4. 1-2µL was then input into PCR reactions using Q5 2X master mix (NEB) and 3 gene-specific primers (for simultaneous WT and KO product amplification). Reactions were carried out according to manufacturer-specified recommendations. F0 founders were first identified and bred into C57BL/6j to establish germline transmission (F1). Separate F1 heterozygotes for each individual lincRNA deletion were then outbred for multiple generations (to C57BL/6j) to reduce off-target effects. Handlr and Atcayos null alleles were generated at Jackson Labs using the same targeting strategy but in a homogenous C57BL/6j background.

#### *Transverse aortic constriction cardiac hypertrophy models*

Operations were performed in the Gladstone under IACUC protocols and monitored by the UCSF LARC. Experiments were performed as described by Duan et al<sup>12</sup>. For transverse aortic constriction (TAC), 12-20 week old male mice were anaesthetized with

ketamine/xylazine and mechanically ventilated. After thoracotomy, TAC was executed between the left common carotid and the brachiocephalic arteries using a 7-0 silk suture and 27-gauge needle. After surgery, pressure overload was confirmed by Doppler probe measurement of flow velocity at the carotid artery. Echocardiography was performed at baseline, 1 week, 4 weeks, 6 weeks, and 8 weeks after operation to measure left ventricle (LV) fractional area change (FAC). LV areas were obtained from two-dimensional measurements at the end-diastole and end-systole. At 8 weeks post-surgery, mice were sacrificed for analysis. First, heart, lung, and body weights were measured. Subsequently, a 10-20mg concentric short axis slice of the left ventricle was collected and preserved in RNAlater reagent (ThermoFisher). Heart sections were disrupted in PureZOL (Bio-Rad) on a TissueLyser II (Qiagen). RNA was then purified with Aurum purification kit (BioRad). qRT-PCR was performed using TaqMan chemistry including FastStart Universal Probe Master (Roche), labeled probes from the Universal Probe Library (Roche), and gene-specific oligonucleotide primers run on a 7900HT (ThermoFisher) cyclor with absolute quantification. Gene expression levels were normalized to *cycloB* and *Actb* internal controls.

#### *E8.25 RNA isolation and qPCR analysis*

At E8.25, embryos were liberated from the uterus and dissected from extraembryonic tissues and membranes. Only embryos displaying late cardiac crescent formation before heart tube expansion and cavitation were kept and deemed E8.25. The anterior half of each embryo was washed twice in cold PBS and transferred into Trizol (Gibco), while the posterior half was washed in PBS and stored at -20°C for genotyping. RNA

from Trizol samples was precipitated using standard protocols and further purified/condensed using Qiagen RNeasy MinElute columns. 250ng RNA was reverse transcribed using the AffinityScript Reverse Transcription kit (Agilent) using 200ng random hexamer and/or 100ng dT<sub>20</sub> primers, where appropriate. RT-qPCR was subsequently performed with 5.0ng cDNA and 500nM gene-specific primers in PowerUP SYBR Green master mix (Thermo Fisher). Reactions were run on a 7900HT (ThermoFisher) cyclor with absolute quantification. Gene expression levels were normalized to Actb internal controls using the  $\Delta$ Ct method.

### *E15.5 Histology*

At E15.5, embryos were liberated from the uterus and dissected from extraembryonic tissues and membranes. Whole hearts were removed, rinsed twice in D-PBS, and fixed overnight in 4% paraformaldehyde. Each heart was then paraffin embedded and sectioned at an oblique transverse plane. Hematoxylin and eosin staining and imaging were performed in the Gladstone Histology Core (UCSF).



## Results

### ***A small subset of annotated intergenic long noncoding RNAs display cardiac specific expression and epigenetic regulation in vitro.***

We hypothesized that, like many canonical genes, a subset of lincRNAs would be specifically expressed in the cardiac lineage. We also predicted that those most critical for heart formation would function early in its development. To find candidate lincRNAs, we performed a bioinformatic screen of previously described RNA-seq data sets (Chapter 1 and 2) from differentiations of mouse ESCs into cardiomyocytes.

Additionally, we integrated parallel histone modification ChIP-seq data<sup>13</sup>. We chose to focus on Noncode version 4.0<sup>9</sup>-annotated transcripts which were lowly expressed in ESCs (FPKM < 0.5), while strongly upregulated in cardiac mesoderm or cardiac progenitors (FPKM >1.0). Given our prior analyses of the transcriptome during cardiac differentiation, we decided RNA that transcribed antisense from protein coding gene promoters were likely representative of a basic component of those genes, instead of discrete noncoding genes themselves. Therefore, we filtered for RNA transcripts whose transcriptional starts sites (TSS) began greater than 1 kilobase (kb) from the TSS of known protein-coding genes. To avoid spurious transcripts, we required candidates be spliced and then further refined the list to those displaying histone H3 lysine-4 trimethylation (H3K4me3) and H3 lysine-27 acetylation (H3K27Ac) at their promoters. After removing annotated transcripts that actually spliced into nearby protein coding genes (A930006K02Rik into *Ifnar1*), we were surprised to find these criteria narrowed candidates to only nine total lincRNAs for study out of 114,104 considered transcripts

(Fig1A).

The lincRNA *Rubie* (**R**na **U**pstream **B**mp4 in the **I**nnear **E**ar, Gm15219) was known to co-express with *Bmp4* after E15.0 in the mouse inner ear, and its perturbed splicing was previously implicated in vestibular malformation and consequent circling behavior<sup>13</sup>. However, our candidate screen revealed it to be expressed much earlier in the developing cardiac mesoderm as well. As in the inner ear, its expression *in vitro* overlapped the TGF- $\beta$  signaling protein *Bmp4*<sup>14</sup>, and these genes, separated by approximately 176kb, co-occupied a strongly interacting region within the same topologically associated domain (TAD)<sup>15</sup> (Fig1B). Interestingly, in prior CHIP experiments (Chapter 2), we also discovered stable BRG1/BAF DNA binding 5' to *Rubie*, as well as at the TSS of *Bmp4*. This further implicated active epigenetic regulation within this domain.

*Hand2*, a transcription factor critical for heart development<sup>16</sup>, was previously shown to be regulated by antisense transcription of the noncoding RNA *Upperhand* (*Uph*) away from its promoter<sup>17</sup>. Our search revealed *5033428122Rik* as an independent lincRNA approximately 8 kilobases downstream of *Hand2*, which we named *Handlr* (**H**and2-**A**ssociated **L**inc**R**na). *Handlr* displayed numerous splice forms, but 3' rapid amplification of cDNA ends (RACE) of E9.5 cDNA revealed a single predominant 5-exon, polyadenylated isoform that varied from its annotated structure (Fig2B). *Handlr*'s expression overlapped *Hand2* *in vitro* (Fig2C), but these genes sat near a TAD border (Fig2D) and were divided by a CTCF insulation site<sup>18</sup>.

Seven additional annotated lincRNAs met the criteria for subsequent analyses. We discovered *Atcayos* (2310050B05Rik) transcription to span the important cardiomyocyte metabolic regulator *Nmrk2*<sup>19</sup> and precede its expression in differentiating cardiac progenitors and cardiomyocytes (Fig2E, F). Furthermore, our previous eCLIP experiments could detect SMARCE1 RNA binding events within *Atcayos*, suggesting a regulatory role for its transcript. Also, *E130006D01Rik*, named *HrtLincR5* (**HeaRT LincRna** of chromosome 5), was expressed within an *Mn1*-interacting DNA domain<sup>20</sup> approximately 275kb downstream of this transcriptional coactivator (Fig3A). This transcript displayed highly stereotypic splicing and was only detected at the cardiac progenitor stage of differentiation (Fig3B). *Gm12829*, named *HrtLincR4* (on chromosome 4) was correlatively expressed within a genomic domain in frequent contact with *Trabd2b*, a Wnt protein binding metalloprotease<sup>21</sup> (Fig3C). We also found SMARCE1 RNA binding within this lincRNA, albeit antisense to *HrtLincR4*, along with stable BRG1/BAF DNA and SMARCE1 RNA interactions within *Trabd2b*. In addition, its expression was only transiently detected within at most an 18-hour window at the cardiac mesoderm (cMES) stage of differentiation (Fig3D). *C430049B03Rik*, named *HrtLincRX* (on X chromosome) was highly expressed early in our differentiation model and contained a miRNA cluster in its 3' tail that had previously been shown to drive cardiomyocyte specification<sup>22</sup> (Fig4A, B). This lincRNA also lied approximately 12.5kb downstream of- and overlapped expression with- the important placental gene *Plac1*<sup>23</sup>. Finally, *5033406O09Rik*, *9630002D21Rik*, and *2810410L24Rik* also fulfilled the criteria of our screen (Fig4C).

All nine lincRNAs contained regions with highly homologous sequence to human and/or mammalian genomes (Figure1-4). To assess the protein coding potential of these candidates, we employed multiple tests. First, we evaluated PhyloCSF<sup>24</sup> codon scores in all three frames for each transcript. Whereas extended stretches of positively-scoring codons could be observed in *Bmp4* and the micropeptide-containing *Apela*<sup>25</sup>, we found very little evidence for protein coding potential in the lincRNA cohort, with one exception. A 28 amino acid reading frame in the second exon of *HrtLincR4* was predicted to have coding potential (Fig5A), which will need to be validated in future experiments. However, *HrtLincR4*, as well as *Rubie*, *Handlr*, *Atcayos*, *HrtlincR5*, and *HrtlincRX*, displayed negative coding-non-coding indices<sup>26</sup> (CNCI) similar to the known lincRNA *Neat1* (Fig5B). Also, five of these six lincRNAs were strongly enriched in the nucleus, where *HrtlincR5* RNA molecules were relatively evenly distributed between the nucleus and cytoplasm (Fig5C). Furthermore, these six lincRNAs could generate cDNA using oligo dT primers at least as efficiently as *Actb* and the polyadenylated lincRNA *Neat1*<sup>27</sup>, suggesting their status as polyadenylated noncoding transcripts (Fig5D).

