Integrative Characterization of Human Long Non-Coding RNAs

Dissertation Advisers: Prof. Aviv Regev & Prof. John L. Rinn Nataly Moran Cabili

Integrative Characterization of Human Long Non-Coding RNAs

Abstract

Since its early discovery as a messenger, RNA has been shown to play a diverse set of regulatory, structural and even catalytic roles. The more recent understanding that the genome is pervasively transcribed stimulated the discovery of a new prevalent class of long non coding RNAs (lncRNAs). While these are lower abundant and relatively less conserved than other class of functional RNAs, lncRNAs are emerging as key players in different cellular processes in development and disease.

Determining the function of individual lncRNAs, however, still remains a challenge. Recent advances in transcriptome profiling and RNA imaging allow for an unprecedented analysis of lncRNA transcripts. In this thesis I applied RNA sequencing (RNA-Seq) and single molecule RNA florescent *in situ* hybridization (FISH) to systematically characterize lncRNAs as a class and as individuals.

First, I used an integrative approach to define a reference catalogue of over 8,000 human large intergenic noncoding RNAs (lincRNAs). This catalogue unifies previously existing annotation sources with transcripts we assembled from RNA-Seq data collected from ~4 billion RNA-Seq reads across 24 human tissues and cell types. I characterized each lincRNA by a panorama of more than 30 properties, including sequence, structural, transcriptional, and orthology features.

We find that lincRNA expression is strikingly tissue specific compared to coding genes, and that they are typically co-expressed with their neighboring genes, albeit to a similar extent to that of pairs of neighboring protein-coding genes. We identify ~1000 lincRNAs that have a syntenic expressed ortholog in another species. We distinguish an additional sub-set of transcripts that have high evolutionary conservation but may encode open reading frames, and may serve either as lincRNAs or as small peptides. Our integrated, comprehensive, yet conservative, reference catalogue of human lincRNAs reveals the global properties of lincRNAs and facilitates experimental studies and further functional classification of these genes.

To approach other fundamental aspects of lincRNAs biology, including their subcellular localization, abundance and variation at a single cell resolution, we then systematically applied single molecule imaging. We used single-cell single-molecule RNA FISH to survey 61 lncRNAs and catalog their abundance and cellular localization patterns in three human cell types. I chose our lncRNA set by a variety of properties such as conservation and tissue specific expression across a spectrum of expression abundance. Notably, we reveal challenges in applying single molecule RNA-FISH to reliably measure lncRNAs and establish an integrated imaging and computational strategy to overcome these.

We found that lncRNAs have diverse sub-cellular localization patterns, ranging from strictly nuclear localization to almost exclusive cytoplasmic localization, with the majority localized primarily in the nucleus. We observed that the low abundance of lncRNAs as measured in bulk cell populations cannot be explained by high expression in a small subset of ‘jackpot’ cells among a majority of non-expressing cells. Simultaneous analysis of lncRNAs and mRNAs from corresponding divergently transcribed loci showed that divergent lncRNAs do not present a distinct localization pattern and are not always co-regulated with their neighbor. Overall, our study highlights important differences and similarities between lncRNAs and mRNAs. The rich set of localization patterns we observe suggest a broad range of potential functions for lncRNAs.

In summary, this work highlighted key properties of lncRNAs as a class and individually. It demonstrated the application of two emerging technologies for the specific study of this low abundant class of lncRNAs, revealed challenges, and provided strategies to approach them. Importantly, our catalogs provide valuable resources that can assist in generating testable hypothesis for mechanistic studies that would further elucidate lncRNAs functions.

Table Of Contents

Acknowledgments

As I come to close this chapter in my life, I would like to thank many unique people, with whom I was lucky to interact and who made my time at graduate school a wonderful experience.

First, I would like to express a deep gratitude to my advisers, ‘my academic mom and dad’, Aviv Regev and John Rinn. Thank you both for enabling me to grow and believe in myself as an independent scientist, and for always being there when I needed. Thank you John, for your inspirational enthusiasm and dedication, for creating new opportunities for me, and providing endless intellectual and emotional support. Thank you Aviv, for being an inspiring role model as a person, woman and scientist, providing insightful comments and ideas, teaching me to write and finding a solution to every problem. I would also like to thank my dear collaborator and ‘adopted adviser’, Arjun Raj. Thank you Arjun for a true mentorship: this has been an enriching learning experience. I am very fortunate to have you all as my mentors.

I thank my ‘informal’ mentors Manuel Garber, Mitchell Guttmann, Martin Sauvageau, and Sarah Calvo, who have patiently guided me through different steps of this endeavor, generously sharing their time and knowledge. Thank you Martin, for your dear friendship and for teaching me how to become an experimentalist. Thank you Sarah, for being an inspiration and for teaching me all the ‘little things’ I use every day.

