Sam68 and circRNA biogenesis in early development

Maria Carla Antonelli

TESI DOCTORAL UPF / YEAR 2020

Directors de la tesi

Dr. Gian Gaetano Tartaglia

Dr. Elias George Bechara El Halal

POMPEU FABRA UNIVERSITY

Department of Experimental and Health Sciences

CENTRE FOR GENOMIC REGULATION

Department of Bioinformatics and Genomics

A nonno Gino
ed al futuro nuovo arrivato Teodoro,

Acknowledgements

During these years I had the chance to get to know many people that became part of my life here in Barcelona and I must say that each one of them left a sign in his own way. I am very grateful to Gian for giving me the opportunity to start this adventure in his laboratory. Gian, thank you for being so spontaneous and empathic, always supportive and with a contagious smile.

I want to thank one of the most important people that characterized my experience in Barcelona. Dear Elias, your advices and guidance have been essential for my phD. I learnt a lot from your kindness, generosity, patience and charisma. Our relationship evolved into a beautiful friendship and I am very grateful for this.

A big thank goes to Alessandro for making my lab days full of laughs and gossip. I was lucky to share many years with you, even though you never heard my advices (*ma la smetti?*) and continued driving me crazy. Science was a lot more fun with people like you and now that we don't share a daily lab life I miss both of you every single day.

I would like to thank Fatima and the Gebauer lab for welcoming me into their lives for my last year at CRG, for all the advices and discussion. I wish we could have shared more time together! To Annagiulia, thank you for your support during these last months of coffee breaks and weekends spent in the lab before the thesis submission! To Rosario and Raffaella, thank you for the time spent

together in Barcelona, for all the advices and the celebrations we could share (especially the amazing dinners)!

I also want to thank the Tartaglia lab and all its members past and present: Nieves, Fernando, Alex, Magda, Michele, Riccardo, Davide, Mimma, Stefanie, Benedetta, Teresa, Iona, Andrea, Mireia. With all of you I share great memories, from lab retreats to dinners and parties. I have to admit that being the "last one" of Tartaglia Lab to remain in CRG has not been easy. I missed the daily adventures in our loud laboratory and every time I go there it feels so empty without you. Only now I realized how lucky we were!

My PhD journey started with the loss of two of the most important people in my life: my grandmothers, Nonna Emilia e Nonna Erminia. I would like to dedicate my thesis to them too, for all the memories we shared and for everything they thought me. I wish they could be still here to celebrate my PhD with me. I also want to thank my grandfather, Nonno Gino. He follows every academic accomplishment with huge enthusiasm and I am proud to say that he came to visit me during my international adventures here in Barcelona at the ripe old age of 90. He was so excited to visit CRG with me!

A big thank goes to my family, to my parents Mario and Patrizia and to my sisters Eugenia, Claudia and Erminia, for all the unconditional support in every choice I made in my career, from moving to Germany for my internship, through an Erasmus in Ireland and all

the interviews that finally led to a PhD here in Barcelona. I am lucky to be part of such a big and loud family, where we always support each other and I am happy I had the opportunity to show you Barcelona with my eyes during the many trips you organized to visit me. And even if I am not mentioning all of you, I want to say that I am proud to be part of such a great family!

During these years I had to face some health issues, from being completely blocked in a bed not being able to stand for more than 15 seconds to another major surgery with a lot (A LOT) of post-surgery complications. I could had never made it mentally without the help of my family and friends. To my 90's girls Jackie, Elena, Bea and Ati, I think I will always feel that I never thanked you enough for your presence during those times. You came to visit me every day in order to never make me feel alone, filling my day with all kind of activities, with dinners, movies. But besides this unfortunate part of my PhD journey, I shared with you the best part of my life in Barcelona. From improvised breakfasts with Jackie, to a 10 days stay in Elena's apartment before finding my own, including brunches with Bea and 3 years of flat-sharing with Ati. These acknowledgments will never be enough to show you how much all the memories I have with you in Barcelona mean to me.

Y quiero agradecer quien, hace dos años, ha llegado en mi vida inesperadamente. Samuel, gracias por el apoyo, por cuidarme, apoyarme y acompañarme en cada momento. Has sido la sorpresa mas grande de mi vida aquí en Barcelona y agradezco de verdad

habernos encontrado, por que tu llegada ha hecho que me conociera mas a mi misma, enriqueciéndome. Gracias por compartir cada día conmigo anécdotas, recuerdos y tradiciones de tu tierra y tu familia, que aun estén tan lejanos, siento cercanos. Después de todo lo que hemos vivido en tan solo 2 años juntos (incluyendo una pandemia mundial), se que podemos hacer frente a cualquier cosa ☺.

Abstract

During the last decades the non-coding portion of the genome became subject of intense research due to its contribution to the complexity of biological and pathological processes. CircRNAs are a class of non-coding RNAs able to influence gene expression during development and their biogenesis relies on the activity of several RBPs. Indeed, the Signal Transduction and Activation of RNA (STAR) family members Quaking and Sam68 have been shown to be involved in the biogenesis of circRNAs in specific biological contexts, but whether their contribution is extended to embryonic development is unknown. Our evidences suggest that Sam68 is involved in circRNA biogenesis during differentiation towards cardiac lineage, providing a new layer of regulation during embryonic development.

Preface

The key to decipher the complexity of biological and pathological processes relies on both coding and non-coding portions of the genome. The non-coding complement includes intronic regions, repeats, linear long RNAs, highly structured tRNAs and rRNAs, small RNAs and, more recently, circular RNAs (circRNAs). The latter became subject of intense research due to their ability to influence gene expression during development. Contrary to the increasing knowledge on circRNA-mediated regulation during brain and heart development, the complex landscape of circRNAs coordinating earlier developmental stages is still far from being characterized. Several RNA Binding Proteins (RBPs) have been shown to be implicated in the biogenesis of circRNAs, especially the Signal Transduction and Activation of RNA (STAR) family members Quaking and Sam68, but whether their contribution is extended to embryonic development is unknown. In order to fill this gap, we have identified the Sam68 protein interactome during different time points of mESCs differentiation. The GO analysis shows that 50% of the whole interactome is composed by RBPs and all steps of RNA metabolism are well represented. We demonstrate by Size Exclusion Chromatography (SEC) that Sam68 strongly interacts with another member of the STAR family (Slm2), suggesting that some of Sam68 activities might rely on cooperation between STAR members.

Since the canonical roles of Sam68 in RNA processing have been widely characterized, we focused on protein partners that could

extend Sam68 functions. Indeed, we show an interaction between Sam68 and Ilf2-Ilf3, essential circRNA biogenesis factors and regulators of pluripotency in mESCs. To further investigate our hypothesis, we performed a genome-wide circRNA analysis upon Sam68 KO condition and our results highlights a significant downregulation of circRNAs, suggesting the implication of Sam68 in the regulation of circRNA biogenesis during differentiation.

Moreover, in order to demonstrate the direct involvement of Sam68 in circRNAs formation, we investigated Sam68 binding profile using a bioinformatic tool developed in our laboratory that estimates the binding propensity between protein-RNA pairs and validating the predicted binding propensity by performing a transcriptome-wide mapping of Sam68 mRNA target using iCLIP2.

Our data revealed that Sam68 preferentially binds in intronic regions and that circRNA-forming transcripts are consistently present among its mRNA targets.

In conclusion, our evidences suggest an important contribution for Sam68 in circRNA biogenesis during differentiation, providing a new layer of regulation during embryonic development.

Resume

Durante las últimas décadas, la parte no codificante del genoma se ha convertido en objeto de intensa investigación debido a su contribución a la complejidad de los procesos biológicos y patológicos. Los CircRNAs son una clase de RNA no codificante capaces de influir en la expresión génica durante el desarrollo y su biogénesis se basa en la actividad de varias proteínas que ligan el RNA llamadas RNA-Binding Proteins (RBPs). Se ha demostrado que varias RBPs están implicadas en la biogénesis de los circRNAs, especialmente Quaking y Sam68, miembros de la familia STAR (Signal Transduction and Activation of RNA) pero se desconoce si su contribución se extiende al desarrollo embrionario. Mi proyecto de doctorado tiene como objetivo desenredar esta brecha definiendo el papel de Sam68 en la biogénesis de circRNAs durante la diferenciación de células madre embrionarias (mESCs).

Table of Contents

1. Introduction

1.1 RNA as a key-regulator in the cell

RNA was considered to function as the intermediary between DNA and protein and the concept of "one gene-one protein" has been adopted for decades. This concept has intensively evolved and nowadays, RNA is considered the regulation center of the eukaryotic cell. However, its complex contribution to the control of cellular functionality during every biological process is still far from being fully understood. The canonical RNA classification relies on what for many years was considered as the most important feature of this molecule: the capacity to encode for a protein. However, the extensive portion of the genome that does not have this ability revealed a hidden layer of regulation, suggesting that the key to decipher the complexity of biological and pathological processes was to investigate the non-coding RNA world[1].

1.2 Types of RNA: coding vs non-coding

The coding feature of RNA is normally associated to mRNAs. The life of mRNAs starts with the transcription of pre-mRNAs in the nucleus where successive steps of processing occur to transform the pre-mRNAs into mature messenger RNAs (mRNAs). This maturation phenomenon is orchestrated by complex machineries including the 5' capping, the 3' cleavage/polyadenylation and the pre-mRNA splicing (Figure 1). The mature mRNA is then exported

to the cytoplasm to reach ribosomes in order to be translated into proteins.

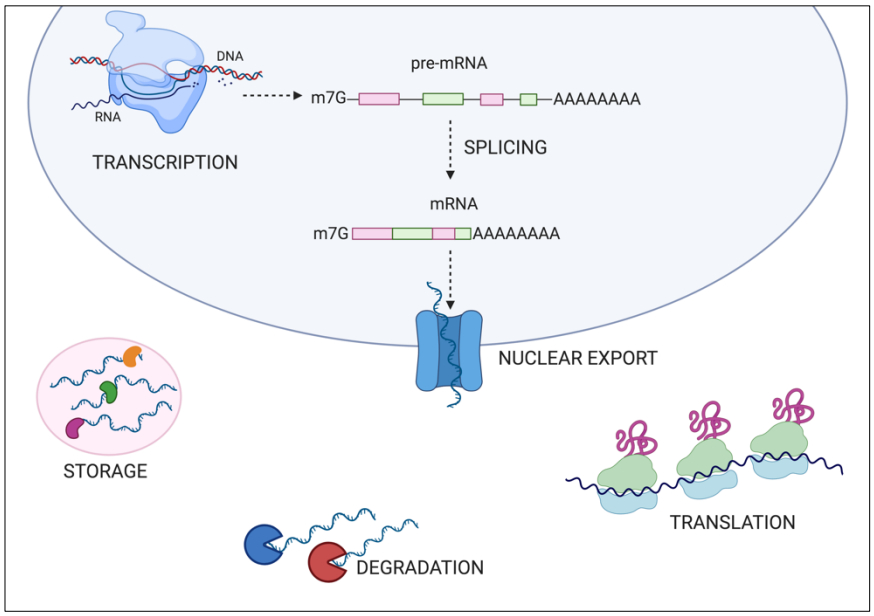


Figure 1. mRNA Life Cycle. Adapted from Ye et al., 2014.

While, the non-coding side of RNA has many dynamic sub-classes whose functions are in constant evolution together with the high throughput bioinformatic tools developed to analyze the genome[2]. Non-coding RNAs are classified in "long" and "small" according to their size and both are responsible for the control of gene expression adopting different mechanisms and have been described to participate in many processes. Figure 2 shows the most studied classes of non-coding RNAs[3].

Nevertheless, only a small portion of non-coding genome has been functionally characterized so far and a deeper knowledge of its roles in the coordination of gene expression is required in order to untangle

the RNA regulatory networks acting in both physiological and pathological contexts.

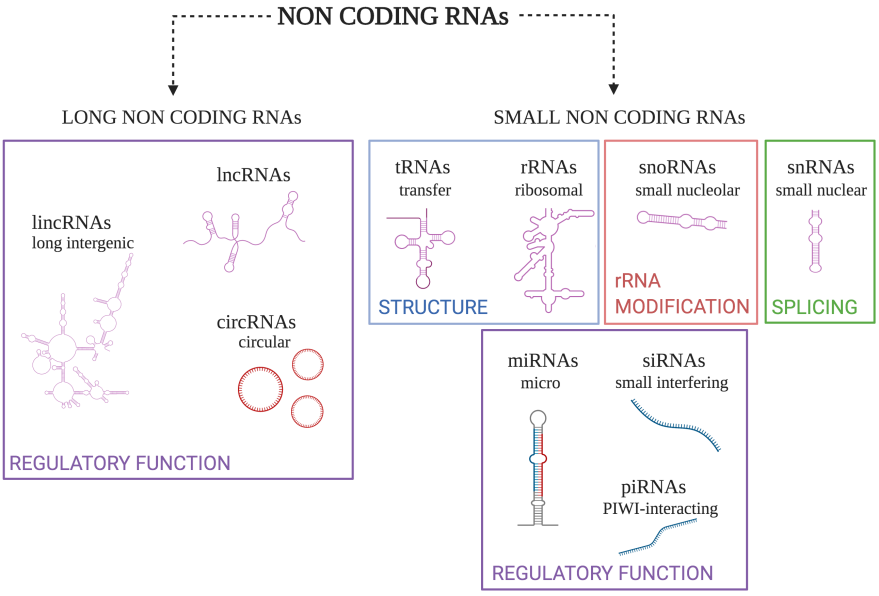


Figure 2. Most studied non-coding RNA classed. Adapted from Mattick et al., 2006

1.3 Small and long non-coding RNAs

In general, the class of small non-coding RNAs includes molecules shorter than 200 nucleotides. Well known examples of structural small non-coding RNAs (ncRNAs) with established molecular functions are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), key players in the translational process. Other essential small ncRNAs involved respectively in splicing and RNA modification mechanisms are small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs)[3].

MicroRNAs (miRNAs) are also included in this class and are crucial regulators of animal development, cell differentiation and homeostasis[4]. Indeed, depletion of the miRNA biogenesis factors Drosha and Dicer leads to embryo lethality[5]. Notably, the different categories of nc-RNAs are transcribed by different RNA polymerases: RNA polymerase I (RNAPI) transcribes rRNA genes, RNA polymerase II (RNAPII) transcribes mRNAs, miRNAs, snRNAs, and snoRNA genes, and RNA polymerase III (RNAPIII) transcribes tRNAs and 5S rRNA genes[6].

Long non-coding RNAs are RNA molecules longer than 200 nucleotides responsible to regulate gene expression and signaling pathways in various biological functions and disease processes which became objects of intense research in the last two decades[7]. One of the most studied is the lncRNA Xist. During female mammalian development, Xist mediates X-chromosome inactivation (XCI) in order to achieve dosage compensation of sex chromosomal genes between females (XX) and males (XY). Xist promotes gene silencing, recruiting chromatin modifying factors and coordinating a structural reorganization of the X chromosome[8,9].

Another example of how lncRNAs are important in pathological context is the LncRNA HOTAIR which has increased expression in cancer cells and induces metastasis formation by interacting with the Polycomb repressive complex 2 (PRC2) in order to cause cell invasion[10].

Recent evidences revealed the existence of another class of lncRNAs named circular RNAs, in virtue of their circular structure, those circRNAs might be important in the regulation of gene expression[11].

1.4 Role of ncRNAs in development and pathologies

A normal embryo development is the result of precise spatio-temporal activation of essential genes in embryonic cells. RNA-sequencing data obtained from pre-implantation embryos revealed that some lncRNAs are associated with the early stage of development. In mice, promoter-associated noncoding RNAs (pancRNAs) modulates the expression of their cognate gene during the zygotic activation. An example of this mechanism is the pancRNA Il17d (Interleukin 17d). When decreased, Il17D causes downregulation of its cognate mRNA, leading to embryo lethality[12]. On the other hand, several single-cell RNA sequencing experiments have shown that lncRNAs have a development-stage specific expression pattern. Some of the lncRNAs are expressed in a very narrow time-frame or in specific cells[13]. During preimplantation development, the lncRNA *Xist* promotes the X-inactivation in 4-cell stage embryo[14] and *Kcnq1ot1*, a paternally expressed nc-RNA expressed since 2-cell stage, regulates the establishment of imprinting in Kcnq1 domain[15]. Moreover, a recent study identified the lncRNA *Handsdown* as an important regulator of cardiac gene program in early mouse development[16].

LncRNAs have also been associated with cancer, neurological disorders and cardiovascular diseases. MIAT lncRNA is involved in myocardial infarction and retinal development, while Mhy7-as regulates the expression ratio of the sarcomeric proteins Mhy6 and Mhy7[17].

1.5 The rising star of the RNA family: Circular RNAs

Circular RNAs (circRNAs) are a class of long non-coding RNAs with a covalently closed ring structure expressed in a tissue-specific and cell-specific manner[18]. They have been detected for the first time by electron microscopy 40 years ago[19] when the RNA field considered them byproducts of aberrant splicing events and later on they have been identified as abnormally spliced transcripts called "scrambled exons"[20]. During the last years, circRNA finally became object of high interest because of their potential to be considered as disease biomarkers due to their specific expression pattern[21].

1.6 RNA-binding proteins and CircRNA biogenesis

RNA Binding Proteins (RBPs) are associated factors that escort the RNAs throughout their lifetimes. Some of the RBPs remain stably bound while others are subject to dynamic exchange. RBPs are able to recognize and bind to specific RNA sequences and structures through their RNA binding domains (RBDs) (Figure 3). In humans, more than 40 different RBDs are found, providing an exhaustive regulation of diverse pathways. Two RBDs, the RNA recognition motif (RRM) which is represented in nearly 500 different human genes and the heterogeneous nuclear ribonucleoprotein K-homology (KH) domain, are present in almost all RBPs for single stranded RNA recognition[22] (Table 1). Other motifs as the double-stranded RNA binding domain (dsRBD), zinc fingers, RGG boxes, and the Pumilio homology domain in PUF proteins are also found[23,24].

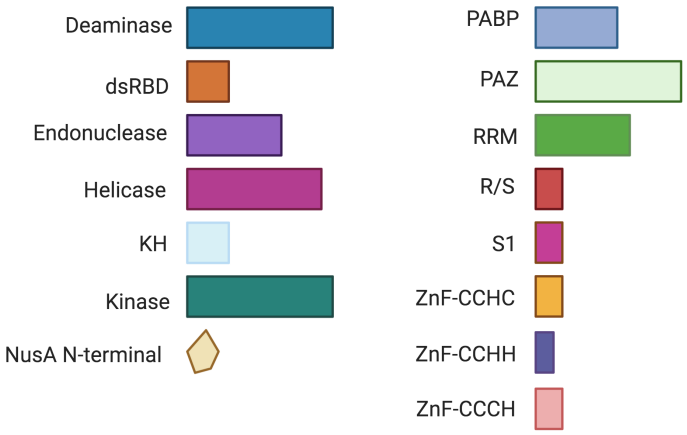


Figure 3. Most common RNA binding domains. Adapted from Lunde et al., 2007.

Domain	Topology	RNA Recognition Surface	Protein-RNA interactions
RRM	αβ	Surface of β–sheet	Interacts with 4 nucleotides of ssRNA through stacking, electrostatics and hydrogen bonding
KH	αβ	Hydrophobic cleft formed by variable loop between β2 and β3 and GXXG loop; Type II: Same as type I, except variable loop is between α2 and β2	Recognizes 4 nucleotides of ssRNA through hydrophobic interactions between non-aromatic residues and the bases; sugar-phosphate backbone contacts from GXXG loop, and hydrogen bonding to bases

Table 1. Topology, RNA Recognition Surface and Protein-RNA interactions of RRM and KH domain. Adapted from Lunde et al., 2007.

Recent RNA interactome capture experiments have shown that the human genome encodes more than 1500 RBPs that are able to interact with RNAs via either canonical or non-canonical RNA Binding Domains[25]. RBPs play a crucial role in the generation of circular RNAs that are obtained from pre-mRNA transcripts through a type of alternative splicing named back-splicing (Figure 4). This process relies on the canonical spliceosomal machinery[11] and the canonical splice sites[26].

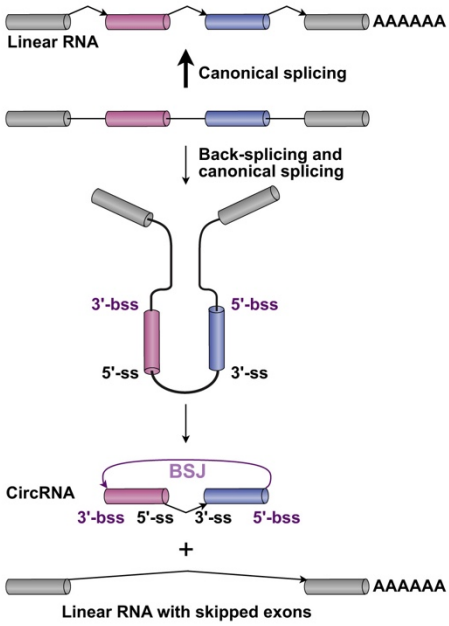


Figure 4. The mechanism of back-splicing. Adapted from Li et al., 2018 Mol. Cell

Generation of circRNAs occurs when a downstream splice donor site of one exon is covalently ligated to the splice acceptor site of an upstream exon, creating the typical signature of circRNAs, the back-spliced junction (BSJ). Advances in genome-wide approaches helped to untangle the mechanism of exon circularization, providing evidences that it depends on flanking intronic complementary sequences[27]. However, the only presence of the same cis-elements is not able to explain how circRNA expression is modulated at the same loci across cell types and tissues, suggesting the presence of other trans-acting factors influencing circRNA processing[21]. Indeed, biogenesis of circRNAs has been reported to be enhanced by RNA

binding proteins (RBPs) with known roles in regulation of alternative splicing (e.g. Muscleblind, FUS, Quaking) (Figure 5): in flies and in humans, the splicing factor Muscleblind regulates the production of a circRNA (circ-mbl) deriving from its own pre-mRNA transcript.

This regulatory function in circRNA biogenesis depends on the presence of functional MBL binding sites in the flanking intronic sequences[11]. Another RNA binding protein which has been shown to control circRNA biogenesis in in-vitro derived motor neurons by influencing back-splicing events is FUS, a known splicing regulator whose mutations are causally linked to amyotrophic lateral sclerosis (ALS). FUS regulates circRNA formation binding to introns flanking circularizing exons. Notably, alterations of FUS nuclear levels directly correlate with circRNAs abundance[28].

Furthermore, the member of the STAR family Quaking influences formation of circRNAs during the epithelial to mesenchymal transition, a cell transformation which is essential in embryo development, wound healing and invasive metastasis[29]. Quaking binds to introns flanking circularizing exons, facilitating the formation of the BSJ by dimerization[30]. Moreover, circRNA production has been suggested to be crucial upon viral infection, where immune response factors such as Ilf3 influence circRNA biogenesis by stabilizing intronic RNA pairs in the nucleus and binding to circRNP in the cytoplasm.

During the innate immune response against viral infection, Ilf3 isoforms (90 kDa and 110 kDa) are promptly released from circRNPs complexes in the cytoplasm allowing their binding to viral mRNAs in order to inhibit viral replication[31].

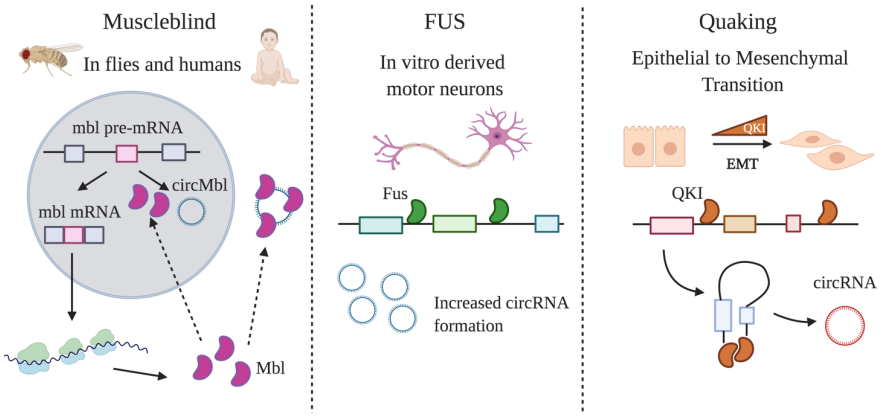


Figure 5. Muscleblind, FUS and Quaking and their involvement in circRNA biogenesis. Adapted from Ashwal-Fluss 2014, Errichelli et al., 2017 and Conn et al., 2015.

1.7 CircRNAs properties and functions

CircRNAs control gene expression through distinct molecular mechanisms[18] (Figure 6).

1.7.1 CircRNAs and Splicing

Even though the back-splicing reaction is less favorable, circRNA biogenesis can compete with the canonical splicing, causing lower levels of the linear mRNA counterpart[11,27,32]. A class of circRNAs (EIciRNAs) has been found to associate with the Pol II machinery. They present introns retained between the circularized exons, mostly localize in the nucleus, where they associate with U1 snRNP enhancing the transcription of their parental gene[33]. Interestingly, it has been reported that circRNAs can bind to their cognate locus,

forming a RNA:DNA hybrid that cause transcriptional pausing, coordinating the production of alternative mRNA isoforms[34].

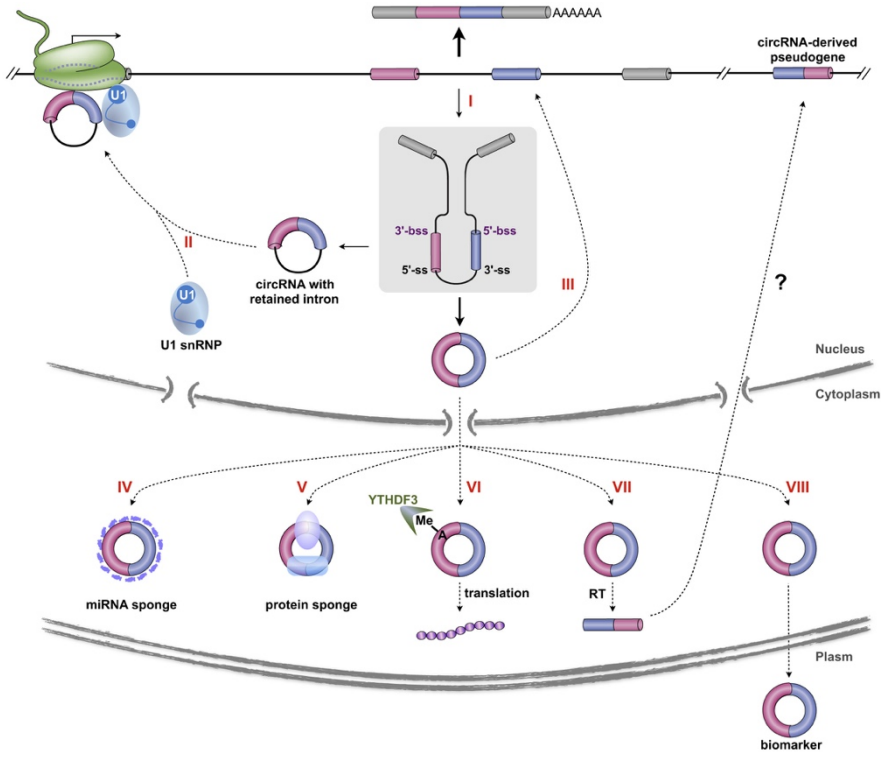


Figure 6. Potential roles of circRNAs in gene expression.
Review from Li et al., 2018.

1.7.2 CircRNAs and molecular sponges

Furthermore, recent evidences revealed another important role for circRNAs in acting as miRNA sponges. It has been proposed that miRNAs are sequestered by circRNAs through their complementary binding sites, therefore modulating miRNA-dependent gene expression regulation. The antisense of Human CDR1 (cerebellar degeneration-related protein 1 transcript) is a highly conserved circRNA in the mammalian brain which has over 60 binding sites for

mir-7. Upon CDR1as reduction, expression of mRNAs containing miR-7 binding sites is decreased, suggesting that CDR1as competes against miR-7 targeting influencing gene expression[35,36].

In addition to act as miRNA sponges, circRNA could sequester RBPs to form circ-RNPs complexes in the cytoplasm. NF90 and NF110 (known as well as ILF3 isoforms) are released from circRNPs complexes in the cytoplasm during the innate immune response against virus infection, allowing their binding to viral mRNAs in order to prevent viral replication[31]. The identification of circRNAs that could act as RBP sponges has been enhanced by bioinformatic tools that analyze RPB binding sites enrichment in circRNA sequences. The bioinformatic tool CircInteractome is able to identify potential RBP sponges by integrating available databases of CLIP, miRNA and RBP datasets. For example, the circRNA *hsa_circ_0024707* has 85 predicted binding sites for AGO2 and could potentially function as AGO2 super-sponge. FMRP and HuR could be sequestered by the mature circRNA *hsa_circ_0000020* given its multiple binding sites[37].

1.7.3 CircRNAs and Translation

Unlike linear mRNA molecules, circRNAs lack a 5' end 7-methylguanosine (m7G) cap structure and a 3' poly(A) tail, which would suggest they can be translated into peptides only in a cap-independent manner through the usage of IRESs (Internal Ribosome Entry Site). Strikingly, recent studies performed by ribosome foot-printing demonstrated that a subset of circRNAs is associated to ribosomes and translated into peptides in a cap-independent

manner[38]. Furthermore, an interesting example of protein-coding circRNA is the *circ-ZNF609*, involved in myoblast proliferation and known to associate with heavy polysomes. Indeed, the 5' UTR of *circ-ZNF609* can promote an IRES-dependent translation[39].

Additionally, Given the many properties of circRNAs, many efforts were made to discover a possible role in shaping genome architecture, being source of pseudogenes. Taking advantage of their unique back-spliced junction, a computational pipeline (CIRCpseudo) was developed and was able to identify 33 circRNA-derived pseudogenes in the mouse genome[40], adding another level of circRNA-mediated regulation.

Finally, the intrinsically stable ring structure of circRNAs gives them the capacity to avoid canonical RNA decay machineries and the stability to be transported to extracellular fluids by exosomes[33], a feature that converts them in promising biomarkers for several diseases.

1.8 CircRNAs during development

Single cell RNA-transcriptome analysis identified 2891 circRNAs expressed in mouse preimplantation embryos with a dynamic expression pattern during this developmental process[41]. A later study performed in the same group detected 10,032 circRNAs from 2974 hosting genes in human pre-implantation embryos, suggesting a possible role for circRNAs in the regulation of gene expression during mammalian early embryonic development[21,42].

Notably, *circBIRC6* and *circCORO1C* have been functionally linked with the maintenance of pluripotency in human embryonic stem cells.

CircBIRC6 acts as a sponge for miR-34a and miR-145, avoiding the suppression of the pluripotent factors NANOG, OCT4, and SOX2[43,44].

Moreover, recent evidences have demonstrated that circRNAs are important during brain and heart development. Indeed, this class of lncRNA is enriched in brain tissues and it has been shown that a subset of circRNAs is differentially expressed in the mouse hippocampus at different developmental stages[45]. During cardiomyocytes differentiation, circRNA expression changes dynamically over the time-course and the GO terms analysis revealed that this subset of circRNAs mapped to genes involved in heart development and anatomical structural development[46]. *CircSLC8A1-1* and *circTTN-275* figure among those circRNAs with increasing expression over the differentiation time-course. The gene Slc8A1 encodes for a Na^{+}/Ca^{++} exchanger essential for the cardiomyocytes electrophysiological features and the circRNA derived from exon 2 is the most abundant circRNA in human hearts. Titin (TTN) is a large abundant component of the striated muscle and has the largest number of circRNA isoforms in the heart (402 isoforms). Mutation of the TTN locus are linked to dilated cardiomyopathy[47]. Although an increasing amount of information regarding circRNA-mediated regulation during brain and heart development has been described during the last years, the complex landscape of circRNAs coordinating earlier developmental stages is still far to be fully characterized.

