A comprehensive analysis of Med12 controlled (l)ncRNAs

and

characterization of a novel Sall1 antisense transcript

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

by

Bruno Filipe Teixeira Pereira

August 2019

All the work here mentioned was perform at the Max-Planck Institute for Molecular Genetics, in the Department of Developmental Genetics, under the supervision of Dr. Heinrich Schrewe, from November 2014 to June 2019.

1st Reviewer: Dr. Heinrich Schrewe

2nd Reviewer: Prof. Dr. Sigmar Stricker

Data of Defense: 2020.02.11

À minha esposa e ao meu filho

Contents

1. Introduction

1.1. Gene expression in eukaryotes

Since the first discovery over 70 years ago that DNA is capable of transferring genetic information between different strains of bacteria (Avery et al. 1944), DNA has been regarded as one of the most important molecules in cells. Despite the identical genomic information in practically all its cells, an organism is composed of multiple different cell types and tissues with diverse functions. The difference between cell types arise from the differences in gene expression between them (Moore et al. 2013). There are several layers acting in concert to control the expression of genes, which act in concert despite the difference in their functional mechanism.

One layer of gene expression control lies within the chemical modifications deposited on the nucleosome, the basic organizational unit of DNA, composed of the DNA wrapped around a protein complex formed by dimers of four different histones (H2A, H2B, H3 and H4). These histones contain a N-terminal that protrude from the nucleosome and can undergo a plethora of modifications such as methylation, acetylation and phosphorylation. Histone modifications are associated with activation or repression of the modified genomic regions. Thus, acetylation of the residue lysine 7 of H3 (H3K27ac) promotes a more permissive chromatin state, which allows transcription to occur. Histone modifications associated with active transcription are usually restricted to gene regulatory regions and active gene bodies (Shogren-Knaak et al. 2006). On the other, hand regions of highly condensed chromatin (heterochromatin) with suppressed transcription are often associated with trimethylation of lysine 27 of H3 (H3K27me3), a histone modification deposited by Polycomb repressor complex 2 (PRC2) (Boyer et al. 2006, Pauler et al. 2009).

The enrichment of certain histone modification in the genome can be used to identify regions with different condensation degrees. Euchromatin, which is less condensed and more permissive to gene expression is enriched in H3K27me3, while heterochromatin, more condensed and where transcription is repressed, is associated with higher levels of H3K9me3 (Becker et al. 2017). Some of the most extensively studied histone modifications and their association with a gene transcriptional status are summarized in Figure 1 (Iglesias-Platas et al. 2016).

The DNA can also be directly modified by addition of a methyl group at the 5’ position of cytosine residues, which predominantly occurs at cytosine-guanine dinucleotides (CpG). DNA methylation is associated with a higher condensation of chromatin and transcriptional repression, playing an important role in processes such as gene imprinting and silencing of repetitive DNA sequences (Suzuki et al. 2008).

Three different DNA methyl transferases (DNMTs) are responsible for the deposition and maintenance of DNA methylation: DNMT1, DNMT3A and DNMT3B (Lyko 2018).

Transcription factors (TFs) are proteins capable of binding DNA in a sequence-specific manner. They can bind to both proximal promoters of genes and more distal regulatory regions such as enhancers. TFs act mainly by promoting the recruitment of the transcription initiation machinery to the promoter of their target genes. In addition, these factors guide histone modifiers, which are complexes capable of adding and removing modifications in histone residues and chromatin remodelling complexes, thereby regulating target gene expression.

Many TFs are expressed in a tissue-specific manner and are key regulators in establishing tissue identity. For example, T-cell acute lymphocytic leukemia protein 1 (Tal1) is a key hematopoietic TF essential for the generation of hematopoietic precursors cells in the mouse (Robb et al. 1995). Sal-like 1 (Sall1) is another key regulator and depletion of this protein leads to mice born without kidneys or with severe defects of these organs (Nishinakamura et al. 2005).

Deletion of Lung Kruppel-like factor (Lklf) in mice leads to major defects in lung development, not affecting remaining organs formation (Wani et al. 1999). The vast majority of TFs act in combination with other TFs and other regulators of transcription, such as the above mentioned histone modifiers. This

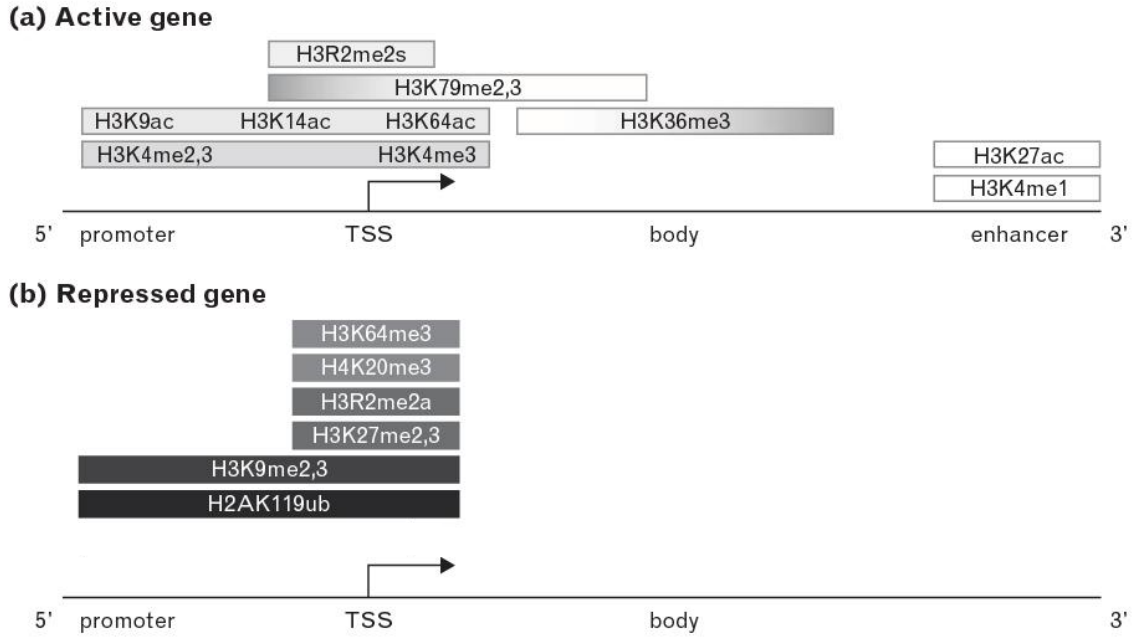


Figure 1- Histone modifications found at gene locus
Histone marks that are associated with **a)** active genes and enhancers and **b)** repressed genes (Iglesias-Platas et al. 2016)

mode of action allows TFs to have different roles depending on their interaction partners and the chromatin status. Serum response factor (SRF), for example, is a ubiquitously expressed TF which interaction with Nuclear factor of activated T-cells (NAFT) and Homocysteine-responsive endoplasmic reticulum-resident (HERP1) is important for the proper differentiation of smooth muscle, while its interaction with NK3 Homeobox 1 (NKX3-1) plays a role in proper differentiation and function of the prostate (Doi et al. 2005, Gonzalez Bosc et al. 2005, Zhang et al. 2008).

1.1.1. Transcription initiation mechanism

One of the major processes regulating spatial and temporal gene expression is the recruitment of RNA polymerase to the transcription start site (TSS) of genes and the subsequent gene transcription. In eukaryotes, three distinct polymerases have been identified, each involved in the transcription of specific classes of genes (Roeder et al. 1969). RNA polymerase I (Pol I) produces the vast majority of cellular RNAs, which consist of ribosomal RNA (rRNA). Pol III specializes in the transcription of small RNAs like transfer RNAs (tRNAs) and 5S rRNA. All messenger RNAs (mRNAs) and other non-coding RNAs (ncRNAs) are transcribed by Pol II (Khatter et al. 2017).

Purified Pol II is capable of melting double-stranded DNA, transcribe it and proofreading the synthesized RNA. However, this purified enzyme does not recognize promoter sequences nor does it respond to regulatory signals. These additional functions are performed by general transcription factors (GTFs) that associate with Pol II, namely TFIIA, TFIIB, TFIID, TFIIF, TFIIE and TFIIH. Together with Pol II, the GTFs form the pre-initiation complex (PIC), which is assembled near a gene TSS (Cramer et al. 2000).

This assembly is a process that occurs in a stepwise manner, involving the recruiting of additional co-activators, as depicted in Figure 2. Binding of the core promoter by TFIID comprises the first step in PIC assembly (Orphanides et al. 1996). The core promoter consists of the minimal DNA sequence necessary to specify basal transcription and can contain elements such as TATA-box, B recognition element (BRE), initiator element (Inr) and downstream promoter element (DPE). TFIID, one of the GTFs, is composed of the TATA-binding protein (TBP) and 14 TBP-associated factors (TAFs) and is able to identify and bind the different elements present at the promoter. The TBP subunit recognizes and binds the TATA-box sequence and is sufficient to drive expression from promoters that contain this sequence.

However, most metazoan promoters lack TATA-box and as such other elements are recognized by different subunits of TFIID. TAF1 and TAF2 subunits bind Inr elements, while DPE are recognized by TAF6 and TAF9 (Burke et al. 1997, Chalkley et al. 1999). TFIIA and TFIIB interact specifically with TBP bound to

DNA. TFIIA stabilizes the TBP-DNA interaction and protects it from negative regulators (Weideman et al. 1997). Apart from binding to BRE, TFIIB also promotes selective binding of Pol II to TFIID at the promoter by interacting with Pol II and TBP (Leuther et al. 1996). Binding of Pol II and TFIIF to the assembled GTFs stabilizes the whole complex. With Pol II bound to the promoter region, the DNA strand is guided to the cleft of Pol II which results in recruitment of TFIIE by TFIIF, a crucial step for binding of TFIIH (Kornberg 2001). TFIIH is composed of 10 subunits, including a Pol II carboxyl-terminal domain (CTD) kinase-cyclin pair (Kin28-Cyclin H) and two DNA helicases (xeroderma pigmentosum type B (XPB) and XPD) (Schultz et al. 2000) that facilitate promotor melting, leading to transcription initiation (Douziech et al. 2000). Although Pol II and GTFs complex assembly is sufficient for proper transcription initiation *in vivo,* this complex fails to respond to activators bound at distal regulatory regions. This responsiveness is conferred by a macromolecular complex termed Mediator. The Mediator complex integrates regulatory signals from a single or multiple TFs bound to enhancers and transduces the information to Pol II and the remaining GTFs.

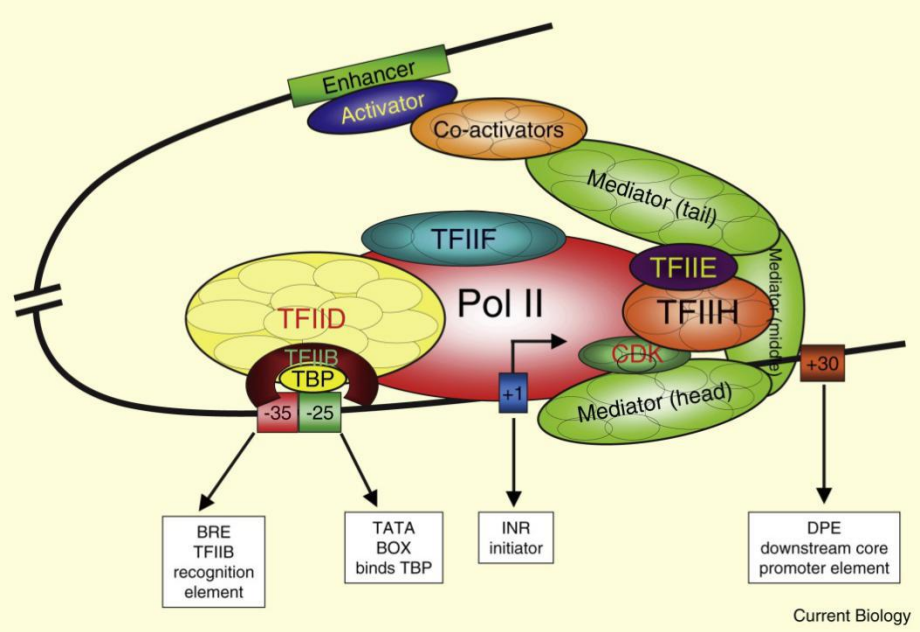


Figure 2 – Schematic summary of PIC
Core elements are present at -35, -25 and +1 bp from TSS. PIC assembly near the TSS is initiated by binding of TBP to the TATA box, which is then followed by step wise bind of TFIIB, Pol II/TFIIF, TFIIE and TFIIH. PIC assembly can be initiated without a TATA box, by binding of TAFs from the TFIID to other promoter elements. PIC assembly is stimulated by TFs bound to distal enhancer, which recruit different coactivators, like Mediator, that in turn bridges TFs at enhancers and the general Pol II machinery. (Krishnamurthy et al. 2009).

With the advent of Cryogenic electron microscopy (cryo-EM), it became possible to generate atomic models of complex structures. Using this method, it was possible to confirm the composition and interactions between GTFS and Pol II that had been described in a series of previous studies (Hantsche et al. 2017). Additionally, new aspects of the interactions were observed. Within the factors, mobile domains were identified that adopt a specific location within the core Mediator. Essential regions for cell viability were observed in fixed structures while non-essential, non-conserved regions were often mobile. The position of TFIID at the promoter was determined to be located on the side of Pol II lobe and protrusion, allowing contacts between this GTF and the downstream DNA, which explains the contribution of TFIID in promoter recognition. TFIIH, a GTF not included in previous crystalized structures of the PIC, was analysed by cryo-EM as well. Consistent with its role in DNA melting, it was shown to be located on the downstream DNA.

1.1.2. Enhancers in gene activation

Enhancers were originally described as small sequences of DNA capable of driving target gene expression regardless of distance and orientation. Additional features are characteristic of these elements, such as an open state of the surrounding chromatin, the presence of TF binding motifs and enrichment of bound co-activators (Bulger et al. 2011). With the development of genome-wide methods, such as chromatin immunoprecipitation sequencing (ChIP-seq), these techniques have permitted a systematic annotation and prediction of enhancers. They rely on the combination of multiple features, such as accessible chromatin (as assessed by DNase hypersensitivity assay (DNase-seq)), the presence of different histone marks such as H3K27ac, a higher ratio of H3K4me1 compared to H3K4me3, binding of co-activators such as p300 and clustered binding of multiple TFs (assessed by Chip-seq) (Calo et al. 2013). These methods have allowed the prediction of a high number of enhancers through the genome (400,000 to ~1 million in human cells) (Shlyueva et al. 2014). Enhancer activity has been linked to transcription from these elements by early reports (Tuan et al. 1992). More recently, conclusive evidence has demonstrated the transcription of non-polyadenylated, non-coding transcripts originating from enhancers, termed enhancer RNAs (eRNAs) (Kim et al. 2010), which were found to be expressed in a cell-type specific manner.

1.2. The Mediator complex

Mediator was first discovered in the yeast *S. cerevisiae* in two independent studies. While the Young laboratory was studying extragenic suppressors of a pol II CTD mutation (Thompson et al. 1993), the Kornberg laboratory was identifying factors necessary for the response to activator proteins in a Pol II and GTFs reconstituted system (Flanagan et al. 1991). Although these studies demonstrated that yeast Mediator is essential for TF-dependent gene activation *in vitro*, there was no evidence for conservation of the complex in other species. However, several activating cofactors that were described based on a variety of functional assays were later identified as homologs of Mediator subunits.

The Positive cofactor complex (PC2), isolated in Hela cells as a complex capable of stimulating transcription upon addition of synthetic activator GAL4-AH (Kretzschmar et al. 1994), was further characterized in a later study and shown to consist of Mediator subunits (Malik et al. 2000). Likewise, the TRAP complex (thyroid hormone receptor-associated proteins), described as a group of nuclear proteins that associate with human thyroid hormone receptor alpha in the presence of thyroid hormone (Fondell et al. 1996), was also identified as a Mediator related complex. Due to the different assays used to identify these complexes, diverse nomenclatures were used to name the subunits of Mediator, such as Med14, which is also known as TRAP170, ARC150 (Activator-recruited cofactor), CRSP2 (Cofactor required for Sp1 transcriptional activation) in humans and as Rgr1 in *S. cerevisiae*. A common nomenclature was created in 2004 in order to facilitate communication between researchers and to unify description of homologous proteins between different species (Bourbon et al. 2004).

Mediator is a highly conserved structure, present in all eukaryotes. Despite the conservation in structure and general function of Mediator, the conservation of the individual subunits between species is poor. Similarity in subunit sequence between human and yeast ranges between 12-42%, while eight subunits are not present in yeast (Poss et al. 2013). The divergence observed for orthologous subunits can be explained by the computational prediction of at least one intrinsically disordered region (IDR) in over 70% of the yeast and human subunits. These IDR evolve more rapidly than structured sequences and contribute to Mediator interaction with a vast number of TFs and to its intrinsic flexibility (Toth-Petroczy et al. 2008).

In yeast, Mediator contains 25 subunits (30 in mammals) that are organized in four different modules, designated as head, middle, tail and kinase modules (Tsai et al. 2013, Tsai et al. 2014). One of the first applications of mass spectrometry (MS) in the study of Mediator revealed two stable forms in yeast. The smaller one, which consists of a 21 subunit complex (26 in mammals) composed by the head, middle and

tail modules, is called the core Mediator (matching the PC2 complex previously identified). The larger form (previously characterized as the TRAP complex) is constituted of 25 subunits (29 in mammals), including the kinase module in addition to the core Mediator (Liu et al. 2001). Early immunoprecipitation and fractionation studies indicated that MED26 (CRSP70), part of the core Mediator, dissociates the complex upon kinase module binding (Taatjes et al. 2002). However, more recent evidence obtained using MS-based multidimensional protein identification technology (MudPIT) suggests that a smaller fraction of Mediator contains both MED26 and the kinase module (Sato et al. 2004).

Due to the large size of the whole complex most structural studies on Mediator relied on portions of the complex. However, developments in cryo-EM allowed the analysis of the complete Mediator-PIC complex (Robinson et al. 2016). This work allowed to complement the structures previously described, to position the Mediator in relation to the PIC complex and to determine the position of the tail module, which had so far only been predicted (Figure 3). Chemical cross-linking applied to Mediator-PIC complex revealed 71 cross-links between the two complexes. As demonstrated by the surface map, the interactions between Mediator and PIC are mostly restricted to head and middle module (Figure 3).

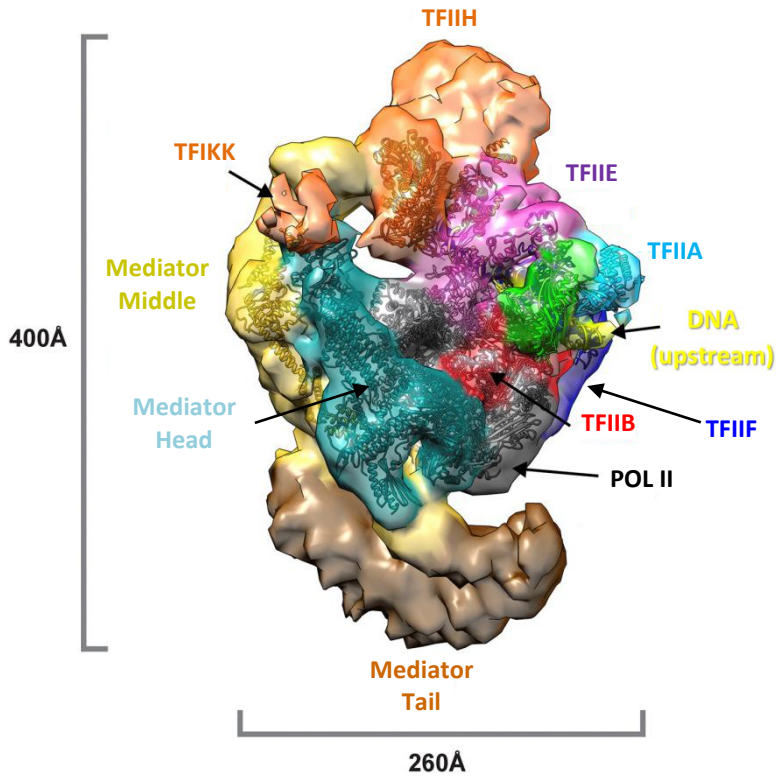


Figure 3 - Cryo-EM Structure of the Mediator-PIC Complex
Surface representation of Mediator-PIC cryo-EM data. Adapted from (Robinson et al. 2016)

1.2.1. Functions of the Mediator complex

The structure of Mediator reflects a functional organization. The head and middle modules are necessary for Mediator function in regulating transcription. When these two modules are complemented with MED14, they reconstituted a functional basal Mediator *in vitro*. On the other hand, tail and kinase modules interact with TFs, which allow Mediator to be responsive to regulatory signals (Cevher et al. 2014).To regulate transcription, Mediator is recruited to enhancers in metazoans and to upstream activating sequences (UAS) in yeasts *via* specific TFs. Following recruitment, Mediator can then coordinate PIC assembly at promoters (Figure 4a), an observation supported by different *in vitro* studies (Ebmeier et al. 2010, Lin et al. 2011, Chen et al. 2012). Chromatin immunoprecipitation sequencing (ChIP-seq) data for diverse Mediator subunits revealed that Mediator occupies predominantly enhancers, with a low occupancy at promotors (Andrau et al. 2006). Depletion of Kin28, the TFIIH subunit responsible for phosphorylation of Pol II CTD, results in a promoter escape defect. This results in an arrest of Pol II at the promoter. When Kin28 was depleted, it was observed that Mediator stabilized at promoters, confirming that the complex indeed associates with promoters *in vivo* (Jeronimo et al. 2014). These data reveal that, despite its important role on PIC assembly, Mediator occupancy at promoters is transient.

A recent study by the Roeder laboratory has shown that in yeast, Mediator is recruited to UAS with all Mediator modules detected at these elements (Petrenko et al. 2016). At promoters, however, no kinase module could be detected, even in Kin28-depleted mutants where Mediator stabilization at promoters was observed (Jeronimo et al. 2014). The lack of kinase module at promoters is consistent with a previous model in which the kinase module sterically blocks Mediator interaction with Pol II (Elmlund et al. 2006). These observations suggest that Mediator undergoes a conformational change during transcription activation, ejecting the kinase module upon PIC integration, which in turn allows interaction between Pol II and Mediator (Figure 4b). In contrast, mouse cell ChIP-seq data for Med12 subunit, which is part of the kinase module, obtained from provided evidence that the kinase module can be found at promoters, although at lower levels than at enhancers (Kagey et al. 2010). Sequential ChIP experiments have demonstrated that a single Mediator complex connects UAS and core promoters. This observation links Mediator to DNA looping, a process that brings enhancers and promoters to close proximity (Reavey et al. 2015). This hypothesis is corroborated by observation of interactions between Mediator and cohesin in murine embryonic stem cells (mESCs), using co-immunoprecipitation assays.

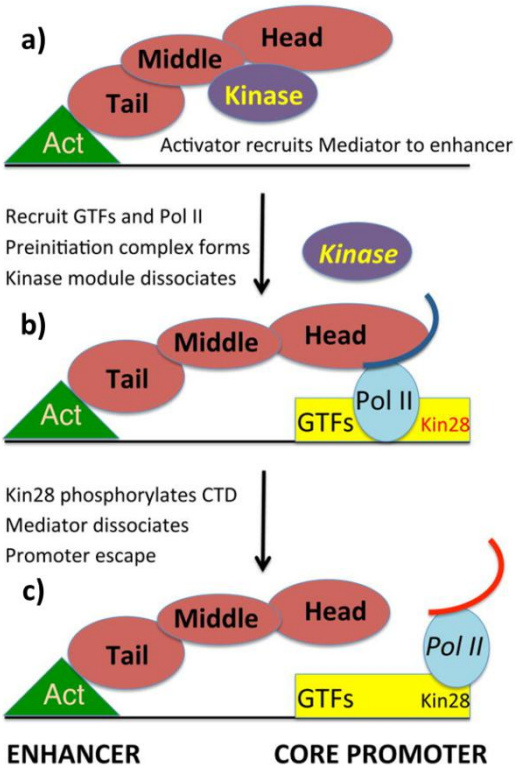


Figure 4 - Conformational change of Mediator during transcription activation
a) Activator proteins (Act) bound to enhancers recruit Mediator through interactions with its subunits. Most of subunits capable of interacting with TFs reside in the tail module. The recruited Mediator contains the kinase module. **b)** TFs recruit the GTFs and Pol II to the core promoter, which also includes the protein Kin28 of the TFIIH. This recruitment leads to the dissociation of the kinase module and interaction between the head Module and Pol II (blue line indicates CTD of Pol II). Conformational Change of Mediator leads to the PIC formation and consequent transcription initiation. **c)** Phosphorylation of Pol II CTD (red line) by Kin28 leads to dissociation of Mediator, promoter escape from Pol II and to transcription initiation (Petrenko et al. 2016).

Additionally, ChIP-seq data for Mediator and cohesin revealed both complexes bound at enhancers and promoters of a variety of active genes. DNA looping between enhancer and promoter of these co-bound genes was further characterized using chromosome conformation capture (3C), which reveals interaction frequency between genomic regions. Upon Med12 or Structural maintenance of chromosomes 1a (Smc1a) depletion, there was a decreased frequency of the interactions between promotor and enhancer of genes bound by these two proteins, further supporting the role of Mediator in DNA looping (Kagey et al. 2010).

In order to assess the impact of the kinase module on Mediator occupancy at enhancers and promoters, Med13-depleted mutants were used for ChIP-seq analysis. These mutants assembled the Mediator complex without the kinase module (Jeronimo et al. 2016). ChIP-seq data for Med15 (tail) and

Med7 (middle), obtained from Kin28/Med13-depleted mutants, revealed that, although the middle and tail module occupancy at promoters was only slightly affected, their occupancy at UAS drastically increased. These observations support a model were kinase module antagonizes Mediator-UAS interaction.

Additionally, Cyclin-dependent kinase 8 (Cdk8), the kinase subunit of the kinase module, is capable of phosphorylating several TFs which often results in their degradation. This degradation further supports the mentioned model, since - by degrading TFs - the kinase module prevents these factors from recruiting Mediator to the enhancer of their target genes (Fryer et al. 2004, Poss et al. 2016).

Tailless Mediator, generated in yeast by Med15 and Med3 deletions, was depleted from UAS. Depletion of Mediator from UAS in these mutants supports the hypothesis of TFs recruiting Mediator through interactions with subunits of the tail module. On the other hand, in metazoans, several TFs recruit Mediator through the interaction with modules other than the tail (Ito et al. 1999, Zhou et al. 2006). Although no Mediator was found above background levels at promoters in tailless yeast mutants, cells were viable and showed only minor gene expression defects. Depletion of Kin28 in Med15/Med3 mutants revealed a smaller occupancy of promoters by Mediator.

These results led to the conclusion that Mediator is capable of binding promoters without being first recruited to UAS, contrary to the previous model in which Mediator first had to be recruited to UAS. It also revealed that Mediator's essential function is to act on promoters by contributing to PIC assembly, since expression profile was only slightly affected although no Mediator was present at UAS, (Jeronimo et al. 2016). This function is further supported by the observation of several contacts between Mediator and PIC components, such as RNA Pol II, TFIIB, TFIIE, TFIIF and TFIIH (Plaschka et al. 2015).

Thus, the Mediator complex plays an important role in PIC assembly, with additional functions in other steps throughout gene transcription. In order for transcription to initiate, the PIC complex must be disrupted, a process known as promoter escape. Despite multiple contacts between Pol II and Mediator, it is the CTD-Mediator interaction that allows the entry of Mediator into the PIC complex, with the affinity for the other contacts at least two orders of magnitude lower. The importance of this specific insertion was confirmed by the lack of detectable affinity by Mediator for Pol II without CTD. In yeast, a seven residue sequence (heptad) is repeated 26 times in the CTD of Pol II (52 repeats in mammals). Through contacts with TFIIH and the remaining GTFs, Mediator stabilizes TFIIH in the complex. Additionally, Mediator, upon binding to PIC, creates a path that guides CTD to the TFIIK submodule. The Kin28 kinase of TFIIK phosphorylates all heptads, disrupting Mediator-CTD contacts, which results in Pol II release and promoter escape (Robinson et al. 2016).

For several years, transcription initiation was considered the rate limiting step during transcription. However, pausing of Pol II after transcription of 30-60 nucleotides was verified in multiple genes (Core et al. 2008), revealing that for these genes, pause release was the limiting step. This phenomenon can result from intrinsic factors, such as specific residues in the Pol II protein which mutation increase pause rate, certain DNA sequences that cause an abrupt reduction of melting temperature, stabilizing RNA-DNA duplexes, or even the transcribed RNA, which can fold into secondary structures that interfere with elongation (Herbert et al. 2006, Kaplan et al. 2012, Hein et al. 2014). However, pausing can also be induced by factors such as DRB-sensitivity-inducing factor (DSIF) and Negative elongation factor (NELF), which - in a purified system - were sufficient to induce pausing of Pol II (Adelman et al. 2012). In order to release Pol II, Positive transcription elongation factor b (P-TEFb) phosphorylates DSIF and NELF, neutralizing their repressive effect. P-TEFb has been shown to be recruited in a CDK8-dependent manner, inducing gene expression by releasing Pol II into active elongation (Donner et al. 2010). Additionally, by interaction with Med26 and CDK9, P-TEFb induces expression of target genes in mESCs. In an *in vitro* reconstituted system, purified Mediator complex was able to promote expression of genes where DSIF induced Pol II pausing, even in the absence of P-TEFb. This revealed the possibility of Mediator not only recruiting this factor to release Pol II pausing, but to act together with P-TEFb in its pausing release function.

The Mediator complex has been described as a molecular bridge, transducing information from activators at enhancer to Pol II and GTFs at promoter. More recently, Mediator has been associated with a new class of enhancers, termed super enhancers (SE). While traditional enhancers typically span a few hundred base pairs , SEs are clusters of enhancers, spanning up to 50 kb (Whyte et al. 2013). SEs share many features with typical enhancers, including presence of Mediator and histones modifications, such as H3K27ac and H3K4me, however, at much higher levels, typically over one order of magnitude. The same increase was observed for other enhancer associated factors such as TFs, co-factors, chromatin regulators, Pol II and eRNAS. The high levels of these features are due not only to the size of the domains but also to the higher density of the enhancers that constitute SEs. These super enhancers are typically in control of key genes in cell-type-specific processes, such as most genes that control ESCs pluripotency.

Genes under the control of SEs are expressed at higher levels than genes under the control of typical enhancers. However, SEs are more sensible to loss of co-activators, such as Mediator. Upon Mediator depletion in ESCs, the genes more affected were the ones associated with SEs, revealing the critical role of the complex in these clusters. TFs usually contain at least one of each of two different domains: a DNA-binding domain (DBD) and an activation domain (AD). Although a deep understanding of structure

and function has been documented for DBDs, ADs are less well understood. These domains contain intrinsically disordered regions (IDRs), which make their structural characterization difficult. The low complexity of ADs contributes to the recent observation that GNC4, a yeast master TF, binds to Med15 at multiple sites and in different conformations (Tuttle et al. 2018). The current "lock and key" model, according to which two proteins bind in defined complementary conformations, does not explain this kind of "fuzzy" complexes. Additionally, several observations regarding TFs are not explained by the "lock and key" model: hundreds of TFs with different ADs interact with a small subsets of coactivators, such as Mediator. Furthermore, ADs are interchangeable between different TFs and with the IDR present in most coactivators. Another model consolidating all of these observations was recently suggested (Hnisz et al. 2017). According to this model, transcriptional regulation is impelled by phase-separated condensates that allow a compartmentalization and concentration of the transcriptional machinery and regulators. This physical separation explains the high density of molecules detected in SEs, which is further supported by the observation that enhancer elements within these clusters are in close spatial proximity with the other enhancers and with the regulated promoters. In eukaryotic cells, membraneless organelles, such as nucleoli and Cajal bodies, allow concentration of essential reactions through formation of such phase separated droplets.

This model was recently corroborated by the observation that IRDs of both MED1 and Bromodomain-containing protein 4 (BRD4) were capable of generating phase separated droplets in an *in vitro* assay, which were able to concentrate the transcriptional machinery in a nuclear extract (Sabari et al. 2018). Another study by the same group revealed that MED1 and MED15 were able to generate phase separated droplets *in vitro* (Boija et al. 2018). Additionally, when MED1 was combined with several known interactors (e.g. NANOG, GATA binding protein 2 (GATA2), SRY-box 2 (SOX2)), the TFs were found concentrated together with the subunits, despite the capacity of these TFs to form droplets on their own. Furthermore, residues required for target gene activation for Octamer-binding transcription factor 4 (OCT4) and GCN4 were also required for phase separation with Mediator. These studies demonstrated that Mediator not only plays an important role in phase separation, but also that this separation is important for the proper regulation of transcription.

Several TFs have been shown to interact with Mediator and a recent study has increased the binding repertoire of Mediator by discovering a class of ncRNAs that interact with the complex. While studying the function of ncRNAs-activating (ncRNA-a), the Shiekhattar laboratory developed a luciferase reporter assay to analyse the transcriptional control of the genes analysed. Of the tested factors with known activator or enhancer functions, only Mediator subunits decreased the reporter signal when depleted

(Lai et al. 2013). Depletion of Mediator subunits in human embryonic kidney 293 cells (HEK293) resulted in a decreased expression in a subset of the ncRNAs-a and in all of their targets genes. Mediator and Pol II were detected at the promoter of ncRNAs-a target genes, with a loss of occupancy observed upon Mediator subunit or ncRNAs-a depletion. By fractionation of affinity-purified Mediator and UV-crosslink RNA Immunoprecipitation (UV-RIP) the authors could additionally detect a physical interaction between ncRNAs-a and Mediator (Lai et al. 2013).

1.2.2. Mediator subunit functions

Mediator subunits can be classified into two groups: those that are required for the overall function of the complex and those that interact with specific TFs and other transcriptional regulators. The first group is required for the expression of all Pol II transcripts and includes, for instance, subunits with important structural functions that maintain the structural integrity of the complex. The second group consists of subunits that act as an interface through which transcription regulators are able to recruit Mediator to their target genes. Subunits of this group are involved in the transcription of a subset of genes, namely those under the control of the TFs. One such example are nuclear receptors that interact with Mediator through Med1. Depletion of this subunit in mouse fibroblasts resulted selective loss of a subset of nuclear receptor dependent pathways (Chen et al. 2007).

Loss of structural integrity of Mediator results in an impairment of Mediator function. This effect was observed for several subunits with important structural roles, such as Med17 and Med14. When Med17 (srb4) was depleted in yeast, the head and kinase modules dissociated, ceasing all Pol II transcription (Linder et al. 2006). Depletion of the Drosophila homolog of Med17 (dTRAP80) resulted in a similar effect (Boube et al. 2000). In a study where head or middle subunits from yeast Mediator were expressed in bacteria, the subunits assembled into the respective modules. However, when both modules were purified and combined, they did not form a bimodular complex and as such were not capable of stimulating basal transcription in an *in vitro* assay (Cevher et al. 2014). Only upon addition of Med14 the two modules were able to associate and strongly induce basal transcription. These data, together with the observation that deletion of the C-terminus of Med14 leads to a loss of the tail module from the Mediator complex (Li et al. 1995), highlight the essential role of Med14 in Mediator structure .

Different studies showed that of the 25 subunits in yeast, 10 are essential for cell viability (Med4, Med6, Med7, Med8, Med10, Med11, Med14, Med17, Med21 and Med22) and depletion of each tested

subunit in mice resulted in embryonic lethality (Soutourina 2018). Besides the Mediator subunits' essential role for embryonic development, specific functions of subunits have been identified.

Through Med23, Mediator is recruited to a subset of genes under the control of ELK1. Deletion of this subunit in mouse embryonic fibroblasts (MEFs) prevents adipogenesis in culture, suggesting Med23 as an endpoint of the insulin signalling pathway (Wang et al. 2009). Mediator isolated from Med1 (Trap220) KO MEFs was stable and transcriptionally active, revealing that Med1 is not essential for the complex integrity. This subunit is the interface through which Mediator is recruited by different nuclear receptors and its depletion in MEFS affected nuclear receptor-dependent gene expression. The Roeder laboratory concluded that this was a Med1 specific effect and not a general Mediator defect, since expression mediated by p53 and VP16, which recruit Mediator through Med25, was not affected (Ito et al. 2000, Vojnic et al. 2011, Lee et al. 2018).

A subunit of Mediator complex can contain binding sites for different transcriptional regulators, as *in vitro* studies in yeast have demonstrated. VP16 and Gal4 bind Med15 (Gall11) and while both bind the same region in this subunit, Gal4 is capable of binding to an additional site. Med15 mutants lacking these binding regions showed no interaction with the mentioned activators, which affected expression of VP16 and Gal4 target genes (Park et al. 2000). Although certain TFs bind to a specific subunit like mentioned above, others can interact with multiple subunits. One example of the latter is RE1 Silence TF (REST), a key TF in neuronal gene repression in non-neuronal cells, which recruits repressors such as G9a histone methyl transferase to its target genes. Co-immunoprecipitations assays in human cells revealed independent interactions between REST and two different Mediator subunits: Med9 and Med26 (Ding et al. 2009). While depletion of either subunit with small interfering RNA (siRNA) did not affect REST-Mediator association, a double depletion did. Quantification of REST repressed genes expression in HeLa cells revealed that only upon depletion of both subunits was the repressive effect disrupted. These data suggest a synergetic effect of both subunits in modulating REST repression.

Another cooperative effect, between Med15 and Med16, during heat shock response in yeast has been observed (Kim et al. 2013). When yeasts were subjected to above optimal temperatures, different heat shock genes (HSG) started to be expressed. Expression of these genes was induced by Hsf1 recruitment of Mediator to the promoter of HSG. Deletion of either Med15 or Med16 decreases recruitment of Mediator while a double deletion completely prevents it.

1.2.3. The Kinase module

The kinase module is reversibly coupled to Mediator and it is its presence that distinguishes the two main stable forms of Mediator. In yeast, the kinase module is composed of 4 subunits: Med13, Med12, Cdk8 and Cyclin C (CycC). In vertebrates, due to duplication events, three of the subunits acquired paralog pairs, namely Med13-like (Med13L), Med12L and Cdk19. The kinase module is always composed of 4 subunits, including CycC and one paralog of each pair. By expressing and capturing kinase subunits with a HaloTag in human cells followed by MS, it was possible to determine the composition of the kinase module and its interaction with the core Mediator (Daniels et al. 2013). It was observed that paralogs were mutually exclusive, but would interact with any of the other pairs, which allows for up to 8 different kinase module compositions. As mentioned before, while in yeast the Med26 subunit is not found in Mediator with kinase module, in vertebrates, a small fraction of Mediator contains both kinase and Med26. Interestingly, the inclusion of Med26 was only observed when Med13L was present in the kinase module. The small fraction of Mediator-kinase-Med26 complexes observed in vertebrates arises from the Med13L subunit present in these. This observation correlates with the lack of such complexes in yeast, since no Med13L homolog exists.

Multiple studies led to the initial characterization of the kinase module as a negative regulator of Mediator functions. It was observed that only upon dislocation of the kinase module Mediator was able to interact with Pol II (Elmlund et al. 2006). Additionally, the kinase module subunit Cdk8 phosphorylates several substrates, including Cyclin H, part of TFIIH, preventing transcription activation (Akoulitchev et al. 2000). Cdk8 also phosphorylates a variety of TFs, targeting them to degradation and preventing their target genes induction (Fryer et al. 2004). In fruit flies, Med12 (kohtalo) and Med13 (skuld) suppress a subset of Hedgehog target genes, which are important for normal differentiation in the eye (Janody et al. 2003), while in *C. elegans* both play a role in the repression of Wnt target genes involved in asymmetric cell division (Yoda et al. 2005).Med12 (dpy-22) is additionally involved in the repression of Ras target genes, which play a crucial role during vulval fate specification in C. elegans (Moghal et al. 2003).

However, more recent studies revealed the involvement of kinase module in transcription activation. In mESC, Cdk8 phosphorylates receptor-activated Smads, which leads to recruitment of Yes associated protein 1 (YAP). Recruitment of YAP results in expression of canonical Bone morphogenetic protein (BMP) pathway genes involved in repression of neuronal differentiation (Alarcon et al. 2009). More evidence on the activating role of the kinase module comes from the *Drosophila* TF Pygopus, which recruits Mediator through interactions with Med12 and Med13, activating Wnt targets (Carrera et al. 2008). Using a genome-wide RNA interference (RNAi) screen, it was observed that Med12 and Med13 act as co-

activators of the TFs RUNX and GATA during crystal cell differentiation in *Drosophila* blood cells. This TF activation is independent of Cdk8 and CycC, although all four subunits are involved in the emergence and proliferation of this cell lineage (Gobert et al. 2010).

As mentioned before, the kinase module is able to dissociate from Mediator. A study from the group of Dylan Taatjes provided evidences that the kinase module exists as a stable and active complex in human cells and performs regulatory functions independent of Mediator (Knuesel et al. 2009). The stable kinase module was isolated from Hela nuclear extracts and confirmed by MS to be constituted by all expected four subunits without representation of core subunits. Recombinant expression of kinase subunits assembled in a complex with a 1:1:1:1 stoichiometry. Cdk8, as part of this recombinant kinase module, phosphorylated H3 *in vitro* in a histone octamer core. However, when DNA was wrapped around the histone core CDK8 was not able to modify H3. Only upon kinase association with the core, Mediator did Cdk8 phosphorylate H3 in the chromatin context. These data confirmed that the kinase module can have functions independent of Mediator, but that the Mediator complex modulates its activity. It could additionally be shown that Med12, but not Med13, is important for Cdk8 activity. This fact was evidenced by the lack of activity when the kinase module was assembled without Med12, while a kinase module lacking Med13 showed kinase activity comparable to the four subunit complex (Knuesel et al. 2009). This role of Med12 is not observed in yeast since the Cdk8/CycC pair maintains its activity independently of both Med12 and Med13 (Myer et al. 1998).

1.2.4. The MED12 subunit

MED12 is one of the most interesting subunits of Mediator, being the one of the largest subunits and the only one described as a genetic hub due to its role in a variety of cell regulatory processes (Lehner et al. 2006). In all mammals, the MED12 gene is located on the X chromosome and contains 45 exons spanning over 25 kb and coding for a protein of around 230 kDa (Knuesel et al. 2009). Although expressed in all tissues, expression levels are tissue and age-dependent, reaching a peak at embryonic day 7 (E7.0) in mice and decreasing until birth to low but stable levels (Philibert et al. 1999). The role of MED12 in the overall function of Mediator has not been widely investigated, but a number of studies have demonstrated several gene-specific roles for this subunit.

Its role in early mouse development was described for the first time in a study by the Schrewe laboratory using embryos generated from mESCs, expressing either 5% of Med12 ($Med12^{hypo}$) or completely devoid of Med12 ($Med12^{null}$) (Rocha et al. 2010). The developmental defects observed in

these embryos together with the analysis of target gene expression, allowed determining the essential role of Med12 in both the canonical Wnt and Wnt/PCP (Planar cell polarity) pathways. While $Med12^{null}$ embryos failed to complete gastrulation and to activate T, a direct Wnt/β-catenin target and died at E7.5. The residual subunit in the $Med12^{hypo}$ allowed the activation of the early processes under Wnt control, but not latter processes. Different phenotypes observed in the $Med12^{hypo}$ embryos, such as neural tube closure, linked Med12 to the Wnt/PCP pathways. And in fact, loss of asymmetric distribution of Pricle1, a key factor in PCP, was detected in neural plate cells in $Med12^{hypo}$ (Rocha et al. 2010).

In order to clarify MED12 function in general transcription control, the Boyer laboratory, using an *in vitro* yeast two-hybrid screen, identified G9a as a binding partner of MED12 (Ding et al. 2008). G9a is a histone methyl transferase which, together with REST, represses neuronal genes in terminal differentiated non-neuronal cells by depositing repressive H3K9 mono- and demethylation. MED12 and G9a co-immunoprecipitated in cellular extracts and serial immunoprecipitations revealed that, together with REST, they form a trimeric complex in mammals. Depletion of MED12 by RNAi decreased G9a expression and Mediator associated H3 methyl transferase activity, culminating in de-repression of REST target genes. On the other hand, none of these effects was observed upon deletion of other subunits of the Mediator complex, such as CDK8 or MED23. In addition, MED12 depletion did not affect expression of REST or any of its corepressors, supporting a specific role for MED12 in G9a dependent REST targeted repression. Med12 interaction with Sox10 in mouse glia cells was identified. Glia-specific deletion of Med12 led to defects in glia terminal differentiation, since Sox10 failed to recruit Mediator to its target genes and induce their expression (Vogl et al. 2013).

The same study also observed an interaction between Sox10 and Med12L. This paralog is poorly characterized due to a lower expression than MED12 but also to a lack of available tools such as specific anti-MED12L antibodies. Thus, Med12l and Sox10 interaction was not further investigated. In human cells, MED12 binds GLI Family Zinc Finger 3 (GLI3), a target of Sonic Hedgehog (SHH) pathway and acts as a repressor of GLI3 transactivation activity (Zhou et al. 2006). Another MED12 binding partner is β-Catenin, the canonical Wnt signaling effector protein. This subunit is an essential interface for targeted recruitment of the Mediator and thus for proper expression of β-catenin targets genes (Kim et al. 2006).

The interaction of GLI3 and β-Catenin with MED12 and their respective role in SHH and Wnt signaling has been shown to be conserved in other organisms (Boube et al. 2000, Rocha et al. 2010).

Using zebrafish mutants, it was possible to observe a role for Med12 in several different developmental processes. Med12 mutants showed defects in neuronal and endoderm development including defects on formation of a subset of neuronal types, chondrogenesis, normal neural crest cell

development, organogenesis of the liver, kidney and pancreas. Most of these defects can be explained by the inability of TFs like Sox9, Sox32 and Foxa2 to induce expression of their target genes (Hong et al. 2005, Rau et al. 2006, Wang et al. 2006).

Hematopoietic-specific Med12 deletion in mice resulted in reduction of bone marrow and of hematopoietic stem and progenitor cells (HSPC), which led to the animals death within four days after deletion. Depletion of Med12 reduced H3K27ac in lineage specific enhancers, to which Med12 strongly binds and their consequent de-activation (Aranda-Orgilles et al. 2016). Depletion of other kinase subunits did not affect HSPC function. Furthermore, in these mutants all of the cellular Med12 was found in association with Mediator, revealing a role for Med12 within the core Mediator that is independent of the kinase module. In the same study, several phenotypes observed in murine cells were also observed in human cells, suggesting a conserved Med12 function in hematopoiesis. This observation is in agreement with a previous report where a single point missense mutation in zebrafish med12 resulted in myelopoiesis and late hematopoiesis defects (Keightley et al. 2011). MED12 can function in Mediator independently of the kinase module and new studies additionally support Mediator-independent functions. A fraction of MED12 was found to reside in the cytosol in different human tumorous lines, where it inhibits glycosylation of TGF-β2R (Huang et al. 2012). Upon MED12 depletion, TGF-β was strongly activated conferring resistance to several anti-cancer drugs.

As described above, MED12 plays a crucial role in various processes and pathways and in fact, in a study performed in *C. elegans*, where 65.000 pairs of genes were evaluated in their ability to genetically interact, Med12 has been described as one of the six hub genes, due to its genetic link to multiple different developmental pathways (Lehner et al. 2006). Several studies mentioned before support a conservation of MED12 genetic hub activity in additional species.

1.2.5. MED12 in human pathologies

MED12 has been associated with different human pathologies including tumours. Mutations in MED12 have also been associated with different X chromosome linked intellectual disability syndromes (XLID), specifically Opitz-Kavegia (FG), Lujan-Fryns and Ohdo Maat-Kievit-Brunner (MKB) type (Risheg et al. 2007, Schwartz et al. 2007, Vulto-van Silfhout et al. 2013). The mutations associated with FG (R961W and G958E) and with Lujan syndromes (N1007S) are in close proximity, situated in exon 21 and 22 of MED12, respectively. In turn, Ohdo mutations (G1148H, S1165P and H1729N) are more spread through the gene, located in exons 24, 25 and 37. While there are defects common in all of these syndromes,

such as intellectual disability, neonatal hypotonia and craniofacial defects (including prominent forehead, down slanting palpebral, high narrow palate and micrognathia), others are more specific. Agenesis of the corpus callosum is found in FG and Lujan syndromes, anal defects are present in the FG patients, hypernasal voice is a characteristic of Lujan syndrome and blepharophimosis arises from Ohdo mutations. Although the etiology of these syndromes is known, there is a lack of information regarding their pathogenesis. However, recent studies have highlighted the impact of these mutations in the normal function of MED12. MED12 missense mutations associated with FG and Lujan syndromes do not disrupt its interaction with G9a but prevent REST repression rescue in a reporter assay. These data, together with the observation that these mutants support β-catenin transactivation, suggest that the mutations described specifically disrupt MED12-REST corepressor function (Ding et al. 2008). Another study demonstrated that FG and Lujan mutations disrupt MED12 mediated repression of GLI3-dependent SHH (Zhou et al. 2012). These mutations resulted in a disruption of CDK8 recruitment through Mediator to GLI3 target genes promoters, leading to their over activation in response to SHH signalling.

Whole genome analyses have identified mutations on MED12 with a low frequency in colorectal, breast and ovarian carcinomas (<1%). A more pronounced frequency was found in hormone-associated cancers, such as prostate cancer and adrenocortical carcinoma, with 5% of analysed tumours containing MED12 mutations (Barbieri et al. 2012, Assie et al. 2014) and up to 7% in uterine leiomyosarcomas and chronic lymphocytic leukemia (Kampjarvi et al. 2012, Wu et al. 2017).

Strikingly, mutations in exon 2 were found in 70% of the studied benign uterine leiomyomas (Makinen et al. 2011) and in almost 60% of benign breast fibroadenoma (Lim et al. 2014). Interestingly, all of the found mutations were in frame, with the vast majority consisting of missense mutations. This observation supports Med12 as an essential protein, since the lack of frameshift mutations is very likely due to the embryonic lethality caused by this type of mutations.

Additionally, Mediator has been implied in activation of ncRNAs target genes and, although it could be shown that Mediator and ncRNAs-a interact, the subunits involved in this interaction were not identified. FG associated mutations did not affect MED12 association with Mediator, however, through UV-RNA Immunoprecipitation (UV-RIP) data, a decreased interaction between Mediator and ncRNA-a was observed. Therefore, MED12 has been suggested as the main interface between the Mediator complex and ncRNA-a (Lai et al. 2013).

1.3. Non-coding RNAs

Despite the known existence of genes that generate functional RNAs without the need to be translated into proteins, such as the classical tRNAs and rRNAs, for several years most of the genome had been considered as "junk DNA" without any functional role. However, the completion of the Human Genome Project revealed that only 2% of the genome encode for proteins, revealing that the vast majority of the genome was not translated into proteins.

With the development of next generation sequencing and gene prediction algorithms, it was shown that more than 70% of the genome can be transcribed into RNA, with estimates that up to 93% of the genome can be transcribed (Djebali et al. 2012). In fact, the number of ncRNA genes exceeds the number of protein coding genes. However, despite the evidence that these genes are transcribed, for the vast majority their function is still unknown.

Genes described as ncRNAs can be classified by their length as small ncRNAs (sncRNAs) or long ncRNAs (lncRNAs), based on an arbitrary cut-off of 200 nt (Djebali et al. 2012). Micro RNAs (miRNAs) are the most studied class of sncRNAs. They originate from a pre-miRNA that is cleaved by Drosha and Dicer to form a mature miRNA of approximately 22 nt. They regulate gene expression post-transcriptionally, usually by recruiting the RNA-induced silencing complex (RISC) and, through base pairing, direct this complex to their target genes, resulting in their degradation (Ha et al. 2014). One miRNA can bind multiple targets either by perfect or imperfect base pairing. Conversely, the transcripts of one gene can be targeted by multiple miRNAs.

Other known classes of sncRNAs include small nuclear RNAs (snRNAs), which are part of the spliceosome (Yean et al. 2000), small nucleolar RNAs (snoRNAs), that target mainly rRNA for methylation or pseudouridylation (Kiss 2001) and piwi-interacting RNAs (piRNAs), which interact with PIWI proteins, guiding them to transposable elements locus in order induce a repressive state and prevent their expression (Le Thomas et al. 2013).

1.3.1. Long Non-coding RNAs

The most diverse class of ncRNAs are lncRNAs, with thousands of genes identified. Similarly to mRNAs, lncRNAs are transcribed by Pol II, most are 5' capped, polyadenylated and spliced. The epigenetics marks associated with coding genes (such as H3K4me3 enrichment at TSS and H3K36me3 along gene body of active genes) are also found in lncRNA loci, although usually less prominent. Despite being expressed at lower levels than mRNAs, lncRNAs show an even higher tissue specificity than coding genes (Cabili et al.

2011). This specificity makes lncRNAs suited to act as fine-tuners of individual genes in very specific developmental processes. Although usually tissue specific, spatial conservation is lower than mRNAs and changes in specificity between species are common. One such example is the lncRNA H19 X-linked (H19X), which in humans and mice is expressed mainly in placental tissues, while in the opossum its expression is restricted to the testis (Necsulea et al. 2014) .

As the cost of sequencing decreases, it becomes feasible to sequence more deeply, which allows the detection of lowly expressed lncRNAs. As such, over 16,000 human lncRNAs are already part of gene databases such as Gencode, with studies indicating the existence of over 60,000 lncRNAs among different human tumours, tissues and cells lines (Iyer et al. 2015). Despite the evolution of methods and algorithms that allow identifying new lncRNAs, their functional characterization as proven to be far more difficult. Therefore, their genomic location is often used as criteria to describe them. lncRNAs can originate from intergenic regions (termed long intergenic ncRNAS (lincRNAS)) or regions near known genes, from which they can be either exonic or intronic, convergent or divergent, in a sense or antisense orientation (Djebali et al. 2012).

A common method to characterize protein coding genes is identification of sequence conservation between different species. However, the majority of lncRNAs are poorly conversed even in close relatives. Nonetheless, there are lncRNAs that show a striking conservation, such as MALAT, one of the most conserved lncRNAs. This gene is conserved throughout the jawed vertebrates with functions in gene splicing and transcriptional regulation of a small subset of genes (Gutschner et al. 2013).

Another example of conservation is TUNA, a gene involved in normal brain development that shows outstanding high exonic conservation across vertebrates. It contains an element of almost 200 bp that shows conservation above 80% between zebrafish, mouse and humans, even exceeding conservation of most coding genes (Lin et al. 2014). In turn, HOTAIR and GAS5 are two lncRNAs with functions in repressing HoxD genes and other imprinted genes and in maintenance of mESC pluripotency network and induced pluripotent stem cells (iPSCs) reprogramming, respectively (Li et al. 2013, Tu et al. 2018).

Despite their conserved function, their sequence is not conserved. These observations demonstrate that sequence conservation is not sufficient to defined lncRNAs conservation and that other dimensions are necessary. One such dimension is secondary structure, with studies showing conserved secondary structures among characterized lncRNAs. However most methods for resolving lncRNAs secondary structures are base on *in silico* predictions, which are limited to regions that are already moderately conserved (Gorodkin et al. 2011). Additionally, random RNA sequences can also fold into stable structures, demonstrating that presence of a complex and stable secondary structure cannot be used as

evidence for a functional role (Schultes et al. 2005). A recent study, where a new statistical test for RNAs secondary structure was developed, concluded that there was no evidence of conservation for the proposed structures for several known lncRNAs such as X-Inactive Specific Transcript (XIST) or HOX Transcript Antisense RNA (HOTAIR) (Rivas et al. 2016).

