SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE

XXXI Ciclo
(A.A. 2017/2018)

Characterisation of lnc-G4, a long noncoding RNA that regulates skeletal muscle differentiation through translational repression of G-quadruplex containing mRNAs

Dottorando
Davide Mariani

Docente guida
Prof. Irene Bozzoni

Tutore
Dr Gianluca Canettieri

Coordinatore
Prof. Fulvio Cruciani

Cover image: differentiated C2C12 mouse myoblasts after immunostaining for Myosin Heavy Chain (red); cell nuclei are stained with DAPI (blue).

INDEX

SUMMARY

Long non-coding RNAs are crucial regulators of the fine tuning of gene expression. Their role has been widely studied especially in developmental processes such as skeletal muscle differentiation.
In particular, a novel cytoplasmic long non-coding RNA, called lnc-G4, has a relevant role in promoting murine C2C12 myoblast differentiation. The analysis of the interactors of this long non-coding RNA showed its ability to base-pair with many mRNAs thanks to a repeated element embedded in its sequence; among the interactors, we focused on MLX mRNA, which encodes for a myogenic transcription factor. We demonstrated that lnc-G4 directly interacts with the three splicing isoforms of MLX mRNA, while it is able to specifically inhibit the translation of only MLX γ isoform; this translational regulation could depend on the recruitment of the RNA helicase DHX36. Interestingly, the effect of lnc-G4 on MLX γ regulates the subcellular localization of the other isoforms, and this has an impact on the transcriptional activation of MLX targets.
Taken together, these evidences suggest that lnc-G4 could be a key factor in post-transcriptional gene regulation during the early phases of myogenesis through the translational regulation of MLX γ. Moreover, lnc-G4 interacts with other mRNAs, and the regulation mechanism could be extended to many other targets.

1. INTRODUCTION

1.1. A brief history of non-coding RNA

In every representation of the central dogma of molecular biology (first stated by Francis Crick in 1958), RNA lies at the heart of the genetic information flow; its central role came to light in 1961, when Jacob and Monod established for the first time the concept of messenger RNA (mRNA) as the connection between genetic information and protein synthesis (Jacob & Monod 1961). Later in the '60s, a population of *heterogeneous nuclear RNA* (hnRNA) was identified in higher eukaryotes, leading to the formulation of extensive RNA-dependent regulatory networks that could also be at the bases of complex organisms' evolution (Britten & Davidson 1969).
Despite these early discoveries and hypotheses, a protein-centric view of molecular biology persisted for the following decades, in which the regulation of gene expression was ascribed to the exclusive action of transcription factors. However, as it was written by Thomas Cech and Joan Steitz:

> "The seeds of a revolution are invariably sown decades before it erupts. And so it is with the revolution in noncoding RNAs."

In fact, between 1970 and 2000, the way to think about RNA biology was overthrown, and the established rules were step by step replaced by new rules; some examples are here reported:

1. The description of the first protein-independent RNA catalysis (Kruger et al. 1982) opened the ribozyme field, counteracting the idea that only proteins can have an enzymatic activity; ribosomal RNA itself was shown to be responsible for the catalysis of peptide bond formation.

2. The discovery of introns and of the splicing machinery (in which small nuclear RNAs are major components) showed that RNA transcription is not a straight-forward process, and later the evidences of alternative splicing demonstrated that exon rearrangement is responsible for the complexity of higher eukaryotes transcriptome.

3. The well-established model of proteins as unique regulators of transcription and translation was unhinged by the characterization of the first microRNAs (Lee et al. 1993) and the description of the RNA interference pathway.

Together, these milestones of RNA biology shed light on the pleiotropic and multivalent role of this molecule, that has even been indicated as the primordial constituent of life according to the "RNA world hypothesis" (Gilbert 1986).
Moreover, during the '90s it became clear that there was a correlation between the number and size of introns and intergenic sequences and the developmental complexity of organisms, and that these sequences could have evolved in order to express a wide spectrum of trans-acting regulatory RNAs. In fact, soon after the Human Genome Project conclusion, it was observed that the transcriptional activity of human chromosomes 21 and 22 was an order of magnitude higher than accounted for by the predicted and characterized exons (Kapranov et al. 2002).
The non-coding RNA revolution was completed by the advent of genome tiling arrays and the evolution of deep sequencing technologies that characterized the last decade. In particular, the Encyclopaedia of DNA Elements (ENCODE) project produced the first complete view of regions of transcription, binding of transcription factors and chromatin dynamics of the human genome, showing that 80.4% of it participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type (Dunham et al. 2012). However, it was also observed that 74.7% of the human genome is transcribed in at least one of the 15 cell lines analysed (Djebali et al. 2012), while the estimated 20.687 protein-

coding genes cover only the 2.94% of the sequence (ENCODE). With the increase of the number of publicly available sequencing data, pervasive transcription has been widely accepted as a common feature of eukaryotic genomes spanning from yeast to mammals.
These recent studies have enlighted the complexity of transcriptomes, where the so called "junk DNA" regions are not transcriptionally silent and give rise to a plethora of non-coding RNAs which could originate a repertoire of new functions.
Nevertheless, even if the existence of a huge amount of non-coding transcripts is incontrovertible nowadays, the question is whether such transcriptional activities serve any biological function.

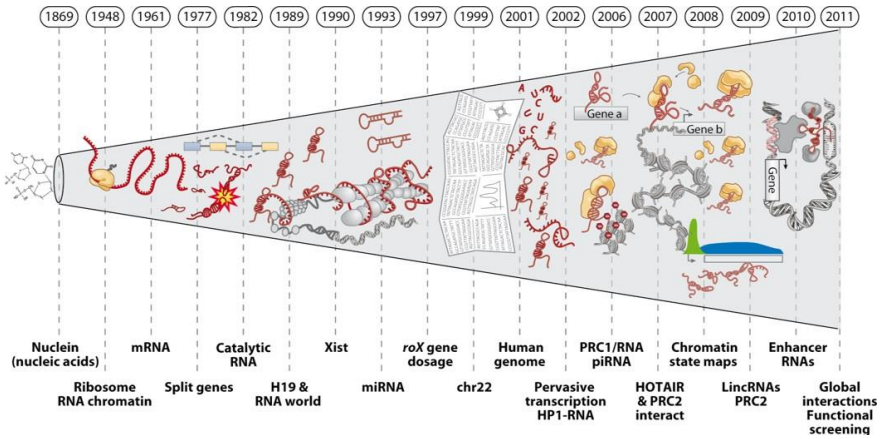


Figure A – Graphical representation of the most important milestones of molecular biology of RNA (from Rinn and Chang 2012)

1.2. Long non coding RNAs: definition and features

Long non-coding RNAs (lncRNAs) are a class of transcripts that mirrors the length and structure of mRNAs, including 5' cap, 3' polyA tail and splicing dynamics, though they do not have a clear coding potential. The operative definition of lncRNAs describes them as RNA molecules larger than 200 nt, in order to clearly distinguish these transcripts from small regulatory RNAs (Rinn & Chang 2012).

Before the advent of deep sequencing, a few dozen lncRNAs that play a role in development, imprinting (e.g. H19) and X chromosome inactivation (e.g. Xist) had been identified by biochemical approaches and genetic screenings.
Nowadays, RNA sequencing allows a comprehensive visualization of the transcriptome thanks to the reconstruction of each transcript isoform at single-nucleotide resolution (Garber et al. 2011). On the other side, the analysis of epigenetic marks of Polymerase-II initiation (H3K4me3) and elongation (H3K36me3) by ChIP-seq approach consents the mapping of novel transcriptional units outside form known protein-coding loci (Guttman et al. 2009).
The combination of these high-throughput techniques led to the identification and annotation of more than 15000 lncRNAs encoded in the human genome, and at least 12000 in the mouse genome (Uszczynska-Ratajczak et al. 2018).
Despite the structural similarities between them and mRNAs, lncRNAs possess unique features regarding their genomic organization and their pattern of expression. First of all, lncRNAs can be classified based on their localization and orientation within the genome in five different categories:

1. **Sense lncRNAs**, which originate from the same strand of protein coding genes, and may overlap them partially or completely;
2. **Antisense lncRNAs**, which are transcribed from the opposite strand of a protein-coding gene and overlap at least one coding exon. Antisense transcription regards as many as the 87% of coding transcripts in the mouse genome (Ma et al. 2013)
3. **Divergent lncRNAs**, that share their promoter with protein-coding genes and are transcribed in opposite direction;
4. **Intronic lncRNAs**, which are entirely embedded in introns of protein-coding genes without any preference in directionality;
5. **Intergenic lncRNAs**, also called lincRNAs, that are encoded by totally independent transcriptional units,

and are usually 5 kb away from other genes (Guttman et al. 2009)

As a main trend, the expression level of lncRNAs is lower than the one of mRNAs. Several studies have estimated the abundance of the two species starting from RNA-seq data, quantifying differences of the median expression level spanning from 3-fold (Guttman et al. 2010) to 10-fold (Guttman et al. 2009).
Moreover, the vast majority of lncRNAs exhibit expression patterns that are limited to one or a few tissues, independently from their expression ranges. In particular, 78% of human lincRNAs are tissue-specific, relative to only ~19% of coding genes (Cabili et al. 2011)
Tissue specificity of lncRNAs goes hand in hand with their role as molecular signals, since their transcription occurs at a very specific time and place to integrate developmental cues and interpret cellular context (Wang & Chang 2011).

The reconstruction of full-length gene structures of lncRNAs allows to study their evolutionary sequence conservation. Excluding the Telomeric repeat-containing RNA Terra, which is strongly conserved from yeast to mammals, mammalian lncRNAs lack any known orthologue in species outside the vertebrates (Ulitsky & Bartel 2013). Generally, lncRNA sequences appear to be more conserved than mRNA introns, while they are less conserved than coding sequences (Rinn & Chang 2012). The low sequence conservation is due to a high evolutionary rate: for example, within rodents only 60% of the lncRNAs (compared to >90% of mRNAs) expressed in *Mus musculus* liver have alignable counterparts expressed in the livers of *Mus castaneus* and rat (Kutter et al. 2012). Despite undetectable sequence conservation, genomic location and intron-exon structure are more prone to be maintained in different organisms, as well as the secondary structure of the lncRNA (Ulitsky & Bartel 2013). Finally, the fourth dimension of lncRNA conservation is at the level of the function, so that different molecules in different organisms can be responsible for the same phenotypic trait or molecular ac

A single lncRNA is able to interact with other nucleic acids through base-pairing, and to be recognised by protein interactors thanks to secondary structures and specific binding sequences; this is mainly due to the modular structure of these molecules, that allows them to be the central core of interaction networks between DNA, RNA and protein effectors (Guttman & Rinn 2012).
Summarising current evidences, the functionality of lncRNAs depends on three main mechanistic themes:

1. Decoy activity: lncRNAs can sequester other factors, either proteins or RNAs, and thus compete for their binding to other interactors. For example the lncRNA Gas5, which is induced upon starvation, is able to bind the glucocorticoid receptor through a hairpin sequence motif and thus competing for binding and activation of metabolic genes (Kino et al. 2010).
2. Scaffolding activity: lncRNAs can act as adaptors to bring together two or more proteins into discrete complexes (Spitale and Tsai 2011). A well described mechanism of scaffolding RNA is TERC, the RNA component of the telomerase complex, which is responsible for template synthesis, binding and catalytic activity of the protein component TERC (Collins 2008). In fact, mutations that affect the RNA scaffold structure result in impaired telomerase activity and in the onset of dyskeratosis congenita (Chen & Greider 2004).
3. Guide activity: many lncRNAs are required for the localization of protein complexes, and to target them on specific DNA or RNA sequences. For example, lincRNA-p21 is induced in a p53-dependent manner after DNA damage, and is able to binding the nuclear factor hnRNP-K and mediate its recruitment on the promoters of genes that have to be transcriptionally repressed (Huarte et al. 2010).

These general mechanisms of action of long noncoding RNAs allow them to act both as transcriptional or post-transcriptional regulation, and this distinction is mainly depending on their preferential localization within the nucleus or in the cytoplasm.

1.3. Nuclear long non-coding RNAs

A significant fraction of lncRNAs is preferentially localised in the nucleus, and in several cases they have been shown to be crucial regulators of transcription and RNA processing, but also of chromatin architecture and organization of nuclear domain. The recruitment of lncRNAs on chromatin depends on protein-mediated interactions, RNA-RNA base pairing or RNA-DNA direct interface by formation of triple helixes (Rinn & Chang 2012).
The identification of their molecular mechanisms has been strongly promoted by newly developed techniques, like CHIRP (Chu et al. 2012) or RAP (Engreitz et al. 2015), that allow the capture of specific RNAs and the detection of their chromatin association sites in a genome-wide manner. lncRNAs can act on chromatin organization by recruiting epigenetic remodelers: Oct4P4 lncRNA interacts with the SUV39H1 to direct the silencing of the Oct4 gene through deposition of H3K9me3 and HP1α on its promoter, leading to reduced mESC self-renewal (Scarola et al. 2015).
The orchestration of chromatin remodelling is not the only way by which lncRNAs affect transcription. For example, the lncRNA PACER positively regulates the expression of COX2 by sequestering the p50 repressive subunit of NF-kB away from its promoter (Krawczyk & Emerson 2014).
Moreover, long noncoding RNAs are major organisers of nuclear structure. The best characterised example is Xist, that is not only responsible for X-chromosome inactivation by recruiting the PRC2 repressive complex, but is also interacts directly with the Lamin B receptor in order to relocalize the entire inactivated X chromosome next to the nuclear lamina (Chen et al. 2016).
Finally, nuclear lncRNAs are also necessary to nucleate and maintain specific nuclear domains. NEAT1 RNA transcription is essential for the structure of paraspeckles (nuclear bodies in which specific Alu- containing mRNAs are retained) (Mao et al. 2011), while MALAT1 has been shown to localize serine/arginine (SR) splicing factors to a compartment called nuclear speckles, where they can be stored and modified by phosphorylation (Bernard et al. 2010). These few examples of functional nuclear-retained lncRNAs

show how their ability to interact with both protein and nucleic acid interaction is at the base of several different mechanisms that allow to regulate transcription at different levels.

1.4. Cytoplasmic long non-coding RNAs

Perhaps the most common misperception of lncRNAs is that they are mainly localized in the nucleus. Studied based on subcellular RNA-seq are often biased since the distribution is quantified relative to mRNA abundance. Keeping in mind that some lncRNAs might act in the nucleus before making their way to the cytoplasm, the current picture is that many lncRNAs spend most of their time in the cytoplasm, which is frequently their site of action (Ulitsky & Bartel 2013). In the cytoplasm, lncRNAs take part to the multiple layers of post-transcriptional regulation by affecting the half-life and the translation of messenger RNAs; in this way, lncRNAs can quickly regulate gene expression by directly intervening on already transcribed mRNAs, a feature that assumes great importance in development and differentiation.
Starting from mRNA stability modulation, there are several examples of cytoplasmic lncRNAs that are able to induce specific mRNA decay. One of the pathways on which they act is STAU1-mediated decay (SMD), that involves the degradation of translationally active mRNAs whose 3' UTRs contain stem structures bound by STAU1. However, not all SMD targets that possess comparable structures in their 3'UTR: in some cases, a STAU1 binding site can be formed by Alu sequence-mediated base-pairing between so called half-STAU1-binding site lncRNAs (1/2 sbsRNAs) and their target mRNAs, committing them to degradation (Gong & Maquat 2011). Another example of cytoplasmic lncRNA-mediated decay is represented by the DNA-damage induced RNA Gadd7; this transcript associates to TDP43 protein, interfering with its interaction with the cyclin-dependent kinase 6 (Cdk6) mRNA, thus leading to Cdk6 mRNA instability and G1 phase arrest of the cell cycle (Liu et al. 2012).
On the other side, cytoplasmic lncRNAs can have a role as mRNA stabilizer, in both protein dependent and independent manner.

In the former case, a good example is represented by TINCR, a lncRNA involved in epidermal differentiation which is able to interact with several mRNAs of genes associated with epidermal barrier formation, thanks to a unique, 25-nt sequence motif called TINCR-box (Kretz et al. 2013). Notably, TINCR-mediated stabilization relies on the direct binding of the already cited STAU1 protein, that is this case has a positive effect on mRNA stability.
In other cases, the base-pairing of the RNA alone is sufficient to trigger the regulation mechanism. For example, the conserved noncoding BACE1-AS antisense transcript stabilises BACE1 mRNA by forming a sense-antisense duplex, as demonstrated by RNase protection assay (Faghihi et al. 2008); the binding of the lncRNA has been shown to mask the miR-485-5p binding site on BACE1 mRNA, preventing its miRNA-induced repression (Faghihi et al. 2010). In this case, the presence of the RNA alone is sufficient to trigger the regulation mechanism.

Another layer of regulation in which lncRNAs seem to have a major role in translational control. A paradigmatic example is lincRNA-p21, already cited for its nuclear role; in the cytoplasm, this transcript is associated with CTNNB1 and JUNB mRNAs by imperfect base-pairing in their 3'UTR regions. The formation of lincRNA-p21–mRNA complex recruits the translational repressors RCK as well as Fragile X mental retardation protein (FMRP), thus reducing ribosome occupancy on the mRNAs (Yoon et al. 2012). Translational repression mediated by lncRNAs has also a validated role in cancer progression: treRNA stimulates tumour invasion *in vitro* and metastasis formation *in vivo* by downregulating the translation of the epithelial marker E-cadherin (Gumireddy et al. 2013). Even in this case, the scaffolding activity of the lncRNA recruits protein regulators, such as hnRNP K, FXR1 and 2, to form a novel ribonucleoprotein complex required to exert its function.
An example of positive regulation of translation is represented by Uchl1-AS, that is an antisense transcript to mouse ubiquitin carboxy-terminal hydrolase L1 (Uchl1), sharing a 5′ overlapping sequence that contains an embedded inverted SINEB2 element with its sense transcript (Carrieri et al. 2012). Under stress signalling

control, Uchl1-AS migrates to the cytoplasm, interacts with the mRNA via the SINEB2 sequence, and recruits the protein coding transcript to the heavy polysomes increasing its translation.

In the crowded environment of the cytoplasm, not only RNA-protein, but also RNA-RNA crosstalk between different RNA species becomes an important layer of regulation. In particular, a huge variety of transcript contain miRNA binding sites; competition for the interaction with shared microRNAs is then a possibility for a new layer of regulation (Tay et al. 2014). In fact, the expression of lncRNAs containing miRNA binding sites in specific tissues and windows of time can impact the stoichiometry of microRNA-mRNA interaction, as demonstrated by several studies.
PTENP1, a noncoding transcript expressed from a pseudogene, regulates the tumour suppressor PTEN expression both at transcriptional level, by recruiting DNMT3 and EZH2 on its promoter, and post-transcriptional level, acting as a decoy for PTEN-targeting miRNAs (Johnsson et al. 2013). Moreover, sequestration of miRNAs has been shown to be important for embryonic stem cell maintenance: the lncRNA linc-RoR is an effective sponge for miR-145, whose targets OCT4, SOX2 and NANOG are core regulators of pluripotency (Wang et al. 2013).
Another peculiar example of competing endogenous RNA is represented by ciRS-7, a circular RNA that contains 73 selectively conserved target sites for miR-7 in which a mismatch in the central part of the duplex prevents miRNA-mediated endocleavage (Hansen et al. 2013).
Finally, in the recent years several lncRNAs have been identified as regulators of post-translational modification of cytoplasmic proteins. Just to cite an example, lincRNA-p21 has an additional role in hypoxic condition response, binding both the transcription factor HIF-1α and the ubiquitin ligase VHL, preventing HIF-1α ubiquitination and degradation and promoting glycolysis under hypoxia (Yang et al. 2014).
Taken together, all the cited examples show how cytoplasmic lncRNAs are core components of a complex network of RNA-protein and RNA-RNA interactions that consent them to exert

important functions in the post-transcriptional regulation of gene expression, especially during stimuli response and developmental programs.

1.5. Not all long non-coding RNAs are strictly non-coding

As it has been reported above, pervasive translation of complex eukaryotes genomes gives rise to thousands of non-coding transcripts, beside the already known protein-coding gene products. However, canonical open reading frames (ORFs) have historically been identified as sequencing coding for peptides longer than 100 codons, based on the assumption that short peptides are unable to fold into secondary structure and thus to perform biological functions. Recently, the introduction of Ribosome Profiling, a powerful high-throughput technique based on deep sequencing of ribosome protected fragments after RNase footprinting, have permitted the quantitative study of translation by estimating the ribosomal engagement of transcripts (Ingolia et al. 2009). Several Ribosome Profiling works have showed that ribosomes occupy many *bona fide* non-coding regions of transcriptomes, including lncRNAs, thus enlarging the complexity of the mammalian translatome. These data have been confirmed by the analysis of the distribution of thousands of lncRNAs in polysome profiling after sucrose fractionation (Van Heesch et al. 2014). Taken together, these studies suggest that many lncRNAs could be translated in order to produce small, previously unidentified peptides, despite of their predicted low coding potential. There are in fact examples of transcripts, formerly identified as long non-coding RNAs, which are in fact coding for short functional peptides (Rion & Rüegg 2017). The first to be identified was the Drosophila tarsal-less/polished rice transcript, which encodes peptides of 11 to 32 amino acids whit a role in post-translational regulation of the Shavenbaby transcription factor (Kondo et al. 2010). More recently, other short regulatory peptide have been identified in mammals; for instance, Myoregulin and DWORF are ultraconserved micropeptides that fold into transmembrane α-helixes and interact with the SERCA calcium

channel, regulating calcium uptake in skeletal muscle (Anderson et al. 2015; Nelson et al. 2016). Nevertheless, these examples could be interesting exceptions, while it is still unclear at what extent lncRNAs could be translation templates. In fact, ORFs in lncRNAs might act as upstream ORFs (uORFs) to prevent ribosome scanning or translation in downstream regions of the transcripts, thereby enabling the lncRNAs to perform noncoding functions in the cytoplasm without interference from the ribosome. lncRNA ORFs might also tether factors to ribosomes or modulate the stability of the lncRNA by influencing RNA decay pathways, some of which depend on translation, such as non-sense mediated decay (Ulitsky & Bartel 2013).