1.9 CircRNAs in diseases

A growing number of evidences is demonstrating that misregulation of circRNAs is functionally linked with several human diseases such as cancer, neurological disorders and heart conditions[44]. Figure 7 represents the characterized circRNAs in the different physiological systems. In non-small cell lung carcinoma, *circRNA_100876* correlates with tumor cell growth and progression, providing a marker for tumor stage determination.

CircRNA *ciRS-7* acts as a miRNA sponge for miRNA-7 in Alzheimer's disease, avoiding the downregulation of ubiquitin-protein ligase A (UBE2A), an important phagocyte of amyloid peptides plaques. In Parkinson Disease, *circCDR1* is a negative regulator of mir-7, a direct inhibitor of the synuclein protein[48].

Moreover, other circRNAs are also associated with important cardiac pathologies. A circRNA produced from the lncRNA ANRIL is associated with arteriosclerosis risk and the *circRNA HRCR* acts as a mir-223 sponge in order to inhibit heart failure and cardiac hypertrophy[46]. The circRNA *hsa_circ_0124644* is highly expressed in the peripheral blood of cardiovascular artery disease (CAD) patients, suggesting its use as a potential specific biomarker for CAD.

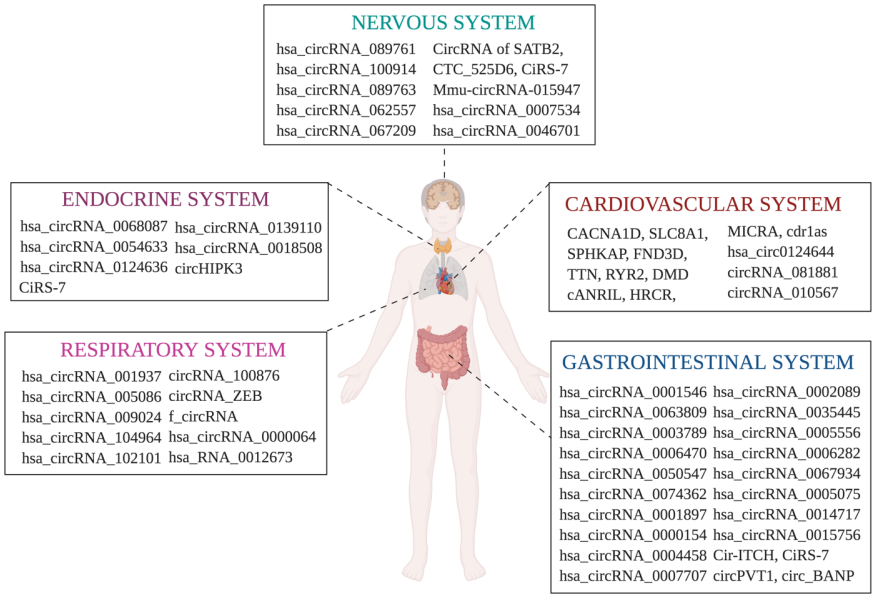


Figure 7. Landscape of characterized circRNAs in nervous, cardiovascular, gastrointestinal, endocrine and respiratory systems. Adapted from Lee et al., 2019.

1.10 RBPs during early stage of development

While the transcriptional and epigenetic regulatory network controlling the establishment of different germ layers has been widely characterized[49–51], less attention has been directed towards post-transcriptional control of the pluripotent state. In this scenario, the idea that RBPs might play an important role is gaining more interest.

1.10.1 Alternative splicing and pluripotency

An example of how an RBP and alternative splicing can influence pluripotency regards Muscleblind-like RNA binding proteins MBNL1 and MBNL2 in the AS regulation of the transcription factor (TF) Foxp1. This protein has two isoforms that differ for the DNA binding motif and only one is able to enhance transcription of pluripotency-associated genes[52]. Indeed, MBNL1 and MBNL2 negatively regulate cassette exon alternative splicing events that define the switch between differentiated cells and embryonic stem cells, suggesting a central role for splicing regulators in pluripotency maintenance[53].

In human ESCs and neural progenitors, RBFOX contributes to self-renewal and lineage specification acting as an upstream splicing regulator of many splicing factors. RBFOX depletion in hESCs caused cell death, while its absence in neural progenitors did not cause any phenotype, suggesting a different set of targets for the two cell types[54].

1.10.2 3'UTRs and pluripotency

Another crucial element that define the post-transcriptional and post-translational regulatory network controlled by RBPs in the switch between pluripotency and differentiation is the length of the 3' UTR of a transcript. During pluripotency maintenance it has been shown that proximal polyadenylation sites are favored compared to the distal ones, resulting in shorter 3'UTRs in ESCs compared to differentiated cells. Notably, the RBP Fip1 promotes the choice of a shorter polyadenylation site which implies fewer binding sites for microRNAs and other RBPs, increasing mRNA stability[55]. Poly(A) tail-length regulation represents another critical post-transcriptional mechanism that modulates the choice between pluripotency maintenance and differentiation. CCR4-NOT complex regulates the pluripotent state by promoting deadenylation and degradation of differentiation-related genes[56].

1.10.3 m^6A and pluripotency

Post-transcriptional RNA modifications, such as N^6-methyladenosine (m^6A), are another important tool that influence pluripotency. The most prevalent chemical modification enriching stop codons at 3'UTRs in eukaryotic mRNAs is m^6A. The fate of a modified mRNA changes according to the position of the m^6A modification: if enriched at the transcriptional start site, it promotes translation, whereas it decreases mRNA stability when accumulated in internal positions along the transcript[57]. In the context of pluripotency, depletion of the methyltransferases responsible for

m^{6}A mRNA methylation (Mettl3 and Mettl14) impairs self-renewal capacity in mESCs[58].

1.10.4 RNA export and pluripotency

In addition to alternative splicing, RNA modifications and the choice of alternative polyadenylation sites, RNA's fate is also influenced by nuclear transport. RBPs shuttle RNA molecules between nucleus and cytoplasm in order to reach the translation machinery. Thoc2 and Thoc5 are members of the THO complex and their depletion has been shown to lead to nuclear accumulation of a subset mRNAs essential for pluripotency maintenance, including Nanog, Sox2, Klf4, and Esrrb[59].

1.11 RBPs in heart development

During embryonic development, the heart is the first organ becoming functional around day 7.5 in the mouse and at the third week in the human embryo[60].

The development of the heart is a complex process tightly regulated by different players. Among them, RNA binding proteins have been shown to participate in every step of heart morphogenesis, from the establishment of cardiac lineages to maturation of the heart after birth[61].

RNA binding motif (RBM) proteins are a subgroup of loosely related RRM-containing proteins involved in splice site selection and non-sense mediated decay during cardiomyocyte differentiation and myofibril development. Indeed, RBM24 is highly expressed in both human and mouse embryonic stem cells upon differentiation into

cardiomyocytes[62,63]. RBM20 regulates the splicing of the Titin gene, crucial for the biomechanical properties of the heart. Mutations of the Titin locus have been linked to cardiomyopathies characterized by altered myofibril structure and function. Moreover, evidences have shown that RBM20 is essential for the production of a subset of circRNAs originating from the I-band of the Titin gene: in presence of functional RBM20, the Ig repeats, N2A, and PEVK regions contained in the I band of Titin are spliced out by RBM20, providing the substrate for circRNA formation. However, mutations or absence of RBM20 induce the inclusion of these regions in the linear transcript, preventing circRNA formation. Moreover, the analysis of RBM20 binding sites in the identified titin circRNAs showed a 5-fold increase compared to the control. These evidences confirm that splicing-related RBPs can exert an important role in heart development through circRNA metabolism regulation[64].

Notably, the Signal Transduction and Activation of RNA family has also been linked to heart development: Quaking mutations are lethal for the mouse embryo due to Quaking functions in heart and vasculature development[65] and we recently observed a defect in the electrophysiological features of cardiomyocytes in absence of Sam68, which led to the discovery of a new role for Sam68 in the regulation of the expression and the alternative splicing of cardiac-related transcripts (Dasti et al., submitted).

1.12 STAR family

The Signal Transduction and Activation of RNAs (STAR) family is composed by five highly conserved RBPs[66]: Src associated mitosis

68 (Sam68, also known as Khdrbs1), Quaking (QK), Khdrbs2 (Slm1), Khdrbs3 (Slm2) and Splicing factor 1 (SF1). Their conserved HNRNPK Homology (KH) domain located at their N-terminus part is able to bind RNA and it is flanked by two Quaking domains (QUA1 and QUA2), responsible of homodimerization[67,68]. All the members of the STAR family are involved in the regulation of the RNA metabolism, controlling alternative splicing, translation efficiency, transport, RNA stability and localization within the cell. Moreover, STAR RBPs are subjected to both post-transcription and post-translation modifications leading to the generation of several isoforms and fine-tuning their subcellular localization and RNA binding affinities.

1.13 Sam68

The most well characterized member of the STAR family is Sam68s (SRC-associated in mitosis 68 KDa), known as well as Khdrbs1 (KH domain containing, RNA binding, signal transduction associated 1)[69,70]. Sam68 is a 443 amino acids protein, which encode for a protein of 68KDa. It is ubiquitously expressed and its localization in the cell is prevalently nuclear. Its protein structure includes a proline-rich region, a SH3 domain, WW-binding sites, RGG boxes and a bipartite nuclear localization signal (NLS) flanked by tyrosines strings in the C-terminal[71,72]. Its RNA binding ability is due to the presence of a KH domain, flanked by QUA domains, responsible for homodimerization. Sam68 binds A/U rich motifs (UAAA and UUAA), identified by SELEX experiments[70,73] and it controls many aspects of RNA metabolism in response to signaling pathways, from

transcription to alternative splicing, but also translation and nuclear export (Figure 8).

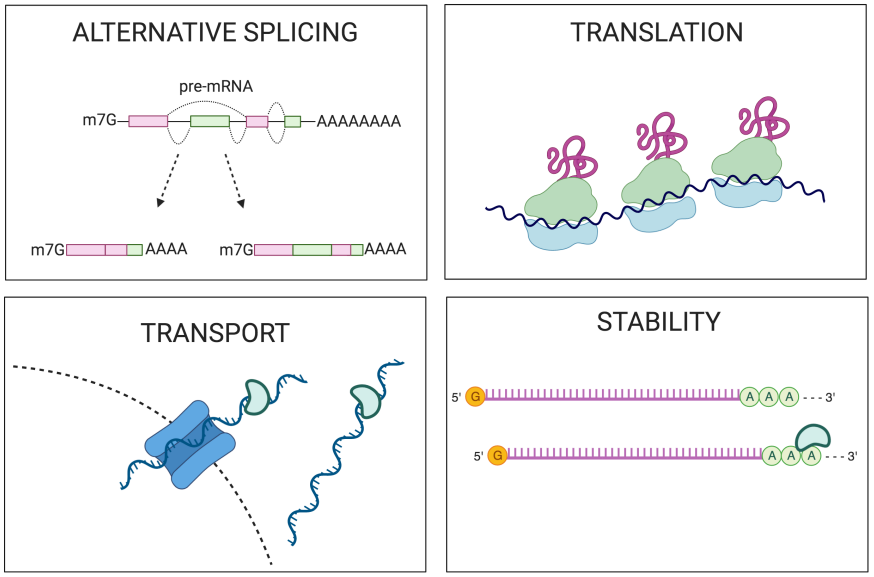


Figure 8. Sam68 Molecular Functions. Adapted from Paronetto et al., 2007; Modem 2005 and Li et al., 2017.

In fact, Sam68 regulates alternative splicing of CD44[74], a cell surface molecule involved in cell adhesion, proliferation and invasive cell migration. The apoptosis factor BCl-x is also regulated at alternative splicing level by Sam68[75]. Surival of Motor Neurons (SMN) transcript is linked to Spinal muscular atrophy (SMA) and its alternative splicing is regulated by Sam68, which binds to SMN2 pre-mRNA inducing the recruitment of the splicing repressor hnRNP A1[76,77]. Another interesting alternative splicing target of Sam68 is cyclin D1: in prostate cancer cells Sam68 enhances the splicing of the isoform with higher oncogenic potential[78].

During development, Sam68 has been shown to be involved in neurogenesis, adipogenesis and spermatogenesis and, as previously described in this paragraph, many of its functions are related to its interactors, both proteins and mRNA targets (Figure 9).

During neurogenesis Sam68 controls the splicing of Neurexin, a presynaptic protein in the mouse brain[79,80].

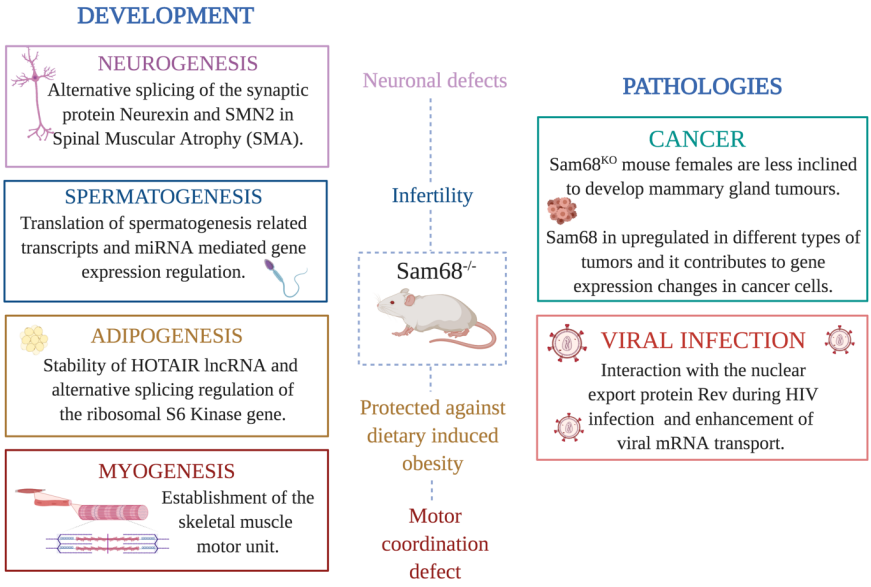


Figure 9. Sam68 Functions in development and pathologies. Adapted from Pagliarini et al., 2015; Paronetto et al., 2009; Li et al., 2017; De Paola et al., 2020; Frisone et al., 2015; Modem 2005.

Sam68 is also involved in the adipogenic differentiation by enhancing HOTAIR lncRNA stability and through the alternative splicing regulation of the ribosomal S6 Kinase gene (S6K1)[81].

Moreover, Sam68 interacts with miRNA processing proteins Drosha and Dicer and it is involved in miRNA-mediated gene expression regulation in spermatogenesis, influencing miRNA expression[82,83].

Sam68 has been found to interact with the translation factor eIF4F and regulates the translation of transcripts associated with male germ cell differentiation. It enhances the expression of spermatogenesis related transcripts, which reflect the infertility of $Sam68^{KO}$ mice phenotype[84].

The role of Sam68 in nuclear export is described from retroviral mRNA processing studies[85]. During HIV infection, the nuclear export protein Rev mediates translocation of viral RNAs and Sam68 has been shown to interact with Rev, synergizing its function. Indeed, Sam68 reduced expression resulted in decreased Rev function and inhibition of HIV production.

1.14 STAR family during development

$Sam68^{-/-}$ and Quaking mutant mice models are available, but less is known about the other members of the STAR family. $Sam68^{-/-}$ mice are infertile due to the inability at enhancing the translation of spermatozoa related transcripts[84]. and the females are less inclined to develop mammary gland tumours[86]. Moreover, a motor coordination defect is observed[87] and recent evidences showed that Sam68 in important for the establishment of the skeletal muscle motor unit[88].

Quaking mutant mice grow with a severe defect in the myelination, while Sam68:$Slm1^{DKO}$ mice show defects in cerebellar morphogenesis[89]. Slm1 and Slm2 are important regulators of neuronal alternative splicing expressed in specific tissues (testis, brain). There is a mutually exclusive expression pattern when Slm2 is deleted in vivo: the absence of Slm2 controls alternative splicing of Slm1, increasing the stability of Slm1 mRNA[90].

Furthermore, Slm2 has a role in cortical network activity and influences splicing patterns of the presynaptic protein Neurexin in the mouse brain[80]. Both Sam68 and Slm2 are able to bind Neurexin pre-mRNA, but only SLM2 is proficient to regulate its splicing depending on the binding sites density[91].

However, despite the known contribution of STAR RBPs during muscle, vasculature and brain development there is still no information regarding their role during the early stage of development, when the formation of the embryo occurs.

1.15 STAR proteins and circRNAs

Besides the canonical roles of STAR proteins in RNA processing previously described, new evidences shed light on a new layer of regulation controlled by two members of the STAR family, Quaking and Sam68.

Quaking binds to intronic regions and it is known to promote circRNA biogenesis during the epithelial to mesenchymal transition (EMT), taking advantage of its dimerization process that brings in close proximity the circularizing exons (Figure 10). Interestingly, the insertion of Quaking binding sites into linear RNA is able to induce exon circularization[30].

Recent studies demonstrated that the other member of the STAR family Sam68 is also involved in the pre-mRNA circularization process. Sam68 binds in proximity of *Alu* sequences in the Survival of Motor Neurons (SMN) locus, linked to Spinal Muscular Atrophy (SMA) disease, triggering biogenesis of circRNAs.

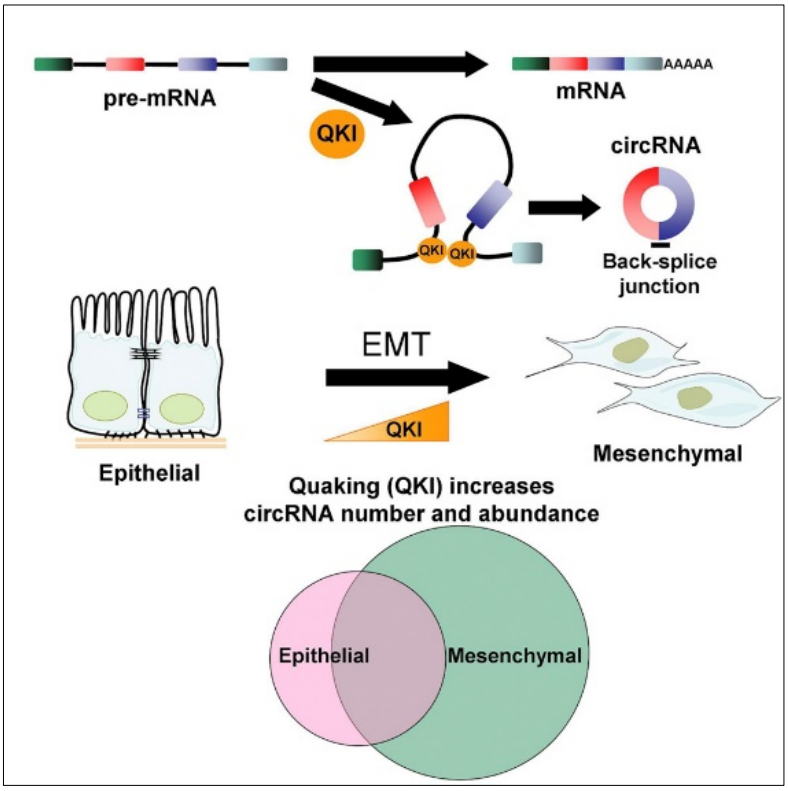


Figure 10. QKI in circRNA biogenesis, from Conn et al., 2015

The ability of Sam68 to promote circularization is due to the fact that the SMN locus presents high density of *Alu* sequences in inverted orientation (IR*Alus*), suggesting that Sam68 might use this signature mark to promote circRNA production[92]. However, whether Sam68 might control circRNA biogenesis by taking advantage of IR*Alus* in other physiological contexts has to be further investigated.

1.16 Sam68 in heart development

Recently, in our laboratory we identified Sam68 as a new player in early heart development. When Sam68 is depleted in mouse Embryonic Stem Cells (ESCs), a defect in the electrophysiological features of cardiomyocytes is observed. The transcriptomic analysis revealed an up-regulation of cardiac-related transcripts upon Sam68 KO, in addition to an alteration of alternative splicing (AS) events. Among the transcripts altered at AS level we found Slc8a1, a sodium calcium pump highly expressed during heart development and known to generate circRNAs (Dasti et al., submitted). Interestingly, we found that Sam68 directly regulates the expression of GATA4, a master regulator of mesodermal development, the embryonic germ layer which gives rise to cardiac tissues.

Together with QKI, Sam68 regulates the expression and the alternative splicing of cardiac-related transcripts, however, the mechanism underlying the involvement of Sam68 in heart development are not fully understood yet and, given the increasing amount of informations regarding circRNA functions in heart development and the recent discovery that Sam68 has a role in pre-mRNA circularization, we wondered if there could be any link between Sam68, circRNAs and heart development. Notably, one of the tools used to study molecular mechanisms regulating the early stage of development are mouse Embryonic Stem Cells (mESCs).

1.17 A model to study the early stage of development

Embryonic Stem Cells (ESCs) derived from the inner cell mass (ICM) of the mammalian blastocyst have two main characteristics

that define them: self-renewal and pluripotency. In fact, they have the ability to give rise to an identical daughter cell and, under specific stimuli, to differentiate into all cell types of the three germ layers (endoderm, ectoderm and mesoderm)[93].

To better investigate the mechanisms regulating the early stage of development, ESCs are widely used. Interestingly, under specific culture conditions, ESCs can form Embryonic bodies (EBs), complex round-shaped 3D structures in which the cells differentiate spontaneously into the three germ layers, recapitulating the embryo development and the differentiation progress[94,95]. Therefore, EBs represent a robust in vitro model to investigate the mechanisms that coordinate the pluripotency state.

2. Objective of the study

During the last decades the non-coding portion of the genome became subject of intense research due to its contribution to the complexity of biological and pathological processes[48]. CircRNAs are a class of non-coding RNAs able to influence gene expression during development and their biogenesis relies on the activity of several RBPs. Indeed, the Signal Transduction and Activation of RNA (STAR) family members Quaking and Sam68 have been shown to be involved in the biogenesis of circRNAs in specific biological contexts, but whether their contribution is extended to embryonic development is unknown.

My PhD project aims to untangle this gap by defining the role of Sam68 in circRNA biogenesis during mESCs differentiation. In order to achieve that, we analyzed the Sam68 protein interactome and focused on protein interactors that could provide information about uncharacterized functions of Sam68. Then, we analyzed the differential expression of circRNAs in Sam68$^{-/-}$ cells and defined its mRNA targets to better understand its specific contribution to circRNA metabolism.

3. Methods

3.1 Cell lines

The project was carried out with E14 mESCs. Cells were plated at a density of ~1,5X10^{4}cells/cm^2 on gelatin pre-coated culture dish and were passed every two days after dissociation with Trypsin-EDTA 0,05% (2530054, Gibco) and maintained in Knock-Out DMEM with 15% Embryonic Stem Cells qualified FBS (16141079, Gibco), 1% Penicillin-Streptomycin (15140122, Gibco), 1% L-Glutamine (25030081, Gibco), 1% Sodium Pyruvate (11360070, Gibco), 1% NEAA (11140068, Gibco), β-mercaptoethanol (31350010, Invitrogen) and leukemia inhibitory factor LIF (ESG-1106 Millipore).

3.2 Generation of GFP-Flagged cell line by CRISPR-Cas9

The KI E14 cell line for Sam68 was realized by the Tissue Engineering Facility of the Center for Genomic Regulation in collaboration with the facility of Biomolecular Screening and Protein Technology. E14 cells were transfected with pSpCas9 459-sgRNA63 which codifies for Cas9, the sgRNA and the HDR linearized donor vector for the insertion of GFP sequence at the N-terminus of Sam68. containing the GFP tag. The transfected cells were selected with Puromycin and checked under the microscope to observe GFP fluorescence. The selected clones were amplified, tested by PCR and sequenced to verify GFP integration in the precise desired position.

3.3 Generation of KO cell line by CRISPR-Cas9

The KO E14 cell line for Sam68 and Quaking was realized by the Tissue Engineering Facility of the Center for Genomic Regulation in collaboration with the facility of Biomolecular Screening and Protein Technology. Briefly, E14 WT cells were transfected with the bicistronic plasmid pSpCas9 PX459 codifying for both the Cas9 and the sgRNA targeting the first exon of either Sam68 or Quaking. Afterwards, the transfected cells were selected for the puromycin resistance. The Cas9 activity of each sgRNA was then verified by T7 exonuclease assay. After the puromycin selection the cells were FACS sorted in order to get colonies derived by a single clone. Consequently, the clones were singularly sequenced and the *bona fide* KO ones selected in order to carry out the research project.

3.4 Embryonic Body Assay

E14 cells were dissociated with Trypsin-EDTA 0,05% (2530054, Gibco) and diluted in Knock-Out DMEM with 10% Embryonic Stem Cells qualified FBS (16141079, Gibco) supplemented with 1% Penicillin/Streptomycin, 1% L-Glutamine, 1% Sodium Pyruvate, 1% NEAA to a concentration of 50`000 cells/ml and plated on culture dish in hanging drops to obtain a final concentration of 1000 cells/drop. Embryonic bodies were collected at Day 3 and gently moved to a low-attachment plate. KO-DMEM 10% FBS Medium was changed every two days until Day 10.

3.5 Protein Extraction, SDS page and WB

A variable amount of Lysis buffer (150mM NaCl, 25mM Tris pH 7.4, 5mM EDTA, 1mM DTT, 0,5% NP40, protease inhibitor) was added to cell pellet according to its size followed by disruption of cell membranes by either pipetting or vortexing and incubation on ice for 10 minutes. To assure membrane disruption, lysates were homogenized by using 1 ml syringe. The whole operation was repeated three times in total and the lysates were centrifuged at maximum speed (>13.000 rpm). Pellet was discarded and the supernatant containing the proteins was quantified by measuring protein concentration at the spectrophotometer as absorbance at 595nm in Bradford reagent.

A defined amount of protein was boiled at 98°C in 4x LDS Protein Buffer added with antioxidant (DTT) and samples were loaded on NUPAGE 4%-12% bis-tris gel (NP0321BOX, Invitrogen) immersed in MOPS buffer (NP0001, Invitrogen). The run was performed at 150V for 1 hour and the proteins were dry-blotted on a nitrocellulose membrane with the iBlot™ 2 NC transfer stacks (IB23002, Invitrogen) using the P3 program for 7 minutes corresponding to 20V. The nitrocellulose membrane was blocked with 5% milk in T-TBS (TBS buffer with 0,1% Tween20) for 1h at room temperature (RT) shaking and then incubated with the primary antibody in 2,5% milk in T-TBS for either 1h at RT or overnight at +4°C always shaking. After the incubation with the primary antibody the membrane was washed 3 times for 5 minutes with T-TBS shaking and then incubated with the corresponding HRP-conjugated secondary antibody in T-TBS with 2,5% milk for 1 hour at RT. The

membrane was washed 3 more times for 5 minutes with T-TBS. The signal was revealed with the Immobilon Classico Western HRP substrate (WBLUC0100, Millipore) and the membrane developed with the Amersham 600 (GE Healthcare and Life Sciences) or iBright 1500 (Invitrogen).

3.6 GFP trap®_Magnetic Agarose Beads

GFP-Trap®_Magnetic Agarose Beads protocol (gtma-100, Chromotek) for Immunoprecipitation of GFP- Fusion Proteins from mammalian cell extract was performed according to manufacturer instructions. Cell pellets were resuspended in 200 µl ice-cold lysis buffer by pipetting or using a syringe. Tubes were placed on ice for 30 minutes with extensive pipetting every 10 minutes in order to disrupt membranes. Cell lysates were centrifuged at 20.000x g for 10 min at +4°C and transferred to pre- cooled tubes. 300 ⌈l of dilution buffer were added to lysate. For each sample, 25 ⌈l of GFP-Trap®_MA beads were diluted into 500 µl ice-cold dilution buffer and magnetically separated until the supernatant was clear. Equilibrated beads were washed twice. Lysate was added to equilibrated GFP-Trap®_MA and the samples were immunoprecipitated for 2 hours at 4°C. Beads were magnetically separated until supernatant was clear and washed in 500 ⌈l of dilution buffer. Wash was repeated twice and the beads were Resuspend in 100 µl 2x SDS-sample buffer. Samples were boiled for 10 minutes at 95°C shaking in order to dissociate immunocomplexes from the beads and the efficiency of the GFP-trap pull-down was verified by SDS-Page and WB.

3.7 Protein Immunoprecipitation

A volume of 100 µl per sample of protein G Dynabeads (10004D, Life Technologies) was washed twice with 900µl of Lysis buffer (150mM NaCl, 25mM Tris pH 7.4, 5mM EDTA, 1mM DTT, 0,5% NP40, protease inhibitor and 100U/ml of RNAase inhibitor SUPERase in AM2696, Invitrogen). Protein G Dynabeads were resuspended in 100µl of lysis buffer with 4µg of antibody against the protein of interest or IgG for the negative control and incubated at +4°C on a rotating wheel. Cell pellets were resuspended in 1ml of lysis buffer on ice and DNase (AM2238, Invitrogen) was added in order to avoid DNA agglomerates. Samples were incubated on ice for 10 minutes and centrifuged at 13000 rpm for 10 minutes at +4°C. Lysates were mixed with equilibrated beads and incubated 2 hours at +4°C on a rotating wheel. After that, beads were washed three times with Wash Buffer (150mM, 25mM Tris pH 7.4, 5mM EDTA, 1mM DTT and 0,5% NP40), with the last one performed with PBS. 20% of the volume was resuspended in 4x NuPAGE LDS buffer and H_2O, boiled for 10 minutes at 98°C shaking and used for SDS Page-WB analysis.

3.8 Proteomic Analysis

Pull-down of Sam68-GFP complexes at different time points (D0, D3, D10) was performed according to GFP trap protocol.

Snap-frozen samples in five biological replicates were sent to the Buchmann Institute of Molecular Life Sciences (Frankfurt, Germany) where the group of our collaborator Dr. Martin Vabulas proceeded with the proteomic analysis. According to the information

uploaded on the PRoteomics IDEntifications (PRIDE) database, samples were processed by on-beads LysC/Trypsin digestion and released peptides were fractionated into 3 SCX fractions. The MS data were analyzed by Dr. Giulia Calloni using the software environment MaxQuant version 1.5.3.30 (Cox et al. 2008). Proteins were identified by searching MS and MS/MS data against the mouse complete proteome sequences from UniProtKB, version of January 2016, containing 50189 sequences. Carbamido-methylation of cysteines was set as fixed modification. N-terminal acetylation and oxidation of methionines were set as variable modifications. Up to two missed cleavages were allowed. The initial allowed mass deviation of the precursor ion was up to 4.5 ppm and for the fragment masses it was up to 20 ppm. The 'match between runs' option was enabled to match identifications across samples within a time window of 2 min of the aligned retention times. The maximum false peptide and protein discovery rate was set to 0.01. Protein matching to the reverse database or identified only with modified peptides were filtered out. Relative protein quantitation was performed using the LFQ algorithm of the Maxquant with a minimum ratio count of 1 (Cox et al. 2014). Bioinformatic data analysis was performed using Perseus (version 1.5.2.6)[96].

The proteins with minimum four valid values in at least one group (pulldown/background) of five biological replicates were considered as quantified and used for downstream analysis. Proteins enriched in the pulldown over background control at different time points during cell differentiation were identified by two-sample t-test at different permutation-based FDR cutoffs (0.001, 0.01 and 0.05) and s0 = 0.1.

3.9 Size Exclusion Chromatography

Sam68-GFP cellular lysates were fractionated over a Superose 6 HR 10/30 column (Amersham Biosciences) and Sam68 KO mESCs cells were used as negative control. The collected fractions were first screened by immunoblot in order to identify all the Sam68 positive fractions and then further analysed by WB.