Development of biochemical assays capable of elucidating the RNA complex folding pattern such as SHAPE-MaP (Selective 2'-hydroxyl acylation analysed by primer extension and mutational profiling) or icSHAPE (*in vivo* click selective 2'-hydroxyl acylation and profiling experiment) allow to confirm the predictions of the current *in silico* methods. These experimentally validated structures can then be used as training sets to further improve reliability of computed predictions (Smola et al. 2015, Spitale et al. 2015).

One of the criteria used to define the coding potential of a gene is the presence of a long open reading frame (ORF), with small ORFs (sORFs) being usually discarded as non-significant. However, in multiple lncRNAs sORFs were detected, leading to protein products shorter than 100 amino acids. Due to the presence of these sORFs, dozens of lncRNAs have been re-annotated as protein coding (Olexiouk et al. 2018). This information reveals that lack of long ORFs is not sufficient to discard any coding potential and further criteria must be used.

1.3.1.1. lncRNAs functions

Due to their size and flexibility, lncRNAs can fold into complex structures capable of interacting with proteins. Through base pairing, lncRNAs can additionally interact with DNA and other RNAs, allowing formation of complexes containing RNA, DNA and proteins. Due to their versatility, lncRNAs have been discovered to affect, either positively or negatively, gene expression at epigenetic, transcriptional, post-transcriptional and translational levels (Figure 5).

One of the first and most widely studied lncRNA is XIST. This lncRNA is the main organizer of X-chromosome inactivation (XCI), a crucial process through which mammalian females, who possess two X-chromosomes, can prevent double dosage of genes transcribed from this chromosome (Penny et al. 1996). This lncRNA coats the chromosome from which it is transcribed, acting is cis and through recruitment of different repressor complexes that induce a chromosome-wide inactivation. Interestingly, when Xist was ectopically expressed from autosomal chromosomes, a chromosome wide repression was observed. This demonstrated that Xist-mediated chromosome-wide repression is not restricted to the X-chromosome and that Xist expression was sufficient to induce repression in cis (Wutz et al. 2000).

A number of other lncRNAs play a role during XCI, mainly by modulating Xist expression, such as Jpx and Ftx, which are two activators of Xist. Jpx acts as an indirect activator by binding the Xist repressor CTCF, a factor with important functions in 3D genome structure regulation and enhancer insulation, and preventing its repressor activity (Sun et al. 2013). In turn, Ftx transcript does not act on Xist, but it is the transcription from its locus that leads to Xist expression in *cis* (Furlan et al. 2018). In order to ensure that Xist is only transcribed from the inactivated chromosome, Tsix, a lncRNA antisense of Xist, is expressed from the active chromosome and represses Xist in *cis* (Lee et al. 1999). XCI is an essential process that reveals the importance of lncRNAs and the different functions performed by these transcripts. Additional lncRNAs have been proposed to exert their functions through multiple mechanisms.

ecCEBPA (extra coding CCAAT enhancer binding protein alpha) is a non-polyadenylated lncRNAwhich is transcribed in the sense orientation of its coding neighbour CEBPA, encompassing the full CEPBPA mRNA (Di Ruscio et al. 2013). It has been proposed that this lncRNA prevents the binding of DNMT1 (DNA methyl transferase 1) to ecCEBPA locus, which in turn allows the expression of CEBPA by keeping its promoter free of methylation (Figure 5a). In fact, ecCEBPA interacts with DNMT1 and depletion of the lncRNA with short hairpin RNAs (shRNAs) led to an increased methylation at CEBPA promoter. This increase led to down-regulation CEBPA, while the opposite effect was observed upon ecCEBPA overexpression (Di Ruscio et al. 2013).

Enhancer RNAs (eRNAs) are a class of ncRNAs that can be further categorized in 2 classes: 1D, that are transcribed unidirectionally and are indistinguishable from other lncRNAs and 2D, shorter ncRNAs that are non-polyadenylated and non-spliced (Schmitz et al. 2016). Together with the already mentioned ncRNA-a, which have enhancer-like functions, eRNAs are described as acting in the formation of DNA loops (Figure 5b). As an example, depletion of either ncRNA-7 or ncRNA-3 led to a decreased DNA looping between their locus and the target promoters (Lai et al. 2013). However, another study concluded that depletion of Arc and Gadd45b associated eRNAs did not affect the DNA looping at these loci (Schaukowitch et al. 2014). These data suggest that not all eRNAS are involved in stabilization of DNA lopping. Due to the complex structures that lncRNAs are capable of folding into, they can act as scaffolds, allowing the assembly of different protein complexes. PRC2 is a repressive complex that deposits the repressive H3K27me3 mark and Lysine-specific histone demethylase 1A (LSD1) is a demethylase that mediates H3K4me2 demethylation. HOTAIR, a lincRNA involved in HOXD genes silencing, has been shown to act as a scaffold, bringing both PRC2 and LSD1 to the loci of target genes (Figure 5 d-e) (Tsai et al. 2010). RNase treatment abrogated the interaction between these two complexes. Additionally, HOTAIR knockdown led to a loss of occupancy of PRC2 and LSD1 from HOXD promoters and consequent

gain of H3K4me2 and loss of H3K27me3. Although the lncRNAs described so far were mostly nuclear, as further evidenced by their function, a large portion of lncRNAs are exported to the cytoplasm by mechanisms which remain unclear. Outside of the nucleus, lncRNAs can mediate post-transcriptional regulation, such as competing endogenous RNAs (ceRNAs). This class of transcripts competes with other RNAs, mainly mRNAs, for miRNAs binding sites (Figure 5f). One example is linc-MD1, a muscle specific lncRNA that sequesters mir133, preventing repression of MAML1 and MEF2C (Cesana et al. 2011). Another example is the cytoplasmic circular RNA (circRNA) CDR1as, which contains 63 binding sites of mir7. (Memczak et al. 2013) This circRNA acts as a sponge, removing mir7 from solution, which prevents degradation of the miRNA target mRNAS. Through splicing mediation lncRNAs also play a role in gene regulation, a mechanism observed for MIAT, a lncRNA which overexpression or depletion changed the ratio of Wnt7b isoforms 201 and 202 (Aprea et al. 2013). TINCR is a cytoplasmic lncRNA and one of the most upregulated lncRNAs during epidermal differentiation. Consistent with its cellular location, it binds several mRNAs containing a "TINCR box", leading to its target mRNAs' stabilization. On the other hand, the half-STAU1 lncRNA decreases stability of its target genes (Figure 5h). Despite being STAU1-mediated, TINCR and half-SRAU1 mechanism of action have opposite effects on their target mRNAS (Kretz et al. 2013, Wang et al. 2013).

Multiple lncRNAs have also been reported to have a role in different pathological settings. By far the highest association is with cancer. Analysis of thousands of tumor samples has revealed that despite the same number of protein coding genes and lncRNAS are found misregulated in different tumour samples, more than half the lncRNAs showed specificity for only one type of tumor (Yan et al. 2015). A large number of lncRNAs has been associated with cancer due to their misregulation in tumour compared to healthy tissues and due to their tissue specificity. However, for multiple lncRNAs, their role in cancers was analysed in more detail. The TF p53 is one of the best studied tumour suppressors that acts by activating genes responsible for cell cycle arrest and apoptosis upon oncogenic signalling (Real 2007). The lncRNA lincRNA activator of enhancer domain (LED) is activated by p53 and induces expression of p53 target genes. Accordingly, this lncRNA was found downregulated in different cancers (Leveille et al. 2015). The gene lincRNA-p21 is another lncRNA activated by p53. Depletion of this lncRNA leads to de-repression of Polycomb repressed genes, increasing the rate of cell proliferation (Dimitrova et al. 2014). High rates of survival are associated with high expression of this gene. While the mentioned lncRNAs act as tumour suppressors, others have been associated with proliferation of cancer. MALAT1 depletion in breast cancer cells led to a reduction of cell proliferation. Together with HuR, a RNA-binding protein with known roles in cancer progression, this lncRNA controls the dedifferentiation process of breast cancer

cells (Jadaliha et al. 2016, Latorre et al. 2016). High levels of MYC, a gene involved in tumorigenesis, are not sufficient to promote tumour development and the lncRNA PVT1 is also necessary. This was shown by depleting PVT1 in colorectal cancer cells, which failed to induce xenograft tumours (Tseng et al. 2014).

Although the majority of pathologies-associated lncRNAs has been linked to cancer, multiple others

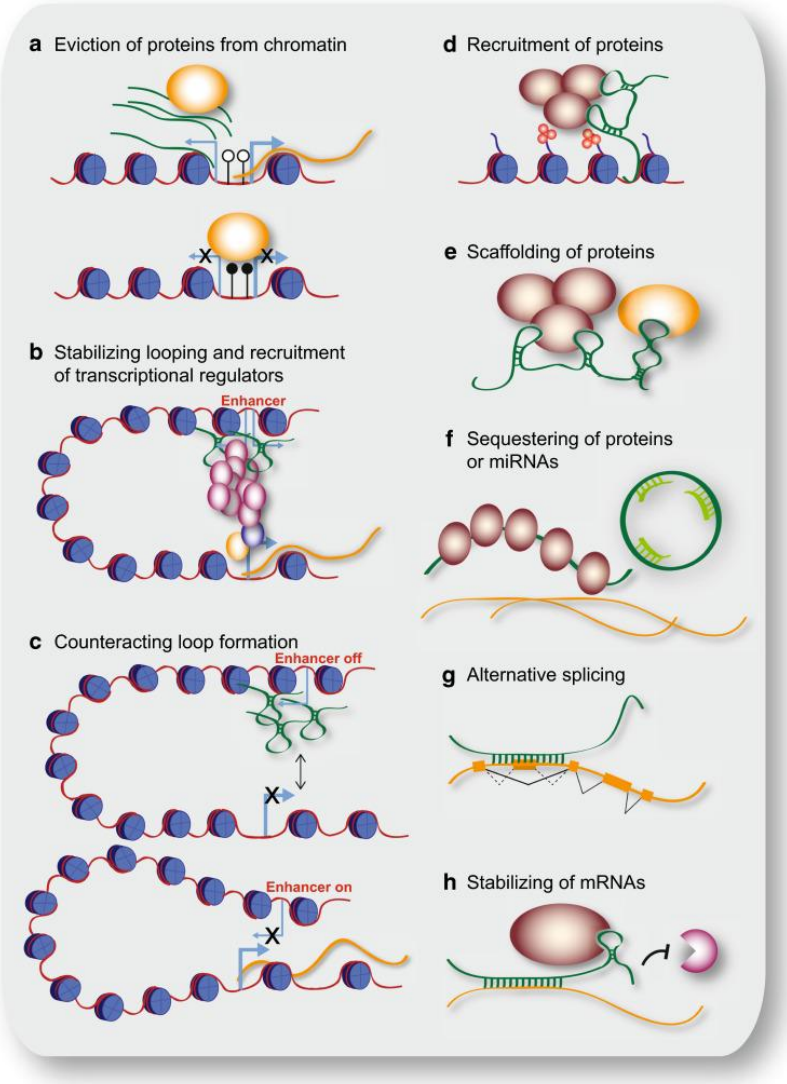


Figure 5 - lncRNA mechanisms of actions
These transcripts able to a) evict repressor proteins from target locus; b) stabilize loop formation, resulting in the target gene expression; c) repress formation of enhancer and promoter loop; d) guide proteins to target locations by interaction with both proteins and target DNA sequences; e) bind different protein complexes and acts as a scaffold, which allows a concert action of these complexes; f) bind miRNA or proteins, reducing their concentration and attenuating or preventing their function; g) bind a target mRNA and affect its splicing pattern; h) target mRNAs and through recruitment of proteins such as STAU1 , prevent mRNA degradation (Schmitz et al. 2016).

are associated with different diseases, such as autism. Autism spectrum disorder (ASD) is caused by diverse genetic factors. Two different studies, analysing patient and control cortices samples, revealed dozens of lncRNAs misregulated in disease samples, most of which were enriched in brain compared to other tissues. Additionally, a subset of these lncRNAs was co-expressed with protein coding genes with known roles in this disorder. Several of these lncRNAs were also in close proximity of genes with a characterized function in ASD (Parikshak et al. 2016, Gudenas et al. 2017).

Alzheimer's disease (AD) is caused by accumulation of extracellular plaques in brain tissues. These plaques are composed of amyloid beta (Aβ), which is formed upon cleavage of amyloid precursor protein (APP), a process performed by β-site APP cleaving enzyme 1 (BACE1) (Faghihi et al. 2008). Multiple lncRNAs have been implicated in AD. BACE1 antisense (BACE1-AS) transcript is a conserved lncRNA that induces BACE1 expression, enhancing Aβ production. It also binds BACE1 mRNA in order to increase its stability, further increasing BACE1 levels (Mulder et al. 2010, Dash et al. 2014, Liu et al. 2014). Neuroblastoma differentiation marker 29 (NDM29) is another lncRNA that also promotes BACE1 function and Aβ production (Massone et al. 2012). The lncRNA 51A is also associated with AD, with its overexpression resulting in increased Aβ formation.

LncRNAs have also been associated with other less common pathologies. Prader Willi Syndrome (PWS) is caused by a loss of genes activity from a 5-6 Mb from the paternally driven chromosome 15. This syndrome is caused by either deletion of this region from the chromosome inherited from the father (70%) or by inheriting two copies of chromosome 15 from the mother (25%) (Cassidy et al. 2012). From this region, five lncRNAs are expressed, part of a new class designated sno-lncRNAs, which include a snoRNA at their 3' and 5' end. These sno-lncRNAs contain multiple binding sites for Fox2, a gene that regulates the splicing pattern of a number of target genes. By binding Fox2, the sno-lncRNA titrate the coding gene levels, maintaining the proper splicing patterns that allow a normal development (Yin et al. 2012).

1.3.1.2. Methods for functional characterization of lncRNAs

Despite the diverse functions already attributed to a number of lncRNAs, their functional characterization is not a trivial undertaking. Depending on their genomic position but also gene structure, different methods routinely applied for characterization of protein coding genes may or not be suitable. As mentioned before, certain lncRNAs are transcribed and exert their function in *trans,* affecting their targets regardless of their genomic location. In turn, genes with function in *cis* act on their neighbouring

genes. Function can also be attributed to transcription of the and not to the transcript itself, which makes it possible that transcription of a lncRNA has a *cis* effect while the transcript has a different function in *trans*. Although some lncRNAs have a very specific expression pattern, this doesn't necessarily translate into important biological function (Oliver et al. 2015). Additionally, certain lncRNAs overlap protein coding genes, which complicate genomic manipulations without affecting the neighbour gene. All of these features make functionally characterization of lncRNAs a challenging undertaking. Nonetheless, different methods have been successfully used to identify lncRNA functions.

On lncRNAs distant from other genes or regulatory elements, direct deletion of the full gene can be a viable option to perturb its normal expression. Additionally, CRISPR-Cas9 mediated editing is faster and more precise than through classical methods such Talen or Zinc finger mediated edition. In a previous study, mutants for 18 lncRNAs were generated by replacing their locus with a lacZ reporter cassette, while keeping the regulatory region intact. This strategy allowed a simultaneous deletion of the lncRNA and assessment of its expression pattern *in vivo* (Sauvageau et al. 2013). However, this kind of approach might result in the disruption of local regulatory elements, especially in long deletions.

Another approach is the insertion of a stop cassette near the lncRNA transcription start site (TSS). This method reduces the risk of unintentional disruption of regulatory elements and has been successfully used characterize lncRNAs, such as Fendrr, an essential lncRNA for normal embryonic development (Grote et al. 2013). Although usually effective, there are reports of this approach resulting only in moderate downregulation of the targeted gene (Kraus et al. 2013). Overexpression of a lncRNA, which can be achieved by different methods, also allows studying its function. Injection of either viral vectors or *in vitro* transcribed lncRNAs, as well as insertion of BAC constructs that overexpress a specific lncRNA, have been used in functional studies (Ulitsky et al. 2011, Grote et al. 2013, Lv et al. 2018). However, these approaches are only suitable to study lncRNAs with effects in *trans*. Overexpression from the endogenous locus by insertion of a stronger promoter can be used to assess *cis* effects. Nevertheless, this method comes with the same risks as already described for direct genomic editing.

Methods that do not rely on genomic modification have also been used to study lncRNAs. Through RISC, siRNAs mediate target RNA degradation by base pairing, affecting both mRNA and lncRNAs. It's an effective method for lncRNA depletion and most of a transcript of interested can be targeted by the siRNAs. In turn, this method has shown a high risk of false positives due to off target effects, since specificity depends on a "seed" sequence of only 7-8 nt (Birmingham et al. 2006). Antisense oligonucleotides (ASOs) are a more reliable alternative to the use of siRNAs. These single stranded DNA

oligos of usually 15-20 nt are frequently chemically modified in order to increase efficiency or nuclease resistance (Meng et al. 2015).

Engineering of the Cas9 protein has led to the development of additional methods such as CRIPSR interference (CRISPRi). Mutation on both nuclease domains of Cas9 results in a dead Cas9 (dCas9), which can still target genomic regions, yet is not able to cleave DNA. Positioning of dCas9 near a gene TSS can prevent Pol II elongation and TFs binding by physical blockade. Additionally, fusion of dCas9 with repressor domains further contributes to gene repression. The Krüppel associated box (KRAB) domain has been successfully fused with dCas9 and results in deposition of the repressor mark H3K9me3. This method efficiency is restricted to a window of -50 to +300bp of the gene TSS, which limits the risk of affecting expression of neighbouring genes (Qi et al. 2013). The engineered dCas9 can also be fused to activator domains such as VP64, p65 or even multiple domains in tandem, which can induce an overexpression of the endogenous gene by several orders of magnitude (Chavez et al. 2016).

As mentioned before, different methods have different advantages and disadvantages, so care must be taken when selecting the most appropriate for the lncRNAs of interest.

1.4. High-throughput RNA sequencing

The thousands of described lncRNAs have been discovered by analysing transcriptome data obtained through high-throughput RNA sequencing (RNA-seq). Since first described over a decade ago (Emrich et al. 2007), RNA-seq has become a cornerstone of molecular biology, with its main application being genome-wide differential gene expression (DGE) analysis. Incredible technological advances have allowed sequencing of millions of fragments simultaneously, however, the workflow for DGE has not substantially changed. First, RNA is extracted, followed by rRNA depletion or poly-A RNA enrichment. cDNA is then synthesized and a sequencing library is generated by ligation of a specific adapter to the cDNA. The library is then sequenced and obtained reads aligned to the genome. Reads overlapping the transcripts are quantified and normalized between samples. Finally, statistical tests are performed in order to detect significant changes on gene expression between different groups.

Short read sequencing represents the most common method to generate RNA-seq data. Before cDNA generation, RNA is fragmented, resulting in a library with fragments of 150-400 nt. These are then sequenced from one end (single-end reads) or from both ends (paired-end reads). When using Illumina technology, sequencing of these fragments results in reads as small as 50 nt. With this approach, robust results are obtained, with high intra- and inter-platform correlation (Li et al. 2014). However, this

method shows limitations especially in assigning reads to a specific isoform in genes with multiple splice variants, which is critical in longer transcripts or highly variable isoforms (Djebali et al. 2012). Long-read sequencing overcomes these limitations by sequencing complete transcript sequences. With these methods, ambiguity of reds mapping is reduced, longer transcripts can be identified, leading to a more complete index of different transcripts and increases accuracy of splice-junction identification. The use of long-read sequencing has allowed discovering a number of novel transcripts and detecting full length homologs (Thomas et al. 2014). For both short and long read sequencing, RNA must be reversed transcribed to cDNA. However, it has been recently demonstrated that using Oxford Nanopore sequencing, it was possible to sequence RNA directly in a method termed direct RNA-seq (dRNA-seq), removing the bias generated by the reverse transcriptases and polymerases used on other methods (Garalde et al. 2018).

Long-read methods present clear advantages, such as allowing full transcript sequencing from the poly-A tail to the 5' cap and direct detection of isoforms with the need to re-construct them. However, these methods also display several limitations. They show a lower throughout (around three order of magnitude lower compared to short-read methods) which complicate the analysis of lowly expressed genes, as is the case for most lncRNAs. Additionally, long-read sequencing technologies produce reads with one or two orders of magnitude higher error rates (Weirather et al. 2017). Since longer reads are obtained, it is necessary to obtain intact RNA and cDNA from it. However, due to shearing from handling samples and due to RNA degradation, such full size transcripts are not always present.

The processing of samples and methods to generate sequencing libraries have also evolved throughout the years (Cartolano et al. 2016). RNA-seq was initially developed for analysis of polyadenylated transcripts. However, not only is this kind of approach biased for the 3' end of transcripts, it also does not allow to study the many transcripts without a poly-A tail, such as miRNAs and some eRNAs. In order to overcome this limitation, rRNA depletion removes this highly expressed class of ncRNAS (up to 95% off total RNA in a cell), allowing also to study all transcripts in a samples, including ncRNAs not polyadenylated (Morlan et al. 2012)T. he sequencing library can additionally be prepared in a stranded or unstranded fashion. Using stranded libraries, information about the strand of origin of the sequenced reads is retained. This strand information allows detecting antisense transcripts and correctly quantifying gene expression of overlapping genes (Zhao et al. 2015). How the reads are generated for each fragment can also be varied. The cDNA fragments can be sequenced from one or both of their ends, without any changes necessary to the library preparation method. For DGE analysis, single-end sequencing is enough to obtain robust results. However, paired-end reads help to disambiguate read

mappings and is preferred for alternative exons quantification and for *de novo* transcript assembly for identification of new genes (Alamancos et al. 2014).

Having sequenced the cDNA fragments it is necessary to use computational tools for the next steps of the analysis. The first step is to align the generated reads to a reference genome. Since the sequences are derived from RNA, they may span exons. Besides mapping, most used tools also perform spliced alignment by allowing gaps in reads (Dobin et al. 2013, Kim et al. 2015). The vast majority of reads will be equally mapped using any of the splice aligner available. However, for a subset of the reads, the available tools will handle them differently, an effect more noticeable for reads with multiple possible genomic locations. The most common practice is to exclude these reads from the analysis, which can cause bias in the results. Some methods consider all possible locations as valid or only the ones with higher score, discarding all others. In case of multiple alignments with the same score, some tools keep them all while others randomly select among them. Other tools calculate an estimate of uncertainty, which are used in the next steps of the analysis (McDermaid et al. 2018).

The mapped reads are attributed to transcripts in order to determine transcripts abundance. Due to the different approaches available to perform this task, the choice of used tool during this step is the one that has a great impact on the final results (Robert et al. 2015). Assigned reads need to be normalized due to differences in read depth in different samples or to technical variability (Risso et al. 2014) but also to transcript size , since longer transcripts will have more reads mapped to them even if there are expressed at a similar level as a shorter transcript. For this normalization step, most computational approaches rely on two assumptions: that most genes remain unchanged between samples and that mRNA level is the same in all samples. The final step consists on modelling differential expression in order to identify genes with a significantly altered expression between groups. Despite the number of tools available for this last step, the choice of which tool to use will have the least impact in the final results.

1.5. Aim of the project

Previous work has shown that Med12 is essential for canonical Wnt and Wnt/PCP pathways. Embryos generated from Med12 depleted mutant cells died during early gastrulation and revealed loss of expression of β-catenin target genes. Additionally, embryos generated from a hypomorfic mutant showed striking defects in neural tube closure, heart development, axis truncation and malformations of the branchial arches. Another study has revealed that MED12 depletion affected the expression of a subset of lncRNAs in human cells, linking Med12 to activation of expression of lncRNAs.

The aim of this project was to evaluate the impact of Med12 depletion on embryonic stem cells by analysing their transcriptome and to identify misregulated genes important for the processes disrupted in embryos generated with these cells. For this, previously generated transcriptome data for Med12 mutant cells was used to evaluate the expression changes caused by variations of Med12 expression levels. Furthermore, expression of ncRNAs was evaluated in the different mutant cells in order to identify lncRNAs whose expression was dependent of Med12. Besides annotated genes, putative novel genes were additionally assembled and studied. Finally, having identified lncRNAs that were dependent of Med12 for normal expression, gene structure, expression pattern and function were assessed for candidate lncRNAs.

2. Material and Methods

2.1. Constructs generation

2.1.1. CRISPR-Cas9 vector guide sequence cloning

To generate double strand breaks (DSB) at targeted locations in the genome of ES cells, the CRISPR-Cas9 method was used. This method relies on the Cas9 nuclease that generates DSB in regions complementary to the associated single guide RNA (sgRNA). By changing the sequence of the sgRNA, the Cas9 can be guided to the desired locus, where it will generate a DSB. In this study, the CRISPR-Cas9 vector px459, a generous gift from Dr. Dario Lupianez (Max-Delbrück-Centrum für Molekulare Medizin, Berlin), was used. Guide sequences were designed using Benchling online tool (www.benching.com) and selected for the ones with least change of off-target effects (off-target score < 1.0 for individual genomic locations). For each sequence a forward and reverse primer were ligated in order to generate a double strand guide sequence with overhangs compatible with px459 vector digested with BbsI (Supplementary Table 1). Briefly, for primer pair ligation, 1 µl of 100 µM of each primer, 1 µl 10x T4 ligase buffer (Promega) in a 10µl volume were heated up to 95°C for 5 min followed by a slow cooldown to 4°C (1% ramp decrease) in a Master Cycler Pro S (Eppendorf). 1 µl of 1:250 dilution of annealed primer pairs were ligated to 50ng of BbsI digested px459 vector using 2 µl 10x T4 ligase buffer and 1 µl T4 Ligase (Promega) in a total volume of 20 µl and incubated at RT for 30 min. Insertion was confirmed by Sanger sequencing.

2.1.2. β-galactosidase donor vector generation

In order to generate the vector donor for insertion of a β-galactosidase (β-gal) cassette into LN-BP18 locus, a double strand gene block was ordered (Integrated DNA Technologies) consisting of the guide sequence and respective protospacer adjacent motif (PAM) that the Cas9 targeted on LN-BP18, a 120 base pairs (bp) homology arm matching the sequence immediately before the induced double strand break (DSB), a spacer region with restriction sites for cassette insertion, another 120 bp homology arm matching the sequencing immediately after the induced DSB and again the guide sequence and the respective PAM. Flanking the whole construct with the same guide sequences used to target LN-BP18 allowed Cas9 to also target the donor vector and excise the donor sequence from vector backbone. This gene block was cloned into pBluescript SK vector (Stratagene). A β-gal mouse codon optimized cassette, a generous gift from Dr. Frederic Koch (Max Planck Institute for Molecular Genetics) and the 3x polyadenylation (pA) signal from pCCALL vector were amplified by polymerase chain reaction (PCR). The

resulting amplicons were digested with restriction enzymes which sites were included in the primers used for PCR amplification. Digested fragments were inserted into pBluescript-gene-block vector linearized with proper enzyme. The whole construct sequence was confirmed by Sanger sequencing.

2.2. Mouse strains and animal husbandry

The mouse strains used for this study were purchased from the Harlan Laboratories (Harlan Winkelmann GmbH, Borchen, Germany). The outbred strain CD1® was derived from the Swiss-Webster mice at the beginning of the 20th century and a stock established at the Institute for Cancer Research (ICR) in Philadelphia, PA (USA). Offspring from this strain were transferred to the Charles River Breeding Laboratories, in Willington, MA (USA), from which a breeding stock was obtained by the Harlan Sprague Dawley, Inc. (Hsd:ICR(CD-1®)). The mice were housed at specific pathogen free conditions at the animal facility of the Max Planck Institute for Molecular Genetics and kept under a 12h cycle of light and dark at 22°C and a relative humidity of 55 ± 10%. They were fed a pelleted, irradiated diet (ssniff M-Z®, Soest, Germany) composed of 22% raw protein, 4.5% raw fat, 3.9% raw fibre and 6.8% raw ashes. For timed mattings, day of plug was assumed to be E0.5. All animal experiments were approved by the Berlin State Office for Safety at Work, Health protection and Technical Safety (Landesamt für Gesundheit und Soziales, LAGeSo) and carried out in accordance with the German animal welfare act (Tierschutzgesetz, TSchG).

2.3. Whole mount in situ hybridization

Whole-mount in situ hybridization (WISH) was performed following the protocol described in the MAMEP website (http://mamep.molgen.mpg.de/index.php) with slight changes.

2.3.1. Fixation of mouse embryos

Timed-pregnant mice were euthanized via cervical dislocation. The uterus was removed and the embryos were dissected in cold PBS and fixed overnight at 4°C in 4% PFA/PBS solution on a roller mixer. The next day, the fixative was removed by two washes with cold PBS for 10 min and the embryos were dehydrated through a graded methanol series (25% MetOH/PBS, 50% MetOH/PBS, 75% MetOH/PBS, 1x for 10 min each, 100% MetOH 2x for 10 min). All steps were performed at 4°C on a roller mixer with pre-cooled solutions. The fixed embryos were stored in 100% MetOH at -20°C until used.

2.3.2. Preparation of labelled probes

All probes were prepared by PCR amplification following the protocol described in the MAMEP database. Probes templates for Cufflinks predicted genes were generated by PCR amplification using complementary DNA (cDNA) from mouse embryonic stem cells (mESC) and subcloned into pPCRII-TOPO (Invitrogen). Antisense in situ probes were generated as described on the MAMEP website using either T7 or SP6 RNA polymerase (Promega) and labelled with Digoxigenin11dUTP (DIG). Using an anti-DIG antibody coupled to alkaline phosphatase (AP), the location of the probes can be determined. Sall1 probe was generated from vector UH010 of the MAMEP database. Probes against predicted genes corresponded to the total transcript length isolated, described in Figure 9. Probe against Sall1 probe was complementary to nucleotides 3741-4418 of Sall1 mRNA.

2.3.3. Processing of mouse embryos

The desired number of fixed embryos of the same stage was pooled into 2 ml microtubes (Sarstedt). If not stated otherwise, all subsequent steps comprised a 10 min incubation at 4°C on a roller mixer. The embryos were rehydrated (1x 75% MetOH/PBST, 1x 50% MetOH/PBST, 1x 25% MetOH/PBST, 2x PBST), bleached in a 6% H_2O_2/PBST solution (30 min for E10.5, 45 min for E11.5) and washed 3x in PBST. The specimens were digested with 10 µg/ml proteinase K/PBST (13 min for E10.5, 17 min for E11.5) in order to allow for better penetration of the labelled probe. The digestion process was stopped by incubation with 2 mg/ml glycine/PBST solution followed by two washes with PBST.

The embryos were refixed in 0.2% glutaraldehyde/4% PFA/PBST for 30 min at room temperature (RT) while rolling, washed twice with PBST at RT and preincubated with hybridization solution (50% formamide, 5x SSC pH 5.0, 1% SDS, 0.05 µg/ml yeast RNA (Sigma-Aldrich), 0.05 µg/ml heparin (Sigma-Aldrich), in RNase-free H2O) for 15 min at RT.

Embryos were sorted according to the probe used and prehybridized for 2h at 68°C to reduce unspecific background staining. 2 ml of preheated hybridization solution (68°C) was added to the embryos together with 10 µl of denatured probe (5min at 95°C) per ml of hybridization solution and incubated overnight at 68°C in an oven with rocking function (Hybaid Shake'n'Stack; Thermo Fisher Scientific).

2.3.4. Antibody incubation

Hybridization solution was discarded and embryos washed twice at 68°C for 30 min with preheated Solution 1 (50% formamide, 5x SSC pH 5.0, 1% SDS) followed by two washes for 30 min at 68°C with Solution 3T (50% formamide, 2× SSC pH 5.0, 0.1% Tween-20). Embryos were then incubated for 20 min at RT with 50% solution 3T: 50% MABT pre-heated to 68°C, followed by a 15 min wash in MABT + 2mM Levamisole. Embryos were incubated in blocking solution (2% blocking reagent (Roche), 20% Serum (Gibco), 2mM levamisole, in MABT) for 1h at RT. 0.5 µl of anti-DIG antibody (Roche) was added and embryos incubated overnight at 4°C.

The next day, the embryos were washed twice for 15 min, twice for 30 min and at least six times for 1h at RT in the dark in MABT. To reduce background staining, the specimens were incubated overnight at 4°C in MABT in the dark. .

2.3.5. Staining

The next day, the embryos were washed 2 times for 15 min and 1 time for 40 min at RT with freshly prepared NTMT (100 mM Tris-HCl pH 9.5, 100 mM NaCl, 50 mM MgCl2, 0.1% Tween-20, 1 µM levamisole). Embryos were then stained with BM-Purple (Roche) and incubated at RT, rocking in the dark. The staining intensity was monitored periodically under a binocular. Once an appropriate staining was obtained, the reaction was stopped by washing the embryos once in NTT (100 mM Tris-HCl pH 9.5, 100 mM NaCl, 0.1% Tween-20) and several times in PBST at RT. The stained embryos were postfixed in 4% PFA/PBST and stored in the dark at 4°C.

2.3.6. Imaging

Whole-mount specimens were photographed with a SteREO Discovery.V12 microscope (Zeiss) and an AxioCam Color camera (Zeiss) using the AxioVision 4.6 software (Zeiss).

2.4. ES cell culture

2.4.1. Culture procedure

All the ESCs used in this study were male. They were derived from JM8A1.N3 with a C57BL/6N genetic background (Pettitt et al. 2009) or from a G4 hybrid line (129S6/C57BL6) (George et al. 2007). The procedures were performed under sterile conditions in a laminar flow hood (HERAsafe; Heraeus). ES cells

were seeded onto a monolayer of mitotically inactivated primary embryonic fibroblasts, i.e., feeder cells, in a gelatine-coated 6 cm cell culture dish (Corning) and incubated at 37°C in a humidified 7.5% CO2 incubator (HERAcell 150; Heraeus). The cells were grown in ES cell medium composed of Dulbecco's Modified Eagle's Medium (DMEM containing 4,500 mg/ml glucose, without sodium pyruvate; Sigma-Aldrich), 15% (v/v) ES cell-qualified, heat-inactivated fetal bovine serum (FBS; Gibco), 2 mM L-glutamine (Sigma-Aldrich), 50 U/ml penicillin (Sigma-Aldrich), 50 µg/ml streptomycin (Sigma-Aldrich), 1% 100× non-essential amino acids (Sigma-Aldrich), 0.1 mM β-mercaptoethanol (Sigma-Aldrich), 1% 100x nucleosides (Sigma-Aldrich). 1000 U/ml murine leukemia inhibitory factor (LIF; Chemicon) were added to keep the ES cells in an undifferentiated state. The medium was exchanged daily for 3 days or until plate was 90% confluent. At this point cells were splitted by trypsinization. Before trypsinization, the ES cells were grown in fresh medium for at least 2 h. The medium was aspirated and the cells were carefully washed twice with cell-culture grade D-PBS (Lonza). 1 ml trypsin/EDTA solution (Gibco) was added and the cells were incubated at 37°C for 10 min in order to disrupt cell-cell contacts. The enzyme was inactivated by the addition of 2 ml ES cell medium before pipetting vigorously up and down to produce a single cell suspension. To determine cell density, 10 µl of cell culture and 10µl of 0.4% Trypan Blue Stain were mixed then 10 µl were loaded into a LUNA Cell Counting Slide and viable cells counted in LUNA Automated Cell Counter (Logos Biosystems). Cells were either platted in a fresh gelatinized plate with a monolayer of feeders or aliquots were prepared by freezing cells in ES medium with 20% FCS and 10% DMSO.

2.4.2. ES cells transformation

In order to generate the different mutant ESCs, $3.0x10^5$ ESC were seeded onto with a monolayer of $4.0x10^5$ feeder cells in a gelatine-coated 6well plate (Corning) and incubated for 24h at 37°C. To transform ESC, Lipofectamine 2000 (Invitrogen) was used according to the manufacturer's instructions. Briefly 8 µg of DNA were mixed with 125ul OptiMEM and incubated with 125 µl of a 1:4 lipofectamine/OptiMEM solution for 15 min at RT. Afterwards, the combined solution was added to the cells, acting for 5 hours at 37°C.

Cells were then splitted as described below into 4x6 cm culture dishes (Corning) containing puromycin resistant feeders in the following ratios 1/12, 2/12, 4/12 and 5/12. Selection with 2 µg/ml of puromycin (invivoGen) started the following day for 48h, followed by 24h using 1 µg/ml of puromycin and then approximately 96h without selection, until colonies were visible. Throughout the whole procedure ES medium was exchanged daily. For Sall1, LN-BP18 TSS1 and TSS2 excision 4 µg of each

CRISPR-Cas9 vector was used, while for knock-in of the β-gal-3xpA cassette into LN-BP18, 2 µg of CRISPR-Cas9 vector and 6 µg of donor vector were used.

2.4.3. Colonies picking

Fresh ES cell medium was added to the cells 3-4h prior to picking. The cells were washed twice with D-PBS before covering the colonies with a layer of fresh D-PBS. Individual colonies were picked using disposable 10 µl pipette tips under a stereo microscope (MZ8; Leica) and transferred to the wells of a round-bottomed 96-well plate (Corning) containing 50 µl cold trypsin/EDTA solution. After all colonies had been picked, the 96-well plate was placed in the 37°C incubator for 10 min. 100 µl ES cell medium were added per well to inactivate the trypsin. The colonies were disaggregated with a multi-channel pipette and transferred to the wells of a gelatinized, flat-bottomed 96-well plate (Corning) containing a monolayer of feeder cells ($1x10^6$ feeder cells/plate) and grown in regular ES cell medium.

2.4.4. Splitting and freezing

After the cells had been grown 2-3 days, they were washed twice with D-PBS before incubating them at 37°C with 50 µl trypsin/EDTA for 10 min. The trypsinization was stopped by adding 100 µl bicarbonate-free DMEM (Sigma-Aldrich) supplemented with 10 mM HEPES (Sigma-Aldrich) and 20% FBS (v/v). The cells in the so-called 'DNA Original Plate' were disaggregated and 50 µl of the 150 µl were transferred to the wells of a round-bottomed 96-well plate containing 50 µl 2x concentrated ES cell freezing medium (bicarbonate-free DMEM (Sigma-Aldrich), 10 mM HEPES, 20% FBS, 20% DMSO). The contents of this so-called 'Master Plate' were mixed well by pipetting. Another 50 µl of the remaining 100 µl cell suspension were transferred to the wells of a gelatinized, flat-bottomed 96-well plate (Corning) containing 200 µl ES cell medium ('DNA Replica Plate') and the contents were again mixed. The 'Master Plate' was sealed, placed inside a styrofoam box and frozen at -80°C. 200 µl ES cell medium was added to the remaining 50 µl cell suspension in the 'DNA Original Plate' and the contents were mixed.

The cells in the 'DNA Original Plate' and the 'DNA Replica Plate' were grown to confluency at 37°C. DNA was isolated and subsequently used for analysis by PCR.

2.4.5. Screening of ES clones by PCR

In order to identify clones with successful transformation, a PCR screen was used. The procedure followed the protocol previously described (Ramirez-Solis et al., 1993). ES cells were grown to confluency in the wells of a 96-well plate (in the so-called 'DNA Original Plate' and the 'DNA Replica Plate' (2.4.4) in a humidified incubator at 37°C and 7.5% CO_2. The cells were carefully washed twice with D-PBS and 50 µl prewarmed lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM EDTA pH 8.0, 10 mM NaCl, 0.5% sarcosyl) containing 1 mg/ml proteinase K (Roche)) were added per well. The plate was placed inside a humidified chamber and incubated overnight at 60°C. The next day, 100 µl ice-cold 75 mM NaCl/100% EtOH was added without mixing. The plate was allowed to stand on the bench for 30 min to precipitate the DNA as a filamentous network on the bottom of the wells. The plate was then carefully inverted to discard the solution and excess liquid was blotted on a paper towel. The wells were rinsed three times by adding 200 µl 70% EtOH. After the final wash, the precipitated DNA was allowed to dry on the bench. The 'DNA Replica Plate' was sealed and stored at -20°C. The 'DNA Original Plate' was used for PCR screen after resuspending DNA in 50 µl 1xTE buffer. For PCR, PrimeSTAR HS DNA Polymerase (Takara) was used for primer pairs 3 and 4, while for the remaining pairs Taq DNA polymerase (Invitrogen) was used (Supplementary Table 2). Amplification mixture was prepared as described in Supplementary Table 3 and the run followed the conditions in Supplementary Table 4. For primer pairs 3 and 4, since a different polymerase was used, the denaturation temperature was changed to 98°C. the PCR reactions were then loaded on a 2% agarose gel (Biozym) using 1:30000 SYBR Safe DNA gel staining (Invitrogen) to visualize the double strand DNA using a Gel Doc XR+ (BioRad).

2.4.6. Screening of ES clones by Southern blot

PCR screen of JM8A1.N3 (JM8) ESC clones that had successfully integrated the β-gal cassette into the LN-BP18 locus did not provide clear results. As such, Southern blot analysis was performed. The procedure followed the protocol described in (Ramirez-Solis et al., 1993). ES cells were grown to confluency in the wells of a 96-well plate (in the so-called "DNA Original Plate" and the "DNA Replica Plate" (section 2.4.4) in a humidified incubator at 37°C and 7.5% CO2.

The cells were carefully washed twice with D-PBS and 50 µl prewarmed lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM EDTA pH 8.0, 10 mM NaCl, 0.5% sarcosyl) containing 1 mg/ml proteinase K (Roche) were added per well. The plate was placed inside a humidified chamber and incubated o/n at 60°C. The next day, 100 µl ice-cold 75 mM NaCl/100% EtOH were added without mixing. The plate was allowed to stand

on the bench for 30 min to precipitate the DNA as a filamentous network on the bottom of the wells. The plate was then carefully inverted to discard the solution and excess liquid was blotted on a paper towel. The wells were rinsed 3 times by addition of 200 µl 70% EtOH. After the final wash, the precipitated DNA was allowed to dry on the bench. The 'DNA Replica Plate' was sealed and stored at -20oC. The "DNA Original Plate" was used for restriction enzyme digestion.

2.4.6.1. DNA digestion and electrophoresis

A restriction digest mix was prepared containing 1 U/µl BamHI (Promega), for identification of correct integration of β-gal, plus 1xBuffer E (Promega), 1 mM spermidine (Promega), 100 µg/ml BSA (Promega), 100 µg/ml RNase A (Promega). The mix was prewarmed to 37°C and 30 µl were added to each well without mixing. The reaction was incubated at 37°C for 2-4 h in a humidified chamber before mixing the content of the wells. The 37°C incubation was continued o/n in a humidified chamber. The next day, 6 µl of 6x gel loading buffer were added to each well and the DNA was electrophoretically separated in a 0.7% agarose gel in 1x TBE electrophoresis buffer for 3 h at 80 V. The next day, the gel was documented with the Gel Doc XR+ (BioRad).

2.4.6.2. DNA bloting

After electrophoretic size separation, the gel was pretreated in order to facilitate transfer of large DNA fragments. First, the DNA was partially depurinated by soaking the gel twice in 0.25 N HCl for 10 min. The gel was washed in ddH_2O for 5 min before denaturing the DNA by placing the gel in a bath of 0.5 N NaOH on a moving platform for 40 min at RT. The DNA was then blotted o/n onto a preequilibrated nylon Zeta-Probe GT membrane (Bio-Rad) by capillary transfer. The DNA was UV-crosslinked to the membrane using 5000 µJ/cm^2 radiation (Stratalinker 2400; Stratagene). The membrane was washed twice in 2xSSC pH 7.0 and directly hybridized.

2.4.6.3. Probe labelling

For the generation of the radioactively labelled probe, a fragment of 550 bp, downstream of exon 4 was amplified by PCR from WT ESCs using the primers 5'-CTGACCACGATCCCCTCATC and 3'GGCAGGACAATCAGCCATCT. 25 ng of purified PCR product were diluted in 12 µl of ddH_2O and denatured for 5 min at 95°C. After the sample was cooled down on ice for 2 min, dATP, dGTP, dTTP, Klenow enzyme and random primers were added (Rediprime II Random Prime Labelling System;

Amersham) and the components mixed. All following procedures were performed in an isotope laboratory facility according to the manufacturer's instructions. 5 µl of 50 µCi (α-^{32}P) dCTP (Redivue; Amersham) were added and the mixture was incubated for 30 min at 37°C to allow for the labeling reaction catalysed by the Klenow fragment of the DNA polymerase I. The radioactive sample was pipetted onto a Sephadex G-50 MicroSpin Column (GE) to separate the labelled probe from unincorporated radioactive nucleotides, according to manufacturers' instructions. The labelled DNA probe was denatured for 5 min at 95°C, cooled down on ice for 5min and added to the hybridization buffer (see below).

2.4.6.4. Hybridization and detection

The DNA-blotted membrane was prehybridized in a glass bottle with 20 ml prewarmed ExpressHyb Hybridization Solution (Clontech) for 30 min in a hybridization oven (Hybaid Shake'n'Stack; Thermo Scientific) at 68°C with constant rotation. After prehybridization, the denatured, labelled probe was added to the solution and the membrane was hybridized o/n at 60°C with constant rotation. The next day, the membrane was washed in two times for 20 min in 2x SSC 0.1% SDS at 68°C. The membrane was them sealed inside a plastic bag. The radioactively labelled DNA was exposed in a phosphoimager (Amersham) for 24h before detection. In case of lower signal, exposure on extended.

2.4.7. Depletion of feeder cells from ESC culture

Before RNA and protein extraction or *in vitro* differentiation of mutant ESC, feeder cells used to support ESC growth were removed in order to prevent contamination of results since the feeder cells used were derived from WT mouse embryonic fibroblasts. As such, ESC were cultured in a 6 cm plate following procedures described in section 2.4.1. After reaching confluency, feeder cells were remove with feeder removal microbeads (Macs) following manufacturers' instruction. Briefly, a single cell suspension was obtained. The cell suspension was then filtered using a 30 µm filter (Sysmex) and cells were resuspended in 80 µl of cold ES medium. 20 µl of beads suspension was mixed with cells and incubated rocking at 4°C for 15min. Afterwards, 400 µl of cold ES medium was added to cells and the whole cell suspension loaded into equilibrated LS columns (Macs). Cells were eluted by washing columns with ES medium and used directly.

2.5. *In vitro* differentiation of ESCs into mesoderm

Since LN-BP18 was lowly expressed in ESCs and was found to also be expressed in the caudal end of E8.5 mice embryos, ESCs were *in vitro* differentiated into mesoderm transcriptionally similar to the one in E8.5 caudal end (Koch et al. 2017). The differentiation protocol was performed as described previously with slight modifications (Gouti et al. 2014). The procedures were performed under sterile conditions in a laminar flow hood. Briefly, to 12 well plates, 0.8 ml of 0.025 mg/ml Synthemax II-SC (Corning) were added per well and incubated at RT for 2h, after which they were aspirate. $2.0x10^5$ ESCs were plated into each well.

Cells were incubated for 48h cells with N2B27 medium (DMEM/F12 medium (Gibco) and neurobasal medium (Gibco) (1:1), 1x N2 (Thermo Fisher Scientific), 1x B27 (Thermo Fisher Scientific), 2 mM L-glutamine, 40 µg/ml BSA, 0.1 mM 2-mercaptoethanol, 50 U/ml penicillin, 50 µg/ml streptomycin) supplemented with 10ng/ml bFGF (Peprotech), 24h with N2B27 medium supplemented with 10ng/ml bFGF and 5 µM CHIR99021 (CHI, Calbiochem) and 48h with N2B27 medium with 5 µM Chiron. Medium was changed daily and cells incubated at 37°C, 7.5% CO_2.

For every time point, duplicate samples were prepared in order to extract RNA and perform X-gal staining. RNA was extracted as described in section 2.7.1.

2.5.1. X-Gal staining for assessing β-gal reporter cassette activity

In order to check β-gal reporter cassette activity in generated mutants, X-Gal staining was done in ESCs, *in vitro* differentiated cells and in embryos. Cells were washed twice with cell-culture grade D-PBS and fixed with cold 4% PFA/PBS at RT for 5 min. Embryos were dissected in PBS and fixed in 4%PFA/PBS at 4 °C for one hour. Samples were then rinsed three times at RT in rinse buffer (5 mM EGTA, 0.01% deoxycholate, 0.02% NP40, 2mM MgCl2, in PBS). Meanwhile, the staining buffer was prepared containing 5 mM potassium ferricyanide (Merck) and 5 mM potassium ferrocyanide (Merck), in rinse buffer. To this X-gal (40 mg/ml in dimethylformamide) was then added to a final concentration of 1 mg/ml and the prepared staining buffer was then filtered. Staining was done o/n at 37 °C in the dark. After staining, samples were washed 3 times with PBS and refixed in 4%PFA/PBS. After one final wash with PBS, samples were imaged with a SteREO Discovery.V12 microscope and an AxioCam Color camera using the AxioVision 4.6 software.

2.6. Nuclear and cytoplasmic fractionation

To determine LN-BP18 cellular localization nuclear and cytoplasmic fractions for ESCs were prepared. $1x10^6$ WT ESCs were lysed in Low salt buffer (50 mM TRIS-HCl, 150 mM NaCl, 0.4% Triton X-100, 5% glycerol, 1x cOmplete (Roche) and 1 U/µl RNasin (Promega)) and mixed for 10min at 4°C. Lysate was centrifuged and the supernatant constituting the cytoplasmic fraction collected. For RNA extraction, RNeasy micro kit was used on the nuclear pellet and on cytoplasmic solution as described in section 2.7.1. For protein, the extraction nuclear pellet was lysed with 150 µl western lysis buffer (1% SDS, 10 mM TRIS-HCl, 2 mM EDTA) and 50 µl of 4x NuPAGE LDS buffer (Thermo Fisher Scientific) was added to nuclear lysate and to 150µl of cytoplasmic fraction.

2.7. Real-time quantitative PCR analysis

2.7.1. RNA extraction

Total RNA was isolated from embryonic tissues or cells using the RNeasy Micro Kit (Qiagen) according to the manufacturers' instructions with slight modifications. Briefly, cells or tissues were lysed in 1 vol. RLT Buffer, 1 vol. 70% EtOH was added to the lysate and the contents mixed by pipetting. The sample was transferred to an RNeasy column placed in a collection tube and centrifuged and the flow-through discarded. The column was washed with RW1 Buffer. An on-column DNase digestion step was performed, supplemented with an additional 1µl of DNase I (Roche) and incubated at table top 30 min. The column was washed with, RW1 Buffer, RPE Buffer and with 80% EtOH. The RNA was eluted with RNase-free water.

The RNA concentration was measured by UV spectrophotometry (ND-1000 Nanodrop)and the samples were either used directly or stored at -80°C.

2.7.2. Reverse transcription of RNA

The SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen) was used for reverse transcription of RNA into cDNA. Briefly, 1-5 µg RNA were diluted in a total volume of 8 µl DECP-H2O. 1 µl of either random hexamers, oligo dT or of a gene specific primer and 1 µl dNTP mix (10 mM) were added per reaction and the samples were incubated at 65°C for 5 min before placing them on ice for 1 min. A reaction mixture was prepared containing 2 µl 10× RT Buffer, 4 µl MgCl2 (25 mM), 2 µl 0.1 M DTT, 1 µl

SuperScript III and 1 µl RNaseOUT per sample. The reaction mixture was added to the RNA and incubated at 25°C for 10 min, at 50°C for 50 min and at 85°C for 5 min.

The reactions were cooled on ice before adding 1 µl RNase H to each tube and incubating them for 20 min at 37°C. The cDNA samples were used directly or stored at -20°C.

2.7.3. Quantitative real-time PCR

Quantitative real-time PCR (qPCR) was performed with the above yielded cDNA using GoTaq qPCR master mix (Promega). This mix contains polymerase, dNTPs, buffer and SYBR Green I Dye. For each reaction, a mixture of 10 µl GoTaq qPCR master mix, 5.0 µl of primer mix (1µM), 4.8 µl DECP-H_2O and 0.2 µl cDNA were transferred to the wells of a MicroAmp Fast Optical 96-Well Reaction Plate (Applied Biosystems).

Each sample was run in technical triplicates in a StepOne Real-Time PCR System according to the program listed in Supplementary Table 5. The results were analysed using the StepOnePlus Software. The housekeeping gene Pmm2 was used as internal control for normalization of each cDNA sample and expression normalized to a control sample using the ΔΔCt method. For data derived from embryonic tissue, no internal control was used and samples were normalized to the sample “rest” by calculating the ΔCt.

2.8. Isolation of putative new lncRNAs

To verify putative lncRNAs expression in ESCs, RNA was extracted from wild type (WT) ESCs as mentioned in section 1.8.1. The cDNA was generated according to the protocol on section 2.7.2 ,using 5 µg of total RNA and random hexamers. Obtained cDNA was used directly as a template for PCR with a combination of different primers along the identified genes. For PCR, Taq DNA polymerase was used with the conditions in Supplementary Table 3 and Supplementary Table 4, using an elongation step of 2 minutes instead of 30 seconds. PCR was then loaded on a 2% agarose gel using 1:20000 SYBR Safe DNA gel staining to visualize the double strand DNA. Amplified fragments were extracted from the gel using a MinElute PCR Purification Kit (Qiagen) following manufacturers’ instructions. Briefly, 1 vol. of agarose containing the PCR amplicons was dissolved in 3 vol. of QG buffer. Dissolved agarose was then loaded into a Minelute column and washed with QG and PE buffer. Finally it was eluted using 25 µl of EB buffer. The eluted amplicons were cloned into pCR2.1-TOPO vector using TOPO TA cloning kit (Invitrogen)

according to the manufacturer's instruction, using 4 µl of DNA. Insertion was confirmed by Sanger sequencing.

2.9. Rapid Amplification of cDNA Ends (RACE)

To determine LN-BP18 gene structure, its transcription start and end sites (TSS and TES, respectively) were identified using 5´ RACE System for Rapid Amplification of cDNA Ends and 3´ RACE System for Rapid Amplification of cDNA Ends (Thermo Fisher Scientific), respectively, following the manufacturer's instructions and using the primers in Supplementary Table 6. Briefly, 5 µg of total RNA were used for cDNA generation using a gene specific primer (GSP) in the 5' RACE (primers 1, 4 or 7) or oligo dT for 3' RACE. With the generated cDNA, LN-BP18 transcripts were amplified using a GSP (primer 2, 5, 7 or 9) and an adapter oligo supplied with the kit. Then a nested PCR was performed using an adjacent GSP (primers 3, 6, 8 or 10) and the same adapter. Obtained amplicons were extracted, cloned and sequenced as described in section 2.8. Primers 1, 2 and 3 were used to identify TSS1.

Primers 4, 5 and 6 were used to confirm TSS1. Primers 7 and 8 were used to identify TSS2 and primers 9 and 10 for determining the TES of LN-BP18 (Supplementary Table 8 - Primers used during RACE protocol).

2.10. LN-BP18 Isoforms identification

LN-BP18 initial isolation revealed the presence of two different isoforms. In order to identify additional isoforms, RNA was extracted from WT and Med12^{null} ESCs, as well as from forelimbs of E11.5 WT embryos as mentioned in section 2.7.1. cDNA was generated according to the protocol on section 2.7.2 ,using 5 µg of total RNA and a GSP located at the identified TES, random hexamers or oligo dT. PCR amplification of LN-BP18 isoforms as described in Supplementary Table 3 and Supplementary Table 4, using different annealing temperature and with an elongation step of 2 min. However, since the obtained signal was weak and unspecific fragments were also amplified, the obtained solution was used as a template for a nested PCR. For this additional PCR, 2.5 µl of a 1:100 dilution of the original PCR reaction was used as a template, using primers adjacent to the ones used in the first round of amplification. Amplicons were extracted, cloned and sequenced as described in section 2.8.

