UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO SCIENZE DELLA VITA E DELL'AMBIENTE

Corso di Laurea Magistrale in Biologia Molecolare e Applicata

Progettazione di un sistema CRISPR/Cas13 per identificare lncRNA funzionali nel cancro

Design of a CRISPR/Cas13 system for functional lncRNAs identification

Tesi di Laurea Magistrale
di:
Riccardo Trozzo

Relatore:
Chiar.mo Prof.
Francesco Piva

Correlatore:
Chiar.mo Prof.
Roland Rad

Sessione Febbraio 2022

Anno Accademico 2021/2022

SUNTO IN ITALIANO

Questo lavoro consiste nella progettazione di un sistema CRISPR/Cas13 per l'identificazione di lncRNA funzionali nel cancro tramite HTS (screening ad alta capacità).

I lncRNA sono una classe di acidi ribonucleici di lunghezza superiore ai 200 nucleotidi ancora poco caratterizzata. La loro disregolazione è stata associata a diverse malattie, tra cui diversi tipi di cancro. Questo le rende delle molecole di potenziale ampia importanza nella ricerca sui tumori, anche per quanto riguarda un possibile risvolto clinico. Per questo motivo, studi di genomica funzionale possono essere di grande aiuto per la caratterizzazione della funzione di questa classe di molecole. Questi studi sono normalmente portati avanti con sistemi atti alla disattivazione di specifici geni, come ad esempio i più recenti sistemi CRISPR, di cui CRISPR/Cas13 è un promettente sistema in grado di sopprimere l'espressione di specifici RNA basandosi sulla complementarietà di sequenza. Per questo motivo abbiamo sviluppato un sistema CRISPR/Cas13 per l'utilizzo in screening ad alta capacità (HTS).

Abbiamo generato un plasmide che contiene una cassetta di espressione con la proteina CasRx (Cas13 da Ruminococcus flavefaciens) che viene integrata all'interno del genoma. Abbiamo utilizzato questo vettore per creare 60 linee cellulari di cancro che esprimono in maniera omogenea e stabile la cassetta di espressione per CasRx selezionando quelle con la migliore efficienza di silenziamento di RNA per essere usate successivamente per eseguire screening ad alta capacità. Abbiamo anche ottimizzato gli RNA guida (RNA utilizzati per riconoscere il target del silenziamento genico) utilizzando un tipo di DR (direct repeat) modificata e due sequenze di riconoscimento diverse per ogni sgRNA (single guide RNA). Inoltre, abbiamo generato un plasmide che è stato poi utilizzato per creare un topo transgenico da utilizzare in futuri esperimenti in vivo.

Date le caratteristiche dei lncRNA e data la loro difficile e ridotta caratterizzazione, abbiamo deciso di mettere insieme trascritti da tutti i database disponibili contenenti lncRNA, sia quelli curati manualmente come ENSEMBL che quelli costruiti computazionalmente come NONCODE. Abbiamo aggiunto anche informazioni sulla conservazione evolutiva dei lncRNA utilizzando trascritti provenienti da 6 articoli che hanno studiato la conservazione evolutiva di questa classe di trascritti. L'aggiunta di questa informazione è importante al fine di poter poi eseguire degli esperimenti in vivo sul modello animale generato precedentemente, in quanto solo lo studio di trascritti conservati può avere valenza clinica anche sull'uomo. Data la similarità di gruppi di trascritti, provenienti dallo stesso locus genico e che originano probabilmente dallo stesso trascritto iniziale per processamento dell'RNA, abbiamo creato, tramite diversi script, delle famiglie di lncRNA che raggruppano gli RNA simili tra di loro in un'unica famiglia, che verrà poi utilizzata per disegnare gli RNA guida.

Successivamente, è stato creato un programma che, a partire da un software di scoring per sgRNA di CRISPR/Cas13 ideato da un altro laboratorio, permette di disegnare sgRNA a partire da un input in regioni genomiche o in sequenze. Questo programma permette all'utente di scegliere molti parametri su cui basarsi per disegnare le guide, come, per esempio, la loro distanza nel trascritto.

Tra i quasi 100'000 lncRNA nella nostra lista, ne abbiamo selezionati 24172 basandoci sul livello di espressione. In particolare, abbiamo mappato RNA-seq provenienti da tutte le linee cellulari derivate da tumori solidi in CCLE (Cancer Cell Line Encyclopedia) e abbiamo selezionato un gruppo di lncRNA più espressi in media in tutte le linee cellulari, un gruppi in specifici tessuti e un gruppo in specifiche linee cellulari, selezionando poi anche alcuni tra i più espressi trascritti conservati. Abbiamo quindi usato il programma descritto precedentemente per

disegnare una libreria di sgRNA con la score più alto possibile e seguendo dei parametri di ampia distanza delle sgRNA all'interno del trascritto targettizzato. Inoltre, uno step di rimozione di off-targets ci ha permesso di eliminare tutte quelle guide che mappano trascritti multipli o geni codificanti proteine. Abbiamo utilizzato il programma anche per disegnare gli opportuni sgRNA da utilizzare come controllo, parte fondamentale dell'esperimento.

Per validare la nostra libreria di sgRNA abbiamo eseguito un esperimento di HTS per osservare il comportamento dei controlli e delle sgRNA.

Il sistema realizzato è risultato promettente per lo studio futuro della funzionalità di lncRNA nel cancro. In particolare, l'utilizzo di CRISPR/Cas13 invece di altri sistemi come CRISPR/Cas9 o RNAi (RNA interference) ci permette di limitare l'effetto off-target e soprattutto di targettizzare in modo semplice lncRNA che agiscono nel nucleo (altri sistemi come RNAi non lo permettono). Data la caratteristica espressione tipo cellulare specifica di molti lncRNA, la generazione di 60 linee cellulari tumorali esprimenti stabilmente CRISPR/Cas13 per eseguire esperimenti HTS è di grande utilità per uno studio di ampia portata che mira ad elucidare in maniera più chiara il funzionamento dei lncRNA nei tumori. L'utilizzo di trascritti da molteplici database e il mantenimento delle informazioni sulla conservazione ci permettono di minimizzare il bias dovuto alla grande differenza e poca concordanza che ritroviamo nei database di lncRNA e la selezione dei trascritti più espressi risulta in una valida libreria di sgRNA utile alla ricerca di lncRNA funzionali.

Ci sono poche librerie di sgRNA rivolte allo studio dei lncRNA, di cui soltanto tre utilizzano il sistema CRISPR/Cas13. Di queste, una ha come target soltanto 25 lncRNA e altri due sono specifici per RNA circolari (crRNA). Questo rende la nostra libreria di sgRNA la più grande disponibile al momento non solo per quanto riguarda il sistema CRISPR/Cas13, ma in generale,

visto che la libreria precedentemente più grande per lncRNA ha avuto come target 16401 trascritti.

Inoltre, la nostra libreria di sgRNA comprende 54 dei 76 lncRNA finora ritenuti funzionali nel cancro e quelli che non sono presenti non possono essere targettizzati per via delle loro caratteristiche (e.g. sono antisenso ad un trascritto codificante proteine).

Infine, l'esperimento HTS eseguito in questo studio suggerisce che la nostra libreria di sgRNA è ben disegnata, nonostante ci sia ancora spazio per miglioramenti.

Acknowledgements

I would like to thank my Italian advisor, Professor Francesco Piva, for giving me the possibility of carrying out my thesis work in a foreign University and for constantly following me during this period with invaluable advice.

I must express my sincere gratitude to my advisor, Professor Roland Rad, for welcoming me to his laboratory at the Technical University of Munich and for supervising all the work, giving relentless support throughout all the stages of this journey.

My deepest thanks go to my mentor Juan José Montero Valderrama for his inestimable contribution in making this work possible. I'm profoundly grateful for his constant support and valuable advice without which this thesis would not be possible and for his profound belief in my abilities. I couldn't have asked for a better mentor, both scientifically and personally.

A profound thank goes to bioinformaticians Olga Baranov, Niklas de Andrade Kraetzig and Stephen Clayton for always giving helpful advice and for helping solve informatics problems.

A special appreciation goes to my colleague Ekaterina Zhigalova for teaching me invaluable laboratory skills, especially in cell culture, and for being a big part of this project. Besides her, I'd also like to express my gratitude to all my laboratory colleagues: Miguel Silva, Alexander Belka, Najib Ben Khaled, Julia Eichinger, Anja Grotloh, Devin Jones, Rupert Oellinger, Sebastian Widholz, Sebastian Mueller, Christine Klement, Roman Maresk, Thorsten Kaltenbaker, Anja Pfauss, David Salier, Katharina Collins, Jessica Loeprich and all the other lab members for being great colleagues and great friends.

Thanks from the deepest of my heart to my family for always supporting me in every decision and for always being by my side.

Last but not least, I'd like to thank my friends for always being there in the best and worst moments.

Contents

Chapter 1

Scope of the Study

LncRNAs are a novel class of non-coding transcripts longer than 200 bp whose role as important cell regulators is emerging in a growing number of studies. Despite the low amount of well-characterized lncRNAs, already several of them have been linked to disease and, in particular, to cancer. The lack of well developed and concordant annotations and databases along with the low level of conservation are hindering important discoveries about the function of this class of transcripts by making it difficult to design and perform both in vitro and in vivo experiments. The recent advent of the CRISPR/Cas technologies are opening up plenty of possibilities to perform faster, better and cheaper functional studies via genetic perturbation. Among these techniques, whole-genome scale perturbation is a promising and cost-effective way to assess functionality on the genomic scale and is starting to be used not only in protein-coding genes but also in lncRNAs. Our study aims to provide a robust and scalable screening platform to investigate non-coding RNAs functionality in diverse tumours by using the CRISPR/Cas13 RNA-targeting system and by designing a large and up-to-date sgRNA library. This will allow the generation

of a lncRNA vulnerability atlas across solid tumours and the identification of novel functional lncRNAs.

Chapter 2

Introduction

2.1 lncRNAs

2.1.1 LncRNAs and Their Characteristics

Non-coding transcripts have been long not considered due to the focus on protein-coding genes given by the common knowledge that RNA transcripts are normally translated into proteins. However, the Human Genome Project [1] showed that the genome undergoes what is called pervasive transcription, which means that the majority of the DNA is transcribed. This gives rise to a transcriptome which, not considering ribosomal RNAs, consists mostly of non-coding RNAs, which according to the ENCODE project are summing up to 54% of the transcribed sequences. Of this heterogeneous landscape of non-coding sequences, the most abundant consists of lncRNAs. LncRNAs are transcripts longer than 200nt that do not encode any protein or peptide (they lack an open reading frame). They are mainly transcribed by RNA polymerase II and the majority of them are polyadenylated. However, some of these transcripts do not follow this biogenesis path and are transcribed by RNA

polymerase III. In comparison with the mRNAs of the protein-coding genes, lncRNAs are less stable and are prevalently localized in the nucleus [2, 3]. They are also enriched in two exons transcripts and have a median length of less than 1000bp [4], resulting in them being on average shorter and less complex than protein-coding RNAs. They are usually classified based on their localization in the genome in: Antisense lncRNAs, which are, as the name suggests, arising from the antisense strand of a protein-coding gene; Long intergenic non-coding RNAs or LincRNAs, which are not overlapping in any way with protein-coding genes; Sense overlapping lncRNAs, having some degree of overlapping on the same strand with coding genes and Sense intronic lncRNAs, arising from the same strand of introns of coding genes. LincRNAs are found to represent the most abundant group [2, 5]. In addition, some lncRNAs have been shown to be precursors of small RNAs. Zfas1, for example, gives rise to three different small nucleolar RNAs or snoRNA while h19 is thought to be a precursor of a microRNA, making the boundary between lncRNAs and small RNAs less defined [6, 7].

2.1.2 Functions of lncRNAs

LncRNAs have been found to have a diverse set of functions, both at genetic and epigenetic levels regulating transcriptional, post-transcriptional, translational and post-translational stages [8]. LncRNAs can interact with different proteins and promote their binding to specific regions in order to mediate expression activation or suppression (Fig.2.1*a*). For example, the well-known lncRNA MALAT-1 [9, 10] was shown to mediate the expression of the LTBP3 gene by

recruiting the transcription factor Sp1 on the LTBP3 promoter [11].

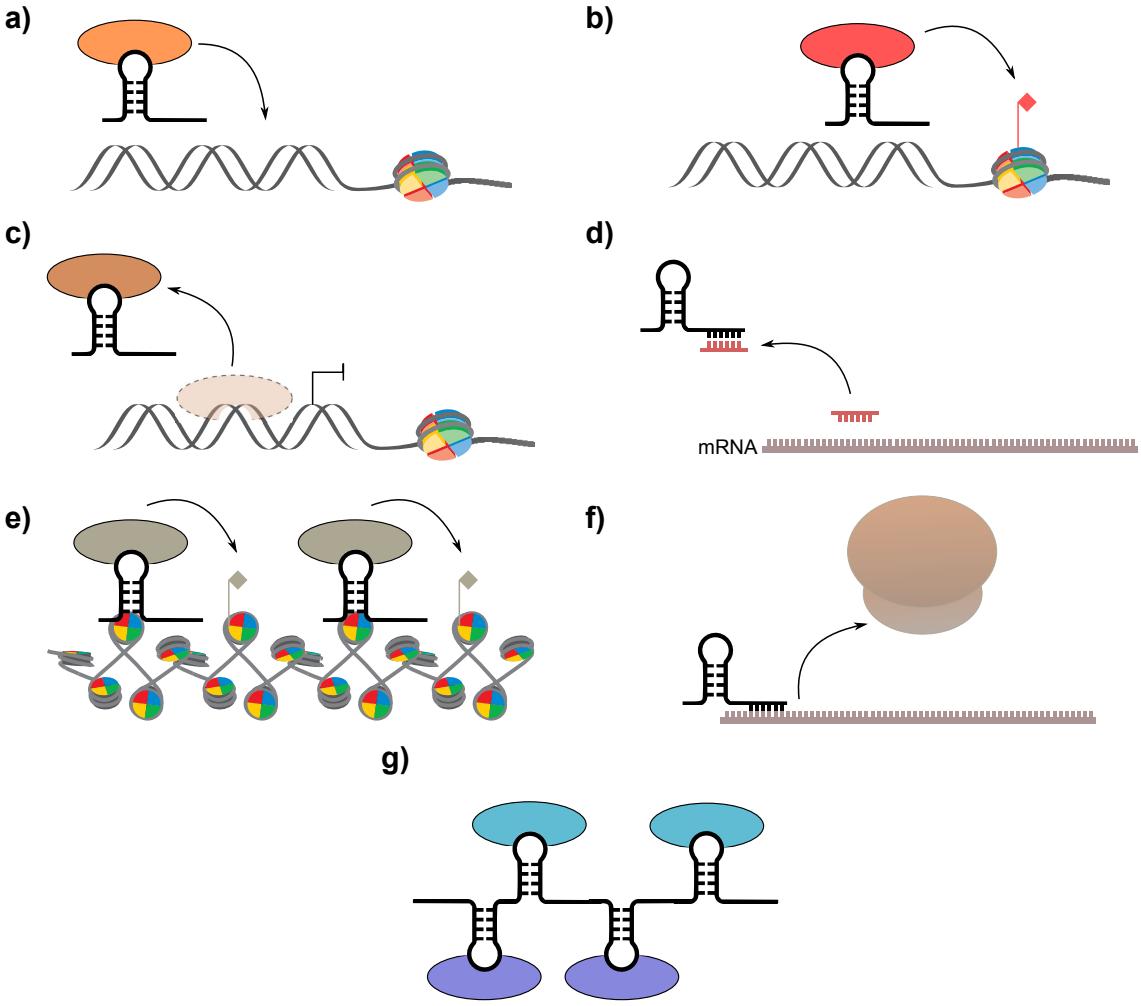

FIGURE 2.1: Different roles of lncRNAs: **(a)** lncRNAs can interact with proteins to promote their binding on specific loci **(b)** lncRNAs can act as guides, they promote the binding of chromatin modification proteins on the DNA **(c)** lncRNAs can act as a decoy, they bind proteins that would otherwise bind in DNA regulatory sequences, sequestering resulting in alteration of transcription **(d)** lncRNAs can act as miRNA sponges, sequestering miRNA and preventing them to bind their target mRNA **(e)** lncRNAs can act promote extensive chromatin remodeling **(f)** lncRNAs can promote mRNA translation by assisting their binding to ribosomes **(g)** lncRNAs can act as Scaffold, they recruit several proteins forming complexes with the chromatin that have a structural role

LncRNAs can also recruit chromatin modifiers to specific locations in the

genome (Fig.2.1*b*). The lncRNA ANRASSF1 is able to recruit the repressing epigenetic complex PRC2 at the RASSF1A promoter, resulting in the down-regulation of the gene [12]. LncRNAs can act as decoy (Fig.2.1*c*) in order to sequester transcription factors or other proteins inhibiting their functions [13]. The lncRNA PANDA was shown to interact with transcription factor NF-YA upon DNA damage, competing with its binging to pro-apoptotic promoters to mediate cell survival [14]. There are also non-coding transcripts with the function of miRNA sponges (Fig.2.1*d*). The non-coding isoform of PNUTS, lncRNA-PNUTS, binds on the micro RNA miR-205, acting as a sponge and leading to the upregulation of the ZEB gene proteins during early stages of EMT transition [15]. Extensive chromatin remodeling is also a well documented function of this class of transcripts (Fig.2.1*e*). One of the first characterized lncRNAs, XIST, is required for X chromosome inactivation during mammalian development. The XIST lncRNA is transcribed selectively from the X copy that will be silenced, and acts by binding numerous sites on the chromosome body recruiting chromatin modulators that lead to the complete inactivation of most of its genes [16, 17].

At the post-transcriptional level lncRNAs can bind specific mRNAs by sequence complementarity in order to module their translation (Fig.2.1*f*). The antisense non-coding transcript of the coding gene Ubiquitin Carboxyterminal Hydrolase L1 (Uchl1) was shown to promote, under rapamycin-induced mTOR inactivation, the overexpression of its complementary sense transcript, by recruiting the mRNA to polysomes without affecting the messenger level [18].

Besides regulatory functions, there are several non-coding transcripts involved in structural functions, serving as scaffolds to build ribonucleic-protein

complexes (Fig.2.1*g*). One of these lncRNAs is the nuclear enriched abundant transcript 1 (NEAT1). NEAT1 serves as the main scaffold in the formation of paraspeckles, complexes of lncRNAs and proteins present in the nucleus that have a role in the regulation of the expression of specific genes [19].

2.1.3 LncRNA Role in Disease and Cancer

Aberrant expression of lncRNAs has been linked to several diseases, ranging from cardiovascular diseases [20] to neurodegenerative diseases like Alzheimer [21]. Of particular interest is the association of lncRNAs atypical expression patterns and cancer. The switch from RNA Microarrays to High-throughput sequencing led to a big increase in the identification of lncRNAs deregulated in cancer phenotypes and, even though functional characterization struggles to keep up, numerous lncRNAs are being found to be dysregulated in cancers. A lot of these studies, though, are carried out using bioinformatic approaches to find correlations between lncRNA over or under expression, somatic mutation or gene dosage changes by copy number alterations and cancer [22]. These methods are however not sufficient to confirm the functionality of these transcripts or to derive a tumour suppressor or oncogenic role. To clearly elucidate the role of these transcripts in tumours it is necessary to perform functional studies with methods like RNAi [23, 24], or the more recent and reliable CRISPR-Cas system [25] of which CRISPR-Cas13 is a promising tool to perform loss or gain of function studies. In vivo studies on animal models are also fundamental, but are hindered by the low level of conservation of lncRNAs, so the field remains widely unexplored with only a few knock-down mouse

models already generated [26]. However, even with all these limitations, some lncRNAs have been found to play important roles in the regulation of tumour suppressor genes and oncogenes [27]. For example, the lncRNA *lincRNA-p21* was found important in transcriptional response dependent on p53. As we already said, lncRNAs have also a role in chromatin modulation, therefore, they have been linked to cancer-related chromatin alterations. One example is the well-characterized lncRNA HOTAIR. This lncRNA is capable of mediating the formation of a complex composed of the Polycomb Repressive Complex 2 (PRC2) and the complex LSD1 that mediates epigenetic modifications [28] and it's found to be overexpressed in several human cancers and related to tumorigenesis and metastasis [29]. Another lncRNA linked to cancer is PTENP1. This lncRNA act as a decoy for other non-coding transcripts belonging to the class of microRNAs thus positively regulating the expression level of PTEN [30] and it was found to be altered in melanoma, prostate and colon cancer [31]. Given all these examples, it must be underlined that only a few lncRNAs so far were found functional in cancer [27]. Part of the reason for this, as already mentioned, is the lack of functional studies and conservation of these transcripts coupled with the fact that the majority of the lncRNAs are poorly annotated [32].

2.1.4 lncRNA Databases and Annotations

LncRNAs annotations and databases are much more discordant and underdeveloped than the ones of protein-coding genes. The main reasons for this are their low level of expression compared to coding genes, which makes it difficult to confidently assign reads to transcripts and their poor evolutionary

conservation [33–38] and cell-type specificity, which makes it hard to detect lncRNAs that are arising from just a few samples and translates to a poor link between sequence and function. The lack of open reading frames (ORFs) is another fundamental factor, while ORFs and other sequence features can be used to easily identify protein-coding genes, this is not possible for non-coding transcripts.

There can be two types of annotation: Automatic and Manual. Automatic annotation is carried out by computers using algorithms that are performing transcript assembly based on RNA-seq reads. Due to the short length of these reads, the assembly of transcripts is not accurate [39] leading to the identification of incomplete or wrong constructs. Manual annotation, on the contrary, is, as the name suggests, put together manually by researchers following specific criteria. This results in more complete and accurate transcripts collections. The resulting difference is that automatic annotations are producing a much higher number of transcripts but with less accuracy, while manual annotations are producing smaller databases but with more accurate lncRNAs. This can be seen comparing the size of various databases. Integrative databases (i.e. databases that use both manual and automatically reconstructed transcripts) are quite large, with the largest being NONCODE [40] presenting a collection of 96411 lncRNA genes, while manually curated databases, which consist of GENCODE [5] developed from the ENBL and RefSeq [41] curated from the NCBI, are relatively smaller, with GENCODE having 15,512 lncRNAs. The difference is not only in the database size, resulting in a tradeoff between quality and size, but also in the number of lncRNAs that the databases share. According to Uszczynska-Ratajczak et al. [32] the majority of transcripts are not

shared between databases, even in the manual curated ones less than 50% off the lncRNAs are shared.

2.1.5 Evolutionary conservation of LncRNAs

LncRNAs are known to be poorly evolutionary conserved but information on conservation is fundamental because of its tight link with functionality [33]. This characteristic, alongside the low expression level, the high tissue and cell-type specificity and the lack of an ORF that cause non-coding sequences to be unconstrained (unlike highly constrained protein-coding sequences), makes evolutionary conservation analysis of lncRNAs very challenging. Due to this, only a few extensive studies on conservation have been performed. These studies were taken on a variety of tissue types derived from different organisms from hominids like the chimpanzee, to mice and non-mammal animals, including samples from different developmental stages in a study from Sarropoulos et al. [34]. Since there is no consensus on how to perform this type of analysis, different groups chose different methods to decide whether or not two transcripts have to be considered conserved. Necsulea et al. [33] and Sarropoulos et al. [34] used RNA-seq reads assembly to determine non-coding transcripts and reconstructed homologous families using sequence similarity. Hezroni et al. [38] and Pervouchine et al. [36] also used RNA-seq assemblies, but to determine homologous families they added, besides lncRNA sequence similarity, whole-genome alignments requiring also synteny as criteria. Washietl et al. [35] use lincRNAs from GENCODE [4] and assess orthology using information on synteny by genome-wide alignments. Chen et al. [37] uses a

custom-made pipeline to assemble novel lncRNAs from RNA-seq data and use syntenic regions first followed by transcript-genome and transcript-transcript sequence similarity to assess evolutionary conservation. These conservation analyses found the lncRNAs to have a low level of sequence constraint. The lncRNAs that were found conserved at various degrees, though, showed a higher level of expression conservation [36] along with conservation of tissue specificity [35]. Of notice, Sarropoulos study, which included developmental lncRNAs, also found that conserved transcripts are more likely to have a dynamic expression during development. The poor consensus in the method used to build their collections together with the limitations in the number of samples analyzed and the high lncRNA tissue specificity results in a poor overlap between the conserved sets. However, by putting the different datasets together the individual limitations of the different methods used are bypassed and the resulting collection of conserved lncRNAs given by this papers remains an invaluable tool for inferring functionality to integrate with other data.

