UNIVERSITEIT GENT

FACULTY OF MEDICINE AND HEALTH SCIENCES

Exploring lncRNAs in cancer: tools for discovery and characterization of cancer associated lncRNAs

This thesis is submitted as fulfillment of the requirements for the degree of Doctor in Biomedical Sciences by Pieter-Jan Volders, 2015

Promotor
prof. dr. ir. Jo Vandesompele

Co-promotor
prof. dr. Kris Gevaert

Supervisor
dr. ir. Pieter Mestdagh

Center for Medical Genetics Ghent (CMGG)
Cancer Research Institute Ghent (CRIG)
Bioinformatics Institute Ghent N2N (BIG N2N)

Ghent University Hospital, Medical Research Building
De Pintelaan 185, 9000 Ghent, Belgium
+32 9 332 1951
PieterJan.Volders@ugent.be

Promoter:

prof. dr. ir. Jo Vandesompele
Department of Pediatrics and medical genetics, Ghent University, Belgium

Co-promoter:

prof. dr. Kris Gevaert
Department of Biochemistry, Ghent University, Belgium

Supervisor:

dr. ir. Pieter Mestdagh
Department of Pediatrics and medical genetics, Ghent University, Belgium

Members of the examination committee:

prof. dr. Elfride De Baere (chairman)
Department of Pediatrics and medical genetics, Ghent University, Belgium

prof. dr. Petra Van Damme (secretary)
Department of Biochemistry, Ghent University, Belgium

dr. ir. Gerben Menschaert
Department of Mathematical Modelling, Statistics and Bio-informatics, Ghent University, Belgium

prof. dr. ir. Katleen De Preter
Department of Pediatrics and medical genetics, Ghent University, Belgium

prof. dr. ir. Stein Aerts
Department of Human Genetics, KU Leuven, Belgium

dr. Guillaume Smits
ULB Genetics Center, Université libre de Bruxelles, Belgium
HUDERF - Queen Fabiola Brussels Children Hospital, Belgium

prof. dr. Maité Huarte
Center for Applied Medical Research, University of Navarra, Spain

The research described in this thesis was conducted at the Center for Medical Genetics, Ghent University, Ghent, Belgium.

This work was supported by the BIG N2N, Bioinformatics Institute Ghent From Nucleotides to Networks

I. INTRODUCTION

I.1. LONG NON-CODING RNA

The central dogma of molecular biology 'DNA makes RNA makes protein' is no more. Protein coding genes are no longer the largest class of genetically encoded entities in the human genome, as a myriad of long non-coding RNAs (lncRNAs) have been sequenced and await functional characterization.

The publication of the first draft of the human reference genome in 2001[1,2] marked the beginning of the post genomic era. With the reference genome at hand, researchers began to explore the transcriptome, the part of genome that is transcribed into RNA. Until recently, messenger RNAs (mRNA) were thought to be the most prevalent and important entities of the human transcriptome. After transcription and splicing (Figure 1), these transcripts are exported to the cytoplasm where they serve as templates for protein synthesis. About 22,000 genes in the human genome produce mRNA transcripts, collectively referred to as protein coding genes as their sequences encode the amino acid sequences for a protein. Non-coding genes on the other hand are not translated into protein; the RNA transcript itself or a derivate RNA molecule forms the functional product of the corresponding gene. Several thousands of non-coding genes have been discovered and functionally described. The majority of the functionally annotated non-coding RNA transcripts are further processed into short RNA molecules with diverse functions in genetics. MicroRNA (miRNA) and small nucleolar RNA (snoRNA) for instance, are two well described classes of non-coding RNA[3].

One of the most surprising discoveries brought about by the spectacular advancements in (RNA) sequencing technology is the extensive transcription arising from regions previously regarded as genomic wasteland. Both large-scale collaborative efforts, such as the ENCODE (Encyclopedia of DNA Elements)[4] project and the FANTOM (Functional ANnoTation Of the Mammalian genome)[5] consortium, and smaller projects by individual research groups[6-8] have expanded the known human transcriptome several times in numbers of transcripts. The great majority of

these novel transcripts are long (> 200 nt), multi-exonic and without conserved open reading frames (ORFs)[6,9]. As such they give rise to a new genetic class called long non-coding RNAs (lncRNAs). Recently, RNA sequencing of over 7,000 human tumor samples revealed more than 90,000 distinct lncRNA genes, making this the largest genetic class in the human genome[10].

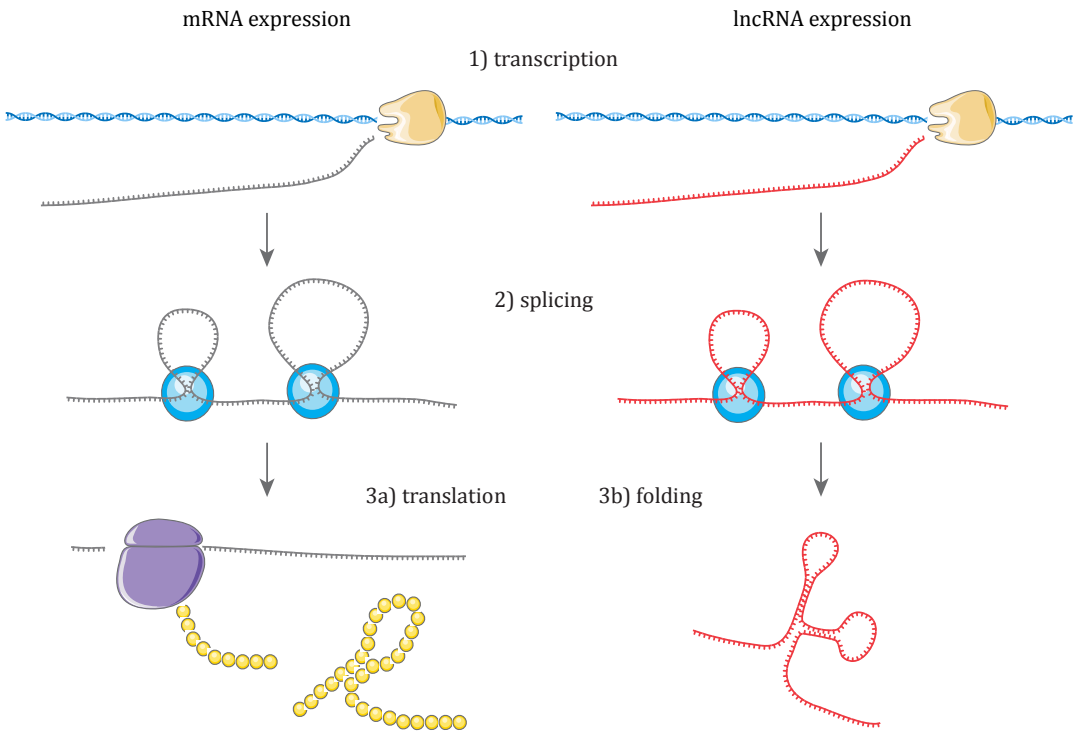


Figure 1: Protein coding mRNA versus lncRNA expression. Both mRNA and lncRNA are transcribed by RNA polymerase II and can undergo splicing. mRNA subsequently binds with ribosomes and is translated to protein. While a protein is here the functional entity and not the mRNA, lncRNA is functional in itself often through complex secondary structures.

LncRNA genes resemble their protein coding counterparts in genetic structure. Although lncRNA transcripts are on average slightly smaller, they can also be multi-exonic and subject to alternative splicing[6]. Even from an epigenetic perspective they are very similar. Their promoter and gene body regions exhibit the same chromatin modifications associated with RNA polymerase II transcribed protein coding genes[11]. Compared to mRNA, the expression levels of lncRNA are typically lower yet more

tissue and cell-type specific[9]. Interestingly, the most extensive lncRNA expression is found in the testes[6]. While protein coding genes are evolutionary well conserved, lncRNA genes appear to be the result of more recent evolutionary adaptations as their sequence conservation scores are often much lower[9]. Most importantly, lncRNAs lack ORFs that exhibit the evolutionary pattern typically observed for protein coding ORFs or any other evidence of protein coding potential[6,9]. Moreover, on the subcellular level, lncRNAs are found to be more nuclear enriched compared to protein coding mRNAs[4]. An observation that is highly suggestive for a function that does not require ribosomal translation.

I.2. LNCRNA FUNCTIONS

In sharp contrast to the extraordinary rate at which new lncRNAs are being reported, the rate at which they are functionally characterized is low. Currently, less than 200 human lncRNA genes have been functionally studied (181 according to lncrnadb.org[12] and 191 according to genenames.org[13]). A surprisingly low number considering more than 4,400 lncRNA papers have been published to date. Indeed, the majority of publications is focused on just a handful of lncRNAs[13]. Nevertheless, the biological processes in which lncRNAs are known to be involved are numerous and diverse.

DOSAGE COMPENSATION: EARLY EVIDENCE FOR FUNCTIONAL LNCRNAS

The cellular process where lncRNAs made their debut is X-chromosome inactivation (XCI). Without a doubt this is currently the best and most extensively described process that shows how lncRNAs can play crucial roles in a cell. XCI is the mechanism that balances the gene expression on the X-chromosome between sexes in mammals. The role of a non-coding RNA in XCI has been recognized since 1992 when a 17 kb lncRNA was found to be exclusively expressed from the inactive X chromosome[14]. Referred to as the X-inactive-specific transcript (XIST), this lncRNA appeared to coat the entire inactive X-chromosome (Xi)[15]. Currently it is recognized that XCI requires the interplay between several lncRNAs expressed from the same genomic locus (X-inactivation center). The TSIX lncRNA is expressed on the active X-chromosome (Xa) and prevents inactivation by epigenetic silencing of XIST[16]. In the

absence of TSIX expression XIST recruits chromatin modifying complexes. Together with XIST, these complexes spread across the entire chromosome, eventually silencing the Xi[17].

EMBRYONIC DEVELOPMENT AND CELL DIFFERENTIATION

The discovery of pervasive lncRNA transcription in the human HOX loci already suggests an important role for lncRNAs in development and differentiation[18]. This role is further confirmed by the finding that the majority of lncRNAs expressed in mouse embryonic stem (ES) cells have implications in ES cell transcriptional regulation and are associated both with cell differentiation and pluripotency[19]. Indeed, several lncRNAs have been discovered that aid in maintaining the pluripotent state and prevent differentiation into specific lineages. The lncRNA LINC-ROR for instance, was found to be crucial for induced pluripotent stem cell and ES cell survival. LINC-ROR is under transcriptional control of the key pluripotency transcription factors SOX2, OCT4 and NANOG and probably promotes survival by inhibiting the p53 response pathway[20]. In addition to pluripotency, various lncRNAs have been implicated in embryonic development. Morpholino-mediated inhibition of the conserved lncRNAs cyrano and megamind in zebrafish embryos results in severe but distinct phenotypes. While the cyrano morphant showed many developmental defects resulting in an overall body malformation, megamind knockdown resulted in specific brain and eye defects, pointing to a role in brain development[21]. Also in other species lncRNAs were found with roles in embryonic development. In mouse embryogenesis, two lncRNAs have been implicated in cardiac development. The poetically named lncRNA Braveheart (BVHT) has been identified and studied in mouse ES cells. BVHT was found to be a key regulator in a cardiac gene network and its expression is required for cardiac cell fate[22]. *In vivo* evidence for the crucial role of lncRNAs in mammalian development was found with the lncRNA FENDRR (Fetal-lethal noncoding developmental regulatory RNA). Not only were homozygous FENDRR mutants embryonic lethal, they showed an impaired heart and body wall development[23].

IMPLICATIONS IN DISEASE

The currently reported functions of lncRNAs go well beyond dosage compensation and development. They have been associated with processes as diverse as immune response[24], paraspecle formation[25] and growth arrest[26]. Given their central role in many cellular pathways it should be no surprise that dysregulation of lncRNAs is often associated with disease. LncRNAs have already been implicated in an variety of diseases[27] including COPD[28], AIDS[29], Alzheimer's disease[30], cardiovascular diseases[31], autoimmune diseases[32] and cancer. In addition, lncRNAs have been associated with rare disorders such as the Beckwith-Wiedemann syndrome[33], Angelman syndrome[34], Klinefelter's syndrome[35] and blepharophimosis syndrome[36].

I.3. LNCRNAS IN CANCER

Numerous reports on lncRNA involvement in cancer have been published to date, far more than any other disease or process[27]. These reports include both oncogenic and tumor suppressive lncRNAs, implicated in many different cancer types (Table 1). Like cancer itself, their individual mode of action is diverse. In fact, distinct lncRNAs have been associated with all of the hallmarks of tumor biology proposed by Hanahan and Weinberg[37,38].

HOTAIR is one of the most frequently described oncogenic lncRNAs to date. Elevated expression is found in breast cancer[39], colorectal cancer[40] and pancreatic cancer[41], and is associated with poor prognosis and metastasis. Transcribed from the HOXC locus, HOTAIR regulates gene expression in trans (unlike other lncRNAs in that locus) on many genomic loci including HOXD (the mechanism is further described in section I.4)[18]. Among the genes regulated by HOTAIR are several genes with important roles in different aspects of cancer biology[39,41].

Table 1: A summary of lncRNAs with a published role in cancer. With new cancer associated lncRNAs reported on a weekly basis, this list is far from complete.

Role	LncRNA	Cancer type
Oncogenic	CCAL	Colorectal cancer[42]
	FAL1	Ovarian cancer[43]
	HOTAIR	Breast cancer[39], colorectal cancer[40], pancreatic cancer[41]
	MALAT1	Lung cancer[44,45], hepatocellular cancer[46], bladder cancer[47]
	PCGEM1	Prostate cancer[48]
	PVT1	Gastric cancer[49], ovarian & breast cancer[50]
Tumor suppressive	GAS5	Breast cancer[51]
	MEG3	Pituitary adenoma[52,53], meningioma[54], hepatocellular cancer[55], thyroid cancer[56]
	PTENP1	Prostate cancer[57]
	TUSC7	Colorectal cancer[58]

Mechanisms of lncRNA deregulation in cancer

Almost every cell in the body of a multicellular organism has the potential to abandon its task and develop into a tumor. Cancer cells are characterized by their ability to evade the intrinsic mechanisms that prevent uncontrollable growth and dedifferentiation[59]. These characteristics result from genetic changes that are either acquired during the lifetime of an individual (somatic) or inherited (germline). Somatic mutations arise from exogenous mutagenic exposures and mitotic DNA copy errors and accumulate over time[60]. While the majority of the mutations are so-called passenger mutations that will not confer growth advantage, some mutations will by chance affect one of the many cancer genes in the genome. These mutations are driver mutations and they promote progression of a cell into a cancer cell.[61]. Mutations can be small, affecting only a single basepair or larger such as translocations and copy-number variations. Somatic copy-number aberrations (SCNAs) are extremely common in cancer and due to their size the earliest studied

type of genetic change[62]. By studying the copy-number profile of cancer cells, researchers have discovered many important oncogenes and tumor suppressor genes. In doing so, they contributed to development of improved therapeutics and treatments[63]. While some cancer genes are rarely affected or affected only in specific cancer types, others are broadly affected in many patients and entities. MYC and CDKN2A/B for instance, exhibit copy-number chances in as many as 30% of all human tumors[64].

Even though SCNAs affecting protein coding genes in cancer have been extensively studied, lncRNAs have been largely overlooked in this regard. This can be explained by the use of outdated genomic annotation in the design and analysis of the used platforms. To overcome this problem, some researchers have repurposed existing DNA microarray platforms and re-annotated the probe content with current lncRNA annotation[43,65]. These efforts led to the identification of new prostate cancer associated lncRNAs[65] and the discovery of the oncogenic lncRNA FAL1 (focally amplified lncRNA on chromosome 1)[43].

Single nucleotide polymorphisms (SNPs) constitute a second class of well-studied alterations in the cancer genome. Especially due to the many genome-wide association studies (GWAS), numerous SNPs have now been linked to specific diseases including cancer. Interestingly, about half of the cancer GWAS hits fall outside of known protein coding loci. While some likely affect cis-regulatory regions of nearby genes, many of those may be transcribed as lncRNAs[66]. Early evidence of SNPs affecting the function of lncRNAs in cancer has been reported for the lncRNA ANRIL[67]. In fact, ANRIL was identified as a major genomic hotspot in GWAS. The 126 kb gene spans several SNPs that are associated with a variety of diseases including gliomas and basal cell carcinomas. More recently, re-annotation of SNP and GWAS databases provided many more SNPs that potentially affect lncRNAs[68]. To detect disease-causing SNPs in protein coding genes, the discrimination between synonymous SNPs (those that do not alter the protein sequence) and non-synonymous SNPs is typically made. It is important to note that this classification is not usable for lncRNA SNPs. While some authors have tried to predict the effect of

SNPs on the secondary RNA structure[69], the field currently lacks established tools that can be used for these predictions.

Given that the mature RNA is the functional form of lncRNA genes, measurements of lncRNA expression therefore closely represent the levels of the active molecule. Quite a few lncRNAs exhibit deregulated expression in cancer that is often reported as predictive for disease severity and progression[41,45,47,70]. As such, whole-transcriptome analysis of cancer tissue led to the identification of several differentially expressed lncRNAs, for instance PCAT-1 in prostate cancer[71]. Recently, analysis of the combined transcriptomes of thousands of cancer and normal samples revealed 8,000 lineage or cancer-associated lncRNA genes[10].

LNCRNAS TO THE CLINIC

Although lncRNA therapeutics is still in the early stages of development, their tissue[4] and cancer[10] specific expression makes them ideal candidates both as biomarkers and targets for therapy. In addition, lncRNAs are often found to be involved in epigenetic regulation of many target genes[72] and, as such, can exert broad effects on gene expression and cell functioning.

Obviously, targeting oncogenic lncRNAs would be an interesting therapeutic approach. Down-regulation of lncRNA expression *in vivo* can be achieved by antisense technology such as antisense oligonucleotides (ASOs)[44]. While the first antisense drugs showed poor performance in clinical trials[73], second generation drugs are well on their way to the clinic with many drugs efficiently targeting both protein coding and non-coding RNAs[74]. Degradation of an oncogenic lncRNA is however not required to disturb its undesired activity. As many lncRNAs require interaction with protein partners such as Polycomb repressive complex 2 (PRC2) to exhibit their function, a disruption of this interaction is sufficient to block lncRNA functionality. Such a steric hindrance could be achieved by both small molecules and ASOs, the latter being currently under development by the Massachusetts-based company RaNA Therapeutics[75].

In particular lncRNAs that act in cis on tumor suppressor genes make interesting potential drug targets. As antisense lncRNAs typically repress transcription from the

opposite strand, targeting them could induce expression on the sense strand. Researchers have demonstrated this concept *in vivo* for the BDNF locus, as targeting the BDNF-AS transcript with ASOs resulted in a sevenfold increased expression of BDNF[76]. Given the frequent occurrence of antisense transcription[77], this could be an approach to reactivate specific tumor suppressor genes in cancer.

LncRNAs have been proposed as potent biomarkers in cancer as the expression levels of several lncRNAs have been shown to be indicative for disease severity or progression[41,45,47,70]. Furthermore, several lncRNAs are present in bodily fluids at detectable levels. For instance, the hepatocellular carcinoma associated lncRNA HULC (highly upregulated in liver cancer) shows great potential as biomarker since its expression is highly correlated with tumor grade. HULC RNA is detectable in plasma of patients which makes testing fast and safe[70]. Even more convenient testing is possible through exosomal lncRNA in urine. Linc-p21 for instance, has a higher exosomal concentration in the urine of patients with prostate cancer[78]. Although the majority of lncRNA based biomarkers are still under development, one lncRNA already found its way to the clinic as a biomarker in prostate cancer. The Progensa PCA3 assay, based on the concentration of the prostate cancer associated 3 (PCA3) lncRNA in urine, was recently approved by the FDA as biomarker for prostate cancer[79]; urinary PCA3 levels are predictive for positive biopsies and outperform other biomarkers[80].

I.4. MECHANISMS OF LNCRNA FUNCTION

LNCRNAS AS GUIDES OR SCAFFOLDS FOR CHROMATIN MODIFICATION

Chromatin remodeling is an important mechanism for the regulation of gene expression at specific loci. Euchromatin is an open structure whereby the DNA is accessible enabling active transcription, whereas heterochromatin is more condensed with little or no transcription possible. The transition between the two states depends heavily on modifications of specific amino acids in the N-terminal tails of histones, often referred to as the histone code. The effect of the modification depends on the histone (H2A, H2B, H3 or H4), type of modification (commonly acetylation or methylation) and the specific amino acid that is modified, making the

histone code extensive and complex[81,82]. A feature associated with transcriptionally silenced chromatin is trimethylation of histone H3 on lysine 27 (H3K27me3) brought about by PRC2. This multiprotein complex localizes to specific sites in the genome where H3K27 is methylated through its enzymatic subunits EZH1 and EZH2[83]. PRC2 and other chromatin-modifying complexes are often ubiquitously expressed and target a wide range of genes, while the epigenetic pattern heavily depends on cell type and condition. What determines the specificity of these enzymatic complexes has long remained unclear, but is now thought to be mediated by specific lncRNAs[72,83]. Several lncRNAs have been found to associate with and guide PRC2, notable examples are XIST[84], HOTAIR[18] and ANRIL[85]. Interestingly, some lncRNAs target PRC2 in cis to nearby genomic loci[84] while others work in trans[18] on many loci spanning several different chromosomes.

In addition to PRC2, lncRNAs have been found to function as a guide for other chromatin-modifying complexes as well. HOTTIP for instance, regulates gene activation by interacting with WDR5, directly recruiting the MLL H3K4 methylase complex to maintain H3K4me3 (H3K4 methylation is associated with active transcription)[86]. HOTAIR is an intriguing example, as this lncRNA appears to interact with both demethylase and methyltransferase complexes. The 5' domain of HOTAIR interacts with components of PRC2 while the 3' domain binds lysine specific demethylase 1A (LSD1) a component of CoREST/REST repressor complexes. LSD1 mediates H3K4me2 demethylation. In the absence of HOTAIR, H3K4me2 gain and H3K27me3 loss is observed in the HOXD locus, suggesting that HOTAIR functions both as a molecular scaffold and as a guide of PRC2 and LSD1[87] (Figure 2a).

LncRNAs with a guide function are thus believed to associate both with protein and DNA and function as a bridge between both. RNA-protein interactions are fairly common; in fact RNA-binding proteins are one of the most abundant human protein classes with over 1,500 members[88]. RNA-protein interactions require a complex three-dimensional structure of both the RNA and the protein and generally involve conformational changes to either or both interaction partners[89]. The exact mechanism of RNA-DNA interaction however, remains unclear. Local RNA:DNA hybrid triplexes are a possible explanation[90], but there is little evidence to support a

widespread role for such a mechanism. Preliminary results from chromatin isolation by RNA Purification (ChIRP)[91] have revealed that genomic lncRNA binding sites are small, numerous and sequence specific. As chromatin marks typically span several kilobases, this indicates that chromatin-modifying complexes are recruited by lncRNAs to specific loci and subsequently spread out bilaterally[92]. However, more research is needed to uncover the mechanism behind lncRNA-chromatin association.

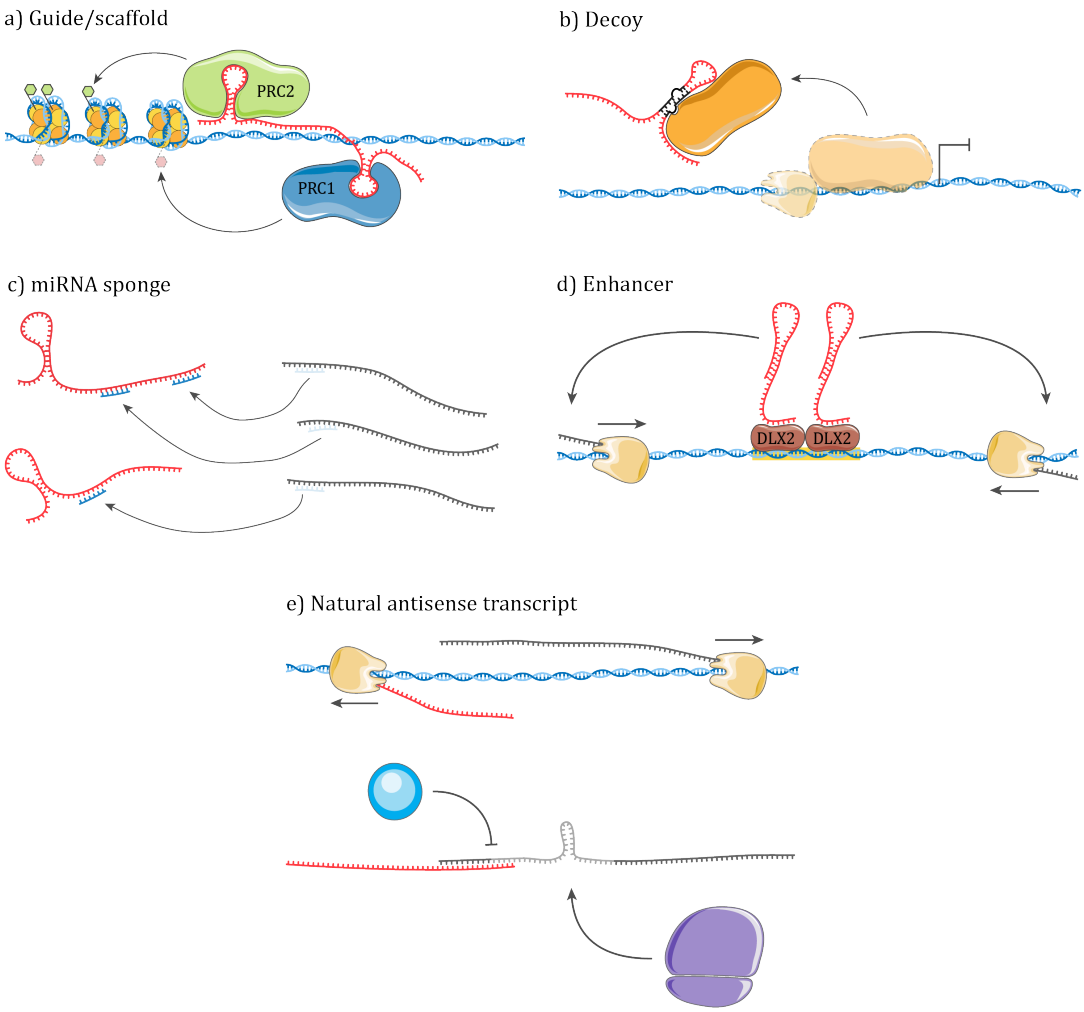


Figure 2: LncRNA modes of action. **a)** HOTAIR functions as a guide and as a molecular scaffold for histone-modifying complexes LSD1 and PRC2. In this way, it silences its target genes. **b)** The lncRNA GAS5 acts as a decoy for the glucocorticoid receptor transcription factor trough a hairpin resembling its genomic target. **c)** The PTEN pseudogene PTENP1 shares a set of miRNA target sites with its ancestor. By binding the miRNAs it prevents downregulation of PTEN. **d)** DLX5 and DLX6 share an enhancer region that harbors EVF2 lncRNA. EVF2 induces DLX5/6 gene expression is cis by interaction with other DLX proteins. **e)** In the absence of the ZEB2 NAT, the internal ribosome entry site is removed from the ZEB2 transcript by splicing and translation is inhibited. Only when the ZEB2 NAT anneals with the splice-site efficient translation is possible.

DECOY LNCRNAS

If lncRNAs can bind with proteins to enhance their function, it is not hard to imagine they can inhibit protein functions as well. The lncRNA GAS5 (growth arrest–specific 5) for instance, functions as a repressor of the glucocorticoid receptor (GR) when cells undergo growth arrest due to starvation. A specific domain located in the stem of a hairpin in the mature GAS5 RNA highly resembles genomic glucocorticoid response elements (GRE) normally found in the regulatory regions of glucocorticoid-responsive genes. GR recognizes and binds this domain on GAS5 and can no longer carry out its normal function as a transcription factor[26] (Figure 2b). Other transcription factors are inhibited by lncRNAs as well. The transcription factor NF-YA, known for the induction of pro-apoptotic genes downstream of p53, is blocked by a lncRNA from the CDKN1A locus named PANDA (P21 associated ncRNA DNA damage activated)[93].

In addition to transcription factors, lncRNAs have been show to function as decoys for DNA methyltransferases. In this way, they can prevent gene silencing in cis as was shown for the CEBPA locus. Together with CEBPA mRNA, a lncRNA spanning the locus in sense is transcribed. This 4.5 kb RNA is termed extra-coding CEBPA (ecCEBPA) as it spans the entire coding region of CEBPA. It interacts with DNA methyltransferases 1 (DNMT1) through a stem-loop structure and protects the CEBPA locus from genomic methylation by the methyltransferase. This mechanism, whereby a sense spanning lncRNA prevents silencing of an actively transcribed locus, may exist for many more genes[94].

COMPETING ENDOGENOUS RNAS

MicroRNAs (miRNAs) comprise an extensively studied class of small (21-25 nt) non-coding RNAs. They restrain the translation of their target mRNAs typically by incomplete basepairing with the 3’ UTR region of the target at specific seed regions. The target RNA is recognized and degraded by a miRNA-loaded RISC protein complex, ultimately leading to a decrease in the protein abundance of the target[95]. LncRNAs can interfere with the function of miRNAs, as was first shown for the PTEN pseudogene PTENP1. Since the 3’ UTR of PTENP1 is highly homologous to that of PTEN, they share several miRNA seeds. Regulatory miRNAs that would normally

target PTEN bind to PTENP1 instead. PTENP1 thus acts as a miRNA decoy and prevents downregulation of the PTEN tumor suppressor[57] (Figure 2c). LncRNAs that carry out this kind of post-transcriptional regulation are also referred to as competing endogenous RNAs (ceRNA) or miRNA sponges. Other examples of lncRNAs belonging to this subgroup are LNCMD1[96] and LINC-ROR[97].

A very peculiar subtype of lncRNA that must be mentioned here is circular RNA (circRNA). Members of this recently discovered class of RNA, such as ciRS-7, have been shown to function as highly efficient miRNA sponges due to their resistance to conventional miRNA destabilization[98,99].

However, these individual examples are most likely oversimplifications as crosstalk between different RNA species through miRNA binding sites is likely common. It is hypothesized that a large number of RNAs, both coding and non-coding, compete for the same set of miRNAs, thus forming a large-scale regulatory network[100].

LNCRNAS CONTROL TRANSCRIPTION IN CIS

Several lncRNAs have been found to directly regulate the expression of other genes in the same locus[101]. Furthermore, a large number of lncRNAs reside in annotated enhancer regions[102,103]. An example is EVF2, a lncRNA transcribed in the enhancer region of DLX5 and DLX6. EVF2 combines with DLX2, a protein encoded in a different DLX gene cluster. Together, the EVF2-DLX2 complex promotes the transcription of the DLX5/6 gene cluster (Figure 2d). Using incremental deletions, the functional domain of EVF2 was narrowed down to a 300 bp region that corresponds to an ultraconserved region in the genome[104].

A different subclass of lncRNAs that regulates gene expression in cis is that of the natural antisense transcripts (NATs). NATs are typically defined as transcripts that overlap in part with a protein coding transcript but are transcribed from the opposite DNA strand[105]. For the majority of protein coding loci, antisense transcription is observed, making this a large but poorly understood subclass of lncRNAs. Although several NATs function by directing epigenetic mechanisms already described in previous sections, they can also reduce transcription of the sense strand via a mechanism determined by their orientation. The transcription

collision model states that the act of transcription on the antisense strand rather than its product inhibits transcription of the sense strand[77]. It is unclear however how many NATs follow this model of transcriptional repression. Due to their (partial) sequence complementarity, NATs can form RNA duplexes with the sense transcript and interfere with splicing and RNA editing[77]. A well-studied example of this kind of post-transcriptional regulation is found in the ZEB2 locus. In the absence of the NAT, which is transcribed from a different promoter downstream of the ZEB2 promotor, the ZEB2 gene cannot be translated. Only when the ZEB2 NAT is expressed together with the ZEB2 transcript, efficient translation occurs. This particular NAT complements a splice site in the 5' region of the ZEB2 transcript and induces intron retention upon annealing. The intron contains an internal ribosome entry site required for efficient translation and expression of the ZEB2 protein[106] (Figure 2e).

Although antisense transcription seems to be a prevalent feature of eukaryote genes, the NATs are likely a diverse group of lncRNAs that enhance or reduce transcription of the sense transcript.

In addition to NATs, many transcript loci produce Promoter Upstream Transcripts (PROMPTs). These unstable RNA transcripts are produced up to 2.5 kilobases upstream of active transcription start sites. The function PROMPTs is currently unresolved although it is speculated that they affect nearby gene expression by competition for transcription machinery[107,108].

LncRNA subclassification

It is apparent that several lncRNAs have a function that is mechanistically dependent on their relative position and orientation to adjacent protein coding genes. As a result, this is often used to subclassify lncRNAs into distinct classes[109-112]. Although different authors have been using slightly different definitions, the following five classes are typically distinguished:

Antisense. The lncRNA is transcribed from the opposite strand to the protein coding gene. Overlap can be complete or partial.

Intronic. The entire lncRNA transcribed is contained within an intron of a protein coding gene. Sometimes a further distinction is made according to the relative orientation to the protein coding gene.

Bidirectional. The lncRNA and the protein coding gene are divergently transcribed with the start positions within a few hundred basepairs of each other.

Sense overlapping. The lncRNA and protein coding gene overlap and reside on the same strand. As such they share a portion of their sequence.

Intergenic. LncRNA that no dot overlap with a protein coding gene on either strand.

I.5. LNCRNA CONSERVATION

In the past, sequence conservation has proven to be a valuable selector for functionality. The low conservation of lncRNAs compared to mRNAs[8,113] has thus led some researchers to regard them as just functionless artifacts of transcription[114]. Notably, the amount of non-coding DNA in an organism's genome is highly correlated with biological complexity[115]. As such, one could argue that the differences found in protein coding genes alone cannot explain the distinction between higher evolved species and more primitive ones, and that recently evolved evolutionary adaptations can only be explained by recently evolved genetic entities found in the non-coding part of the genome. In addition, basepair conservation scores such as PhastCons[116] or PhyloP[117] might not be suitable measures to assess the true functional conservation of a non-coding gene. For instance, although the function of XIST is conserved between human and mouse, its sequence shows poor overall conservation[118]. The position of the XIST-specific tandem repeats (key elements in the function of XIST) however showed striking resemblance between the two species. This suggests that not the sequence but the position and pattern of tandem repeats is responsible for the functional conservation of the gene[118]. In addition, a large-scale study on zebrafish lncRNAs showed that for the majority of zebrafish lncRNAs, sequence similarity to mammalian lncRNAs is absent or limited to a short region of high conservation. However, in two lncRNA knockout models, the phenotype could be rescued by adding the mouse or human ortholog[21]. This demonstrates again that functional conservation of lncRNAs not necessarily requires

sequence conservation. As a result, other measures for conservation have been developed and used to study lncRNA evolution. When examining the conservation of splice sites, it was found that more than 70% of the human lncRNAs is conserved within placental mammals and 15% dates back even further[119] (Figure 3 left panel). The conservation of lncRNA promoter sequences has been studied as well, and evolutionary selection to an extent comparable to that of protein coding genes could be detected for the majority of lncRNA promoters[120]. Of particular interest in this regard is a large-scale RNA sequencing effort to explore the evolution of lncRNAs based on their transcripts[121]. In this study, the transcriptomes of 8 organs from 11 tetrapod species (ranging from Xenopus to human) were sequenced. With only a limited number of species-specific lncRNAs, over 80% was found to be primate-specific (Figure 3 right panel). Although the number is relatively small, 425 lncRNAs (3%) appear to have originated more than 300 million year ago. Interestingly, these ancient lncRNAs have promoters enriched with homeobox transcription factor binding sites, suggesting a role in embryogenesis. Although the different methods used to assess lncRNA conservation each have their differences and particularities, they all agree that the great majority of lncRNAs is conserved to a larger extent than initially presumed only based on sequence conservation.

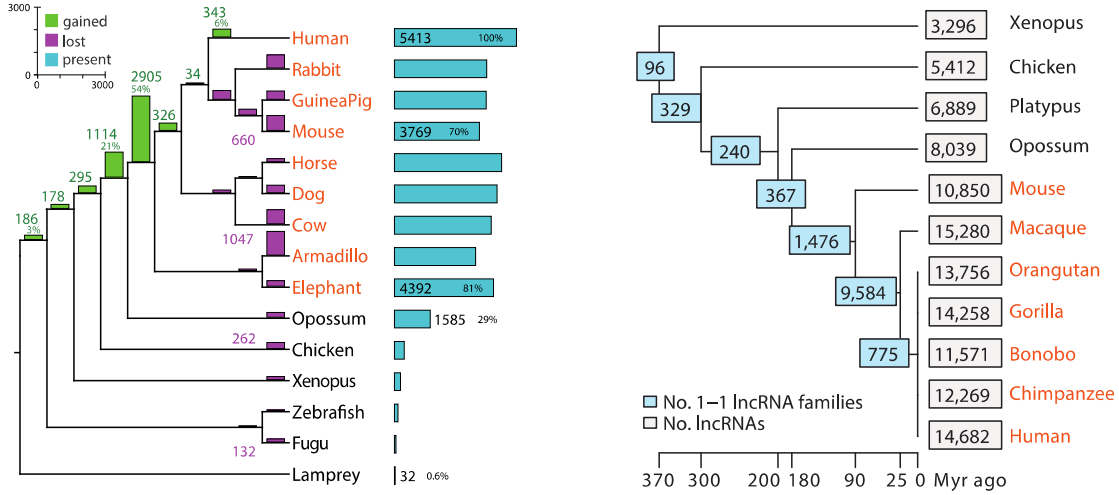


Figure 3: Two different methods shed different lights on lncRNA evolution. Splice-site conservation (left panel) suggests that most lncRNAs were already present at the divergence of placental mammals (orange). Of the 5,413 human lncRNAs examined in this study, 2,905 (54%) have emerged at this divergence and a substantial number are even older. Transcript sequence similarity (right panel) however, suggests most lncRNAs are more recent evolutionary adaptations. Here, only 1,476 (10%) lncRNA families are found to be specific for placental mammals while 9,584 (65%) are primate-specific. Even though the methods disagree on the evolutionary age of lncRNAs, they both yield a large number of conserved lncRNAs. Adapted from Nitsche *et al.*[119] and Necsulea *et al.*[121].

I.6. STUDYING LNCRNA STRUCTURE

It is often speculated that lncRNAs perform their function through a specific and complex secondary and tertiary structure. Unfortunately, lncRNA structure and its relation to function is currently poorly understood and functional reports are mostly limited to small domains of the lncRNA. For instance, the tandem repeat regions in XIST are not only the most conserved part of the gene; they have also been shown to form an intricate stem-loop structure[122]. Components of PRC2 can specifically interact with this structure, suggesting that this domain functions in PRC2 recruitment. The most conserved part of MALAT1 corresponds to a cloverleaf-like structure at the 3' end of the transcript. Further processing of the transcript results in cleavage, producing a small tRNA-like RNA called the mascRNA (MALAT1 associated small cytoplasmic RNA). The function of mascRNA however remains unclear[123].

Secondary RNA structures can be studied using either *in silico* or *in vitro* methodologies. Several algorithms have been implemented to predict the most probable conformation of nucleic acids in the cellular environment. A popular approach is the use of dynamic programming to find the set of base pairings that result in the structure with the lowest free energy. Quite a few programs are implementations of such an algorithm, including UNAFold[124], RNAstructure[125] and the ViennaRNA suite[126]. It is important to note that these algorithms are prone to false positives on long (> 200nt) RNA sequences and such MFE structures are unreliable. Therefore, it is recommended to use a sliding window approach and focus on local structures[127]. *In vitro* approaches are primarily based on determining the positional susceptibility to certain chemical modifications or nucleases. For instance in Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) [128], the 2'-hydroxyl group the RNA ribose component is chemically modified. Single stranded regions, loops ore bulges are more sensitive to this reaction. To analyze RNA structures transcriptome-wide, SHAPE can be followed by deep RNA sequencing. This recently developed approach is referred to as SHAPE-Seq[129]. Similarly, Parallel Analysis of RNA Structure (PARS) is based on deep RNA sequencing of nuclease treated RNA. Two different nucleases are employed: the RNase V1,

which specifically targets single stranded RNA and the S1 nuclease, which cleaves double stranded RNA[130].

I.7. LNCRNA CODING POTENTIAL

Although a task of discriminating protein coding from non-coding RNA seems trivial, it has proven to be a topic of much debate.

IN SILICO PREDICTION OF CODING ORFS

A plethora of computational methods that aim to distinguish protein coding from non-coding sequences have been developed. Each method investigates features in the sequence or evolution of coding ORFs that set them apart from non-coding sequences to score transcripts of unknown coding potential. Often, some form of machine learning is involved in the feature selection and scoring. Notable examples are CPC[131], CONC[132], PORTRAIT[133], CPAT[134], PLEK[135], iSeeRNA[136] and PhyloCSF[137]. There is a striking similarity in the feature sets these algorithms use. Without assessing and comparing every algorithm in detail, a selection of interesting features is explored in the following paragraphs.

By chance, a random progression of nucleotides can contain a short canonical ORF, but long ORFs are unlikely to be incidental. Indeed, the (relative) ORF size is often found to be the most powerful individual discriminator on typically used benchmarking sets[132,136]. However, one might wonder if this is not just reflecting a bias in current annotation. This is a plausible explanation as research and annotation groups have been focusing primarily on transcripts containing ORFs larger than 300 nucleotides (100 amino acids)[138]. Currently, proteomics groups are shifting their focus to small (less than 100 amino acids) but functional proteins, often called micropeptides[139]. Although further research on this topic is needed, it is not unthinkable that algorithms using current annotation as training or benchmarking data and ORF size as a scoring feature will generate false negatives and are unsuitable for detecting novel micropeptides in unannotated transcripts.

A second feature used by several coding potential prediction programs is similarity to annotated protein sequences[131,132]. Bioinformatics tool such as BLASTX allow efficient querying of protein databases using nucleotide input sequences[140]. The ORF

sequence can thus be used and the number of hits and the corresponding scores can serve as training parameters for the prediction model[131,132]. A high homology to the sequences of known proteins may however be unsuitable to detect coding ORFs in novel transcripts. ORFs encoding proteins that are biologically distinct from already reported proteins, such as ORFs of micropeptides, will likely be classified as non-coding. In addition, since most annotated protein coding genes are represented in protein databases as well, typically used benchmarking datasets will result in an overestimation of the sensitivity obtained from this feature.

Although coding and non-coding RNA share the same alphabet, they speak a different language. The nucleotide composition of putative coding ORFs is thus expected to differ from that of ORFs arisen by random variation. In addition to single nucleotide distributions, k-mer distributions (with k ranging from 1 to 6) have been proven to be informative as well[134,135]. Trimer distributions are especially important as those reflect the codon usage within the ORF. While several different codons can be translated to the same amino acid, some codons seem preferred over others. This codon usage bias is one of the oldest described features of coding sequences[141]. Overall, the nucleotide composition entails a powerful set of features based on intrinsic properties of coding ORFs with little bias to current annotations.

Over the course of evolution, synonymous nucleotide substitutions are more common since they do not alter the function of the protein. This evolutionary pattern can be observed in the codon substitution frequencies across multispecies whole genome alignments. The phyloCSF algorithm[137] makes use of these codon substitution frequencies to estimate the likelihood that a given ORF represents a conserved coding sequence. PhyloCSF was able to detect novel proteins in many use cases and it has proven to be applicable even for the detection of small proteins including micropeptides[142].

In conclusion, a wide selection of programs or methods to assess coding potential in unannotated transcripts is available. The performance of several programs is excellent, with sensitivities and specificities well exceeding 95% on the used training and test data[134]. However, training data are based on the current annotation of

protein coding genes, which shows a strong bias toward proteins larger than 100 amino acids[143] with high evolutionary conservation. Even though it is uncertain how many short or evolutionary recent proteins remain to be discovered, current coding potential prediction methods are perhaps not the most suitable means to answer this question. As a result, *in silico* predictions show very little coding potential in present-day lncRNA annotations.

RIBOSOME PROFILING: RIBOSOME OCCUPANCY AS AN INDICATOR FOR TRANSLATION

Advances in next-generation sequencing allowed the development of numerous methods to study specific entities in the genome and transcriptome[144]. Ribosome profiling (also known as ribosome footprinting or Ribo-Seq) is one of those methods and has gained much attention in recent years. Here, enzymatic degradation of RNA not associated with ribosomes, followed by deep sequencing of the 28-30 nucleotide ribosome protected fragments is used to map ribosome occupancy to single basepair resolution[145]. Ribosome profiling experiments revealed significant ribosome occupancy on non-AUG ORFs, upstream ORFs (uORFs) and lncRNAs[146]. These findings, although controversial, have a great impact on our current understanding of translation and on the numbers of protein coding genes (and therefore also lncRNA genes) in the genome. The discovery of ribosome occupancy on the majority of lncRNA transcripts[147] has been a topic of much debate in recent years. Different research groups have come to conflicting, even opposite conclusions on the number of true non-coding lncRNAs. To separate coding from non-coding ORFs using ribosome profiling data, several authors have developed metrics based on specific properties of translated RNA (Table 2). Although all of these methods acknowledge the existence of coding ORFs in genes currently annotated as lncRNA, most authors agree that the number of misclassified lncRNAs is low. Some authors however, remain convinced that ribosome occupancy indicates that the majority of lncRNAs is translated and as such are protein coding genes[148].