### **A cohort of screened cardiac lincRNAs display dynamic expression in vivo in the developing mouse heart.**

Next, we looked to understand the spatiotemporal expression patterns in the developing embryo for each of the nine candidate lincRNAs. Therefore, we performed *in situ* hybridization experiments to label each transcript between E7.25 and E10.5. Strikingly, the expression patterns observed *in vitro* were largely predictive of those observed *in vivo*. *Rubie* was first observed in the E7.75 embryo, where, similar to *Bmp4*<sup>28</sup>, it strongly

demarcated the extraembryonic boundary and flanked the eventual heart field. From E8.0 to E8.5, its expression became less focused, spreading throughout the developing cardiac crescent and heart tube, respectively. By E8.75 *Rubie* transcription began migrating away from the heart, where at E9.5, it strongly resided in posterior mesoderm as well as the otic vesicle (Fig6A). *Handlr* was transcribed in the developing heart tube by E8.5, where its expression was strongly detected at E9.5 throughout both first and second heart fields (Fig6B). Additionally, *Handlr* expression was detected within ventral mesoderm at this time. These patterns overlapped what was previously shown for *Hand2* at this developmental stage<sup>29</sup>, suggesting common regulation between *Hand2* and *Handlr*.

*Atcayos*, as predicted by *in vitro* expression patterns, was weakly expressed during early stages of heart tube formation, while it was dramatically upregulated after E9.5 in the developing ventricles, as well as cranial structures and somitic mesenchyme.

HrtLincR5 was broadly expressed throughout the mesoderm, including the nascent cardiac crescent, at E8.25. However, as expected by its short-lived *in vitro* expression patterns, HrtLincR5 was only weakly detected *in vivo* by E9.5 throughout noncardiac mesenchyme (Fig6D).

From E8.25 through E9.5, *HrtLincR4* displayed strong expression in developing pharyngeal mesoderm, just dorsal to the developing cardiac crescent (Fig6E). Given its highly transient expression within differentiating cardiac mesoderm *in vitro*, these data suggested HrtLincR4 to be quickly specified to 2<sup>o</sup> heart field and/or adjacent tissues during the onset of cardiac lineage commitment. HrtLincRX was strongly expressed by

E7.5 during cardiac lineage formation in anterior mesoderm at the extraembryonic boarder, as well as in extraembryonic tissues. At E8.25, it was strongly expressed in the cardiac crescent, amniotic membranes, and the developing allantois. While expression of the miR322/503 cluster was previously shown to be cardiac-specific<sup>22</sup>, this lincRNA was widely expressed throughout the heart, forelimb, and somitic mesoderm at E9.5 and E10.5 (Fig6F). This suggested divergent regulation and/or compounding roles for *HrtLincRX* versus its miRNA components. We could not effectively validate the expression of *5033406O09Rik*, *9630002D21Rik*, or *2810410L24Rik* beyond diffuse, low levels in the developing mouse embryo (Fig6G). These experiments established the striking expression patterns of numerous tissue-specific lincRNAs identified from our screen of *in vitro* cardiac differentiation. Therefore, we aimed to test developmental importance of *Rubie*, *Handlr*, *Atcayos*, *HrtLincR5*, *HrtLincR4*, and *HrtLincRX* expression during embryonic development.

**Cas9 ablation of lincRNA promoter regions *in vivo* identifies local gene regulatory roles.**

In order to determine the requirement for the six lincRNAs that displayed compelling *in vivo* expression, we generated knockout mouse lines through pronuclear Cas9 mRNA and tru-sg<sup>10</sup> RNA injections. For each knockout, paired tru-sgRNAs were co-injected to induce 2-3kb deletions flanking the respective lincRNA transcriptional start site (TSS, Fig7A), which successfully generated heritable alleles for all six target regions. After outbreeding, we crossed heterozygotes and harvested the anterior half of E8.25 embryos for RT-qPCR (Fig7B). We found that these deletions ablated downstream

transcription of each lincRNA, respectively (Fig.7C-H). Therefore, these RNA molecules required promoter-centric RNA polymerase recruitment and/or transcriptional progression and were likely not simply the products of spurious expression from the region. As these lincRNAs were nuclear enriched, we hypothesized they were involved in transcriptional regulation within their local genomic environments. To test this, we measured expression of neighboring protein-coding genes sharing the same respective topologically associated domains (TAD). While *Rubie* was previously implicated in BMP4 signaling in the inner ear<sup>13</sup>, its requirement for proper *Bmp4* expression had not been established. We found that loss of *Rubie* within the nucleus resulted in significant reduction of *Bmp4* expression in the anterior half of the developing embryo during cardiac specification. Furthermore, the amount of transcribed *Rubie* was directly correlated with *Bmp4* levels in this region at the same time point. This effect was maintained even within equivalent underlying genotypes, whereby *Rubie* and *Bmp4* transcript levels were still significantly correlated among *Rubie*<sup>+/-</sup> offspring (Fig7C). These data allowed us to conclude that either the act of *Rubie* transcription and/or its physical RNA molecule were responsible for its observed regulation of *Bmp4*.

Despite proximity to- and co-expression with *Handlr*, *Hand2* activation was not dependent on *Handlr* lincRNA (or its underlying promoter DNA sequence, Fig.7D). We hypothesized that the CTCF boundary between these genes played a role in this segregation. We also could not find a correlation between *Mn1*'s expression to *HrtLincR5* (Fig.7F). In contrast, *Nmrk2* and *Trabd2b* expression was dependent on *Atcayos* (Fig.7E) and *HrtLincR4* (Fig.7G), respectively. Given each of these lincRNA's

interaction with the BAF complex, we postulated that RNA binding within these genes' transcripts was involved in this dependency. Furthermore, *Plac1* transcription was significantly and inversely correlated to *HrtLincRX* levels, whereby loss of *HrtLincRX* resulted in approximately a 2-fold expression increase in *Plac1*. However, using IntaRNA software<sup>30</sup>, we calculated stable RNA-RNA interactions between all three miRNAs constituents of its 3' tail (miRNA-322, miRNA-351, miRNA-503) and the 5'- and 3'-untranslated regions (UTR) of *Plac1*. Therefore, this relationship could likely be explained by the loss of inhibitory miRNA binding to *Plac1* noncoding regions (Fig7H). Nonetheless, these data indicated a potential role for HrtLinRX and its miRNAs in demarcating embryonic from extraembryonic mesoderm during gastrulation and early cardiogenesis.

***lincRNA cohort ablation in vivo results in mild phenotypic effects on mouse embryo development.***

In order to determine the requirement of our lincRNA cohort for viable embryonic development *in vivo*, we bred heterozygotes for each gene and examined ratios of expected offspring that survived to weaning. We could not establish any reduction in viability within null progeny for any lincRNA (Fig8A-F). However, *Rubie* knockout did sporadically recapitulate the circling behavior described by Roberts et al, which they observed as a result of aberrant *Rubie* splicing<sup>13</sup>. This provided evidence that the subtle, yet significant, reduction of *Bmp4* in null embryos was developmentally relevant. Despite their clear expression within the developing heart, we concluded that none of the lincRNAs were individually required for viable development in the FVBn; C57BL/6j



mixed background.

*Handlr* and *Atcayos* were the only lincRNAs present in the adult heart (Fig 9A, B), whereby *Atcayos*' very high expression was reduced by approximately 50% after induction of cardiac hypertrophy via transverse aortic constriction (TAC; Fig9B)<sup>12</sup>.

Therefore, we performed TAC experiments on *Handlr* and *Atcayos* null males and compared their responses to wild type (WT) litter mates. At baseline, we were surprised to find *Handlr* null adults had significantly increased fractional area shortening (FAC) than littermate WT controls. However, loss of *Handlr* did not invoke a significant change in hypertrophic response, whereby this elevated contractility was not sustained greater than 1 week after TAC. We also could not detect any noticeable changes in the expression of the canonical hypertrophic response genes *Bnp*, *Anf*, or *Acta1* due to *Handlr* knockout. Nor was heart or lung hypertrophy significantly altered compared to the WT genetic background (Fig9A). In addition, despite its strong expression in the adult heart, loss of *Atcayos* also did not induce dramatic TAC response effects in comparison to littermate controls (Fig9B).

***Compound heterozygotes reveal genetic interaction between Rubie and Bmp4 in patterning the developing right ventricular outflow tract.***

Despite the lack of overt lethality in lincRNA-deficient offspring, we next worked to examine whether proper morphological heart development was altered in *Handlr* and *Rubie* null embryos, the only conditions that produced noticeable physiological effects. Therefore, we harvested E15.5 hearts and examined transverse histological sections to

establish any change to chamber septation, myocardial trabeculation and/or compaction, or ventricular outflow tract (OFT) development. Although *Handlr*<sup>-/-</sup> adults exhibited increased ventricular fractional shortening over WT controls, we could not associate this functionality with overt changes in cardiac anatomy (Fig10A, B). Due to overlapping expression patterns between *Hand2* and *Handlr* in the developing heart, we next tested *Hand2*<sup>+/-</sup> x *Handlr*<sup>+/-</sup> crosses to eliminate one allele of either *Hand2* or *Handlr* per chromosome. However, neither *Hand2* heterozygosity nor *Hand2* / *Handlr* compound heterozygosity resulted in any clear effects on heart morphogenesis (Fig10C, D). In addition, we did not notice any elevated lethality in *Hand2*<sup>+/-</sup> / *Handlr*<sup>+/-</sup> offspring (data not shown, n = 63).