3.10 iCLIP2

iCLIP2 protocol[97] was carried out according to Buchbender et al., 2019.

3.10.1 UV-C crosslinking

Cells were washed with 5 ml PBS and 6 ml of ice-cold PBS were added to cells growing in a 10 cm plate. The plate was then placed on ice-plate covered with a layer of PBS. The lid was removed and the cells were irradiated once with 150 mJ/cm2 in a UV-C crosslinker (CL-1000 Ultraviolet Crosslinker, UVP) at 254 nm. Cells were harvested by scraping and cell suspension was transferred in a microtube, spinned at 1,500 xg (4,000 rpm) for 1 min at 4°C to pellet cells. Snap frozen pellets were stored at -80°C until use.

3.10.2 Bead preparation

A volume of 100 μl of protein G Dynabeads per experiment (10004D, Life Technologies) was added to a microtube. Beads were washed twice with lysis buffer (50 mM Tris-HCl, pH 7.4, 100 mM NaCl, 1%

Igepal CA-630, 0.1% SDS, 0.5% Sodium deoxycholate, 1:100 (v/v) Protease inhibitors) and resuspended in 100 μl of lysis buffer with 2-10 μg antibody per experiment. Bead-antibody conjugation occurred by rotating tubes at room temperature for 30-60 min (until lysate was ready). Beads were washed once with high-salt wash buffer (50 mM Tris-HCl, pH 7.4, 1 M NaCl, 1% Igepal CA-630, 0.1% SDS, 0.5% Sodium deoxycholate), twice with lysis buffer and resuspended in 100 μl lysis buffer with added protease inhibitor (P8340-5ML, Merck Sigma).

3.10.3 Lysis and partial RNA digestion

Cell pellets were resuspended in 1 ml lysis buffer (with protease inhibitors). An optimized RNase I dilution (1/200) (AM2295, Life Technologies) in lysis buffer was prepared and added to the lysate together with 2 μl Turbo DNase (AM2238, Life Technologies).
RNA was digested for 3 min shaking at 37°C and 1,100 rpm. After incubation the reaction was transferred to ice for 3 minutes and centrifuged at 4°C at full speed 21,100 xg (14,800 rpm) for 10 min and the supernatant transferred to a new 1.5 ml tube. A volume of 500 μl of the lysate was loaded onto a Proteus mini clarification spin column (42225.01, SERVA Electrophoresis GmbH) and centrifuged at 4°C at 16,000 xg (12,900 rpm) for 1 min. Flow-through was moved to a new tube. The step was repeated with the second half of the lysate.

3.10.4 Immunoprecipitation

Beads were added to cell extracts and the sample were incubated on a rotating wheel for 2 hours at 4°C. Supernatant was discarded and the beads were washed twice with high-salt wash buffer and twice with PNK buffer (20 mM Tris-HCl, pH 7.4, 10 mM MgCl2, 0.2% Tween-20) in order to be resuspended in 1 ml PNK buffer.

3.10.5 RNA 3´ end dephosphorylation

Supernatant was discarded and the beads were resuspended in 20 μl of the following mixture:

- 15 μl Water
- 4 μl 5X PNK buffer, pH 6.5 (350 mM Tris-HCl, pH 6.5, 50 mM MgCl25 mM Dithiothreitol)
- 0.5 μl RNase inhibitor (N2615, Promega GmbH)
- 0.5 μl T4 PNK enzyme with 3' phosphatase activity (M0201S. NEB)

Samples were incubated for 20 min at 37°C in a thermomixer at 1,100 rpm and then washed once with PNK buffer, twice with high-salt wash buffer, incubated for at least 2 min on ice and finally washed again twice with PNK buffer.

3.10.6 First adapter ligation to the 3' end of the RNA

Supernatant was carefully removed and the beads were resuspended in 20 μl of the following mix:

- 8 μl Water
- 5 μl 4X ligation buffer (200 mM Tris-HCl, pH 7.8, 40 mM MgCl2, 4 mM Dithiothreitol)
- 0.5 μl RNase inhibitor
- 1.5 μl Pre-adenylated L3-App (20 μM)
- 4 μl PEG400 (202398-500G, Merck Sigma)
- 1 μl T4 RNA ligase (M204L, NEB)

Samples were incubated overnight at 16°C in a thermomixer at 1,100 rpm and a volume of 500 μl PNK buffer was added. Samples were washed twice with 1 ml high-salt wash buffer (incubated at least 2 min on ice), twice with 1 ml PNK buffer and left in 1 ml of the second wash.

3.10.7 SDS-PAGE and nitrocellulose transfer

Samples were loaded on a 4-12% NuPAGE Bis-Tris gel (NP0322BOX, Life Technologies) according to the manufacturer's instructions (Life Technologies) immersed in 1X MOPS-SDS running buffer (NP0001, Life Technologies). The run was performed for 50 min at 180 V. Protein-RNA complexes were dry transferred from the gel to a nitrocellulose membrane with 20x Transfer buffer (NP0006-1, Life Technologies).

3.10.8 RNA isolation

Protein-RNA complexes were isolated out of the nitrocellulose membrane and the selected regions were cut in several pieces and placed into 1.5 ml tubes. A volume of 10 μl proteinase K (3115828001, Merck Sigma) in 200 μl PK buffer (100 mM Tris-HCl, pH 7.4, 50 mM NaCl, 10 mM EDTA) was added to the nitrocellulose pieces (all should be submerged) and the tubes were incubated in a thermomixer for 20 min at 37°C and 1,100 rpm.

Another 200 μl of PK + urea buffer (7M) were added and the samples were incubated for further 20 min at 37°C and 1,100 rpm.

The solution was collected and added together with 400 μl phenol/chloroform to a 2 ml Phase Lock Gel Heavy tube (733–2478, VWR International GmbH). Tubes were incubated for 5 min at 30°C shaking at 1,100 rpm (do not vortex). Phases were separated by spinning for 5 min at 16,000 xg (12,900 rpm) at room temperature.

The aqueous layer was transferred into a new tube and spin again for 1 min in order to be transferred again into a new tube. Precipitation was performed by adding 0.75 μl GlycoBlue (AM9516, Life Technologies) and 40 μl 3 M sodium acetate pH 5.5. After mixing 1 ml 100% ethanol was added and the tubes were mixed by inverting several times and placed overnight at -20°C. Tubes were centrifuged for 20 min at 21,100 xg (14,800 rpm) at 4°C in order to remove the supernatant and the pellet was washed with 0.9 ml of 80% ethanol, and centrifuged again for 5 min. After air-drying it for 3 min, the pellet was resuspended in 5 μl water and transferred to a PCR tube.

3.10.9 Reverse transcription (RT)

The following reagents were added to the resuspended pellet:

1 μl primer *RToligo* (0.5 pmol/μl)

1 μl dNTP mix (10 mM)

RT thermal programme:

70°C 5 min

25°C hold

until the RT mix (see below) is added, mix by pipetting:

<u>RT mix</u>

7 μl Water

4 μl 5X RT buffer

1 μl 0.1 M DTT

0.5 μl RNase inhibitor

0.5 μl SuperScript III (18080085, Life Technologies)

25°C 5 min

42°C 20 min

50°C 40 min

80°C 5 min

4°C Hold

A volume of 1.65 μl 1 M NaOH was added and samples were incubated at 98°C for 20 min. Then 20 μl 1 M HEPES-NaOH pH 7.3 was added.

3.10.10 Second linker ligation to the 3' end of the cDNA

MyONE clean-up

A volume of 10 µl MyONE Silane beads (37002D, Life Technologies) was considered per sample. Beads were magnetically attracted and washed with 500 µl RLT buffer (79216, Qiagen) and resuspended in 125 µl RLT buffer. A volume of 125 µl washed beads was added to each sample and mixed with an additional 150 µl of 100% ethanol.

After 5 min at room temperature, the samples were mixed once more by pipetting and incubated a second time for 5 min. Beads were magnetically attracted and the supernatant was discarded. Beads were resuspended in 1 ml 80% ethanol and transferred to a new tube. Again, beads were washed twice with 80% ethanol and incubated for 30 seconds at room temperature. The mix was briefly spin in a microcentrifuge and the supernatant was discarded. Beads were air-dried for 5 min at room temperature, resuspended in 5 µl water and incubate for 5 min at room temperature.

Second linker ligation

The second adapter was added together with DMSO to the cDNA-bead solution and the mix was heated for 2 min at 75°C and immediately kept it on ice for >1 min:

2 µl L##clip2.0 (10 μM)
1 µl 100% DMSO (from M0531L, NEB)

The following ligation master-mix was prepared on ice:

- 0.3 μl Water
- 2.0 µl 10X NEB RNA ligase buffer (with DTT)
- 0.2 µl ATP, 100 mM
- 9.0 µl 50% PEG 8000
- 0.5 µl High conc. RNA Ligase (M0437M, NEB)

To ensure homogeneity, the ligation master-mix was mixed by vigorous stirring, pipetting and flicking. The mix was briefly centrifuged in a microcentrifuge. A volume of 12 μl of ligation master-mix was added to 8 μl sample-linker mix and mixed it thoroughly. 1 µl of RNA ligase was added to each sample (final volume: 21 μl), followed by stirring and the sample was incubated overnight at room temperature, shaking at 1,100 rpm.

MyONE clean-up

Fresh MyONE Silane beads were added to the cDNA-bead mix. Therefore, 5 µl of fresh MyONE Silane beads were magnetically attracted, discarding the supernatant and washed with 500 µl of RLT buffer. Magnetic attraction was repeated and in order to resuspend the beads in 60 μl RLT per sample. A volume of 60 µl beads was added to the cDNA-bead slurry and mixed. Then 60 µl of 100% ethanol were added and the solution was mixed by pipetting

carefully. After 5 min at room temperature, the sample was mixed once more by pipetting and the 5 min incubation step was repeated a second time. The beads were magnetically attracted and the supernatant was discarded. The beads were resuspended in 1 ml of 80% ethanol and the mix was transferred to a new tube. This step was repeated other two times incubating the mix 30 seconds at room temperature in 80% ethanol. The mix was spin in a microcentrifuge and after the beads were magnetically attracted, the supernatant was discarded. After air-drying the beads for 5 min at room temperature, they were resuspended in 23 µl of water and the mix was incubated for 5 min at room temperature. The beads were magnetically attracted and the eluate was added to the PCR mix of the next step.

3.10.11 First PCR (cDNA pre-amplification)

The following PCR mix was prepared:

- 22.5 μl cDNA
- 2.5 μl Primer mix of P5Solexa_s and P3Solexa_s, 10 μM each
- 25.0 μl 2X Phusion HF PCR MasterMix (M0531L, NEB)

The following PCR program was run:

98°C 30 s
98°C30 s
65°C 30 s } 8-15 cycles
72°C 30 s
72°C 3 min
16°C Hold

To check the size of the obtained libraries before proceeding with purification, a small percentage of PCR product after cycle optimization was loaded on a Novex™ TBE Gels 6% (Invitrogen, EC62652BOX). The run was performed at 150 V for 30 minutes and the gel was stained with SYBR Gold (Life Technologies, S11494) and visualized by Chemi-Blot using UV. If the size of the obtained fragments was higher than 155 nt (ensuring the insertion of at least 20 nt between linkers), a bead-based purification was performed.

3.10.12 First ProNex size selection to remove primer-dimers

To remove excess primer dimers, samples were size selected with ProNex Chemistry (NG2001, Promega GmbH). ProNex Chemistry beads were equilibrated at room temperature for 30 min and resuspended by vigorous vortexing. For 50 μl of sample, 147.5 μl of ProNex Chemistry beads were considered. This is a 1:2.95 v/v ratio of sample to beads. The solution was mixed by pipetting 10X up and down. The samples were incubated at room temperature for 10 min and then placed on a magnetic stand for 2 min in order to discard the supernatant. Beads were left on the magnetic stand and added with 300 μl ProNex Wash Buffer and incubated for 30-60 s before removal. Samples were air-dried for ca. 8-10 min (>60 min) until cracking started and the beads were removed from the magnetic stand and eluted in 23 μl ProNex Elution Buffer or water. All samples were resuspended by pipetting, and let them stand for 5 min at room temperature. After returning the samples to the magnetic stand for 1 min, the eluted cDNA was moved to a clean tube.

3.10.13 Second PCR amplification and PCR cycle optimisation

The following PCR mix was prepared:

- 3.5 μl water
- 1 μl cDNA
- 0.5 μl Primer mix of P5Solexa_s and P3Solexa_s, 10 μM each
- 5 μl 2X Phusion HF PCR MasterMix (M0531L, NEB)

The following PCR program was run:

98°C 30 s
98°C30 s ⎫
65°C 30 s ⎭ 10-17 cycles
72°C 30 s
72°C 3 min
16°C Hold

Cycle optimisation is essential in order to reach a proper concentration for sequencing and verify the quality of the library. Indeed, an optimized cycle that falls between 6 and 11 indicates a very good library, while 10-13 required cycles means that the library contains a consistent amount of PCR duplicates. When the fragment size shifts into higher size ranges it means that there is over-amplification. To remove residual primer dimers, samples were size selected for the second time with ProNex Chemistry (NG2001, Promega GmbH) (see above).

3.11 iCLIP2 Sequencing and Analysis

The size and the quality of purified libraries were verified on the Bioanalyzer and the sequencing was performed using 50 or 75 bp single reads on a HiSeq 2500 sequencer or MiSeq.

The analysis was performed by the Bioinformatics facility (Sarah Bonnin).

The samples were sequenced in 3 lanes divided as shown below:

Lane 1: Sam D0 + SamD10

Lane 2: QKI D0 + QKI D10

Lane 3: IgG D0 + IgG D10

The reads were then demultiplexed and analysed with the DESeq2 pipeline. The adapter was trimmed off the raw sequences (fastq files) with Skewer (version 0.2.2). The quality of both raw and trimmed reads was assessed with the FastQC tool.

The reads that aligned to the rRNAs or the tRNAs (coordinates obtained from the UCSC table browser), were removed (using bowtie2 with the following parameters: --no-unal --mp 3 -L 10 -N 1 --local -D 10 -R 3).

The reads were then aligned to the Gencode M24 version of the genome (mm10/GRCm38) using the STAR mapper (version 2.5.3a) (with parameters: --outFilterScoreMinOverLread 0.2 --outFilterMatchNminOverLread 0.2 --outFilterMismatchNmax 999 --outFilterMultimapNmax 1 --seedSearchStartLmax 15 --alignEndsType Extend5pOfRead1 --outReadsUnmapped Fastx --outSJfilterReads Unique --quantMode GeneCounts). The 2th column of the resulting STAR counts file that corresponds to an “unstranded” protocol was kept.

Reads were de-duplicated using the UMI-tools, more specifically the umi_tools dedup tool with the "--method=unique" parameter.

PureCLIP (version 1.3.1) was used to detect positions with significant crosslink signals in the deduplicated BAM files. To do so, triplicates from a same experimental condition were merged prior to running PureCLIP. PureCLIP was run with option -ld (allows for higher precision to compute emission probabilities). For STAR mapping, UMI-tools and PureCLIP, we followed some of Busch et. al recommandations.

3.12 RNA Extraction

Total RNA was extracted either with the Maxwell 16 LEV simplyRNA Cells Kit (AS1270, Promega) or manually with QIAGEN RNAeasy kit (ID: 74106, QIAGEN) according to manufacturer instructions.

3.13 RNA Sequencing

The quality and quantity of the extracted RNA was verified using either Bioanalyzer or Nanodrop. Depletion of rRNA and library preparation was performed with the TruSeq stranded total RNA Library prep Human/Mouse/Rat (20020596, Illumina). The sequencing was performed using 2x125bp paired-ends reads on a HiSeq 2500 sequencer with HiSeq v4 chemistry. The obtained reads were demultiplexed and analyzed with the DESeq2 pipeline.

3.14 CircRNA Analysis

Trimmed fastq files were mapped again to the Mus musculus reference genome using bwa-mem: mapping results were then fed into the CIRI2 (version 2.0.6) pipeline for detection and quantification of circular RNAs[98].

The "junction_reads" results from CIRI2 were used to assess circular RNA differential expression between experimental groups.

The analysis provided the number of counts of the circRNA and the number of counts of the linear counterpart. To specifically select circRNA altered at the biogenesis level, we only considered circRNAs that had a FDRpVal < 0.01 (column "FDR_circ") whose linear counterparts were not differentially expressed (column FDRlin > 0.01) or that were differentially expressed in the opposite direction of the circRNA (column FDRlin < 0.01 and column logFC_lin has the opposite sign of column logFC_circ).

3.15 Reverse Transcription

A variable amount of RNA (from 500ng up to 1μg) was retro transcribed to cDNA using the NZY first-strand cDNA synthesis kit (MB125, NZYtech) according to manufacturer instructions. Residual RNA was then digested with the RNase H, a component of the kit.

3.16 Semiquantitative PCR

Semi Quantitative Polymerase Chain Reaction (PCR) was performed by mixing a defined amount of cDNA with H_2O up to a volume of 10μL. A volume of 12.5μL of Phusion High Fidelity PCR Master

Mix (M0531, NEB) was then added. To amplify the DNA a specific thermocycler program was used according to the melting temperature of the primers as well as to the specificity of each DNA target. The products of the PCR reaction were run at 100 V on an agarose gel (from 0,5 up to 3% depending on the amplimer size) in TBE buffer.

4 Results

4.1 Sam68 protein networks

The STAR member Sam68 controls many aspects of RNA metabolism such as alternative splicing[75–77,80], mRNA translation[84], localization and RNA transport[85]. In order to untangle the complexity of all the mechanisms involving our protein of interest in the context of early development of mouse embryonic stem cells (mESCs), we decided to analyze both its protein and RNA partners.

We took advantage of the technology present in CRG facilities and generated a E14 mESCs Sam68-GFP cell line.

Mouse embryonic stem cells have the ability to give rise to an identical daughter cell and, under specific stimuli, to differentiate into all cell types of the three germ layers (endoderm, ectoderm and mesoderm)[93]. Interestingly, ESCs can form Embryonic bodies (EBs) when cultured without the cytokines that maintain them in undifferentiated state (e.g. Leukemia inhibitory factor LIF). The cells forming EBs are allowed to spontaneously differentiate into the embryonic germ layers, providing a robust in vitro model that recapitulate the embryo development and the differentiation progress[94,95].

To identify the Sam68 protein network, we performed an immunoprecipitation followed by MS analysis and determined Sam68 interactors across the differentiation time points (D0, D3, D10).

4.1.1 Generation of Sam68-GFP knock-in mESCs

The generation of Sam68-GFP knock-in cell line was carried out in collaboration with the Tissue Engineering CRG Facility and the Biomolecular Screening and Protein Technology CRG Unit.

As summarized in Figure 1A, E14 WT cells were transfected with the bicistronic plasmid pSpCas9 459–sgRNA63 and the HDR linearized donor vector for the insertion of GFP sequence at the N-terminus of Sam68. The transfected cells were selected with puromycin and visualized under a fluorescent microscope to look for GFP signal. DNA was extracted from the selected clones and was analyzed by PCR to confirm the GFP insertion.

As shown Figure 1B, the insertion of GFP correspond to a band of 2391 bp. From the PCR analysis emerged that the clones 2, 24 and 16 were possible heterozygous. To further confirm the GFP insertion, the possible homozygous clones (8, 10, 20 and 30) were sequenced (Figure 1D) and a WB analysis was performed in order to ensure the translation of Sam68-GFP with the expected size of 100 KDa (Figure 1C). We discarded clone 10 since the sequencing revealed that it was heterozygous and decided to select clone 8 and 20 to proceed with embryonic bodies assay and the MS analysis.

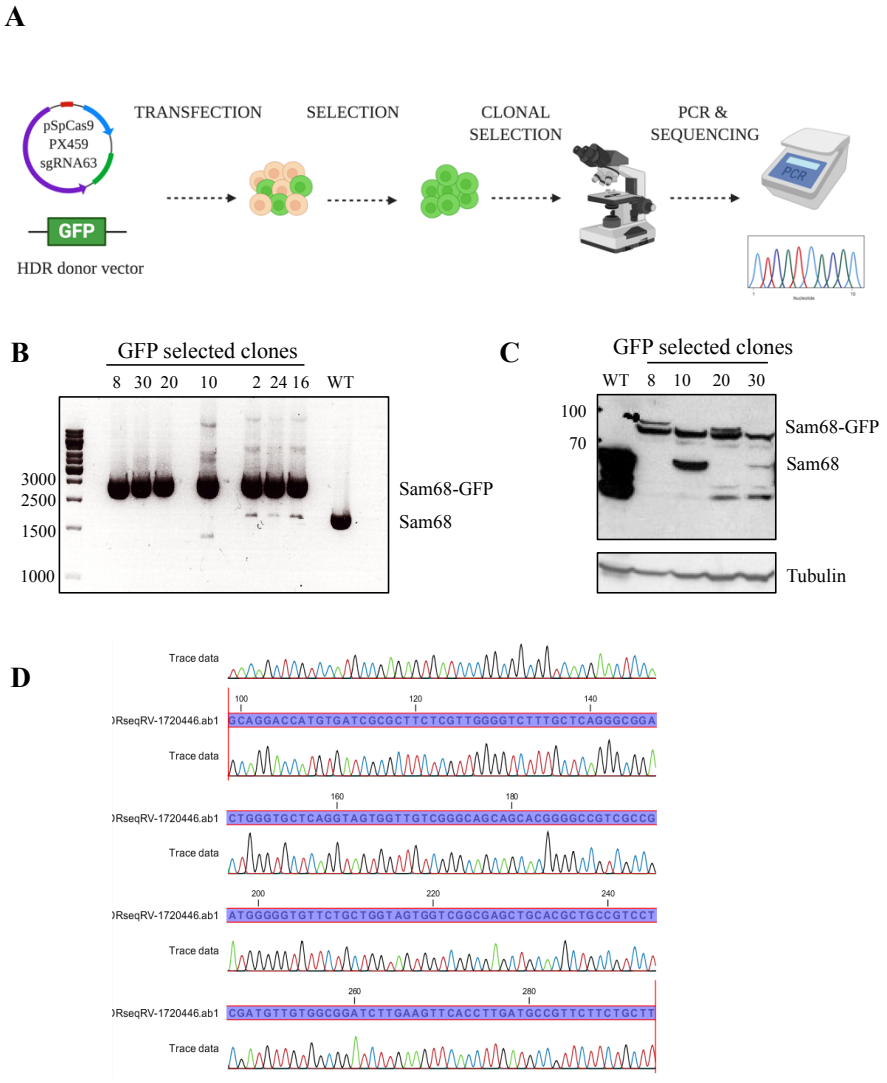


Figure 1A. Schematic workflow of Sam68-GFP KI cell line generation by CRISPR-Cas9. **1B**. PCR analysis of GFP positive clones after puromycin selection. GFP insertion correspond to a band of 2391 bp. **1C.** WB analysis of selected clones. Sam68-GFP expected size is 100 KDa. **1D.** Sequencing capture of the CRISPR-Cas9 generated Sam68 GFP mESCs.

4.1.2 Immunoprecipitation of Sam68 complexes

Embryonic Bodies (EBs) are round shaped 3D structures in which the cells differentiate spontaneously into the three germ layers, recapitulating the embryo development and the differentiation progress[94,95]. As shown in Figure 2A, we performed an Embryonic Bodies (EBs) assay with Sam68-GFP mESCs, considering three time points representing respectively pluripotent state, early stage of differentiation and late stage of differentiation (D0, D3, D10). Sam68-GFP interactors were immunoprecipitated by GFP trap method according to manufacturer instruction.

Once the efficiency of the GFP trap protocol was confirmed (Figure 2A), the same experiment was repeated on five biological replicates at 3 different time points we considered (D0, D3, D10) (Figure 2C). GFP infected E14 mESCs with stably expressed GFP were used as control background (BG).

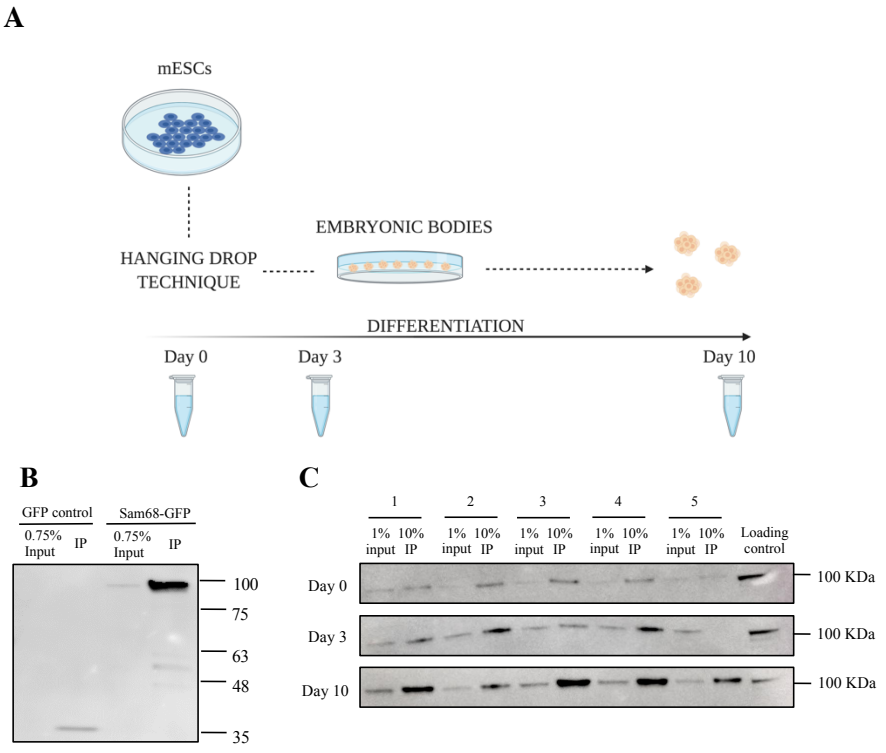


Figure 2A. Schematic workflow of Embryonic Bodies Assay **2B**. WB showing IP of Sam68-GFP (100 KDa) compared to input. GFP cells used as a control. **2C**. IP of Sam68-GFP in five biological replicates at different time points (Day 0, Day 3, Day 10)

4.1.3 Proteomic analysis

The MS data were analyzed by Dr. Giulia Calloni (Institute of Molecular Life Sciences – Frankfurt, Germany) using the software environment MaxQuant version 1.5.3.30[99]. Proteins were quantified (Figure 3) and the ones enriched in the pulldown (SamGFP mESCs) over background control (GFP mESCs) at different time points during mESCs differentiation were identified by two-sample t-test at different permutation-based FDR cutoffs (0.001, 0.01 and 0.05) and s0 = 0.1.

Figure 3. Bar plot showing protein quantification for each time point and replicate. Sam68-GFP samples (PD) are indicated in green, while the orange refers to the GFP cells (BG).

The Pearson correlation (Figures 4 and 5) obtained from the samples showed a high degree of reproducibility between our biological replicates in Sam68-GFP mESCs cells (referred as PD) and in our negative control (BG). To ensure data quality, from the downstream analysis we decided to exclude the samples with an average correlation coefficient inferior to 0.8.

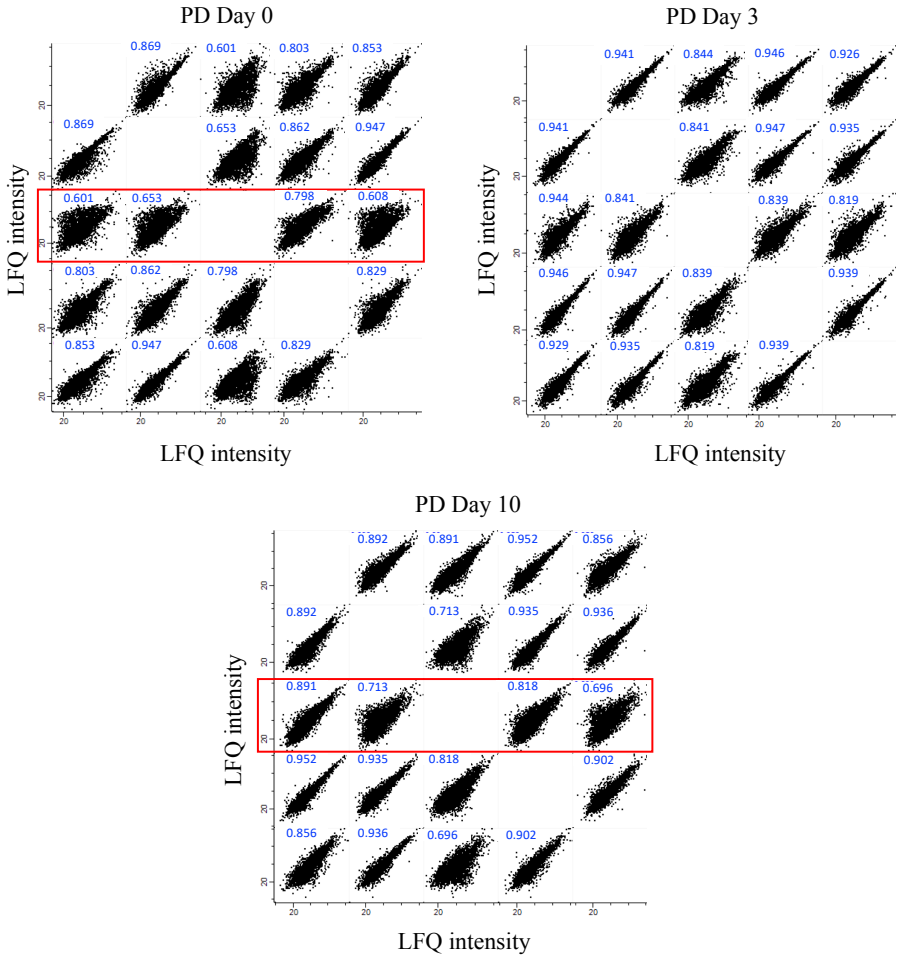


Figure 4. Pearson correlation analysis obtained from LFQ (Label Free Quantification) intensities of Sam68-GFP pull-down (PD) between biological replicates at D0, D3, D10. Samples excluded from the analysis are indicated in red.

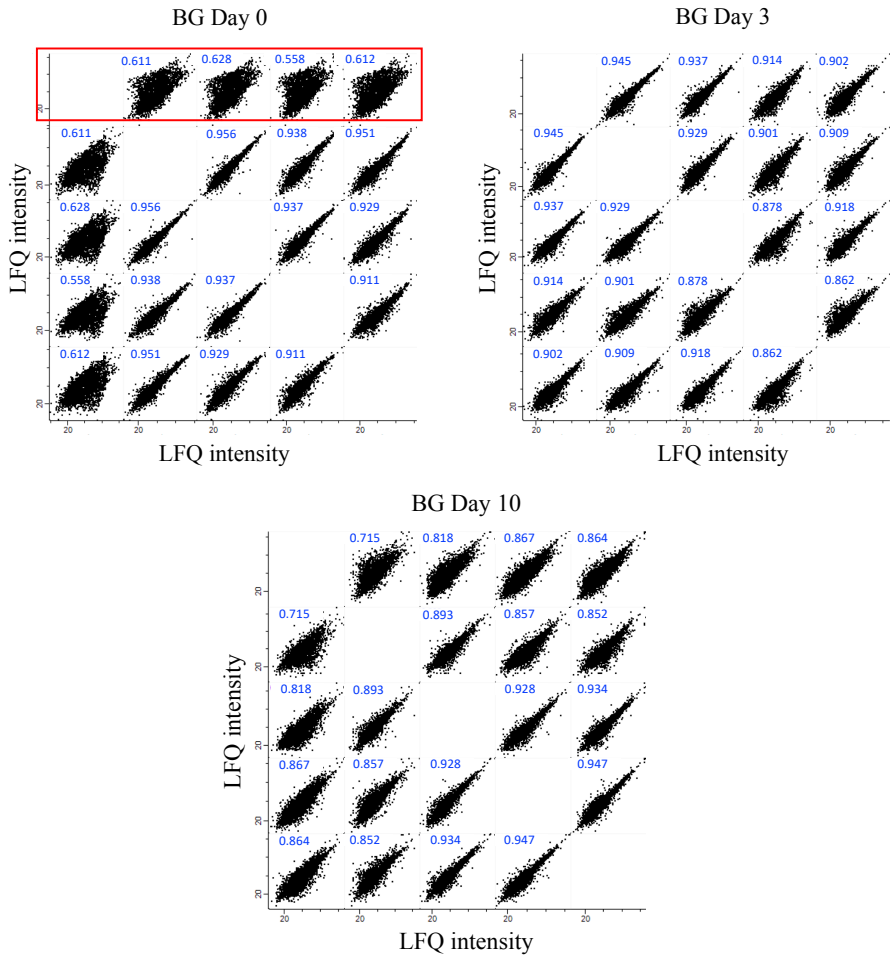


Figure 5. Pearson correlation analysis obtained from LFQ (Label Free Quantification) intensities of GFP-transfected pull-down used as back-ground (BG) between biological replicates at D0, D3, D10. Samples excluded from the analysis are indicated in red.