I thank my dear collaborators and colleagues: Cole Trapnell, Loyal Goff, Margaret Dunagin, Patrick McClanahan, Barbara Tazon-Vega, Sabine Loewer, David Kelley, David Shechner, Ezgi Hacisuleyman, Casey Gifford, Chiara Gerhardinger, Michael Morse, Or Zuk, and all members of the Regev, Rinn and Raj labs and ‘inhabitants of the 6th floor’. Thank you for taking the time to teach me and help me with different aspects of my science. Thank you for also making it fun.

I thank the members of my dissertation and examining committees: Angela DePace, Martha Bulyk, Chris Burge, Brad Bernstein, Fritz Roth, Pam Silver, Galit Lahav and Mike Springer. Thank you, for your time, support and insights.

I thank the Howard Hughes Medical Institute International Graduate Research Fellowship, and the Systems Biology Department for financial support. I especially would like to thank Sam Reed for making everything work out so smoothly.

I thank my classmates and friends for making lab feel like a second home: Tal Shay, Michal Rabani, Moran Yassour, Danny Lieber, Ashley Wolf, Max Staller and the ‘Rinn lab family’. Rinn lab, thank you, it was a great privilege to be part of our hunting team.

I thank my loving and supporting family that is always there for me. To my parents, Dorit and Avi, for being my source of inspiration and motivation and for making everything in life possible. To my mom for teaching me to be dedicated, careful and thorough, and to my dad for teaching me about true friendship. To my sister, Karin, that is inspiring in how she accomplishes every dream in spite of all challenges. To my parents in law, Osnat and Yossi, for always being there and supporting me being a career mom. To my grandparents, Aryeh and Flora for always cheering, and Nehama for being my first teacher, and such a dedicated one.

Finally, I thank my loving husband, Amitay, for unlimited support and encouragement, for teaching me how to ‘take things easy’, and for being an exceptional father to Arielle. And last but not least, thank you, Ariellie, for being our sunshine and an endless source of joy. You make everything have a whole different meaning.

To my loving grandparents,

Nehama and Shlomo,

Flora and Aryeh

Chapter 1 | Introduction

1.1 RNA as a functional molecule

The central dogma of biology describes RNA as a messenger carrying the genetic information encoded in the DNA to the factories decoding it to protein [1]. Yet, this initial perception of RNA as a passive molecule is misleading, because ever since the discovery of the messenger RNA (mRNA) in the early 1960s [2], RNA was shown to play key roles in diverse processes including as a catalyst. In fact, RNA's unique capability to form structure and to base pair with nucleic acid was already demonstrated within the process of protein translation with the discovery of transfer RNA (tRNA) [3, 4] and the more recent appreciation that the ribosomal RNA itself carries catalytic activity [5-7].

In the 1960s another large set of heteronuclear RNAs (hnRNAs), that resembled mRNA but did not enter the ribosome, was discovered [8, 9]. This fraction was first elucidated with the discovery of small nuclear RNA (snRNAs) [10], later shown to be involved in splicing [11], and the discovery of splicing [12] and of small nucleolar RNA (snoRNA) [11], which are involved in ribosome biogenesis, in the 1970s. snoRNA and snRNA were discovered based on their nuclear localization , are conserved from yeast to mammals, and use base pair complementarity as part of their target recognition [13]. Interestingly, snRNAs are transported to the cytoplasm to be assembled into a ribonucleoprotein complex and then shuttle back to the nucleus to participate in splicing [13].

The 1980s provided a new perspective with the discoveries of catalytic RNAs: the RNaseP in bacteria [14], and the self-splicing RNA in Tetrahymeana [15][16], which were later followed by discoveries of several riboswitches [17]. Together these supported the idea of an ancient "RNA world" in which RNA was the sole molecule capable of preserving genetic information and catalyzing chemical reactions [18].

In the 1990s, the regulatory roles of RNA were started to be appreciated with the discovery of the first microRNA in C. elegance, lin-4, that repressed lin-14 through short sequence base pairing with the 3' untranslated region (UTR) of its mRNA target [19]. let7 was later shown to act similarly [20]. Genome scale efforts in the early 2000s then revealed an abundant class of such tiny RNA, termed microRNAs [21-23], now known to destabilize their mRNA target by base-pairing with the 3' UTR of the mRNA through a short seed sequence on the microRNA, while being bound to a protein complex containing Argonaute proteins [24]. microRNAs, which were shown to be conserved in animals and plants, have critical roles in diverse processes in development and disease by the same shared mechanism [24]. The piwi-interacting RNAs (piRNAs) are another class of more recently discovered small RNAs that 'guide' Argonaute mediated silencing of transposons in the testis [25].

The early 1990's also brought the idea that RNA may come "in many more flavors" with the discovery of two long RNA molecules that had every property of an mRNA yet did not appear to encode a protein. These were therefore referred as long non coding RNAs (lncRNAs). H19, the first lncRNA discovered, showcases many of the themes used to this day to discover new lncRNAs. H19 was inadvertently discovered during a differential hybridization screen that first detected this highly abundant RNA in fetal liver [26]. Immediately it had become a mystery: it did not have any open reading frame (ORF) that was conserved in its mouse orthologs, it was cytoplasmic but did not associate with ribosomes, and polyclonal antibodies raised against the in vitro product of its longest ORF did not detect any proteins in vivo. While the functional role of this imprinted gene [27] remained a mystery until later when it was shown it is in fact a microRNA host gene [28-30] , it suggested that many other such lncRNAs could exist.