Our experiments found that *Bmp4* expression was dependent on the amount of *Rubie* transcript present in the nucleus. Numerous studies have established the requirement of proper BMP4 dosage for normal septation of the atria, ventricles, and outflow tract (OFT), as well as viable embryo development<sup>31,32</sup>. Therefore, we tested the hypothesis that compound haploinsufficiency of *Bmp4* and *Rubie* together would result in an exacerbated onset of resulting phenotypes. After breeding *Bmp4*<sup>fl/fl</sup> x *Rubie*<sup>+/-</sup>; *Actb-Cre*<sup>+</sup> in the FVB/n; C57BL/6j mixed genetic background, we did not observe any lethality in *Bmp4*<sup>+/-</sup> (*Actb-Cre*<sup>+</sup>) offspring. However, we did find a modest yet sustained ~20% reduction in recovered pups carrying the *Bmp4* / *Rubie* compound heterozygote genotype (Fig 10F; n = 186). When we again looked at E15.5 hearts, loss of a single *Bmp4* allele did not induce abnormal cardiac phenotypes (Fig10G). However, *Bmp4*<sup>+/-</sup> (*Actb-Cre*<sup>+</sup>); *Rubie*<sup>+/-</sup> offspring exhibited incidences of OFT distortion out of the right

ventricle beyond its typical boundary. In these cases, the origins of the pulmonary artery migrated laterally toward the left ventricular OFT and aortic valve (Fig10H). While we were unable to clearly establish communication between the pulmonary and aortic outflow systems in these incidents, the data indicated a genetic interaction between *Rubie* and *Bmp4* in patterning secondary heart field derivatives. Interestingly, these compound heterozygotes did not display any significant reduction in *Bmp4* expression over the benign *Rubie*<sup>-/-</sup> genotype (data not shown). This suggested that genetic interactions between *Rubie* and *Bmp4* could be the result of more complex underlying mechanisms than simply through effects on BMP4 dosage.

## Discussion

With thousands of uncharacterized noncoding transcriptional elements expressed throughout the genome, efforts must be taken to better understand the functional relevance of unstudied lincRNA genes. Towards this need, these experiments were meant to identify and test the requirement for cardiac progenitor-specific lincRNAs in the developing embryo. We discovered a surprisingly sparse set of annotated genes that contained epigenetically-regulated promoter signatures and cardiac-specific expression *in vivo*. Ablation of these transcripts in the developing mouse revealed modulatory roles of *Rubie*, *Atcayos*, *HrtLincR4*, and *HrtLincRX* within their local genomic environments. In particular, we showed for the first time the requirement of *Rubie* expression for normal *Bmp4* dosage. However, despite clear transcription in the developing heart, none of these lincRNAs, including *Rubie*, was required for cardiac morphogenesis or embryo viability. When we generated compound heterozygotes for *Rubie* and *Bmp4*, though, a slight yet sustained reduction in recovered offspring was observed. Furthermore, this dual haploinsufficiency resulted in incidents of perturbed right ventricular outflow tract orientation.

The subtle effects created by ablation of this cardiac-specific group are in agreement with the results most often obtained by others' efforts to knockout lincRNAs<sup>33</sup>. However, several of these lincRNAs did function within the nucleus to impact gene expression in their local environments, including *Rubie*'s influence on *Bmp4*. Consequently, future experiments are needed to dissect the physical mechanisms that underlie these effects. More so, we hypothesize that, in the majority of lincRNA-centric experiments, overt

phenotypic impacts will be observed only after the molecular components and context of each respective transcript are targeted as a whole.

## Description of Figures

**Figure 3.1.** Epigenetically regulated cardiac lincRNAs and genomic characterization of lincRNA Rubie. A.) Criteria for lincRNA identification and resulting 9 candidates; asterisk, name assigned by Bruneau Lab. B.) UCSC Genome Browser tracks of Rubie RNA-seq and overlaid histone H3 ChIP-seq at ESC, MES, CP, and CM stages of *in vitro* differentiation; ESC, embryonic stem cell, MES, mesoderm, CP, cardiac progenitor, CM, cardiomyocyte; blue, ESC, green, MES, orange, CP, red, CM; K4me3, histone H3 lysine 4 trimethylation; K27me3, histone H3 lysine 27 trimethylation; K27Ac, histone H3 lysine 27 acetylation; RefSeq annotation in blue. C.) Quantified expression of Rubie and Bmp4 at each differentiation stage. D.) 3D Genome Browser\* Hi-C heatmap from Dixon, et al. of chromosome interactions around Bmp4 and Rubie loci; TAD, topologically associated domain.

**Figure 3.2.** Genomic characterization of Handlr and Atcayos lincRNAs. A.) UCSC Genome Browser tracks of Handlr RNA-seq and overlaid histone H3 ChIP-seq at ESC, MES, CP, and CM stages of *in vitro differentiation*; Ensembl annotation in red; actual exon structure of predominant Handlr transcript in black with blue stars. B.) Electrophoregram of Handlr 3' RACE products from E9.5 mouse cDNA and Sanger sequence of predominant RNA transcript; polyA tail highlighted. C.) Quantified expression of Handlr and Hand2 at each differentiation stage. D.) 3D Genome Browser\* Hi-C heatmap from Dixon, et al. of chromosome interactions around Handlr and Hand2 loci; TAD, topologically associated domain. E.) UCSC Genome Browser tracks of Atcayos and Nmrk2 RNA-seq and overlaid histone H3 ChIP-seq at ESC, MES, CP,

and CM stages of *in vitro* differentiation; Ensembl annotation in red. F.) Quantified expression of *Atcayos* and *Nmrk2* at each differentiation stage. ESC, embryonic stem cell, MES, mesoderm, CP, cardiac progenitor, CM, cardiomyocyte; blue, ESC, green, MES, orange, CP, red, CM; K4me3, histone H3 lysine 4 trimethylation; K27me3, histone H3 lysine 27 trimethylation; K27Ac, histone H3 lysine 27 acetylation

**Figure 3.3.** Genomic characterization of *HrtLincR4* and *HrtLincR5* lincRNAs. A.) UCSC Genome Browser tracks of *HrtLincR4* RNA-seq and overlaid histone H3 ChIP-seq during cardiac differentiation *in vitro*; B.) Quantified expression of *HrtlincR4* and *Trabd2b* at each differentiation stage. C.) 3D Genome Browser\* Hi-C heatmap from Dixon, et al. of chromosome interactions around *HrtlincR4* and *Trabd2b* loci; TAD, topologically associated domain. D.) UCSC Genome Browser tracks of *HrtLincR5* RNA-seq and overlaid histone H3 ChIP-seq during cardiac differentiation *in vitro*; E.) Quantified expression of *HrtlincR5* and *Mn1* at each differentiation stage. ESC, embryonic stem cell, MES, mesoderm, cMES, cardiac mesoderm; CP, cardiac progenitor, CM, cardiomyocyte; blue, ESC; green, MES; orange, CP; red, CM; K4me3, histone H3 lysine 4 trimethylation; K27me3, histone H3 lysine 27 trimethylation; K27Ac, histone H3 lysine 27 acetylation; Ensembl annotations in red.

**Figure 3.4.** Genomic characterization of *HrtLincRX*, *5033406O09Rik*, *9630002D21Rik*, and *2810410L24Rik* lincRNAs. A.) UCSC Genome Browser tracks of *HrtLincRX* RNA-seq and overlaid histone H3 ChIP-seq during cardiac differentiation *in vitro*; RefSeq annotation, including 3' miRNA cluster, in blue. B.) Quantified expression of *HrtLincRX* and *Plac1* at each differentiation stage. C.) UCSC Genome Browser tracks of

5033406O09Rik, 9630002D21Rik, 2810410L24Rik RNA-seq and overlaid histone H3 ChIP-seq during cardiac differentiation *in vitro*, respectively, as well as quantified expression at each differentiation stage. ESC, embryonic stem cell, MES, mesoderm, cMES, cardiac mesoderm; CP, cardiac progenitor, CM, cardiomyocyte; blue, ESC; green, MES; orange, CP; red, CM; K4me3, histone H3 lysine 4 trimethylation; K27me3, histone H3 lysine 27 trimethylation; K27Ac, histone H3 lysine 27 acetylation; Ensembl annotation in red, RefSeq annotations in blue.

**Figure 3.5.** Molecular characterization of lincRNA cohort. A.) UCSC Genome Browser tracks of PhyloCSF codon scores for all three frames of known protein coding genes (*Bmp4*, *Apela*) and lincRNA cohort; scale, -15 to +15; positive score indicates higher coding potential; green, (+) strand; red, (-) strand. B.) Left: Coding-nonCoding-Index (CNCI) scores for lincRNA cohort. C.) Nuclear vs Cytoplasmic enrichment of lincRNA cohort compared to *Actb* and known nuclear-enriched lincRNA *Neat1*; \*,  $p < 0.05$ ; \*\*\*,  $p < 0.005$ ; n.s., not significant; Student's 2-tailed t-test. D.) Efficiency of RT-qPCR amplification from dT<sub>20</sub>- or random hexamer-primed cDNA for lincRNA cohort compared to *Actb* and known polyadenylated lincRNA *Neat1*. Data presented as mean +/- SEM.

**Figure 3.6.** LincRNA expression patterns *in vivo*. A.) *in situ* hybridization staining for Rubie from E7.5 through E9.5. B.) *in situ* hybridization staining for Handlr at E8.5 and E9.5. C.) *in situ* hybridization staining for Atcayos at E9.5 and E10.5. D.) *in situ* hybridization staining for HrtLincR5 at E8.25 and E9.5. E.) *in situ* hybridization staining for HrtLincR4 at E8.5 and E9.5. F.) *in situ* hybridization staining for HrtLincRX from E7.5 through E10.5. G.) *in situ* hybridization staining for 5033406O09Rik, 9630002D21Rik,



and 2810410L24Rik, respectively, at various developmental timepoints.