2.11. Western Blot

Proteins were extracted by lysing cells in western lysis buffer (1% SDS, 10 mM TRIS-HCl, 2 mM EDTA). After incubation on ice for 10 min, benzonase (Sigma-Aldrich) was added and samples incubate at 37°C for 30 min in order to degrade DNA and RNA. Protein concentration was calculated using BCAProtein Assay Kit (Thermo Fisher Scientific). To 3 volumes of volumes of protein lysate, 1 volume of 4x NuPAGE LDS buffer (Thermo Fisher Scientific) and dithiothreitol to a final concentration of 50mM were added. Samples were at 95°C and cooled down to RT before being loaded into NuPAGE 4-12% Bis-Tris gel (Thermo Fisher Scientific) and ran in MOPS buffer (Thermo Fisher Scientific).

After the run proteins were transferred to a PVDF membrane (BioRad) pre-activated with methanol using a BioRad Minigel Blotting System in transfer buffer (25mM Tris base, 192 mM glycine, 10% methanol) at 4°C using 100 V, 400 mA for 1 hour. The membrane was then blocked in 5% skim milk (Sigma-Aldrich) in TBS-T (10mM Tris pH 8, 150mM NaCl, 0.01% Tween-20) at RT, cut into appropriated pieces and incubated with primary antibody diluted in 1% milk in TBS-T for 1h at RT. After 3 washes of 5min in TBS-T, membranes were incubated 1h at RT with secondary antibodies coupled to HRP diluted 1:5000 in 1% milk in TBS-T. Membranes were washed three times in TBS-T and two times in TBS for 5min each after which the signal was detected with Amersham ECL substrate (Health Care RPN3243) in a Fusion SL Advance (Vilber Lourmat).

2.12. RNA-seq analysis

2.12.1. Libraries preparation

RNA was extracted from 2 biological replicates of WT, $Med12^{flox}$, $Med12^{hypo}$ and $Med12^{null}$ mESC using RNeasy Micro kits as mentioned in section 2.7.1. The RNA was quantified using the Qubit RNA HS Assay (Life Technologies) and the integrity was verified using Bioanalyser RNA Pico chips (Agilent) using a Bioanalyser (Agilent). rRNA was depleted of 500 ng total RNA using Ribo-Zero Magnetic Kit (Ilumina) according to manufacturer's instructions with small changes. 90 µl of magnetic beads were used for each sample and washed in 90 µl of H_2O followed by resuspension on 35 µl of resuspension buffer and addition of 0.5 µl RiboGuard RNase Inhibitor. 2 µl of Ribo-Zero reaction buffer and 4 µl Ribo-Zero rRNA Removal solution were added to 16 µl of the RNA suspension. rRNA depleted RNA was cleaned using RNeasy Micro columns (Quiagen) and used to generate strand-specific RNA-seq libraries with ScriptSeq V2 RNA-Seq Library Preparation Kit (Ilumina) according to manufacturer's instructions. The cDNA was purified with MinElute PCR Purification Kit (Qiagen) and indexed using ScriptSeq Index PCR Primer

(Ilumina) according to manufacturer's instructions for 15 cycles of amplification. Libraries were then purified with AMPure XP beads (Beckman Coulter). Quantification of all RNA-seq libraries was performed using the Qubit high sensitivity DNA assay (Life Technologies) and the size distribution was verified using the DNA HS Bioanalyser chips (Agilent). Libraries were pulled together and sequenced in a single lane of the HiSeq 2500 (Ilumina). Approximately 45 million pair end reads of length 75 bp were obtained for each of sample

2.12.2. Bioinformatics analysis

For differential gene expression analysis performed in section 3.1, reads quality was evaluated with FastQC and mapped against mouse genome (mm10) with Hisat2 (Kim et al. 2015) with the options *-dta-cufflinks --no-discordant --no-mixed*. Resulting .sam files were converted to .bam using samtools and supplied to Cufflinks (Trapnell et al. 2012). Cufflinks was used with the *de novo* assembly, instructing the algorithm to assembled transcripts that were not part of the supplied annotation. The software was used with options *--no-effective-length-correction --library-type fr-firststrand and* supplying also a mask file for rRNA and tRNA genes. Since this assembly step had to be done in separately for each sample Cuffmerge, part of the Cufflinks package, was used to create a unique non-redundant annotation file containing all Refseq transcripts plus the *de novo* assembled ones with the default parameters.

Quantification was done using Cuffdiff, another tool from the same suite, which allowed quantification of transcripts even without replicates, using options *-no-effective-length-correction*, *–multi-read-correct* and supplying once more the rRNA and tRNA mask file.

Genes were considered expressed if their Fragments per kilo base per million mapped reads (FPKM) in any of the three samples was at least 2.0 and misregulated if |log2FC| >= 1 in $Med12^{hypo}$ or $Med12^{null}$ compared to WT. In order to be able to calculate lof2FC in samples with FPKM = 0 in one of the samples, the log2FC was calculate after adding 1.0 to both FPKM values in comparison. To generate files for visualization of data on IGB genome browser, reads mapped to the positive or negative strand were saved in separate .bam files and .wig files generated for mm10 mouse genome using genomeCoverageBed with options *–bga –trackline* and converted to .bigwig with the tool wigToBigWig with default parameters.

For analysing the generated RNA-seq data in section 3.5, read quality was analysed with FastQC which allowed detecting an error during sequencing of base 40 of the second mate reducing the quality of the remaining sequence. As such the last 36 nucleotides of mate 2 were trimmed with Seqtk. While for mate

1 the whole read sequence was used, for mate 2 only the first 39 nucleotides were mapped against mouse genome (mm10).

In order to compare the three aligners, Tophat2 (Kim et al. 2013), Hisat2 (Kim et al. 2015) and STAR (Dobin et al. 2013) were used with default settings with the first replicate of WT sample. After mapping, reads not a proper pair and or with a mapping score below 10 were excluded using bamtool filter option with parameters *"mapQuality": >= 10 "isProperPair" : "true"* and remaining reads quantified with samtools. Based on the final amount of reads, Star aligner was the chosen tool.

Star was used with default parameters, GenCODE annotation file M20 as the transcriptome database and mapped reads were filtered as mentioned above. *De novo* transcripts were assembled using Cufflinks and resulting files were merged in a single non-redundant list using Cuffmerge. Newly assembled genes not described in the M20 annotation were filtered and only transcripts with more than one exon and longer than 200 bp were kept for further analysis. For genes with multiple transcripts, if at least one of the isoforms fulfilled the mentioned requirements then all transcripts for that gene were kept for differential analysis. Filtered transcripts were added to the M20 annotation as well as the final LN-BP18 structure (Figure 14a). Reads overlapping the described exons in the complemented annotation were counted for each sample using Htseq (Anders et al. 2015) with options *–count --mode=intersection-nonempty --stranded=yes --minaqual=10 --type='exon' --idattr='gene_id'*.

In order to perform differential gene analysis between samples, count tables generated by Htseq were supplied to Deseq2 (Love et al. 2014), using a Galaxy instance maintained by the Freiburg University (http://galaxy.uni-freiburg.de/), with default parameters, using all samples as different factor levels for one single factor. Using R (https://www.r-project.org/), files generated for each pairwise comparison were combined.

Genes with a p.adj < 0.05 and |log2FC| >= log2 (1.5) in $Med12^{hypo}$ or $Med12^{null}$ compared with $Med12^{flox}$ were considered misregulated. Density plots were generated using R. IGB (Freese et al. 2016) was used to visualize and export browser views.

Functional enrichment analysis of misregulated genes, gene ontology (GO) term enrichment analysis was performed using DAVID (Huang da et al. 2009, Huang da et al. 2009) was used, supplying a list of misregulated genes for functional analysis and either a list of expressed genes (section 3.1) or a list of all genes with valid p.adj in at least one of the log2FC comparisons (section 3.5) as the background.

For heatmap generation, a file with gene names in the first column and expression values for each sample in the remaining three columns was generated. This file was supplied to Gene Cluster (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Using this tool, expression values

were log transformed and centered by mean. Genes were clustered using centered correlation as similarity metric and centroid linkage as clustering method. Generated heatmaps were visualized using Tree View (Saldanha 2004).

2.12.3. Embryonic tissues RNA-seq data analysis

In order to assess gene expression during embryonic development, a set of 67 transcriptome datasets were obtained from Encode database (Supplementary Table 10). Read count tables for each sample were generated with HtSeq as described in section 0. Expression values for each gene were calculated in FPKM using Ballgown (Frazee et al. 2015) with default settings. Bar plots were generated with R.

2.12.4. Med12 Chromatin Immunoprecipitation analysis

In order to assess if any of the misregulated transcripts identified on section 3.5.1 could be a direct target of Med12, Med12 ChIP-seq data was obtained from a previous study (Kagey et al. 2010). Fastq files for two Med12 ChIP-seq replicates and one control with the run identifiers SRR058985, SRR058986 and SRR058997, respectively, were downloaded from the Gene Expression Omnibus (GEO) database and mapped to mm10 genome using Star with default parameters. Obtained .bam files for both Med12 replicates were merged with samtools. MACS2 (Zhang et al. 2008) was used with option *callpeak* to call peaks on the merged Med12 ChIP sample enriched again control, which resulted in 2269 peaks.

In order to check what genomic features overlapped with the peaks, annotation files containing the promotor region (-2kb to +1kb of TSS) and gene body (+1kb of TSS to end of gene) were created by manipulating genomic coordinates of annotation file using R.

With the Galaxy tool bedtools Intersect intervals (Quinlan et al. 2010) with option *–wb*, overlaps between Med12 peaks and promotor regions were found as long as at least 1bp was overlapping on either strand. Peaks found on promotors were removed from the peak list with Galaxy tool bedtools SubtractBed. Remaining peaks were used to find overlaps with the gene body region using the same strategy as for the promotor and again the overlapping peaks were removed from the peak list. Finally for the peaks not overlapping either promotors or gene bodies, the nearest non overlapping gene either upstream or downstream of the peak was found using Galaxy tool Fetch closest non-overlapping feature.

Venn diagrams were generated using the VennDiagram package for R.

3. Results

3.1. Transcriptome analysis and *de novo* transcript assembly using Med12 mutant embryonic stem cells data

As it was shown in previous studies by the Schrewe laboratory (Rocha et al. 2010, Rocha et al. 2010), Med12 is essential for canonical Wnt and Wnt/PCP pathways and its depletion resulted in striking phenotypes during mouse development. Disruption of canonical Wnt pathway was confirmed by observed downregulation of β-catenin target genes and by defects in processes where these genes are involved, such as axis truncation and malformation of the branchial arcs. Other observed phenotypes are associated with Wnt/PCR pathway such as defects in neural tube closure, closure of the ventricular septum and open palate. MED12 mutations also affect the normal development in humans, with patients suffering from different Med12 associated X-linked intellectual disabilities (XLID) syndromes revealing multiple defects, such hypoplastic heart, agenesis of the corpus callosum and maxillary hypoplasia (Graham et al. 2013). Mutations on Med12 have also been found in a number of different cancer types, such as prostate cancer and chronic lymphocytic leukemia, but also in the majority of analysed benign uterine leiomyomas and benign breast fibroadenoma, making the study of Med12 important also in a clinical setting.

In order to evaluate the impact of Med12 deficiency in a more homogeneous population, Med12 mutant embryonic stem cells (ESCs) that were previously generated in order to study the function of Med12 in embryonic development were used. These clones were $Med12^{hypo}$, in which Med12 expression was reduced by 95% and $Med12^{null}$, which were completely devoid of Med12 (Figure 6a) (Rocha et al. 2010).

The analysis of gene expression in embryos generated from these mutant cells was done mostly by whole mount in situ hybridization (WISH), which despite allowing the detection of the spatial and temporal expression of target genes in a whole embryo, it is a method that lack sensitivity and that does not allow precise quantification of gene expression. To overcome these limitations, all the transcripts expressed in the Med12 mutant ESC were quantified using RNA-seq data, which allows not only quantifying gene expression more precisely and measuring small variations in expression between samples, but also detecting lowly expressed genes. In order to verify genes misregulated in these ESC mutants, a WT ESC control was also included in the analysis. This dataset included data generate for one sample of each genotype. With the analysis of these RNA-seq dataset, disturbed gees with a role in pathways and biological processes that would explain the observed defects in the embryos generated

from these mutant ESCs were expected to be found. Furthermore, previous studies have shown the existence of lncRNAs whose expression was dependent of Med12 in human cells (Lai et al. 2013). In order to identify additional lncRNAs regulated by Med12 in ESC, non-coding annotations were also included in the pipeline analysis.

One important factor in RNA-seq is variability, resulting from technical and biological variation. For this reason, replicates are necessary to perform statistical tests on gene expression from which to measure variation. As such, although analysis of these data could provide insights into Med12 influence on overall gene expression, for certain genes, the differences observed between samples might be not statistically significant.

Cufflinks is an algorithm that can perform *de novo* assembly of transcripts that are not present in the annotation database. Before expression quantification, Cufflinks was used in order to find possible new transcripts not described in the used RefSeq transcriptome data (Pruitt et al. 2007). This algorithm predicted thousands of new transcripts. For several of them, the predicted structure was assembled with either only a few reads supporting the structure, or the whole locus was transcribed but only a part of it was predicted by Cufflinks. In order to minimize this kind of artefacts, only transcripts longer than 200 nt were selected. This allowed discarding all the very small transcripts, which accounted for hundreds of new transcripts. Additionally, transcripts with only one exon were discarded. This allowed keeping only transcripts for which Cufflinks identified a spliced read, which should only occur on true transcribed genes. Such filtering allowed to keep only the transcripts with higher probability to reflect true genes. After gene expression quantification, the thousands of newly assembled transcripts by Cufflinks were

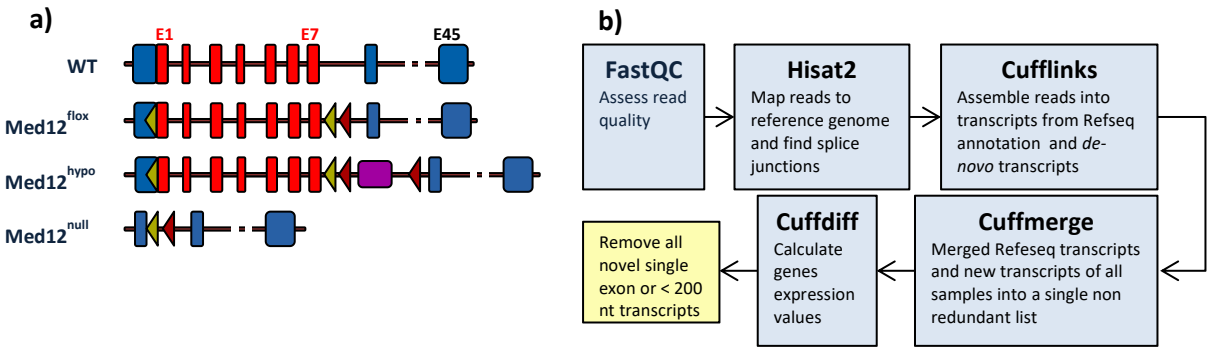


Figure 6 – Med12 mutant ESC transcriptome analysis
a) Schematic representation of Med12 mutant alleles of the ESCs used and generated in a previous study (Rocha et al. 2010). LoxP sites, yellow triangles; frt sites, red triangles; exons, blue boxes; exons removed upon Cre expression, red boxes; neomycin resistance cassette, purple box; **b)** Flowchart of tools used and their function for RNA-seq analysis.

filtered (Figure 6b), reducing the number to almost 400 new genes. Out of over 23.000 genes analysed,

less than half were expressed above 2.0 FPKM (fragments per kilobase million) in one of the three samples. Genes that did not meet this criterion were considered as not expressed, resulting in 11,000 expressed genes of which almost 1,700 were misregulated over 2-fold in the $Med12^{hypo}$ and/or the $Med12^{null}$ mutant. The vast majority of these were coding genes, with 14% either annotated as non-coding or new Cufflinks genes.

3.1.1. Analysis of misregulated protein coding genes

Analysis of misregulated genes in Med12 deficient ESC was first focused on the annotated protein coding genes, in order to correlated with results obtained from a previous study which made use of these cells (Rocha et al. 2010). Misregulated coding genes were clustered based on their expression in the three samples; WT, $Med12^{hypo}$ and $Med12^{null}$ (Figure 7), which revealed groups of misregulated genes with a similar expression dynamic in all the samples. Heatmaps of these clustered genes were generated and gene expression colour coded according to their Z-score, a normalized expression value that allows more easily to compare genes with different orders of expression values. Despite the formation of five clusters, genes were observed to be either upregulated (clusters 1, 3 and 5) or downregulated (clusters 2 and 4) in both mutants, but with different degrees of misregulation between samples.

Using functional enrichment analysis, genes that share a biological function and that are over-represented in a specific dataset can be identified. The Database for Annotation, Visualization and Integrated Discovery (DAVID) tools allow to identify enriched Gene ontology (GO) terms in a given dataset. These GO terms describe gene function, identifying, among other aspects, the biological process in which a gene is involved, such as “neural tube closure”. To understand the biological significance of the generated clusters for the misregulated coding genes, GO term enrichment analysis was performed for each of the five identified clusters using DAVID. For this analysis, misregulated genes on each cluster were supplied as individual lists and the full list of expressed genes in any of the three samples was used as background (Table 1). Over 30 genes encoding for ribosomal proteins (such as Ribosomal protein S7 (RPS17) or RPL22) were upregulated in the mutant cells and assigned to cluster 1. As such, terms linked to translation were among the most enriched hits in this cluster. When looking at downregulated genes in both mutants (cluster 2 and 4) several are involved in diverse organs development processes such as heart development (e.g. Spalt like transcription factor 1 (Sall1), Gli family zinc finger 2 (Gli2) and TEA Domain Transcription Factor 1 (Tead1)).

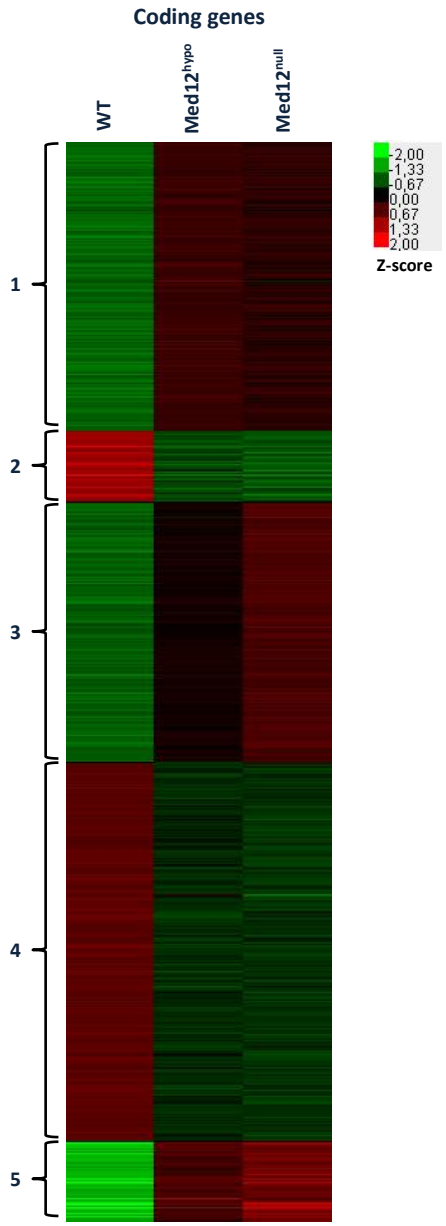


Figure 7 - Clustering of misregulated protein coding genes
Clustering was based on genes expression in the different samples analysed and revealed genes similarly misregulated in both Med12 mutants. The expression of each gene was expressed using Z-score, a normalized expression value, in order to compare genes with different expression levels. Genes with lower expression values are coloured in green and with higher expression coloured in red.

Table 1 - Gene ontology enrichment analysis on 5 clusters identified for misregulated genes in Med12 mutant ESCs.
Selected hits are presented in this table. A more complete list of terms is shown on Supplementary Table 12.

Cluster	Go term	p-value
1	translation	1.30E-07
	ribosomal small subunit assembly	1.30E-05
	cardiac muscle contraction	4.20E-02
2	regulation of transcription from Pol II promoters	2.30E-05
	heart development	1.20E-03
	axon guidance	6.60E-03
	palate development	7.40E-03
3	outer dynein arm assembly	9.50E-04
	outflow tract septum morphogenesis	5.20E-02
4	regulation of transcription from Pol II promoters	6.20E-06
	axon guidance	6.30E-05
	patterning blood vessels / embryonic hemopoiesis	1.10E-03
	heart development	2.70E-03
5	angiogenesis	2.90E-04
	brain development	4.10E-02
	neuron differentiation	8.90E-02

Genes such as Roundabout guidance receptor 1 (Robo1) and Gastrulation brain homeobox 2 (Gbx2) are among the downregulated genes that function in axon guidance, a process critical for the proper closure of the neural tube. In cluster 5, upregulated genes in the mutants are involved in different processes important for brain development (e.g. Secretin (SCT), Collagen type IV alpha 1 chain (Col4a1) and Cbp/P300-interacting transactivator 1 (Cited1)), such as angiogenesis, an important process for normal development of most organs with upregulated genes important for this process such as Wnt7b (Wnt family member 7Bb) and Heart and neural crest derivatives expressed 1 (Hand1). Genes important for proper neuron differentiation were also assigned to cluster 5, including Inhibitor of differentiation 1 (Id1) and DNA damage inducible transcript 4 (DDIT4). While both Med12 mutant cells show a similar variation on gene expression, as observed by cluster 1, 2, 4 and 5, in cluster 3 are genes for which the remaining 5% of Med12 in the $Med12^{hypo}$ mutant (Rocha et al. 2010) was enough to keep them partially repressed. Genes in this cluster were slightly upregulated in the $Med12^{hypo}$ mutant and even more upregulated in the $Med12^{null}$. Multiple genes in this cluster were involved in cell movement (outer dynein arm plays a role in the beating movement of cilia or flagella). Interestingly, enrichment of genes involved

in heart development and axon guidance reflected two phenotypes observed on reported embryos: heart defects and neural tube closure defects. The fact that the identified misregulated genes in Med12 mutant cells and their known functions validate the defects observed in the reported embryos generated from the same cells, confirms the observations of this analysis.

A previous study has concluded that Med12 was important for proper maintenance of Nanog target genes expression, since knockdown of Med12 in mESC by small interfering RNAs resulted in downregulation of Nanog and consequent misregulation of its target genes, such as Sox2, Oct4 and Dkk1 (Tutter et al. 2009). However a more recent paper, where Med12 was depleted in mESC by direct genomic editing, no effect on Nanog target genes was observed (Rocha et al. 2010). The role of Med12 in pluripotency was also not supported by the analysed RNA-seq data from Med12 mutant ESCs, since no effect on the mentioned Nanog target genes or on Nanog was observed (data not shown)

3.1.2. Analysis of the non-coding transcriptome

Previous studies have linked MED12 to the regulation of a multitude of coding genes, however few have analysed the effect of Med12 on the expression of non-coding genes. One such study observed a downregulation of a subset of lncRNAs upon depletion of MED12 in HeLa cells. During analysis the transcriptome data of $Med12^{hypo}$ and $Med12^{null}$ mutant ESCs, in order to further verify the effect of Med12 depletion on ncRNA expression, genes annotated as non-coding were analysed. Furthermore, new transcripts were assembled based of the mapped reads with Cufflinks, allowing identification of putative new genes.

Over 200 ncRNAs were identified as misregulated in one of the mutant cells analysed, including 38 putative new genes. Clustering of misregulated non-coding genes revealed that several were differently affected in the two mutants (Figure 8a), contrary to what was observed for the coding genes (Figure 7). Distribution of log2FC for misregulated coding genes (Figure 8b, left) confirmed the similar pattern on both mutants, with a slightly stronger upregulation of genes on $Med12^{null}$. However, when only the non-coding genes were analysed, the difference between the two mutants became more evident (Figure 8b, right). Although GO term enrichment analysis is a useful method to detect enrichment of coding genes in a dataset that function on the same process, for studying ncRNAs, this method usually does not provide any significant results. This is due to the fact that although thousands of ncRNAs have been described, only for a small subset of these has their function been characterized. As such, the GO terms associated with non-coding genes are far less than with coding genes. However, when this method was applied to

the misregulated ncRNAS identified on Med12 deficient ESCs, enrichment of genes associated with the process of X chromosome inactivation (XCI) was observed (Figure 8c).

While males have only one copy of the X chromosome, females have two. In order to prevent a double dosage of genes of this chromosome, one of the X chromosomes in females is randomly inactivated through XCI. This is one of the few processes where multiple lncRNAs are involved that has been deeply

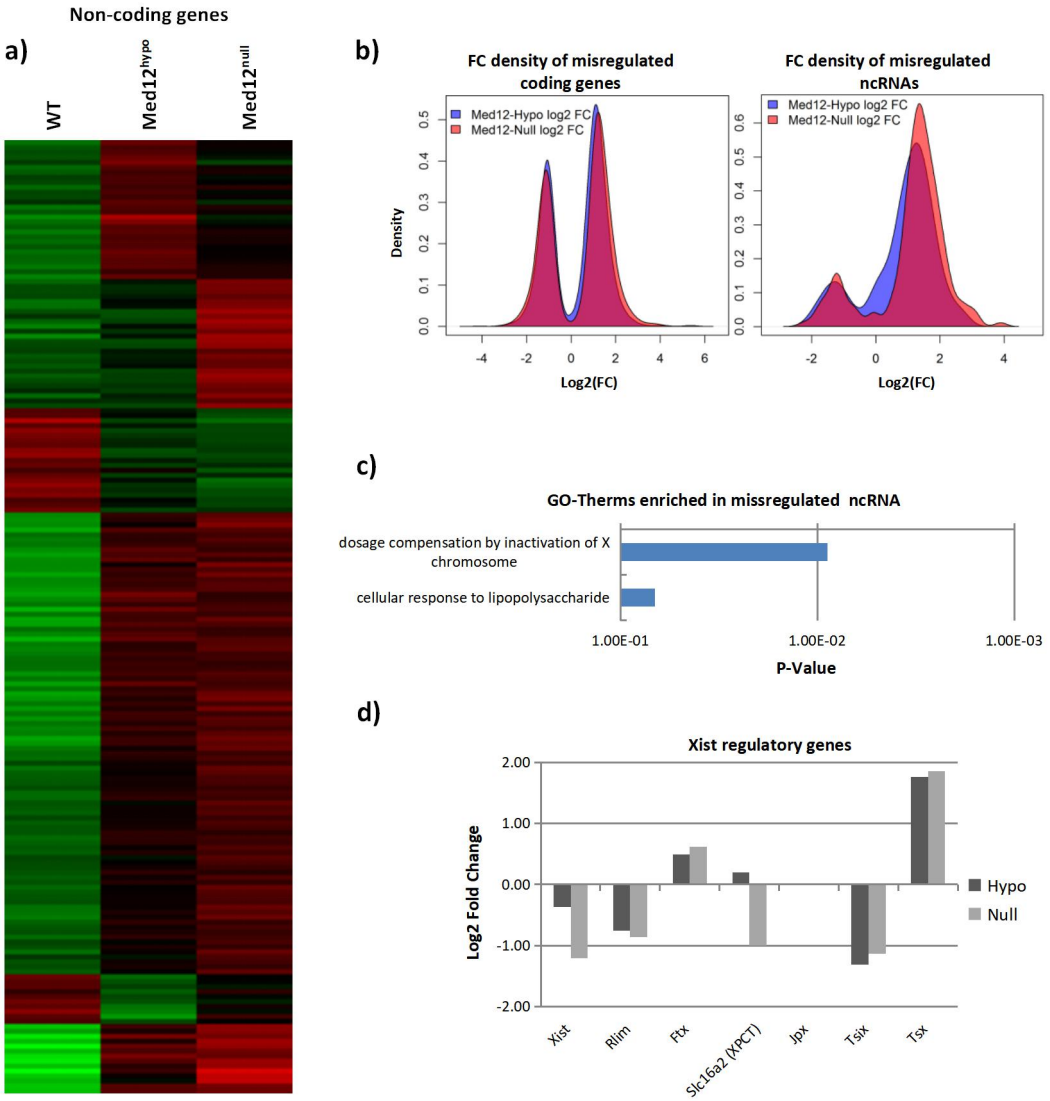


Figure 8 - Analysis of ncRNAs in Med12 mutant cells transcriptome data

a) Clustered heatmap of misregulated lncRNAs expression in Med12 mutants and WT samples. Z-scores represent normalized expression values, in order to compare genes with different expression levels. Genes with lower expression values are coloured in green and with higher expression coloured in red; **b)** density plots of log2FC distribution on both Med12 mutants compared to the WT sample for misregulated protein coding genes (left) and ncRNAs (right) showing that the difference between mutants was more accentuated for ncRNAs; **c)** GO term enrichment analysis terms for misregulated ncRNAs; **d)** log2FC for Xist and other genes with a role on XCI calculated from the RNA-seq data.

studied and characterized, explaining the association of the GO terms with genes in this process (Penny et al. 1996, Lee et al. 1999, Sun et al. 2013). To verify the effect of Med12 depletion on XCI, log2FC for Xist the main factor in XCI and its known regulators was calculated from the RNAseq data (Figure 8d). Xist and multiple of its regulators were misregulated in at least one of the mutants, suggesting a potential role for Med12 in this process. However, the ESCs used in this study were derived from male blastocysts, as such XCI it should not be an active process in these cells. To confirm the role of Med12 in XCI, a different system where XCI is active should be used, such as female Med12 deficient ESC.

Analysis of the Med12 depleted ESC transcriptome data revealed hundreds on misregulated ncRNAs, including potential novel genes. These results confirm Med12 role as an important regulator of multiple lncRNAs expression in ESCs.

3.1.3. Identification and validation of putative novel non-coding genes

Of the almost 400 Cufflinks putative new lncRNAs, 38 were found misregulated in $Med12^{hypo}$ and/or $Med12^{null}$. For these, the structure predicted by Cufflinks was visually inspected and compared to the mapped reads in all samples. This allowed excluding loci where the mapped reads were not sufficient to sustain the Cufflinks predictions. These included repetitive elements, which explained why an accumulation of reads was observed, regions where reads were mapped at very low level, or long genomics regions with reads mapped throughout the whole region, without the accumulation of reads that usually occur at exons of mature RNAs. Of the 38 Cufflinks predictions, 11 displayed a promising predicted structure that resembled a spliced gene (Figure 9, black track). For these, primers were designed along the predicted exons and in nearby regions that, by assessing mapped reads distribution on the locus, could be part of the predicted genes but were not be identified by Cufflinks. The primers were used in all possible combinations for every predicted transcript.

Since there was no data confirming the presence of a poly-A tail in any of the transcripts, random hexamers were used to reverse transcribe RNA extracted from WT G4 ESCs. The use of random hexamers allows to generate cDNA from all the RNA molecules present in a sample, contrary to oligo dT that only allows to reverse transcribe RNA molecules with a poly-A tail. PCR amplified fragments were separated by size using gel electrophoresis and the individual bands for each gene purified from the agarose gel, cloned into pCR2.1-TOPO vector and the sequence of inserted fragments determined by Sanger sequencing. With this approach all of the 11 predicted genes were confirmed to be expressed in ESCs and, for a subset of them, multiple isoforms could be isolated (Figure 9, blue track). Most of the

identified exons matched the mapped reads in at least one of the samples (Figure 9, red tracks), confirming the obtained data from RNA-seq. Additionally, gene structures obtained by this experimental procedure (Figure 9, blue track) were very similar to the predicted structures by Cufflinks (Figure 9, black track), demonstrating the reliability of this tool in predicting potential new genes.

As it was shown for other putative lncRNAs, small peptides might arise from genes previously characterized as a non-coding gene (Nelson et al. 2016). To evaluate the possibility of such small peptides being encoded by the isolated novel transcripts, their coding potential was assessed using CPAT. This tool calculates the coding probability of transcripts based on their sequence, with a cut-off determined using a training set supplied (Wang et al. 2013). For mouse transcripts, the optimal threshold used for defining non-coding transcripts was determined by the developers of CPAT to be 0.44, meaning that transcripts with a lower score are not likely to code for proteins. For all the transcripts analysed, their coding probability was below the threshold, supporting their non-coding character (Table 2).

Overall, the analyses allowed to identify 11 putative novel lncRNAs expressed in ESCs, in whose expression was Med12 dependent, since their expression was misregulated in Med12 deficient ESCs.

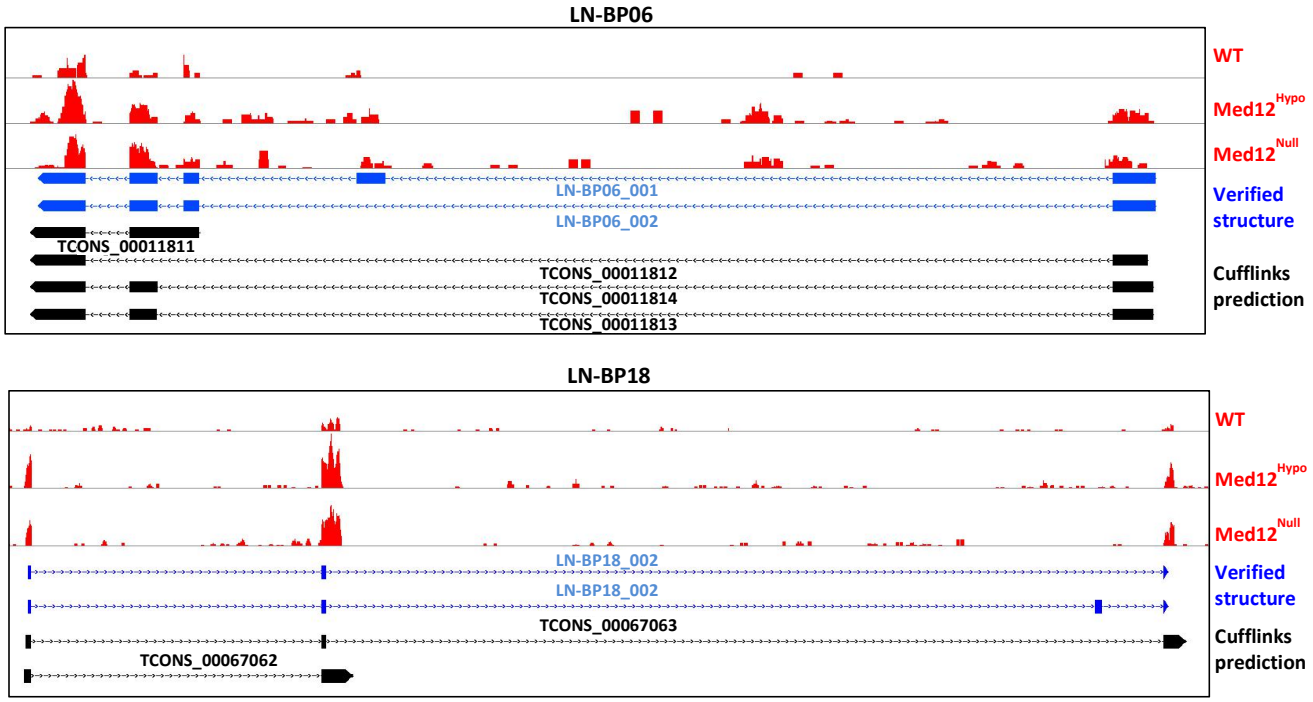


Figure 9 - Misregulated putative novel lncRNAs

Browser view of experimentally obtained gene structures for two of the identified misregulated putative lncRNAs predicted by Cufflinks. Mapped reads from the 3 samples, red track; structure obtained for the genes after PCR amplification, blue track; gene structure predicted by Cufflinks based on mapped reads, black track.

Table 2 - coding potential of misregulated putative novel lncRNAs
For each gene, only the isoform with the highest coding probability is shown; "mRNA size" represents the total length of the transcripts after maturation, "ORF size" the size of the longest open reading frame (ORF) found in any of the 3 reading frames on the same strand as the gene. Transcripts with a coding probability below 0.44 do not code for any protein according to the CPAT algorithm.

Transcript	mRNA size	ORF size	Coding probability
LN-BP02_1	676	162	0.06
LN-BP06_1	873	147	0.02
LN-BP07_1	410	159	0.06
LN-BP11_1	616	240	0.07
LN-BP13_1	932	282	0.12
LN-BP13_6	311	132	0.03
LN-BP17_1	704	156	0.03
LN-BP18_1	289	78	0.04
LN-BP23_1	704	174	0.16
LN-BP26_1	828	99	0.03
LN-BP31_1	911	114	0.03

3.1.4. Characterization of expression of novel non-coding genes in embryos

As shown, all of the 11 Cufflinks predicted novel genes were confirmed to be expressed in ESCs. In order to study their potential expression pattern during mouse embryonic development, whole mount in situ hybridization (WISH) was performed for all 11 lncRNAs, using E9.5, E10.5 and E11.5 WT embryos. The isolated transcripts (Figure 9 and Supplementary Figure 1) were used to generate antisense RNA probes. These probes were labelled with Digoxigenin11dUTP (DIG) and their location determined with anti-DIG antibody. As showed in Figure 10a, no distinct specific pattern could be identified for the tested genes. Although no specific staining was observed for these genes, there was some signal obtained that very likely resulted from trapping of the probe or of the staining solution. The unspecific staining resulted from trapping in the embryos brain, as observed to LN-BP13 or LN-BP31, from staining solution deposited in the mouse surface, as is the case of the signal observed in the limbs and neural tube of embryo used for LN-B06 detection or from trapping in the otic vesicles, a signal obtained for several of the analysed embryos (Figure 10a). However, a specific signal could be observed for LN-BP18, with a faint expression domain observed on both limbs and also on the caudal end of the embryos (Figure 10b, bottom row).

Multiple lncRNAs have been shown to regulate the expression of their neighbouring coding genes (Anderson et al. 2016, Paralkar et al. 2016). Furthermore, genes with similar expression patterns suggest a shared function and/or regulation. Sall1 is a transcription factor (TF) critical for normal development of

the kidney, a tissue where it modulates Wnt signalling (Sato et al. 2004, Kiefer et al. 2010). It is also involved in neural tube closure and in ESC differentiation (Bohm et al. 2008, Karantzali et al. 2011). Interestingly, LN-BP18 and Sall1 are two divergent genes 9 kb apart, located on chromosome 8. To verify if these genes were co-expressed, probes against Sall1 were used for WISH. The data obtained showed that these two neighbouring genes were expressed in similar tissues, specifically in forelimbs, hind limbs and in the caudal end (Figure 10b). Publicly available transcriptome data generated by the Wold laboratory for 12 tissues (forebrain, midbrain, hindbrain, neural tube, heart, kidney, liver, lung, intestine, stomach, limb and thymus) across up to 8 developmental time points, from E10.5 to postnatal day 0, were downloaded in order to evaluate the *in vivo* expression of these two genes. The observed expression domains identified by WISH were confirmed by the transcriptome data, corresponding to tissues with high expression of both genes (Figure 10c). Additionally, new expression domains were identified, namely the kidney and neural tube and throughout the brain, albeit at lower levels. Furthermore, in most of the analysed tissues, Sall1 expression was 6-fold higher than that of LN-BP18.

Despite being expressed in ESC, all except one of the 11 analysed transcripts revealed an expression pattern in E10.5 mouse embryos. For LN-BP18, expression was detected in the limbs and in the caudal end of embryos, similar to the detected expression pattern of its coding neighbour Sall1.

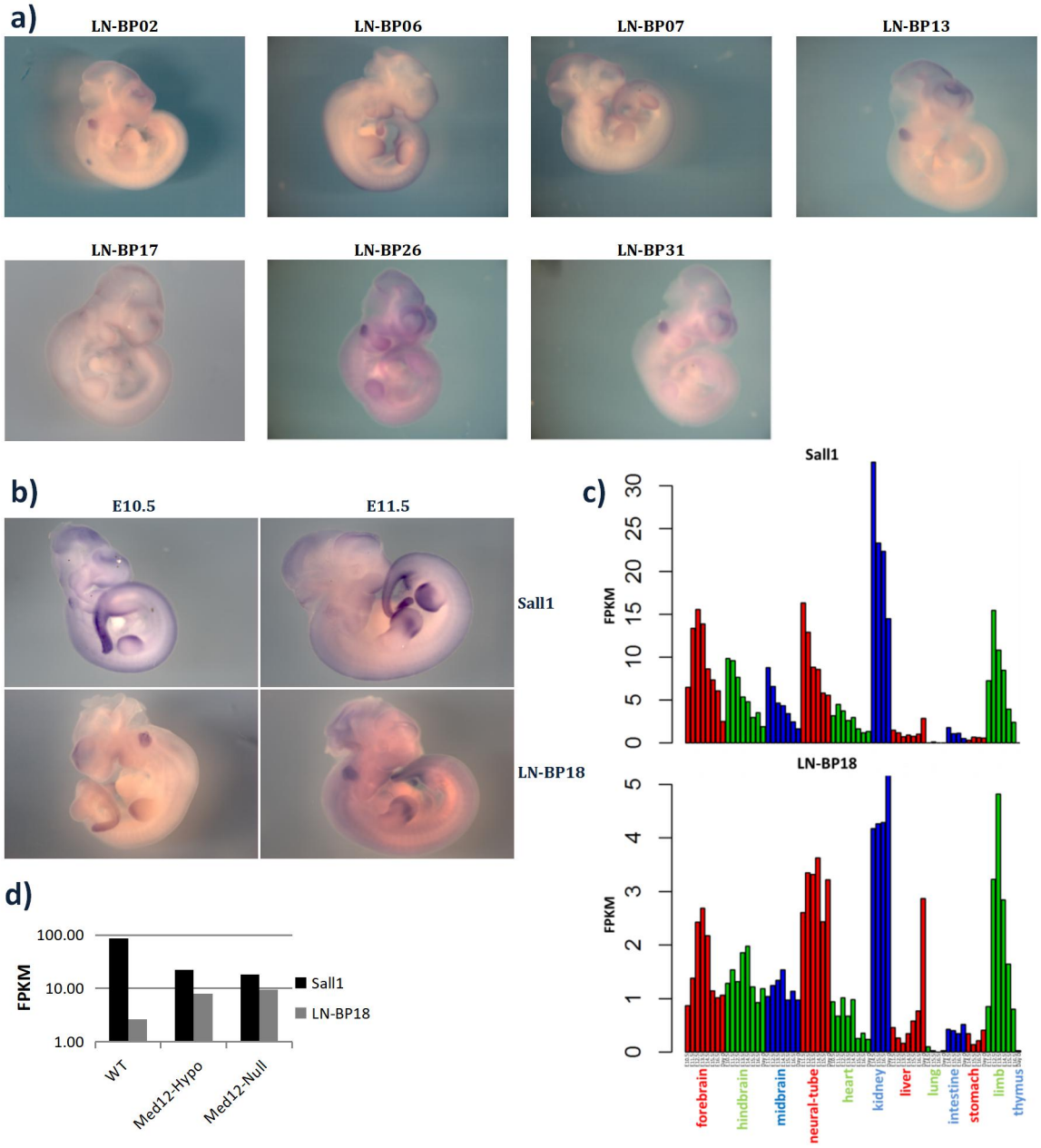


Figure 10 - *In vivo* expression of a misregulated putative novel lncRNAs

a) Whole mount *in situ* hybridization (WISH) performed on E11.5 WT mouse embryos using probes against the putative new ncRNAs identified and **b)** against Sall1 and LN-BP18 in E10.5 and 11.5 CD1 embryos revealing similar pattern for these two genes; **c)** Sall1 and LN-BP18 expression quantified in 12 different tissues across eight developmental stages of mouse embryos, downloaded from Encode database, **d)** LN-BP18 and Sall1 expression from the RNA-seq data from the three different ESCs analysed.

3.2. Characterization of the long non-coding gene LN-BP18

As mentioned before, LN-BP18 was referred to as a novel gene since it was not described in the Refseq database used for the analysis of transcriptome data from Med12 mutant cells. A search on the GenCODE database (https://www.gencodegenes.org/), one of the most complete databases for ncRNAs with data generated from different cell types and tissues, did not reveal any information regarding LN-BP18. However, a predicated transcript generated by automated computational analysis, termed Gm3134 (Figure 11, black track) has been described in the NCBI database (https://www.ncbi.nlm.nih.gov/refseq/) in the same genomic region, on mouse chromosome 8. This gene, annotated as a ncRNA, was automatically assembled by analyses of the mouse ENCODE transcriptome data (Yue et al. 2014), which was generated from 123 mouse cell types and primary tissues. Analyses of this transcriptome data showed a very low expression for this ncRNA in all the analyzed tissues. However a small enrichment was described in the central nervous system, kidney and liver. Associated with this gene was a single study with a specific mention of this locus (Thiagarajan et al. 2011). In this study, the only observation was the expression of a Sall1 divergent transcript with similar expression pattern to Sall1. Specifically, using RNA-seq and histological sections *in situ* hybridization, both were found to be expressed in early nephrons of E15.5 mice embryos. No further characterization was performed on this divergent transcript. Thus, the only information available supporting expression of a Sall1 divergent transcript was the prediction for Gm3134 and the observation that there was transcription originating from this locus in mouse embryonic kidneys.

Comparing the experimentally obtained structure of LN-BP18 with the predicted for Gm3134 revealed big discrepancies (Figure 11). According to the automatically generated data, Gm3134 had 9 exons which could be spliced into 12 different isoforms, ranging from 370 bp to over 3.7 kb. The whole gene spanned a genomic region of 31 kb on chromosome 8, including 5 kb which overlapped with the coding region of Sall1 (Figure 11, black track). On the other hand, experimental data revealed 4 exons for LN-BP18, including one not predicted for Gm3134. For LN-BP18, two small isoforms were identified with sizes below 300 bp, spanning 16 kb and starting 9 kb downstream of Sall1 (Figure 11, blue track).

In order to validate LN-BP18 gene structure and compare it with Gm3134 predictions, chromatin immunoprecipitation sequencing (ChIP-seq) datasets for different histone modifications in ESC were obtained from previously published studies (Mikkelsen et al. 2007, Marson et al. 2008, Creyghton et al. 2010). H3K27ac and H3K4me3 are two histone marks associated with active promoters. Since an active transcription start site (TSS) is free of nucleosomes so that the pre-initiation complex (PIC) can assemble,

usually these “active marks” accumulate in the nucleosomes flanking this region. This sort of “double peak” enrichment was detected near the first exon of LN-BP18, supporting the Cufflinks prediction (Figure 9). Additionally, in Med12^{hypo} and Med12^{null}, an upregulation was observed for this locus. Analysing the reads mapped to the positive strand (Figure 11, top two red tracks), it could be observed that the regions where more reads were mapped in the Med12^{null} mutant, comparing to the WT sample, corresponded to the identified LN-BP18 exons. The TSS of several predicted isoforms for Gm3134 resided within the first intron of Sall1, in a region where “active marks” H3K4me3 and H3K27ac were present, supporting this region as a potential active TSS. However Sall1 TSS was also near this region, making the distinction between Sall1 and Gm3134 TSS associated histone marks difficult. The analysed data generated from ESC did not support the NCBI prediction since no transcription was observed for most of the predicted exons. This can be due to the prediction being based on mouse ENCODE transcriptome data, a dataset generated from 123 cell types and tissues. As such, it was possible that some, if not most, of the predicted features to not be supported by data derived from ESCs.

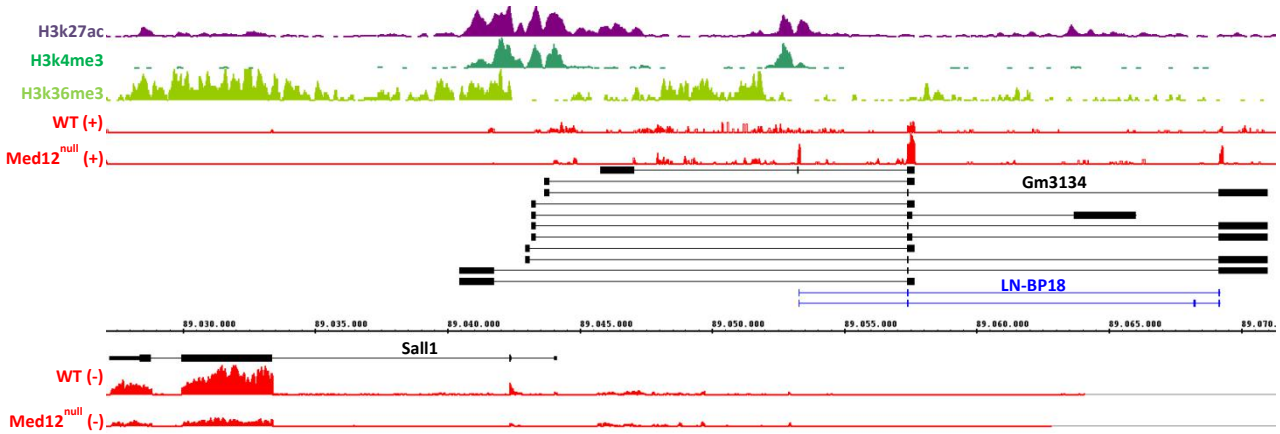


Figure 11 - Gm3134 locus and predicted structure
Histone modifications associated with active genes (H3K27ac and H3K4me3) are found at the transcription start site (TSS) for LN-BP18 experimentally verified structure (Figure 9) (first exon, blue track) and also the TSS of several of the Gm3134 predicted isoforms (first exons, black track). Histone modification H3K36me3 is found along the gene body of expressed genes and is associated with Sall1 gene body and to a lower degree to LN-BP18 gene. LN-BP18 exons (blue boxes) correspond to the genomic locations where an increased number of reads were mapped to in the Med12^{null} mutant compared to the WT sample (red tracks). Most of Gm3134 exons (black boxes) show no signs of expression in mESC. In red tracks, (+) indicates reads mapped to the positive strand and (-) to the negative strand.

3.2.1. Characterization of LN-BP18 gene structure

The transcriptome data obtained for Med12 mutants ESCs supported the gene structure of LN-BP18 predicted by Cufflinks. However, it did not support the NCBI prediction for Gm3134, since no transcription could be attributed to the predicted exons that did not overlap the LN-BP18 exons. Additionally, the identified LN-BP18 gene structure did not match the predicted structure from NCBI, including an exon not present in Gm3134. Furthermore, before functional analysis of a lncRNA can be done, it is necessary to understand its transcript characteristics, such as transcription start site (TSS), transcription end site (TES), possible isoforms, presence or lack of polyadenylated tail (poly-A tail). As such, a more in depth characterization of this gene was performed.

3.2.1.1. Identification of LN-BP18 transcription start site

The identification of the TSS of a gene can be achieved by using 5' rapid amplification of cDNA ends (5' RACE) approach. With this method a gene specific primer (GSP) is used to generate cDNA only from the gene of interest. Through terminal deoxynucleotidyl transferase, a poly C tail is added to the 3' of the newly synthesized single strand cDNA, which is used as a binding site for an adapter primer. This allows PCR amplification of transcripts without the prior knowledge of the original RNA 5' end. Three different GSP on the last two exons common to both isoforms of LN-BP18 (Figure 11 blue track) were used on WT ESC cDNA. The TSS could be identified as part of a new exon located in the first intron of Sall1, 10 kb upstream of the original prediction and overlapping two of Gm3134 predicted TSS (Figure 12, TSS1). A second round of 5' RACE using GSP located on the newly identified exon confirmed the location of LN-BP18 TSS, termed TSS1. However, as discriminated on the previous section, several lines of evidence supported the existence of the Cufflinks predicted TSS. Furthermore, the RNA-seq data obtained from Med12 mutant ESCs (Figure 11, top red tracks) did not support transcription originating from TSS1. To verify if the Cufflinks prediction TSS could be a true TSS, as supported by the mentioned data, a 5' RACE was repeated using primers targeting the predicted first exon. The new data indicated the presence of an additional TSS located on this exon, which started 170 bp earlier than previously identified and was designated as TSS2. Both identified TSSs of LN-BP18 were supported by the presence of H3K4me3 and H3K27ac (Figure 11). In order to further validate the presence of these two TSSs, publicly available data was obtained for ESCs (Figure 12, green tracks) and differentiated tissues (Figure 12, blue tracks). ChIP-seq data for different histone modifications in ESCs revealed H3K4me3 and H3K27ac peaks around both LN-BP18 TSSs, as observed before.

During transcription by Pol II, a cap structure is added to the 5' end of the transcripts. This cap, consisting of an inverted 7-methyl guanosine, is added to the first residue of the nascent RNA and protects it from exonucleases degradation and is important for different processes, such as nuclear export and RNA splicing (Lewis et al. 1997). Cap analysis gene expression (CAGE) is a method that enables the sequencing of the 5' end of RNAs and allows the identification of TSSs in a genome-wide scale. CAGE peaks were found closely to both TSSs of LN-BP18 (Figure 12, red peaks), which together with the mentioned histone marks strongly support the true existence of both TSSs. Additionally DNase I hypersensitive sites sequencing (DNase), which is used to identify regions of open chromatin and RNA polymerase II (Pol II) ChIP-seq data revealed that, in ESC, both TSS are accessible and that RNA Pol II binds them. On the other hand, when analysing the data originating from differentiated tissues, a loss of both H3K4me3 and H3K27ac was observed on the TSS2 together with loss of Pol II binding and closure of chromatin. These data obtained from the mentioned diverse methods supported the existence of two different LN-BP18 TSSs. There is now clear evidence of active transcription in TSS1 in differentiated and ESCs cells, since activating histone marks (H3K27ac and H3K4me3) can be found flanking this site, the

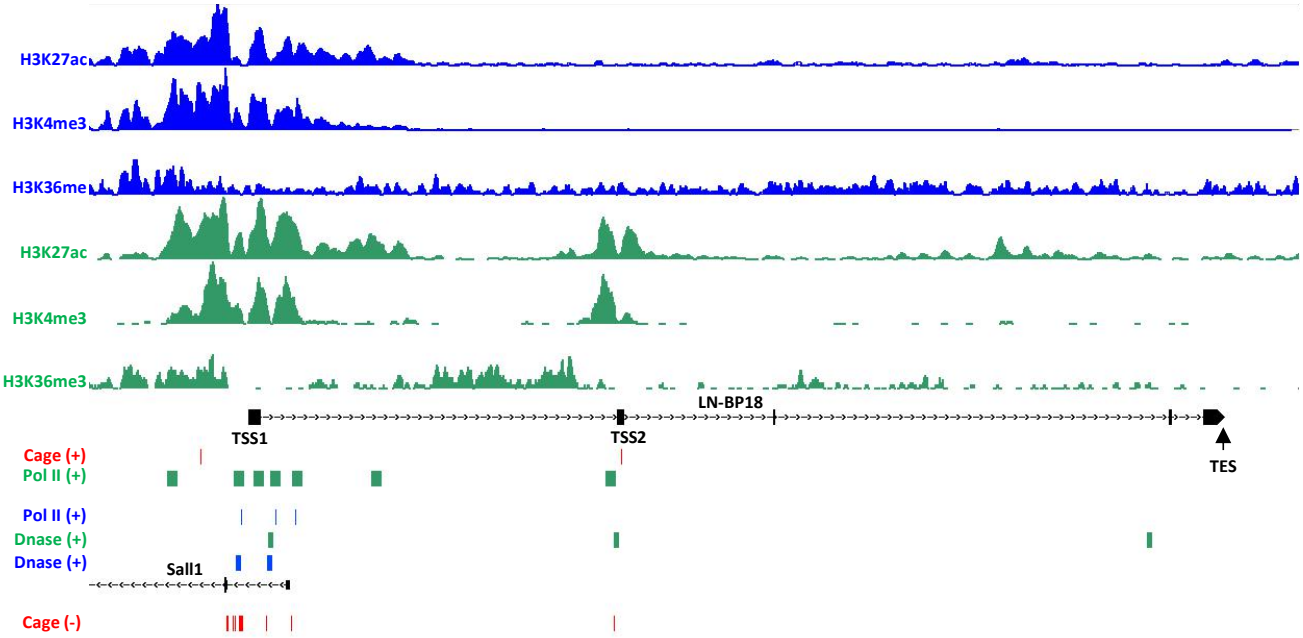


Figure 12 - LN-BP18 TSS2 is ESC specific
LN-BP18 locus with both TSS indicated, together with signal from histone marks found at active promoters (H3K27ac and H3K4me3). H3K36me3 is found on gene body of active genes. Pol II ChIP-seq and I data was also obtained for both ESC and differentiated tissues and indicate regions with active or poised transcription and accessible chromatin, respectively. Data in green generated from ESC and in blue from differentiated tissue, with E14.5 mouse embryo brain used as a representative tissue. CAGE data was stranded and generated from different mouse tissues and cell lines. CAGE data, red track

chromatin is accessible (DNase I), Pol II is found binding in this region and a CAGE peak is found is close proximity. However, for TSS2, clear evidence was found only in ESCs

3.2.1.2. Identification of LN-BP18 transcription end site

Having identified the different TSSs for LN-BP18, the next step was to determine the transcriptional end site (TES) of LN-BP18 using a 3' RACE protocol. This protocol can only be applied to transcripts with a poly-A tail since it relies on the use of an oligo dT primer binding to this tail. As mentioned before, the majority of lncRNAs possess a poly-A tail, however multiple functional lncRNAs, such as ecCEBPA which interacts with DNA methyl transferase 1 (DNMT1) and keeps its own promoter free of methylation, do not (Lai et al. 2014). To verify if LN-BP18 was modified with a poly-A tail, cDNA was generated using an oligo dT primer instead of random hexamers. The use of oligo dT only generates cDNA from genes that are modified with a poly-A tail. The cDNA generated from polyadenylated RNA was used as template for PCR amplification of LN-BP18 transcripts, which resulted in their successful amplification, suggesting the presence of a poly-A tail on LN-BP18.