2.2 CRISPR/Cas Systems and Their Use in Molecular Biology

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) sequences associated with Cas proteins are primordial immune mechanisms found in most bacteria and archaea genomes with the scope of neutralizing Bacteriophages infections by cleaving their viral nucleic acids. This system is adaptive, meaning that the microorganisms possessing it undergo a process of immunization that

occurs via the integration of the new infecting nucleic acid in the CRISPR loci, in order to have an immediate response on following infections [42–44]. These specific characteristics of the CRISPR-Cas system make it a unique tool to perform a different range of genetic perturbation applications. The CRISPR-Cas system is characterized by having a CRISPR locus, where several target sequences, named spacers, are separated by short repeating sequences called direct repeats and a Cas module, which comprises one or more proteins that are capable of recognizing the spacers and use them as a template to perform cleavage of complementary nucleic acids. There are two classes of CRISPR-Cas and several types for each class. Specifically, class 1 systems are characterized by the presence of several Cas proteins acting together as a complex to perform the nucleotide cleavage and are divided into types I, III and IV. Class 2 CRISPR-Cas are defined, instead, by having only one Cas protein that alone has the nuclease enzymatic activity and is divided into type II, V and VI [45]. Type II and V contain respectively the Cas9 and the Cas12a protein (formerly named Cpf1) and have been extensively characterized. Since target recognition is based on Watson-Crick base-base complementarity, making the system very easy to design, Class 2 CRISPR/Cas was harnessed to perform a variety of DNA targeting experiments. For example, CRISPR/Cas9, arguably the most widely used system from the CRISPR/Cas family, consist of a Cas protein named Cas9 which has nuclease activity and is able to disrupt the DNA forming a DSB in a region of the genome that is complementary to the spacer; a crRNA (crisprRNA) that is made of the direct repeats followed by the target sequences, a tracrRNA (trans-activating crRNA) which forms a hybrid with the crRNA that is necessary for the Cas9 protein to be active and produce the DSB [46, 47]. To perform a cut

into the DNA the system requires a *Protospacer Adjacent Sequence* (PAM) that is NGG (where N stays for "any base") following the target sequence. In the engineered version of the CRISPR/Cas9 the tracrRNA and the crRNA are fused together forming one single molecule [46] (2.2 *(a)*).

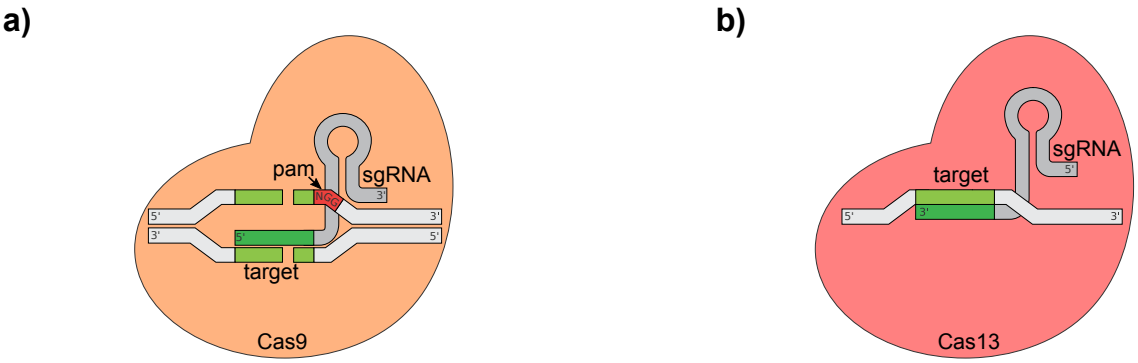

FIGURE 2.2: Different CRISPR/Cas systems: **(a)** CRISPR/Cas9 requires PAM sequence and creates DSBs in the DNA **(b)** CRISPR/Cas13 doesn't require a PAM sequence and acts directly on the RNA resulting in its knock-down

Besides the types I-V, Type VI was only recently described [48–50] and it has the novel characteristic of targeting exclusively RNA instead of DNA, opening up new possibilities for the study of RNA biology. Within type VI there are four different subtypes [49–52]. All of them were used to assess their potential as molecular biology tools for knock-down of RNAs, with the type VI-D resulting in being the most promising [52]. In respect to typeVI-A to typeVI-C, Cas13d is significantly smaller and doesn't show to be dependent on PFS (Protospacer Flanking Sites) [52]. From this sub-type [53] engineered and evaluated the activity of different orthologues, finding their *Ruminococcus flavefaciens XPD3002* Cas13 fused to a nuclear localization signal (NLS) as the most efficient [53]. Cas13d sequence is not highly conserved in respect to its orthologues a-c, except for the RNA cleaving domains, that in all the proteins belonging to type VI

consist of two HEPN RNases. The Cas13d is found normally in an inactive apo conformation which changes to a surveillance conformation upon binding of the crRNA thanks to the recognition of a short hairpin. When the surveillance conformation recognizes the target RNA a cleavage competent complex is formed [54]. There is no requirement for a PAM sequence unlike we find in Cas9 giving much more room for targeting sequence design. Another advantage is that the seed region covers almost all the length of the crRNA, resulting in a system less subjected to an off-target effect. Cas13 proteins are able to process pre-crRNA into mature guides, giving the possibility to use multiple spacers [54] (2.2 *(b)*).

2.3 Functional Screenings Methods for lncRNAs

2.3.1 How Functional Screenings Work

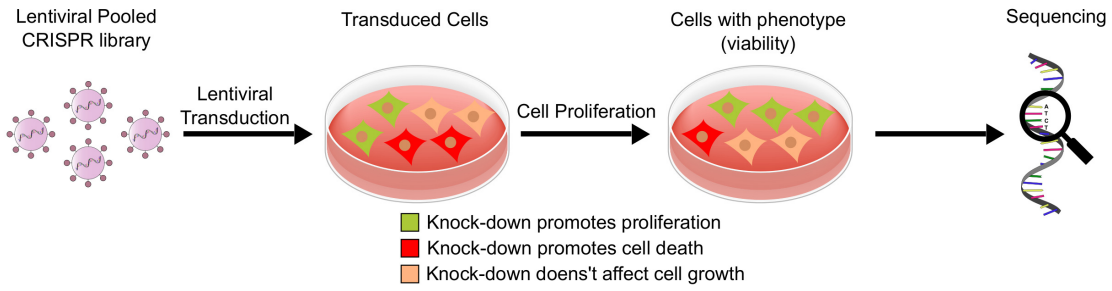

FIGURE 2.3: Pooled Screenings: **(a)** gRNA library is delivered to the cells through lentiviral transduction. After selection, cell proliferation is carried on for several days. Knock-down of a transcript can promote cell death, cell proliferation or can leave the phenotype unaffected. At the end of the experiment, the cell pool is sequenced to measure the phenotype (proliferation)

Functional genomics screening techniques are valuable molecular biology tools that allow to elucidate gene functionality in a variety of conditions (e.g. cancer). They consist of delivering to cells in 2D culture a CRISPR guide RNA (gRNA) library, making sure that each cell is infected by only one guide by using a low multiplicity of infection (MOI). This delivery will trigger a loss of function of the gene or a gain of function, based on the system used for the screening (e.g. RNAi), and measurement of the phenotype (e.g. viability) is then performed (2.3 *(a)*). A typical type of screen is one based on cellular fitness. It consists in using a knock-down or knock-out system in order to have a loss of function and what is measured is the change in cell viability: a reduction in the number of cells that contain a specific gRNA, indicates an important role of the targeted gene for keeping the cell alive, this is considered a drop out in the screening. On the other hand, an increase in the number of cells shows that this gRNA is affecting a gene with negative effects on cell proliferation (e.g. a tumour suppressor) and is considered an enrichment in the screening (Fig.2.3*a*) [55]. The majority of the screens are performed with the CRISPR/Cas9 system, which is the best-characterised system. These screenings are usually performed on protein-coding genes either by knock-out (CRISPR ko), gene activation (CRISPRa) or gene inactivation (CRISPRi) [55].

2.3.2 Screens for lncRNAs

While protein-coding genes screenings have been extensively performed, functional screenings on lncRNAs, especially in a cancer context, have been performed only in a handful of studies, and there is little consensus on the

method that is best to use [56]. A broad range of techniques are available that make it possible to perform large scale screenings of lncRNAs, but many of them bear big burdens.

RNAi RNAi is a method that allows targeting directly the RNA molecule by using a dsRNA with complementarity to the RNA that has to be knocked down [57, 58]. Its function is only transient, but genome integration can also be obtained by using short hairpin RNAs (shRNA), which are, instead, integrated into the genome by viral infection, and therefore have a stable long term expression [59, 60]. Though several screenings were performed on lncRNAs using RNAi [61–65], this system yields two main drawbacks. First, it has a significant off-target effect that makes it difficult to discern whether the observed phenotype is caused by the targeted lncRNA [66–68]; second, its action is mostly carried on in the cytoplasm, leaving behind the majority of lncRNAs that exercise their function in the nucleus [69] (Fig.2.4*(a)*).

ASOs Another method to target directly the transcripts is using antisense oligonucleotides (ASO). ASOs are synthetic short (20bp) DNA molecules that are complementary to an RNA target. The binding of the ASO to the target transcript can be recognized by the RNAse-H that destroys the target resulting in its knock-down [70]. The biggest screening performed with ASOs [71], though, has targeted only 285 transcripts and was performed only for a short time (48h), given to the transient nature of the method (since DNA cannot be encoded genetically). Hence, ASOs use is preferred for the validation of targets rather than in functional screenings (Fig.2.4*(b)*).

CRISPR CRISPR/Cas systems can be used both for targeting the transcripts indirectly, by targeting the genome [72] or directly by targeting the RNA molecule [49]. CRISPR/Cas9 is a common system used to perform protein-coding genes screenings [73], and was therefore used also for lncRNA discovery [74]. Targeting lncRNAs with Cas9, though, brings a major limitation. In fact, in coding genes screenings this technique is used to disrupt the target's ORF, thus knocking it out. The lack of an ORF that characterizes lncRNAs makes it impossible to use this approach in the same way, therefore, the possible workarounds are to perform deletions or to target splice-sites. However, disrupting a lncRNA sequence in this way does not necessarily lead to a dropout, because the mutated region may not affect its function. The Cas9 protein has also been used in its dead Cas9 (dCas9) version, where the nuclease domain is not active anymore, fused with proteins that are able to perform chromatin modifications. The CRISPR interference (CRISPRi) system, for example, consist of a dCas9 that is fused with a repressor domain, leading to silencing of the gene [75]. The CRISPR activation (CRISPRa) system, on the other hand, pairs the dCas9 with an activator domain, increasing the expression of the targeted lncRNA. Both these systems have also been used successfully in functional screenings [25, 76, 77]. Nevertheless, CRISPR/Cas9 is targeting directly or indirectly the genome, either by inducing double stranded breaks (DSBs) or by modulating chromatin, resulting in the possibility of interfering with other DNA elements close to the target or overlapping with it, making the use of this technique far from optimal [78]. CRISPR/Cas13, on the other hand, gives us the possibility to directly target the transcript of interest [49], solving all the issues that are connected to DNA targeting systems. Furthermore, compared

with RNAi, CRISPR/Cas13 doesn't show to have a problematic level of off-target effect [79] making it a first choice tool for functional screening of lncRNAs (Fig.2.4*(c)*).

By far, only three screenings were performed using CRISPR/Cas13 for lncRNA studies. The first one was performed on a small subset of very long intergenic lncRNAs (vlincRNAs), and it was used to assess anticancer drug treatments response [80]. The other two CRISPR/Cas13 screenings were performed on circular RNAs (circRNAs). They were designed to target the back-splicing junctions typical of circRNAs in order to selectively knock them down [81, 82].

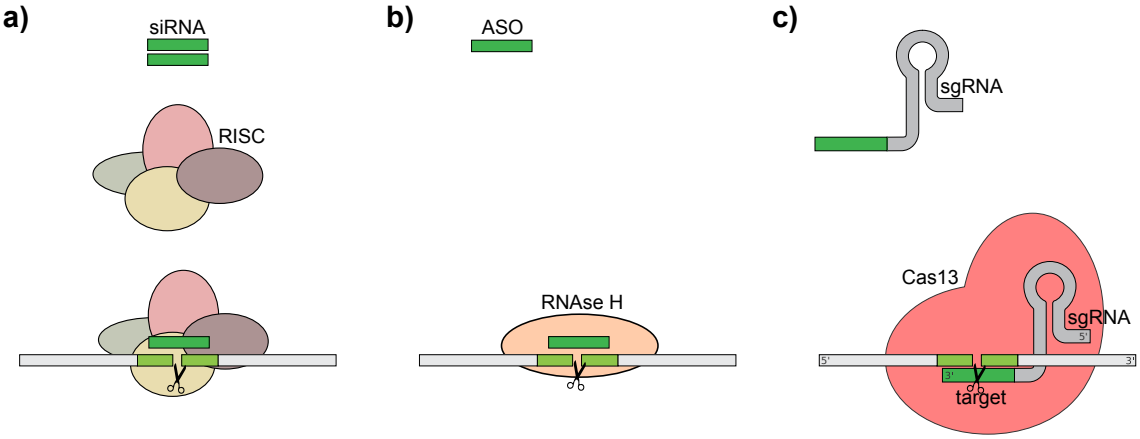

FIGURE 2.4: lncRNA Perturbation Methods: **(a)** RNA knock-down with siRNA: the RISC complex binds the delivered siRNA and recruit a complementary RNA that is then cleaved by a nuclease domain **(b)** RNA knock-down with ASOs (Antisense Oligonucleotides): the delivered ASO recruits an RNAseH that perform the cleavage of the complementary RNA **(c)** RNA knock-down with CRISPR/Cas13: the Cas13 protein recognises the guide RNA and binds the complementary RNA mediating its cleavage

Chapter 3

Materials and Methods

3.1 Cell Culture

3.1.1 Cell Maintenance

All cell lines were cultured in DMEM (ThermoFisher Scientific) or RPMI (ThermoFisher Scientific) medium with 10% FCS Superior (Sigma-Aldrich) and 1% penc/strep (ThermoFisher Scientific). Passaging of the cells was made at 75-85% confluency. Incubation conditions were 37°C with 5% CO_2.

3.1.2 Transfection

For all transfection experiments, we used a Lipofectamine®3000 (ThermoFisher Scientific) protocol. In Tube 1 we put 140 µL of OptiMEM and 7.2 µL of Lipofectamine®3000. In Tube 2 we put 140 µL of OptiMEM (ThermoFisher Scientific), 4.8 µL of P3000 reagent given with the Lipofectamine®3000 kit and 2.24 µg of plasmid DNA, which consists of pre-mixed Cas13 transposon/transposase plasmids in a ratio of 5:1. We combined the tubes with

a 1:1 ratio followed by an incubation period of 5 minutes at room temperature. The mix was then added to one well of a 6-well plate. All the ratios were multiplied times the number of wells used. The cells were then selected using two different Blasticidin concentrations, respectively 5 µg/mL and 10 µg/mL in case the higher one would result too toxic. We used as control a non-transfected version of the cell line. The selection was carried on until all the cells in the control well were dead.

3.1.3 Monoclonal Cell Line Generation

To generate monoclonal cell lines that have CRISPR/Cas13 integrated into the genome and constitutively expressed we performed a serial dilution with cells that integrated the Cas13 protein. First, we cultured the cells in a 10 cm dish collecting two times 10 mL of the conditional medium. After having obtained the conditional medium the cells were trypsinized, followed by a centrifugation step to remove all the trypsin. Then the cell pellet was resuspended in 10 mL medium. From this 1 mL of the medium was taken and diluted again to 10 mL. From this second dilution 200 µL were transferred to a well of a 96-well plate. A subsequential 1/2 dilution along the vertical axis was then performed, followed by a subsequential 1/2 dilution along the horizontal axis. This procedure was repeated to obtain four 96-well plates. These were incubated for a minimum of two weeks up to 4-5 weeks for the monoclonal cells to grow enough. The clones that were grown enough from the 96-well plates were trypsinized and transferred to a new 96-well plate. Here they were kept growing until reaching full confluency, usually an additional one or two weeks.

3.1.4 Lentiviral Production

Lentiviral production was carried on using HEK293T cells. On day one the cells were seeded on a 10 cm dish with a density of 1-1.5 million cells in DMEM with 5% FCS and 1% penc/strep. On day two the medium was changed to DMEM without the addition of antibiotics. The transfection mix was prepared using 18 µL of TransIT®(Mirus) transfection reagent with 270 µL of OptiMEM, followed by an incubation time of 5 minutes at room temperature. The packaging plasmids pMD2.G (AddGene # 12259), which supplies the envelope gene, and psPAX (AddGene # 12260), which supplies the gag and pol genes, and the sgRNA was added followed by 30 minutes of incubation period at room temperature. On day 3 the medium was replaced with the growth medium of the cell line to be transfected. On day 4 the first virus batch was collected in a tube using a 0.22 µm filter and the medium was replaced. On day 5 the second batch of the virus was collected in the same way and the cells were discarded.

3.1.5 Lentiviral Transduction

For cell transduction on day 1, we seeded the cells in order for them to reach 65-75% confluency on the following day. On day 2 we added the first virus batch along with 4 mL off growth medium and Polybrene reagent with a concentration of 10 µg/mL. We mixed thoroughly and then added the virus dropwise to the wells. On day 3 we performed a second round of transduction using the second virus batch and the same procedure used in the first round. On day 4 we changed the medium with growth medium + 10% FCS + 1% penc/strep.

3.1.6 Flowcytometry

Cells for flow cytometry analysis were prepared by trypsinization until complete detachment and addition of FACS media, consisting of PBS with 2MM EDTA. Data were analysed with the Cytoflex LX machine from Beckman-Coulter.

3.2 Determination of the multiplicity of infection (MOI)

The MOI determination was performed on two groups of the same cell line. Virus supernatant was added in various concentrations to 70% confluent cells on 15cm dishes. Media with the virus was removed 24h after infection and a 2.5 µg/mL of puromycin was added with new media to the first group while no puromycin was added to the second, a control with no transduction where puromycin was added was used to monitor selection progression. The selection was carried out until no cells were left alive on the control plate. Cells on both groups were counted and the division between the cells in group 1 and group 2 gave us the MOI. We selected the amount of virus necessary to have a MOI of 0.3.

3.3 Pooled CRISPR/Cas13 depletion screen

Cells were transduced with 400 cells per sgRNA coverage and MOI 0.3 via spin infection. The day after transduction cells were split and seeded in dishes with an appropriate puromycin concentration selection. The selection was carried out for 4 days and then the cells were trypsinized and mixed, and afterwards, they were

split into 2 replicate at 200 cells per sgRNA coverage. After replicates seeding normal passaging was then carried out for 3 weeks always maintaining 200 cells per sgRNA coverage.

3.4 Cloning

3.4.1 Gibson Cloning

All Gibson clonings were performed by digesting for 4h the plasmid with the appropriate enzyme to obtain the backbone (plasmid without insert). The backbone was then directly PCR purified in case of digestion with one enzyme while when digested with two enzymes it was run on 1% agarose gel to then cut and purify only the fragment of interest using NEB Monarch®purification kit (New England BioLabs). The Fragments were obtained by PCR. The reaction was set by using always 50-100ng of the vector with 2-3 fold of insert fragments. H_2O up to 5 µL was added to the vector and fragments mix with 5 µL of 2xGA Master Mix®(New England BioLabs).

3.4.2 Gateway Cloning

Gateway cloning was performed by adding 150ng of entry vector to 11 µL of a 150ng solution of the destination vector, adding water up to 8 µL. LR clonase II®(ThermoFisher Scientific) was thawed for 2 minutes on ice and vortexed briefly. Then, 2 µL of the LR clonase II®were added to the tube with entry and destination vectors followed by vortexing and microcentrifugation. The reaction was incubated at 25°C for 1h.

3.4.3 Bacterial Transformation

All bacterial transformations were done using chemical transformation. 20 µL of KCM and 5 µL of vector were mixed on ice, adding water up to 100 µL. 100 µL of *E.coli* Stbl3 strain competent cells were added and mixed on ice. Incubation of 20 minutes on ice and 10 minutes at room temperature followed. We added then 1 mL of LB medium without antibiotic and incubated in continuous shaking for 1h at 37°C at 750rpm. We then performed a centrifugation step and plated the bacteria on agar dishes with antibiotic resistance according to the transformed plasmid.

3.4.4 Colony PCR

Colony PCR to validate cloning was performed by picking 96 clones from the agar dish and resuspending them in 20 µL of H_2O on a 96-well PCR plate. 4.6 µL of this bacteria suspension was added to a 5.4 µL Kapa2G PCR Master Mix (Roche) and mixed. The mix was then run on a 1.5% Agarose Gel.

3.4.5 Plasmid Miniprep

Plasmid Miniprep was performed using NEB Monarch®Plasmid Miniprep kit (New England BioLabs) and validated by Sanger Sequencing.

3.4.6 Plasmid Midiprep

Plasmid Midiprep was performed using a Plasmid Midiprep endonuclease-free kit (Qiagen) and validated by Sanger sequencing. and then the cells were

harvested for sequencing.

3.5 Bioinformatics

3.5.1 Fusion of similar lncRNAs into lncRNA families

selection of the lncRNAs Genomic coordinates of lncRNAs from all the databases were downloaded from the RNAcentral website. LncRNAs genomic coordinates from the 6 papers on evolutionary conservation of lncRNAs [33–38] were downloaded in excel format from the download section of their publications. The different custom formatted excel tables from each paper were converted with custom scripts in formatted bed file (chromosome; start; end; name; score; strand) tables as defined by Ensembl bed format specifications. From these tables, only the lncRNAs found to be evolutionarily conserved were retained. In the case of [35], where ENSEMBL gene IDs were given as identifiers of conserved regions, all the transcripts arising from each gene were retrieved and considered conserved. All genomic coordinates were referred to genome build hg19 while data from RNAcentral was downloaded for the most recent hg38 build. Therefore all the hg19 coordinates were changed to hg38 with a custom script using the command line version of the liftover tool from UCSC genome browser [83]. All the bed files were then merged into a single bed file retaining the information on their database of origin and conservation.

transcripts fusion steps First, the merge tool from BedTools was used in order to create macro groups of transcripts that are overlapping at least 50 bp.

Then with custom scripts, we refined these macro groups according to several characteristics. First gene body only was considered and the macro groups were split into groups of transcripts overlapping for at least 25% of the length of the smallest transcript. Then, considering exons overlap, within each of these groups transcripts that share at least one exon overlapping by 60% of the length of the longer exon were merged. All the calculations were made for each strand separately since things overlapping in the same strand shouldn't be put together. Afterwards, small single-exon transcripts overlapping for at least 30% to an exon of a biger transcript and multi-exon transcripts overlapping for at least 90% of their whole exome one or more exons of much bigger transcripts were merged. Since all these steps require a lot of exon-exon comparisons and thus they are computationally very demanding, each step was preceded by sorting of the file by coordinates and pairwise non-exhaustive iterative comparisons of subsequent exons were made. After the non-exhaustive comparisons reduced the number of exons to parse, exhaustive comparisons between all the exons in each group were made. This made the scripts much faster. To trim transcripts overlapping coding genes body the GTF file of the complete GRCh38.12 transcriptome annotation was downloaded from GENCODE [5] and the coding genes were selected. The genomic localization of those coding transcripts was used to trim lncRNAs overlapping coding exons on the opposite strand and lncRNAs overlapping coding exons and introns on the same strand. We eliminated transcripts that were trimmed for more than 60% of their original length.

3.5.2 sgRNA Library Design Tool

The sgRNA library design tool takes the input file in one of the three possible formats (ENSEMBL ID, genomic coordinates, sequence) and retrieves the transcript sequence. It then trims the sequence at both ends according to the *trim range* parameter. It feeds this sequence to the Wessel et al. algorithm [84] to generate the scoring of all the possible 23mers (or another user-defined guide length). It performs off-target filtering of spacers that map more than a single transcript or that map a coding gene. This step is done with the bowtie2 tool using a custom bowtie2 transcriptome reference with all our lncRNAs and all the coding transcripts from GENCODE. To retain only the unique mappers, the tool *view* from the samtools [85] was used, retaining only reads with maximum *MAPQ* score. It then selects a number of spacers equal to the multiplication of the *number of guides* times the *number of spacers* parameters that fulfil the *minimum distance* and *global minimum distance* parameters, avoiding to design guides complementary to exon-exon junctions (this is done by removing guides that are not mapping completely to a single exon, because of their overlap to multiple exons). All the spacers are extended according to the *extension* parameter. All the spacers containing the enzyme cutting recognition sequence provided by the user are also filtered out.

3.5.3 RNA-Seq Data Analysis

RNA-Seq data from CCLE (Cancer Cell Line Encyclopedia) was downloaded using SRA-toolkit. To obtain the counts used to select the list of transcripts to design the library, the raw reads were processed with Trimmomatic to remove

Illumina Adapter sequences, leading and trailing bases with a quality lower than 25 and keeping reads of is a minimum length requirement of 50bp is met. Trimmed reads were aligned to the GRCh38 genome build using STAR and the counts obtained with feature counts by allowing multi-overlap of feature, using our reference transcripts along with the coding transcript collection from Ensembl. The coding transcripts were obtained by filtering the transcript collection from Ensembl for the coding features. For the second RNA-Seq analysis raw reads were quantified with Salmon using standard parameters and the entire genome as a decoy. TPM count tables were obtained by dividing the read counts by the length of the transcript in kilobases (RPK). Each RPK value was then divided by the sum of all the RPKs divided by 1'000'000.