Besides the discussion on the interpretation of ribosome footprints from lncRNA transcripts, one may also pose questions on the usefulness of ribosome profiling for the discovery of new and functional proteins. For instance, uORFs have been recognized as regulatory elements for several years. Although they function by

association with ribosomes, they probably do not encode functional proteins[149,150]. The finding that the ribosome footprints of lncRNA ORFs resemble those of uORFs more than those of coding ORFs (Figure 4) further supports the hypothesis that the ribosome footprints on lncRNA ORFs resemble regulatory rather than coding events[151]. Regulatory ORFs on lncRNAs may for instance control the steady state[152] or subcellular localization[153] of the transcript.

Despite the recent advances in ribosome profiling analysis, the debate on the role of lncRNA ORFs will probably last until functional studies show that such ORFs are nonessential for the function of the lncRNA or that proteomics shows that they encode functional and stable peptides.

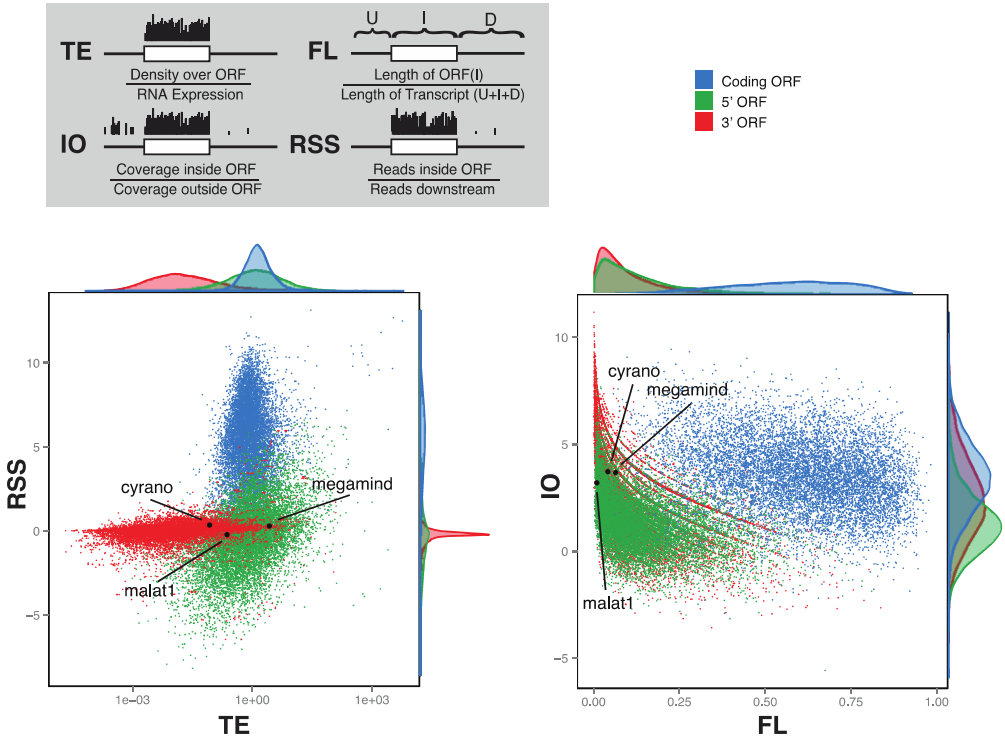


Figure 4: Although ribosome occupancy is found on both coding and non-coding ORFs, several metrics can clearly distinguish these profiles. Coding ORFs have lower ribosome occupancy after the stop codon, resulting in a higher ribosome release score (RSS) compared to ORFs in the UTRs. Furthermore, 3' ORFs exhibit a lower translation efficiency (TE). Adapted from Chew *et al.*[151]

Table 2: Overview of metrics and methods used by different authors to distinguish coding and non-coding ORFs based on ribosome profiling. Different authors draw contrasting conclusions from the ribosome occupancy observed on lncRNA. While some report evidence for translation for the great majority of lncRNAs, others find only small numbers of true translation events.

Metric	Publication	Definition	Conclusions
Translational efficiency (TE)	Ingolia *et al.*, 2011[147] Ruiz-Orera *et al*,., 2014[154]	Ratio of number of ribosome protected reads and RNA-seq reads for an ORF.	The majority of lncRNAs contain regions of high translation comparable to protein coding genes.
Translation initiation sites (TIS)	Lee *et al.*, 2012[155]	Lactimidomycin was used to specifically stall initiating ribosomes. A TIS is a position in which initiating ribosomes are enriched above a measured background.	A limited number (4%) of the analyzed non-coding RNA loci show evidence of translation.
Ribosome release score (RRS)	Guttman *et al.*, 2013[156]	Ratio between the normalized number of reads that are contained within the putative ORF and the normalized number of reads contained within the putative 3' UTR.	RRS nicely separates translated RNAs and lncRNAs
Translated ORF classifier (TOC)	Chew *et al.*, 2013[151]	Random forest classifier combining TE, RRS, the ratio of bases covered within an ORF	Less than 10% of mouse lncRNA loci are classified as coding. Interestingly, most

		versus outside and the relative ORF size. Trained on annotated protein coding genes and classifies ORFs as coding, leader-like and trailer-like.	lncRNA ORFs resemble upstream ORFs of coding genes.
ORFscore	Bazzini *et al.*, 2014[157]	The periodicity of ribosome movement is detectable in ribosome profiling data. The proportion of codons with in-frame reads is here compared to a uniform distribution using a modified chi-squared statistic.	Less then 1% of the analyzed lncRNA transcripts contain a translated ORF.
Fragment length organization similarity score (FLOSS)	Ingolia *et al.*, 2014[148]	Measures the disagreement between the observed and expected fragment length distribution.	The vast majority of lncRNAs (90%) were classified with protein coding genes .

MASS SPECTROMETRY: SEE IT TO BELIEVE IT

Shotgun proteomics is the method of choice for high throughput analysis of proteins in complex mixtures by identification of individual peptides. In this approach, proteins are first enzymatically digested, producing complex mixtures of peptides. Next, peptides are separated often using liquid chromatography (LC). Typically, individual peptides are ionized and then analyzed in a two-step process called tandem mass spectrometry (MS/MS). First, the mass-to-charge ratio (m/z) of the entire peptide ion is measured followed by fragmentation of this peptide ion and acquisition of an m/z spectrum of its fragment ions (MS/MS spectrum). In this way a MS/MS spectrum is obtained for every peptide[158,159]. From this spectrum the peptide's amino acid composition can be read using computational methods. Commonly, a database search method is used whereby MS/MS spectra are compared to theoretical spectra generated from known protein sequences[158]. To use this approach for the discovery of novel proteins, protein sequence databases must be extended with predicted protein sequences. In addition, computationally intensive *de novo* sequencing can predict the sequence from an MS/MS spectrum[158].

Several research groups have turned to shotgun proteomics to evaluate putative ORFs on lncRNAs. Therefore, predicted ORFs based on transcriptomes or genome sequences are added to the protein search space. The first such effort, termed Pinstripe, employed non-redundant peptides from the public PRoteomics Identifications DatabasE (PRIDE). The peptide sequences were mapped to a custom transcriptome based on RNA sequencing of 16 human tissues. From this transcriptome the authors reported 736 canonical open reading frames (ORFs) supported by three or more PRIDE peptides compared to over 32,000 non-coding loci[160]. A significant fraction, although the authors admit their method is likely to generate considerable number of false positives. Indeed, more extensive approaches using (re)processing of MS/MS spectra have come up with much smaller numbers. Slavoff *et al*. combined proteomic and transcriptomic analysis of the K562 human leukemia (CML) cell line. Their database consisted of RefSeq mRNA transcripts and three-frame translated transcripts obtained from RNA sequencing. By analyzing MS/MS spectra against this database they identified 90 micropeptides, 8 of which

are encoded by lncRNAs. These 8 micropeptides represent just 0.4% of the 1,866 lncRNA transcripts detected using RNA sequencing[161]. Similar approaches that instead made use of ribosome profiling to define the *in silico* peptidome have been developed as well[162-164]. In mouse and human cell lines these approaches came up with respectively 83 and 22 novel micropeptides.

Overall, the numbers of (small) proteins encoded by lncRNAs that are detected by mass spectrometry appear to be rather limited and not in line with the high numbers suggested by some ribosome profiling studies. Several explanations for this discrepancy are possible. First of all, ribosome occupancy alone may not be a good indicator for active translation, as was already indicated by specific ribosome profiling efforts[156]. Secondly, it is possible that the translation events produce unstable proteins that are readily degraded and as such undetectable by proteomics. In addition, it is possible that the translation events are rare and generate only low amounts of protein that are difficult to detect using proteomics.

It is apparent that the discrimination between coding and non-coding RNA is far from trivial. Even though the advent of ribosome profiling promised better insight in the translation of the transcriptome it leaves much room for interpretation. The finding that ribosome occupancy on lncRNAs more closely resembles that of non-coding ORFs in mRNA UTRs along with the low numbers of detected proteins by means of proteomics points to a true non-coding role for lncRNAs.

I.8. LNCRNA ANNOTATION IN REFERENCE DATABASES

As lncRNA is currently well accepted as a genetic subclass by the genetic research community, lncRNA annotations are slowly finding their way to the international reference databases. These curated annotations represent a more established subset of lncRNAs based on several lines of evidence. Both the European Ensembl[165] initiative and their American counterpart RefSeq[166] make use of a combination of automated annotation pipelines and manual curation. RefSeq classifies RNA sequences as either coding (NM_* records) or non-coding (NR_*) records. Therefore, the non-coding records are not limited to lncRNAs but include non-coding transcripts such as transcribed pseudogenes or non-coding isoforms of protein-coding genes.

Ensembl makes use of a more elaborate classification schema consisting of many biotypes. Interestingly, lncRNAs are subclassified in 11 distinct biotypes including lincRNA (long interspersed non-coding RNA)[167].

I.9. CONCLUSION

Although the first lncRNA was discovered in 1990[168], it took several decades for geneticists to grasp the true scale of this genetic class. Recent advancements in next-generation sequencing technology uncovered tens of thousands of lncRNA loci in the human genome. Even though the great majority remains to be functionally studied, some common themes seem to be emerging.

By an assortment of molecular mechanisms, lncRNAs can affect gene expression both in cis and in trans. As such, they are involved in many cellular processes and play a role in different genetic diseases. Our current understanding of lncRNA evolution and conservation is poor and restricted to preliminary research that looks beyond sequence conservation to identify lncRNA orthologs. Similarly, the assessment of lncRNA coding potential remains elusive, with conflicting reports coming from authors using slightly different analysis methods. All in all, lncRNAs have proven to be a class of genes with intriguing features. Nevertheless many secrets remain that will likely continue to unfold over the next few years.

I.10. REFERENCES

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3. Eddy, S. R. Non-Coding Rna Genes and the Modern Rna World. *Nature Reviews Genetics* **2,** 919–929 (2001).
4. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
5. The FANTOM Consortium. The Transcriptional Landscape of the Mammalian Genome. *Science* **309,** 1559–1563 (2005).
6. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* (2011). doi:10.1101/gad.17446611
7. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics* **9,** (2013).
8. Nielsen, M. M. *et al.* Identification of expressed and conserved human noncoding RNAs. *RNA* **20,** 236–251 (2014).
9. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22,** 1775–1789 (2012).
10. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* (2015). doi:10.1038/ng.3192
11. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
12. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research* **43,** D1079–D1085 (2015).
13. Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research* **43,** D168–D173 (2015).
14. Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71,** 527–542 (1992).
15. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137 (1996).
16. Lee, J. T., Davidow, L. S. & Warshawsky, D. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.* **21,** 400–404 (1999).
17. Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338,** 1435–1439 (2012).
18. Rinn, J. L. *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Gastroenterology* **129,** 1311–1323 (2007).
19. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* (2011). doi:10.1038/nature10398

20. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42,** 1113–1117 (2010).
21. Ulitsky, I., Shkumatava, A., Jan, C., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–1550 (2011).
22. Klattenhoff, C. A. *et al.* Braveheart, a Long Noncoding RNA Required for Cardiovascular Lineage Commitment. *Gastroenterology* (2013). doi:10.1016/j.cell.2013.01.003
23. Grote, P. *et al.* The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. *Developmental Cell* **24,** 206–214 (2013).
24. Gomez, J. A. *et al.* The NeST Long ncRNA Controls Microbial Susceptibility and Epigenetic Activation of the Interferon-γ Locus. *Cell* **152,** 743–754 (2013).
25. Nakagawa, S., Naganuma, T., Shioi, G. & Hirose, T. Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *The Journal of Cell Biology* **193,** 31–39 (2011).
26. Kino, T., Hurt, D. E., Ichijo, T., Nader, N. & Chrousos, G. P. Noncoding RNA Gas5 Is a Growth Arrest and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Science signaling* **3,** ra8–ra8 (2010).
27. Chen, G. *et al.* LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Research* (2012). doi:10.1093/nar/gks1099
28. De Smet, E. G., Mestdagh, P., Vandesompele, J., Brusselle, G. G. & Bracke, K. R. Non-coding RNAs in the pathogenesis of COPD. *Thorax* thoraxjnl–2014–206560 (2015). doi:10.1136/thoraxjnl-2014-206560
29. Zhang, Q., Chen, C.-Y., Yedavalli, V. S. R. K. & Jeang, K.-T. NEAT1 Long Noncoding RNA and Paraspeckle Bodies Modulate HIV-1 Posttranscriptional Expression. *mBio* **4,** e00596–12–e00596–12 (2013).
30. Tan, L., Yu, J.-T., Hu, N. & Tan, L. Non-coding RNAs in Alzheimer's Disease. *Mol Neurobiol* **47,** 382–393–393 (2013).
31. Uchida, S. & Dimmeler, S. Long Noncoding RNAs in Cardiovascular Diseases. *Circulation Research* **116,** 737–750 (2015).
32. Wu, G.-C. *et al.* Emerging role of long noncoding RNAs in autoimmune diseases. *Autoimmunity Reviews* **14,** 798–805 (2015).
33. Reik, W. *et al.* Allelic methylation of H19 and IGF2 in the Beckwith-Wiedemann syndrome. *Hum. Mol. Genet.* **3,** 1297–1301 (1994).
34. Meng, L. *et al.* Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. *Nature* **518,** 409–412 (2015).
35. Poplinski, A., Wieacker, P., Kliesch, S. & Gromoll, J. Severe XIST hypomethylation clearly distinguishes (SRY+) 46,XX-maleness from Klinefelter syndrome. *European Journal of Endocrinology* **162,** 169–175 (2010).
36. De Baere, E. *et al.* Identification of BPESC1, a novel gene disrupted by a balanced chromosomal translocation, t(3;4)(q23;p15.2), in a patient with BPES. *Genomics* **68,** 296–304 (2000).
37. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70 (2000).

38. Gutschner, T. & Diederichs, S. The Hallmarks of Cancer: A long non-coding RNA point of view. *rnabiology* **9,** 0––1 (2012).
39. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).
40. Kogo, R. *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Research* **71,** 6320–6326 (2011).
41. Kim, K. *et al.* HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* **32,** 1616–1625 (2013).
42. Ma, Y. *et al.* Long non-coding RNA CCAL regulates colorectal cancer progression by activating Wnt/β-catenin signalling pathway via suppression of activator protein 2α. *Gut* gutjnl–2014–308392 (2015). doi:10.1136/gutjnl-2014-308392
43. Hu, X. *et al.* A Functional Genomic Approach Identifies FAL1 as an Oncogenic Long Noncoding RNA that Associates with BMI1 and Represses p21 Expression in Cancer. *Cancer Cell* **26,** 344–357 (2014).
44. Gutschner, T. *et al.* The non-coding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research* **73,** canres.2850.2012–1189 (2012).
45. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22,** 8031–8041 (2003).
46. Lai, M.-C. *et al.* Long non-coding RNA MALAT-1 overexpression predicts tumor recurrence of hepatocellular carcinoma after liver transplantation. *Med Oncol* **29,** 1810–1816 (2011).
47. Ying, L. *et al.* Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. *Mol. BioSyst.* **8,** 2289–2294 (2012).
48. Srikantan, V. *et al.* PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 12216–12221 (2000).
49. Ding, J. *et al.* Expression and clinical significance of the long non-coding RNA PVT1 in human gastric cancer. *OncoTargets and therapy* **7,** 1625–1630 (2014).
50. Guan, Y. *et al.* Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin Cancer Res* **13,** 5745–5755 (2007).
51. Mourtada-Maarabouni, M., Pickard, M. R., Hedge, V. L., Farzaneh, F. & Williams, G. T. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* **28,** 195–208 (2009).
52. Zhang, X. *et al.* A Pituitary-Derived MEG3 Isoform Functions as a Growth Suppressor in Tumor Cells. *The Journal of Clinical Endocrinology & Metabolism* **88,** 5119–5126 (2003).
53. Zhou, Y., Zhang, X. & Klibanski, A. MEG3 noncoding RNA: a tumor suppressor. *J Mol Endocrinol* **48,** R45–53 (2012).
54. Zhang, X. *et al.* Maternally Expressed Gene 3, an Imprinted Noncoding RNA Gene, Is Associated with Meningioma Pathogenesis and Progression. *Cancer Research* **70,** 2350–2358 (2010).
55. Braconi, C. *et al.* microRNA-29 can regulate expression of the long non-coding RNA gene MEG3 in hepatocellular cancer. *Oncogene* **30,** 4750–

4756 (2011).
56. Wang, C., Yan, G., Zhang, Y., Jia, X. & Bo, P. Long non-coding RNA MEG3 suppresses migration and invasion of thyroid carcinoma by targeting of Rac1. *Neoplasma* (2015). doi:10.4149/neo_2015_065
57. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465,** 1033–1038 (2010).
58. Liu, Q. *et al.* LncRNA loc285194 is a p53-regulated tumor suppressor. *Nucleic Acids Research* **41,** 4976–4987 (2013).
59. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
60. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
61. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
62. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61,** 759–767 (1990).
63. Stuart, D. & Sellers, W. R. Linking somatic genetic alterations in cancer to therapeutics. *Cell regulation* **21,** 304–310 (2009).
64. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
65. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology* **20,** 908–913 (2013).
66. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **108,** 2419–2425 (2013).
67. Pasmant, E., Sabbagh, A., Vidaud, M. & Bièche, I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25,** 444–448 (2011).
68. Gong, J., Liu, W., Zhang, J., Miao, X. & Guo, A.-Y. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Research* **43,** D181–6 (2015).
69. Sabarinathan, R. *et al.* RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. *Human Mutation* **34,** 546–556 (2013).
70. Xie, H., Ma, H. & Zhou, D. Plasma HULC as a Promising Novel Biomarker for the Detection of Hepatocellular Carcinoma. *BioMed Research International* **2013,** 1–5 (2013).
71. Prensner, J. R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology* **29,** 742–749 (2011).
72. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11667–11672 (2009).
73. Stein, C. A. Does antisense exist? *Nat Med* **1,** 1119–1121 (1995).
74. Burnett, J. C. & Rossi, J. J. RNA-Based Therapeutics: Current Progress and Future Prospects. *Chemistry & Biology* **19,** 60–71 (2012).
75. RaNA Therapeutics. RaNA Therapeutics, Inc. and Santaris Pharma A/S announce agreement to develop RNA-targeted medicines that selectively activate protein expression. *ranarx.com* (2013). at

<http://ranarx.com/wordpress/wp-content/uploads/RaNA-Santaris-Press-Release-FINAL-7-8-13.pdf>
76. Modarresi, F. *et al.* Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nature Biotechnology* **30,** 453–459
77. Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nature Reviews Molecular Cell Biology* **10,** 637–643 (2009).
78. Işın, M. *et al.* Exosomal lncRNA-p21 levels may help to distinguish prostate cancer from benign disease. *Front. Gene.* **6,** 168 (2015).
79. Center for Devices & Health, R. Recently-Approved Devices - PROGENSA® PCA3 Assay - P100033.
80. Haese, A. *et al.* Clinical Utility of the PCA3 Urine Assay in European Men Scheduled for Repeat Biopsy. *European Urology* **54,** 1081–1088 (2008).
81. Lodish, H. *et al. Molecular Cell Biology*. (W.H.Freeman & Co Ltd, 2007).
82. David Allis, C., Jenuwein, T. & Reinberg, D. *Epigenetics*. (Cold Spring Harbor Laboratory Press, 2007).
83. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–349 (2011).
84. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300,** 131–135 (2003).
85. Kotake, Y. *et al.* Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor gene. *Oncogene* **30,** 1956–1962 (2011).
86. Yang, Y. W. *et al.* Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. *eLife* **3,** e02046 (2014).
87. Tsai, M.-C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329,** 689–693 (2010).
88. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature Reviews Genetics* **15,** 829–845 (2014).
89. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural & Molecular Biology* **20,** 300–307 (2013).
90. Martianov, I., Ramadass, A., Barros, A. S., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445,** 666–670 (2007).
91. Chu, C., Quinn, J. & Chang, H. Y. Chromatin Isolation by RNA Purification (ChIRP). *JoVE* (2012). doi:10.3791/3912
92. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Molecular Cell* **44,** 667–678 (2011).
93. Hung, T. *et al.* Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat. Genet.* **43,** 621–629 (2011).
94. Di Ruscio, A. *et al.* DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503,** 371–376 (2013).
95. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics* **5,** 522–531 (2004).
96. Cesana, M. *et al.* A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA. *Gastroenterology* **147,**

358–369 (2011).
97. Wang, Y. *et al.* Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal. *Developmental Cell* **25,** 69–80 (2013).
98. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20,** 1666–1670 (2014).
99. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495,** 384–388 (2013).
100. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146,** 353–358 (2011).
101. Ørom, U. A. *et al.* Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell* **143,** 46–58 (2010).
102. De Santa, F. *et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol* **8,** e1000384 (2010).
103. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
104. Feng, J. *et al.* The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes & Development* **20,** 1470–1484 (2006).
105. Khorkova, O., Myers, A. J., Hsiao, J. & Wahlestedt, C. Natural antisense transcripts. *Hum. Mol. Genet.* **23,** R54–63 (2014).
106. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes & Development* **22,** 756–769 (2008).
107. Preker, P. *et al.* PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Research* **39,** 7179–7193 (2011).
108. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322,** 1851–1854 (2008).
109. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
110. Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E. & Mattick, J. S. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Research* **39,** D146–D151 (2010).
111. Mattick, J. S. & Rinn, J. L. Discovery and annotation of long noncoding RNAs. *Nature Structural & Molecular Biology* **22,** 5–7 (2015).
112. Mathew W Wright, E. A. B. Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature. *Human genomics* **5,** 90 (2011).
113. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28,** 503–510 (2010).
114. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14,** 103–105 (2007).
115. Taft, R. J. & Mattick, J. S. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology* **5,** P1–24 (2003).
116. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect,

worm, and yeast genomes. *Genome Research* **15,** 1034–1050 (2005).
117. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20,** 110–121 (2010).
118. Nesterova, T. B. *et al.* Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Research* **11,** 833–849 (2001).
119. Nitsche, A., Rose, D., Fasold, M., Reiche, K. & Stadler, P. F. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *RNA* **21,** 801–812 (2015).
120. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research* **17,** 556–565 (2007).
121. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).
122. Maenner, S. *et al.* 2-D Structure of the A Region of Xist RNA and Its Implication for PRC2 Association. *PLoS Biol* **8,** e1000276 (2010).
123. Wilusz, J. E., Freier, S. M. & Spector, D. L. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135,** 919–932 (2008).
124. Markham, N. R. & Zuker, M. in *Bioinformatics* **453,** 3–31 (Humana Press, 2008).
125. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11,** 129 (2010).
126. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6,** 26 (2011).
127. Bernhart, S. H. & Hofacker, I. L. Think global, fold local. *tbi.univie.ac.at* at <https://www.tbi.univie.ac.at/~ulim/berniebsvposter.pdf>
128. Low, J. T. & Weeks, K. M. SHAPE-directed RNA secondary structure prediction. *Methods* **52,** 150–158 (2010).
129. Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research* **42,** e165–e165 (2014).
130. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).
131. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35,** W345–W349 (2007).
132. Liu, J., Gough, J. & Rost, B. Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. *PLoS Genetics* **2,** e29 (2006).
133. Arrial, R. T., Togawa, R. C. & Brigido, M. M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **10,** 239 (2009).
134. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41,** e74–

e74 (2013).
135. Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15,** 311 (2014).
136. Sun, K. *et al.* iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC GENOMICS* **14 Suppl 2,** S7 (2013).
137. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** I275–I282 (2011).
138. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Computational Biology* **4,** e1000176 (2008).
139. Crappé, J., Van Criekinge, W. & Menschaert, G. Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteomics* **3,** 128–137 (2014).
140. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25,** 3389–3402 (1997).
141. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology* **151,** 389–409 (1981).
142. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160,** 595–606 (2015).
143. Frith, M. C. *et al.* The Abundance of Short Proteins in the Mammalian Proteome. *PLoS Genetics* **2,** e52 (2006).
144. Shendure, J. & Aiden, E. L. The expanding scope of DNA sequencing. *Nature Biotechnology* **30,** 1084–1094 (2012).
145. Ingolia, N. T. in *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis* **470,** 119–142 (Elsevier, 2010).
146. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics* **15,** 205–213 (2014).
147. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147,** 789–802 (2011).
148. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports* **8,** 1365–1379 (2014).
149. Medenbach, J., Seiler, M. & Hentze, M. W. Translational Control via Protein-Regulated Upstream Open Reading Frames. *Cell* **145,** 902–913 (2011).
150. Morris, D. R. & Geballe, A. P. Upstream open reading frames as regulators of mRNA translation. *Mol. Cell. Biol.* **20,** 8635–8642 (2000).
151. Chew, G.-L. *et al.* Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140,** 2828–2834 (2013).

152. Tani, H., Torimura, M. & Akimitsu, N. The RNA Degradation Pathway Regulates the Function of GAS5 a Non-Coding RNA in Mammalian Cells. *PLoS ONE* **8,** e55684 (2013).
153. de Turris, V., Nicholson, P., Orozco, R. Z., Singer, R. H. & Mühlemann, O. Cotranscriptional effect of a premature termination codon revealed by live-cell imaging. *RNA* **17,** 2094–2107 (2011).
154. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3,** e03523 (2014).
155. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E2424–E2432 (2012).
156. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154,** 240–251 (2013).
157. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33,** 981–993 (2014).
158. Nesvizhskii, A. I. in *Mass Spectrometry Data Analysis in Proteomics* **367,** 87–120 (Humana Press, 2006).
159. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
160. Gascoigne, D. K. *et al.* Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28,** 3042–3050 (2012).
161. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology* **9,** 59–64 (2013).
162. Menschaert, G. *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell Proteomics* **12,** 1780–1790 (2013).
163. Koch, A. *et al.* A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics* **14,** 2688–2698 (2014).
164. Crappé, J. *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research* **43,** e29–e29 (2015).
165. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Research* **40,** D84–90 (2012).
166. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research* **40,** D130–D135 (2012).
167. Gene and transcript types. *vega.sanger.ac.uk* at <http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html>
168. Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10,** 28–36 (1990).

II. RESEARCH OBJECTIVES

Even though the lncRNA research field is still young, it is growing at an immense pace. An ever-increasing number of lncRNAs is being reported in literature as the number of research labs that shift their focus to lncRNA grows. The rapidly changing lncRNA annotations in numerous lncRNA resources however, are a burden to the field as this diversification impedes scientific communication. The established genetic databases struggle to keep up with lncRNA literature or take a conservative position and await further research. As both bioinformatics and wet-lab applications rely on lncRNA annotation, lncRNA researchers find it difficult to use the currently available platforms. To address this issue, we developed LNCipedia, a public lncRNA resource that aims to provide the most complete and up-to-date view on the lncRNome (**research paper 1**). In order to fulfill this promise, LNCipedia has been updated on several occasions and currently holds over five times the initial number of entries (**research paper 2**). Without functional validation, it is not straightforward to distinguish between coding and non-coding RNA, as was already thoroughly discussed in the introduction. To address this issue, we have evaluated several methods to assess the coding potential of lncRNA transcripts. In collaboration with the research lab of Prof. Martens, we devised a strategy to query large-scale proteomics datasets for putative protein products of lncRNAs. While we already introduced this method in the LNCipedia publications, we aim to publish a commentary paper as well, in which we comment on the discrepancy observed between our results and some ribosome profiling studies (**research paper 3**).

With a lncRNA database at hand, we were able to develop a number of platforms to functionally study lncRNAs in screening experiments. Gene expression microarrays are often the method of choice for high throughput expression profiling experiments. As commercial platforms typically lack extensive lncRNA annotation we developed and subsequently updated a custom gene expression microarray covering both protein coding and lncRNA transcripts (**case study 1**). In addition, we designed a unique platform to detect small and focal copy-number aberrations targeting lncRNA genes. Since oncogenes and tumor suppressor genes are frequent targets of genetic

amplifications or deletions respectively, we reasoned that screening the cancer genome for genetic aberrations on lncRNAs is a valuable method to identify novel cancer associated lncRNAs. As such, we screened a panel of 80 cancer cell lines using our platform and found many putative cancer associated lncRNAs (**research paper 4**). The ability to transiently impede gene expression *in vitro* is invaluable for functional genomic research. While for protein coding genes this can easily be achieved by various means, their lower expression and nuclear localization make lncRNA harder to target. We evaluated the potential of ASOs for lncRNA knockdown. In addition, we developed a tool to assess the potential of an ASO using its thermodynamic properties and target RNA structure (**research paper 5**).

III. RESULTS

Cataloging lncRNAs

- **Research paper 1**: Volders et al., *LNCipedia: a database for annotated human lncRNA transcript sequences and structures*, **Nucleic acids research** (2013)
- **Research paper 2**: Volders et al., *An update on LNCipedia: a database for annotated human lncRNA sequences*, **Nucleic acids research** (2015)
- **Research paper 3**: Volders and Verheggen et al., *Non-coding after all: Large-scale proteomics reprocessing suggests limited translation of lncRNAs.*

Tools to functionally study lncRNAs

- **Case study 1:** *Development of combined mRNA and lncRNA expression profiling platforms.*
- **Research paper 4**: Volders et al., *Targeted genomic screen reveals focal long non-coding RNA copy number alterations in cancer cells.*
- **Research paper 5**: Volders et al., *Potent antisense oligonucleotide selection for lncRNA knockdown.*

III.1. **RESEARCH PAPER 1**: LNCIPEDIA: A DATABASE FOR ANNOTATED HUMAN LNCRNA TRANSCRIPT SEQUENCES AND STRUCTURES

Pieter-Jan Volders, Kenny Helsens, Xiaowei Wang, Björn Menten, Lennart Martens, Kris Gevaert, Jo Vandesompele and Pieter Mestdagh

Nucleic acids research (2013)

http://nar.oxfordjournals.org/content/41/D1/D246

Contributions: Apart from the PRIDE reprocessing pipeline, the candidate contributed in whole or in part to design of the work, data acquisition, analysis and interpretation and drafting of the manuscript. For the PRIDE reprocessing pipeline and the matching paragraphs in the manuscript, the candidate contributed to the concept of the analysis, the interpretation of the data and the revision of the text.

D246–D251 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 5 October 2012*
doi:10.1093/nar/gks915

LNCipedia: a database for annotated human lncRNA transcript sequences and structures

Pieter-Jan Volders[1], Kenny Helsens[2,3], Xiaowei Wang[4], Björn Menten[1], Lennart Martens[2,3], Kris Gevaert[2,3], Jo Vandesompele[1,*] and Pieter Mestdagh[1,*]

[1]Center for Medical Genetics, Ghent University, [2]Department of Medical Protein Research, VIB, [3]Department of Biochemistry, Ghent University, 9000 Ghent, Belgium and [4]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA

Received August 15, 2012; Accepted September 10, 2012

ABSTRACT

Here, we present LNCipedia (http://www.lncipedia.org), a novel database for human long non-coding RNA (lncRNA) transcripts and genes. LncRNAs constitute a large and diverse class of non-coding RNA genes. Although several lncRNAs have been functionally annotated, the majority remains to be characterized. Different high-throughput methods to identify new lncRNAs (including RNA sequencing and annotation of chromatin-state maps) have been applied in various studies resulting in multiple unrelated lncRNA data sets. LNCipedia offers 21 488 annotated human lncRNA transcripts obtained from different sources. In addition to basic transcript information and gene structure, several statistics are determined for each entry in the database, such as secondary structure information, protein coding potential and microRNA binding sites. Our analyses suggest that, much like microRNAs, many lncRNAs have a significant secondary structure, in-line with their presumed association with proteins or protein complexes. Available literature on specific lncRNAs is linked, and users or authors can submit articles through a web interface. Protein coding potential is assessed by two different prediction algorithms: Coding Potential Calculator and HMMER. In addition, a novel strategy has been integrated for detecting potentially coding lncRNAs by automatically re-analysing the large body of publicly available mass spectrometry data in the PRIDE database. LNCipedia is publicly available and allows users to query and download lncRNA sequences and structures based on different search criteria. The database may serve as a resource to initiate small- and large-scale lncRNA studies. As an example, the LNCipedia content was used to develop a custom microarray for expression profiling of all available lncRNAs.

INTRODUCTION

Long non-coding RNAs (lncRNAs) constitute a recently discovered class of non-coding RNAs that grew in size drastically during the past few years. LncRNA genes give rise to long (>200 bp) and often multiexonic transcripts that are supposed not to get translated to protein, as commonly assessed by means of *in silico* prediction algorithms (1). In comparison with their protein-coding counterparts, lncRNA genes are poorly conserved (2) and are more numerous in biologically complex species (3). Although only a fraction of the lncRNA genes has been characterized experimentally, lncRNAs seem to function as transcriptional regulators through direct interaction with chromatin-modifying proteins and transcription factors (1,4,5).

LncRNAs with experimentally validated functions or expression patterns have been named accordingly. Notable examples are XIST (X inactive-specific transcript) (6), HOTAIR (HOX transcript antisense RNA) (7) and HULC (highly up-regulated in liver cancer) (8). The HUGO Gene Nomenclature Committee currently uses several schemes to name lncRNAs with an unknown function. LncRNAs that reside on the opposite strand to (antisense) or in an intron of (intronic) a protein-coding gene are named after the protein-coding gene with suffixes '-AS' and '-IT', respectively. Intergenic lncRNAs are numbered and get the prefix 'LINC' (9).

Recent advances in non-coding RNA research have led to the creation of several lncRNA resources. LncRNAdb focuses on lncRNA transcripts with well-described functions in literature (10), whereas the ncRNA database

*To whom correspondence should be addressed. Tel: +32 9 3326979; Fax: +32 9 3326549; Email: pieter.mestdagh@ugent.be
Correspondence may also be addressed to Jo Vandesompele. Tel: +32 479 353563; Fax: +32 9 3326549; Email: joke.vandesompele@ugent.be

(ncRNAdb) provides RNA sequences and annotation from different sources (11). The NONCODE database (12) contains a larger collection of human long non-coding RNAs (33 829) obtained from different sources and by different experimental procedures (13). Rfam provides structures and annotation of well-known RNA families along with predictions of new members of these families (14). However, it does not provide information for an individual lncRNA. Although each of these resources provides valuable information, database unification and integration of lncRNA transcript sequence details with a broad set of bioinformatics tools and a universal lncRNA gene building and naming scheme is currently lacking. Here, we present LNCipedia, a catalogue of 21 488 lncRNA transcripts that were clustered into genes and named accordingly, and they were analysed using multiple bioinformatics tools, revealing insights in lncRNA structure, experimentally verified (lack of) protein coding potential, function and regulation. We believe such a database facilitates human lncRNA research and communication among scientists.

DATABASE DEVELOPMENT

The sources used in the data collection step are listed in Table 1. The most recent version of each source at the time of development has been included. The sequences and annotations are extracted and stored in a mongoDB database using custom Perl scripts. To this purpose, import scripts for different file formats, such as FASTA, BED and GFF, have been developed. Redundant transcripts are grouped in a single record, while maintaining all annotation from the original sources. The web interface for LNCipedia is build using the Mojolicious Perl web framework and offers different ways of querying the data (Figure 1). LNCipedia will be updated when newer versions of the lncRNA sources are released or if new sources become available. In addition, researchers are encouraged to submit new transcript sequences or annotations trough lncipedia.org.

Of note, each of the input sources uses a different naming scheme. LncRNA researchers have previously used the gene symbol of the nearest protein coding gene to refer to a given lncRNA (15). Based on this strategy, we have implemented a universal lncRNA nomenclature to ease communication among researchers. Different lncRNA transcripts are considered to belong to the same gene if they share at least one (partially) overlapping exon and reside on the same DNA strand. In this way, transcripts are clustered into genes. These lncRNA genes are then named after the HUGO symbol of the nearest protein-coding gene on the same strand using the following scheme: 'lnc-HUGO-#'. The lncRNA genes are numbered, starting with the lncRNA gene closest to the protein-coding gene. A second number is added to denote the different transcript variants starting with the most upstream transcript, for example, lnc-MYCN-1:1 denotes transcript 1 from gene lnc-MYCN-1 (Figure 2).

Table 1. The different sources of lncRNA transcripts used for LNCipedia at the time of development[a]

Source	Version	Number of transcripts
Ensembl (biotype = lincRNA)	Version 64	9069
Human bodymap lincRNAs (2)		14 279
LncRNAdb (10)	September 2011	134
Total number of unique transcripts		21 488

[a]The database will be updated with new transcripts when new versions of the sources are released.

INTEGRATED ANALYSIS TOOLS

LncRNA-protein interactions are, in part, mediated by the secondary structure of the lncRNA. The Vienna RNA package (16,17) consists of a set of algorithms for predicting and analysing RNA secondary structures. We applied the RNAfold algorithm to generate a secondary structure plot and dot plot with pair probabilities. Both of these images are processed with the provided relplot.pl script to obtain a structure plot with colour annotated base pair probabilities. The output postscript (.ps) images are converted to the graphics interchange format (.gif) for display in web browsers.

Structural RNAs, such as miRNAs, have a significantly lower minimum free energy of folding compared with randomly shuffled sequences (18). The Randfold algorithm implements the randomization test and returns the mean free energy of folding and *P*-value for every RNA sequence. Hence, a significant *P*-value denotes a high propensity in the sequence towards a stable secondary structure.

Recently, it has been shown that lncRNAs can act as a miRNA sponge by binding specific microRNAs and, thus, interfering with their role as negative regulators of gene expression (5,19,20). We include miRNA seed predictions for every lncRNA to allow researchers to evaluate possible miRNA–lncRNA interactions. miRNA seed predictions were performed using the MirTarget2 algoritm (21).

PROTEIN CODING POTENTIAL

Assessment of protein coding potential is an important aspect in the study of non-coding RNAs. LNCipedia reports the outcome of two different protein coding potential prediction algorithms. The Coding Potential Calculator (CPC) applies a support vector machine classifier to the output of open reading frame analysis and Basic Local Alignment Search Tool search (22). CPC returns the predicted status of the transcript (coding/non-coding) and a coding potential score. We applied version 0.9 of the CPC software and report the predicted status and the coding potential score for every transcript. Another popular strategy for detection of

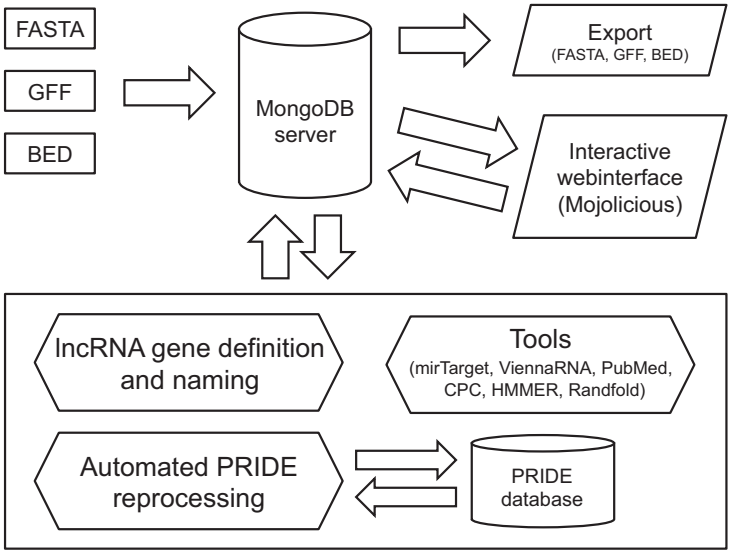


Figure 1. LNCipedia is generated in a multistep process that comprises importing, naming, analysis and visualization of lncRNA genes. Import scripts for the FASTA, BED and GFF file formats process lncRNA transcripts and detect redundancy. LncRNA naming is preceded by the creation of lncRNA transcript clusters and requires information on the nearest protein-coding gene on the same DNA strand. Every lncRNA transcript is subsequently analysed using multiple algorithms, and the results are appended to the database. A web-interface build using Perl enables lncRNA visualization and database querying.

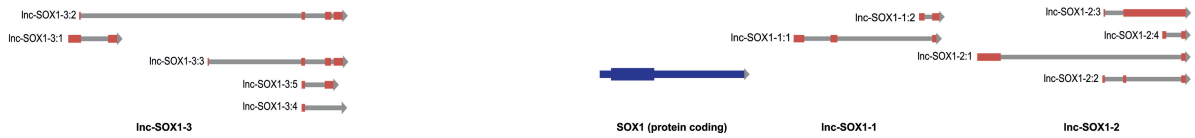


Figure 2. The SOX1 protein-coding gene locus contains three lncRNAs on the same DNA strand, numbered according to their distance in relation to SOX1. LncRNA transcripts are numbered according to their order in the gene, starting with the most upstream transcript.

coding sequences is based on known protein domains. The HMMER3 suite provides software based on hidden Markov models for sequence based homology searches (23). It is often used in combination with the Pfam protein families database (24). Using the hmmscan algorithm, we searched for Pfam protein domains in the RNA sequence. All six reading frames were translated *in silico*, and the number of hits in 5′ to 3′ and 3′ to 5′ direction are reported.

A unique feature of LNCipedia is the incorporation of an automated reprocessing pipeline that relies on publicly available fragmentation spectra from the PRIDE database at EMBL-EBI (25) to detect potentially coding lncRNAs. The concept behind this feature is that mass spectrometry based proteomics data may contain serendipitously recorded mass spectra derived from translated lncRNAs. As standard identification strategies in proteomics are based on searching these spectra against protein sequence databases, such as UniProtKB/Swiss-Prot (26), they are implicitly unable to detect coding forms of lncRNAs, as they are not present in these databases. To uncover such potential traces of coding lncRNAs, the spectra, thus, need to be re-searched against a purpose-built database that comprises a combination of the possible translations of known lncRNAs, the known proteins for that organism as obtained from a traditional sequence database and corresponding decoy sequences for both these constituent databases for quality control and FDR estimation purposes (27). A spectrum can, thus, be matched against a lncRNA, a known protein, or a decoy sequence. The known proteins must be included to prevent relatively low-scoring matches of spectra against lncRNAs to be picked up where a much better match for that spectrum can be found for a known protein.

We have implemented such a pipeline by using the SearchGUI tool (28) to run the X!Tandem (29) search algorithm. All results are then collated and filtered at 1% FDR by the PeptideShaker algorithm (http://code.google.com/p/peptide-shaker). The pipeline infers the original search parameters, such as mass errors and post-translational modifications both directly from the PRIDE database and by using the PRIDE automatic spectrum annotation pipeline (http://code.google.com/p/pride-asa-pipeline). All the tools and algorithms used are freely available as open source.

The pipeline has so far been ran on 149 PRIDE experiments from at least 15 different tissues, yielding 81 579 peptide-to-spectrum matches (PSMs) against the

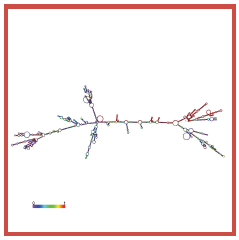
Incipedia.org

A comprehensive compendium of long non-coding RNAs | Home | Database | Search | Download | About | Contact

Transcript: lnc-SMUG1-3:6

Basic information

Incipedia transcript ID: lnc-SMUG1-3:6
Incipedia gene ID: lnc-SMUG1-3
Location: chr12:54356092-54368740
Strand: -
Transcript size: 2421 bp
Exons: 7
Sources: Ensembl release 64 - Sep 2011
Alternative transcript names: ENST00000424518
Alternative gene names: ENSG00000228630; HOTAIR



RNA sequence:

CCAGTTCTCAGGCGAGAGCCGCGGCTGACAGGGTCTGGGACAGAAGGAAAGCCCTCCAGCCTCCAGGCCCTGCCTTCTGCCTGCACATTCTGCCCTGATTTCCGGAACCTGGAAGCC
TAGGCAGGCAGTGGGGAACTCTGACTCGCCTGTGCTCTGGAGCTTGATCCGAAAGCTTCCACAGTGAGGACTGCTCCGTGGGGGTAAGAGAGCACCAGGCACTGAGGCCTGGGAGTT
CCACAGACCAACACCCCTGCTCCTGGCGGCTCCCACCCGGGACTTAGACCCTCAGGTCCCTAATATCCCGGAGGTGCTCTCAATCAGAAAGGTCCTGCTCCGCTTCGCAGTGGAATG
GAACGGATTTACAAGCCTGCAGTAGGGGAGTGGGGAGTGGAGAGAGGGAGCCCAGAGTTACAGACGCGGCGCCAGAGGAAGGAGGGGCGTCTTTATTTTTTAAGGCCCCAAAGAGT
CTGATGTTTACAAGACCAGAAATGCCACGGCGCGCGTCCTGCCAGAGAAAAGGCTGAAATGGAGGACCGGCGCCTTCCTTATAAGTATGCACATTGCGCAGAGAAGTGCTGCAACCTA
AACCAGCAATTACACCCAAGCTCGTTCGGGCCTAAGCCAGTACCGACCTCGTAGAAAAAGCAACCACCAAGCTAGACAGACAGACCCAGACGAGGGAAGAGAGCGCCCAGACCAAGCGTG
AAAGCGAACCACGCAGAGAAATGCAGGCAAGGGAGCAAGGCGGCAGTTCCCGGAACAAACGTGGCAGAGGGCAAGACGGGCACTCACAGACAGACAGGTTTATGTATTTTATTTTTTA
AAATCTGATTTGGTGTTCCATGAGGAAAAGGGAAAATCTAGGGAACGGAGTACAGAGAGAATAATCCGGGTCCTAGCTCGCCACATGAACGCCCAGAGAACGCTGGAAAAACCTGA
GCGGGTGCCGGGGCAGCACCCGGCTCGGGTCAGCCACTGCCCCACACCGGGCCCACCAAGCCCCGCCCCTCGCGGCCACCGGGGCTTCCTTGCTCTTCTTATCATCTCCATCTTTAT
GATGAGGCTTGTTAACAAGACCAGAGAGCTGGCCAAGCACCTCTATCTCAGCCGCGCCCGCTCAGCCGAGCAGCGGTCGGTGGGGGGACTGGGAGGCGCTAATTAATTGATTCCTTT
CGACTCTAAAATATGCGGGCGTCTACACGGAACCCATGGACTCATAAACAATATATCTCTTGCGCGTGAGTGCACTCTCTCTCAAATAATTTTTCCATAGCCAAATCTCAGAGGGTTC
TGGATTTTTAGTTGCTAAGGAAAGATCCAAATGGGACCAATTTTAGCAGCCCAAACAGATCCCGTTCAGTGTCAGAAAATGCTTCCCCAAAGGGGTTGGGAGTGTGTTTTGTTGGAA
AAAAGCTTGGGTTATAGGAAAGCCTTTCCCTGCTACTTGTGTAGACCCAGCCCAATTTAAGAATTACAAGGAAGCGAAGGGGTTGTGTAGGCCGGAAGCCTCTCTGTCCCGGCTGGAT
GCAGGGGACTTGAGCTGCTCCGGAATTTGAGAGGAACATAGAAGCAAAGGTCCAGCCTTTGCTTCGTGCTGATTCCTAGACTTAAGATTCAAAAACAAATTTTTAAAAGTGAAACCAG
CCCTAGCCTTTGGAAGCTCTTGAAGGTTCAGCACCCACCCAGGAATCCACCTGCCTGTTACACCCCTCTCCAAGACACAGTGGCACCGCTTTTCTAACTGGCAGCACAGAGCAACTCT

Structure:

Protein coding potential

CPC coding potential score: -1.19011 (noncoding) ?
HMMER Pfam domains in 3' to 5' reading frames: 0 ?
HMMER Pfam domains in 5' to 3' reading frames: 0

PRIDE database search

Number of hits in the PRIDE database: 0 ?