With the assumption of the presence of a poly-A tail added to this gene, the 3' RACE protocol could be applied. However the obtained TES matched a stretch of 27 adenines present in the last exon (Figure 13a, black "A" in tandem). This led to the hypothesis that the oligo dT, used on the first step of the 3' RACE protocol to generate cDNA, did not bind the poly-A tail added to processed RNAs, but bound instead the stretch of adenines found. From a single RNA molecule, this would results in the generation of two different cDNA fragments; one from the TSS up to the oligo dT that bound the adenines in tandem in the last exon and the other comprising the sequence between both oligo dT (Figure 13a). To confirm this hypothesis, cDNA generated with either oligo dT or random hexamers was used as template for PCR using a primer on exon 1 and a reverse primer complementary to a downstream region of the identified stretch of adenines (Figure 13a, red arrows). As shown in Figure 13b, when using oligo dT no product was obtained, contrary to when random hexamers were used to generate the cDNA.

These data supported the hypothesis that the oligo dT used in the 3' RACE bound prematurely to a stretch of adenines on the last exon, which resulted in a truncated cDNA. Despite the premature binding of the oligo dT on this secondary region, a different oligo dT should have still bound the LN-BP18 poly-A tail and produced cDNA until the reverse transcriptase was released by the premature bound oligo. As such, when using GSPs downstream of the premature binding region, they should allow identifying the cDNA generated from the poly-A tail using 3' RACE, which in turn would identify LN-BP18 TES. In fact,

using these primers (Figure 13a, blue arrows), the majority of products obtained terminated shortly after 600bp of the beginning of the last exon, marking this as the LN-BP18 TES site.

Using 3' RACE it was possible to determine LN-BP18 TES. This method can only be applied to transcripts with a poly-A tail, suggesting that LN-BP18 transcripts are modified with such a tail. On the identified TES a polyadenylation signal could be identified, further supporting the presence of a poly-A tail on LN-BP18 transcripts.

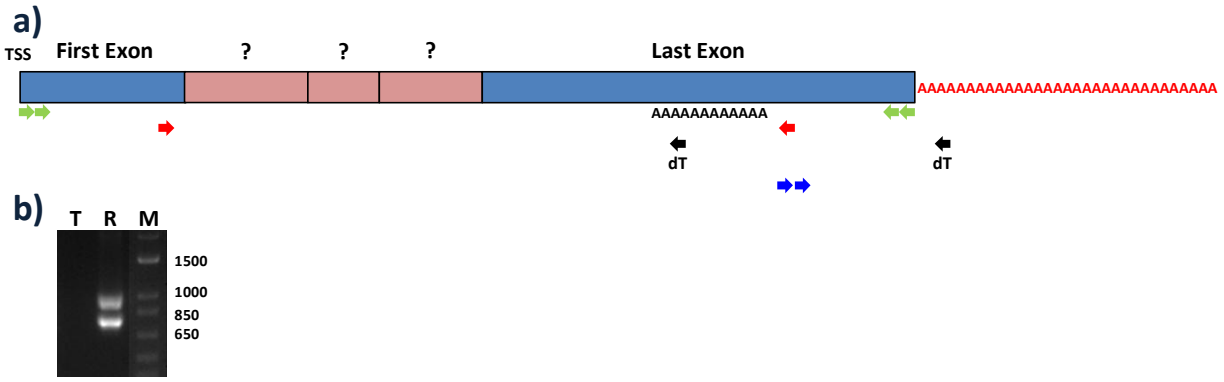


Figure 13 – LN-BP18 gene structure characterization and isoform identification
a) Schematic representation of LN-BP18 mature RNA. Known LN-BP18 exons, blue boxes; exons in unknown configurations resulting from alternative splicing, red boxes; poly-A tail, red "A" in tandem, stretch of adenines present at the last exons, black "A" in tandem; oligo dT used for cDNA generation from tailed RNA, black arrows; GSP used for isoform identification, green arrows; GSP used to identify LN-BP18 TES, blue arrows; GSP used on PCR from figure b), red arrows; **b)** PCR amplification of LN-BP18 transcripts using cDNA generated with either oligo dT (T) or random hexamers (R). GSP were located in the first exon and downstream of the identified stretch of adenines found during the first round of 3' RACE (red arrows, a)). Sizes of DNA marker used (M) are indicated in bp.

3.2.2. Identification of LN-BP18 isoforms

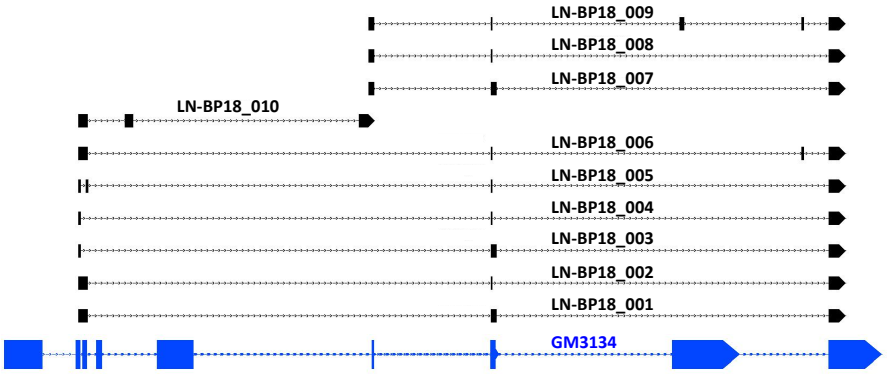
The first observations while studying this lncRNA was the presence of multiple isoforms. However, since this first observation, its known gene structure became more complex, with new exons and different TSSs identified. Additionally, according to the Gm3134 prediction based on data from different tissues and cell lines, up to 12 isoforms originated from this gene. To experimentally confirm the presence of different isoforms, cDNA was generated from WT and Med12^{null} ESC RNA, in order to isolate TSS2 transcripts and from forelimbs dissected from E11.5 WT embryos to amplify TSS1 transcripts. RNA was reverse transcribed with SuperScript III using a primer at the TES, random hexamers or oligo dT. In order to amplified the different LN-BP18 isoforms, primers near either TSS and at the TES were used and

the amplified fragments sequenced by Sanger sequencing. This led to the identification of 7 exons, revealing splice variants for three of the exons and nine different isoforms in total (Figure 14a). Fragment LN-BP18_010 does not represent a complete isoform, since it was identified using primers on exons 1 and 3. As such, no information regarding the remaining transcript was obtained. However, it was still included as it was the only instance where exon 2 and a long variant of exon 3 could be isolated.

Comparison of LN-BP identified isoforms with the Gm3134 predicted structure showed that although similar, there were major differences between the two gene annotations. Exon 3, 4, 5 and 7 of LN-BP18 overlapped some of the Gm3134 exons with some differences in the full exon length, more notorious for exons 5 and 7 of LN-BP18. The length of these two exons was of 170 and 600 bp for LN-BP18, while for Gm3134 were 2300 and 1800 bp, respectively. Exon 1 of LN-BP18 also overlapped some of the predicted exons. However the full length exon one was not part of the Gm3134 prediction. Finally, the identified exons 2 and 6 have not been predicted for Gm3134. On the other hand, no evidence was found for three of the Gm3134 exons.

The isoforms identified were more complex and diverse than the original two identified during the first experiments detecting LN-BP18 expression in ESCs (Figure 9). As such, the isoforms were evaluated for their coding potential, using again CPAT, in order to assess the probability of small peptides being encoded by this lncRNAs, as previously observed for other lncRNAs (Nelson et al. 2016). None of the identified isoforms possessed any coding potential, as evidence by the coding probability below the threshold (0.44) (Figure 14b). PhyloCSF is another method to identify genomic regions that are likely to represent protein coding sequences. This algorithm compares the alignment of 29 different mammal datasets identifying evolutionary signatures, such as high frequency of synonymous codon substitution, across all 6 reading frames. This algorithm has been successfully used to identify peptides originating from presumably non-coding transcripts, such as Gm34302 gene and its small 34aa long encoded peptide with roles in heart contractibility (Nelson et al. 2016). PhyloCSF data further confirmed the lack of coding potential of LN-BP18, as no evolutionary signatures were detected throughout the whole LN-BP18 locus in any of the reading frames. On the other hand, a clear signal from the coding sequence of Sall1 was detectable (Figure 14c). Due to the complex splicing pattern observed for LN-BP18, which included differently spliced variants for a subset of the exons, there was a chance that some of the identified isoforms resulted from artefacts or anomalous splicing events. In order to further confirm the observed splicing patterns, acceptor and donor splice sites of all exons were identified and for the vast majority matched the consensus sites previously observed in mouse transcripts (Senapathy et al. 1990).

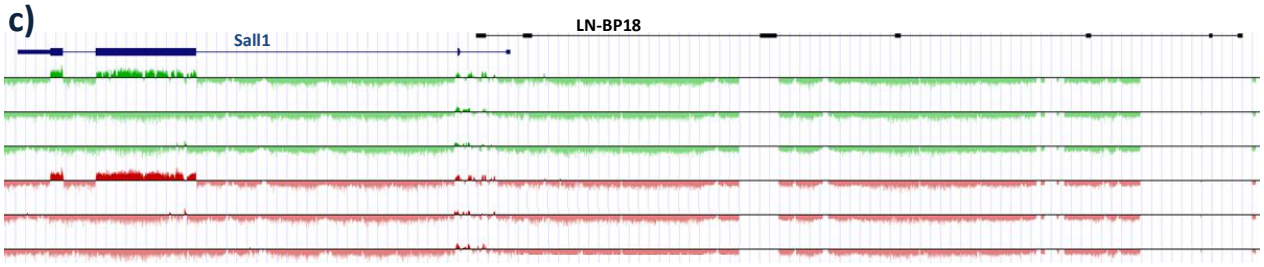
a)



b)

Isoform	Transcript Size	ORF Size	Coding probability
001	1156	276	0.09
002	1017	138	0.01
003	926	276	0.09
004	787	138	0.01
005	873	138	0.01
006	1118	138	0.01
007	1027	276	0.09
008	888	138	0.01
009	1161	138	0.01
010	1215	213	0.03

c)



d)

Acceptor/Donor sites

5'	Exon	3'
	Ex1sh	CCATTTCGTGAGTGT
CTGCCGGGCCTCCCTCCCGCCTGCCCGCCGCCTGGCTCGGA	Ex1	AGGAAAGGTTTGGGG
TGGTCTCCGGCTCTCTCCCGCATCTCGCTCTTCCCTTCTCT	Ex2	AACACACCAACGTGC
CACACACACACACACACACACACACACACACTCAGCCTTTA	Ex3	TCTGCAGGTCTGATG
TGTGCTCCGCCTCTCTTCCAGGGTAGTAAAAAGGCAGAGAC	Ex4sh	CTCCGGGGTGAGTGC
TGTGCTCCGCCTCTCTTCCAGGGTAGTAAAAAGGCAGAGAC	Ex4ln	CATTGAGGTTGTATG
AGACAGGATTCCTTCCTGCAGGTGTTTTTCAGCCCTACTAC	Ex5	CCTTCTGGTGAGTAG
ATAAATGTAATTTGTTTACAGAGACAAGCTTGGAGAGAGAG	Ex6	AAAGCAGGTAAAGTA
AGTCACCTTGTCCTTCTACAGAACTCTGCTGCCTGGTCCAG	Ex7	

Figure 14 - LN-BP18 isoforms and coding potential

a) LN-BP18 isoforms identified using primers on either TSS and on the TES. LN-BP18_010 resulted from amplification with primers on the first and third exons and as such was not a full transcript. Gm3134 predicted structure in blue is also evidenced; **b)** evaluation of coding potential of all 10 isoforms described using CPAT; **c)** PhyloCSF tracks from USCS browser for all 6 reading frames. In green, positive strand; in red, negative strand; **d)** acceptor and donor splice sites from all LN-BP18 identified exons. In red are the border nucleotides between exon and intron that correspond to the known mouse consensus acceptor and donor sites. “Ex1sh” represents the exon 1 found on isoforms 003 and 004; “Ex1” represents the second splice variant of exon 1 in isoform 005. While the 5’ sequence of this exon is exclusive for this splice variant, the 3’ end is common to isoforms 001, 002 and 006. 3’ sequence of “Ex2” matches the one found on fragment 010. For splice variants of exon 3 found on isoforms 007-009, no 3’ sequence analysis was performed since this corresponded to the second transcription start site. “Ex4sh” is found on isoforms 002,004,005,006,008 and 009 and “Ex4ln” in isoforms 001,003,007.

The only exceptions found were a short variant of exon 1, exon 2 and the long exon 3, found only on either LN-BP18_010 or LN-BP18_005 (Figure 14d), revealing that these might be due to aberrant splicing variants.

Additionally, BLAST searches were performed with the full mature transcript sequence, the DNA sequence of the biggest open reading frame (ORF) found within each isoform or the amino acid encoded by it, however no significantly similar sequences were identified. Despite no similar transcripts was found on other species, an uncharacterized lncRNA divergent from Sall1 is also predicted in human. The five transcripts predicted for this lncRNA, termed AC087564.1, were aligned against LN-BP18 isoforms and identity between all transcripts calculated, with sequences for Fendrr and Hotair murine genes also included as controls. The observed percentage of identity between LN-BP18 and AC087564.1 transcripts was around 30%, lower than all alignments performed (Figure 15). Identity between these two lncRNAs was similar to the identity with the used control genes. This revealed that there was no sequence conservation between LN-BP18 and AC087564.1. LN-BP18 was slightly more similar to Hotair and Fendrr transcripts, possible because all of these transcripts were found in mouse while AC087564.1 is a human gene. A low conservation of transcripts between species is not exclusive of conserved function, as observed for Gas5, a lncRNA that shows poor sequence conservation but its function as a regulator for self-renewal and pluripotency is conserved (Tu et al. 2018). The fact that AC087564.1 is divergent of Sall1, despite Sall1 location in a different chromosome (chromosome 8 in mouse and chromosome 16 in human), strongly suggested a similar function as LN-BP18. As such, it would be interesting to verify if any function discovered for LN-BP18 is conserved for AC087564.1.

Amplification by PCR of LN-BP18 using both ESC and tissue derived cDNA allowed to identify nine full isoforms. Different splice variants were observed for three of the seven exons of LN-BP18, revealing a complex splicing pattern for this lncRNA.

AC087564.1_002	100	54	38	36	33	29	28	29	29	27	27	28	27	29	27	30	32	32
AC087564.1_004	54	100	37	37	35	31	28	30	31	30	30	30	30	30	31	35	33	32
AC087564.1_005	38	37	100	44	41	35	33	32	33	33	33	33	33	33	33	36	36	37
AC087564.1_001	36	37	44	100	58	32	32	31	32	32	33	32	34	33	34	32	34	35
AC087564.1_003	33	35	41	58	100	32	28	26	28	30	28	27	29	27	27		35	34
Hotair	29	31	35	32	32	100	40	40	41	38	39	41	40	40	41	39	40	41
BP18_009	28	28	33	32	28	40	100	92	100	84	77	78	77	73	78	47	42	41
BP18_007	29	30	32	31	26	40	92	100	100	91	85	89	77	80	80	47	41	41
BP18_008	29	31	33	32	28	41	100	100	100	100	96	96	83	83	83	47	42	43
BP18_004	27	30	33	32	30	38	84	91	100	100	93	84	86	84	94	47	42	41
BP18_005	27	30	33	33	28	39	77	85	96	93	100	87	81	81	88	47	43	43
BP18_003	28	30	33	32	27	41	78	89	96	84	87	100	85	93	91	47	44	43
BP18_006	27	30	33	34	29	40	77	77	83	86	81	85	100	94	100	47	44	44
BP18_001	29	30	33	33	27	40	73	80	83	84	81	93	94	100	100	47	44	43
BP18_002	27	31	33	34	27	41	78	80	83	94	88	91	100	100	100	47	45	44
Fendrr_001	30	35	36	32		39	47	47	47	47	47	47	47	47	47	100	51	74
Fendrr_002	32	33	36	34	35	40	42	41	42	42	43	44	44	44	45	51	100	82
Fendrr_003	32	32	37	35	34	41	41	41	43	41	43	43	44	43	44	74	82	100

Figure 15 – Sequence similarity between LN-BP18 and the human AC087564.1
Percentage of sequence identity between the different isoforms of LN-BP18 and AC087564.1. Green denotes a high identity between transcripts and red a low identity.

3.2.3. Expression analysis of LN-BP18

Having identified this new lncRNA gene structure, the normal expression pattern was more finely dissected, in order to complement the whole mount *in situ* hybridization (WISH) and tissue RNA-seq data analysed in section 3.1.4. The cellular localization of a lncRNA can provide hints of its functions, e.g., a lncRNA which recruits repressor complexes to target genes will be enriched in the nucleus, while lncRNAs involved in mRNA stabilization and translation will be enriched in the cytosol. To determine LN-BP18 cellular localization, nuclear and cytoplasmic fractions were obtained for WT ESCs. LN-BP18 levels were then quantified by real-time quantitative PCR (qPCR), which allows performing a relative quantification of gene expression. Using a dye that binds to double-stranded DNA, the fluorescent signal reflects the amount of double-stranded PCR product that is generated during the reaction.

Analysing the isoforms identified for LN-BP18, exon 4 was always the second exon of the transcripts with either exon 1 or 3 as the first exon, depending from which TSS the transcript originated from. As such, the quantification of the splicing between exon 1 and exon 4 and between exon 3 and exon 4 was

used to assess transcription from TSS1 and TSS2, respectively. Since in the majority of isoforms originating from either TSS, exon 4 was proceeded by exon 7, quantification of this splicing was use as a measure of the overall LN-BP18 expression. Since for exon 1 there were two different variants, two different TSS1 quantifications could be used, matching what is found in LN-BP18_003 or in LN-BP18_001. The same is observed for exon 4, which possessed two variants and as such, for assessment of LN-BP18 overall levels, two different splicings could be quantified, the one found on LN-BP18_001 and LN-BP18_002. An average of the relative quantification for both splicing events from each TSS was calculated.

Cellular localization was evaluated as a ratio between the nuclear and cytoplasmic level, as previously applied to determine the nuclear location of the lncRNA Fendrr (Grote et al. 2013). Western blots using antibodies against histone 3 (H3) and Gapdh, a nuclear and cytoplasmic marker, respectively, confirmed the efficiency of fractionation, with the majority of H3 detected in the nuclear fraction while no Gapdh was detectable in the cytoplasm (Figure 16a). Expression quantification by qPCR of LN-BP18 and different controls was performed on these fractions. Xist is a key factor during XCI and as such is enriched in the nucleus. Nanog and Sox2 are highly expressed in ESC and their mRNA needs to be translated in order for the TFs to maintain the pluripotency network and so they are enriched in the cytoplasm, where the translation machinery is. With this approach, no clear enrichment could be found in any of the two fractions for LN-BP18 (Figure 16b).

WISH data revealed expression of LN-BP18 during embryonic development in forelimbs, hind limbs and in the caudal end (Figure 10b), which was corroborated by tissues transcriptome data (Figure 10c). In order to determine the *in vivo* expression of LN-BP18 while distinguishing both TSSs, E11.5 WT embryos were dissected into 5 different tissues, according to the domains observed form the WISH data: forelimbs, hind limbs, tail, head and then the remaining tissues as a single sample designated as “rest”.

Quantification by qPCR on these tissues confirmed an enrichment of LN-BP18 in limbs and tail, with expression 4-fold higher in these tissues. No significant difference was observed between the expression levels in the head and in the remaining tissues (rest). Additionally, isoforms originating from the TSS2 were not detectable in any of the tissues (Figure 16d). The RNA-seq data revealed a misregulation of LN-BP18 when Med12 was depleted in ESCs. To verify if other Med12 mutants, which did not alter Med12 levels, could affect this lncRNA expression, a Med12-Opitz ESC mutant, previously generated by Dr. Heinrich Schrewe using male cells, which expressed the Med12 R961W mutation associated with FG syndrome, was used. Med12 levels were unaffected by the mutation and previous studies have shown that it does not affect Med12 interaction with Mediator (Lai et al. 2013).

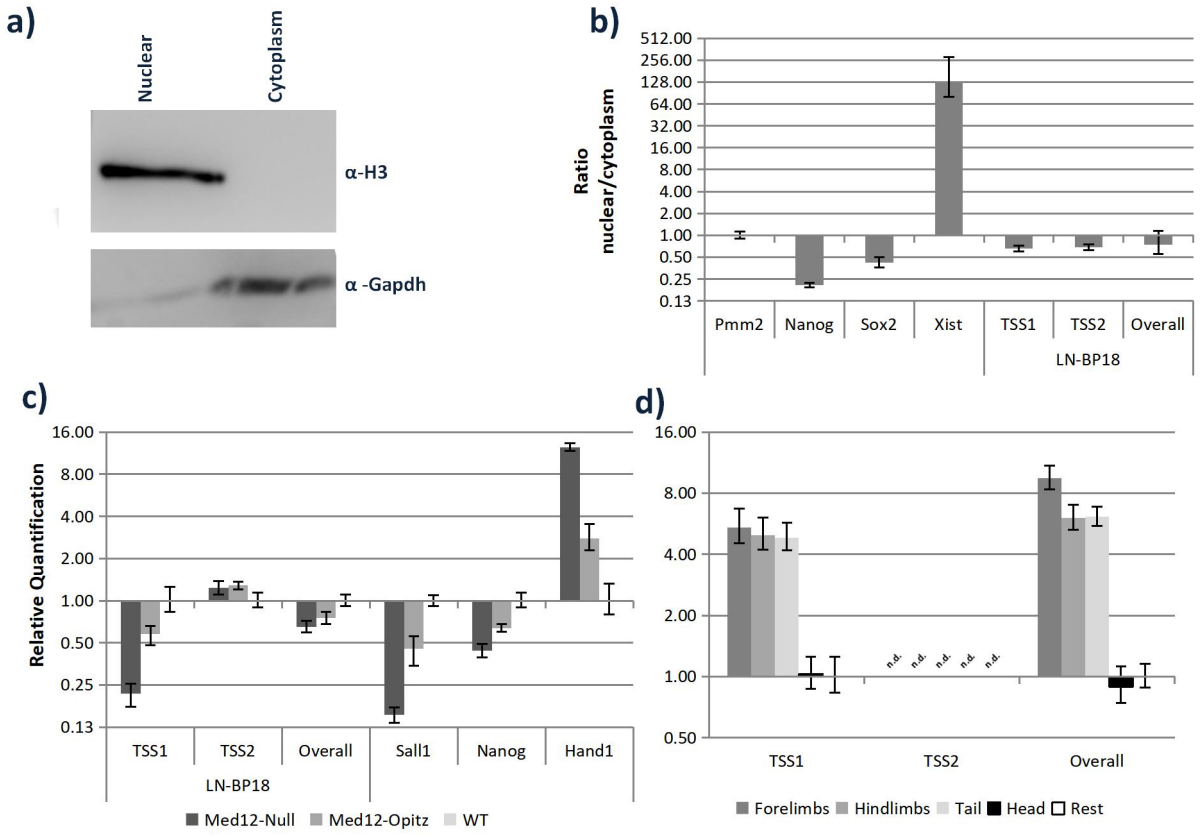


Figure 16 - LN-BP18 *in vivo* expression
a) Western blot with proteins extracted from nuclear and cytoplasmic fraction of WT ESCs using anti-H3 and anti-Gapdh antibodies; **b)** quantification of LN-BP18 and both nuclear and cytoplasmic markers using qPCR and expressed as a ratio between nuclear and cytoplasmic fractions normalized to Pmm2; **c)** Relative quantification of overall (Ex4sh-7 and Ex4ln-7), TSS1 (Ex1sh-4 and Ex1ln-4) and TSS2 (Ex3-4) LN-NP18 transcripts and possible related genes in WT, Med12-null and Med12-Opitz ESCs using qPCR. Pmm2 was used as internal control and expression values normalized to WT sample; **d)** Relative quantification of LN-BP18 in different tissues from E11.5 embryos. Expression was normalized to the sample rest; n.d., expression not detectable (n=3, SD).

Analysis by qPCR was used to quantify expression of LN-BP18 and Sall1, together with other markers, in Med12^{null} and Med12-Opitz mutants.LN-BP18 TSS1 and Sall1 were downregulated in both mutants, with expression 4-fold lower in the Med12^{null} mutant (Figure 16c), with a more moderated effect observed Med12-Opitz cells, with a 2 fold decrease in the expression of both genes. Despite the observed effect on LN-BP18 TSS1 expression, TSS2 and overall levels remained unchanged. Analysis of mutant ESCs where LN-BP18 was misregulated is useful since it might indicate possible contexts where its function might be relevant. However, to determine its function, mutants with targeted perturbations on LN-BP18 expression were necessary.

3.3. Generation of LN-BP18 mutants mESC

Having characterized the locus and transcripts of this novel lncRNA, the next important step would be a functional analysis of this gene. In contrast to protein coding genes, where disruption of the open reading frame leads to its deficiency, non-coding genes need different strategies, such as inactivation of the TSS, insertion transcription stop signals or insertion of the coding sequence of a reporter gene. All of the three strategies mentioned were used to inactivate LN-BP18 expression in mouse embryonic stem cells.

3.3.1. Beta-galactosidase reporter line generation

As it was observed by whole mount in situ hybridization (WISH), LN-BP18 was lowly expressed in vivo. As such, a LN-BP18-β-galactosidase (LN-BP18-β-gal) reporter mutant was generated, since this reporter presents a higher sensitivity than the previously used WISH. One strategy already mentioned to study the function of lncRNAs is the replacement of the whole locus by a β-gal cassette, leaving the regulatory region intact.

This method allowed a simultaneous perturbation of the gene and the generation of a reporter line (Sauvageau et al. 2013). However, since LN-BP18 was expressed in an antisense overlapping orientation of Sall1, deletion of this locus would also disrupt the Sall1 gene structure. As such, instead of replacing the whole locus, a β-gal cassette which was codon optimized for mouse, followed by a triple pA signal was inserted into exon 4 of LN-BP18. Exon 4 was targeted for insertion since it was present in all identified isoforms (Figure 14a) originating from either TSS, allowing targeting transcripts regardless from which TSS they were transcribed. Since the inserted cassette contained no promoter and the regulatory region of LN-BP18 was left intact, the reporter expression should be driven by the same regulators of LN-BP18 and expression observed in the same cells as the lncRNA. Additionally, due to the triple pA signal inserted, transcription was expected to terminate after the reporter cassette. Insertion of β-gal and triple pA signal was achieved through homologous recombination and in order to increase recombination frequency, a double strand break (DSB) was induced on exon 4 through CRISPR-Cas9. Homology arms of 120 bp matching the immediately upstream (5’ homology) and downstream (3’ homology) genomic sequence of the DSB were used (Figure 17a). Previous data by other groups have shown that homology arms of 120 bp are sufficient to ensure proper homologous recombination (data not shown).

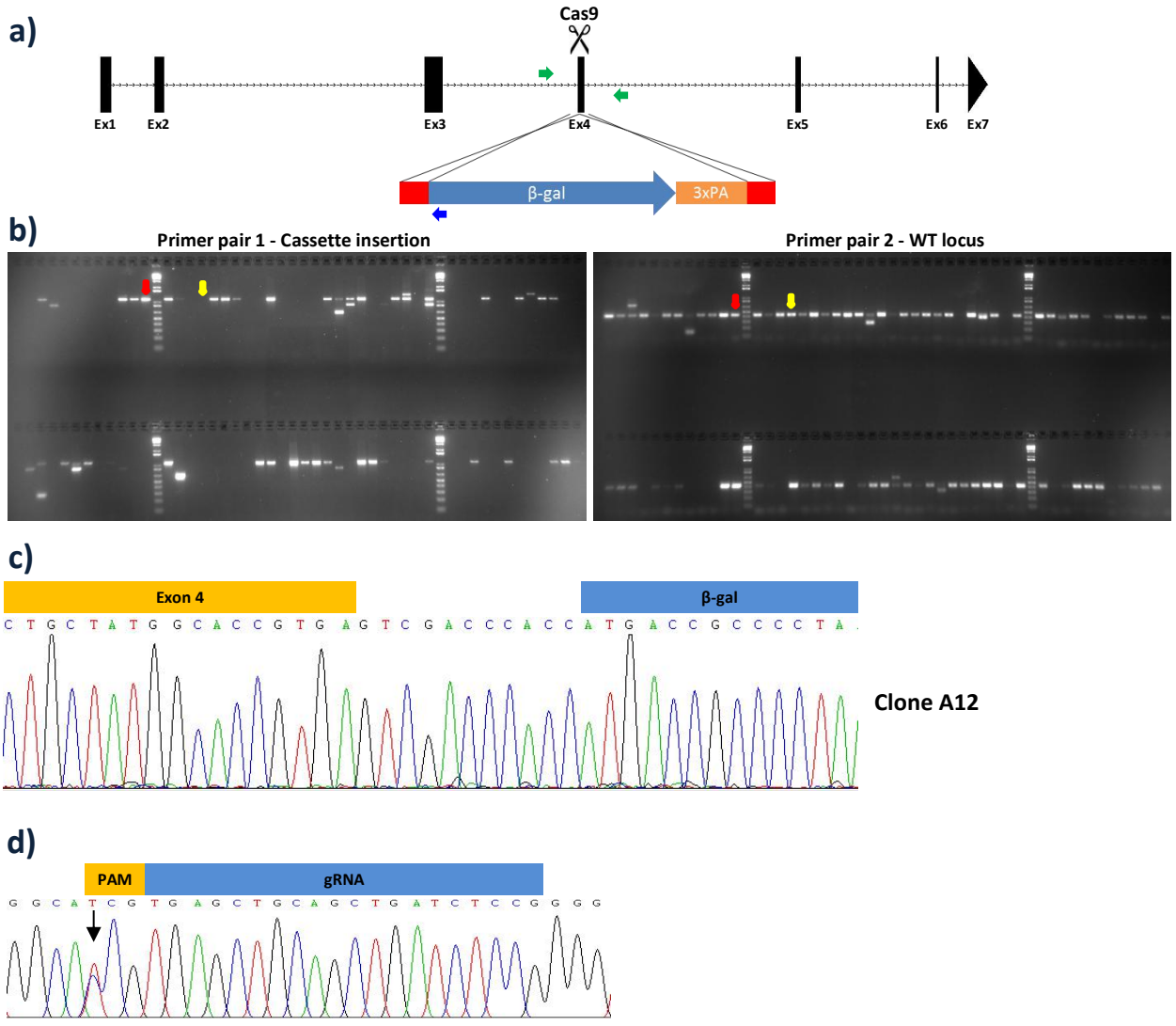


Figure 17 - CRISPR-Cas9 strategy used for β-gal knock-in

a) Schematic representation of strategy used to generate LN-BP18 β-gal reporter line**.** Homology arms of 120 bp, red boxes; primers used for WT clones screening by PCR, green arrows; primer used for screening the cassette insertion, blue arrow; **b)** Screening of 96 clones using PCR. For primer pair 1, only clones with insertion of β-gal cassette in the correct LN-BP18 locus should give a signal. With primer pair 2, clones with at least 1 WT locus should generate a signal. Example of heterozygous clones, red arrow; example of WT clone, yellow arrows; **c)** chromatogram of 5' border of β-gal insertion for clone A12 evidencing correct insertion; **d)** Sequencing of Cas9 target region for double strand break generation during homologous recombination, with a polymorphism evident on the PAM sequence of the G4 hybrid cell used (black arrow).

Through PCR, the β-gal cassette and the triple pA signal were amplified from a vector kindly gifted by Dr. Frederik Koch (Max Planck Institute for Molecular Genetics, Berlin) and from pCCALL vector, respectively and homology arms amplified from G4 WT genomic DNA. These four elements were cloned into pBluescript SK vector as a single cassette consisting of 5' homology-β-gal-3xpA-3' homology (Figure 17a). This whole cassette was flanked by the same genomic region and the protospacer adjacent motif (PAM) sequence targeted by Cas9 which allowed the nuclease do excise the cassette from the pBluescript upon ESC transformation.

Transformation of ESC was performed as described in section 2.4.2. Briefly, 3.0x105 G4 129S6/C57BL6 hybrid cells (George et al. 2007) were seeded on a monolayer of feeder cells and cultured for 24h. To transform cells, Lipofectamin 2000 was used according to manufacturer's instruction using 2 µg of CRISPR-Cas9 vector and 6 µg of donor vector. Cells were cultured for seven days using puromycin selection for the first three days. 96 colonies were picked into a 96 well plate and cultured to confluency. Afterwards, 1/3 of cells were frozen as a "Master Plate" and the remaining 2/3 were cultured in two 96well plates until confluency and used to extract DNA according to the protocol described in section2.4.5. The clones were screened by PCR using the strategy indicated in Figure 17a. For WT screening, primers flanking exon 4 were used (Figure 17a, green arrows), while for successful cassette insertion, the forward primer for WT screen and a reverse primer in the β-gal cassette were used (Figure 17a, left green arrow and blue arrow, respectively). With this screening strategy, WT clones should generate signal only when using primer pair 2 (Figure 17b, yellow arrow), heterozygous clones, with insertion in only one of the alleles, should generate signal with both pairs (Figure 17b, red arrow) and homozygous clones, with insertion in both alleles should give signal only with pair 1.

Transformed clones were screened for insertion of reporter cassette using the mentioned PCR strategy. Correct transformation of these clones was further confirmed by Sanger sequencing of amplified PCR fragments (Figure 17c). Curiously, all the successfully transformed G4 clones were heterozygous, with no homozygous insertion obtained despite the high efficiency of transformation (>30% of screened clones). When designing guide RNAs for the Cas9 nuclease, the reference mouse genome used was generated from a C57Bl6 strain. However the G4 cells used for the transformation were derived from a F1 hybrid cross between C57Bl6 and 129S6, which meant that the maternal inherited chromosomes were derived from 129S6 strain while paternal inherited chromosomes were derived from C57Bl6 (George et al. 2007). Public genomic sequencing data revealed a mutation on the PAM sequence used by the designed guide RNA on the 129S6 allele. The mutation, which rendered targeting of this allele by Cas9 nuclease impossible, was confirmed by Sanger sequencing (Figure 17d).

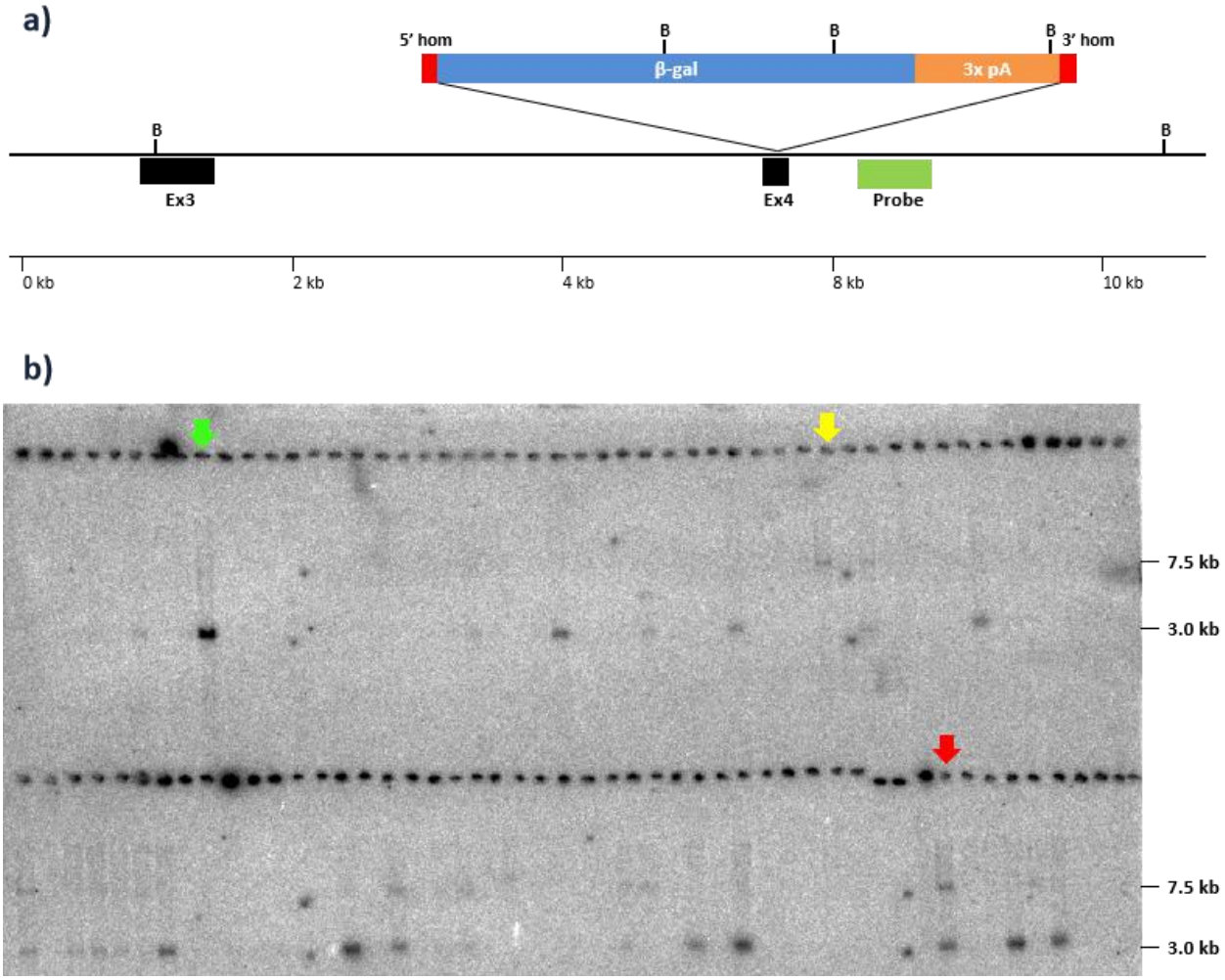


Figure 18 - Screening of JM8-LN-BP18-β-gal clones
a) Schematic representation of β-gal cassette insertion in LN-BP18 exon 4. Due to BamHI restriction sites present in the cassette, a smaller fragment is detected by southern blot when using the indicated probe. In the WT locus, a fragment of 7.5kb is detected, while on the transformed allele, a 3 kb fragment is detected. Southern probe used to screen clones, green box; BamHI restriction sites in the LN-BP18 locus and β-gal cassette, bold "B"; **b)** southern blot result from 96 screened clones. Example of homozygous clone, green arrow; heterozygous clone, red arrow; WT clone, yellow arrow.

As such, only the paternal allele could be edited, which explained the generation of only heterozygous clones in G4 cells.

Since the established protocol was successful in transforming the C57Bl6 allele, JM8 cells derived from C57BL/6N background were transformed in order to generate not only heterozygous but also homozygous reporter cells. The same protocol used to transform, select and screen transformed G4 mutants was repeated for the JM8 cells. However, the PCR screen was not very informative since unspecific amplicons were obtained and results were not concordant between different screens

performed (data not shown). As such, another screening method was used and successfully transformed cells were screen by Southern blot, by digesting genomic DNA with BamHI and using an external probe of the 3' homology arm. Positive recombination led to the detection of a 3.0 KB fragment, while on WT allele a 7.5 kb fragment was detected. 192 JM8 clones were picked and treated as described for G4 clones. Extracted DNA was digested with BamHI and resulting fragments size separated in an agarose gel by electrophoresis. DNA was then transferred to a nylon membrane and a radioactively labelled probe was used to detect the shift in size of the fragment, resulting from the cassette insertion (Figure 18a, green box). With this method, multiple potential heterozygous and homozygous JM8 mutant cells were identified (Figure 18b).

3.3.1.1. LN-BP18 expression analysis in β-gal reporter mutant cells

Having successfully edited the LN-BP18 locus in ESCs, the next step was to evaluate the activity of the β-gal reporter. As described in section 2.4.4, for each plate of picked colonies, two distinct DNA plates were prepared. While one plate was used for preparing DNA, which allowed screening clones with southern blot, the other was used for X-gal staining, which allowed assessing activity of the β-gal reporter. Briefly, G4 cells were washed with cell-culture grade D-PBS, fixed with 4% PFA/PBS for 5min and incubated with staining solution at 37°C o/n. Cells were then washed again with cell-culture grade D-PBS and re-fixed with 4% PFA/PBS before detecting the resulting signal. However, no reporter signal was detectable on any of the G4 heterozygous clones (data not shown). Since LN-BP18 expression was very low in ESCs it is possible that the low expression was the cause for the lack of reporter activity. As such, a different condition was used to test the activity of the reporter cassette.

In a recent paper, fluorescence-activated cell sorting (FACS) was used to isolate neuro-mesodermal progenitors cells (NMPs) that co-expressed Brachyury (T) and Sox2 reporters. Sorted cells were used to generate transcriptome data and interestingly, LN-BP18 was found to be expressed in T+, Sox2+ and double T+/Sox2+ cells (Figure 19a) (Koch et al. 2017). Based on this finding, a five days in vitro differentiation protocol (Figure 19b) previously published was used to differentiate ESCs into NMPs (Gouti et al. 2014). At day 3 of the differentiation protocol the differentiated cells were similar to the T+/Sox2+ double positive cells found In vivo at the caudal end of E8.5 embryos and at day 5 they were similar to the T+ cells (Gouti et al. 2014, Koch et al. 2017). Two different G4 derived clones (C5 and D1), heterozygous for the β-gal cassette insertion, were differentiated following this protocol. Briefly, clones were expanded, feeder depleted and seeded into a 12 well plate coated with synthemax.

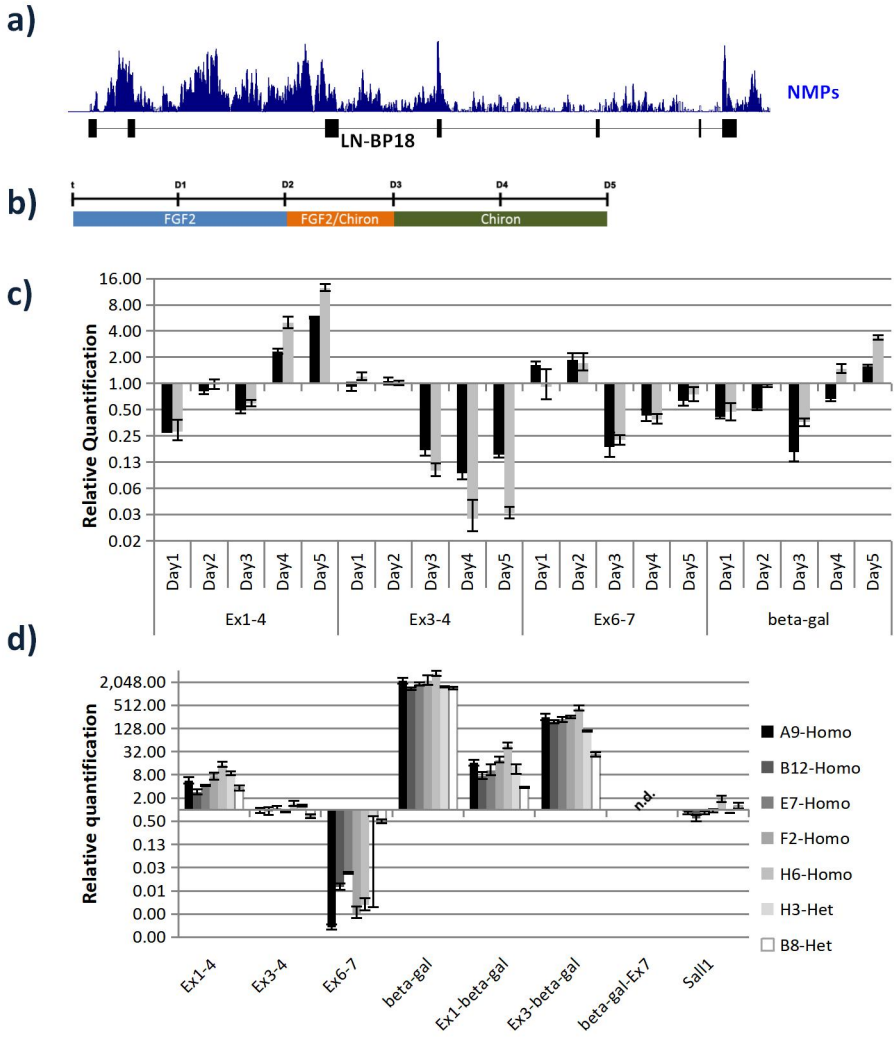


Figure 19 - Characterization of LN-BP18 β-gal reporter ESCs mutants

a) Genome browser track showing RNA-seq data of $T^+/Sox2^+$ double positive cells FACS cells from caudal end E8.5 embryos showing LN-BP18 expression on these cells. Similar expression was also observed on single T^+ or $Sox2^+$ positive cells (Koch et al. 2017); **b)** differentiation protocol scheme followed to differentiate ESCs into mesodermal. Cells at day 3 of differentiation are similar to the $T^+/Sox2^+$ double positive found at caudal end of E8.5 embryos (Gouti et al. 2014); **c)** LN-BP18 and β-gal cassette expression quantified by qPCR through the differentiation course for two G4 LN-BP18-β-gal heterozygous clones. Expression was normalized to ESCs cultured for 24h with ES+LIF medium; **d)** quantification of LN-BP18 and β-gal expression on C57Bl6-LN-BP18-β-gal heterozygous and homozygous clones with qpCR. Expression was normalized to WT control. Since no β-gal transcripts could be detected for the WT control, the minimum detectable values was attributed in order to obtain a relative value for the mutants. On both qPCRs presented, Pmm2 was used as an internal control; n.d., expression not detectable; (n=3, SD).

For each clone 12 wells were prepared and after every 24 h of culture, a well of each clone was used for X-gal staining and another for RNA extraction for qPCR quantification. Cells cultured for 24h in the same conditions but using ES+LIF medium were used as the control to which expression levels were normalized. X-gal staining at each of the five time points analysed yielded no detectable signal (data not shown).Despite the lack of signal from the X-gal staining, by qPCR it was possible to observe an upregulation of the cassette expression by day 5 (Figure 19d). Additionally, the qPCR data revealed that, upon differentiation the expression of LN-BP18 isoforms originating from TSS2 (Figure 19d, Ex3-4) were decreased. Simultaneously, transcription of isoforms originating from TSS1 increased throughout the course of differentiation (Figure 19d, Ex1-4).

Expression analysis by qPCR of LN-BP18 and β-gal cassette expression levels on C57Bl6 reporter ESCs revealed that transcription of isoforms originating from TSS2 were not affected. In heterozygous clones there was a 2-fold downregulation of transcripts containing exons downstream of the targeted exon 4, consistent with premature termination of one of the alleles. In turn, in homozygous clones, this downregulation was stronger, further demonstrating the efficiency of the stop cassette. The β-gal cassette was transcribed as part of the LN-BP18 RNA, since transcripts containing either exon 1 or 3 followed by the β-gal (Ex1-beta-gal and Ex3-beta-gal, respectively) could be detected (Figure 19d). No transcript was detected containing both the β-gal and exon 7 (beta-gal-ex7) in any of the mutants, revealing an abortion of transcription after the triple pA signal inserted.

Analysis of Sall1 expression additionally confirmed that the insertion did not affect the normal Sall1 levels (Figure 19d), confirming that the inserted cassette affected the expression of only LN-BP18. As such, any possible misregulated genes or phenotypes detected using these cells should be driven by LN-BP18 deficiency. However, expression from TSS1 transcripts was upregulated in all mutant clones. Despite their upregulation, TSS1 transcripts were on average 13-fold lower expressed than TSS2 transcripts (data not shown).

3.3.1.2. Analysis of LN-BP18-β-gal embryos

Despite no activity of the β-gal reporter could be detected in cultured mutant cells or in *in vitro* differentiated G4 heterozygous clones, expression of the cassette could be detected in all successfully transformed clones (Figure 19d). As such, different G4 heterozygous clones were used to generate embryos by tetraploid complementation assay. Preliminary data from one E11.5 embryo generated from the ESC clone C5 has revealed that the construct generated is functional, with the expression of β-gal

driven by the intact LN-BP18 regulatory region. A strong signal was obtained in the mesenchyme of forelimbs and hind limbs buds as well as in the caudal end (Figure 20), confirming the expression domains identified through whole mount *in situ* hybridization (WISH) (Figure 10b).

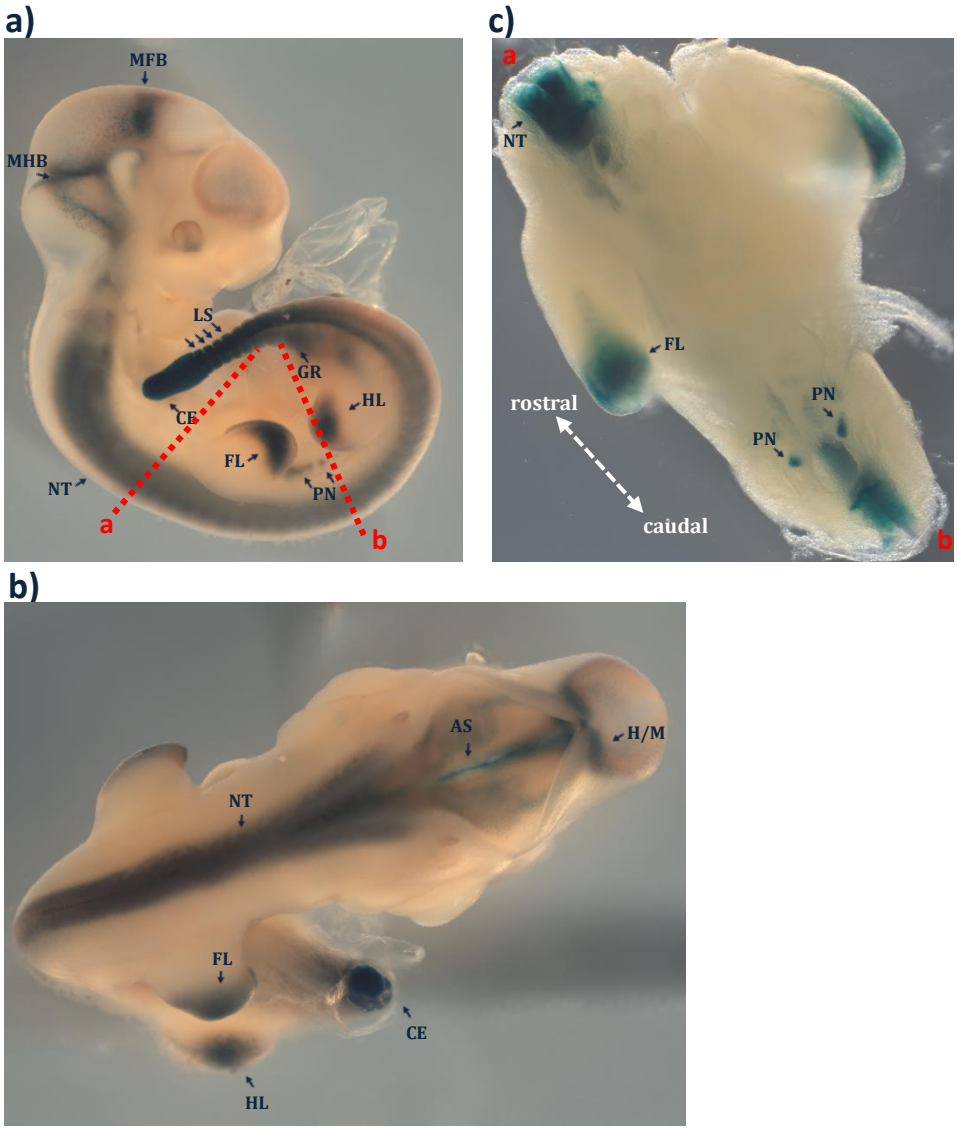


Figure 20 - X-Gal staining of E11.5 embryo generated from a G4 derived LN-BP18-β-Gal heterozygous clone

a) Lateral view. Embryo was dissected using the cutting planes indicated with red dash to generate figure c); **b)** dorsal view of the embryo; **c)** ventral view of the dissected embryo. A strong reporter activity was obtained in the: caudal end (CE); late somites (LS); mesenchyme of forelimbs (FL) and hind limbs (HL) bud; neural tube (NT); pronephros (PN); in the midbrain-hindbrain boundary (MHB) and midbrain-forebrain boundary (MFB) and axial structures (AS) that could represent notochord, floor plate or both. A weaker and more diffused signal was detected in the genital region (GR) and also at the heart (not shown).

A strong signal was also observed in the recently formed somites, with a weaker signal detected in older somites. A specific and strong was detectable in the isthmus, which separates hindbrain from midbrain and also in the boundary between midbrain and forebrain, confirming the expression in these three sections of the brain, as was indicated from the embryonic tissue RNA-seq data (Figure 10c). This RNA-seq data also indicates that the higher expression of LN-BP18 is found in embryonic kidneys, an indication supported by the β-gal reporter, where expression is found in the pronephros, which correspond to the first stage of kidney development. A clear signal was also observed on the dorsal side of the embryo, in the neural tube, which was identified from the RNA-seq as one of the tissues with higher expression of LN-BP18 (Figure 10c).. A clear signal was also observed in axial structures, which could be either notochord or floor plate or even both. A weaker expression was observed for other regions such as the heart and genital region. The observed speckled pattern, especially in regions with low signal, resulted from a nuclear localization signal on the β-gal cassette.

LN-BP18 and Sall1 have been found to be co-expressed, as observed with WISH, RNA-seq and now with a reporter construct for LN-BP18, since Sall1 was previously reported to be expressed in the same tissues as the observed signal from the LN-BP18-β-gal reporter embryo (Buck et al. 2001).

3.3.2. LN-BP18 gene inactivation by excision of its transcription start sites (TSS)

Although the β-gal reporter line generated for LN-BP18 affected the normal transcription of the lncRNA, due to the stop cassette (3x pA) inserted downstream of the β-gal sequence, the first four exons could still be expressed (Figure 19d). In the case of the isoform LN-BP18_001 (Figure 14a), this corresponded to 40% of the total isoform. Therefore, to inactivate the LN-BP18 gene, a different approach using CRISPR-Cas9 to completely excise either of the TSS, was used. To excise the desired regions, two different guides were designed for each excision, flanking the targeted region. Cas9 generated DSB at this two sites and repair of these breaks by non-homologous end joining resulted, in some cells, in the excision of the whole region within the two breaks. Due to positioning of TSS1 within Sall1 intron, this TSS was excised using guides that kept the Sall1 donor and acceptor splicing sites intact. For TSS2, a 2.3kb region was deleted, including the full exon 3 (Figure 21a). A double deletion could also be performed in order to completely abolish LN-BP18 expression, however, individual deletion of each allowed assessing the impact of TSS on LN-BP18 expression but also of potential interactors, such as Sall1.