Shortly after the discovery of H19, a screen to detect genes escaping X inactivation (XIC) revealed XIST (X inactive specific transcript) [31], a lncRNA that coats the inactive X chromosome [32, 33] and is required for XIC [34, 35], which soon became a paradigm for understanding lncRNA mechanism of action [36]. A few other lncRNAs including: TSIX [37], Air [38], Tug1 [39], NRON [40], Kcnq1ot1 [41] and HOTAIR [42] were discovered in the following decade. Yet, the genome scale efforts pursued at that time suggested these were only the tip of the iceberg.

1.2 The discovery of a prevalent class of lncRNAs

Emerging technologies reveal pervasive transcription of mammalian genomes

The completion of sequencing of the human genome revealed that its vast majority (>95%) does not encode protein [43, 44]. Rather, its ~20,000 protein coding genes are dispersed within repetitive elements (that in total account for 50%-70% of the human genome [45]) and other regulatory sequence. Surprisingly, subsequent high-throughput genome scale efforts revealed that as much as ~70-90% of mammalian genomes is transcribed at some time point during development [46]. This was first observed using unbiased transcriptome profiling by hybridization to tilling microarrays targeting one or several chromosomes [47-49] and then by deep sequencing of large pools of cDNA clones (usually in their truncated forms as expressed sequence tags (ESTs) to increase throughput and cost effectiveness) [46, 50].

The extreme rarity of most of these new transcripts [50], however, questioned whether this transcription is in fact accidental. In support of this view, it was shown that as many as 90% of RNA Polymerase II initiation events are non-specific [51]. Moreover, these transcribed regions suffered from low cross-species conservation rates and therefore were less likely to be functional [52]. However, it was suspected that the large pool of cryptic transcripts does include a subset of RNA molecules that may not encode new proteins but might indeed be functional. The existence of short functional RNA, such as microrRNAs, as well as lncRNAs, such as XIST and imprinting lncRNAs, was already clear [53]. It remained unclear, however, how many functional ncRNA were still to be discovered. To do so one first needed a 'map' of potential transcript models to explore.

Finding needles in a haystack – early discoveries of lncRNA sets

Most of the early studies discovering these new transcripts used tiling microarrays in which cDNA from poly-adenylated or ribo-depeleted RNA, is hybridized to microarrays comprised of short probes (~50bp) tiling large continuous genomic regions at high resolution (~10bp) [54, 55]. Computational analysis was then applied to define peaks of hybridization signal across the tiled region and subsequently to identify the boundaries of transcribed regions. This approach was applied to identify lncRNAs across mammalian genomes [47-49] and at fine resolution across the HOX genes clusters [42]. This hybridization approach, however, suffered from low reproducibility, low dynamic range and low resolution which was essential to define exon boundaries of spliced transcripts [56] [57].

A significant advancement in the discovery of a lncRNA subset that was likely to be functional was in the pioneering work by *Guttman et al.* [58]. In this work chromatin state maps coupled with tiling arrays and novel algorithms were used to identify large intergenic (or intervening) non coding RNAs (lincRNAs) that carried the chromatin signature of actively transcribed genes [58]. Chromatin state maps are generated by chromatin immune-precipitation followed by next generation sequencing (ChIP-Seq) to detect the genomic regions marked by a specific chromatin modification [59]. Using such maps, researchers revealed the chromatin state of different genomic loci such as promoters, actively transcribed genes, repressed regions, enhancers and others [60, 61].

By scanning for 'K4-K36' domains, that is histone 3 lysine 4 trimethylation (H3K4me3) across the promoter followed by H3K36me3 along the downstream transcribed region, in gene deserts, ~3500 novel noncoding gene loci were identified in the mouse [58] and human [62]. Hybridization of poly-A RNA to tiling microarrays tiling 350 of the ~1600 intergenic K4-K36 domains in mouse, confirmed the expression of ~70% of these transcripts and defined their exons. Interestingly, these exonic regions demonstrated elevated evolutionary conservation despite very low coding potential, which suggested they were likely to be functional ncRNAs. Importantly, a significant number of these were later shown to have phenotypic effect in response to perturbation [63-66].

Overall, the key advancement in the 'K4-K36' domains approach is that it demarcated genomic regions that were likely to be homogeneously and significantly expressed across the population and could be effectively detected by a cost effective hybridization based approach. This was due to the use of a statistical approach (scan statistics) to define a robust signal, and since it relied on chromatin state maps which reflect the aggregate of a binary signal per cell (as each cell can produce a limited amount of signal per genomic loci) rather than on RNA measurements (as RNA is an amplified signal from one genomic locus). While this approach identified transcripts that were likely to be functional based on their conservation, it was still limited as it did not define the complete and accurate transcript model and exon boundaries, information that is essential for subsequent functional investigation.