Secondary structure information

RNAfold image: download
Randfold minimum free energy: -825.83
Randfold P-value: 0.001

Targetting miRNAs

MirTarget2 predictions:

MicroRNA	MirTarget2 score ?
hsa-miR-3688-3p	93.51
hsa-miR-1251	87.25
hsa-miR-202-5p	82.56
hsa-miR-26b-3p	81.72
hsa-miR-892a	80.28

Available literature

- Guil et al., 2012
- Niinuma et al., 2012
- Kogo et al., 2011
- Schorderet et al., 2011
- Geng et al., 2011
- Kaneko et al., 2010
- Tsai et al., 2010
- Gupta et al., 2010

Figure 3. The transcript page in the web interface provides a clear overview of information available on a specific lncRNA transcript.

custom-built protein sequence database that includes UniprotKB/Swiss-Prot and LNCipedia translations (Supplementary Figure S1). Within these PSMs, there were just 14 matches that could provide evidence for translation of LNCipedia entries. However, after close inspection of the FDR of the PSMs that passed our quality criteria, we noticed that although the PSMs from UniProtKB/Swiss-Prot have an expected FDR of 0.9%, the subset of PSMs from translated LNCipedia entries comes with an overwhelming FDR of 166% (Supplementary Figure S2). As such, there are only vague suggestions so far that any of these entries can effectively be translated.

As the PRIDE database is growing exponentially, and additional lncRNA transcript discovery is ongoing, searches for potentially coding lncRNAs need to be carried out anew at regular intervals to stay up-to-date with the growing amount of public data. We, therefore, envision running the full pipeline on all applicable PRIDE data at a set interval of 3 months; thus, periodically updating the knowledge on which lncRNAs might have coding potential. The output of each reprocessing effort will be used to annotate the LNCipedia, and past results will be kept available as well.

Besides this recurrent re-analysis of the relevant publicly available proteomics data, we also plan to extend the statistical approach used to evaluate the identification of a lncRNA by including information about the consistency with which such an identification is found across (unrelated) PRIDE experiments. Indeed, a relatively poor match in any individual experimental data set that, however, keeps returning across many such data sets, may well be a real indication that translation is taken place for that lncRNA.

LNCIPEDIA ACCES

LNCipedia is publicly available through a web interface at http://www.lncipedia.org. The interface allows users to query lncRNAs by name, chromosomal region or (partial) sequence. Several statistics are calculated that allow the user to evaluate different parameters regarding lncRNA secondary structure and regulation (Figure 3). The entire LNCipedia collection is available for download in the FASTA, GFF or BED format.

LncRNA researchers can contribute to LNCipedia by contacting the authors. In addition, registered users can modify existing records (updating aliases and adding PubMed literature records) directly using a web interface.

LNCRNA EXPRESSION ARRAY

The LNCipedia content can prove useful when designing large-scale screening experiments, such as lncRNA gene expression profiling. As a proof of concept, we have developed a custom lncRNA gene expression array using the Agilent Sureprint 60 k platform. In addition to roughly 33 000 probes for protein coding genes, we selected 23 042 probes for lncRNA transcripts in LNCipedia covering 97% of all LNCipedia transcripts with at least one probe (Agilent MicroArray Design ID: 039714). The performance of the expression array was evaluated using RNA sample titrations according to the MicroArray Quality Control standards (30). Adequate titration response of the lncRNA probes is shown in Supplementary Figure S3.

CONCLUSION AND FUTURE DIRECTION

Three important features are unique to LNCipedia: gene definitions and usage of a universal nomenclature for lncRNA transcripts, PRIDE analysis for detection of lncRNAs that may code for small peptides and miRNA seed predictions for lncRNA transcripts. These, along with the other tools available, are expected to make LNCipedia a powerful resource for human lncRNA research.

With the advances in RNA sequencing technology, more lncRNA genes are expected to get discovered. The authors will update LNCipedia when new sequences are reported in the literature or in other sources. In addition, new features will be developed to increase the interactive capabilities of LNCipedia. In this way, the lncRNA community will be able to upload and maintain records in the database. LNCipedia has the potential to become a community resource for lncRNA transcript information and annotation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–3 and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the equal contribution of J.V. and P.M.

FUNDING

Ghent University Multidisciplinary Research Partnership ‘Bioinformatics: from nucleotides to networks’ (to P.J.V., L.M., K.G., J.V.); National Institutes of Health [R01GM089784 to X.W.]; Flemish Fund for Scientific Research Flanders (FWO) (to P.M.); Ghent University Special Research Fund (BOF) (to J.V.). Funding for open access charge: Ghent University.

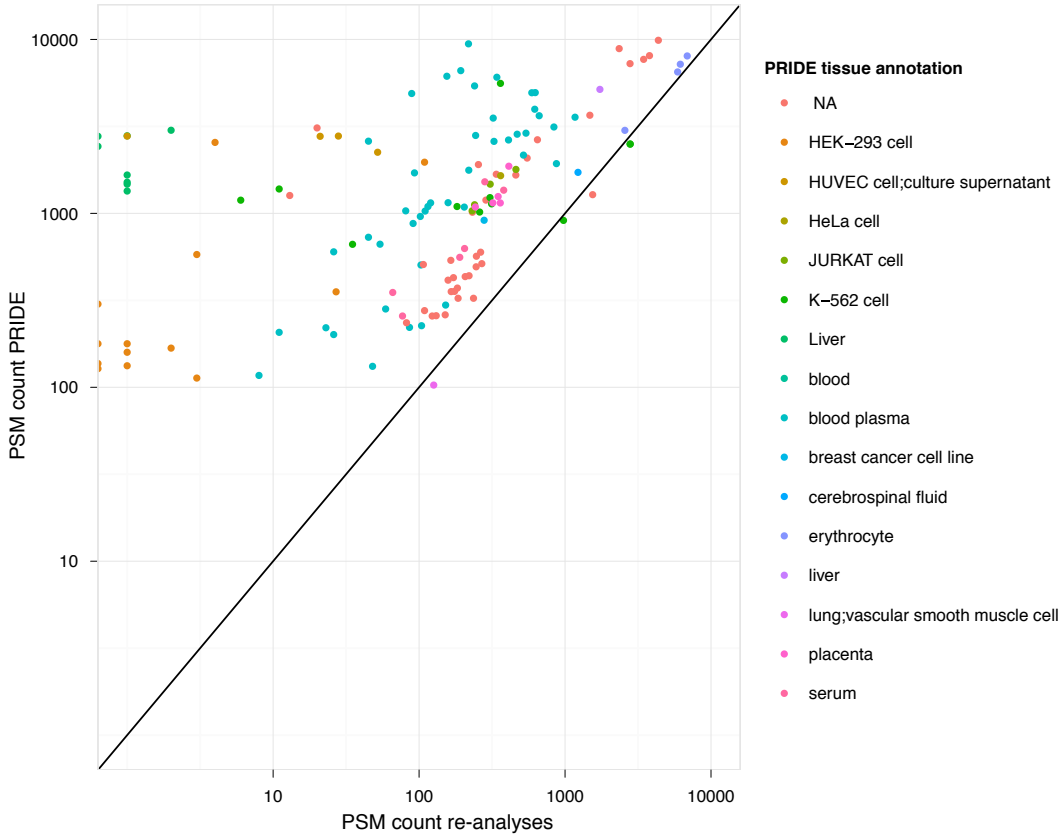
Conflict of interest statement. None declared.

REFERENCES

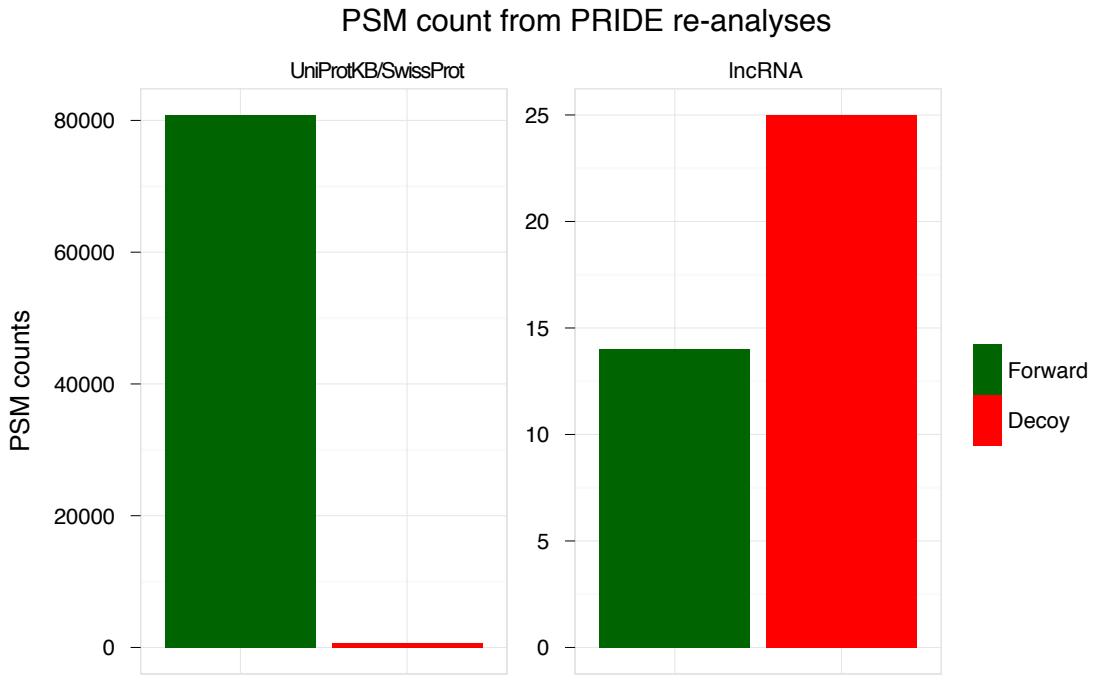
1. Mercer,T. and Dinger,M. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet*., **10**, 155–159.
2. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*., **25**, 1915–1927.
3. Taft,R.J. and Mattick,J.S. (2003) Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biol*., **5**, P1–P24.
4. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.

5. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
6. Brown,C.J., Ballabio,A., Rupert,J.L., Lafreniere,R.G., Grompe,M., Tonlorenzi,R. and Willard,H.F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38–44.
7. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
8. Panzitt,K., Tschernatsch,M.M., Guelly,C., Moustafa,T., Stradner,M., Strohmaier,H.M., Buck,C.R., Denk,H., Schroeder,R., Trauner,M. *et al.* (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology*, **132**, 330–342.
9. Wright,M.W. and Bruford,E.A. (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics*, **5**, 90–98.
10. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2010) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.
11. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2007) Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res.*, **35**, D162–D164.
12. Liu,C., Bai,B., Skogerbø,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.
13. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbo,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al.* (2011) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.
14. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2010) Rfam: wikipedia, clans and the 'decimal' release. *Nucleic Acids Res.*, **39**, D141–D145.
15. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
16. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
17. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. Chem. Mon.*, **125**, 167–188.
18. Bonnet,E., Wuyts,J., Rouzé,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
19. Cesana,M., Cacchiarelli,D., Legnini,I., Santini,T., Sthandier,O., Chinappi,M., Tramontano,A. and Bozzoni,I. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Gastroenterology*, **147**, 358–369.
20. Kretz,M., Webster,D.E., Flockhart,R.J., Lee,C.S., Zehnder,A., Lopez-Pajares,V., Qu,K., Zheng,G.X., Chow,J., Kim,G.E. *et al.* (2012) Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev.*, **26**, 338–343.
21. Wang,X. and El Naqa,I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.
22. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
23. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
24. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
25. Martens,L., Hermjakob,H., Jones,P., Adamski,M., Taylor,C., States,D., Gevaert,K., Vandekerckhove,J. and Apweiler,R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.
26. Sadygov,R.G., Cociorva,D. and Yates,J.R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods*, **1**, 195–202.
27. Vaudel,M., Burkhart,J.M., Sickmann,A., Martens,L. and Zahedi,R.P. (2011) Peptide identification quality control. *Proteomics*, **11**, 2105–2114.
28. Vaudel,M., Barsnes,H., Berven,F.S., Sickmann,A. and Martens,L. (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, **11**, 996–999.
29. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.
30. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.

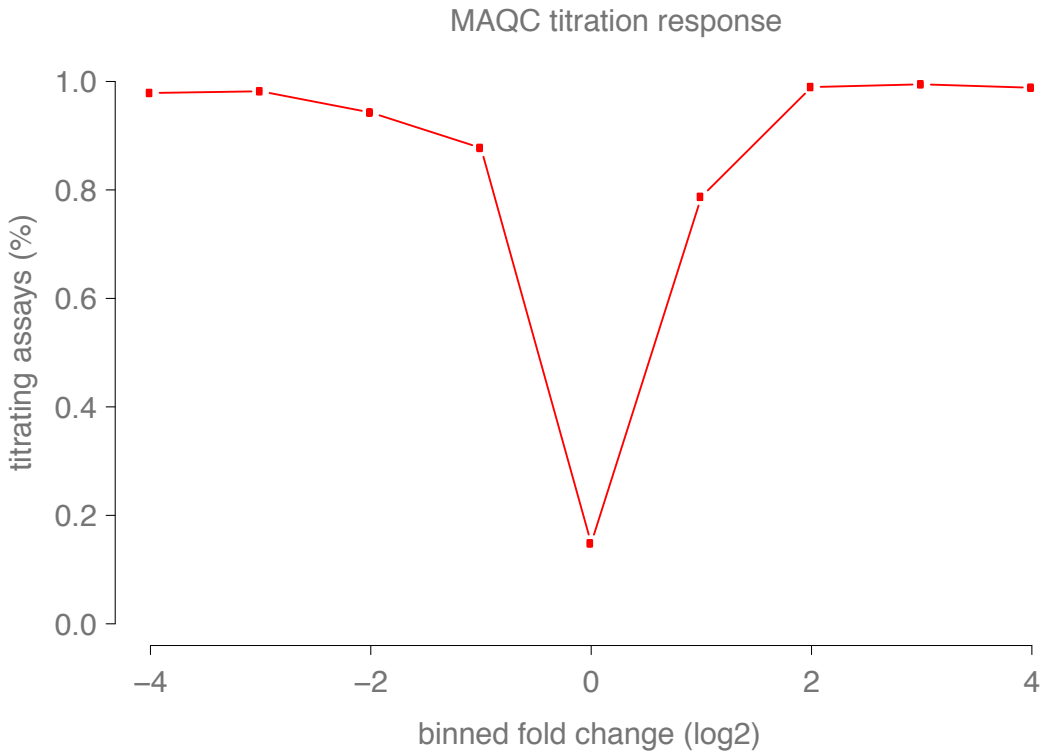
Supplemental Figures



Supplemental Figure 1. Scatterplot of re-analysis of 149 PRIDE experiments. The X-axis and the Y-axis show the number of peptide-to-spectrum matches (PSMs) *per* experiment by our automated re-analysis and as deposited in PRIDE, respectively. For most experiments, our analysis yields roughly half the amount of PSMs annotated in PRIDE. One reason for this is that our approach applies a stringent 1% FDR cutoff, while such stringency is not required when depositing an experiment into PRIDE. Furthermore, our protein sequence database is considerably larger than UniProtKB/Swiss-Prot since it contains a translated version of Incipedia. This inherently leads to larger e-values, and thus less PSMs in our stringent results set.



Supplemental Figure 2. PSM counts after re-analysis of 149 PRIDE experiments with an FDR limit of 1%. PSMs from UniProtKB/Swiss-Prot are called ‘false’ (left bar chart) and from lncRNA translations are dubbed ‘true’ (right bar chart). Decoy hits, indicative of the amount of false positives, are given in red, while normal hits are given in green. Note that while the left bar chart with UniprotKB/Swiss-Prot hits shows an expected FDR of 1%, the right bar chart with PSMs from lncipedia translations shows a much larger FDR of 166%.



Supplemental Figure 3. MAQC titration response of lncRNA probes. LncRNA expression was measured for samples A (Universal human reference RNA, Agilent Technologies), B (Human brain total RNA, Ambion), C (25% A + 75% B) and D (75% A + 25% B). The percentage of lncRNA probes that follow the monotonic titration response (Y-axis) is plotted in function of the binned log2-fold change (X-axis) between samples A and B. Titration response was calculated according to Shippy et al., Nature Biotechnology, 2006.

Supplemental Methods

Data is read from the PRIDE database after filtering applicable experiments by taxonomy, number of spectra and consistent taxonomic origin of the reported proteins. The data is then analyzed to detect applicable search engine settings, notably the precursor and fragment ion mass tolerances as well as the (variable) modifications to consider. Allowed missed cleavages are set to 1. PeptideShaker is run in automatic mode to filter the proposed peptide-to-spectrum matches hits at the 1% false discovery rate as calculated through the decoy database searching built-in to SearchGUI.

III.2. **RESEARCH PAPER 2**: AN UPDATE ON LNCIPEDIA: A DATABASE FOR ANNOTATED HUMAN LNCRNA SEQUENCES

Pieter-Jan Volders, Kenneth Verheggen, Gerben Menschaert, Klaas Vandepoele, Lennart Martens, Jo Vandesompele and Pieter Mestdagh

Nucleic acids research (2015)

http://nar.oxfordjournals.org/content/43/D1/D174

Contributions: Apart from the PRIDE reprocessing pipeline, the candidate contributed in whole or in part to design of the work, data acquisition, analysis and interpretation and drafting of the manuscript. For the PRIDE reprocessing pipeline and the matching paragraphs in the manuscript, the candidate contributed to the concept of the analysis, the interpretation of the data and the revision of the text.

D174–D180 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 05 November 2014*
doi: 10.1093/nar/gku1060

An update on LNCipedia: a database for annotated human lncRNA sequences

Pieter-Jan Volders[1], Kenneth Verheggen[2,3], Gerben Menschaert[4], Klaas Vandepoele[5,6], Lennart Martens[2,3], Jo Vandesompele[1] and Pieter Mestdagh[1,*]

[1]Center for Medical Genetics, Ghent University, Ghent 9000, Belgium, [2]Department of Medical Protein Research, VIB, Ghent 9000, Belgium, [3]Department of Biochemistry, Ghent University, Ghent 9000 Belgium, [4]Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent 9000, Belgium, [5]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent 9000, Belgium and [6]Department of Plant Systems Biology, VIB, Ghent 9000, Belgium

Received August 29, 2014; Revised October 13, 2014; Accepted October 15, 2014

ABSTRACT

The human genome is pervasively transcribed, producing thousands of non-coding RNA transcripts. The majority of these transcripts are long non-coding RNAs (lncRNAs) and novel lncRNA genes are being identified at rapid pace. To streamline these efforts, we created LNCipedia, an online repository of lncRNA transcripts and annotation. Here, we present LNCipedia 3.0 (http://www.lncipedia.org), the latest version of the publicly available human lncRNA database. Compared to the previous version of LNCipedia, the database grew over five times in size, gaining over 90 000 new lncRNA transcripts. Assessment of the protein-coding potential of LNCipedia entries is improved with state-of-the art methods that include large-scale reprocessing of publicly available proteomics data. As a result, a high-confidence set of lncRNA transcripts with low coding potential is defined and made available for download. In addition, a tool to assess lncRNA gene conservation between human, mouse and zebrafish has been implemented.

INTRODUCTION

Over the past decade long non-coding RNAs (lncRNAs) have emerged as a large class of functional non-coding RNAs (ncRNAs) (1). Defined as ncRNA transcripts longer than 200 nucleotides, lncRNAs have been shown to function mainly as transcriptional regulators by interaction with other biomolecules, such as proteins (2–4) and microRNAs (5). They are involved in a wide range of processes including cardiac development (6), dosage compensation (7,8) and cancer (2,9–10). Several specialist databases concerning lncRNA have been developed. Well-known examples are lncRNAdb, which focuses on lncRNAs with described functions (11), and NONCODE (12,13). In addition to these general lncRNA databases, databases that describe specific lncRNA subclasses have been compiled as well. LncRNAdisease contains lncRNAs with published disease associations (14) while lncRNAs targeted by microRNAs can be found in DIANA-LncBase (15).

Distinguishing coding from ncRNA sequences is an important step, both in the ncRNA and the protein research field. Classic approaches are based on either open reading frame (ORF) length, ORF conservation or structural protein domains (16). Recent computational methods make use of more complex features or machine learning approaches. Notable examples are the Coding-Potential Calculator (CPC), Coding-Potential Assessment Tool (CPAT) and PhyloCSF. CPC utilizes a support vector machine trained on features that describe long, high-quality ORFs with sequence similarity (BLASTX) to known proteins (17). CPAT is a logistic regression model that only uses sequence-derived features, such as ORF size, codon and hexamer usage bias (18). In contrast to CPC and CPAT, PhyloCSF employs codon substitution frequencies in whole-genome multi-species alignments and maximum likelihood trees to distinguish between coding and non-coding loci (19).

ORF length is either directly or indirectly used in all these computational prediction methods yet ORFs yielding short peptides (<100 amino acids) are difficult to predict. The discovery of functional peptides shorter than 100 amino acids, like the Drosophila gene tarsal-less (tal), thus raised the possibility that several lncRNAs are actually misclassified protein-coding genes encoding micropeptides (20,21). As small ORFs can also occur by chance in long transcripts, many well-described lncRNAs harbor non-functional ORFs (22). In addition to small ORFs, the *in silico* prediction of coding ORFs is further complicated by the existence of non-canonical (non-AUG) start codons (23).

*To whom correspondence should be addressed. Tel: +32 9 3326979; Fax: +32 9 3326549; Email: Pieter.Mestdagh@UGent.be

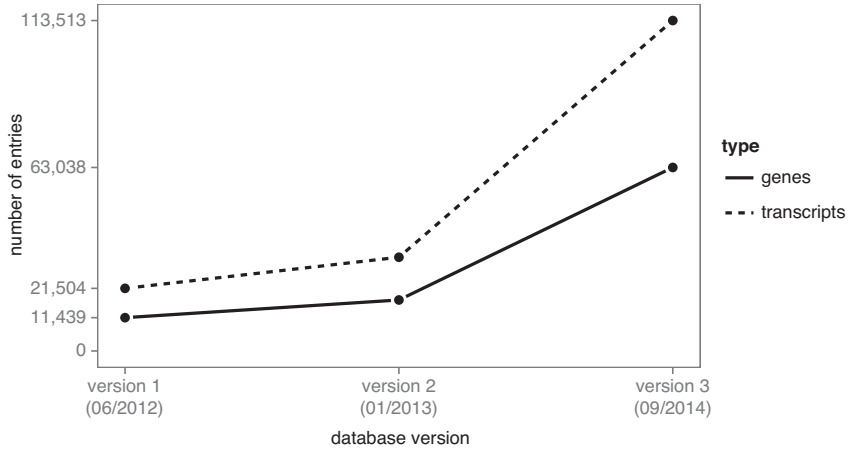


Figure 1. LNCipedia has grown substantially since itsfi rst release. Thefi rst version (41) was based on sequences and annotation from three different sources and was made available to the public in 2012. For the 2013 release of LNCipedia (unpublished), no additional sources were used, but the different sources were updated to the most recent version. For version 3.0 of LNCipedia, both new sources were added and existing sources were updated.

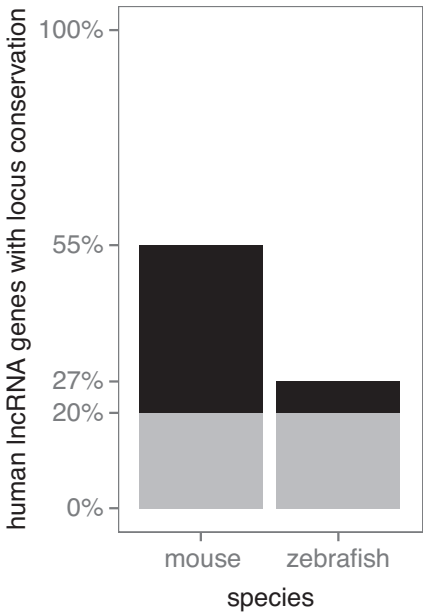


Figure 2. Many lncRNA loci are conserved in mouse or zebrafish. Locus conservation is a novel tool to determine the orthologous locus of a human lncRNA in another species. When the order of thefl anking protein-coding genes is conserved in another species, the lncRNA locus is considered conserved. The majority of the conserved loci in zebrafish are also conserved in mouse, this fraction is depicted in gray.

Experimental procedures to detect translated ORFs and their products have been developed as well. One such method is referred to as ribosome profiling and is based on deep sequencing of ribosome-protected mRNA fragments. Although many ncRNAs show ribosome occupancy, by using initiation-specific translation inhibitors in combination with ribosome profiling, researchers were able to map translation initiation sites (TIS) with base pair resolution and improve the detection of true ORFs (23,24). Other researchers were able to use the periodicity of ribosome movement on the mRNA to define actively translated ORFs (25). In addition to ribosome profiling, mass spectrometry has been applied in the search for novel peptides arising from lncRNAs (26,27). Several authors report small numbers of (micro) peptides arising from lncRNAs using either ribosome profiling or mass spectrometry. The debate on the putative function and total number of these peptides is still ongoing (26–28).

Here, we report on LNCipedia 3.0, the latest version of our publically available lncRNA database. In version 3.0, our major improvement is the evaluation of protein-coding potential with state-of-the-art algorithms and data sets. As such we have generated a high-confidence data set that excludes lncRNAs with possible protein-coding potential. In addition, a new tool to assess the conservation of lncRNA genes has been implemented. The database content has been updated and now contains over five times the number of transcripts compared to thefi rst version.

MATERIALS AND METHODS

Locus conservation

The upstream and downstream protein-coding genes that flank a human lncRNA gene are queried in the public Ensembl (29) MySQL database (version 73). For both genes, the orthologs in mouse and zebrafish are obtained using the Ensembl Compara API (version 73). If any pair of orthologs are neighboring genes, the locus is reported as conserved.

PhyloCSF

Whole-genome alignments of 46 species are obtained from the UCSC website (30) and processed using the PHAST (31) package (version 1.3) to obtain the required input format for PhyloCSF (19). To validate our workflow, we benchmarked PhyloCSF with transcripts annotated in Ensembl

(version 75). Transcripts with biotype 'lincRNA' or 'antisense' (20 320 transcripts) serve as negative set while transcripts with biotype 'protein_coding' and an annotated coding sequence (36 959 transcripts) serve as positive set.

TIS

Ribosome profiling sequencing data of HEK-293 cells treated with cycloheximide (CHX) and lactimidomycin (LTM) were processed (24). Two technical replicates of both treatments were pooled (Bioproject http://www.ncbi.nlm.nih.gov/bioproject/PRJNA171327: runs SRR618770 and SRR618771 for CHX and runs SRR618772 and SRR618773 for LTM).

The reads werefi rst clipped to remove their 3′ cloning adaptor sequence using the FASTX-Toolkit (fastx_clipper tool). Unclipped and clipped reads shorter than 25 nt were discarded. The remaining reads were mapped using the RNA-seq STAR aligner (32), sequentially using indices based on the following sequences: (i) Phix genome (widely used as a quality control for Illumina sequencing runs), (ii) *Homo sapiens* rRNA (Refseq IDs NR_003285.2, NR_003286.1, NR_003287.1, NR_023363.1) and (iii) the human reference genome (downloaded from the igenomes repository http://support.illumina.com/sequencing/sequencing_software/igenome.ilmn, using the *H. sapiens* genome build GRCh37 and Ensembl annotation version 70). The human STAR index was built taking into account the splice site annotation from Ensembl. Only uniquely mapped reads that are between 28 and 35 nt long were retained. Footprint alignments were assigned to a specific P-site nucleotide based on the fragment length (the 5′ offset is set to respectively 12, 13 or 14 for profiles with length $\leq$ 30 nt, 31–33 nt, or $\geq$ 34 nt (23)).

PRoteomics IDEntifications (PRIDE) reprocessing

The processing pipeline consists of three major modules. Thefi rst module is based on the PRIDE automated spectrum annotation pipeline (pride-asap) (33), and is used to reverse engineer the original search parameters from submitted data. The key parameters extracted by pride-asap in this stage are the allowable mass errors, the post-translational modifications (PTMs) to consider, and the enzyme used. Recent developments in this module have greatly improved the PTM inference by considering the modifications found in the PSI-mod (34) and Unimod (35) databases, as well as the frequency of occurrence of these modifications. Two thresholds are calculated based on this information, with thefi rst one serving as a lower threshold to exclude very low abundance modifications while the second threshold is used to determine whether a sufficiently abundant modification is to be considered as either variable orfi xed. A second development has been the impromptu determination of the protease used in the original experiment. Instead of assuming the use of trypsin, the pride-asap module now calculates the most likely enzyme based on all reported peptide sequences reported in PRIDE for that experiment. Overall, these updates to the module allow a reduction in search space to consider, providing faster processing times and leaving less room for false-positive matches.

The second module handles the peptide-to-spectrum matching, relying on SearchGUI (36) to automatically run multiple search engines in parallel; in this case OMSSA (37) and X!Tandem (38). SearchGUI is configured to use the target/decoy approach (39), where both the original (target) sequence database is searched, but also a reversed (decoy) version of that database. Matches from the latter can then be used to determine a false discovery rate (FDR) (39).

The third andfi nal module uses PeptideShaker (http://peptide-shaker.googlecode.com) and the compomics-utilities library (40) to collect, process and analyze the results generated by SearchGUI.

RESULTS

LNCipedia 3.0 content

LNCipedia 1.0 (41) combined sequences and annotation from three different public resources, namely, Ensembl (29,42), Human body map lincRNAs (43) and the lncRNA database (11). In LNCipedia version 3.0, we have complemented these resources with four additional public data sets (Table 1). Two of these data sets are obtained from databases (44,45), and two from lncRNA research articles describing RNA sequencing workflows and reporting on novel lncRNAs (46,47). As with LNCipedia 1.0, redundant transcripts are merged into the same record. The result of this extension and integration of sources is that LNCipedia 3.0 represents a more than 5-fold increase in transcript content over version 1.0 (Figure 1). The majority of these transcripts (80%) is found in new loci and as such give rise to novel lncRNA genes.

In LNCipedia 1.0 we introduced a universal lncRNA nomenclature to overcome the confusion caused by the use of different identifiers by different authors and databases. As was suggested by others, we named lncRNAs after neighboring protein-coding genes on the same strand (48). In LNCipedia 3.0, we hold true to this strategy. Existing genes are expanded when novel transcripts have overlapping exons and new genes are created when a transcript does not share exonic sequence with any existing gene.

Locus conservation

The identification of orthologous lncRNAs is an important step for animal modeling and functional research across species. Conservation of gene order is a straightforward metric often used in comparative genomics. We applied the concept of gene order conservation to determine the orthologous locus of a lncRNA in another species. Using the Ensembl Compara API, we have assessed the conservation in the order of thefl anking protein-coding genes. Currently, orthologs for non-coding genes are not as well annotated as for protein-coding genes,fl anking non-coding genes were therefore not taken into account. When the order is conserved in mouse or zebrafish we report the locus as conserved. In this way, wefi nd locus conservation for 55% of the human lncRNA genes in mouse, and for 27% in zebrafish (Figure 2). The majority of the conserved loci in zebrafish are also conserved in mouse, as one would expect. While locus conservation is no proof for the functional con-

Table 1. Overview of data sources contributing to lncRNA content in LNCipedia 3.0

Source	Version	Number of transcripts
Ensembl (42)	75	23 498
Refseq (44)	March 2014	6917
Nielsen *et al.* (46)		7656
Hangauer *et al.* (47)		5339
NONCODE (45)	4	93 164
LNCipedia (41)	1.0	21 504
Total number of unique transcripts		113 513

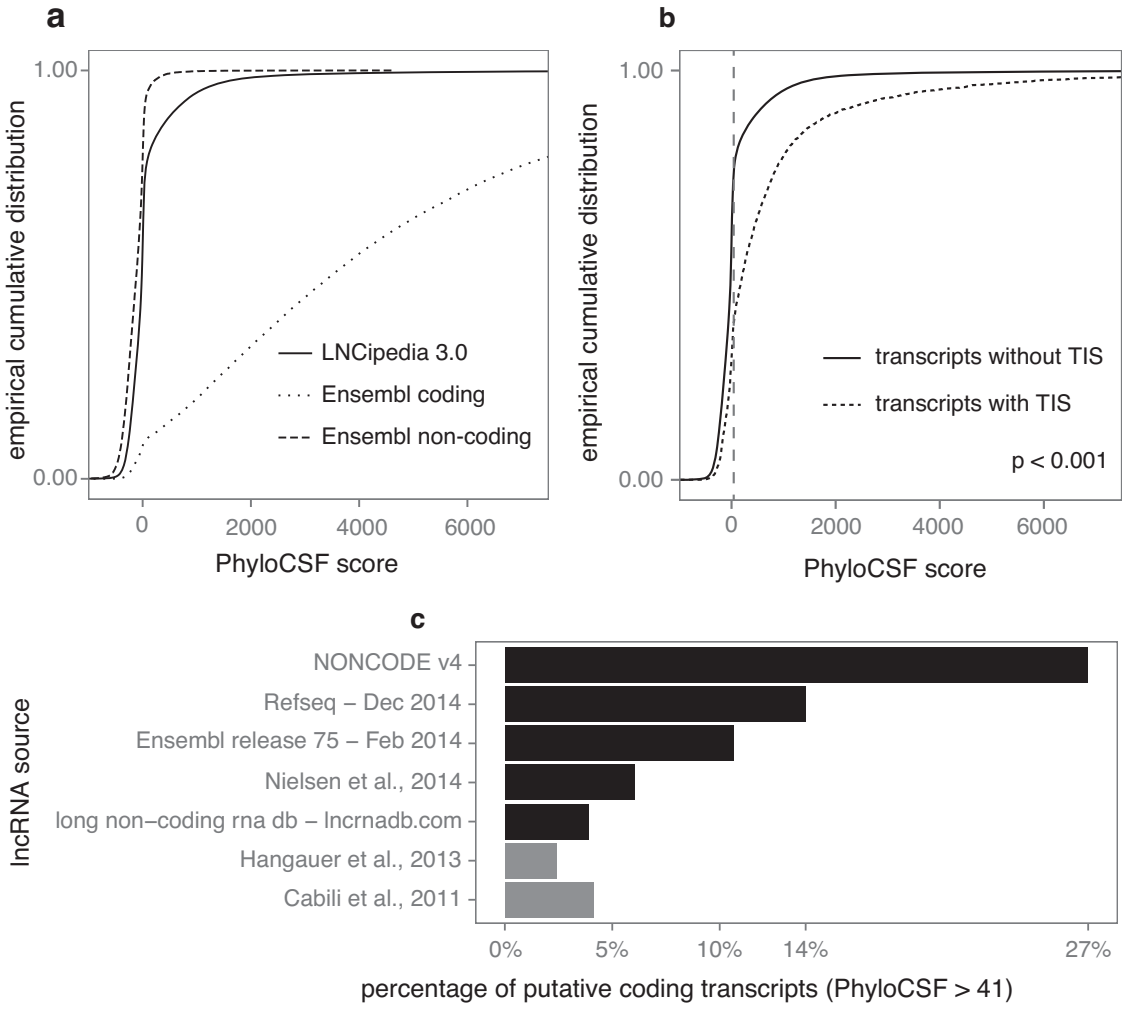


Figure 3. Different methods suggest contamination of coding sequences in lncRNA data sets. (a) PhyloCSF benchmarking and score distributions. We can observe a considerable difference between the score distributions of coding and non-coding transcripts in the Ensembl data set. In addition, while the great majority of LNCipedia is presumably non-coding, it also contains a fraction of transcripts with a PhyloCSF score in the coding range. (b) Transcripts with a TIS have a significantly higher PhyloCSF score (Mann–Whitney U test) compared to other transcripts. (c) Several public lncRNA resources suffer from considerable contamination with protein-coding sequences. The percentage of transcripts with PhyloCSF score greater than 41 is shown for the different sources in LNCipedia 3.0. Two sources alreadyfi ltered with PhyloCSF are depicted in gray. In the case of RefSeq, only entries with property "biomol_ncrna_lncrna" were considered.

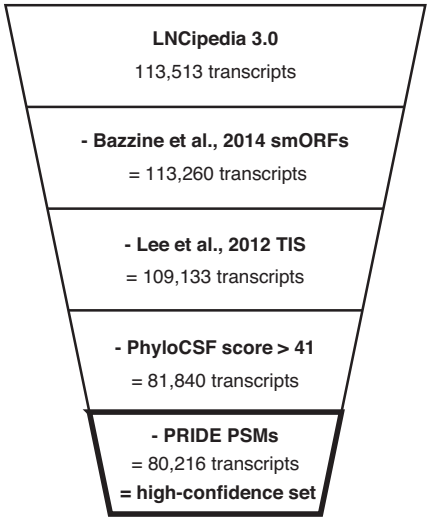


Figure 4. Transcripts with a likely coding potential are removed in the definition of a high-confidence set. Transcripts containing small ORFs (25), TIS (24), PhyloCSF score greater than 41 or PSMs with an identification confidence higher than 90% are excluded.

servation of the lncRNA itself, it may serve afi rst step in finding the orthologous lncRNA.

Protein-coding potential

For collection of lncRNA transcript sequences, we rely on public data sets that are often contaminated with small numbers of transcripts harboring coding ORFs (25,26). While we already presented several measures to assess this problem (41), we further expanded these with state-of-the-art tools and included additional lncRNA transcript data sets. One such measure is the PhyloCSF (19) score. We have benchmarked PhyloCSF using Ensembl transcripts and we have determined 41 as an optimal threshold for the PhyloCSF score resulting in a precision of 95% and sensitivity of 91% (Supplemental Material and Figures). From the empirical cumulative distribution (Figure 3a) it is apparent that LNCipedia most likely contains a considerable fraction of protein-coding sequences. When applying our precomputed cutoff, these transcripts add up to about 26% of the collection. Figure 3c shows the distribution of these putative coding transcripts among the different sources used for LNCipedia. It is clear that some lncRNA data sets suffer more from contamination of coding sequences than others. Strikingly, nearly 50% of Refseq annotated non-coding sequences are predicted to be coding according to the PhyloCSF score cutoff. It is no surprise that the lowest number of coding sequences is observed in Cabili *et al.* and Hangauer *et al.* as these studies applied PhyloCSF as afilter in their workflow.

Another measure to assess protein-coding potential is the use of ribosome profiling to map TIS. When we map the TIS observed in HEK-293 (24) to LNCipedia entries, we find 4154 trancripts with at least one TIS. Of note, these transcripts have significantly higher PhyloCSF scores (Figure 3b), which is a good validation of both methods.

PRIDE

Similar to the rapid growth of LNCipedia, the submission of mass spectrometry data to the PRIDE repository has flourished as well (49). While these increased collections of lncRNAs and mass spectrometry data provide even more means to detect potentially coding lncRNAs, they also require much more compute power to process. The only way to analyze these data in a timely fashion is to make use of parallelization on a compute cluster or through grid computing (50). We have therefore set up such a grid environment based on dedicated hardware running a collection of Linux virtual machines, allowing us to re-analyze the full human complement of PRIDE in under a week.

At the time of writing, the pipeline has been run on 2493 PRIDE experiments, containing 39 463 035 fragmentation mass spectra and covering all 68 annotated human tissues in the public repository. This resulted in a total of 8 064 657 peptide-to-spectrum matches (PSMs), of which 747 305 were matched to lncRNAs in LNCipedia (393 859 matched the target database and 353 446 matched the decoy database). Of these PSMs, 18 929 target sequences (representing 2040 transcripts, from 1770 genes) had an identification confidence higher than 90% (in contrast to only 2001 decoy sequences that had such a high confidence). Of note, the estimation of the FDR remains a complex issue in these very broad searches (51,52), and care should be taken to interpret these results. Indeed, as supplementary Figures S1 and S2 illustrate, while the confidence compares reasonably well with the estimated FDR, especially at higher confidences (higher than 90%), the evolution of the FDR toward the higher confidences is very different between the UniProtKB-SwissProt-derived identifications and the lncRNA matches.

No significantly higher PhyloCSF score was found for transcripts containing PSMs with identification confidence higher than 90%. In addition, no significant overlap is observed between the set of transcripts identified in PRIDE and the sets containing TIS and smORFs. This observation illustrates the very unique nature of the PRIDE analysis and strongly suggests its ability to detect coding potential not predicted by other methods.

HIGH-CONFIDENCE SET

Since LNCipedia contains a non-negligible number of putative coding transcripts, we propose afi ltering strategy to create a stringent or high-confidence data set. Four groups of putative coding transcripts are removed (Figure 4, Supplementary Figure S3). The first group consists of 253 lncRNAs containing small ORFs (smORFs) (25). Bazzini *et al.* developed an approach to detect smORFs using ribosome profiling whereby the periodicity of ribosome movement on actively translated ORFs is used to distinguish coding from non-coding sequences. A second approach to apply ribosome profiling in the quest for novel coding RNAs has been described by Lee *et al.* (24). Using LTM, a ribosome inhibitor specific to initiating ribosomes, TIS were mapped in HEK-293 cells. Note that 4127 lncRNA transcripts containing at least one TIS are thus withdrawn. While these transcripts have a good change to give rise to peptides, it is important to consider that a negative result

does not guarantee the opposite. The transcript may not be expressed or translated in the sample. The nextfi ltering step is based on PhyloCSF (19). As discussed earlier, this algorithm can distinguish between coding and non-coding sequences with high accuracy. As such, 27 293 transcripts with a PhyloCSF score higher than 41 are discarded. Finally, the 2040 PSM containing transcripts from the PRIDE reprocessing pipeline are excluded as well. The resulting set of 80 216 transcripts (71% of LNCipedia 3.0) representing 48 028 genes (76%) is referred to as 'high-confidence set' and is available for download on the LNCipedia website.

CONCLUSION AND FUTURE DIRECTION

With over 90 000 new transcripts, LNCipedia content increased 5-fold since its first publication in 2012. This makes it to our knowledge the largest publicly available human lncRNA resource. Furthermore, we improved the evaluation of coding potential with state-of-the-art algorithms, published data sets and an improved PRIDE reprocessing pipeline. In addition, we have developed a locus conservation analysis tool, which can aid in the search for lncRNA orthologs or prioritarization of lncRNAs for animal studies.

As in the previous years, LNCipedia will be updated when new lncRNA data sets are available. With the arrival of a new human reference genome (GRCh38), an important improvement to the database will be remapping chromosomal positions to this new reference genome. We will also continue to automatically run searches against the ever-growing contents of the PRIDE database on a routine basis. Furthermore, we will improve the specificity of the PRIDE searches by taking possible contamination from viral sequences into account.

In conclusion, LNCipedia 3.0 provides significant improvements over the previous version in terms of data content and data annotation.

AVAILABILITY

LNCipedia 3.0 can be accessed trough a web interface at www.lncipedia.org. Exports are available in FASTA, GFF, GTF or BED format for both the entire lncRNA collection and the high-confidence set. In addition, Integrative Genome Viewer (IGV) users have the option of loading an IGV optimized data set directly in the application. As in version 1.0, the database can be queried by chromosomal position or (partial) sequence. We encourage the lncRNA research community to contribute to LNCipedia by submitting newly discovered lncRNAs and by adding PubMed literature records to existing entries using the web interface.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to acknowledge Jasper Anckaert, Stephanie Letellier and Justine Nuytens for their contribution in the development of this LNCipedia version.