As for the generation of the LN-BP18 reporter mutants, G4 cells were cultured, transformed, selected, frozen and DNA extracted using the same protocol (section 0). In the generation of TSS-KO mutants, 4 µg

of each CRISPR-Cas9 vector was used to transform G4 cells and for each transformation, 96 clones were picked and screened by PCR using primers flanking the deleted region and flanking the 5' induced DSB (Figure 21c and d, red arrows). With the used PCR conditions (Supplementary Table 4), primers flanking the excised regions should not generate any signal due to the large size of the flanked region. However, upon excision of the TSS region, the fragment that these primers amplified was smaller, allowing its amplification and detection by PCR (Figure 21b, "KO signal"). The use of primers flanking one of the excision borders generated signal only on the WT allele, since upon excision of TSS, the binding site for the primer within this regions was lost (Figure 21b,"WT signal").

The PCR used to screen TSS1-KO mutants generated unspecific amplicons and included in some cases contamination by feeders DNA that complicated discrimination of heterozygous and homozygous clones. As such, for promising candidates, clones were expanded, feeder depleted as described in section 2.4.7 and DNA extracted. This DNA free of feeder contamination was used to screen TSS1-KO clones (Figure 21b, left pictures). For TSS2-KO screening results, clean results were observed without any feeder contamination (Figure 21b, right pictures). Successful excision of the desired genomic region was confirmed by Sanger sequencing in screened clones (Figure 21c, d). Although for some clones, sequencing results demonstrated break points precisely at the induced DSB (e.g. TSS1 clone 4E, Figure 21c), in others, slightly different break points were observed, such as for clone TSS2 3C (Figure 21d). For this clone, the 5' break point was 2 bp upstream of the DSB, while the 3' break point was 4 bp upstream of the DSB.

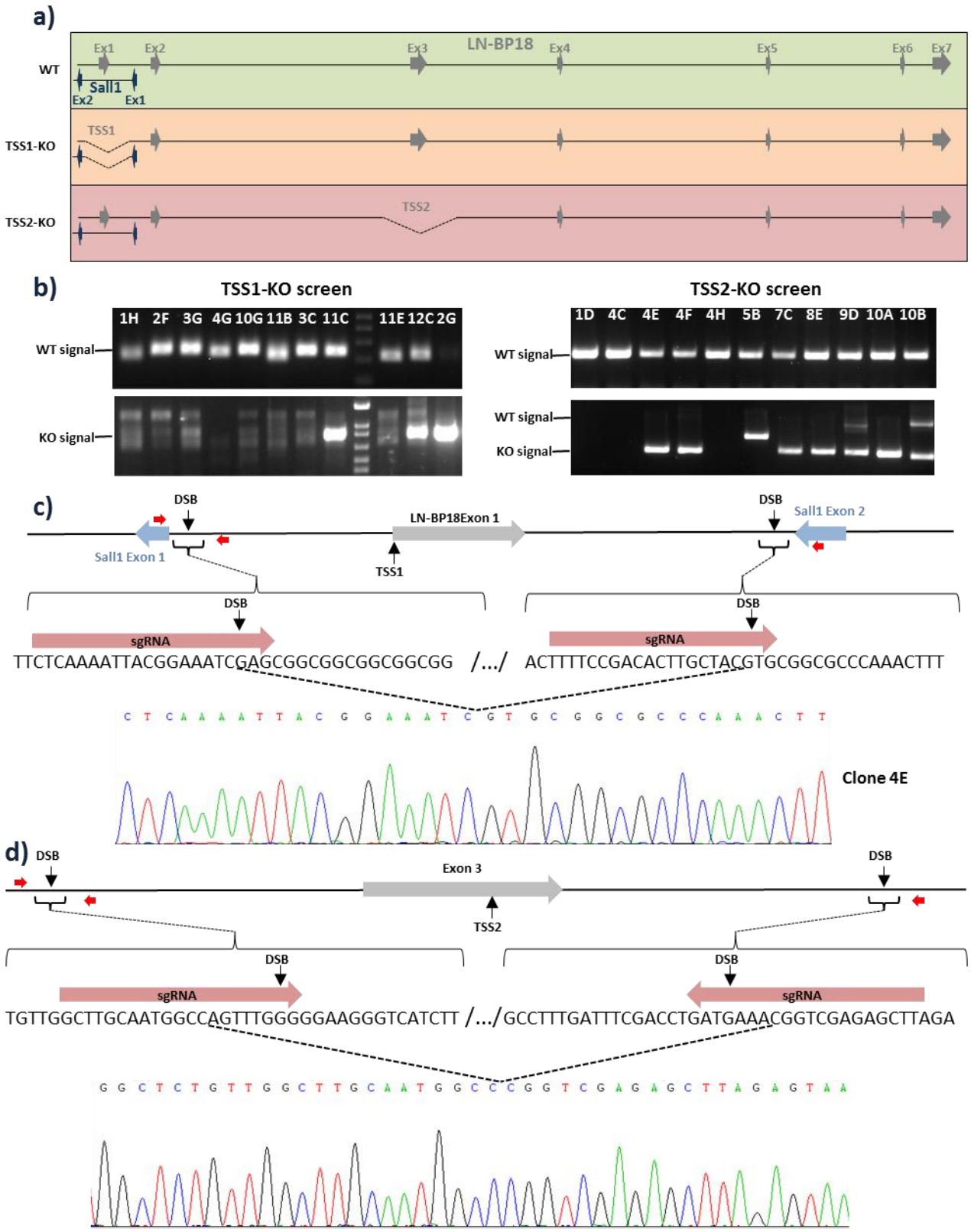


Figure 21 - Excision of LN-BP18 TSS in mouse embryonic stem cells

a) Schematic representation of LN-BP18 and Sall1 locus, evidencing the excisions performed to remove either of LN-BP18 TSS; **b)** PCR screen of TSS mutant clones. PCR used to detect excision of TSS1 (KO signal) generated several fainter unspecific bands. A clear and strong signal was obtained for clone 11C, 12C and 2G, **c)** and **d)** schematic representation of double strand breaks (DSB) performed by Cas9 in order to excise TSS1 or TSS2. The chromatogram confirms that the clones had their respective TSS removed, demonstrating either **c)** the predicted KO locus, with the break points precisely at the induced DSB or **d)** different break points from the predicted ones but still very close to the induced DSB. Primers used for screening potential mutant clones, red arrows.

3.3.2.1. Characterization of LN-BP18 expression in TSS-KO mutant ES cells

For successfully transformed clones, quantification of LN-BP18 expression by qPCR was performed. For this, clones were expanded, feeder depleted and their RNA extracted as described before. 5 µg of RNA were reversed transcribed into cDNA and used for qPCR quantification using splice forms specific primer sets. To normalize the amount of input material on each sample, the housekeeping gene Pmm2 was used and all samples were quantified in three technical replicates. Results from qPCR revealed that for five of the eight LN-BP18-TSS1-KO homozygous mutant clones, no LN-BP18 transcripts originating from TSS1 could be detected (Figure 22a). In the two of the remaining homozygous clones and in heterozygous clones transcripts containing the longer exon 1 (Ex1ln-4), such as LN-BP18_001, could not be detected. On the other hand, transcripts, such as LN-BP18_003, that contained a shorter exon 1 (Ex1sh-4) were downregulated but still detectable. Clone 1H was the only analysed clone where an upregulation was observed for TSS1. This discrepancy with the results of the remaining clones revealed this was an abnormal clone, or that some contamination with uncharacterized ESC occurred. As such this clone was not further studied. Transcription from TSS2 was not affect in most homozygous and heterozygous clones, with a downregulation observed in a homozygous mutant ESC (Clone 11E). For this homozygous clone and for one of the heterozygous clones the overall LN-BP18 level was downregulated, while for remaining clones it was either unaffected or slightly downregulated (Figure 22a).

In TSS2 mutants, there was a downregulation of transcripts originating from this TSS and downregulation of the overall expression of LN-BP18 was also detected, contrary to the effects of TSS1 deletion (Figure 22c). Importantly, Sall1 expression was not affected in TSS2 clones (Figure 22c) neither in TSS1 clones, even on homozygous TSS1-KO where most of its intron was deleted. This revealed that its regulatory region was left intact and that splicing was not affected (Figure 22b). The TSS1 homozygous KO clone 11E was the only homozygous clone with a downregulation of TSS2 and overall LN-BP18 expression, and was additionally the clone with the lowest expression of Sall1. These data revealed possible genomic alterations not intendent that could not be detected through the screening methods

applied. As such and as concluded for clone 1H, clone 11E was discarded from further analysis. To test the effect of LN-BP18 TSS inactivation on other two important pluripotency genes, Nanog and Oct4, were selected for expression analysis. Unlike Sall1, expression of these genes was upregulated approximately 2-fold in TSS1 mutant clones (Figure 22b). However, when TSS2 was inactivated, this effect was not observed (Figure 22c). For TSS1 KO cells, plotting LN-BP18 overall levels against Nanog and Oct4, a positive correlation was observed, with an R squared of 0.410 and 0.387, respectively (Figure 22e).

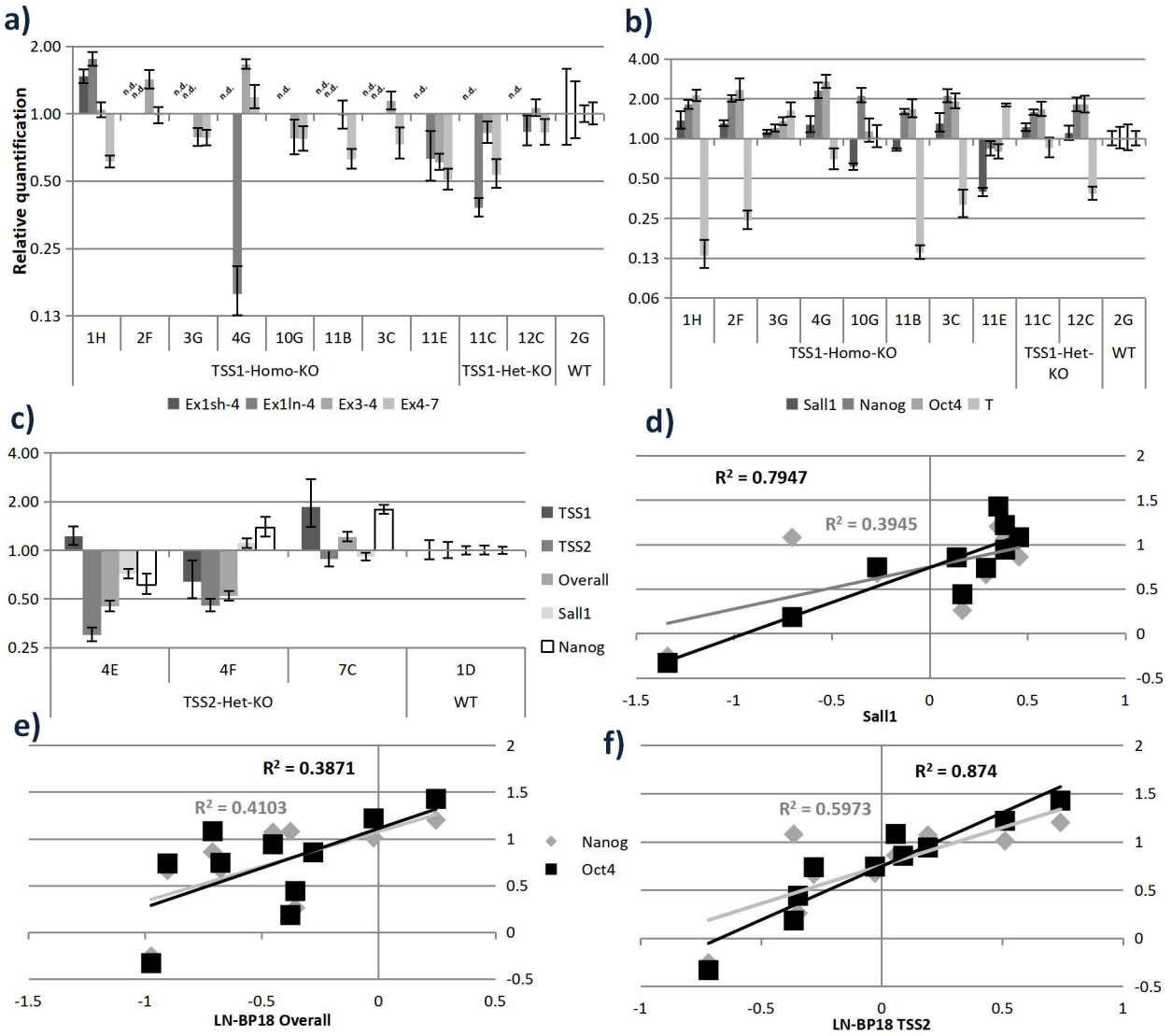


Figure 22 - Excision of LN-BP18 TSS through CRSIPR-Cas9

Quantification by qPCR normalized to Pmm2 expression and compared to WT sample of **a)** of LN-BP18 overall, TSS1 and TSS2 transcripts and **b)** Sall1, pluripotency and differentiation markers in LN-BP18-TSS1 mutants.; **c)** relative quantification in TSS2 mutant cells of the indicated genes; **d)** correlation between ΔΔCT for Nanog and Oct4 against Sall1, **e)** LN-BP18 overall or **f)** TSS2 transcripts in LN-BP18-TSS1 mutant ESCs. Pmm2 was used as an internal control and expression normalized to WT sample (n=3, SD); n.d., expression not detectable in sample.

Additionally, when correlating expression of the two TFs to the expression of isoforms originating from the TSS2, the coefficient was even higher, with a value of 0.660 and 0.867 for Nanog and Oct4, respectively (Figure 22f). A similar correlation between the mentioned key pluripotency TFs and Sall1 was also observed (Figure 22d). This correlation between genes expression indicated a possible co-regulation of these genes or that one of the genes activates the expression of the other.

To better understand the relation between LN-BP18 and Sall1 and untangle their possible role in pluripotency regulators expression, Sall1 depleted mutants were generated.

3.4. Generation of Sall1 depleted mESC

As already demonstrated, LN-BP18 and Sall1 were expressed in similar tissues and both were affected by a decrease of Med12 levels (Figure 10b-d). However, LN-BP18 is already lowly expressed in ESC and as such, the generated LN-BP18 mutant ESCs, which result in a decrease the already low expression of the lncRNA might not generate any observable effects in ESCs. However, Sall1 is highly expressed in ESC, where it has been identified as an important factor in pluripotency maintenance (Karantzali et al. 2011). As such, should Sall1 depletion increase LN-BP18 expression, this effect would be easier to quantify than a decrease of the lncRNA expression. Additionally, Sall1 depletion could provide insight into these antisense genes relation and their function in ESCs.

To generate Sall1-KO mutant ESCs, guide RNAs flanking most of its coding sequence were designed in order to excise the majority of the protein sequence (Figure 23a). As before, G4 cells were transformed with CRISPR-Cas9 vectors and individual clones were screened for the excision of the coding sequence of the Sall1 gene. Contrary to LN-BP18 generated mutant cells, where the levels of the mutated gene could only be quantified on the RNA level, with Sall1 mutant cells the protein level can also be quantified using a western blot. For this, cells were expanded and protein extracted as described in section 0. The protein levels of the housekeeping gene β-actin were used to normalize the amount of input used.

Quantification of Sall1 protein levels in mutant clones allowed the identification of heterozygous and homozygous deletions. While on heterozygous clones Sall1 protein levels were 4-fold lower when compared to the WT samples (clone 8E and 12E), in homozygous clones (clones 9G and 12D) no Sall1 protein could be detected (Figure 23b). Sanger sequencing of the homozygous clones locus showed that, in clone 12D both alleles had a deletion of the coding region, with a ~400bp shift downstream in one of the alleles (Figure 23a and c, “Deletion 2”).

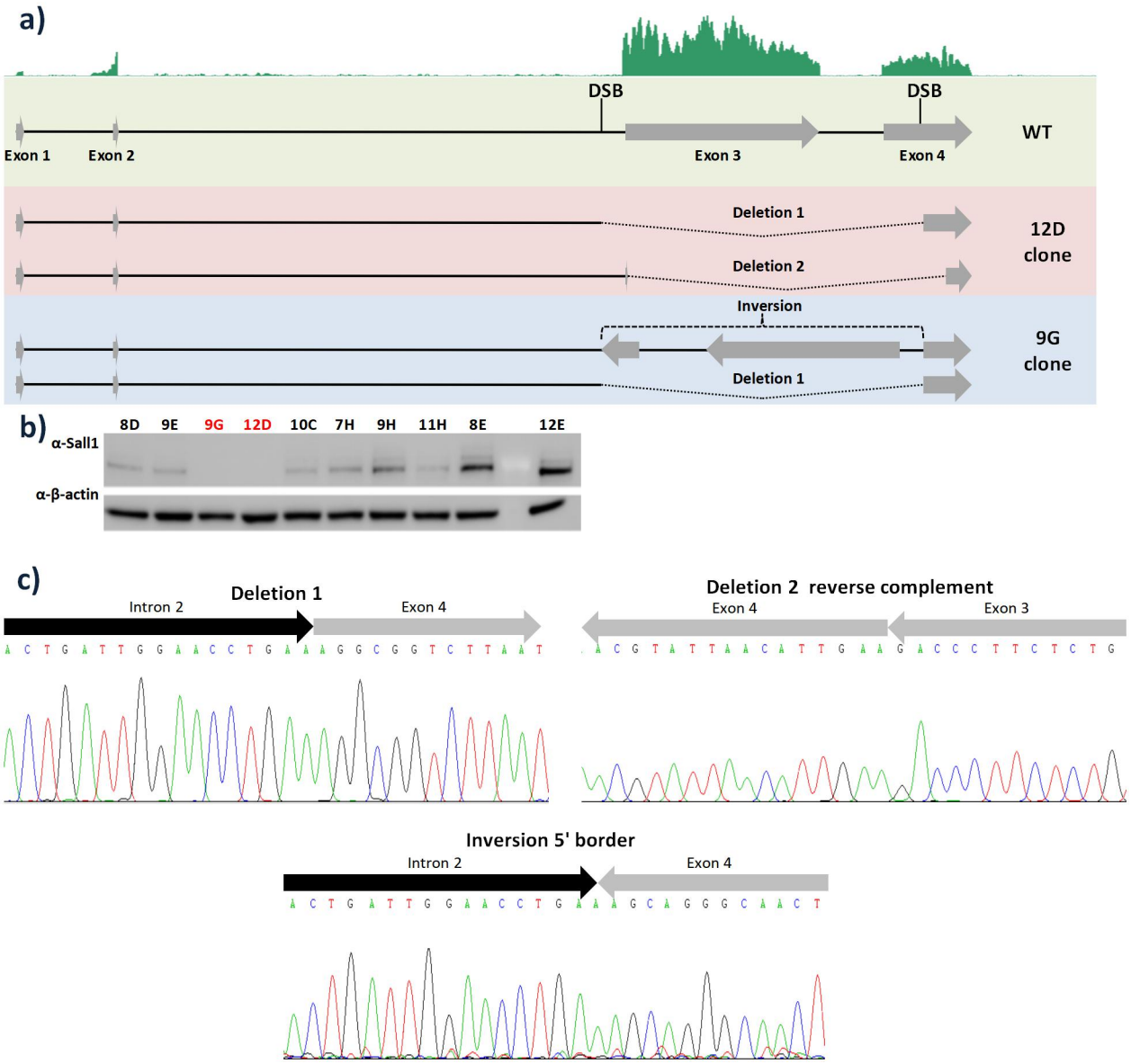


Figure 23 - Generation of Sall1 depleted ESCs mutant cells

a) Schematic representation of Sall1 locus in WT genome (green background) and in two homozygous clones, with indication of the induced double strand breaks (DSB) in order to excise the majority of Sall1 coding sequence. Top green track: RNA-seq data for Sall1 obtained for WT sample from section 3.1, **b)** Western blot on protein extracts of Sall1 deficient ESC using anti-Sall1 and anti-β-actin antibodies. β-actin was used as a house keeping gene in order to normalized the amount protein extract used; Sall1 homozygous mutant clones are highlighted in red; **c)** Chromatogram obtain with Sanger sequencing of the mutant clones genome. "Deletion 1" represents the predicted deletion, with break point matching precisely the induced DSB. "Deletion 2" was the deletion found on homozygous mutant clone 12D, where the break points are located ~400bp downstream of both induced DSB. "Inversion 1" was the genomic event observed for one of the alleles of clone 9G, where instead of excising the fragment within the DSB, the whole fragment was inverted.

Interestingly, in clone 9G while one of the alleles had the expected deletion (Figure 23a and c, “Deletion 1”)), on the other a complete inversion occurred (Figure 23and c, “Inversion”), which also resulted in a potential null allele, as suggested by the lack of detectable Sall1 protein in this clone (Figure 23b).

3.4.1. Characterization of Sall1 depleted mutants

A previous study by the Kretsovali laboratory demonstrated that Sall1 synergizes with Nanog in the activation of this key pluripotency TF target genes. It was also shown that Sall1 repressed mesodermal and ectodermal differentiation markers in ESCs (Karantzali et al. 2011). The authors reduced Sall1 expression by 2-fold using siRNAs, resulting in a similar downregulation of Nanog, while other important pluripotency factors, such as Oct4, where not affected. Through Sall1 overexpression during embryoid body differentiation, the researchers also observed a repressor effect on several mesodermal and ectodermal markers. This effect was further confirmed for Hand1 and T, two early mesoderm markers, by their 2-fold increased expression in ESC upon Sall1 downregulation. To verify if these reported effects were also detectable in the Sall1 mutants generated in this project, RNA was extracted from the Sall1 mutant ESC clones, reverse transcribed into cDNA and expression levels of T, Hand1, Nanog and Oct4 were evaluated by qPCR (Figure 24a, b). Contrary to what has been described on the report by the

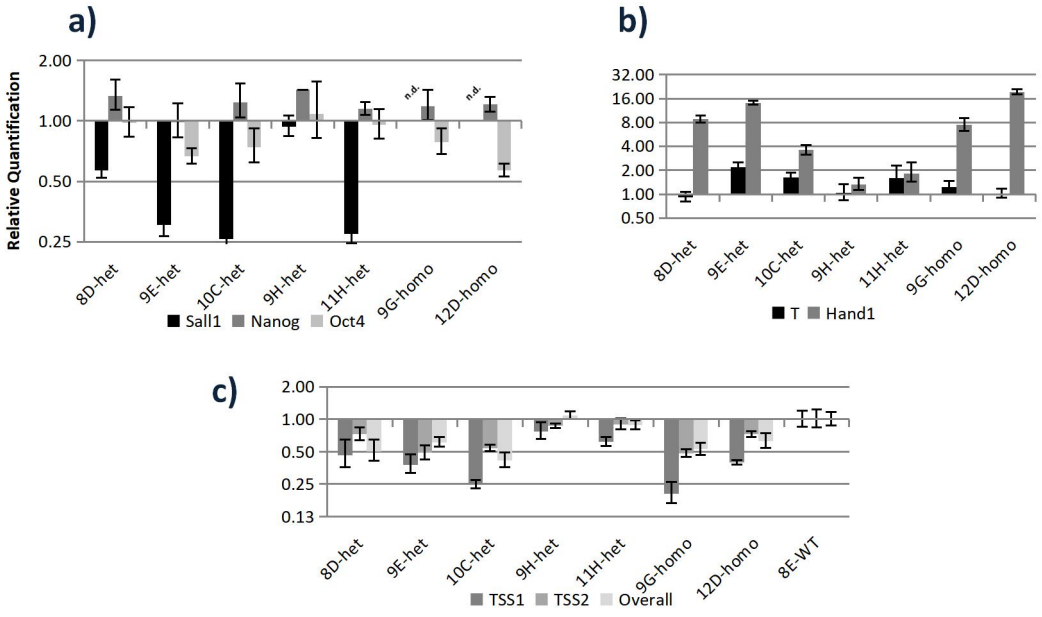


Figure 24 - qPCR quantification in Sall1 depleted ESCs
a) Relative quantification of Sall1, Nanog, **b)** differentiation markers previously reported to be affected by Sall1 depletion and **c)** LN-BP18. Pmm2 was used as an internal control and expression normalized to a WT sample (n=3, SD); n.d., not detectable in sample.

Kretsovali group, no significant changes on normal Nanog levels were detected even in cells with devoid

of Sall1, with T expression was also not effected in these mutant cells (Figure 24a). However a clear change on Hand1 expression was detected, being upregulated in five of the seven mutant clones. Interestingly, the reported effects of Sall1 depletion on Nanog and Hand1 expression levels could be recapitulated on Med12 depleted ESCs. In the Med12^{null} mutant, Sall1 and Nanog were 4-fold and 2-fold downregulated, respectively, while Hand1 was 8-fold upregulated.

LN-BP18 was quantified in the Sall1 depleted cells. While TSS1 transcripts were downregulated over 2-fold, the overall and TSS2 levels were only mildly downregulated (Figure 24c), suggesting an activating effect of Sall1 on LN-BP18, acting on transcription originating from TSS1.

3.5. Identification of lncRNAs targets of Med12

Analysis of transcriptome data previously generated by the Schrewe group for Med12^{null}, Med12^{hypo} and WT ESCs revealed Med12 as an important regulator of ncRNAs expression in ESC (section 3.1). With these data, over 200 non-coding genes were found misregulated, including several putative novel lncRNAs, such as LN-BP18. However, as already mentioned, technical and biological variability is a critical aspect of RNA-seq data analysis, since if not account for, the expression values obtained are not reliable. This variability is even more important to take into account when analysing lowly expressed genes, as is the case of most lncRNAs, or when identifying genes with only a slight variation in expression between the different conditions. This might result a high percentage of false positives and false negatives when classifying genes as misregulated, as observed for LN-BP18. Although analysis of the original RNA-seq datasets identified LN-BP18 as upregulated in both Med12 mutant ESCs, with a 3-fold upregulation in its expression (Figure 10d), qPCR quantification in Med12^{null} cells showed that this lncRNA overall levels were slightly downregulated and that transcription from TSS1 was 4-fold downregulated.

In order to identify lncRNAs that might be regulated by Med12 and considering the lack of statistical confidence observed for the original data analysed, new RNA-seq data was generated, using the same mutant ESCs. Additionally, the Med12^{flox} mutant ESC was also included in the generated data. This mutant has been previously showed to express Med12 at normal levels and mice generated with it showed no distinct phenotype and were fertile (Rocha et al. 2010).

Two different biological replicates were used for each sample, which allowed accounting for the technical and biological variability by performing statistical tests between replicates. All cells were cultured in 6cm plates until confluent, feeder depleted and RNA extracted with RNeasy Micro kits. rRNA was depleted from 500 ng total RNA using Ribo-Zero Magnetic Kit and strand specific cDNA was

generated with ScriptSeq V2 RNA-Seq Library Preparation Kit. To each sample a different index was added and all samples were pooled and analysed in a lane of a HiSeq 2500 system. Around 45 million paired end reads with a length of 75 nt were obtained for each sample. Using a longer read size (75 nt in the new data compared to 50 nt in the original data) increased the likelihood of mapping a spliced reads, an important aspect when using tools as Cufflinks to identified putative novel genes.

In paired ended sequencing, each fragment is sequenced from both ends. This results in two sequences for each fragment, called mates. While the first mate results in 5' to 3' sequencing of the top strand of the fragment, the second mate results in the 5' to 3' sequencing of the bottom strand of the same fragment. In order to assess the quality of the generate transcriptome data, FastQC was used on all samples. This tool allowed detecting and error during the sequencing cycle 40 of the second mate, rendering all the sequences obtained after position 39 unreliable (Figure 25b). As such only the first 39nt of mate 2 were used.

Since the analysis of the original RNA-seq data, multiple aligners have been created and newer versions released for older tools. Since one of the main goals of this dataset was to study lncRNAs, which are usually lowly expressed, the more reads were mapped to the genome, the more reliable the expression results for these genes would be. As such, in order to evaluate if a different aligner perform better than the Hisat2 used to map the original RNA-seq data, Tophat2 (Kim et al. 2013) and Star (Dobin et al. 2013), two of the most commonly used aligner were evaluated together with Hisat2. The Tophat2 it is an aligner routinely used in-house and developed to be used together with Cufflinks in a pipeline while STAR is one of the most commonly used aligners, with benchmark tests identifying this tool as one of the aligners with higher sensitivity and precision (Baruzzo et al. 2017).Using one of the WT replicates, the tools were evaluated regarding the number of mapped reads in a proper pair with a mapping score above 10 (Figure 25a). Among the tested tools, STAR was the one which mapped more reads fulfilling the filtering criteria with the added bonus of also being the fastest tool (Figure 25a).

The fact that more reads were kept, indicated a potentially higher sensitivity for STAR, which allows mapping more reads and thus making it easier to detect genes with a low expression, such as lncRNAs. For this reason, STAR was the chosen aligner for this analysis. *De novo* transcript assembly was once again performed before differential gene analysis. Cufflinks *de novo* assembled transcripts longer than 200bp and with more than one exon were kept, discarding all others, resulting in the addition of almost 420 novel predicted genes. The novel transcripts were filtered before gene quantification since if maintained in the annotation, these transcripts would be accounted for the read depth normalization step and as such removing them before this step allowed for a more accurate normalization.

Cuffdiff is a tool specialized in transcripts quantification, not being the most appropriate tool for gene quantification. In some cases, this tool also combines multiple genes into a single entry. This merging behaviour occurs when mapped reads allow assembly of a single transcript that spans genes in close proximity. One such example is the case of the genes Cd68, Eif4a1, Senp3, Tnfsf12 and Tnfsf13, which are in close proximity and as such were assigned to one entry with a single expression value, not allowing quantification of each gene separately. As such, Deseq2 was used instead (Love et al. 2014). This is one of the most commonly used tools for differential gene analysis, performing a series of statistical tests for each gene in every pairwise comparison. Principal component analysis revealed that both replicates for

a)

Tool	Time (min)	Reads after filtering	%
Tophat2	1040	25,874,590	28.85
Hisat2	146	70,043,830	78.10
Star	74	71,579,432	79.82

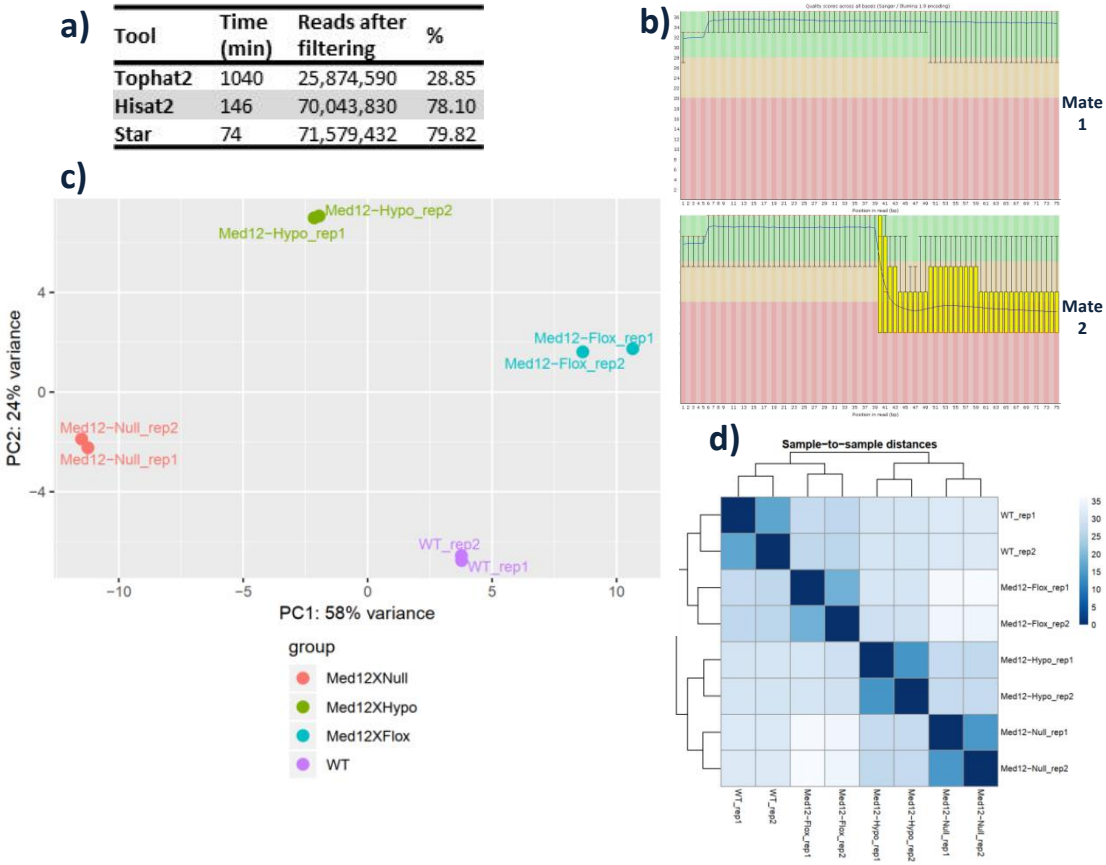


Figure 25- Med12 mutant cells transcriptome data quality control
a) Comparison of performance for the three tested aligners using WT_rep1 sample. Time necessary and amount of reads in proper pair with a score above 10 after mapping were evaluated; **b)** Fasqc report for WT_rep1, representative of all samples, revealing a sequencing error at cycle 40 of second mate; **c)** principal component analysis on all samples revealed that all replicates clustered together and away from other samples; d) sample to sample distances, evidencing the most similar samples.

each sample clustered together and separately from the remaining samples, as expected (Figure 25c). Sample to sample distances further showed that WT and Med12^{flox} were the more similar samples, with

Med12^{hypo} and Med12^{null} more similar to one another than to any of the remaining samples (Figure 25d). Due to the similarity between WT and the Med12^{flox} the latter was used as the control sample (Figure 6a). This allowed to normalize expression to possible effects resulting from the transformation protocol used to generate these cells and from the presence of the LoxP sites on the Med12 locus, since it has been show before that the Cre-Lox system might generate unintended effects throughout the genome (Harno et al. 2013).

3.5.1. Analysis of misregulated genes in Med12 depleted mESC

Of the 55,000 genes included in the quantification, for 5,000 genes Deseq2 statistical test could be successfully applied. These tests could not be applied to the remaining genes since they were either lowly expressed or not expressed in all samples. Over 2,000 genes had their expression effected in one of the mutants with changes above 50% in expression compared to the Med12^{flox} control, with a false discovery rate (FDR) < 0.05 (Figure 26a). Although the majority of misregulated genes were protein coding, 200 were non-coding genes, 50 of which were new transcripts predicted by Cufflinks (Figure 26b). Contrary to what as observed in the original RNA-seq analysis and as previously demonstrated by qPCR data (Figure 16c), the overall LN-BP18 level were not significant affected by Med12 deficiency in this new RNA-seq data.

Comparing the log2FC distribution revealed that genes were similarly misregulated in both mutants, with upregulated genes showing a higher fold change on Med12^{null} (Figure 26b). This suggested that the 5% remaining Med12 in the Med12^{hypo} mutant was sufficient to partially repress some of the affected genes. However, restricting the distribution to misregulated ncRNAs, the differences were more subtle, with genes more downregulated in the Med12^{null}, contrary to the original data analysed (Figure 8b). GO term enrichment analysis revealed cell differentiation as the most enriched hit on misregulated coding genes (Figure 26e), due to misregulated genes such as Forkhead box N4 (Foxn4) and Homeobox protein NK-6 homolog B (Nkx6-2) and genes also misregulated in the original RNA-seq data such as GLI family zinc finger 2 (Gli2) and Neurogenic locus notch homolog protein 3 (Notch3). This enriched term, together with regulation of apoptotic process (e.g. Fibroblast growth factor receptor substrate 2 (Frs2) and Early growth response 1 (Egr1)), cell adhesion (e.g. NUAK family kinase 1 (Nuak1) and Filamin binding LIM protein 1 (Fblim1)) and cell proliferation (e.g. Cyclin dependent kinase inhibitor 1B (Cdkn1) and Fibroblast growth factor 5 (Fgf5)) represent diverse processes that act in concert for the proper development of different organs and tissues. Other enriched terms indicated more specific

developmental process, such as lung (e.g. Gli3 and Paired like homeodomain 2 (Pitx2)) and limb development (e.g. Wnt family member 3 (Wnt3) and SPARC related modular calcium binding 1 (Smoc1)), with the last being one of the expression domains identified for LN-BP18 and Sall1.

Different pathways where Med12 plays clear roles were also identified, due to enrichment of genes with function on canonical Wnt or BMP signalling pathways (Rocha et al. 2010, Huang et al. 2012), further confirming the confidence of the identified hits. The broad terms enriched on Med12 depleted mutants were expected since this gene has been identified as one of the few "hub" genes, on which multiple processes converge. As such, depletion of this gene affected a variety of developmental processes, as confirmed by the diverse phenotypes associated with $Med12^{hypo}$ mutants (Rocha et al. 2010).Several of the GO terms enriched in the misregulated coding genes were also enriched for the misregulated coding genes of the original RNA-seq data (Table 1), such as angiogenesis and axon guidance, further supporting to the results presented in section 3.1.1. Canonical Wnt signalling was one of the enriched GO terms identified in the analysis of misregulated protein coding genes (Figure 26e). However, several of the Wnt targets that had been found to be originally misregulated, such as T, Axin2 and Sall4, in the newer data their expression is not affect by Med12 depletion (data not shown). These observations reveal that, for some individual genes, the results previously obtained are not supported by the most recent data. However, the disturbance of important pathways and biological process were supported.

Interestingly, GO terms were enriched in misregulated ncRNAs, including DNA methylation, with Ftx as one of the genes identified (Figure 26f). Ftx is one of the Xist regulators in XCI, a process in which several of its regulators were found misregulated in the previously mentioned RNA-seq data (section 3.1). From the analysed Xist regulatory genes, only Ftx was found misregulated (Figure 26d), contrary to what was observed on the previously analysed data (Figure 8d). Despite no clear effect of Med12 on XCI, with only one of the main regulators of this process misregulated, the data suggested a role for Med12 in the proper expression of regulators of this process.

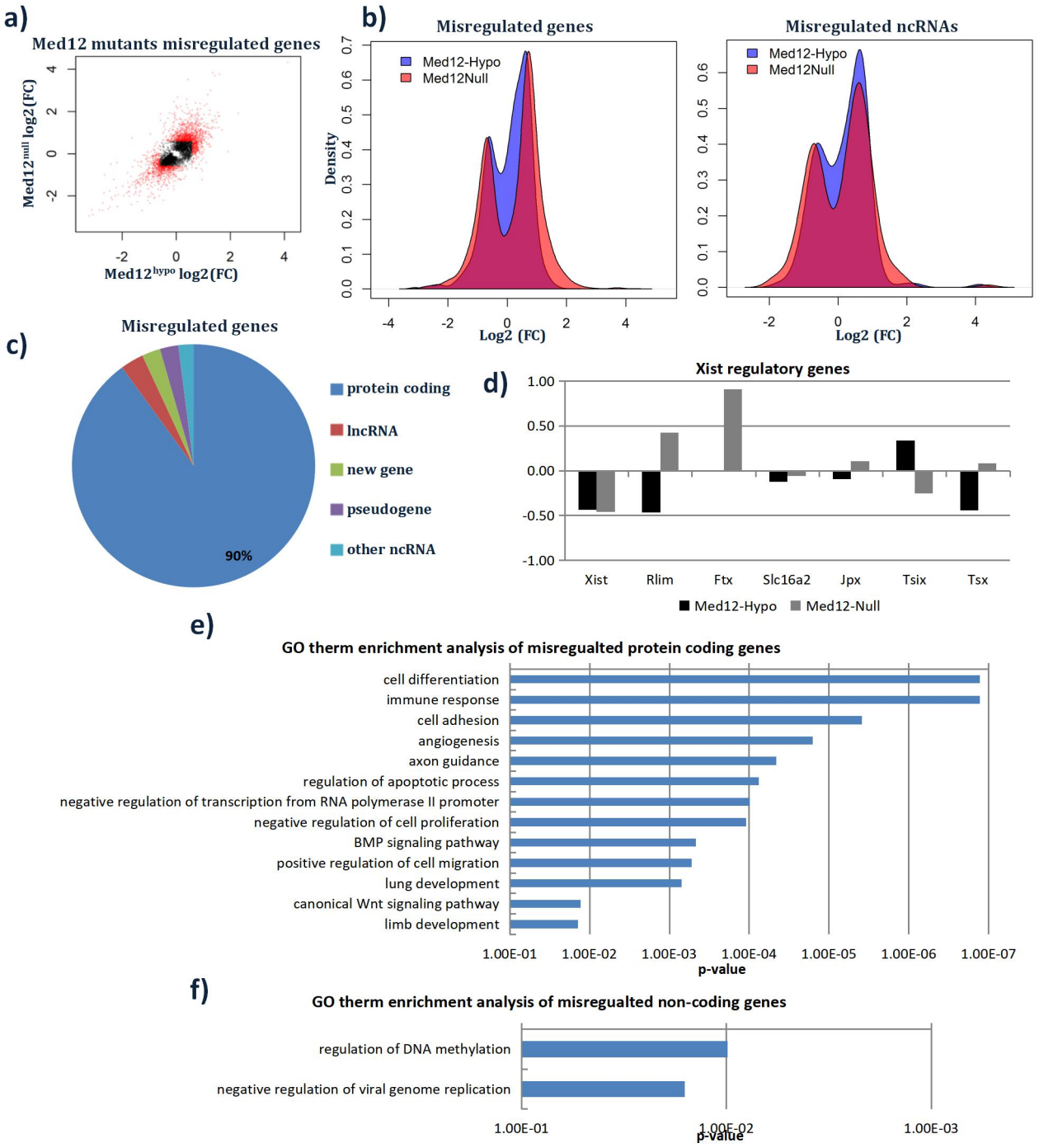


Figure 26 - Med12 mutants RNA-seq analysis

a) Scatter plot of Med12^{hypo} and Med12^{Nul} log2FC for the 5000 genes with a successful statistical test; in red misregulated genes; **b)** gene type of misregulated genes; **d)** log2FC on Med12 mutants of genes associated with XCI; **c)** density plots of log2FC distribution on both mutants for all misregulated genes (left) or misregulated non-coding genes (right); **e)** GO term enrichment analysis non redundant top hits for misregulated coding genes or **f)** non-coding genes.

3.5.2. Med12 Chromatin Immunoprecipitation data analysis

Analysis of the Med12 mutant cells revealed 150 annotated non-coding genes as well as 50 misregulated putative new genes predicted by Cufflinks as misregulated in at least one of the mutant samples. Due to the role of MED12 in a plethora of different processes and pathways, demonstrated by the misregulation of almost 2000 protein coding genes in the analysed Med12 deficient cells, for most of the misregulated lncRNAs there was a high chance that their expression was indirectly affected by Med12 depletion and the observed effect was due to misregulation of one or several regulators of these non-coding genes. If expression of a lncRNA is directly affected by reduced levels of Med12, then this subunit should be found binding at its promoter or gene body. To identify putative direct targets of Med12, ChIP-seq data for Med12 in ESC from a previous study was analysed (Kagey et al. 2010). Reads were mapped with STAR and peaks enriched in the Med12 immunoprecipitated sample against the input were determined using MACS2, identifying over 2,200 peaks, representing Med12 binding sites in the genome. Identified peaks were assigned to either promoter, gene body or to intergenic regions (Figure 27a). While the majority of peaks were found associated with protein coding genes in any of the three regions, significant differences were observed between this regions. In the other regions half of the identified genes were protein coding, this number increases to 76% when looking at genes bodies (Figure 27b). One explanation for this increase could be the presence of more and larger introns on protein coding genes, resulting in more peaks found on them. Interestingly, some of the Med12 peaks were associated with Cufflinks assembled genes, supporting the hypothesis that these might be Med12 direct targets. Log2FC for both mutants was plotted for all misregulated genes associated with Med12 peaks (Figure 27c).Genes with peaks in promoter, gene body or in both could be found either up or downregulated, with a small enrichment for downregulated genes By calculating the density of log2FC distribution of misregulated genes with peaks in promoter or in gene body it could be observed if Med12 binding affected these genes differently upon the subunit depletion. When looking at all the misregulated genes, there was a higher density of upregulated genes in the Med12 mutants (Figure 26b, left). On the other hand, when restricting this analysis to genes with Med12 peaks in the gene body, no significant difference could be observed in the density of up and downregulated genes (Figure 27e, top left). Furthermore, the majority of genes where Med12 was found binding at their promoter were downregulated (Figure 27e, top right). Focusing analysis on the misregulated ncRNAs, the vast majority were not bound by Med12 at their locus (Figure 27d), suggesting an indirect action of Med12 on their expression.

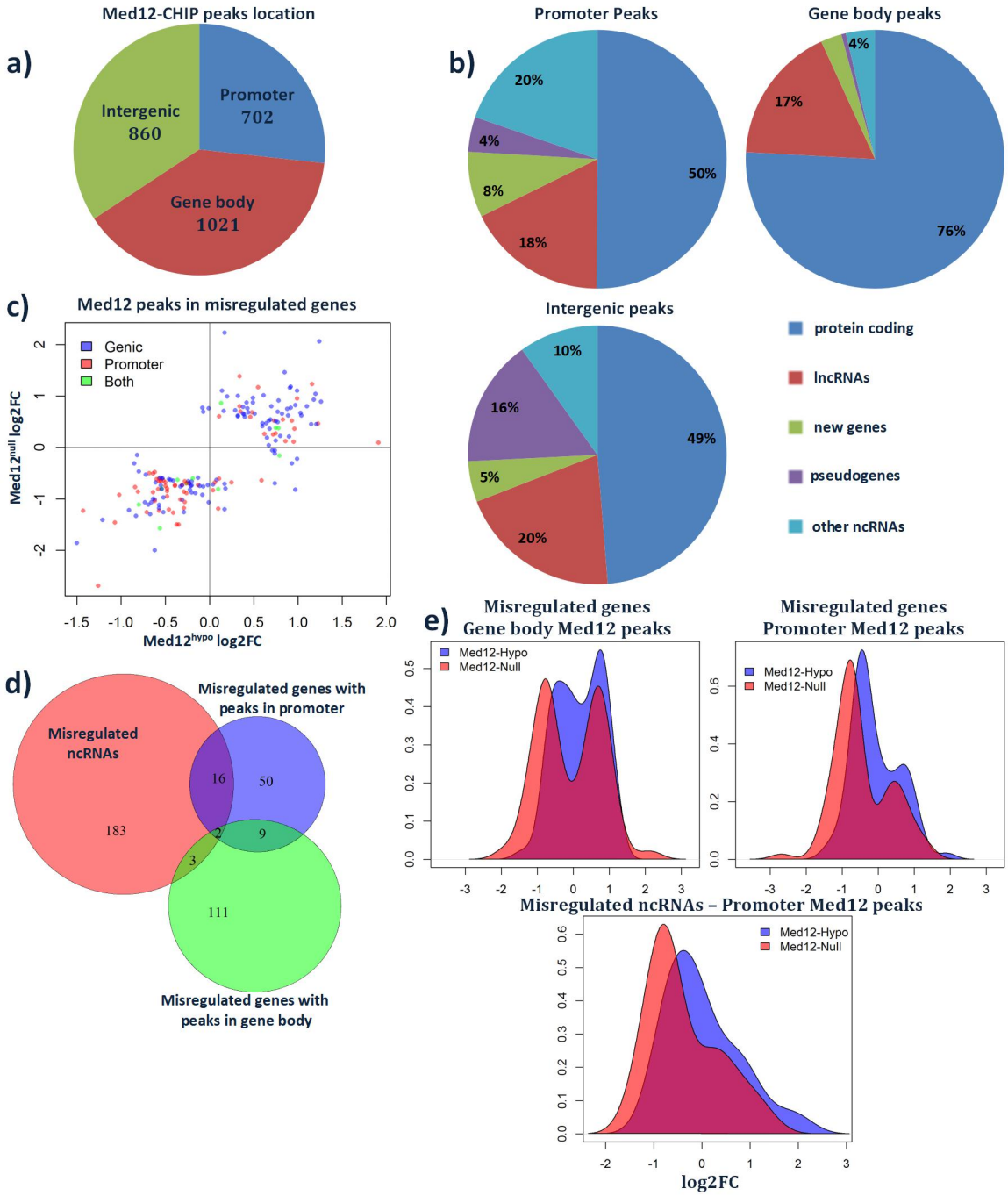


Figure 27 - Analysis of Med12 ChIP-seq public data

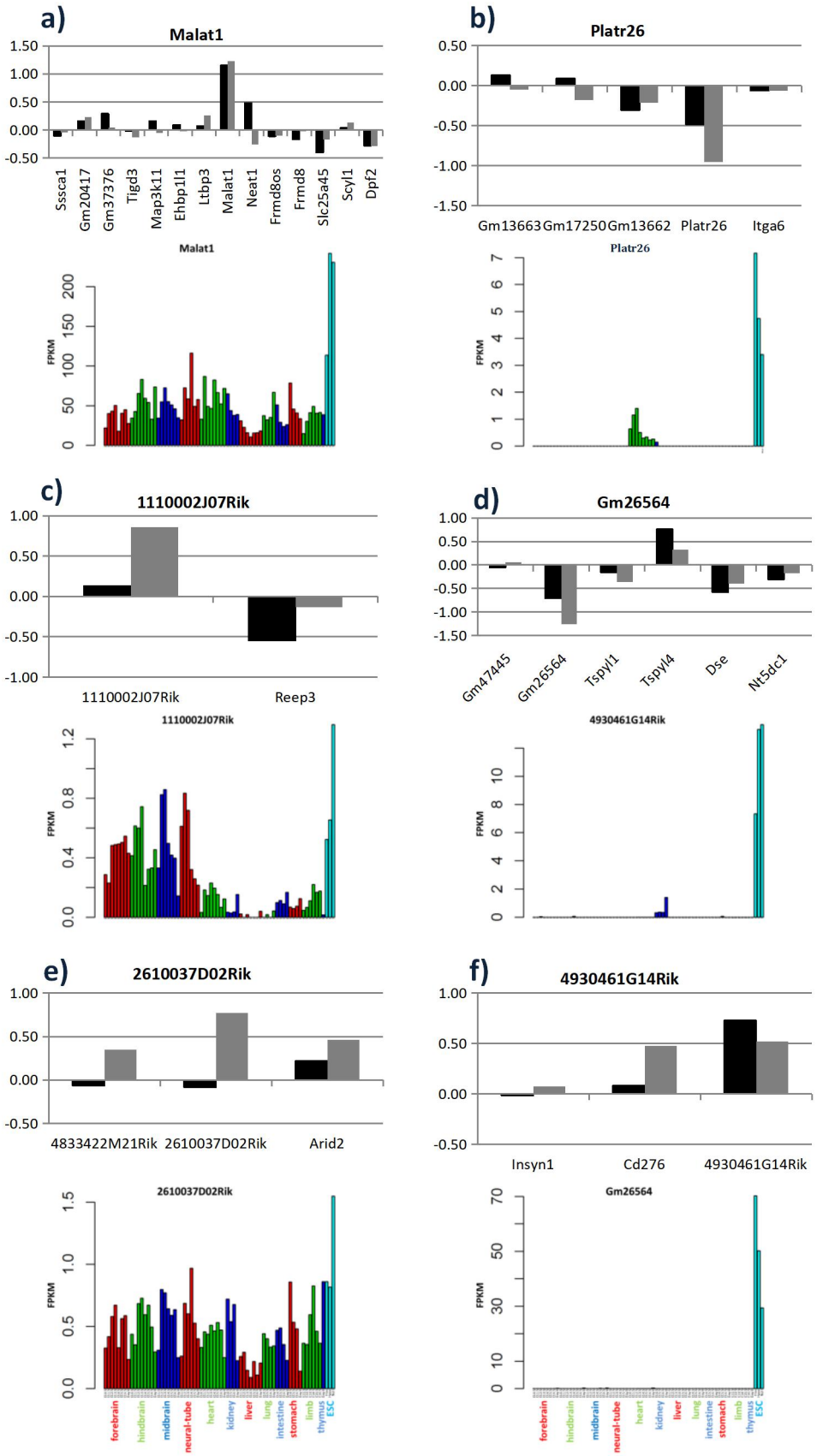
a) Peak association with different gene regions: promoter (-2kb/+1kb TSS); gene body (-1kb TSS to TES); Intergenic peaks were associated with the closest gene; **b)** gene type of genes with peaks on promoter (top left), gene body (top right) or closest to an intergenic peak (bottom); **c**) log2FC of misregulated genes with Med12 binding at promoter and/or gene body; **d)** venn diagram showing overlap between misregulated ncRNAs and misregulated genes with Med12 peaks at their promoter and/or gene body ; **e)** log2FC distribution for misregulated genes with Med12 peaks in promoter (top left) or gene body (top right) and for misregulated ncRNAs with Med12 peaks at promoter (bottom).

However for the ones that were associated with Med12 binding at their promoter, they were mostly downregulated, revealing these genes as good candidates for being direct targets of Med12 (Figure 27e, bottom). For these ncRNAs, evidence suggested that Med12 was necessary for their activation, consistent with the data reported previously on a subset of lncRNA (Lai et al. 2013).

3.5.3. Characterization of lncRNAs putative Med12 targets

Of the 50 misregulated ncRNAs that were predicted by Cufflinks, four had Med12 binding to their promoter. None of these four genes showed any coding probability by CPAT analysis (data not shown). Additionally, eight annotated lincRNAs were also misregulated in Med12 depleted ESCs and this subunit bound their promoter or gene body. Among these was the already mentioned Malat1 and Platr26, a lncRNA associated with pluripotency.