**Figure 3.7.** Cas9 ablation of cardiac lincRNAs in vivo and effects on local gene expression. A.) lincRNA TSS/promoter ablation strategy; TSS, transcriptional start site; tru-sgRNA, truncated single guide RNA. B.) Schematic for RT-qPCR on anterior half of E8.25 embryo; A, anterior, P, posterior, red line, bisection point. C.) Left: gDNA PCR genotyping electrophoregram of Rubie alleles and resulting Rubie and Bmp4 expression in anterior E8.25 embryos; Right: correlation between Rubie transcript expression and Bmp4 expression for all genotypes and only Rubie<sup>+/-</sup> (heterozygotes) only, respectively. D.) gDNA PCR genotyping electrophoregram of Handlr alleles and resulting Handlr and Hand2 expression in anterior E8.25 embryos. E.) gDNA PCR genotyping electrophoregram of Atcayos alleles and resulting Atcayos and Nmrk2 expression in anterior E8.25 embryo. F.) gDNA PCR genotyping electrophoregram of HrtLincR5 alleles and resulting HrtLincR5 and Mn1 expression in anterior E8.25 embryos. G.) gDNA PCR genotyping electrophoregram of HrtLincR4 alleles and resulting HrtLincR4 and Trabd2b expression in anterior E8.25 embryos. H.) Left: gDNA PCR genotyping electrophoregram of HrtLincRX alleles and resulting HrtLincRX and Plac1 expression in anterior E8.25 embryos; Right: correlation between HrtLincRX transcript expression and Plac1 expression; IntaRNA 2.0 binding prediction between HrtLincRX 3' miRNAs and Plac1 UTRs. \*, p < 0.05; \*\*\*, p < 0.005; n.s., not significant; Student's 2-tailed t-test. Data presented as mean +/- SEM.

**Figure 3.8.** Requirements for lincRNA cohort for viable development. A.) Left: Offspring recovered at weening from Rubie<sup>+/-</sup> x Rubie<sup>+/-</sup> cross vs expected Mendelian ratios;

Right: Representative sporadic circling behavior only observed in *Rubie*<sup>-/-</sup> offspring. B.) Offspring recovered at weening from *Handlr*<sup>+/-</sup> x *Handlr*<sup>+/-</sup> cross vs expected Mendelian ratios. C.) Offspring recovered at weening from *Atcayos*<sup>+/-</sup> x *Atcayos*<sup>+/-</sup> cross vs expected Mendelian ratios. D.) Offspring recovered at weening from *HrtLincR5*<sup>+/-</sup> x *HrtLincR5*<sup>+/-</sup> cross vs expected Mendelian ratios. E.) Offspring recovered at weening from *HrtLincR4*<sup>+/-</sup> x *HrtLincR4*<sup>+/-</sup> cross vs expected Mendelian ratios. F.) Male offspring recovered at weening from *HrtLincRX*<sup>+/-</sup> x *HrtLincRX*<sup>+/-</sup> cross vs expected Mendelian ratios.

**Figure 3.9.** TAC hypertrophy models in *Handlr* and *Atcayos* null mice. A.) Top left: RNA-seq expression of *Handlr* in adult heart before and after transaortic constriction (TAC) from Duan et al, 2017\*; Top right: Fractional area shortening of *Handlr*<sup>-/-</sup> and wildtype (WT) littermate controls at baseline and after TAC; \*, p < 0.05; \*\*, p < 0.01; Student's t-test; Bottom left: RT-qPCR of canonical hypertrophic response genes at 8 weeks after TAC; Heart and lung organ weights at 8 weeks after TAC; n=5-7. B.) Top left: RNA-seq expression of *Atcayos* in adult heart before and after TAC from Duan et al, 2017\*; Top right: Fractional area shortening of *Atcayos*<sup>-/-</sup> and WT littermate controls at baseline and after TAC; Bottom left: RT-qPCR of canonical hypertrophic response genes at 8 weeks after TAC; Heart and lung organ weights at 8 weeks after TAC; n=7-9. n.s., not significant. Data presented as mean +/- SEM.

**Figure 3.10.** Effect of lincRNA ablation, *Hand2 / Handlr* compound heterozygosity, and *Rubie / Bmp4* compound heterozygosity on heart development. A-E,G-H.) Oblique transverse hematoxylin and eosin (H&E) histological sections of cardiac ventricular and

outflow tract (OFT) morphogenesis, respectively, at E15.5. A.) Representative wild type (WT) morphology. B.) Representative *Handlr<sup>-/-</sup>* morphology. C.) Representative *Hand2<sup>+/-</sup>* morphology. D.) Representative *Hand2<sup>+/-</sup>; Handlr<sup>+/-</sup>* morphology. E.) Representative *Rubie<sup>-/-</sup>* morphology. F.) Top: gDNA PCR genotyping electrophoregram of *Rubie* and *Actb-Cre* transgene alleles. Bottom: Offspring recovered at weening from *Rubie<sup>+/-</sup>; Actb-Cre<sup>+</sup>* x *Bmp4<sup>fl/fl</sup>* mating vs expected Mendelian ratios G.) Representative *Bmp4<sup>+/-</sup>* (*Actb-Cre<sup>+</sup>*) morphology. H.) Representative *Bmp4<sup>+/-</sup>; Rubie<sup>+/-</sup>* (*Actb-Cre<sup>+</sup>*) morphology in 2 separate individuals. RV, right ventricle; LV, left ventricle; OFT, outflow tract; scale bar, 300 $\mu$ m; arrows, mal-formed OFT orientation.

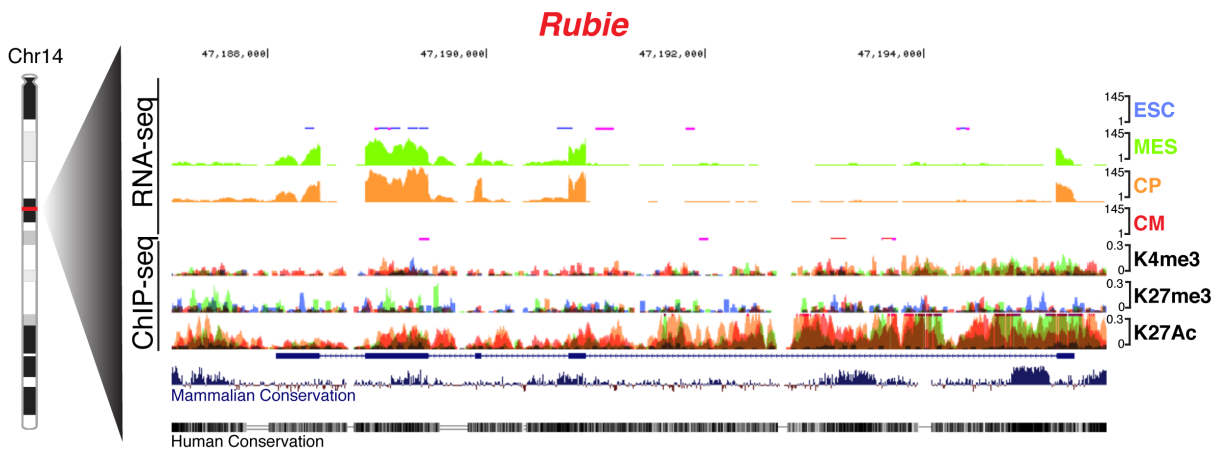
A.

***lincRNA Criteria***

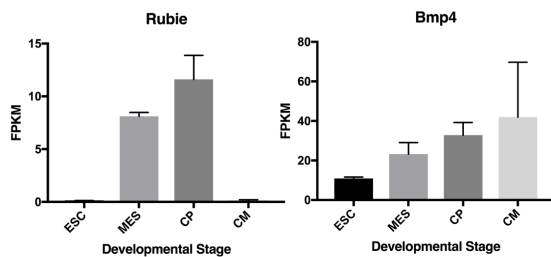
1. Noncode v4 annotated
2. *lincRNA* TSS > 1kb from protein coding TSS
3. Splices > 1 (discrete from nearby genes)
4. FPKM < 0.5 in ESCs
5. FPKM > 1.0 in cardiac progenitors
6. Promoter H3K4me3 in cardiac progenitors
7. Promoter H3K27Ac in cardiac progenitors

| <b>Annotated Name</b> | <b>Short Name</b> | <b>Location(mm9)</b>      | <b>Neighbor</b> |
|-----------------------|-------------------|---------------------------|-----------------|
| Gm15219               | Rubie             | chr14:47188052-47189230   | Bmp4            |
| 5033428I22Rik         | Handlr*           | chr8:59810914-59819479    | Hand2           |
| 2310050B05Rik         | Atcayos           | chr10:80657435-80673622   | Nmrk2           |
| E130006D01Rik         | HrtLincR5*        | chr5:112163303-112190748  | Mn1             |
| Gm12829               | HrtLincR4*        | chr4:114353409-114360200  | Trabd2b         |
| C430049B03Rik         | HrtLincRX*        | chrX:50406289-50410367    | Plac1           |
| 5033406O09Rik         | -                 | chr11:120046713-120050065 | -               |
| 9630002D21Rik         | -                 | chr12:113180201-113182695 | -               |
| 2810410L24Rik         | -                 | chr12:72678951-72687808   | -               |

B.



C.



D.