FUNDING

Multidisciplinary Research Partnership 'Bioinformatics: From Nucleotides to Networks' Project of Ghent University [01MR0310W to P.V. and K. Vandepoele]; Fund for Scientific Research Flanders [FWO; to P.M. and G.M.]; The European Union 7th Framework Program 'PRIME-XS' [262067 to L.M.]; Ghent University [to K. Verheggen]. Funding for open access charge: Multidisciplinary Research Partnership 'Bioinformatics: From Nucleotides to Networks' Project of Ghent University [01MR0310W].
Conflict of interest statement. None declared.

REFERENCES

1. Mercer,T. and Dinger,M. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet*, **10**, 155–159.
2. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.-C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
3. Margueron,R. and Reinberg,D. (2011) The Polycomb complex PRC2 and its mark in life. *Nature*, **469**, 343–349.
4. Tsai,M.-C., Manor,O., Wan,Y., Mosammaparast,N., Wang,J.K., Lan,F., Shi,Y., Segal,E. and Chang,H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.
5. Cesana,M., Cacchiarelli,D., Legnini,I., Santini,T., Sthandier,O., Chinappi,M., Tramontano,A. and Bozzoni,I. (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Gastroenterology*, **147**, 358–369.
6. Klattenhoff,C.A., Scheuermann,J.C., Surface,L.E., Bradley,R.K., Fields,P.A., Steinhauser,M.L., Ding,H., Butty,V.L., Torrey,L., Haas,S. *et al.* (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Gastroenterology*, **152**, 570–583.
7. Penny,G.D., Kay,G.F., Sheardown,S.A., Rastan,S. and Brockdorff,N. (1996) Requirement for Xist in X chromosome inactivation. *Nature*, **379**, 131–137.
8. Lee,J.T., Davidow,L.S. and Warshawsky,D. (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.*, **21**, 400–404.
9. Panzitt,K., Tschernatsch,M.M.O., Guelly,C., Moustafa,T., Stradner,M., Strohmaier,H.M., Buck,C.R., Denk,H., Schroeder,R., Trauner,M. *et al.* (2007) Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology*, **132**, 330–342.
10. Gibb,E.A., Brown,C.J. and Lam,W.L. (2011) The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer*, **10**, 38.
11. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2010) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.
12. Liu,C., Bai,B., Skogerbø,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.
13. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbo,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al.* (2011) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.
14. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2012) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
15. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Reczko,M., Maragkakis,M., Dalamagas,T.M. and Hatzigeorgiou,A.G. (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.*, **41**, D239–D245.
16. Dinger,M.E., Pang,K.C., Mercer,T.R. and Mattick,J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.
17. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of

transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
18. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.-P. and Li,W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
19. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, I275–I282.
20. Galindo,M.I., Pueyo,J.I., Fouix,S., Bishop,S.A. and Couso,J.P. (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.*, **5**, e106.
21. Crappé,J., Van Criekinge,W. and Menschaert,G. (2014) Little things make big things happen: a summary of micropeptide encoding genes. *EuPA Open Proteom.*, **3**, 128–137.
22. Dinger,M.E., Gascoigne,D.K. and Mattick,J.S. (2011) The evolution of RNAs with multiple functions. *Biochimie*, **93**, 2013–2018.
23. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.
24. Lee,S., Liu,B., Lee,S., Huang,S.-X., Shen,B. and Qian,S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424–E2432.
25. Bazzini,A.A., Johnstone,T.G., Christiano,R., Mackowiak,S.D., Obermayer,B., Fleming,E.S., Vejnar,C.E., Lee,M.T., Rajewsky,N., Walther,T.C. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.
26. Gascoigne,D.K., Cheetham,S.W., Cattenoz,P.B., Clark,M.B., Amaral,P.P., Taft,R.J., Wilhelm,D., Dinger,M.E. and Mattick,J.S. (2012) Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics*, **28**, 3042–3050.
27. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2012) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
28. Chew,G.-L., Pauli,A., Rinn,J.L., Regev,A., Schier,A.F. and Valen,E. (2013) Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development*, **140**, 2828–2834.
29. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
30. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
31. Hubisz,M.J., Pollard,K.S. and Siepel,A. (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformat.*, **12**, 41–51.
32. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
33. Hulstaert,N., Reisinger,F., Rameseder,J., Barsnes,H., Vizcaíno,J.A. and Martens,L. (2013) Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J. Proteom.*, **95**, 89–92.
34. Montecchi-Palazzi,L., Beavis,R., Binz,P.-A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
35. Creasy,D.M. and Cottrell,J.S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.
36. Vaudel,M., Barsnes,H., Berven,F.S., Sickmann,A. and Martens,L. (2011) SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, **11**, 996–999.
37. Geer,L.Y., Markey,S.P., Kowalak,J.A., Wagner,L., Xu,M., Maynard,D.M., Yang,X., Shi,W. and Bryant,S.H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.
38. Fenyö,D. and Beavis,R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, **75**, 768–774.
39. Vaudel,M., Sickmann,A. and Martens,L. (2012) Current methods for global proteome identification. *Expert Rev Proteomics*, **9**, 519–532.
40. Barsnes,H., Vaudel,M., Colaert,N., Helsens,K., Sickmann,A., Berven,F.S. and Martens,L. (2011) Compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformat.*, **12**, 70.
41. Volders,P.-J., Helsens,K., Wang,X., Menten,B., Martens,L., Gevaert,K., Vandesompele,J. and Mestdagh,P. (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, **41**, D246–D251.
42. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
43. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
44. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
45. Xie,C., Yuan,J., Li,H., Li,M., Zhao,G., Bu,D., Zhu,W., Wu,W., Chen,R. and Zhao,Y. (2013) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98–D103.
46. Nielsen,M.M., Tehler,D., Vang,S., Sudzina,F., Hedegaard,J., Nordentoft,I., Orntoft,T.F., Lund,A.H. and Pedersen,J.S. (2014) Identification of expressed and conserved human noncoding RNAs. *RNA*, **20**, 236–251.
47. Hangauer,M.J., Vaughn,I.W. and McManus,M.T. (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.*, **9**, e1003569.
48. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
49. Vizcaíno,J.A., Côté,R.G., Csordas,A., Dianes,J.A., Fabregat,A., Foster,J.M., Griss,J., Alpi,E., Birim,M., Contell,J. *et al.* (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.*, **41**, D1063–D1069.
50. Verheggen,K., Barsnes,H. and Martens,L. (2014) Distributed computing and data storage in proteomics: many hands make light work, and a stronger memory. *Proteomics*, **14**, 367–377.
51. Vaudel,M., Burkhart,J.M., Sickmann,A., Martens,L. and Zahedi,R.P. (2011) Peptide identification quality control. *Proteomics*, **11**, 2105–2114.
52. Colaert,N., Degroeve,S., Helsens,K. and Martens,L. (2011) Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res.*, **10**, 5555–5561.

Supplemental methods: benchmarking PhyloCSF

Pieter-Jan Volders

12 augustus 2014

Data input

```
non_coding_cut <- read.table("non_coding_cut.txt", header=F)
coding_cut <- read.table("coding_cut.txt", header=F)
lncipedia_cut <- read.table("lncipedia_cut_replaced.txt", header=F,
sep = "\t")

non_coding_cut$group = 'non_coding'
coding_cut$group = 'coding'
lncipedia_cut$group = 'lncipedia'

non_coding_cut$source = 'ensembl'
coding_cut$source = 'ensembl'
lncipedia_cut$source = 'lncipedia'

all_phylocsf = rbind(non_coding_cut, coding_cut, lncipedia_cut)

all_ensembl = rbind(non_coding_cut, coding_cut)
colnames(all_ensembl)[1:2] = c("filename", "score")
```

Plotting

```
library(ggplot2)
ggtheme = theme(
  axis.text.x = element_text(colour='gray50'),#, angle = 90, hjust =
1, vjust = 0.5),
  axis.text.y = element_text(colour='gray50'),
  panel.background = element_blank(),
  panel.grid.minor = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_rect(colour='gray50', fill = NA),
  strip.background = element_blank()
)


ggplot(all_phylocsf, aes(V2, linetype = factor(group))) +
  stat_ecdf(n=5000, geom="line") +
  coord_cartesian(xlim=c(-1000, 7500)) +
  ggtheme
```

Use cutoff to increase sampling in the range of the plot

```
all_phylocsf[all_phylocsf$V2 > 5000, 'V2'] = 5000
all_phylocsf[all_phylocsf$V2 < -5000, 'V2'] = -5000
```

```
ggplot(subset(all_phylocsf, source=='ensembl'), aes(x=V2, fill=group
)) +
  geom_density(alpha=.3) +
  #stat_bin(aes(color=group),binwidth=10, geom="line", position='dod
ge') +
  coord_cartesian(xlim=c(-4000, 4000)) +
  #facet_wrap(~ source, ncol=1) +
  ggtitle("Ensembl PhyloCSF scores") +
  xlab("PhyloCSF score") +
  scale_fill_discrete(name="Ensembl biotype",
                      breaks=c("coding", "non_coding"),
                      labels=c("protein_coding", "lincrna"))+
  ggtheme
```

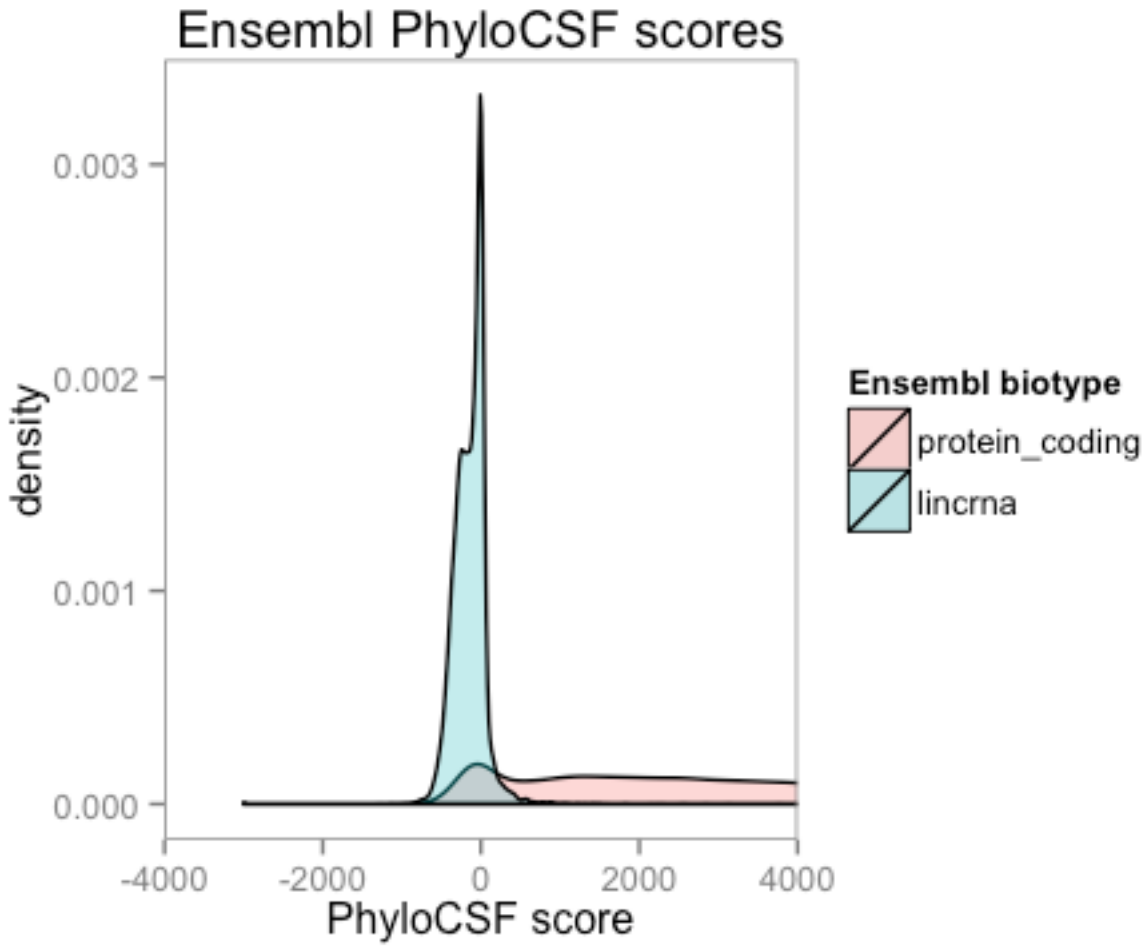


Figure 1: Density plot of the PhyloCSF scores for Ensembl transcripts. Only a small fraction of the protein-coding transcripts have a PhyloCSF score in the same range as lncRNA transcripts.

ROC analysis to determine optimal cutoff

```
all_ensembl$label = 0
all_ensembl[all_ensembl$group == 'coding', 'label'] = 1
library(ROCR)

pred = prediction(all_ensembl$score, all_ensembl$label)

perf = performance(pred,"tpr","fpr")
plot(perf, print.cutoffs.at = c(0, 40, 100, 500))
```

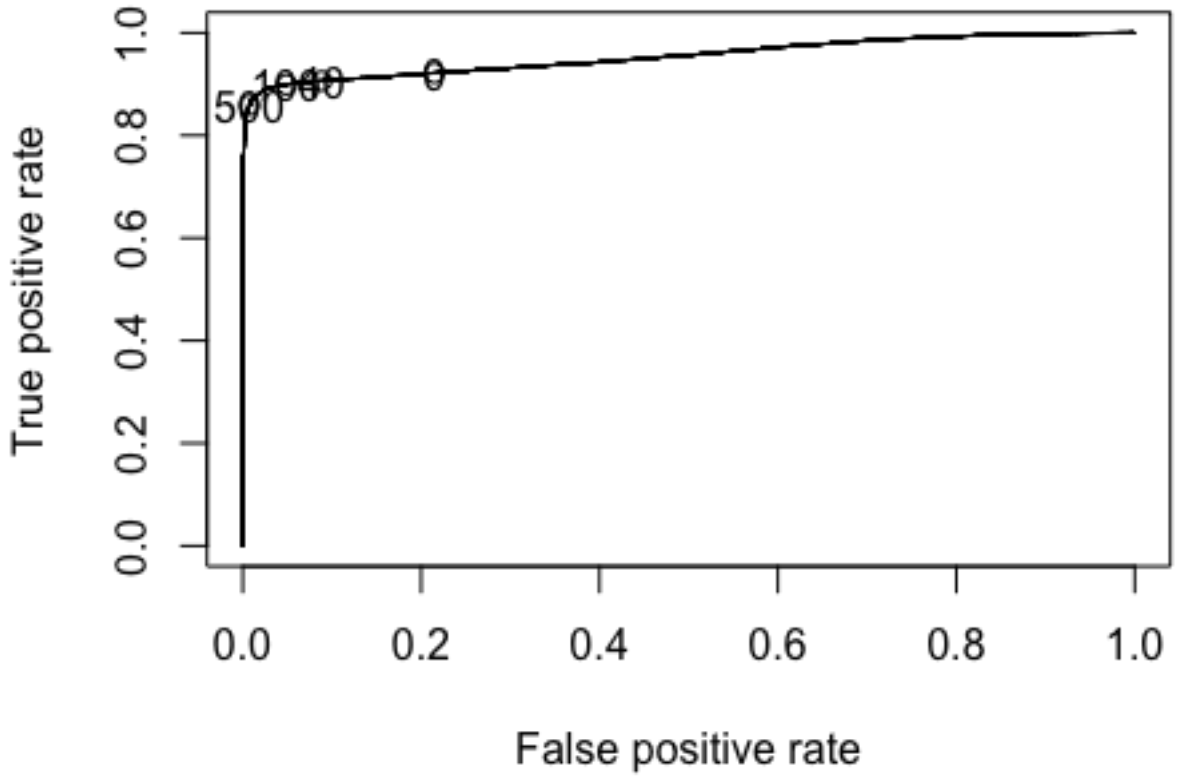


Figure 2: Receiver operating characteristic (ROC) curve of the PhyloCSF score cutoff

```
perf = performance(pred,"acc","cutoff")
plot(perf, xlim = c(0, 5000))
```

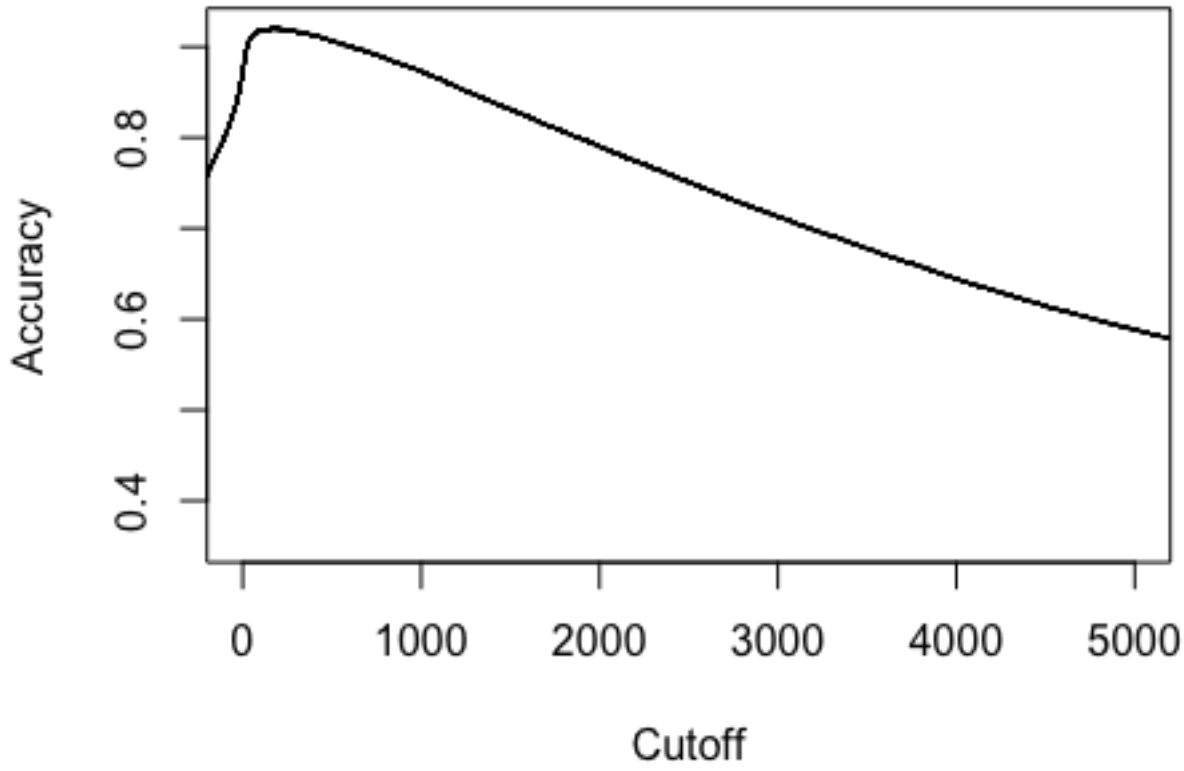


Figure 3: Accuracy plot. A cutoff off 180.7105 will yield the highest accuracy.

```
n = which.max(perf@y.values[[1]])
max_acc_cutoff = perf@x.values[[1]][n]

perf <- performance(pred, "prec", "rec", "cutoff")
plot(perf)
```

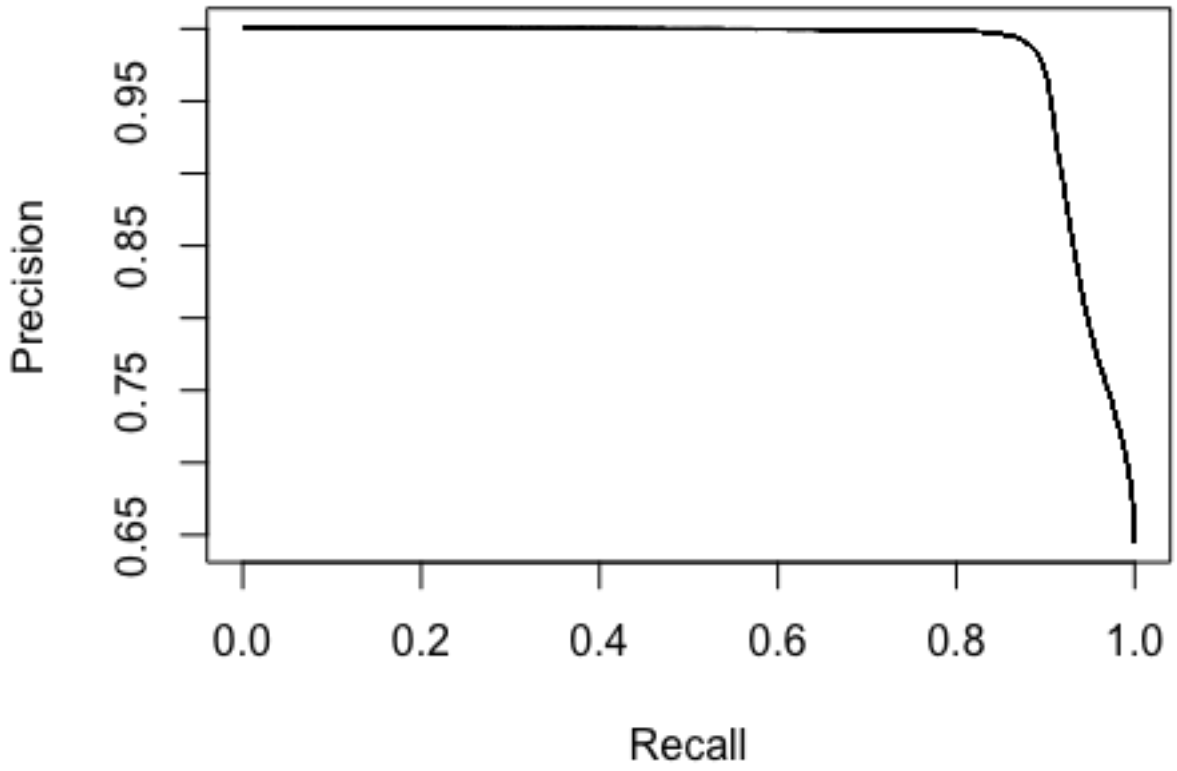


Figure 4: Precision/recall curve

Precision of 95%

```
library(dplyr)

precision_recall = data.frame(
  Precision = perf@y.values[[1]],
  Recall = perf@x.values[[1]],
  Cutoff = perf@alpha.values[[1]]
)
precision_recall_sorted = tbl_df(precision_recall) %>%
  filter(Precision > 0.95) %>%
  arrange(Precision)
```

A cuttoff of 41.2019 will result in a Precision of 95.0013% and a Recall (Sensitivity) of 90.6058%

```
perf = performance(pred,"sens", 'spec', 'cutoff')
n= length(perf@y.values[[1]])
sens_spec = data.frame(
  Cutoff = c(perf@alpha.values[[1]], perf@alpha.values[[1]]),
  Performance = c(perf@y.values[[1]], perf@x.values[[1]]),
  Measure = c(rep('Sensitivity',  n), rep('Specificity',  n))
```

```
)

n = which.min(abs(perf@y.values[[1]]-perf@x.values[[1]]))
sens_spec_cutoff = perf@alpha.values[[1]][n]

sens_spec_intersect = perf@x.values[[1]][n]

ggplot(sens_spec, aes(Cutoff, Performance, color=Measure)) +
  geom_line() +
  coord_cartesian(xlim = c(-1000, 5000)) +
  geom_vline(xintercept = sens_spec_cutoff)+
  geom_hline(yintercept = sens_spec_intersect) +
  ggtheme
```

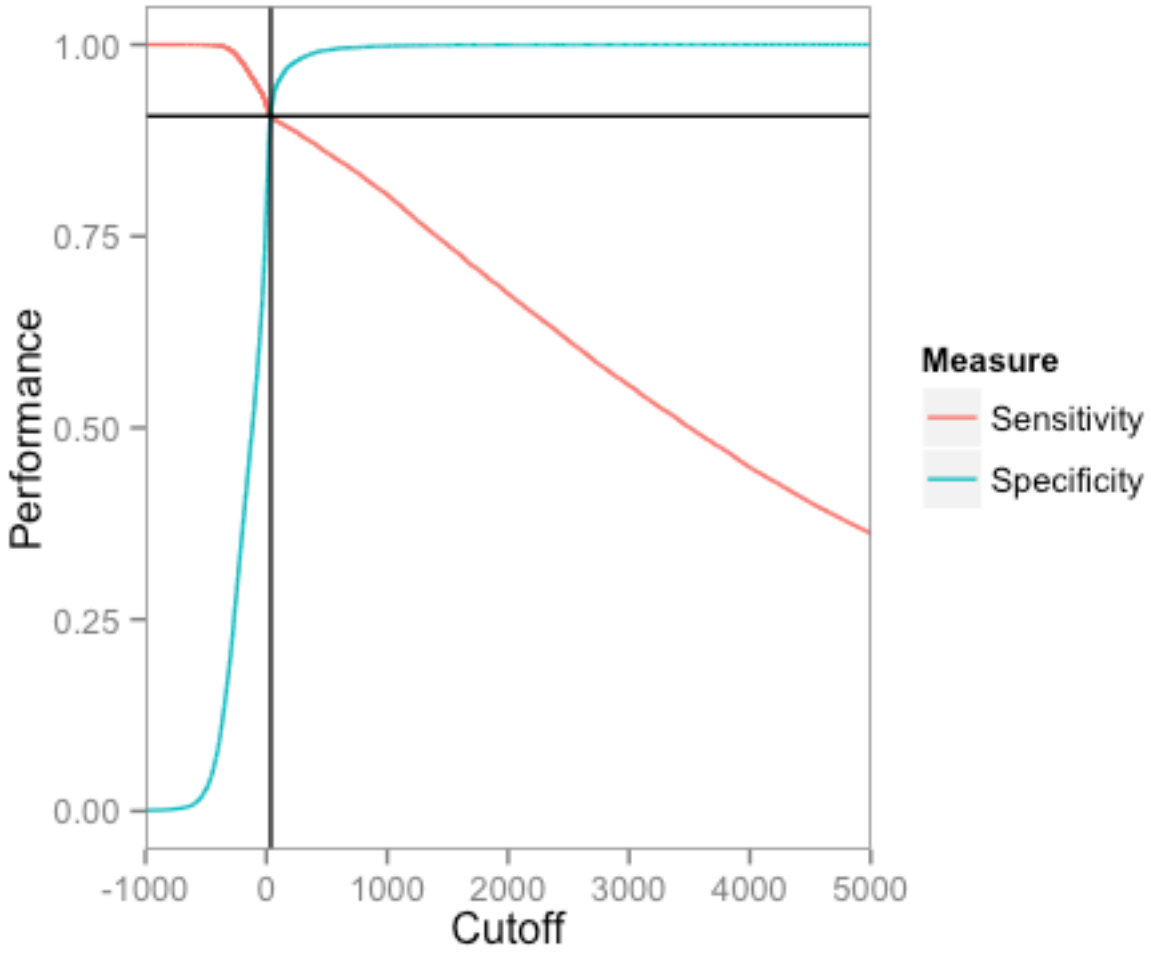


Figure 5: Sensitivity/specificity plot. The sensitivity and specificity plots intersect at cutoff 36.6852, resulting in a sensitivity and specificity of 0.9069

```
lncipedia_nr_coding = sum(lncipedia_cut$V2 > precision_recall_sorted
[1,'Cutoff'])
lncipedia_nr        = length(lncipedia_cut$V2)
```

Using a cutoff of 41.2019, we observe 29497 putative coding transcripts in LNCipedia. This corresponds to 26.2677% of the database.

III.3. **RESEARCH PAPER 3**: NON-CODING AFTER ALL: LARGE-SCALE PROTEOMICS REPROCESSING SUGGESTS LIMITED TRANSLATION OF LNCRNAS

Kenneth Verheggen, Pieter-Jan Volders*, Kris Gevaert, Pieter Mestdagh, Gerben Menschaert, Petra Van Damme, Jo Vandesompele, Lennart Martens*

**equally contributing authors*

Contributions: The candidate contributed in part to design of the work and performed data acquisition, analysis and interpretation of the data shown under "*LncRNA expression and composition show no indication of coding role*". Apart from the section "*The Influence of Protein Composition On Detectability by Mass Spectrometry*", the candidate drafted the manuscript.

NON-CODING AFTER ALL: LARGE-SCALE REPROCESSING OF PROTEOMICS DATA SUGGESTS LIMITED TRANSLATION OF LNCRNAS

KENNETH VERHEGGEN[*,1,2,5], *PIETER-JAN VOLDERS*[*,3,5], *PIETER MESTDAGH*[3,5], *GERBEN MENSCHAERT*[4], *PETRA VAN DAMME*[1,2], *KRIS GEVAERT*[1,2], *LENNART MARTENS*[1,2,5,#], *JO VANDESOMPELE*[3,5]

[1] Medical Biotechnology Center, VIB, Ghent 9000, Belgium

[2] Department of Biochemistry, Ghent University, Ghent 9000 Belgium

[3] Center for Medical Genetics, Ghent University, Ghent 9000, Belgium

[4] Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Ghent 9000, Belgium

[5] Bioinformatics Institute Ghent, Ghent University, Ghent 9000, Belgium

** equally contributing authors*

 # Corresponding author: Prof. Dr. Lennart Martens, A. Baertsoenkaai 3, B-9000 Gent, Belgium, lennart.martens@vib-ugent.be, tel: +32 9 264 93 58, fax: +32 9 264 94 84

ABSTRACT

Over the past decade, long non-coding RNAs (lncRNAs) have emerged as novel functional entities of the eukaryotic genome. However, the scientific community remains divided over the amount of true non-coding transcripts among the large number of unannotated transcripts identified by recent large scale and deep RNA-sequencing efforts. Here, we systematically exclude possible technical reasons underlying the absence of lncRNA-encoded proteins in mass spectrometry datasets, strongly suggesting that the large majority of lncRNAs is indeed not translated.

INTRODUCTION

Advances in sequencing technologies have uncovered pervasive transcription of the eukaryotic genome outside of annotated protein-coding loci. Most of these novel transcripts are long (> 200 nucleotides), lack large open reading frames (ORFs) and homology to annotated protein-coding genes[1]. Termed long non-coding RNAs (lncRNAs), these transcripts comprise a vast, diverse and largely unexplored class of RNA, outnumbering any other class of genetic entities in the human genome[2]. Those that have been studied in detail play important roles in a wide range of cellular processes during normal development and in homeostasis and disease, including cancer[3].

Similar to lncRNAs, short open reading frame (sORF)-encoded polypeptides (SEPs) or micropeptides have gained increased attention over the past few years. While classical bioactive peptides are enzymatically cleaved from longer protein precursors, micropeptides are small peptides (< 100 amino acids) directly translated from single sORFs. So far, only a limited number of these micropeptides have been discovered and functionally characterized[4].

The coding potential of newly discovered RNA transcripts is typically assessed by means of prediction algorithms[5-7]. While each algorithm has its own strengths and weaknesses, they are all biased to current annotations and may thus be unsuitable for the detection of small or non-conserved proteins including micropeptides.

Although the advent of ribosome profiling[8] (sequencing of ribosome protected RNA fragments) promised to provide evidence for (the lack of) translation of expressed ORFs, much is still open to interpretation. Numerous studies report substantial ribosome occupancy of lncRNA transcripts[9-12]. The striking similarities in the pattern and size of ribosome protected fragments covering protein-coding transcripts and lncRNAs have led some researchers to conclude that up to 90% of the lncRNA transcriptome bears coding ORFs[10]. Other researchers report much more conservative numbers[11-14]. For instance, if the relative abundance of ribosomes before and after stop codons (termed ribosome release) is used to discriminate between protein-coding and non-coding transcripts, only a few novel coding ORFs

are found[11]. When taking into account the phased movement of ribosomes across translated ORFs, only a small number of novel peptides arising from transcripts annotated as lncRNAs[13] are identified. Different research groups have thus developed different metrics and methodologies to detect coding ORFs in ribosome profiling data. Without a consensus, the true coding potential of lncRNA transcripts remains open to speculation.

Mass spectrometry is often considered as the gold standard in detection and characterization of proteins or peptides. So far, few studies have turned to mass spectrometry to study micropeptides and lncRNA-encoded proteins. Reported numbers vary from less than 100 up to 1,600 in human[15-18]. Compared to the more than 60,000 reported lncRNA genes[2,15], these numbers are fairly low and definitely much lower than those reported by various ribosome profiling studies.

This discrepancy in the reported amounts of potentially coding lncRNAs is the source of spirited discussion in the field. Indeed, a resolution of this conflict has direct relevance for further investigations into the biological roles of lncRNAs.

The most direct observation of coding lncRNAs is the actual detection by mass spectrometry based proteomics of the encoded proteins. As such, the absence of large amounts of detected lncRNA-derived proteins strongly hints at a limited coding potential for lncRNAs. The main criticism of this approach however, is that mass spectrometry-based proteomics is somehow biased against the detection of lncRNA products.

Here, we therefore examine the possible biases of mass spectrometry to detect and characterize lncRNA-encoded proteins based on a detailed yet exhaustive reprocessing of very large amounts of public proteomics data. Our findings clearly show that there are no obvious technical reasons why mass spectrometry would have largely missed (micro)peptides originating from non-coding RNA transcripts, thus eliminating the possibility that mass spectrometry would be biased against the detection of putative lncRNA-encoded proteins.

THE INFLUENCE OF PROTEIN COMPOSITION ON DETECTABILITY BY MASS SPECTROMETRY

Mass spectrometry enables high-throughput protein identification in complex samples. However, there is some controversy regarding the limitations of this technique in terms of detectability of peptides and thus, by extension, proteins. Several potential causes have been proposed, including biases due to the size of the protein sequence, the amino acid composition, the abundance, and the half-life of proteins[19-21]. Here, we investigate these presumed issues and identify potential reasons as to why certain predicted ORF products evade detection. The applied strategy revolves around the reprocessing of publicly available data in PRIDE[22], one of the world's leading mass spectrometry repositories[23]. Sequence database searches were performed using an automated reprocessing pipeline, consisting of pride-asap[24] for the detection of data set specific parameters, SearchGUI[25] to match the fragmentation mass spectra against peptides derived from protein sequence databases, and PeptideShaker[26] to integrate the identifications and control these at a 1% false discovery rate at the peptide-to-spectrum match level (see Supplementary Material for details).

A first potential factor that may contribute to a detection bias is the size of a protein. In order to analyse this, publicly available submissions of human projects to PRIDE were searched against the human complement of the UniProtKB/SwissProt[27] protein sequence database using our reprocessing pipeline. The resulting set of proteins was ranked according to sequence length. A simple spectral count over all PRIDE assays in which a protein was identified, was used to indicate the number of times the protein was observed. Q8WZ42, the megadalton protein titin, represented by its canonical isoform of 34,350 residues, was identified 298 times in 183 assays. This indicates that large proteins are picked up despite their length, as is to be expected due to the relatively higher number of potential MS/MS-identifiable peptides following enzymatic cleavage of larger proteins. At the same time, short proteins are also frequently identified across a broad range of assays (Table 1). It is noteworthy that out of 20,207 human entries in UniProtKB/SwissProt, only 36 –(mainly) tissue or

cell specific– proteins (0.18%) are smaller than the shortest reported protein sequences in Table 1. These numbers provide a strong indication that protein length is not likely a major determining factor in protein detectability by mass spectrometry using standard sampling protocols.

protein	gene name	length (AA)	average MW (da)	spectral count	assay count
P62328	TMSB4X	44	4921.46	787	287
P63313	TMSB10	44	4894.48	366	229
Q8N4H5	TOMM5	51	6035.31	88	70
P62891	RPL39	51	6275.49	109	52
Q59GN2	RPL39P5	51	6322.59	107	51
Q5VTU8	ATP5EP2	51	5806.87	53	43
P56381	ATP5E	51	5648.57	53	43
Q96IX5	USMG5	58	6326.38	112	86
P62861	FAU	59	6647.86	248	141
P13640	MT1G	62	6647.86	71	47

Table 1: **The ten shortest human proteins identified by reprocessing of the reprocessed PRIDE data.**

A second feature that could impose a bias on protein detection using mass spectrometry is the amino acid sequence composition. The existence of such a potential bias was investigated by comparing the composition of peptides that have been identified at high confidence with the composition of *in silico* generated peptide sequences. A theoretical digest of the human UniProtKB/SwissProt database was therefore created using dbtoolkit[28] with tryptic cleavage rules, allowing for two missed cleavages. Both empirical peptides from the reprocessing of the human data in PRIDE and *in silico* obtained peptide sequences from the *in silico* digest of UniProtKB/SwissProt were filtered to sizes between 5 and 30 amino acids, which is the common range of observed peptide lengths in practice[29]. The amino acid composition of both theoretical and observed peptides was then calculated by counting the occurrence rate of an amino acid per position in the sequence (Figure 1). There is a high positive correlation between both datasets (Spearman $\rho = 0.952$, p

< 0.01), hinting that there is no reason to assume that the composition of proteins identified by the reprocessing of PRIDE and those generated by *in silico* digestion is very different. The somewhat higher occurrence rates for R and K in the experimental data are most likely related to the fact that these residues are strong bases and therefore strongly promote ionization. The explanation for the slightly lower occurrence rate of S in the experimental data can be related to the fact that S can be phosphorylated *in vivo*, and the somewhat lower efficiency in the detection of phosphorylated residues.

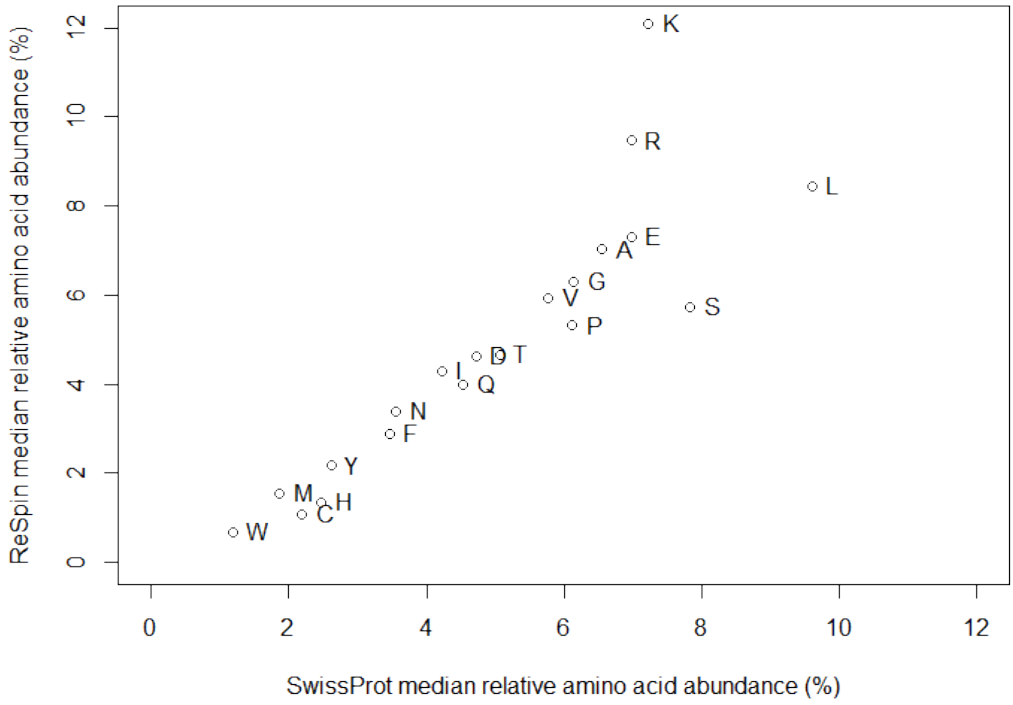


Figure 1: Comparison between theoretical (UniProtKB/SwissProt) and observed (reprocessed PRIDE data) peptide sequence amino acid composition for human data from PRIDE and UniProtKB/SwissProt.

Another important property that can affect detection by mass spectrometry is protein (and thus peptide) abundance in the sample. Although there are examples of successful enrichment protocols[30], the detection of products of rare translation events is not straightforward. In order to investigate the influence of the abundance on the detectability of proteins by mass spectrometry, we first make use of the study

by Anderson and Hunter[31] that reports empirically obtained protein quantification values in human blood plasma. Reprocessing of the subset of PRIDE data sets derived from human blood was carried out, and their estimated abundances were mapped to the values reported by Anderson and Hunter (Figure 2). While it is clear that the lowest abundant proteins are not detected, the abundance range of human plasma is quite extreme at eleven orders of magnitude, of which at least eight are covered reliably in the PRIDE data. This analysis thus shows that mass spectrometry based proteomics is only biased against the very least abundant proteins.

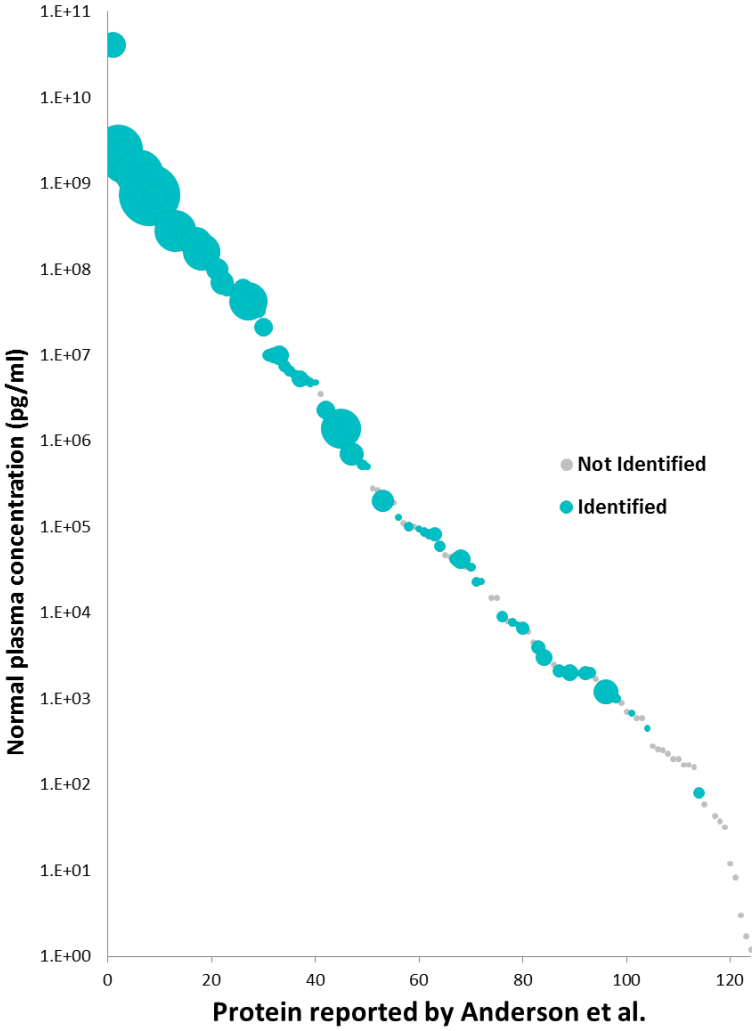


Figure 2: Reprocessing results for PRIDE data sets derived from human blood plasma mapped onto the abundance values reported by Anderson and Hunter. The size of a bubble corresponds to the number of PRIDE assays in which that protein was identified.

Another possibility for detection bias is provided by the half-life of a protein as rapidly degraded proteins may escape detection as well. In order to assess a possible

bias based on protein half-life, we make use of the study by Schwanhäusser *et al.*, where half-life values for murine proteins are reported[32]. Because PRIDE also contains murine data, extensive reprocessing of these murine data sets against the mouse complement of the UniProtKB/SwissProt database was performed and the reprocessed identifications were mapped to the originally reported half-life data (Figure 3). This analysis reveals that the PRIDE data cover the entire half-life range, indicating no influence of protein half-life values on detectability.

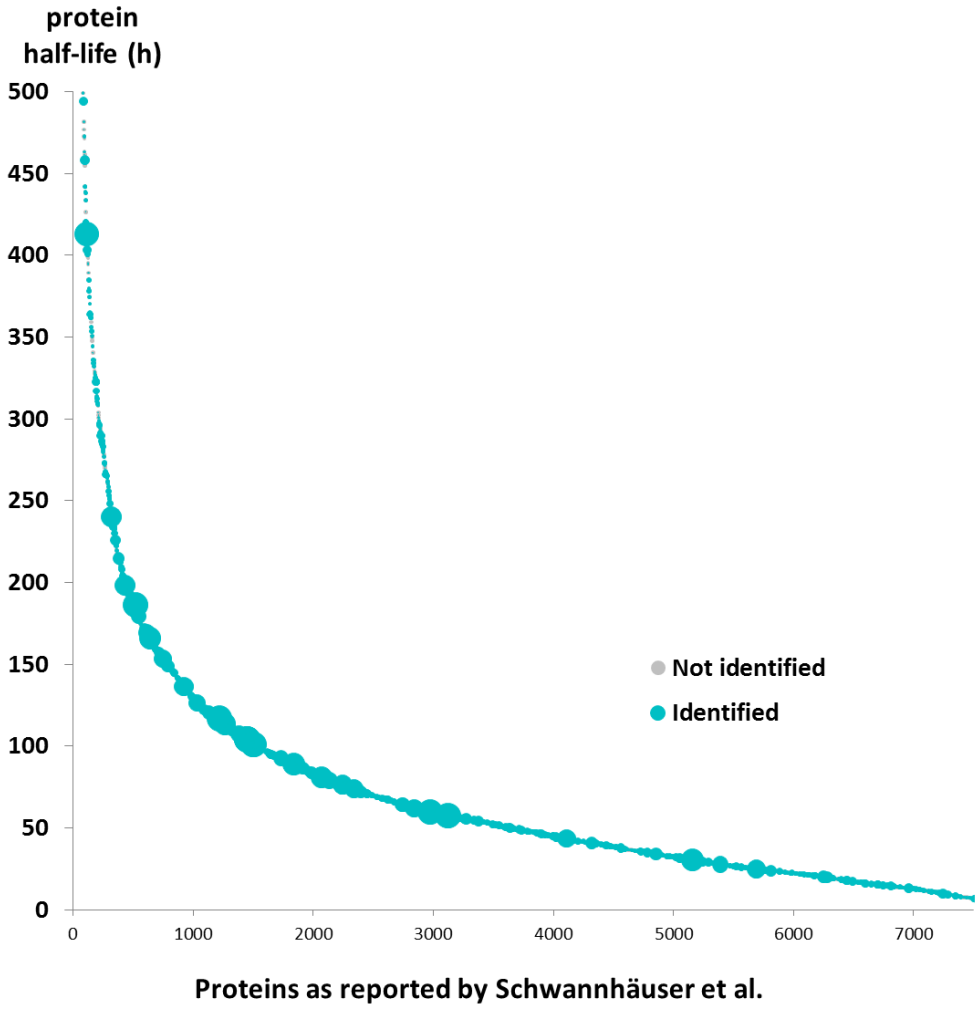


Figure 3: Reprocessing results for all PRIDE murine data mapped onto the half-life values reported by Schwanhäusser *et al*. The size of the bubble corresponds to the number of PRIDE assays in which the protein was identified.