These genes represent a starting point for future experiments for identifying and characterizing lncRNAs regulated by Med12. RNA-seq data for embryonic tissues obtained from public databases indicated that the several of these candidate lncRNAs are expressed mainly on ESC. Additionally the expression pattern in mouse embryos of these Med12 target candidates and the log2FC of genes within 100 kb of candidate genes was compiled in Figure 28. Misregulated genes within this window represent putative targets for the candidate lncRNAs, since for multiple lncRNAs it has been shown that they affect the expression of neighbouring genes (Anderson et al. 2016, Paralkar et al. 2016).


a)
Malat1
1.50
1.00
0.50
0.00
-0.50
Sssca1
Gm20417
Gm37376
Tigd3
Map3k11
Ehbp1l1
Ltbp3
Malat1
Neat1
Frmd8os
Frmd8
Slc25a45
Scyl1
Dpf2
Malat1
FPKM
0
50
100
150
200
b)
Platr26
0.50
0.00
-0.50
-1.00
-1.50
Gm13663
Gm17250
Gm13662
Platr26
Itga6
Platr26
FPKM
0
1
2
3
4
5
6
7
c)
1110002J07Rik
1.00
0.50
0.00
-0.50
-1.00
1110002J07Rik
Reep3
1110002J07Rik
FPKM
0.0
0.4
0.8
1.2
d)
Gm26564
1.00
0.50
0.00
-0.50
-1.00
-1.50
Gm47445
Gm26564
Tspyl1
Tspyl4
Dse
Nt5dc1
4930461G14Rik
FPKM
0
2
4
6
8
10
e)
2610037D02Rik
1.00
0.50
0.00
-0.50
4833422M21Rik
2610037D02Rik
Arid2
2610037D02Rik
FPKM
0.0
0.5
1.0
1.5
forebrain
hindbrain
midbrain
neural-tube
heart
kidney
liver
lung
intestine
stomach
limb
thymus
ESC
f)
4930461G14Rik
1.00
0.50
0.00
-0.50
Insyn1
Cd276
4930461G14Rik
Gm26564
FPKM
0
10
30
50
70
forebrain
hindbrain
midbrain
neural-tube
heart
kidney
liver
lung
intestine
stomach
limb
thymus
ESC

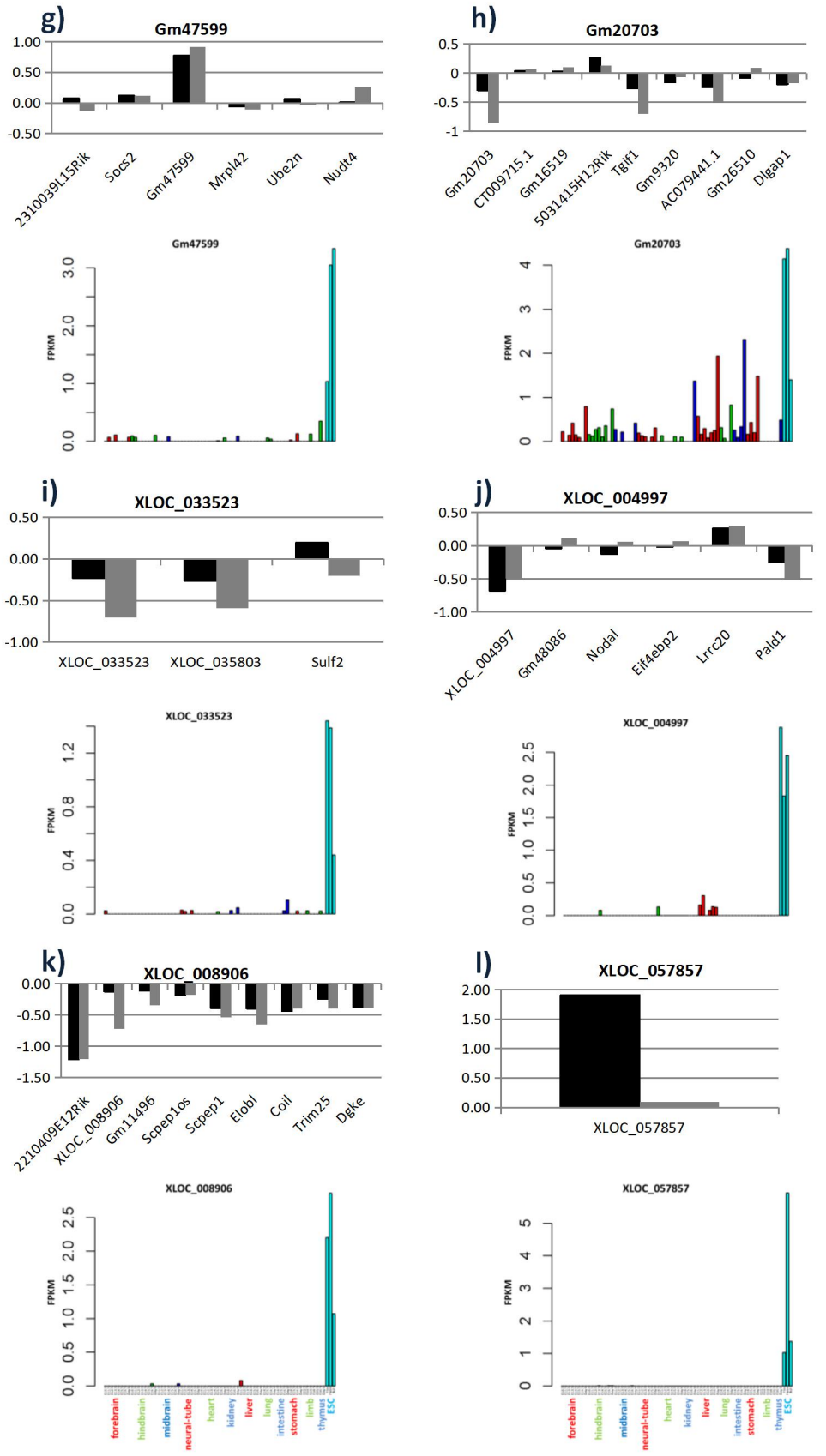

g)
Gm47599
1.00
0.50
0.00
-0.50
2310039L15Rik
Socs2
Gm47599
Mrpl42
Ube2n
Nudt4
Gm47599
FPKM
3.0
2.0
1.0
0.0
h)
Gm20703
0.5
0
-0.5
-1
Gm20703
CT009715.1
Gm16519
5031415H12Rik
Tgif1
Gm9320
AC079441.1
Gm26510
Dlgap1
Gm20703
FPKM
4
3
2
1
0
i)
XLOC_033523
0.50
0.00
-0.50
-1.00
XLOC_033523
XLOC_035803
Sulf2
XLOC_033523
FPKM
1.2
0.8
0.4
0.0
j)
XLOC_004997
0.50
0.00
-0.50
-1.00
XLOC_004997
Gm48086
Nodal
Eif4ebp2
Lrrc20
Pald1
XLOC_004997
FPKM
2.5
2.0
1.5
1.0
0.5
0.0
k)
XLOC_008906
0.00
-0.50
-1.00
-1.50
2210409E12Rik
XLOC_008906
Gm11496
Scpep1os
Scpep1
Elobl
Coil
Trim25
Dgke
XLOC_008906
FPKM
2.5
2.0
1.5
1.0
0.5
0.0
forebrain
hindbrain
midbrain
neural-tube
heart
kidney
liver
lung
intestine
stomach
limb
thymus
ESC
l)
XLOC_057857
2.00
1.50
1.00
0.50
0.00
XLOC_057857
XLOC_057857
FPKM
5
4
3
2
1
0
forebrain
hindbrain
midbrain
neural-tube
heart
kidney
liver
lung
intestine
stomach
limb
thymus
ESC

Figure 28 - Candidate direct Med12 target lncRNAs
Log2FC determined for Med12^{null} and Med12^{hypo} for all genes up to 100kb away from the identified candidate lncRNA (top) and candidate lncRNA expression in 13 different tissues across 8 developmental stages on mice embryos , downloaded from Encode database and in the analysed ESCs mutants (bottom) for **a)** Malat1; **b)** Platr26; **c)** 1110002J07Rik; **d)** Gm26564; **e)** 2610037D02Rik; **f)** 4930461G14Rik; **g)** Gm47599; **h)** Gm20703; **i)** XLOC_033523; **j)** XLOC_004997; **k)** XLOC_008906**; l)** XLOC_057857**.** For plots of log2FC, genes are displayed in order they appear in the selected genomic window, from 5' to 3' of the chromosome. Genes without any log2FC calculated for any of the samples were omitted. For genes with valid log2FC for only one of the samples, in the remaining sample log2FC was assumed to be 0.

4. Discussion

4.1. Med12 depletion in mESCs reflects the phenotypes observed in mutant embryos

Mediator, a conserved multi protein complex involved in the transcription of RNA polymerase II (Pol II) target genes, acts as a molecular bridge that transduces information from enhancers to the general transcription machinery at target gene promoters (Krishnamurthy et al. 2009). Among the 30 subunits that compose this complex, the kinase module subunit Med12has been described as a genetic hub, due to its function in multiple developmental pathways (Lehner et al. 2006). In humans, mutations of this gene have been associated with a variety of human pathologies such as cancer (Assie et al. 2014, Lim et al. 2014) and intellectual disabilities syndromes. Patients with these syndromes show characteristics resulting from abnormal development of diverse tissues, such as syndactyly, hypoplastic heart defects, agenesis of the corpus callosum, anal atresia and diverse craniofacial defects (Risheg et al. 2007, Schwartz et al. 2007, Vulto-van Silfhout et al. 2013). In mouse, embryos generated from $Med12^{null}$ (producing no Med12 protein) mutant embryonic stem cells (ESCs) died during early gastrulation, with severe disruption of canonical Wnt pathway (Rocha et al. 2010). Embryos generated with $Med12^{hypo}$ mutant ESC (with 5% of normal Med12 expression) developed further and displayed striking defects, such as neural tube closure defects, axis truncation and cardiac malformations, with the latter indicated as the probable cause of lethality. Whole mount *in situ* hybridization (WISH) revealed a perturbed activation of multiple β-catenin target genes in these embryos, correlating with the observed defects and demonstrating the essential role of Med12 in the activation of Wnt targets in the mouse (Rocha et al. 2010).

To confirm the observations from the mentioned study and to evaluate the impact of Med12 depletion in a more homogeneous sample, $Med12^{hypo}$ and $Med12^{null}$ mutant ESCs, from which embryos were previously generated were analysed (Rocha et al. 2010). The effects of Med12 depletion in these cells were evaluated with RNA-seq, a method that allowed to study all expressed genes and simultaneously to quantify their expression even for lowly expressed genes. RNA-seq data for these Med12 mutant ESCs and for WT ESCs were previously generated by the Schrewe laboratory. Analysis of this transcriptome data revealed over 1,400 misregulated coding genes in $Med12^{hypo}$ and/or $Med12^{null}$ mutants. Clustering of these genes showed that both Med12 mutant cells were very similar in their expression data, with misregulated genes either up- (clusters 1, 3 and 5) or downregulated (clusters 2 and 4) in both samples (Figure 7). Within the identified clusters, gene ontology (GO) enrichment analysis on the misregulated genes allowed to identify loci with similar functions within each cluster. This analysis

revealed genes associated with positive and negative regulation from Pol II transcripts (Table 1), consistent with the known functions of Mediator. Interestingly, it was observed an enrichment of misregulated genes involved in heart development, such as motile sperm domain containing 3 (Mospd3) and Roundabout guidance receptor 1 (Robo1), with the latter also involved in axon guidance (Pall et al. 2004, Blockus et al. 2016). The two GO terms mentioned are linked to of the defects observed in $Med12^{hypo}$ embryos: abnormal heart development and neural tube closure defects (Rocha et al. 2010). This observation revealed that strong effects of Med12 depletion could be detected in the analysed data.

The study of the Schrewe group concluded that Med12 was important for the proper activation of Wnt target genes. This was demonstrated by downregulation of multiple Wnt targets in $Med12^{null}$ embryos and by the compromised response to canonical Wnt signalling in $Med12^{hypo}$ ESC (Rocha et al. 2010). Analysis of the transcriptome data allowed to confirm misregulation of several Wnt targets, such as T, Axin2, Ccnd1 and Myc, that had been found downregulated in the reported embryos. Sall4 and Sall1 are two genes important for proper neural tube closure which were found misregulated in the mutant cells, suggesting these genes as potentially involved in the neural tube defects observed on the reported mutant embryos (Bohm et al. 2008). As could be shown by the misregulation of the mentioned genes, the transcriptome data confirmed disruption of β-catenin activation of Wnt target genes upon Med12 depletion.

Even though the RNA-seq data analysed was generated from ESC, several genes with roles in development were found downregulated (Figure 7, clusters 2 and 4). Downregulation of genes that in principle should not be expressed in ESC can have two explanations. On one hand, these genes can have a basal expression level in ESC but, due to Med12 depletion, the Mediator role in basal expression can be disturbed and so further reduce expression (Lacombe et al. 2013). On the other hand, these genes can be required for normal development and have additional functions in ESC. Indeed, Sall1 is an important gene for proper neural tube closure and kidney development and is additionally associated with pluripotency maintenance (Bohm et al. 2008, Kiefer et al. 2010, Karantzali et al. 2011).

Protein coding genes and ncRNAs were differently affect by the remaining 5% Med12 expressed in the $Med12^{hypo}$ mutant. This could be observed in the performed clustering, with protein coding genes similarly misregulated in both mutant cells (Figure 7). By contrast, clustering of the non-coding genes, revealed clusters with genes differently misregulated in the two mutants (Figure 8a). This was evident, for example for the first two clusters, where genes were upregulated in only one of the mutant cells. By plotting the distribution of log2FC for misregulated genes, it was further confirmed the similar effect of both Med12 mutations in the expression of coding genes and the difference in the ncRNAs expression

(Figure 8b). Furthermore, through this plot it was possible to observe that the majority of misregulated non-coding genes were upregulated in both Med12 mutant cells and that more genes were upregulated in the Med12^{null} compared to Med12^{hypo}. Additionally, a number of misregulated non-coding genes in the Med12^{null} mutant were not affected by the decrease of Med12 expression in the Med12^{hypo} mutant, as observed by the higher density of genes in this mutant with a log2FC between -1 and +1 (Figure 8b, plot on the right). The difference observed between coding and non-coding genes in both mutants revealed that ncRNAs are more sensible to the remaining 5% Med12 in the Med12^{hypo} mutants.

While the reported embryos generated from Med12^{null} ESCs died at E7.5, the Med12^{hypo} embryos survived up to E10.5 (Rocha et al. 2010). This difference in embryonic phenotype was due to the 5% of Med12 still expressed by the Med12^{Hypo}, and this level was suggested to be enough to activate early processes controlled by Wnt signalling but not for the later processes. In contrast to this hypothesis, multiple Wnt targets were similarly misregulated in both Med12 mutant ESCs. Several reasons might explain the observed difference in the generated embryos and the transcriptome data from ESCs. It is possible that, while the mentioned Wnt targets had a similar expression in both ESCs mutants, the factors activated by β-catenin at later developmental stages would be differently affected by the residual Med12. It is also possible that due to the different expression profiles between ESCs and the differentiated tissues in the embryos, the same Wnt targets would be differently regulated in both cases. Finally, one of the observations that resulted from the transcriptome data analysis of mutant ESc was the similar expression of protein coding genes in both Med12 mutants, while non-coding genes were more sensible to the different Med12 levels. This suggests non-coding genes as a potential source for the difference in phenotypes observed on the reported embryos.

Despite the known role of Med12 in multiple developmental pathways, one controversial role was in the maintenance of pluripotency in ESCs. In a study where Med12 was knocked down in ESCs using small interfering RNAs (siRNAs), Nanog was similarly downregulated which resulted in the downregulation of pluripotency genes and upregulation of differentiation markers (Tutter et al. 2009). By contrast, previous analysis of Med12^{hypo} mutant ESC revealed that the reported genes were not misregulated in this mutant (Rocha et al. 2010). This observation was supported by the RNA-seq analysis of Med12 mutant ESC performed in this section, since no variation in Nanog or its target genes was detected (data not shown). The discrepancies observed between the two studies very probably arise from the difference in the method used to knock down Med12 expression. In the study from the Schrewe laboratory, the Med12 genomic locus was directly manipulated with precise editing (Rocha et al. 2010). However, in the study from the Kadam group siRNAs were used, that are known to potentially have severe off targets effects,

which could result in the downregulation on unspecific genes (Tutter et al. 2009). One way to confirm this hypothesis would be to perform a rescue experiment. For this, the same siRNA would be used in ESC expressing a mutated Med12 mRNA that is resistant to siRNAs inactivation, an approach successfully used in previous studies (Jiang et al. 2004). If the expression of Nanog and its target genes was still perturbed in this experiment, then it could be concluded that the defects were due to unspecific targeting by the siRNAs.

4.2. Med12 regulates expression of putative novel long non-coding genes

As mentioned before, the Mediator complex is required for the proper expression of almost all Pol II transcripts (Petrenko et al. 2017). Accordingly, depletion of its subunit Med12 affected expression of over 1,400 protein coding genes in ESCs. Pol II mediates transcription of all mRNAs but also of most ncRNAs (Bunch 2018). As such, analysis of the transcriptome of Med12 depleted ESCs also revealed misregulation of over 200 non-coding genes. The majority of these were already annotated, such as Platr3, a lncRNA that has been associated with pluripotency maintenance and that is upregulated 2-fold in both mutants. Tsix and Xist are two extensively studied lncRNAs that were found misregulated in both mutants, with the latter being the main factor in X chromosome inactivation (XCI). XCI is the process through which double dosage from genes in the X chromosome is prevented by random inactivation one of the copies in female cells. Tsix is a repressor of Xist that acts in *cis*, ensuring that only one of the X chromosome copies is silenced (Penny et al. 1996, Lee et al. 1999). Other genes involved in this process, such as Slc1a2 and Tsx, were also misregulated, supporting the identification of XCI as the most enriched GO term in misregulated non-coding genes (Figure 8c). Misregulation of several genes with a role in XCI in ESC mutants with decreased expression of Med12, suggests a potential role for this subunit in the regulation of XCI (Figure 8d). XCI prevents X chromosome double dosage, a problem arising in female cells which contain two copies of the X chromosome. In male cells that contain only one copy of the X chromosome, as is the case of the G4 hybrid cells used in this study, this process should not be active. In order to follow up this potential role for Med12 in XCI a more adequate system should be used, such as Med12 depletion in female ESC, where XCI is induced upon differentiation.

The fact that only two GO terms were found enriched in the misregulated non-coding genes, while for protein coding genes an extensive list was obtain can be explained by the reduced number of ncRNAs that are associated with GO terms. Although thousands of non-coding genes have been described over the years, only for a small subset was their function determined. As such no Go term is associated to the

vast majority of identified ncRNAs, explaining not only the reduced GO term list obtained but also the high p-value associated with them (Table 1).

Multiple other non-coding genes without a known function were also misregulated in at least one of the Med12 mutant cells (Figure 8a), including putative new genes assembled during this study. Assembly of novel genes was performed by Cuffllinks which identified thousands of new transcript (Trapnell et al. 2012). The assembled transcripts were filtered in order to select only those with a higher probability of representing true expressed transcripts. For that reason, only predictions longer than 200 nt and with multiple exons were kept in the analysis. Of the 400 predictions kept after filtering, 38 were found misregulated in at least one of the Med12 mutant ESC. A visual inspection of Cufflinks predictions and mapped reads distribution in the analysed clones for these 38 genes was performed. For 11 of the novel genes, mapped reads supported the Cufflinks predictions and as such their expression was tested *in vivo* using cDNA from WT ESCs. Using primers along the predicted exons, transcripts were amplified for all genes and sequenced (Figure 9 and Supplementary Figure 1). The isolation of these transcripts demonstrated that they represented real genes expressed in ESCs and confirmed the predictions made by Cufflinks, including the multiple isoforms for some of the predictions (Figure 9 and Supplementary Figure 1). The fact that all selected novel genes where expressed *in vivo* confirmed that the filtering criteria applied to the transcripts predicted by Cufflinks, followed by the visual inspection comparing them to the mapped reads was sufficient to select predictions that reflected real genes and not artefacts. The identified gene structure for the 11 novel genes was in most cases similar to the Cufflinks predicted one. As such, although the vast majority of predicted assemble transcripts failed the filtering criteria, for the selected ones Cufflinks reliably predicted their structure.

Recent reports have shown that small functional peptides can originate from genes previously characterized as lncRNAS (Nelson et al. 2016). In order to verify if any of the 11 novel genes could represent a protein coding gene, the coding probability of the isolated isoforms was verified. Using the tool CPAT, which calculates coding probability from the transcripts sequence, transcripts with a coding probability below 0.44 are very unlikely to code for any peptide (Wang et al. 2013). This was the case case for all analysed transcripts, supporting the hypothesis that these represented novel lncRNAs (Table 2).

Having confirmed the expression of these 11 novel lncRNAs in ESCs (Figure 9 and Supplementary Figure 1) their possible expression pattern during embryonic development was assessed using whole mount *in situ* hybridization (WISH). For this method, WT mouse embryos at E9.5, E10.5 and E11.5 were analysed, using the isolated transcripts (Figure 9 and Supplementary Figure 1) as template for generating

antisense RNA probes. For the majority of analysed genes, no specific expression pattern was observed in the tested embryos (Figure 10a). The lack of a detectable specific pattern could be explained in numerous ways. On one hand, the expression could be very low, which is usually observed for lncRNAS, and resulting signal too weak to be detected (Cabili et al. 2011). On the other hand, the transcripts could be expressed at different developmental stages other than the ones analysed. Furthermore, their expression domain could be restricted to a small subset of cells and as such the signal generated was too low. These genes could also be expressed in an internal tissue which is not easily observed when using whole embryos. Finally, the identified transcripts could be ESC specific and not expressed in any differentiated tissues of the embryo. Despite the lack of a specific signal, some signal was observed in the analysed embryos (Figure 10a). However, it was not specific and resulted from either trapping of the reagents in diverse mouse structures, such as observed in the brain vesicles or in the otic vesicles, or from deposition of the staining subtract on the embryos surface. Nevertheless, for one of the novel genes designated as LN-BP18, a specific expression in limbs and caudal end of E10.5 and E11.5 embryos was detected (Figure 10b, bottom row). This new gene was located on chromosome 8, 9 kb downstream of Sall1 which was closest gene. Since multiple lncRNAs were reported to have an effect in the expression of their closest neighbour, Sall1 expression was also analysed (Anderson et al. 2016, Ritter et al. 2019). In the analysed Med12 mutant cells LN-BP18 was upregulated 3-fold and Sall1 was 4-fold downregulated and WISH performed for Sall1 transcripts revealed a similar expression pattern as LN-BP18 (Figure 10b and d). Using publicly available RNA-seq data for diverse embryonic tissues at multiple developmental stages, the spatial and temporal expression of these two genes was further analysed. These data revealed that they were co-expressed in the analysed tissues, with the highest expression detected in embryonic kidneys. Both genes were also enriched in limbs and neural tube and to a lesser extent in forebrain, midbrain and hindbrain. Genes that are co-expressed across developmental time suggest a shared function and/or regulation (Figure 10c). In accordance with this possibility, both genes were misregulated in Med12 deplete ESCs (Figure 10d), supporting a interaction between them.

Despite the fact that LN-BP18 has not been characterized before, a previous study revealed that its locus was co-expressed with Sall1 in early nephrons of E15.5 mice embryos (Thiagarajan et al. 2011). However no attempt was done in identifying the gene in this locus, or in testing its expression in different tissues and/or developmental stages. Furthermore, the only mention of a gene in this locus is a NCBI automatic gene prediction termed Gm3134, which was based on data from over 120 mouse cells and tissues (Yue et al. 2014), without any form of experimental validation. As such, despite hints suggesting the existence of LN-BP18 transcript, no clear proof had been previously produced. Chromatin

immunoprecipitation sequencing (ChIP-seq) data allows identifying the genomic location of the desired protein. By using antibodies specific for a certain histone modification, their presence can be determined genome-wide. In active genes, their TSS is usually flanked by nucleosomes containing the histone marks H3K4me3 and H327ac. Analysis of publicly available Chip-seq data for these histone modifications supported the LN-BP18 structure that was experimentally obtained, since H3K27ac and H3K4me3 where found flanking the predicted transcription start site (TSS) of LN-BP18 (Figure 11, blue track). These marks were also found flanking the predicted TSS for Gm3134, also supporting this prediction. However, except for the exons that overlapped the LN-BP18 exons, no transcription was detected for the Gm3134 exons in the Med12 mutant and WT cells analysed (Figure 11, red tracks). As already mentioned, the Gm3134 prediction was based in data generated from of dozens of different cells and tissues, making it possible that Gm3134 represented the true LN-BP18 structure in other cells. However, due to the discrepancies between the predicted gene structure of Gm3134, and the verified gene structure of LN-BP18, these two genes were considered as distinct throughout this thesis.

Interestingly, while mutations on MED12 have been associated with different syndromes, SALL1 mutations are associated with the autosomal dominant disorder Townes-Brocks syndrome (TBS) (Kohlhase 1993). This syndrome is characterized by anorectal, ear and thumbs malformations and frequently by heart and kidneys defects. Mental retardation is another recurrent defect, occurring in about 10% of the patients. Most of enumerated TBS defects have also been associated with Med12 related XLID disorders, revealing a possible disruption of similar genes in both cases (Risheg et al. 2007, Schwartz et al. 2007, Vulto-van Silfhout et al. 2013). The overlapping characteristics described for the different syndromes, the effect of Med12 depletion on the expression of Sall1 and LN-BP18 in ESCs and the observation that the expression profile of these two genes was very similar led to a more detailed characterization of this novel lncRNA and of its possible interaction with Sall1.

4.3. LN-BP18 presents a complex gene structure

The gene structure of LN-BP18 predicted by Cufflinks (Figure 9, black track) was experimentally validated using WT ESC cDNA (Figure 9, blue track), with the four exons identified spliced in two different variants. However, RNA-seq data generated from Med12 mutant ESCs suggested that some of the exons were actually longer then what was identified (Figure 9, red tracks). Additionally, one of the identified exons was not supported by the analysed transcriptome data, since no reads were mapped to it. As such, a more detailed characterization of this novel gene structure was performed.

lncRNAs share multiple features with mRNA, such as addition of a cap to their 5' end, that confers stability to the RNA and protects it from exonucleases degradation. They can additionally be processed and a poly adenylated tail (poly-A tail), which consists of hundreds of adenine bases in tandem, added to their 3' end. Most of the identified lncRNAs are also spliced, with an average of ~2.3 isoforms per gene found for human lncRNAs (Guttman et al. 2009, Cabili et al. 2011). In order to identify the full length of LN-BP18 transcript, the transcription start site (TSS) and transcription end site (TES) were determined. To identify the TSS of LN-BP18, a 5' RACE approach was applied to WT ESC cDNA. With this method the TSS was identified in an additional exon 10 kb upstream of the previously identified first exon, in the first intron of Sall1 (Figure 12, "TSS1"). Further 5' RACE experiments using primers in the newly identified exon confirmed the presence of this TSS. Additional 5' RACE experiments identified an alternative TSS, located slightly upstream of the first exon predicted by Cufflinks (Figure 12, "TSS2"). These data suggested two alternative TSSs for LN-BP18. In order to validate the identified TSSs, additional data was analysed. As mentioned before, the TSS of active genes is usually flanked by nucleosomes containing the histone marks H3K4me3 and H3K27ac. Additionally, in order for the genes to be expressed the pre-initiation complex (PIC) must be assembled near the TSS and chromatin in this region must be accessible to allow binding of the different factors. This accessibility can be verified using DNaseI hypersensitivity sites sequencing (DNase-seq) which identifies open chromatin regions based on their sensitivity to DNaseI digestion. With cap analysis gene expression (CAGE), the 5' end of RNAs is sequenced and mapped genome-wide. With these data, the genomic location from where transcription started can be identified. Data from the Fantom consortium combines CAGE data for multiple mouse cells and tissues and the peaks present in these data represented potential TSS. Furthermore, since lncRNAs are transcribed by Pol II, ChIP-seq data for this factor indicates where transcription is active or poised. Using the mentioned data, obtained from public databases for ESC (Figure 12, green coloured data) and different embryonic tissues from which E14.5 brain tissue data was representative (Figure 12, blue coloured data), both identified TSS for LN-BP18 were characterized. The TSS1 of LN-BP18 was flanked by an enrichment of H3K4me3 and H3K2ac histone modifications, with Pol II found binding in this TSS. Additionally, this region consisted of open chromatin, as assessed by DNase sequencing and a CAGE peak was present near the TSS1. All of these features were observed in ESC and also in differentiated tissues, confirming TSS1 as a true TSS in all analysed cell types. In contrast, although the same features were found confirming TSS2, this was only observed in in ESC. In differentiated tissues, no enrichment of active histone marks was observed flanking this TSS, Pol II was not bound in this region and the chromatin was condensed. These features suggest that while TSS1 is active in all cell types, TSS2 is active

only in ESC, with transcription from TSS2repressed upon differentiation. These observations suggest that the expression of LN-BP18 observed in different tissues by WISH or by RNA-seq data results from transcription exclusively from the TSS1.

Having identified two distinct TSS for LN-BP18, its TES was identified with a 3' RACE approach using RNA extracted from WT ESC. The first step in this method is to reverse transcribe RNA using an oligo dT primer. As such, only polyadenylated RNAs can be used with this approach. In order to confirm the presence of a poly-A tail in LN-BP18 transcripts, RNA extracted from WT ESC was used to generate cDNA with oligo dT primer. With this cDNA, gene specific primers (GSP) for LN-BP18 were used in order to amplify fragments of LN-BP18. Amplification of LN-BP18 fragments from cDNA generated using oligo dt primer, which should reverse transcribe only polyadenylated RNA, suggested the presence of a poly-A tail in LN-BP18 transcripts (data not shown). Using the 3' RACE approach with WT ESCs RNA, a stretch of 27 adenines in tandem was located in the last exon (Figure 13a). PCR amplification using primers flanking this repetitive region on cDNA generated with random hexamers or with oligo dT primer revealed that the oligo dT primer used during the first step of the 3' RACE method bound to two distinct region: the poly-A tail and the stretch of adenines present in the sequence of the last exon (Figure 13a, black arrows). This caused the generation of two cDNA fragments from LN-BP18 transcripts: one from the poly-A tail up to the stretch of adenines identified and another from this repetitive region to the 5' end of the original RNA. This hypothesis was supported by the lack of amplified fragments using cDNA generated with oligo dT and primers flanking the stretch of 27 adenines (Figure 13a, red arrows and Figure 13b). Repeating the 3' RACE experiment using primers downstream of the repetitive region (Figure 13a, blue arrows) allowed to identify the true TES of LN-BP18 (Figure 12). A consensus polyadenylation signal (AATAAA) was identify in this genomic location, further supporting the identified TES and the presence of a poly-A tail in LN-BP18 transcripts (Proudfoot et al. 1976).

The initial isolation of LN-BP18 transcripts from ESC revealed the presence of two distinct isoforms (Figure 9). However, after this initial analysis additional exons and two distinct TSSs have been identified. Analysis of active transcription marks indicated that TSS2 was only active in ESCs (Figure 12), suggesting that all expression of LN-BP18 observed in embryonic tissues, such as in forelimbs, originated exclusively from TSS1. Thus, to identify additional isoforms, cDNA was generated from RNA extracted from ESCs and from forelimbs. Using these cDNAs, LN-BP18 transcripts were amplified by PCR using primers in either TSS and in the TES (Figure 13a, green arrows). Since the initial PCR run resulted in low signals and in the amplification of unspecific fragments, a nested PCR was performed using the original PCR as template and primers adjacent to the first set of primers. Amplified fragments were sequence and their different

splice variants identified. With this methodology, two additional exons and nine full isoforms were identified (Figure 14a). For three of the seven exons (Exon 1, 3 and 4) different splice variants were also identified, with shorter and longer versions of the same exons identified in different isoforms. Fragment LN-BP_010 was obtained using primers in exons 1 and 3 and does not represent a full isoform since there was no information regarding the reaming exons present after exon3. However this fragment was the only instance were exon 2 was identified, and as such as still included.

Analysis of all isoforms by CAPT revealed a lack of coding potential for all isoforms (coding probability < 0.44) (Figure 14b). In order to confirm this observation, PhyloCSF data was also analysed. These data identify evolutionary signatures by aligning different mammal datasets and has been successfully used to identify small peptides originating from genes characterized as lncRNAs (Nelson et al. 2016). A clear signal was obtained in the coding regions of Sall1 (Figure 14c). In contrast, no signal was obtained in the whole locus of LN-BP18, supporting the lack of coding potential for this gene. Since multiple isoforms were identified for LN-BP and these were amplified using a nested PCR, which allows amplification of very lowly expressed transcripts, it was possible that some of identified isoforms resulted from abnormal splicing events that would usually be targeted for degradation (Wery et al. 2016). One way of verifying this possibility is to analyse the splicing acceptor and donor sites in the identified exons (Senapathy et al. 1990). In exons without a consensus donor and/or acceptor sites, the hypothesis that they resulted from aberrant splicing events is supported. Analysing the identified exons, majority of them contained the mouse consensus acceptor and donor sites (Figure 14d). The exceptions were a smaller variant of exon 1 which was only found in isoform LN-BP18_005, exon 2 that is present only in the fragment LN-BP18_010 and the acceptor site for exon 3 only found in the same fragment. These exons were identified in a single transcript, contrary to the remaining exons that were found in multiple isoforms. These data suggest that isoform LN-BP18_005 and the fragment LN-BP18_010 are not true LN-BP18 isoforms but that in fact are aberrant splice variants. The lack of an identified isoform containing both exon 1 and exon 3 reveals that transcripts originating from TSS1 do not retain exon 3, where the TSS2 is located.

Comparing the predicted Gm3134 and LN-BP18 revealed a similar structure, with a predicted splice pattern as complex as what was observed for LN-BP18 (Figure 14). The majority of exons were common between LN-BP18 and Gm3134, although in some cases the described length was very different. These discrepancies were observed for exon 5 of LN-BP18 which was 170 bp and in Gm3134 was predicted to 1800 bp long. Other exons were present in only one of the genes, such as exon 6 of LN-BP18. The similarity between the predicted Gm3134 and the experimentally validated LN-BP18 revealed that the automatic predictions can be very similar to the true gene structure, hence representing a good starting

point for the study of the uncharacterized genes. Although some of Gm3134 features were not observed, such as some exons, it is possible that they could be validated using tissues other than the ones used in this study. As mentioned before, human lncRNAs show an average of ~2.3 isoforms (Cabili et al. 2011). As such, the splicing pattern identified for LN-BP18 is far more complex that what is usual for lncRNAs. Yet, previous reports have identified lncRNAs with an even more complex pattern, such as GNG12 antisense 1 (GNG12-AS1), a lncRNA with 10 exons that can be spliced into 38 different isoforms (Niemczyk et al. 2013). This reveals that the complex splicing pattern of LN-BP18 was above average but still within what has been found for other lncRNAs.

In order to identify possible conserved LN-BP18 transcripts in other species, BLAST searches using the full sequence of the isoforms, together with identified open reading frames (ORFs) or putative proteins coded by the ORFS found on each isoform revealed no homologous transcripts in other species. However, an uncharacterized lncRNA termed AC087564.1 has been predicted in humans. Similarly to LN-BP18, this predicted ncRNA is divergent of Sall1. Alignment of the isoforms of this predicted ncRNA and isoforms identified for LN-BP18 confirmed the lack of conservation between the two genes. The calculated identity between the isoforms of both genes was on the same levels as compared to completely unrelated lncRNAs, such as Fendrr and Hotair (Figure 15). Multiple lncRNAS have demonstrated poor sequence conservation but conserved function. Indeed, Gas5 is a lncRNA that acts as a regulator of self-renewal and pluripotency in mouse ESC and in human induced pluripotent stem cells, despite poor sequence conservation between the two species (Tu et al. 2018). The divergent position regarding Sall1 supports a relation between LN-BP18 and AC087564.1, which might translate in similar functions despite the lack of sequence conservation. However, to confirm this hypothesis a more detailed characterization of AC087564.1 is necessary.

4.4. LN-BP18 is dynamically expressed *in vivo*

The generated whole mount *in situ* hybridization (WISH) data allowed to detect LN-BP18 expression in forelimbs, hind limbs and in the caudal end of E10.5 and E11.5 mice embryos. Publicly available transcriptome data confirmed that LN-BP18 was enriched in these tissues and revealed additional expression in kidney and in neural tube, with lower expression in brain. A silar expression pattern was observed for Sall1, co-expression and/or similar functions (Figure 10b, c). Considering the identified TSSs for LN-BP18 and the observation that TSS2 showed a ESCs specific activation, the lncRNA expression was evaluated in different contexts. First it was quantified in nuclear and cytoplasmic fractions of ESCs, since

the cellular localization of a lncRNA can provide hints for their function. For example, Xist plays a role in transcription repression and so is found enriched in the nucleus (Penny et al. 1996). When LN-BP18 expression was quantified in these two fractions no clear enrichment could be observed for transcripts originating from any of the TSSs (Figure 16b). Since the analysed RNA-seq data revealed a very low expression of LN-BP18 in ESCs, with a PFKM of 2.6 (Figure 10d), it is possible that expression levels were not sufficient to properly detect enrichment in any of the two analysed fractions. It was also possible the LN-BP18 was located at similar levels in both fractions, with different function in each of them. To properly identify the cellular location of LNBP18, a more robust and reliable approach would be to use RNA fluorescence in situ hybridization (RNA-FISH) with labelled probes against LN-BP18 (Femino et al. 1998). With this method, it would be possible to determine LN-BP18 localization on ESCs despite its low expression, as previously applied to other lowly expressed lncRNAs (Ritter et al. 2019).

By qPCR quantification LN-BP18 was found to be enriched 4 to 8-fold in forelimbs, hind limbs and the caudal end of E11.5 embryos compared to the remaining analysed tissues (Figure 16d), confirming the expression pattern detected with WISH (Figure 10b). Interestingly and in accordance with the hypothesis that TSS2 represented a ESCs specific start site, no transcription initiating from TSS2 could be detected in any of the embryonic tissues analysed (Figure 16d). This confirmed that in differentiated tissues, TSS1 is the main start site used for LN-BP18 transcription. Quantification of LN-BP18 in Med12^{null} ESCs, revealed that LN-BP18 overall levels and transcription from TSS2 where not affected by Med12 deficiency. In contrast, transcripts from TSS1 were 4-fold downregulated in these cells (Figure 16c). The strong downregulation of LN-BP18 TSS1 while its overall levels and TSS2 transcription remained unchanged revealed that the main TSS for LN-BP18 expression in ESC is the TSS2, contrary to what was observed in differentiated tissues (Figure 16d). This observation was confirmed by the 40-fold lower abundancy of TSS1 transcripts compared to TSS2 transcripts in G4 cells, as calculated by the qPCR data obtained (data not shown). In mutant ESC expressing a mutated version of Med12 associated with Opitz-Kaveggia syndrome (R961W), a similar but milder effect was observed. In this clone, LN-BP18 TSS1 transcripts were 2-fold downregulated while no change in transcription from TSS2 was detected (Figure 16c). The downregulation of TSS1 but not of TSS2 transcripts in the different Med12 mutant ESCs analysed suggested different mechanisms regulating expression from each TSS. These data suggested that activation of LN-BP18 transcription involved a mechanism mediated by Med12, while activation of TSS2 transcript was Med12-independent. Furthermore, Sall1 was also found downregulated in the Med12 mutant ESCs, with a stronger downregulation observed in the Med12^{null} cells compared to Med12-Opitz cells. While TSS2 transcripts were not affected, TSS1 and Sall1 were similarly downregulated in both

mutants. This suggested a co-regulation between LN-BP18 TSS1 and Sall1, which might correlate with shared function.

The downregulation of Sall1 observed in Med12-Opitz mutant ESCs (Figure 16c) suggested a role for Sall1 in at least some of the defects observed in patients with Med12 associated pathologies, as is the case of the Opitz-Kaveggia syndrome. Sall1 functions on heart development, limb morphogenesis and is critical for proper kidney formation (Nishinakamura et al. 2005, Kawakami et al. 2009, Morita et al. 2016). In Med12 associated intellectual disabilities, defects in all of these tissues have been described, further supporting a disrupted function of Sall1 in these patients. Several mutant ESC and mouse expressing mutated Med12 versions associated with Opitz-Kaveggia, Lujan-Fryns and Ohdo intellectual disabilities syndromes have been generated by the Schrewe group and represent an extremely useful tool to further study the mechanism behind the observed phenotypes in human patients. In Med12-Opitz mutant cells, LN-BP18 transcription from TSS1 was also found misregulated. The effects caused by the Med12 mutated protein on LN-BP18 expression suggest a possible functional role for this lncRNA in pathologies. The confirmation of LN-BP18 misregulation in the other mentioned Med12 mutant ESCs and especially in mouse lines expressing these mutated Med12 would strengthen the claim of such functional role.

The methods used to evaluate LN-BP18 expression, present their own limitations. Although WISH allows determining the spatial and temporal expression of a gene, it lacks sensitivity which makes it less useful for studying genes lowly expressed, as is the case of LN-BP18 and lncRNAs in general. In contrast, qPCR has a great sensitivity and allows to quantify even lowly expressed transcripts, however its capacity to detect specific expression domains during embryo development is dependent on how finely dissected the analysed sample is. The drawbacks for qPCR are also true for data generated by RNA-seq. The generation of a reporter line, in which a gene that generates a signal quantifiable even at low levels and under the control of the same regulatory elements as LN-BP18, allows to overcome the mentioned limitations. Furthermore, by generating embryos with these reporter cells it is possible to assess the temporal and spatial expression pattern of a gene with high sensitivity by detecting the signal generated by the reporter cassette. To generate a reporter line for LN-BP18, a beta-galactosidade (β-gal) reporter cassette was knocked-in into exon 4 of LN-BP18. Additionally, a triple polyadenylation signal (3x pA) was included downstream of the reporter cassette. The generated transgene should be expressed only when LN-BP18 expression is induced, reflecting this lncRNA true expression pattern. Due to the insertion of the 3x pA, transcription of the remaining LN-BP18 transcripts should be prevented, resulting in the expression of only truncated transcripts. This disturbance of the normal LN-BP18 might result in phenotypes in embryos generated from these reporter cells. Using CRISPR-Caas9 to insert this reporter

cassette into LN-BP18 exon 4 through homologous recombination (Figure 17a) in G4 hybrid cells allowed the generation of heterozygous ESC clones, which showed successful transformation of one of the alleles (Figure 17b and c).

X-gal staining of heterozygous reporter ESCs cultures revealed no evidence of active reporter (data not shown), even though expression of the reporter cassette expression could be detected by qPCR (Figure 19d). The absence of reporter activity could have resulted from lack of necessary elements for proper assembly of translational machinery or from adoption of complex structures by transcript which prevented proper ribosomal function. The reporter could also be active but at such small levels that could not be properly detected. To test the activity of the reporter cassette, additional systems were tested. A previous study has generated RNA-seq data for mesodermal progenitors cells (NMPs) that co-expressed Brachyury (T) and Sox2, as well as for cells expressing either T or Sox2 (Koch et al. 2017). These populations were sorted from the caudal end of E8.5 embryos and analysis of their transcriptome allowed to detect expression of LN-BP18 (Figure 19b).

To verify if reporter activity could be detected in these cells, a previously published protocol was used to differentiate ESCs (Figure 19b). By day 3 of the *in vitro* differentiation protocol cells were transcriptionally similar to the NMP population and by day 5 were similar to cells expressing T but not Sox2 in the caudal end of E8.5 embryos (Gouti et al. 2014, Koch et al. 2017). This protocol was applied to two different β-gal heterozygous clones and expression quantified throughout the course of differentiation. Despite the 2-fold upregulation of the reporter cassette after five days of differentiation (Figure 19c), no reporter activity was observed after X-Gal staining in any time point (data not shown). These data supported the hypothesis of an inefficient translation of the cassette, since the expression of the cassette was verified to be upregulated. However it could still be that despite the observed 2-fold increased expression, the expression was still too low to be detected. Despite the lack of reporter activity, quantification of LN-BP18 expression throughout the differentiation protocol revealed that transcription from TSS1 increased from day 4 up to 8-fold by day 5. On the other hand, TSS2 transcription decreased up to 10-fold from day 3 to day 5 of differentiation. This dynamic confirmed that TSS2 is active mainly in ESCs and that upon differentiation, transcription is repressed from this site while from TSS1 is activated.

In the G4 (129S6/C57BL6) hybrid cells used, a mutation in the PAM sequence was identified in the 129S6 allele (Figure 17d). This mutation prevented the Cas9 nuclease to generate the intended DSB in the 129S6 allele and as such, using the G4 cells only heterozygous clones could be obtained. Since the transformation protocol was successful in transforming the allele with C57BL6 background, JM8 ESC were used and the same protocol applied. In these cells, that have a C57BL6 background, homozygous

and heterozygous transformed clones were successfully generated (Figure 18b). Analysis by qPCR on these clones revealed that transcription after the 3x pA was 6-fold decreased in homozygous clones, while on heterozygous clones it was 2-fold decreased (Figure 19d). These data support the high efficiency of the inserted 3x pA signal in stopping transcription. Since the stop cassette was inserted into exon 4, and the primers used to quantify TSS1 and TSS2 isoforms as well as the overall LN-BP18 levels were located before the inserted cassette, it was still possible to quantify expression of these transcripts. In the reporter clones, no changes were observed on LN-BP18 TSS2 transcription and on Sall1 expression. In contrast, a 6-fold upregulation of TSS1 transcripts was observed (Figure 19e). In ESCs, even with this upregulation, the TSS1 transcripts accounted for less than 6% of the total LN-BP18 transcripts. However, since TSS1 is active mainly in differentiated cells, it is possible that upon differentiation the levels of LN-BP18 originating from TSS1 in the reporter line would be higher than normal due to this effect. As such a proper WT control should be used to account this possible variation.

Analysis of cultured mutant ESC transformed with the β-gal reporter construct revealed no activity of the reporter. Additionally, no reporter activity was detectable when selected G4 derived heterozygous clones were *in vitro* differentiated into mesodermal progenitors cells (NMPs) or paraxial mesoderm, cell types which express LN-BP18 (Figure 19a). However, qPCR data for clones successfully transformed with the β-gal reporter cassette revealed that the inserted cassette was expressed as part of LN-BP18 transcript (Figure 19d). In order to verify if the expressed cassette was functional *in vivo*, embryos were generated from G4 derived LN-BP18-β-gal heterozygous clones through tetraploid complementation assay. Preliminary data obtained for one E11.5 embryo generated from clone C5 (Figure 17b) revealed that the inserted cassette is functional, since a specific signal could be detected in this embryo (Figure 20). Confirming the data from the WISH analysis (Figure 10b), a strong expression was observed in the mesenchyme of forelimb (FL) and hind limb (HL) buds and in the caudal end (CE), where a strong signal was also observed in late somites (S) (Figure 20). The strong reporter activity was restricted to late somites, since the signal in early somites was significantly weaker suggesting a potential role in axis elongation. Expression in the caudal end confirms the observation that LN-BP18 was expressed in cells isolated from the caudal end of E8.5 embryos expressing T, Sox2 or both (Figure 19a) (Koch et al. 2017). Expression of LN-BP18 in NMPs further revealed that LN-BP18 is expressed as early as E8.5 at the caudal end and possibly in the other tissues where expression was detected. In agreement with the RNA-seq data for different embryonic tissues (Figure 10c), expression was identified at the forebrain, midbrain and hindbrain, specifically at the boundaries that separate these structures (Figure 19 , MHB and MFB). Sall1 is an essential regulator of kidney development and since the data described before indicated that

Sall1 and LN-BP18 are co-expressed (Figure 10b and c), LN-BP18 was also expected to be expressed in these organs, a hypothesis supported by the RNA-seq data (Figure 10c). Although at E11.5 embryos the kidneys have not started to form, a staining occurred at the intermediate mesoderm, more specifically at pronephros. These structures correspond to the first of three stages of kidney development in mammals, confirming the co-expression of Sall1 and LN-BP18 in these organs. As kidneys, the known role of Sall1 in neural tube closure (Bohm et al. 2008) and the observed co-expression of this gene and LN-BP18 suggested expression of the lncRNA in this tissue, supported by RNA-seq data, in which a higher expression of LN-BP18 observed in this tissue (Figure 10c). Expression of LN-BP18 in the neural tube was confirmed with the reporter embryo, with a clear signal observed along neural tube (Figure 19, NT). A clear signal was also obtained in axial structures, which can originated from notochord, floor plate or both (Figure 19 , AS). However, due to their similar position and shape in the embryo, it is not possible to distinguish in which of these structures the signal is generated. For this distinction, a more detailed analysis is necessary, such as embedding the embryo and performing transversal sections of the spinal cord. Imaging these sections should present clear evidence of which of these two structures revealed reporter activity. Other expression domains with a lower signal have been additionally identified. These included the genital region (Figure 19, GR) and the heart (not shown), with the latter being supported by the RNA-seq data analysed (Figure 10c). Although no expression data are presented in this study to support expression in the genital region, Sall1 has been shown to be expressed in the genital tubercle in E11.5 mice embryos (Buck et al. 2001). Additionally, anorectal malformation are a common defect observed in patients with Townes-Brocks syndrome (TBS) (Townes et al. 1972). Together with the observed co-expression of Sall1 and Pn-BP18, These data support LN-BP18 expression in the genital region of embryos. No clear phenotype was observed in this mutant, despite the disruption of LN-BP18 expression due to the presence of a stop cassette downstream of the β-gal cassette. It is possible that since this was an heterozygous clone, LN-BP18 expression of the WT allele was enough to mitigate any possible defects arising from the disruption of the transformed allele. In order to verify this hypothesis or if no defects would be observed even if both alleles were disrupted, embryos should be generated from the JM8 derived LN-BP18-β-gal homozygous reporter cells.

Disruption of LN-BP18 expression from TSS1 in Med12 mutant cells (Figure 16c) and the observed expression in multiple tissues where defects were detected in previously reported Med12 deficient embryos (Figure 20) (Rocha et al. 2010) suggest a potential role for LN-BP18 in these defects. Analysis of LN-BP18 expression in the Med12 deficient embryos would allow verifying if this lncRNA is misregulated in these embryos, supporting a functional role for LN-BP18 and mediation of its expression by Med12.

4.5. LN-BP18 correlation with coding genes suggest a possible role in pluripotency

While the β-gal reporter line generated for LN-BP18 disrupted its expression, up to 40% of the lncRNA was still expressed since the stop cassette was inserted in exon 4. In order to generate cells that do not express LN-BP18, different mutants were created by complete excision of one of the two TSSs using CRISPR-Cas9 (Figure 21a). Analysis of LN-BP18 expression in TSS1-KO clones revealed one of the clones as abnormal (1H), since it was the only clone with an increased expression of TSS1 transcripts. This might have results from some contamination with other cells, or from unpredicted genomic rearrangements in the LN-BP18 locus that were not detected. As such, this clone was excluded from further analysis. Clone 11E was also excluded from the analysis since it was the only clone with a notorious downregulation of LN-BP18 TSS2 transcripts and Sall1 expression (Figure 22a and b). For the remaining homozygous KO clones, no transcripts originating from TSS1 could be detected. In heterozygous deletions, some of the isoforms (e.g. LN-BP18_001) were downregulated 2-fold and others (e.g. LN-BP18_003) were not detectable (Figure 22a). As observed before, in ESCs the TSS2 was the main active transcription site and transcripts from TSS1 were near the detection limit. As such it is possible that a small decrease of these transcripts would result in levels below the detection limit. This would lead to a lack of quantifiable signal even in the heterozygous clones. The deletion in these clones specifically affected TSS1 transcripts since there was no significant effect on the TSS2 transcripts in ESCs (Figure 22a). As TSS2 is the main transcription site for LN-BP18 in ESCs, no effect on overall levels of LN-BP18 was observed in TSS1 mutant ESC. For TSS2 deletion, due to the small efficiency of transformation, only heterozygous clones were obtained. On these, the opposite effect was observed. No changes were detected in TSS1 transcripts, while TSS2 and overall levels were 2-fold decreased. For both TSS1 and TSS2 mutants, no changes were observed on Sall1 expression (Figure 22c). This suggested that LN-BP18 does not affect Sall1 expression, since halving LN-BP18 levels by deleting one copy of TSS2 had no effect on Sall1 expression (Figure 22b).

Curiously, upon TSS1 deletion, expression of Nanog and Oct4, two important pluripotency genes, was 2-fold increased in these clones (Figure 22b). In TSS1-KO mutant cells, a positive correlation was observed between LN-BP18 overall expression and Nanog and Oct4 gene expression (Figure 22e). The correlation between the genes suggested that they were either co-regulated or that one of the genes regulated the others. An even stronger correlation was observed between these two pluripotency genes and TSS2 transcripts (Figure 22f). Since these transcripts are ESC specific, their correlation with two

important pluripotency regulators suggested a function in the pluripotency network for LN-BP18 transcripts originating from TSS2. This hypothesis would explain why these isoforms are repressed upon differentiation, since during this process the pluripotency network is suppressed in favour of differentiation programs. Sall1 has been associated with maintenance of pluripotency and its depletion with siRNAs resulted in Nanog downregulation in ESCs (Karantzali et al. 2011). In TSS1-KO mutant ESCs, Sall1 expression was also correlated Nanog and Oct4 levels (Figure 22d). Since both LN-BP18 and Sall1 were correlated with pluripotency regulators, this suggested a shared function for these antisense neighbours in the maintenance of pluripotency network in Esc. In contrast, it is also possible that only one acts on the pluripotency network while simultaneously activating the expression of its neighbour.

4.6. Sall1 depletion in ESC suggests an activation function on LN-BP18 expression and supports a role for the lncRNA in pluripotency

Sall1 has been linked to pluripotency in a previous study by the Kretsovali laboratory (Karantzali et al. 2011). In this study, depletion of Sall1 by siRNAs in ESCs resulted in a 2-fold downregulation of Nanog, a key regulator of pluripotency. A 2-fold upregulation of Hand1 and T, two important differentiation markers was also observed upon Sall1 depletion. This study suggested a double role for Sall1 in the ESC. It induced Nanog expression and by consequence expression of Nanog target genes in ESCs, while also repressing differentiation markers. In order to untangle the effect of Sall1 and LN-BP18 in pluripotency maintenance and to better understand the relation between these two genes, Sall1 deficient ESCs were generated by excision of most of Sall1 coding region using the CRISPR-Cas9 method (Figure 23b). Despite the very precise induction of double strand breaks (DSB) by the Cas9 nuclease, the repair mechanism can repair these breaks in different ways, resulting in variations between clones, as observed by the presence of bands of different sizes when screening LN-BP18-β-gal transformed colonies (Figure 17b). Additionally, both alleles of Sall1 homozygous KO clone 12D had a deletion of similar size, but for one of them a shit of the deleted region around 400bo was observed (Figure 23a and c). For the second Sall1 homozygous clone, it was observed a complete inversion of the whole region that should have been excised, a phenomenon observed in previous studies and in some cases taken advantage of in order to generate inversions or even duplications in the genome (Korablev et al. 2017). These observations revealed that a proper screening strategy is necessary in order to detect the occurrence of such anomalous events. Expression analysis in the generated Sall1-KO ESCs revealed a downregulation of Sall1 by 2 to 4-fold in heterozygous mutants (Figure 24a). In contrast to what was reported by the Kretsovali

group, no effect was observed in Nanog expression (Figure 24a). Even on homozygous mutated clones, where no Sall1 could be quantified, Nanog levels were unchanged (Figure 23b, Figure 24a). Additionally, despite the report that Sall1 depletion did not affect Oct4 expression, a small downregulation of this gene was observed in the generated mutants. Sall1 repressive effect on differentiation markers could be confirmed for Hand1, with levels of this gene upregulated over 4-fold upon Sall1 depletion. However, for T only a small increase, below 2-fold, was observed (Figure 24a, b). In the Kretsovali study, downregulation of Sall1 was achieved by siRNAs, which are known to potentially have severe off-target effects. However, the Sall1 depleted mutants here described were generated with a site specific deletion, with a careful selection of guide sequence which showed the less probability of off-target effects. As such, it is possible that the detected effects on T and Nanog expression levels observed in the Kretsovali study resulted from off-target effects by the used siRNAs. The unspecific targeting by the used siRNA was further supported by the observation that depleting Sall1 in ESCs did not affect Nanog and T expression. Med12 mutant ESCs showed a downregulation of Sall1 and expression changes in Nanog and T similar to the described in the Kretsovali study (Figure 16c). This observation suggested that the siRNA targeted unspecifically possible regulators of Sall1 that could also regulate expression of Nanog. Additionally, clones without Sall1 grew at similar rates as WT cells, without any notorious defects in morphology, and without evidence of differentiation (data not shown), suggesting that Sall1 was not essential for the maintenance of pluripotency.