3.5.4 Screening controls design

Non-targeting controls were designed by first generating 10000 random 23mers and filtering out all the ones that were mapping any location in the transcriptome reference with up to 3 mismatches with bowtie aligner [86]. 600 of these 23mers were selected. Another set of 7mers was randomly generated and randomly used to elongate the 600 23mers to reach 30bp of size which is the size needed by our CasRx system. Always essential and never essential genes are genes that are always dropping out (always essential) in screenings from *depmap* or that are never dropping out (never essential) in screenings from *depmap*. A list of 300 always essential controls and 300 never essential controls was selected. Using BiomaRT [87] we retrieved the *MANE Select* (Matched Annotation from NCBI and Ensembl) transcript (A selection of representative transcripts for each gene from

ENSEMBL and NCBI [88, 89]) for each of the controls and we fed the sequences to our guides design program.

3.5.5 Screening analysis

The sgRNA sequencing counts were obtained with the Mageck count tool from the *MAGeCK* utility by using only the first sgRNA of two to perform the internal mapping. The screening analysis was performed using the mageck count with the mageck mle tool from the *MAGeCK* software. The counts from *MAGeCK* were used to calculate a custom LFC score. The LFC for each gRNA and each replicate was calculated (normalized count of the treatment divided normalized counts of the control) followed by applying log2 to the resulting frequency. After, the best two gRNA per gene were selected. From this selection, the mean was obtained to generate the final LFC score.

Chapter 4

Results

4.1 Construction and validation of the CasRx Plasmid

4.1.1 Construction of the Plasmid

One of the main limitations regarding the application of the CRISPR-Cas13 system is the amount of protein that can be expressed in the target cell or tissue. The previous in vitro approaches used transient delivery [53], but in our hands, similar plasmid constructs fail using permanent lentiviral delivery. In previously published work this problem was bypassed by performing a clonal selection of single cells with high expression of the Cas13 protein [80, 84]. Also, different Cas13 systems have been tested but it was shown that the Cas13d protein from the Ruminococcus flavefaciens XPD3002 (also named CasRx) has the most robust expression and the highest specificity in mammal cells. This protein shows the greatest knock-down efficiency compared to other members of the Cas13d family and has a better performance in off-target activity

compared to shRNA [53, 81]. Another important aspect to take into consideration is that lncRNAs mostly present a nuclear localization. Therefore, for our purposes, the CasRx system needs to be engineered so that it translocates into the nucleus. In order to use the CRISPR-Cas13 system to target specific lncRNAs, we developed a plasmid system that fulfils all the aforementioned requirements and we tested its performance.

To construct the plasmid (Fig.4.1*a*) we started by cloning the humanized sequence for the *CasRx* protein to an entry vector with ampicillin resistance and attL sequences. The fragment containing the sequence for the *RfCas13d* protein and the Nuclear Localization Sequence (NLS) was obtained by PCR amplification from another previously published plasmid (Addgene #109049) [53]. The technique used for cloning was *Gibson assembly*[90], therefore the primers used for the amplification were designed to have 3' and 5' ends homologous to the ones of the entry vector. After having obtained the new plasmid we used it as a backbone for another *Gibson assembly* to introduce blasticidin resistance in order to be able to perform selection in mammal cells later on. The insert for this new assembly was obtained again by PCR of a Cas9 lentiviral vector having Blasticidin resistance, and the insert obtained has the *P2A* sequence followed by the sequence for the *Blasticidin* resistance, with 3' and 5' ends homologous to the vector in which they have to be inserted by *Gibson*. The last step to obtain the final vector was done by *Gateway Cloning* with another plasmid already present in the lab that contained the *CAG* promoter, the *attR1* and *attR2* sequences and the *LTRs* for the transposase integration, used for recombination by *Gateway* with the *attL1* and *attL2* sites present in our construct.

The engineered plasmid (Fig.4.1*b*) consist of a backbone containing

Ampicillin resistance with an insert having several modules: a CAG promoter; the RfCas13d or CasRx protein sequence; an SV40 Nuclear Localization Sequence (NLS); a P2A sequence; a Blasticidin resistance; attB sites and LTR sequences. The *CAG promter* is a synthetic promoter that produces high levels of expression [91] and it was found to be more stable and stronger than other commonly used promoters [92]. It is made by the fusion of the CMV (Cytomegalovirus) enhancer, the β-actin promoter from chicken and a chimeric intron made from chicken β-actin and rabbit β-globin [91]. Using this promoter we tried to solve the need for high expression of the CasRx protein needed for the system to work properly. The CAG promoter notwithstanding was shown to be poorly compatible with lentiviral delivery [93], therefore we decided to add transposon recognition sequences (LTRs) in order to overcome this problem by delivering the plasmid with transposition. The *RfCas13d* is the sequence that codes for the *Cas13* protein, which is the one that will produce the break in the targeted RNA leading to its knock-down. The RfCas13d protein is bound to a *SV40 NLS* signal, which is a *Nuclear Localization Signal* taken from *Simian Virus 40* to improve its knock-down capacity and allows us to target nuclear RNA transcripts as shown by [53]. The *blasticidin* resistance is used for the selection: by adding *blasticidin* to the growth medium, only the cells that have effectively integrated the plasmid are surviving, while the others are being killed by the Antibiotic drug. Between the *RfCas13d* element and the *blasticidin resistance* element there is a *P2A* sequence. This sequence induces the separation of the CasRx protein and the blasticidin resistance, both expressed under the same promoter, therefore in the same mRNA, thus avoiding the use of multiple promoters that would result in longer and more bulky plasmids [94]. The *attB*

sites are specific homologous recombination sequences that can be used for Gateway Cloning [95]. The LTRs are sequences recognised by a transposase that define the region of the construct that will be integrated by the transposase into the cells that will then be used for experiments. This construct was delivered by transfection. During the procedure two plasmids are reaching the cells, the first one consist of the construct just described, the second one is a plasmid containing a transposase. This plasmid was already present in the lab and encodes for the *piggyBac transposase*. Once transcribed, the transposase protein will recognize the *LTR* sites in the CasRx plasmid and integrate it into the genome of the targeted cells, thus providing constitutive expression of the Cas13 protein (Fig.4.1*c*).

To test the system, first, we transfected two different murine pancreatic ductal adenocarcinoma cell lines (PDAC) cells previously established in our laboratory, PPT8442 and PPT16992, with the transposon vector encoding the *CasRx* protein and the *hypbase* transposase plasmids. With this we expected the *CasRx* cassette (CAG promoter, CasRx and blasticidin resistance) to be delivered and integrated into the genome thanks to transposition performed by the PiggyBac transposase, allowing the permanent expression of the CasRx protein in the cells under the CAG promoter. Then we selected for the CasRx integration, using 10 µg/µL blasticidin. In this way, only the cells that integrated the Cas13 into their DNA survived, whereas in the case of the negative control, in which no transposase plasmid was transfected, no cell survived after selection (Fig.4.1*e*). This proves that our transposition-based delivery method is functional.

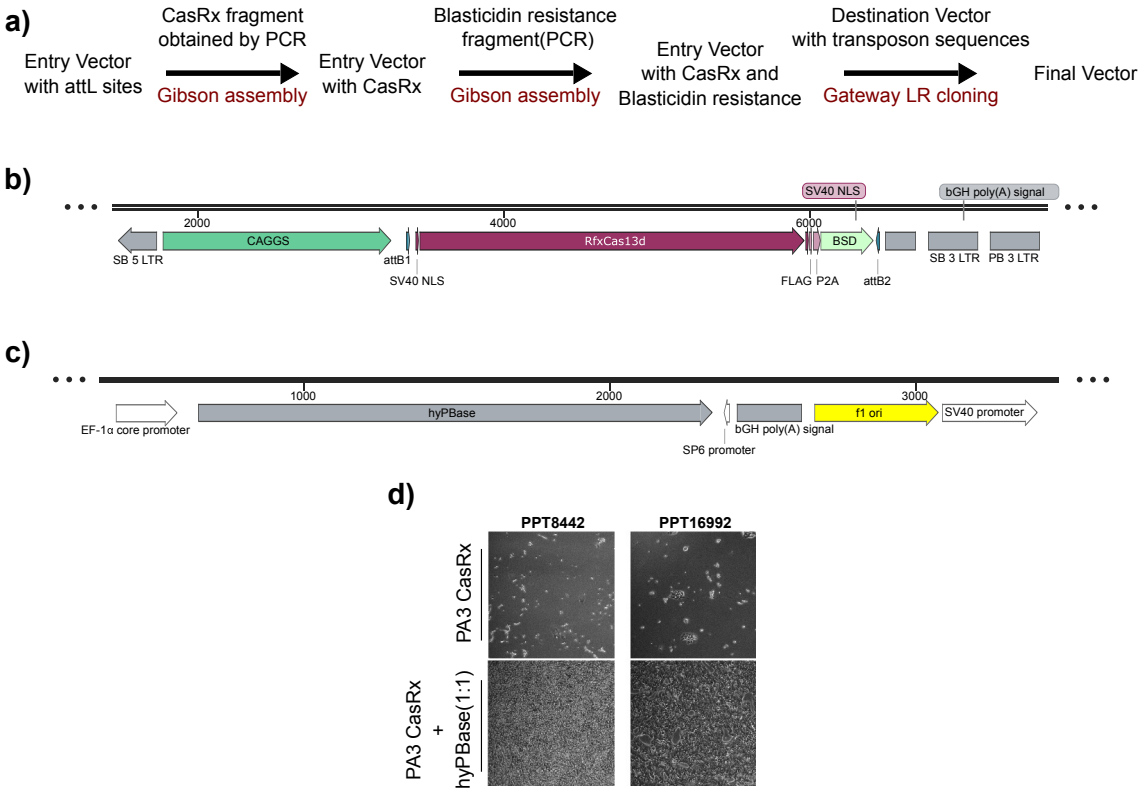

FIGURE 4.1: Cloning of the Cas13 vector: **(a)** flow chart showing the cloning steps to obtain the final CasRx vector: the CasRx fragment is inserted by Gibson assembly into the entry vector, the blasticidin resistance is then added with another Gibson Assembly and the resulting vector is then cloned by Gateway into a destination transposable vector to obtain the final vector **(b)** CasRx vector important elements, namely the CAGGS promoter, the RfCasRx protein sequence, P2A cleavage sequence, blasticidin resistance, SV40 nuclear localization sequence, attB sites and LTR sequences **(c)** Hypbase vector important elements, namely the transposase protein sequence **(d)** Microscope picture of PPT8442 and PPT16992 cell transfected with CasRx vector vs with CasRx vector and hypbase vector, the cells without the transposase vector don't survive selection with puromycin

4.1.2 Validation and Optimization

Once the CasRx vector was created, we optimized the system on the side of the sgRNA. The plasmid for the sgRNA was already present in the lab, the only thing that needed to be changed is the stuffer, that is the actual sequence of the

sgRNA plus the direct repeats and a final polyT stretch for RNApolimeraseIII termination, everything downstream a U6 promoter (Fig.4.2*a*). According to [84], a modified version of the sequence for the Direct Repeat (DR) of the sgRNA leads to a better performance in silencing for the Cas13 system. They tried different modifications of the sgRNA stem-loop and found that the disruption of the first base pair of the proximal stem results in improved knock-down efficiency, especially for low silencing efficiency guides. Thus, we performed an experiment to assess whether in our settings the system works better with the same modification of the *Direct Repeat* stem. First, we transfected the cells with *CasRx* and *hypbase* plasmids, in order to deliver the CasRx protein so that it would be constitutively expressed by integration in the genome thanks to transposition performed by the *PiggyBac* transposon. Then we selected, using blasticidin, only the cells that integrated the CasRx into their genome using non-transfected cells as control. Once the cells were selected (all the cells in the control well are dead) we proceeded with transduction of the GFP plasmid, which contains a GFP protein that is fluorescent and emits green light upon excitation. The GFP plasmid contains a gene that confers hygromycin resistance, so we used this antibiotic to select cells that correctly integrated the GFP plasmid. This double selection resulted in cells that have both CasRx and GFP plasmids integrated. Then, we transduced three separate batches of cells with different sgRNAs, one containing the standard version of the *Direct Repeat*; one the modified version and the last one a non-targeting sequence as the control and subsequently we measured GFP fluorescence level by flow-cytometry (Fig.4.2*b*). We performed the experiments on the *PPT8442* and *PPT16992 PDAC* (Pancreatic Ductal Adenocarcinoma) cell lines previously established in our

laboratory from solid tumours and we tried this approach with two different GFP targeting sequences as sgRNAs. From the results it can be seen (Fig.4.2*c* and Fig.4.2*d*) that by using the modified version of the *DR*, we can gain in knock-down efficiency, therefore we implemented the new version of the *Direct Repeat* in all our further experiments and in the guide design.

Our aim is to use two spacers in every sgRNA array in order to have a higher chance of having a good targeting sequence and a higher chance of targeting a functional isoform of each lncRNA, allowing to reduce the number of sgRNAs required and the size of the library. Therefore, we tested whether this setting could negatively influence the knock-down efficiency because of the bulkyness of the plasmid or the maturation steps performed. In order to do this, we performed an experiment on different cell lines comparing the use of one spacer per array with two spacers per array (Fig.4.2*e*). As a target, we used again *GFP*. We first established CasRx cell lines by monoclonal expansion (Fig.4.2*f*), following the procedure described in Chapter 3, and selected then the clones with the best knock-down efficiency measured by flow cytometry. We then performed the silencing in a total of 4 selected clones per cell line. Each clone was independently infected by a non-targeting guide; a sgRNA with only one spacer targeting GFP and sgRNA with two spacers both targeting GFP and the silencing was again measured by flow citometry and normalized by the control. The results (in Fig.4.2*d* and Fig.4.2*e* the cell lines *PPT16992* and *PPT16900* are shown) show that in all the selected clones and in both the cell lines the use of two spacers inside an array gives a knock-down efficiency comparable to the use of only one spacer, therefore we could proceed with the use of two spacers.

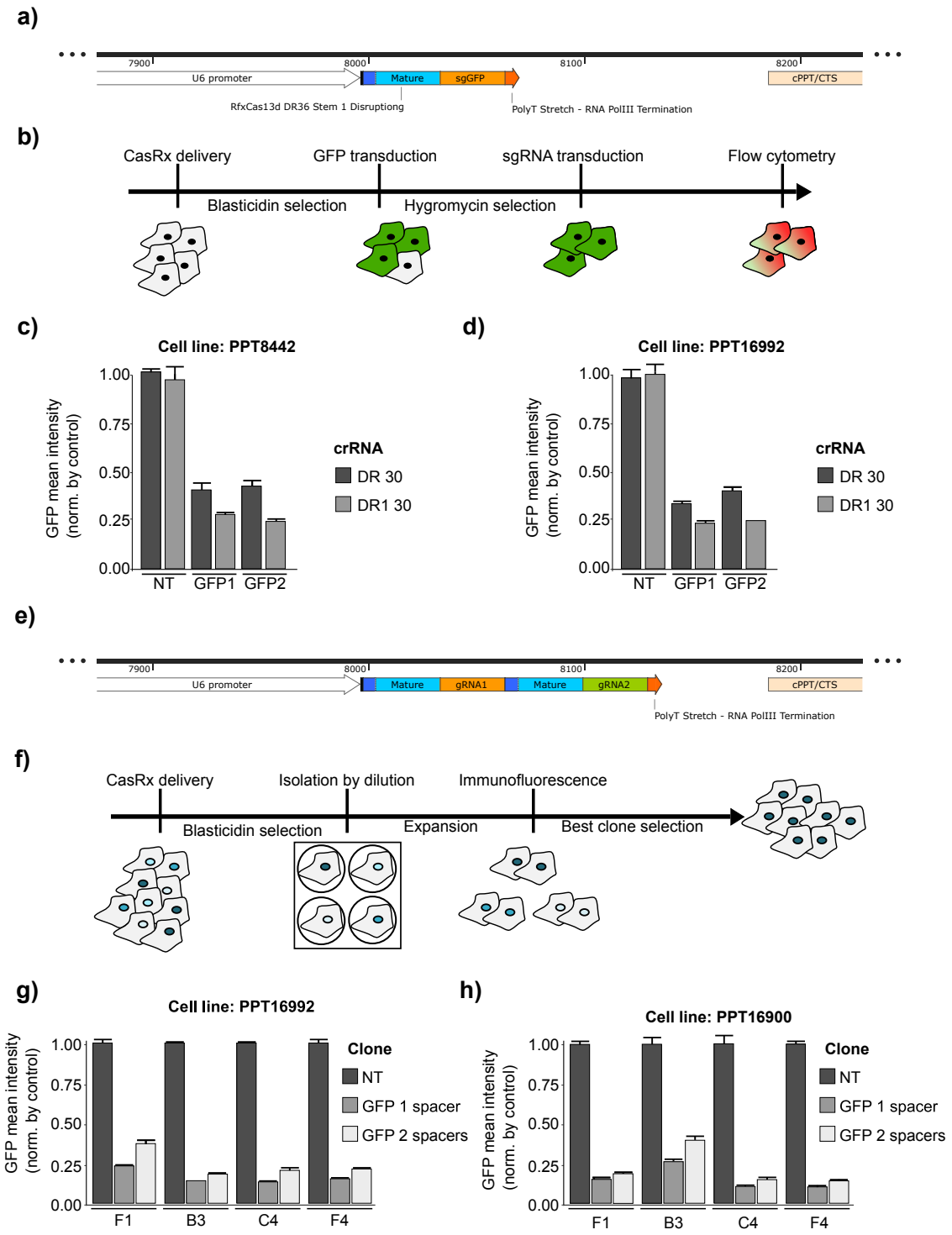

FIGURE 4.2: Optimization of the CasRx system delivery: caption on next page

FIGURE 4.2: Optimization of the delivery: **(a)** image showing the most important elements of the guide RNA plasmid with 1 spacer, namely the U6 promoter and the guide with the mature segment and the GFP targeting sequence highlighted **(b)** Flow chart showing the procedure used to test the two different direct repeat versions: the cells are transfected with CasRx and the transposon plasmid followed by blasticidin selection. The cells are then transduced with GFP and are selected with hygromycin. Finally, the sgRNA is transduced and the GFP intensity is measured with flow cytometry **c** GFP mean intensity normalized by non-targeting control of DR 30 and DR1 30 both with GFP1 gRNA and GFP2 gRNA in cell line PPT8442 **d** GFP mean intensity normalized by non-targeting control of DR 30 and DR1 30 both with GFP1 gRNA and GFP2 gRNA in cell line PPT16992 **(e)** image showing the most important elements of the guide RNA plasmid with 2 spacers, namely the U6 promoter and the guide with the two mature segments and the GFP targeting gRNAs sequences highlighted **f** flow chart showing the clonal selection procedure to select CasRx clones: CasRx is delivered to the cells with the transposase followed by blasticidin selection. A subsequent dilution is then performed and single clones are selected and expanded. The clones are transduced with GFP and the GFP intensity is assessed **(g)** GFP mean intensity normalized by the non-targeting control of the guide with one spacer (GFP 1 spacer) and the guide with 2 spacers (GFP 2 spacers) in the 4 best CasRx clones of cell line PPT 16992 **(h)** GFP mean intensity normalized by the non-targeting control of the guide with one spacer (GFP 1 spacer) and the guide with 2 spacers (GFP 2 spacers) in the 4 best CasRx clones of cell line PPT 16900

4.1.3 Generation of a CasRx transgenic mouse animal model

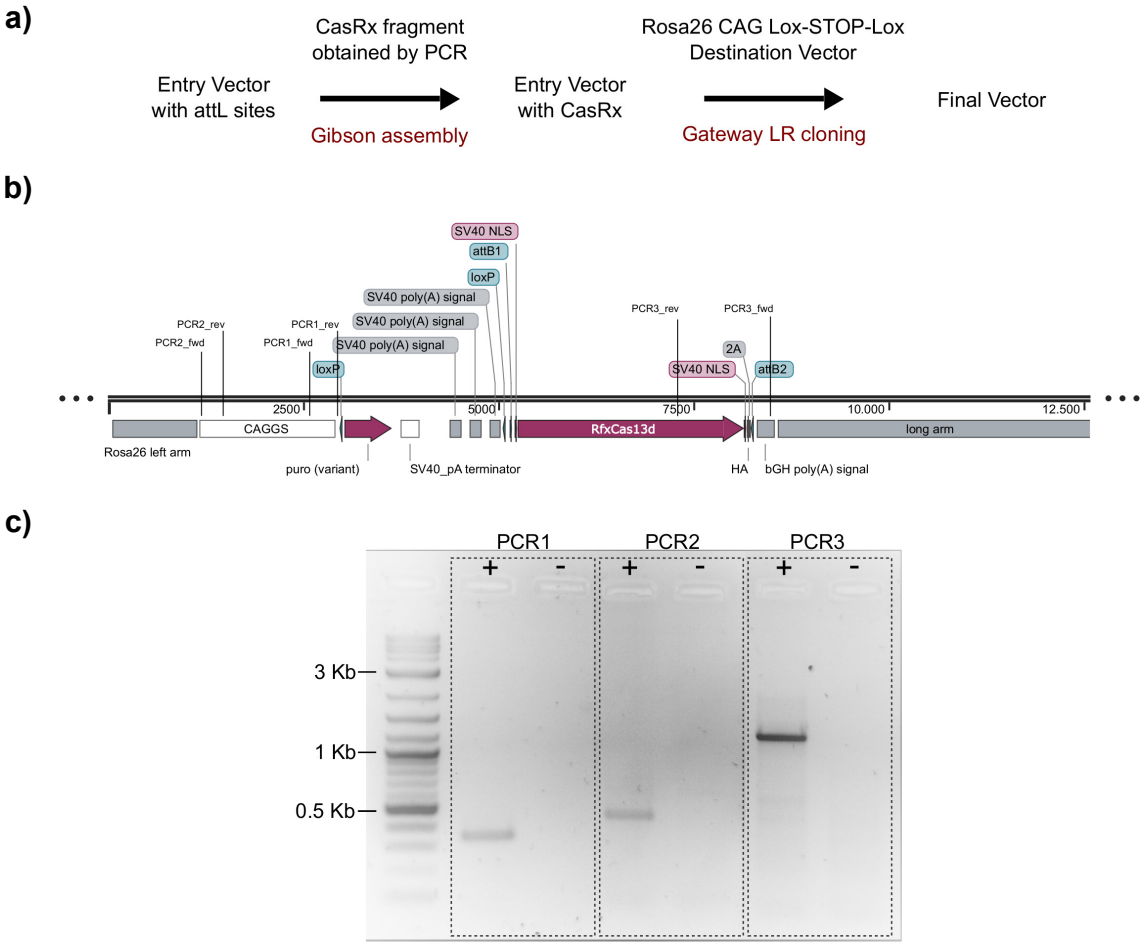

FIGURE 4.3: Generation of the CasRx mouse Model: **(a)** flow chart showing the cloning steps performed to obtain the Rosa CasRx vector; the entry vector with CasRx obtained in the previous experiment is cloned by Gateway in a Rosa26 destination vector specific for integration in the animal Rosa locus **(b)** image showing the important elements of the Rosa CasRx vector with the primers used to validate the PCR animal **(c)** Agarose gel electrophoresis of the PCR products obtained using primers PCR1, PCR2 and PCR3(showed in the vector in figure b) amplifying two portions of the CAGGS promoter and a portion of the CasRx protein

A future goal of the project is the use of CasRx animal models to perform in vivo experiments in order to validate interesting lncRNAs that might be functional in cancer and involved in tumorigenesis. This type of animal model will allow

removing the current barriers for the study of lncRNAs, generating clinical relevant data, which through preclinical studies will allow identifying lncRNAs as new therapeutic targets. For this purpose, we generated a targeting plasmid that was then used to generate a CasRx transgenic animal. This plasmid was designed to integrate the CasRx cassette into the *Rosa26* locus. The *Rosa26* locus is a widely used locus for integration and expression of transgenes in mice. Its advantage lies in the fact that it is a widely expressed locus in all cell types, it doesn't undergo silencing and its modification doesn't translate in loss of functional sequences [96]. The plasmid was obtained by Gateway cloning with the *CasRx* entry vector already cloned before (Fig4.1*a*) and a Rosa26 targeting destination vector alredy available in the laboratory (Fig4.3*a*). The vector obtained presents the following elements: CAG promoter; puromycin resistance; SV40 termination sequence; RfCas13d protein sequence; loxP sites; Rosa26 left and right homology arms. The Rosa26 left and right arms serve for the integration by homologous recombination of the construct into the Rosa26 locus. The loxP sites and the SV40 terminator sequence from the loxP-STOP-loxP (LSL) system. This system allows the conditional expression of the Cas13 protein only upon activation of the Cre recombinase which exceeds the fragment between the loxP sites, eliminating the SV40 termination sequence, thus allowing the expression of the Cas13 protein by the CAG promoter. When Cre is activated also the puromycin resistance sequence is excided (Fig4.3*b*).