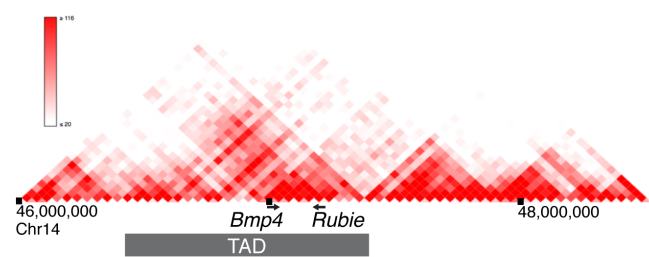


Figure 3.1. Epigenetically regulated cardiac *lincRNAs* and genomic characterization of *lincRNA Rubie*



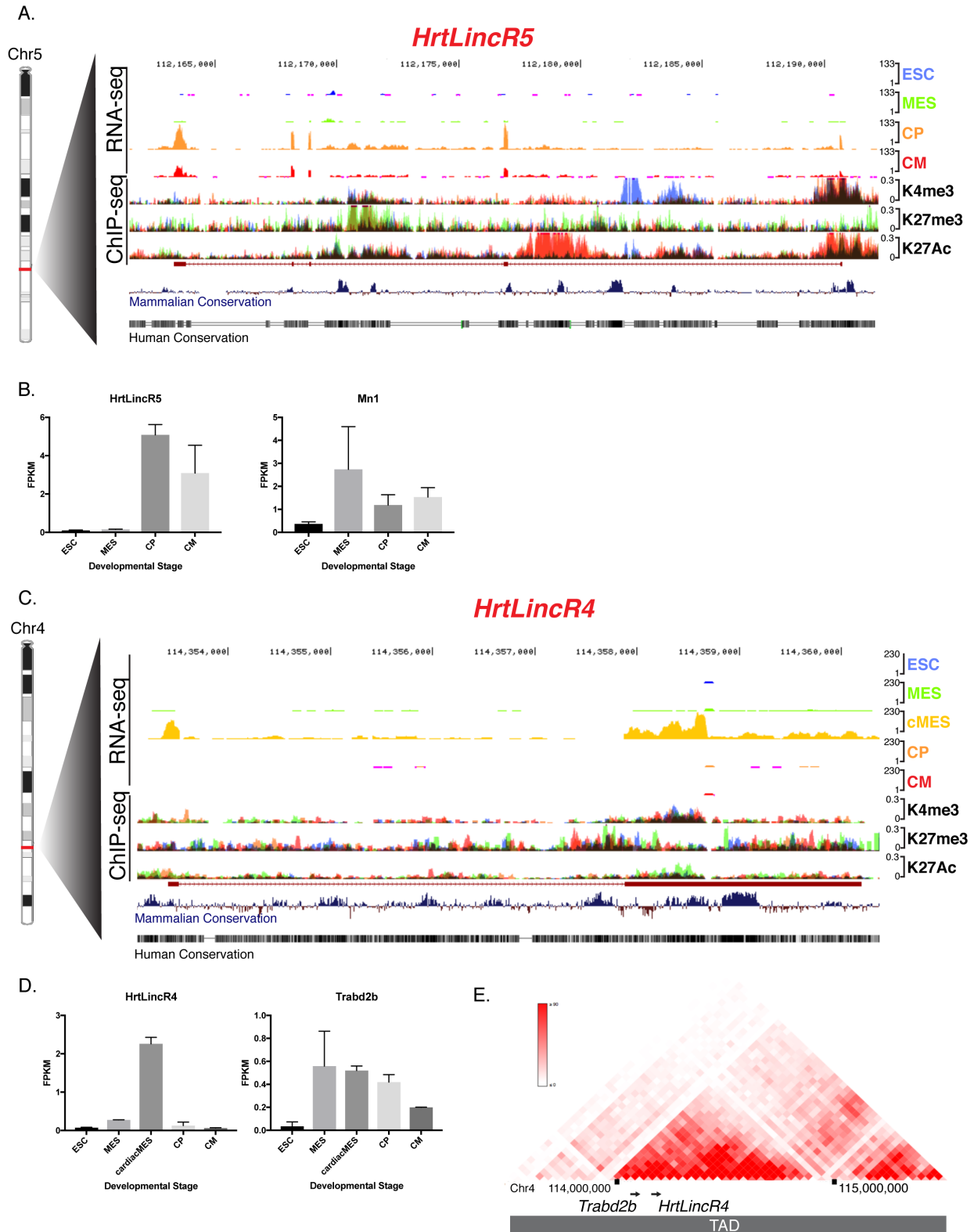


Figure 3.3. Genomic characterization of HrtLincR4 and HrtLincR5 lincRNAs

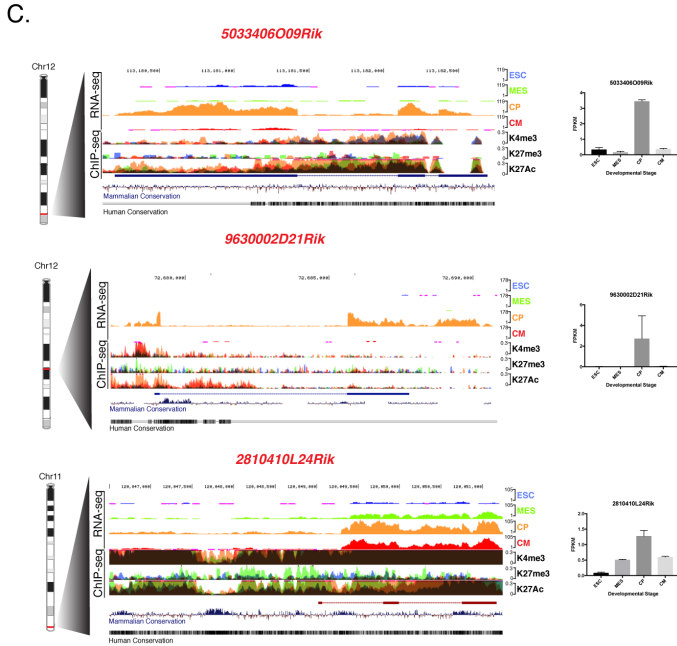
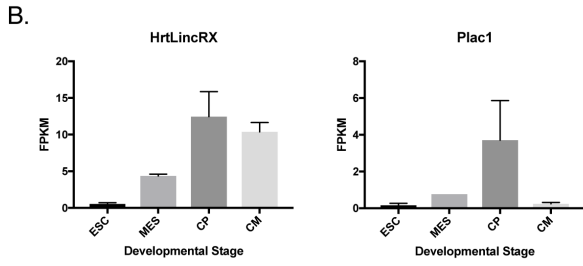
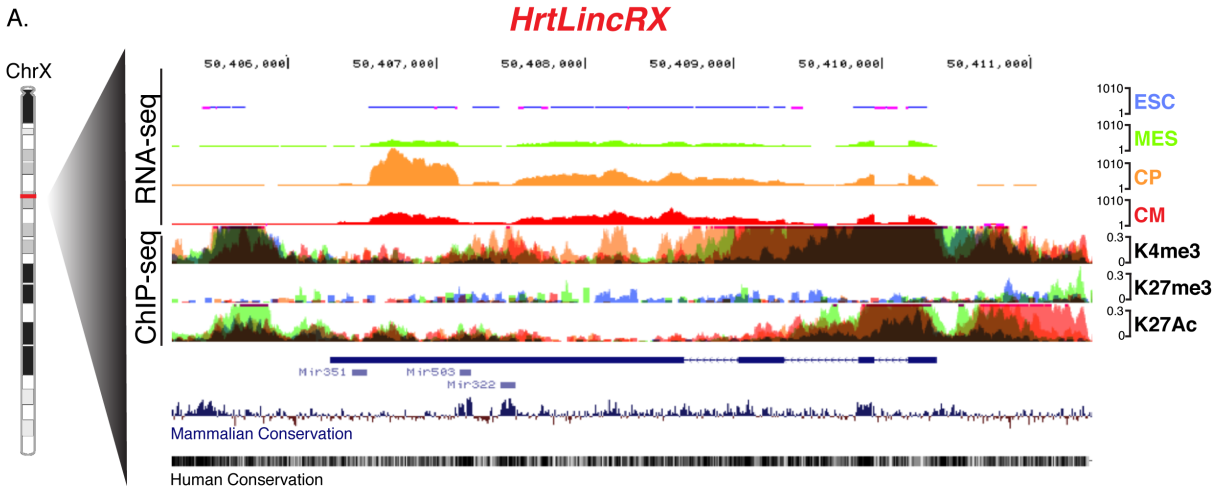
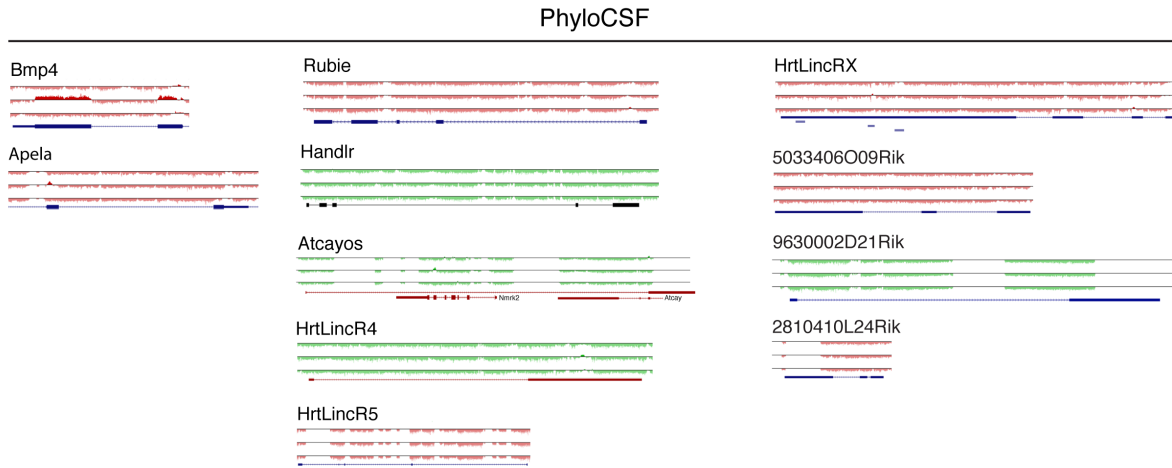
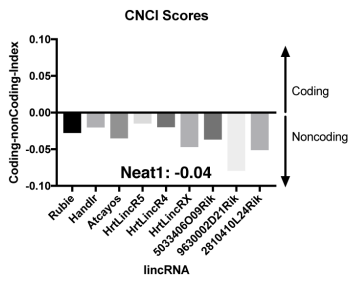


Figure 3.4. Genomic characterization of HrtLincRX, 5033406O09Rik, 9630002D21Rik, and 2810410L24Rik lincRNAs

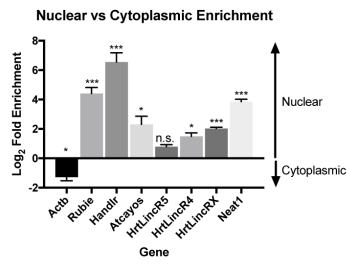
A.