The samples that were not considered in our analysis are indicated by a red rectangular (Figure 4 and 5). Two replicates were excluded from the PD set and in our control background GFP cells we applied the same principle and eliminated the samples not respecting our criteria (one replicate).

The MS analysis identified 194 significant interactors (FDR<0.01) at the 3 time points considered (D0, D3, D10): the number of Sam68 protein partners increases during the differentiation timeline, although some of them are shared between the time points, as indicated in the Venn diagram, where Sam68 is highlighted in red (Khdrbs1) (Figure 6). It is worth mentioning that Sam68 expression increases during differentiation (Figure 2C) and the MS data were normalized to avoid the bias due to Sam68 higher enrichment in D10 compared to D0 and D3.

As shown in Figure 7, the volcano plots represent for each time point Sam68 protein partners divided according to the different FDR cutoffs (0.001, 0.01 and 0.05) and highlighted by different colors.

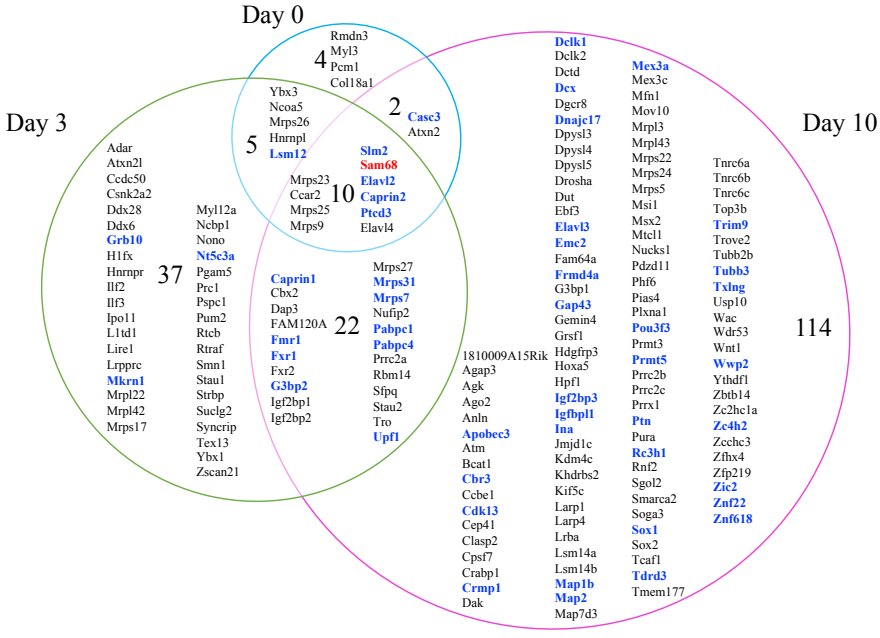
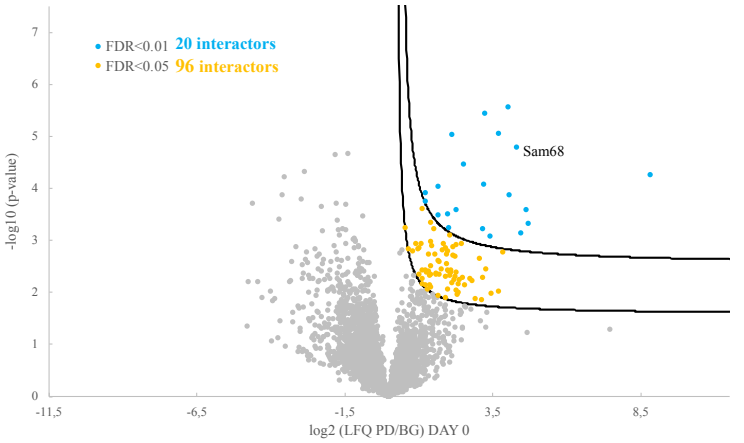
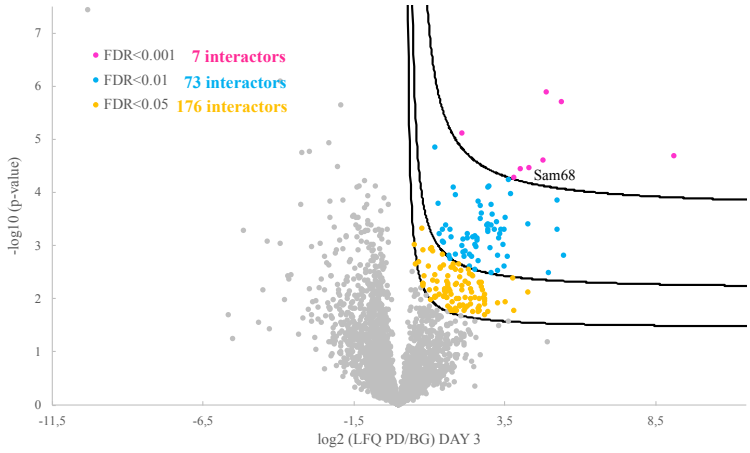


Figure 6. Venn diagram showing Sam68 interactors during D0, D3, D10 (FDR 0.01). In blue are highlighted interactors with FDR<0.001 at D3 and D10.


FDR<0.01 20 interactors
FDR<0.05 96 interactors
Sam68
-log10 (p-value)
log2 (LFQ PD/BG) DAY 0


FDR<0.001 7 interactors
FDR<0.01 73 interactors
FDR<0.05 176 interactors
Sam68
-log10 (p-value)
log2 (LFQ PD/BG) DAY 3

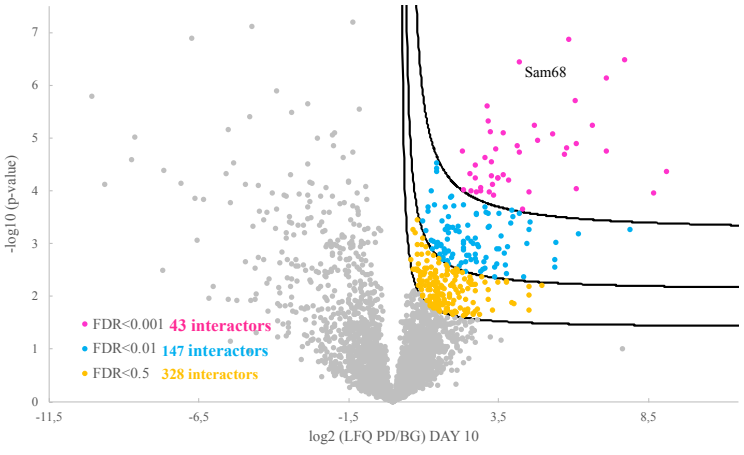

Sam68
FDR<0.001 43 interactors
FDR<0.01 147 interactors
FDR<0.5 328 interactors
-log10 (p-value)
log2 (LFQ PD/BG) DAY 10

Figure 7. Volcano plots showing Sam68 interactors during D0, D3, D10. Each FDR threshold is indicated by a different color: interactors with FDR<0.5 are in yellow, while blue and pink indicate respectively protein partners with FDR<0.01 and FDR <0.001.

The GO terms analysis for molecular function obtained with the tool developed in our laboratory (CleverGO) revealed that 50% of Sam68 interactome is composed by RBPs involved in all the steps of the RNA metabolism. Among the functional interaction network, we found interactors involved in mRNA splicing (Sfpq and Nono)[100,101], transport and translation (e.g. Stau2, Caprin1)[102,103], m6A readers associated with translation stability and translational control (e.g. Igf2bp1, Igf2bp2, Igf2bp3)[104].

Indeed, the most significant GO terms obtained from the analysis are related to *3´UTR* and *5´UTR* binding, as well as *polyA* and *polyU* binding, reflecting the known roles of Sam68 in RNA metabolism[72,105]. Between Sam68 protein partners we could retrieve some whose interaction has already been described in other physiological contexts. Indeed, in our proteome we find components of the miRNA pathway (Drosha, Ago2, Dgcr8): Sam68 was already linked to miRNA processing machinery during spermatogenesis, where it interacts with Drosha and Dicer[83].

Moreover, we detected the interaction with the heterogeneous nuclear ribonucleoproteins HnrnpL, important player in transcriptional regulation during myogenic differentiation[106].

In order to extract more information from our MS results, we decided to increase the stringency of our MS analysis and select only Sam68 interactors that showed a p-value lower than 0.01 and a fold-change

(FC) higher than 2 compared to the control. The data were represented in a heatmap showing the grade of interaction with Sam68 for each time point (Figure 8). The obtained clustered visualization allowed us to identify three different classes of interactors:

- core components, constantly interacting with Sam68 during differentiation, represented in darker blue;
- time-point specific interactors, acting only in a specific stage of the differentiation process;
- specialized interactors, transiently present in one or more days but not as core components.

Among the core components we found Pum, a translational regulator with the ability to bind to 3'UTRs[107], the cell proliferation and apoptosis regulator Ccar[108], Elavl4 and Elavl2, involved in neocortical development and neuronal excitability[109,110], Casc3, mitochondrial ribosomal proteins associated with breast cancer[111] Mrps23, Mrps27 and Ptcd3, a mitochondrial translation-related protein important in tumor progression[112].

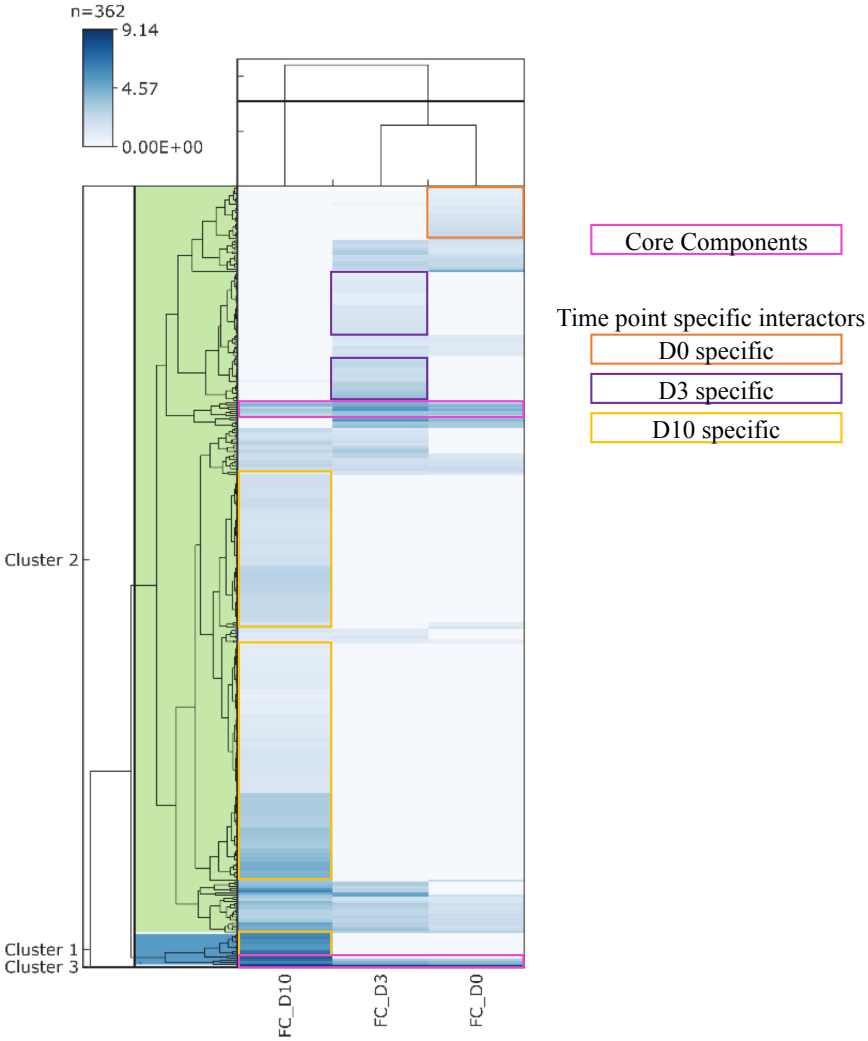


Figure 8. Heat map (FDR<0.01- FC>2) showing Sam68 interactors during D0, D3, D10 time points of mESCs differentiation.

Interestingly, the strongest interaction that we identified in the whole proteome is between Sam68 and Slm2, another member of the STAR family. Slm2, together with Sam68, shares a very important role in alternative splicing during neural development. Indeed, both STAR members are able to bind Neurexin pre-mRNA, but only Slm2 is

proficient to regulate its splicing depending on the binding sites density[79,91].

As Sam68 shuttles between the nucleus and the cytoplasm, we wanted to investigate whether Slm2 is always present in Sam68 containing complexes. To do so, we took advantage of the Size Exclusion Chromatography (SEC) technology available in the Biomolecular Screening and Protein Technology CRG Unit.

Sam68-GFP cellular lysates were fractionated over a Superose 6 column and Sam68$^{-/-}$ mESCs cells were used as negative control. As shown in Figure 9, the collected fractions were first screened by immunoblot in order to identify all the Sam68 positive fractions and then further analysed by western blot. In WT conditions, we could identify both a small (150 KDa) and a large (>670 KDa) Sam68 containing complex. Interestingly, Slm2 was detected only in some of the collected fractions.

Furthermore, in order to analyse if Sam68 interactions were RNA-dependent, the samples were treated with RNAse before fractionation and we observed that the treatment caused a change in the SEC profile in addition to a partial precipitation of Sam68 before column injection (Figure 10), suggesting that RNA is an essential scaffolding element in Sam68 interaction network.

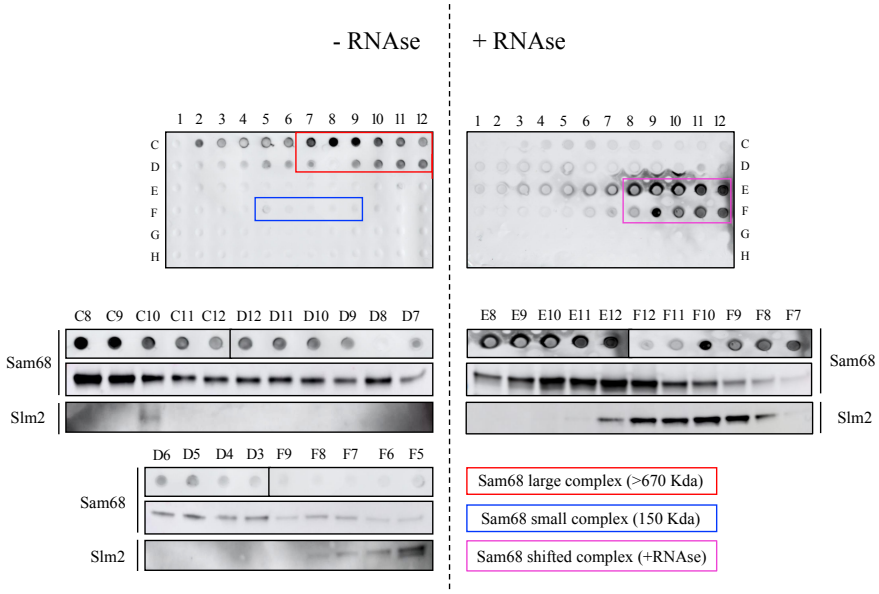


Figure 9. Dot blot analysis of mESCs lysates in absence or presence of RNAse fractionated over a Superose 6 column. Areas marked in red correspond to Sam68 large complex (>670 KDa), blue color indicates Sam68 small complex (150 KDa). In RNAse condition, Sam68 shifted complex is highlighted in pink.

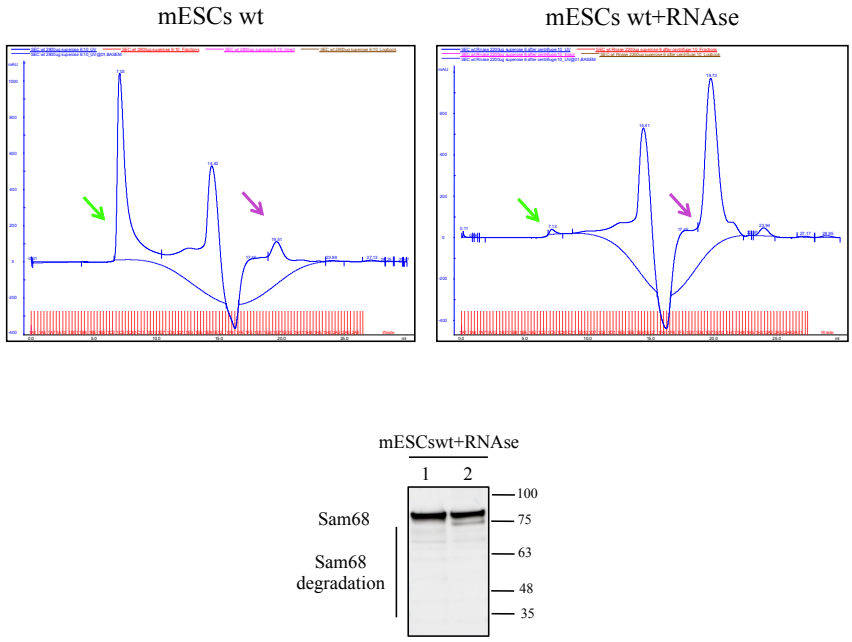


Figure 10. Size Exclusion Chromatography (SEC) profiles of mESCs lysates in wt and wt+RNAse conditions. The arrows indicate the shift due to RNAse treatment. WB of the precipitations residues obtained after RNAse treatment performed before injection.

4.2 Sam68 interaction with circRNA biogenesis factors

The canonical roles of Sam68 in RNA processing have been widely studied and characterized in the last decades[72,105], hence, we decided to focus our attention on interactors that could extend the regulative functions of Sam68 in RNA metabolism.

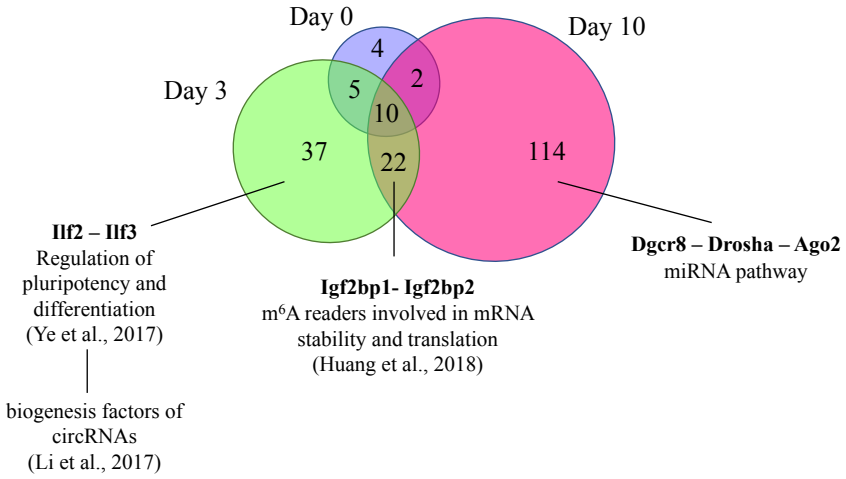


Figure 11. Venn diagram showing Sam68 interactors involved in circRNA metabolism.

As shown in Figure 11, from our list of candidates there were some elements that captured our attention: at Day3 time point we found two component of the Interleukin enhancer-binding factor family (Ilf2 and Ilf3), involved in the biogenesis of circRNAs upon viral infection and in the regulation of pluripotency and differentiation in mESCs[93], the m6A readers Igf2bp proteins that are important for mRNA stability and translation[104] as well as components of the

miRNA pathway (Dgcr8 – Drosha – Ago2)[83], present at Day10. All these interactors are part of the circRNA metabolism[18]. CircRNAs are a class of ncRNAs able to influence gene expression that are becoming subject of intense research[113]. Indeed, Ago2-mediated cleavage has been proposed as a possible circRNA degradation mechanism[114], while m^6A modification is enriched in circRNAs influencing their stability[115].

CircRNA production has been suggested to be an essential element upon viral infection: immune response factors such as Ilf3 influence circRNA biogenesis by stabilizing intronic RNA pairs in the nucleus and binding to circRNP in the cytoplasm. During the innate immune response against virus infection, Ilf3 isoforms (90 kDa and 110 kDa) are released from circRNPs complexes in the cytoplasm allowing their binding to viral mRNAs to inhibit viral replication[31].

The STAR member Quaking is a known circRNA biogenesis factor during the epithelial to mesenchymal transition (EMT) and a new role for Sam68 in the pre-mRNA circularization process has been recently described for the SMN locus[30,92]. Therefore, we wondered whether Sam68 was actively involved in circRNA metabolism during mESCs differentiation by interacting with key biogenesis factor such as Ilf2 and Ilf3.

To support the hypothesis of a possible role of Sam68 in circRNA biogenesis, we decided to investigate this interaction by performing a pull down (PD) of Ilf3 in presence and absence of RNAse.

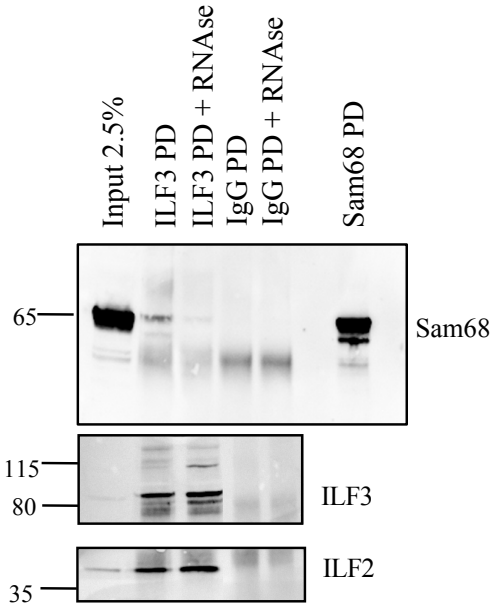


Figure 12. Western blots representing the pull-down (PD) of Ilf3. The upper membrane indicates the presence of Sam68 in absence and presence of RNAse. IgG were used as a negative control and Sam68 IP was used as positive control for the detection of the interaction. The middle and lower membranes confirm the pull-down of Ilf3 and the known interaction with Ilf2.

The membrane was revealed with antibodies against Ilf3 (90 and 110 KDa), Ilf2 (45 KDa) and Sam68. IgG conjugated beads were used as negative control and a pull down of Sam68 was considered as positive control. As reported in figure 12, Sam68 resulted to be much more abundant in the protein extract compared to Ilf2 and Ilf3 endogenous expression, as it is also shown from the input band. Indeed, in previous efforts performed to confirm the interaction, we immunoprecipitated Sam68 and it was very difficult to visualize the circRNA biogenesis factors bands given their low expression.

Notably, we confirmed the interaction between Sam68 and Ilf3 at D3 time point. Comparing the IgG signal between the samples, it is noticeable that in the Ilf3 PD+RNase lane Sam68 band is weaker, suggesting that Sam68 signal might have been stronger if the loaded amount was the same as the other samples. Moreover, RNase treatment seems to destabilize this interaction, suggesting that the presence of RNA balances the cooperation between Sam68 and the circRNA biogenesis factor Ilf3.

4.3 CircRNA analysis

The interaction between Sam68 and the circRNA biogenesis factor Ilf3 led us to further investigate the involvement of Sam68 in circRNA biogenesis, wondering whether Sam68 expression could affect circRNAs production at a genome-wide level.

We took advantage of RNA sequencing data previously produced in our laboratory (Dasti et al., submitted) in order to analyse circRNAs abundance upon Sam68 and QKI depletion during mESCs differentiation (Figure 13). The circRNA analysis was performed in collaboration with Magdalena Arnal and Alessio Colantoni (Italian Institute of Technology- Genoa, Italy) using a specific pipeline for circRNA identification.

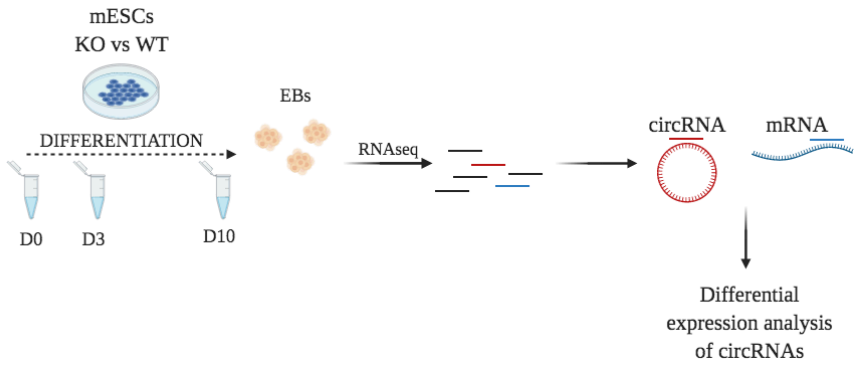


Figure 13. CircRNA differential expression analysis workflow.

To detect circRNAs from our RNA sequencing data, we analysed the reads mapping to back-spliced junctions as well as those mapping to their linear counterpart and focused on the reads mapping to circRNAs that were differentially expressed upon Sam68 and QKI depletion but with unaltered level of the linear mRNA. These circRNA are the ones altered at biogenesis level, since their expression is not due to the corresponding linear RNA changes.

4.3.1 Sam68 circRNA analysis

Strikingly, the differential expression analysis revealed that, in the absence of Sam68, circRNAs expression level is significantly decreased, suggesting the implication of Sam68 in the regulation of circRNA biogenesis during differentiation.

TIME POINT	CircRNAs		FDR
	UP	DOWN	
SAM68 D0	2	339	< 0.01
SAM68 D3	1	580	< 0.01
SAM68 D10	3	278	< 0.01

Table 1. The table represents the number of circRNAs upregulated and downregulated retrieved from the circRNA analysis in Sam68$^{-/-}$ mESCs.

The Table 1 reports the number of circRNAs with altered biogenesis when Sam68 is depleted from mESCs, revealing a consistent number of downregulated circRNAs.

As shown in Figure 14, we plotted the log2 fold change values of the identified circRNAs against the values corresponding to their linear counterpart.

In absence of Sam68, the circRNA biogenesis defect is mostly appreciable at D0 and D3, where the distributions of the data showed that most of differentially expressed circRNAs did not have alteration in their linear counterpart.

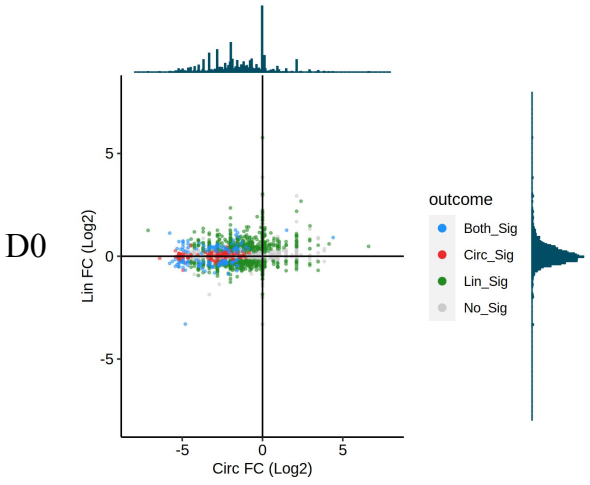
Notably, at D10, circRNAs seemed to follow their cognate linear RNA expression, as indicated from the changes in the data distribution.

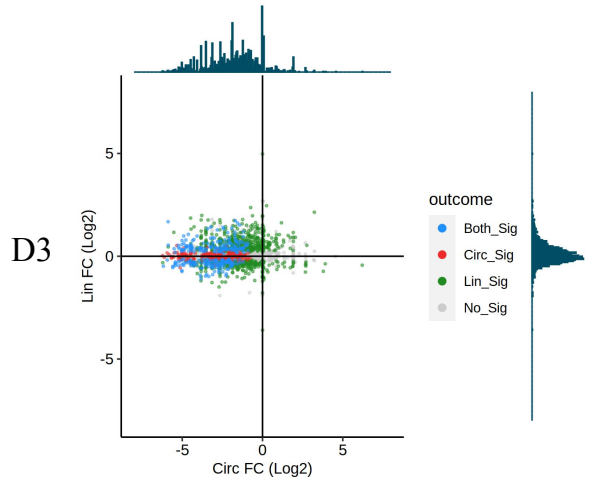
We have shown that Sam68 depletion severely affects the cardiomyocytes differentiation and function. We therefore focused on the circRNAs that are highly expressed in mouse cardiomyocytes. We observed a high overlap between the Sam68 regulated circRNAs and the cardiomyocytes ones. The scatter plot (Figure 15) represents

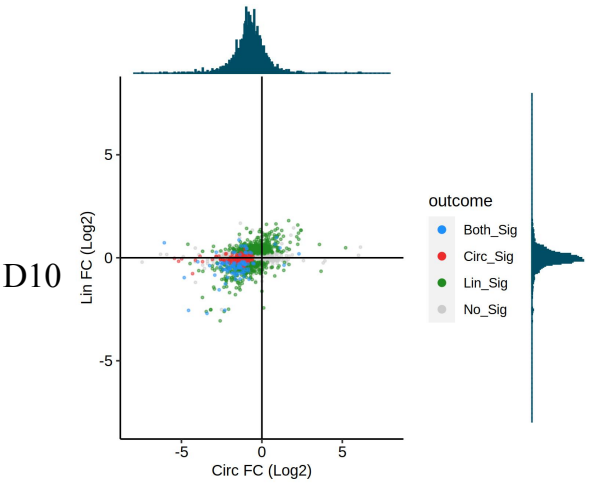
in red the circRNA hosting genes highly expressed in mouse heart compared to the total circRNA analysis outcome.

The normalization against linear expression is still ongoing and will be essential for the identification of altered circRNAs with specific functions in cardiomyocytes.

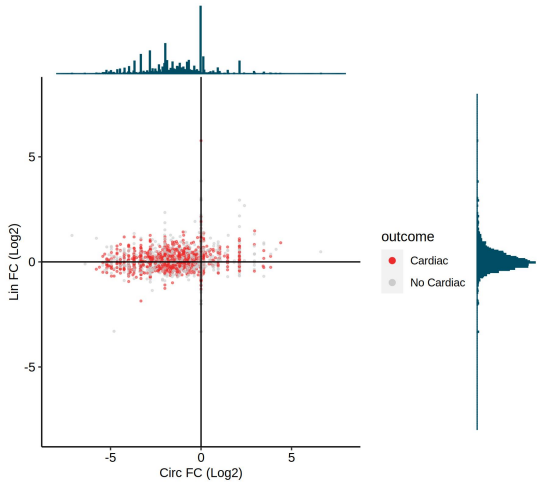
Figure 14 (following page). Scatter plots (FDR<0.01) of circRNA analysis in Sam68$^{-/-}$ cells representing circ FC (log2) values of the identified circRNAs against the values corresponding to their linear counterpart (Lin FC Log2). Red dots correspond to circRNAs with unaltered linear cognate expression, while green dots indicate circRNA with differentially expressed linear RNA counterpart.