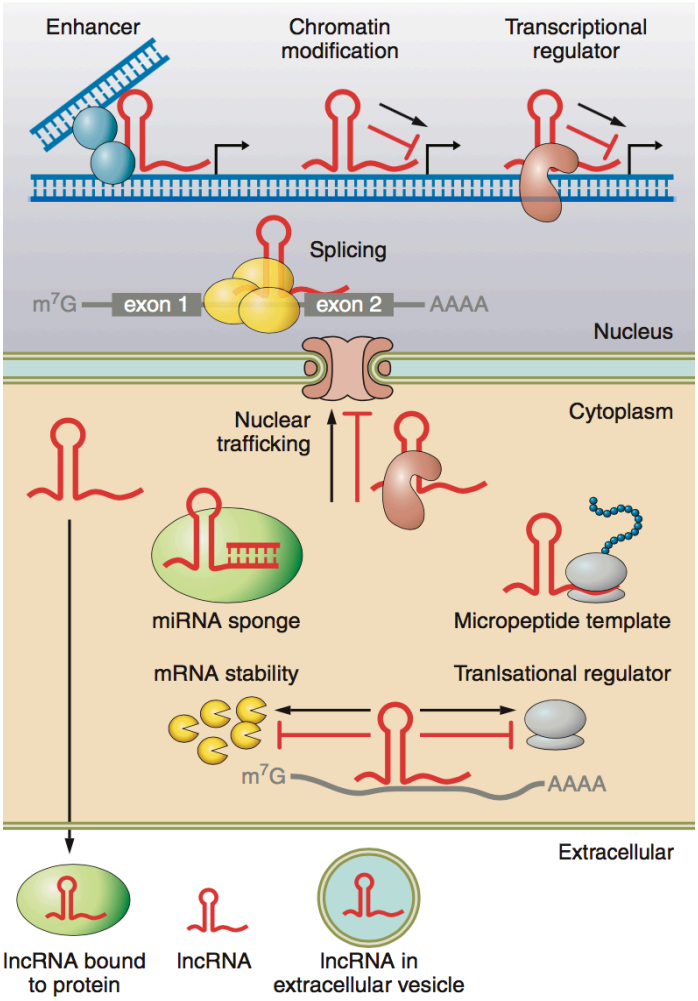


Figure B – Graphical summary of different mechanisms of action of lncRNAs in diverse cellular compartments; together with the examples that were already discussed and presented in the previous paragraphs, the picture shows that lncRNAs can also be embedded in extracellular vesicles or circulate freely in biological fluids, with a speculative role on cell-to-cell communication (adapted from Beerman et al, 2016)

1.6. Molecular regulation of skeletal muscle differentiation

In mammalian embryos, the formation of skeletal muscle of trunk and limbs starts from the somites, which are derivatives of the antero-posterior segmentation of paraxial mesoderm. In particular, somitic cells subsequently distribute into a ventral mesenchymal sclerotome, the precursor of vertebral column and ribs, and a dorsal dermomyotome, that gives rise to both dorsal dermis and all the skeletal muscles of limbs and trunk. Head muscles formation follows a different path, taking origin from the mesodermal core of branchial arches (Buckingham and Rigby 2014). However, myogenesis is not limited to embryonic development and occurs also in adult vertebrate as a response to muscle injury. Muscle regeneration is the formation of new muscle fibers on the template of the extracellular matrix following injury; this function is exerted by satellite cells, somite-derived myogenic progenitors located under the basal lamina that surrounds the multinucleated fibre. Once they are activated by injury signalling, satellite cells leave their niche and undergo asymmetric divisions to form new fibers and maintain a stem cell pool in the same time.

Transcriptional control is fundamental in both embryonic and adult skeletal muscle differentiation, and it mainly depends on a core unit of four Myogenic Regulatory Factors (MRFs) that belong to MyoD family (Buckingham and Rigby 2014). The role and molecular circuitries in which these basic helix-loop-helix (bHLH) factors are involved have been widely elucidated in the last 30 years: for instance, overexpression of MyoD in non-muscle cells leads to activation of myogenic program and suppression of other cell fates (Weintraub et al. 1991). Mutations of the MRFs in mouse models have shown that MyoD, Myf5 and Mrf4 act as essential myogenic determination factors, while Myogenin is a differentiation factor necessary to control the differentiation program (Moncaut et al. 2013).
The upstream controllers of MRFs activations are Pax3 and Pax7, members of the family of paired domain transcription factors. Pax3

expression begins in the presomitic paraxial mesoderm, and it identifies migrating myogenic progenitors that have not yet activated the determination genes (Buckingham & Relaix 2007).In somite multipotent cells, the balance between Pax3 and Foxc2 is necessary for self-renewal, but a relative increase in Pax3 level in the trigger for Myf5 activation and therefore for the beginning of myogenic program. On the other side, Pax7 is expresses in the central domain of the dermomyotome; while it is not necessary for pre-natal myogenesis, it becomes the dominant factor in adult progenitors. Pax7 is necessary for satellite cell quiescence, since it is an activator of the HLH inhibitor Id3 (Kumar et al. 2009), but also to trigger the initiation of differentiation by direct MyoD activation (Hu et al. 2008) and chromatin remodelling of Myf5 locus through the recruitment of Wdr5– Ash2L–MLL2 histone methyltransferase (HMT) complex (McKinnell et al. 2008).
The induction of skeletal muscle differentiation requires the downregulation of Pax3/Pax7 and the subsequent activation of MRF network. Briefly, Myf5 expressing cells, called myoblasts, undergo a proliferative phase under FGF and Notch signalling, until MyoD activation leads to cell cycle exit. MRF factors act as obligate heterodimers with E proteins, thus binding E-boxes containing promoters and enhancers of a wide plethora of genes controlling muscle structure, contraction and metabolism. MRF-dependent gene activation requires also a complete chromatin remodelling. For instance, MyoD interacts with the BAF60c subunit of the SWI/SNF chromatin-remodelling complex and recruits it on the promoters of inactive genes creating poised chromatin domains ready for rapid transcription in response to differentiation stimuli (Forcales et al. 2012). MRFs do not act in isolation but they work in concert with other transcription factors, like members of the Mef2 family (Molkentin et al. 1995). After the proliferation arrest and lineage-specific transcriptional activation, myoblast transition towards the myocyte state is characterised by a reduction in cell motility and the precise orientation and positioning of muscle cells in response to chemotactic gradients (Griffin et al. 2010). The formation of myotubes, multinucleated syncytia with contraction ability, requires extensive actin cytoskeleton rearrangement; moreover, myoblast

fusion requires the presence of the small, muscle-specific peptide Myomixer, that mediated plasma membrane fusion through the association with the fusogenic protein Myomaker (Bi et al. 2017).

1.7. Noncoding RNAs in skeletal muscle differentiation

Transcriptional regulation of the onset of skeletal muscle formation is modified by post-transcriptional mechanisms that affect the presence and function of the transcription factors concerned (Buckingham and Rigby 2014). This fine regulation of differentiation timing mainly depends on a complex network of long and small non-coding RNAs. First of all, myogenic microRNAs are fundamental to repress self-renewal factors and to prevent early expression of late myogenic markers. For example, Pax3/Pax7 are targets of miR-206, and their downregulation contributes to the induction of differentiation; miR-1 maintains the repression during the late phases of the developmental program (Horak et al. 2016). Moreover, miR-31 is required for translational repression of late myogenic markers such as dystrophin (Cacchiarelli et al. 2011).
Together with the cited microRNA circuitries, also lncRNAs work as crucial regulators of skeletal muscle differentiation thanks to their tissue-specific expression and their fine temporal regulation, acting at different layers of gene expression both in the nucleus and the cytoplasm. In some cases, the way of action of myogenesis-involved lncRNAs is strictly connected to microRNAs; H19, the first long noncoding RNA to be identified, harbours both canonical and non-canonical binding sites for microRNAs of the let-7 family, and its depletion causes precocious muscle differentiation (Kallen et al. 2013). Furthermore, linc-MD1 is a muscle-specific, MyoD-induced competing endogenous RNA that governs the timing of myogenesis. linc-MD1originates from the miR-206/133b locus and acts as a decoy for miR-133 and miR-135, derepressing their mRNA targets, including MAML1 and MEF2C, which encode crucial myogenic factors required for the activation of muscle- specific genes (Cesana et al. 2011). Moreover, the interplay between linc-MD1 and the protein HuR in the early phases of differentiation prevents the Drosha cleavage of the lncRNA, allowing its sponging activity;

when this feedforward regulatory loop is interrupted from external stimuli, linc-MD1 is processed for miR-133b production triggering the progression to late differentiation stages (Legnini et al. 2014).
Recently, a study based on RNA-seq characterized 30 novel lncRNA species, including previously unannotated transcripts, which undergo modulated expression during *in vitro* differentiation of C2C12 murine myoblasts (Ballarino et al. 2015). These newly identified transcripts have been classified based on their expression dynamics in growth condition (GM) versus differentiating condition (DM), their subcellular localization between nucleus and cytoplasm and their expression in tissues withdrawn from 2-month-old wild type and *mdx* mice (*mdx* mouse is the principal animal model for Duchenne Muscular Dystrophy, carrying a point mutation in exon 23 of DMD gene; it is characterised by increased muscle damage and regeneration). Some of this new lncRNAs have been further characterised in their molecular mechanism of action in the last years. For instance, Charme is a nuclear lncRNA that takes contact with the *nctc* locus, which contains Igf2 and the troponins Tnnt3 and Tnni2 genes, and is essential for its organization; the depletion of the lncRNA causes the disassembly of the chromatin domain and the downregulation of the genes contained therein.
Moreover, Charme$^{-/-}$ mouse shows severe cardiac remodelling and has shorter lifespan (Ballarino et al. 2018).
Another case is represented by lnc-31, a cytoplasmic lncRNA that is mainly expressed in growing condition and is required to sustain myoblast proliferation. lnc-31 interacts with both Rock1 mRNA and YB-1 protein, and increases Rock1 translation thanks to a localized effect of YB-1 stabilization and to the counteraction of YB-1 proteasome degradation (Dimartino et al. 2018).

1.8. G-quadruplexes in RNA biology

The identification of nucleic acid structures with a high guanine abundance became apparent in the early ’60s, and later in the ‘80s it was discovered that these formations were at the base of telomere formation (Henderson et al. 1987). Nowadays, structural analyses

have given the G-quadruplex name to these peculiar DNA and RNA assemblies.
G-quadruplexes formed by guanine-rich nucleic acid sequences are non-canonical structures organized by stacking of tetrads or G-quartets, in which four guanines are assembled in a planar arrangement by Hoogsteen hydrogen bonding. In order to fold into a G-quadruplex structure, the linear sequence of DNA or RNA must contain four guanine stretches of 2-5 G, intercut with variable loops from 1 to 7 nt. The stability of this structure depends on the presence of a monovalent cation, typically potassium, that "coordinates" the structure by positioning within or in between the plane of the tetrads (Millevoi et al. 2012).
The human genome contains as many as 376,000 G-quadruplex-forming sequences which are not randomly distributed (Todd et al. 2005). At the DNA level, quadruplexes are enriched at the telomere, where they inhibit the activity of telomerase, and they are associated with ~40% of human promoter, supporting their role in transcriptional regulation. Nevertheless, an increasing number of evidences in the last few years support the view that RNA G-quadruplexes are also crucial, intrinsic biological regulators. Additionally, RNA G-quadruplexes have been demonstrated to be more stable than the DNA counterpart, with faster association dynamics and slower dissociation (Saccà et al. 2005). Increasing evidences show that RNA quadruplexes at the 5' and 3' ends and/or other locations in the RNAs have functional role in pre-mRNA processing (including splicing and polyadenylation), RNA stability and turnover, mRNA translation and subcellular targeting.
Starting from the regulation of mRNA splicing, G-quadruplexes placed in the vicinity of splicing sites of a growing list of genes, can act as cis- regulatory elements of alternative exon inclusion or exclusion. For instance, a G-quadruplex located in intron 3 of the TP53 gene favours splicing of the adjacent intron 2; mutation of the G-quadruplex leads to intron 2 retention (Marcel et al. 2011).
More interestingly, computational analysis revealed that G-quadruplexes forming motifs are overrepresented in the 5′-UTRs, which are key elements in the initiation of mRNA translation. The presence of these secondary structures in the 5′-UTR, can interfere

with translation either by cis-acting effects (steric blockade) or through the binding of trans-acting mRNA binding proteins that may impact on the recruitment of eIFs to the mRNAs (e.g., eIF4F to the cap) or also affect the scanning process or the recognition of the initiation codon. Nonetheless, G-quadruplexes located in 5′-UTR are able to activate translation: G-quadruplex-forming motifs are present in IRES elements, and a mutational analysis has shown that the G-quadruplex within the VEGF IRES is essential for IRES function (Morris et al. 2010).
Even if there is a positional bias toward the presence of G-quadruplexes in the 5′-UTR of mRNAs, these structural elements can also be found in 3′-UTR and open reading frames (ORF).
The involvement of G-quadruplex-binding proteins is particularly evident for mRNA translational regulation, where G-quadruplex structures needs to be unfolded to relieve translational repression. In particular, the RNA helicase DHX36, also named RHAU, is a crucial actor for its global G-quadruplex unwinding activity and its extreme specificity in the recognition of these structures (Lattmann et al. 2011). This helicase has an important role in mRNA G-quadruplex resolution, but also in AU-elements mediated decay for the regulation of mRNA turnover. A good example of this mechanism is represented by the regulation of Nkx2.5 cardiac-specific mRNA: in this case, DHX36 is able to promote translation of Nkx2.5 through the resolution of a G-quadruplex in the 5’ UTR on the mRNA, while on the other side it induces the decay of the transcript by binding a AU-rich element in the 3’ UTR (Nie et al. 2015).

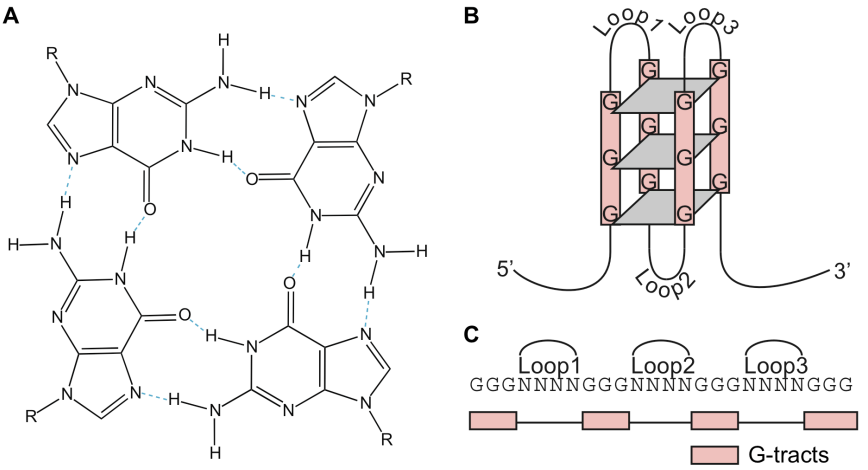


Figure C – A) Structure of a G-quartet. The planar ring of four hydrogen-bonded guanines is formed by guanines from different G-tracts, which are separated by intervening loop regions in the intra-molecular G4 DNA structure.
B) Schematic of an intra-molecular G4 DNA structure consisting of three G-quartets. Inter-molecular G4 DNA structures can also form from two or four strands.
C) The G4 DNA motif sequence used in this study with four G-tracts of three guanines separated by loop regions.

2. AIM OF THE THESIS

In 1988, Francis Crick wrote:

> "Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself."

Crick has always been a pioneer, or a sort of "prophet" of molecular biology future directions. In particular, he always had a complete and wide vision of organisms as complex systems that are finely regulated at the molecular level.
The technological advances made in the last years revolutionized the protein-centric paradigm of gene expression regulation. Nowadays, the eukaryotic genome is considered as a "RNA machine" in which different species of non-coding RNAs cooperate to modulate gene expression in a very fine way.
In this landscape, long non-coding RNAs have been identified and described as essential regulators of developmental and differentiation processes. Thanks to their versatility, modularity and cell- and tissue-specific expression, these molecules have been selected during evolution to exert a wide variety of functions and increase the complexity of gene expression modulation.
Long non-coding RNAs are core regulators of every step of mRNA processing, from transcription to translation and turnover, by acting both in nuclear and cytoplasmic compartments. Even if a strong effort has been made to reveal and enlighten the ways of action of nuclear lncRNAs, the molecular mechanisms by which these molecules influence mRNA life in the cytoplasm are still poorly understood, with few complete examples.
The aim of my thesis is to add knowledge on the role of cytoplasmic long non-coding RNAs in post-translational regulation during skeletal muscle differentiation.
Since many years, the research group of Prof. Irene Bozzoni, in which I prepared my thesis, is interested in understanding the mechanisms of regulation of skeletal muscle differentiation in both

physiological and pathological conditions, particularly focusing on the involvement of non-coding RNA species.
In particular, during my thesis I worked on the characterization of long non-coding RNAs differentially expressed during skeletal muscle differentiation of C2C12 murine myoblasts, previously identified by the group of Dr. Monica Ballarino (Ballarino et al. 2015). Among the cytoplasmic candidates, we selected lnc-G4 for further investigation due to its muscle-specific expression, its abundance and the interesting phenotype on muscle differentiation after its knockdown. Using a RNA pulldown approach coupled with high-throughput techniques, I identified and validated the protein and RNA interactors of this long non-coding RNA.
My work allowed me to describe a molecular mechanism in which lnc-G4 physically interacts with the γ isoform of MLX mRNA in a G-quadruplex containing region. This interaction is sufficient to repress MLX γ translation, and to thus to modulate the subcellular localization of the total level of MLX protein.
Moreover, I showed that the C-rich region embedded in lnc-G4 could be able to interact with many other G-quadruplex containing mRNAs, thus increasing the number of putative target of lnc-G4 activity during skeletal muscle differentiation.

3. RESULTS

3.1. Criteria for lnc-G4 selection

As previously described, we started our work with a candidate selection from the lncRNAs identified from Ballarino et al. 2015, according to the following criteria:

- Prevalent cytoplasmic localization, because we were interested on the characterization of novel cytoplasmic lncRNA-mediated mechanisms of gene expression regulation;
- Expression level: we focused on quite abundant molecules, selecting the lncRNAs with a relevant expression (quantified as a FPKM value between 7 and 50) in at least one time point of C2C12 differentiation;
- Tissue specificity: we selected lncRNAs that are upregulated or exclusively expressed in skeletal muscle tissues of *wt* and *mdx* mice.

According to these criteria, 6 out of 30 lncRNAs were chosen for further characterization.
The identified candidates were then subjected to a siRNA screening in order to select molecules with an interesting phenotype on myoblast proliferation and/or differentiation *in vitro*.
The phenotypic screening part was performed by dr. Sama Shamloo, a former PhD student of the laboratory; the details of this preliminary work will not be presented and discussed in this manuscript.
My thesis project started form the best candidate identified in this screening, afterwards named lnc-G4.

3.2. Characterization of lnc-G4

Lnc-G4 is a multiexonic long non-coding RNA of 1409 nt which is transcribed from the negative strand of an intergenic locus of *Mus musculus* chromosome X (Fig. 1A). It can be classified as an intergenic long non-coding RNA (lincRNA) since it is expressed from a totally independent transcriptional unit, far more than 100 kb from the flanking genes (*Bcor* and *Atp6ap2* respectively).
The timing of expression of lnc-G4 was analysed during time-course experiments of C2C12 differentiation, from proliferating myoblasts (condition named GM, Growth Medium) to terminally differentiated myotubes (condition named DM5, after 5 days from switch to Differentiation Medium). While the transcript is almost not detectable in GM condition, lnc-G4 expression dramatically increases during the early phases of C2C12 differentiation, with a peak in DM2, and drops down during the late phases of the differentiation process (Fig. 1B). This expression pattern is coherent with the RNA sequencing data published by Ballarino et al. 2015, where lnc-G4 expression reaches the highest FPKM value (17,25) in the DM3 sample.
Interestingly, the expression pattern of lnc-G4 recapitulates the one of Myogenin, a master regulator of skeletal muscle differentiation. To evaluate a putative role of Myogenin as a transcriptional regulator of lnc-G4 expression, publicly available ChIP-seq datasets from the mouse ENCODE project have been analysed (Fig. 1C). In Myogenin ChIP-seq experiments performed in C2C12 myoblasts and differentiated myotubes, there is a strong recruitment of the transcription factor on the putative lnc-G4 promoter; the maximum peak is reached in the early phases of myogenesis (60 hours), while Myogenin is less enriched in terminally differentiated myotubes (7 days).

The subcellular localization of lnc-G4 was then determined by nuclear-cytoplasmic fractionation of C2C12 cells at day 2 of differentiation, when lnc-G4 is expressed at its maximum. Around 70% of the transcript is localized in the cytoplasm, suggesting a prevalent cytoplasmic function for the lncRNA (Fig. 2A); GAPDH

and its precursor pre-GAPDH were used as control of fractionation efficiency.