The transgenic mouse was created by a collaborator. To get the transgenic mouse, the construct was injected in the male pronucleus of a zygote obtained by the mating of two mice. This zygote was put into a female mouse with progesterone-induced *pseudo-pregnancy* and the offspring were screened with

PCR to check whether they contain the transgene or not. The PCR confirming the transgenic genotype can be viewed in Fig.4.3*c*. We used three different sets of primers, two of them amplifying in the CAG promoter sequence and the third one amplifying in the CasRx sequence region. The length of the observed fragments corresponds to the expected length of the PCR products, confirming the genotype of the transgenic animal.

4.2 Generation of a panel of human cancer cell lines engineered with the CasRx system

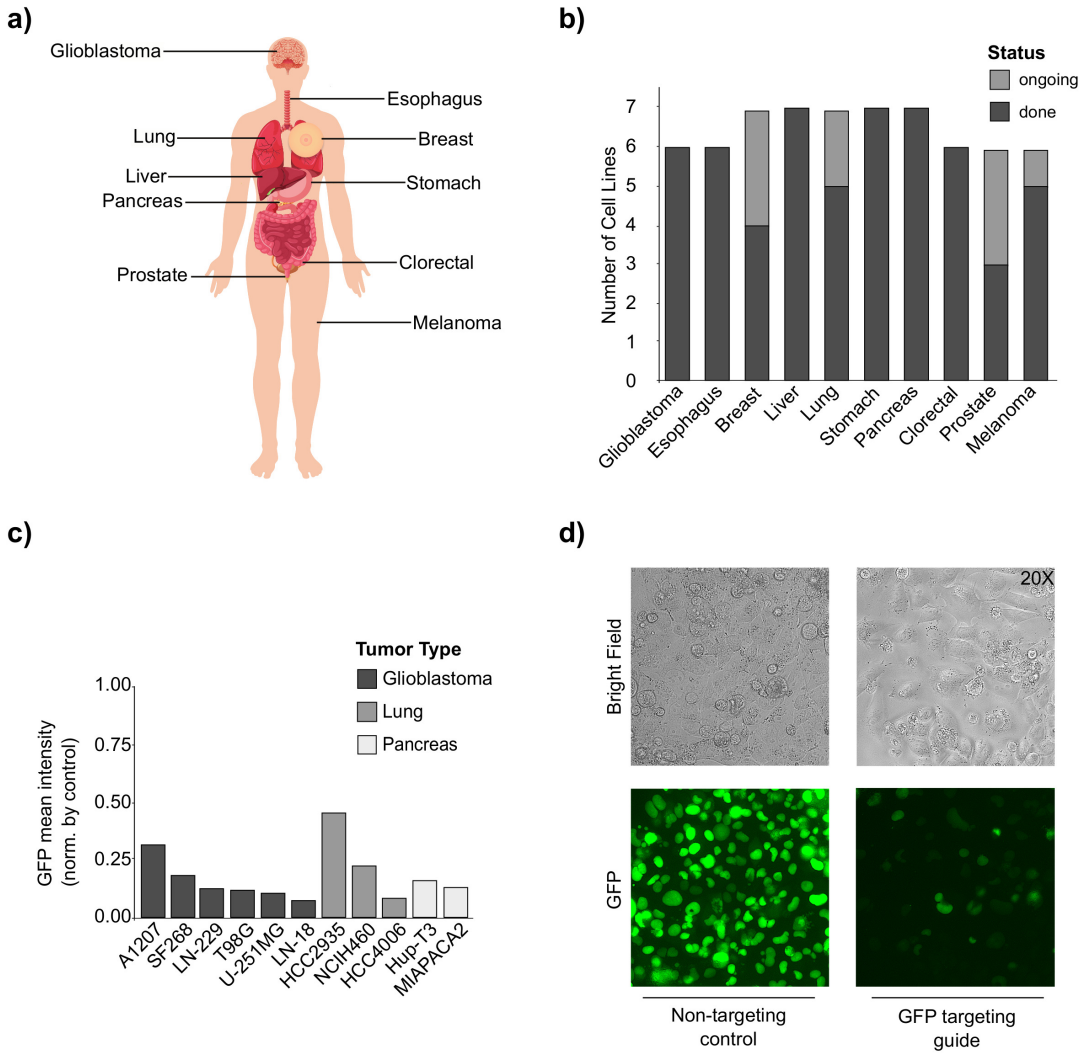

FIGURE 4.4: Establishment of the cell lines: **(a)** Image showing the tissues/organs from where our cancer cell lines are derived **(b)** Bar graph showing the status of the CasRx cell lines generation **(c)** Bar graph showing GFP mean intensity of the final validation knock-down experiment normalized by the control in the cell lines I personally generated **(d)** Microscope picture showing the comparison between the non-targeting and GFP targeting guides in MIAPACA2 final validation experiment

It is known that lncRNAs are heavily tissue type specific [97] and in several recent functional genomic screens, lncRNA hits were found only in single-cell lines, supporting this characteristic of the non-coding transcriptome and suggesting that a great number of lncRNAs might be functional in a cell-line or cell-type-specific fashion [27]. This means that performing a screen in just a few cell lines might leave out potential important hits. Therefore, we decided to generate cell lines for further screenings based on different tumor types (Fig.4.4*a*). We selected tumours affecting ten different human organs: Breast Carcinoma; Colorectal Adenocarcinoma; Esophagus Carcinoma; Glioblastoma; Liver Carcinoma; Lung Carcinoma; Melanoma; Pancreatic Ductal Adenocarcinoma; Prostate Carcinoma and Stomach Adenocarcinoma. We chose these tumours among others because of their higher incidence and mortality, thus increasing the clinical impact of our further discoveries. We excluded haematological tumours since the screening technique is quite different between adherent and not adherent cell lines. In summary, we selected the top 10 more clinically relevant solid tumours. For each of the tumour types listed above, we aimed to have from five to six human cell lines with high CasRx activity. To be on the safe side we selected 8-9 cell lines for each type, discarding afterwards the ones with low knock-down efficiency. All the cells were chosen from the CCLE database. The Cancer Cell Line Encyclopedia (CCLE) is a project that started in 2008 with the aim of genetic characterization of 1000 cancer cell lines. It is a database of Cancer cell lines obtained from patients on which a set of different genomic analyses, such as RNA-Seq or ChIP-seq, were performed, and this data is openly available for download [98]. There were three main criteria for a cell line to be selected. First, it should have sequencing data on the CCLE database;

second, it should already have a successful screening performed on the DepMap portal, meaning that a genetic perturbation system can work in this cell line; third, that we are able to get access to the cell line.

The generation of the CasRx cell lines is a time consuming and laborious process consisting of multiple steps: after thawing and expanding the cells, they were transfected with the *CasRx* plasmid and with the transposon plasmid at the same time using lipofectamine 3000, thus resulting in genome integration of the CasRx cassette. The transfection was followed by the selection of the cells that integrated the construct with blasticidin. We used non-transfected cells as a negative control. Once the cells were selected and we ended up only with cells containing the CasRx protein we could proceed with clonal expansion (Clonal expansion was performed as described in chapter 3) which results in cell line clones that express the CasRx all in the same manner. Once the clones were generated we performed a preliminary validation and a final validation. The preliminary validation had the purpose of giving a general idea on which of the many clones could have high knock-down efficiency, thus restricting the number of clones to expand and in which the more laborious final validation is performed. To perform the preliminary validation we infect the cells with both GFP plasmid with hygromycin resistance and GFP targeting sgRNA plasmid without selection. This results in a heterogeneous pool of cells where we find cells without any plasmid, cells expressing only GFP or only the sgRNA and cells expressing both the GFP and the sgRNA plasmid. Given that the sgRNA expresses also the *far-red* Fluorescent Protein we can detect using flow cytometry both green and far-red. This allows us to compare cells with only GFP and no sgRNA (by gathering only cells emitting green) with cells with both GFP and the

sgRNA (emitting both green and far-red) and measure the amount of knock-down of the green signal. With this, we can select the 4-6 best silencing clones with which to perform the final validation step. This final validation step consists in transducing the CasRx cells with the GFP plasmid and selecting for roughly one week using Hygromicin to have almost all of the cells expressing GFP. After that, we split the individual clones into two wells. The first well was transduced with a sgRNA targeting GFP and the second one with a non-targeting sgRNA. By flow cytometry, we measured the intensity of GFP in the well with the non-targeting guide and in the well with the GFP targeting one in the far-red positive cells (sgRNA plasmid has far-red protein expression). The comparison of the GFP intensity between both conditions allows us to calculate the level of knock-down performed by the CasRx system. This process allows us to select the clone with the best silencing, which is the one that can then be used to perform screenings. Once the clone is selected, it is further expanded and frozen in liquid nitrogen for further use. At the current state, almost all the cell lines have been generated (Fig.4.4*b*), only breast, lung, prostate and melanoma are missing one or two cell lines to be completed, which is planned to be finished in the next two months. Of the 53 established CasRx cell lines, I personally generated twelve of these (Fig.4.4*c*), of which five are Glioblastoma lines (A1027, SF268, LN-229, T98G, U-251MG, LN-18), three are Lung Carcinoma lines (HCC2935, NCIH460, HCC4006) and two are Pancreatic Ductal Adenocarcinoma cell lines (Hup-T3, MIAPACA2). The GFP mean intensity after targeting resulted in very low levels, corresponding to a knock-down efficiency almost in every case higher than 75%. Only one of the cell lines, namely the Lung Carcinoma line HCC2935 showed a knock-down efficiency closer to 50%.

Moreover, Glioblastoma and Pancreas Cas13 cell lines show a more homogeneous knock-down level, while in Lung we find quite heterogeneous numbers. As an example, in Fig.4.4*d* a fluorescence microscope picture of the final validation of the MIAPACA2 Pancreas cell line is shown. It can be seen that the GFP signal undergoes extensive silencing, producing a highly noticeable and clear difference between the non-targeting control and the GFP targeting sample. The knock-down efficiency that was measured for the shown sample is 87%.

4.3 Similar lncRNAs fusion into custom lncRNAs families

After the generation of high knock-down efficiency CasRx cell lines, our next goal was to generate a CRISPR sgRNA library targeting different lncRNAs. In comparison with previous lncRNA CRISPR screenings, that generated libraries targeting well-annotated transcripts expressed in a few numbers of cell lines [25, 74, 99], we wanted to generate a more unbiased library that can be used across as many solid tumour types as possible. Also, to increase the ratio of new discoveries, we didn't focus on small databases with few but well-annotated transcripts, but we tried to include as many different lncRNAs as possible. For this reason, we created our own collection of lncRNA families based on transcripts coming from different databases. We started by selecting all the RNA sequences included in RNAcentral. *RNAcentral* is a database maintained by the European Bioinformatic Institute and the Wellcome Foundation that collects non-coding transcripts from several databases and integrates them, facilitating

their use for research; from all these transcripts we selected the lncRNAs [100]. The main databases in *RNAcentral* are LncBook [101], GeneCards [102], LNCipedia [103], NONCODE [104], ENSEMBL [88] and RefSeq [105]. Most of these databases are collections of lncRNAs without manual curation, so they have the advantage of having a greater amount of transcripts, but the lack of manual curation can also mean that several of these transcripts are junk RNAs, in most of the cases automatically assembled by an algorithm by mistake, and often presenting redundancy. ENSEMBL and RefSeq, instead, are much smaller databases, but all the transcripts have been manually curated and well annotated, and consequentially enriched in more complex transcripts, hence we can say that our collection of lncRNAs present both these types of transcripts, allowing us to screen well-annotated and curated transcripts along with several novel transcripts that, even though less accurate, could more easily lead to the discovery of novel functions. Adding up to these RNAs, a set of evolutionary conserved non-coding transcripts were taken from six different papers analyzing lncRNA conservation [33–38]. Each of these papers analyses lncRNAs conservation spanning several million years of evolution and could contain important information about the functionality of the lncRNAs since it is known from the literature that lncRNAs are poorly conserved and the ones that have been found to have important roles in several cellular processes are instead more conserved than the general trend [33]. Besides this, in order to perform relevant in vivo experiments with mouse models, we need to have mouse-human conserved transcripts. From each of the listed papers, we selected all the lncRNAs that were considered conserved during human genome evolution according to their evolutionary conservation analysis. These RNAs, besides

adding transcripts to our collection, are also used to investigate for conservation of our lncRNAs. Several of the transcripts from this big collection are identical to each other or share a high amount of similarity (e.g. high percentage of exon overlapping), suggesting them coming from the same original transcript undergoing alternative splicing or RNA processing in general. Based on these observations, we thought that lncRNAs presenting these characteristics should be grouped together and be considered a unique family, in order to more easily screen them. Since no previous script or program was fulfilling our needs, we created a series of scripts that perform different steps of merging the transcripts based on shared characteristics. All the scripts are written mainly in *R* and partly in *bash* and are run on a computer cluster that has 96 cores and 256 Gb of RAM. To maximize the power of the cluster we parallelized the calculations by running the scripts for each chromosome and each strand separately.

The main criteria upon which two or more transcripts are fused together is the sharing of exons. Since a lot of transcripts are presenting overlap of exons that are on the extremes of their transcript body but are clearly separated and do not belong to the same transcript family, we could not use this as the sole criteria of merging, therefore, there are two main steps before considering exons overlap, that is carried on taking into consideration the whole transcript body, without dealing with individual exons and introns. First, the transcripts are grouped based on the overlap of the gene body: if the transcripts overlap for more than 50 bp they are assigned to the same group. This step is simply carried out using the *merge* option of the BedTools tool [106]. Then our custom scripts come into play. The first script is still considering only the transcript bodies and is refining the groups, splitting them into more sub-groups based on

overlapping percentages. This step is carried on by comparing always two consecutive transcripts at a time, using sorted genomic locations to speed up the process. If the shortest transcript is covered by the longest one for at least 25% of its length, then the two transcript bodies are considered to be part of the same sub-group. The value of 25 was decided after several test runs with different percentages where 25 was the value that was fusing the transcript bodies in a closer way to what we would have done manually, a higher number would be too strict, separating transcripts that should have been merged, and a lower number too loose, merging things that were not supposed to be together. An example of the first fusion step performed by our custom script is shown in Fig.4.5*a*, where we can see the orange and green transcripts being considered separately even if they are sharing one exon because the gene bodies are only overlapping in their extremes. In these subgroups, there are still several transcripts that have nothing to do with each other, namely, transcripts that are in the introns of others and that, more in general, though having an overlapping transcript body, don't have enough exon overlapping to be considered part of the same family. Running the scripts for the two different strands separately also avoids overlapping transcripts coming from different strands clashing with each other. At this point, we run the third script. This script is not considering the transcript body anymore, but the exons of each transcript within a sub-group of transcripts. Each exon is compared two by two with all the other exons in the same sub-group in an exhaustive manner, and if there is an overlap of at least 60% of the exon length of the bigger exon then the two transcripts that have the compared exons are merged together. Also in this step several test runs with different percentage values were run, both considering once the shortest exon

and once the longest, after manual evaluation of how the different parameters were behaving we decided to use 60% as a percentage of exon overlap and to choose the longest exon to apply this percentage to. Once these fusion steps were done we had to fix some problems arising due to the fact that we were using the bigger exon as a comparison for the percentage. First, we found that some monoexonic transcripts that were overlapping to one exon of a bigger transcript were still maintained separated. We, therefore, wrote another script to fuse these monoexonic transcripts back to the transcript they were overlapping to, using a minimum overlap of 30% (Fig.4.5*(b)*). Second, a similar problem arised from transcripts with more than one exon, but whose exons were also overlapping almost completely other exons of other transcripts. A script also solved this problem by fusing them (Fig.4.5*(c)*). This script used an overlap percentage of 90% as a criterion to merge the transcripts applied to total transcript exome.

Since in general the average number of reads covering lncRNAs is much lower than the one covering protein-coding genes [97] overlapping with exons of coding genes, especially in an unstranded RNA-seq setting, in the same or opposite strand would lead to disproportionate read counts in lncRNAs, overestimating the expression of those lncRNAs. Therefore we sought to trim our transcripts given the amount of overlapping with coding genes exons. We removed all the bases that were completely overlapping with an antisense protein-coding gene and then controlled how much of the original transcript length was left. If the transcript was trimmed for more than 60% we eliminated that transcript family for our list (Fig.4.5*d*).

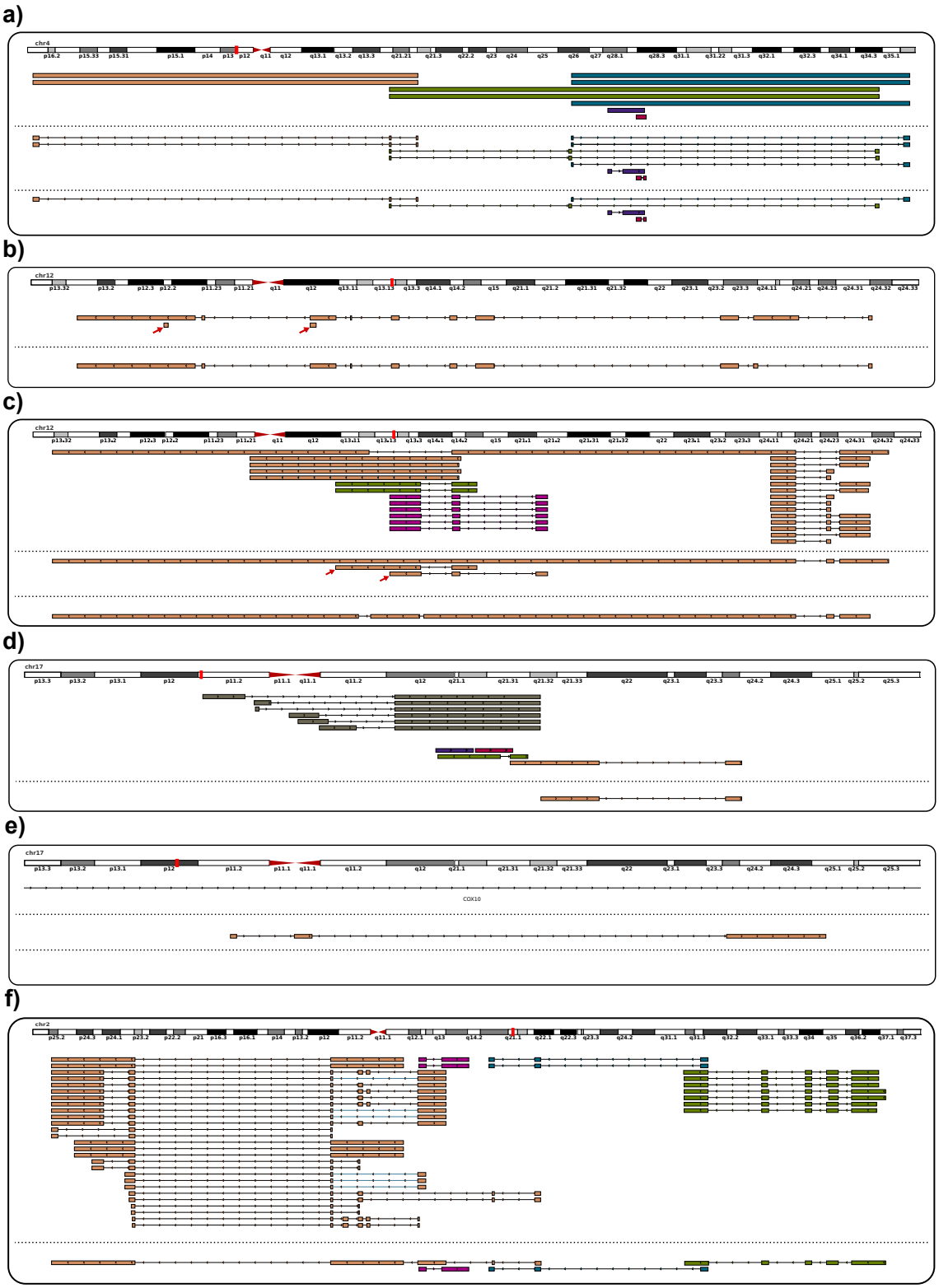

FIGURE 4.5: Transcripts fusion: Legend on following page

FIGURE 4.5: Transcripts fusion: **(a)** genome viewer image showing the merging step of the lncRNAs into different macro-groups based on the overlap of at least 25% of the entire gene body and the merging of the transcripts in this macro group based on exon similarity **(b)** genome viewer image showing the removal of monoexonic transcripts at least 30% overlapping with an exon of another transcript **(c)** genome viewer image showing the removal of multiexonic transcripts with an overlap of at least 90% with one or multiple exons of another transcript **(d)** genome viewer image showing the trimming of transcripts that are overlapping with coding genes exons on any of the strands **(e)** genome viewer image showing the trimming of transcripts that are overlapping coding genes introns on the same strand **(f)** genome viewer image showing a representation of a region in chr2 with different transcripts after performing all the merging steps

Since the CasRx system works in the nucleus, besides targeting exons, it could possibly target also introns before splicing. Because of this, the targeting lncRNAs transcribed in the same location and orientation as the introns or the exons of a coding gene can give false positives, since potentially we could be also targeting the coding gene. Consequently, for lncRNAs overlapping coding genes, we applied the same strategy that we used for anti-sense transcripts, with the difference that this time also base pairs overlapping with introns are trimmed (Fig.4.5*e*). An example of the result produced by all these scripts in a region in q13.13 in chr12 is shown in Fig.4.5*f*. We can clearly distinguish four main transcript groups that are split apart by the algorithm according to gene body overlap and then fused within each group according to exons overlap, obtaining four merged transcripts.

After obtaining this list we went through a step of manual curation. Specifically, we checked transcripts that had become very long or that gained a high amount of exons after fusion. They were mainly coming from Necsulea

paper since they have a collection of very complex transcripts. We found 33 of these transcripts and removed them maintaining the conservation information in the family they were part of.

We then performed some quality control and analysis on our list of lncRNA families. First, we wanted to check the complexity of our custom lncRNAs families by checking the number of transcripts for each number of exons (Fig.4.6*a*). We can see that the most common transcripts in our list are, respectively, biexonic and monoexonic transcripts, with the majority of the lncRNAs not exceeding 10 exons. The same goes for the transcript length (measured by summing up the exons lengths), where we can see that the majority of our collection doesn't exceed 2500 bp of length (Fig.4.6*b*). Both of this information is in line with the current knowledge about lncRNAs since it's now well known that these types of RNAs usually have a low number of exons (generally one or two) and are on average 1-2 kb long [107]. Since each of the new lncRNA families maintains the information about the database of origin of all the transcripts that contributed to form it, we analyzed the list using *Venn Diagrams* and *UpSet plots* [108] to have an idea about which databases contributed the most to our collection and from which of the papers were the conserved transcripts coming. Concerning the databases, most of the lncRNAs are coming from LncBook, GeneCards and LNCipedia, with lncBook alone contributing for the majority (Fig.4.6*e*). This is consistent with the fact that those are the three biggest databases. Taking apart LncBook we can see that a big amount of our lncRNA families are shared between all the different databases (we considered that a lncRNA family is shared when at list a lncRNA from a database is fused with one from another or several others databases. Take into

consideration that this doesn't mean that the databases have exactly the same transcript, but just highly similar ones). Regarding conserved lncRNAs, also here the major contribution is carried over by the publications with the higher number of transcripts, namely Necsulea and Sarropoulos [33, 34], with only a small amount of overlap between the two of them (Fig.4.6*c* and Fig.4.6*d*). This shows the low consensus, and therefore low overlapping of sets, between what is considered conserved in the different papers, highlighting the difficulties still present in evolutionary analysis of lncRNAs in comparison to those of coding genes. In total, in our collection of families the non-conserved lncRNAs represent the majority, precisely the 82.3%, while the remaining 17.7% is representing evolutionary conserved transcripts. Of these, 14.1% are obtained by merging transcripts from both one or more papers and one or more databases, and 3.7% being exclusively unique from the papers (Fig.4.6*f*). These last small group of transcripts could be of great interest if found functional, since they have never been characterized before in any other study, thus representing potentially new lncRNAs.