D0
Lin FC (Log2)
Circ FC (Log2)
5
0
-5
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig

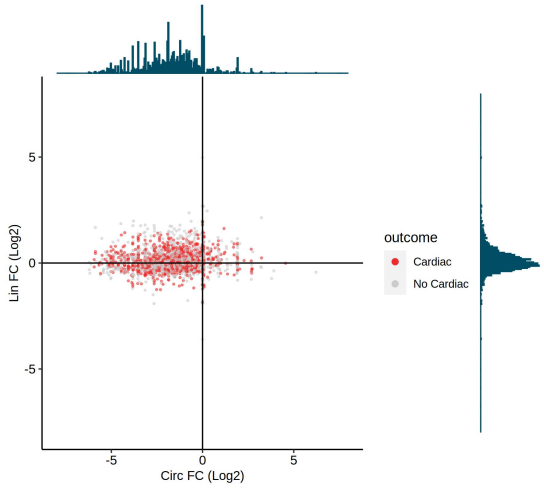

D3
Lin FC (Log2)
Circ FC (Log2)
5
0
-5
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig


D10
Lin FC (Log2)
Circ FC (Log2)
5
0
-5
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig

D0


Lin FC (Log2)
Circ FC (Log2)
outcome
Cardiac
No Cardiac

D3


Lin FC (Log2)
Circ FC (Log2)
outcome
Cardiac
No Cardiac

D10

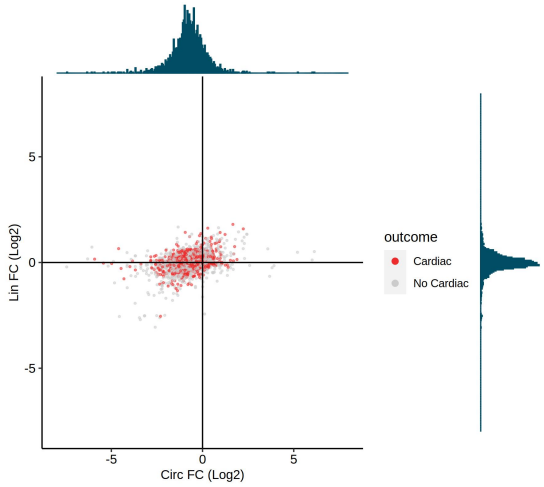

Lin FC (Log2)
Circ FC (Log2)
outcome
Cardiac
No Cardiac

Figure 15. Scatter plots (FDR<0.01) representing the overlap between differentially expressed circRNAs upon Sam68$^{-/-}$ cells with circRNAs highly expressed in cardiomyocytes from Tan et al., 2018.

4.3.2 QKI circRNA analysis

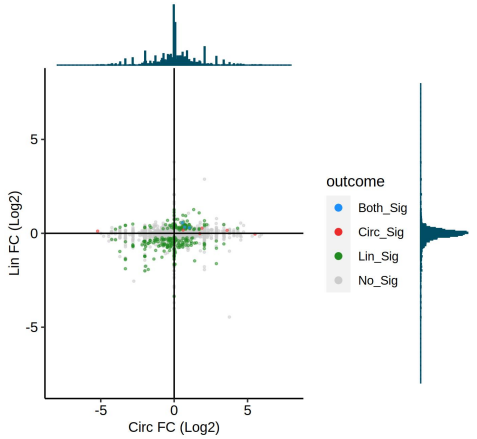
As previously described, the STAR family member QKI is a well-known circRNA biogenesis factor during epithelial to mesenchymal transition (EMT). Indeed, QKI binds introns flanking circRNA-forming exons and facilitates the formation of back-spliced junctions because of its dimerization ability[30].

TIME POINT	CircRNAs		FDR
	UP	DOWN	
QKI D0	29	7	< 0.01
QKI D3	7	1	< 0.01
QKI D10	9	14	< 0.01

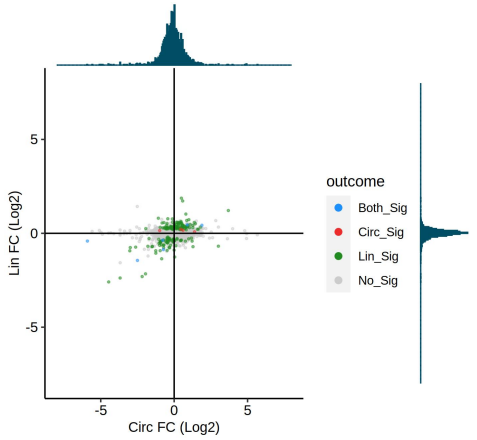
Table 2. The table represents the number of circRNAs upregulated and downregulated retrieved from the circRNA analysis in QKI$^{-/-}$ mESCs.

However, despite its known circRNA biogenesis activity in a determinate biological context such as epithelial to mesenchymal transition (EMT), the differential expression circRNA analysis performed upon QKI depletion at different time points of mESCs differentiation did not show any specific trend related to the RBP absence (Table 2), suggesting that QKI does not affect circRNA biogenesis during mESCs differentiation (Figure 16).

D0


Lin FC (Log2)
Circ FC (Log2)
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig

D3


Lin FC (Log2)
Circ FC (Log2)
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig

D10

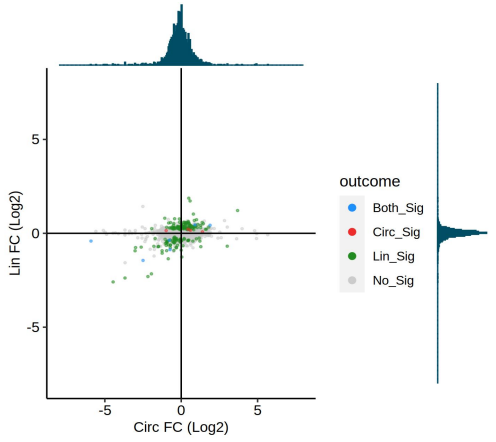

Lin FC (Log2)
Circ FC (Log2)
outcome
Both_Sig
Circ_Sig
Lin_Sig
No_Sig

Figure 16. Scatter plots (FDR<0.01) of circRNA analysis in QKI$^{-/-}$cells representing circ FC (log2) values of the identified circRNAs against the values corresponding to their linear counterpart (Lin FC Log2). Red coloured dots correspond to circRNAs with unaltered linear cognate expression, while green dots indicate circRNA with differentially expressed linear RNA counterpart.

4.4 Sam68 mRNA targets

Our circRNA analysis showed a significant downregulation of circRNA at biogenesis level in Sam68$^{-/-}$ mESCs, however, in order to demonstrate the direct involvement of Sam68 in circRNAs formation, it was essential for us to investigate its binding profile.

The STAR member Quaking promotes circRNA production by binding to intronic regions and bringing in close proximity circularizing exons thanks to its dimerization ability[30]. Sam68 has a similar binding ability and has been recently discovered to be involved in the pre-mRNA circularization of the SMN locus, taking advantage of the high density of *Alu* sequences present in this precise locus[92]. Nevertheless, whether Sam68 could induce circRNA biogenesis in other physiological contexts was not known yet, hence, the investigation of Sam68 mRNA targets represented a crucial step to define its possible role in circRNA formation during differentiation.

In order to investigate the binding profile of Sam68, we first analysed its interaction propensity towards circRNA-forming transcripts expressed in mouse cardiomyocytes[46] using catRAPID, a bioinformatic tool developed in our laboratory that estimates the binding propensity between protein-RNA pairs; then, to validate the predicted binding propensity, we performed a transcriptome-wide

mapping of Sam68 mRNA targets using the iCLIP2 (individual-nucleotide resolution UV cross-linking and immunoprecipitation) technique. This recently updated protocol represents a high-resolution method able to identify specific protein-RNA complexes using immunoprecipitation followed by SDS-PAGE and recovery of RNA sequences[97].

4.4.1 Predicted mRNA targets with catRAPID

CatRAPID is an algorithm able to estimate the binding propensity of protein-RNA pairs by combining secondary structure, hydrogen bonding and van der Waals contributions[116].

CatRAPID was previously used by our laboratory to predict the binding propensity between Sam68 and Gata4, a master regulator of mesodermal differentiation. We were particularly interested in confirming this interaction because of a defective phenotype of cardiomyocytes differentiation functionally related to Sam68 depletion in mESCs. The algorithm could indeed predict the binding, narrowing down the nucleotides involved. The interaction between Sam68 and the predicted nucleotides range Gata4 was confirmed by RNA immunoprecipitation (RIP) and electrophoretic mobility shift assay (EMSA) (Dasti et al., submitted).

CatRAPID analysis was performed in collaboration with Alexandros Armaos (Italian Institute of Technology – Genoa, Italy).

In order to investigate the binding propensity of Sam68 against circRNA genes expressed in mouse cardiomyocytes[46] and its binding specificity, a pool of 1799 RBPs[117] was used as control. To ensure specificity of Sam68 and QKI binding towards the selected targets,

we considered all the possible interactions between this RBP pool and the mouse transcriptome, which were around 1.7M interactions. We calculated the normalised catRAPID score for these 1.7M interactions by multiplying the catRAPID score with the RBP propensity of each RBP. Then, we built as background the distribution with the normalised catRAPID scores of that RNA and the RBPome. Hence, the specificity of the interaction between an RNA from the list and SAM68 or QKI will depend on how high this interaction ranks with respect to the background distribution of that specific RNA (Figure 17).

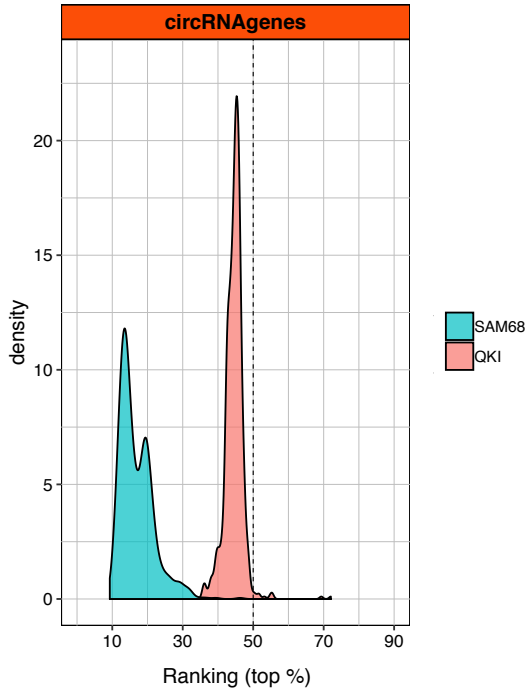


Figure 17. Comparison between Sam68 and QKI binding specificity to circRNA genes against 1799 RBPs retrieved from RBPome published in Hentze et al., 2018.

There are no clear differences in the specificity of the interactions of these RNAs with QKI, since the values rank always modestly high (top 50%). However, for SAM68 we see much more specificity that ranges between top 20% for the circRNAs genes, meaning that Sam68 binding propensity to transcripts that are the source of circRNAs in mouse cardiomyocytes is very significant. We also decided to consider different features of the target transcripts in terms of length, number of exons, 3'UTR and 5'UTR length. Our catRAPID analysis showed high scores for the interaction propensity between Sam68 and cardiac-related circRNA genes. Compared to the background, those transcripts appeared to be longer, with higher number of exons and longest 3'UTRs and 5'UTRs (Figure 18).

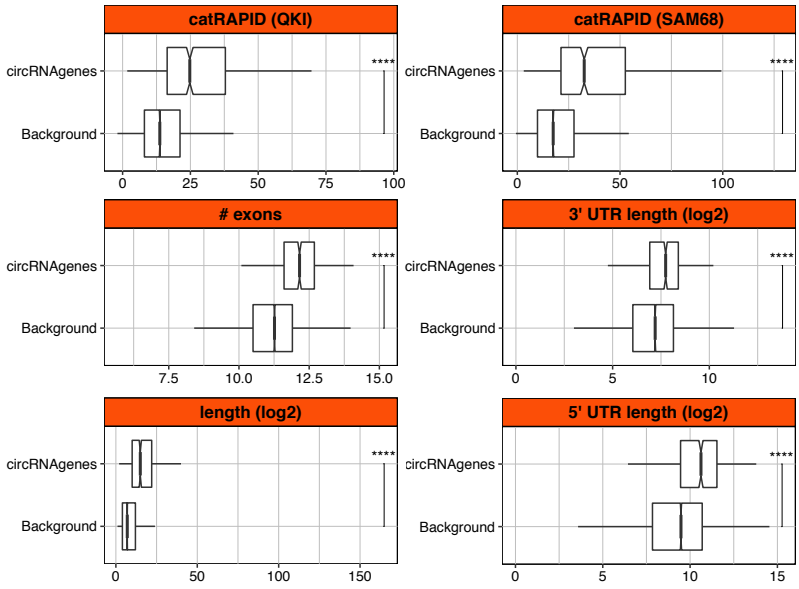


Figure 18. CatRAPID analysis of QKI and Sam68 binding propensity towards circRNA genes highly expressed in mouse cardiomyocytes and analysis of circRNA transcripts features in terms of length, number of exons, 3'UTR and 5'UTR length. Background set consists in randomly sequences of the same size as the positive set selected 50 times from the mouse transcriptome.

4.4.2 Sam68 and QKI mRNA targets with iCLIP2

The bioinformatic tool CatRAPID predicted a high propensity for Sam68 to directly bind transcripts involved in circRNA production in mouse cardiomyocytes. To experimentally validate Sam68 binding to its mRNA targets, we performed Sam68 and QKI iCLIP2 on wt E14 mESCs. IgG were used as negative control to remove the non-specific binding background. After inducing a covalent binding between our RBPs of interest (Sam68 and QKI) and their RNA targets by UV-crosslinking, we performed an optimized RNAse treatment on the protein lysates with the aim to obtain RNA fragments ranging between 50-200 nucleotides. We immunoprecipitated the protein-RNA complexes and, as shown in Figure 19, we cut from the membrane the portion above Sam68 molecular weight in order to proceed with RNA fragments recovery.

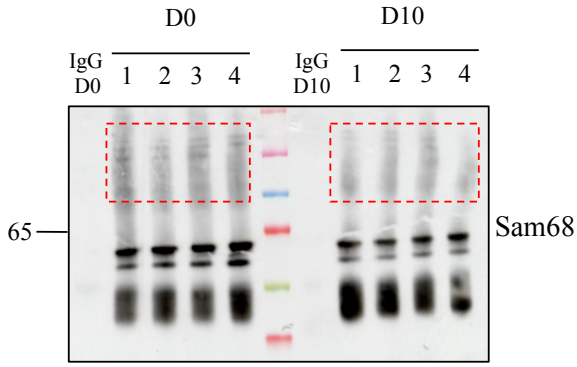


Figure 19. Western blot of Sam68 pull-down at D0 and D10 time points for iCLIP2 experiment. IgG were used as a negative control. The red area indicates the portion of the membrane that was cut in order to recover the protein-RNA complexes.

RNA fragments were converted in cDNA by reverse transcription and a second adapter was added. To minimise sample-loss, we performed a cDNA pre-amplification before the PCR and checked the size of the obtained iCLIP2 libraries. A small percentage of PCR product (after cycle optimization) was loaded on a 6% TBE Gel (Figure 20). If the size of the obtained fragments was higher than 155 nt (ensuring the insertion of at least 20 nt between linkers), a bead-based purification was performed. In order to remove the excess of primer dimers, samples were size-selected with ProNex Chemistry.

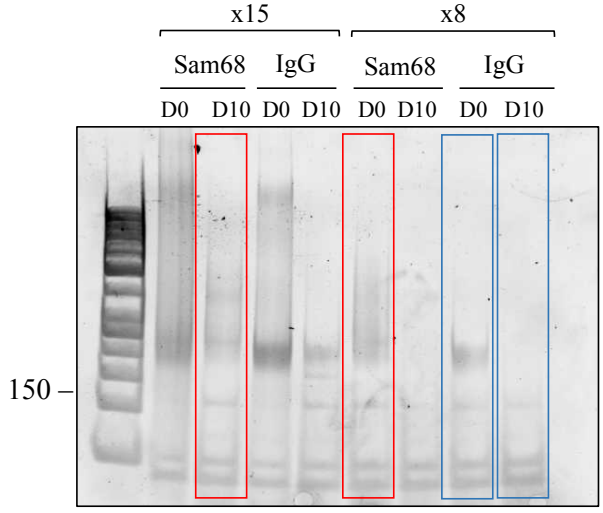


Figure 20. TBE 6% gels representing pre-amplification optimization considering either 15 or 8 cycles. In red and blue are marked the samples selected for the purification and sequencing for Sam68 and IgG.

To remove residual primer dimers, samples were size selected for the second time with ProNex Chemistry and the size and the quality of purified libraries were verified on the Bioanalyzer before sequencing. The iCLIP2 sequencing analysis was carried out by the Bioinformatics facility of CRG (Sarah Bonnin).

Sample	Reads	% Uniquely Aligned
IgG D0	25 M	58-66
IgG D10	0,815 M	51-58*
Sam68 D0	21 M	71-77
Sam68 D10	17 M	66-81*
QKI D0	12 M	67-72
QKI D10	15 M	72-78*

Table 3. Table representing number of reads and % of uniquely aligned sequences in iCLIP2 samples. Values marked with * are obtained excluding one triplicate from the average calculations.

Peak calling was performed using PureCLIP [118] and the samples were merged by replicate in order to avoid losing signal.

The number of reads and the percentage of uniquely aligned reads obtained for each condition are indicated in Table 3.

The iCLIP2 analysis identified the genomic binding sites of Sam68 and QKI, demonstrating that, in genic regions, both RBPs preferentially binds introns (Figure 21). A smaller percentage of binding occurs in exon 3'UTR, 5'UTR and promoter regions.

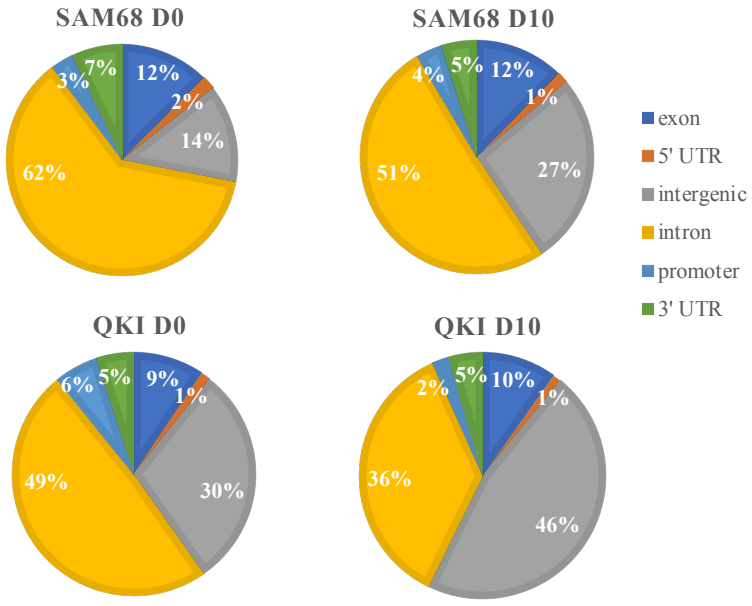


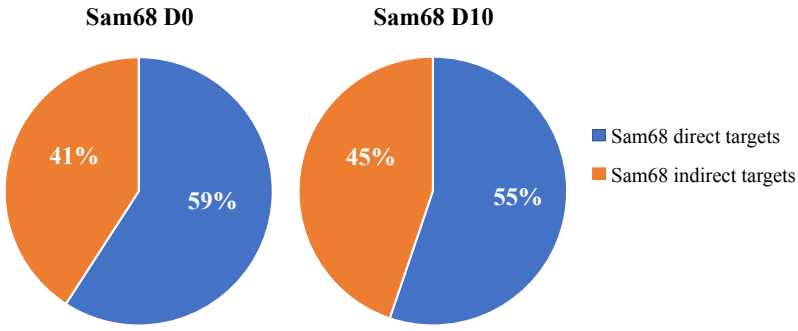
Figure 21. Pie charts representing genomic binding preferences for Sam68 and QKI, at D0 and D10 time points during mESCs differentiation.

We integrated the circRNA analysis data with the iCLIP2 experimental output and verified how many of the downregulated circRNAs retrieved upon Sam68 depletion were direct targets of our RBP of interest.

Strikingly, as shown in Figure 22A, at D0 the 59% of the identified circRNAs is a direct target of Sam68, while at D10 there is a slight decrease of circRNA directly bound. As a proof of concept for Sam68 binding specificity of circRNA direct targets, some examples retrieved from the Genome Browser are represented in Figure 22B.

Our next step is to identify the Sam68 binding motif and pattern in circRNA mRNAs.

A



B

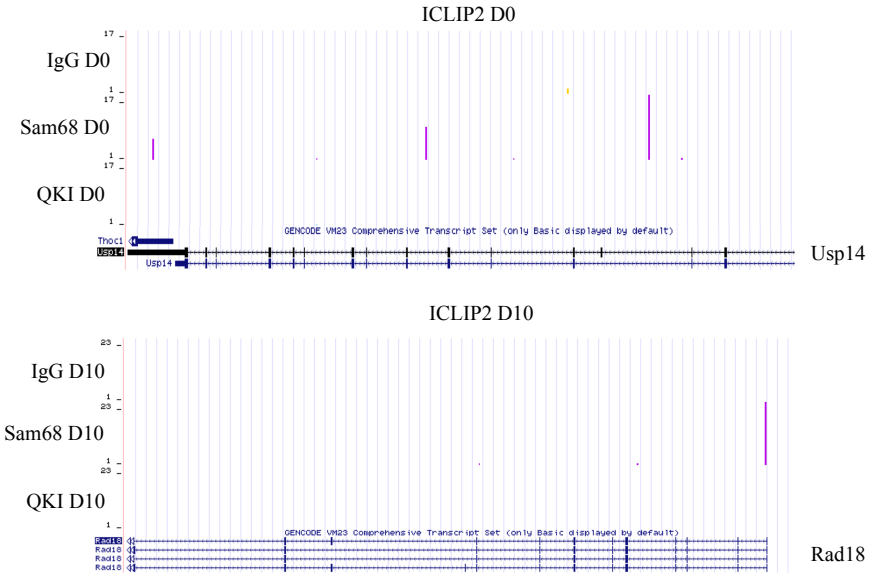


Figure 22A. Pie charts indicating percentages of Sam68 direct targets among circRNA genes highly expressed in mouse cardiomyocytes at D0 and D10 time points during mESCs differentiation. **22B.** Genome browser captures of Sam68 and QKI iCLIP2 experiment on direct mRNA targets in common with the CircRNA analysis. Sam68 binding is highlighted in purple.

4.4.3 CatRAPID vs iCLIP2

Given the availability of iCLIP2 data, we wanted to further confirm that the bioinformatic predictor catRAPID was able to reproduce our experimental output and, in order to achieve that, we considered the results from iCLIP2 as a positive set and used as a negative set a background of the same size obtained from the random selection of transcripts in the mouse transcriptome.

AUC (Area Under the Curve) scores were calculated for each condition (Sam68 D0, Sam68 D10, QKI D0 and QKI D10) and are represented in the bar plot below (Figure 23), confirming that our predictor catRAPID can reproduce iCLIP2 experiment in big extent.

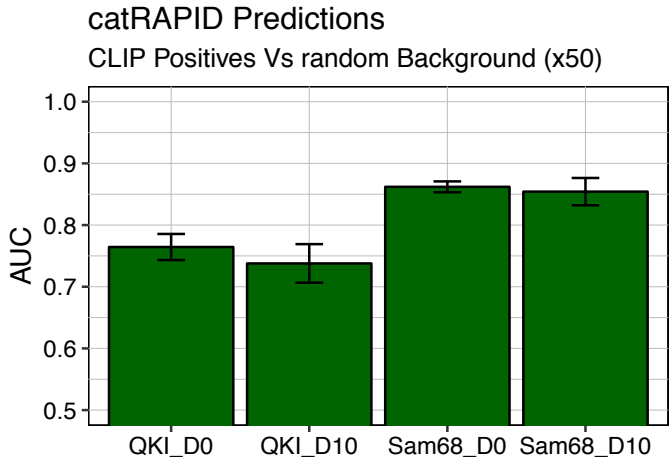


Figure 23. Bar-plots representing catRAPID predictions for Sam68 and QKI in agreement with iCLIP2 results. A background of the same size obtained from the random selection of transcripts in the mouse transcriptome is used as negative control.

5. STAR proteins regulate cardiomyocyte differentiation by enhancing Gata4 translation

RNA binding proteins (RBPs) are key players in post-transcriptional events and could play a central role in the balance between pluripotency and differentiation. The Signal Transduction and Activation of RNA (STAR) family is a group of RBPs involved in mammalian development. Notably, we found that two STAR proteins, Sam68 and Quaking (QKI), modulate mouse embryonic stem cells (mESCs) cell cycle, self-renewal and cardiomyocyte differentiation. We discovered that Sam68, which acts antagonistically to QKI, directly regulates the expression of Gata4, a master regulator of mesodermal development, the embryonic germ layer which gives rise to cardiac tissues.

STAR proteins regulate cardiomyocyte differentiation by enhancing Gata4 translation

Alessandro Dasti[1-2], Maria Carla Antonelli[1], Magdalena Arnal[1-2], Alexandros Armaos[1-2], Sarah Bonnin[1], Thomas Graf[1-3], Maria Paola Paronetto[4,5], Tian V Tian[1-6], Elias Bechara[1-2], Gian Gaetano Tartaglia[1-2-3-7].

1- Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain
2- Department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy
3- Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, 08010, Barcelona, Spain
4- Institute for Treatment and Research (IRCCS) Fondazione Santa Lucia, Rome, Italy
5- Department of Movement, Human and Health Sciences, University of Rome "Foro Italico," Rome, Italy
6- Vall d'Hebron Institute of Oncology (VHIO), C/ Nazaret 115-117, 08035 Barcelona, Spain
7- Department of Biology "Charles Darwin", Sapienza University of Rome, P.le A. Moro, Rome, 00185

Summary

RNA binding proteins (RBPs) control many post-transcriptional events and could play a central role in regulating the cell stem state. The Signal Transduction and Activation of RNA (STAR) family is a group of proteins involved in mammalian development. Using CRISPR-Cas9 generated knock-out lines, we found that two STAR proteins, Sam68 and Quaking (QKI), govern mouse embryonic stem cells (mESCs) cell cycle, self-renewal and cardiomyocyte differentiation. RNA sequencing and ribosome profiling experiments combined with advanced computational models allowed us to prioritize Sam68 and QKI interactions involved in the regulation of cardiogenic factors. We discovered that Sam68, which acts antagonistically to QKI, binds directly to the cardiogenic-specific transcription factor Gata4 enhancing its translation. By unveiling details of QKI and Sam68 molecular networks, we shed light on how these RBPs act as master regulators in mESC differentiation.

Introduction

Embryonic stem cells (ESC) are an outstanding laboratory tool to study the early phases of mammalian development beside being promising in both therapy and disease modeling[1,2]. The mechanisms underlying the self-renewal and pluripotency properties of the ESCs have been subject of intensive studies in recent years[3–5]. Upon differentiation stimuli, the ESCs actively exit from the pluripotent state by disrupting the core pluripotency transcription factors (TFs) network and by expressing lineage-specific transcription factors that in turn activate gene expression programs that lead to differentiation[6]. Beside transcription factors and chromatin remodelling factors, recent advances in *omics* techniques indicate the existence of a post-transcriptional regulation layer that can modulate pluripotency and lineage commitment[7–9]. More precisely, RNA Binding Proteins (RBPs) have been shown to be major players regulating the different steps of the RNA metabolism in this process[10–14]. Specifically, FOX2, SON, SFRS2, MYC, GCN5, ZCCHC24, and RBM47 facilitate pluripotency-specific alternative splicing (AS) of their target genes[15–18]. In contrast, MBNL1, MBNL2, and SFRS11 promoted differentiation-specific AS patterns for a large number of splicing events[19]. Furthermore, RBM24 enhanced cardiac-specific splicing favoring the cardiomyocyte differentiation of ESCs[20]

Another family of RBPs including KHDRBS1, also known as Sam68, and Quaking (QKI), was shown to be involved in mammalian development. Sam68 plays a critical role during male germ cells development by enhancing the translation of RNAs that regulate the spermatogenesis[21,22]. Sam68 is also important in fine tuning the switch between self-renewal and differentiation of the neuronal progenitor cells (NPCs)[23]. In parallel, QKI is indispensable for correct axonal myelination by controlling both AS, localization and transport of myelinogenesis-associated RNAs[24]. The Sam68$^{-/-}$ mouse model is associated with male infertility and female mammary gland tumor[25,26]. In contrast, the QKI$^{-/-}$ mouse model is not viable due to severe cardiovascular defects suggesting an important role of this protein during cardiac tissue development[27–29].

In this study, we decided to investigate the role of Sam68 and QKI in the early stages of mammalian development. We show that the depletion of either proteins by CRISPR-Cas9 has distinct effects on ESCs self-renewal and proliferation. Using RNA sequencing we identified several pathways involved in heart development to be mis-regulated in Sam68$^{-/-}$ or QKI$^{-/-}$ cells that failed to form *bona fide* cardiomyocytes. Combining computational analysis and ribosome profiling technique we found hundreds of transcripts, including many implicated in proper cardiomyocyte differentiation, which are directly bound and/or whose translation is modulated by Sam68 or QKI. We specifically detected the transcription factor Gata4 to be a direct target of Sam68 and that its translation is regulated by the latter. Our findings revealed a key circuit of posttranscriptional gene regulation operated by RBPs that can be disrupted during cardiomyocyte differentiation.

Results

Sam68 and Quaking positively regulate cell cycle and self-renewal in ESCs

ESCs (E14) depleted of either Sam68 or QKI were generated using the CRISPR-Cas9 technology by targeting the first exon of the genes which resulted in complete abolishment of the proteins compared to wild type cells (**Figure 1A**). Although they belong to the same RBP family, the absence of any of the two proteins did not affect the expression levels of the other (**Figure S1A**). In addition, the KO stem cell colonies showed normal domed-shape morphology with net borders and did not have any apparent anomaly (**Figure S1B**).

To dissect the potential involvement of Sam68 and QKI in pluripotency and self-renewal, both WT and KO cells were first synchronized and their proliferation was monitored through 72 hours. Depletion of either protein resulted in a pronounced decrease in the proliferation rate (**Figure 1B**). Consequently, in the clonogenic assay, the KO cells presented a significantly reduced area (**Figures 1C** and **1D**) and number (**Figures 1C** and **1E**), indicating that both Sam68 and QKI positively sustain the self-renewal of the mESCs. This observation is consistent with a previous study associating the role of Sam68 in the self-renewal of neuronal progenitor cells (NPCs)[23].
On the contrary, in the exit from pluripotency assay, a Rex1-dGFP reporter cell line was used[30]. Sam68 and QKI were knocked-down (KD) by lentiviral infection (Figure S1C) and the dynamics of pluripotency exit was measured by FACS after growing the cells in differentiation-permissive medium. The difference between the KD and the control cell lines was not statistically significant suggesting that none of the two proteins is involved in this process (**Figure 1F**). As an ultimate step in the cellular characterization of the KO cell lines, *in vivo* teratoma assay was performed. The teratomas harbored differentiated cells derived from the three germ layers suggesting that the depletion of either protein does not alter the pluripotent capacity of the mESCs *in vivo* (**Figure 1G**). Of note, an increase in mesoderm-derived cells in Sam68$^{-/-}$ line was observed (**Figure S1D**) suggesting a role of Sam68 in the mesodermal differentiation.