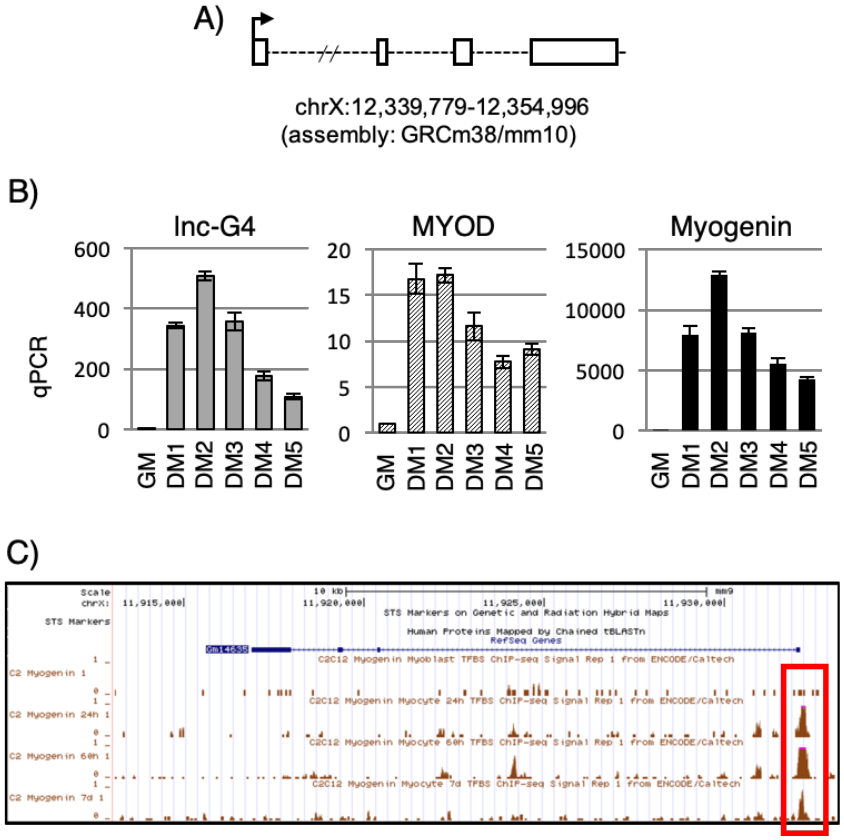
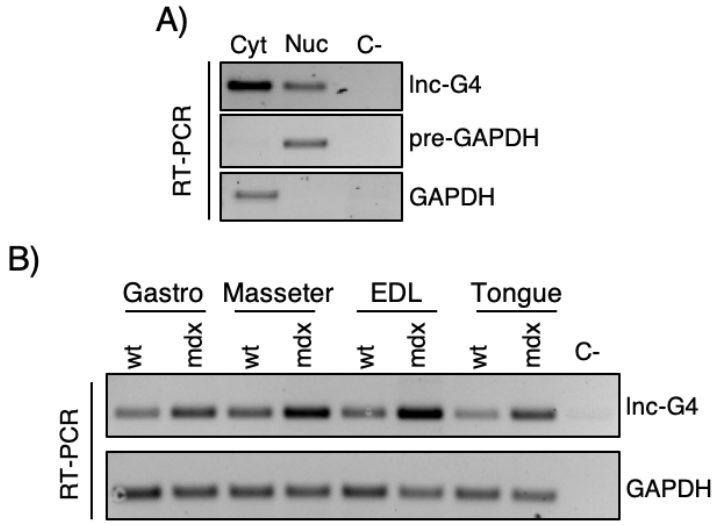


Figure 1 – A) Schematic representation of lnc-G4 locus genomic coordinates and structure. B) qPCR quantification of lnc-G4 expression during complete C2C12 differentiation, compared to the expression of the major myogenic markers MYOD and Myogenin. C) UCSC screenshot showing Myogenin binding site (highlighted in red) in the putative promoter of lnc-G4, identified in Myogenin ChIP-seq data in different time-points of C2C12 differentiation.

In order to validate the skeletal muscle specificity of lnc-G4, already analysed in Ballarino et al. 2015, we checked the expression of the lncRNA in different murine skeletal muscle samples. We selected

gastrocnemius, masseter, *extensor digitorum longus* (EDL) and tongue of both two-month-old wild type (*wt*) and *mdx* in order to analyse a sample set of muscles with different fibre composition coming from different body districts. As shown in Fig. 2B, lnc-G4 is expressed in all the analysed samples and it is upregulated in *mdx* dystrophic condition, in which muscle regeneration is strongly activated. Taken together, these data characterize lnc-G4 as a mostly cytoplasmic lncRNA with a strong skeletal muscle specificity, mainly expressed in the early phases of C2C12 differentiation.

Figure 2 – A) RT-PCR performed on RNA resulting from subcellular fractionation of C2C12 at day 2 of differentiation, showing that most of lnc-G4 transcript is



localized in the cytoplasm. Pre-GAPDH and GAPDH were used as controls of fractionation efficiency. B) RT-PCR analysis of lnc-G4 expression in skeletal muscle samples from two-month-old wt and mdx mice. Gastro=gastrocnemius, EDL=extensor digitorum longus. GAPDH was used as loading control.

3.3. lnc-G4 downregulation affects proper myogenesis

In order to identify the potential role of lnc-G4 in the establishment of proper muscle differentiation, a RNA interference loss-of-function approach has been chosen. Knockdown of lnc-G4 has been performed in C2C12 cells at day 2 of differentiation, using two different specific siRNAs, obtaining a consistent downregulation of ~80% of the lncRNA level as measured by qPCR (Fig. 3A).
The impact of lnc-G4 downregulation on myogenesis was evaluated through Myosin Heavy Chain (MHC) immunostaining on C2C12 cells treated with scramble and the two different specific siRNAs. As shown in Fig. 3B, lnc-G4 depletion had a clear phenotypic effect on C2C12 differentiation, with a lower expression of MHC protein marker and a reduced formation of mature, correctly oriented myotubes with respect to the scramble-treated condition.

To quantitatively analyse the effect of lnc-G4 downregulation on skeletal muscle differentiation, two different parameters were taken in account:

- Fusion index, defined as the fraction of the total number of fibre-embedded nuclei to the total number of nuclei; lnc-G4 depletion causes a reduction of the fusion index of 50%.
- MHC^+ mononucleated index, defined as the fraction of the mononucleated MHC^+ cells to the total of MHC^+ cells; lnc-G4 depletion causes a marked increase of cells that already express Myosin Heavy Chain without beginning the fusion process.

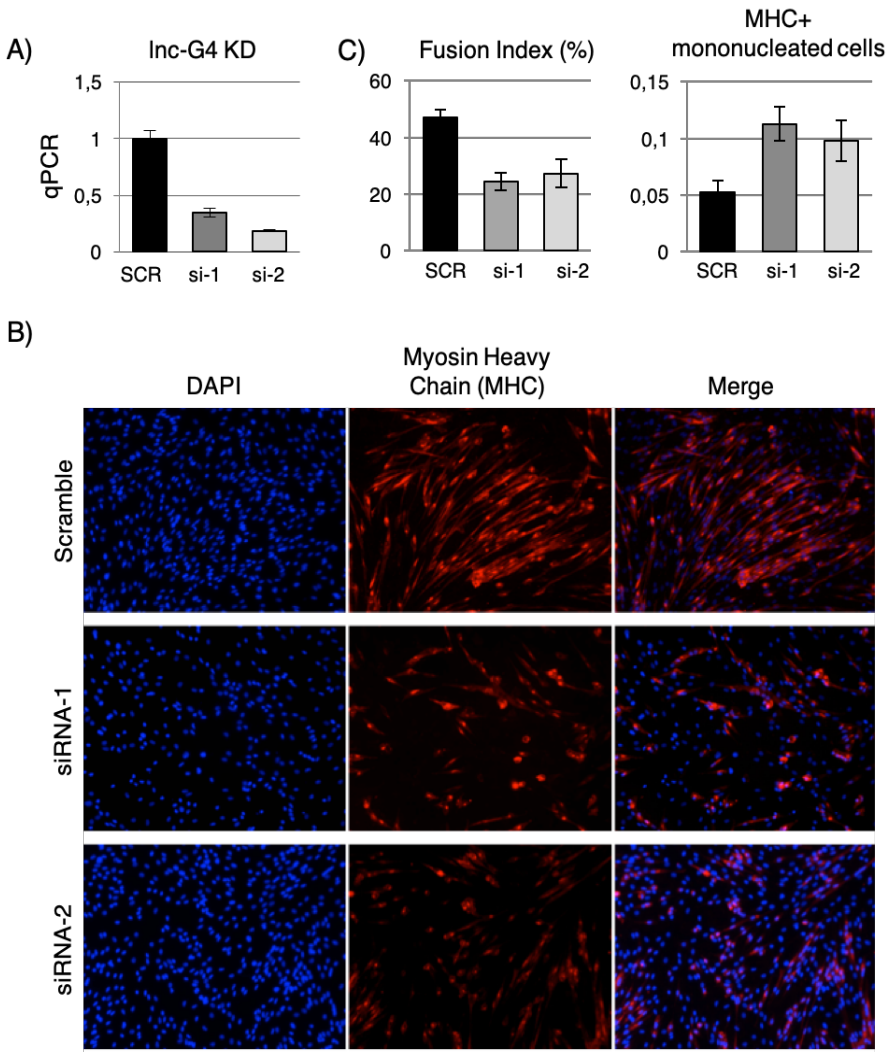


Figure 3 – A) qPCR quantification of lnc-G4 knockdown efficiency after specific siRNA transfection, analysed in C2C12 at day 2 of differentiation. lnc-G4 levels are normalised on GAPDH. B) Immunostaining for Myosin Heavy Chain on C2C12 at day 2 differentiation treated with scramble and two specific siRNAs against lnc-G4. DAPI was used for total nuclear count. C) Fusion index and MHC^{+}mononucleated cells measurement on MHC immunostaining.

To better understand and characterize the molecular pathways modulated after lnc-G4 knockdown, RNA isolated from C2C12 cells transfected with scramble or specific siRNAs was subjected to Illumina RNA-sequencing. The results were analysed by Dr. Alessio Colantoni (Sapienza University of Rome) with a pipeline focused on the identification of differential expression (details are reported in Material and Methods).
The bioinformatics analysis for showed that the expression of more than 50 genes was affected by lnc-G4 depletion. Gene Ontology analysis revealed the upregulation of genes involved in myoblast proliferation, chemotaxis and cytokine production and response, together with the downregulation of gene products involved in regulation of muscle contraction, muscle metabolism and calcium homeostasis, coherently with the observed phenotype (Fig. 4A). Moreover, a subpopulation of the most downregulated and upregulated genes was subjected to qPCR validation of the RNA-sequencing results. In particular, Figure 4B shows that lnc-G4 knockdown with both siRNAs do not affect early myogenic markers expression (Myogenin, MYOD), while it negatively affects the expression of late myogenic transcription factors (MEF2C), muscle-specific proteins (MCK, Dystrophin), and contraction markers (LRRN1, TNNC2, CRABP2). Furthermore, lnc-G4 downregulation positively impacts on the expression of a subset of cytokines (CXCL5, CCL2, CCL7) that have a validated role in myoblast proliferation and survival.

Taken together, these findings suggest that lnc-G4 strongly promotes skeletal muscle differentiation, and its depletion impacts on myocyte fusion ability and mature myotube formation. The phenotypic effect of lnc-G4 downregulation is recapitulated by the transcriptomic changes analysed by RNA-sequencing.

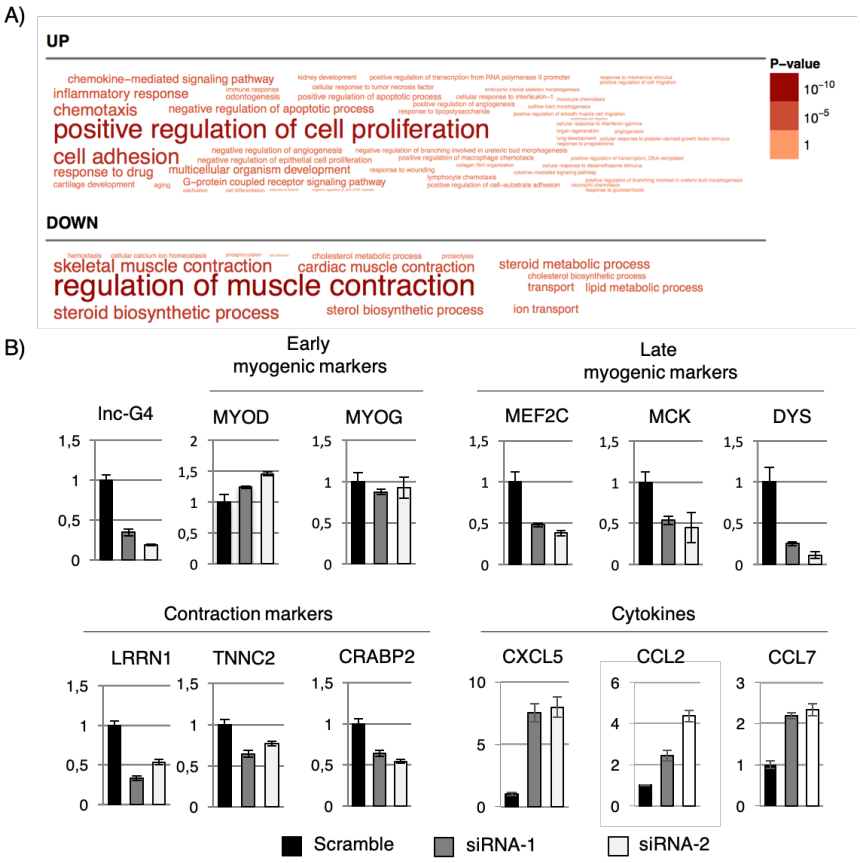


Figure 4 – A) Word-cloud output of the Gene Ontology term enrichment analysis; divided in upregulated and downregulated genes. B) qPCR validation of mRNAs whose expression in affected by lnc-G4 knockdown with two different siRNAs; targets are classified in early myogenic markers, late myogenic markers, skeletal muscle markers and myogenesis-relevant cytokines. Presented data are normalized on GAPDH as internal control.

3.4. lnc-G4 has a complex molecular interactome

As treated in the Introduction, cytoplasmic long non coding RNAs can influence gene expression thanks to many different mechanisms of post-transcriptional regulation. However, all these diverse ways of action rely on the ability of lncRNAs to interacts with other RNA species, especially mRNAs and microRNAs, and with a wide variety of RNA binding proteins.
The molecular interactome of lnc-G4 has been inspected with a novel endogenous RNA pulldown approach, coupled with high-throughput techniques in order to identify mRNA and protein interactors. Briefly, DNA antisense biotinylated probes designed against lnc-G4 sequence have been used to isolate the lncRNA from a cytoplasmic extract of C2C12 at day 2 of differentiation (Fig. 5A). Eight 20-nt probes have been designed on the complete sequence of lnc-G4, excluding repeated parts and regions which were predicted to be strongly structured in order to increase the accessibility; the selected probes were subsequently divided in two different sets of four probes each, and RNA pulldown was performed independently with both the sets (Fig. 5B). A third set of four probes, designed against *LacZ* mRNA sequence, was used as a negative control since it is not expressed in C2C12 cellular system. As shown in Figure 5C, the obtained pulldown efficiency for lnc-G4 was around 10% of the input with both the probe sets. The developed RNA pulldown approach allowed the isolation of both mRNA and protein putative interaction, and their subsequent identification through RNA sequencing and mass spectrometry, respectively.

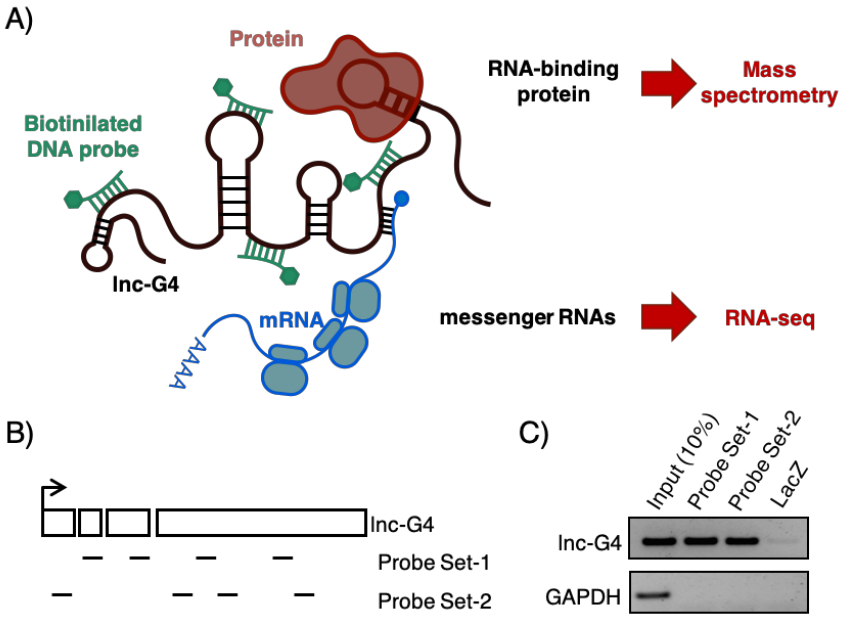


Figure 5 – A) Schematic representation of the RNA pulldown workflow. B) Schematic representation of the position of the DNA biotinylated probes used for lnc-G4 pulldown. C) RT-PCR analysis of lnc-G4 enrichment in the RNA pulldown performed with the two probe sets; GAPDH was used as negative control.

3.4.1. lnc-G4 co-precipitates with several mRNAs

RNA samples retrieved after lnc-G4 pulldown through Probe-set 1 and LacZ oligonucleotides were analysed by next generation RNA sequencing in order to identify mRNAs associated with the lncRNA. The sequencing resulted in a list of 61 mRNAs, identified with a bioinformatics pipeline built by Dr. Alessio Colantoni (explained in Materials and Methods). The complete list of identified transcripts is available in Table 1.

	Gene name	P-Value		Gene name	P-Value		Gene name	P-Value
	lnc-G4	9,50E-09	21	Acad8	7,90E-04	42	Asb1	3,79E-03
1	Rps9	7,36E-05	22	Snord72	8,37E-04	43	Eef1g	3,79E-03
2	Mybph	1,80E-03	23	Rpl37	8,37E-04	44	Dclk1	3,79E-03
3	Mir503	4,94E-03	24	Glis3	8,54E-04	45	Cdkn1a	3,89E-03
4	Mir322	4,94E-03	25	Tysnd1	8,66E-04	46	Eef2	4,71E-03
5	Six1	1,14E-05	26	Rps8	9,86E-04	47	Rpl18	4,82E-03
6	Uba52	5,20E-06	27	Rps7	1,07E-03	48	Rn7sk	5,76E-03
7	Kxd1	5,20E-05	28	Myh3	1,41E-03	49	Tnnc1	5,91E-03
8	Arf5	1,14E-04	29	Rpl8	1,45E-03	50	Rpl19	5,93E-03
9	Atf2	1,46E-04	30	Acsl6	1,54E-03	51	Myl1	6,02E-03
10	Ndrg4	2,60E-04	31	Cdc26	1,55E-03	52	AC092404.1	7,54E-03
11	Slc15a4	5,50E-04	32	Mlx	1,65E-03	53	Rpl13	8,11E-03
12	Rpl18a	1,25E-04	33	Myo1c	1,85E-03	54	Mir7079	8,11E-03
13	Snora68	1,25E-04	34	Tnnt3	1,88E-03	55	Vti1b	8,14E-03
14	Rps20	1,62E-04	35	Cnpy2	1,89E-03	56	Mylpf	8,51E-03
15	Spire1	2,04E-04	36	Gm22009	2,08E-03	57	Cd63	8,62E-03
16	Coq2	2,70E-04	37	Ahnak	2,12E-03	58	Rps6	9,07E-03
17	mt-Co1	3,08E-04	38	Eef1a1	2,12E-03	59	Rps27a	9,20E-03
18	Zfp532	3,42E-04	39	Rbm45	2,18E-03	60	Sqstm1	9,21E-03
19	Hist1h2bg	4,38E-04	40	Lars	2,27E-03	61	Rps19	9,21E-03
20	Myeov2	6,59E-04	41	Usp10	2,58E-03			

Table 1. List of mRNAs enriched in lnc-G4 pulldown identified from the RNA sequencing. RNAs are listed based on their p-value (calculated based on the sequencing replicates and the enrichment towards the LacZ negative control) from lower to higher value. Lnc-G4 is the first result since it is the most enriched transcript in its own pulldown.

Moreover, a bioinformatic prediction of the base-pairing interaction between the identified mRNAs and lnc-G4 was performed with the IntaRNA bioinformatics tool, which is specific for the prediction of RNA-RNA interactions among long RNA molecules (Mann et al. 2017). In particular, only interactions with a free energy lower or equal to -20 Kcal/mol have been considered, in order to select only the most stable and consistent base-pairings.
Interestingly, 22 out of 62 mRNAs were predicted to interact in a specific region of 250 nt at the beginning of lnc-G4 exon 4, within a C-rich sequence which is part of a LINE-L2 repeated element (Fig. 6A). Among them, we selected the 7 most promising interactors (i.e. the candidates with the best interaction energy) for further analysis. RT-PCR validation of the selected mRNA candidates identified from the RNA sequencing was performed on different replicates of lnc-G4 pulldown performed with both the probe sets, confirming the efficient precipitation of the identified RNAs; GAPDH was used as negative control (Fig.6B).