B.



C.



D.

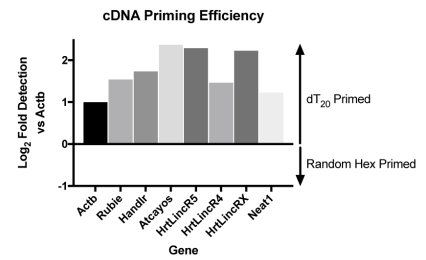


Figure 3.5. Molecular characterization of lincRNA cohort



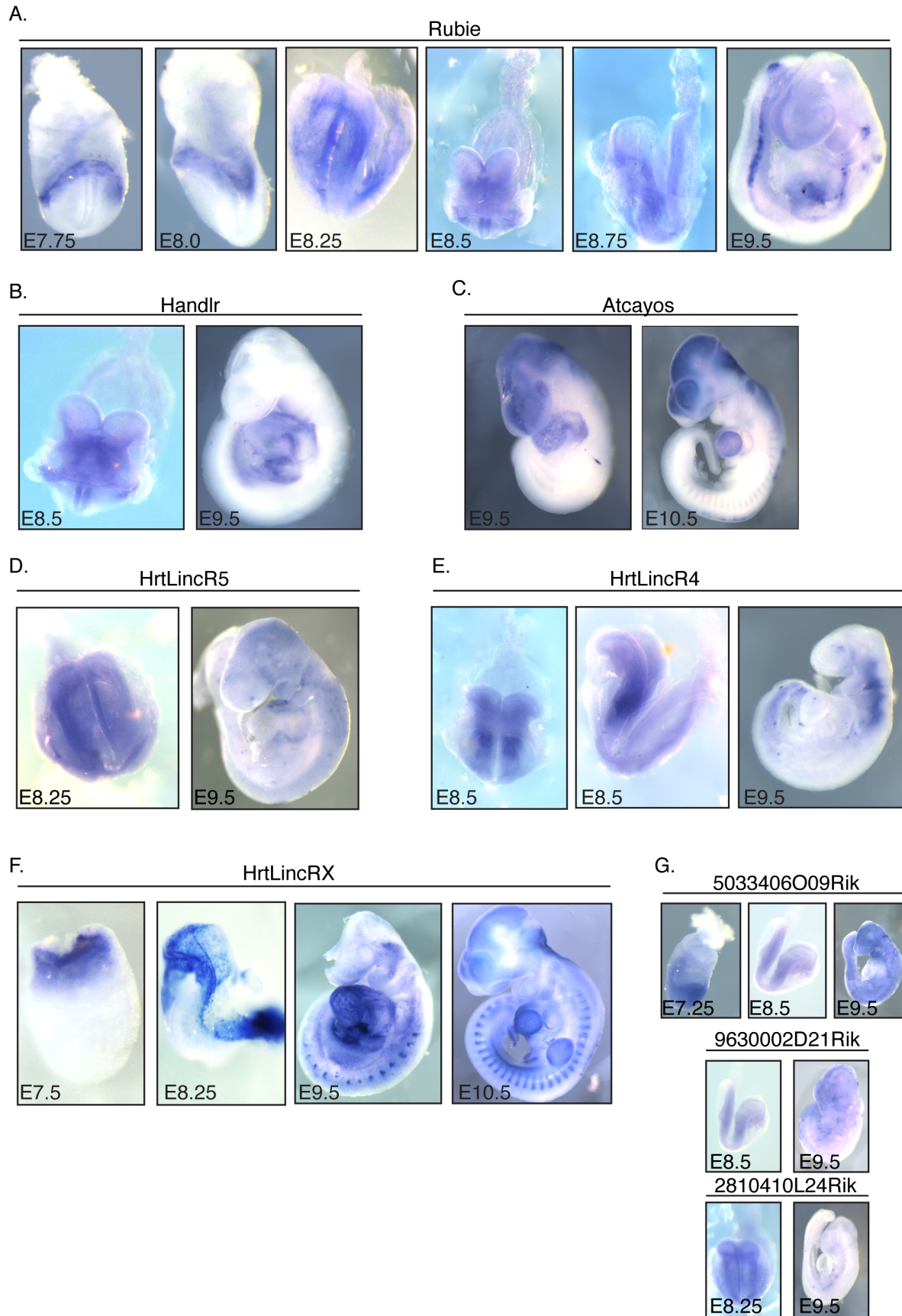


Figure 3.6. LincRNA expression patterns in vivo

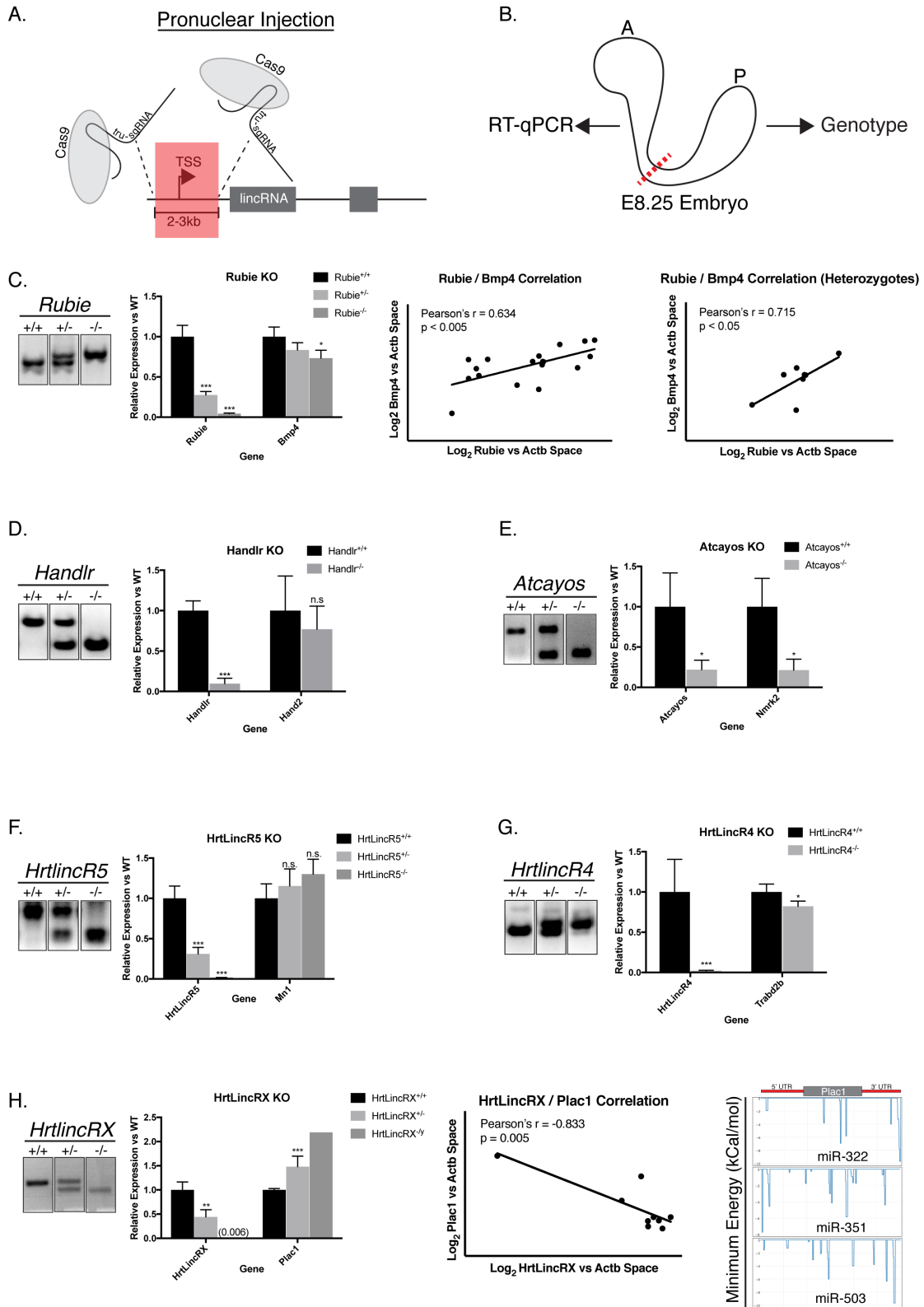


Figure 3.7. Cas9 ablation of cardiac lincRNAs in vivo and effects on local gene expression