In parallel to these studies, efforts to sequence full cDNA clones were on going by large consortia and annotation groups such as the FANTOM (Functional Annotation of The Mouse) project [46] and GENCODE (the reference annotation of ENCODE, the Encyclopedia Of DNA Elements, in human) [67]. The first catalog of potential lncRNAs in mouse was generated by the FANTOM consortium and included 3652 high confidence transcripts distinguished from >34,000 novel transcripts that were detected by cDNA cloning followed by Sanger sequencing across mouse tissues [46, 68, 69]. These were termed macroRNAs and were shown to be under purifying selection [70, 71] and differentially expressed in several developmental systems in the mouse [72-74]. Parallel sets were later created for human but were still limitedly explored [75, 76].

While providing most accurate gene models for newly discovered transcripts, back in 2009 catalogs containing these models did not detail the tissues or cell type in which they were expressed, limiting the accessibility to study them by the wide community. Moreover, full cDNA cloning had lower throughput and was more costly compared to short tag sequencing and tiling arrays and catalogs relying on such approach were still partial with respect the number of cell types sampled and the detection of low abundant molecules. A major step forward was the appearance of high-throughput RNA sequencing (RNA-Seq).

RNA-Seq – transcriptome exploration within reach

Massively parallel cDNA sequencing, or RNA-Seq, uses massively-parallel sequencing technologies and provides a cost effective experimental procedure to explore the entire transcriptome in an unbiased manner and fine resolution [77]. The availability of millions of RNA sequencing reads and new algorithms offered far more precise measure of the transcriptome relative to hybridization based approaches. Briefly, an RNA sample (usually fractionated as polyA, or depleted of rRNA) is converted to cDNA, fragmented and tagged with adaptors on one or both ends. cDNA is then amplified (with some exceptions) and fragments are sequenced from either one or both ends (single- or paired- end reads, respectively) on a next generation sequencer to obtain short sequence reads (read length varies based on the technology used and are usually within the range of 35-400 bp).

This unbiased sample of the transcriptome can be used to both identify the transcribed transcripts (including previously unknown transcripts) and to quantify their abundance. This requires powerful computational methodologies for read alignment, ab-initio (or de-novo) transcript assembly and estimation of transcript abundance and differential expression [78]. Next, I will highlight the key methods for each task used in our work to annotate lincRNAs from RNA-Seq data.

Fully leveraging the power of RNA-Seq to identify novel transcripts requires a fast aligner that will map short reads to the genome, including reads that span exon-exon junctions termed spliced reads. Tophat [79], provides such a solution in two main steps. First it uses Bowtie [80], a fast un-spliced reads aligner, to map all reads that map to the reference genome without allowing large gaps. Bowtie is significantly faster than traditional unspliced aligners as it uses the Burrows-Wheeler transformation (BWT) [81] that compacts the genome into a database that can be searched efficiently. Briefly, BWT is an efficient string compression technique that permutes a string in a reversible manner such that the new string contains repeats of identical characters and the same character composition as in the input string. Next, Tophat identifies all potential exon spliced junctions based on the islands of already mapped reads, and performs a second fast alignment step to map the spliced reads (**Figure 1.1**).

To define the map of all expressed transcripts in a given sample one needs to assemble the reads into transcripts. This task of transcript reconstruction can be approached by either genome independent [82] or guided approaches [83, 84] (**Figure 1.2a**). Since the latter approach is more computationally efficient and a reference genome is available for our organisms of interest, it is mostly adopted for the task of lncRNA annotation. Originally, two such methods, named

Cufflinks [84] and Scripture [83], were developed to reconstruct full transcripts given a spliced read alignment to a reference genome taking advantage of longer read length (~75bp). The main difference between the two approaches is that Cufflinks identifies a minimal set of isoforms that can explain the alignment, while Scripture reports all isoforms (**Figure 1.2b**). This main difference is relevant when solving the subsequent task of isoform abundance estimation [78].

Scripture, originally designed to distinguish low expressed transcript from noise, uses a scan statistics to identify all isoforms that are significantly expressed on a particular genomic locus [83]. In brief, Scripture transform the genome alignment to a graph topology that represents all the possible connections of bases that occur continuously (non-spliced reads) or non-continuously (spliced reads) and assigns weights to these connections based on read density. It then scans this graph using a sliding window and applies a scan statistics to identify all connected paths that are significantly represented. These correspond to significantly expressed transcripts.

Cufflinks was designed to detect the most likely isoforms and estimate their abundance, simultaneously [84]. It uses a conceptually similar graph representation of the read alignment and relies on graph theory to provide a maximum likelihood explanation of the data that corresponds to a minimal set of isoforms in a given locus. Cufflinks may include more lowly expressed transcripts in its output compared to Scripture. Importantly, it addresses both the transcript reconstruction and abundance estimation tasks at once and fully utilizes the power of paired end reads which can inform connections between distal exons despite the limited read length.