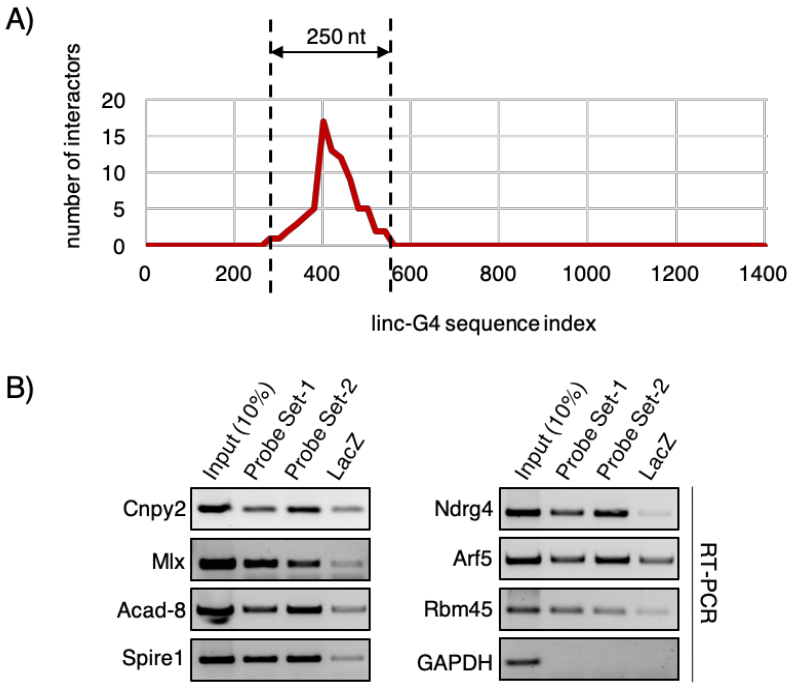


Figure 6. A) Graph showing the number of interactors for lnc-G4 20-nt sequence slots. The majority of interactions fall in a 250-nt region at the beginning of lnc-G4 exon 4. B) RT-PCR validation of mRNA candidate enrichment in lnc-G4 pulldown performed with the two probe sets; GAPDH was used as negative control.

The different pull-down efficiency among different mRNAs can be due to both differences in the length of the base-pairing regions between the lncRNA and the mRNAs and in the energy of interaction: ta good and strong interaction promotes a higher efficiency of pulldown. Due to the technical limitations of working with many candidates, we chose MLX mRNA for further investigation, because of its good enrichment and its well-studied role in regulation of gene expression during myogenesis.

3.4.2. lnc-G4 directly interacts with MLX mRNA

MLX, which stands for MAX-like protein X, is a basic helix-loop-helix component of the MYC/MAX network of transcription factors. The factors involved in this network integrate diverse cellular stimuli, like nutrient sensing and metabolic stress, in order to regulate proliferation and/or differentiation by dimerization and binding to E-box elements in target promoters (Diolaiti et al. 2015). Interestingly, the output resulting from the activation of this network of transcription factors depends of the dynamics of dimerization: for example, MLX can heterodimerize with MXD proteins to triggers proliferation arrest and differentiation, or it can bind to MONDO-A/B factors for glucose sensing activity. In the particular panorama of skeletal muscle differentiation, MLX has been shown to promote myogenesis in a glucose-depending manner: the heterodimerization with the MONDO-A glucose-sensing factors causes MLX translocation to the nucleus, with subsequent activation of several cytokine genes by binding to the carbohydrate responsive elements in their promoters (Hunt et al. 2015). Moreover, MLX is expressed in three different splicing isoforms:

- MLX α, which lacks exon 3;
- MLX β, which contains exon 3;
- MLX γ, which has an extended exon 1.

Using the already cited IntaRNA bioinformatic suite, we predicted the possible base-pairing interaction between lnc-G4 and the three different annotated isoform of MLX mRNA. We were able to find two different regions of interaction, named Region 1 and Region 2 respectively. In particular, Region 1 is restricted to the MLX γ isoform, since it is included in the specific part of exon 1 within its coding sequence, while Region 2 is common to the three isoforms since it falls into the common 3'UTR sequence (Fig. 7A). The presence of all the three isoforms in lnc-G4 pulldown was validated through RT-PCR analysis using specific primers able to distinguish between the different transcripts (Fig. 7B).
In order to check whether the binding between lnc-G4 and MLX mRNA isoforms was direct or not, we optimized the RNA pulldown protocol including a AMT-crosslinking step. AMT (4'-aminomethyl-4,5',8-trimethylpsoralen) is a chemical compound which is able to form covalent bonds amid base-pairing RNA molecules after UV-irradiation at 365 nm; the covalent crosslinking is revertible under UV-irradiation at 254 nm. As shown in Figure 7C, MLX total mRNA was enriched with both probe sets in lnc-G4 pulldown performed with AMT crosslinking, while GAPDH was used as negative control. Moreover, RT-PCR analysis showed that all the three MLX mRNA isoforms were enriched in AMT-crosslinked pulldown (Fig 7D).
Taken together, these data indicate that MLX mRNA is a valid interactor of lnc-G4; in particular, lnc-G4 directly bind to the three isoforms of MLX mRNA, presumably thanks to the two predicted binding regions.

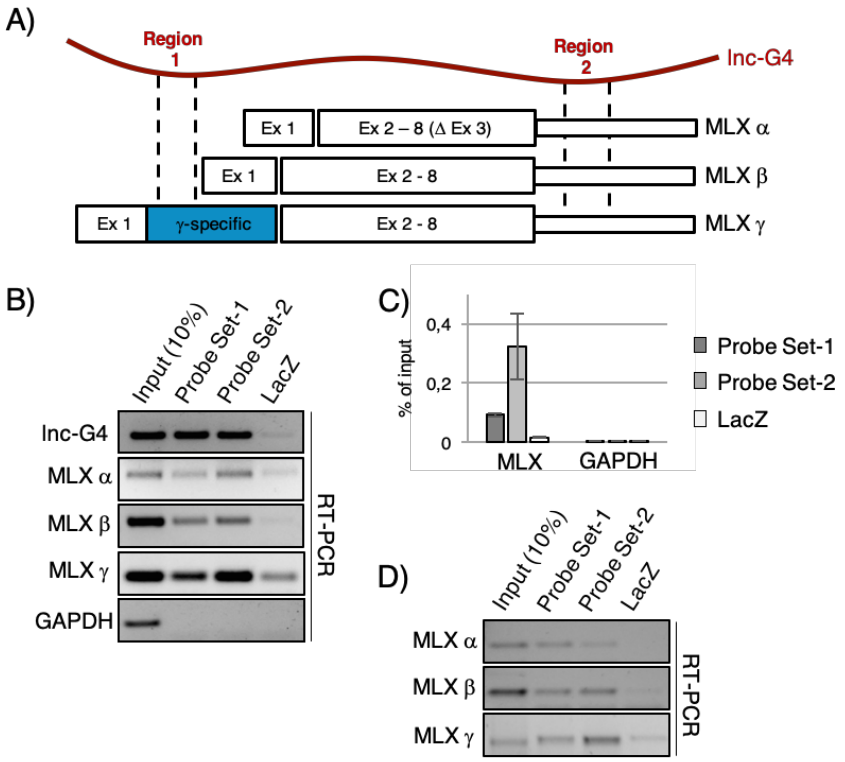


Figure 7 - A) Schematic representation of the respective position of interacting regions 1 and 2 on lnc-G4 and MLX isoforms transcripts. B) RT-PCR validation of the enrichment of the three MLX isoforms in lnc-G4 pulldown performed with the two probe sets; GAPDH was used as negative control. C) qPCR analysis of total MLX mRNA enrichment in AMT-crosslinked lnc-G4 pulldown; data are expressed in percentage of Input RNA, GAPDH is used as negative control. D) RT-PCR validation of the enrichment of the three MLX isoforms in AMT-crosslinked lnc-G4 pulldown.

3.4.3. lnc-G4 interacts with DHX36 RNA helicase, and forms a molecular complex together with MLX γ mRNA

Protein samples isolated from lnc-G4 native pulldown were analysed by mass spectrometry in order to identify putative lnc-G4 protein partners. Specifically, enriched proteins were cleaved into peptides with trypsin and analysed by high-resolution LTQ Velos Pro instrument. We were able to identify 375 proteins in all of the samples (172 in Probe-set 1 sample, 199 in Probe-set 2 sample, and 221 for the LacZ). Among the proteins that were common in lnc-G4 pulldown performed with the two sets of probes, only 14 proteins were significantly enriched compared to LacZ control (Table 2).

	Probe set 1			**Probe set 2**			**LacZ**		
	Score	**Coverage**	**# Peptides**	**Score**	**Coverage**	**# Peptides**	**Score**	**Coverage**	**# Peptides**
PUR-B	108,3	49	9	87,3	52	8	24,6	29	3
DHX36	56	17	10	26,9	9	5	0	0	0
CSRP1	37,3	38	4	26,4	28	3	18	20	2
RFA1	42,6	14	6	21,8	11	4	3,2	2	1
CPSF7	14,3	5	1	27,1	8	2	6	5	1
PFKAM	20,7	8	4	16,3	8	4	4,3	2	1
ALBU	8,1	5	2	18,7	5	2	3,7	2	1
CAZA2	7,2	12	2	13,7	19	3	3,5	6	1
RS11	9,4	11	1	17,7	11	1	5,2	11	1
Hnrnp0	16,4	7	1	3,9	5	1	0	0	0
RS8	8,8	7	1	8,6	7	1	0	0	0
MYL4	3,7	10	1	7,7	23	2	0	0	0
CCD30	3,1	2	1	3,4	2	1	0	0	0
SE1L1	3,5	2	1	3,2	2	1	0	0	0

Table 2 – List of proteins enriched in lnc-G4 pulldown with the two probe sets with respect to LacZ control. Score=sum of individual peptide counts. Coverage=percentage of protein sequence covered by identified peptides. #Peptides=number of distinct and uniquely assignable peptides.

Due to its good enrichment in both Probe-set 1 and 2 pulldowns and its complete absence in LacZ control, we chose DHX36 for further validation. DHX36, also known as RHAU (RNA helicase associated with AU-rich element) and G4R1, is a member of the DExH/D family of ATP-dependent RNA helicases. Thanks to its flexible, partially unstructured N-terminal domain, DHX36 has a strong specificity for unwinding both DNA and RNA G-quadruplex structures. It has been shown that DHX36 is able to regulate both stability and translation efficiency of G-quadruplex containing mRNAs (Nie et al. 2015); its role in post-transcriptional gene regulation made it a good candidate to study.
First of all, we validated the presence of DHX36 in different replicates of lnc-G4 pulldown through Western Blot, confirming the results of the mass spectrometry (Fig. 8A). To further confirm the interaction between DHX36 and lnc-G4, we performed RNA immunoprecipitation of the protein and checked for the enrichment of lnc-G4 through RT-PCR. Interestingly, we could demonstrate that also MLX is immunoprecipitated together with DHX36, while GAPDH was used as a negative control (Fig. 8B). This evidence allowed us to speculate that lnc-G4, MLX mRNA and DHX36 are part of the same cytoplasmic molecular complex.
To understand whether lnc-G4 is required to recruit DHX36 on MLX mRNA, we performed DHX36 immunoprecipitation on extracts of C2C12 cell transfected with scramble and lnc-G4 siRNA. The results showed that, in scramble condition, both lnc-G4 and MLX total mRNA were enriched in DHX36 pulldown, compared to the positive control WBP4 (Lattmann et al. 2011) and the negative control GAPDH; however, no significative changes in MLX enrichment were detected after lnc-G4 knockdown with the specific siRNA (Fig. 8C).

Since lnc-G4 had been shown to interact with the three MLX mRNA isoforms, we verified their enrichment in DHX36 immunoprecipitation in both scramble and lnc-G4 depletion conditions. The RT-PCR analysis revealed that DHX36 specifically interacts with the MLX γ isoform, while α and β isoforms resulted to be not enriched; also in this case, lnc-G4 knockdown did not affect the enrichment (Fig. 8D).
Taken together, these data describe DHX36 as a good protein interactor for lnc-G4; moreover, DHX36 is also able to specifically interact with MLX γ isoform in a lnc-G4 independent manner.

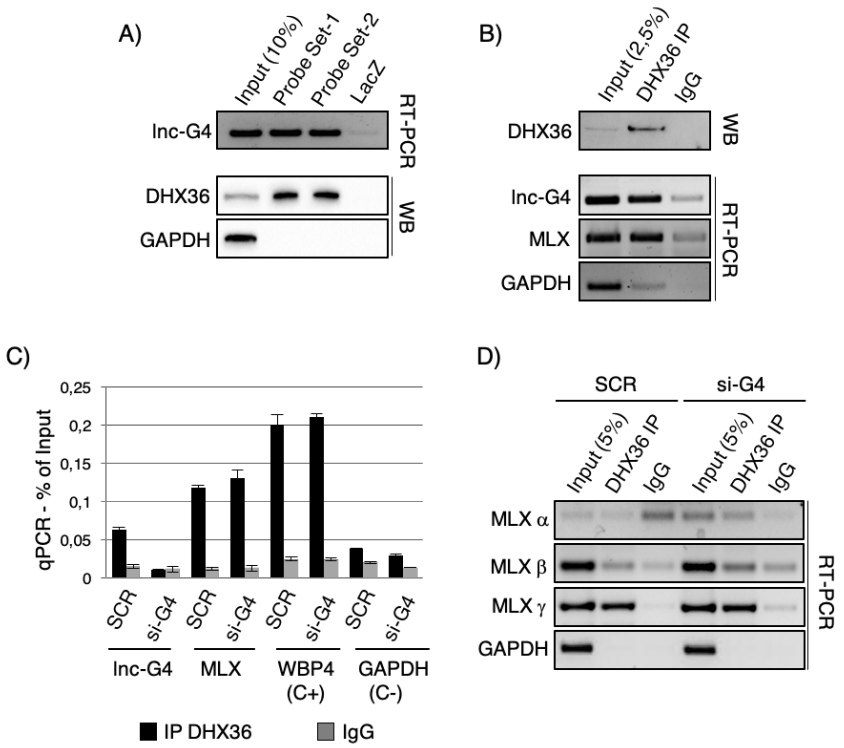


Figure 8 – A) WB showing the effective enrichment of DHX36 in lnc-G4 pulldown; GAPDH was used as negative control. B) WB showing the effective immunoprecipitation of DHX36, and subsequent RT-PCR analysis of the enrichment of lnc-G4 and MLX total mRNA; GAPDH was used as negative control. C) qPCR quantification RNA enrichment in DHX36 immunoprecipitation in scramble and lnc-G4 depletion. WBP4 was used as a positive control, GAPDH as a negative control. Results are expressed in percentage of Input RNA. D) RT-PCR analysis of MLX isoforms enrichment in DHX36 immunoprecipitation in control versus lnc-G4 depletion condition; GAPDH was used as negative control.

3.5. lnc-G4 and DHX36 do not affect MLX mRNA stability and the total protein levels

Previously shown data demonstrate a reciprocal relationship between lnc-G4, MLX mRNA isoforms and the RNA helicase DHX36, which are coexisting in the same molecular complex.
In order to understand the role of both lnc-G4 and DHX36 in the post-transcriptional regulation of MLX, we investigated the behaviour of MLX mRNA isoforms in C2C12 cells at day 2 of differentiation in condition of lnc-G4 and DHX36 knockdown. As shown in Figure 9A, the strong depletion of lnc-G4, performed with two different siRNAs, and of DHX36 did not affect the level of MLX isoforms transcripts, at least at the steady state, suggesting no effect on the mRNA stability.
To check whether lnc-G4 and DHX36 could regulate MLX translation efficiency, we also controlled the conduct of MLX protein in C2C12 cells at day 2 of differentiation in condition of lnc-G4 and DHX36 knockdown. Even in this case, it was not possible to appreciate any significative change in the overall MLX protein level after the good depletion of both lnc-G4 and DHX36 (Fig. 9B). However, the commercially available antibody for MLX detection was not able to detect and discriminate the three protein isoforms corresponding to α, β and γ transcripts, loosing information about a hypothetical isoform-specific regulation of translation driven by lnc-G4 interaction. To evaluate and verify this hypothesis, the following experiments have been performed using the luciferase reporter system described below.

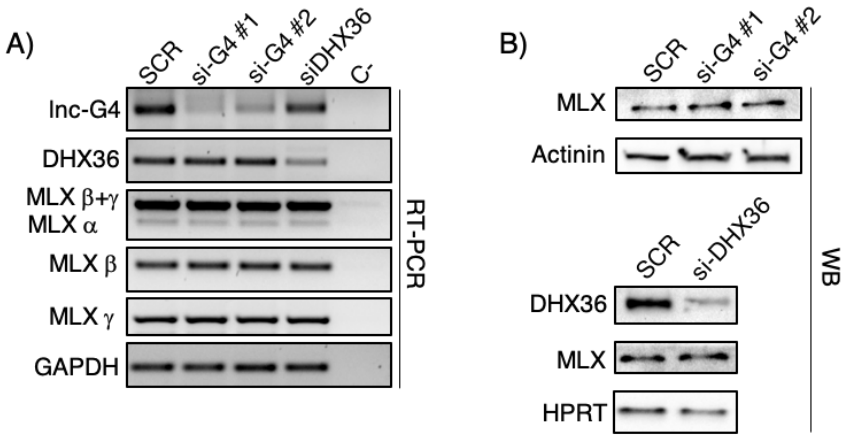


Figure 9 – A) RT-PCR analysis of MLX isoforms modulation after lnc-G4 and DHX36 depletion. Lnc-G4 and DHX36 were amplified to check the effective knockdown; GAPDH was used as a negative control. B) Western Blot analysis of MLX total protein level after lnc-G4 and DHX36 depletion. DHX36 hybridization was used to check the effective knockdown; Actinin and HPRT respectively were used as negative controls.

3.6. lnc-G4 specifically regulates MLX γ translation

The further characterization of the role of lnc-G4 in the regulation of MLX isoforms required the construction of reporter constructs. To do so, we built a luciferase construct in which the two predicted regions of interaction between MLX mRNAs and lnc-G4 were cloned upstream and downstream the *Renilla* luciferase coding sequence. Specifically, the exon 1 of MLX γ, containing the MLX γ-specific region of interaction, was cloned at the 5' of *Renilla* luciferase, in frame to its coding sequence, while the common 3'UTR was cloned at the 3' of *Renilla* luciferase.
The effect of lnc-G4 on the described construct was evaluated performing luciferase assay on C2C12 proliferating cells, which normally do not express lnc-G4, after transfection of an empty vector or of a lnc-G4 overexpressing plasmid. As represented in Figure 10A,

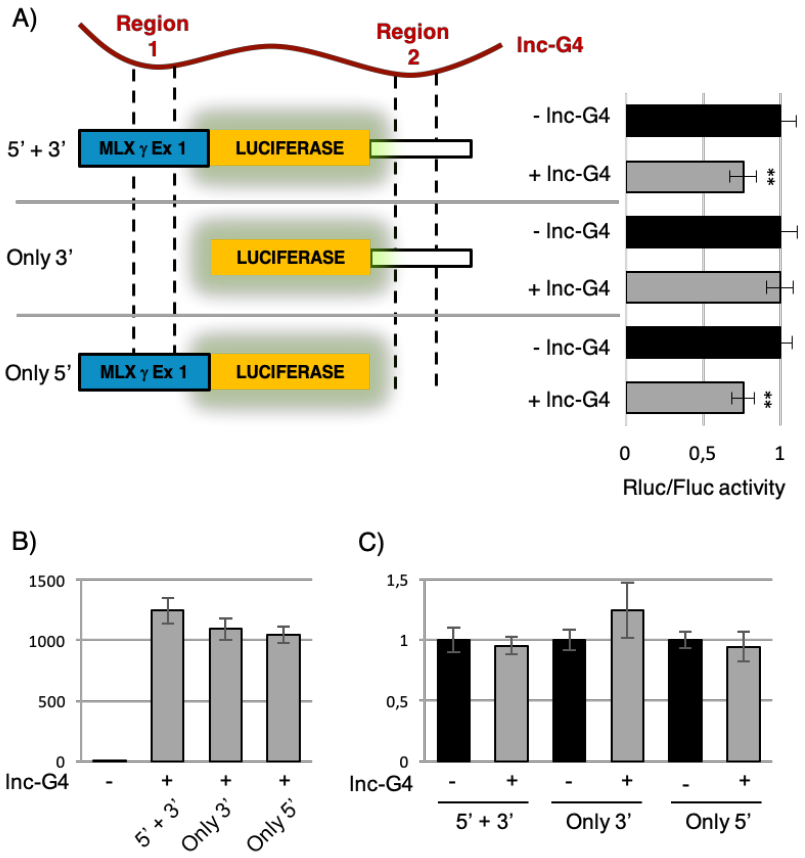


*Figure 10 – A) Schematic representation of Renilla luciferase reporter constructs used to study the translational regulation of MLX γ isoform. In the same panel, graph showing the result of luciferase assay performed with the 5'+3', only 3' and only 5' mutants of MLX-Rluc reporter; results are expressed as ratio between RLuc and the control luciferase Fluc. **=p<0,05. B) qPCR measurement of lnc-G4 overexpression. C) qPCR measurement of the levels of MLX-Rluc reporter mRNA.*

the data show a significative downregulation of ~25% of luciferase activity in presence of lnc-G4.
In order to understand if both the regions of interaction are required for lnc-G4 mediated regulation, we generated mutants of the luciferase construct:

- a first mutant, called "only 3'", contains only the MLX 3'UTR with the isoform-common region of interaction;
- a second mutant, called "only 5'", contains only the MLX-γ specific region at the 5' of the coding sequence.