Transcriptional and post-transcriptional landscape driven by Sam68 and QKI during ESCs differentiation

To assess the roles of Sam68 and QKI during early mammalian development, an embryoid body (EB) differentiation assay was set and RNA isolated from three biological replicates in different time points was used to perform total RNA-sequencing. Sam68 and QKI protein levels were checked throughout the differentiation in wild type and knockout cells (**Figure S2A**). Significant and consistent changes in gene expression were detected compared to wild type as shown in the principal component analysis (PCA) (**Figures S2B** and **S2C**). Notably, the expression of all the classes of coding and non-coding RNAs is affected at each time-point of differentiation (**Figure 2A**). Moreover, the extent of the deregulation

is higher in Sam68$^{-/-}$ compared to QKI$^{-/-}$ for both protein coding (**Figure 2B**) and non-coding RNAs (**Figure S2D**) (Lncenc1 and?), suggesting a broader role of this protein in regulating all the RNA types during ESCs differentiation. Interestingly, the overlap between differentially expressed RNAs for the two proteins in each time-point is rather limited (**Figure 2C**), revealing distinct regulatory functions of these factors in the differentiation towards various lineages.

At the post-transcriptional level, both Sam68 and QKI were previously described to regulate the AS in different developmental processes[24,31–33]. The global changes of the AS in Sam68$^{-/-}$ and QKI$^{-/-}$ cells were measured. Despite the large number of deregulated events in each cell line at each time point, the overlap between the two KO lines was fewer than 5%, indicating divergent roles of the two proteins in the AS regulation in pluripotent and differentiating mESCs (**Figure 2D**). This pattern of regulation has been depicted for members of the SR RBPs family[34,35] and the RBM family[36]. Surprisingly, in the absence of Sam68 the majority of the altered AS events belong to the category of intron retention, in line with what was previously described for specific Sam68 targets[37,38] whereas cassette exon and alternative 3' or 5' splice site choices appeared to prevail in QKI$^{-/-}$ (**Figure 2E**) suggesting different mechanism through which Sam68 and QKI regulate the AS.

Sam68 and QKI regulate cardiac-specific RNAs and cardiomyocyte differentiation

Gene Ontology (GO) analysis on the differentially expressed genes at day 10 of EBs was performed. In addition to genes involved in neurogenesis and vasculature development as previously described[24,33,39] (**Figures S3A** and **S3B**), an over-representation of cardiac related terms was observed suggesting a novel involvement of these factors in the cardiac development and in a variety of biological processes. Surprisingly, while the majority of cardiac-related RNAs were upregulated upon the depletion of Sam68, a complete opposite trend was observed in QKI$^{-/-}$ EBs (**Figure 3A**).

The general downregulation of cardiomyocyte's structural proteins (e.g. Actn2, Mybphl, Nebl, Myh6, Actc1, Tmod1, Tnni3, Tnnc1, Tnnt3)[20,40–45], functional proteins, (e.g. ion channels Ryr2 and Scn5A)[46–48], and cardiac-specific splicing regulators (Rbm20, Rbm24)[20,49] observed in QKI$^{-/-}$ was further confirmed by RT-PCR (**Figure 3B**). This downregulation can be explained by the significant down-regulation of master cardiac TFs Gata4, Gata6 and Mef2c, detected in the earlier stages of differentiation in the QKI KO line (**Figure 3C**). Indeed, these TFs are responsible for the expression of many of the genes that are down-regulated at day 10 of EBs differentiation[50,51].
Conversely, in absence of Sam68, both significant up-regulation of Gata4, Nkx2-5 and Tbx18 and down-regulation of Isl1 were observed (**Figure 3D**). This general imbalance of cardiac progenitor specific TFs[52] is likely to be the reason underlying the general up-regulation of transcripts encoding structural (Tnnt2, Myl2, Myl4, Actc1)[53,54] and functional proteins (Cacnb2[55], Ryr2) of the cardiac cells (**Figure 3D**).

The deregulation at the expression level came along with an altered pattern of AS in both Sam68$^{-/-}$ and QKI$^{-/-}$ cell lines. More specifically, the absence of QKI affected the correct inclusion of several cassette exons of structural and functional cardiac components (**Figure S3C**) whereas the absence of Sam68 led to the aberrant splicing of different exons of the calcium channel subunit, Cacna1c, which is highly important for the excitability of the cardiomyocytes (**Figure S3D**)[56].

To check the functional defect signature of these molecular abnormalities EBs were differentiated into *bona fide* cardiomyocytes and the formation of the consequent spontaneous beating *foci* was quantified after 8 days. In the absence of either QKI or Sam68 a significant reduction of beating *foci* compared to the WT control was observed (**Figure 3E**), indicating an important role of these two proteins in this specific differentiative path. This phenocopies the features of the *in vitro* Gata4$^{-/-}$ derived cardiomyocytes[57]. In addition, the *foci* generated in the absence of Sam68 showed an impairment in their beating activity by contracting faster compared to the WT counterparts (**Figure 3F**). This phenotype could be related either to the aberrant expression of Ryr2 or the altered AS of Cacna1c or a combination of both. On the contrary, QKI KO beating *foci* showed a dramatic impairment to make impossible any quantification. To sum up, all these data indicate that both Sam68 and QKI are important factors in the regulation of the cardiomyocyte differentiation of mESCs through the control of cardiac-related RNAs expression, AS and their subsequent translation.

Sam68 directly binds GATA4 mRNAs and regulates its translation

The up-regulation of Gata4 mRNA in Sam68$^{-/-}$ was expected to be accompanied by more abundant cardiomyocyte differentiation[58], however the observed phenotype rather reflected its absence[57]. This discrepancy could be due to translational regulation. Indeed, both Sam68 and QKI were previously described to regulate translation[21]. To assess the global variation in translation of Sam68$^{-/-}$ and QKI$^{-/-}$, ribosome profiling was performed at day 10 of EBs. This technique snapshots the actively translated mRNAs[59]. The majority of the obtained clusters aligned to coding sequences indicating high-quality libraries (**Figure S4A**). Cardiac transcripts whose translation was significantly affected were found in Sam68$^{-/-}$ but not in QKI$^{-/-}$ suggesting that the mechanism by which QKI controls cardiac differentiation is translation independent (**Figures 4A** and **S4B**). Sam68 exerts a translation repressive effect on ion channels (Kcnj8)[60], structural proteins (Myh6) and TFs (Nkx2-5, Osr1)[61]. Conversely, it enhances the translation of the mesoderm specifier transcription factor Sox17[62], Ihh, a mediator of the Indian Hedgehog signaling pathway indispensable for the embryonic development of the heart[63] and most importantly of Gata4 (**Figure 4A**) which was further confirmed by western blot and immunofluorescence (**Figures 4B, 4C** and **S4C**).

To unravel whether this functional regulation was the result of a direct binding, the interaction propensity of the affected cardiac mRNAs was calculated for Sam68 and QKI compared with a repertoire of 1799 RBPs as a control using the *cat*RAPID

algorithm[64,65](**Figure 4D**). In fact, cardiac mRNAs interaction propensities were higher for Sam68 than QKI, falling in the top 20% of all RBPs. In addition, Sam68 shows an interaction to Gata4 mRNA around 90% higher than the rest of the transcriptome (**Figure 4E**). This interaction is predicted to occur in a region spanning nucleotides 500 to 1000 of the Gata4 mRNA corresponding to part of the 5'untranslated region (UTR) and the first exon of the transcript (**Figure 4F**). This prediction was confirmed by both electrophoretic mobility shift assay (EMSA) (**Figure 4G**) and RNA Immunoprecipitation (RIP; **Figure S4D**). To conclude, during mESCs differentiation towards the cardiomyocyte lineage, Sam68, unlike QKI, binds directly and control the translation of the cardiogenic transcription factor Gata4.

Discussion

Here we identified two STAR members that have overlapping, but yet distinguishable roles in both embryonic stem cells pluripotency and differentiation. Our results indicate that both proteins positively regulate mESCs proliferation, self-renewal and are indispensable for proper differentiation towards the cardiomyocyte lineage by regulating expression, alternative splicing and translation of transcription factors, structural and functional cardiac-specific proteins. In particular, we show that Gata4 is a direct target of Sam68 which regulates its physiological protein levels.

Sam68 and QKI in self-renewal and proliferation

Clonogenic assay demonstrated that both Sam68 and QKI are involved in proliferation and self-renewal of mouse embryonic stem cells. These proteins have been reported to regulate cell proliferation[29,66] and our findings go in line with previously reported prolonged G2-M phase in Sam68 depleted chicken fibrobalsts[67] and an altered cell proliferation in absence of QKI in colon cancer cells[68]. One of the molecular mechanisms orchestrating the reduced self-renewal capacities of both $QKI^{-/-}$ and $Sam68^{-/-}$ mESCs could be attributed to the decreased levels of Wnt3 and its direct downstream target Lef1, known to promote self-renewal[69]. Indeed, we observed the downregulation of both Wnt3 in Sam68 deficient cells and Lef1 at the transcription and translation levels in $QKI^{-/-}$ and $Sam68^{-/-}$ respectively (**Figures S1E** and **S1F**). On the other hand, mESCs use glycolysis as a main source of ATP production and switch to oxidative phosphorylation upon differentiation[70]. This metabolic switch mechanism was shown to connect Sam68 to self-renewal in neuronal progenitor cells[23]. In addition, our results demonstrate that the non-coding RNA, Lncenc1, recently shown to be important for the expression of glycolysis-associated genes[71] to be consistently downregulated in both $QKI^{-/-}$ and $Sam68^{-/-}$ cells (**Figures S1E** and **S1F**).

Role in cardiac differentiation and role of other RBPs

RNA-sequencing at both early and late stages of embryoid bodies (EBs) differentiation in Sam68 and QKI depleted cells demonstrated their implication in the nervous tissue development as shown[23,29]. Strikingly, we also revealed novel pathways that were not previously described. Among these we found terms containing genes encoding important functional and structural cardiomyocytes-specific proteins that are regulated in opposite directions in Sam68 and QKI KO cells. This suggests that not only both proteins are involved in the development of the heart but that probably they exert different functions in this biological context. This might not be surprising as it has been shown that RBPs belonging to the same family can exert antagonistic functions in the same process[35,36]. In detail, QKI acts early on the core TFs network at day 3 of EBs differentiation. Although at this stage these TFs are lowly expressed[72], we observed a dramatic downregulation of Gata4 and Gata6 mRNAs. This downregulation was further extended to several genes that codify for structural, functional and cardiac-specific splicing regulators (Rbm24 and Rbm20) at day 10. The latter could be, at least partially, the reason underlying the aberrant alternative splicing landscape occurring in the cardiac specific transcripts[20,49]. These extensive defects in the very early stages of differentiation, might be the reason underlying the lack of beating *foci* formation as well as the lethality of the $QKI^{-/-}$ embryos. Conversely, in the case of $Sam68^{-/-}$ cells, the majority of the defects were observed at D10 of EBs differentiation. In fact, we detected a dramatic upregulation of the mRNA levels of cardiac transcription factors (Gata4, Nkx2-5 and Tbx18) that form a cascade that drives precursor cells towards a cardiomyocyte cell identity[73,74]. In addition, Tbx18 is known to be crucial for the formation of the sinoatrial node cells that initiate the electric impulse and stimulate the contraction of the organ[52]. To the same extent, the depletion of Sam68 caused aberrant upregulation of structural proteins (Tnnt2, Actc1) and the ion channels (RyR2 and Cacnb2) important for the cardiomyocyte depolarization and their consequent contractions[75,76]. Moreover, Cacna1c, another important ion channel subunit[77], showed altered pattern of several cassette exons inclusions in absence of Sam68. All together, these data can explain the abnormal beating activity that we observed upon Sam68 depletion.

Sam68 and Gata4

Pre-mRNA splicing, localization and stability were shown to be the molecular mechanisms through which QKI regulates neuronal development[24,78–81]. Our results show that, unlike QKI, Sam68 can exert its role in cardiomyocyte differentiation through translation regulation as well. Sam68 as translation enhancer was previously described in male germ cells differentiation[21,22]. Here we showed, for the first time, that the translation regulation mechanism through Sam68 is conserved in the development of the cardiac tissue. Nevertheless, our results presented a puzzling case where Gata4 showed sharp up and downregulation of its mRNA and protein levels respectively in the absence of Sam68. Moreover, we have shown that Sam68 binds directly to Gata4 mRNA and

acts as an enhancer of its translation. Two possible hypotheses can be at the basis of this mis-regulation, the nuclear export and the transcription-translation feedback loop. One of the known functions of Sam68 is the nuclear export of the HIV transcripts into the cytosol of the infected cells in order to be translated[82]. By a similar mechanism and in normal conditions, Sam68 would enhance the nuclear export of Gata4 mRNA consequently allowing its proper translation. Conversely, in the absence of Sam68, the majority Gata4 mRNA might be retained in the nucleus, thus being less available to actively translating ribosomes leading to low protein production. In parallel, a positive transcription-translation feedback loop previously described for circadian oscillators[83] can also explain this puzzling scenario: In the absence of Sam68, Gata4 mRNA is poorly translated, this induces an increase in Gata4 mRNA transcription by either a transcription enhancer cofactor or through direct binding of GATA4 to its own promoter[84]. Furthermore, the translation of Gata4 is crucial for the proper spermatocyte differentiation[85] which is highly compromised in Sam68$^{-/-}$ mice[21,22]. To date, this is the first evidence of post-transcriptional regulation of Gata4 mRNA by a specific RNA binding protein. Consequently, this can shed more light on the understanding of the Sam68$^{-/-}$ infertility phenotype through Gata4 regulation. To sum up, Sam68 and QKI regulate a circuit of genes involved in the development of the cardiovascular system that could be the cause of the embryonic lethality in QKI$^{-/-}$ and the high perinatal mortality of Sam68$^{-/-}$.

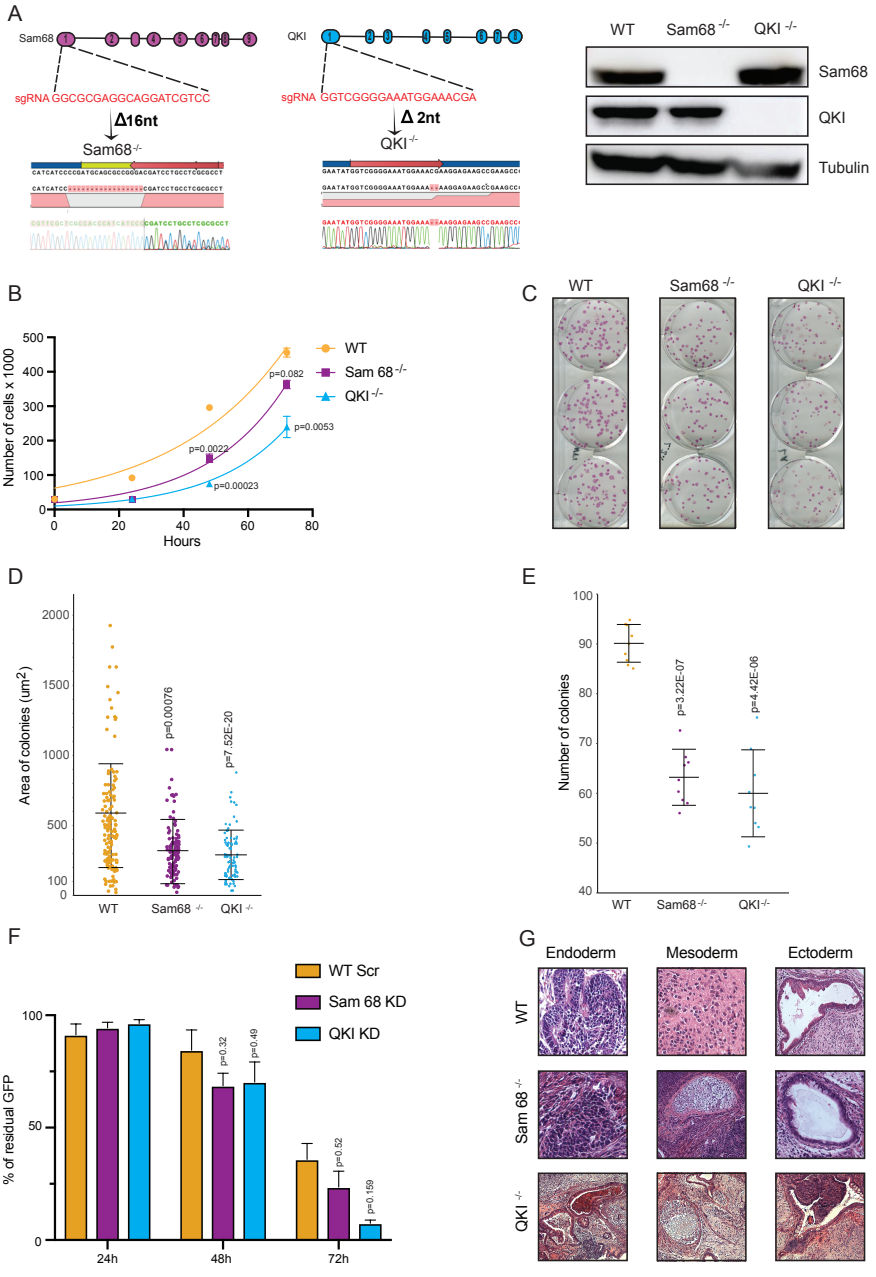


Figure 1. Sam68 and Quaking positively regulate cell cycle and self-renewal in ESCs

Figure 1. Sam68 and Quaking positively regulate cell cycle and self-renewal in ESCs

A. CRISPR-Cas9 design strategy targeting the first exon of either Sam68 or QKI genomic locus in order to generate E14 mESC KO lines. The induced frame shift in both cases is shown in the Sanger sequencing of the respective cell lines. On the right: Western blot validating the absence of either protein.
B. Proliferation assay of WT, Sam68$^{-/-}$ and QKI$^{-/-}$ mESCs. Each time point is the average of three independent biological replicas. The p-values obtained by student t test are indicated for each significative point.
C. Clonogenic assay pictures of WT, Sam68$^{-/-}$ and QKI$^{-/-}$ mESCs
D. Area of the colonies in μm^2. Each column represents the results obtained by performing the experiment in three biological replicas, p-values for both Sam68 $^{-/-}$ and QKI$^{-/-}$ are indicated.
E. Total number of colonies obtained after plating 200 cells per well. Results are represented for the three biological triplicates
F. Exit from pluripotency. The bar plot shows the percentage of residual GFP (pluripotent cells) at different time points.
G. Teratoma assay indicating the presence of terminally differentiated cells derived from ectoderm, mesoderm and endoderm

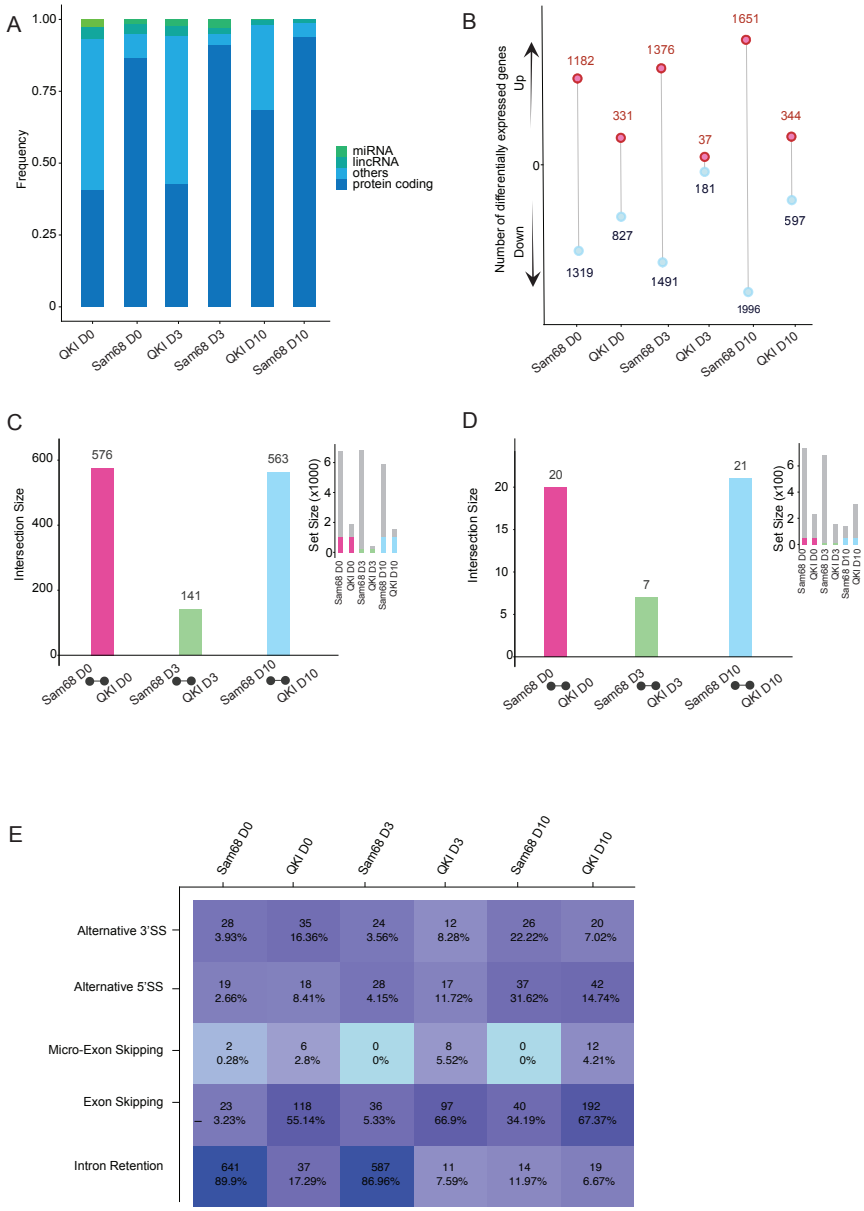


Figure 2. Transcriptional and posttranscriptional landscape driven by Sam68 and QKI during ESCs differentiation

Figure 2. Transcriptional and posttranscriptional landscape driven by Sam68 and QKI during ESCs differentiation

A. Different classes of differentially expressed RNAs in either Sam68$^{-/-}$ or QKI$^{-/-}$ mESCs at day 0, day 3 and day 10 of EBs differentiation. Only RNAs with ± 1,5 fold change and an adjusted p value<0,01 are represented.
B. Numbers of differentially expressed protein coding RNAs at each time point for each KO cell line
C. Common deregulated coding RNAs at each time point of both Sam68$^{-/-}$ and QKI$^{-/-}$ cells. The size of the set is represented on the right.
D. Common aberrant alternative splicing events in the two cell lines at day 0, day 3, day 10 of EBs differentiation. The size of the set is shown on the right
E. Heatmap showing the different types of aberrant AS events in the absence of either QKI or Sam68 during EBs differentiation of mESCs. Only events with a DPSI>+25% or <-25% and an adjusted p value<0,01 are considered.

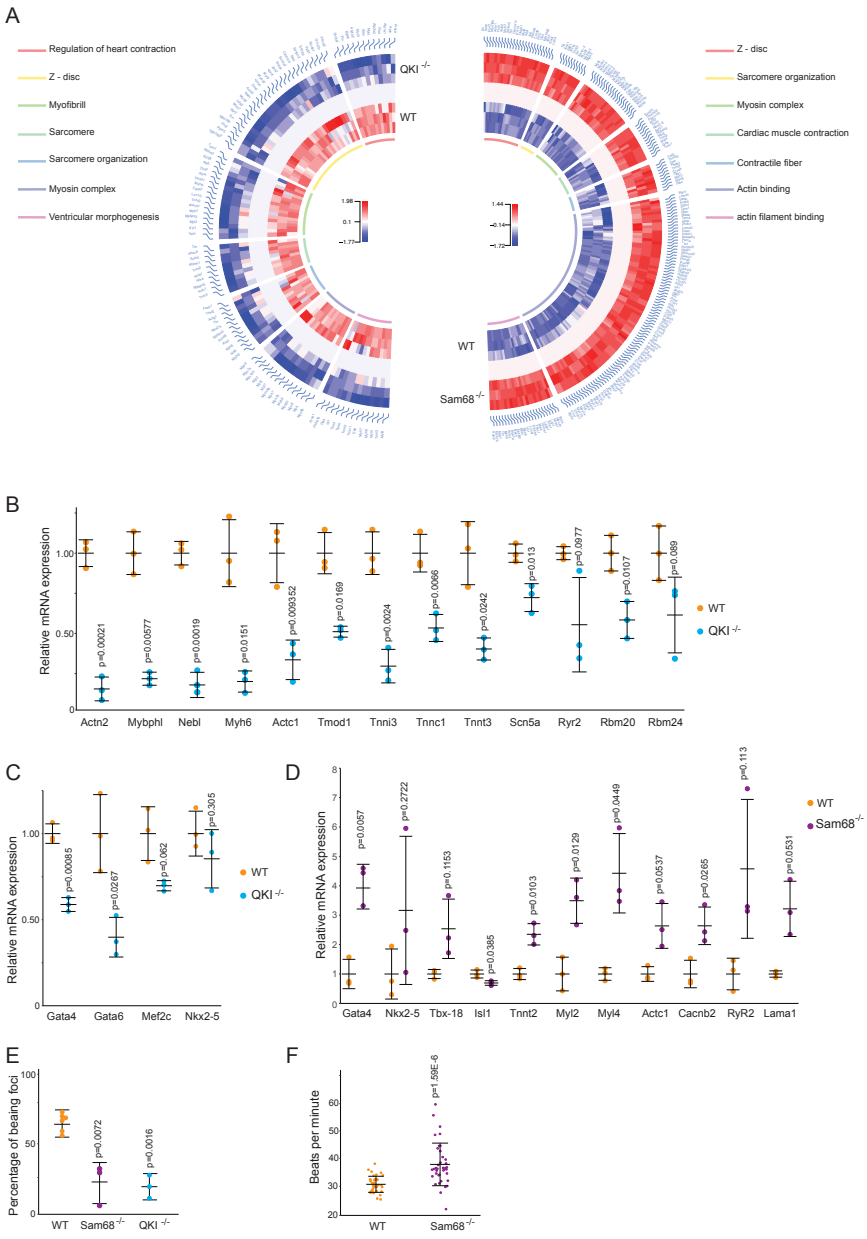


Figure 3. Sam68 and QKI regulate cardiac-specific RNAs and cardiomyocyte differentiation

Figure 3. Sam68 and QKI regulate cardiac-specific RNAs and cardiomyocyte differentiation

A. Circle plot of selected GO gene-sets related with cardiac development at day 10 of EBs differentiation. RNAs belonging to this category are mainly downregulated for QKI$^{-/-}$ cells (on the left) and upregulated for Sam68$^{-/-}$ cells (on the right). GO classes are indicated in the respective color legends
B. RNA-sequencing validation of cardiac related transcripts in QKI$^{-/-}$ EBs at day 10 of differentiation via real-time PCR shows severe downregulation of these RNAs.
C. Real-time PCR shows a significant downregulation of Gata4 and other cardiogenic transcription factors in QKI$^{-/-}$ cells at day 3 of EBs differentiation
D. RNA-sequencing validation via real-time PCR of the altered expression of cardiac-related mRNAs in Sam68$^{-/-}$ cells at day 10 of EBs differentiation. P-values are indicated for each validated transcript.
E. Percentage of beating *foci* generated after 8 days of cardiomyocyte differentiation. KO lines were compared to WT cells and the experiment was repeated in 3 independent replicas
F. Quantification of the beating activity of WT and Sam68$^{-/-}$ beating *foci* expressed in beats per minute

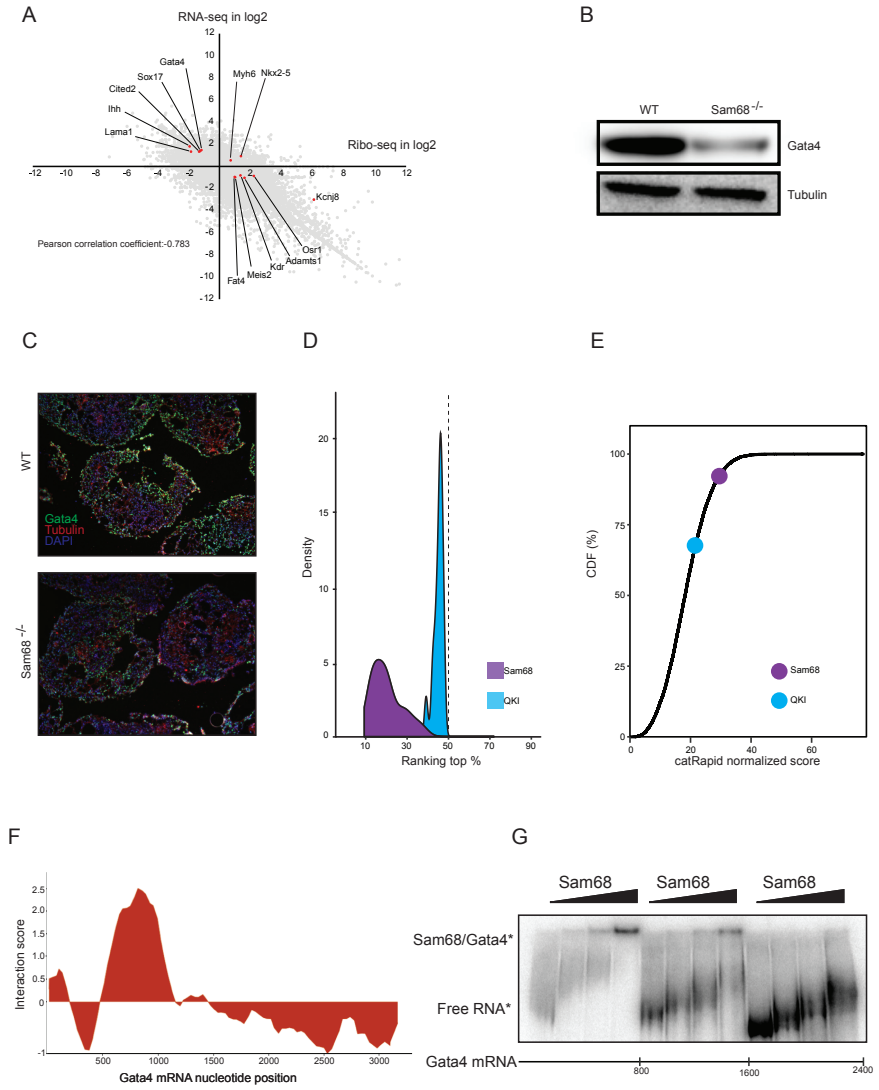
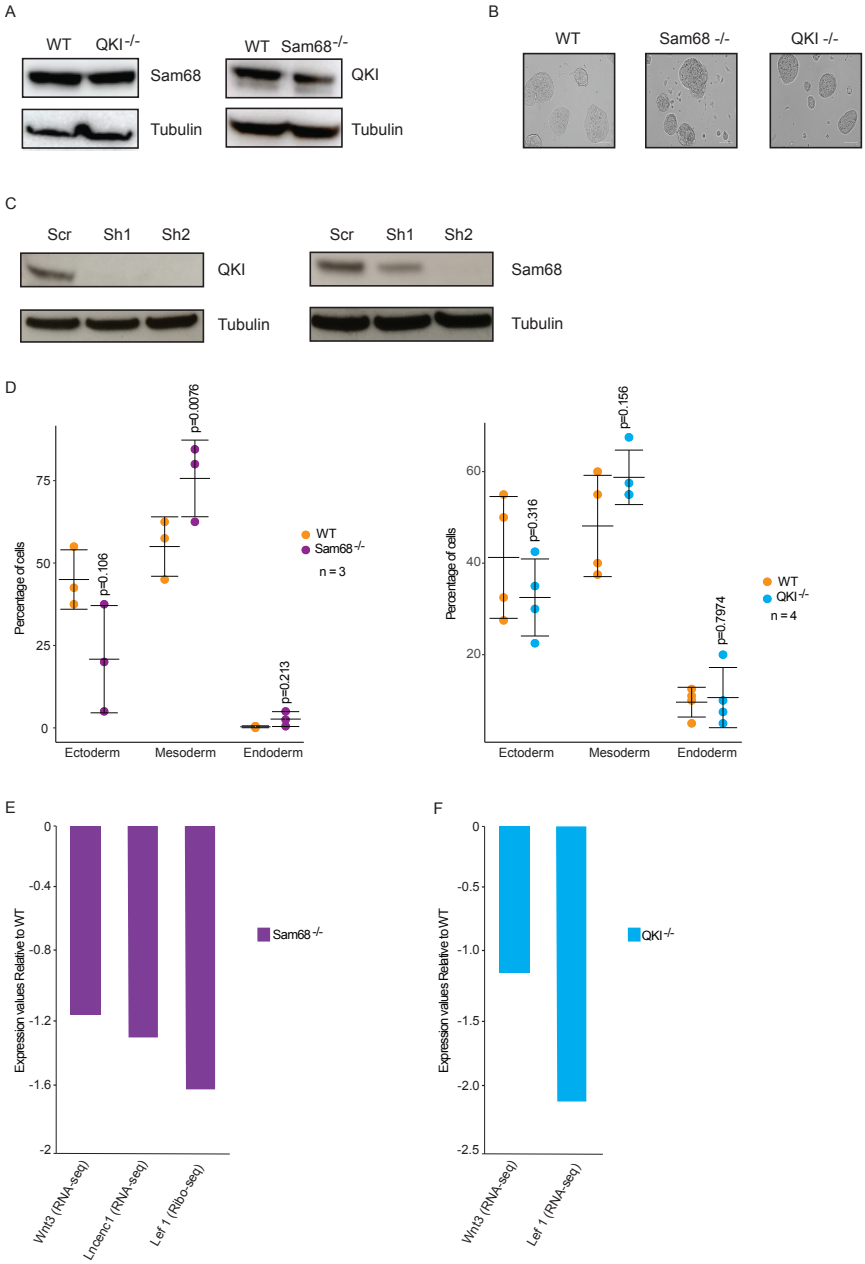


Figure 4. Sam68 directly binds GATA4 mRNAs and regulates its translation

Figure 4. Sam68 directly binds GATA4 mRNAs and regulates its translation

A. Transcription-translation correlation plot in Sam68$^{-/-}$ EBs at day10 of differentiation. Cardiac-related RNAs whose translation is significantly affected are highlighted.
B. Western blot showing a dramatic reduction of GATA4 protein levels in Sam68$^{-/-}$ EBs.
C. Immunofluorescence showing reduced GATA4 (in green) levels in the absence of Sam68.
D. Plot showing the interaction propensities of Sam68 and QKI to cardiac-related transcripts whose translation is affected in absence of Sam68 in differentiating EBs compared to a pool of 1799 RBPs.
E. Plot showing the cumulative distribution function (CDF) of the *cat*RAPID score calculated for QKI and Sam68 to the mRNA mouse transcriptome with RNAs. The dots represent the binding propensity of either QKI or Sam68 to Gata4 RNA. The RNA considered in this plot range from 3000 to 4000 nucleotides (i.e. Gata4 length to avoid potential biases arising from transcript sizes).
F. Interaction propensity between Sam68 protein and Gata4 RNA as predicted by *cat*RAPID.
G. EMSA shows that Sam68 binds to the first part of Gata4 RNA as demonstrated by the shift of the Sam68-Gata4 RNA complex.