In addition, we calculated the N-terminal instability index of human proteins as described by Guruprasad *et al.*[33]. This metric is based on the dipeptide composition

of a protein and provides a crude estimation of protein half-life when large-scale experimental data are lacking, as is the case for human proteins. The underlying assumption is that a protein's half-life correlates negatively to its relative instability. We therefore compared the calculated instability indices for all proteins in the human complement of UniProtKB/SwissProt with those calculated for the identified proteins from the human data sets in PRIDE. Only a minor deviation is revealed between the instability index distributions of observed and theoretical proteins, providing additional proof that the degradation rate of a protein is of little, if any, influence on its detectability.

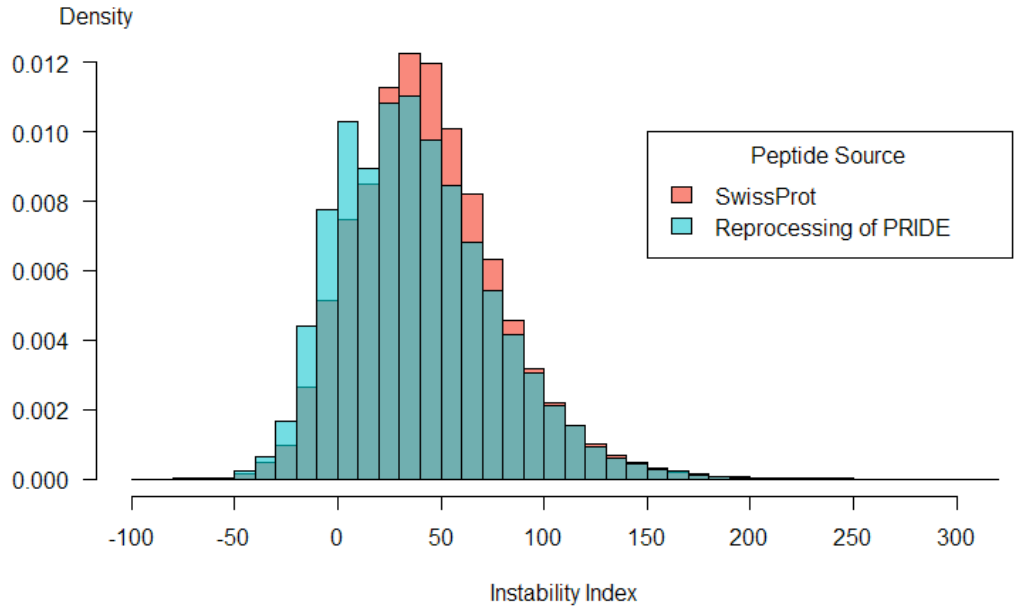


Figure 4: The instability index distributions of human UniProtKB/SwissProt proteins, and of identified proteins from reprocessed human data sets in PRIDE.

LNCRNA EXPRESSION AND COMPOSITION SHOW NO INDICATION OF CODING POTENTIAL

The expression profile of lncRNAs differs extensively from that of protein coding mRNAs (Figure 4a). LncRNAs are generally expressed at a lower level and are more abundant in the nucleus. While mRNAs are transported to the cytoplasm for ribosomal translation, several lncRNAs have a documented function in the nucleus[34]. As such, the nuclear enrichment of lncRNAs suggests a non-coding role for the majority of the lncRNA transcripts.

We have observed that very low protein abundance can hamper the detection by mass spectrometry (**Error! Reference source not found.**2) and lncRNAs are expressed at lower levels compared to mRNAs. Because expression level is a good predictor for protein concentration[35], one might speculate that lncRNAs give rise to proteins at concentrations below the mass spectrometry detection limit. To examine this issue, we first compared lncRNA and mRNA expression levels in the GENCODE v7 dataset1 (see Supplementary Material for details). While the average expression level of lncRNAs is below that of protein coding genes, the expression range is very similar (Figure 4b). In addition, a substantial number of lncRNAs are expressed at levels similar to typical mRNA transcripts. To evaluate the protein detectability as a function of its mRNA expression, we compared mRNA expression levels to the normalized spectral abundance factor (NSAF+)[36] of the corresponding protein. The expression level is defined as the maximally observed RPKM (reads per kilobase per million mapped reads) for a particular mRNA across 11 cell lines in the GENCODE dataset. The maximally observed NSAF+ for each protein from the 4,413 assays in PRIDE that originate from these cell lines is reported. The NSAF+ and RPKM show a low but significant correlation (Spearman $\rho = 0.32$, p-value < 0.01), which is particularly apparent in the higher expression ranges (Figure 4c). Importantly, even though low abundant proteins are more difficult to detect, detected proteins cover the entire expression range. Thus, should lncRNAs give rise to proteins, their concentrations should be detectable by mass spectrometry.

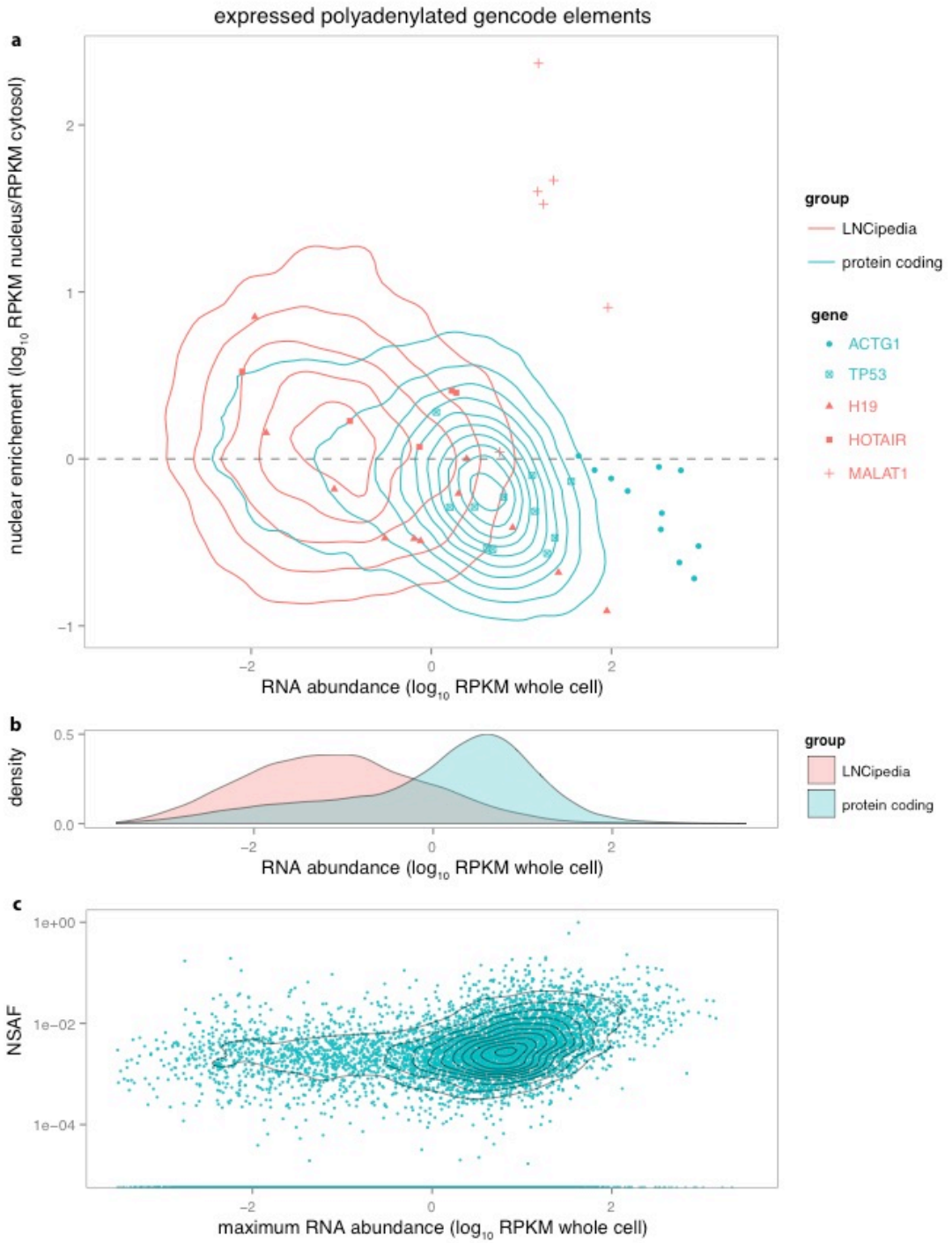


Figure 4: LncRNA and mRNA expression profile and detectability. a) Two-dimensional kernel density plot of lncRNA and mRNA expression levels and subcellular localization. The enrichment of nuclear over cytosolic expression *versus* the expression in the whole-cell extract is shown. Selected lncRNA and protein coding genes are depicted. Especially low abundant lncRNAs show nuclear enrichment compared to mRNAs (adapted from Djebali *et al.*[1]) b) Whole-cell expression distribution for lncRNAs and mRNAs. Although lncRNAs are generally expressed at lower levels, a substantial overlap is observed. c) Normalized spectral abundance factor (NSAF) of the detected protein as a function of its RNA expression level. While mRNA expression and NSAF are moderately correlated, the entire range of expression is clearly covered and thus detectable with mass-spectrometry.

The fact remains that most (if not all) lncRNAs contain canonical ORFs. While predictions classify these as non-coding (hence the annotation as lncRNA), it is conceivable that these ORFs represent recent evolutionary adaptations and are thus difficult to detect by *in silico* analyses. To evaluate if lncRNA ORFs are evolutionary retained or products of random nucleotide progression, we examined the relative size of these ORFs. By using the reverse complement of the sequence as a control, it is obvious that mRNA ORFs are much larger than random ORFs in the reverse complement sequence (see Supplementary Material for details). In contrast, lncRNA ORFs do not differ in size from randomly occurring ORFs, suggesting that they are indeed the product of random nucleotide progression. In addition, it was previously shown that lncRNA ORFs do not show the within-species substitution patterns expected of recently evolved proteins[11].

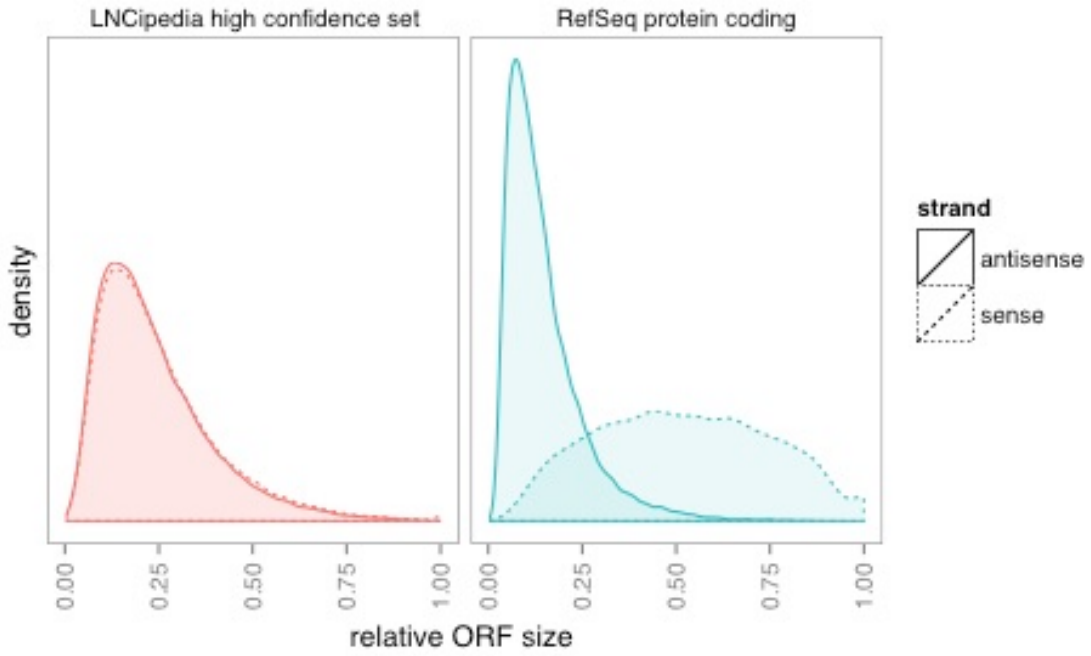


Figure 5: Relative size of the largest canonical ORF in mRNA and lncRNA transcripts. Using the reverse complement sequence as a control, it is apparent that lncRNA (as opposed to mRNA) ORFs are not larger than what would be expected from random nucleotide progression.

CONCLUSION

Investigations into the proportion of coding lncRNAs have resulted in very different estimates. RNA-based analyses, including ribosome profiling, has led to very high estimates, while the more direct measurement of lncRNA-derived proteins *via* mass spectrometry has turned up only a small percentage of putatively coding lncRNAs. In order to help resolve this discrepancy, we here performed a detailed yet thorough analysis across the very large amounts of publicly data available for the human and murine proteomes to eliminate possible biases of mass spectrometry based proteomics in detecting lncRNA-derived proteins. Our analyses reveal that the detection of proteins by mass spectrometry displays only limited bias, relating to proteins with very low abundance and/or very short sequence lengths (shorter than 44 amino acids). Nevertheless, it should be noted that specialized methods can circumvent the observed protein detection biases. Targeted sampling of less studied tissues may still reveal the existence of lncRNA-encoded, tissue specific[1] translation products. Short translation products can be picked up using peptidomics approaches[37], and enrichment protocols[30] can boost yet unseen (micro-)peptides above the mass spectrometry detection threshold. Our analyses thus also delineate useful methods and protocols for comprehensive analysis strategies that are tailored towards finding yet unfound putative protein products from lncRNAs.

Even though mass spectrometry has its limitations in the detection of very low abundant or very small proteins, we firmly demonstrate here that these limitations alone cannot explain the discrepancy between the observed number of lncRNA-encoded proteins and the predicted number by various ribosome profiling studies. In addition, we show that the putative protein products of lncRNA ORFs do not differ in protein sequence length or composition from currently well-detectable proteins. It is thus unlikely that the majority of the current lncRNA annotation consists of miss-classified protein coding genes. These findings confirm that ribosome association alone is insufficient to define novel coding ORFs, as was already suggested by some ribosome profiling studies.

REFERENCES

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
2. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* (2015). doi:10.1038/ng.3192
3. Mercer, T. & Dinger, M. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* (2009).
4. Crappé, J., Van Criekinge, W. & Menschaert, G. Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteomics* **3,** 128–137 (2014).
5. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** I275–I282 (2011).
6. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research* **41,** e74–e74 (2013).
7. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35,** W345–W349 (2007).
8. Ingolia, N. T. in *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis* **470,** 119–142 (Elsevier, 2010).
9. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147,** 789–802 (2011).
10. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports* **8,** 1365–1379 (2014).
11. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154,** 240–251 (2013).
12. Chew, G.-L. *et al.* Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140,** 2828–2834 (2013).
13. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33,** 981–993 (2014).
14. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E2424–E2432 (2012).
15. Volders, P.-J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research* **43,** D174–80 (2015).
16. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology* **9,** 59–64 (2013).
17. Menschaert, G. *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell Proteomics* **12,** 1780–1790 (2013).
18. Crappé, J. *et al.* Combining in silico prediction and ribosome profiling in a

genome-wide search for novel putatively coding sORFs. *BMC GENOMICS* **14,** (2013).
19. Brewis, I. A. & Brennan, P. Proteomics technologies for the global identification and quantification of proteins. *Adv Protein Chem Struct Biol* **80,** 1–44 (2010).
20. Klie, S. *et al.* Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.* **7,** 182–191 (2008).
21. Leary, D. H., Hervey, W. J., Deschamps, J. R., Kusterbeck, A. W. & Vora, G. J. Which metaproteome? The impact of protein extraction bias on metaproteomic analyses. *Mol. Cell. Probes* **27,** 193–199 (2013).
22. Vizcaíno, J. A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* **41,** D1063–9 (2013).
23. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **32,** 223–226 (2014).
24. Hulstaert, N. *et al.* Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J Proteomics* **95,** 89–92 (2013).
25. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11,** 996–999 (2011).
26. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology* **33,** 22–24 (2015).
27. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **42,** D191–8 (2014).
28. Martens, L., Vandekerckhove, J. & Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **21,** 3584–3585 (2005).
29. Vandermarliere, E., Mueller, M. & Martens, L. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom Rev* **32,** 453–465 (2013).
30. Mustafa, G. M., Larry, D., Petersen, J. R. & Elferink, C. J. Targeted proteomics for biomarker discovery and validation of hepatocellular carcinoma in hepatitis C infected patients. *World J Hepatol* **7,** 1312–1324 (2015).
31. Anderson, L. & Hunter, C. L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell Proteomics* **5,** 573–588 (2006).
32. Schwanhäusser, B. *et al.* Corrigendum: Global quantification of mammalian gene expression control. *Nature* **495,** 126–127 (2013).
33. Guruprasad, K., Reddy, B. V. & Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4,** 155–161 (1990).
34. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81,** 145–166 (2012).
35. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
36. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **5,** 2339–2347 (2006).

37. Schulz-Knappe, P., Schrader, M. & Zucht, H.-D. The peptidomics concept. *Comb. Chem. High Throughput Screen.* **8,** 697–704 (2005).

SUPPLEMENTARY MATERIAL

1.1. REPROCESSING OF PUBLICLY AVAILABLE DATA IN PRIDE

The reprocessing of the public data was performed in four steps: (i) the acquisition of data through the PRIDE webservice (Reisinger *et al.*, 2015), (ii) the extraction of spectra and search parameters using pride-asap (Hulstaert *et al.*, 2013), (iii) the execution of sequence database searches using SearchGUI (Vaudel *et al.*, 2011), and (iv) the collation of identifications using PeptideShaker (Vaudel *et al.*, 2015).

The acquisition of the required input data was executed using the PRIDE webservice (www.ebi.ac.uk/pride/ws/archive/). A query was launched to find all complete (including legacy) human projects in PRIDE, resulting in 639 projects that contained a total of 6438 assays for all human data, and 11 projects comprising 361 assays for all mouse data at the time of retrieval.

Spectra were extracted in the Mascot Generic File (MGF) format from the assay files retrieved from PRIDE using pride-asap. Any assay that contained no useable tandem mass spectra (e.g., an assay with only MS1 data) were discarded at this point. This filtering step lead to 4413 retained assays for human data, and 71 retained assays for mouse data. For human plasma, 811 assays could be found amongst the 4413. were Simultaneously, search parameters were extracted from the reported identifications in PRIDE using pride-asap. These parameters include precursor and fragment ion mass accuraccies, the most probable modifications and their occurrence rate, the used digestion enzyme, and the amount of missed cleavages. In the uncommon event that no identifications were reported in the PRIDE assay, default settings were used ; a precursor and fragment ion mass accuracy of 0.6 da, carbamidomethyl-cysteine as a fixed modification, acetylation of lysine and oxidation of methionine as variable modifications, and digestion with trypsin, allowing for 2 missed cleavages.

The obtained spectra were re-analyzed in SearchGUI using the parameters extracted by pride-asap. Three search engines were enabled in SearchGUI: X!Tandem (Craig and Beavis, 2004), MyriMatch (Tabb *et al.*, 2007), and MS-GF+ (Kim and Pevzner, 2014), with matching performed against the human or mouse complement of the

UniProtKB/SwissProt protein sequence database (release 2015_05) for human and mouse PRIDE data, respectively. These databases were automatically appended with their reversed protein sequences as decoys by SearchGUI.

The SearchGUI output was fed into PeptideShaker to collate the identifications from the three search engines, and to control FDR at 1% at the peptide-to-spectrum match (psm) level. A final list of identifications was then exported by PeptideShaker in CSV format.

1.2. LNCRNA EXPRESSION AND SUBCELLULAR LOCALIZATION

Processed RNA sequencing datasets from the GENCODE project (Djebali *et al*., 2012) were obtained from the Gene Expression Omnibus (GEO) website. The dataset comprises RNA sequencing in eleven cell lines and three cell fractions. Average RPKM values are extracted from the public datasets. Further analysis is performed using the statistical environment R and the ggplot2 and dplyr packages. Using Ensembl identifiers, the transcripts are classified as *protein coding* if the corresponding Ensembl biotype is protein_coding, or LNCipedia 3.1 (Volders *et al*., 2015) if the transcript corresponds to a lncRNA. Only transcripts that are expressed in the three cell fractions are retained. The nuclear enrichment is calculated as the rpkm in the nuclear fraction of the rpkm in the cytosolic fraction. To compare mRNA expression and protein abundance, the normalized spectral abundance factor (NSAF) (Zybailov *et al*., 2006) is calculated in PeptideShaker. Only experiments from the set of eleven cell lines are considered. The resulting files were further processed using custom java code and the NSAF values are matched to the corresponding mRNA transcript using Ensembl and UniProtKB/SwissProt identifiers. Only the highest RPKM and NSAF values across the eleven cell lines are used for visualization. All data visualization is performed with the ggplot2 R package.

accession	cell line	fraction	Filename
GSM758559	GM12878	cell	GSM758559_hg19_wgEncodeCshlLongRnaSeqGm12878CellPapGeneGencV7.txt
GSM758560	GM12878	cytosol	GSM758560_hg19_wgEncodeCshlLongRnaSeqGm12878CytosolPapGeneGencV7.txt
GSM758563	HUVEC	cell	GSM758563_hg19_wgEncodeCshlLongRnaSeqHuvecCellPapGeneGencV7.txt
GSM758564	A549	cell	GSM758564_hg19_wgEncodeCshlLongRnaSeqA549CellPapGeneGencV7.txt
GSM758565	HUVEC	nucleus	GSM758565_hg19_wgEncodeCshlLongRnaSeqHuvecNucleusPapGeneGencV7.txt
GSM758566	H1-hESC	cell	GSM758566_hg19_wgEncodeCshlLongRnaSeqH1hescCellPapGeneGencV7.txt
GSM758568	HepG2	nucleus	GSM758568_hg19_wgEncodeCshlLongRnaSeqHepg2NucleusPapGeneGencV7.txt
GSM758569	HUVEC	cytosol	GSM758569_hg19_wgEncodeCshlLongRnaSeqHuvecCytosolPapGeneGencV7.txt
GSM758570	H1-hESC	cytosol	GSM758570_hg19_wgEncodeCshlLongRnaSeqH1hescCytosolPapGeneGencV7.txt
GSM758574	H1-hESC	nucleus	GSM758574_hg19_wgEncodeCshlLongRnaSeqH1hescNucleusPapGeneGencV7.txt
GSM758575	HepG2	cell	GSM758575_hg19_wgEncodeCshlLongRnaSeqHepg2CellPapGeneGencV7.txt
GSM758576	HepG2	cytosol	GSM758576_hg19_wgEncodeCshlLongRnaSeqHepg2CytosolPapGeneGencV7.txt
GSM765386	GM12878	nucleus	GSM765386_wgEncodeCshlLongRnaSeqGm12878NucleusPapGeneGencV3c.txt
GSM765387	K562	nucleus	GSM765387_wgEncodeCshlLongRnaSeqK562NucleusPapGeneGencV3c.txt
GSM765388	MCF-7	cell	GSM765388_wgEncodeCshlLongRnaSeqMcf7CellPapGeneGencV3c.txt
GSM765399	NHEK	nucleus	GSM765399_wgEncodeCshlLongRnaSeqNhekNucleusPapGeneGencV3c.txt
GSM765400	NHEK	cytosol	GSM765400_wgEncodeCshlLongRnaSeqNhe

			kCytosolPapGeneGencV3c.txt
GSM765401	NHEK	cell	GSM765401_wgEncodeCshlLongRnaSeqNhekCellPapGeneGencV3c.txt
GSM765402	HeLa-S3	cell	GSM765402_wgEncodeCshlLongRnaSeqHelas3CellPapGeneGencV3c.txt
GSM765403	HeLa-S3	nucleus	GSM765403_wgEncodeCshlLongRnaSeqHelas3NucleusPapGeneGencV3c.txt
GSM765404	HeLa-S3	cytosol	GSM765404_wgEncodeCshlLongRnaSeqHelas3CytosolPapGeneGencV3c.txt
GSM765405	K562	cell	GSM765405_wgEncodeCshlLongRnaSeqK562CellPapGeneGencV3c.txt
GSM840137	K562	cytosol	GSM840137_hg19_wgEncodeCshlLongRnaSeqK562CytosolPapGeneGencV7.txt
GSM981244	IMR90	cytosol	GSM981244_hg19_wgEncodeCshlLongRnaSeqImr90CytosolPapGeneGencV10.txt
GSM981245	MCF-7	nucleus	GSM981245_hg19_wgEncodeCshlLongRnaSeqMcf7NucleusPapGeneGencV10.txt
GSM981246	A549	cytosol	GSM981246_hg19_wgEncodeCshlLongRnaSeqA549CytosolPapGeneGencV10.txt
GSM981247	A549	nucleus	GSM981247_hg19_wgEncodeCshlLongRnaSeqA549NucleusPapGeneGencV10.txt
GSM981248	IMR90	nucleus	GSM981248_hg19_wgEncodeCshlLongRnaSeqImr90NucleusPapGeneGencV10.txt
GSM981249	IMR90	cell	GSM981249_hg19_wgEncodeCshlLongRnaSeqImr90CellPapGeneGencV10.txt
GSM981250	SK-N-SH	nucleus	GSM981250_hg19_wgEncodeCshlLongRnaSeqSknshNucleusPapGeneGencV10.txt
GSM981251	SK-N-SH	cytosol	GSM981251_hg19_wgEncodeCshlLongRnaSeqSknshCytosolPapGeneGencV10.txt
GSM981252	MCF-7	cytosol	GSM981252_hg19_wgEncodeCshlLongRnaSeqMcf7CytosolPapGeneGencV10.txt
GSM981253	SK-N-SH	cell	GSM981253_hg19_wgEncodeCshlLongRnaSeqSknshCellPapGeneGencV10.txt

***Table 2**: Processed RNA sequencing datasets obtained from GEO*

1.3. LncRNA ORF statistics

LncRNA and protein coding transcript sequences are obtained from LNCipedia (3.1, high confidence set) and RefSeq (NM_* records only) respectively. Using Perl scripting, all canonical open reading frames (ORF) are determined both in the transcript sequence and in the reverse complement. Further analysis is performed using the statistical environment R and the ggplot2 and dplyr packages. For each transcript and its reverse complement, the largest canonical ORF is selected. The distribution of the sizes of these ORFs (relative to the transcript length) is visualized with the ggplot2 R package.

III.4. **CASE STUDY 1:** DEVELOPMENT OF COMBINED MRNA AND LNCRNA EXPRESSION PROFILING PLATFORMS

Pieter-Jan Volders, Pieter Mestdagh, Björn Menten & Jo Vandesompele

Contributions: The candidate contributed to the concept and performed the design of the platform, supervised data acquisition and performed data analysis and interpretation.

DEVELOPMENT OF COMBINED MRNA AND LNCRNA EXPRESSION PROFILING PLATFORMS

PIETER-JAN VOLDERS, PIETER MESTDAGH, BJÖRN MENTEN & JO VANDESOMPELE

INTRODUCTION

Long non-coding RNA (lncRNA) constitute a large and diverse class of non-coding RNA genes. Although several lncRNAs have been functionally annotated, the majority remains to be characterized. LNCipedia (http://www.lncipedia.org) is the largest public compendium of lncRNAs with and without a known function[1]. While microarrays are a popular choice for gene expression profiling studies, the commercial platforms lack probes covering lncRNAs. Hereby we describe the development of several custom gene expression platforms for detection of both mRNA and lncRNA expression using Agilent SurePrint technology.

EXPRESSION ARRAY VERSION 1 (MARCH 2012)

The SurePrint G3 Human Gene Expression Microarrays offered by Agilent Technologies are a popular choice for gene expression profiling studies both in Center for Medical Genetics Ghent and other laboratories. While this platform covers all human protein coding genes it only measures a limited number of lncRNAs for which little or no annotation is available. Therefore, we designed a custom gene expression microarray that covers all lncRNAs in LNCipedia[1] in addition to the mRNA content of the 60k gene expression microarray. Therefore, we used the same 33,128 mRNA probes as the Agilent SurePrint G3 Gene Expression Microarrays. By removing the poorly annotated lncRNAs from the commercial platform and reducing the number of replicate probes, we were able to free enough space to fit all the content

on the 8x60k platform. Using Agilent's eArray application[a], we designed 23,042 lncRNA specific probes covering 96% of all LNCipedia transcripts available at the time. The final microarray design consists of the Agilent mRNA probe groups, the LNCipedia lncRNA probe group and the recommended reference and replicate probes. Microarrays using this design are available from Agilent (Agilent MicroArray Design ID: 039714).

The performance of the expression array was evaluated using RNA sample titrations according to the MicroArray Quality Control (MAQC) standards[2]. Adequate titration response of the lncRNA probes is shown in Figure 1.

Currently, a total of 1134 hybridizations have been performed at the CMGG using this design as part of experiments belonging to many different research units and projects. That makes our custom array by far the most popular choice for gene expression profiling of human samples in our lab. In addition, our work attracted the interest of Agilent and was presented as a customer success story on their website[3].

EXPRESSION ARRAY VERSION 2 (MID 2013)

With the release of LNCipedia version 2.1[4] mid 2013 we revisited the expression array design. Given the substantial increase in lncRNA content, only 75% of the lncRNA genes in LNCipedia 2.1 were covered by the initial version of the expression array. We therefore set out to update the design and create a second gene expression profiling platform. As we again opted for the 8x60k layout, we could not cover every lncRNA transcript with a unique probe, as the number of spots on the array is insufficient. To cover all possible lncRNAs in LNCipedia nonetheless, we opted to select probes in overlapping exons and thus minimize the number of probes required. Our algorithm preferentially selects probes that target the highest number of transcripts in a certain locus. Using this strategy, we selected 25 961 probes covering 95% of the lncRNA genes and 90% of all lncRNA transcripts present in LNCipedia 2.1 (Table 2). (Agilent MicroArray Design ID: 050524).

[a] https://earray.chem.agilent.com/earray/

Again the Microarray Quality Control (MAQC) reference RNA samples were used to evaluate the array and two different labeling kits. The Low Input Quick Amp labelling kit was found to be the method of choice as it significantly increases the detection rate for lncRNAs (Figure 2).

The updated design has been used for 216 hybridizations in the CMGG to this date. In addition, our design has been shared with several scientific collaborators including the UGent spin-off company Biogazelle which has currently used it to profile over 350 samples.

COLLABORATION WITH AGILENT TECHNOLOGIES AND DEVELOPMENT OF A COMMERCIAL PLATFORM

Following the success of the custom microarray, we contacted Agilent to evaluate the commercial potential of our design. After positive evaluation by Agilent, the company decided to use this design as a basis for the development the next version of their SurePrint G3 platform. In close collaboration, the design was further optimized based on the combined expertise of our lab and the company. The Agilent SurePrint G3 Gene Expression Microarrays for Human version 3 was announced in June 2015 with a press release on the Agilent website and is currently available as a catalogue product[5].

REFERENCES

1. Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41,** D246–D251 (2013).

2. Canales, R. D. *et al.* Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology* **24,** 1115–1122 (2006).

3. Agilent Technologies. *A Custom Gene Expression Array for More Than 21,000 lncRNAs*. *chem.agilent.com* (2013). at <http://www.chem.agilent.com/library/casestudies/Public/Sucess%20Story%20HD%20newsletter%205991-2384EN.pdf>

4. Volders, P.-J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research* gku1060 (2014). doi:10.1093/nar/gku1060

5. Agilent Technologies. Agilent | Agilent Technologies Launches Updated Gene Expression Microarrays for Human, Mouse and Rat Models. *agilent.com* (2015). at <http://www.agilent.com/about/newsroom/presrel/2015/08jun-ca15025.html>

6. Shippy, R. *et al.* Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology* **24,** 1123–1131 (2006).

FIGURES AND TABLES

Table 1: Comparison of the first and second expression array using LNCipedia 2.1 as a reference.

	Version 1.1 (2012)	Version 2.0 (2013)
Nr of mRNA probes	33,128	33,128
Nr of lcnRNA probes	23,042	25,961
Genes covered (LNCipedia 2.1)	13,220 (75.5%)	16,635 (95.0%)
Transcripts covered (LNCipedia 2.1)	23,949 (74.4%)	28,949 (90.0%)

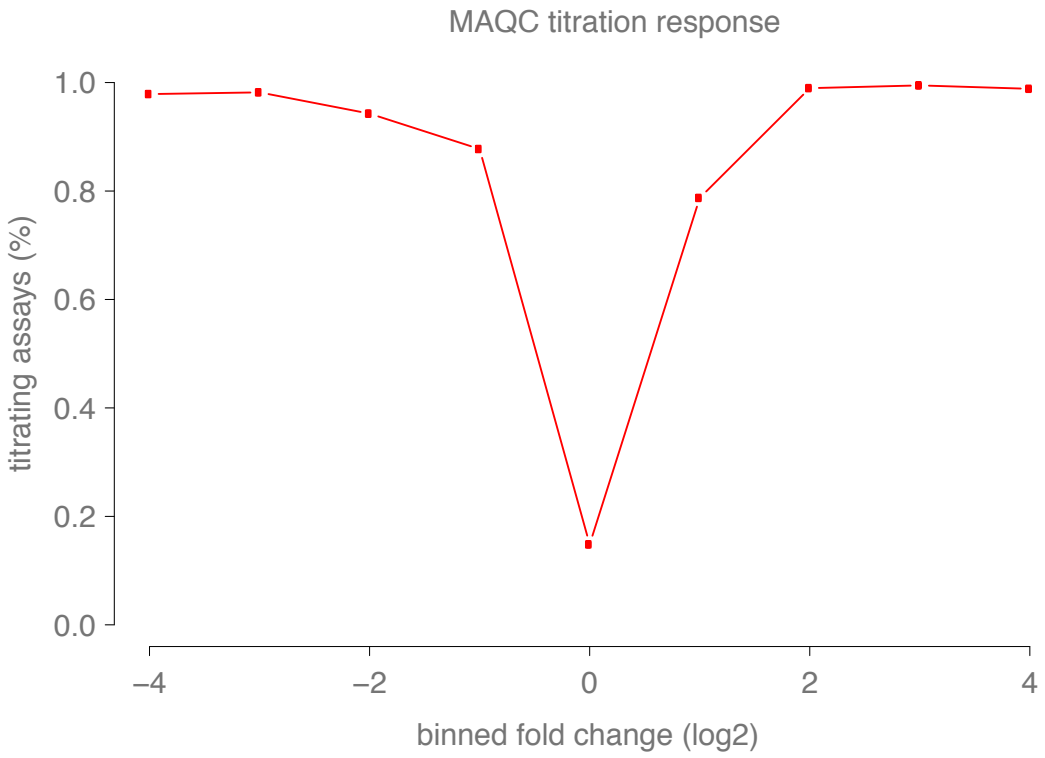


Figure 1: MAQC titration response of lncRNA probes. LncRNA expression was measured for samples A (Universal human reference RNA, Agilent Technologies), B (Human brain total RNA, Ambion), C (25% A + 75% B) and D (75% A + 25% B). The percentage of lncRNA probes that follow the monotonic titration response (Y-axis) is plotted in function of the binned log2-fold change (X-axis) between samples A and B. Titration response was calculated according to Shippy et al., 2006[6].

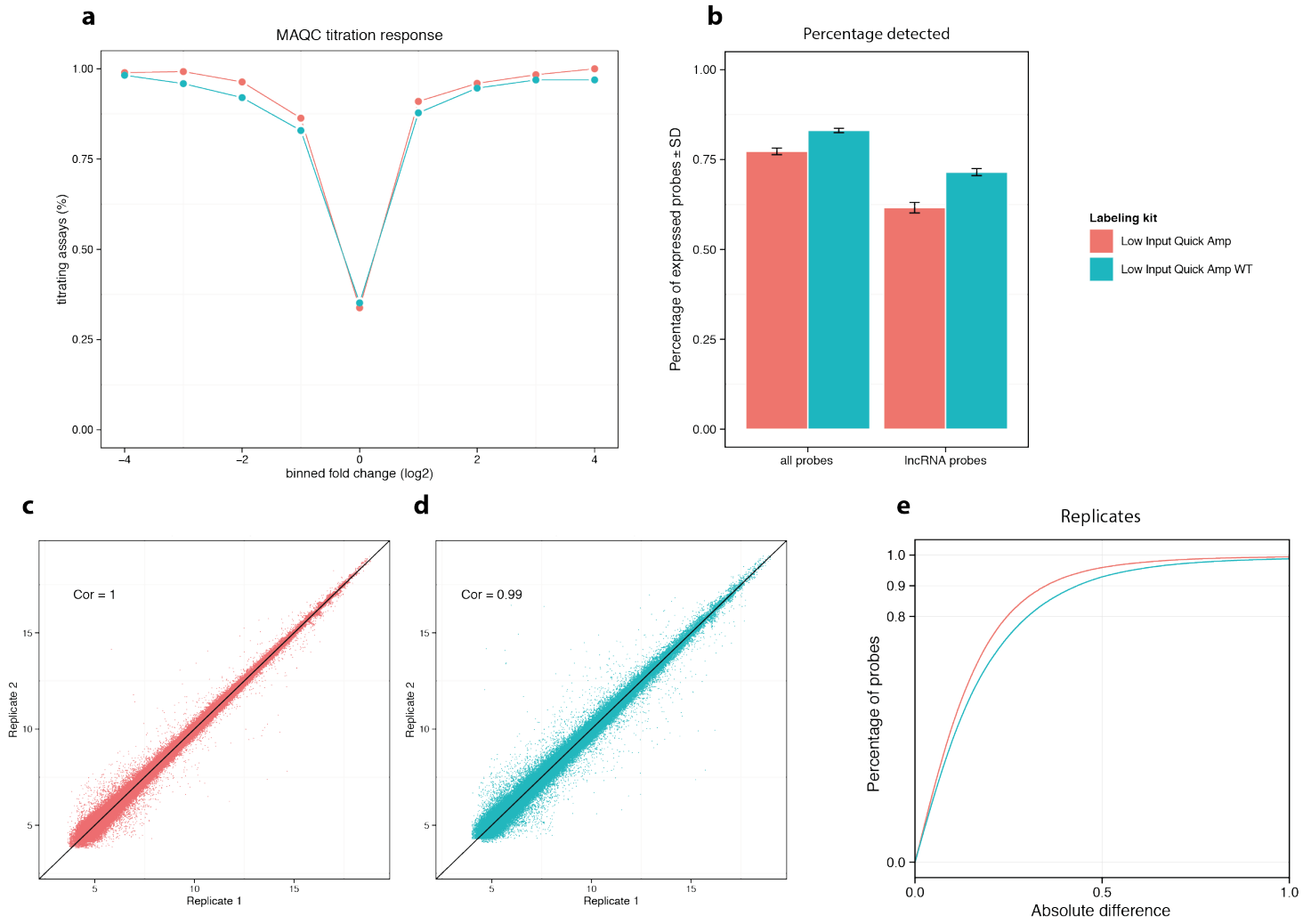


Figure 2: a) The titration response is excellent when using both labeling kits. b) More transcripts can be detected with the whole-transcriptome kit. The percentage of probes that is detected above the background (darkcorner probe + 1) in at least one of the MAQC samples is depicted for the two labeling kits. LncRNAs are found to be expressed at a lower level compared to mRNAs, as is often reported in literature. The number of detectable probes is higher when using the WT kit, especially for lncRNA probes. c,d) A very good correlation is observed for replicates. All samples have been analyzed twice, the Pearson correlation for the expression (probe-level) is reported. e) Labeling with poly-A kit slightly decreases the absolute difference between replicates.

III.5. **RESEARCH PAPER 4**: TARGETED GENOMIC SCREEN REVEALS FOCAL LONG NON-CODING RNA COPY NUMBER ALTERATIONS IN CANCER CELLS

Pieter-Jan Volders, Pieter Mestdagh, Steve Lefever, Björn Menten and Jo Vandesompele

Contributions: The candidate contributed in part to design of the work, contributed to and supervised data acquisition and performed data analysis and interpretation. The candidate drafted the manuscript.

TARGETED GENOMIC SCREEN REVEALS FOCAL LONG NON-CODING RNA COPY NUMBER ALTERATIONS IN CANCER

PIETER-JAN VOLDERS, PIETER MESTDAGH, STEVE LEFEVER, BJÖRN MENTEN AND JO VANDESOMPELE

ABSTRACT

The landscape of somatic copy-number alterations (SCNAs) affecting long non-coding RNAs (lncRNAs) in human cancer remains largely unexplored. While the majority of lncRNAs remains to be functionally characterized, several have been implicated in cancer development and metastasis. Considering the plethora of lncRNAs genes that is currently reported, it is conceivable that several lncRNAs might function as oncogenes or tumor suppressor genes.

We devised a strategy to detect focal lncRNA SCNAs using a custom DNA microarray platform probing 20 418 lncRNA genes. By screening a panel of 80 cancer cell lines, we detected numerous focal aberrations targeting one or multiple lncRNAs without affecting neighboring protein-coding genes. These focal aberrations are highly suggestive for a tumor suppressive or oncogenic role of the targeted lncRNA gene. Although functional validation remains an essential step in the further characterization of the involved candidate cancer lncRNAs, our results provide a direct way of prioritizing candidate lncRNAs involved in cancer pathogenesis.

INTRODUCTION

The cancer genome is marked by large numbers of genetic and non-genetic alterations. The greater majority of those are somatic. Only a small fraction of the somatic mutations, the so-called driver mutations, contribute to cancer development by activating or inactivating specific cancer genes. The remainder are passenger mutations that do not confer growth advantage but were acquired at

some point during cancer cell proliferation[1]. Differentiating between driver and passenger mutations is of the biggest challenges in the quest for new cancer genes and putative therapeutic targets. While somatic alterations can be as small as a single nucleotide substitution, insertion or deletion, somatic copy-number alterations (SCNA) affect the largest fraction of the genome[2]. In some cases, SCNA affect entire or partial chromosome arms. The ability to detect these genetic alterations using (molecular) cytogenetic methods has made large SCNA historically the best studied cancer associated genetic alterations. Many well-known oncogenes and tumor suppressor genes have been initially identified as targets of recurrent genomic amplifications or deletions, respectively. Notable examples are tumor suppressor genes PTEN[3] and RB1[4] and oncogenes HER2 (ERBB2)[5] and the MYC-family of transcription factors[6,7]. The resulting diagnostic and therapeutic successes have made cancer SCNA subject of many studies. Additionally, the advent of genome-wide array comparative genome hybridization (array-CGH) platforms that enable robust identification of small SCNAs greatly improved our knowledge of the cancer genome[8-10].

As cancer genetics until now mainly focused on protein-coding genes, not much is known on SCNAs affecting non-coding RNA genes in cancer. In recent years, our knowledge on the non-coding genome has expanded enormously. This is especially the case for the class of long non-coding RNAs (lncRNAs), consisting of genes with transcripts larger than 200 nucleotides that do not encode proteins. In the past 5 years, ten thousands of human lncRNAs have been reported and catalogued, making this the largest genetic class in the human genome[11]. While the bulk of lncRNAs remains to be functionally annotated, they have been implicated in many important normal cellular processes such as dosage compensation[12], chromatin remodeling[13], and cell differentiation[14]; when deregulated, they play a role in disease as well, including cancer[15].

The discovery of cancer associated lncRNAs such as HOTAIR[16], MALAT1[17] and PVT1[18] uncovered an important role for lncRNAs in oncogenesis. The reason for the current hiatus in our knowledge on lncRNA SCNAs is the fact that the majority of lncRNA annotations are very recent. Most commercially available platforms are based on

older genomic annotations (with no probes for lncRNAs, or probes for as yet unannotated lncRNAs) or lncRNAs are simply overlooked in the data analysis. Indeed, recurrent SCNAs outside of protein coding regions have been reported[2,19]. To overcome this problem, existing DNA microarray platforms have been repurposed and probe content was reannotated with current lncRNA annotation[20,21]. One such effort resulted in the discovery of the oncogenic FAL1 (focally amplified lncRNA on chromosome 1) lncRNA in ovarian cancer[21]. While the potential of this approach lies in its ability to make use of the large amount of publically available DNA microarray data, the used platforms have several disadvantages for the discovery of putative cancer associated lncRNAs. Whole cancer genome sequencing has the potential in principle to circumvent these limitations, but the method is still relatively expensive, and challenging in terms of data-analysis. Consequently, public databases (e.g. TCGA) are mainly populated with targeted exome sequencing datasets, again focusing on protein coding genes.