The experiments performed with the "only 3'" mutant showed no significative changes in luciferase activity in absence or presence of lnc-G4; *per contra*, experiments carried out with the "only 5'" mutant present a significative downregulation of ~25% of luciferase activity in lnc-G4 overexpression, compared to the empty vector control.
To exclude an effect on RNA stability and confirm that the observed downregulation is due to a translational modulation, qPCR analysis was performed to check the effective lnc-G4 overexpression (Fig. 10B) and the levels of the luciferase construct mRNA (Fig. 10C) in the different conditions of the experiments. While lnc-G4 expression was comparable among the different conditions, the levels of MLX-luciferase construct mRNAs did not show any significant changes.
Taken together, these data demonstrate that lnc-G4 has a role in translational modulation of MLX γ, the only isoform to contain the 5' interacting region.

To further characterize the lnc-G4 mediated regulation of MLX γ, and to evaluate whether the base-pairing between lnc-G4 and MLX γ mRNA is responsible for this regulation, luciferase assays have been repeated with *ad hoc* mutants of the "only 5'" reporter constructs.
First of all, a mutant carrying a 75-bp deletion corresponding to the 5' interacting region, thus abolishing the base-pairing between lnc-G4 and the reported transcript, has been generated (named Δ-mutant). As shown in Figure 11, luciferase assay performed with Δ-

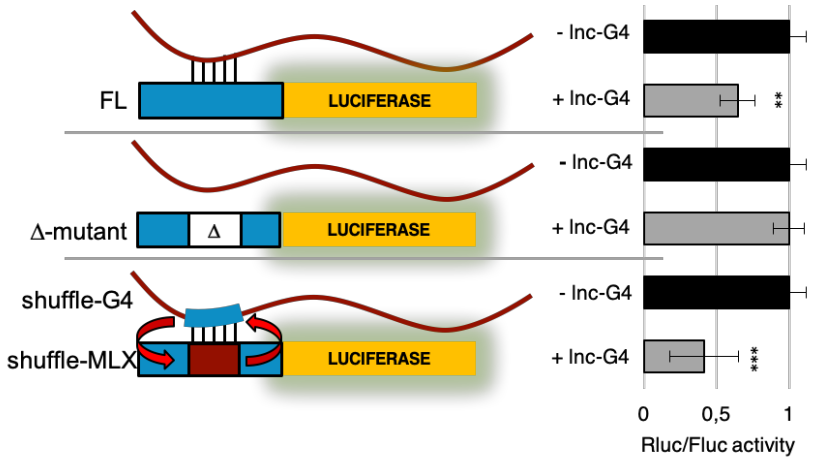


*Figure 11 –Schematic representation of Renilla luciferase reporter mutants used to study the importance of base-pairing in translational regulation of MLX γ isoform. In the same panel, graph showing the result of luciferase assay performed with the full length reporter (FL), the deletion mutant (Δ-mutant) and the shuffle mutants of lnc-G4 overexpression plasmid and MLX-Rluc reporter; results are expressed as ratio between RLuc and the control luciferase Fluc. **=p<0,05, ***=p<0,01*

mutant did not show any significative variation of luciferase activity in lnc-G4 overexpression, compared with the empty vector control. This evidence suggest that the base-pairing region is necessary for the translational regulation of the luciferase construct.
Furthermore, in order to better investigate the regulation mechanism, we generated two complementary constructs, respectively named shuffle-G4 and shuffle-MLX, in which the two regions of interaction have been exchanged. This complementary mutation is able to disrupt the transcripts sequence, but also to restore the base-pairing between the molecules.
Luciferase assay performed with the two shuffle-mutants showed a significant downregulation of more than 50% of the luciferase activity in lnc-G4 overexpression, compared with the empty vector control. Even in this case the RNA levels have been controlled by qPCR in order to exclude changes in reporter transcripts levels and to confirm the exclusive translational effect (data not shown).

Taken together, these data demonstrate that lnc-G4 regulates MLX γ translation in a base-pairing dependent manner, and that the base-pairing is both necessary and sufficient to trigger the translational regulation.

3.7. MLX γ interacting region can fold into a G-quadruplex structure

The previous experiments have shown that the base-pairing interaction between lnc-G4 and MLX γ is functional and essential to exert the translational regulation, so our interest fell on investigating better the characteristics of this specific region of the transcript.
Since MLX γ is efficiently enriched in DHX36 immunoprecipitation respect to the other isoforms, and since the MLX γ specific part of exon 1 is a G-rich region that base-pairs with lnc-G4 C-rich region, we could hypothesize the presence of a RNA G-quadruplex contained within MLX γ mRNA. To evaluated this possibility, we used the QGRS mapper bioinformatics tool (Kikin et al. 2006); briefly, this tool is able to predict the presence of G-quadruplexes within a given sequence and assign them a G-score based on the length of the G-stretches and of the interposed loops. We could identify a stable G-quadruplex structure, with a G-score of 63, inside the MLX γ specific region of exon 1; moreover, the putative G-quadruplex forming sequence overlaps the interacting region with lnc-G4 (Fig. 12A).
To validate the existence of the identified G-quadruplex, we resorted to an *in vitro* folding assay(Booy et al. 2018). Briefly, we designed RNA oligonucleotides with the sequences of lnc-G4 and MLX γ interacting regions; the oligonucleotides have been allowed to fold in presence of potassium ions (that are able to stabilize G-quadruplex structures) and then loaded on native acrylamide gels.

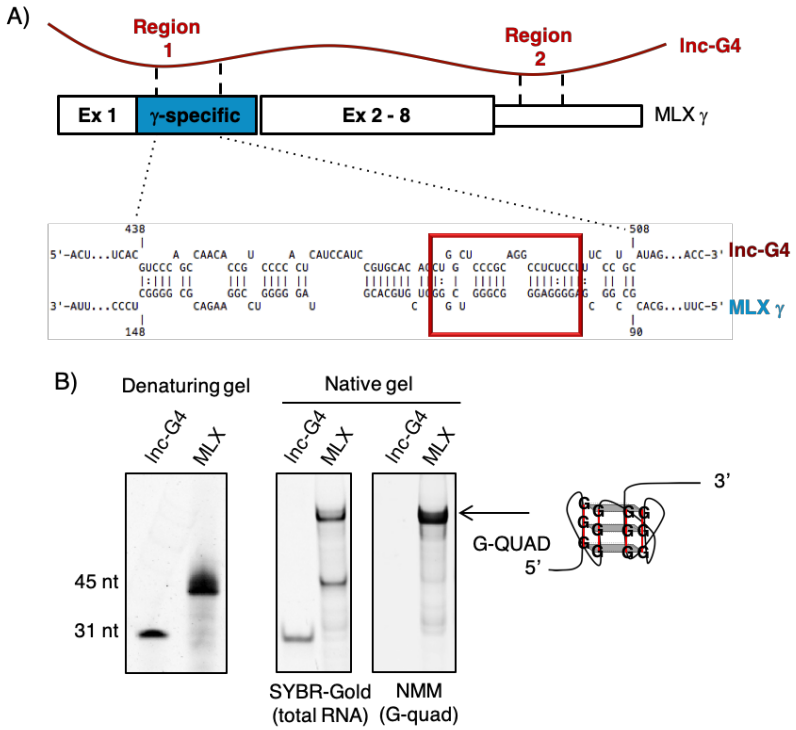


Figure 12 – A) Schematic representation of the base-pairing between lnc-G4 region 1 and MLX γ exon 1; the red box highlights the putative G-quadruplex forming region. B) In vitro folding assay to validate the G-quadruplex structure in MLX γ exon1; Sybr-Gold was used for total RNA detection, while N-methyl mesoporphyrin IX (NMM) was used for specific G-quadruplex marking.

Total RNA was subsequently stained with Sybr-Gold, while G-quadruplexes were identified with the specific N-methyl mesoporphirin IX dye (Arthanari et al. 1998).
RNA oligonucleotides were also loaded on acrylamide denaturing gel in order to disrupt secondary structures and compare the electrophoretic migration with the native condition.
As shown in Figure 12B, only two bands corresponding to the linearized oligonucleotides were detectable in denaturing condition. *Per contra*, a supershift band was identifiable when MLX γ oligonucleotide was loaded onto the native gel, while no supershift

band was detectable for the lnc-G4 oligonucleotide. Moreover, the N-methyl mesoporphyrin IX staining demonstrated that the supershift band identified after MLX γ oligonucleotide migration on native acrylamide gel corresponded to a *bona fide* G-quadruplex. Taken together, these data demonstrate that MLX γ sequence is able to fold into a stable G-quadruplex structure that could be important for the translational regulation; interestingly, the G-quadruplex forming sequence is embedded inside the lnc-G4 binding region, so the interaction between the two molecules could require the intervention of DHX36 to unwind the secondary structure and allow the interaction.

3.8. lnc-G4 dependent MLX γ modulation affects the subcellular localization of total MLX protein

According to scientific literature, the three MLX protein isoforms are ubiquitously expressed, but their reciprocal ratio is tissue specific. Moreover, MLX α, β and γ are able to form different heterodimers depending on the relative abundance of the isoforms, thus affecting the subcellular localization. In fact, MLX γ contains both nuclear and cytoplasmic localization signals, and its heterodimerization with α and β isoforms trigger their translocation from the cytoplasm to the nucleus (Meroni et al. 2000, schematic representation in Fig. 13A). Given this possibility, we hypothesized that lnc-G4 could modulate the nuclear or cytoplasmic accumulation of MLX total protein through the specific translational regulation of the only MLX γ isoform. To check this hypothesis, we performed immunofluorescence experiments using a MLX-specific antibody in C2C12 at day 2 of differentiation in different conditions (Fig. 13 C). In cells transfected with a scramble control, MLX shows a strong cytoplasmic accumulation with formation of aggregated clusters, while the detectable signal in the nucleus is lower. On the contrary, lnc-G4 knockdown led to the loss of cytoplasmic granulation, with

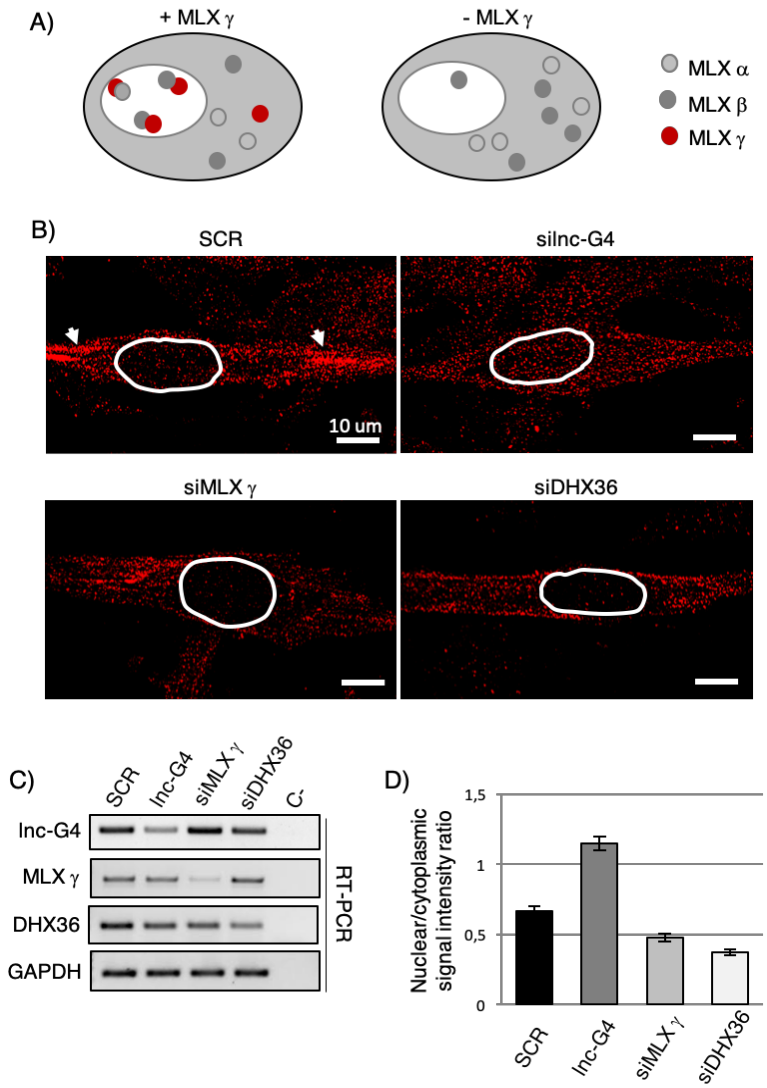


Figure 13 – A) Schematic representation of MLX γ role in subcellular trafficking of the other isoforms. B) MLX immunofluorescence in C2C12 at day 2 of differentiation in knockdown of lnc-G4, MLX γ and DHX36; the white circle indicates the nucleus position (stained with DAPI, data not shown). C) RT-PCR control of effective knockdown of lnc-G4, MLX γ and DHX36 D) Quantification of nuclear/cytoplasmic fluorescence intensity ratio.

a higher intensity of fluorescence signal within the nucleus and a nuclear/cytoplasmic fluorescence intensity ratio of ~1. This experimental evidence is coherent with the upregulation of MLX γ protein levels and subsequent higher efficiency of nuclear shuttling of the other MLX isoforms.
Moreover, we used MLX immunofluorescence to control whether direct modulation of MLX γ levels could affect the subcellular localization of the total protein. In this case, MLX γ depletion with a specific siRNA caused a significant decrease of nuclear/cytoplasmic fluorescence intensity ratio, meaning that signal intensity in the nucleus was lower respect to the scramble control. Finally, the same system was used to check the effect of DHX36 knockdown on MLX subcellular localization; even in this case a significant decrease of nuclear/cytoplasmic fluorescence intensity ratio was measured. This is coherent with a hypothesized positive effect of DHX36 on MLX γ translation, while its knockdown causes a less efficient translation and phenocopies the absence of MLX γ isoform obtained by RNA interference.
The effect of lnc-G4, MLX g and DHX36 knockdown on MLX protein subcellular localization has been evaluated both qualitatively, through simple image visualization (Fig. 13B), and quantitatively, through quantification of the signal intensity and expression of the nuclear/cytoplasmic ratio (Fig. 13D).
Taken together, these data show that the translational modulation of the single MLX γ isoform operated by lnc-G4 is sufficient to intervene on the subcellular localization of total MLX protein.

Since MLX has a validated role in directing skeletal muscle differentiation (Hunt et al. 2015), we checked if there was any dynamic change of its subcellular localization during the timing of skeletal muscle differentiation. To do that, we performed MLX immunofluorescence on C2C12 cells at different time points of differentiation. Interestingly, we were able to show that MLX had higher levels of nuclear signal both in proliferation (GM) and in the late phases of myogenesis (DM3/4), while the protein is more accumulated in the cytoplasm in the early phases of differentiation

(DM2). This observation is coherent with lnc-G4 expression profile, which has a peak in the early phases of differentiation (DM2, precisely), so the highest level of expression of lnc-G4 corresponds to the maximal translational repression of lnc-G4, with subsequent lower level of nuclear/cytoplasmic fluorescence signal intensity ratio (Fig 14A and B).

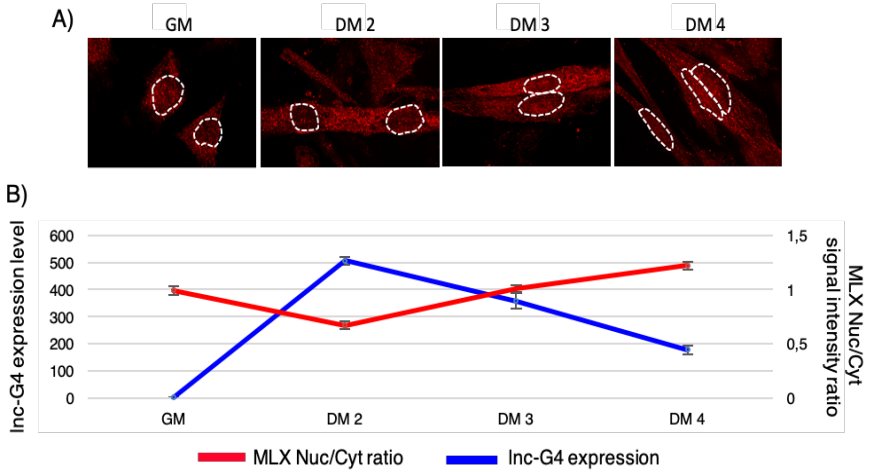


Figure 14 – A) MLX immunofluorescence in C2C12 at different time points of differentiation; the white circle indicates the nucleus position. B) Quantification of nuclear/cytoplasmic MLX fluorescence intensity ratio at the different time points of differentiation, compared to lnc-G4 expression profile during C2C12 myogenesis.

3.9. lnc-G4 could regulate the translation of many other mRNAs

Nowadays, high-throughput techniques such as Ribosome Profiling (Ingolia et al. 2009) allow the transcriptome-wide study of translation by analysing the mRNA translation efficiency and not the protein accumulation. Briefly, Ribosome Profiling is based on global translational blockade with ciclohexymide, followed by partial RNase I digestion and recovery of ribosome-protected mRNA fragments, which can then be used for library preparation

and subsequent sequencing. The alignment of the obtained reads gives an insight on the ribosomal occupancy on a given mRNA, thus inferring its translation levels. Moreover, Ribosome Profiling obtained data can be normalized on RNA-seq data, that give information about the mRNA abundance and stability, in order to obtain an accurate measurement of translational efficiency (Fig 15A).
Since we could demonstrate a role for lnc-G4 in MLX g translational regulation, and since the mRNA interactome of lnc-G4 had been shown to be very complex, we postulated the possibility that lnc-G4 could affect and modulate the translation of other mRNAs.
To verify this hypothesis, I spent part of my PhD at the Max Delbrück Centre for Molecular Medicine in Berlin, Germany, in the laboratory of Dr. Markus Landthaler. During this period of collaboration, I performed Ribosome Profiling on C2C12 cells at day 2 of differentiation transfected with both scramble and lnc-G4 specific siRNA, generating the NGS libraries. After reads alignment and ribosome occupancy quantification, the Translation Efficiency was calculated normalizing the obtained data on the RNA sequencing data produced during the phenotypic characterization of lnc-G4 effect on skeletal muscle differentiation.
The graph shown in Figure 15B resumes the analysis results. 175 genes were found to be significantly translationally modulated after lnc-G4 knockdown: for 55 of them the translational efficiency was decreased, meaning that lnc-G4 could promote their translation, while 120 of them had an increased translational efficiency, suggesting a role of lnc-G4 as a translational repressor.
These preliminary data, that obviously require further analysis and validation, suggest a wider role for lnc-G4 in translational regulation of a large subset of mRNAs.

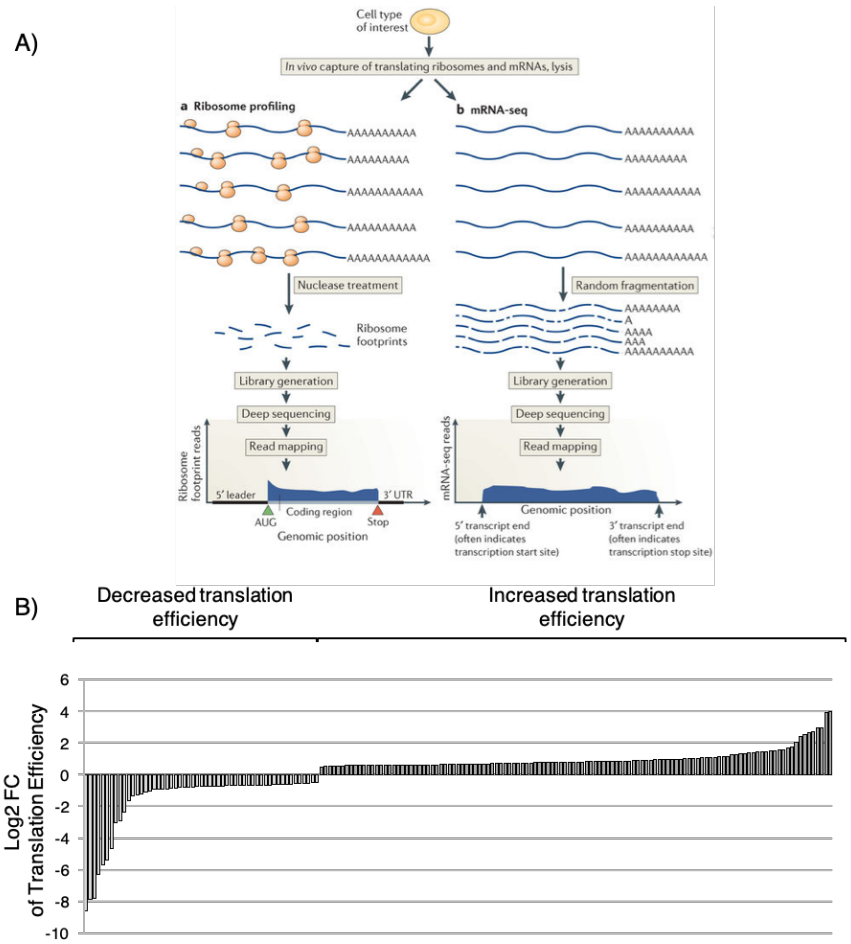


Figure 15 – A) Schematic representation of Ribosome Profiling workflow. B) Graph showing the results from the Ribosome Profiling in scramble versus lnc-G4 knockdown; results are expressed in terms of Translation Efficiency, which is normalized on the mRNA abundance estimated from RNA sequencing data.