LN-BP18 overall levels, as well as TSS1 and TSS2 transcription were quantified in Sall1 depleted clones, with a slight downregulation observed for all three (Figure 24c). A stronger effect in the TSS1 transcripts was observed, with these 2-fold downregulated. As observed before, deletion of LN-BP18 TSS1, together with most of Sall1 first intron did not affect the expression of the coding gene (Figure 22b). Additionally, when TSS2 was deleted from one allele, decreasing the levels of LN-BP18, Sall1 expression was also unaffected (Figure 22c). However, the deletion performed for Sall1 depletion, which was 9 kb away from LN-BP18 TSS1, resulted in a mild downregulation of this lncRNA, an effect observed for both TSSs and its overall levels (Figure 24c). These data suggested Sall1 as an activator of LN-BP18. However, the LN-BP18 expression in WT ESCs was very low and as such, it is possible that only mild effects resulting from Sall1 deficiency could be detectable. The detected low expression of LN-BP18 in Sall1 deficient ESCs can be due to the basal expression level of this gene without a regulator activating it, or due to a weak active transcription induced by a different factor. Additionally, as quantified by qPCR in Med12^{null} mutants, upon depletion of Med12, Sall1 and LN-BP18 TSS1 were 6 and 4-fold downregulated, respectively, while TSS2 and overall LN-BP18 levels unaffected (Figure 16c). This suggested that Sall1 activates LN-BP18

expression in a mechanism that is dependent of Med12. Additionally it reveals that Sall1 acts mainly on TSS1 transcripts. This activating function would explain the co-expression of Sall1 and LN-BP18 in all observed tissues and the fact that only TSS1 was active in differentiated cells. It would also explain why in all analysed tissues Sall1 expression was on average 6-fold higher than LN-BP18 (Figure 10c), while in WT ESC it was almost 40-fold higher (Figure 10d). This suggests that different mechanisms act in activating transcription for each TSS, with Sall1 potentially activating transcription from TSS1 in a Med12 dependent mechanism. In contrast, the lack of effect on TSS2 transcription in Med12 deficient cells (Figure 16c) suggested a Med12-independent mechanism activating TSS2 transcription.

4.7. Identification of lncRNAs as candidates of Med12 target genes

Analysis of the initial RNA-seq data (section 3.1) for $Med12^{hypo}$ and $Med12^{null}$ mutant ESC confirmed several observations of a previous study where embryos were generated from these cells (Rocha et al. 2010), such as downregulation of canonical Wnt targets and misregulation of genes with a role in heart development and neural tube closure, two striking phenotypes observed in these embryos. From analysis of this RNA-seq data, hundreds of non-coding genes were also found misregulated in these clones, including a novel gene designated as LN-BP18. While from RNA-seq data a 2-fold upregulation of LN-BP18 was observed in $Med12^{null}$ mutants, qPCR data from the same cells revealed that the overall levels of this lncRNA were not affected (Figure 10d and Figure 16c). As mentioned in the beginning of this study, a critical aspect of RNA-seq data is to account for biological and technical variation through the use of replicates. However, in the original RNA-seq data analysed only a replicate per clone was generated. Without replicates from with to assess variation of expression and perform statistical tests on the obtained gene expression, no confidence could be assigned to the observed variations in gene expression between different mutants. For highly expressed genes that showed a big variation in their expression between samples, the chance that such variation would be due to biological or technical variation was very low. However for genes with a less evident change in expression or that were lowly expressed, the chance of wrongly classifying them as misregulated was higher. As such, in order to confirm the results obtained from the analysis of the original RNA-seq data and to generate a list of lncRNAs that are putative targets of Med12, new RNA-seq data was generated using two replicates for each Med12 deficient ESC. Additionally, transcriptome data for the $Med12^{flox}$ clone (Figure 6a), a mutant that showed normal Med12 expression and mice generated from it showed no evident phenotype, was also generated (Rocha et al. 2010). With these new samples, misregulated genes could be identified

considering the fold change between samples but also the false discovery rate (FDR) could be calculated for comparison of gene expression. Additionally, while the original RNA-seq data consisted of paired-end reads of length 50 nt, for the new data paired-end reads of length 75 nt were generated. This longer length increased the likelihood of obtained and successfully mapping spliced reads, which are critical for more robust *de novo* gene assembly. Analyses of this new RNA-seq data also followed a different pipeline than originally used. First, STAR was the aligner used, since with this tool more reads were kept after removing reads not in a proper pair or with a mapping score below 10 (Figure 25a). Retaining more reads after filtering increased the chances of detecting expression and significant changes in lowly expressed genes. GenCODE is one of the databases with a more complete index of identified and predicted non-coding genes (https://www.gencodegenes.org/). Thus, it was included in this analysis, replacing the original RefSeq annotation used. Despite the extensive list of ncRNAs used, Cufflinks was again used to assemble new genes not present in the annotation since it has been shown to reliably predict novel genes, as demonstrated for LN-BP18. After removing small predicted transcripts or with a single exon, this *de novo* assembly resulted in the addition of 420 putative new genes into the analysis. Differential gene analyses was performed with Deseq2, a tool more reliable and sensitive than Cuffdiff and that also associated a false discovery rate (FDR) value to the calculated fold change between samples (Love et al. 2014).

Since sample-to-sample distances revealed that WT and Med12^{flox} were the more similar samples, the latter was used as control for differential gene analysis (Figure 25d). This choice allowed to account for possible effects from the transformation protocol or the presence of Loxp sites used for generation of Med12^{null} and Med12^{hypo} mutants, an effect described in previous studies (Qiu et al. 2011, Heffner et al. 2012). With this new pipeline, 150 annotated non-coding genes and 50 Cufflinks predictions were found misregulated with $|log2FC| >= log2 (1.5)$ and $FDR < 0.05$. Expression of ncRNAs was more similar between both mutants than the expression of coding genes (Figure 26b), contrary to the original data (Figure 8b). The discrepancy between the two datasets could have stemmed from the variability not accounted for on the data without replicates or from the lower threshold used to classify genes as misregulated on this section. On one hand, since with the newer data it was possible to have a statistical confidence associated with the fold change obtained for each gene, a lower threshold could be used. On the other hand, using smaller thresholds for the data in section 3.1 would have resulted in a big number of genes being wrongly classified as misregulated when the fold changes observed resulted from the associated biological or technical variation. Additionally, a new pipeline was used on the new analyses. If this new pipeline had been applied to the original data, differences original gene expression would also

have been observed, since different tools generate slightly different results even when used on the same data. However this new pipeline was performed with more robust and sensitive tools. As such, even though differences would arise from simply using a different pipeline analysis, a higher confidence in the newer results would be obtained.

Over 1800 protein coding genes were found significantly misregulated in $Med12^{hypo}$ and/or in $Med12^{null}$. Gene ontology (GO) term enrichment analysis in the misregulated coding genes revealed several terms that were also enriched in the original RNA-seq data, such as axon guidance and angiogenesis (Table 1 and Figure 26e), supporting the main conclusion of the analysis of the original data. Downregulation of Wnt targets was observed in the originally data and while some, such as Ccnd1 and Tbx6, were also found downregulated in the new data, no significant change of expression was observed for other targets, including T, Axin2 and Sall4. However, Wnt pathway was one of enriched GO terms in misregulated genes observed for the new data (Figure 26e). These data confirm the observation that in both RNA-seq datasets, misregulated genes are associated with the same pathways and biological processes, despite the lack of concordance for some of these genes individually. As such, the new data supports the conclusions reached after analysing the misregulated coding genes in the original data.The RNA-seq data for Med12 mutant cells confirmed the downregulation of Wnt targets observed in embryos generated from these cells.

In a recent study from the Schrewe group (data not published), gastruloids were formed using $Med12^{flox}$, $Med12^{hypo}$ and $Med12^{null}$ mutant ESC according to a previously published protocol (Baillie-Johnson et al. 2015). These gastruloids consisted of ESC aggregates that formed a three dimensional structure and that were able to differentiated and organize into structures resembling what is found in embryos (van den Brink et al. 2014). After activating Wnt signalling, the round gastruloids formed from $Med12^{flox}$ started elongating and germ layer specification into mesoderm was observed in the pole. T expression was detected in this reagions, similarly to what was verified for embryonic axis elongation. Although less efficiently, $Med12^{hypo}$ mutant gastruloids were also able to form these elongated structures. However T expression was severely reduced at the pole. $Med12^{null}$ gastruloids failed to elongate and were retained in their round structure, with no T expression detect. This behaviour resembled what was found in $Med12^{null}$ embryos (Rocha et al. 2010), with no expression of T detected, demonstrating a similar effect for Med12 deficiency both *in vivo* and *in vitro*. One of the GO terms found enriched on misregulated coding genes in Med12 mutant ESC was cell adhesion, a process confirmed to be disturbed in $Med12^{null}$ gastruloids, which formed multiple smaller aggregates, contrary to the $Med12^{flox}$ and $Med12^{hypo}$clones, for which a single aggregate was formed. Data obtained from gastruloids

not only confirmed the developmental role of Med12 previously observed in mouse embryos, but also confirmed observations from the analysis of Med12 deficient ESC transcriptome data.

One aspect not analysed in this study was the presence of a Med12 paralog, designated as Med12L. This gene as not been studied as extensively as Med12. However it is known that its highly similar to Med12 and that its integration in the Mediator complex is mutually exclusive to that of Med12 (Daniels et al. 2013). It is possible that their similar protein sequence and similar position within the Mediator allows some functional overlap between this paralogs. This is supported by a recent study where mutations in Med12L have been identified in individuals with intellectual disability (Nizon et al. 2019). These patients showed additionally abnormalities in the corpus callosum and mild defects in facial morphology, features previously associated with Med12 mutations in humans (Risheg et al. 2007, Schwartz et al. 2007).

The expression of Med12L was found upregulated over 1 fold in $Med12^{hypo}$ and over 2-fold in $Med12^{null}$ mutants, consisting with a potential compensation of Med12 deficiency by Med12l. In order to study the interaction between this two genes and to verify if Med12l does indeed compensate - up to certain degree - Med12 depletion or Med12 represses Med12l expression, which would result in the increased expression of Med12L upon Med12 depletion, will required the generation of Med12l deficient mutants or even double Med12/Med12l deficient mutants.

Go term enrichment analysis revealed two biological processes enriched in the misregulated ncRNAs, repression of viral genome and regulation of methylation (Figure 26f). Interestingly Ftx, which activates Xist by preventing DNA methylation at its promoter, was one of the lncRNAs misregulated in $Med12^{null}$. However, despite inducing Xist expression, in the $Med12^{null}$ mutant, Xist was not misregulated (Figure 26d), suggesting that Ftx activates Xist through a Med12 dependent mechanism. However, Ftx upregulation might have had no effect on Xist expression because this lncRNA acts on X chromosome inactivation in female cells (Penny et al. 1996) and as mentioned before, all the cells used in this study were male.

Over 200 ncRNAs were significantly misregulated in at least one of the Med12 mutant ESC, confirming the importance of Med12 in the proper expression of non-coding genes. However, due to Med12 role in multiple pathways, most of the effects were most likely indirect, resulting from misregulation of regulators of these lncRNAs. In order to identify lncRNAs in which misregulation was directly mediated by Med12, ChIP-seq data for this subunit was obtained for ESCs from a previous study (Kagey et al. 2010). The rationale behind this analysis was if expression of a lncRNA is mediated directly by Med12, then this subunit should be binding at their promoter and/or gene body, which can be identified through ChIP-seq.

It was possible that Med12 regulated expression by bounding to a distal enhancer, however, in regions with several genes in close proximity of this potential enhancer, assigning the enhancer to the proper target gene would not be trivial. The majority of the identified Med12 binding regions were assigned to coding genes, with almost 500 regions assigned to different non-coding genes (Figure 27b). Despite the majority of ncRNAs with Med12 bound at either their promoter or gene body were not misregulated, around 10% of misregulated ncRNAs contained at least one Med12 binding region (Figure 27d). This suggested that Med12 mediated their expression. Interestingly, while analysing all classes of misregulated genes with Med12 binding at their gene body, these genes were found either up or downregulated with similar frequencies. However, almost 70% of genes with Med12 at promoter where found downregulated, a similar fraction as when analysing only misregulated ncRNAs (Figure 27e). This suggested that when Med12 is found at promoter of genes, it frequently mediates activation of their expression.

Twelve lncRNAs were identified as good candidates for direct targets of Med12. They were significant misregulated in Med12 deficient ESC and ChIP-seq data confirmed Med12 bind at their promoter and/or gene body. Among these, four were Cufflinks predicted transcripts that showed no coding potential as assessed by CPAT (data not shown). In order to evaluate these lncRNAs expression *in vivo*, RNA-seq data generated from different embryonic tissues and across multiple developmental stages was analysed. Knowing the temporal and spatial expression for these genes allows for better design the systems in which to analyse them in future studies. Furthermore, since a number of lncRNAs act by regulating the expression of their closest neighbours (Anderson et al. 2016, Paralkar et al. 2016), genes up to 100kb away from the candidates Med12 targets were identified and their log2FC for both mutants plotted (Figure 28). Misregulated genes within this window represent a starting point for identification of possible targets of the lncRNAS. Interestingly, most of these lncRNAs are enriched in ESCs and for several no expression was detected in analysed tissues. Cufflinks predictions, identified by their name “XLOC...” were found exclusively in ESCs and even is these cells their expression was quite low which might explain why they have not been previously identified. Among the twelve potential Med12 targets, some lncRNAs have been preciously studied. Pltr26 has been associated with pluripotency in mESCs, an association supported by the expression profile of this gene (Bergmann et al. 2015). However, despite its enrichment in ESCs, its expression was also identified, even though lower levels, in embryonic heart. In this organ, Platr26 expression as not been describe before and might represent a different function performed by this lncRNA. Besides this lncRNA and Malat1, none of the identified genes have been previously studied,

representing different predictions by the Fantom Consortium (e.g. 4930461G14Rik) or by NCBI (e.g.Gm265649).

5. Outlook

Various LN-BP18 mutant embryonic stem cells (ESCs) have been generated during this study. Expression analysis of these ESCs might allow the identification of possible functions of this lncRNA. As shown, LN-BP18 was lowly expressed in ESCs but showed a higher expression in embryonic tissues such as pronephros, neural tube and limbs. Due to its higher expression in these tissues, perturbation of its expression could provide important information regarding its expression and function tissues. As such, analysis of LN-BP18 TSS1 mutant embryos generated by tetraploid complementation assay, should result in embryos where no LN-BP18 is expressed (homozygous KO clones) or is downregulated (heterozygous KO clones), allowing to assess its function.

With the same method, embryos were generated from the LN-BP18-β-gal reporter ESCs and the *in vivo* expression of this lncRNA could be determined in a more sensitive way. Sectioning of these embryos would allow generating a more detailed expression pattern for LN-BP18 in the developing embryo. Additionally, by analysing the expression of LN-BP18 in additional mutant embryos, such as the generated Sall1-KO mutants, specifically in the tissues where expression of LN-BP18 was observed using the LN-BP18-β-gal reporter, the functional relation between these two genes could be determined

In mutants where LN-BP18 TSS1 transcripts were downregulated, an upregulation of Nanog and Oct4 was observed. If expression of LN-BP18 TSS1 isoforms is inversely correlated with expression of these important pluripotency regulators, then upon upregulation of these non-coding transcripts, a downregulation of the pluripotency regulators is expected. In order to confirm this hypothesis, transformation of ESC with constructs on which different LN-BP18 isoforms are under the control of a strong promoter, would allow an overexpression of these transcripts and assessment of their effect in the expression of Nanog, Oct4 and additional pluripotency regulators, as well as other potential interactors such as Sall1. Additionally, generation of embryos form cells overexpressing these isoforms would allow detecting possible consequences of abnormally high levels of LN-BP18 in mouse development. Furthermore, since generated data indicated that LN-BP18 TSS2 activity is restricted in ESC and since no effect was observed on the expression of pluripotency markers upon its downregulation, it is possible that upregulation of these transcripts would have quantifiable effects on pluripotency associated genes. As such, embryo generating from cells overexpressing TSS2 transcripts would allow assessing the impact of ectopic expression of LN-BP18 TSS2 transcripts in embryonic development. In order to detect potential target genes of LN-BP18, chromatin isolation by RNA purification (CHIRP-seq) could be used. By generating a biotin tagged version of LN-BP18, RNA-DNA complexes can be crosslinked,

fragmented and purified using the biotin tag on LN-BP18 to identify binding sites for this lncRNA, which would represent putative targets.

In the final section of this study, 12 different annotated and novel lncRNAs were identified as misregulated upon Med12 depletion in ESCs and Med12 binding sites were identified in their promoter/gene body, strongly suggesting that the expression of these genes was mediated by Med12. The majority of these genes are predictions based on automatic gene assembly, and as such have not been studied. Thus, similarly to what was performed to LN-BP18, their expression pattern and gene structure needs to be identified. Then different mutations need to be generated in order to study their function. Their expression pattern has been assessed in different embryonic tissues, representing a starting point for future experiments analysing their expression pattern but also to further confirm any identified expression domains. The neighbours of these candidate genes that were also misregulated in Med12 deficient cells have been identified and assessing their expression upon perturbation of candidate genes expression can confirm if these neighbours are indeed target genes.

These candidate genes represent lncRNAs which normal expression is dependent of Med12. On the other hand, the detection of lncRNAs that interact with Med12 has not been addressed in this study. To identify such non-coding transcripts, a method such as RNA Immunoprecipitation (RIP) can be used. A previous attempt in performing RIP against Med12 was not successful, since the available anti-Med12 antibodies are not specific enough for this kind of method. To overcome this problem, different ESC and mouse lines expressing Med12-tagged versions have been generated by the Schrewe group and represent a useful tool to study Med12 interacting partners, both proteins and ncRNAs.

6. Summary

The function of the Mediator subunit Med12 on gene regulation has been widely studied and its interaction and regulation of protein coding genes broadly documented. However, only recently has its interaction with non-coding genes been verified. Analysis of transcriptome data from Med12 deficient embryonic stem cells (ESCs) revealed hundreds of misregulated protein coding genes, including multiple Wnt targets and genes involved in the developmental processes that were found affected in embryos previously generated with these cells.

In addition to the protein coding genes, multiple misregulated non-coding genes were found during the analysis of transcriptome data generated from Med12 mutant cells, including several putative novel transcripts. Among these, an uncharacterized long non-coding (lnc)RNA was found to be differentially expressed cells, tissues and mouse embryos. This gene, designated as LN-BP18, encodes for antisense transcripts of Sall1, a gene also misregulated in the analysed cells. In humans, mutations in this gene are associated with Townes-Brocks syndrome (TBS), which shows several overlapping characteristics with MED12-associated X-linked intellectual disability syndromes. These features led to the deeper characterization of LN-BP18.

Detailed gene and transcript analyses of this novel lncRNA led to the identification of two distinct transcription start sites (TSSs), termed TSS1 and TSS2. While TSS1 was active in all analysed tissues, TSS2 was found active only in ESCs. *In vitro* differentiation of ESCs confirmed this observation, with expression of transcripts originating from TSS1 increasing throughout the differentiation protocol, while the opposite dynamic occurring for TSS2 transcripts. Characterization of the gene structure revealed a complex splicing pattern, with its 7 exons spliced into 9 different isoforms, including spliced variants for three of the exons. Multiple analyses confirmed the lack of coding potential of all identified isoforms. BLAST searches revealed no homologous transcripts in other species, however, a non-conserved predicted lncRNA was described in human, which was also present in a divergent configuration relative to SALL1, suggesting a potential functional similarity to LN-BP18 despite the low sequence similarity.

Expression analyses of the different mutant ESCs generated revealed a dynamic expression of LN-BP18. TSS2 transcripts, which were only detected in ESCs and not in embryonic tissues, showed a positive correlation with different pluripotency markers. This correlation, together with the ESC-specific activation of this TSS, suggested a potential role in the pluripotency network for isoforms originating from TSS2.

Sall1 and LN-BP18 TSS1 transcripts were downregulated in Med12 depleted ESCs. Additionally, in Sall1-depleted cells, LN-BP18 was downregulated, with a strong effect observed for the TSS1 transcripts compared to TSS2. These observations, together with the co-expression of these two genes in embryonic tissues, suggested LN-BP18, specifically the TSS1, as a target of Sall1 activation. This activation is potentially Med12-dependent, since the effect on LN-BP18 expression was stronger upon Med12 depletion than in Sall1 deplete cells.

A heterozygous LN-BP18-β-galactosidase reporter mutant ESC line was generated to detect expression of LN-BP18 in a more sensitive way. Expression of the reporter gene identified in addition to embryonal limb and caudal end expression seen by whole mount *in situ* hybridization (WISH), a clear expression in the pronephros, somites, neural tube, forebrain/midbrain- and midbrain/hindbrain-boundaries, demonstrating the importance of this reporter line for studying LN-BP18 expression and function during development.

Finally, RNA-seq data from Med12 depleted ESC mutant cells was analysed together with public Med12 chromatin immunoprecipitation sequencing (ChIP-seq) data from ESCs. This analysis allowed identifying 12 lncRNAs, both annotated as well as new predictions, representing candidate lncRNAs whose expression is mediated by Med12. Compiled information for these genes presented here, offer insight into possible systems to analyse these genes in future studies as well as putative targets.

Data from this thesis describe the genetic structure and expression of a previously uncharacterized lncRNA. These data, together with the different mutants generated of this gene establish the ground work for future studies clarifying the functions of LN-BP18 in ESCs, but also during embryonic development.

7. Zusammenfassung

Die Funktion der “Mediator” Untereinheit Med12, bei der Genregulation wurde intensiv erforscht und die Interaktionen von Med12 und seine Rolle bei der Regulation von proteinkodierenden Genen sind gut dokumentiert. Erst kürzlich aber wurde die Interaktion von Med12 mit nicht - proteinkodierenden (non-coding) Genen gefunden. Die Analyse von Transkriptomdaten von Med12 defizienten embryonalen Stammzellen (embryonic stem cells, ESC) zeigten hunderte von deregulierten proteinkodierenden Genen, darunter zahlreiche Wnt Zielgene und Gene, die an bestimmten Entwicklungsprozessen beteiligt sind, die auch in Med12 defizienten Embryonen beeinträchtigt sind.

Neben diesen proteinkodierenden Genen wurden auch zahlreiche non-coding Gene gefunden, darunter auch einige mit bisher unbekannten Transkripten. Eine bisher uncharakterisierte lange nichtcodierende RNA (long noncoding RNA, lncRNA) war in Zellen, Geweben und Mausembryonen differentiell exprimiert. Dieses Gen, LN-BP18, generiert ein antisense Transkript zum Gen Sal1 welches in Med12 Mutanten ebenfalls misreguliert ist. Im Menschen sind Mutationen in Sal1 mit dem Townes-Brocks Syndrom (TBS) assoziiert, das mehrere Merkmale von mit Med12 assoziierten Syndromen X-chromosomaler mentaler Retardierung aufweist. Diese Umstände waren ausschlaggebend für eine nähere Charakterisierung von LN-BP18.

Eine detaillierte Analyse des Gens und seiner Transkripte führte zur Identifizierung von zwei unterschiedlichen Transkriptionsstarts (transcription start sites, TSS), die als TSS1 und TSS2 bezeichnet wurden. Währen TSS1 in allen analysierten Geweben aktiv war, zeigte TSS2 nur in ES-Zellen Aktivität. Die Untersuchung von in vitro differenzierten ES-Zellen bestätigte diesen Befund. Hier zeigten sich, dass die Aktivität der TSS1 während der Differenzierung zunahm, während TSS2 reduzierte Aktivität zeigte. Die Charakterisierung des Gens zeigte ein komplexes Splicingmuster, die 7 Exons werden zu 9 verschiedene Isoformen kombiniert. Ausserdem gibt es Spleißvarianten von drei Exons. Verschiedene Analysen bestätigten die Abwesenheit einer proteinkodierenden Sequenz in allen identifizierten Isoformen. BLAST Analysen zeigten keine homologen Transkripte in anderen Spezies. Allerdings wurde eine nicht konservierte lncRNA im Menschen beschrieben, die ebenfalls in entgegengesetzter Richtung zu Sal1 transkribiert wird, was trotz der geringen Sequenzähnlichkeit auf eine mögliche funktionelle Gemeinsamkeit mit LN-BP18 hinweisen könnte.

Die Expressionsanalyse der verschiedenen mutanten ES-Zellinien zeigte ein dynamisches Expressionsmuster von LN-BP18. TSS2 Transkripte, die nur in ES-Zellen, nicht aber in embryonalen Geweben detektiert wurden zeigten eine positive Korelation mit verschiedenen Markern für Pluripotenz.

Dies, sowie die ES-zellspezische Aktivierung dieser TSS deutete auf eine mögliche Rolle dieser Transkripte im Pluripotenz-Netzwerk hin.

Sal1 und LN-BP18 zeigten verringerte Expression in Med12 defizienten ES-Zellen. Außerdem war LN-BP18 in Sal1 defizienten Zellen herunterreguliert. Dies war für TSS1 Transkripte ausgeprägter als für TSS2 Transkripte. Dies, und die Co-Expression der beiden Gene in embryonalen Geweben deutete darauf hin, dass die TSS1 von LN-BP18 durch Sal1 reguliert wird. Diese Aktivierung ist möglicherweise Med12 abhängig, da der Effekt auf die LN-BP18 expression bei Med12 defizienten Zellen stärker war als bei Sal1 Defizienz.

Um die Expression von LN-BP18 genauer zu untersuchen, wurde eine heterozygote LN-BP18-β-Galactosidase Reporterlinie generiert. In Mäusen dieser Linie zeigte sich die Expression des Reportergens außer in den embryonalen Gliedmaßen und dem kaudalen Ende, die auch bei in situ Hybridisierung sichtbar war, im Pronephros, den Somiten, dem Neuralrohr, der Vorderhirn-Mittelhirn- sowie der Mittelhirn-Hinterhirngrenze. Dies zeigt die Nützlichkeit dieser Reporterlinie bei der Untersuchung der LN-BP18 Expression und seiner Funktion während der Entwicklung.

Schließlich wurden RNA-seq Daten von Med12 defizienten ES-Zellen zusammen mit öffentlich zugänglichen Med12 „chromatin immunoprecipitation sequencing“ (ChIP-seq) Daten analysiert. Hier wurden 12 LncRNAs gefunden, die teilweise noch nicht annotiert waren. Diese sind möglicherweise durch Med12 reguliert. Hier durchgeführte erste Untersuchungen dieser Gene ermöglichen eine Eingrenzung Ihres Funktionszusammenhanges für eine zukünftige funktionelle Analyse.

Diese Arbeit beschreibt die Genstruktur und Expression einer bisher uncharakterisierten lncRNA. Diese Analysen und die verschiedenen Mutanten, die in dieser Arbeit erzeugt wurden legen den Grundstein für eine zukünftige funktionelle Analyse von LN-BP18 in ES-Zellen und während der Embryonalentwicklung.

8. Acknowledgments

My first and biggest acknowledgment goes to my supervisor Dr. Heinrich Schrewe. I must thank him for his continuous support, encouragement in improving my skills in different areas and to become a better scientist, guidance in all the moments that I felt lost in my project and above all for sharing his incredible enthusiasm for science.

I thank Prof. Dr. Bernhard Herrmann for supporting me as a PhD student in his department. I also thank Dr. Andreas Mayer for being part of my Thesis Advisory Committee. I thank as well Prof. Dr. Sigmar Stricker for the useful input and for agreeing to be my supervisor at the Frei Universität.

I would also like to thank Andrea König, Gaby Bläß and Manuela Scholze-Wittler for all the amazing technical help and for making possible all the work performed in the Department of Developmental Genetics. I acknowledge Susanne Braun and Davitasvili Shirly for their assistance during the characterization of LN-Bp18 gene structure. I thank Frederic Koch, Pavel Tsaytler and Jesse Veenvliet for the all the invaluable help, interesting discussions, and useful ideas. I also thank Dennis Schifferl and Michael Gerloff for their help in multiple experiments. I acknowledge Nina Bailly, Aleksandra Arczewska and Michael Robson for help with the thesis. I would like to all the member of the Developmental Genetics Department for contributing to an amazing working atmosphere and for all the help in the everyday problems in the laboratory.

Todo este percurso só foi possível graças a todo o apoio dos meus pais, ao seu incentivo para fazer o que gosto e à compreensão da minha escolha de emigrar. Obrigado a todos os amigos e família por sempre me fazerem sentir bem vindo em Portugal e aos amigos em Berlim por tornarem a vida na Alemanha muito mais divertida.

Mas principalmente estou grato à Patricia, por todo o amor e apoio, por sempre me incentivar a melhorar e a dar o melhor de mim e por ser o meu porto seguro nos maus momentos.

E ao Francisco, que este trabalho te incentive a seguir os teus sonhos e que sirva de exemplo que se trabalhares para cumprires os teus objetivos, consegues fazer tudo.

9. Bibliography

1. Adelman, K. and J. T. Lis (2012). "**Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans**." Nat Rev Genet 13(10): 720-731.
2. AkoulitchevRobinson , S., S. Chuikov and D. Reinberg (2000). "**TFIIH is negatively regulated by cdk8-containing mediator complexes**." Nature 407(6800): 102-106.
3. Alamancos, G. P., E. Agirre and E. Eyras (2014). "**Methods to study splicing from high-throughput RNA sequencing data**." Methods Mol Biol 1126: 357-397.
4. Alarcon, C., A. I. Zaromytidou, Q. Xi, S. Gao, J. Yu, S. Fujisawa, A. Barlas*, et al.* (2009). "**Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways**." Cell 139(4): 757-769.
5. Anders, S., P. T. Pyl and W. Huber (2015). "**HTSeq--a Python framework to work with high-throughput sequencing data**." Bioinformatics 31(2): 166-169.
6. Anderson, K. M., D. M. Anderson, J. R. McAnally, J. M. Shelton, R. Bassel-Duby and E. N. Olson (2016). "**Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development**." Nature 539(7629): 433-436.
7. Andrau, J. C., L. van de Pasch, P. Lijnzaad, T. Bijma, M. G. Koerkamp, J. van de Peppel, M. Werner*, et al.* (2006). "**Genome-wide location of the coactivator mediator: Binding without activation and transient Cdk8 interaction on DNA**." Mol Cell 22(2): 179-192.
8. Aprea, J., S. Prenninger, M. Dori, T. Ghosh, L. S. Monasor, E. Wessendorf, S. Zocher*, et al.* (2013). "**Transcriptome sequencing during mouse brain development identifies long non-coding RNAs functionally involved in neurogenic commitment**." EMBO J 32(24): 3145-3160.
9. Aranda-Orgilles, B., R. Saldana-Meyer, E. Wang, E. Trompouki, A. Fassl, S. Lau, J. Mullenders*, et al.* (2016). "**MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis**." Cell Stem Cell.
10. Assie, G., E. Letouze, M. Fassnacht, A. Jouinot, W. Luscap, O. Barreau, H. Omeiri*, et al.* (2014). "**Integrated genomic characterization of adrenocortical carcinoma**." Nat Genet 46(6): 607-612.
11. Avery, O. T., C. M. Macleod and M. McCarty (1944). "**Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii**." J Exp Med 79(2): 137-158.
12. Baillie-Johnson, P., S. C. van den Brink, T. Balayo, D. A. Turner and A. M. Arias (2015). "**Generation of Aggregates of Mouse Embryonic Stem Cells that Show Symmetry Breaking, Polarization and Emergent Collective Behaviour In Vitro**." Jove-Journal of Visualized Experiments(105).
13. Barbieri, C. E., S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J. P. Theurillat, T. A. White*, et al.* (2012). "**Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer**." Nat Genet 44(6): 685-689.
14. Baruzzo, G., K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald and G. R. Grant (2017). "**Simulation-based comprehensive benchmarking of RNA-seq aligners**." Nature Methods 14(2): 135-139.
15. Becker, J. S., R. L. McCarthy, S. Sidoli, G. Donahue, K. E. Kaeding, Z. Y. He, S. Lin*, et al.* (2017). "**Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes**." Molecular Cell 68(6): 1023-+.
16. Bergmann, J. H., J. J. Li, M. A. Eckersley-Maslin, F. Rigo, S. M. Freier and D. L. Spector (2015). "**Regulation of the ESC transcriptome by nuclear long noncoding RNAs**." Genome Research 25(9): 1336-1346.

17. Birmingham, A., E. M. Anderson, A. Reynolds, D. Ilsley-Tyree, D. Leake, Y. Fedorov, S. Baskerville*, et al.* (2006). "**3' UTR seed matches, but not overall identity, are associated with RNAi off-targets**." Nat Methods 3(3): 199-204.
18. Blockus, H. and A. Chedotal (2016). "**Slit-Robo signaling**." Development 143(17): 3037-3044.
19. Bohm, J., A. Buck, W. Borozdin, A. U. Mannan, U. Matysiak-Scholze, I. Adham, W. Schulz-Schaeffer*, et al.* (2008). "**Sall1, Sall2, and Sall4 Are Required for Neural Tube Closure in Mice**." American Journal of Pathology 173(5): 1455-1463.
20. Boija, A., I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li*, et al.* (2018). "**Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains**." Cell 175(7): 1842-+.
21. Boube, M., C. Faucher, L. Joulia, D. L. Cribbs and H. M. Bourbon (2000). "**Drosophila homologs of transcriptional mediator complex subunits are required for adult cell and segment identity specification**." Genes Dev 14(22): 2906-2917.
22. Bourbon, H. M., A. Aguilera, A. Z. Ansari, F. J. Asturias, A. J. Berk, S. Bjorklund, T. K. Blackwell*, et al.* (2004). "**A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II**." Mol Cell 14(5): 553-557.
23. Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine*, et al.* (2006). "**Polycomb complexes repress developmental regulators in murine embryonic stem cells**." Nature 441(7091): 349-353.
24. Buck, A., A. Kispert and J. Kohlhase (2001). "**Embryonic expression of the murine homologue of SALL1, the gene mutated in Townes-Brocks syndrome**." Mechanisms of Development 104(1-2): 143-146.
25. Bulger, M. and M. Groudine (2011). "**Functional and Mechanistic Diversity of Distal Transcription Enhancers**." Cell 144(3): 327-339.
26. Bunch, H. (2018). "**Gene regulation of mammalian long non-coding RNA**." Molecular Genetics and Genomics 293(1): 1-15.
27. Burke, T. W. and J. T. Kadonaga (1997). "**The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila**." Genes Dev 11(22): 3020-3031.
28. Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev and J. L. Rinn (2011). "**Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**." Genes Dev 25(18): 1915-1927.
29. Calo, E. and J. Wysocka (2013). "**Modification of Enhancer Chromatin: What, How, and Why?**" Molecular Cell 49(5): 825-837.
30. Carrera, I., F. Janody, N. Leeds, F. Duveau and J. E. Treisman (2008). "**Pygopus activates Wingless target gene transcription through the mediator complex subunits Med12 and Med13**." Proc Natl Acad Sci U S A 105(18): 6644-6649.
31. Cartolano, M., B. Huettel, B. Hartwig, R. Reinhardt and K. Schneeberger (2016). "**cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing**." PLoS One 11(6): e0157779.
32. Cassidy, S. B., S. Schwartz, J. L. Miller and D. J. Driscoll (2012). "**Prader-Willi syndrome**." Genetics in Medicine 14(1): 10-26.
33. Cesana, M., D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano*, et al.* (2011). "**A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA**." Cell 147(2): 358-369.
34. Cevher, M. A., Y. Shi, D. Li, B. T. Chait, S. Malik and R. G. Roeder (2014). "**Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit**." Nat Struct Mol Biol 21(12): 1028-1034.

35. Chalkley, G. E. and C. P. Verrijzer (1999). "**DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator**." EMBO J 18(17): 4835-4845.
36. Chavez, A., M. Tuttle, B. W. Pruitt, B. Ewen-Campen, R. Chari, D. Ter-Ovanesyan, S. J. Haque*, et al.* (2016). "**Comparison of Cas9 activators in multiple species**." Nat Methods 13(7): 563-567.
37. Chen, W. and R. G. Roeder (2007). "**The mediator subunit MED1/TRAP220 is required for optimal glucocorticoid receptor-mediated transcription activation**." Nucleic Acids Research 35(18): 6161-6169.
38. Chen, X. F., L. Lehmann, J. J. Lin, A. Vashisht, R. Schmidt, R. Ferrari, C. Huang*, et al.* (2012). "**Mediator and SAGA have distinct roles in Pol II preinitiation complex assembly and function**." Cell Rep 2(5): 1061-1067.
39. Core, L. J., J. J. Waterfall and J. T. Lis (2008). "**Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters**." Science 322(5909): 1845-1848.
40. Cramer, P., D. A. Bushnell, J. Fu, A. L. Gnatt, B. Maier-Davis, N. E. Thompson, R. R. Burgess*, et al.* (2000). "**Architecture of RNA polymerase II and implications for the transcription mechanism**." Science 288(5466): 640-649.
41. Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna*, et al.* (2010). "**Histone H3K27ac separates active from poised enhancers and predicts developmental state**." Proceedings of the National Academy of Sciences of the United States of America 107(50): 21931-21936.
42. Daniels, D. L., M. Ford, M. K. Schwinn, H. e. l. e. n. Benink, M. D.Galbraith, R. Amunugama, R. Jones*, et al.* (2013). "**Mutual Exclusivity of MED12/MED12L, MED13/13L, and CDK8/19 Paralogs Revealed within the CDK-Mediator Kinase Module**." Journal of Proteomics & Bioinformatics 6(2): 1-6.
43. Dash, R., T. B. Emran, M. M. Uddin, A. Islam and M. Junaid (2014). "**Molecular docking of fisetin with AD associated AChE, ABAD and BACE1 proteins**." Bioinformation 10(9): 562-568.
44. Di Ruscio, A., A. K. Ebralidze, T. Benoukraf, G. Amabile, L. A. Goff, J. Terragni, M. E. Figueroa*, et al.* (2013). "**DNMT1-interacting RNAs block gene-specific DNA methylation**." Nature 503(7476): 371-376.
45. Dimitrova, N., J. R. Zamudio, R. M. Jong, D. Soukup, R. Resnick, K. Sarma, A. J. Ward*, et al.* (2014). "**LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint**." Mol Cell 54(5): 777-790.
46. Ding, N., C. Tomomori-Sato, S. Sato, R. C. Conaway, J. W. Conaway and T. G. Boyer (2009). "**MED19 and MED26 are synergistic functional targets of the RE1 silencing transcription factor in epigenetic silencing of neuronal gene expression**." J Biol Chem 284(5): 2648-2656.
47. Ding, N., H. Zhou, P. O. Esteve, H. G. Chin, S. Kim, X. Xu, S. M. Joseph*, et al.* (2008). "**Mediator links epigenetic silencing of neuronal gene expression with x-linked mental retardation**." Mol Cell 31(3): 347-359.
48. Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer*, et al.* (2012). "**Landscape of transcription in human cells**." Nature 489(7414): 101-108.
49. Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer*, et al.* (2012). "**Landscape of transcription in human cells**." Nature 489(7414): 101-108.
50. Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut*, et al.* (2013). "**STAR: ultrafast universal RNA-seq aligner**." Bioinformatics 29(1): 15-21.
51. Doi, H., T. Iso, M. Yamazaki, H. Akiyama, H. Kanai, H. Sato, K. Kawai-Kowase*, et al.* (2005). "**HERP1 inhibits myocardin-induced vascular smooth muscle cell differentiation by interfering with SRF binding to CArG box**." Arterioscler Thromb Vasc Biol 25(11): 2328-2334.
52. Donner, A. J., C. C. Ebmeier, D. J. Taatjes and J. M. Espinosa (2010). "**CDK8 is a positive regulator of transcriptional elongation within the serum response network**." Nat Struct Mol Biol 17(2): 194-201.

53. Douziech, M., F. Coin, J. M. Chipoulet, Y. Arai, Y. Ohkuma, J. M. Egly and B. Coulombe (2000). "**Mechanism of promoter melting by the xeroderma pigmentosum complementation group B helicase of transcription factor IIH revealed by protein-DNA photo-cross-linking**." Mol Cell Biol 20(21): 8168-8177.
54. Ebmeier, C. C. and D. J. Taatjes (2010). "**Activator-Mediator binding regulates Mediator-cofactor interactions**." Proc Natl Acad Sci U S A 107(25): 11283-11288.
55. Elmlund, H., V. Baraznenok, M. Lindahl, C. O. Samuelsen, P. J. Koeck, S. Holmberg, H. Hebert*, et al.* (2006). "**The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II**." Proc Natl Acad Sci U S A 103(43): 15788-15793.
56. Emrich, S. J., W. B. Barbazuk, L. Li and P. S. Schnable (2007). "**Gene discovery and annotation using LCM-454 transcriptome sequencing**." Genome Research 17(1): 69-73.
57. Faghihi, M. A., F. Modarresi, A. M. Khalil, D. E. Wood, B. G. Sahagan, T. E. Morgan, C. E. Finch*, et al.* (2008). "**Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase**." Nature Medicine 14(7): 723-730.
58. Femino, A., F. S. Fay, K. Fogarty and R. H. Singer (1998). "**Visualization of single RNA transcripts in situ**." Science 280(5363): 585-590.
59. Flanagan, P. M., R. J. Kelleher, 3rd, M. H. Sayre, H. Tschochner and R. D. Kornberg (1991). "**A mediator required for activation of RNA polymerase II transcription in vitro**." Nature 350(6317): 436-438.
60. Fondell, J. D., H. Ge and R. G. Roeder (1996). "**Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex**." Proc Natl Acad Sci U S A 93(16): 8329-8333.
61. Frazee, A. C., G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg and J. T. Leek (2015). "**Ballgown bridges the gap between transcriptome assembly and expression analysis**." Nat Biotechnol 33(3): 243-246.
62. Freese, N. H., D. C. Norris and A. E. Loraine (2016). "**Integrated genome browser: visual analytics platform for genomics**." Bioinformatics 32(14): 2089-2095.
63. Fryer, C. J., J. B. White and K. A. Jones (2004). "**Mastermind recruits CycC:CDK8 to phosphorylate the Notch ICD and coordinate activation with turnover**." Mol Cell 16(4): 509-520.
64. Furlan, G., N. Gutierrez Hernandez, C. Huret, R. Galupa, J. G. van Bemmel, A. Romito, E. Heard*, et al.* (2018). "**The Ftx Noncoding Locus Controls X Chromosome Inactivation Independently of Its RNA Products**." Mol Cell 70(3): 462-472 e468.
65. Garalde, D. R., E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic*, et al.* (2018). "**Highly parallel direct RNA sequencing on an array of nanopores**." Nat Methods 15(3): 201-206.
66. George, S. H., M. Gertsenstein, K. Vintersten, E. Korets-Smith, J. Murphy, M. E. Stevens, J. J. Haigh*, et al.* (2007). "**Developmental and adult phenotyping directly from mutant embryonic stem cells**." Proc Natl Acad Sci U S A 104(11): 4455-4460.
67. Gobert, V., D. Osman, S. Bras, B. Auge, M. Boube, H. M. Bourbon, T. Horn*, et al.* (2010). "**A genome-wide RNA interference screen identifies a differential role of the mediator CDK8 module subunits for GATA/ RUNX-activated transcription in Drosophila**." Mol Cell Biol 30(11): 2837-2848.
68. Gonzalez Bosc, L. V., J. J. Layne, M. T. Nelson and D. C. Hill-Eubanks (2005). "**Nuclear factor of activated T cells and serum response factor cooperatively regulate the activity of an alpha-actin intronic enhancer**." J Biol Chem 280(28): 26113-26120.
69. Gorodkin, J. and I. L. Hofacker (2011). "**From structure prediction to genomic screens for novel non-coding RNAs**." PLoS Comput Biol 7(8): e1002100.
70. Gouti, M., A. Tsakiridis, F. J. Wymeersch, Y. Huang, J. Kleinjung, V. Wilson and J. Briscoe (2014). "**In vitro generation of neuromesodermal progenitors reveals distinct roles for wnt signalling in the specification of spinal cord and paraxial mesoderm identity**." PLoS Biol 12(8): e1001937.

71. Graham, J. M. and C. E. Schwartz (2013). "**MED12 Related Disorders**." American Journal of Medical Genetics Part A 161(11): 2734-2740.
72. Grote, P., L. Wittler, D. Hendrix, F. Koch, S. Wahrisch, A. Beisaw, K. Macura*, et al.* (2013). "**The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse**." Dev Cell 24(2): 206-214.
73. Gudenas, B. L., A. K. Srivastava and L. Wang (2017). "**Integrative genomic analyses for identification and prioritization of long non-coding RNAs associated with autism**." PLoS One 12(5): e0178532.
74. Gutschner, T., M. Hammerle and S. Diederichs (2013). "**MALAT1 -- a paradigm for long noncoding RNA function in cancer**." J Mol Med (Berl) 91(7): 791-801.
75. Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte*, et al.* (2009). "**Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**." Nature 458(7235): 223-227.
76. Ha, M. and V. N. Kim (2014). "**Regulation of microRNA biogenesis**." Nat Rev Mol Cell Biol 15(8): 509-524.
77. Hantsche, M. and P. Cramer (2017). "**Conserved RNA polymerase II initiation complex structure**." Curr Opin Struct Biol 47: 17-22.
78. Harno, E., E. C. Cottrell and A. White (2013). "**Metabolic pitfalls of CNS Cre-based technology**." Cell Metab 18(1): 21-28.
79. Heffner, C. S., C. H. Pratt, R. P. Babiuk, Y. Sharma, S. F. Rockwood, L. R. Donahue, J. T. Eppig*, et al.* (2012). "**Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource**." Nature Communications 3.
80. Hein, P. P., K. E. Kolb, T. Windgassen, M. J. Bellecourt, S. A. Darst, R. A. Mooney and R. Landick (2014). "**RNA polymerase pausing and nascent-RNA structure formation are linked through clamp-domain movement**." Nat Struct Mol Biol 21(9): 794-802.
81. Herbert, K. M., A. La Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick and S. M. Block (2006). "**Sequence-resolved detection of pausing by single RNA polymerase molecules**." Cell 125(6): 1083-1094.
82. Hnisz, D., K. Shrinivas, R. A. Young, A. K. Chakraborty and P. A. Sharp (2017). "**A Phase Separation Model for Transcriptional Control**." Cell 169(1): 13-23.
83. Hong, S. K., C. E. Haldin, N. D. Lawson, B. M. Weinstein, I. B. Dawid and N. A. Hukriede (2005). "**The zebrafish kohtalo/trap230 gene is required for the development of the brain, neural crest, and pronephric kidney**." Proc Natl Acad Sci U S A 102(51): 18473-18478.
84. Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "**Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**." Nucleic Acids Res 37(1): 1-13.
85. Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "**Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**." Nat Protoc 4(1): 44-57.
86. Huang, S., M. Holzel, T. Knijnenburg, A. Schlicker, P. Roepman, U. McDermott, M. Garnett*, et al.* (2012). "**MED12 controls the response to multiple cancer drugs through regulation of TGF-beta receptor signaling**." Cell 151(5): 937-950.
87. Iglesias-Platas, I. and D. Monk (2016). "**Nongenomic regulation of gene expression**." Curr Opin Pediatr 28(4): 521-528.
88. Ito, M., C. X. Yuan, S. Malik, W. Gu, J. D. Fondell, S. Yamamura, Z. Y. Fu*, et al.* (1999). "**Identity between TRAP and SMCC complexes indicates novel pathways for the function of nuclear receptors and diverse mammalian activators**." Mol Cell 3(3): 361-370.
89. Ito, M., C. X. Yuan, H. J. Okano, R. B. Darnell and R. G. Roeder (2000). "**Involvement of the TRAP220 component of the TRAP/SMCC coactivator complex in embryonic development and thyroid hormone action**." Mol Cell 5(4): 683-693.

90. Iyer, M. K., Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette*, et al.* (2015). "**The landscape of long noncoding RNAs in the human transcriptome**." Nat Genet 47(3): 199-208.
91. Jadaliha, M., X. Zong, P. Malakar, T. Ray, D. K. Singh, S. M. Freier, T. Jensen*, et al.* (2016). "**Functional and prognostic significance of long non-coding RNA MALAT1 as a metastasis driver in ER negative lymph node negative breast cancer**." Oncotarget 7(26): 40418-40436.
92. Janody, F., Z. Martirosyan, A. Benlali and J. E. Treisman (2003). "**Two subunits of the Drosophila mediator complex act together to control cell affinity**." Development 130(16): 3691-3701.
93. Jeronimo, C., M. F. Langelier, A. R. Bataille, J. M. Pascal, B. F. Pugh and F. Robert (2016). "**Tail and Kinase Modules Differently Regulate Core Mediator Recruitment and Function In Vivo**." Mol Cell 64(3): 455-466.
94. Jeronimo, C. and F. Robert (2014). "**Kin28 regulates the transient association of Mediator with core promoters**." Nat Struct Mol Biol 21(5): 449-455.
95. Jiang, Y. and D. H. Price (2004). "**Rescue of the TTF2 knockdown phenotype with an siRNA-resistant replacement vector**." Cell Cycle 3(9): 1151-1153.
96. Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier*, et al.* (2010). "**Mediator and cohesin connect gene expression and chromatin architecture**." Nature 467(7314): 430-435.
97. Kampjarvi, K., N. Makinen, O. Kilpivaara, J. Arola, H. R. Heinonen, J. Bohm, O. Abdel-Wahab*, et al.* (2012). "**Somatic MED12 mutations in uterine leiomyosarcoma and colorectal cancer**." Br J Cancer 107(10): 1761-1765.
98. Kaplan, C. D., H. Jin, I. L. Zhang and A. Belyanin (2012). "**Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo**." PLoS Genet 8(4): e1002627.
99. Karantzali, E., V. Lekakis, M. Ioannou, C. Hadjimichael, J. Papamatheakis and A. Kretsovali (2011). "**Sall1 regulates embryonic stem cell differentiation in association with nanog**." J Biol Chem 286(2): 1037-1045.
100. Kawakami, Y., Y. Uchiyama, C. R. Esteban, T. Inenaga, N. Koyano-Nakagawa, M. Marti, M. Kmita*, et al.* (2009). "**Sall genes regulate region-specific morphogenesis in the mouse limb by modulating Hox activities**." Developmental Biology 331(2): 498-498.
101. Keightley, M. C., J. E. Layton, J. W. Hayman, J. K. Heath and G. J. Lieschke (2011). "**Mediator subunit 12 is required for neutrophil development in zebrafish**." PLoS One 6(8): e23845.
102. Khatter, H., M. K. Vorlander and C. W. Muller (2017). "**RNA polymerase I and III: similar yet unique**." Curr Opin Struct Biol 47: 88-94.
103. Kiefer, S. M., L. Robbins, K. M. Stumpff, C. Lin, L. Ma and M. Rauchman (2010). "**Sall1-dependent signals affect Wnt signaling and ureter tip fate to initiate kidney development**." Development 137(18): 3099-3106.
104. Kim, D., B. Langmead and S. L. Salzberg (2015). "**HISAT: a fast spliced aligner with low memory requirements**." Nat Methods 12(4): 357-360.
105. Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "**TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**." Genome Biol 14(4): R36.
106. Kim, S. and D. S. Gross (2013). "**Mediator recruitment to heat shock genes requires dual Hsf1 activation domains and mediator tail subunits Med15 and Med16**." J Biol Chem 288(17): 12197-12213.
107. Kim, S., X. Xu, A. Hecht and T. G. Boyer (2006). "**Mediator is a transducer of Wnt/beta-catenin signaling**." J Biol Chem 281(20): 14066-14075.
108. Kim, T. K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin*, et al.* (2010). "**Widespread transcription at neuronal activity-regulated enhancers**." Nature 465(7295): 182-U165.

109. Kiss, T. (2001). "**Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs**." EMBO J 20(14): 3617-3622.
110. Knuesel, M. T., K. D. Meyer, A. J. Donner, J. M. Espinosa and D. J. Taatjes (2009). "**The human CDK8 subcomplex is a histone kinase that requires Med12 for activity and can function independently of mediator**." Mol Cell Biol 29(3): 650-661.
111. Knuesel, M. T., K. D. Meyer, A. J. Donner, J. M. Espinosa and D. J. Taatjes (2009). "**The Human CDK8 Subcomplex Is a Histone Kinase That Requires Med12 for Activity and Can Function Independently of Mediator**." Molecular and Cellular Biology 29(3): 650-661.
112. Koch, F., M. Scholze, L. Wittler, D. Schifferl, S. Sudheer, P. Grote, B. Timmermann*, et al.* (2017). "**Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors**." Dev Cell 42(5): 514-526 e517.
113. Kohlhase, J. (1993). **Townes-Brocks Syndrome**. GeneReviews((R)). M. P. Adam, H. H. Ardinger, R. A. Pagon et al. Seattle (WA).
114. Korablev, A. N., I. A. Serova and O. L. Serov (2017). "**Generation of megabase-scale deletions, inversions and duplications involving the Contactin-6 gene in mice by CRISPR/Cas9 technology**." Bmc Genetics 18.
115. Kornberg, R. D. (2001). "**The eukaryotic gene transcription machinery**." Biol Chem 382(8): 1103-1107.
116. Kraus, P., V. Sivakamasundari, S. L. Lim, X. Xing, L. Lipovich and T. Lufkin (2013). "**Making sense of Dlx1 antisense RNA**." Dev Biol 376(2): 224-235.
117. Kretz, M., Z. Siprashvili, C. Chu, D. E. Webster, A. Zehnder, K. Qu, C. S. Lee*, et al.* (2013). "**Control of somatic tissue differentiation by the long non-coding RNA TINCR**." Nature 493(7431): 231-235.
118. Kretzschmar, M., G. Stelzer, R. G. Roeder and M. Meisterernst (1994). "**RNA polymerase II cofactor PC2 facilitates activation of transcription by GAL4-AH in vitro**." Mol Cell Biol 14(6): 3927-3937.
119. Krishnamurthy, S. and M. Hampsey (2009). "**Eukaryotic transcription initiation**." Curr Biol 19(4): R153-156.
120. Lacombe, T., S. L. Poh, R. Barbey and L. Kuras (2013). "**Mediator is an intrinsic component of the basal RNA polymerase II machinery in vivo**." Nucleic Acids Research 41(21): 9651-9662.
121. Lai, F., U. A. Orom, M. Cesaroni, M. Beringer, D. J. Taatjes, G. A. Blobel and R. Shiekhattar (2013). "**Activating RNAs associate with Mediator to enhance chromatin architecture and transcription**." Nature 494(7438): 497-501.
122. Lai, F. and R. Shiekhattar (2014). "**Where long noncoding RNAs meet DNA methylation**." Cell Research 24(3): 263-264.
123. Latorre, E., S. Carelli, I. Raimondi, V. D'Agostino, I. Castiglioni, C. Zucal, G. Moro*, et al.* (2016). "**The Ribonucleic Complex HuR-MALAT1 Represses CD133 Expression and Suppresses Epithelial-Mesenchymal Transition in Breast Cancer**." Cancer Res 76(9): 2626-2636.
124. Le Thomas, A., A. K. Rogers, A. Webster, G. K. Marinov, S. E. Liao, E. M. Perkins, J. K. Hur*, et al.* (2013). "**Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state**." Genes Dev 27(4): 390-399.
125. Lee, J. T., L. S. Davidow and D. Warshawsky (1999). "**Tsix, a gene antisense to Xist at the X-inactivation centre**." Nat Genet 21(4): 400-404.
126. Lee, M. S., K. Lim, M. K. Lee and S. W. Chi (2018). "**Structural Basis for the Interaction between p53 Transactivation Domain and the Mediator Subunit MED25**." Molecules 23(10).
127. Lehner, B., C. Crombie, J. Tischler, A. Fortunato and A. G. Fraser (2006). "**Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways**." Nat Genet 38(8): 896-903.