The occurrence of SCNAs is inversely proportional to their size, with small SCNAs being more common than larger ones[2,19]. However, smaller SNCAs are covered by fewer probes making them more difficult to detect reliably. It is reasonable to assume that a substantial number of SCNAs are overlooked in this way. As SCNA recurrence is often used to prioritize putative cancer genes, more samples will be required to compensate for the undetected small SCNAs. Secondly, reliably detectable and thus larger SCNAs will contain multiple genes, possible including protein-coding genes, making it harder to identify lncRNA cancer genes.

Here we present a targeted and cost-effective approach to identify focal lncRNA SCNA based on a custom DNA microarray covering 20 418 lncRNA transcripts and their flanking protein coding genes. We show the ability of this platform to detect focal aberrations that only affect lncRNA exons and not encompass their flanking protein coding genes. By analyzing the DNA of 80 cancer cell lines covering 11 cancer subtypes we reveal that lncRNAs are frequently targeted by focal aberrations in human cancer. In addition, we have generated a dataset with putative oncogenic and tumor suppressor lncRNAs for future functional studies.

METHODS

LncRNA exon database

LncRNA transcript annotation was obtained from LNCipedia[22] (version 1.0) and stored in a MongoDB NoSQL database. Protein coding transcript annotation was obtained from Ensembl's[23] biomart (version 64, September 2011) and stored in the same format. For every lncRNA transcript, the nearest upstream and downstream protein coding transcript was determined. To interface with the MongoDB dataset, both perl scripts and mongo shell scripts were employed. Using MongoDB's MapReduce functionality, a non-redundant exon collection was built starting from the collection of non-redundant transcripts.

Array CGH platform design

Array CGH probe design was performed using Agilent Technologies eArray software[a]. A BED file of all non-redundant exons was generated from the exon database and uploaded into eArray for probe design. Since our criterion to have 2 probes per exon was initially not met, the exon boundaries were extended and the corresponding BED files were uploaded as well. Exon boundaries are extended with 100 bp, 300 bp and 500 bp. In addition, less stringent selection parameters were used for the 500 bp extended exon. In this way, 5 probe datasets were generated and stored in a separate MongoDB collection. From this collection, 2 probes per exon (neighborhood) were selected with preference for the probes closest to the exon. Overlapping transcripts were taken into account to avoid duplicate probe selection. For transcripts with fewer than 5 exons, additional probes were selected until the transcript was covered by at least 10 probes. For the flanking protein coding genes, probes were designed for the 2 exons closest to the lncRNA. From this set, the 2 probes nearest to the lncRNA were selected. The resulting set of 166 417 unique probes was uploaded to eArray and supplemented with normalization and QC probe groups recommended by Agilent Technologies. Agilent Technologies subsequently manufactured the final design in the 4x180k format.

Cancer cell line DNA and RNA

[a] https://earray.chem.agilent.com/earray/

The National Cancer Institute (NCI) provided DNA and RNA samples for all cell lines in the NCI 60 cancer cell line panel. The neuroblastoma and T-ALL cell lines were available in house, RNA extraction was performed with the miRNeasy Mini Kit (QIAGEN) and DNA extraction with the QIAamp DNA Mini Kit (QIAGEN).

ARRAY CGH

400 ng of genomic DNA was labeled with Cy3-dCTP (GE Healthcare, Belgium) using a Bioprime array CGH genomic labeling system (Invitrogen, Belgium). In parallel, Kreatech gender-matched controls were labeled with Cy5-dCTP. Samples were hybridized on the custom array CGH arrays for 40 h at 65 °C. After washing, the samples were scanned at 5 µm resolution using a DNA microarray scanner G2505B (Agilent Technologies). The scan images were analyzed using the feature extraction software 9.5.3.1 (Agilent Technologies). Segmentation is achieved using the circular binary segmentation algorithm in the DNACopy R package. Visual inspection and creation of the copy number profile plots is performed with 'arrayCGHbase'[24].

SEGMENT ANALYSIS AND FILTERING

Segment position and statistics are stored in a MongoDB collection. A perl script is used to combine the segment annotation with lncRNA and protein coding gene annotation in other collections and implement the filtering process. First, only segments that overlap lncRNA exons are retained. Next, segments with an absolute average log-ratio less than 1.5 are discarded as are segments contained within segmental duplications (UCSC genomicSuperDups track) or segments that overlap with more than 3 known variants (database of genomic variants[25]). The absolute log-ratio of the nearest segments covering the flanking protein coding genes should be 0.5 lower than the segment covering the lncRNA (corresponding to about 1 copy less). A more stringent subset of segments is obtained by requiring the absolute log-ratio of the nearest segments covering the flanking protein coding genes to be less than 0.35 (copy number neutral).

RT-QPCR VALIDATION

QPCR assays are designed based on the chromosomal locations of the altered segment covering the lncRNA and the nearest exons of the two flanking protein coding genes. Primer design is performed using Primer3[26], primers spanning

common SNPs are excluded. Specificity is evaluated using BiSearch[27]. All qPCR reaction are prepared using Bio-Rad's SsoAdvanced Universal SYBR Green Supermix in 5 µl (2.5 µL mastermix, 0.25 µl of each forward and reverse primers (250 nM final concentration) and 2 µl DNA (5 ng)). QPCR plates are analyzed on the LightCycler480 (Roche) using 2 min activation at 95 °C, followed by 45 cycles of 5 sec at 95 °C, 30 sec at 60 °C and 1 second at 72 °C, and a melt curve analysis.

Calculation of normalized relative quantities was done using the qbase+ software version 2.6 (Biogazelle) and the open source statistical environment R (version 3). The Cq values corresponding to the altered segment are normalized to those corresponding to the flanking protein coding genes and scaled to the control sample (Human Genomic DNA, Roche). Downstream analysis and data visualization was achieved using R and third party modules (plyr, ggplot2).

RESULTS

A TARGETED PLATFORM TO DETECT FOCAL COPY NUMBER CHANGES IN LNCRNA GENES

LncRNAs are underrepresented on commercial array CGH platforms and the mean chromosomal distance between the probes on these arrays makes them unsuitable to detect small aberrations that only involve (part of) a single lncRNA gene (Figure S1, Supplementary material).

In order to detect small and focal SCNAs that only affect lncRNA exons, we designed a custom 180k CGH array covering intergenic lncRNA exons and the nearest exons of their flanking protein coding genes. To this purpose, we constructed a database with 52 324 non-redundant exons derived from all transcripts listed in LNCipedia (Figure 1, Figure S2 and Figure S3). The database was subsequently extended with protein coding gene annotation from Ensembl. Next, we designed probes using the genomic sequence of the lncRNA exons and the two nearest exons of the flanking protein coding genes. By removing duplicate probes in overlapping exons and selecting additional probes for transcripts with fewer exons, we were able to cover the majority (94%) of the transcripts with at least 10 probes (Figure S4). Only 1.2% of

lncRNAs could not be covered by any probe. For 95% of the lncRNA transcripts we succeeded in designing 2 probes for each flanking protein coding exon.

To assess the quality of our custom aCGH platform, we compared the profiles for 60 cancer cell lines (NCI-60 subset) to publically available profiles of two different array CGH platforms. The average log ratio in 1 Mb bins was calculated and correlated between the different platforms. These correlations were compared with correlations among unrelated cell lines (Figure S5 and Figure S6). Correlation between the same cell lines across different platforms was high (median Pearson's correlation = 0.70), validating the quality of our profiles. As expected, cell lines derived from the same individual (such as NCI/ADR-RES and OVCAR-8) are also highly correlated (Pearson's correlation = 0.74). In addition, this analysis revealed problems with 2 DNA samples (HCT-15 and CAKI-1) as the obtained profiles showed poor correlation with publically available profiles. This poor correlation remained unresolved by repeating the hybridization. As such, results from these two cell lines should be interpreted with care.

FREQUENT FOCAL LNCRNAS COPY NUMBER ALTERATIONS IN CANCER CELL LINES

To explore focal lncRNA SCNAs in cancer, we analyzed DNA from 80 cancer cell lines covering 11 cancer subtypes with our custom DNA microarray (Table 1). An extensive filtering was performed on the resulting segments to shortlist focal lncRNA SCNA alterations. To be considered a lncRNA SCNA, a segment should (1) overlap with exonic lncRNA sequence, (2) not be contained within known segmental duplications, (3) overlap with at most 3 known variants and (4) have an absolute average log-ratio that is larger than 1.5 (reflecting homozygous deletions and gene amplifications). In the case of a copy number gain, an additional requirement was that the segment includes the entire transcript. Finally, to withhold a focal SCNA (5), the segment cannot overlap any of the flanking protein coding gene exons. To attribute for anomalies in the circular binary segmentation process that generates multiple segments for the same genetic alteration, we pose further requirements on the segments spanning the flanking protein coding genes. A SCNA is only considered focal if the difference between its absolute average log-ratio and that of the segment spanning the nearest exon of the flanking protein coding genes is at least

0.5. Using these settings, 173 focal SCNAs affecting 136 lncRNAs in at least one cell line were identified (Figure 2). The majority of these lncRNAs (111) is affected in a single cell line, 16 are affected in 2 cell lines, 7 in 3 cell lines, 1 in 4 cell lines and 1 in 5 cell lines. By confining the relative difference in log-ratio between the segment covering the lncRNA and the segment covering the flanking protein coding genes, it is possible to retain focal SCNA that are part of larger ones (for instance a large hemizygous deletion that contains a smaller homozygous deletion). A more stringent subset of 76 lncRNA SCNA is obtained if we require that the flanking protein coding gene does not show any copy number change (Figure S7).

RT-qPCR CONFIRMS THE MAJORITY OF FOCAL ABERRATIONS

We devised a unique strategy to validate the selected focal lncRNA SCNAs using qPCR. Assays were designed targeting the genomic locus of the aberration and the nearest exons of the flanking protein coding genes. By comparing the Cq value of the lncRNA locus and the flanking coding exons, we can accurately assess the difference in copy number between the two. Using this strategy, we evaluated 88 events (Figure 3). For 66 of these (75%) an altered copy number status compared to at least one of the two flanking assays could be confirmed, of which 43 (49%) showed the expected relative difference in Cq values with both flanking assays and were thus validated as focal aberrations. The validation rate is higher for the amplifications than for the deletions (56% and 48%, respectively). The validation rate drastically increases when we limit our analysis to the subset of segments with an absolute average log-ratio larger than 2.5. In that case, 58 out of 64 (91%) events are confirmed copy number alterations. The fraction of confirmed focal aberrations remains similar (53%).

MOST NOVEL LNCRNA ABBERATIONS DO NOT CORRESPOND TO COMMON SOMATIC VARIANTS

As our custom platform differs considerably from other array CGH platforms, it not unlikely that the newly found SCNAs actually comprise uncharted germline copy-number variants that may exist in a normal population and do not contribute to cancer. To assess this possibility, we performed an RT-qPCR experiment for five validated loci on DNA from 192 healthy individuals. Neither homozygous deletions

nor high order amplifications could be detected for any lncRNA in any of the samples (Figure S8). Of note, for one lncRNA heterozygous deletions were found in 12 individuals (6%).

DISCUSSION

Even though the number of samples we examined is limited and confined to cancer cell lines, we were able to detect a large number of SCNA that specifically affect lncRNA exons. This suggests that similarly to protein-coding genes, lncRNAs are frequently targeted by SCNAs in cancer. After rigorous filtering focused on novel highly aberrant segments that not encompass protein coding genes, we report 136 such events, including 25 that are recurrent. Of those, 76 events were marked as focal based on the copy number of the flanking protein coding genes. Since the cancer genome harbors many large SCNAs, it is important to also consider the events where the flanking protein coding genes are not strictly copy number normal. As long as the lncRNA itself is focally affected by a second event as well.

Our strategy uncovered several cancer-associated lncRNAs. For instance, the known oncogene lnc-MYC-2 (PVT1) was detected as a recurrent focal aberration (Figure 2, Figure S3). PVT1 has been implicated in several cancer types including gastric cancer[28], ovarian cancer and breast cancer[18]. PVT1 copy number was found to be co-gained in more than 98% of cancers with a MYC copy number increase[29]. Our work not only confirms frequent amplification of PVT1 in cancer, but also reveals that PVT1 amplifications can be focal. Another interesting accordance with previous studies is found in a large-scale pan-cancer study on SCNAs[19]. Although the authors mainly focus on SCNAs affecting protein coding genes and use limited lncRNA annotation, they report one lncRNA, lnc-DCTD-5 (LINC00290), as the sole member of a frequently deleted region. Our results reveal a recurrent and focal deletion in ovarian and breast cancer cell lines, suggesting a role in cancer (Figure 2).

The validation rate determined with qPCR was strongly dependent on the log-ratio cutoff applied to the segments, with an absolute average log-ratio larger than 2.5 showing high validation rates for lncRNA copy number status. The relatively high

cutoff is likely to be related to the unique design of our platform. As the probes are confined to small genomic loci (lncRNA exons) is it not unimaginable that the observed signal-to-noise ratio is different compared to typical designs. In addition, qPCR may not be the most appropriate method to detect hemizygous copy number changes. Even with a stringent log-ratio cutoff (2.5), only 50% of the events could be confirmed to be truly focal. This suggests that the limited number of probes on the flanking protein coding genes is insufficient to define the breakpoints of the segments in some cases.

Nevertheless, even when taking the validation rate into account, our research finds about 100 lncRNAs affected by focal SCNA. As the majority of these events are likely no germline copy-number variants, these SCNAs harbor interesting candidates for further research.

CONCLUSION

We developed and applied a unique array CGH platform capable of detecting small and focal lncRNA SCNAs. We have screened a panel of 80 cancer cell lines and shortlisted 136 lncRNA genes with a putative role in cancer. Among this list are several lncRNAs that have been implicated in cancer, validating our approach. Since the great majority of the lncRNAs on our platform have yet to be functionally studied, this finding suggests that our research provides many new cancer related lncRNA genes. We present a set of lncRNA genes to the lncRNA and cancer research community as novel candidate cancer lncRNA genes for further functional exploration.

REFERENCES

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
2. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
3. Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275,** 1943–1947 (1997).
4. Friend, S. H. *et al.* A human DNA segment with properties of the gene that

predisposes to retinoblastoma and osteosarcoma. - PubMed - NCBI. *Nature* **323,** 643–646 (1986).
5. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235,** 177–182 (1987).
6. Nau, M. M. *et al.* Human small-cell lung cancers show amplification and expression of the N-myc gene. *Proceedings of the National Academy of Sciences of the United States of America* **83,** 1092–1096 (1986).
7. Little, C. D., Nau, M. M., Carney, D. N., Gazdar, A. F. & Minna, J. D. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *, Published online: 10 November 1983; | doi:10.1038/306194a0* **306,** 194–196 (1983).
8. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research* **64,** 3060–3071 (2004).
9. Network, T. C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
10. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–898 (2007).
11. Volders, P.-J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research* gku1060 (2014). doi:10.1093/nar/gku1060
12. Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338,** 1435–1439 (2012).
13. Kogo, R. *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Research* **71,** 6320–6326 (2011).
14. Ulitsky, I., Shkumatava, A., Jan, C., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–1550 (2011).
15. Gutschner, T. & Diederichs, S. The Hallmarks of Cancer: A long non-coding RNA point of view. *rnabiology* **9,** 0—1 (2012).
16. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).
17. Gutschner, T. *et al.* The non-coding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research* **73,** canres.2850.2012–1189 (2012).
18. Guan, Y. *et al.* Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin Cancer Res* **13,** 5745–5755 (2007).
19. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
20. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology* **20,** 908–913 (2013).
21. Hu, X. *et al.* A Functional Genomic Approach Identifies FAL1 as an Oncogenic Long Noncoding RNA that Associates with BMI1 and Represses p21 Expression in Cancer. *Cancer Cell* **26,** 344–357 (2014).
22. Volders, P.-J. *et al.* LNCipedia: a database for annotated human lncRNA

transcript sequences and structures. *Nucleic Acids Research* **41,** D246–D251 (2013).
23. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Research* **30,** 38–41 (2002).
24. Menten, B. *et al.* arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics* **6,** 124 (2005).
25. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. - PubMed - NCBI. *Nucleic Acids Research* **42,** D986–D992 (2013).
26. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132,** 365–386 (2000).
27. Tusnády, G. E., Simon, I., Váradi, A. & Arányi, T. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Research* **33,** e9–e9 (2005).
28. Ding, J. *et al.* Expression and clinical significance of the long non-coding RNA PVT1 in human gastric cancer. *OncoTargets and therapy* **7,** 1625–1630 (2014).
29. Tseng, Y.-Y. *et al.* PVT1 dependence in cancer with MYC copy-number increase. *Nature* (2014). doi:10.1038/nature13311

Table 1: Overview of cell line panel and cell line origin

cancer subtype	#	cell lines	origin
breast	6	MCF7, MDA-MB-231, HS578T, BT-549, T47D, MDA-MB-468	NCI
CNS	6	SF-268, SF-295, SF-539, SNB-19, SNB-75, U251	NCI
colon	7	COLO205, HCC-2998, HCT-116, HCT-15, HT29, KM12, SW-620	NCI
leukemia	6	CCRF-CEM*, HL-60, K-562, MOLT-4*, RPMI-8226, SR	NCI
melanoma	9	LOXIMVI, MALME-3M, M14, SK-MEL-2, SK-MEL-28, SK-MEL-5, UACC-257, UACC-62, MDA-MB-435	NCI
non-small cell lung	9	A549, EKVX, HOP-62, HOP-92, NCI-H226, NCI-H23, NCI-H322M, NCI-H460, NCI-H522	NCI
ovarian	7	IGROV1, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, SK-OV-3, NCI-ADR-RES	NCI
prostate	2	PC-3, DU-145	NCI
renal	8	786-0, A498, ACHN, CAKI-1, RXF-393, SN12C, TK-10, UO-31	NCI
T-cell acute lymphoblastic leukemia	8	Jurkat-DSMZ, ALL-SIL, DND-41, HPB-ALL, TALL-1, LOUCY, MOLT-16, PEER	DSMZ
neuroblastoma	12	CLB-GA, IMR-32, NB-1, NGP, N206, SHEP, SH-SY5Y, SK-N-SH, SK-N-BE-2c, CHP-134, SK-N-AS, CHP-902R	CMGG
11 subtypes	80		

* MOLT-4 and CCRF-CEM are T-ALL cell lines in the NCI60 panel.

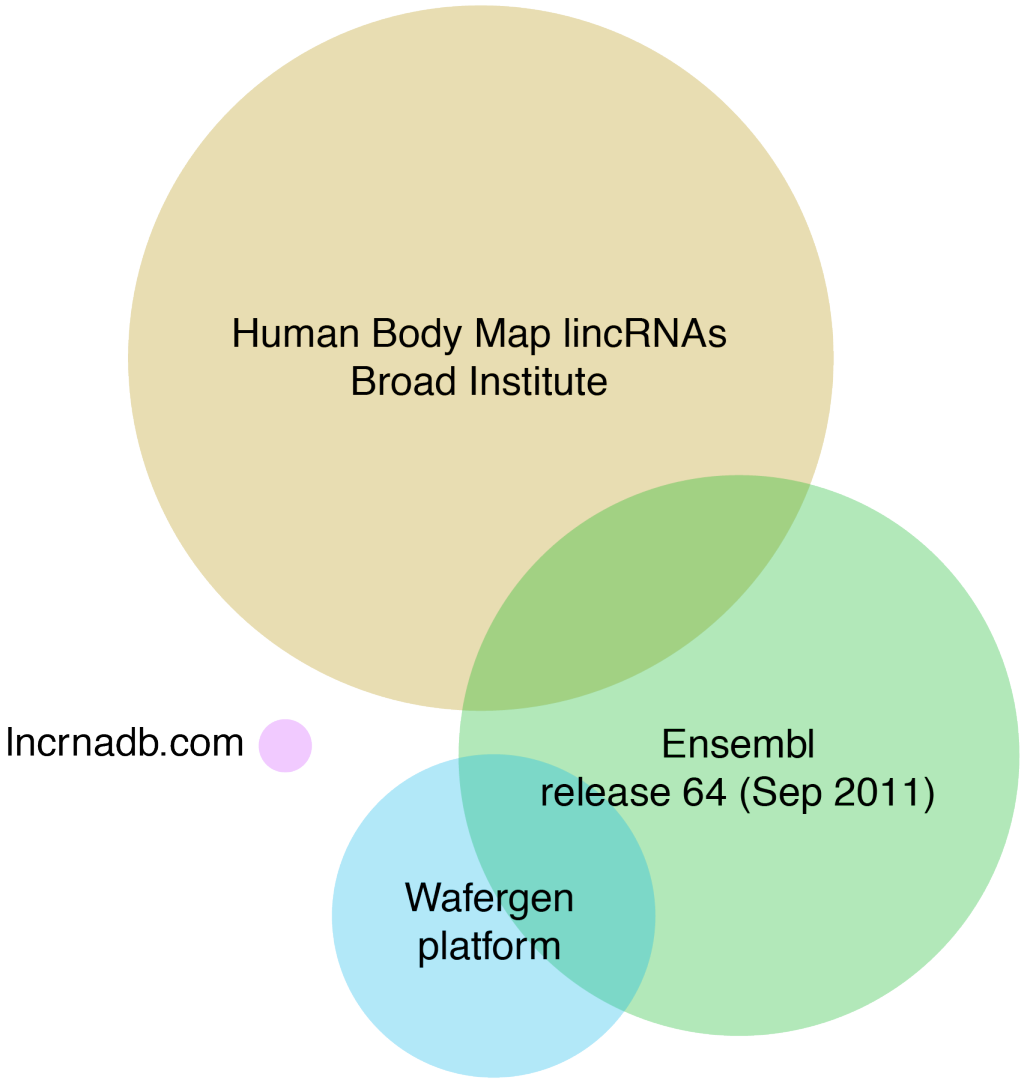


Figure 1: Euler diagram of the different lncRNA sources. The circle diameter and overlap correspond to the number of lncRNAs. The sources include several lncRNA databases and the lncRNAs on the Wafergen SmartChip Human LncRNA1 Panel (http://www.wafergen.com/products/smartchip-panels/smartchip-human-lncrna1-panel/)

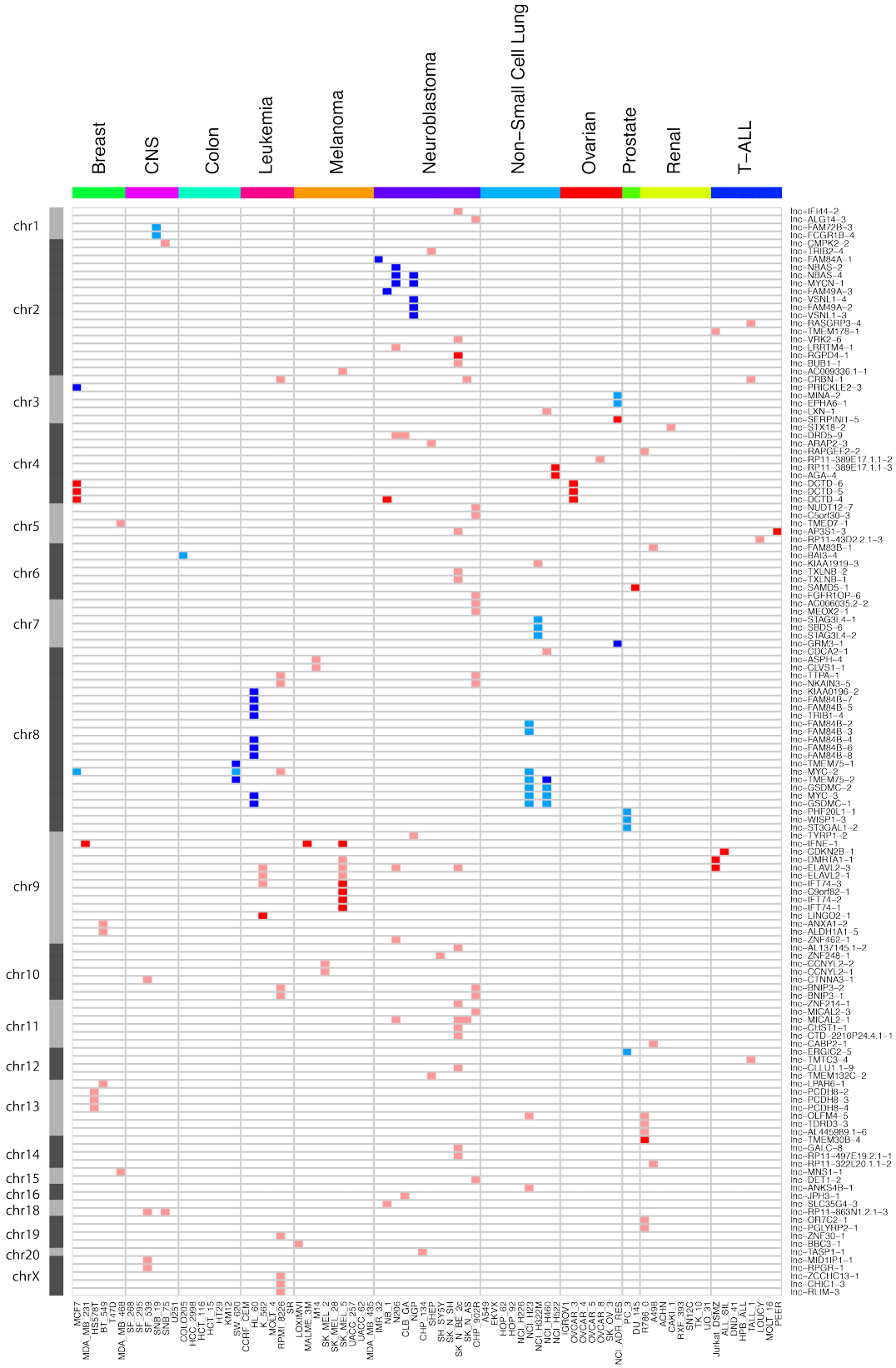


Figure 2: Overview of the lncRNA genes affected by focal SCNAs after extensive filtering. Red represents copy number loss (log-ratio < 1.5) in that cell line while blue corresponds to copy number gain (log-ratio > 1.5). Dark red and blue correspond to copy number changes with absolute log-ratio above 2.5.

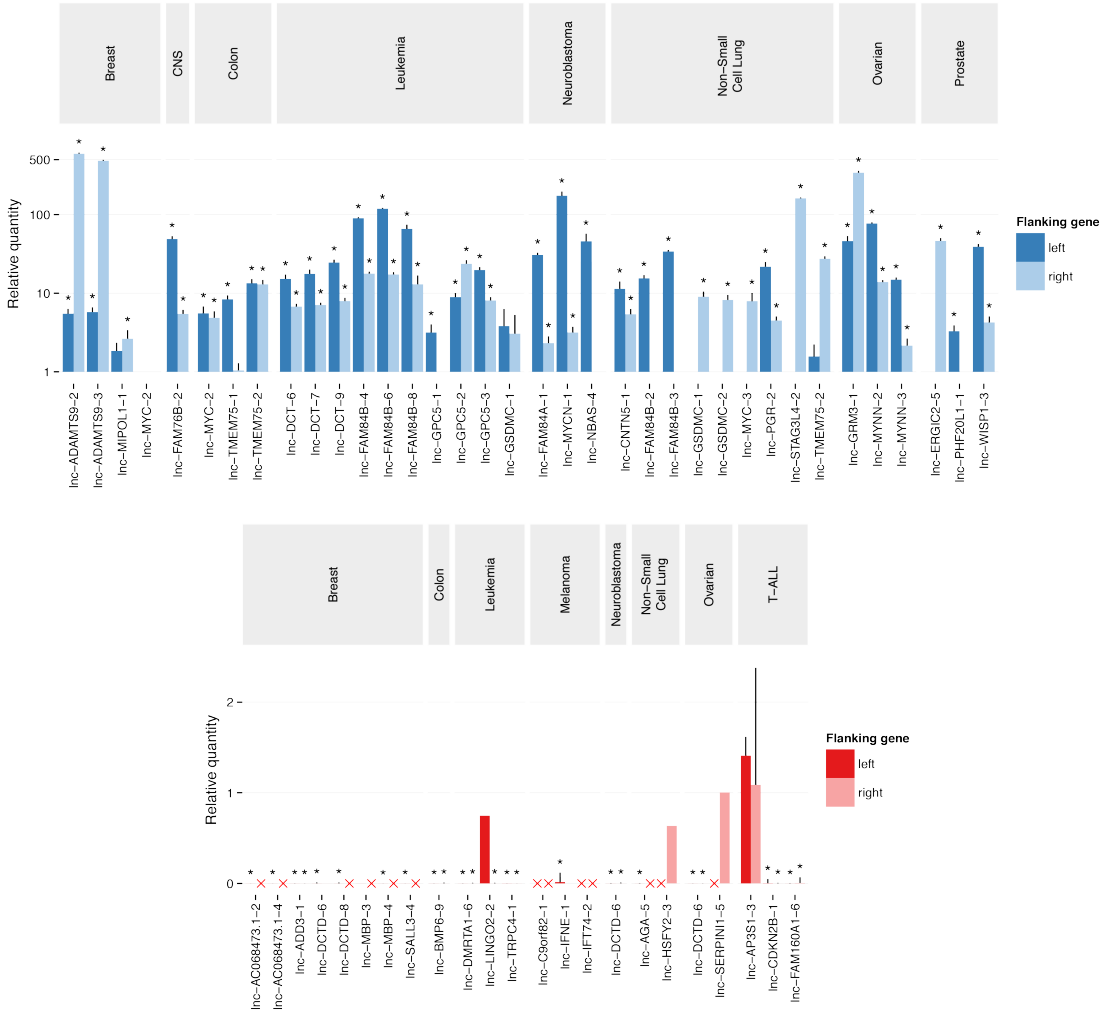


Figure 3: RT-qPCR validation of the putative focal SCNAs. The Cq value of the aberration is normalized to the Cq value of each of the flanking regions. A copy number gain (blue) is considered confirmed and focal when the relative quantity to both flanking regions is higher than 1. Similarly, a copy number loss (red) is considered confirmed and focal when the relative quantity to both flanking regions is less than 1. Red crosses represent Cq values > 35, corresponding to a homozygous deletion of the flanking regions. Stars represent significant (p-value < 0.05) differences from 1.

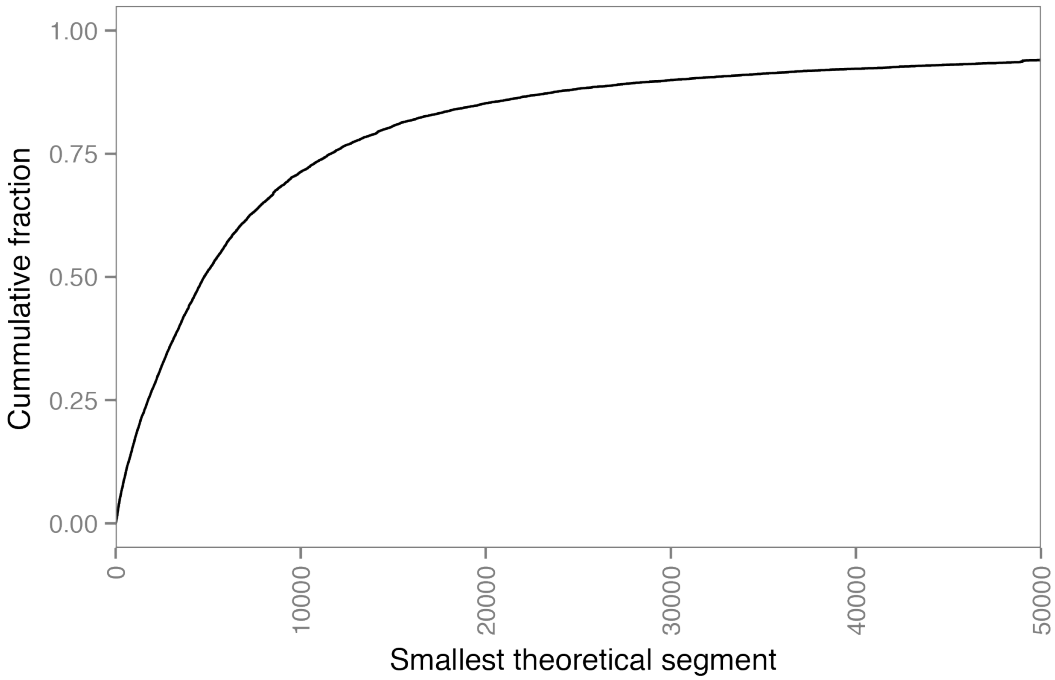


Figure S1: The smallest theoretical segment that covers each lncRNA using the Genome-Wide Human SNP Array 6.0. A theoretical segment is the distance between the two closest probes covering lncRNAs.

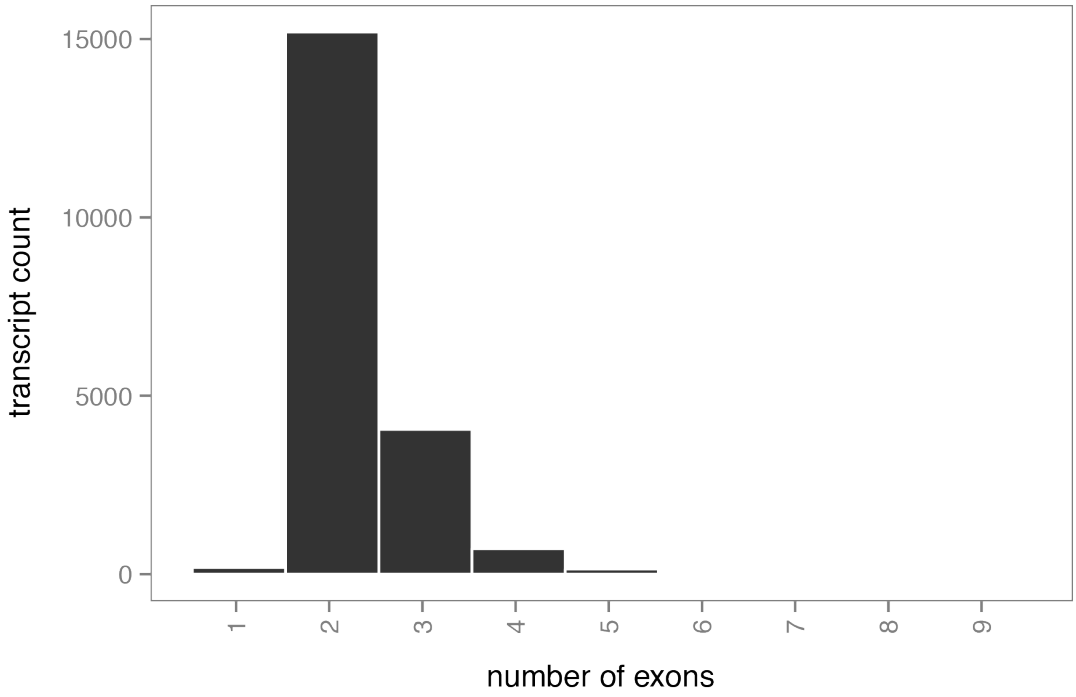


Figure S2: Distribution of the number of exons for the lncRNA transcripts in our dataset

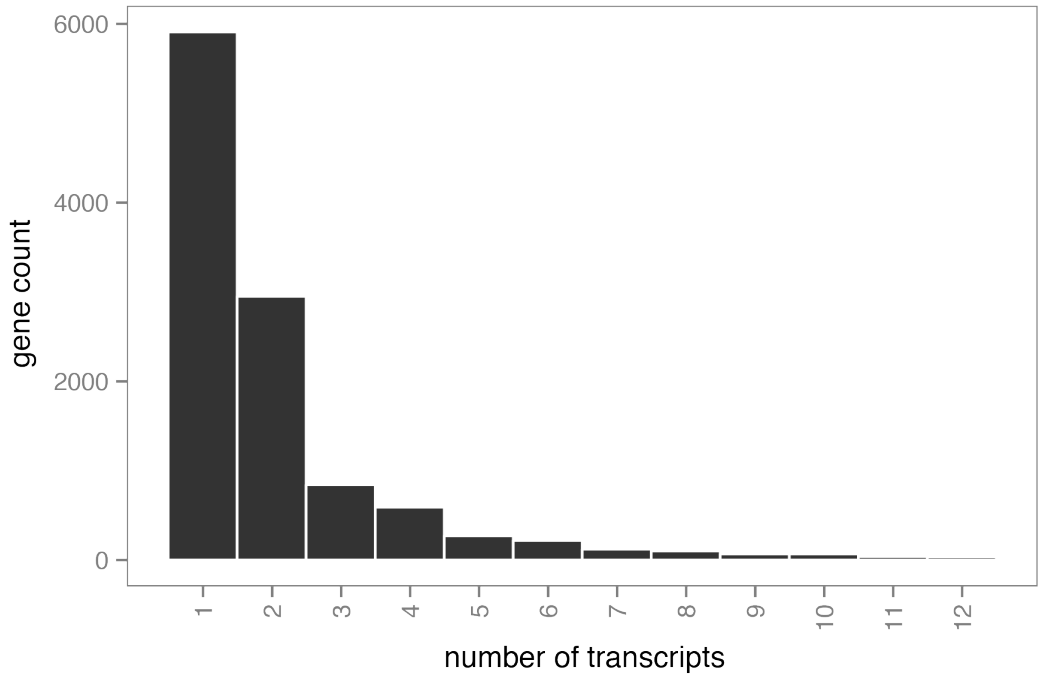


Figure S3: Distribution of the number of transcripts per lncRNA gene (locus).

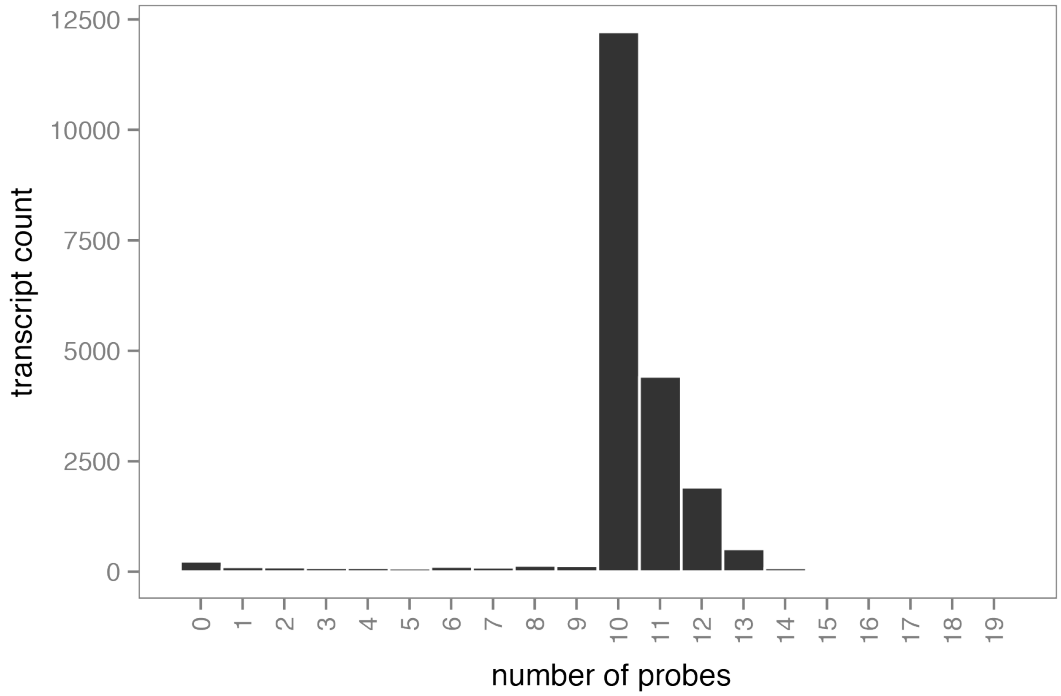


Figure S4: The required number of probes for a transcript depends on the number of exons. For transcripts with five exons or less, the required number of probes is 10. For larger transcripts, the required number is two times the number of exons. Only for a small fraction of transcripts, the pipeline failed to design at least 10 probes.

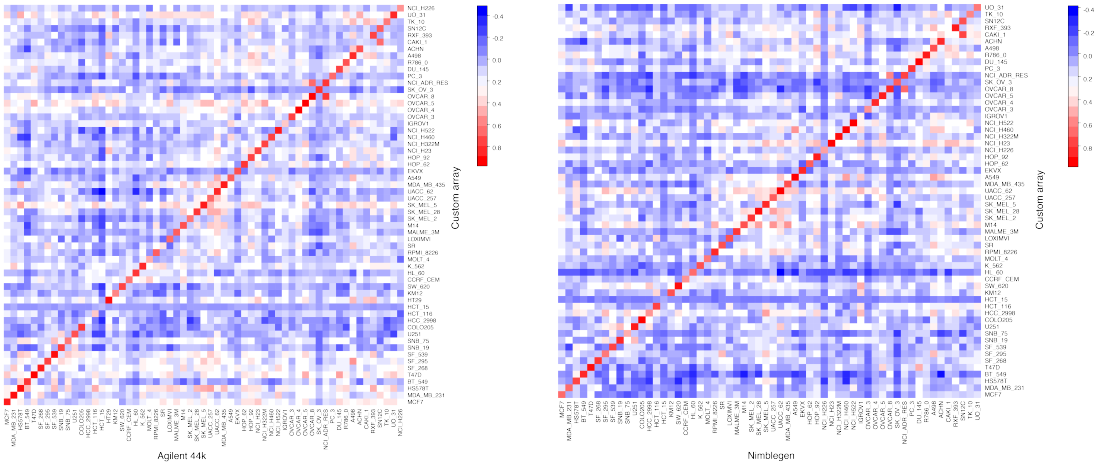


Figure S5: Comparison of the global copy number profiles (averaged in 1Mb bins) with publically available profiles of two different array CGH platforms (Agilent 44K and Nimblegen 385k). Pearson correlation of all samples is depicted. Blue corresponds to no correlation while red is a high correlation. Excellent correlation if observed for our platform. As expected, cell lines derived from the same individual (such as NCI/ADR-RES and OVCAR-8) are also highly correlated. In addition, this analysis revealed problems with 2 DNA samples (HCT-15 and CAKI-1) that were unresolved by repeating the hybridization.

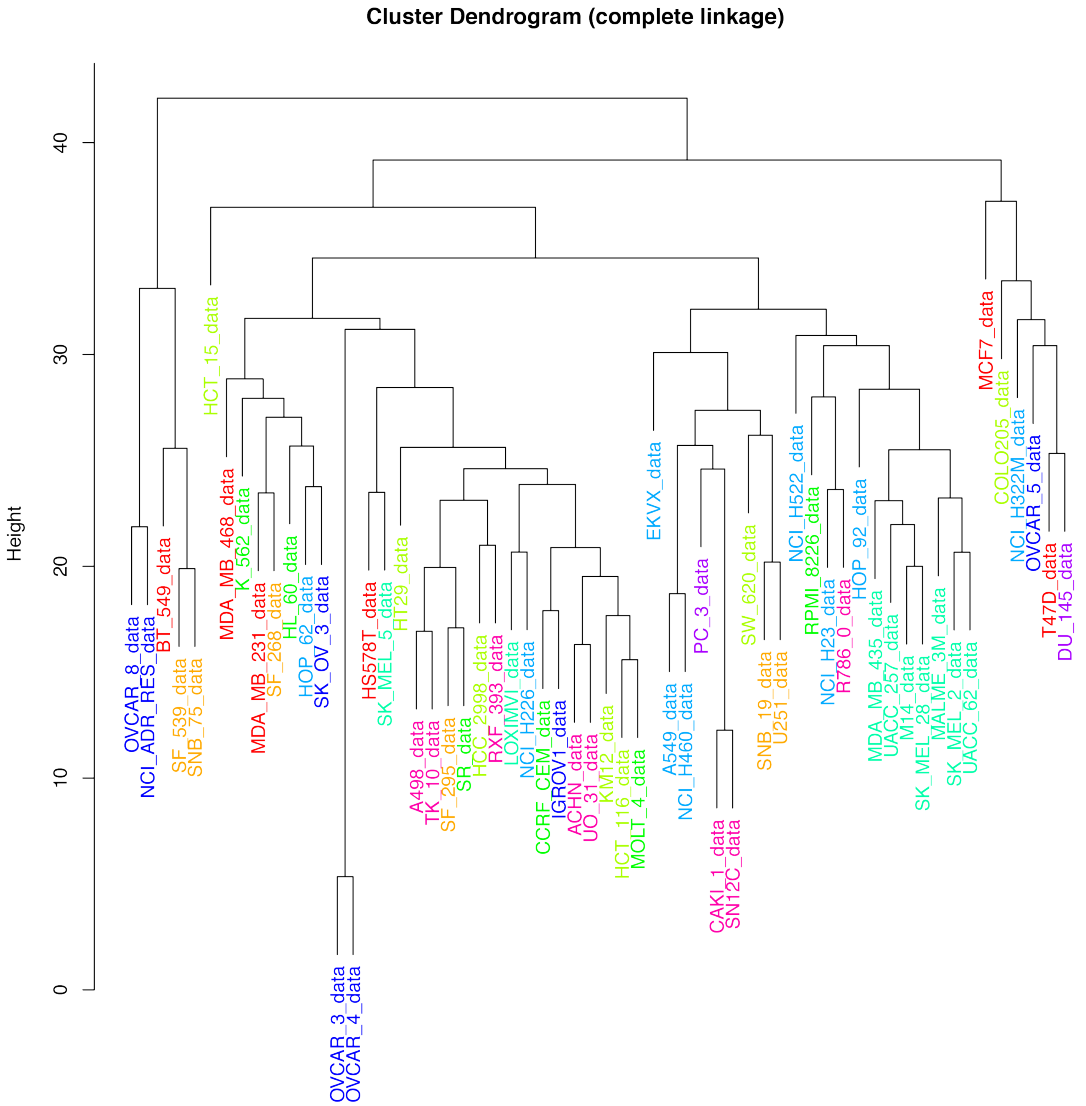


Figure S6: Complete linkage tree of the cell line based on their global copy number profiles. Cell lines derived from the same individual cluster closely together. On a cancer subtype level, only melanoma samples (green) cluster together.