To further analyze our collection of lncRNAs we decided to use RNA-Seq data from CCLE. We downloaded the data and then used a custom pipeline that processes it by trimming it first with Trimmomatic [109] and then by mapping it to the GRCh38 genome with the splice aware aligner *STAR* [110]. To retrieve the raw counts for our lncRNAs we used *feature counts* [111]. By default, *feature counts* assign a read to a transcript only if it can be uniquely assigned to it, so reads that fall in regions of RNAs that are overlapping with others, which is not uncommon in our case, are discarded. Nevertheless, we need that information to be kept, so we decided to run *feature counts* with different settings. In our run,

if a read falls in a region where two transcripts are present, this read is assigned two times, one for each transcript. In this way, in regions where two transcripts are slightly overlapping and one is highly expressed, also the other transcript expression level will be brought up artificially. This might be seen as a problem, but also allows us to enrich all the transcripts present in highly transcribed loci. Once we obtained the count table we performed a PCA (Principal Component Analysis) (Fig.4.7*a*) and we calculated sample distances in order to generate a heatmap to show how the samples are clustering together (Fig.4.7*b*). The first thing we noted in the PCA is that non-solid tumours (e.g. AML) are forming a clear cluster far away from the solid-tumour samples (Supplementary Figure 1). Given this and since we aim to create a lncRNA library for CRISPR screening of solid tumours, we removed from the CCLE data all the non-solid tumour types, all being from blood or immune cells. This prevented us from having a bias driven by the non-solid tumour data in our selection process, enriching for RNAs that might not be expressed in solid cancer types. The PCA, even without non-solid tumours, still struggles to cluster all the tumour types clearly. It is noticeable that samples from the same tumour type are close together, but a lot of overlap between the tumour types is still present, we need to consider, however, that the *principal component 1* and the *principal component 2* of the PCA represent only a very small fraction of the total variance. In addition to this, we found that in a paper from a different laboratory making a PCA of the data from CCLE [112] they obtained a cluster similar to ours. The sample clustering shown by the heatmap is clearer and well defined. We can clearly distinguish different tumour types clustering together, and even though there are samples that are sparse and not clustering with the same tumour type, several of them are

coming from tissues close together that share a similar ontogeny, one example being samples from Esophagus clustering with cells derived from Lung.

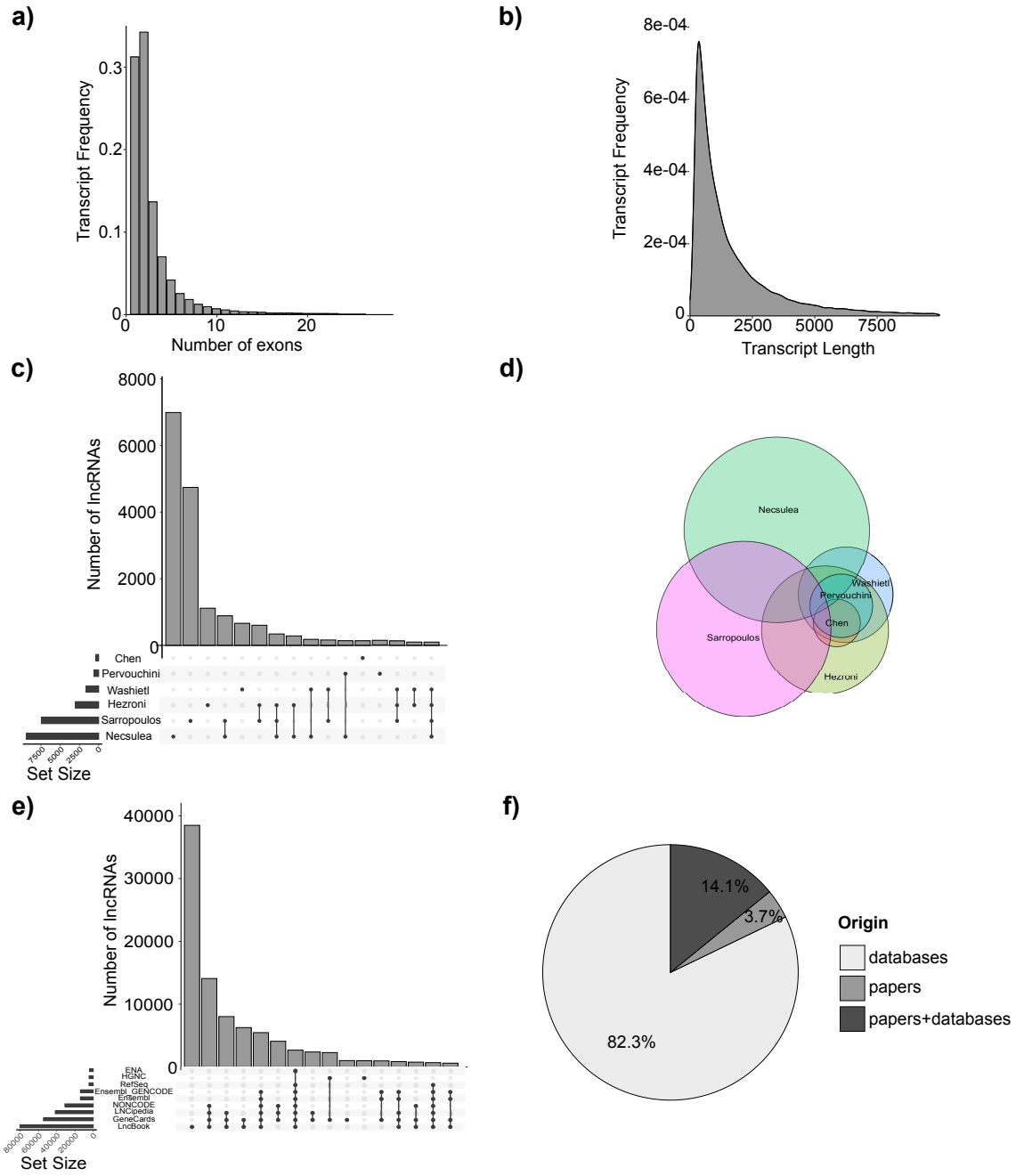

FIGURE 4.6: Analysis of the lncRNA list: **(a)** bar graph showing the frequency of transcripts by number of exons of our transcript collection **(b)** bar graph showing the frequency of transcripts by transcript length of our transcript collection **(c)** UpSet plot showing the overlap between transcripts coming from the six papers on evolutionary conservation **(d)** Venn diagram showing the overlap between transcripts coming from the six papers on evolutionary conservation **(e)** UpSet plot showing the overlap between transcripts coming from the major databases in *RNA central* **(f)** Pie Plot showing the origin of the fused transcripts in percentage on the total amount

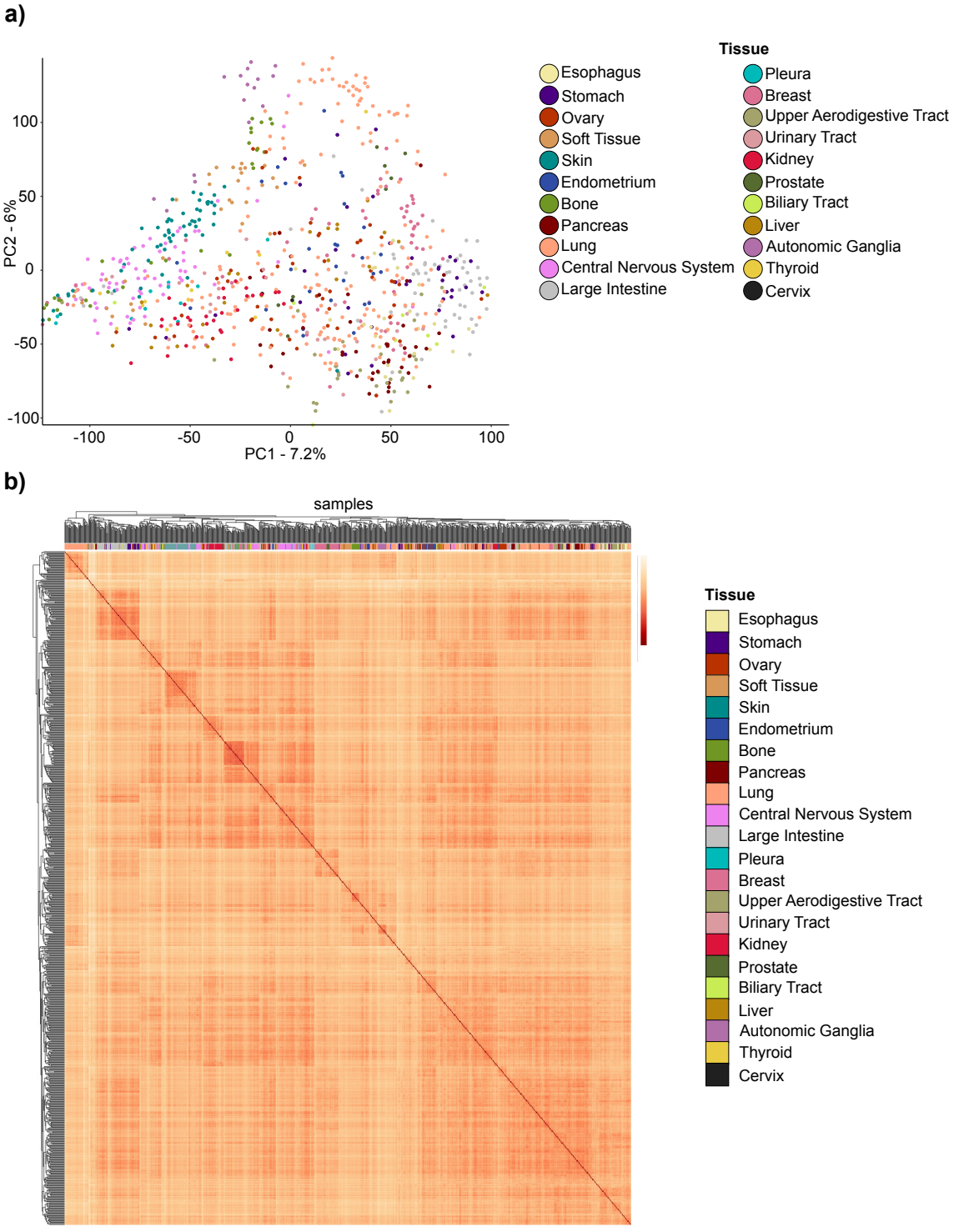

FIGURE 4.7: PCA plot and Heatmap clustering samples: **(a)** PCA plot showing the different CCLE samples by tissue type after removal of non-solid tumor types **(b)** Heatmap showing sample-to-sample distance clusters after removal of non-solid tumor types

Since several transcripts were in repetitive regions or were overlapping too heavily with others and to improve the quality of our collection of lncRNA families, before the selection of the lncRNA families to screen, we performed an off-target filtering with bowtie2 [113] to remove all the possible 30mers (sgRNA length) that were mapping multiple locations. This resulted in losing almost every 30mer in heavily overlapping transcript or RNAs laying in highly repetitive regions. We used this off-target filtering step in order to perform a final manual curation step on the transcripts that were losing too many guides due to overlapping with other sequences. To do this we selected all the transcripts that were losing more than 90% of their guides, filtered only the ones that were having this loss due to overlapping with another transcript and merged them based on how similar they were to each other. With this manual curation step we checked and merged 427 lncRNAs, we hence obtained a new improved collection of transcript families.

4.4 Generation of a CasRx sgRNA Library Design Tool and it's general principles

Once we obtained our list of transcripts we had to select how many of them to target and how to design the sgRNAs. While CRISPR/Cas9 screenings have a well-established methodology in functional genomics studies [73], such a big scale screening of lncRNAs with the CRISPR/Cas13 system is a novelty in the field [56], therefore, there aren't many tools available for the design of guides for the Cas13 system. In a recent paper published in Nature Biotechnology, Wessel

et al. [84] performed a screening using GFP targeting gRNAs to assess the parameters affecting gRNA knock-down efficiency. They then used these parameters (e.g. secondary structure Minimum Free Energy) to feed a Random Forest algorithm, which they found to be the best machine learning algorithm in terms of accuracy of prediction for this purpose. They thus developed a guide prediction tool capable, given an RNA sequence, of finding the best possible sgRNAs to target that specific transcript. The tool gives back to the user a ranked list of all possible sgRNAs targeting the desired transcript with a score based on the random forest algorithm, the higher is the score the higher is the predicted knock-down capability of the guide. In order to be able to perform a more complex design of guide RNAs, we wrote a program starting from the [84] source code, in the form of both a command-line tool for more experienced users and of a graphical application realized with R shiny for researchers that are not comfortable with bash scripting or more in general with bioinformatics. This tool allows a much more complex design of sgRNAs. As the first step, the program is running part of the code from [84] in order to predict the score of all the possible sgRNAs for the desired lncRNA and return a list of ranked guides. From this list, specific guides are chosen according to different design parameters (Fig.4.8*a*).

The program requires the user to give as input one of three possible file types: a list containing ENSEMBL IDs of transcripts; a CSV (comma separated) file containing a name in the first column and sequence in the second column or a CSV file containing genomic coordinates in the format IDs-chr-start-end-strand (Table4.1). The file has to be loaded and the type of file has to be prompted by the user using a specific multiple selection window

(Fig.4.8*a*.I). In the case a file with genomic ranges is provided, the user has a

TABLE 4.1: File Formats Allowed

ENSMUST00000223470
ENSMUST00000201192
ENSMUST00000183905

Name1	ATTCCGTCTACG...
Name2	ATTCCGTCTACG...
Name3	ATTCCGTCTACG...

ID1	chr1	33542065	33542117	-
ID1	chr1	33539171	33539356	-
ID2	chr7	22769545	22769987	+

panel to decide whether to retrieve the sequences for those regions from the mouse build GRCm38 or the latest human build GRCh38 (Fig.4.8*a*.II). Subsequentially, the user has to decide the number of guides (Fig.4.8*a*.III) and of spacers (Fig.4.8*a*.IV) for each guide that the program has to design. Usually, in genomic screenings three to four guides targeting the same transcript are designed, this is made to have more statistical power when the downstream analysis of the screening is done. If we want to perform screenings with an array format (i.e. multiple spacers per sgRNA targeting different regions of the same transcript) we will specify more than one spacer per guide. The user can then decide the length in bp (base pairs) of the sgRNA (Fig.4.8*a*.V) and also have an extension parameter (Fig.4.8*a*.VI) to prompt the number of base pairs up to which the sgRNA will be extended, this bases will be added to the 3' end of the predicted sgRNA. This is because when we use multiple spacers for each sgRNA the protein performs a maturation step consisting in cutting part of the guide RNA and part of the direct repeat, so providing an extension will just add the base pairs that the protein needs to cut in order to create the mature guide [53, 114]. Next, to target evenly every transcript, the user can decide how far

away from each other each of the sequences used as spacers should be. There is the possibility to do this separately for the distance of spacers within each guide, that we named *local distance* (Fig.4.8*a*.VIII), and the distance between all the spacers from different guides, called *global distance* (Fig.4.8*a*.IX). Both types of distance, local and global can be provided either in *base pairs* or in percentage of transcript length (Fig.4.8*a*.VII). Giving the value in percentage is of great importance if there is a big variation of length in the transcripts to be screened, so that every transcript is covered homogeneously by the targeting guides, independently of its length. In addition to this, the minimum distance is always the length of the spacer, so even if the distance prompted by the user is zero it is set automatically to match the spacer length, thus avoiding any kind of overlapping between the guides. Once the local and global minimum distances are set, there is the possibility to tell the program to avoid designing guides over the exon-exon junctions (Fig.4.8*a*.X). This is only possible when the file provided is in the genomic coordinates format, because only in that format the information about the position of the splice site is maintained. Furthermore, we need to avoid the off-target effect. Even knowing that the Cas13 protein is a much more specific system than RNA-interference we decided to give the possibility to remove from the guide pool all the guides presenting fully matching off-targets. To achieve this, we use bowtie2, which allows us to fast map short sequences onto a reference genome/transcriptome. The reference transcriptome was obtained by fusing all the sequences of our collection of lncRNAs with all the coding transcripts and non-coding transcripts arising from coding genes loci from the Ensembl database. In this way not only we can avoid targeting sequences that are mapping in multiple locations in the genome, but

we also avoid targeting regions where more than one lncRNA is present, meaning that for each of the lncRNAs present in our list, we only target regions that are unique for it and not overlapping with other transcripts. This step serves also as a filtering step, removing all those transcripts that are impossible to design due to heavy overlap with others. To filter for multimappers the user has to tick the corresponding box in the program (Fig.4.8*a*.XI). Considering that the library has to be cloned and that the methods to clone it can differ and can usually involve the use of restriction enzymes, the user has also the possibility to tell the program to avoid using guides that have a specific restriction enzyme cutting sequence. This will prevent the enzymes used in the cloning process to cut also the guide RNAs resulting in non-functional guides. Every kind of canonical DNA sequence can be provided (Fig.4.8*a*.XII), thus this functionality is not restricted to restriction enzyme cutting sequences filtering but can be used to filter any kind of sequence. At last, it is also possible to adjust the range of the transcript at the extremes to avoid designing guides to close to the 5' and/or 3' ends of the RNA sequence (Fig.4.8*a*.XIII). This could be used, for example, to avoid targeting the TSS (transcription start site). The command-line alternative to the program works the same way, but each of the arguments must be given as options of the bash script in the shell. Considering that in our and most likely in other cases the number of targets can reach a big number, we heavily parallelized every process in the program, that is capable to use as many CPU cores as the system in which we run it allows. This makes our program extremely fast, especially if used in HPC clusters.

a)

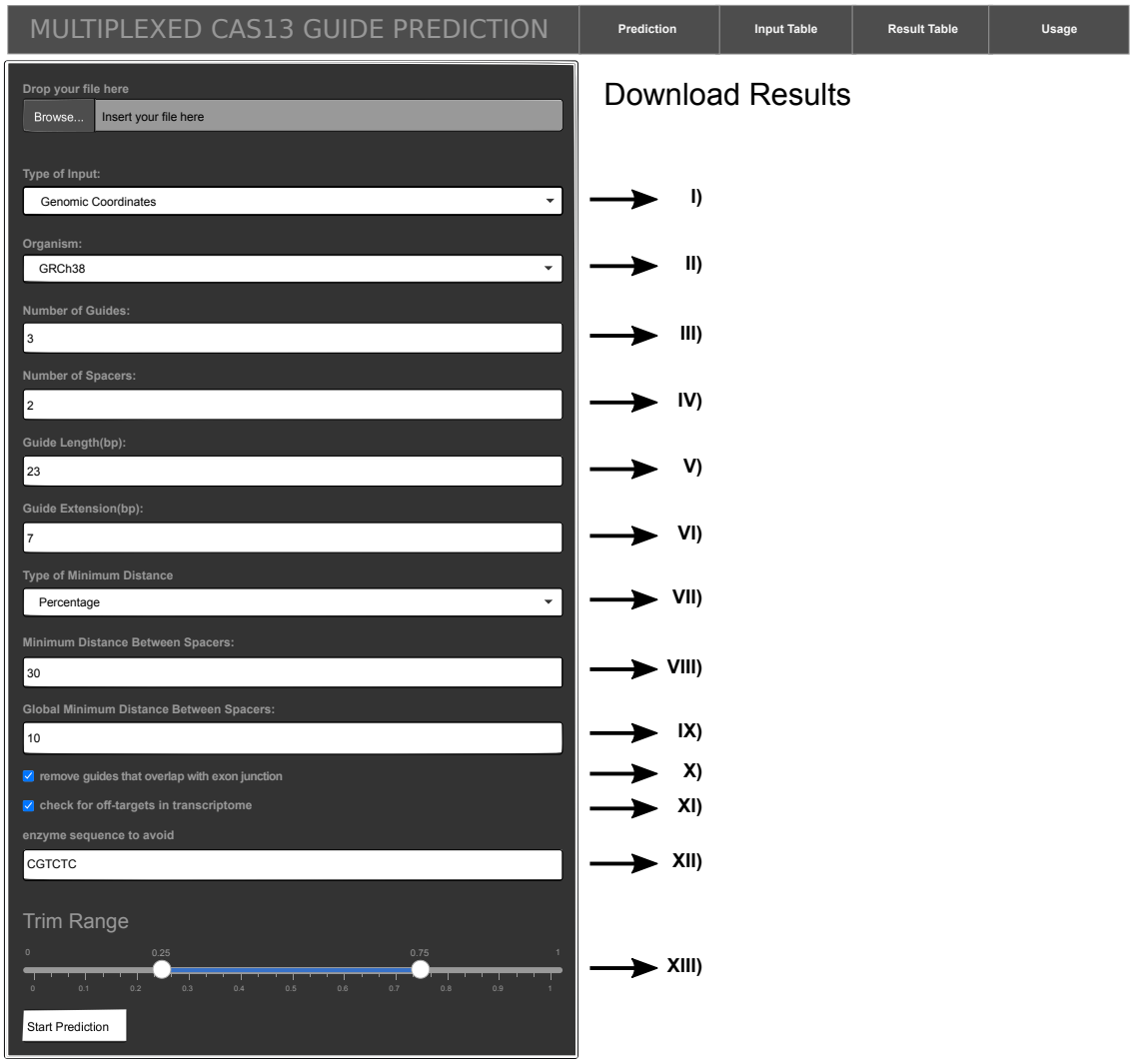

FIGURE 4.8: Guide Design: **(a)** image showing the main page of the R Shiny app with a number that connects it to the description of the function in the text: I) the user can load the file and select the format in which the file is provided II) the user can select which genome to use to design the library III) the user can select the number of guides that the program will design IV) the user will select the number of spacers for each of the guides of the array V) the user can select the guide length VI) the user can extend the guide by a specific amount of bases in case the system performs a maturation step VII) the user can select if the distance between the spacers is calculated in base pairs(bp) or in percentage VIII) the user can specify the minimum distance between the spacers within each guide IX) the user can specify the minimum global distance between all the spacers X) the user can specify if he wants the program to remove guides that are falling in exon-exon junctions XI) the user can specify if he wants the program to perform an off-target filtering step XII) the user can specify an enzyme cutting sequence or any sequence and the spacers containing that sequence will be removed XIII) the user can specify the range of the transcript on which he wants to design guides

4.5 Pan-cancer CasRx sgRNA library design

While coding genes are only slightly more than 20'000 and can be targeted all together in one single screen, our final collection contains 97817 lncRNA families. Consequently, we had to select which would be worth screening and therefore which ones to design the guides against. We followed mainly an expression criterion and subsequentially a conservation one. We decided to opt for these two criteria because only transcripts that are expressed can be functional and can be targeted and transcripts that have clear sequence constraints are possibly more likely to be functional [35]. With this criteria, therefore, we expect to enrich our list for functional transcripts.

To select the lncRNAs based on conservation we used the counts obtained previously with *feature counts*, normalizing them using TPM value. We chose TPM over other commonly used methods such as RPKM or FPKM because TPM allows a sample by sample comparison, while RPKM and FPKM do not.

$$TPM = A * \frac{1}{\sum A} * 10^6 (A = \frac{gene\ count * 10^3}{gene\ length(bp)}) \tag{4.1}$$

This is because in a given sample the sum of all TPMs is the same, therefore when we compare two samples we compare fractions of the same size, unlike in the other proposed methods [115, 116]. We decided to first select 30161 lncRNAs based solely on expression level but on three different selection steps. First, we selected the 10300 most expressed transcripts by average TPM value considering all samples (excluded non-solid tumours like stated before). Second, we selected 10320 transcripts by average TPM value by tissue. The selection procedure

consisted in averaging the expression values for each tumour type and going then type by type picking the most expressed non-coding transcripts. Last, since it is well known that a high amount of lncRNAs are cell-type-specific, we also selected other 9540 RNAs highly expressed by single cell type. We hence obtained 30161 lncRNAs. To this number we added another 3663 lncRNAs by first selecting only conserved lncRNAs that were remaining after taking the first 30161 and then applying the same criteria applied before, therefore we selected 2001 lncRNAs based on total average TPM value, 1014 based on tissue average TPM value and 648 based on cell-line-specific TPM value. After this procedure of selection, we ended up with a list of 33824 lncRNAs on which to design the guide RNAs against.

To design the guides for our selection we used 4 different parameter settings of our design program. First, we decided to design a dynamic number of guides, based on the transcript complexity, calculated by multiplying the number of exons of the transcript times the length of the transcript. We then used the logarithm of this number and divided it into six different equidistant regions on the density plot, assigning to each of them from two to seven guides (Fig.4.9*a*). In this way, the least complex transcripts (i.e. short transcripts with a low number of exons) are targeted by two sgRNAs, while very complex transcripts (i.e. long transcripts with a high amount of exons) are targeted with seven guides. This is done because complex lncRNA families are likely arising from highly complex loci with several different transcripts contributing to the same family, by increasing the number of guides targeting these loci we also increase the number of original transcripts that are knocked down in our screening, and we also have a higher change of getting high knock-down efficiency guides.