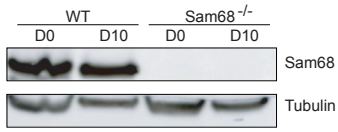


Supplementary Figure 1

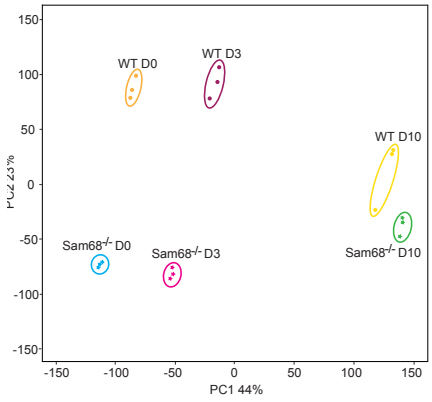
Supplementary Figure 1

A. WB showing absence of compensation at the protein level of one member of the family in the KO background of the other one.
B. KOs and WT mESCs colonies in bright-field microscopy.
C. WB showing the KD protein levels of either QKI or Sam68 in Rex1-dGFP cells.
D. Quantification of the teratoma derived tissues obtained by injecting WT and either Sam68$^{-/-}$ (on the left) or QKI$^{-/-}$ (on the right) mESCs in SCID mice.
E. mRNA expression and RNA translation efficiency of mESCs pro-proliferative factors in Sam68$^{-/-}$ mESCs compared to the WT control.
F. mRNA expression and RNA translation efficiency of mESCs pro-proliferative factors in QKI$^{-/-}$ mESCs compared to the WT control.

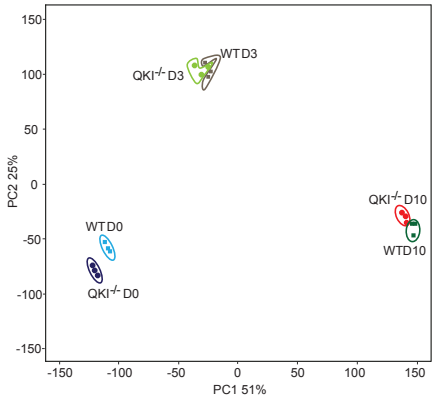
A



B



C



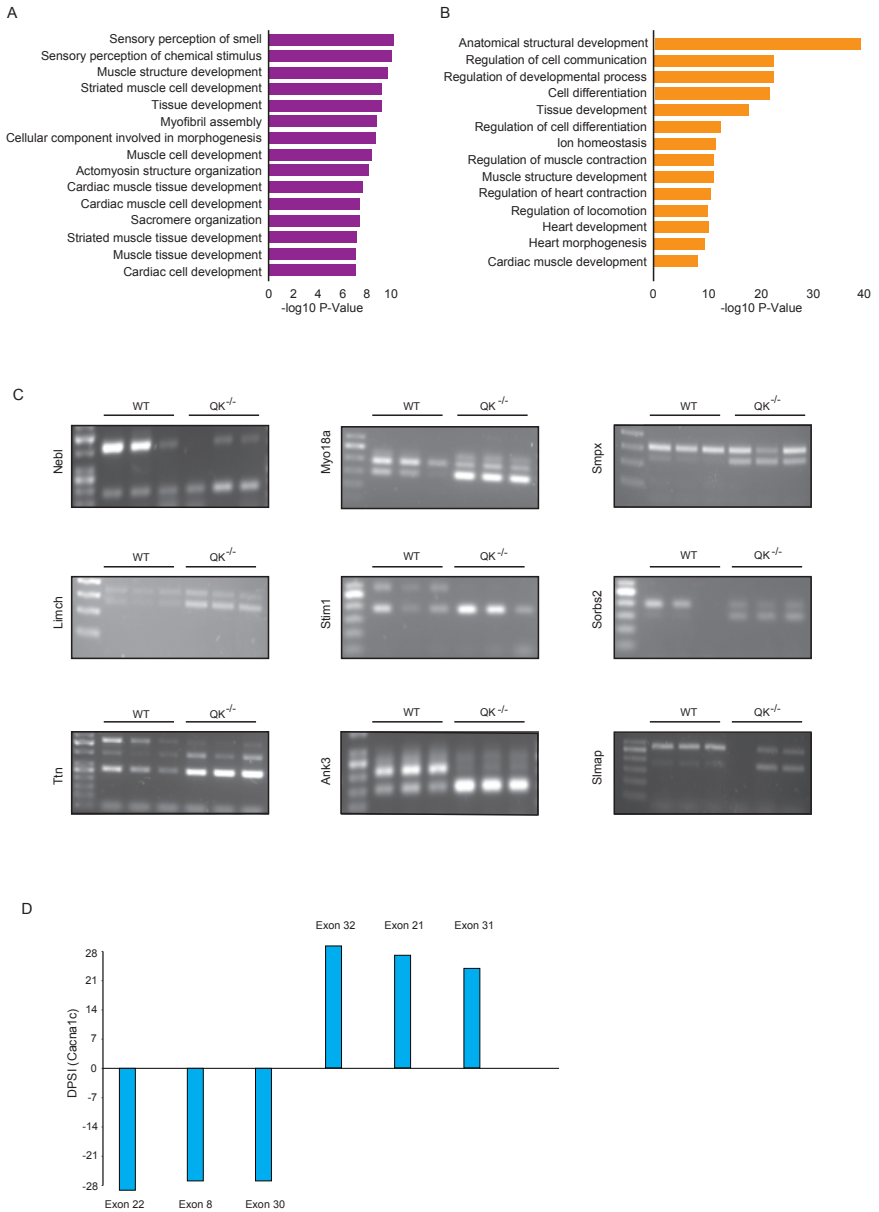
D

	Protein Coding	lnRNAs	miRNAs	Others
Sam68$^{-/-}$ D0	2501	51	19	112
QKI$^{-/-}$ D0	1158	242	155	3456
Sam68$^{-/-}$ D3	2867	5	7	10
QKI$^{-/-}$ D3	218	104	7	1569
Sam68$^{-/-}$ D10	3647	14	1	49
QKI$^{-/-}$ D10	941	19	0	46

Supplementary Figure 2

Supplementary Figure 2

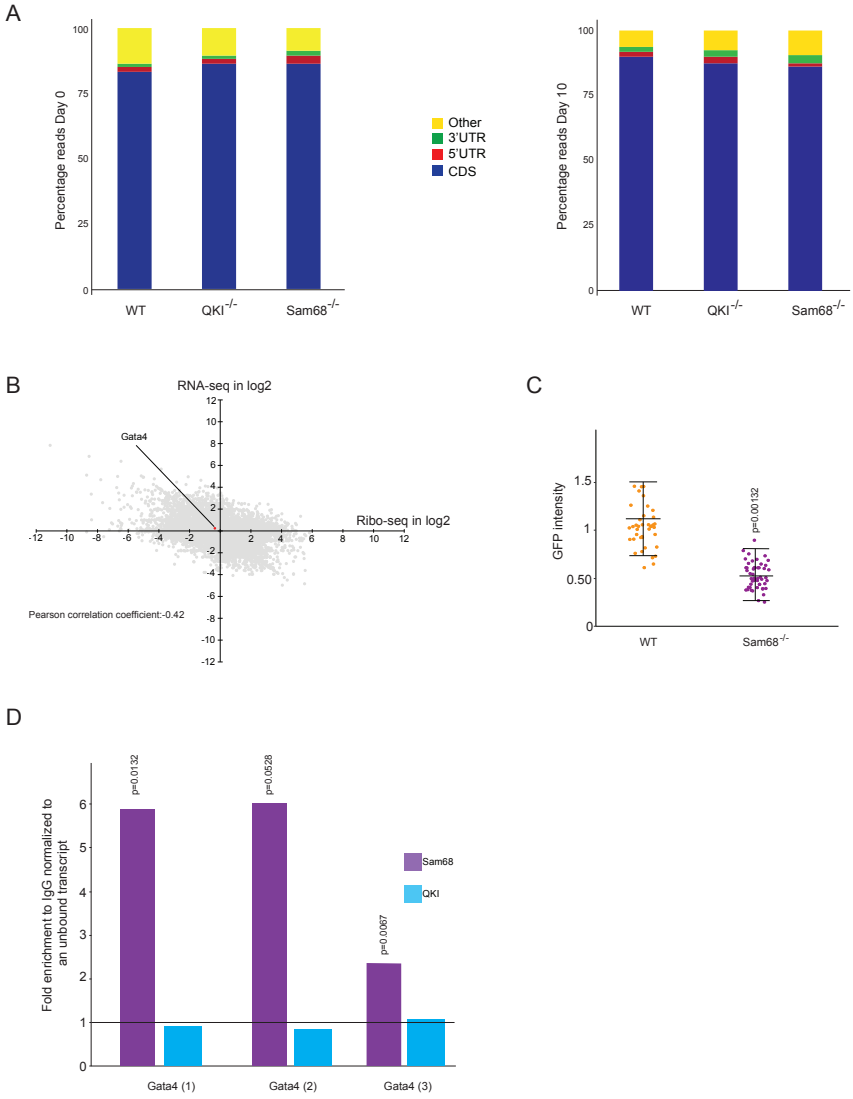
A. WB showing the expression levels of Sam68 in WT and Sam68 KO undifferentiated mESCs and EBs at day 10 of differentiation.
B. Principal component analysis of the RNA-sequencing WT and Sam68$^{-/-}$ samples.
C. Principal component analysis of the RNA-sequencing WT and QKI$^{-/-}$ samples.
D. Table showing the absolute numbers of differentially expressed RNAs for each RNA category. Only RNAs with $\pm$ 1,5 fold change and with an adjusted p value<0,01 are represented.



Supplementary Figure 3

Supplementary Figure 3

A. Gene ontology analysis of the most upregulated genes in Sam68 KO embryoid bodies at day 10 of differentiation shows overrepresentation of cardiac-related terms.
B. Gene ontology analysis of the RNA-sequencing of the most downregulated genes in QKI KO embryoid bodies at day 10 of differentiation shows overrepresentation of cardiac-related terms.
C. Validation of aberrant cassette-exons AS on cardiac-related RNAs in $QKI^{-/-}$ EBs at day 10 of differentiation.
D. Bar plot showing several aberrant AS cassette-exon events occurring on the Cacna1c transcript in $Sam68^{-/-}$ EBs at day 10 of differentiation.



Supplementary Figure 4

Supplementary Figure 4

A. Bar plot showing that the majority of the ribosome profiling reads at day0 (left) and day10 (right) of EBs differentiation align to the codifying sequence of the RNAs demonstrating the good quality of the generated libraries.
B. Plot that shows the correlation between transcription and translation changes in QKI KO EBs at day 10 of differentiation.
C. Quantification of immunofluorescence shown in figure 4C.
D. RNA immunoprecipitation of Sam68 and QKI to detect Gata4 binding. The bar plot shows the real-time PCR demonstrating the enrichment of the Gata4 transcript compared to IgG and normalized to a non-bound RNA (Rplp0).

REFERENCES

1. Rowe, R. G. & Daley, G. Q. Induced pluripotent stem cells in disease modelling and drug discovery. *Nat. Rev. Genet.* (2019) doi:10.1038/s41576-019-0100-z.

2. Rikhtegar, R. *et al.* Stem cells as therapy for heart disease: iPSCs, ESCs, CSCs, and skeletal myoblasts. *Biomed. Pharmacother. Biomedecine Pharmacother.* **109**, 304–313 (2019).

3. Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W. & Roder, J. C. Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 8424–8428 (1993).

4. Smith, A. A glossary for stem-cell biology. *Nature* **441**, 1060–1060 (2006).

5. Martin Gonzalez, J. *et al.* Embryonic Stem Cell Culture Conditions Support Distinct States Associated with Different Developmental Stages and Potency. *Stem Cell Rep.* **7**, 177–191 (2016).

6. Loh, K. M., Lim, B. & Ang, L. T. Ex uno plures: molecular designs for embryonic pluripotency. *Physiol. Rev.* **95**, 245–295 (2015).

7. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11**, 75–87 (2010).

8. Mitchell, S. F. & Parker, R. Principles and properties of eukaryotic mRNPs. *Mol. Cell* **54**, 547–558 (2014).

9. Lu, R. *et al.* Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature* **462**, 358–362 (2009).

10. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241–245 (2013).

11. Atlasi, Y., Mowla, S. J., Ziaee, S. A. M., Gokhale, P. J. & Andrews, P. W. OCT4 spliced variants are differentially expressed in human pluripotent and nonpluripotent cells. *Stem Cells Dayt. Ohio* **26**, 3068–3074 (2008).

12. Gabut, M. *et al.* An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* **147**, 132–146 (2011).

13. Lu, Y. *et al.* Alternative splicing of MBD2 supports self-renewal in human pluripotent stem cells. *Cell Stem Cell* **15**, 92–101 (2014).

14. Tahmasebi, S. *et al.* Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12360–12367 (2016).

15. Cieply, B. *et al.* Multiphasic and Dynamic Changes in Alternative Splicing during Induction of Pluripotency Are Coordinated by Numerous RNA-Binding Proteins. *Cell Rep.* **15**, 247–255 (2016).

16. Yeo, G. W. *et al.* An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.* **16**, 130–137 (2009).

17. Lu, X. *et al.* SON connects the splicing-regulatory network with pluripotency in human embryonic stem cells. *Nat. Cell Biol.* **15**, 1141–1152 (2013).

18. Hirsch, C. L. *et al.* Myc and SAGA rewire an alternative splicing network during early somatic cell reprogramming. *Genes Dev.* **29**, 803–816 (2015).

9. Toh, C.-X. D. *et al.* RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. *Cell Rep.* **15**, 2597–2607 (2016).

20. Yang, J. *et al.* RBM24 is a major regulator of muscle-specific alternative splicing. *Dev. Cell* **31**, 87–99 (2014).

21. Paronetto, M. P. *et al.* Sam68 regulates translation of target mRNAs in male germ cells, necessary for mouse spermatogenesis. *J. Cell Biol.* **185**, 235–249 (2009).

22. Paronetto, M. P. & Sette, C. Role of RNA-binding proteins in mammalian spermatogenesis. *Int. J. Androl.* **33**, 2–12 (2010).

23. La Rosa, P. *et al.* Sam68 promotes self-renewal and glycolytic metabolism in mouse neural progenitor cells by modulating Aldh1a3 pre-mRNA 3'-end processing. *eLife* **5**, e20750 (2016).

24. Wu, J. I., Reed, R. B., Grabowski, P. J. & Artzt, K. Function of quaking in myelination: regulation of alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.***99**, 4233–4238 (2002).

25. Richard, S. *et al.* Ablation of the Sam68 RNA binding protein protects mice from age-related bone loss. *PLoS Genet.* **1**, e74 (2005).

26. Richard, S. *et al.* Sam68 haploinsufficiency delays onset of mammary tumorigenesis and metastasis. *Oncogene* **27**, 548–556 (2008).

27. Lu, Z. *et al.* The quakingviable mutation affects qkI mRNA expression specifically in myelin-producing cells of the nervous system. *Nucleic Acids Res.***31**, 4616–4624 (2003).

28. Hardy, R. J. *et al.* Neural cell type-specific expression of QKI proteins is altered in quakingviable mutant mice. *J. Neurosci. Off. J. Soc. Neurosci.* **16**, 7941–7949 (1996).

29. Larocque, D. *et al.* The QKI-6 and QKI-7 RNA binding proteins block proliferation and promote Schwann cell myelination. *PloS One* **4**, e5867 (2009).

30. Wray, J. *et al.* Inhibition of glycogen synthase kinase-3 alleviates Tcf3 repression of the pluripotency network and increases embryonic stem cell resistance to differentiation. *Nat. Cell Biol.* **13**, 838–845 (2011).

31. Paronetto, M. P., Achsel, T., Massiello, A., Chalfant, C. E. & Sette, C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *J. Cell Biol.* **176**, 929–939 (2007).

32. Matter, N., Herrlich, P. & König, H. Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature* **420**, 691–695 (2002).

33. Chawla, G. *et al.* Sam68 regulates a set of alternatively spliced exons during neurogenesis. *Mol. Cell. Biol.* **29**, 201–213 (2009).

34. Pandit, S. *et al.* Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. *Mol. Cell* **50**, 223–235 (2013).

35. Änkö, M.-L. *et al.* The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol.* **13**, R17 (2012).

36. Bechara, E. G., Sebestyén, E., Bernardis, I., Eyras, E. & Valcárcel, J. RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation. *Mol. Cell* **52**, 720–733 (2013).

37. Valacca, C. *et al.* Sam68 regulates EMT through alternative splicing–activated nonsense-mediated mRNA decay of the SF2/ASF proto-oncogene. *J. Cell Biol.* **191**, 87–99 (2010).

38. Huot, M.-É. *et al.* The Sam68 STAR RNA-Binding Protein Regulates mTOR Alternative Splicing during Adipogenesis. *Mol. Cell* **46**, 187–199 (2012).

39. Bonnet, A. *et al.* Quaking RNA-Binding Proteins Control Early Myofibril Formation by Modulating Tropomyosin. *Dev. Cell* **42**, 527-541.e4 (2017).

40. Sjöblom, B., Salmazo, A. & Djinović-Carugo, K. Alpha-actinin structure and regulation. *Cell. Mol. Life Sci. CMLS* **65**, 2688–2701 (2008).

41. Luther, P. K. *et al.* Understanding the Organisation and Role of Myosin Binding Protein C in Normal Striated Muscle by Comparison with MyBP-C Knockout Cardiac Muscle. *J. Mol. Biol.* **384**, 60–72 (2008).

42. Chu, M., Gregorio, C. C. & Pappas, C. T. Nebulin, a multi-functional giant. *J. Exp. Biol.* **219**, 146–152 (2016).

43. Matsson, H. *et al.* Alpha-cardiac actin mutations produce atrial septal defects. *Hum. Mol. Genet.* **17**, 256–265 (2008).

44. Jin, J.-P., Zhang, Z. & Bautista, J. A. Isoform diversity, regulation, and functional adaptation of troponin and calponin. *Crit. Rev. Eukaryot. Gene Expr.* **18**, 93–124 (2008).

45. Li, M. X. & Hwang, P. M. Structure and function of cardiac troponin C (TNNC1): Implications for heart failure, cardiomyopathies, and troponin modulating drugs. *Gene* **571**, 153–166 (2015).

46. Sun, B. *et al.* The cardiac ryanodine receptor, but not sarcoplasmicreticulum Ca^{2+} -ATPase, is a major determinant of Ca^{2+} alternans in intact mouse hearts. *J. Biol. Chem.* **293**, 13650–13661 (2018).

47. Yang, H.-T. *et al.* The ryanodine receptor modulates the spontaneous beating rate of cardiomyocytes during development. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9225–9230 (2002).

48. Remme, C. A. Cardiac sodium channelopathy associated with *SCN5A*mutations: electrophysiological, molecular and genetic aspects: Cardiac sodium channelopathy associated with *SCN5A* mutations. *J. Physiol.* **591**, 4099–4116 (2013).

49. Beraldi, R. *et al.* Rbm20-deficient cardiogenesis reveals early disruption of RNA processing and sarcomere remodeling establishing a developmental etiology for dilated cardiomyopathy. *Hum. Mol. Genet.* **23**, 3779–3791 (2014).

50. Xin, M. *et al.* A threshold of GATA4 and GATA6 expression is required for cardiovascular development. *Proc. Natl. Acad. Sci.* **103**, 11189–11194 (2006).

51. Schlesinger, J. *et al.* The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5, Srf, Histone Modifications, and MicroRNAs. *PLoS Genet.* **7**, e1001313 (2011).

52. Wiese, C. *et al.* Formation of the sinus node head and differentiation of sinus node myocardium are independently regulated by Tbx18 and Tbx3. *Circ. Res.* **104**, 388–397 (2009).

53. Peng, W. *et al.* Dysfunction of Myosin Light-Chain 4 (MYL4) Leads to Heritable Atrial Cardiomyopathy With Electrical, Contractile, and Structural Components: Evidence From Genetically-Engineered Rats. *J. Am. Heart Assoc.* **6**, (2017).

54. Sheikh, F., Lyon, R. C. & Chen, J. Functions of myosin light chain-2 (MYL2) in cardiac muscle and disease. *Gene* **569**, 14–20 (2015).

55. Chernyavskaya, Y., Ebert, A. M., Milligan, E. & Garrity, D. M. Voltage-gated calcium channel CACNB2 (β2.1) protein is required in the heart for control of cell proliferation and heart tube integrity. *Dev. Dyn. Off. Publ. Am. Assoc. Anat.* **241**, 648–662 (2012).

56. Zhang, Q., Chen, J., Qin, Y., Wang, J. & Zhou, L. Mutations in voltage-gated L-type calcium channel: implications in cardiac arrhythmia. *Channels* **12**, 201–218 (2018).

57. Narita, N., Bielinska, M. & Wilson, D. B. Cardiomyocyte differentiation by GATA-4-deficient embryonic stem cells. *Dev. Camb. Engl.* **124**, 3755–3764 (1997).

58. Laemmle, L. L., Cohen, J. B. & Glorioso, J. C. Constitutive Expression of GATA4 Dramatically Increases the Cardiogenic Potential of D3 Mouse Embryonic Stem Cells. *Open Biotechnol. J.* **10**, 248–257 (2016).

59. Ingolia, N. T. Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.* **470**, 119–142 (2010).

60. Delaney, J. T. *et al.* A KCNJ8 mutation associated with early repolarization and atrial fibrillation. *Eur. Eur. Pacing Arrhythm. Card. Electrophysiol. J. Work. Groups Card. Pacing Arrhythm. Card. Cell. Electrophysiol. Eur. Soc. Cardiol.***14**, 1428–1432 (2012).

61. Zhou, L. *et al.* Tbx5 and Osr1 interact to regulate posterior second heart field cell cycle progression for cardiac septation. *J. Mol. Cell. Cardiol.* **85**, 1–12 (2015).

62. Liu, Y. *et al.* Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 3859–3864 (2007).

63. Varjosalo, M. & Taipale, J. Hedgehog: functions and mechanisms. *Genes Dev.* **22**, 2454–2472 (2008).

64. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol. Cell* **63**, 696–710 (2016).

65. Livi, C. M., Klus, P., Delli Ponti, R. & Tartaglia, G. G. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinforma. Oxf. Engl.* **32**, 773–775 (2016).

66. Babic, I., Cherry, E. & Fujita, D. J. SUMO modification of Sam68 enhances its ability to repress cyclin D1 expression and inhibits its ability to induce apoptosis. *Oncogene* **25**, 4955–4964 (2006).

67. Li, Q.-H. *et al.* Retardation of the G2-M phase progression on gene disruption of RNA binding protein Sam68 in the DT40 cell line. *FEBS Lett.***525**, 145–150 (2002).

68. He, B. *et al.* MicroRNA-155 promotes the proliferation and invasion abilities of colon cancer cells by targeting quaking. *Mol. Med. Rep.* **11**, 2355–2359 (2015).

69. Xu, Z. *et al.* Wnt/β-catenin signaling promotes self-renewal and inhibits the primed state transition in naïve human embryonic stem cells. *Proc. Natl. Acad. Sci.* **113**, E6382–E6390 (2016).

70. Khacho, M. & Slack, R. S. Mitochondrial and Reactive Oxygen Species Signaling Coordinate Stem Cell Fate Decisions and Life Long Maintenance. *Antioxid. Redox Signal.* **28**, 1090–1101 (2018).

71. Sun, Z. *et al.* The Long Noncoding RNA Lncenc1 Maintains Naive States of Mouse ESCs by Promoting the Glycolysis Pathway. *Stem Cell Rep.* **11**, 741–755 (2018).

72. Sargent, C. Y., Berguig, G. Y. & McDevitt, T. C. Cardiomyogenic Differentiation of Embryoid Bodies Is Promoted by Rotary Orbital Suspension Culture. *Tissue Eng. Part A* **15**, 331–342 (2009).

73. Chan, S. S.-K. *et al.* Mesp1 Patterns Mesoderm into Cardiac, Hematopoietic, or Skeletal Myogenic Progenitors in a Context-Dependent Manner. *Cell Stem Cell* **12**, 587–601 (2013).

74. Bondue, A. *et al.* Mesp1 Acts as a Master Regulator of Multipotent Cardiovascular Progenitor Specification. *Cell Stem Cell* **3**, 69–84 (2008).

75. Hedley, P. L. *et al.* The genetic basis of Brugada syndrome: a mutation update. *Hum. Mutat.* **30**, 1256–1266 (2009).

76. Napolitano, C., Splawski, I., Timothy, K. W., Bloise, R. & Priori, S. G. Timothy Syndrome. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1993).

77. Zhu, Y., Luo, J., Jiang, F. & Liu, G. Genetic analysis of sick sinus syndrome in a family harboring compound CACNA1C and TTN mutations. *Mol. Med. Rep.* (2018) doi:10.3892/mmr.2018.8773.

78. Zhang, Y. & Feng, Y. Distinct molecular mechanisms lead to diminished myelin basic protein and 2',3'-cyclic nucleotide 3'-phosphodiesterase in qk(v) dysmyelination. *J. Neurochem.* **77**, 165–172 (2001).

79. Li, Z., Zhang, Y., Li, D. & Feng, Y. Destabilization and mislocalization of myelin basic protein mRNAs in quaking dysmyelination lacking the QKI RNA-binding proteins. *J. Neurosci. Off. J. Soc. Neurosci.* **20**, 4944–4953 (2000).

80. Larocque, D. *et al.* Nuclear retention of MBP mRNAs in the quaking viable mice. *Neuron* **36**, 815–829 (2002).

81. Hayakawa-Yano, Y. & Yano, M. An RNA Switch of a Large Exon of Ninein Is Regulated by the Neural Stem Cell Specific-RNA Binding Protein, Qki5. *Int. J. Mol. Sci.* **20**, 1010 (2019).

82. He, J. J., Henao-Mejia, J. & Liu, Y. Sam68 functions in nuclear export and translation of HIV-1 RNA. *RNA Biol.* **6**, 384–386 (2009).

83. Hurley, J. M., Loros, J. J. & Dunlap, J. C. Circadian Oscillators: Around the Transcription–Translation Feedback Loop and on to Output. *Trends Biochem. Sci.* **41**, 834–846 (2016).

6. Discussion

Recently, the non-coding portion of the genome became subject of intense research due to its regulatory role in the complexity of biological and pathological processes[119]. Key functions of non-coding RNAs include catalysis of reactions, scaffolding of molecular interactions and action as regulatory signals[120].

Among the numerous classes of ncRNAs, circRNAs are receiving increasing attention due to their peculiar circular structure and tissue-specific expression. Notably, circRNAs are able to influence gene expression during development and their biogenesis relies on the activity of several RBPs[18,31]. Indeed, the Signal Transduction and Activation of RNA (STAR) family members Quaking and Sam68 have been shown to be involved in the biogenesis of circRNAs in specific biological contexts[30,92], but whether their contribution is extended to embryonic development is unknown.

The STAR member Sam68 regulates many aspects of RNA processing such as alternative splicing[75–78,80], mRNA translation[84], localization and RNA transport[85]. Remarkably, many of its functions occur in different biological contexts and are linked to its broad interplay capacity.

Hence, to understand the many layers of regulation mediated by Sam68 during mouse embryonic stem cells differentiation, we decided to further investigate its interactome, both at protein and RNA level.

6.1 Sam68 protein networks

To untangle all the regulatory functions involving Sam68 in the context of early development of mouse Embryonic Stem Cells (mESCs), we analysed Sam68 proteome during differentiation. Our results revealed that 50% of Sam68 whole interactome is composed by RBPs and all the steps of the RNA metabolism are well represented among the functional interaction network: we found interactors involved in mRNA splicing (Sfpq, Nono)[100,101], transport and translation (e.g. Stau2, Caprin1)[102,103]. Among Sam68 protein partners we could retrieve some previously described interactors in other physiological contexts: indeed, during spermatogenesis, Sam68 interacts with the miRNA processing players Drosha and Dicer[83]. Interestingly, in our network we found Drosha, in addition to Ago2 and Dgcr8, suggesting that Sam68 might have a role in miRNA processing during mESCs differentiation. Moreover, we detected the interaction with the heterogeneous nuclear ribonucleoproteins HnrnpL, important player in transcriptional regulation during myogenic differentiation[106].