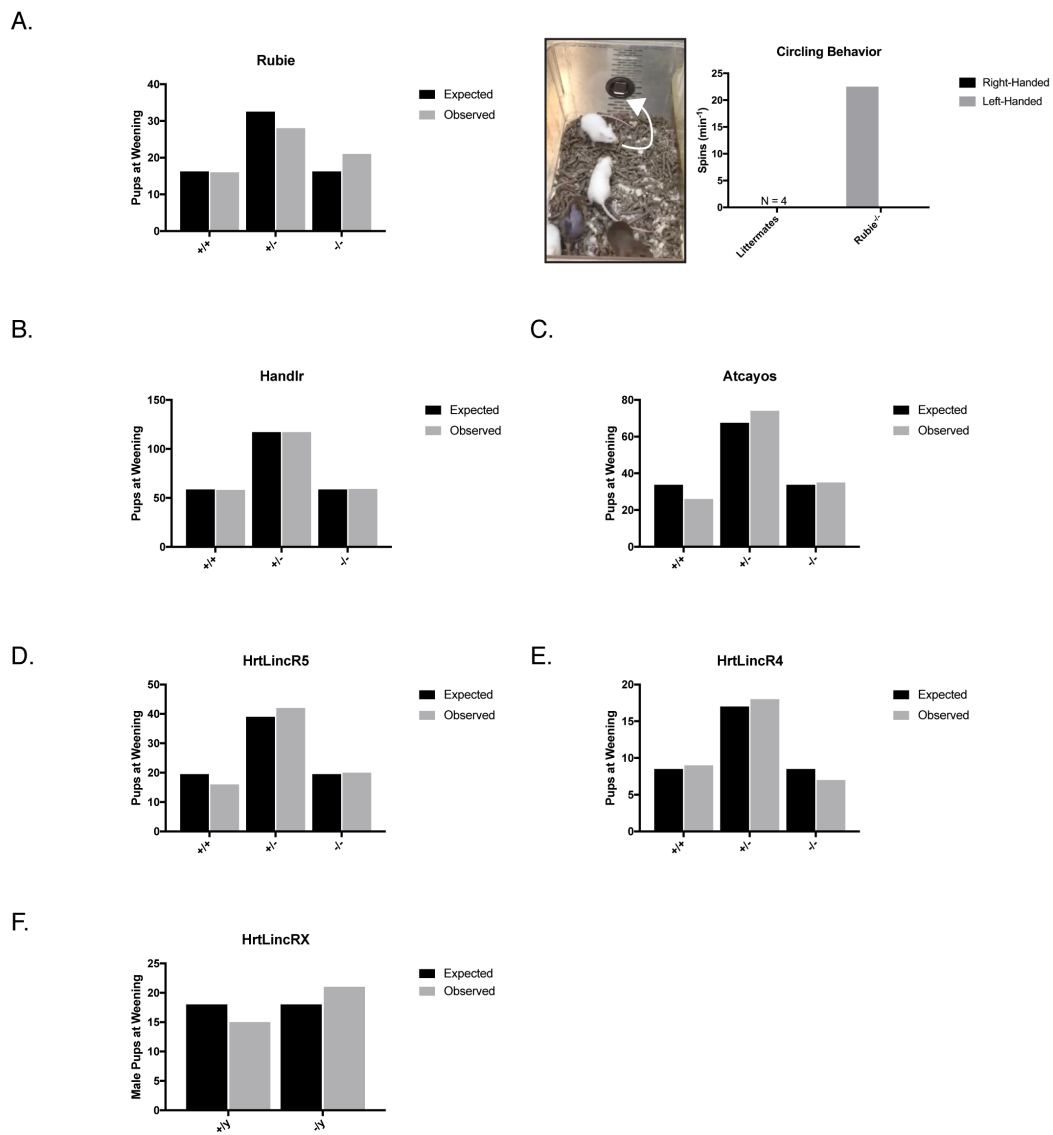
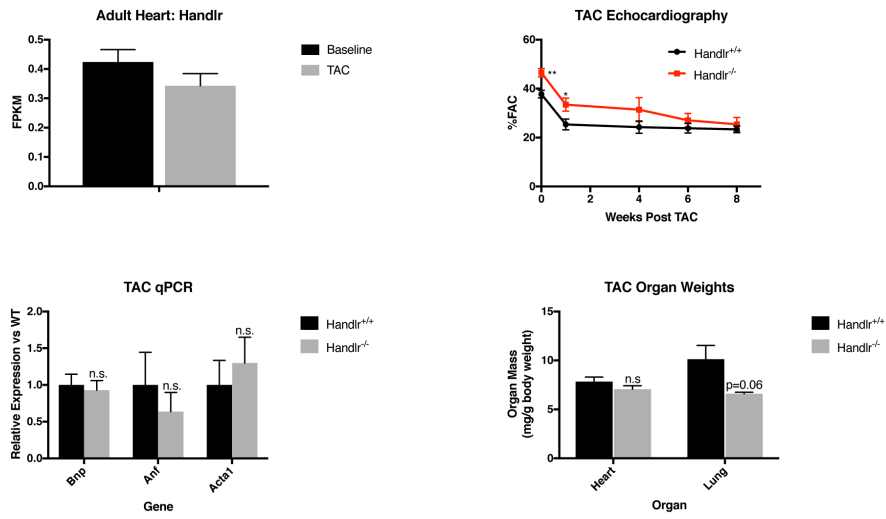


Figure 3.8. Requirements for lincRNA cohort for viable development

A.



B.

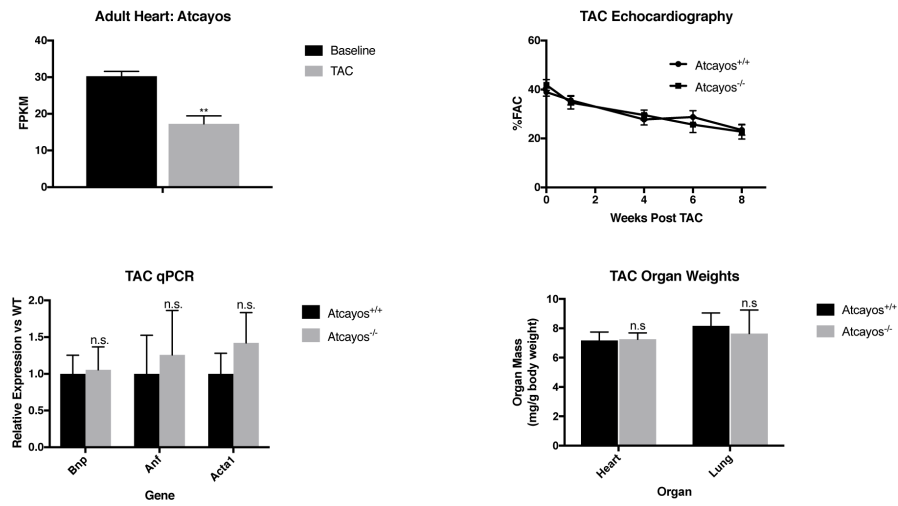


Figure 3.9. TAC hypertrophy models in Handlr and Atcayos null mice

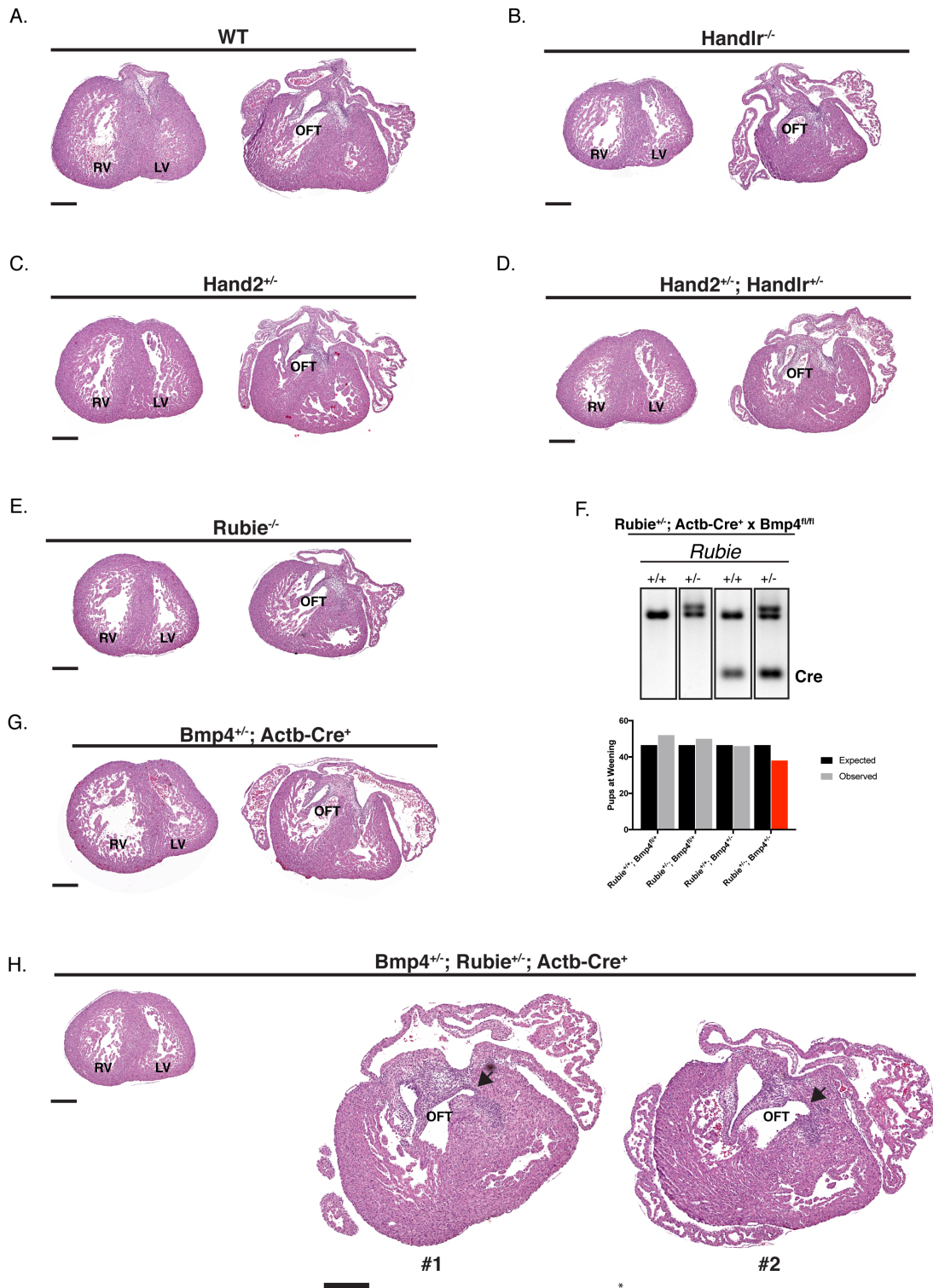


Figure 3.10. Effect of lincRNA ablation, Hand2 / Handlr compound heterozygosity, and Rubie / Bmp4 compound heterozygosity on heart development