Given a set of transcript models one can then use RNA-Seq to quantify their abundance across samples as well identify transcripts that are differentially expressed across samples. Initially, the RPKM (read per kilobase of transcript per million mapped reads) metric was introduced as a simple approach to estimate gene abundance by counting all the reads aligned to the gene and normalizing this sum by the transcript’s length and the library coverage [85]. This approach, however, turned out to be naïve for the task of quantifying the abundance of isoforms of the same gene, as this requires a systematic approach to model the uncertainty in assigning reads to each of the isoforms. Since many of the reads align to exonic regions that are shared between isoforms they cannot be assigned uniquely [78].

Later and more advanced methods such as Cufflinks [84, 86], mixture of isoforms (MISO) [87] and RSEM (RNA-Seq by Expectation Maximization) [88] estimate abundance at the isoform level and model the uncertainty in read assignment by constructing a 'likelihood function' which models the sequencing process and identifies isoform abundance estimates that best explain the reads obtained in the experiment (i.e. a maximum likelihood estimate (MLE)). These methods apply different statistical frameworks to find a MLE, and provide a confidence measure on this estimate. This is especially important in the case of low abundant isoforms for which the uncertainty is greater.

The combination of different de-novo assembly approaches provides a better understanding of a newly explored transcriptome as one can establish classification and scoring schemes that categorizes each transcript (e.g. one can apply a scan statistics on a minimal isoform set reported by Cufflinks as an additional measure of abundance). Of note, large isoform complexity in a given region imposes computational challenges that require enhanced computational resources to obtain a solution. Therefore, when using RNA-Seq ab initio assemblers that report a maximal isoform set, it may be necessary to pre-filter such sets prior to the isoform quantification step, as incorrect or miss-assembled transcripts introduce uncertainty and complexity which may result with infeasible computation [78]. Overall, isoform based methods provide a confidence measure on their predictions, which is particularly informative for low abundant transcripts such as lncRNAs.

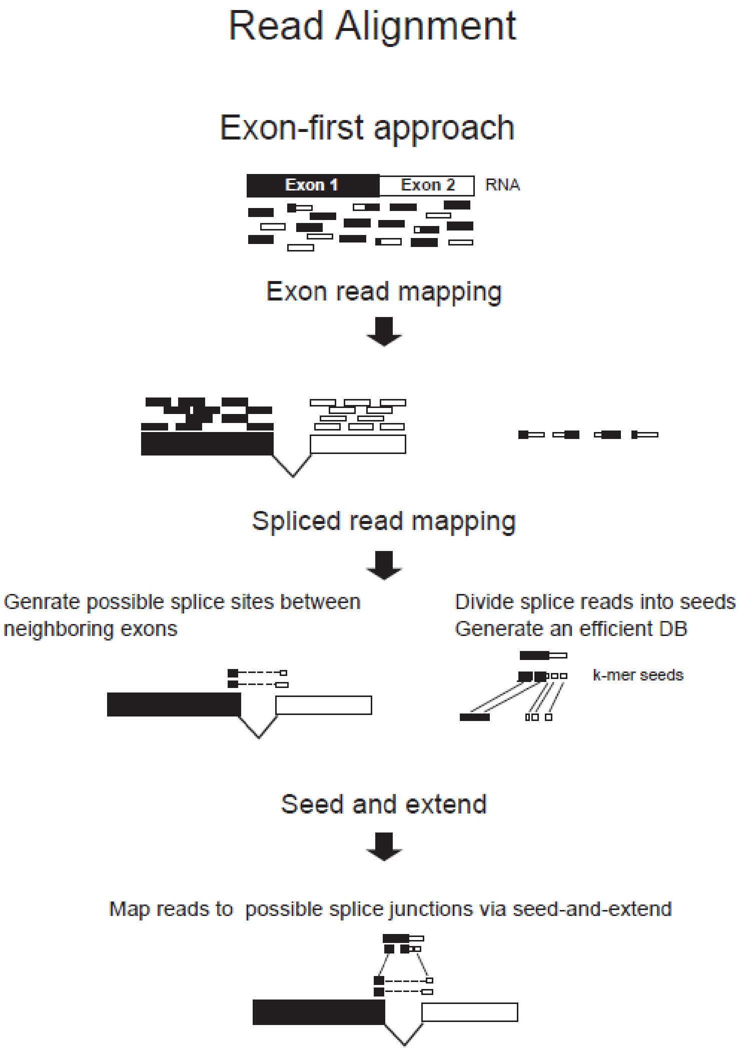

Read Alignment
Exon-first approach
Exon 1
Exon 2
RNA
Exon read mapping
Spliced read mapping
Genrate possible splice sites between neighboring exons
Divide splice reads into seeds
Generate an efficient DB
k-mer seeds
Seed and extend
Map reads to possible splice junctions via seed-and-extend

Figure 1.1| An exon first approach for read alignment. Adapted with permission from *Garber et al.* [78]. The illustration describes reads originating from a two exon transcript. Portions of the reads are labeled based on the exon of origin. In the first step unspliced reads are fully mapped to the genome (top). The remaining spliced reads are then divided into smaller pieces which are stored in an efficient database (DB). Genomic regions that were covered by exonic reads are used to generate possible splice sites between neighboring exons. A seed and extend step is then applied to map to spliced reads by mapping the k-mer seeds to the possible splice junctions and extending these from this initial mapping to align the full read from the two sides of the exon junction.