From our MS analysis we identified that the stronger interaction occurs between Sam68 and the other member of the STAR family Slm2. Both STAR members are involved in alternative splicing during neurogenesis: both RBPs are able to bind Neurexin pre-mRNA, but only Slm2 is proficient to regulate its splicing depending on the binding sites density[79,91]. Slm2 is mostly expressed in specific-tissues (testis, brain), Sam68 is ubiquitous and its expression reflects its versatile post-transcriptional regulation functions in different biological contexts[72,105,121]. Although Sam68 is able to exert the same

molecular functions of Slm2, the cooperation/competition balance in binding the same targets can be solved considering the tissue-specific expression of the latter and the stoichiometry of interaction revealed by our MS analysis. Indeed, we discovered that Sam68 can interact with two Slm2 proteins at the same time. Nevertheless, the fact that only Slm2 is proficient to regulate *neurexin2* exon 20 (AS4), while Sam68 promotes the inclusion of the same exon only in *neurexin1* isoform suggests that there is no functional redundancy between the two STAR member[89,91]. We investigated the interaction by Size Exclusion Chromatography and noticed that Slm2 was detected only in some of the collected fractions where Sam68 is expressed, suggesting that Slm2 may cooperate with Sam68 only for determinate functions in mESCs.

Remarkably, one example of complex cooperation/competition relationship is provided by the SR family, splicing regulators that promote exon inclusion by binding to exonic splicing enhancers. Notably, the absence of one SR protein coincides with the loss or compensatory gain in the interaction of other SR proteins at the affected exons[122]. Indeed, given the strong unbalanced interaction between the two STAR members, we could speculate that in mESCs Sam68 might help Slm2 to accomplish its molecular functions by presenting it to the target RNAs.

Our data demonstrate that many of the known functions of Sam68 in RNA processing are linked to its versatile interaction ability with RBPs participating in every step of RNA metabolism.

6.2 Sam68: a new player in circRNA biogenesis?

Focusing our attention on interactors that could extend the regulative functions of Sam68 in RNA metabolism, we identified elements of circRNA processing: circRNAs are a class of ncRNAs able to influence gene expression that are becoming subject of intense research[123]. Advances in genome-wide approaches helped to untangle the mechanism of exon circularization, providing clear evidence that it depends on the presence of flanking intronic complementary sequences[27]. Importantly, the investigation of how circRNA expression is modulated at the same loci across cell types and tissues led to the identification of RNA binding proteins (RBPs) as essential trans-acting factors influencing circRNA processing[21]. Notably, our Sam68 proteome analysis revealed elements of circRNA biogenesis: among those we found two component of the Interleukin enhancer-binding factor family (Ilf2 and Ilf3), involved in the regulation of pluripotency and differentiation in mESCs and, upon viral infection, in the biogenesis of circRNAs[31]. Indeed, Ilf3 influences circRNA biogenesis by increasing intronic RNA pairs in the nucleus and binding to circRNP in the cytoplasm. During the innate immune response against virus infection, Ilf3 isoforms (90 kDa and 110 kDa) are released from circRNPs complexes in the cytoplasm allowing their binding to viral mRNAs to inhibit viral replication. The STAR member Quaking is a known circRNA biogenesis factor during the epithelial to mesenchymal transition (EMT) and a new role for Sam68 in the pre-mRNA circularization process has been recently described for the SMN locus[30,92]. We validated the interaction between Sam68 and Ilf3 by

immunoprecipitation and we discovered that RNase treatment seems to destabilize Sam68 interaction with Ilf3, suggesting that RNA acts as an important scaffolding element.

These data support the hypothesis that Sam68 participates in circRNA processing through the interaction with the circRNA biogenesis complex Ilf2-Ilf3.

The confirmed interaction between Sam68 and the circRNA biogenesis factor Ilf3 led us to further investigate the involvement of Sam68 in circRNA biogenesis at genome-wide level. Using RNA sequencing data previously produced in our laboratory (Dasti et al., submitted), we analyzed circRNAs abundance upon Sam68 and QKI depletion during mESCs differentiation, considering three time points of the embryonic bodies assay (D0, D3, D10).

We focused our attention on circRNAs that were differentially expressed upon Sam68 and QKI depletion but with an unaltered level of the linear mRNA. Indeed, circRNAs with this feature are the ones altered at biogenesis level, since their expression is not due to the corresponding linear RNA changes.

The investigation of circRNA expression upon Sam68 depletion showed a consistent downregulation of circRNAs at biogenesis level, confirming the involvement of Sam68 in the pre-mRNA circularization process during differentiation. Moreover, Sam68 role in circRNA biogenesis is supported by its nuclear localization in mESCs.

6.3 CircRNAs: direct targets of Sam68?

To demonstrate the direct involvement of Sam68 in circRNAs formation, we investigated its binding profile using the protein-RNA interactions predictor CatRAPID[116] and iCLIP2[97]. Notably, the STAR member Quaking promotes circRNA production by binding to intronic regions and bringing in close proximity circularizing exons thanks to its dimerization ability[30]. Given that Sam68 has a similar binding ability and has been recently shown to take advantage of high *Alu* sequences density in the SMN locus to induce pre-mRNA circularization of the latter[92], we wondered if Sam68 could induce circRNA biogenesis in early development.

We analysed its interaction propensity towards circRNA-forming transcripts expressed in mouse cardiomyocytes[46] using catRAPID, a bioinformatic tool developed in our laboratory that estimates the binding propensity between protein-RNA pairs.

Using as control a pool of 1799 RBPs[117], **we showed that Sam68 binding propensity to transcripts that are the source of circRNAs in mouse cardiomyocytes is very significant.**

To experimentally validate our predictions, we performed a transcriptome-wide mapping of Sam68 mRNA targets during D0 and D10 time points. The iCLIP2 analysis identified the genomic coordinates of Sam68 binding, demonstrating that it preferentially binds intronic regions. This result is in line with the analysis of alternative splicing (AS) events affected by Sam68 depletion, where intron retention resulted to be the most affected type of AS event (Dasti et al., submitted). As, previously described, QKI promotes circRNA biogenesis by binding to introns flanking circularizing

exons and (Conn et al., 2015) and given their similar binding activity we could speculate that Sam68 might induce circRNA formation with the same mechanism. Indeed, the integration between circRNA analysis data with the iCLIP2 experimental output highlighted that 59% of the identified circRNAs is a direct target of Sam68, while at D10 there is a slight decrease of circRNAs directly bound.

The analysis of the iCLIP2 data is still ongoing and it is aimed to recover Sam68 RNA binding motif and its specific binding profile, in order to further analyze Sam68 RNA motif enrichment in introns flanking circularizing exons.

6.4 What about QKI?

Quaking promotes circRNA biogenesis during the epithelial to mesenchymal transition (EMT), taking advantage of its dimerization process that brings in close proximity the circularizing exons[30].

However, despite its well-known role as circRNA biogenesis factor, our results regarding the circRNA differential expression analysis performed upon Quaking depletion did not show any specific trend related to the RBP absence, suggesting that Quaking does not affect circRNA biogenesis during mESCs differentiation.

The fact that Quaking does not promote circRNA biogenesis in our cellular model might depend on its localization within the cell. Indeed, in mESCs Quaking is mostly present in the cytoplasm and only a small portion has nuclear localization (Dasti et al., submitted).

Our findings suggest that Quaking is not required for circRNA biogenesis during mESCs differentiation due to its cytoplasmic localization.

6.5 CircRNAs: the missing link between Sam68 and heart development?

In a previous work recently submitted, our laboratory identified Sam68 as a new player in early heart development. Notably, when Sam68 is depleted in mouse Embryonic Stem Cells (ESCs), a defect in the electrophysiological features of cardiomyocytes is observed. Nevertheless, the transcriptomic analysis revealed an up-regulation of cardiac-related transcripts upon Sam68 KO, in addition to an alteration of alternative splicing (AS) events. Among the transcripts altered at AS level we found Slc8a1, a sodium calcium pump highly expressed during heart development and known to generate circRNAs. Interestingly, we found that Sam68 directly regulates the expression of the cardiogenic-factor Gata4. Furthermore, cardiac-related transcripts are regulated by Sam68 also at translational level. (Dasti et al., submitted).

CircRNAs are highly expressed during heart development and are associated with cardiac diseases[48]. Moreover, RBPs with essential roles in splicing during cardiomyocytes and myofibril development are also linked to circRNA production[61].

Notably, in absence of Sam68, we observed a strong defect in circRNA biogenesis at D3 time point, this coincide with the undertake of differentiation paths towards the embryonic germ layers (endoderm, ectoderm, mesoderm) by pluripotent cells.

Nevertheless, our data confirm Sam68 interaction with circRNA biogenesis complex Ilf2-Ilf3 at D3 as well, suggesting that a correct circRNA formation is essential for proper differentiation towards

cardiac lineage. Moreover, in our list of circRNAs with impaired biogenesis we found Trpm7, involved in pathogenesis of atrial fibrillation and whose depletion disrupts adult ventricular function, conduction and repolarization[124]. Among the direct circRNA targets we also identified Rere, Gata4 transcriptional regulator during the epithelial-to-mesenchymal transition (EMT) that occurs in cardiac development, when endocardial cells migrate in the space between endocardium and myocardium to form the atrioventricular septum[125]. Another interesting circRNA target is Pum2, an RNA binding protein that regulates the translational network in myocardial fibrosis together with the STAR member QKI[126].

To conclude, the involvement of Sam68 in circRNA biogenesis could further explain its role in heart development and the phenotype of impaired cardiomyocytes, providing a new layer of Sam68-mediated regulation during mESCs differentiation.

7. Conclusions

The work carried out during my PhD studies at the Centre for Genomic Regulation (CRG) of Barcelona has been compiled under the form of a thesis with the title "Sam68 and circRNA biogenesis in early development".

The thesis represents my personal contribution to the understanding of the role of Sam68 in the regulation of circRNA biogenesis during mouse embryonic stem cells differentiation specifically towards the cardiomyocyte lineage.

The obtained results can be summarized as follows:

4. Sam68 proteome network during mESCs differentiation is composed by RBPs involved in all the steps of RNA metabolism;
5. Sam68 strongly interacts with Slm2, another member of the STAR family;
6. Sam68 directly interacts with the circRNA biogenesis complex Ilf2-Ilf3, with a direct binding towards Ilf3;
7. Sam68 depletion in mESCs induces circRNAs downregulation at biogenesis level;
8. Sam68 binding propensity is very specific towards circRNA-related genes highly expressed in mouse cardiomyocytes;

9. Quaking is not required for circRNA biogenesis during differentiation due to its cytoplasmic localization;

10. CircRNAs with a defect in biogenesis in Sam68$^{-/-}$ mESCs are direct targets of Sam68;

11. CatRAPID can reproduce iCLIP2 experimental output.

8. Annex

8.1 Antibodies

Anti-GFP SAB4301138 Sigma

Anti-Sam68: C-20 sc-333 Santa Cruz

Anti-Sam68: H4 sc-514468 Santa Cruz

Anti-Quaking: N-20 sc-103851 Santa Cruz

Anti-panQuaking: N147/6 MABN624 Millipore

Anti-Slm2 ab68515 Abcam

Anti-ILF3 612155 BD Transduction Laboratories

Anti-ILF2 EB07784 Everest Biotech

Anti-beta Tubulin: ab6046 Abcam

Normal Rabbit IgG: sc-3888 Santa Cruz

Protein-G-HRP conjugated: ab97046 Abcam

Goat anti-rabbit-HRP conjugated: P0448 Dako

8.2 iCLIP2 primers

Name	Sequence (IDT)
RToligo	GGATCCTGAACCGCT
L3-App	/rApp/AGATCGGAAGAGCGGTTCAG/ddC/
L01clip2.0	/5Phos/NNNNATCACGNNNNNAGATCGGAAGAGCGTCGTG/3ddC/
L02clip2.0	/5Phos/NNNNCGATGTNNNNNAGATCGGAAGAGCGTCGTG/3ddC/
L03clip2.0	/5Phos/NNNNTTAGGCNNNNNAGATCGGAAGAGCGTCGTG/3ddC/
L04clip2.0	/5Phos/NNNNTGACCANNNNNAGATCGGAAGAGCGTCGTG/3ddC/
L05clip2.0	/5Phos/NNNNACAGTGNNNNNAGATCGGAAGAGCGTCGTG/3ddC/
L06clip2.0	/5Phos/NNNNGCCAATNNNNNAGATCGGAAGAGCGTCGTG/3ddC/
P5Solexa_s	ACACGACGCTCTTCCGATCT
P3Solexa_s	CTGAACCGCTCTTCCGATCT
P5Solexa	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCT
P3Solexa	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCT GAACCGCTCTTCCGATCT

9. Bibliography

1. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
2. Nam, J. W., Choi, S. W. & You, B. H. Incredible RNA: Dual functions of coding and noncoding. *Molecules and Cells* vol. 39 367–374 (2016).
3. Mattick, J. S. & Makunin, I. v. Non-coding RNA. *Human molecular genetics* vol. 15 Spec No 1 (2006).
4. Gebert, L. F. R. & MacRae, I. J. Regulation of microRNA function in animals. *Nature Reviews Molecular Cell Biology* vol. 20 21–37 (2019).
5. Pauli, A., Rinn, J. L. & Schier, A. F. Non-coding RNAs as regulators of embryogenesis. *Nature Reviews Genetics* vol. 12 136–149 (2011).
6. Carter, R. & Drouin, G. Structural differentiation of the three eukaryotic RNA polymerases. *Genomics* **94**, (2009).
7. Fatica, A. & Bozzoni, I. Long non-coding RNAs: New players in cell differentiation and development. *Nature Reviews Genetics* vol. 15 7–21 (2014).
8. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131–137 (1996).
9. Loda, A. & Heard, E. Xist RNA in action: Past, present, and future. *PLoS Genetics* vol. 15 e1008333 (2019).

10. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
11. Ashwal-Fluss, R. *et al.* CircRNA Biogenesis competes with Pre-mRNA splicing. *Molecular Cell* **56**, 55–66 (2014).
12. Hamazaki, N., Uesaka, M., Nakashima, K., Agata, K. & Imamura, T. Gene activation-associated long noncoding RNAs function in mouse preimplantation development. *Development (Cambridge)* **142**, 910–920 (2015).
13. Zhang, K., Huang, K., Luo, Y. & Li, S. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. *BMC Genomics* **15**, 845 (2014).
14. Okamoto, I. *et al.* Evidence for de novo imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* **438**, (2005).
15. Lewis, A. *et al.* Epigenetic dynamics of the Kcnq1 imprinted domain in the early embryo. *Development* **133**, (2006).
16. Ritter, N. *et al.* The lncRNA Locus Handsdown Regulates Cardiac Gene Programs and Is Essential for Early Mouse Development. *Developmental Cell* **50**, (2019).
17. Maass, P. G., Luft, F. C. & Bähring, S. Long non-coding RNA in health and disease. *Journal of Molecular Medicine* **92**, (2014).
18. Li, X., Yang, L. & Chen, L.-L. The Biogenesis, Functions, and Challenges of Circular RNAs. *Molecular Cell* **71**, (2018).

19. Hsu, M. T. & Coca-Prados, M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells [24]. *Nature* vol. 280 339–340 (1979).
20. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64**, 607–613 (1991).
21. Dong, R., Ma, X. K., Chen, L. L. & Yang, L. Increased complexity of circRNA expression during species evolution. *RNA Biology* **14**, 1064–1074 (2017).
22. Messias, A. C. & Sattler, M. Structural Basis of Single-Stranded RNA Recognition. *Accounts of Chemical Research* **37**, (2004).
23. Stefl, R., Skrisovska, L. & Allain, F. H. -T. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO reports* **6**, (2005).
24. Sontheimer, E. J. Assembly and function of RNA silencing complexes. *Nature Reviews Molecular Cell Biology* **6**, (2005).
25. Castello, A. *et al.* System-wide identification of RNA-binding proteins by interactome capture. *Nature Protocols* **8**, (2013).
26. Starke, S. *et al.* Exon circularization requires canonical splice signals. *Cell Reports* **10**, 103–111 (2015).
27. Zhang, X. O. *et al.* Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
28. Errichelli, L. *et al.* FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons. *Nature Communications* **8**, (2017).

29. Nieto, M. A. Epithelial plasticity: A common theme in embryonic and cancer cells. *Science* vol. 342 (2013).

30. Conn, S. J. *et al.* The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160**, 1125–1134 (2015).

31. Li, X. *et al.* Coordinated circRNA Biogenesis and Function with NF90/NF110 in Viral Infection. *Molecular Cell* **67**, 214-227.e7 (2017).

32. Kelly, S., Greenman, C., Cook, P. R. & Papantonis, A. Exon Skipping Is Correlated with Exon Circularization. *Journal of Molecular Biology* **427**, 2414–2417 (2015).

33. Li, Y. *et al.* Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Nature Publishing Group* (2015) doi:10.1038/cr.2015.82.

34. Conn, V. M. *et al.* A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nature Plants* **3**, (2017).

35. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).

36. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, (2013).

37. Dudekula, D. B. *et al.* CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs. *RNA Biology* **13**, (2016).

38. Pamudurti, N. R. *et al.* Translation of CircRNAs. *Molecular Cell* **66**, 9-21.e7 (2017).

39. Legnini, I. *et al.* Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Molecular Cell* **66**, 22-37.e9 (2017).

40. Dong, R. *et al.* CircRNA-derived pseudogenes. (2016) doi:10.1038/cr.2016.42.

41. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biology* **16**, (2015).

42. Dang, Y. *et al.* Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biology* **17**, (2016).

43. Yu, C.-Y. *et al.* The circular RNA circBIRC6 participates in the molecular circuitry controlling human pluripotency. *Nature Communications* **8**, (2017).

44. di Agostino, S. *et al.* Circular RNAs in Embryogenesis and Cell Differentiation With a Focus on Cancer Development. *Frontiers in Cell and Developmental Biology* **8**, (2020).

45. You, X. *et al.* Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature Neuroscience* **18**, 603–610 (2015).

46. Tan, W. L. W. *et al.* A landscape of circular RNA expression in the human heart. *Cardiovascular Research* **113**, 298–309 (2017).

47. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *New England Journal of Medicine* **366**, 619–628 (2012).

48. Lee, E. C. S. *et al.* The roles of circular RNAs in human development and diseases. *Biomedicine & Pharmacotherapy* **111**, (2019).

49. Ng, H. H. & Surani, M. A. The transcriptional and signalling networks of pluripotency. *Nature Cell Biology* vol. 13 490–496 (2011).

50. Watanabe, A., Yamada, Y. & Yamanaka, S. Epigenetic regulation in pluripotent stem cells: A key to breaking the epigenetic barrier. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 368 (2013).

51. Wells, C. A. & Choi, J. Transcriptional Profiling of Stem Cells: Moving from Descriptive to Predictive Paradigms. *Stem Cell Reports* **13**, (2019).

52. Gabut, M. *et al.* An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* **147**, 132–146 (2011).

53. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, (2013).

54. Yeo, G. W. *et al.* An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology* **16**, (2009).

55. Lackford, B. *et al.* Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO Journal* **33**, 878–889 (2014).

56. Zukeran, A. *et al.* The CCR4-NOT deadenylase activity contributes to generation of induced pluripotent stem cells.

Biochemical and Biophysical Research Communications **474**, (2016).

57. Schwartz, S. *et al.* Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5' sites. *Cell Reports* **8**, 284–296 (2014).

58. Wang, Y. *et al.* N6 -methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology* **16**, 191–198 (2014).

59. Wang, L. *et al.* The THO complex regulates pluripotency gene mRNA export and controls embryonic stem cell self-renewal and somatic cell reprogramming. *Cell Stem Cell* **13**, 676–690 (2013).

60. Brand, T. Heart development: molecular insights into cardiac specification and early morphogenesis. *Developmental Biology* **258**, (2003).

61. Blech-Hermoni, Y. & Ladd, A. N. RNA binding proteins in the regulation of heart development. *International Journal of Biochemistry and Cell Biology* vol. 45 2467–2478 (2013).

62. Xu, E. *et al.* RNA-Binding Protein RBM24 Regulates p63 Expression via mRNA Stability. *Molecular Cancer Research* **12**, (2014).

63. Miller, M. T., Higgin, J. J. & Tanaka Hall, T. M. Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. *Nature Structural & Molecular Biology* **15**, (2008).

64. Khan, M. A. F. *et al.* RBM20 Regulates Circular RNA Production From the Titin Gene. *Circulation Research* **119**, (2016).

65. Justice, M. J. & Hirschi, K. K. The Role of Quaking in Mammalian Embryonic Development. in (2010). doi:10.1007/978-1-4419-7005-3_6.

66. Vernet, C. & Artzt, K. STAR, a gene family involved in signal transduction and activation of RNA. *Trends in Genetics* vol. 13 479–484 (1997).

67. Wu, J., Zhou, L., Tonissen, K., Tee, R. & Artzt, K. The quaking I-5 protein (QKI-5) has a novel nuclear localization signal and shuttles between the nucleus and the cytoplasm. *Journal of Biological Chemistry* **274**, 29202–29210 (1999).

68. Maguire, M. L. *et al.* Solution structure and backbone dynamics of the KH-QUA2 region of the Xenopus STAR/GSG quaking protein. *Journal of Molecular Biology* **348**, 265–279 (2005).

69. Fumagalli, S., Totty, N. F., Hsuan, J. J. & Courtneidge, S. A. A target for Src in mitosis. *Nature* **368**, 871–874 (1994).

70. Taylor, S. J. & Shalloway, D. An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis. *Nature* **368**, 867–871 (1994).

71. Lukong, K. E. & Richard, S. Sam68, the KH domain-containing superSTAR. *Biochimica et Biophysica Acta - Reviews on Cancer* vol. 1653 73–86 (2003).

72. Sánchez-Jiménez, F. & Sánchez-Margalet, V. Role of Sam68 in Post-Transcriptional Gene Regulation. *International Journal of Molecular Sciences* **14**, (2013).

73. Lin, Q., Taylor, S. J. & Shalloway, D. Specificity and determinants of Sam68 RNA binding. Implications for the biological function of K homology domains. *Journal of Biological Chemistry* **272**, 27274–27280 (1997).

74. Babic, I., Jakymiw, A. & Fujita, D. J. The RNA binding protein Sam68 is acetylated in tumor cell lines, and its acetylation correlates with enhanced RNA binding activity. *Oncogene* **23**, 3781–3789 (2004).

75. Paronetto, M. P., Achsel, T., Massiello, A., Chalfant, C. E. & Sette, C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *Journal of Cell Biology* **176**, 929–939 (2007).

76. Pedrotti, S. *et al.* The splicing regulator Sam68 binds to a novel exonic splicing silencer and functions in SMN2 alternative splicing in spinal muscular atrophy. *EMBO Journal* **29**, 1235–1247 (2010).

77. Pagliarini, V. *et al.* SAM68 is a physiological regulator of SMN2 splicing in spinal muscular atrophy. *Journal of Cell Biology* **211**, 77–90 (2015).

78. Paronetto, M. P. *et al.* Alternative splicing of the cyclin D1 proto-oncogene is regulated by the RNA-binding protein Sam68. *Cancer Research* **70**, 229–239 (2010).

79. Iijima, T. *et al.* SAM68 regulates neuronal activity-dependent alternative splicing of neurexin-1. *Cell* **147**, 1601–1614 (2011).

80. Ehrmann, I. *et al.* The Tissue-Specific RNA Binding Protein T-STAR Controls Regional Splicing Patterns of Neurexin Pre-mRNAs in the Brain. *PLoS Genetics* **9**, (2013).

81. Li, N., Hébert, S., Song, J., Kleinman, C. L. & Richard, S. Transcriptome profiling in preadipocytes identifies long noncoding RNAs as Sam68 targets. *Oncotarget* **8**, 81994–82005 (2017).

82. Subramanyam, D. & Blelloch, R. From microRNAs to targets: Pathway discovery in cell fate transitions. *Current Opinion in Genetics and Development* vol. 21 498–503 (2011).

83. Messina, V. *et al.* The RNA Binding Protein SAM68 Transiently Localizes in the Chromatoid Body of Male Germ Cells and Influences Expression of Select MicroRNAs. *PLoS ONE* **7**, (2012).

84. Paronetto, M. P. *et al.* Sam68 regulates translation of target mRNAs in male germ cells, necessary for mouse spermatogenesis. *Journal of Cell Biology* **185**, 235–249 (2009).

85. Modem, S., Badri, K. R., Holland, T. C. & Reddy, T. R. Sam68 is absolutely required for Rev function and HIV-1 production. *Nucleic Acids Research* **33**, 873–879 (2005).

86. Richard, S. *et al.* Sam68 haploinsufficiency delays onset of mammary tumorigenesis and metastasis. *Oncogene* **27**, 548–556 (2008).

87. Lukong, K. E. & Richard, S. Motor coordination defects in mice deficient for the Sam68 RNA-binding protein. *Behavioural Brain Research* **189**, 357–363 (2008).

88. de Paola, E. *et al.* Sam68 splicing regulation contributes to motor unit establishment in the postnatal skeletal muscle Review Timeline: Transaction Report. doi:10.26508/lsa.201900637.

89. Iijima, T., Iijima, Y., Witte, H. & Scheiffele, P. Neuronal cell type-specific alternative splicing is regulated by the KH domain protein SLM1. *Journal of Cell Biology* **204**, 331–342 (2014).

90. Traunmüller, L., Bornmann, C. & Scheiffele, P. Alternative splicing coupled nonsense-mediated decay generates neuronal cell type-specific expression of SLM proteins. *Journal of Neuroscience* **34**, 16755–16761 (2014).

91. Danilenko, M. *et al.* Binding site density enables paralog-specific activity of SLM2 and Sam68 proteins in Neurexin2 AS4 splicing control. *Nucleic Acids Research* **45**, 4120–4130 (2017).

92. Pagliarini, V. *et al.* Sam68 binds Alu-rich introns in SMN and promotes pre-mRNA circularization. *Nucleic Acids Research* **48**, 633–645 (2020).

93. Ye, J. & Blelloch, R. Regulation of pluripotency by RNA binding proteins. *Cell Stem Cell* vol. 15 271–280 (2014).

94. Itskovitz-Eldor, J. *et al.* Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Molecular medicine (Cambridge, Mass.)* **6**, 88–95 (2000).

95. Yang, M. J. *et al.* Novel method of forming human embryoid bodies in a polystyrene dish surface-coated with a temperature-responsive methylcellulose hydrogel. *Biomacromolecules* **8**, 2746–2752 (2007).

96. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **11**, (2016).

97. Buchbender, A. *et al.* Improved library preparation with the new iCLIP2 protocol. *Methods* 1–16 (2019) doi:10.1016/j.ymeth.2019.10.003.

98. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biology* **16**, (2015).

99. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008).

100. Knott, G. J., Bond, C. S. & Fox, A. H. The DBHS proteins SFPQ, NONO and PSPC1: A multipurpose molecular scaffold. *Nucleic Acids Research* vol. 44 3989–4004 (2016).

101. Yarosh, C. A. *et al.* TRAP150 interacts with the RNA-binding domain of PSF and antagonizes splicing of numerous

PSF-target genes in T cells. *Nucleic Acids Research* **43**, (2015).

102. Heber, S. *et al.* Staufen2-mediated RNA recognition and localization requires combinatorial action of multiple domains. *Nature Communications* **10**, (2019).

103. Solomon, S. *et al.* Distinct Structural Features ofCaprin-1 Mediate Its Interaction with G3BP-1 and Its Induction of Phosphorylation of Eukaryotic Translation InitiationFactor 2α, Entry to Cytoplasmic Stress Granules, and Selective Interaction with a Subset of mRNAs. *Molecular and Cellular Biology* **27**, 2324–2342 (2007).

104. Huang, H. *et al.* Recognition of RNA N 6 -methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature Cell Biology* **20**, 285–295 (2018).

105. Frisone, P. *et al.* SAM68: Signal Transduction and RNA Metabolism in Human Cancer. *BioMed Research International* **2015**, (2015).

106. Zhao, Y. *et al.* MyoD induced enhancer RNA interacts with hnRNPL to activate target gene transcription during myogenic differentiation. *Nature Communications* **10**, (2019).

107. Miles, W. O., Tschöp, K., Herr, A., Ji, J. Y. & Dyson, N. J. Pumilio facilitates miRNA regulation of the E2F3 oncogene. *Genes and Development* **26**, 356–368 (2012).

108. Muthu, M., Cheriyan, V. T. & Rishi, A. K. CARP-1/CCAR1: A biphasic regulator of cancer cell growth and apoptosis. *Oncotarget* **6**, 6499–6510 (2015).

109. DeBoer, E. M. *et al.* Prenatal deletion of the RNA-binding protein HuD disrupts postnatal cortical circuit maturation and behavior. *Journal of Neuroscience* **34**, 3674–3686 (2014).

110. Ince-Dunn, G. *et al.* Neuronal Elav-like (Hu) Proteins Regulate RNA Splicing and Abundance to Control Glutamate Levels and Neuronal Excitability. *Neuron* **75**, 1067–1080 (2012).

111. Klæstad, E. *et al.* MRPS23 amplification and gene expression in breast cancer; association with proliferation and the non-basal subtypes. *Breast Cancer Research and Treatment* **180**, 73–86 (2020).

112. Huang, Y. *et al.* Elevated expression of PTCD3 correlates with tumor progression and predicts poor prognosis in patients with prostate cancer. *Molecular Medicine Reports* **18**, 3914–3922 (2018).

113. Holdt, L. M., Kohlmaier, A. & Teupser, D. Molecular roles and function of circular RNAs in eukaryotic cells. *Cellular and Molecular Life Sciences* **75**, 1071–1098 (2018).

114. Hansen, T. B. *et al.* MiRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO Journal* **30**, 4414–4422 (2011).

115. Zhou, C. *et al.* Genome-Wide Maps of m6A circRNAs Identify Widespread and Cell-Type-Specific Methylation Patterns that Are Distinct from mRNAs. *Cell Reports* **20**, (2017).

116. Livi, C. M., Klus, P., Delli Ponti, R. & Tartaglia, G. G. CatRAPID signature: Identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics* **32**, 773–775 (2016).

117. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* vol. 19 327–341 (2018).

118. Busch, A., Brüggemann, M., Ebersberger, S. & Zarnack, K. iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites. *Methods* **178**, (2020).

119. Morris, K. v. & Mattick, J. S. The rise of regulatory RNA. *Nature Reviews Genetics* **15**, (2014).

120. Fu, X.-D. Non-coding RNA: a new frontier in regulatory biology. *National Science Review* **1**, (2014).

121. Venables, J. P. *et al.* T-STAR/ETOILE: A novel relative of SAM68 that interacts with an RNA-binding protein implicated in spermatogenesis. *Human Molecular Genetics* **8**, 959–969 (1999).

122. Pandit, S. *et al.* Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. *Molecular Cell* **50**, (2013).

123. Holdt, L. M., Kohlmaier, A. & Teupser, D. Molecular roles and function of circular RNAs in eukaryotic cells. *Cellular and Molecular Life Sciences* **75**, 1071–1098 (2018).

124. Sah, R. *et al.* Timing of Myocardial *Trpm7* Deletion During Cardiogenesis Variably Disrupts Adult Ventricular Function, Conduction, and Repolarization. *Circulation* **128**, (2013).

125. Kim, B. J. *et al.* RERE deficiency leads to decreased expression of GATA4 and the development of ventricular septal defects. *Disease Models & Mechanisms* **11**, (2018).

126. Chothani, S. *et al.* Widespread Translational Control of Fibrosis in the Human Heart by RNA-Binding Proteins. *Circulation* **140**, (2019).