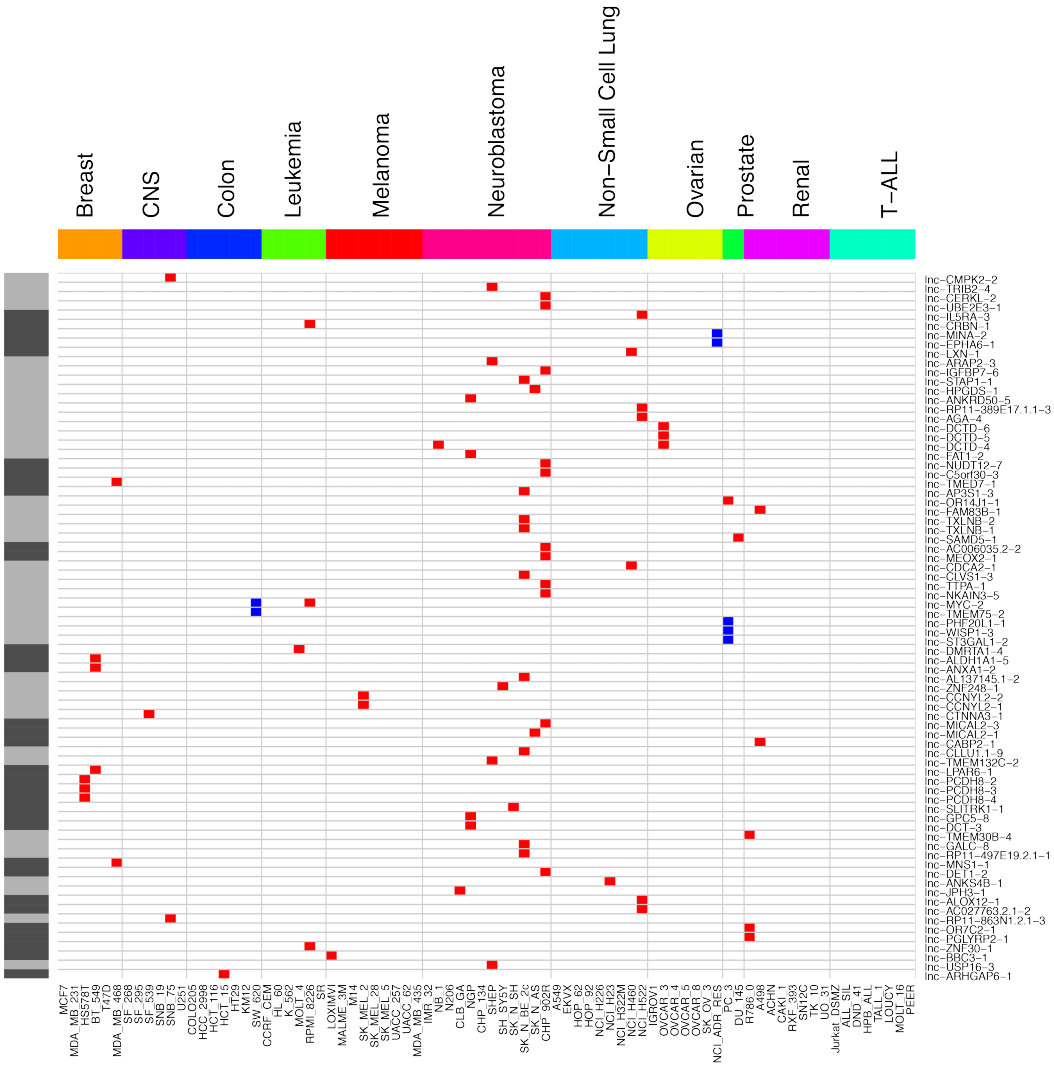


Figure S7: Overview of the lncRNA genes affected by focal SCNAs and copy number normal flanking protein coding genes. Red represents copy number loss (log-ratio < 1.5) in that cell line while blue corresponds to copy number gain (log-ratio > 1.5).

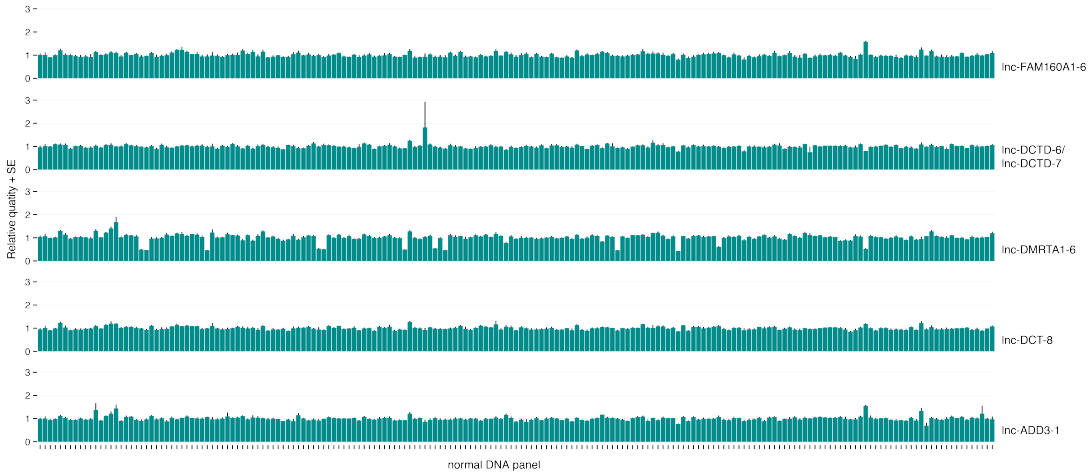


Figure S8: RT-qPCR validation of 5 selected loci on 192 DNA samples of healthy individuals. Neither homozygous deletions nor high order amplifications could be detected for any lncRNA in any of the samples. However, heterozygous lnc-DMRTA1-6 deletions were present in 12 samples (6%).

III.6. **RESEARCH PAPER 5**: POTENT ANTISENSE OLIGONUCLEOTIDE SELECTION FOR LNCRNA KNOCKDOWN

Pieter-Jan Volders, Justine Nuytens, Pieter Mestdagh and Jo Vandesompele

Contributions: The candidate contributed in part to design of the work, contributed to and supervised data acquisition and performed data analysis and interpretation. The candidate drafted the manuscript.

POTENT ANTISENSE OLIGONUCLEOTIDE SELECTION FOR LNCRNA KNOCKDOWN

PIETER-JAN VOLDERS, JUSTINE NUYTENS, PIETER MESTDAGH AND JO VANDESOMPELE

ABSTRACT

In this work, the potential of antisense oligonucleotides (ASOs) for transient knockdown of long non-coding RNAs (lncRNAs) is evaluated. Several features of both the oligo and the target RNA are examined for their effect on the observed knockdown in an experiment comprising 90 oligos targeting 6 human lncRNAs. Of these features, the affinity of the oligo for the target, the secondary structure of the target RNA and dimerization of the oligos are statistically significantly correlated to the knockdown efficacy. We trained a generalized additive model on these features that shows 78% accuracy in predicting functional ASOs. In addition, the effect of two nucleic acid analogs (2'-O-methyl and locked nucleic acid) is evaluated. The use of nucleic acid analogs decreases the required dose and increases the stability of the knockdown over time. Furthermore, the design and evaluation of non-targeting controls oligos is discussed. Together, this research shows that ASOs can be useful and efficient tools for studying the role of lncRNA, and we propose a strategy for the selection and evaluation of ASOs.

INTRODUCTION

Advances in sequencing technology have unveiled extensive and genome-wide transcription outside of known protein-coding loci. Transcripts (>200 nt) with poor coding potential comprise the majority of this transcription and have given rise to a new group of RNAs termed long non-coding RNAs (lncRNAs)[1-3]. Although the function of various lncRNAs has been studied in detail, the great majority remains to be fully characterized[4].

Perturbation of gene expression is an important aspect of functional genomics research and antisense-based tools have proven to be valuable for this purpose. From the lab, where they are used to study the function of the target gene, they are finding their way to the clinic, where overexpression of disease associated genes is attenuated by administration of antisense drugs[5]. Oligonucleotides with sequence complementarity to a known target can inhibit the function of the

target via three mechanisms: RNA interference (RNAi), ribonuclease H (RNase H) mediated degradation and steric hindrance[6].

RNase H is an endonuclease that specifically cleaves DNA/RNA duplexes and can be trigged by small DNA molecules called antisense oligonucleotides (ASOs) with sequence complementarity to endogenous RNA. To improve the stability and affinity of antisense oligonucleotides, a wide range of modifications have been developed, giving rise to what is often referred to as "second-generation" oligonucleotides. The improved performance of these ASOs drastically decreases the required dose and number of applications to achieve stable knockdown. These modifications can be classified as modifications of the backbone, sugar modifications and the less common base modifications[7]. The most popular and readily available modification of the backbone is the replacement of the phosphodiester bond with a phosphorothioate (PS) linkage. Full PS-DNA oligos show an improved nuclease resistance, while retaining the ability to induce RNase H mediated cleavage[8]. A broad range of sugar modifications has been developed and many have been shown to improve the efficacy and stability of ASOs[7,9]. Commonly used sugar modifications include 2'-O-Methyl (2'OMe), 2'-O-Methoxyethyl (2'MOE) and locked nucleic acid (LNA)[10]. Many of these sugar modifications[9] are often used in chimeric gapmer conformations where two regions consisting of modified residues flank a central part of unmodified nucleotides. Gapmers have an improved affinity and stability compared to PS-DNA backbone oligos without inhibiting RNase H cleavage[9]. Although RNase H activity is generally desirable, in some cases better results are achieved when fully modified oligos are used instead. These molecules will provide steric hindrance and prevent interaction with other macromolecules or prevent correct processing of the primary RNA transcript.

Antisense strategies that show great effectiveness to silence or functionally impair protein-coding genes may be unsuitable for lncRNAs. While oligos that interfere with ribosome binding are very powerful for mRNA silencing[11], they have poor potential for use against lncRNAs as lncRNAs are by definition not translated to protein. Furthermore, the great majority of lncRNAs are nuclear retained[12] and many have described functions in the nucleus[13,14]. This requires oligonucleotides that can efficiently pass the nuclear membrane and trigger a nuclease that functions in the nucleus.

Until now, researchers have predominantly used small interfering RNAs to achieve transient knockdown of lncRNAs. ASOs however, have several characteristics that make them superior candidates for silencing lncRNAs. Efficient silencing of nuclear retained transcripts has been reported using ASOs on several occasions[15,16]. A potential explanation is that while siRNAs are administered as double stranded RNA molecules, ASOs are of single stranded nature. This makes them smaller than siRNAs and improves delivery to the nucleus, where the great majority of lncRNAs is found[17]. In

contrast to siRNAs, ASOs can mediate knockdown when targeted to intronic sequences as well[18]. As lncRNAs are not translated and the mature RNA transcript is also the functional form of the gene; intronic ASOs prevent formation of the functional transcript and as such improve the chances of observing a phenotype. The lack of a second strand has other advantages as well, for instance it reduces the risk of off-target effects[17].

Successful application of ASOs in vivo in the lncRNA research field has been reported. For instance, ASOs have been used successfully to target the oncogenic lncRNA MALAT1. In one specific study, subcutaneous injection of ASOs in mice effectively inhibited MALAT1 expression in the tumor tissue and decreased lung metastasis[19].

While ASOs have promising features for lncRNA knockdown, open source tools for their design are currently scarce. Here, we develop a design tool and evaluate the efficacy of several chemical modifications.

METHODS

ASO SYNTHESIS

All oligos except LNA gapmers are synthesized by Integrated DNA Technologies (IDT). LNA gapmers are synthesized by Exiqon. All gapmers consist of a fully PS backbone and have nucleic acid analogs at position 1, 2, 3, 14, 15 and 16 (3-10-3 gapmers). Freeze-dried oligos are resuspended in nuclease-free water (Sigma).

CELL CULTURE, TRANSFECTION AND QUANTIFICATION OF KNOCKDOWN

Following trypsinization and dilution, HEK-293T cells were cultured in 96-well plates at 10 000 cells/well density (90 µl). Transfection with DharmaFECT 2 (Dharmacon) is performed 24 hours after seeding. 10 µl of transfection solution is added resulting in a final DharmaFECT 2 concentration of 4% and ASO concentration of 100 nM. Control samples are treated with transfection solution without ASO (reagent only) or not treated at all (cells only).

SK-MEL-28 cells are cultured in 100 µl (10 000 cells/well) and are transfected using Lipofectamine (Life Technologies). 50 µl transfection solution is supplemented after 24 hours so that the Lipofectimine concentration is 1.7% and the ASO concentration 100 nM.

Cell lysis and RNA extraction is performed using the SingleShot SYBR Green Kit (Bio-Rad) according to the manufacturer's instructions. In brief, cells were washed with 125 µl of Ca2+- and Mg2+-free phosphate buffered saline (PBS) and subsequently lysed with 50 µl of SingleShot Cell Lysis Buffer

containing DNase. Lysis reactions were incubated 5 min at room temperature followed by 5 min at 75 °C. cDNA synthesis was carried out on 4 µl of cell lysate in a total volume of 20 µl (20%) using the iScript Advanced cDNA Synthesis Kit for RT-qPCR that is supplied with the SingleShot SYBR Green Kit (both from Bio-Rad).

RT-qPCR

Calculation of normalized relative expression levels was done using the qbase+ software version 2.6 (Biogazelle). Normalization was performed using three stably expressed reference genes (UBC, TBP and YWHAZ). The normalized relative quantities are then scaled to the appropriate control sample to obtain the knockdown percentage. Downstream analysis and data visualization was done using the open source statistical environment R (version 3) and third party modules (plyr, ggplot2, car, reshape2, mgcv, ROCR).

ASO PARAMETER CALCULATION

The affinity of the oligo for the target is estimated by the Gibbs free energy (ΔG) of the hybridization reaction. To calculate the standard Gibbs free energy at 37 °C (ΔG^0_{37}), a custom implementation the nearest neighbor model for nucleic acid hybridization is applied. To this purpose, published nearest-neighbor parameter sets are being used: SantaLucia et al. 2004[20] for DNA/DNA hybridization, Sugimoto et al 1995[21] for DNA/RNA hybridization and Owczarzy et al 2011[22] LNA/DNA hybridization.

To assess the accessibility of the target RNA, the RNAplfold program from the ViennaRNA package (version 2.1.6) is used. Different window sizes (-W), maximum allowed separation of the pairs (-L) and averaging windows (-u) are tested. The accessibility profile is extracted form the _lunp output file.

Oligo dimerization and self-folding are evaluated with respectively the hybrid-min and the quikfold algorithms from the UNAFold package (nucleic acid type is set to DNA).

In order to filter out non-specific oligos, all sequences are aligned to the reference genome using the short read aligner bowtie (version 0.12.7). The used reference genomes are hg19 (GRCh37) for human and mm10 (GRCm38) for mouse. Only alignments without mismatches are considered, oligos with more than one exact match in the reference genome are regarded as non-specific. For non-targeting controls however, alignments up to one mismatch are considered.

MODEL BUILDING AND PARAMETER SELECTION

Efficacy of the oligo is modeled as a function of the different oligo parameters. To account for the distinct origin of the parameters, a non-parametric method called generalized additive models was chosen. All calculations are performed in R (version 3.1.2) and the package mgcv (version 1.8). To

assess if a model with a certain parameter set improves over a model with a different parameter set, the anova.gam function in the mgcv package is used. The quality of the predictions is evaluated using cross-validation and the ROCR package (version 1.0).

NON-TARGETING CONTROLS

Non-targeting controls (NTC) are designed by permuting the sequence of functional oligos until a sequence is obtained with no perfect homology to the reference genome. NTC sequences should have at least 2 mismatch nucleotide compared to the genome sequence.

CELL VIABILITY

Cell viability is evaluated using the CellTiter-Glo luminescent cell viability assay available from Promega. Cells are cultured in 96-well plates as described before. 48 hours after transfection, CellTiter-Glo reagent is supplemented to the wells and the BMG FLUOstar OPTIMA microplate reader is used for read-out.

MRNA EXPRESSION PROFILING

To assess off-target effects, transcriptome-wide gene expression profiling is performed using an in-house developed gene expression array based on the Agilent SurePrint platform. This array measures both lncRNA and mRNA expression. The mRNA probe content is based on the SurePrint G3 Human Gene Expression 8x60K v2 Microarray while the probes for lncRNA are designed using LNCipedia 2.1[4] (http://www.lncipedia.org) as a reference. RNA is extracted using the miRNeasy Micro Kit (Qiagen) and labeled with Agilent's Low Input Quick Amp WT Labeling Kit according to the manufacturers protocol. VSN normalization and further analysis of the expression data is performed with the Bioconductor Limma package (version 3.20.9) in R (version 3.1.2). The Benjamini Hochberg method is used to correct for multiplicity when significantly different expression is assayed.

RESULTS AND DISCUSSION

ANTISENSE OLIGONUCLEOTIDES ARE EFFECTIVE IN SILENCING LNCRNA EXPRESSION

To assess the potential of PS-modified ASOs for knockdown of lncRNA transcripts and to generate empirical data for modeling (see further), a screening experiment was performed. We randomly selected 90 specific 16-mers covering 6 lncRNA genes. Oligonucleotides with multiple occurrences in the reference genome and as such possible off-target effects were rejected. Target lncRNA expression was measured with RT-qPCR in the treated samples and compared to non-targeting control oligos, untreated cells and cells treated with transfection reagent only. Although the success rate differs between the different transcripts, knockdown > 50% is observed with at least some

probes for every transcript (Figure 1). These results establish ASOs as a valuable tool for in vitro lncRNA expression modulation.

PARAMETERS THAT PREDICT KNOCKDOWN EFFICACY

In order to evaluate the relative contribution of different oligo parameters on knockdown efficiency and to identify a predictive parameter set for upfront selection of potent oligos, we used the dataset from the screening experiment for modeling. Thermodynamic properties of the oligo – target interaction have been established as powerful predictors of oligo performance[23]. Therefore, the significance of the Gibbs free energy (ΔG) of oligo – target annealing, oligo dimerization and oligo self-interaction was assessed using generalized additive modeling (GAM), a non-parametric expansion of generalized linear modeling (GLM). GAM confirmed that the ΔG of oligo – target annealing and the ΔG oligo dimerization are significant predictors for the observed knockdown (Figure 2). Oligo self-interaction was the only non-significant parameter and was therefore excluded from further analyses. Next, different versions of the parameters were tested to see if they could improve the model. For the ΔG of oligo – target annealing, two nearest neighbor parameter sets were compared. The parameter set from SantaLucia et al. 2004[20] for DNA/DNA hybridization was found to have a higher predictive value than the Sugimoto et al. 1995[21] set for DNA/RNA hybridization.

Since each transcript has a characteristic secondary structure, one can assume that some regions of the RNA are less accessible to ASO molecules due to intramolecular base pairing. Indeed, for siRNA the importance of the secondary structure of the target site has been documented[24]. The *in silico* assessment of RNA structure is not a trivial task, and the best results for large RNA fragments are obtained when the prediction is limited to small regions[25]. The RNAplfold algorithm[26] computes the probability that a chosen region is free from base pairing and hence available for ASO binding. This probability is referred to as the accessibility of the target site. Using GAM, the accessibility was found to be a significant predictor (Figure 2). Three different parameters in the RNAplfold algorithm were optimized to obtain the highest predictive value. This was achieved by limiting the window size and maximum allowed separation of the pairs to 70 bp. While it seems contradictory that a smaller window size results in an improved accuracy, RNA structure prediction algorithms are prone to false positive interactions on large stretches of RNA and using a sliding window approach is recommended[25]. To further improve the predictive value, the accessibility is averaged over several bases. The third parameter that was optimized determines the size of the window in which the accessibility is averaged. The optimal size was found to be 12 nucleotides. As expected, a high accessibility is associated with more efficient knockdown of the target and the region around potent binding sites shows a significantly higher accessibility (Figure S1).

Of the 90 tested PS-ASOs, 35 target an intron of the transcript. The effect of the ASO position (intronic vs. exonic) is not found to be a significant predictor in this analysis. Several intronic ASOs result in significant knockdown of their target, confirming that ASOs have the capacity to function before splicing. This has important consequences for ASO design, as it widens the search space and increases the number of potential target sites.

The potential of the GAM for upfront selection of oligos is evaluated using a modified leave-one-out cross validation. For all ASOs that target one of the six transcripts, the knockdown is predicted using a GAM build using the ASOs targeting the five remaining transcripts. Although the prediction accuracy differs between the different transcripts (Figure S2), the model shows an overall accuracy of 78% (Figure S3).

DIFFERENT NUCLEIC ACID ANALOGS IMPROVE DURABILITY OF THE KNOCKDOWN

Natural single stranded DNA molecules are rapidly degraded by nucleases and as such show poor stability and knockdown efficacy. These features can be improved dramatically through specific modifications of the DNA structure[8]. Currently a wide range of modifications is being used in research and clinical applications. In this study, the potency of two popular sugar modifications, namely 2'-O-methylation (2'-O-me) and locked nucleic acids (LNA), were evaluated. In the chosen gapmer configuration, the first 3 and last 3 nucleic acids are substituted with 2'-O-me or LNA nucleic acid analogs (3-10-3 conformation). Four PS-DNA ASOs were selected and tested alongside their 2'-O-me and LNA gapmers. Using two different concentrations (10 nM and 100 nM), the knockdown was measured at 5 time points (12, 24, 48, 72 and 96 h post-transfection). Although there is little difference in the measured knockdown at the early time point (24 h) using a 100 nM concentration, it is clear that the use of nucleic acid analogs has a great impact on the knockdown at later time points and lower concentrations (Figure 3, S4). While the effect of pure PS-DNA oligos decreases over time, LNA gapmer mediated knockdown remains stable over the course of the experiment. In addition, it is apparent that, especially for the LNA gapmers, even at the 10 nM concentration, substantial knockdown is observed. This either suggests a higher affinity of the ASO to the RNA or improved activation of RNase H. Indeed, an improved affinity has already been reported for LNA oligos[27]. In addition, this can be examined using nucleic acid thermodynamics as the Gibbs free energy (ΔG) of the hybridization reaction can serve as a measure for the affinity of the ASO to the target. When comparing the ΔG for all possible 16-mers in a given sequence, the ΔG for LNA gapmer – DNA hybridization is on average 7 kJ/mol lower compared to DNA – DNA hybridization (Figure S5). This is a considerable difference when taking into account the ΔG range required for efficient knockdown and supports the hypothesis that a higher affinity is a valuable explanation for the observed difference at lower concentrations.

NON-TARGETING CONTROLS

Given that sequence independent effects of modified DNA are well-established[28], high quality non-targeting controls are indispensable to evaluate the true effect of a targeting oligo. Non-targeting oligos are developed by random permutation of the sequence of functional oligos until a sequence is obtained with no homology to the reference genome. Three different NTCs were designed in this way. To verify that the selected NTCs lack sequence specific effect on the gene expression, transcriptome-wide gene expression profiling was performed using a custom expression microarray. NTC treated samples were compared among each other and untreated samples. No significant up or down regulation could be detected, pointing to a lack of sequence specific off-target effects (Figure S6). To elucidate the sequence independent effects related to the chemistry, PS-DNA NTCs were compared to their 2'-O-methyl gapmer counterparts. Again, no significant differences were detected.

ASO TRANFECTION INDUCES SEQUENCE INDEPENDENT REDUCTION IN CELL VIABILITY

In cancer research, genes are often silenced to probe for their relevance in pathogenesis. As such, the intended cellular phenotype upon knockdown is altered cell viability. To determine whether NTC ASO transfection has impact on cell viability, CellTiter-Glo luminescence was measured 48h post transfection. Two different cell lines were transfected with the NTCs described earlier to elucidate the sequence independent effects. In addition, two different transfection reagents were tested in parallel. Both transfection reagents reduce the cell viability of the two cell lines, although SK-MEL-28 seems more resistant to this toxicity than HEK-293 (Figure 4). In HEK-293, a more pronounced reduction in viability is observed when both transfection reagent and ASO are present. This increased toxicity is observed for all tested ASOs and is thus sequence and chemistry independent. The additional reduction in viability cannot be observed in the SK-MEL-28 cell line. Repetition of this experiment confirmed these findings. Together, these results show the importance of validated NTCs when ASOs are used in phenotyping experiments.

CONCLUSIONS

ASO are successfully applied to reduce the intracellular RNA concentration of their target lncRNA. Random selection of ASOs with perfect complementarity to a lncRNA target of interest is a viable option for ASO selection although the success rate is low; only few random ASOs show a good knockdown performance. Upfront selection of functional ASOs with high potential should reduce the number of oligos to be tested and the cost of the experiment. To test this hypothesis, several parameters of the ASO and its target-site were evaluated for their influence on the potency of the

ASO. The Gibbs free energy (ΔG) of oligo – target annealing, the accessibility of the target site and the ΔG of oligo dimerization were found to be significant predictors of the obtained knockdown. Using these features, a generalized additive model was trained and tested to predict ASO potency in silico. Although PS-DNA oligos showed good knockdown up until two days post transfection, a longer lasting effect can be achieved using 2'-O-me or LNA gapmers. Since sequence independent effects of ASOs are described in literature and confirmed in this work, it is important that high quality NTCs are evaluated in parallel. A strategy to select valid NTC sequences is proposed and 3 NTCs were designed and tested. In conclusion, this work provides lncRNA research community with several tools and strategies that empower them to apply ASOs for the knockdown of their lncRNA of interest.

AVAILABILITY

A web-interface has been developed that allows users to design ASOs for an RNA of interest. All potential ASOs are evaluated using the described GAM model. In addition, non-specific ASOs are automatically removed. The web-interface is available at https://brenner.ugent.be/~janckaer/aso-design (provisory url).

REFERENCES

1. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics* **9,** (2013).
2. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* (2011). doi:10.1101/gad.17446611
3. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* **22,** 1775–1789 (2012).
4. Volders, P.-J. *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Research* **43,** D174–80 (2015).
5. Tamm, I., Dörken, B. & Hartmann, G. Antisense therapy in oncology: new hope for an old idea? *The Lancet* **358,** 489–497 (2001).
6. Kole, R., Krainer, A. R. & Altman, S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat Rev Drug Discov* (2012). doi:10.1038/nrd3625
7. WATTS, J., DELEAVEY, G. & DAMHA, M. Chemically modified siRNA: tools and applications. *Drug Discovery Today* **13,** 842–855 (2008).
8. Milligan, J. F., Matteucci, M. D. & Martin, J. C. Current concepts in antisense drug design. *J. Med. Chem.* **36,** 1923–1937 (1993).
9. Yu, R. Z., Grundy, J. S. & Geary, R. S. Clinical pharmacokinetics of second generation antisense oligonucleotides. *www.expertopin.com/emt* **9,** 169–182 (2013).
10. Grijalvo, S., Aviñó, A. & Eritja, R. Oligonucleotide delivery: a patent review (2010 – 2013). *www.expertopin.com/etp* (2014). doi:10.1517/13543776.2014.915944
11. Dias, N. & Stein, C. A. Antisense Oligonucleotides: Basic Concepts and Mechanisms. *Mol*

Cancer Ther **1,** 347–355 (2002).
12. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
13. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).
14. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 11667–11672 (2009).
15. Ideue, T., Hino, K., Kitao, S., Yokoi, T. & Hirose, T. Efficient oligonucleotide-mediated degradation of nuclear noncoding RNAs in mammalian cultured cells. *RNA* **15,** 1578–1587 (2009).
16. Wheeler, T. M. *et al.* Targeting nuclear RNA for in vivo correction of myotonic dystrophy. *Nature* **488,** 111–115 (2012).
17. Li, C. H. & Chen, Y. Targeting long non-coding RNAs in cancers: Progress and prospects. *The International Journal of Biochemistry & Cell Biology* **45,** 1895–1910 (2013).
18. Vickers, T. A. *et al.* Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *Journal of Biological Chemistry* **278,** 7108–7118 (2003).
19. Gutschner, T. *et al.* The non-coding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research* **73,** canres.2850.2012–1189 (2012).
20. SantaLucia, J., Jr. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33,** 415–440 (2004).
21. Sugimoto, N. *et al.* Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry* **34,** 11211–11216 (1995).
22. Owczarzy, R., You, Y., Groth, C. L. & Tataurov, A. V. Stability and Mismatch Discrimination of Locked Nucleic Acid–DNA Duplexes. *Biochemistry* **50,** 9352–9367 (2011).
23. Matveeva, O. V. *et al.* Thermodynamic criteria for high hit rate antisense oligonucleotide design. *Nucleic Acids Research* **31,** 4989–4994 (2003).
24. Tafer, H. *et al.* The impact of target site accessibility on the design of effective siRNAs. **26,** 578–583 (2008).
25. Bernhart, S. H. & Hofacker, I. L. Think global, fold local. *tbi.univie.ac.at* at <https://www.tbi.univie.ac.at/~ulim/berniebsvposter.pdf>
26. Bernhart, S. H., Hofacker, I. L. & Stadler, P. F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22,** 614–615 (2006).
27. Singh, S. K., Koshkin, A. A., Wengel, J. & Nielsen, P. LNA (locked nucleic acids): synthesis and high-affinity nucleic acid recognition. *Chem. Commun.* 455–456 (1998). doi:10.1039/A708608C
28. Stein, C. A. Does antisense exist? *Nat Med* **1,** 1119–1121 (1995).

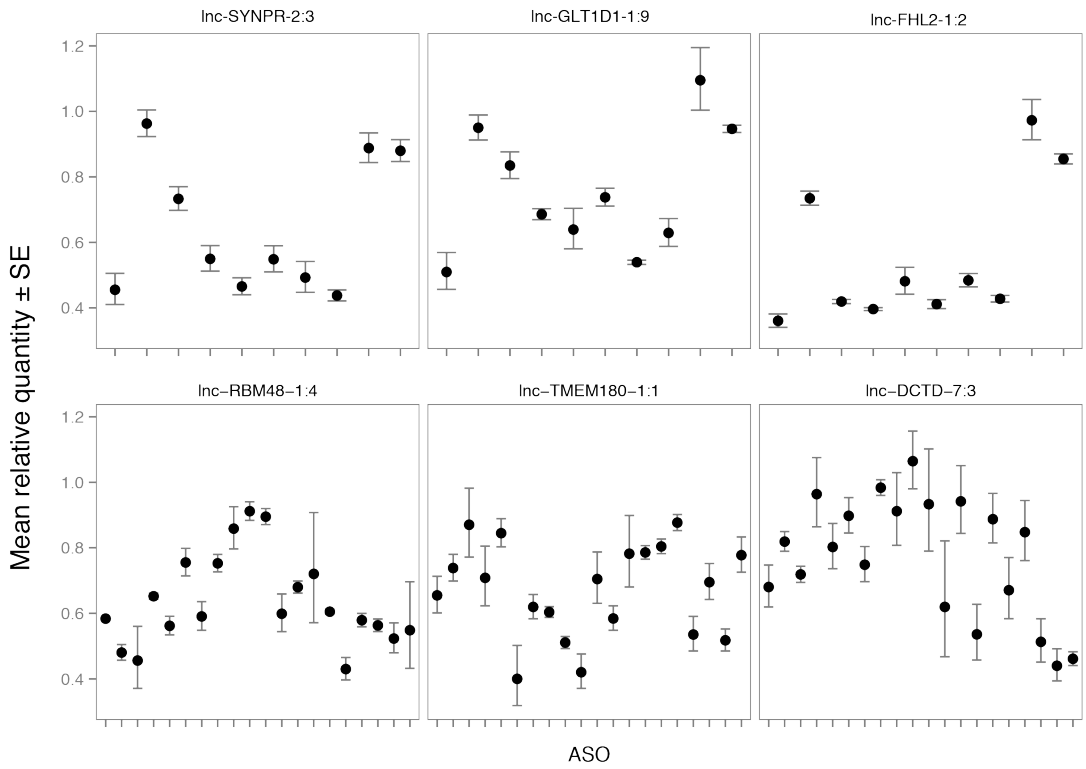


Figure 1: Knockdown at 48 h post transfection for 90 randomly selected PS-ASOs targeting 6 lncRNA transcripts. RNA concentration is measured using RT-qPCR and scaled relative to the sample treated with transfection reagent only. Mean knockdown is plotted along with the standard error calculated from the biological replicates (n=3). Although the results differ between the different lncRNAs, for every lncRNA at least one ASO results in a knockdown > 50%. Oligos are ordered from 5' to 3' along the transcript.

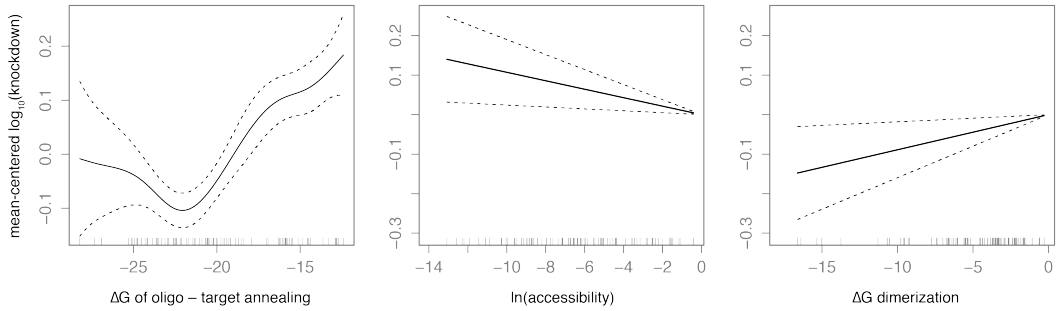


Figure 2: Three features were found to be significant predictors of the knockdown when assessed using a generalized additive model (GAM): the Gibbs free energy (ΔG) of oligo – target annealing (left), the accessibility of the target site (middle) and the ΔG of oligo dimerization (right). While the first feature is being modeled with a GAM spline, the others have been reduced to a simple linear effect.

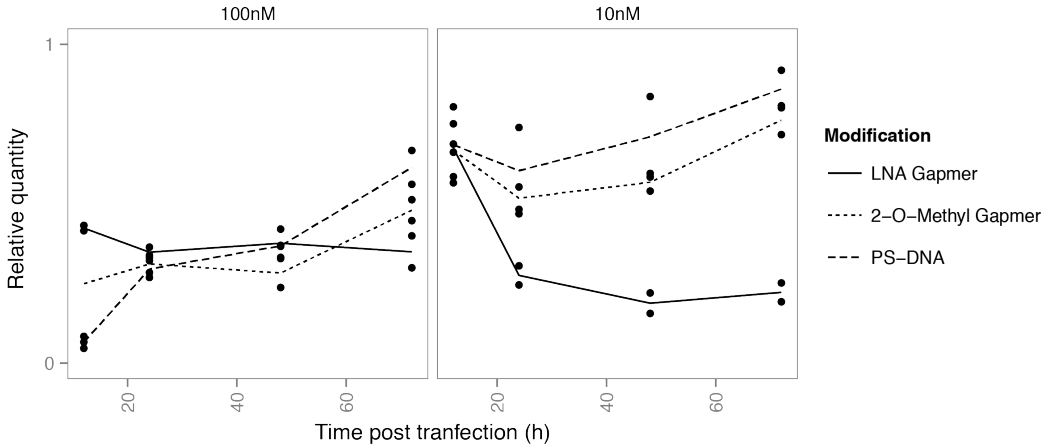


Figure 3: Different nucleic acid analogs increase ASO potency and stability. The same oligonucleotide sequence yields different results when different modifications are being used. Although the differences are subtle for the 100 nM concentration (left) compared to the 10 nM condition (right), the locked nucleic acid (LNA) gapmers (3-10-3 conformation) show substantial improvement of stability of the knockdown.

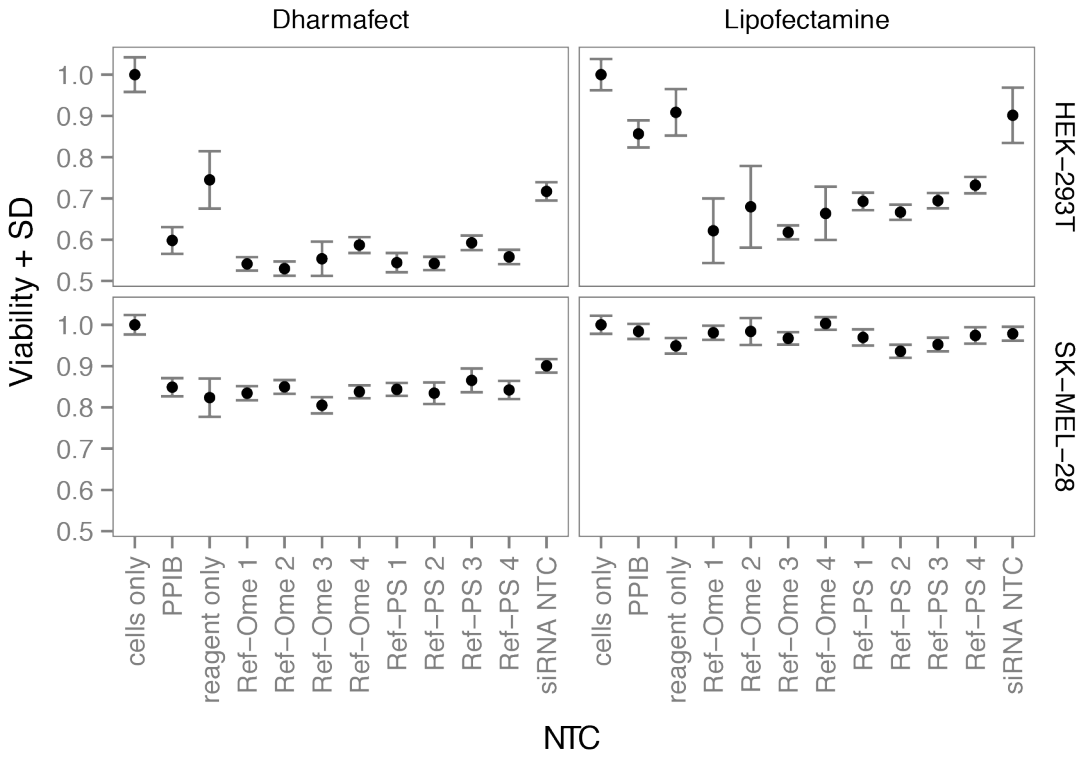


Figure 4: CellTiter-Glo luminescence relative to the non-treated sample as a measure for relative cell viability. The viability is reduced upon transfection with ASOs in a sequence and chemistry independent manner.

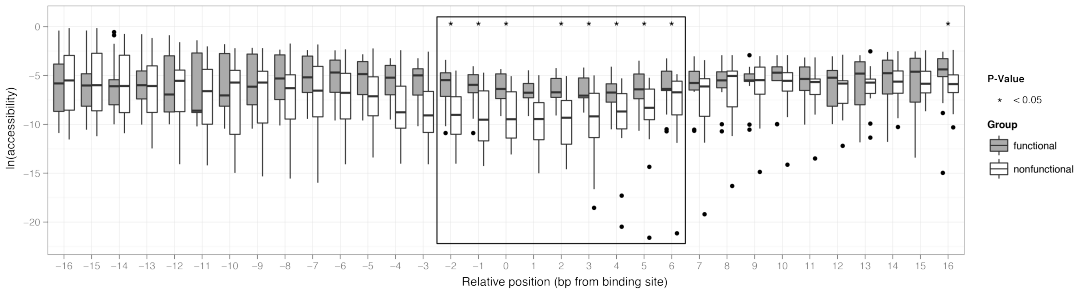


Figure S1: Local RNA accessibility in the region of the target site differs significantly (p<0.05, *) between functional (>50% knockdown) and non-functional ASOs.

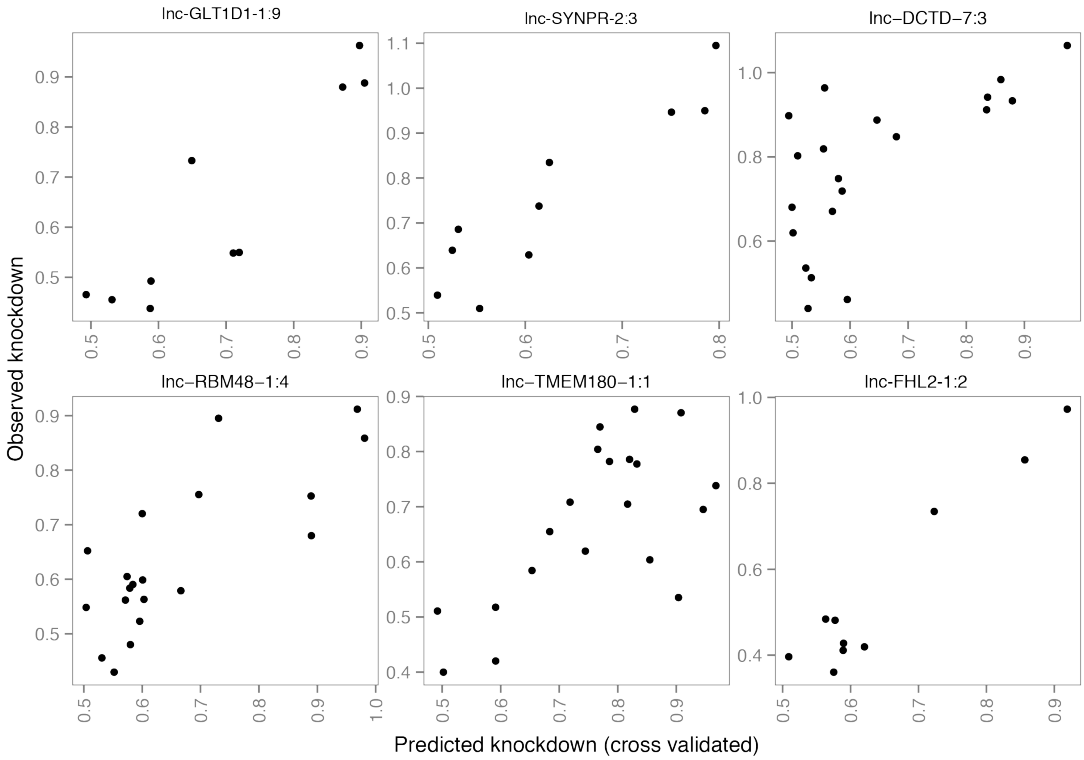


Figure S2: Cross-validation of the predicted knockdown

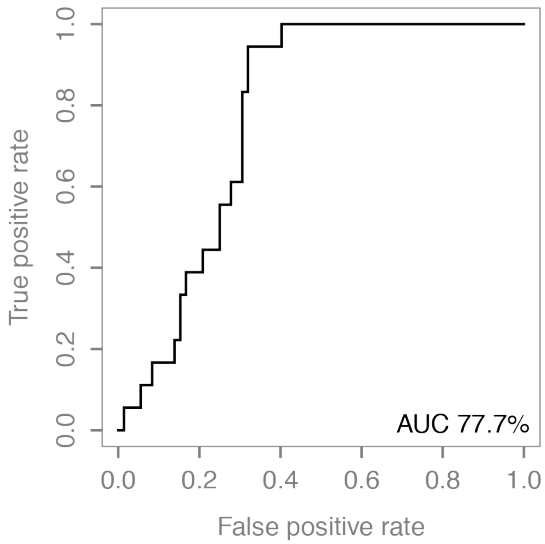


Figure S3: Receiver operating characteristic (ROC) curve representing the ability of the GAM to distinguish functional (>50% knockdown) from non-functional ASOs.

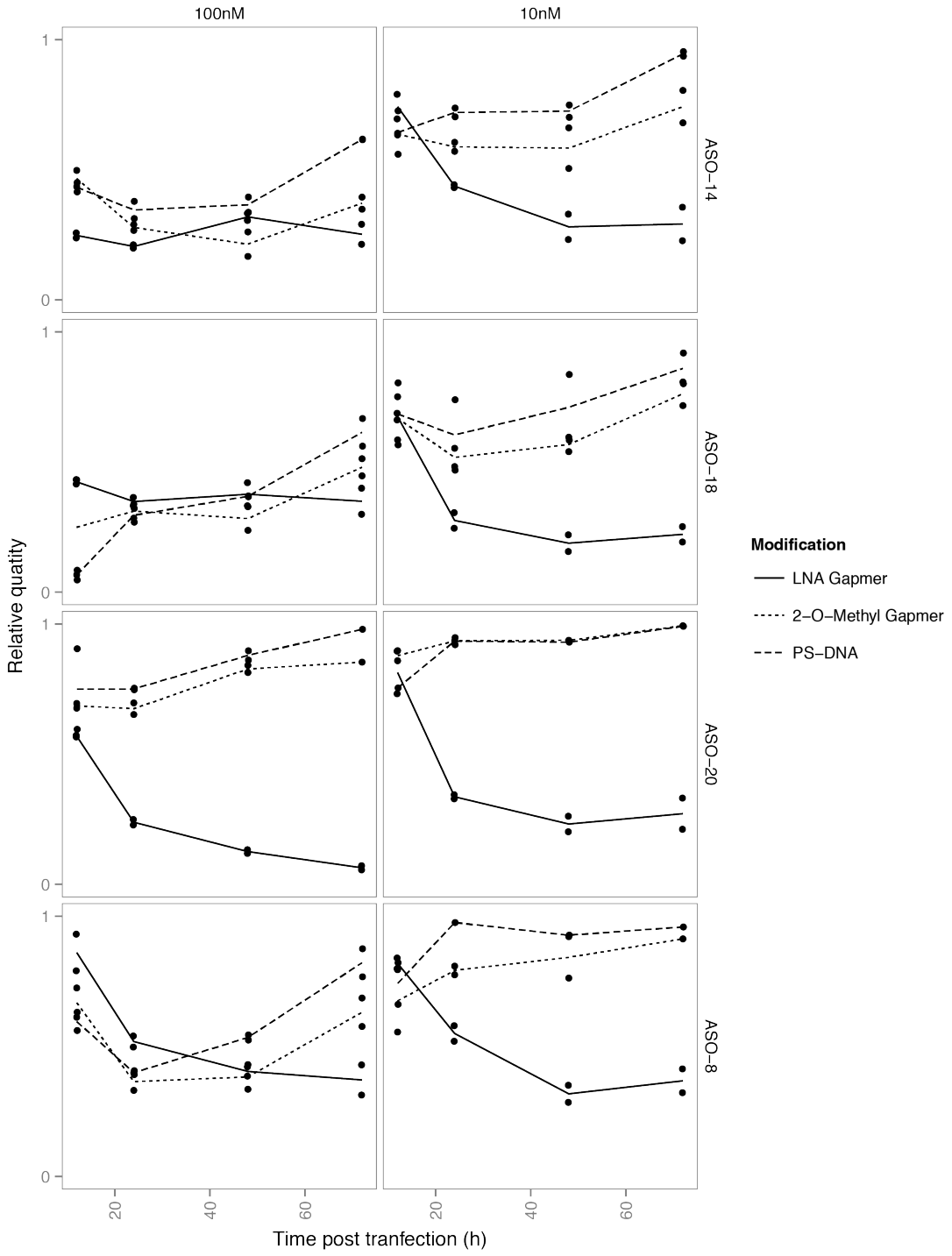


Figure S4: Nucleic acid analogs increase ASO potency and durability, shown for 4 different oligonucleotides. Normalized relative quantities are rescaled to the NTC condition.