3.10. Possible extensibility of lnc-G4 molecular mechanism

As demonstrated from the previously presented experiments, lnc-G4 modulates MLX γ translation thanks to the interaction with a G-quadruplex forming region. Since the region of lnc-G4 with the higher number of predicted interactions with mRNAs is a C-rich region, the binding of sequences which are prone to form G-quadruplex structures could be a key feature of lnc-G4 in order to exert its function. To pursue this goal, we re-analysed the list of putative mRNA interactors identified in the sequencing following lnc-G4 pulldown with the QGRS mapper tool, in order to identify putative G-quadruplex structures. The analysis showed that among the 22 mRNAs which are interacting in the C-rich, 250-nt region of lnc-G4, 21 mRNAs have at least a putative G-quadruplex structure embedded within the interacting region. Identified G-quadruplexes are listed in Table 3.
A preliminary validation of the possible formation of these quadruplex structures has been carried out through the analysis of the enrichment of the candidate transcripts in DHX36 immunoprecipitation. With this approach, we could demonstrate that many candidate mRNAs are significantly interacting with DHX36, which has a strong specificity for G-quadruplex containing mRNAs (Figure 16).
These last presented data are quite preliminary and need further corroborating experiments, but they suggest that the interaction with G-quadruplex forming sequence could be a common feature underlying lnc-G4 mechanism of action.

(Next page) Table 3 – List of G-quadruplex identified in lnc-G4 mRNA interactors, expressed with their G-score, the localization inside the mRNA, the position of the interacting sequence on lnc-G4, the interaction energy.

Gene Name	G_score	Localization	Start on lnc-G4	End on lnc-G4	Energy (kcal/mol)
Acad8	52	CDS	375	468	-309.319
Ahnak	57	3'UTR	257	400	-331.542
Arf5	63	3'UTR	368	474	-349.984
Asb1	54	CDS	379	405	-190.543
Asb1	54	CDS	334	455	-348.598
Cd63	63	5'UTR	372	514	-455.519
Cdkn1a	61	3'UTR	374	475	-354.752
Cnpy2	62	5'UTR-CDS	372	417	-283.902
Eef2	49	3'UTR	372	410	-309.017
Glis3	59	3'UTR	349	443	-254.118
Glis3	67	3'UTR	389	459	-229.127
Lars	49	5'UTR	295	424	-260.651
Mlx	63	CDS	438	508	-196.417
Mlx	63	CDS	438	508	-222.834
Mybph	62	3'UTR	367	475	-35.183
Mybph	50	3'UTR	1356	1383	-217.779
Ndrg4	51	3'UTR	369	417	-228.485
Rbm45	63	5'UTR	394	542	-278.736
Rpl19	60	CDS	374	411	-219.518
Rpl37	62	CDS	372	444	-205.619
Rps19	72	CDS	369	506	-433.249
Six1	63	3'UTR	317	448	-289.877
Spire1	62	CDS	440	542	-199.765
Spire1	62	5'UTR	440	542	-199.765
Uba52	61	5'UTR	368	419	-260.024
Uba52	61	5'UTR	368	419	-260.811
Usp10	56	3'UTR	1347	1385	-29.713
Vti1b	54	5'UTR	369	458	-217.576

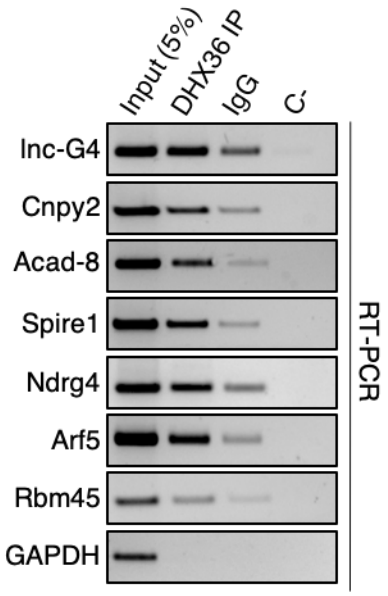


Figure 16 –RT-PCR validation of G-quadruplex containing mRNAs enrichment in DHX36 immunoprecipitation; GAPDH was used as negative control.

4. DISCUSSION

With historic attention on protein-coding genes, many non-coding RNAs have led a life in the dark. Recently, the appreciation of a high number of ncRNA loci by high-throughput technologies has created renewed interest and urged questions about the ‘‘usefulness’’ of such transcription activity (Jensen et al. 2013).
Nowadays there are many examples of regulatory functions exerted by non-coding RNAs, and among this category, long non-coding RNAs are taking the place of honour. The experimental evidences that came out in the last years allowed describing a wide range of functions and molecular mechanisms mediated by lncRNAs, especially in directing and finely modulating complex processes such as cellular differentiation. Long non-coding RNAs are assuming an important position in regulation of myogenesis, both at transcriptional and post-transcriptional level.
In order to investigate additional layers of regulation, we started the molecular and functional characterization of a cytoplasmic long non-coding RNA selected from a set of 30 lncRNAs differentially expressed during skeletal muscle differentiation (Ballarino et al. 2015). In this way we could identify and characterize lnc-G4, so called because of its intrinsic ability to bind G-quadruplex forming sequences, which is exclusively expressed in skeletal muscle tissue of adult mice and it is upregulated in *mdx* dystrophic muscle tissues. This upregulation opens the interest on lnc-G4 role during muscle regeneration, a still poorly characterized process of strong physiological importance, and could shed light on future identification of a human homolog on lnc-G4, that could be involved in the dynamics of skeletal muscle pathologies such as Duchenne Muscular Dystrophy.
The importance of lnc-G4 in the proper onset of skeletal muscle differentiation was widely demonstrated through a RNA interference approach, with both the morphological analysis of myotube formation and the characterization of the transcriptomic changes in C2C12 cells by RNA sequencing. Both the approaches demonstrated the strong importance of lnc-G4 to promote skeletal muscle differentiation and guarantee proper myoblast fusion and

myotube formation. Interestingly, lnc-G4 knockdown does not affect the master regulators of myogenesis, such as MYOD and Myogenin, indicating a role in corroborating the myogenic program after its triggering by the canonical transcription factors network. Nonetheless, lnc-G4 depletion affects the expression of important regulatory proteins and components of mature myotubes contractile structures, indicating its importance in the establishment of late myogenic differentiation following the myoblast fusion step.

As described in the Introduction, cytoplasmic lncRNAs are important directors of interaction dynamics between mRNAs and protein partners. Intriguingly, their scaffold ability coupled with a modular structure allow lncRNAs to form cytoplasmic subcompartments in which specific mRNAs are confined, in order to efficiently regulate their stability and translation even against total cytoplasmic stoichiometry. The efficient native RNA pulldown approach adopted in this work allowed both the identification and the validation of a complex network of lnc-G4 interactors; in particular, the protocol allowed the recovery of endogenous lnc-G4 molecules at the peak of expression (i.e. in C2C12 at day 2 of differentiation), without the need of forcing the system with lnc-G4 overexpression or using the *in-vitro* transcribed RNA as bait.
For what concerns the mRNA interactors, the bioinformatics analysis following RNA sequencing showed the strong enrichment of base-pairing interactions between the identified mRNAs and lnc-G4 within a C-rich region in its exon 4. We could also validate the effective interaction of several candidate mRNAs, even if with variable pulldown efficiency that could be due to many factors, including probes accessibility (the biotinylated probe could compete for the binding of the same region), strength of the interaction, length of the base-pairing portion. This peculiar 250-nt sequence is part of the remains of a repeated LINE L2 element, and constitutes a perfect “interaction platform” for G-rich mRNAs. The importance of repeated elements as mediators of base-pairing between RNA molecules is a common feature of several lncRNA-mediated mechanisms of gene expression regulation (Carrieri et al. 2012; J. Wang, Gong, and Maquat 2013). In particular, LINE L2 elements

are lineage-restricted to rodents; this could suggest the lineage-specific evolution of lnc-G4 in a mouse as a post-transcriptional regulator of a subset of mRNAs. With this hypothesis, it could be possible that different transcripts, specifically evolved in other species, could exert the same mechanism of action on G-quadruplex containing mRNAs, even without any synthenic and sequence conservation.

Moreover, we demonstrated that lnc-G4 interacts with DHX36, a RNA helicase characterized by a specific G-quadruplex recognition domain at its N-terminus. Thanks to our pulldown approach, we could define this interaction network in which a long non-coding RNA, a RNA helicase with G-quadruplex resolvase activity and G-quadruplex containing mRNAs coexist. This idea was even corroborated by the finding of the good enrichment of MLX mRNA in both lnc-G4 pulldown and DHX36 immunoprecipitation. Peculiarly, lnc-G4 is able to interact with all the three isoforms thanks to the second interacting region in their 3' UTR, while the second region of interaction, which has been shown to be functional for the translational regulation, is MLX γ specific.
Furthermore, DHX36 is able to immunoprecipitate only MLX γ isoform, reasonably thanks to the G-quadruplex contained in its extended exon 1. Thanks to these experimental evidences, we could elaborate a working model in which the interaction in the 3' UTR is required for a pre-engagement of lnc-G4 on MLX mRNA; subsequently, when DHX36 open the MLX γ G-quadruplex, making the second interacting region available for the base-pairing, lnc-G4 interacts with the quadruplex-forming sequence and inhibits MLX γ translation.

The complex molecular mechanism that we described is so far limited to a single isoform of MLX mRNA. However, our data demonstrate that the modulation of the protein levels of MLX γ is *per se* sufficient to regulate the nuclear abundance of the total MLX protein. This peculiar way of action constitutes a novel paradigm of fine tuning of gene expression, where the translational regulation of

a single protein isoform impacts on its heterodimerization stoichiometry with other protein partners. Especially in the case of MLX, the modulation of a single isoform could unbalance the complex interaction network that this protein has with the Myc/MAX family of transcription factors, strongly affecting their target genes. In our system, lnc-G4 knockdown increases MLX γ protein levels, thus leading to a higher nuclear import of the total protein; this is consistent with the upregulation of some validated MLX target genes, like the cytokines CXCL5, CCL2 and CCL7, that we see when lnc-G4 is downregulated.
Moreover, through immunofluorescence we demonstrated that MLX protein subcellular trafficking is highly dynamic during C2C12 skeletal muscle differentiation. In particular, our evidences show a strong nuclear depletion of the protein in early phases of myogenesis, with respect to the proliferation condition and the later phases of myogenesis where the protein is more diffused throughout the cell. This regulation of the subcellular abundance of MLX could be related to the protein function, and lnc-G4 action could be a fast, time-saving and energetically "cheap" way to achieve a quick nuclear depletion of MLX. Nevertheless, since the nuclear concentration of MLX is higher in late phases of C2C12 differentiation, lnc-G4 mechanism of action could be a way to temporally sequester MLX in the cytoplasm without shutting down its transcription, in order to keep it available for its function in terminal differentiation.

All together, the evidences presented in this thesis work allowed the formulation of a working model of the dynamics of translational regulation mediated by lnc-G4. As shown in Fig. 17A, the G-quadruplex present in MLX γ mRNA could have *per se* a repressive effect on its translation. In normal conditions, DHX36 is able to open this G-quadruplex structure and positively regulate MLX γ translation (in fact, its depletion impacts MLX γ translation and total MLX subcellular localization). However, in the specific window of time during early C2C12 differentiation in which lnc-G4 is strongly accumulated, MLX γ translation needs to be shut down again

through base-pairing in the G-quadruplex containing region. The effect on the single isoform reflects on the physiology of skeletal muscle differentiation by impacting the subcellular localization of the other isoforms, since MLX γ is required as a nuclear carrier of the other protein isoforms.

This elegant mechanism of action, that couples translational control with nuclear/cytoplasmic trafficking, is however limited to a single isoform of a single mRNA. Nonetheless, this work presents preliminary data suggesting a wider role of lnc-G4 in translational regulation of other mRNAs, and in particular of G-quadruplex containing mRNAs. On one hand, the produced Ribosome Profiling data demonstrate that lnc-G4 downregulation has a strong impact on the translation of a subset of mRNAs, while it doesn't affect their mRNA stability. These data need further analysis to understand if the affected genes are cooperating on a specific pathway, and a subsequent experimental validation work.
On the other hand, many G-quadruplex containing mRNAs are interacting with both lnc-G4 and DHX36. For these candidates it could be possible to hypothesize a similar mechanism of translational regulation, and of course Western Blot and reporter system validation is needed. Moreover, an even wider approach to demonstrate the extensibility of lnc-G4 way of action could involve the crosstalk between Ribosome Profiling data and the identified G-quadruplex containing mRNAs, in order to identify robust candidates to select for further investigation.

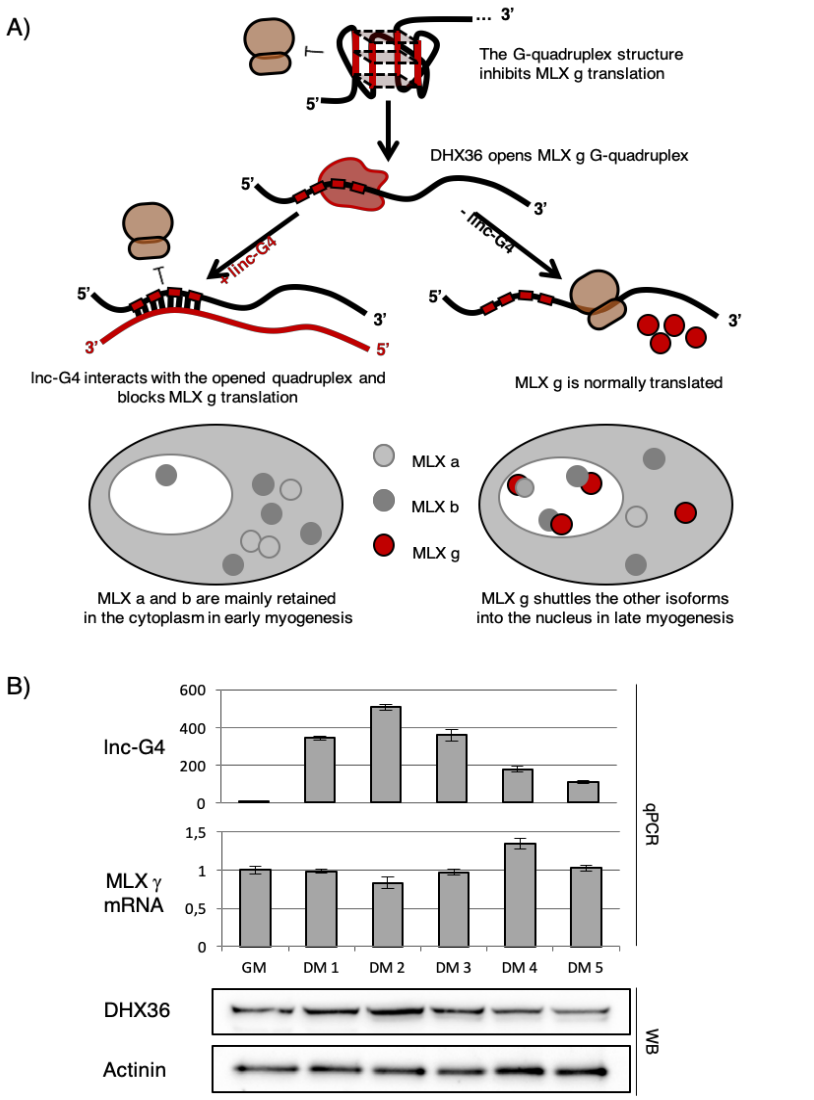


Figure 16 – A) Schematic representation of the working model of lnc-G4 mediated regulation of MLX γ translation and total MLX subcellular localization.
B) Pattern of expression of the components of the working model (lnc-G4, MLX g mRNA and DHX36 protein) during a time-course of C2C12 differentiation. qPCR results are normalized on GAPDH. Actinin was used as a Western Blot loading control.

5. MATERIALS AND METHODS

5.1 Cell culture and treatment methods

C2C12 are murine myoblast cells derived from satellite cells the thigh muscle of a two month old female C3H mouse donor 70h after a crush injury. These cells are a subclone of C2 myoblasts (Yaffe & Saxel 1977) which spontaneously differentiate in culture after serum removal (Burattini et al. 2004). C2C12 are maintained in proliferating conditions, under 60 % of confluence, in Growth Medium (GM; Dulbecco's Modified Eagle's Medium supplemented with 2 mM L-glutamine, 100 U/ml Penicillin, 100 ug/ml Streptomycin, 20% Fetal Bovine Serum).
C2C12 were differentiated by seeding them at >90% confluence and switching to Differentiation Medium (DM; Dulbecco's Modified Eagle's Medium supplemented with 2 mM L-glutamine, 100 U/ml Penicillin, 100 ug/ml Streptomycin, 0.5% Fetal Bovine Serum).

5.1.1 siRNA treatment

To increase the knockdown efficiency, all siRNA transfections have been performed in reverse, by seeding the desired number of C2C12 cells directly of the transfection mix.
For a 3.5 cm culture dish, 5 ul of Lipofectamine RNAiMAX Reagent (Thermo Scientific) were added to 300 ul of Opti-MEM® I Reduced Serum Medium (Gibco); the siRNA was then added at a final concentration of 30 nM for a final volume of 2 ml. After 15 minutes of incubation time at room temperature, the transfection mix was distributed on the culture dish and 200.00 C2C12 cells were seeded in GM.
After 24 hours, cell confluency was checked and the culture medium was replaced with DM, in order to trigger the differentiation. Cells were then collected 48 hours after the induction of differentiation.
Scramble and DHX36 siRNA were purchased from Qiagen (AllStars Negative Control siRNA and Mm Dhx36-4 FlexiTube respectively), while lnc-G4 siRNAs were custom synthetized:

siRNA name	5'-> 3' sequence
siG4-1	GAGAAGCAGAACAAACAGATA
siG4-2	CAGAAACTCTCTGCTTTCGTA

5.1.2 Plasmid transfection for overexpression experiments

For plasmid transfection, 100.000 C2C12 were seeded in a 3.5 cm culture dish and left to proliferate for 24 hours. 1 ug of desired plasmid and 3 ul of Lipofectamine-2000 Reagent (Thermo Scientific) were combined in 300 ul of Opti-MEM® I Reduced Serum Medium (Gibco) and left in incubation for 15 minutes at room temperature. After GM medium replacement, transfection mix was added to the culture dish and left in incubation for 24 hours. Depending of the nature of the experiments, cells were collected in GM or they were differentiated in DM for the appropriate time.

5.2 RNA isolation and analysis

5.2.1 RNA purification

After discarding medium, cells were washed using PBS without Ca^{2+} and Mg^{2+}. 350 µl of TRI-Reagent (Zymo Research), a mono-phasic solution of phenol and guanidine isothiocyanate, were added detaching cells mechanically. Samples were homogenized by vigorous vortexing and then incubated for 5 minutes at room temperature to allow the complete dissociation of nucleoprotein complexes. RNA extraction was then performed with the Direct-zol Miniprep RNA Purification Kit (Zymo Research) according to the manufacturer's indications. After the purification procedure, RNA concentration was measured with NanoDrop 1000 Spectrophotometer.

5.2.2. RNA retrotranscription

For routine experiments, 500 ng of total RNA were retrotranscribed with PrimeScript™ RT Reagent Kit (Takara) according to the manufacturer's protocol, in a final reaction volume of 10 ul.
For low RNA input experiments, such as immunoprecipitation and pulldown, the Superscript VILO cDNA Synthesis Kit was used. In this case, each reaction was set up by mixing 7 ul of RNA sample, 2 ul of 5x VILO Reaction Mix and 1 ul of 10x SuperScript Enzyme Mix, for a final reaction volume of 10 ul. The incubation timing respected the manufacturer's protocol.

5.2.3 Semiquantitative RT-PCR

Semi-quantitative PCRs were performed using MyTaq (Bioline) enzyme (1%), cDNA (15-25 ng equivalent), 10x MyTaq Reaction Buffer, and 1 µl of each 10mM primers; PCR reactions were done for up to 35 cycles at 95°C for 25 s, 55°C for 25 s, 72°C for 25 s.
5 ul of final products were run on 2% w/v agarose gels containing 0.5 ug/ml of Ethidium Bromide and visualized on a GelDoc imager (BioRad).

5.2.4. Quantitative PCR

All quantitative RT-PCRs (qRT-PCRs) were performed in triplicate using PowerUp SYBR Green Mastermix (Thermo Scientific). For each reaction, 5 ng of cDNA equivalent and a final concentration of 300 nM for each primer was used. DNA amplification was obtained with the standard PCR cycling suggested by the manufacturer's protocol and it was monitored with an ABI 7500 Fast qPCR instrument.

5.2.5 RNA sequencing after lnc-G4 knockdown

Illumina TruSeq library preparation was performed by Istituto di Genomica Applicata (Udine, Italy) from 1 µg of total RNA depleted for ribosomal RNA with Ribominus Eukaryote Kit for RNA-seq

(Thermo Scientific). For each sample, 20-40 millions of 100bp long paired-end reads were sequenced. The reads were then preprocessed using Trimmomatic software to remove adapter sequences and to perform quality trimming. Using Bowtie2, the preprocessed reads were then aligned to a database composed of rRNAs in order to discard contaminants. TopHat2 was used to map unmapped reads to mouse genome (mm10 assembly) using the Gencode gene model annotation as a guide. BamTools was used to remove reads mapping to mitochondrial RNA. Read counts were calculated using HTSeq-count differential expression analysis was performed using edgeR software.