### Chapter 3 References

1. Wong, G. K., D. A. Passey and J. Yu (2001). "Most of the human genome is transcribed." Genome Res **11**(12): 1975-1977.
2. Grote, P., L. Wittler, D. Hendrix, F. Koch, S. Wahrisch, A. Beisaw, K. Macura, G. Blass, M. Kellis, M. Werber and B. G. Herrmann (2013). "The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse." Dev Cell **24**(2): 206-214.
3. Han, P., W. Li, C. H. Lin, J. Yang, C. Shang, S. T. Nuernberg, K. K. Jin, W. Xu, C. Y. Lin, C. J. Lin, Y. Xiong, H. Chien, B. Zhou, E. Ashley, D. Bernstein, P. S. Chen, H. V. Chen, T. Quertermous and C. P. Chang (2014). "A long noncoding RNA protects the heart from pathological hypertrophy." Nature **514**(7520): 102-106.
4. Micheletti, R., I. Plaisance, B. J. Abraham, A. Sarre, C. C. Ting, M. Alexanian, D. Maric, D. Maison, M. Nemir, R. A. Young, B. Schroen, A. Gonzalez, S. Ounzain and T. Pedrazzini (2017). "The long noncoding RNA Wisper controls cardiac fibrosis and remodeling." Sci Transl Med **9**(395).
5. Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M.

Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M.  
Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S.  
Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L.  
Huminięcki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M.  
Katoh, Y. Kawasaki, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A.  
Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J.  
Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda,  
S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N.  
Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F.  
Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G.  
Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B.  
Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K.  
Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K.  
Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y.  
Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A.  
Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H.  
Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R.  
D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D.  
A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M.  
Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N.  
Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N.  
Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M.

- Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai and Y. Hayashizaki (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.
6. Quinn, J. J. and H. Y. Chang (2016). "Unique features of long non-coding RNA biogenesis and function." Nat Rev Genet **17**(1): 47-62.
  7. Sati, S., S. Ghosh, V. Jain, V. Scaria and S. Sengupta (2012). "Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci." Nucleic Acids Res **40**(20): 10018-10031.
  8. Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow and R. Guigo (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.
  9. Xie, C., J. Yuan, H. Li, M. Li, G. Zhao, D. Bu, W. Zhu, W. Wu, R. Chen and Y. Zhao (2014). "NONCODEv4: exploring the world of long non-coding RNA genes." Nucleic Acids Research **42**(Database issue): D98-D103.
  10. Fu, Y., J. D. Sander, D. Reyon, V. M. Cascio and J. K. Joung (2014). "Improving CRISPR-Cas nuclease specificity using truncated guide RNAs." Nat Biotechnol **32**(3): 279-284.



11. Yang, H., H. Wang and R. Jaenisch (2014). "Generating genetically modified mice using CRISPR/Cas-mediated genome engineering." Nat Protoc **9**(8): 1956-1968.
12. Duan, Q., S. McMahon, P. Anand, H. Shah, S. Thomas, H. T. Salunga, Y. Huang, R. Zhang, A. Sahadevan, M. E. Lemieux, J. D. Brown, D. Srivastava, J. E. Bradner, T. A. McKinsey and S. M. Haldar (2017). "BET bromodomain inhibition suppresses innate inflammatory and profibrotic transcriptional networks in heart failure." Science translational medicine **9**(390): eaah5084.
13. Wamstad, J. A., J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer and B. G. Bruneau (2012). "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage." Cell **151**(1): 206-220.
14. Roberts, K. A., V. E. Abraira, A. F. Tucker, L. V. Goodrich and N. C. Andrews (2012). "Mutation of Rubie, a novel long non-coding RNA located upstream of Bmp4, causes vestibular malformation in mice." PLoS One **7**(1): e29495.
15. Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature **485**(7398): 376-380.
16. Srivastava, D., T. Thomas, Q. Lin, M. L. Kirby, D. Brown and E. N. Olson (1997). "Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND." Nat Genet **16**(2): 154-160.

17. Anderson, K. M., D. M. Anderson, J. R. McAnally, J. M. Shelton, R. Bassel-Duby and E. N. Olson (2016). "Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development." Nature **539**(7629): 433-436.
18. Martin, D., C. Pantoja, A. Fernandez Minan, C. Valdes-Quezada, E. Molto, F. Matesanz, O. Bogdanovic, E. de la Calle-Mustienes, O. Dominguez, L. Taher, M. Furlan-Magaril, A. Alcina, S. Canon, M. Fedetz, M. A. Blasco, P. S. Pereira, I. Ovcharenko, F. Recillas-Targa, L. Montoliu, M. Manzanares, R. Guigo, M. Serrano, F. Casares and J. L. Gomez-Skarmeta (2011). "Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes." Nat Struct Mol Biol **18**(6): 708-714.
19. Diguët, N., S. A. J. Trammell, C. Tannous, R. Deloux, J. Piquereau, N. Mougnot, A. Gouge, M. Gressette, B. Manoury, J. Blanc, M. Breton, J. F. Decaux, G. G. Lavery, I. Baczko, J. Zoll, A. Garnier, Z. Li, C. Brenner and M. Mericskay (2018). "Nicotinamide Riboside Preserves Cardiac Function in a Mouse Model of Dilated Cardiomyopathy." Circulation **137**(21): 2256-2273.
20. van Wely, K. H., A. C. Molijn, A. Buijs, M. A. Meester-Smoor, A. J. Aarnoudse, A. Hellemons, P. den Besten, G. C. Grosveld and E. C. Zwarthoff (2003). "The MN1 oncoprotein synergizes with coactivators RAC3 and p300 in RAR-RXR-mediated transcription." Oncogene **22**(5): 699-709.
21. The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Research **45**(D1): D158-D169.


22. Shen, X., B. Soibam, A. Benham, X. Xu, M. Chopra, X. Peng, W. Yu, W. Bao, R. Liang, A. Azares, P. Liu, P. H. Gunaratne, M. Mercola, A. J. Cooney, R. J. Schwartz and Y. Liu (2016). "miR-322/-503 cluster is expressed in the earliest cardiac progenitor cells and drives cardiomyocyte specification." Proceedings of the National Academy of Sciences **113**(34): 9551-9556.
23. Jackman, S. M., X. Kong and M. E. Fant (2012). "Plac1 (placenta-specific 1) is essential for normal placental and embryonic development." Mol Reprod Dev **79**(8): 564-572.
24. Lin, M. F., I. Jungreis and M. Kellis (2011). "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions." Bioinformatics **27**(13): i275-i282.
25. Pauli, A., M. L. Norris, E. Valen, G.-L. Chew, J. A. Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, S. Q. Tsai, J. K. Joung, A. Saghatelian and A. F. Schier (2014). "Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors." Science (New York, N.Y.) **343**(6172): 1248636-1248636.
26. Sun, L., H. Luo, D. Bu, G. Zhao, K. Yu, C. Zhang, Y. Liu, R. Chen and Y. Zhao (2013). "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts." Nucleic Acids Research **41**(17): e166-e166.
27. Sasaki, Y. T., T. Ideue, M. Sano, T. Mituyama and T. Hirose (2009). "MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles." Proc Natl Acad Sci U S A **106**(8): 2525-2530.

28. Perea-Gomez, A., W. Shawlot, H. Sasaki, R. R. Behringer and S. Ang (1999). "HNF3beta and Lim1 interact in the visceral endoderm to regulate primitive streak formation and anterior-posterior polarity in the mouse embryo." Development **126**(20): 4499-4511.
29. Charite, J., D. G. McFadden and E. N. Olson (2000). "The bHLH transcription factor dHAND controls Sonic hedgehog expression and establishment of the zone of polarizing activity during limb development." Development **127**(11): 2461-2470.
30. Mann, M., P. R. Wright and R. Backofen (2017). "IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions." Nucleic Acids Res **45**(W1): W435-w439.
31. Dunn, N. R., G. E. Winnier, L. K. Hargett, J. J. Schrick, A. B. Fogo and B. L. Hogan (1997). "Haploinsufficient phenotypes in Bmp4 heterozygous null mice and modification by mutations in Gli3 and Alx4." Dev Biol **188**(2): 235-247.
32. Jiao, K., H. Kulesa, K. Tompkins, Y. Zhou, L. Batts, H. S. Baldwin and B. L. M. Hogan (2003). "An essential role of Bmp4 in the atrioventricular septation of the mouse heart." Genes & Development **17**(19): 2362-2367.
33. Sauvageau, M., L. A. Goff, S. Lodato, B. Bonev, A. F. Groff, C. Gerhardinger, D. B. Sanchez-Gomez, E. Hacisuleyman, E. Li, M. Spence, S. C. Liapis, W. Mallard, M. Morse, M. R. Swerdel, M. F. D'Ecclessis, J. C. Moore, V. Lai, G. Gong, G. D. Yancopoulos, D. Friendewey, M. Kellis, R. P. Hart, D. M. Valenzuela, P. Arlotta and J. L. Rinn (2013). "Multiple knockout mouse models reveal lincRNAs are required for life and brain development." Elife **2**: e01749.

## Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date 06/30/2018