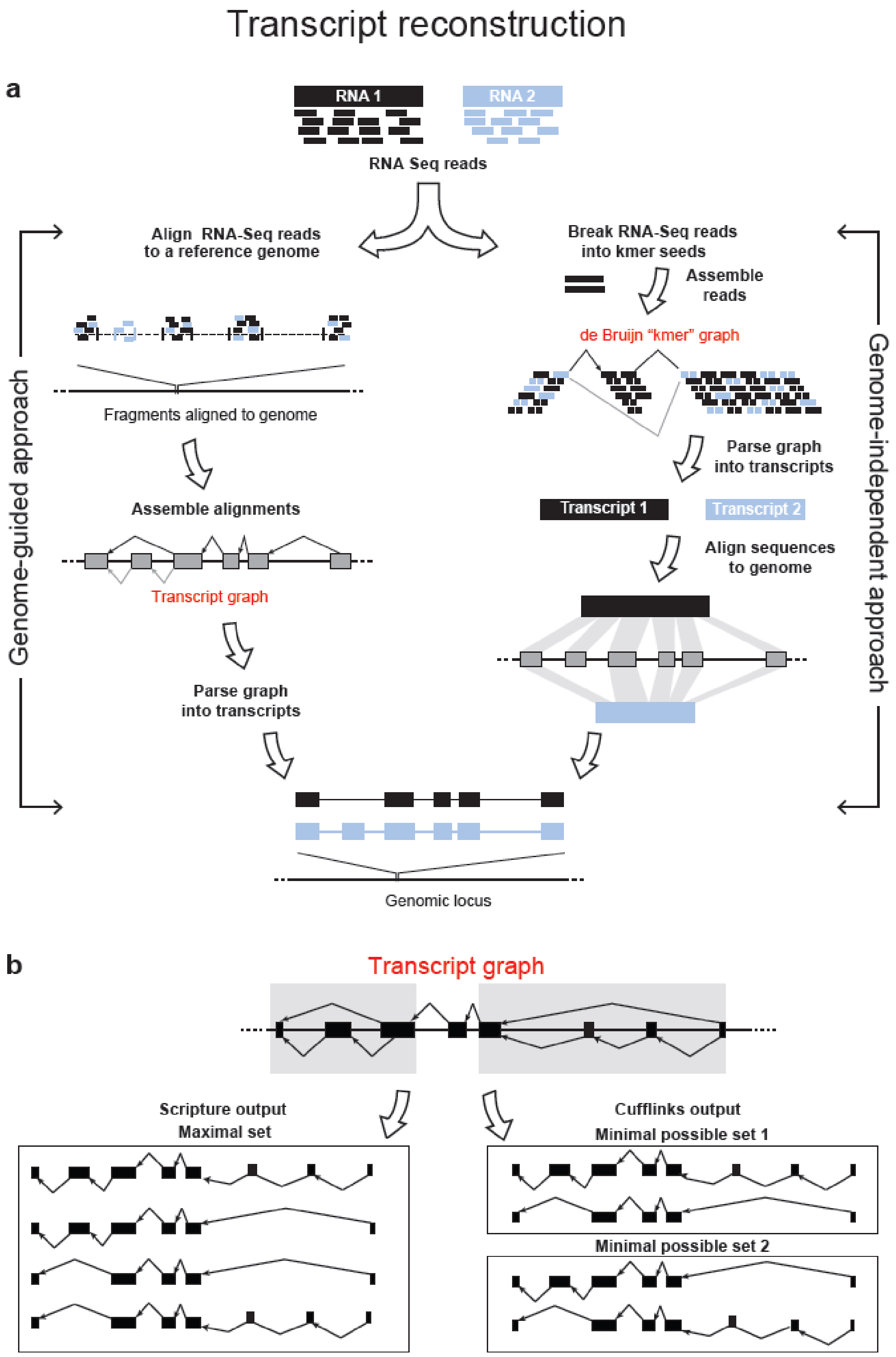


Figure 1.2| Transcript reconstruction approaches. Adapted with permission from *Garber et al.*[78]. **(a)** The illustration describes two isoforms of the same gene (labeled black and gray) and reads originating from each isoform are labeled accordingly. In a genome guided approach (left), reads are first

aligned to the genome. The aligned regions and spliced reads are then used to reconstruct a transcript graph that is then parsed into isoforms annotations. In a genome independent approach the transcripts are divided into short k-mers that are used to generate a de Bruijn graph structure which is in turn parsed to transcript annotations. If a genome is available these can then be aligned to the genome to generate gene annotations. **(b)** A transcript graph can be parsed to a maximal set (left) of transcripts which corresponds to all possible transcripts which are consistent with the spliced reads, or different minimal sets (right) that can explain all the spliced reads by only 2 transcripts.