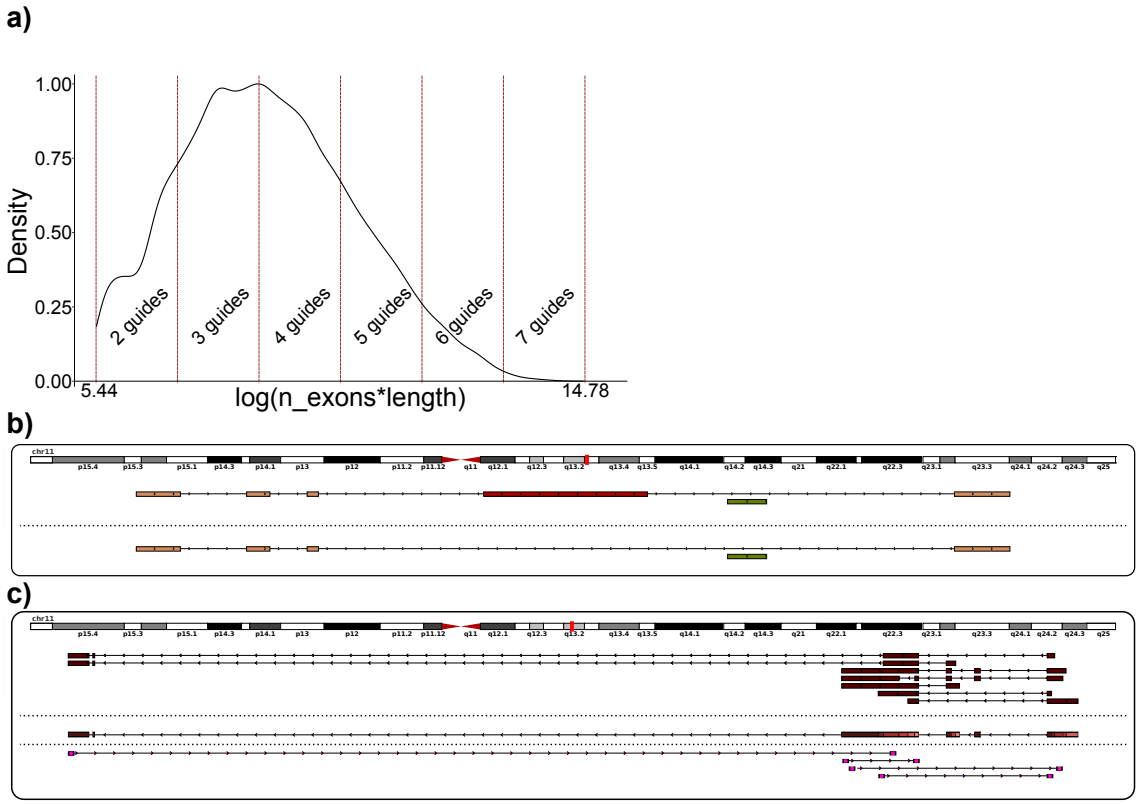

FIGURE 4.9: Guide Design: **(a)** density plot showing the number of exons times length in bp with relative cuts used to assign the number of guides: the assignment is based on a custom score of complexity taking into account transcript length and number of exons and assigns an increasing number of guides to an increasing complexity **(b)** genome viewer image showing the exon removal step based on deviation from the mean expression level. The exon marked in red is removed because its expression is lower than the mean expression minus the standard deviation **(c)** genome viewer image showing a transcript with all the guides that have been designed by our program. The guides are spanning the entire transcript length and most of its exons

Before proceeding with guide design, we filter exons of our transcripts based on expression variance. First, we run *feature counts* this time obtaining exon-level counts instead of transcript level counts. Given the nature of RNA expression, if different exons are part of the same transcript their level of expression should be roughly similar. Having generated our list of transcripts by merging different isoforms and given the non-curated nature of most of the databases from which

we take our transcripts, it is possible that one or more exons of one of our lncRNAs are arising from a low expressed splicing isoform of the transcript, or could also arise by an artefact produced by the *de novo* bioinformatic assembly of transcripts, consequentially, targeting these exons with one of our guides could be an inefficient strategy. In order to reduce this problem, for each RNA we take the raw counts for each exon, normalize them by exon length and then calculate the mean and the standard deviation of the normalized count value. We then remove from the transcripts all the exons that have a normalized count value lower than the mean normalized count value minus one standard deviation (Fig.4.9*b*). At this point we run our design program with the following fixed parameters: Number of Guides: based on complexity; Number of Spacers: 2; Guide Length: 23 bp; Guide Extension: 7 bp; Avoid Exon Junction Overlap; Enzyme Sequence to Avoid: GCTCTC; Trim Range: maintain the full transcript range. The variable parameters are the local and global minimum distances. Specifically, we run the program four times. The first with a *local minimum distance* of 30% and a *global minimum distance* of 10%. Then we maintained the guides for each transcript only if all of them had a score (according to Sanjana's program) of 0.75. Subsequently, we lowered the local and global minimum distances to 25% and 5% respectively. This time we maintained the guides targeting a transcript only if all of them were at least in the 4th quartile of quality according to Sanjana's score. At last, we lowered to 15% and 1% of local and global minimum distances and as the last step, we just used a distance of 30 base pairs, always maintaining the guides only if all the ones from each transcript were in the fourth quartile of quality. We decided to use these criteria because it permits us to retrieve guides that are far away enough from each

other so that they target the transcripts by covering most of their length. In the cases where this leads to a lower quality of the guides, we shortened the length coverage of the transcript maintaining a high quality of the guides, thus having always the best set of guides targeting the transcript for as much of its length as possible. An example of how the guides look like for a general transcript is shown in Fig.4.9*c*. The fused transcript shown has been assigned 4 guides (each of them has 2 spacers) based on the aforementioned criteria. Those guides are spanning the entire merged transcript covering three out of six exons, but the three not covered are small in size. From the original transcripts from which the fusion is derived all of the seven isoforms are targeted with the guides.

To perform a good quality CRISPR screening it is not sufficient to use only the targets we are interested in. In order to have a solid statistical background, a series of controls should be used [26]. First, we need two types of targeting controls. The first type consists of genes that are, according to actual scientific knowledge, always essential, which are genes whose knock-down results in a significant decrease in viability of the cells where these transcripts were knocked down. The second type of targeting control are genes that are known to be never essential. These genes when knocked down will not lead to a decrease in cell viability, since their role is not of particular importance for cell survival. These two controls give us a general idea of whether the screen was successful or not and whether the results of this screen (e.g. the genes that dropped out) are statistically reliable. To obtain these controls we first selected a list of 300 always essential and 300 never essential genes. These are taken based on information from the depmap portal where the always essential genes are genes that are always found to drop out while the never essential genes are genes that are

never found to drop out in cRISPR screenings. These two lists will be used as controls for the screen analysis with *MAGeCK* [117] in order to assess the functionality of the screen and the validity of our list. Once this list was made, we used our program to design Cas13 sgRNAs against the *MANE select* isoforms of those transcripts. *MANE select* is a collection of transcripts shared by RefSeq and Ensembl/GENCODE and contains for each coding gene a representative transcript, which is the main isoform derived from a specific gene. Where we could not find a *MANE select* isoform for our control genes, we derived manually the *ENSEMBL canonical* representative sequence. The third type of control consists of non-targeting sequences. These sequences are designed in such a way that they do not have any target among the transcripts of the cells. To generate these guides we followed a procedure in line with the one described by [118]. We generated ten thousand random 23mers and mapped them to the genome using bowtie and allowing up to three mismatches, we filtered only the sequences that were found not to map in any location of the genome and, among these, randomly selected 600 23mers. We then generated an additional 600 7mers and randomly merged them with the 23mers to obtain 30bp sequences, that is the sgRNA length required by our CRISPR/Cas13 system used with multi-spacer sgRNAs. After the aforementioned procedure, we could design guides according to our criteria for 24172 lncRNAs (The full selection procedure is described in Fig.4.10*a*). On this final list of lncRNAs, we performed a quality control analysis. Considering the criteria upon which we chose this selection we expected to enrich in more complex transcripts. This can be observed in Fig.4.10*b* and 4.10*c* where we clearly see an increase in transcripts of more than 2 exons and in longer transcripts compared to Fig.4.6*a* and *b*.

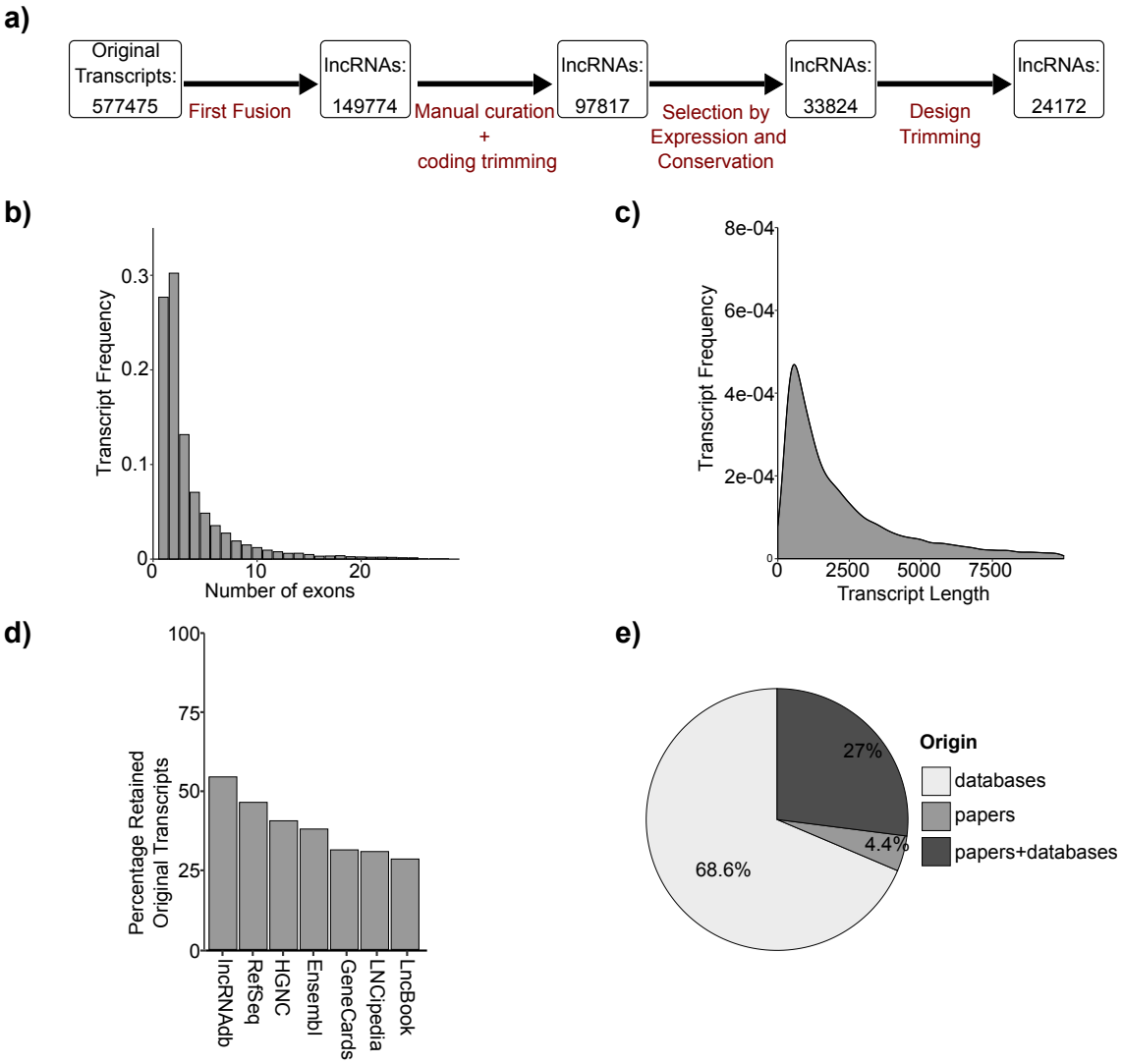

FIGURE 4.10: Analysis of the final lncRNAs selection for which we designed the guides: **(a)** flow chart showing all the steps performed to arrive to the final targeted collection: after the first fusion, we obtain 149774 transcripts. Following manual curation and trimming of transcripts overlapping coding genes, we have 97817 lncRNA families. Of these, following our criteria of selection based on expression and conservation we select 33824 transcripts of which we are able to design guides for 24172 **(b)** bar plot showing the frequency of transcripts by the number of exons. The number of exons is maintained in the same range as the full list **(c)** frequency plot showing the frequency of transcripts by transcript length. The transcript length density is also maintained in the same range as the previous list, with a slightly higher density for longer transcripts **(d)** bar plot showing the percentage of the original transcripts that we target with our collection of guides for each of the most important databases. We are able to maintain about 50% of the transcripts of each database **(e)** Pie Plot showing the origin of the fused transcripts in percentage of the total amount

We also checked which databases were represented more and also here, like in the whole collection, it is mainly correlated to the size of the database, resulting in bigger databases being more represented in our selection (Supplementary Fig 2).

Our lncRNAs families were obtained by merging transcripts from different databases, and since we target our fusion transcripts in a broad manner we expect to target several of the original lncRNAs. Indeed, the number of original transcripts that we target is close to 50% in all the main databases, with lncRNAdb having retained the highest amount and LncBook the lowest (Fig.4.10*d*). Considering evolutionary conservation, the percentage of conserved transcripts in the selection is higher than the one in the broad collection, as well as the percentage of transcripts coming only from papers (Fig.4.10*e*). This is expected since we enrich in conserved transcripts with our selection criteria.

4.6 Characteristic of the Targeted lncRNAs Collection

In order to assess the characteristics of our list and to perform in the future more detailed RNA-seq related analysis (e.g. differential expression analysis) integrating our data with other datasets, and since we aim to quantify expression on a transcript level we decided to use pseudo-alignment methods, in particular Salmon [119], to map RNA-seq reads from CCLE. The classical way of quantifying genes proceeds with the mapping of the reads on the entire genome using splice-aware aligners like *STAR* and then counting the reads falling uniquely into a feature. This approach is valid for gene-level counts, but to compare transcript level counts (which usually present a high amount of

overlap between the features) tools like *salmon* are recommended [120]. The principal characteristic of Salmon is that it performs an alignment using a *quasi-mapping* method directly on the reference transcriptome and using a statistical model of RNA-Seq that takes into account several well-known biases of RNA-Seq, giving in output directly the counts and the TPM values. Those counts are not integer numbers because Salmon can split in an intelligent manner the reads assigning partial reads based on the probability that a read is arising from one transcript or another. To select our lncRNA targeting list we used already counts from *feature counts*, because it was the best approach to rank the transcripts, but to perform analysis, Salmon is recommended. We, therefore, ran Salmon on the RNA-Seq data from CCLE to quantify again the transcripts. First, we wanted to check if the results correlate with the counts obtained from feature counts. With coding genes, the correlation is extremely high (Pearson R: 0.966) and has also a good value for our selection of lncRNAs (Pearson R: 0.696). The correlation line for the lncRNAs is a bit shifted to the right (towards feature counts axis), due to the settings that were used, that is counting two times reads that fall into multiple overlapping features (Supplementary Fig 3). We plotted again a PCA and a heatmap for sample correlation using the lncRNAs count data, resulting in a clustering really similar to the one obtained with feature counts (Fig.4.11*a* and Fig.4.11*b*). Thus, our lncRNAs selection is able to recapitulate the transcriptomic diversity across the different tumour types.

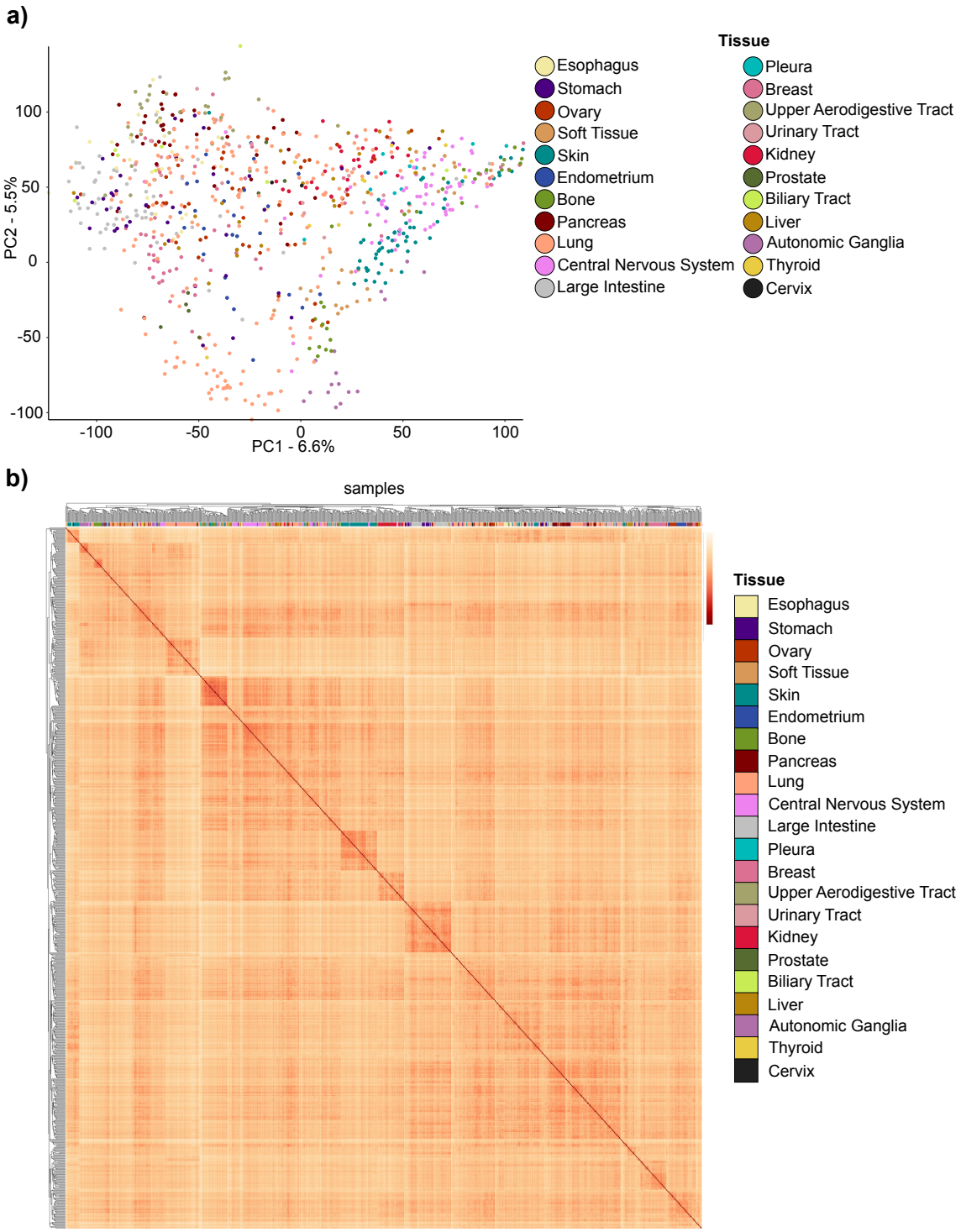

FIGURE 4.11: PCA plot and Heatmap clustering the samples based only on the selected transcripts: **(a)** PCA plot showing the different CCLE samples by tissue type **(b)** Heatmap showing sample-to-sample distance clusters

Once we had Salmon count tables for CCLE data we wanted to check the behaviour of our list regarding the cell lines that we selected to perform the screening. We selected the lncRNAs to be screened considering all cell lines from CCLE except non-solid tumour lines. This is because we aim to end up with a list that can be widely used and functional for different laboratories and purposes. We created a heatmap representing the level of expression of the selected lncRNAs in our cell lines (Fig.4.12*a*). It clearly shows that in each cell line a big part of our selected lncRNAs is expressed. The TPM values in the heatmap also correlate with the selection methods, with the RNAs selected by total average being the most expressed, followed by the tissue-specific and then the cell line-specific ones. We can also observe a slight enrichment of conserved transcripts in the lower part of the heatmap, where expression values are low. This is mostly because we selected the conserved lncRNAs after having already selected 30000 transcripts by their expression, resulting in the remaining transcripts that could be selected having a much lower level of expression. More in detail, we find that the number of expressed lncRNAs is comparable in all the tumour types, with lung expectably having the higher number of expressed transcripts for a bias due to it being over-represented in the CCLE collection (Fig.4.12*b*). In total, considering all of our cell lines together, only the 2.3% of our library is not expressed in any of the cell lines used (we considered to be not expressed transcripts with TPM < 0.001), with the 97.7% being screenable lncRNAs (Fig.4.12*c*). Given all this, we can say that screening our sgRNAs library on this set of cell lines will give us valuable insights into lncRNA functionality.

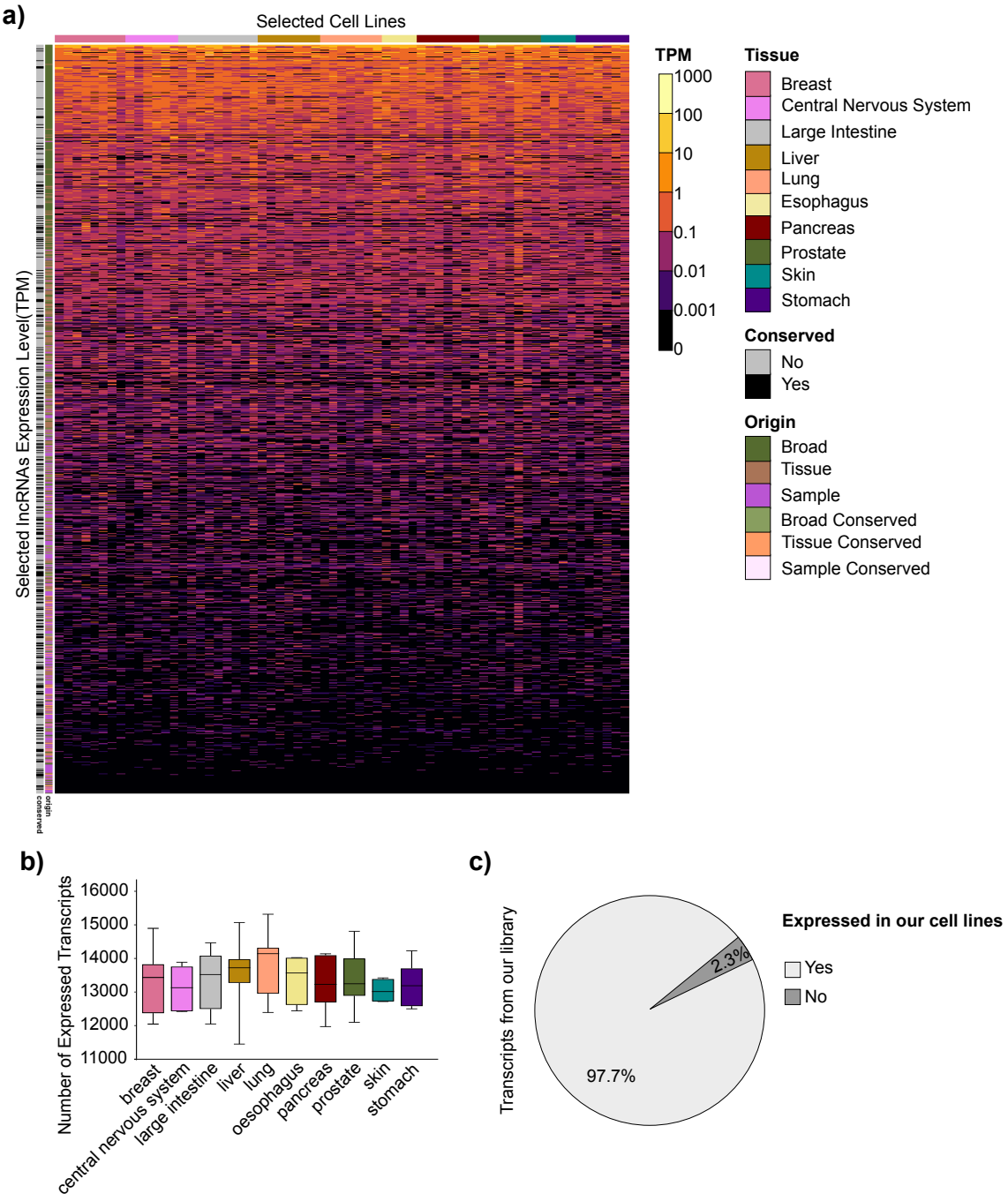

FIGURE 4.12: Final overview: Legend on following page

FIGURE 4.12: Final overview of our transcript selection: **(a)** Heatmap showing the expression of every selected lncRNA in the 60 established cell lines along with the conservation status and the origin of the selected transcript **(b)** Box Plot showing the number of expressed transcript of cell lines in each tissue type **(c)** Pie Plot showing the percentage of the list transcripts that are expressed at Salmon Mappingleast in one of our cell line collection with a TPM > 0.001

4.7 Evaluation of the CasRx lncRNA-panCancer library

To test our library of sgRNAs we performed a first whole-genome screening on the PDAC (Pancreatic Adenocarcinoma) cell line MIAPACA2. We decided to use this cancer cell line because of the promising knock-down efficiency given by the validation tests (Fig.4.3*d*). We performed a dropout screen. By infecting the cells with our sgRNA lentiviral library (each cell is infected with one guide by using a low multiplicity of infection), we are knocking down a specific lncRNA for each infected cell. If this lncRNA is essential for survival in that cell line, the infected cell will die more frequently. On the contrary, if that lncRNA doesn't affect survival the cell will grow as if no knock-down is performed. We can also have the case in which the lncRNA is bad for survival and its knock-down leads to increased proliferation. By letting the cells grow we will have a selective pressure that will cause the cells where the essential RNAs were knocked down to be under-represented. To measure this effect we performed this experiment on two replicates and after three weeks of passaging, the cells were sequenced. The sgRNA library is sequenced to be used as a control. A sgRNA is considered under-represented if its frequency after the experiment is lower than its frequency in the control and vice-versa for the over-represented sgRNAs.