128. Leuther, K. K., D. A. Bushnell and R. D. Kornberg (1996). "**Two-dimensional crystallography of TFIIB- and IIE-RNA polymerase II complexes: implications for start site selection and initiation complex formation**." Cell 85(5): 773-779.
129. Leveille, N., C. A. Melo, K. Rooijers, A. Diaz-Lagares, S. A. Melo, G. Korkmaz, R. Lopes*, et al.* (2015). "**Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA**." Nat Commun 6: 6520.
130. Lewis, J. D. and E. Izaurralde (1997). "**The role of the cap structure in RNA processing and nuclear export**." European Journal of Biochemistry 247(2): 461-469.
131. Li, L., B. Liu, O. L. Wapinski, M. C. Tsai, K. Qu, J. Zhang, J. C. Carlson*, et al.* (2013). "**Targeted disruption of Hotair leads to homeotic transformation and gene derepression**." Cell Rep 5(1): 3-12.
132. Li, S., S. W. Tighe, C. M. Nicolet, D. Grove, S. Levy, W. Farmerie, A. Viale*, et al.* (2014). "**Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study**." Nature Biotechnology 32(9): 915-925.
133. Li, Y., S. Bjorklund, Y. W. Jiang, Y. J. Kim, W. S. Lane, D. J. Stillman and R. D. Kornberg (1995). "**Yeast global transcriptional regulators Sin4 and Rgr1 are components of mediator complex/RNA polymerase II holoenzyme**." Proc Natl Acad Sci U S A 92(24): 10864-10868.
134. Lim, W. K., C. K. Ong, J. Tan, A. A. Thike, C. C. Ng, V. Rajasegaran, S. S. Myint*, et al.* (2014). "**Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma**." Nat Genet 46(8): 877-880.
135. Lin, J. J., L. W. Lehmann, G. Bonora, R. Sridharan, A. A. Vashisht, N. Tran, K. Plath*, et al.* (2011). "**Mediator coordinates PIC assembly with recruitment of CHD1**." Genes Dev 25(20): 2198-2209.
136. Lin, N., K. Y. Chang, Z. Li, K. Gates, Z. A. Rana, J. Dang, D. Zhang*, et al.* (2014). "**An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment**." Mol Cell 53(6): 1005-1019.
137. Linder, T., X. Zhu, V. Baraznenok and C. M. Gustafsson (2006). "**The classical srb4-138 mutant allele causes dissociation of yeast Mediator**." Biochem Biophys Res Commun 349(3): 948-953.
138. Liu, T., Y. Y. Huang, J. L. Chen, H. Y. Chi, Z. H. Yu, J. Wang and C. Chen (2014). "**Attenuated ability of BACE1 to cleave the amyloid precursor protein via silencing long noncoding RNA BACE1-AS expression**." Molecular Medicine Reports 10(3): 1275-1281.
139. Liu, Y., J. A. Ranish, R. Aebersold and S. Hahn (2001). "**Yeast nuclear extract contains two major forms of RNA polymerase II mediator complexes**." J Biol Chem 276(10): 7169-7175.
140. Love, M. I., W. Huber and S. Anders (2014). "**Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**." Genome Biol 15(12): 550.
141. Lv, L., T. Li, X. Li, C. Xu, Q. Liu, H. Jiang, Y. Li*, et al.* (2018). "**The lncRNA Plscr4 Controls Cardiac Hypertrophy by Regulating miR-214**." Mol Ther Nucleic Acids 10: 387-397.
142. Lyko, F. (2018). "**The DNA methyltransferase family: a versatile toolkit for epigenetic regulation**." Nature Reviews Genetics 19(2): 81-92.
143. Makinen, N., M. Mehine, J. Tolvanen, E. Kaasinen, Y. Li, H. J. Lehtonen, M. Gentile*, et al.* (2011). "**MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas**." Science 334(6053): 252-255.
144. Malik, S., W. Gu, W. Wu, J. Qin and R. G. Roeder (2000). "**The USA-derived transcriptional coactivator PC2 is a submodule of TRAP/SMCC and acts synergistically with other PCs**." Mol Cell 5(4): 753-760.
145. Marson, A., S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther*, et al.* (2008). "**Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells**." Cell 134(3): 521-533.

146. Massone, S., E. Ciarlo, S. Vella, M. Nizzari, T. Florio, C. Russo, R. Cancedda*, et al.* (2012). "**NDM29, a RNA polymerase III-dependent non coding RNA, promotes amyloidogenic processing of APP and amyloid beta secretion**." Biochimica Et Biophysica Acta-Molecular Cell Research 1823(7): 1170-1177.
147. McDermaid, A., X. Chen, Y. R. Zhang, C. K. Wang, S. P. Gu, J. Xie and Q. Ma (2018). "**A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation**." Frontiers in Genetics 9.
148. Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier*, et al.* (2013). "**Circular RNAs are a large class of animal RNAs with regulatory potency**." Nature 495(7441): 333-338.
149. Meng, L., A. J. Ward, S. Chun, C. F. Bennett, A. L. Beaudet and F. Rigo (2015). "**Towards a therapy for Angelman syndrome by targeting a long non-coding RNA**." Nature 518(7539): 409-412.
150. Mikkelsen, T. S., M. C. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez*, et al.* (2007). "**Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**." Nature 448(7153): 553-U552.
151. Moghal, N. and P. W. Sternberg (2003). "**A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in C. elegans**." Development 130(1): 57-69.
152. Moore, L. D., T. Le and G. Fan (2013). "**DNA methylation and its basic function**." Neuropsychopharmacology 38(1): 23-38.
153. Morita, Y., P. Andersen, A. Hotta, Y. Tsukahara, N. Sasagawa, N. Hayashida, C. Koga*, et al.* (2016). "**Sall1 transiently marks undifferentiated heart precursors and regulates their fate**." Journal of Molecular and Cellular Cardiology 92: 158-162.
154. Morlan, J. D., K. Qu and D. V. Sinicropi (2012). "**Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue**." PLoS One 7(8): e42882.
155. Mulder, S. D., W. M. van der Flier, J. H. Verheijen, C. Mulder, P. Scheltens, M. A. Blankenstein, C. E. Hack*, et al.* (2010). "**BACE1 Activity in Cerebrospinal Fluid and Its Relation to Markers of AD Pathology**." Journal of Alzheimers Disease 20(1): 253-260.
156. Myer, V. E. and R. A. Young (1998). "**RNA polymerase II holoenzymes and subcomplexes**." Journal of Biological Chemistry 273(43): 27757-27760.
157. Necsulea, A., M. Soumillon, M. Warnefors, A. Liechti, T. Daish, U. Zeller, J. C. Baker*, et al.* (2014). "**The evolution of lncRNA repertoires and expression patterns in tetrapods**." Nature 505(7485): 635-640.
158. Nelson, B. R., C. A. Makarewich, D. M. Anderson, B. R. Winders, C. D. Troupes, F. Wu, A. L. Reese*, et al.* (2016). "**A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle**." Science 351(6270): 271-275.
159. Niemczyk, M., Y. Ito, J. Huddleston, A. Git, S. Abu-Amero, C. Caldas, G. E. Moore*, et al.* (2013). "**Imprinted Chromatin around DIRAS3 Regulates Alternative Splicing of GNG12-AS1, a Long Noncoding RNA**." American Journal of Human Genetics 93(2): 224-235.
160. Nishinakamura, R. and M. Takasato (2005). "**Essential roles of Sall1 in kidney development**." Kidney Int 68(5): 1948-1950.
161. Nizon, M., V. Laugel, K. M. Flanigan, M. Pastore, M. A. Waldrop, J. A. Rosenfeld, R. Marom*, et al.* (2019). "**Variants in MED12L, encoding a subunit of the mediator kinase module, are responsible for intellectual disability associated with transcriptional defect**." Genet Med.
162. Olexiouk, V., W. Van Criekinge and G. Menschaert (2018). "**An update on sORFs.org: a repository of small ORFs identified by ribosome profiling**." Nucleic Acids Res 46(D1): D497-D502.
163. Oliver, P. L., R. A. Chodroff, A. Gosal, B. Edwards, A. F. Cheung, J. Gomez-Rodriguez, G. Elliot*, et al.* (2015). "**Disruption of Visc-2, a Brain-Expressed Conserved Long Noncoding RNA, Does Not Elicit an Overt Anatomical or Behavioral Phenotype**." Cereb Cortex 25(10): 3572-3585.

164. Orphanides, G., T. Lagrange and D. Reinberg (1996). "**The general transcription factors of RNA polymerase II**." Genes Dev 10(21): 2657-2683.
165. Pall, G. S., J. Wallis, R. Axton, D. G. Brownstein, P. Gautier, K. Buerger, C. Mulford*, et al.* (2004). "**A novel transmembrane MSP-containing protein that plays a role in right ventricle development**." Genomics 84(6): 1051-1059.
166. Paralkar, V. R., C. C. Taborda, P. Huang, Y. Yao, A. V. Kossenkov, R. Prasad, J. Luan*, et al.* (2016). "**Unlinking an lncRNA from Its Associated cis Element**." Molecular Cell 62(1): 104-110.
167. Parikshak, N. N., V. Swarup, T. G. Belgard, M. Irimia, G. Ramaswami, M. J. Gandal, C. Hartl*, et al.* (2016). "**Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism**." Nature 540(7633): 423-427.
168. Park, J. M., H. S. Kim, S. J. Han, M. S. Hwang, Y. C. Lee and Y. J. Kim (2000). "**In vivo requirement of activator-specific binding targets of mediator**." Mol Cell Biol 20(23): 8709-8719.
169. Pauler, F. M., M. A. Sloane, R. Huang, K. Regha, M. V. Koerner, I. Tamir, A. Sommer*, et al.* (2009). "**H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome**." Genome Res 19(2): 221-233.
170. Penny, G. D., G. F. Kay, S. A. Sheardown, S. Rastan and N. Brockdorff (1996). "**Requirement for Xist in X chromosome inactivation**." Nature 379(6561): 131-137.
171. Petrenko, N., Y. Jin, K. H. Wong and K. Struhl (2016). "**Mediator Undergoes a Compositional Change during Transcriptional Activation**." Mol Cell 64(3): 443-454.
172. Petrenko, N., Y. Jin, K. H. Wong and K. Struhl (2017). "**Evidence that Mediator is essential for Pol II transcription, but is not a required component of the preinitiation complex in vivo**." Elife 6.
173. Pettitt, S. J., Q. Liang, X. Y. Rairdan, J. L. Moran, H. M. Prosser, D. R. Beier, K. C. Lloyd*, et al.* (2009). "**Agouti C57BL/6N embryonic stem cells for mouse genetic resources**." Nat Methods 6(7): 493-495.
174. Philibert, R. A., S. L. Winfield, P. Damschroder-Williams, C. Tengstrom, B. M. Martin and E. I. Ginns (1999). "**The genomic structure and developmental expression patterns of the human OPA-containing gene (HOPA)**." Human Genetics 105(1-2): 174-178.
175. Plaschka, C., L. Lariviere, L. Wenzeck, M. Seizl, M. Hemann, D. Tegunov, E. V. Petrotchenko*, et al.* (2015). "**Architecture of the RNA polymerase II-Mediator core initiation complex**." Nature 518(7539): 376-380.
176. Poss, Z. C., C. C. Ebmeier, A. T. Odell, A. Tangpeerachaikul, T. Lee, H. E. Pelish, M. D. Shair*, et al.* (2016). "**Identification of Mediator Kinase Substrates in Human Cells using Cortistatin A and Quantitative Phosphoproteomics**." Cell Rep 15(2): 436-450.
177. Poss, Z. C., C. C. Ebmeier and D. J. Taatjes (2013). "**The Mediator complex and transcription regulation**." Crit Rev Biochem Mol Biol 48(6): 575-608.
178. Proudfoot, N. J. and G. G. Brownlee (1976). "**3' Non-Coding Region Sequences in Eukaryotic Messenger-Rna**." Nature 263(5574): 211-214.
179. Pruitt, K. D., T. Tatusova and D. R. Maglott (2007). "**NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**." Nucleic Acids Research 35: D61-D65.
180. Qi, L. S., M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin and W. A. Lim (2013). "**Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression**." Cell 152(5): 1173-1183.
181. Qiu, L. H., J. A. Rivera-Perez and Z. S. Xu (2011). "**A Non-Specific Effect Associated with Conditional Transgene Expression Based on Cre-loxP Strategy in Mice**." Plos One 6(5).
182. Quinlan, A. R. and I. M. Hall (2010). "**BEDTools: a flexible suite of utilities for comparing genomic features**." Bioinformatics 26(6): 841-842.

183. Rau, M. J., S. Fischer and C. J. Neumann (2006). "**Zebrafish Trap230/Med12 is required as a coactivator for Sox9-dependent neural crest, cartilage and ear development**." Dev Biol 296(1): 83-93.
184. Real, F. X. (2007). "**p53: It has it all, but will it make it to the clinic as a marker in bladder cancer?**" Journal of Clinical Oncology 25(34): 5341-5344.
185. Reavey, C. T., M. J. Hickman, K. C. Dobi, D. Botstein and F. Winston (2015). "**Analysis of Polygenic Mutants Suggests a Role for Mediator in Regulating Transcriptional Activation Distance in Saccharomyces cerevisiae**." Genetics 201(2): 599-612.
186. Risheg, H., J. M. Graham, Jr., R. D. Clark, R. C. Rogers, J. M. Opitz, J. B. Moeschler, A. P. Peiffer*, et al.* (2007). "**A recurrent mutation in MED12 leading to R961W causes Opitz-Kaveggia syndrome**." Nat Genet 39(4): 451-453.
187. Risso, D., J. Ngai, T. P. Speed and S. Dudoit (2014). "**Normalization of RNA-seq data using factor analysis of control genes or samples**." Nature Biotechnology 32(9): 896-902.
188. Ritter, N., T. Ali, N. Kopitchinski, P. Schuster, A. Beisaw, D. A. Hendrix, M. H. Schulz*, et al.* (2019). "**The lncRNA Locus Handsdown Regulates Cardiac Gene Programs and Is Essential for Early Mouse Development**." Dev Cell.
189. Rivas, E., J. Clements and S. R. Eddy (2016). "**A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs**." Nature Methods 14: 45.
190. Robb, L., I. Lyons, R. Li, L. Hartley, F. Kontgen, R. P. Harvey, D. Metcalf*, et al.* (1995). "**Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene**." Proc Natl Acad Sci U S A 92(15): 7075-7079.
191. Robert, C. and M. Watson (2015). "**Errors in RNA-Seq quantification affect genes of relevance to human disease**." Genome Biology 16.
192. Robinson, P. J., M. J. Trnka, D. A. Bushnell, R. E. Davis, P. J. Mattei, A. L. Burlingame and R. D. Kornberg (2016). "**Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex**." Cell 166(6): 1411-1422 e1416.
193. Rocha, P. P., W. Bleiss and H. Schrewe (2010). "**Mosaic expression of Med12 in female mice leads to exencephaly, spina bifida, and craniorachischisis**." Birth Defects Res A Clin Mol Teratol 88(8): 626-632.
194. Rocha, P. P., M. Scholze, W. Bleiss and H. Schrewe (2010). "**Med12 is essential for early mouse development and for canonical Wnt and Wnt/PCP signaling**." Development 137(16): 2723-2731.
195. Roeder, R. G. and W. J. Rutter (1969). "**Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms**." Nature 224(5216): 234-237.
196. Sabari, B. R., A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham*, et al.* (2018). "**Coactivator condensation at super-enhancers links phase separation and gene control**." Science 361(6400).
197. Saldanha, A. J. (2004). "**Java Treeview-extensible visualization of microarray data**." Bioinformatics 20(17): 3246-3248.
198. Sato, A., S. Kishida, T. Tanaka, A. Kikuchi, T. Kodama, M. Asashima and R. Nishinakamura (2004). "**Sall1, a causative gene for Townes-Brocks syndrome, enhances the canonical Wnt signaling by localizing to heterochromatin**." Biochemical and Biophysical Research Communications 319(1): 103-113.
199. Sato, S., C. Tomomori-Sato, T. J. Parmely, L. Florens, B. Zybailov, S. K. Swanson, C. A. Banks*, et al.* (2004). "**A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology**." Mol Cell 14(5): 685-691.

200. Sauvageau, M., L. A. Goff, S. Lodato, B. Bonev, A. F. Groff, C. Gerhardinger, D. B. Sanchez-Gomez*, et al.* (2013). "**Multiple knockout mouse models reveal lincRNAs are required for life and brain development**." Elife 2: e01749.
201. Schaukowitch, K., J. Y. Joo, X. Liu, J. K. Watts, C. Martinez and T. K. Kim (2014). "**Enhancer RNA facilitates NELF release from immediate early genes**." Mol Cell 56(1): 29-42.
202. Schmitz, S. U., P. Grote and B. G. Herrmann (2016). "**Mechanisms of long noncoding RNA function in development and disease**." Cell Mol Life Sci 73(13): 2491-2509.
203. Schultes, E. A., A. Spasic, U. Mohanty and D. P. Bartel (2005). "**Compact and ordered collapse of randomly generated RNA sequences**." Nat Struct Mol Biol 12(12): 1130-1136.
204. Schultz, P., S. Fribourg, A. Poterszman, V. Mallouh, D. Moras and J. M. Egly (2000). "**Molecular structure of human TFIIH**." Cell 102(5): 599-607.
205. Schwartz, C. E., P. S. Tarpey, H. A. Lubs, A. Verloes, M. M. May, H. Risheg, M. J. Friez*, et al.* (2007). "**The original Lujan syndrome family has a novel missense mutation (p.N1007S) in the MED12 gene**." Journal of Medical Genetics 44(7): 472-477.
206. Senapathy, P., M. B. Shapiro and N. L. Harris (1990). "**Splice Junctions, Branch Point Sites, and Exons - Sequence Statistics, Identification, and Applications to Genome Project**." Methods in Enzymology 183: 252-278.
207. Shlyueva, D., G. Stampfel and A. Stark (2014). "**Transcriptional enhancers: from properties to genome-wide predictions**." Nature Reviews Genetics 15(4): 272-286.
208. Shogren-Knaak, M., H. Ishii, J. M. Sun, M. J. Pazin, J. R. Davie and C. L. Peterson (2006). "**Histone H4-K16 acetylation controls chromatin structure and protein interactions**." Science 311(5762): 844-847.
209. Smola, M. J., G. M. Rice, S. Busan, N. A. Siegfried and K. M. Weeks (2015). "**Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis**." Nat Protoc 10(11): 1643-1669.
210. Soutourina, J. (2018). "**Transcription regulation by the Mediator complex**." Nat Rev Mol Cell Biol 19(4): 262-274.
211. Spitale, R. C., R. A. Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J. W. Jung, H. Y. Kuchelmeister*, et al.* (2015). "**Structural imprints in vivo decode RNA regulatory mechanisms**." Nature 519(7544): 486-490.
212. Sun, S., B. C. Del Rosario, A. Szanto, Y. Ogawa, Y. Jeon and J. T. Lee (2013). "**Jpx RNA activates Xist by evicting CTCF**." Cell 153(7): 1537-1551.
213. Suzuki, M. M. and A. Bird (2008). "**DNA methylation landscapes: provocative insights from epigenomics**." Nat Rev Genet 9(6): 465-476.
214. Taatjes, D. J., A. M. Naar, F. Andel, 3rd, E. Nogales and R. Tjian (2002). "**Structure, function, and activator-induced conformations of the CRSP coactivator**." Science 295(5557): 1058-1062.
215. Thiagarajan, R. D., N. Cloonan, B. B. Gardiner, T. R. Mercer, G. Kolle, E. Nourbakhsh, S. Wani*, et al.* (2011). "**Refining transcriptional programs in kidney development by integration of deep RNA-sequencing and array-based spatial profiling**." BMC Genomics 12: 441.
216. Thomas, S., J. G. Underwood, E. Tseng, A. K. Holloway and B. B. C. Informatics (2014). "**Long-Read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms**." Plos One 9(4).
217. Thompson, C. M., A. J. Koleske, D. M. Chao and R. A. Young (1993). "**A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast**." Cell 73(7): 1361-1375.

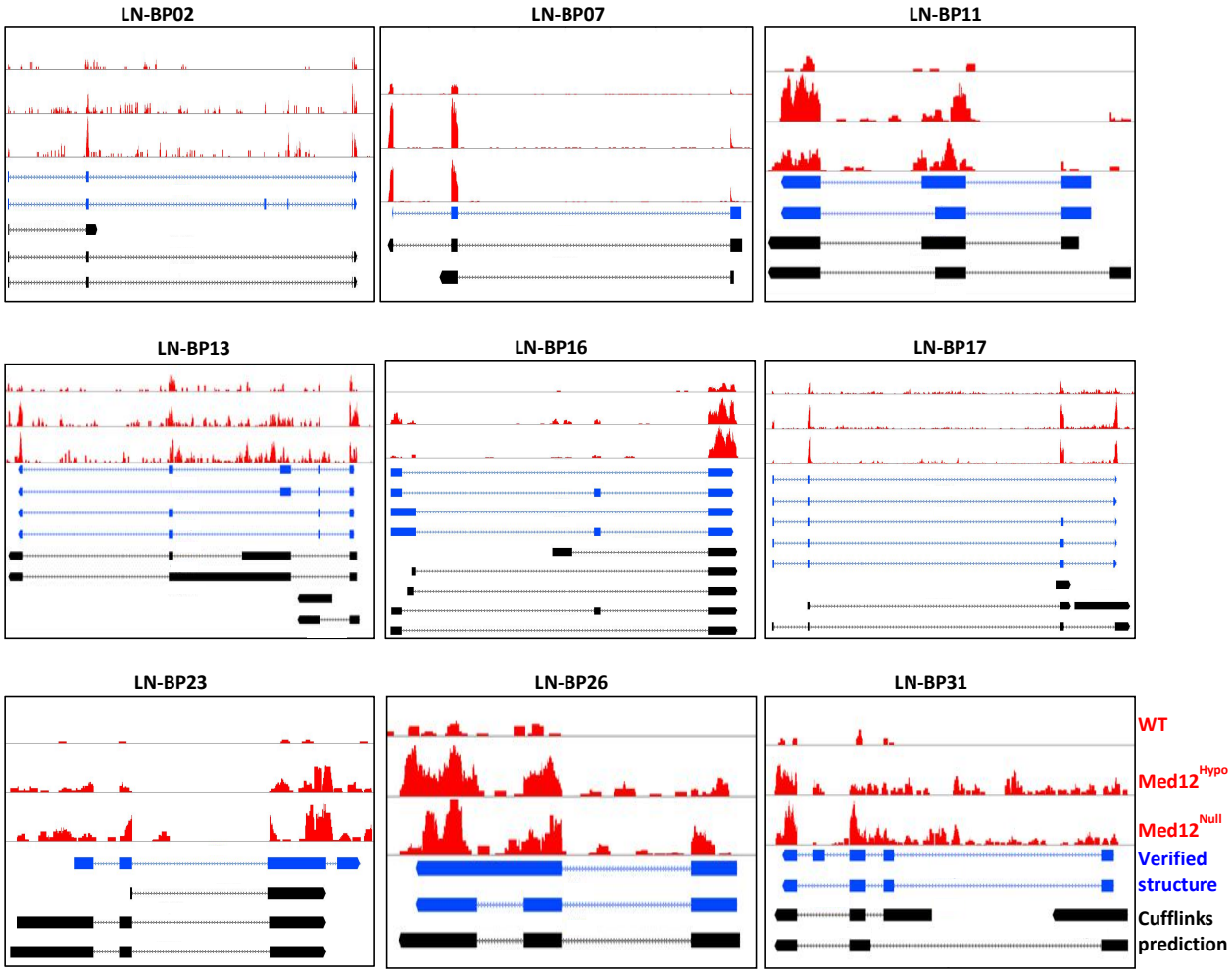
218. Toth-Petroczy, A., C. J. Oldfield, I. Simon, Y. Takagi, A. K. Dunker, V. N. Uversky and M. Fuxreiter (2008). "**Malleable machines in transcription regulation: the mediator complex**." PLoS Comput Biol 4(12): e1000243.
219. Townes, P. L. and E. R. Brocks (1972). "**Hereditary Syndrome of Imperforate Anus with Hand, Foot, and Ear Anomalies**." Journal of Pediatrics 81(2): 321-+.
220. Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel*, et al.* (2012). "**Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**." Nat Protoc 7(3): 562-578.
221. Tsai, K. L., S. Sato, C. Tomomori-Sato, R. C. Conaway, J. W. Conaway and F. J. Asturias (2013). "**A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction**." Nat Struct Mol Biol 20(5): 611-619.
222. Tsai, K. L., C. Tomomori-Sato, S. Sato, R. C. Conaway, J. W. Conaway and F. J. Asturias (2014). "**Subunit Architecture and Functional Modular Rearrangements of the Transcriptional Mediator Complex**." Cell 158(2): 463.
223. Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi*, et al.* (2010). "**Long noncoding RNA as modular scaffold of histone modification complexes**." Science 329(5992): 689-693.
224. Tseng, Y. Y., B. S. Moriarity, W. Gong, R. Akiyama, A. Tiwari, H. Kawakami, P. Ronning*, et al.* (2014). "**PVT1 dependence in cancer with MYC copy-number increase**." Nature 512(7512): 82-86.
225. Tu, J., G. Tian, H. H. Cheung, W. Wei and T. L. Lee (2018). "**Gas5 is an essential lncRNA regulator for self-renewal and pluripotency of mouse embryonic stem cells and induced pluripotent stem cells**." Stem Cell Res Ther 9(1): 71.
226. Tuan, D., S. M. Kong and K. Hu (1992). "**Transcription of the Hypersensitive Site Hs2 Enhancer in Erythroid-Cells**." Proceedings of the National Academy of Sciences of the United States of America 89(23): 11219-11223.
227. Tutter, A. V., M. P. Kowalski, G. A. Baltus, V. Iourgenko, M. Labow, E. Li and S. Kadam (2009). "**Role for Med12 in regulation of Nanog and Nanog target genes**." J Biol Chem 284(6): 3709-3718.
228. Tuttle, L. M., D. Pacheco, L. Warfield, J. Luo, J. Ranish, S. Hahn and R. E. Klevit (2018). "**Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex**." Cell Rep 22(12): 3251-3264.
229. Ulitsky, I., A. Shkumatava, C. H. Jan, H. Sive and D. P. Bartel (2011). "**Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution**." Cell 147(7): 1537-1550.
230. van den Brink, S. C., P. Baillie-Johnson, T. Balayo, A. K. Hadjantonakis, S. Nowotschin, D. A. Turner and A. M. Arias (2014). "**Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells**." Development 141(22): 4231-4242.
231. Vogl, M. R., S. Reiprich, M. Kuspert, T. Kosian, H. Schrewe, K. A. Nave and M. Wegner (2013). "**Sox10 cooperates with the mediator subunit 12 during terminal differentiation of myelinating glia**." J Neurosci 33(15): 6679-6690.
232. Vojnic, E., A. Mourao, M. Seizl, B. Simon, L. Wenzeck, L. Lariviere, S. Baumli*, et al.* (2011). "**Structure and VP16 binding of the Mediator Med25 activator interaction domain**." Nature Structural & Molecular Biology 18(4): 404-U429.
233. Vulto-van Silfhout, A. T., B. B. de Vries, B. W. van Bon, A. Hoischen, M. Ruiterkamp-Versteeg, C. Gilissen, F. Gao*, et al.* (2013). "**Mutations in MED12 cause X-linked Ohdo syndrome**." Am J Hum Genet 92(3): 401-406.
234. Wang, J., C. Gong and L. E. Maquat (2013). "**Control of myogenesis by rodent SINE-containing lncRNAs**." Genes Dev 27(7): 793-804.

235. Wang, L., H. J. Park, S. Dasari, S. Wang, J. P. Kocher and W. Li (2013). "**CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model**." Nucleic Acids Res 41(6): e74.
236. Wang, W., L. Huang, Y. Huang, J. W. Yin, A. J. Berk, J. M. Friedman and G. Wang (2009). "**Mediator MED23 links insulin signaling to the adipogenesis transcription cascade**." Dev Cell 16(5): 764-771.
237. Wang, X., N. Yang, E. Uno, R. G. Roeder and S. Guo (2006). "**A subunit of the mediator complex regulates vertebrate neuronal development**." Proc Natl Acad Sci U S A 103(46): 17284-17289.
238. Wani, M. A., S. E. Wert and J. B. Lingrel (1999). "**Lung Kruppel-like factor, a zinc finger transcription factor, is essential for normal lung development**." J Biol Chem 274(30): 21180-21185.
239. Weideman, C. A., R. C. Netter, L. R. Benjamin, J. J. McAllister, L. A. Schmiedekamp, R. A. Coleman and B. F. Pugh (1997). "**Dynamic interplay of TFIIA, TBP and TATA DNA**." J Mol Biol 271(1): 61-75.
240. Weirather, J. L., M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X. J. Wang, D. Buck*, et al.* (2017). "**Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis**." F1000Res 6: 100.
241. Wery, M., M. Descrimes, N. Vogt, A. S. Dallongeville, D. Gautheret and A. Morillon (2016). "**Nonsense-Mediated Decay Restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded RNA Structure**." Molecular Cell 61(3): 379-392.
242. Whyte, W. A., D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl*, et al.* (2013). "**Master transcription factors and mediator establish super-enhancers at key cell identity genes**." Cell 153(2): 307-319.
243. Wu, B., M. Slabicki, L. Sellner, S. Dietrich, X. Liu, A. Jethwa, J. Hullein*, et al.* (2017). "**MED12 mutations and NOTCH signalling in chronic lymphocytic leukaemia**." Br J Haematol 179(3): 421-429.
244. Wutz, A. and R. Jaenisch (2000). "**A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation**." Mol Cell 5(4): 695-705.
245. Yan, X., Z. Hu, Y. Feng, X. Hu, J. Yuan, S. D. Zhao, Y. Zhang*, et al.* (2015). "**Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers**." Cancer Cell 28(4): 529-540.
246. Yean, S. L., G. Wuenschell, J. Termini and R. J. Lin (2000). "**Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome**." Nature 408(6814): 881-884.
247. Yin, Q. F., L. Yang, Y. Zhang, J. F. Xiang, Y. W. Wu, G. G. Carmichael and L. L. Chen (2012). "**Long Noncoding RNAs with snoRNA Ends**." Molecular Cell 48(2): 219-230.
248. Yoda, A., H. Kouike, H. Okano and H. Sawa (2005). "**Components of the transcriptional Mediator complex are required for asymmetric cell division in C. elegans**." Development 132(8): 1885-1893.
249. Yue, F., Y. Cheng, A. Breschi, J. Vierstra, W. S. Wu, T. Ryba, R. Sandstrom*, et al.* (2014). "**A comparative encyclopedia of DNA elements in the mouse genome**." Nature 515(7527): 355-+.
250. Zhang, Y., R. A. Fillmore and W. E. Zimmer (2008). "**Structural and functional analysis of domains mediating interaction between the bagpipe homologue, Nkx3.1 and serum response factor**." Exp Biol Med (Maywood) 233(3): 297-309.
251. Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum*, et al.* (2008). "**Model-based Analysis of ChIP-Seq (MACS)**." Genome Biology 9(9).
252. Zhao, S. R., Y. Zhang, W. Gordon, J. Quan, H. L. Xi, S. Du, D. von Schack*, et al.* (2015). "**Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap**." Bmc Genomics 16.
253. Zhou, H., S. Kim, S. Ishii and T. G. Boyer (2006). "**Mediator modulates Gli3-dependent Sonic hedgehog signaling**." Mol Cell Biol 26(23): 8667-8682.

254. Zhou, H., J. M. Spaeth, N. H. Kim, X. Xu, M. J. Friez, C. E. Schwartz and T. G. Boyer (2012). "**MED12 mutations link intellectual disability syndromes with dysregulated GLI3-dependent Sonic Hedgehog signaling**." Proc Natl Acad Sci U S A 109(48): 19763-19768.

10. Appendices

10.1. Supplementary Figures



Supplementary Figure 1 - Misregulated novel putative new lncRNAs.
Browser view of experimentally obtained gene structure for the identified misregulated putative lncRNAs predicted by Cufflinks. Mapped reads from the 3 samples, red track; structure obtained for the genes after PCR amplification, blue track; gene structure predicted by Cufflinks based on mapped reads, black track

10.2. Supplementary tables

Supplementary Table 1 - Primers used to generate gRNA for CRISPR-Cas9 targeted DSB generation.
In red are the bases added in order to generate compatible ends for proper ligation into vector. In green are the bases added so that the first base after the U6 promoter that drives gRNA expression is a guanidine, since transcription from this promoter is more efficient if the first base is a guanidine.

Target location	Forward primer	Reverse primer
Sall1 intron 2	CACCGAAACTAATTTGTAGTCGTTC	AAACGAACGACTACAAATTAGTTTC
Sall1 exon 4	CACCGCCGGATTAAGACCGCCTAGC	AAACGCTAGGCGGTCTTAATCCGGC
LN-BP18 exon 4	CACCGCGGAGATCAGCTGCAGCTCA	AAACTGAGCTGCAGCTGATCTCCGC
LN-BP18 5' of TSS1	CACCGTCAAAATTACGGAAATCGAG	AAACCTCGATTTCCGTAATTTTGAC
LN-BP18 Intron 1	CACCGTTTCCGACACTTGCTACGTG	AAACCACGTAGCAAGTGTCGGAAAC
Ln-BP18 intron 2	CACCGTGGCTTGCAATGGCCAGTTT	AAACAAACTGGCCATTGCAAGCCAC
LN-BP18 Intron 3	CACCGAAGCTCTCGACCGTTTCATC	AAACGATGAAACGGTCGAGAGCTTC

Supplementary Table 2 - Primers used for screening engineered mutations

	Genomic Feature	Forward primer	Reverse primer
1	5' B-gal insertion border	ATGCTGCCTCCTGAAATGGT	GAATCTCCACTCGCCGTTCA
2	LN-BP18 Exon 4 WT	ATGCTGCCTCCTGAAATGGT	GCCATACAACCTCAATGGGGA
3	LN-BP18 3' border Deletion TSS1	GCCTGGCTCGGAGATCCTTC	ACCCGGAAGAGGGAGTACAG
4	LN-BP18 TSS1 flanking deletion	CTTCGGGGTCGGATTGGAAA	ACCCGGAAGAGGGAGTACAG
5	LN-BP18 3' border Deletion TSS2	CCCCTACACCTCTCCCAACT	ATCCCGCTTTACACTCGTCG
6	LN-BP18 TSS2 flanking deletion	GCCATGTACGTGTCCTGGTT	ATCCCGCTTTACACTCGTCG
7	Sall1 5' Border deletion WT	TGTGAGGGATAACCATCCTGC	CTTGCTCTTAGTGGGGCGAC
8	Sall1 flanking deletion	TGTGAGGGATAACCATCCTGC	CCTCTAAAGAATTCTGCGTGC

Supplementary Table 3 - PCR reaction

Taq polymerase		Takara polymerase	
Component	**µl**	**Component**	**µl**
Buffer 10x	2.50	5X PrimeSTAR Buffer (Mg^{2+} Plus)	5
MgCl2 (50mM)	0.75	NTP Mixture (2.5 mM each)	2
dNTPs (10mM)	0.50	Primer Fwd (100uM)	0.1
Primer Fwd (100uM)	0.05	Primer Rev (100uM)	0.1
Primer Rev (100uM)	0.05	PrimeSTAR HS DNA Polymerase	0.25
Taq	0.15	H2O	17.05
Template	0.50		
H2O	20.5		

Supplementary Table 4 - PCR run method

Step	Temperature	Time	Cycles
Initial denaturation	94°C	5min	1
Denaturation	94°C	30sec	40
Annealing	60°C	30sec	40
Elongation	72°C	30sec	40
Final elongation	72°C	7min	1

Supplementary Table 5 - qPCR run method

Step	Temperature	Time	Cycles
Initial denaturation	95°C	2min	1
Denaturation	95°C	15sec	40
Elongation	60°C	30sec	40
Melt Curve	95°C	15sec	1
	60°C	15sec	1
	95°C	15sec	1

Supplementary Table 6 - Primers used for qPCR quantification

Gene	Forward primer	Reverse primer
LN-BP18-Ex1sh-4	GGGAGGGAGTCCCGTGG	CACGGTGCCATAGCAGCTC
LN-BP18-Ex1-ln-4	CCAGAGTTCTCCACTCTCAGG	CACGGTGCCATAGCAGCTC
LN-BP18-Ex3-4	GCTGTTGCCCTGAGTTATGG	CACGGTGCCATAGCAGCTC
LN-BP18-Ex4sh-7	GAGCTGCTATGGCACCGTGA	GCCATGCTGCAGTCTTGAA
LN-BP18-Ex4-ln-7	AGGACCCTCCAGCAAGATGA	GCCATGCTGCAGTCTTGAA
LN-BP18-Ex6-7	TCTCCTGTCATATAGTCTGCTGTT	GCCATGCTGCAGTCTTGAA
LN-BP18-Ex1-β-gal	CCAGAGTTCTCCACTCTCAGG	CTTCTGGTCTTCACCCACCG
LN-BP18-Ex3-β-gal	GCTGTTGCCCTGAGTTATGG	CTTCTGGTCTTCACCCACCG
β-gal-LN-BP18-ex7	CTAGCTAGTCTAGGTCGAGCG	GCCATGCTGCAGTCTTGAA
β-gal	GAACGGACTGAGATGTGGCA	AGAAGGGGACCAGCTATCGT
Sall1	CAAGCGAAGCCTCAACATTT	ATCCTTGCTCTTAGTGGGGC
Oct4	TCAGCTTGGGCTAGAGAAGG	TGGGAAAGGTGTCCCTGTAG
Nanog	AAACCAAAGGATGAAGTGCAAG	GGATACTCCACTGGTGCTGAG
Sox2	TACAGCATGATGCAGGAGCAG	TCATGTAGGTCTGCGAGCTG
Hand1	GGCTGAACTCAAAAAGACGGA	CCTTTAATCCTCTTCTCGCCG
T	AACTGGTCTAGCCTCGGAGT	CTCACAGACCAGAGACTGGG
Xist	GGTTCTCTCTCCAGAAGCTAGGAA	TGGTAGATGGCATTGTGTATTATATGG
Pmm2	AGGGAAAGGCCTCACGTTCT	AATACCGCTTATCCCATCCTTCA

Supplementary Table 7 - Primers used for putative new lncRNAs isolation

ncRNA	Forward primer sequence	Reverse primer sequence
LN-BP02	GAAGGGTGGCCTGAAGACT	GTCAGAAGCTAACCACGCAG
	GGGTGTTTGTTCGTAGTGGG	TTATTGGTGTGGGTGTGGGT
LN-BP06	CTGCTGGGATGGGTTCAAAG	GGTAACTGGTCACACTGTATCT
	TGGAGTGGGATGCTGAGATG	GAACAGTTCATAGTGGTAGGGAA
LN-BP07	AAGCTCGCAAACCTTGTCAG	TCCAAAGACAAATCCCGCAG
	GCGTGAGAAGGGAGAGAGAA	ACTTCTTCATTGGGCCTCAGT
LN-BP11	AATGCGCAAAGTCTGGACAG	TGGTCATAGGTTGCCCAGAG
	CCACGCTTTCTGGAAAACCA	GGTCTTGAGGTGAGGGTCTC
LN-BP13	GCGCACACTCCTCAATCAAT	GAAGTGTGTTTGGGGTGGTC
	ATCAAACTCGGGTCCTCTGG	CCTGGCTTTGATCCACATGG
	CAGTGATTTCTGCTCGCCTC	
LN-BP16	GCTACACTGTCCTGCTCTGA	TCTTGTTTTCCGTCACGCTG
	AGTTTCTCTGAGGTGGCTCA	TTATTGGGGGTCCATGGG
LN-BP17	GCCTCTATTCTCTCCTGCGA	AAGCAGGAGTGTGGAAGGTT
	GCTCCTCCTGCATCATGACT	AGGTCCAAGAAGCCAGAAGT
	TCTGAGATCTCTGAGACCCA	
LN-BP18	GCTGTTGCCCTGAGTTATGG	CTGCGACGGTTCATGTGATT
	CTATGGCACCGTGAGCTG	TCGTTTAGTAGCTGCAGAGATT
LN-BP23	CGGACAAGAGGACCAGAAGT	CGGACAAGAGGACCAGAAGT
	TAGCTCCATGAGACCTTGCC	TAGCTCCATGAGACCTTGCC
LN-BP26	ACTGAGTGTTGGGTTGGAGT	CAGATACCCACTGCTCCTCA
	AGTCGTTCCACCTGTGTCTT	GTAGATGCAGGCAACATGGC
LN-BP31	CCCTGTTCAATCTGGAATGCC	GCTTGACACGATGGGTCAC
	CTCCCCGGAAGTGAAGTCTT	GCTAGCCTCCTGTTCTTTTCT

Supplementary Table 8 - Primers used during RACE protocol

Primer	Gene Location	Primer sequence
1	LN-BP18 exon7 #1	CTGCGACGGTTCATGTGATT
2	LN-BP18 exon7 #2	GCCATGCTGCAGTCTTGAA
3	LN-BP18 exon4	CACGGTGCCATAGCAGCTC
4	LN-BP18 exon1 #1	GAGTGGAGAACTCTGGGCGA
5	LN-BP18 exon1 #2	GGCTGCTTTCTGGTAGCTCT
6	LN-BP18 exon1 #3	GAACACTCACGAAATGGGGC
7	LN-BP18 exon3 #1	CTGCAGAACAAACGCTGTGG
8	LN-BP18 exon3 #2	TGTCCATAACTCAGGGCAACA
9	LN-BP18 exon4 #1	GTAAAAAGGCAGAGACTGCTGG
10	LN-BP18 exon4 #2	GAGCTGCTATGGCACCGTGA

Supplementary Table 9 - Antibodies used for Western Blot

Antibody	Host	Supplier	Dilution
α-Sall1	Rabbit	ab31526; Abcam	1:700
α-β-actin	Mouse	A5441, Sigma-Aldrich	1:20000
α-H3	Rabbit	ab1791, Abcam	1:5000
α-GapdH	Rabbit	5174, Cell Signaling	1:1000
α-Rabbit-HRP	Goat	111-035-003, Jackson ImmunoResearch	1:5000
α-Mouse-HRP	Goat	115-035-003, Jackson ImmunoResearch	1:5000

Supplementary Table 10 - Transcriptome data downloaded from the Encode database to assess tissue expression of lncRNAs

Tissue	Sample	Tissue	Sample	Tissue	Sample
Liver E11.5	ENCFF734KYK	**Limb E14.5**	ENCFF888CQF	**forebrain E12.5**	ENCFF604YPA
Liver E12.5	ENCFF527RRL	**Limb E15.5**	ENCFF214NGX	**forebrain E13.5**	ENCFF652LRB
Liver E13.5	ENCFF039JSB	**Lung E14.5**	ENCFF674RZX	**forebrain E14.5**	ENCFF242VUV
Liver E14.5	ENCFF915IMF	**Lung E15.5**	ENCFF099TSL	**forebrain E15.5**	ENCFF370MZR
Liver E15.5	ENCFF257DUT	**Lung E16.5**	ENCFF765KXH	**forebrain E16.5**	ENCFF130CQI
Liver E16.5	ENCFF487NRJ	**Lung day0**	ENCFF471KGO	**forebrainday0**	ENCFF190EHR
Liver day0	ENCFF774ZXX	**neural tube E11.5**	ENCFF485QMJ	**midbrain E11.5**	ENCFF739IFQ
Heart E10.5	ENCFF655GAO	**neural tube E12.5**	ENCFF413LWR	**midbrain E12.5**	ENCFF107QWR
Heart E11.5	ENCFF276KTU	**neural tube E13.5**	ENCFF358GZH	**midbrain E13.5**	ENCFF750QEC
Heart E12.5	ENCFF716KQU	**neural tube E14.5**	ENCFF403FVS	**midbrain E14.5**	ENCFF112LWR
Heart E13.5	ENCFF709XPI	**neural tube E15.5**	ENCFF351DAS	**midbrain E15.5**	ENCFF287UJS
Heart E14.5	ENCFF808WIM	**neural tube day0**	ENCFF685ICX	**midbrain E16.5**	ENCFF864PZZ
Heart E15.5	ENCFF457OAN	**stomach E14.5**	ENCFF643WID	**midbrain day0**	ENCFF398USF
Heart E16.5	ENCFF527FHQ	**stomach E15.5**	ENCFF545CDS	**hindbrain E10.5**	ENCFF844SJF
Heart day0	ENCFF081UNH	**stomach E16.5**	ENCFF929PNC	**hindbrain E11.5**	ENCFF622NTO
Kidney E14.5	ENCFF913EKF	**stomach day0**	ENCFF563UZS	**hindbrain E12.5**	ENCFF961UYK
Kidney E15.5	ENCFF850SDV	**thymus day0**	ENCFF014KTI	**hindbrain E13.5**	ENCFF372BPJ
Kidney E16.5	ENCFF635AXY	**intestine E14.5**	ENCFF694HJE	**hindbrain E14.5**	ENCFF484FNM
Kidney day0	ENCFF769XWI	**intestine E15.5**	ENCFF790QQF	**hindbrain E15.5**	ENCFF215SES
Limb E10.5	ENCFF752XWD	**intestine E16.5**	ENCFF281HBX	**hindbrain E16.5**	ENCFF804RYC
Limb E11.5	ENCFF076RLX	**intestine day0**	ENCFF596ATI	**hindbrain day0**	ENCFF852ZHY
Limb E12.5	ENCFF946JHJ	**forebrain E10.5**	ENCFF046NCT		
Limb E13.5	ENCFF759OCS	**forebrain E11.5**	ENCFF275ARE		

Supplementary Table 11 - Tools used for bioinformatics analysis

Tool	Version
Ballgown	3.9
Bamtools	2.5.1
bedtools	2.27.1
Cpat	2.0.0
Cufflinks	2.2.1
DAVID	6.8
Deseq2	2.11.40.6
FastQC	0.11.8
Fetch closest non-overlapping feature	4.0.1
Gene Cluster	3.0
genomeCoverageBed	2.27.1
Hisat2	2.0.5
IGB	9.0.2
Java TreeView	1.1.6r4
Macs2	2.1.2
R	3.5.1
Samtools	1.9
StepOnePlus Software	2.3
Tophat2	2.1.1
VennDiagram	1.6.20
wigToBigWig	377

Supplementary Table 12 - Go terms enriched in misregulated protein coding genes.

Cluster	Go term	p-value
1	translation	1.26E-07
	ribosomal small subunit assembly	1.34E-05
	oxidation-reduction process	3.01E-04
	cytoplasmic translation	3.49E-04
	ribosomal small subunit biogenesis	1.58E-02
	carbohydrate metabolic process	2.12E-02
	protein homotetramerization	2.70E-02
	monocyte chemotaxis	3.56E-02
	lipid metabolic process	3.60E-02
	immune response	3.61E-02
	cardiac muscle contraction	4.19E-02
2	negative regulation of transcription from RNA polymerase II promoter	8.98E-10
	positive regulation of transcription, DNA-templated	4.60E-06
	positive regulation of transcription from RNA polymerase II promoter	2.25E-05
	cell differentiation	2.42E-04
	positive regulation of cell proliferation	2.48E-04
	regulation of cell proliferation	4.39E-04
	response to mechanical stimulus	5.33E-04
	mammary gland development	9.04E-04
	heart development	1.21E-03
	transcription, DNA-templated	1.38E-03
	regulation of transcription, DNA-templated	1.43E-03
	mammary gland duct morphogenesis	2.12E-03
	response to corticosterone	3.59E-03
	cellular response to peptide	6.45E-03
	axon guidance	6.63E-03
	palate development	7.35E-03
3	outer dynein arm assembly	9.46E-04
	inflammatory response	1.05E-03
	positive regulation of phagocytosis	4.28E-03
	regulation of smooth muscle contraction	5.54E-03
	ion transport	1.02E-02
	wound healing	1.13E-02
	visual learning	2.03E-02
	placenta development	2.46E-02
	response to stimulus	2.61E-02
	peptide cross-linking	3.00E-02

	cilium movement	3.00E-02
	outflow tract septum morphogenesis	5.17E-02
4	regulation of transcription, DNA-templated	2.14E-17
	transcription, DNA-templated	3.63E-17
	positive regulation of transcription from RNA polymerase II promoter	5.79E-13
	positive regulation of transcription, DNA-templated	9.96E-10
	negative regulation of transcription from RNA polymerase II promoter	6.15E-06
	regulation of gene expression	8.24E-06
	axon guidance	6.27E-05
	covalent chromatin modification	2.24E-04
	positive regulation of cell proliferation	4.05E-04
	circadian rhythm	4.29E-04
	patterning of blood vessels	5.99E-04
	embryonic hemopoiesis	1.06E-03
	negative regulation of keratinocyte proliferation	1.06E-03
	negative regulation of transcription, DNA-templated	1.25E-03
	multicellular organism development	1.30E-03
	nervous system development	1.54E-03
	organ morphogenesis	2.20E-03
	cell differentiation	2.44E-03
	rhythmic process	2.62E-03
	heart development	2.72E-03
5	angiogenesis	2.91E-04
	positive regulation of gene expression	4.51E-03
	hepatocyte apoptotic process	4.96E-03
	negative regulation of apoptotic process	1.12E-02
	meiotic cell cycle	2.10E-02
	positive regulation of sequence-specific DNA binding transcription factor activity	2.20E-02
	negative regulation of peptidase activity	2.67E-02
	positive regulation of tumor necrosis factor production	2.86E-02
	cellular response to transforming growth factor beta stimulus	3.48E-02
	protein destabilization	3.48E-02
	collagen-activated tyrosine kinase receptor signaling pathway	3.88E-02
	regulation of nucleocytoplasmic transport	3.88E-02
	brain development	4.08E-02
	response to thyroid hormone	4.82E-02
	placenta development	4.85E-02

negative regulation of RNA polymerase II regulatory region sequence-specific DNA binding	5.76E-02
negative regulation of glycolytic process	6.69E-02
positive regulation of angiogenesis	7.21E-02
phagocytosis, engulfment	7.60E-02
positive regulation of calcium ion import	7.60E-02
positive regulation of MAPK cascade	7.78E-02
spermatogenesis	7.80E-02
neuron differentiation	8.95E-02

10.3. Abbreviations

bp	base pairs
β-gal	beta-galactosidase
CAGE	Cap analysis gene expression
ChIP	chromatin Immunoprecipitation
cDNA	complementary DNA
DSB	double strand break
E	embryonic day
FDR	false discovery rate
FACS	fluorescence-activated cell sorting
FC	fold change
FPKM	fragments per kilo base per million mapped reads
GO	gene ontology
GSP	gene specific primer
kb	kilo bases
lncRNA	long non-coding RNA
mRNA	messenger RNA
mESC	mouse embryonic stem cells
ncRNA	non-coding RNA
ncRNA-a	non-coding RNA-activating
ORF	open reading frame
pA	polyadenylation
PCR	polymerase chain reaction
PIC	pre-initiation complex
PAM	protospacer adjacent motif
qPCR	quantitative real-time PCR
RACE	Rapid Amplification of cDNA Ends
Pol II	RNA polymerase II
RT	room temperature
TBS	Townes-Brocks syndrome
TES	transcription end site
TSS	transcription start site
WISH	whole-mount in situ hybridization
WT	wild type
XCI	X chromosome inactivation
XLID	X chromosome linked intellectual disability syndromes

10.4. List of Tables

10.5. List of Figures

10.6. Errata

Page	Paragraph	Line	Correction
11	2	6	for “lysine 7” read “lysine 27”;
18	2	9	discard “obtained from”;
20	5	6	for “insertion” read “interaction”;
23	2	8	for “resulted selective” read “resulted in selective”;
26	2	8	for “core, Mediator” read “core Mediator”;
27	1	2	for “While Med12^{null}” read “Med12^{null}”;
27	2	3	for “demethylation” read “dimethylation”;
28	4	6	For “G1148R” read “R1148H”
32	5	4	for “acting is *cis*” read “acting in *cis*”;
33	2	1	for “lncRNAwhich ” read “lncRNA which ”;
37	1	1	for “to transcription of the ” read “to the act of transcription”;
39	2	2	for “with the need” read “without the need”;
39	3	7	for “T. he sequencing” read “. The sequencing”;
40	3	5	for “even if there are ” read “even if they are ”;
51	2	7	for “exposure on extended” read “exposure was extended”;
56	1	6	for “were at 95°C” read “were boiled at 95°C”;
58	6	1	for “Functional enrichment” read “For functional enrichment ”;
58	6	2	discard “was used”;
59	2	3	for “section 0” read “section 2.12.2”;
61	1	6	for “Wnt/PCR“ read “Wnt/PCP”;
61	3	8	for “data generate“ read “data generated”;
61	3	9	for “disturbed gees“ read “disturbed genes”;
66	3	3	for “analysis the “ read “analysis of the ”;
70	Table 4	7	For “LN-BP13_6” read “LN-BP16_1”
70	1	11	for “LN-B06 detection “ read “LN-BP06 detection ”;
78	3	1	for “The first observations” read “One of the first observations”;
78	3	8	for “order to amplified” read “order to amplify”;
79	2	1	for “Comparison of LN-BP” read “Comparison of LN-BP18”;
79	3	4	for “encoded by this lncRNAs” read “encoded by this lncRNA”;
87	2	1	for “Briefly, 3.0x105 G4” read “Briefly, 3.0x10^5 G4”;
90:	Fig. 19 caption	4	for “into mesodermal“ read “into paraxial mesoderm”;

93	1	2	for “strong was” read “strong signal was ”;
94	1	2	for “(section 0)” read “(section 3.3.1)”;
96	1	7	for “In the two” read “In two”;
98	4	6	for “section 0” read “section 2.11”;
101	1	1	for “cells with devoid” read “cells devoid”;
101	1	2	for “expression was also” read “expression also”;
102	2	5	for “ detecting and error” read “detecting error”;
102	3	4	for “aligner perform” read “aligner performed”;
102	3	6	for “Tophat2 it is” read “Tophat2 is”;
106	Fig. 26		Figure **b)** and **c)** are swapped.
107	1	14	for “between this regions” read “between these regions”;
107	1	21	for “genes By” read “genes. By”;
108	Fig. 27 caption	1	for “(-1kb TSS to TES)“ read “(+1kb TSS to TES)”;
109	3	3	for “that the several” read “that several”;
113	1	4	for “Med12has” read “Med12 has”;
114	1	5	for “to of ” read “to”;
116	2	12	for “Slc1a2” read “Slc16a2 ”
117	2	3	for “transcript ” read “transcripts”;
120	1	30	for “in in ” read “in”;
121	1	1	for “TSS2repressed ” read “TSS2 repressed ”;
121	2	11	for “region ” read “regions ”;
121	3	9	for “sequence” read “sequenced”;
122	1	4	for “LN-BP_010” read “LN-BP18_010”;
122	1	5	for “reaming” read “remaining”;
122	1	6	for “as such as” read “as such was”;
122	2	7	for “LN-BP” read “LN-BP18”;
123	3	4	for “silar” read “similar”;
123	3	5	discard “co-expression and/or similar functions”;
124	1	6	for “the LN-BP18 ” read “that LN-BP18 ”;
124	1	8	for “LNBP18 ” read “LN-BP18”;
128	1	5	discard “As kidneys”;
128	1	8	for “LN-BP18 observed ” read “LN-BP18 was observed”;
128	1	21	for “Pn-BP18” read “LN-BP18 ”;
130	1	8	for “Esc” read “ESC”;

130	2	7	for “EScs” read “ESCs”;
130	2	13	for “shit” read “shift”;
130	2	13	for “400bo” read “400bp”;
132	2	11	for “from with” read “from which”;
133	1	11	for “obtained” read “obtaining”;
134	3	7	for “reagions” read “regions”;
135	2	9	for “1-fold” read “2-fold”;
135	2	10	for “2-fold” read “4-fold”;
135	2	10	for “consisting” read “consistent”;
135	2	11	for “this two” read “these two”;
136	1	10	for “where” read “were”;
136	2	2	for “bind” read “binding”;
136	2	15	for “preciously ” read “previously”;
138	1	5	discard “tissues”;
140	2	4	for “expressed cells ” read “expressed in cells”;

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsquellen und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Bruno Pereira Berlin, 2019