1.3 lincRNA identification and classification

Classes of lncRNAs by genomic features

Since not much is known about lncRNAs and their functional classes, one approach of bringing order is dissecting them to subsets based on their genomic features. Since it is likely that mechanistic themes are shared across the genes in a set, this classification may facilitate further hypothesis driven functional investigation. These include: (1) natural antisense transcripts, (2) intronic transcripts, (3) pseudogenes, (4) divergently transcribed transcripts and promoter associated transcripts (5) enhancer RNAs and (6) lincRNAs (**Figure 1.3**).

Natural antisense transcripts (NATs) are those transcribed in the antisense direction and overlap exons of another known gene (usually an mRNA) and are frequently enriched around the 3' or 5' of the sense transcript [89-91]. Examples include the imprinting ncRNAs Kcnq1ot1 and Air, as well as Xist/Tsix. Long intronic RNA are those transcribed from the intron of a known gene in either sense or antisense direction [92], such as COLDAIR, a lncRNA expressed from the first intron of the FLC locus, which is involved in regulating plant flowering time [93].

Pseudogenes are relics of genes that lost their coding potential by nonsense, frame shift or other types of mutations [94]. They are usually incorporated randomly throughout the genome by transposition or duplication, and as many as 20% were shown to be transcribed. Recently, studies presented evidence supporting a model in which transcribed pseudogenes serve as microRNA sinks, thus fine tuning microRNA repression targeting their parent gene [95].

Gene promoters were described to transcribe a group of short and long transcripts in addition to and divergent to the gene they control in a phenomena known as divergent transcription [96-99]. These include a set of short transcripts (20-2500nts) that transcribed upstream and downstream divergently relative to the transcription start site mainly bounded by pausing sites of RNA Pol-II. Upstream antisense RNA (uaRNA) or promoter upstream transcripts (PROMPTS) are short, low abundant and usually unstable. In some cases such divergent transcription results in more stable and spliced transcripts often referred as divergent lincRNAs [100, 101]. Divergent transcription also occurs at distal enhancers resulting in enhancers RNAs (eRNAs) [102, 103]. These are usually short and unspliced and were recently shown to contribute to the enhancer function in activating genes in proximity [104, 105].

Large intergenic (or originally intervening) noncoding RNAs (lincRNAs) are the subset of lncRNA that are capped, poly-adenylated and frequently spliced, encoded in intergenic region. That is, they do not overlap the genomic exonic regions encoding protein coding genes and other known types of non-coding transcripts (e.g. microRNAs). The distinction between lncRNAs and lincRNAs was mainly motivated by the need to facilitate the interpretation of sequence conservation and perturbation outcomes, which are first means in screening for functional candidates. This distinction, however, does not imply any differences in the functional mechanism between intergenic and non intergenic lncRNAs, which probably comprise a range of different mechanistic activities (discussed in **Section 1.4**).

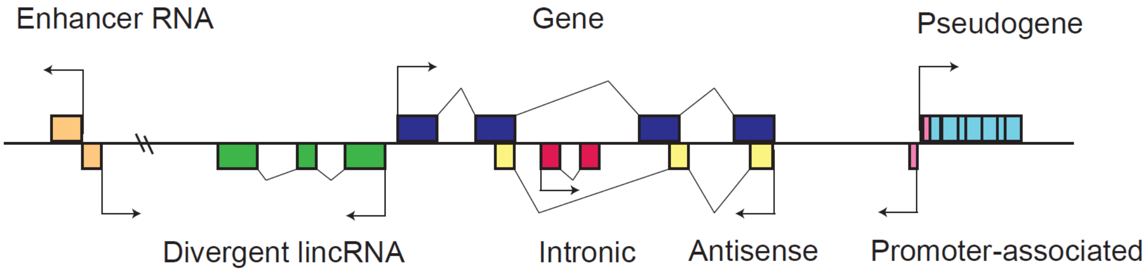


Figure 1.3| Classes of lncRNAs by genomic features. Inspired by [106]. Enhancer RNAs are short and transcribed divergently from an enhancer (peach). Divergent lincRNA (green) which are processed and spliced are transcribed upstream and in the antisense direction from a gene's (blue) promoter. They do not overlap any other exonic sequence. Intronic lncRNAs (red) are transcribed from the intron of a gene in the sense or antisense direction and do not overlap exonic sequence. Antisense transcripts (yellow) are transcribed antisense and share some exonic sequence of the corresponding gene. Promoter associated transcripts (pink) are very short and can be transcribed bidirectionaly upstream and antisense or downstream and sense from a gene's promoter. Pseudogenes (light blue) can be found in the spliced form (as shown here) or the unspliced form of the parent gene and go through mutations such that they lose their coding ability.