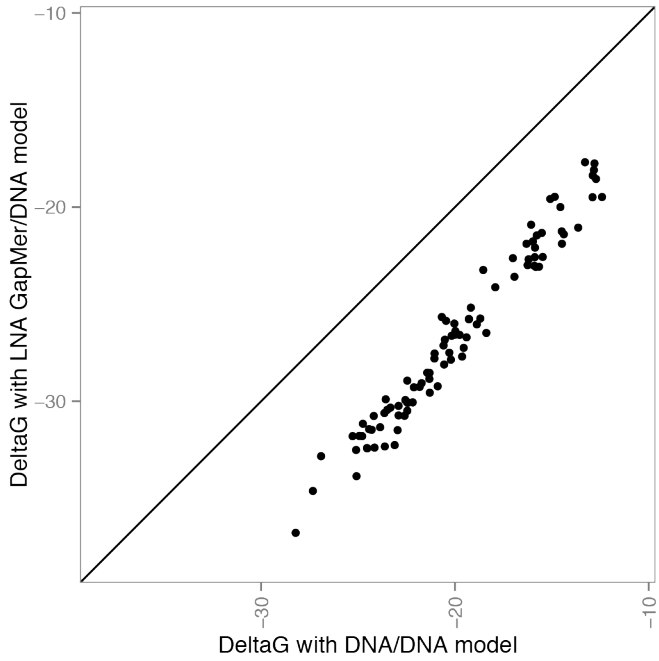


Figure S5: Comparison of the calculated Gibbs free energy (ΔG) of DNA-DNA annealing (SantaLucia et al 2004 parameters) and LNA gapmer-DNA annealing (Owczarzy et al 2011 parameters) for a selection of 16-mers. Although the two are highly correlated, ΔG is on average 7 kJ/mol lower for the LNA gapmers.

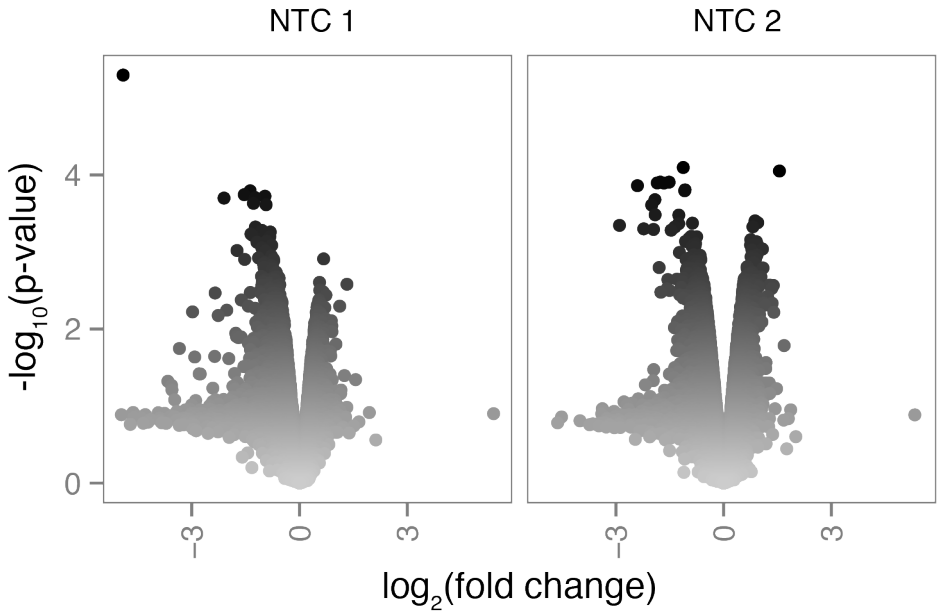


Figure S6: Volcano plot showing the expression of mRNAs and lncRNAs measured by expression array. When the FDR is taken into account (Benjamini Hochberg method), no p-value < 0.05 are observed.

SUPPLEMENTAL MATERIAL

Table S1: Sequences and chromosomal position (hg19) of the used ASOs

ID	target	sequence	chromosome	position
ENST_62_1	lnc-SYNPR-2:3	CCGTCCTGGGACAGCC	chr3	62974917
ENST_62_2	lnc-SYNPR-2:3	ATTTCCGTGTGGCTGA	chr3	62936177
ENST_62_3	lnc-SYNPR-2:3	GTGTCTGGGAGCCAAC	chr3	62939395
ENST_62_4	lnc-SYNPR-2:3	CGTCCTGGGACAGCCA	chr3	62974916
ENST_62_5	lnc-SYNPR-2:3	TGTCTGGGAGCCAACA	chr3	62939394
ENST_62_6	lnc-SYNPR-2:3	AGCAGATTCACTGGGC	chr3	62970735
ENST_62_7	lnc-SYNPR-2:3	GCCAGAAGGTCCTCAG	chr3	63099348
ENST_62_8	lnc-SYNPR-2:3	TACAAGAAGATACAAT	chr3	62974866
ENST_62_9	lnc-SYNPR-2:3	CTCTTATGCCTATAAG	chr3	62937682
ENST_62_10	lnc-SYNPR-2:3	ATTATATAAGAATCTC	chr3	63060671
ENST_25_1	lnc-GLT1D1-1:9	GGGCGCCGGATGCCCA	chr12	129596076
ENST_25_2	lnc-GLT1D1-1:9	GCGCCGGATGCCCACC	chr12	129596074
ENST_25_3	lnc-GLT1D1-1:9	GGCAGGGGGGCTGCAT	chr12	129595561
ENST_25_4	lnc-GLT1D1-1:9	GCTGAGGGCAGCGACG	chr12	129595539
ENST_25_5	lnc-GLT1D1-1:9	GGTCTCCCTCCGAGCA	chr12	129595628
ENST_25_6	lnc-GLT1D1-1:9	GTGGCCAGATGAGGGT	chr12	129597702
ENST_25_7	lnc-GLT1D1-1:9	TCCGTCAGAATGCACA	chr12	129596019
ENST_25_8	lnc-GLT1D1-1:9	GATAATAGAGCAACTC	chr12	129596099
ENST_25_9	lnc-GLT1D1-1:9	TTATATGGGAATTGGT	chr12	129597731
ENST_25_10	lnc-GLT1D1-1:9	CTAGTAAAGATTACTG	chr12	129596329
TCONS_59_1	lnc-FHL2-1:2	GCGTCCGTGAGCTGGG	chr2	106217644
TCONS_59_2	lnc-FHL2-1:2	CGGGGAACACACGCAC	chr2	106213655
TCONS_59_3	lnc-FHL2-1:2	CGGCTGGTGCAACAGG	chr2	106217615
TCONS_59_4	lnc-FHL2-1:2	AGGACATGAAGGCGGA	chr2	106217494
TCONS_59_5	lnc-FHL2-1:2	CAGCGTCCGTGAGCTG	chr2	106217642
TCONS_59_6	lnc-FHL2-1:2	GTGCTGAGCTGTGCAA	chr2	106213611
TCONS_59_7	lnc-FHL2-1:2	GGTGCAACAGGGGTCA	chr2	106217620
TCONS_59_8	lnc-FHL2-1:2	TCGTATATAAAATAAC	chr2	106215652
TCONS_59_9	lnc-FHL2-1:2	AAGAACTACGAATATT	chr2	106217348
TCONS_59_10	lnc-FHL2-1:2	CGGATGAATGTACTTT	chr2	106217506
lnc-RBM48-1	lnc-RBM48-1:4	GCGGCTCCCACATTCC	chr7	92546293
lnc-RBM48-2	lnc-RBM48-1:4	GAAACCAGCCAGGGGT	chr7	92484225
lnc-RBM48-3	lnc-RBM48-1:4	GCAGGGGTGAGACTTG	chr7	92485084
lnc-RBM48-4	lnc-RBM48-1:4	TCTCCTAGGTGTTGCA	chr7	92546228
lnc-RBM48-5	lnc-RBM48-1:4	GGCATATGATGCAGGG	chr7	92485094
lnc-RBM48-6	lnc-RBM48-1:4	CAGGGGTGAGACTTGA	chr7	92485083
lnc-RBM48-7	lnc-RBM48-1:4	CGGCTCCCACATTCCA	chr7	92546292
lnc-RBM48-8	lnc-RBM48-1:4	GGCGCCACGCCCAGTC	chr7	92500869
lnc-RBM48-9	lnc-RBM48-1:4	GCGTCTGGCAGGGGCC	chr7	92527466
lnc-RBM48-10	lnc-RBM48-1:4	TGGGCACGGCATGGGC	chr7	92510450
lnc-RBM48-11	lnc-RBM48-1:4	AGGCATCATCAGCGGC	chr7	92546304
lnc-RBM48-12	lnc-RBM48-1:4	TTTTCACGGTGTGGCA	chr7	92546447
lnc-RBM48-13	lnc-RBM48-1:4	CAGGCCCCCGGATGGC	chr7	92510226
lnc-RBM48-14	lnc-RBM48-1:4	GACATCCTTGGAGAGG	chr7	92485116
lnc-RBM48-15	lnc-RBM48-1:4	TGCGTGGGCTCTGCGA	chr7	92505293
lnc-RBM48-16	lnc-RBM48-1:4	CGAATAATAAAATTCC	chr7	92485009
lnc-RBM48-17	lnc-RBM48-1:4	TTAAGTATTATATGTC	chr7	92496222
lnc-RBM48-18	lnc-RBM48-1:4	CCTTACTTTAATATAA	chr7	92545541
lnc-RBM48-19	lnc-RBM48-1:4	TTTTGAGCTATCTAGG	chr7	92485049
lnc-RBM48-20	lnc-RBM48-1:4	GGGTAAGATTATAATA	chr7	92520172
lnc-DCTD-1	lnc-DCTD-7:3	GCTCCGGTTCAGGGCC	chr4	181985563
lnc-DCTD-2	lnc-DCTD-7:3	CTCGGCCAGCTTTGGC	chr4	181985549
lnc-DCTD-3	lnc-DCTD-7:3	GCCAGCTTTGGCTCCG	chr4	181985553
lnc-DCTD-4	lnc-DCTD-7:3	TGGATTCGCTTGTCTG	chr4	182076832

lnc-DCTD-5	lnc-DCTD-7:3	GCACGGGCGGCCCGCA	chr4	182068810
lnc-DCTD-6	lnc-DCTD-7:3	CAGTGATGGATTCGCT	chr4	182076826
lnc-DCTD-7	lnc-DCTD-7:3	GACCGTGCGCGGTGGC	chr4	181989099
lnc-DCTD-8	lnc-DCTD-7:3	CAGCACGGGCCGGGCT	chr4	182074409
lnc-DCTD-9	lnc-DCTD-7:3	CGTGGCCCAGTGCCGC	chr4	182073201
lnc-DCTD-10	lnc-DCTD-7:3	GCCTGTGGGGGCCGGT	chr4	182034362
lnc-DCTD-11	lnc-DCTD-7:3	GCCCGCTCTGGCAGGC	chr4	182054430
lnc-DCTD-12	lnc-DCTD-7:3	GCGGCAAGCCAGACGC	chr4	182059058
lnc-DCTD-13	lnc-DCTD-7:3	GGCTCCGGTTCAGGGC	chr4	181985562
lnc-DCTD-14	lnc-DCTD-7:3	GCCATCCAGTTGCTGC	chr4	181985612
lnc-DCTD-15	lnc-DCTD-7:3	CAGAATCTCCCCCAGC	chr4	181985531
lnc-DCTD-16	lnc-DCTD-7:3	GCATCATCATACATTA	chr4	181985391
lnc-DCTD-17	lnc-DCTD-7:3	AAGTTGCTAATCCTAT	chr4	182076781
lnc-DCTD-18	lnc-DCTD-7:3	GACTCTATATATATAG	chr4	182003503
lnc-DCTD-19	lnc-DCTD-7:3	TAGAGTATCTATTAAT	chr4	182025590
lnc-DCTD-20	lnc-DCTD-7:3	ATAATCGATATTATTT	chr4	182062869
lnc-1575-1	lnc-TMEM180-1:1	AGCTGAATGTTCGATT	chr10	104211015
lnc-1575-2	lnc-TMEM180-1:1	CCTTATTGTCTGCTGG	chr10	104211076
lnc-1575-3	lnc-TMEM180-1:1	AGTTGGGATGAGTTAT	chr10	104211634
lnc-1575-4	lnc-TMEM180-1:1	GTAACATTCTAGAGTG	chr10	104211773
lnc-1575-5	lnc-TMEM180-1:1	ACTTCACATAGCTATT	chr10	104212437
lnc-1575-6	lnc-TMEM180-1:1	GGTTCCTTATCTACTA	chr10	104212475
lnc-1575-7	lnc-TMEM180-1:1	GTTTTAATCATCCATA	chr10	104212715
lnc-1575-8	lnc-TMEM180-1:1	GACATCACCAGGGTGA	chr10	104212896
lnc-1575-9	lnc-TMEM180-1:1	CTATGACAGTGTTAGA	chr10	104213031
lnc-1575-10	lnc-TMEM180-1:1	ACTAATCACATAATGG	chr10	104215392
lnc-1575-11	lnc-TMEM180-1:1	GAGCAATTAATACTTA	chr10	104215476
lnc-1575-12	lnc-TMEM180-1:1	CAGTCCTTACAGCAGA	chr10	104215568
lnc-1575-13	lnc-TMEM180-1:1	CAACCTTACTGCCATC	chr10	104215695
lnc-1575-14	lnc-TMEM180-1:1	CCCATGACTGCCGGCC	chr10	104210056
lnc-1575-15	lnc-TMEM180-1:1	AGCGCGCGCGGGGCCC	chr10	104210454
lnc-1575-16	lnc-TMEM180-1:1	GCCCCGCCGCGCCACC	chr10	104210649
lnc-1575-17	lnc-TMEM180-1:1	CGGTCCCATGGGCCCC	chr10	104214240
lnc-1575-18	lnc-TMEM180-1:1	GGCAGGACCGCATCCC	chr10	104215959
lnc-1575-19	lnc-TMEM180-1:1	GTTAAGTCTTTCAGTC	chr10	104215833
lnc-1575-20	lnc-TMEM180-1:1	TCAGTAATACTGTAGG	chr10	104211360
NTC 1	NTC	ATGCGACCCCGCCGGA	#	#
NTC 2	NTC	GACACTATCGGACGAG	#	#

Table S2: Chromosomal position (hg19) of the transcripts used in this study

transcript	position
lnc-SYNPR-2:3	chr3:62936146-63099387
lnc-GLT1D1-1:9	chr12:129595538-129597839
lnc-FHL2-1:2	chr2:106209554-106227016
lnc-RBM48-1:4	chr7:92484223-92546488
lnc-DCTD-7:3	chr4:181985270-182076852
lnc-TMEM180-1:1	chr10:104209595-104216051

IV. DISCUSSION AND FUTURE PERSPECTIVES

This PhD is set in the era of emerging lncRNA research. Most lncRNA annotation is very recent; consequently, it is absent from popular platforms and databases. The development of novel tools is thus a prerequisite to study lncRNAs in cancer. Our work comprises both the development and the application of such tools. In this way, this thesis stands out as a unique combination of wet and dry lab research.

IV.1. CATALOGING THE UNKNOWN: CHALLENGES AND REMARKS

The rate at which new lncRNA transcripts are being reported in literature is unseen and a consequence of recent advances in deep RNA sequencing. Yet, the rate at which they find their way to the genomic reference databases such as Ensembl[1], RefSeq[2] and UCSC[3] is much lower as annotators struggle to keep up or hold a more conservative position. A broad collection of sequenced lncRNAs is nonetheless invaluable for high-throughput transcriptomics. We developed LNCipedia to fill this gap and provide extensive and meaningful lncRNA annotation to the community. With an average of over 100 visitors every day (Figure 5), about 500 emails to the authors regarding database exports, over 100 citations (Google scholar) for the original LNCipedia paper and numerous mentions on blogs and social media, we believe LNCipedia has proven to be a relevant resource for lncRNA researchers. While several other lncRNA databases such as NONCODE[4], LncRNA2Function[5], DIANA-LncBase[6] and many more have mushroomed in the past few years, LNCipedia holds a unique position as the largest publicly available human lncRNA resource. As a result, LNCipedia has been recommended as a primary resource in a 2014 review article on lncRNA databases due to its compromise between coverage and depth of annotations[7]. The catalogue of human lncRNAs however is far from complete as is our understanding of the transcriptome in general. In sharp contrast to our initial view on the human genome, we now know that most of the genome is transcribed into distinct RNA transcripts[8-10]. Even more, targeted RNA sequencing of selected genomic loci (CaptureSeq) revealed an immense uncharted complexity of the transcriptome. It uncovers extensive alternative splicing both in and outside of annotated loci including many rare isoforms[11,12]. As the greater majority of these

novel transcripts lack signs of coding potential, it is reasonable to assume that the (long) non-coding transcriptome will continue to grow over the coming years. The crucial question however is not on the number of transcribed non-coding genes but on the number of functional non-coding genes in the genome. Even though the number of lncRNAs with an experimentally derived function is steadily increasing, some people remain skeptical that functional lncRNA genes outnumber protein coding genes in the human genome[13]. Often, the lack of measurable sequence conservation forms the basis for their doubt. Recently, several research groups have studied lncRNA evolution and have each identified a subset of conserved lncRNAs[14-16]. Although their results do show that lncRNAs are more recent adaptations compared to most protein coding genes, they are clearly subject to selective pressure contrary to previous suggestions. In LNCipedia, we apply the concept of locus conservation[17] to aid the identification of lncRNA orthologs. Yet, locus conservation alone is not enough to prove conservation of the gene, let alone its function. In a future release of LNCipedia, we will therefore incorporate more relevant measures and datasets to tackle this problem. Since most research on lncRNA evolution is still in its infancy, the true extent of lncRNA conservation will likely unfold in the near future.

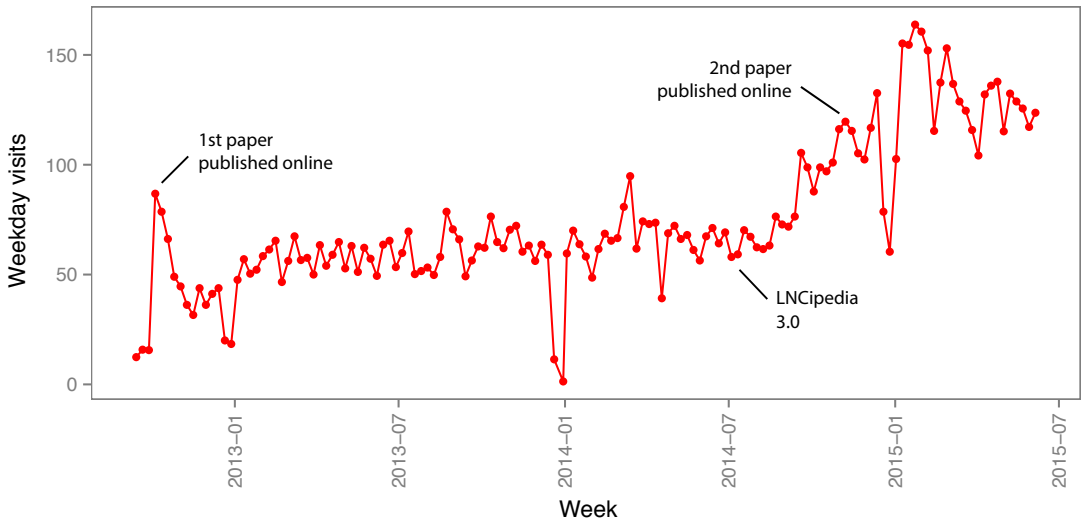


Figure 5: Average number of weekday visits for LNCipedia.org. The number of visits shows a steady incline with visible peaks in the weeks when two papers were published.

LNCRNA CODING POTENTIAL

The ability of a novel RNA sequence to encode a protein can be assessed by *in silico* prediction programs (*vide supra*). In LNCipedia, we include several of these programs, namely CPC[18], HMMER[19] and PhyloCSF[20], and our analyses show poor coding potential for the greater majority of lncRNAs. Nevertheless, a substantial fraction of LNCipedia transcripts show elevated coding potential compared to benchmarking datasets. It is important to note however that while benchmarking datasets are typically composed of intergenic non-coding transcripts, meaning they do not overlap with protein coding sequence, a considerable fraction of lncRNA transcript overlaps protein coding genes, in sense or antisense. As it is currently unclear what fraction of LNCipedia corresponds to these kinds of transcripts, we will look into subclassification of lncRNAs according to their relative position to protein coding genes in a future release of LNCipedia. Further analysis on the coding potential of the intergenic subset can provide a better insight in the true extent of putative protein coding genes. While several research groups have turned to ribosome profiling as a method to distinguish protein coding from non-coding sequence, interpretation of these data is not without pitfalls and complications. Consequently, distinct authors have come to contrary conclusions when examining ribosome occupancy on lncRNAs[21,22]. Therefore, we and other groups have used shotgun proteomics data to detect the putative products of lncRNA ORFs. Our results show that only a minute fraction of LNCipedia lncRNAs (<1.5%) bears ORFs that produce detectable peptides. This is very much in agreement with similar efforts, where also only low numbers of novel peptides were found[23-25]. These low numbers contradict the pervasive translation of lncRNAs that is reported by some ribosome profiling studies[22]. In our commentary paper, we have explored the possibility that this discrepancy is due to the current limitations of mass spectrometry. As such, we have examined several possible explanations why the putative proteins encoded by lncRNA ORFs are less detectable than established protein coding genes. We have come to the conclusion that the most plausible explanation remains that the greater majority of lncRNAs are not translated and their ORFs have thus other functions, if any, than a protein coding one. Since ribosome profiling has gained much attention in its short existence, we are convinced that more research on the interpretation of

ribosome occupancy on (non-) coding ORFs is ongoing and over the next years it will become clear how to use and interpret this kind of data.

The number of lncRNAs that produce detectable peptides according to our PRIDE reprocessing pipeline increased drastically (from 14 to 2,040) in the LNCipedia 3.1 release. This increase can be attributed to both the increase in LNCipedia entries and the increased number of PRIDE experiment evaluated. In addition, the sources that were added to LNCipedia for the 3.1 release show an increased coding potential when assessed with PhyloCSF. The putative novel peptides that are found using the reprocessing pipeline are available on the LNCipedia website for download.

Our work on lncRNA coding potential did not go unnoticed and in the spring of 2015, we were invited to participate in an invitation-only event organized by the European Bioinformatics Institute and the Wellcome Trust. The goal of this retreat was to gather all major genetic annotation groups and devise a consensus framework for validation of novel human coding loci.

IV.2. LNCRNA IN CANCER

Until now the identification of new cancer associated genes by large-scale genetic screening of cancer samples has been mainly restricted to protein coding genes[26,27]. Nevertheless, several lncRNAs have been implicated in cancer, both as tumor suppressor and oncogene[28]. To identify putative novel cancer associated lncRNA genes, we developed a unique platform to detect small and focal copy number aberrations that affect lncRNA exons but do not cover protein coding genes. The copy number profile of cancer cells has often been used to identify new cancer associated genes among protein coding genes[29,30]. These studies clearly show the importance of the size of the aberration as larger aberrations cover multiple genes and thus prevent the clear-cut identification of the cancer gene. Our strategy is unique since our probes are confined to narrow genomic loci and restricted to lncRNA exons and the exons of their flanking protein coding genes. As a result, we were able to detect hundreds of small and focal aberrations pointing to a largely unexplored landscape of cancer-associated lncRNAs. These numbers are very much in agreement with a similar effort that is based on the reannotation of public copy

number profiles[31]. The major and likely underestimated role of lncRNAs in cancer is also suggested from pervasive and differential expression of lncRNAs that has been reported in a recent large-scale RNA sequencing effort[32]. In addition, our list of putative cancer genes also harbors lncRNAs that have previously been associated with cancer by other means, further demonstrating the validity of our approach. Therefore, we selected several of the lncRNAs on our list for further research to elucidate their role in the development of cancer.

IV.3. STUDYING LNCRNA EXPRESSION

While RNA sequencing is the method of choice for *de novo* transcriptome assembly and the discovery of novel transcripts, gene expression microarrays offer an economical alternative for the assessment of global gene expression pattern of known transcripts. Moreover, RNA sequencing and gene expression microarrays generally show good concordance and it is only at high sequencing depth that RNA sequencing outperforms the latter[33]. Recently, a large-scale comparison between RNA sequencing and microarray using 498 primary neuroblastoma samples showed that both platforms performed equally well for clinical endpoint prediction[34]. As such, we designed a custom gene expression microarray based on Agilent's SurePrint G3 platform to measure the expression of LNCipedia lncRNAs and mRNAs. Our custom array and its successor have been used extensively both at the CMGG and other research labs over the past few years. In addition, it formed the basis for the development of the latest version of the commercial SurePrint G3 platform by Agilent Technologies. However, as RNA sequencing becomes more affordable, gene expression microarrays will no longer be the most economical alternative. Furthermore, RNA sequencing yields more information. It is thus foreseeable that RNA sequencing will replace gene expression microarrays for most applications in the near future.

IV.4. LNCRNA PERTURBATION IN VITRO

In vitro perturbation of gene expression is an important aspect of functional genomic research and antisense-based tools have proven to be valuable for this purpose. Antisense strategies that show great effectiveness for protein coding genes however

may be unsuitable for lncRNAs since they are insensitive to inhibition of ribosome association or due to their subcellular localisation[8]. While siRNAs have been predominantly used to achieve transient knockdown of lncRNAs, our work shows great potential in the use of ASOs for this purpose. Random selection of ASOs with perfect complementarity to a lncRNA target of interest is a viable option for ASO selection although the success rate is low. Therefore, we constructed an *in silico* model to predict ASO potency based on thermodynamic properties and the secondary structure of the target. In addition, we have show that 2'-O-me or LNA gapmers can improve the stability of the knockdown and reduce the required concentration. Finally, we devised a strategy to select and evaluate high quality NTCs. As such, our work provides lncRNA researchers with several tools and strategies that empower them to apply ASOs for the knockdown of their lncRNA of interest.

Despite our best efforts and those of other lncRNA research groups, antisense-based lncRNA perturbation remains troublesome and often, sufficient knockdown cannot be achieved. While it is hard to assess the scale of this problem from literature, we have learned from our own experience and contacts with other research groups that *in vitro* perturbation is quite often the most difficult, yet critical step in functional lncRNA research. For that reason, several research groups have turned to genome-editing to create stable knockdown cell lines or model organism. Especially the recently developed CRISPR-Cas systems (Clustered, Regularly Interspaced, Short Palindromic Repeat)[35] have gained the attention of the lncRNA research community. These systems are based on the bacterial Cas9, an endonuclease that forms a complex with specific RNA molecules and subsequently cleaves DNA that is complementary to these RNA molecules. By engineering these so-called guide RNAs (gRNAs) with complementarity to a genomic location of interest, researchers were able to perform genome-editing in a wide array of cell types[35]. While for protein coding genes a small genomic change can be sufficient for knockdown, the lack in our current understanding of lncRNA structure and function makes this impossible for lncRNAs. As such, researchers have opted to remove large parts of the lncRNA or even the entire gene[36].

While genome-editing using CRISPR-Cas is by no means straightforward and its success rate is highly variable[36], it does provide a valuable new tool to study the function of lncRNAs.

IV.5. Concluding remarks

This research has led to the development of a set of tools with substantial relevance to the lncRNA research community. In addition, the application of these tools has led to novel insights into the genetics of lncRNAs in cancer.

IV.6. References

1. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Research* **41,** D48–D55 (2013).
2. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42,** D756–63 (2014).
3. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32,** D493–D496 (2004).
4. Xie, C. *et al.* NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Research* **42,** D98–D103 (2013).
5. Jiang, Q. *et al.* LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC GENOMICS* **16 Suppl 3,** S2 (2015).
6. Paraskevopoulou, M. D. *et al.* DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Research* **41,** D239–45 (2013).
7. Fritah, S., Niclou, S. P. & Azuaje, F. Databases for lncRNAs: a comparative evaluation of emerging tools. *RNA* **20,** 1655–1665 (2014).
8. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
9. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics* **8,** 413–423 (2007).
10. Clark, M. B. *et al.* The Reality of Pervasive Transcription. *PLoS Biol* **9,** e1000625 (2011).
11. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature Biotechnology* **30,** 99–104 (2011).
12. Clark, M. B. *et al.* Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nature Methods* **12,** 339–342 (2015).
13. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most 'Dark Matter' Transcripts Are Associated With Known Genes. *PLoS Biol* **8,** e1000371 (2010).
14. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Research* **41,** 8220–8236 (2013).
15. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports* **11,** 1110–

1122 (2015).
16. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. **505,** 635–640 (2014).
17. Ulitsky, I., Shkumatava, A., Jan, C., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–1550 (2011).
18. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35,** W345–W349 (2007).
19. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7,** e1002195 (2011).
20. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27,** I275–I282 (2011).
21. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **154,** 240–251 (2013).
22. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports* **8,** 1365–1379 (2014).
23. Gascoigne, D. K. *et al.* Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* **28,** 3042–3050 (2012).
24. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature Chemical Biology* **9,** 59–64 (2013).
25. Menschaert, G. *et al.* Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell Proteomics* **12,** 1780–1790 (2013).
26. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
27. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
28. Gutschner, T. & Diederichs, S. The Hallmarks of Cancer: A long non-coding RNA point of view. *rnabiology* **9,** 0—1 (2012).
29. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
30. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
31. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology* **20,** 908–913 (2013).
32. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* (2015). doi:10.1038/ng.3192
33. Consortium, S. M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* **32,** 903–914 (2014).

34. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology* **16,** 133 (2015).
35. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32,** 347–355 (2014).
36. Ho, T.-T. *et al.* Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Research* **43,** e17–e17 (2015).

SAMENVATTING

Lange niet-coderende RNAs (lncRNAs) vormen een nieuwe klasse van genen die talrijker is dan eender welke andere klasse in het menselijk genoom. Aangezien de meeste lncRNA annotatie relatief nieuw is, ontbreekt deze in de populaire databanken en op commerciële platformen. Om dit probleem aan te pakken verzamelde ik humane lncRNA annotatie van diverse bronnen en ontwikkelde ik de publieke lncRNA databank LNCipedia (www.lncipedia.org) in de eerste maanden van mijn doctoraat. Ondertussen is het een referentiedatabank geworden met vele citaties en vermeldingen in het lncRNA onderzoeksveld. Aangezien het debat over het werkelijke aantal niet-coderende lncRNAs nog steeds gaande is, heb ik bijzondere aandacht besteed aan het bepalen van het coderend potentieel van lncRNAs. In samenwerking met andere onderzoeksgroepen van de Universiteit Gent heb ik een strategie bedacht om lncRNAs met coderende open leesramen te detecteren aan de hand van grote publieke massaspectrometrie datasets. Op die manier heb ik LNCipedia kunnen optimaliseren en kunnen aantonen dat slechts een beperkte fractie ervan mogelijks coderende open leesramen bevat. De LNCipedia dataset stelde mij in staat om verschillende platformen te ontwikkelen waarmee we lncRNAs kunnen bestuderen. Het eerste platform is een genexpressie microarray die terzelfdertijd de expressie van lncRNA en proteïne coderend RNA meet. Deze array is uitgegroeid tot het populairste platform in zijn soort aan het Centrum voor Medische Genetica Gent en is reeds gebruikt om meer dan 1000 stalen te profileren. Het tweede platform dat ik ontwikkelde is een DNA microarray met als doel kleine, focale copynumbervariaties te detecteren die specifiek lncRNAs aantasten. Met dit platform heb ik het DNA van 80 kankercellijnen bestudeerd en een groot aantal lncRNAs met een potentiele rol in kanker ontdekt. Om lncRNAs *in vitro* te bestuderen hebben we technieken nodig waarmee we de genexpressie kunnen manipuleren. Daarom evalueerde ik de bruikbaarheid van antisense oligonucleotiden om lncRNAs uit te schakelen en ontwikkelde ik een model dat de werkzaamheid van een oligonucleotide kan voorspellen.

SUMMARY

Long non-coding RNAs (lncRNAs) form a new class of genes that outnumbers any other class of RNAs predicted in the human genome. Since most lncRNA annotation is relatively new, lncRNAs are underrepresented in the established genomic databases and on the commercially available platforms. To address this issue, I collected human lncRNA annotation from different sources and developed the public lncRNA database LNCipedia (www.lncipedia.org) in the first months of my PhD. Since then it has become a reference database with numerous citations and mentions throughout the lncRNA research field. As the debate on the number of true non-coding lncRNAs is still ongoing, I paid particular attention to the assessment of the coding potential of LNCipedia lncRNAs. In collaboration with other research groups at Ghent University, I devised a strategy to use large public proteomics datasets to detect lncRNAs with coding ORFs. In doing so, I have optimized the LNCipedia dataset and showed that only a small number of lncRNAs have coding ORFs. With the LNCipedia catalogue at hand, I was able to design several platforms to study the functional role of lncRNAs. One such platform consists of a custom gene expression microarray to measure the expression of lncRNA and protein coding RNA at the same time. This array quickly grew to become the primary platform for global gene expression profiling at the Center for Medical Genetics Ghent and has currently been used for over 1000 samples. A second platform I developed is a DNA microarray for the detection of small and focal copy number aberrations that affect lncRNA genes. Using this platform, I screened a panel of 80 cancer cell lines and revealed a vast number of putative cancer associated lncRNA genes. To enable *in vitro* lncRNA studies, means to perturb their gene expression are indispensable. Therefore, I examined the use of antisense oligonucleotides for lncRNA knockdown and developed a model to predict the potency of an oligonucleotide.

PERSONAL NOTE

For me, starting a PhD was an obvious choice. Science and technology have been a lifelong fascination and I quickly realized that pursuing a PhD would fulfill the childhood dream of becoming a scientist. Sure, there have been moments of doubt and frustration in the past 4 years, but looking back I am confident I made the right decision.

Jo and Kris, I cannot thank you enough for this amazing opportunity. Jo, not once you doubted my capabilities, even though there were many occasions where I did. I have great admiration for the way you keep motivating people and finding opportunities when all seems lost. Pieter, you are one of the smartest and skillful scientists I know. Even though I learned a lot from you in the past years, you always seem two steps ahead.

Science is a team sport and as such many colleagues contributed to the projects I worked on. Justine, Katrien, Kimberly and Jasper, you made crucial contributions to this thesis. Thank you so much for putting up with my often-chaotic style and terrible sense of humor.

Aan alle vrienden en collega's (wat was ook weer het verschil?) die de afgelopen maanden te weinig (of misschien juist te veel) van mij gehoord hebben: ik maak het snel weer goed, beloofd! Of het nu een bemoedigend sms'je of babbeltje op de bureau, in 't lab of op café was, het deed steeds deugd hoor. Dankzij jullie zijn de afgelopen vier jaar snel voorbij gevlogen.

Voetbal en auto’s gingen helemaal aan mij voorbij als kind; ik bouwde raketten met kartonnen dozen en elektriciteitscentrales met lego. Mama, papa en bompa deelden met plezier in mijn rijke fantasie en interesse in wetenschap. Jullie hebben mij altijd aangemoedigd om te doen en studeren wat me interesseerde. Zonder mijn ouders had ik hier dus zeker niet gestaan, bedankt voor de steun en bedankt om steeds in mij te geloven.

Elke, jij hebt de laatste fase in mijn doctoraat van heel dichtbij meegemaakt. Mijn excuses voor mijn afwezigheid en soms minder goed humeur de afgelopen maanden. Je staat altijd voor me klaar en ik weet dat ik het niet vaak genoeg zeg, maar ik ben er je er wel ongelooflijk dankbaar voor.

Pieter-Jan

CURRICULUM VITAE

PERSONALIA

Pieter-Jan Volders

Adolf Baeyensstraat 106 - 9040 Gent - Tel: 0486/755.665

E-mail: pieterjan.volders@ugent.be - °02-06-1987

PROFILE

Motivated young scientist with a high affinity to bio-informatics and eager to learn and discover. Adapts easily to new environments/methods and loves being challenged. Social and open person who believes a good working atmosphere empowers great results.

EDUCATION

2011 – (2015): PhD in Biomedical Sciences (Ghent University, expected September 2015)

2009 – 2011: **Master of Bioscience Engineering: cell and gene biotechnology**, major in computational biology. With distinction (Ghent University)

2008 – 2009: Preparation program for Master of Bioscience Engineering (Ghent University)

2005 – 2009: Bachelor of Biology. With distinction (Hasselt University)

1998 – 2005: Sciences-Mathematics with option 8h mathematics. (Sint-Jan Berchmanscollege, Genk)

RESEARCH EXPERIENCE

PhD in biomedical sciences:

Ghent University, Faculty of Medicine and Health Sciences, Center for Medical Genetics Ghent, Vandesompele Lab

Functional annotation of long non-coding RNAs in cancer.

Masters thesis:

Ghent University, Faculty of Bioscience Engineering, Lab of Bioinformatics and Computational Genomics (BioBix)

Development of Bioinformatics tools for present-day cancer research.

PEER-REVIEWED PUBLICATIONS (A1)

Volders, Pieter-Jan; Helsens, Kenny; Wang, Xiaowei; Menten, Björn; Martens, Lennart; Gevaert, Kris; Vandesompele, Jo; Mestdagh, Pieter: *LNCipedia: a database for annotated human lncRNA transcript sequences and structures*, **Nucleic acids research** (2013)

Durinck, Kaat; Wallaert, Annelynn; Van de Walle, Inge; Van Loocke, Wouter; Volders, Pieter-Jan; Vanhauwaert, Suzanne; Geerdens, Ellen; Benoit, Yves; Van Roy, Nadine; Poppe, Bruce: *The Notch driven long non-coding RNA repertoire in T-cell acute lymphoblastic leukemia*, **Haematologica** (2014)

Sante, Tom; Vergult, Sarah; Volders, Pieter-Jan; Kloosterman, Wigard P; Trooskens, Geert; De Preter, Katleen; Dheedene, Annelies; Speleman, Frank; De Meyer, Tim; Menten, Björn: *ViVar: A Comprehensive Platform for the Analysis and Visualization of Structural Genomic Variation*, **PloS One** (2014)

Van Peer, Gert; Lefever, Steve; Anckaert, Jasper; Beckers, Anneleen; Rihani, Ali; Van Goethem, Alan; Volders, Pieter-Jan; Zeka, Fjoralba; Ongenaert, Maté; Mestdagh, Pieter: *miRBase Tracker: keeping track of microRNA annotation changes*, **Database** (2014)

Volders, Pieter-Jan; Verheggen, Kenneth; Menschaert, Gerben; Vandepoele, Klaas; Martens, Lennart; Vandesompele, Jo; Mestdagh, Pieter: *An update on LNCipedia: a database for annotated human lncRNA sequences*, **Nucleic acids research** (2015)

Rihani, Ali; Van Goethem, Alan; Ongenaert, Maté; De Brouwer, Sara; Volders, Pieter-Jan; Agarwal, Saurabh; De Preter, Katleen; Mestdagh, Pieter; Shohet, Jason; Speleman, Frank: *Genome wide expression profiling of p53 regulated miRNAs in neuroblastoma*, **Scientific reports** (2015)

Up-to-date list with article citations at http://goo.gl/8FH2AP

CONFERENCE ORAL PRESENTATIONS

Volders P.J.: *Explorative tools for identification of lncRNAs in cancer*, From nucleotides to networks: 2nd annual MRP symposium on Bioinformatics, September 2012, Ghent.

Volders P.J.: *Developing tools for long non-coding RNA research*, N2N: Nucleotides to networks symposium, October 2013, Ghent.

Volders P.J.: *Selecting potent antisense oligonucleotides for lncRNA silencing in vitro using thermodynamic properties and target RNA structure*. Progress Meeting: IUAP P7-03: "Cancer cells and their microenvironment: from gene regulatory networks to therapy", June 2014, Brussels.

Volders P.J.: *LNCipedia: a database for annotated human lncRNA sequences*. Devising a consensus framework for validation of novel human coding loci Retreat, May 2015, Wellcome Trust Conference Centre, Hinxton, UK.

CONFERENCE POSTERS

Volders P.J., Mestdagh P., Menten B., Vandesomepele J.: *Large-scale and targeted genomic screen reveals focal lncRNA copy number alterations in cancer cells*. Keystone symposia: Non-coding RNAs, March 2012, Snowbird, Utah, USA.

Volders P.J., Mestdagh P., Vandesomepele J.: *LNCipedia: a novel databse for annotated lncRNA sequences and structures*. Keystone symposia: Non-coding RNAs, March 2012, Snowbird, Utah, USA.

Volders P.J., Helsens K., Wang X., Martens L., Gevaert K., Vandesompele J., Mestdagh P.: *Lncipedia: a novel database for annotated lncRNA sequences and structures*. From nucleotides to networks: 2nd annual MRP symposium on Bioinformatics, September 2012, Ghent.

Volders P.J., Helsens K., Wang X., Martens L., Gevaert K., Vandesompele J. and Mestdagh P. : *LNCipedia 2.0: a database for annotated human lncRNA transcript sequences and structures*. Noncoding RNAs in Development and Cancer, January 2013, Vancouver, Canada.

Volders P.J., Helsens K., Wang X., Martens L., Gevaert K., Vandesompele J. and Mestdagh P. : *LNCipedia: a database for annotated human lncRNA transcript sequences and structures*. 13th Annual meeting of the Belgian Society of Human Genetics, March 2013, Brussels.

Volders P.J., Helsens K., Wang X., Martens L., Gevaert K., Vandesompele J. and Mestdagh P. : *LNCipedia 2.0: a database for annotated human lncRNA transcript sequences and structures*. IUAP P7/03 progress meeting. Mai 2013, Leuven.

Volders P.J., Helsens K., Wang X., Martens L., Gevaert K., Vandesompele J. and Mestdagh P. : *LNCipedia 2.0: a database for annotated human lncRNA transcript sequences and structures*. EMBO|EMBL Symposium The Non-Coding Genome, October 2013, Heidelberg, Germany.

Volders PJ., Mestdagh P., Vandesompele J. : *Selecting potent antisense oligonucleotides for lncRNA silencing in vitro using generalized additive models*. 4th N2N MRP Bioinformatics symposium, Mai 2914, Ghent.

Volders PJ., Mestdagh P., Vandesompele J.: *Selecting potent antisense oligonucleotides for lncRNA silencing in vitro using thermodynamic properties and target RNA structure*. Progress Meeting: IUAP P7-03: “Cancer cells and their microenvironment: from gene regulatory networks to therapy”, June 2014, Leuven

Volders PJ., Mestdagh P., Vandesompele J. : *Selecting potent antisense oligonucleotides for lncRNA silencing in vitro using thermodynamic properties and target RNA structure*. Non-coding RNA - From Basic Mechanisms to Cancer, June 2014, Heidelberg, Germany.

TEACHING & COACHING

Teaching:

Lesson on database technology in the course Advanced Bio-informatics and Genetic Data-analysis (Master of Science in Biomedical Sciences, prof. Jo Vandesompele), 2013-2015.

Master thesis supervision:

Els Bauwens, Master of Science in de Biochemie en de Biotechnologie, *De rol van lange niet-coderende RNA-genen in kankerontwikkeling*, June 2015

Bachelor thesis supervision

Stephanie Letellier, Bachelor in de Biomedische Laboratoriumtechnologie, *Update van LNCipedia - Een database voor humane long non-coding RNA*, June 2014

Master thesis jury member:

Eline De Schutter, *The epitranscriptome of the colorectal cancer cell line HCT116 measured with m6A-seq*, June 2015

Maxime Menu, Master of Science in Biomedical Sciences, *Creatie en analyse van een humane eiwit expressie atlas*, June 2015

Peer review:

Recurring reviewer for Nucleic Acids Research (impact factor: 8.808)

SKILLS & EXPERTISE

Bio-informatics:

NGS data-analysis, (expression) array analysis, qPCR data processing, primer design, statistics and data visualization (R)

Programming:

Perl, PHP, C++, Objective-C, Javascript, various web-development languages and frameworks

Databases:

MySQL, MongoDB

Wet-lab:

Tissue culture, transfection, qPCR, array comparative genomic hybridization, digital PCR