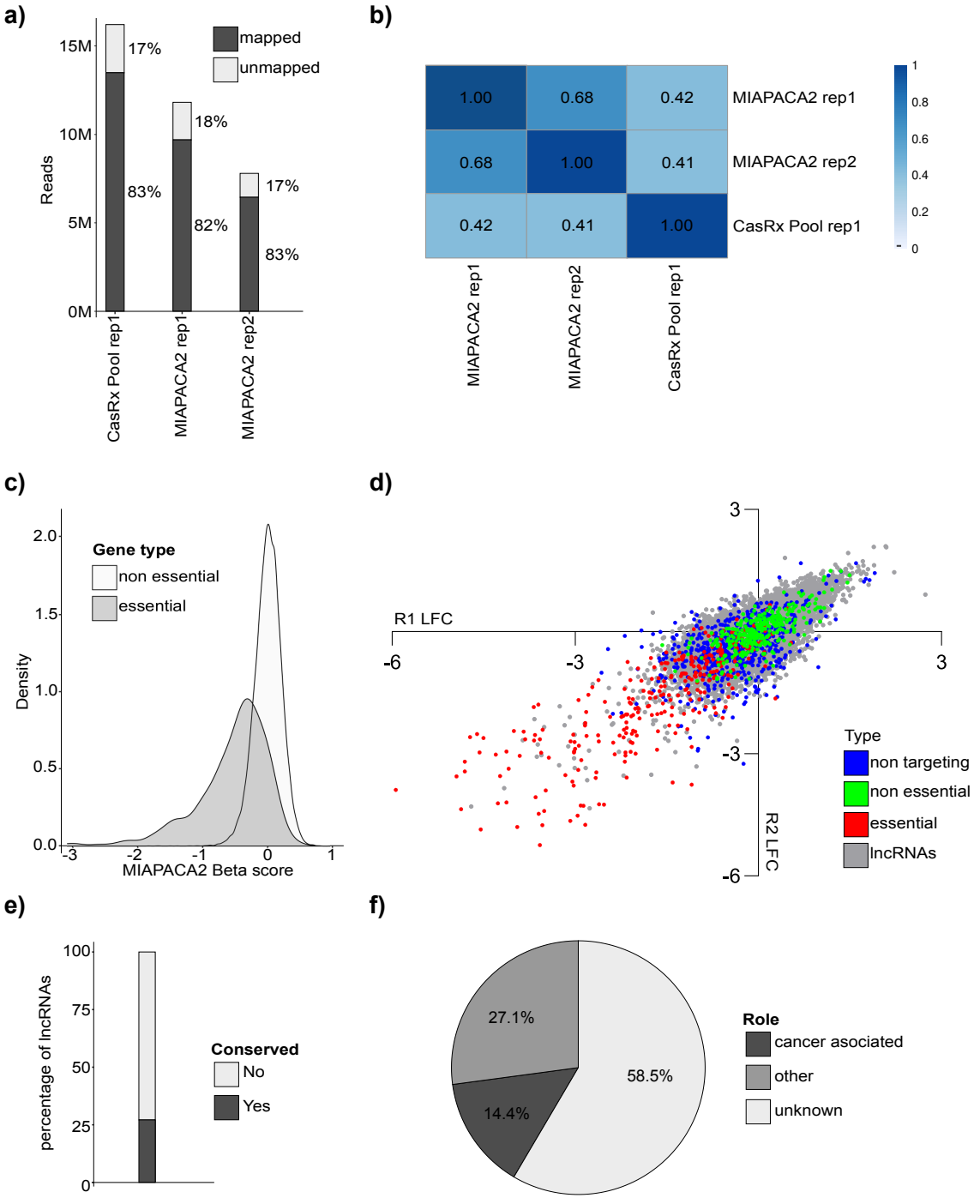

FIGURE 4.13: Analysis of MIAPACA2 CRISPR/CasRx pooled screen: figure legend on next page

FIGURE 4.13: Analysis of MIAPACA2 CRISPR/CasRx pooled screen: **(a)** bar plot showing the number and percentage of reads mapping correctly to the sgRNA reference, we get around 82% mapping reads in each sample **(b)** correlation plot showing sample-sample correlation of RNAseq counts derived from mapping to the sgRNA reference. MIAPACA2 samples are clustering together and are distant from the control **(c)** density plot showing the Beta-score of non-essential control genes and essential control genes. Non-essential genes have an average of 0, which means they are not dropping out, while essential gene density is pronouncedly shifted towards negative beta-scores **(d)** scatter plot showing the LFC calculated with the average of the 2 best gRNAs. The mon targeting genes and the non-essential genes are tightly distributed in the centre while essential genes are shifted to lower LFCs. lncRNAs, as expected, are distributed in a broader way. **(e)** bar plot showing the percentage of conserved transcripts among the lncRNA hits. **(f)** pie plot showing the function of the 118 dropout and enrichment hits. We found several cancer-associated genes and among these genes are well-known lncRNAs like MALAT1 and NEAT1. The majority of the lncRNAs have no known function

We performed the analysis of the sequencing with an in-house pipeline that is a wrapper around the *MAGeCK* tool [117]. The pipeline was made for CRISPR/Cas9 screenings with single guides, but our guides contain two spacers, therefore we adapted the pipeline in order to read and use only the first spacer of the two to do the analysis. Even though the second spacers are not used, this doesn't affect the analysis because the guide pairs are unique and both are dropping out at the same rate.

After performing the counting we had a quality control step. We found a good ratio of mapping reads that is higher than 80% in both the MIAPACA2 replicates and in the control (Fig.4.13*a*). In addition to that, all the samples correlate very well, both the MIAPACA2 samples are clustering together and are more different than control (Fig.4.13*b*). As a quality control proving that the screening works, we checked if the controls behave as expected, that is the non-essential control genes must not have a significant dropout while the essential control genes should result in a lower beta-score, which means that

their depletions results more often in a deadly phenotype, which is exactly what we can see in Fig.4.13*c*. Of note is that since the essential genes are genes that are always dropping out in Cas9 screens, we expect at least a part of them not to drop out with CasRx, since this is a knock-down experiment while Cas9 performs a complete knock-out. Indeed, we can see from Fig.4.13*c* that the difference is not as strong as it would be in a Cas9 dropout screening.

From the *MAGeCK* analysis results, we found only 3 statistically significant dropout lncRNAs hits and 6 enrichments (the cell survives more after the knock-down than expected so that particular gene should have an anti-proliferative role, as, for example, a tumour suppressor). This little amount of significant genes could be due to the fact that each of our targets is actually representing different RNA isoforms therefore some gRNAs could target non expressed isoforms, or could mean that the knock-down efficiency for the sgRNAs is more heterogeneous in the CasRx system than in the Cas9 system. On top of that, *MAGeCK* uses the median of the gRNAs scores in order to calculate the significance of the hit, this would worsen the just highlighted problems since the median is less influenced by the better guides in the pool. This is an expected problem that also other screens found [25, 74]. To overcome this issue, Liu et al. [25], in a screening performed with the CRISPRi system, selected the best 3 gRNAs per gene and used the mean [25]. Considering all this, we opted for selecting the best two gRNAs LFC (Log Fold Change) for each transcript in each replicate and averaged them. We then set an empirical threshold for an LFC to be considered a hit taking into account the LFC values from the never essential genes. This solution is not optimal but can give us an overview of how the screening performs in terms of which kind of hits we find.

In the close future, a more robust statistical method based on this averaged LFC measure has to be properly developed.

$$LFC = log2(\frac{treatment\ guide\ count}{control\ guide\ count}) \tag{4.2}$$

From the plot shown in Fig.4.13*d* we can have a panoramic on the result of this approach. As expected, the non-targeting controls and the non-essential control genes are not dropping out, while the essential control genes have a lower LFC and 131 of them are dropouts. From our targeted lncRNAs we found 118 hits of which 82 are dropouts and 36 are enrichments. Of these hits 27.1% corresponds to conserved lncRNAs (Fig.4.13*e*). We manually annotated this list of dropouts and enrichments and we found that 14.4% of them corresponds to non-coding RNAs that have been associated with cancer, 27.1% are known lncRNAs that are not associated with malignant neoplasms and 58.5% have unknown function. Among the hits with known association with cancer, we find also MALAT1 and NEAT1 that are two of the most well-characterised lncRNAs and present several associations with cancer [121, 122]. In particular, MALAT1 is an evolutionary conserved single exon 8kb lncRNA that is recruited to nuclear paraspeckles [10]; it was suggested as a marker for poor prognosis in patients affected by early-stage non-small lung cancer [123] and was later associated with different cancers [121]. NEAT1 is an essential lncRNA for the formation of paraspeckles [124], its associations with cancer are several and is one of the most studied lncRNAs [122]. In our screening NEAT1 is the best dropout hit. The lncRNA SNHG16, which is the second most dropout hit on our screen after NEAT1, was found to be oncogenic in several cancer types [125]. But most specifically and of

importance for our results, the expression of this lncRNA was found upregulated in pancreatic cancer tissues and this upregulation was linked to poor survival. In addition to this, its knock-down inhibited tumour growth in vitro and in vivo [126], this is in line with our screening results, where knock-down of this lncRNA led to an under-proliferating phenotype. Besides these lncRNAs, we find several others. For example, the dropout LINC00526 was found downregulated in human glioma and correlated with a poor prognosis, and its knockdown was linked to the promotion of glioma cell proliferation [127] while the enriched transcript lnc-IGFBP4-3 was found upregulated in lung cancer and its overexpression promoted Lung Carcinoma cell proliferation [128]. Another dropout in our list, PVT1, was found to be significantly upregulated in pancreatic cancer [129].

Chapter 5

Discussion

Genome-wide CRISPR screenings of coding genes have been widely performed since the advent of this technology. Although they present a certain degree of consensus within the scientific community regarding the key guidelines to follow when performing them and there are several available widely tested sgRNA Cas9 libraries [130], they have been mainly carried out on protein-coding genes, while lncRNA genome-wide screenings are a novelty in the field [56]. Thus, there are only a few libraries available regarding lncRNAs in cancer with several limitations, namely, they have been designed only for a small subset of lncRNAs (25) expressed in one cell line (chronic myeloid leukaemia K562) upon anti-cancer drug treatment [80] and to target specifically circular RNAs (crRNAs), which require a completely different design approach [81]. With our sgRNA library, we seek to overcome this problem by collecting entries from all the most up-to-date lncRNA databases and by broadening the spectrum of our targets to include information from 841 cell lines derived from 22 different cancer types (oesophagus, stomach, ovary, soft tissue, skin, endometrium, bone, pancreas, lung, central nervous system, large intestine,

pleura, breast, upper aerodigestive tract, urinary tract, kidney, prostate, biliary tract, liver, autonomic ganglia, thyroid, cervix) derived from the Therefore, we believe that our library will be of great interest not only for us but for many research laboratories willing to perform genome-wide screening of lncRNAs without having to design their own custom libraries, hence saving time.

The only Cas13 functional screenings performed up to now, were carried out only on 25 specific lncRNAs with the purpose of studying anti-cancer drug response [80] and on circular RNAs [81, 82], while the biggest lncRNA screening related to cancer in terms of the number of targets and amount of screened cell lines knocked-down 16401 lncRNAs reading out a cell proliferation phenotype on 7 different cancer cell lines [25]. This makes our library the biggest lncRNA targeting library available up to date, with a selection of 24172 targets that, being derived from the fusion of several more transcripts has a much higher virtual number of potential targets which we calculated to be 182309.

It was also shown that most of the lncRNAs mainly exercise their function in a tissue-specific fashion [97], raising the question of whether the use of only a handful of cell types would hinder important discoveries. This problem is addressed in our study by establishing a massive amount of CasRx cell lines derived from 10 different tumour types, maximizing our possibility to find several lncRNAs that might be functional in a cancer type-specific manner, but not broadly, and that might have been missed in the lncRNAs screenings already published. All this, along with the fact that the amount of lncRNAs that have been found functional in cancer is low [27], highlights the potential of our screening platform to provide new significant insights into lncRNA functionality.

The lag between protein-coding and non-coding transcriptome annotations is a major drawback for the study of lncRNA functinality [32]. Our fusion approach, by putting together both manual curated and hybrid databases, along with evolutionary conserved non-coding RNAs and an additional collection of conserved non-coding transcripts expressed in different developmental stages [34] is trying to minimize the bottlenecks that the use of a single database would carry. This, jointly with the enrichment in highly expressed transcripts in cancer cell lines and conserved transcripts will translate in a library with a great potential to contain functional transcripts. To underline the robustness of our sgRNA collection, a recent review [131] suggests, in order to design a lncRNA targeting library, an approach that is extremely similar to the one that we used to create ours and also consists in filtering for expressed transcripts via RNAseq data analysis, and possibly add a filter for evolutionary conservation. Another feature that emphasizes the effectiveness of our library is the amount of already characterized lncRNAs that are found functional in cancer and that are present also in our collection. In particular, we found that of 76 lncRNAs that were found functional in cancer [27], 54 (which correspond to 76% of them) are targeted by our pooled sgRNA library. Of the 18 that are not screened by our library, eight are antisense of a protein-coding transcript (e.g. RASSF1-AS1); two are found on the intron of a coding gene in the same strand; one was later found to be a protein-coding gene and four are present in our collection but are not screenable due to repetitive sequences or overlap with other RNAs. All of these aforementioned transcripts cannot be present in our collection, since we removed this type of RNAs due to the impossibility to assess their expression level and therefore rank them properly to select them, or due to the impossibility

of guide design, leaving out only three cancer-related functional lncRNAs (namely DBE-T, DINO and lncPRESS1) that are filtered out because of low expression levels. Nonetheless, it has to be underlined that, though not optimal, the loss of antisense transcripts in our collection is greatly reduced due to our design compared to other studies. In most of the lncRNAs screening performed by far, antisense transcripts are heavily filtered out [56], while we attempt to retain all the antisense transcripts that are screenable with our CRISPR/Cas13 system. To show this, of the 20 cancer-related antisense lncRNAs [27], we filter out only 8, keeping the majority in our sgRNA library.

The experimental setting of our screening, using Cas13 instead of more common RNA-interference techniques will also lead to the better discovery of functional lncRNAs since the limitations of RNA-interference are a higher amount of off-target effect [66], and the difficulty of its expression inside the nucleus, which is where the majority of lncRNAs are thought to perform their main function [69], problems that are overcome by Cas13 because of its ability to target lncRNAs in the nucleus and its limited off-target effects [54]. ASOs, short, synthetic single-strand oligonucleotides that trigger the downregulation of a complementary RNA target are also encoded in the nucleus [70], but the transient nature of their expression makes them not feasible to be used in experiments that need to be carried out for a long time like a CRISPR screen [56], while CasRx doesn't present this problem since it is integrated stably in the nucleus [84]. The use of other CRISPR tools acting on DNA can lead to false hits due to interaction with other elements in the genome [78]. In addition to this, in a cancer setting important genes may undergo copy number alterations and it was found that there is a non-neglectable correlation between copy number

variations and a gene-independent increase in dropout hits using the CRISPR/Cas9 system [132]. By using a transcript targeting system like Cas13 we avoid both problems, possibly reducing the number of false hits.

One major drawback of our approach is certainly the use of only unstranded DNA protocols for the generation of RNA-Seq data. Unstranded RNA-Seq is much less expensive than stranded sequencing, thus bringing researchers that are carrying on massive sequencing projects like CCLE or TCGA to prefer them. This prevents us from investigating thoroughly antisense lncRNAs because we can only screen those antisense RNAs that are falling only in introns of coding transcripts, leaving out the majority of the members of this family of RNAs, which, given its nature, is of great interest and can lead to important discoveries. In addition to this, the use of stranded RNA-Seq protocols can lead to a more clear quantification of the transcripts like lncRNAs that are covering the same genomic regions but on opposite strands, avoiding assigning to these transcripts the same magnitude of counts.

With the MIAPACA2 preliminary screen, we confirmed experimentally that our sgRNAs library works and is well designed and robust. Nonetheless, our fusion approach giving multiple actual transcripts for every guide and the weak consensus among which are the best parameters to design a sgRNA for the CRISPR/Cas13 system require less strict statistics to assign significance to dropout and enrichment hits. This might slightly increase false discovery, but since even with our arbitrary threshold, we are still in the order of a hundred hits and all the hits are going to be validated, this is not a major concern. Since with our approach of calculating the LFC on the average of the selected best two gRNAs, we get 118 hits and with *MAGeCK* we get 9 hits with a false discovery

rate (FDR) of <0.05 and 78 hits with an FDR of <0.25, we would expect that a specific statistic taking into consideration the mean LFC would give a number of hits comparable to other lncRNA screenings already performed with different systems. These screenings find a percentage of hits ranging from 0.29% to 7.6% [25, 74, 99], with most of the cell lines not exceeding 1% of hits and our method finds 118 hits in a number of targeted transcripts in MIAPACA2 that ranges from 12000 to 14000 (Fig.4.12*b*) which result in an average 0.91% hits, in line with the other screening. Developing a better statistical approach based on the average LFC of the best gRNAs and performing the screening on other cell lines will give a better estimation of these numbers since only one cell line doesn't necessarily reflect the behaviour of the system. The detection of several lncRNAs associated with cancer is a good index of the quality of our screening platform. We detect also a significant amount of transcripts whose function is not related to cancer. This doesn't necessarily mean that those transcripts have also a specific unknown role in malignancies. In fact, selecting transcripts based also on conservation means that some of our hits will be functional both in normal and tumour tissues, thus, to discern between those and the ones selectively expressed in cancer cells more bioinformatics analysis (e.g. differential expression) should be carried on. The substantially big amount of transcripts of unknown function (58.5%) strongly underlines the current lack of knowledge about this class of RNAs which our study aims to help fill up.

Given our fusion approach, several guides are targeting different isoforms in the same lncRNA family with possible different functionality. It will be useful to modify the screening analysis pipeline in order to be able to find out which isoforms are more functional and this could also recover some important hits

that are not resulting significant because one of the guides is targeting a non-functional isoform.

5.1 Future Directions

5.1.1 Dropout screen and analysis of 60 cell lines from 10 tumor types

We already performed a test screening on the MIAPACA2 pancreas PDAC cell line. This made sure that our screening system works properly and already gave us insights over which lncRNAs are potentially functional in pancreatic cancer. The next and most prominent step of this research project will be performing the screening on all the sixty cell lines that we established from our selected ten tumour types. This will result in several dropouts, each of them likely being functional lncRNAs in one or more cancers. Based on previous knowledge [56], we expect to find several non-coding transcripts dropping out only in a tumour-type-specific manner, in concordance with the fact that most lncRNAs are also only expressed in a tissue-type-specific or cell-type-specific fashion [97]. Since we first specifically selected also for broadly expressed transcripts (i.e. transcripts whose average expression in all the CCLE samples was the highest), we also expect to find a relevant number of lncRNA dropping out in more than one unique cancer type or cell line, and since it was shown that a lower level of conservation is linked to a more tissue-specific expression [35], we suppose to find also a link between conservation and a less type-specific dropout, which would translate in functionality in a broader amount of cell lines. According to

the current bibliography, once carried out, our screening will represent the biggest and most complete screening performed with the CRISPR/Cas13 system, as well as the one targeting the biggest amount of lncRNAs in the highest number of cell lines. Therefore, this has the potential to find several novel lncRNAs that are functional in cancer cells, broadening the scarce knowledge we have about this new and highly diverse class of transcripts. Furthermore, we think that these findings will be potentially useful also for translational research and clinical development since several lncRNAs could be further studied as targets of cancer therapies [133].

5.1.2 Differential Expression analysis on other tumor types and integration with other Omics data

Big databases like CCLE, TCGA and GTex open up the possibility to analyse unprecedented amounts of data from a variety of tissues and cancer types.

The Cancer Genome Atlas (TGCA) offers a striking collection of more than 20000 molecularly characterized primary cancers, collected and analyzed by several researchers and laboratories around the globe since 2006, offering an unprecedented amount of information on human cancer biology [134]. The Genotype-Tissue Expression database (GTEx) offers a collection of human healthy tissue samples and provides researchers with the possibility to have hundreds of samples for each tissue for RNA-seq data analysis [135].

This has the potential to help to further understand cancer biology and have new insights on the correlation between a broad assortment of genomic data and clinical information (e.g. patient survival) [136]. Therefore, we plan to perform

differential expression analysis comparing cancer samples from TGCA and CCLE with healthy tissue samples from GTeX and to a lesser extent TGCA for each of the tumour types that are used in the screenings and also cross-compare different tissue to further investigate lncRNA specificity. This will give us a complete overview of which of the transcripts from our collection could have clinical relevance in different types of cancer, finding out also which ones are highly tumour-type-specific and which ones have, instead, a broader role in tumour biology.

Besides this, we sought to integrate the RNA-Seq data with other omics data, also provided by TCGA and GTEx, namely Chip-Seq and WGS. This type of information can allow us to further investigate the function of these lncRNAs. We plan to use Chip-Seq data from TCGA and GTex to have a better overview of the change of the Chromatin state in differentially expressed loci. Cancer is caused by a different set of mutations that are positively selected. Mutations found in cancer include point mutations in specific genes, copy-number changes and structural variants [137]. The Pan-cancer Analysis of Whole Genomes [137] provides a set of 2'600 cancer whole genomes from different tumour types that along with TGCA data can be used to infer patterns of mutations related to lncRNAs expression abnormalities.

Another type of analysis that will be interesting to perform in the near future is the inference of Coregulatory Networks among our collection of RNAs. Coregulatory Networks can give insights on which lncRNAs are involved in known regulatory pathways or infer new ones opening a window on the correlation between expression and more complex phenotypes that are arising from the communication and co-regulation of multiple coding and non-coding

elements [138]. While assessing co-regulation between lncRNAs poses numerous challenges, mainly due to the lack of a clean and complete annotation of lncRNAs and their role, the association between lncRNAs and coding genes is feasible and could be particularly useful when investigating pathways involved in cancer development or progression.

5.1.3 In Vivo Validation

The limitations carried by the use of 2D cultured cells are well known and documented [139] and in pre-clinical research, only a few treatments that are found effective in a 2D cell culture setting are also effective in *in vivo* experiments [140]. The use of animal models, nonetheless, is much more expensive and time-consuming and therefore we need *in vitro* research to narrow down the candidates that can be then studied with *in vivo* experiments in animal models. Up to now, only a handful of *in vivo* lncRNA knock-out experiments have been performed in mouse models [141], and most of them didn't give appreciable cancer-related phenotype. For example, the knock-out mice of the relatively well-characterized MALAT1 didn't give any phenotype at all [142]. Therefore, a lot more needs to be done to gain information on the phenotype associated with these transcripts. In order to perform further functional studies of lncRNA dropout hits, we generated an animal model to perform *in vivo* CasRx knock-down. The use of this model to validate our targets will be of great importance since validating the targets in *in vivo* models will give increased relevance to our discoveries. Since the transcripts we will screen *in vivo* will be selected among the conserved ones, we will have a direct impact

on the advancement of understanding of the role of these lncRNAs in humans and it will be easier to translate our findings to clinical advancements.

Chapter 6

Conclusion

We developed a CRISPR/Cas13 perturbation platform for functional genomics studies among with a robust sgRNAs library targeting a broad set of transcripts by merging lncRNAs from all the up-to-date available databases and performing filtering steps to enrich for expressed and conserved transcripts in a very broad range of cancer cell lines. Regarding the number of targeted lncRNAs, our library is the biggest and more complete available. We generated 53 validated CRISPR/CasRx cell lines from ten different tumor types. We proved the validity and functionality of this system by performing a whole-genome preliminar screen. Carrying out the screens for all the cell lines will result in the biggest Cas13 lncRNA screen ever performed, which will possibly translate in a big amount of new information regarding the function of non coding RNAs in cancer that can be associated with clinical relevance in order to give possible therapeutic targets. Besides this, this library will be of great utility as a tool for other reasearchers willing to perform functional genomic screening in a lncRNA context.