5.3 Protein isolation and analysis

5.3.1 Protein extraction

After discarding medium, cells were washed using PBS without Ca^{2+} and Mg^{2+}. 100 ul of standard RIPA buffer, supplied with 1x Complete Protease Inhibitor Cocktail (Roche), were added directly on the plate; cells were scraped and the recovered lysate was incubated on ice for 5 minutes. Cells were thoroughly lysed by rotation for 30 minutes at 4 °C, then cellular debris were pelleted by centrifugation at 13000 rpm, 10 minutes, 4 °C.
Protein concentration was assessed using Bradford Protein Assay (BioRad) at 595 nm, by comparison with a BSA standard curve.

5.3.2 Western Blot

Western Blot samples were prepared by mixing 15-25 ug of proteins with 4x Laemmli Sample Buffer (BioRad) and DTT at a final concentration of 50 mM. After incubation at 70°C for 10 minutes, samples were loaded on 4-15% Mini-PROTEAN TGX Precast acrylamide gel (BioRad) and protein electrophoresis was performed at 150 V in 1x Running Buffer (25 mM TRIS, 192 mM glycine, 0.1% SDS). Separated proteins were transferred on Immobilon-E PVDF 0.45 um membrane (Merck-Millipore) at 80 V for 1 hour in 1x Towbin Transfer Buffer (25 mM TRIS, 192 mM glycine, 20%

methanol). After Ponceau-S staining, membranes were blocked with 5% non-fat dry milk (Difco Skim Milk) for 1 hour and incubated overnight at 4 °C with primary antibody. Hybridized membranes were washed 3x10 minutes in Tris-buffered saline with Tween 20 (Sigma Aldrich) and then incubated with secondary antibody (diluted 1:10000 in 5% milk) for 1 hour at RT, and washed again 3x10 minutes in TBS-T. Protein detection was carried out with LiteAblot®EXTEND Long Lasting Chemiolominescent Substrate (EuroClone) using ChemiDocTM MP System and images were analysed using Image LabTM Software (BioRad).

5.4 Native RNA pulldown

C2C12 cells were plated in two 15 cm plates at a density of $2{,}5x10^6$ cells/plate and grown for 24 hours before switching to DM. After 48 hours of differentiation the cells were washed twice in complete PBS, scraped in 500 ul of Buffer A (TRIS-HCl pH 8 20 mM; NaCl 10 mM; MgCl2 3 mM; NP-40 0,1 %; glycerol 10%; EDTA 0,2 mM; Protease and RNase Inhibitor) and incubated on ice for 15 minutes. The cytoplasmic extract was recovered by centrifugation at 2000 rpm, 10 minutes, 4°C, and protein concentration was assessed with Bradford reagent assay. 1 mg of cytoplasmic extract was used for each pulldown reaction (Odd, Even and LacZ). Preclearing of the extract was performed with 50 ul of Streptavidin MagneSphere Paramagnetic Particles (Promega) equilibrated in Hybridization Buffer (Tris-HCl ph7.4 50mM; NaCl 150mM; MgCl2 1mM; NP-40 0,05%; EDTA 10mM; DTT 1mM; Protease and RNase Inhibitor) for 30 minutes in rotation at room temperature. 1 ul of a mix of four different biotinylated probes (25 uM each) was added to each reaction and incubated for 2 hours in rotation at room temperature.
The probes sequences are hereby listed:

Probe-set	Probe n°	Sequence
Probe-set 1	1	TAGCTAGCTCCAGTGACTAG
	4	TGGGTAAAGTGCTTGATGCA
	6	TGTGCCTTCATGTGGGGAAG
	8	GTCTTTCGAGGATCAAAGGC
Probe –set 2	3	CAGCAGTTAGGTTCCAATTG
	5	ACACGGGCATGGTATACAAC
	7	TAGGAAGCAAGACCGTCATC
	9	TACTGCTCTCATCATTTTGC

100 ul of Streptavidin MagneSphere Paramagnetic Particles (Promega) were added to each reaction and incubated for 30 minutes in rotation at room temperature. After the incubation, the Paramagnetic Particles were washed 4 times in 1 ml of Hybridization Buffer and then resuspended in 200 ul of Hybridization Buffer and divided in two aliquots for RNA and protein analysis.
RNA fraction was recovered by addition of 500 ul of QIAZOL Lysis Reagent directly of the Paramagnetic Particles and then extracted according to manufacturer protocol.
Protein fraction was supplemented with Laemmli Sample Buffer (Bio-Rad) and incubated for 5 minutes at 90 °C.

5.4.1 RNA preparation for sequencing

Total RNA after RNA pull-down was prepared with RNeasy Plus Minikit (Qiagen). TrueSeq library preparation for polyA+ RNAs was performed by Istituto di Genomica Applicata (Udine, Italy) from total enriched RNA of each pull-down. The same procedure was applied to the input RNA. Concentration of all RNA samples was unknown, but the same volume of each sample underwent the RNA-Seq process.

5.4.2 Protein preparation for mass spectrometry

To identify unknow protein partners of lnc-G4, we utilized protein mass spectrometry after purification of RNA-Protein complexes. The protein elutant of pull-downs was loaded to the 4-12% NuPAGE® Bis-Tris gels (Life technologies) and ran up to 2,5 cm into the gel, then the corresponding gele parts to each sample was cuted and sent to IGBMC Plateforme de Protéomique (Strasbourg, France). There, the samples were prepared the samples by manual gel digestion procedure using Trypsin (Promega) and performed mass spectrometry using LTQ Velos Pro instrument with analyze method of TOP20CID, 2h gradient, col C18 15cm. To analyze the results, two different levels of filteration were applied to the samples:
1. Stringent filteration: 1% FDR (False Discovery Rate) and 2 peptides minimum per protein.
2. Flexible filteration : 5% FDR and 1 peptide per protein.
Due to orbitrap troubleshooting, samples were analyzed on the LTQ-Velos Pro system.

5.5 Psoralen-crosslinked RNA pulldown

The protocol was adapted from RICC-seq protocol (Gorbea et al. 2017). C2C12 cells were plated in two 10 cm plates at a density of $1x10^6$ cell/plate and grown for 24 hours before switching to DM. After 48 hours of differentiation the medium was replaced with fresh DM supplemented with 20 ug/ml of 4'-aminomethyl-4,5',8-trimethylpsoralen (AMT, Cayman Chemical) and cells were incubated 30 minutes at 37 °C. After incubation, the cells were washed twice with complete PBS, covered with PBS supplemented with 150 ug/ml AMT and crosslinked at 365 nm at 10 minute intervals for 1 hour. Cells were lysed in 500 ul of Guanidinium Hydrochloride 3M and the lysate was subdivided in 150 ul aliquots; 25 µL of a 20 mg/mL solution of RNase-free proteinase K (Ambion) and 7,5 µL of 20% sodium dodecyl sulfate (SDS) were added to each aliquot, and the samples were incubated at 65°C for one hour with

gentle agitation. RNA was precipitated with conventional phenol/clorophorm extraction; the rest of pulldown protocol was performed as described in Gorbea et al., with minor modifications; the probes used for psoralen pulldown are the same of the native protocol.

5.6 DHX36 immunoprecipitation

Two 10 cm plates of day 2-differentiated C2C12 cells were scraped in PLB Buffer (KCl 100 mM; $MgCl_2$ 5 mM; NP-40 0,5%; DTT 1mM; Protease and RNase inhibitor), collected in a microtube and lised for 15 minutes in rotation at 4°C, then centrifuged at 13000 rpm, 10 minutes, 4°C to remove cellular debris. The supernatant was recovered and protein concentration was quantified by Bradford assay; 0,5 mg of extract was used for each sample (IP and IgG) and was precleared with 40 ul of Protein G Agarose/Salmon Sperm Beads (Millipore) in a final volume of 1 ml of NT2 buffer (TRIS-HCl pH 7.4 50 mM; NaCl 150 mM; $MgCl_2$ 1 mM; NP-40 0,05%; Protease and RNase Inhibitor) in incubation for 2 hours in rotation at 4°C. 10% of the final volume was taken as Input and the remaining precleared lysate was incubated with 5 ug of DHX36 (Rabbit Polyclonal, Proteintech) or rabbit IgG (SantaCruz) antibodies overnight at 4 °C. Subsequently, 80 ul of Protein G agarose beads were added to each sample and incubated for 2 hours at 4 °C. After antibody-protein complex recovery, the beads were washed 4x in NT2 buffer and finally resuspended in 200 ul of NT2 Buffer. To check DHX36 immunoprecipitation efficiency, an aliquot of 50 ul of beads was pelleted, resuspended in 1x Laemmli Sample Buffer (BioRad) and 50 mM DTT and incubated at 70°C for 14 minutes, then the eluate was used for Western Blot.
RNA was recovered from the remaining 150 ul of beads by resuspending them in 500 ul of TRI-Reagent (Zymo Research) and thoroughly vortexing. RNA extraction was carried out with RNeasy Plus Minikit (Qiagen) according to manufacturer's protocol.

5.7 C2C12 Immunostaining

5.7.1 Myosin heavy Chain immunostaining

C2C12 cells grown and differentiated in 3.5 cm culture dishes were washed twice in 1 ml PBS, fixed with in 1,5 ml of 4% paraformaldehyde (PFA) for 15 min at RT and washed twice with 1 ml of PBS. Cells were permeabilized and blocked in 1 ml of PBS with 0.2% Triton X-100 and 1% goat serum. Samples were then stained with in-house made anti-Myosin heavy chain antibody overnight at 4°C, and then with mouse Cy3/AlexaFluor488 secondary antibody for 30 min at RT. Nuclei were labeled with DAPI (4',6-diamidino-2-phenylindole). Acquisition for image-based phenotypic analysis was performed with a widefield Nikon TiE Microscope equipped with a Lumencor SpectraX LED light source and PerfectFocus 3; 4× and 10× objectives were used (Nikon Instruments).

5.7.2 MLX immunostaining

C2C12 cells were cultured on pre-coated glass coverslips (300 ug/ml Collagen Rat Tail, Corning, in PBS) and then were fixed in 4% paraformaldehyde (Electron Microscopy Sciences) in PBS at 4°C for 20 min. Cells were permeabilized with Triton 0.2% for 10 minutes, blocked with 2% BSA/PBS for 20 minutes e and subsequently incubated at 4 °C overnight with anti-MLX (ProteinTech) diluted 1:50 in blocking solution.
After serial washes in 0.1% Triton /PBS, coverslips were incubated in 1%Goat serum/1%Donkey serum/PBS with goat anti-rabbit Cy3 conjugated antibody (1:300; Jackson ImmunoResearch) to detect MLX signal. The incubation was performed for 1 hour at RT. The specificity of immunolabeling was verified in control samples prepared with the incubation buffer alone, followed by the secondary conjugated antibody. The nuclei were stained with DAPI. The images were taken on inverted microscope (Olympus IX73) equipped with a Confocal Imager (CREST X-LIGHT) spinning disk, a CoolSNAP Myo CCD camera (Photometrics) and a

Lumencor Spectra X LED illumination. The acquisitions were performed with 60x NA 1.35 oil-objective (UPLANSApo) and MetaMorph software (Molecular Devices).
The N/C ratio (signal intensity inside nucleus/signal intensity inside cytoplasm) was obtained on 16 bit depth images with a resolution XY of 0,075 micron and with 0,2 micron Z-spacing.
For the quantification of the fluorescence only 1 single optical section was considered a specific range of Intensity balance was manually determined by using MetaMorph or FIJI softwares.
The analysis was taken only in MHC-positive cells by quantifying the mean fluorescence inside composite selections (ROI) delimitating nuclei edges (marked with DAPI staining) and cytoplasm edges (marked with MHC staining) respectively. The quantification of the fluorescence was taken by using FIJI software from about 50 cells on each condition, and the value of the mean N/C ratio were represented as mean±SEM of duplicates.
The images shown in Figure 13 and 14 are a qualitative display of Immunofluorescence analysis to visualize the distribution of MLX signal inside the cells. They are binary images obtained by Threshold processing.
To marks the signal peaks of MLX signals inside the nuclei, a local maxima filter in FIJI sofware was applied.

5.8 Luciferase assay

5.8.1 Constructs generation

For lnc-G4 overexpression, the full sequence of mature lnc-G4 was amplified from C2C12 cDNA and inserted in PCDNA 3.1 (-) expression plasmid using XhoI and NotI Fast restriction enzymes (Thermo Scientific). The “shuffle G4” mutant was generated by inverse PCR with divergent primers.
For MLX-RLuc “only 5’” construct, the full length exon 1 of MLX γ was amplified from C2C12 cDNA. Since the interaction with lnc-G4 is in the coding sequences, we generated a similar system by fusing MLX γ exon 1 with the *Renilla* luciferase coding sequence of

Psicheck2 vector (Promega), in which the ATG start codon had been removed by inverse PCR. We used the In-Fusion HD Cloning Kit (Clontech) to generate the construct in order to maintain the reading frame and to have a fusion protein under the endogenous MLX start codon.
For the "5'+3' " and "only 3'" versions of the plasmid, MLX 3'UTR sequence was amplified form C2C12 cDNA and inserted in Psi-check2 vector (Promega) using XhoI and NotI Fast restriction enzymes (Thermo Scientific).
For the Δ-mutant, a deletion of 75 bp was obtained by inverse PCR with divergent primers from the full-length "only 5'" construct.
The "shuffle MLX" mutant was generated by inverse PCR with divergent primers.

5.8.2 Luciferase assay

All the luciferase experiments were performed using the Dual-Luciferase Reporter Assay System (Promega). 10.000 C2C12 were plated in each well of 12-well multiplates. After 24 hours, cells were transfected with the already described combination of enzymes in the following amounts:

- 20 ng of MLX-Rluc plasmid or its mutants per well
- 250 ng of lnc-G4 expression plasmid or empty vector control per well

Transfection mix was prepared in 200 ul of Opti-MEM® I Reduced Serum Medium (Gibco) with 1,5 ul of Lipofectamine-2000 Reagent (Thermo Scientific) per well. Transfection mix was added to fresh GM medium for a final volume of 1 ml for each well, and left in incubation overnight at 37 °C. The day after, the transfection medium was replaced with fresh GM medium and cells were left in growth for other 24 hours. Cells were washed 2x in PBS and lysed in 100 ul of 1x Passive Lysis Buffer (Promega) supplemented with 1x Protease Inhibitor Cocktail and RNase Inhibitor, for 15 minutes at room temperature on a rocking platform. The lysates were cleared by centrifugation for 2 minutes at 13000 rpm, 4 °C. Luciferase assay was performed in triplicate on 5 µl of lysate with Dual-Luciferase® Reporter Assay System (Promega), according to manufacturers'

instructions; the machine that was used is GloMax®- Multi Detection System (Promega).

5.9 In-gel G-quadruplex staining

RNA oligonucleotide corresponding to MLX γ G-quadruplex sequence and lnc-G4 interacting sequences were customly synthetized (Sigma-Aldrich) with the following sequences:

Name	Sequence
MLX	GCCGGCGAGGGGAGGGCGGGUCGGGUCGUGCACGUAGGGGGUCCG
lnc-G4	GCUGGCUCCCGCAGGCCUCUCCUUUCCCUGC

Synthetic RNAs were resuspended in Resuspension Buffer (20 mM TRIS pH 7.6, 1mM EDTA) at a stock concentration of 200 uM.
For quadruplex formation assay, 1 ul of MLX and lnc-G4 oligonucleotides stocks (for a total amount of 200 pmol) were diluted in a final volume of 10 ul in presence of 100 uM KCl favour G-quadruplex formation and stabilization. For total RNA control, 5 pmol of each oligonucleotide was considered.
Diluted RNAs were heated to 95 °C for 5 minutes, then allowed to passively cool to RT. RNAs were then loaded on native TRIS-Borate EDTA polyacrylamide gels for electrophoretic separation.
For G-quadruplex staining, the gel was immersed in 30 ml N-methyl mesoporphyrin IX Staining Solution (1ug/ml of NMM in 20 mM TRIS pH 7.6, 100 mM KCl, 1 mM EDTA) and incubated for 15 minutes at RT on a rocking platform. The gel was visualized on a Pharos Imager (BioRad) with the Cy3 excitation and emission filters. For total RNA control, the gel was immersed in 30 ml of TBE containing 1x SYBR Gold stain (Thermo Scientific) and incubated for 30 minutes at RT on a rocking platform. The gel was visualized on a GelDoc Imager (BioRad).

5.10 Ribosome Profiling

$1x10^6$ C2C12 cells were reverse transfected with scramble and linc-G4 siRNA as previously described. After 48 hours of differentiation, cells were washed once with cold PBS, and then washed rapidly with cold PBS supplemented with Cicloheximide (100 ug/ml) in order to block translation. Culture dishes were snap-frozen in liquid nitrogen, then moved to dry ice were 150 ul of Lysis Buffer (20 mM TRIS-HCl pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 100 ug/ml cicloheximide, 1% TRITON X-100, 25 U/ml Turbo DNase (Thermo Scientific)) were added to each plate. Cell were scraped and lysates were collected and cleared by centrifuging at 20.000g, 5 minutes, 4 °C.
RNase I digestion was performed by adding 6 ul of RNase I (Ambion, 100 U/ul) to 240 ul of lysate and incubating the mix for 45 minute at RT in slow agitation; the digestion was stopped with 8 ul of SUPERase Inhibitor (Thermo Scientific, 20 U/ul) and the monosomes were recovered by size exclusion chromatography using Illustra MicroSpin Columns S-400 HR (GE Life Sciences), previously equilibrated with the Lysis Buffer, according to manufacturer's protocol.
Recovered samples were mixed with 3 volumes of TRIzol LS (Thermo Scientific) and extracted with Zymo Spin and Concentrator Kit (Zymo Research) according to manufacturer's protocol.
For ribosomal RNA depletion, 5 ug of recovered RNA were treated with Ribo-Zero Gold Kit (Illumina) according to manufacturer's protocol; the recovered RNA was loaded on a 17% TBE/Urea gel in denaturing condition, together with 27nt and 30nt control oligonucleotide, and it was run at 30 W for 90 minutes. The gel was stained with 1x SYBR Gold in TBE, and the band including RNA footprints between 27 and 30 nt was excised. The acrylamide slice was cut in small pieces and RNA was eluted with 350 ul 0.3M NaCl overnight, at 4 °C, shaking at 1200 rpm.
All the other steps of library preparation were performed as described in Ingolia et al. 2009.

5.11 Appendix tables

Table 4 – List of primers used for this thesis work

Primer name	Sequence
lnc-049 Fw	CTAGCAGCAGCACTCACAGC
lnc-049 Rv	CTTGTGCCTCGTTGACAAAA
pre-GAPDH Fw	GTATGTATGGGGAGAGCTGG
GAPDH Fw	TGACGTGCCGCCTGGAGAAA
GAPDH Rv	AGTGTAGCCCAAGATGCCCTTCAG
MYOD Fw	GCAGAATGGCTACGACACC
MYOD Rv	CACTATGCTGGACAGGCAGT
Myogenin Fw	TCCCAACCCAGGAGATCATT
Myogenin Rv	CATATCCTCCACCGTGATGC
MEF2C Fw	ACGAGGATAATGGATGAGCGT
MEF2C Rv	TCACAGTCGCACAGCACG
MCK Fw	TTACACTCTGCCTCCGCACT
MCK Rv	GTACTTGCCCTTGAACTCGC
DMD Fw	AGCTCAACCGTCGATTTGCAGC
DMD Rv	TTCAGCCTCCAGTGGTTCAAGC
LRRN1 Fw	GAAGATCGACAACCCCCACA
LRRN1 Rv	GGACACCGTGAGACACACTT
TNNC2 Fw	AGCGAAGAGGAACTGGCTGAGT
TNNC2 Rv	CGATCTCCTCTTCTGTCACATGC
CRABP2 Fw	GTGGATGGGAGACCCTGTAAG
CRABP2 Rv	TCATTGGTCAGTTCTCGGCTC
CCL2 Fw	CTCACCTGCTGCTACTCATTCAG
CCL2 Rv	GTCAGCACAGACCTCTCTCTTG
CCL7 Fw	CATCCACATGCTGCTATGTC
CCL7 Rv	GCAGACTTCCATGCCCTTCT

Primer Name	Sequence
CXCL5 Fw	CTTGTCCACAATGAGCCTCC
CXCL5 Rv	TTGCGGCTATGACTGAGGAAGGG
CD63 Fw	TGCTCTACGTTCTCCTGCTG
CD63 Rv	GCAATGATGACCACAGGCAA
WBP4 Fw	CTGGGCCAATTCTTTTCTCAGT
WBP4 Rv	CGAAGGGGAACAGAAGAAAGC
Rps7 Fw	AAGCACAGCAGAACAACGTG
Rps7 Rv	ACCCATGGCCCTGAGATTTT
MLX ex4 Fw	GGAGTCCTACAAAGACCGGA
MLX ex5 Rv	GGCAGGTAGGGACAATGGT
MLXγ ex1 Fw	GCAGAGAAGACAGCTCTCACC
MLX ex2 Rv	GCTGTTGTCACTGTAGGCAT
Firefly LUC Fw	TGCAGAAGATCCTGAACGTG
Firefly LUC Rv	CGGTAGACCCAGAGCTGTTC
Renilla LUC Fw	TCGTCCATGCTGAGAGTGTC
Renilla LUC Rv	CTAACCTCGCCCTTCTCCTT
DHX36 Fw	GCTGAGCATCTTCTTGGAGC
DHX36 Rv	AAACCAGCACAGATGACAGC

Table 5 – List of primers used for plasmid generation

Primer Name	Sequence
3'UTR Mlx XhoI Fw	AAAACTCGAGGCAGAGCAGCCAACAAGAG
3'UTR Mlx NotI Rv	AAAAGCGGCCGCCTGGGGAAGGCAGTAGGAA
psiCHECK2 ΔATG Fw	GCTTCCAAGGTGTACGACCC
psiCHECK2 ΔATG Rv	GGTGGCTAGCCTATAGTGAGTC
luc-Mlx 5' Fw	TATAGGCTAGCCACCTGACTGAGCCTCGCCTCTTCCT
luc-Mlx 5' Rv	GTACACCTTGGAAGCGTTGTCACTGTAGGCATACTCGAC
luc-Mlx D75 Fw	TCCCCAACTATCCCCAGAGTC
luc-Mlx D75 Rv	GGGAGACTCTGGGGATAGTT
luc-Mlx mut30 Fw	CCGCAGGCCTCTCCTTTAGGGGGTCCGGAAGACG
luc-Mlx mut30 Rv	GAGCCAGCTGTGCACGGCCGGCGTGCGCGTC
lnc-049 XhoI Fw	AAAACTCGAGACTAGTCACTGGAGCTAGC
lnc-049 NotI Rv	AAAAGCGGCCGCGGGCATGGTATAACAACGT
lnc-049 mut Fw	CGGGCTCGTGCACGTCCCTGCATAGCCAATTATTCC
lnc-049 mut Rv	ACCCGCCCTCCCCTCGATGGATGAGTGGGGA

6. GLOSSARY

Basal lamina: a gel-like layer of extracellular matrix secreted by the epithelial cells, on which the epithelium sits. Basal lamina consists of collagen and a variety of glycoproteins; it helps to connect the epithelial cells to the connective tissue layer and strengthen the epithelial tissues.