Appendix A

Supplementary Figures

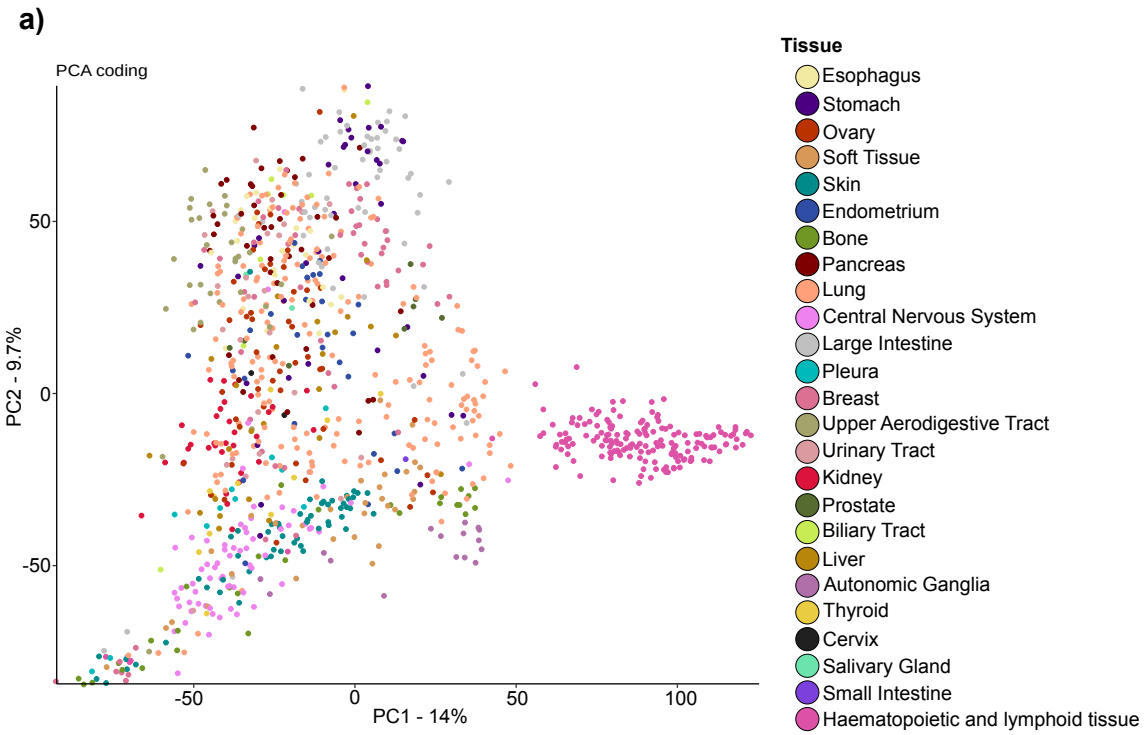

FIGURE A.1: PCA plot of coding genes: **(a)** PCA showing the clustering of coding genes without any filtering. It can be seen how the non-solid tumors are clustering together far away from the solid ones

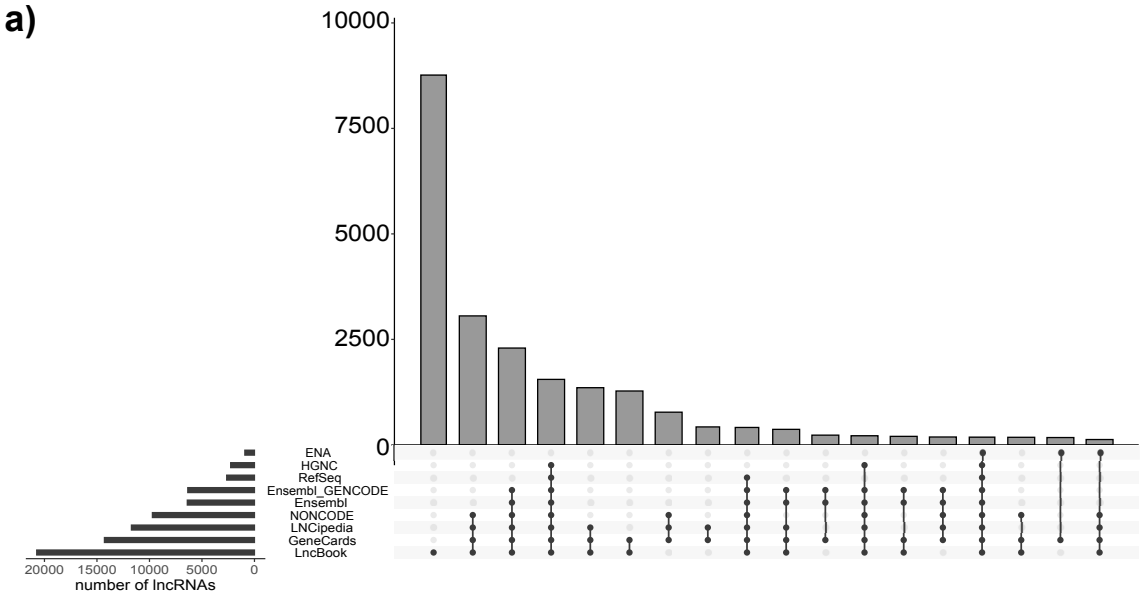

FIGURE A.2: UpSet plot of the selected list: **(a)** UpSet plot after all the filtering and selection steps of the lncRNA collection that will be targeted with our system

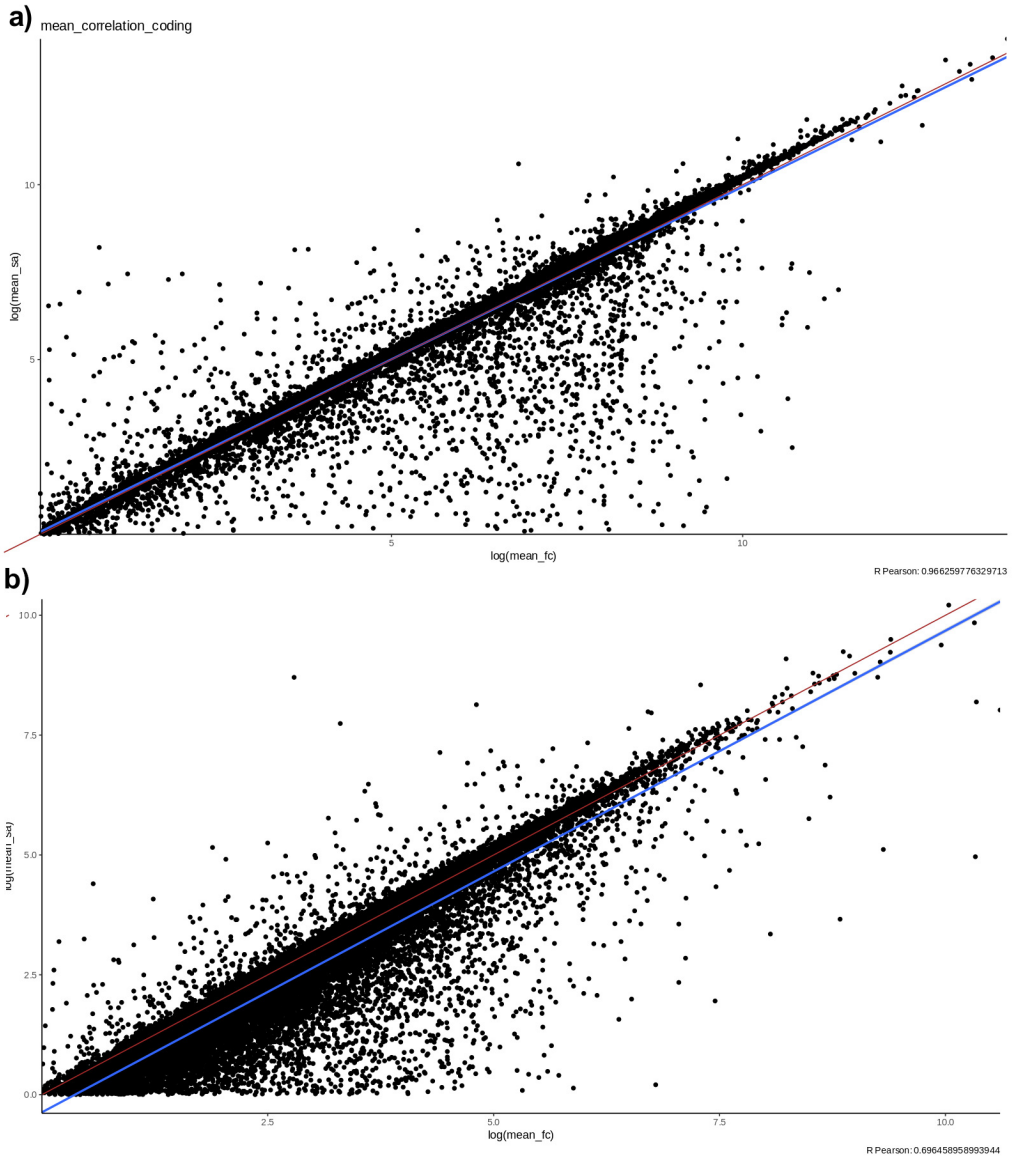

FIGURE A.3: Correlation between Featurecounts and Salmon: **(a)** Scatter plot showing the correlation between the counts obtained with Featurecounts(x axis) and Salmon(y axis) for the coding genes **(b)** Scatter plot showing the correlation between the counts obtained with Featurecounts(x axis) and Salmon(y axis) for the lncRNAs

Bibliography

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).
2. Laurent, G. S., Wahlestedt, C. & Kapranov, P. The Landscape of long noncoding RNA classification. *Trends in Genetics* **31,** 239–251 (May 2015).
3. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biology* **16** (Jan. 2015).
4. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–1774 (Sept. 2012).
5. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47,** D766–D773 (Oct. 2018).
6. Cai, X. & Cullen, B. R. The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA* **13,** 313–316 (Jan. 2007).
7. Askarian-Amiri, M. E. *et al.* SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA* **17,** 878–891 (Apr. 2011).
8. Zhang, X. *et al.* Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *International Journal of Molecular Sciences* **20,** 5573 (Nov. 2019).
9. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science* **304,** 1321–1325 (May 2004).

10. Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8** (Feb. 2007).

11. Li, B. *et al.* Activation of LTBP3 Gene by a Long Noncoding RNA (lncRNA) MALAT1 Transcript in Mesenchymal Stem Cells from Multiple Myeloma. *Journal of Biological Chemistry* **289,** 29365–29375 (Oct. 2014).

12. Beckedorff, F. C. *et al.* The Intronic Long Noncoding RNA ANRASSF1 Recruits PRC2 to the RASSF1A Promoter, Reducing the Expression of RASSF1A and Increasing Cell Proliferation. *PLoS Genetics* **9** (ed Lee, J. T.) e1003705 (Aug. 2013).

13. Morriss, G. R. & Cooper, T. A. Protein sequestration as a normal function of long noncoding RNAs and a pathogenic mechanism of RNAs containing nucleotide repeat expansions. *Human Genetics* **136,** 1247–1263 (May 2017).

14. Hung, T. *et al.* Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature Genetics* **43,** 621–629 (June 2011).

15. Grelet, S. *et al.* A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. *Nature Cell Biology* **19,** 1105–1115 (Aug. 2017).

16. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137 (Jan. 1996).

17. Lee, J. T. & Jaenisch, R. Long-range cis effects of ectopic X-inactivation centres on a mouse autosome. *Nature* **386,** 275–279 (Mar. 1997).

18. Carrieri, C. *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491,** 454–457 (Oct. 2012).

19. Bond, C. S. & Fox, A. H. Paraspeckles: nuclear bodies built on long noncoding RNA. *Journal of Cell Biology* **186,** 637–644 (Aug. 2009).

20. Ishii, N. *et al.* Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *Journal of Human Genetics* **51,** 1087–1099 (Oct. 2006).

21. Zhou, X. & Xu, J. Identification of Alzheimer's disease–associated long noncoding RNAs. *Neurobiology of Aging* **36,** 2925–2931 (Nov. 2015).

22. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics* **47,** 199–208 (Jan. 2015).

23. Beermann, J. *et al.* A large shRNA library approach identifies lncRNA Ntep as an essential regulator of cell proliferation. *Cell Death & Differentiation* **25,** 307–318 (Nov. 2017).

24. Delás, M. J. *et al.* lncRNA requirements for mouse acute myeloid leukemia and normal differentiation. *eLife* **6** (Sept. 2017).

25. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355** (Jan. 2017).

26. Li, L. & Chang, H. Y. Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends in Cell Biology* **24,** 594–602 (Oct. 2014).

27. Huarte, M. The emerging role of lncRNAs in cancer. *Nature Medicine* **21,** 1253–1261 (Nov. 2015).

28. Tsai, M.-C. *et al.* Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science* **329,** 689–693 (Aug. 2010).

29. Zhang, J., Zhang, P., Wang, L., l. Piao, H. & Ma, L. Long non-coding RNA HOTAIR in carcinogenesis and metastasis. *Acta Biochimica et Biophysica Sinica* **46,** 1–5 (Oct. 2013).

30. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465,** 1033–1038 (June 2010).

31. Poliseno, L. *et al.* Deletion of PTENP1 Pseudogene in Human Melanoma. *Journal of Investigative Dermatology* **131,** 2497–2500 (Dec. 2011).

32. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* **19,** 535–548 (May 2018).

33. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (Jan. 2014).

34. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571,** 510–514 (June 2019).

35. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research* **24,** 616–628 (Jan. 2014).

36. Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature Communications* **6** (Jan. 2015).

37. Chen, J. *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biology* **17** (Feb. 2016).

38. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11,** 1110–1122 (May 2015).

39. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10,** 1177–1184 (Nov. 2013).

40. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Research* **46,** D308–D314 (Nov. 2017).

41. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42,** D756–D763 (Nov. 2013).

42. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics* **11,** 181–190 (Feb. 2010).

43. Koonin, E. V. & Makarova, K. S. CRISPR-Cas. *RNA Biology* **10,** 679–686 (Feb. 2013).

44. Barrangou, R. & Marraffini, L. A. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Molecular Cell* **54,** 234–244 (Apr. 2014).

45. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR/Cas systems. *Nature Reviews Microbiology* **13,** 722–736 (Sept. 2015).

46. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (Aug. 2012).

47. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences* **109,** E2579–E2586 (Sept. 2012).

48. Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Molecular Cell* **60,** 385–397 (Nov. 2015).

49. Abudayyeh, O. O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353** (Aug. 2016).

50. East-Seletsky, A. *et al.* Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538,** 270–273 (Sept. 2016).

51. Smargon, A. A. *et al.* Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Molecular Cell* **65,** 618–630.e7 (Feb. 2017).

52. Yan, W. X. *et al.* Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Molecular Cell* **70,** 327–339.e5 (Apr. 2018).

53. Konermann, S. *et al.* Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* **173,** 665–676.e14 (Apr. 2018).

54. Zhang, C. *et al.* Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell* **175,** 212–223.e17 (Sept. 2018).

55. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics* **19,** 67–80 (Dec. 2017).

56. Lucere, K. M., O'Malley, M. M. R. & Diermeier, S. D. Functional Screening Techniques to Identify Long Non-Coding RNAs as Therapeutic Targets in Cancer. *Cancers* **12,** 3695 (Dec. 2020).

57. Fire, A., Albertson, D., Harrison, S. & Moerman, D. Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle. *Development* **113,** 503–514 (Oct. 1991).

58. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391,** 806–811 (Feb. 1998).

59. Yu, J.-Y., DeRuiter, S. L. & Turner, D. L. RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proceedings of the National Academy of Sciences* **99,** 6047–6052 (Apr. 2002).

60. Miyagishi, M. & Taira, K. Development and application of siRNA expression vector. *Nucleic Acids Symposium Series* **2,** 113–114 (Nov. 2002).

61. Nötzold, L. *et al.* The long non-coding RNA LINC00152 is essential for cell cycle progression through mitosis in HeLa cells. *Scientific Reports* **7** (May 2017).

62. Seiler, J. *et al.* The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance. *Nucleic Acids Research* **45,** 5458–5469 (Feb. 2017).

63. Klingenberg, M. *et al.* The Long Noncoding RNA Cancer Susceptibility 9 and RNA Binding Protein Heterogeneous Nuclear Ribonucleoprotein L Form a Complex and Coregulate Genes Linked to AKT Signaling. *Hepatology* **68,** 1817–1832 (Oct. 2018).

64. Tiessen, I. *et al.* A high-throughput screen identifies the long non-coding RNA DRAIC as a regulator of autophagy. *Oncogene* **38,** 5127–5141 (Mar. 2019).

65. Stojic, L. *et al.* A high-content RNAi screen reveals multiple roles for long noncoding RNAs in cell division. *Nature Communications* **11** (Apr. 2020).

66. Smith, I. *et al.* Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLOS Biology* **15** (ed Freeman, T.) e2003213 (Nov. 2017).

67. Jackson, A. L. & Linsley, P. S. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery* **9,** 57–67 (Jan. 2010).

68. Persengiev, S. P., Zhu, X. & Green, M. R. Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA* **10,** 12–18 (Dec. 2003).

69. Lennox, K. A. & Behlke, M. A. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. *Nucleic Acids Research* **44,** 863–877 (Nov. 2015).

70. Monia, B. *et al.* Evaluation of 2'-modified oligonucleotides containing 2'-deoxy gaps as antisense inhibitors of gene expression. *Journal of Biological Chemistry* **268,** 14514–14522 (July 1993).

71. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Research* **30,** 1060–1072 (July 2020).

72. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (Jan. 2013).

73. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343,** 80–84 (Jan. 2014).

74. Zhu, S. *et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nature Biotechnology* **34,** 1279–1286 (Oct. 2016).

75. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152,** 1173–1183 (Feb. 2013).

76. Joung, J. *et al.* Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature* **548,** 343–346 (Aug. 2017).

77. Liu, S. J. *et al.* CRISPRi-based radiation modifier screen identifies long non-coding RNA therapeutic targets in glioma. *Genome Biology* **21** (Mar. 2020).

78. Goyal, A. *et al.* Challenges of CRISPR/Cas9 applications for long non-coding RNA genes. *Nucleic Acids Research,* gkw883 (Sept. 2016).

79. Abudayyeh, O. O. *et al.* RNA targeting with CRISPR-Cas13. *Nature* **550,** 280–284 (Oct. 2017).

80. Xu, D. *et al.* A CRISPR/Cas13-based approach demonstrates biological relevance of vlinc class of long non-coding RNAs in anticancer drug response. *Scientific Reports* **10** (Feb. 2020).

81. Li, S. *et al.* Screening for functional circular RNAs using the CRISPR–Cas13 system. *Nature Methods* **18,** 51–59 (Dec. 2020).

82. Zhang, Y. *et al.* Optimized RNA-targeting CRISPR/Cas13d technology outperforms shRNA in identifying functional circRNAs. *Genome Biology* **22** (Jan. 2021).

83. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12,** 996–1006 (May 2002).

84. Wessels, H.-H. *et al.* Massively parallel Cas13 screens reveal principles for guide RNA design. *Nature Biotechnology* **38,** 722–727 (Mar. 2020).

85. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (June 2009).

86. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).

87. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (Aug. 2005).

88. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* **49,** D884–D891 (Nov. 2020).

89. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research* **50,** D20–D26 (Dec. 2021).

90. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6,** 343–345 (Apr. 2009).

91. Jun-ichi, M. *et al.* Expression vector system based on the chicken beta-actin promoter directs efficient production of interleukin-5. *Gene* **79,** 269–277 (July 1989).

92. Qin, J. Y. *et al.* Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. *PLoS ONE* **5** (ed Hansen, I. A.) e10611 (May 2010).

93. Xia, X., Zhang, Y., Zieth, C. R. & Zhang, S.-C. Transgenes Delivered by Lentiviral Vector are Suppressed in Human Embryonic Stem Cells in A Promoter-Dependent Manner. *Stem Cells and Development* **16,** 167–176 (Feb. 2007).

94. Ibrahimi, A. *et al.* Highly Efficient Multicistronic Lentiviral Vectors with Peptide 2A Sequences. *Human Gene Therapy* **20,** 845–860 (Aug. 2009).

95. Campbell, A. M. Chromosomal insertion sites for phages and plasmids. *Journal of Bacteriology* **174,** 7495–7499 (Dec. 1992).

96. Friedrich, G & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes & Development* **5,** 1513–1523 (Sept. 1991).

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

98. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (Mar. 2012).

99. Liu, Y. *et al.* Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nature Biotechnology* **36,** 1203–1210 (Nov. 2018).

100. And Blake A Sweeney *et al.* RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research* **49,** D212–D220 (Oct. 2020).

101. Ma, L. *et al.* LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Research* **47,** 2699–2699 (Jan. 2019).

102. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54** (June 2016).

103. Volders, P.-J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Research* **47,** D135–D139 (Oct. 2018).

104. Zhao, L. *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research* **49,** D165–D171 (Nov. 2020).

105. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44,** D733–D745 (Nov. 2015).

106. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (Jan. 2010).

107. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (Sept. 2012).

108. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20,** 1983–1992 (Dec. 2014).

109. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (Apr. 2014).

110. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (Oct. 2012).

111. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (Nov. 2013).

112. De Weck, A., Bitter, H. & Kauffmann, A. Fibroblasts cell lines misclassified as cancer cell lines (July 2017).

113. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (Mar. 2012).

114. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (Mar. 2011).

115. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131,** 281–285 (Aug. 2012).

116. Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O. & Coombes, K. A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* **20** (Dec. 2019).

117. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology* **15** (Dec. 2014).

118. Meier, J. A., Zhang, F. & Sanjana, N. E. GUIDES: sgRNA design for loss-of-function screens. *Nature Methods* **14,** 831–832 (Sept. 2017).

119. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14,** 417–419 (Mar. 2017).

120. Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience* **8** (Dec. 2019).

121. Sun, Y. & Ma, L. New Insights into Long Non-Coding RNA MALAT1 in Cancer and Metastasis. *Cancers* **11,** 216 (Feb. 2019).

122. Pisani, G. & Baron, B. NEAT1 and Paraspeckles in Cancer Development and Chemoresistance. *Non-Coding RNA* **6,** 43 (Oct. 2020).

123. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin 4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22,** 8031–8041 (Sept. 2003).

124. Clemson, C. M. *et al.* An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Molecular Cell* **33,** 717–726 (Mar. 2009).

125. Yang, M. & Wei, W. SNHG16: A Novel Long-Non Coding RNA in Human Cancers. *OncoTargets and Therapy* **Volume 12,** 11679–11690 (Dec. 2019).

126. Liu, S., Zhang, W., Liu, K. & Liu, Y. LncRNA SNHG16 promotes tumor growth of pancreatic cancer by targeting miR-218-5p. *Biomedicine & Pharmacotherapy* **114,** 108862 (June 2019).

127. Yan, J. *et al.* Long non-coding RNA LINC00526 represses glioma progression via forming a double negative feedback loop with AXL. *Journal of Cellular and Molecular Medicine* **23,** 5518–5531 (June 2019).

128. Yang, B. *et al.* Overexpression of lncRNA IGFBP4-1 reprograms energy metabolism to promote lung cancer progression. *Molecular Cancer* **16** (Sept. 2017).

129. Wu, B.-Q., Jiang, Y., Zhu, F., Sun, D.-L. & He, X.-Z. Long Noncoding RNA PVT1 Promotes EMT and Cell Proliferation and Migration Through Downregulating p21 in Pancreatic Cancer Cells. *Technology in Cancer Research & Treatment* **16,** 819–827 (Mar. 2017).

130. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34,** 184–191 (Jan. 2016).

131. Pulido-Quetglas, C. & Johnson, R. Designing libraries for pooled CRISPR functional screens of long noncoding RNAs. *Mammalian Genome* (Sept. 2021).

132. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discovery* **6,** 914–929 (June 2016).

133. Arun, G., Diermeier, S. D. & Spector, D. L. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends in Molecular Medicine* **24,** 257–277 (Mar. 2018).

134. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45,** 1113–1120 (Sept. 2013).

135. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (Sept. 2020).

136. Unfried, J. P. *et al.* Identification of Coding and Long Noncoding RNAs Differentially Expressed in Tumors and Preferentially Expressed in Healthy Tissues. *Cancer Research* **79,** 5167–5180 (Aug. 2019).

137. Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (Feb. 2020).

138. Hawe, J. S., Theis, F. J. & Heinig, M. Inferring Interaction Networks From Multi-Omics Data. *Frontiers in Genetics* **10** (June 2019).

139. Costa, E. C. *et al.* 3D tumor spheroids: an overview on the tools and techniques used for their analysis. *Biotechnology Advances* **34,** 1427–1441 (Dec. 2016).

140. Sabroe, I. *et al.* Identifying and hurdling obstacles to translational research. *Nature Reviews Immunology* **7,** 77–82 (Jan. 2007).

141. Gao, F., Cai, Y., Kapranov, P. & Xu, D. Reverse-genetics studies of lncRNAs - what we have learnt and paths forward. *Genome Biology* **21** (Apr. 2020).

142. Nakagawa, S. *et al.* Malat1 is not an essential component of nuclear speckles in mice. *RNA* **18,** 1487–1499 (June 2012).