C2C12: a mouse skeletal myoblast cell line produced in Helen Blau's Lab in Stanford, capable of differentiation to myotubes *in vitro* by serum starvation; they are a useful tool for genetic manipulation to study the molecular pathways underlying the differentiation of myoblasts and osteoblasts.

ceRNA: competing endogenous RNA; a RNA molecule that shares MRE with other transcripts and regulates the expression of those transcripts by competing for the similar miRNA pool.

circRNA: circular RNA; covalently closed circular RNA molecule that are produced from linear precursors by the spliceosome trough a back-splicing reaction, during this reaction the donor splice site is joined to an acceptor splice site located upstream of the donor.

Dermomyotome: is an epithelial cell layer constituting of the dorsal part of the somite lying under the ectoderm.

E proteins: E proteins s are widely expressed transcriptional regulators with very general functions. They are heterodimer partners of the tissue-specific basic Helix-Loop-Helix (bHLH) proteins that are sometimes called class II bHLH proteins.

E-box: An E-box (enhancer box) is a DNA response element found in some eukaryotes that acts as a protein-binding site and has been found to regulate gene expression in neurons, muscles, and other tissues. Its specific DNA sequence, CANNTG (where N can be any

nucleotide), with a palindromic canonical sequence of CACGTG, is recognized and bound by transcription factors to initiate gene transcription. Once the transcription factors bind to the promoters through the E-box, other regulators can bind to the promoter and facilitate transcription.

G-quadruplex: peculiar secondary structure of guanine-rich nucleic acids, both DNA and RNA; separated G-stretches interact through Hoogsteen pairing forming planar tetrads that can stack and generate a stable structure with several regulatory properties.

LINE element: long interspersed nuclear elements are a group of non-LTR (long terminal repeat) retrotransposons which are widespread in the genome of many eukaryotes. Due to the accumulation of random mutations, the sequence of many LINEs has degenerated to the extent that they are no longer transcribed or translated.

***mdx* mouse:** a mouse model widely used for studying Duchenne muscular dystrophy. There is a point mutation in its DMD gene, changing the amino acid coding for a glutamine to a threonine. The muscle cells of mdx mouse produce a small, non-functional dystrophin protein causing a mild form of DMD.

Mesoderm: In all bilaterian animals, the mesoderm is one of the three primary germ layers in the very early embryo. Some of the mesoderm derivatives include the muscle (smooth, cardiac and skeletal), the muscles of the tongue (occipital somites), the pharyngeal arches muscle (muscles of mastication, muscles of facial expressions), connective tissue, dermis and subcutaneous layer of the skin, bone and cartilage, dura mater, endothelium of blood vessels, red blood cells, white blood cells, and microglia, the kidneys and the adrenal cortex.

Molecular evolution: is the process of change in the sequence composition of cellular molecules such as DNA, RNA, and proteins across generations. Molecular evolution is the motor of the origin of non-coding RNAs from previously non-functional genomic regions or following genomic rearrangement events.

MRFs: Myogenic regulatory factors (MRF) are basic helix-loop-helix (bHLH) transcription factors that regulate myogenesis: MyoD, Myf5, myogenin, and MRF4. These proteins contain a conserved basic DNA binding domain that binds the E box DNA motif.

Myoblast: the precursor of a muscle cell; an undifferentiated cell in the mesoderm of the vertebrate embryo.

Myofiber: a long multinucleated single muscle cell formed by differentiation and fusion of myoblasts through myogenesis.

Myogenesis: the formation of muscular tissue during embryonic development or upon injury in adult tissues. During myogenesis muscle fibers would form from the fusion of myoblasts into multi-nucleated fibers called myotubes that will give raise to myofibers.

Myotome: the dorsal part of each somite in a vertebrate embryo that will rise to the skeletal muscle.

Ribosome Profiling: high-throughput technique, developed by Nicholas Ingolia, in which total RNA is digested with RNAse I and the Ribosome Protected Fragments of mRNAs are recovered and used for library preparation and deep sequencing. It allows a deep insight into translation dynamics of a particular system.

RNA helicase: motor enzymes that move directionally along a nucleic acid phosphodiester backbone, separating two annealed RNA strands using energy derived from ATP hydrolysis.

RNA pulldown: molecular technique in which a specific RNA molecule of interest is precipitated with biotinylated probes, in order to analyse and identificate its interacting partners. Psoralen-crosslinked variant of RNA pulldown is specific for the detection of RNA-RNA interactions.

RNA-seq: RNA sequencing; a technique to identify and quantify RNA molecules in biological samples at a given moment in time using next-generation sequencing.

Sarcolemma: also called the myolemma, is the cell membrane of a striated muscle fiber cell.

Satellite cells: precursors of adult skeletal muscles residing in a satellite position around myofibers. They are mitotically quiescent and have the potential to activate upon an external stimuli and provide additional myofibers to their parent muscle fiber or return to a quiescent state.

Translational regulation: level of regulation of gene expression that impacts on the rate of translation of specific mRNA in order to regulate the amount of produced protein; differently from mRNA stability regulation, acting on translation is a way to have quick modulatory response without degrading the mRNA.

7. BIBLIOGRAPHY

Anderson, D.M. et al., 2015. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, 160(4), pp.595–606.

Arthanari, H. et al., 1998. Fluorescent dyes specific for quadruplex DNA. , 26(16), pp.3724–3728.

Ballarino, M. et al., 2018. Deficiency in the nuclear long noncoding RNA *Charme* causes myogenic defects and heart remodeling in mice. *The EMBO Journal*, 37(18), p.e99697.

Ballarino, M. et al., 2015. Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. *Molecular and cellular biology*, 35(4), pp.728–36.

Bernard, D. et al., 2010. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO Journal*, 29(18), pp.3082–3093.

Bi, P. et al., 2017. Control of muscle formation by the fusogenic micropeptide myomixer. *Science*, 356(6335), pp.323–327.

Booy, E.P. et al., 2018. The RNA helicase RHAU (DHX36) unwinds a G4-quadruplex in human telomerase RNA and promotes the formation of the P1 helix template boundary. , 40(9), pp.4110–4124.

Britten, R.J. & Davidson, E.H., 1969. Gene regulation for higher cells: A theory. *Science*, 165(3891), pp.349–357.

Buckingham, M. & Relaix, F., 2007. The Role of *Pax* Genes in the Development of Tissues and Organs: *Pax3* and *Pax7* Regulate Muscle Progenitor Cell Functions. *Annual Review of Cell and Developmental Biology*, 23(1), pp.645–673.

Buckingham, M. & Rigby, P.W.J., 2014. Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis. *Developmental Cell*, 28(3), pp.225–238.

Buckingham, M. & Rigby, P.W.J., 2014. Gene Regulatory Networks and Transcriptional Mechanisms that Control Myogenesis. *Developmental Cell*, 28(3), pp.225–238.

Cabili, M. et al., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development*, 25(18), pp.1915–1927.

Cacchiarelli, D. et al., 2011. MiR-31 modulates dystrophin expression: New implications for Duchenne muscular dystrophy therapy. *EMBO Reports*, 12(2), pp.136–141.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A.R.R., et al., 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, 491(7424), pp.454–7.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A.R.R., et al., 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, 491(7424), pp.454–457.

Cesana, M. et al., 2011. A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. *Cell*, 147(2), pp.358–369.

Chen, C.K. et al., 2016. Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science*, 354(6311), pp.468–472.

Chen, J.L. & Greider, C.W., 2004. Telomerase RNA structure and function: Implications for dyskeratosis congenita. *Trends in Biochemical Sciences*, 29(4), pp.183–192.

Chu, C., Quinn, J. & Chang, H.Y., 2012. Chromatin Isolation by RNA Purification (ChIRP). *Journal of Visualized Experiments*, (31), p.e3912.

Collins, K., 2008. Physiological assembly and activity of human telomerase complexes. *Mechanisms of Ageing and Development*, 129(1–2), pp.91–98.

Dimartino, D. et al., 2018. The Long Non-coding RNA lnc-31 Interacts with Rock1 mRNA and Mediates Its YB-1-Dependent Translation. *Cell Reports*, 23(3), pp.733–740.

Diolaiti, D. et al., 2015. Functional interactions among members of the MAX and MLX transcriptional network during oncogenesis. *BBA - Gene Regulatory Mechanisms*, 1849(5), pp.484–500.

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Engreitz, J., Lander, E.S. & Guttman, M., 2015. RNA antisense purification (RAP) for mapping RNA interactions with chromatin. *Methods in Molecular Biology*, 1262, pp.183–197.

Faghihi, M.A. et al., 2010. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biology*, 11(5), pp.1–13.

Faghihi, M.A. et al., 2008. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nature Medicine*, 14(7), pp.723–730.

Forcales, S. V. et al., 2012. Signal-dependent incorporation of MyoD-BAF60c into Brg1-based SWI/SNF chromatin-remodelling complex. *EMBO Journal*, 31(2), pp.301–316.

Garber, M. et al., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), pp.469–477.

Gilbert, W., 1986. Origin of life - The RNA world. *Nature*, 319, p.618.

Gong, C. & Maquat, L.E., 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature*, 470(7333), pp.284–288.

Gorbea, C., Mosbruger, T. & Cazalla, D., 2017. A viral Sm-class RNA base-pairs with mRNAs & recruits microRNAs to inhibit apoptosis. *Nature*, 550(7675), pp.275–279.

Griffin, C.A. et al., 2010. Chemokine expression and control of muscle cell migration during myogenesis. *Journal of Cell Science*, 123(18), pp.3052–3060.

Gumireddy, K. et al., 2013. Identification of a long non-coding RNA-associated RNP complex regulating metastasis at the translational step. *The EMBO journal*, 32(20), pp.2672–84.

Guttman, M. et al., 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5), pp.503–510.

Guttman, M. et al., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235), pp.223–7.

Guttman, M. & Rinn, J.L., 2012. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385), pp.339–346.

Hansen, T.B. et al., 2013. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441), pp.384–388.

Van Heesch, S. et al., 2014. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biology*, 15(1), pp.1–12.

Henderson, E. et al., 1987. Telomeric DNA Oligonucleotides Form Novel Intramolecular StrucWes Containing Guanine-Guanine Base Paiirs. *Cell*, 51, pp.899–908.

Horak, M., Novak, J. & Bienertova-Vasku, J., 2016. Muscle-specific microRNAs in skeletal muscle development. *Developmental Biology*, 410(1), pp.1–13.

Hu, P. et al., 2008. Codependent Activators Direct Myoblast-Specific MyoD Transcription. *Developmental Cell*, 15(4), pp.534–546.

Huarte, M. et al., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3), pp.409–419.

Hunt, L.C. et al., 2015. The glucose-sensing transcription factor MLX promotes myogenesis via myokine signaling. *Genes & development*, 29(23), pp.2475–89.

Ingolia, N.T. et al., 2009. Supporting material for : Genome-wide in vivo analysis of translation with sub-codon resolution by ribosome profiling. *Science (New York, N.Y.)*, 324(5924), pp.218–224.

Jacob, F. & Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3), pp.318–356.

Jensen, T.H., Jacquier, A. & Libri, D., 2013. Dealing with pervasive transcription. *Molecular Cell*, 52(4), pp.473–484.

Johnsson, P. et al., 2013. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature Structural and Molecular Biology*, 20(4), pp.440–446.

Kallen, A.N. et al., 2013. The Imprinted H19 LncRNA Antagonizes Let-7 MicroRNAs. *Molecular Cell*, 52(1), pp.101–112.

Kapranov, P. et al., 2002. Large-Scale Transcriptional Activity in Chromosomes 21 and 22. *Science*, 296(5569), pp.916–919.

Kikin, O., Antonio, L.D. & Bagga, P.S., 2006. QGRS Mapper : a web-based server for predicting G-quadruplexes in nucleotide sequences. , 34(21), pp.676–682.

Kino, T. et al., 2010. Noncoding RNA Gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Science Signaling*, 3(107), pp.1–16.

Kondo, T. et al., 2010. Small Peptides Switch the. *Science*, 329, pp.336–339.

Krawczyk, M. & Emerson, B.M., 2014. p50-associated COX-2 extragenic RNA (PACER) activates COX-2 gene expression by occluding repressive NF-κB complexes. *eLIFE*, 3, pp.1–21.

Kretz, M. et al., 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, 493(7431), pp.231–5.

Kruger, K. et al., 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31, pp.147–157.

Kumar, D. et al., 2009. Id3 Is a Direct Transcriptional Target of Pax7 in Quiescent Satellite Cells. *Molecular Biology of the Cell*, 20, pp.3170–3177.

Kutter, C. et al., 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genetics*, 8(7), p.e1002841.

Lattmann, S. et al., 2011. The DEAH-box RNA helicase RHAU binds an intramolecular RNA G-quadruplex in TERC and associates with telomerase holoenzyme. , 39(21), pp.9390–9404.

Lee, R.C., Feinbaum, R.L. & Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), pp.843–854.

Legnini, I. et al., 2014. A Feedforward Regulatory Loop between HuR and the Long Noncoding RNA linc-MD1 Controls Early Phases of Myogenesis. *Molecular Cell*, 53(3), pp.506–514.

Liu, X. et al., 2012. Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. *The EMBO journal*, 31(23), pp.4415–27.

Ma, L., Bajic, V.B. & Zhang, Z., 2013. On the classification of long non-coding RNAs. *RNA Biology*, 10(6), pp.925–933.

Mann, M., Wright, P.R. & Backofen, R., 2017. IntaRNA 2 . 0 : enhanced and customizable prediction of RNA – RNA interactions. *Nucleic Acids Research*, 45, pp.435–439.

Mao, Y.S. et al., 2011. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nature Cell Biology*, 13(1), pp.95–101.

Marcel, V. et al., 2011. G-quadruplex structures in TP53 intron 3: Role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesis*, 32(3), pp.271–278.

McKinnell, I.W. et al., 2008. Pax7 activates myogenic genes by recruitment of a histone methyltransferase complex. *Nature Cell Biology*, 10(1), pp.77–84.

Meroni, G. et al., 2000. Mlx, a new Max-like bHLHZip family member: the center stage of a novel transcription factors regulatory pathway? *Oncogene*, 19(29), pp.3266–3277.

Millevoi, S., Moine, H. & Vagner, S., 2012. G-quadruplexes in RNA biology. *WIREs RNA*, 3(4), pp.495–507.

Molkentin, J.D. et al., 1995. Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. , 63, pp.1125–1136.

Moncaut, N., Rigby, P.W.J. & Carvajal, J.J., 2013. Dial M(RF) for myogenesis. *FEBS Journal*, 280(17), pp.3980–3990.

Morris, M.J. et al., 2010. An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES. *Journal of the American Chemical Society*, 132(50), pp.17831–17839.

Nelson, B.R. et al., 2016. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, 351(6270), pp.271–275.

Nie, J. et al., 2015. Post-transcriptional Regulation of Nkx2-5 by RHAU in Heart Development. *Cell Reports*, 13(4), pp.723–732.

Rinn, J.L. & Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81, pp.145–166.

Rion, N. & Rüegg, M.A., 2017. LncRNA-encoded peptides: More than translational noise? *Cell Research*, 27(5), pp.604–605.

Saccà, B., Lacroix, L. & Mergny, J., 2005. The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Research*, 33(4), pp.1182–1192.

Scarola, M. et al., 2015. Epigenetic silencing of Oct4 by a complex containing SUV39H1 and Oct4 pseudogene lncRNA. *Nature Communications*, 6, p.e7631.

Spitale, R.C., Tsai, M.C. & Chang, H.Y., 2011. RNA templating the epigenome: Long noncoding RNAs as molecular scaffolds. *Epigenetics*, 6(5), pp.539–543.

Tay, Y., Rinn, J. & Pandolfi, P.P., 2014. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483), pp.344–352.

Todd, A.K., Johnston, M. & Neidle, S., 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Research*, 33(9), pp.2901–2907.

Ulitsky, I. & Bartel, D.P., 2013. LincRNAs: Genomics, evolution, and mechanisms. *Cell*, 154(1), pp.26–46.

Uszczynska-Ratajczak, B. et al., 2018. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics*, 19(9), pp.535–548.

Wang, J., Gong, C. & Maquat, L.E., 2013. Control of myogenesis by rodent SINE-containing lncRNAs. *Genes and Development*, 27(7), pp.793–804.

Wang, K.C. & Chang, H.Y., 2011. Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell*, 43(6), pp.904–914.

Wang, Y. et al., 2013. Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal. *Developmental Cell*, 25(1), pp.69–80.

Weintraub, H. et al., 1991. Muscle-specific transcriptional activation by MyoD. *Genes and Development*, 5(8), pp.1377–1386.

Yang, F. et al., 2014. Reciprocal Regulation of HIF-1α and LincRNA-p21 Modulates the Warburg Effect. *Molecular Cell*, 53(1), pp.88–100.

Yoon, J.H. et al., 2012. LincRNA-p21 Suppresses Target mRNA Translation. *Molecular Cell*, 47(4), pp.648–655.

8. LIST OF PUBLICATIONS

- Dimartino, Colantoni, Ballarino, Martone, Mariani, Danner, Bruckmann, Meister, Morlando and Bozzoni, "The long non-coding RNA *linc-31* interacts with *Rock1* mRNA and mediates its YB-1 dependent translation", Cell Reports, Volume 23, Issue 3, P733-740, April 17, 2018

- Mariani, Martone, Setti, Capparelli, Shamloo, Colantoni, Dimartino, Morlando, Santini, Bozzoni, "A long non-coding RNA regulates skeletal muscle differentiation through translational repression of G-quadruplex containing mRNAs", manuscript in preparation

9. ACKNOWLEDGEMENTS

Prima di ringraziare chi ha condiviso con me questi anni, voglio ricordare chi purtroppo non ha potuto farlo. Sono sicuro che mio padre avrebbe seguito la discussione di questa tesi di dottorato indossando una delle tanto odiate cravatte, ma con quella sua espressione seria e al tempo stesso orgogliosa che mi è rimasta nel cuore. A lui dedico queste pagine, e con esse tutto il percorso che ha portato alla loro scrittura.

Ringrazio la professoressa Irene Bozzoni, che mi ha dato la possibilità di essere parte di un gruppo di ricerca così affiatato e capace di stare al passo con i tempi nella non sempre facile situazione della ricerca scientifica italiana.
Un ringraziamento particolare e doveroso va alla dottoressa Julie Martone, colei che in questi anni mi ha insegnato tutto ciò che conosco di questo mestiere, a partire dalle prime PCR della mia vita. Grazie per la pazienza, le risate, gli screzi, le merende e ogni momento di gioia e fallimento che abbiamo condiviso e condivideremo (perché puoi stare sicura che non ti libererai facilmente di me).
Grazie a Adriano, Francesca, Michela, Francesco e alle *new entry* Silvia e Flavia: è vero che quando siamo al completo Lab9 è un po'sovraffollato, però considero un privilegio avere voi come compagni di viaggio. Grazie a Gaia, l'amica un po' pazza che tutti dovrebbero avere a portata di scrivania, per come da anni ci supportiamo e sopportiamo a vicenda. Grazie anche a tutti i membri passati e presenti dei laboratori Bozzoni, perché è grazie a loro se queste mura diventano una sorta di seconda casa.
Grazie alla mia famiglia, che mi ha sempre incoraggiato e lasciato libero di trovare la mia strada, pur consapevoli che mi avrebbe portato lontano da casa; grazie per la presenza costante in ogni momento di questa avventura e in generale in tutta la mia vita.

E, ultimo ma non ultimo,grazie anche a Daniele, che ha vissuto con me una parte di questa avventura con grande pazienza e interesse; mi auguro di avere in futuro tante altre occasioni di inserirlo nei ringraziamenti.