Annotation of lincRNAs

While new transcripts are constantly being identified, the next challenge is to systematically classify them into classes to facilitate their selection for further functional investigation. This process includes two main steps: (1) identification of full length transcript models; and (2) evaluation of coding potential. To date catalogs of lincRNAs are available for human, mouse, zebrafish, frog, nematode, *Arabidopsis*, maize and *Plasmodium* (see [107] for review). These catalogs vary in the input data and the computational methods used for classification, which may be constrained by cost and their availability for the specific species at the time generated.

As previously described, catalogs of lncRNAs transcripts are based on 3 types of data: (1) tiling arrays, which are no longer used due to their low resolution and accuracy; (2) cDNA full length sequencing, which currently provides the most reliable full transcript models, but is costly and may be limited in coverage; (3) RNA-Seq, which has high coverage but is somewhat limited in its ability to correctly reconstruct very low abundant transcripts due to low frequency of reads (especially spliced) spanning these loci. This, however, can be improved by sequencing higher numbers of longer reads. More recently, RNA-Seq based transcripts model are refined by integrating other sequencing based approach that define the transcript's 3' and 5' boundaries in the transcript reconstruction process (such as Capped Analysis of Gene Expression (CAGE) [108] and 3P-Seq, a method to annotate polyadenylation sites [109, 110]).

As of 2014, cDNA based lncRNA catalog generated by annotation groups such as RIKEN [46, 111], REFSEQ [112], GENCODE, HAVANA (Human and Vertebrate Analysis and Annotation) and Ensembl [113], are available for human and mouse and are quite comprehensive. However, it is still difficult to recover from the data in the form it is currently provided in what context each transcript is expressed. This was still lucking back at 2009 when we started our work in generating the first comprehensive RNA-Seq based lincRNA catalog in human, and thus motivated it.

Distinguishing between protein coding and noncoding genes

A second challenge and perhaps a more difficult one is determining whether a novel transcript encodes a protein [107]. Methods that approach such task usually rely on one or several of the following premises: (1) long noncoding sequence are unlikely to randomly contain ORFs longer than 200 bp (200 bp is approximately two standard deviations above the average length of an ORFs found in 500 bp of random sequence ; [114]); (2) The nucleotide sequence of functional ORFs is constrained by non-random codon usage [115]; (3) Evolutionary conserved ORFs are characterized by relaxation of constrain on the third position of codons; (4) Coding sequence usually comprise known protein coding domains documented in protein database; (5) Coding transcripts are bound by the ribosome when translated.

Most of the cDNA based lncRNA catalogs available are manually curated based on ORF length cutoffs, conservation estimates such as those from CRITICA [116], and other characteristics of the ORF [68, 70, 117, 118]. In species that are less explored from a genomic perspective, ORF length cutoffs are mainly used (Arabidopsis [119], Maize [120], *Plasmodium falciparum [121]*). Many of the available RNA-Seq based catalogs apply an automated procedure such as the coding potential calculator (CPC) [101, 110, 122]. CPC is a support vector machine algorithm that estimates coding potential by evaluating features on transcript and ORF length as well as similarity to sequences in known protein databases but without using any measure of evolutionary conservation [123]. Coding Potential Assessment Tool (CPAT) takes a similar approach and uses an alignment free logistic regression model on different transcript features [124].

Other work, including ours, applied a combination of searches against protein domain databases and coding potential estimates by phyloCSF [125], a supervised method to evaluate the conservation of ORFs based on multiple sequence alignment across large clades (such as mammals) [83, 100, 126] . RNAcode is another similar but unsupervised approach that has been used for annotation [127] [122]. The latter approaches are more sensitive to detect short conserved ORFs that may encode peptides (such as in [128]), while the former approaches are more likely to detect long and newly evolving ORFs. More recently, a lincRNA classification also included analysis of ribosome profiling and polysome association data that provide an estimate of the transcript being translated by the ribosome [122]. Such data, however, should be analyzed with care as discussed in **Section 1.5**.

1.4 Functional mechanism of lncRNAs: lessons from case studies

While the vast majority of lincRNA catalogs still remain unexplored, there is a rapidly growing number of studies describing the involvement of individual lncRNAs in diverse biological systems, development and disease. Their mechanism of action, however, is still mostly un-deciphered. While much of the early work mainly suggested the involvement of lncRNAs in chromatin regulation, it is now becoming clear that the set of genes currently classified as lncRNAs comprise multiple subsets with diverse functions. Overall their potential mechanisms can largely be divided to 3 groups: (1) those that rely on the act of their own transcription or their nascent transcript, (2) those that require the mature transcript but depend on their site of transcription, (3) and those that are independent from their site of transcription. Transcripts in the first two groups are considered as acting in *cis*, as they act at their site of transcription, while the third act in *trans*. Here, I will briefly survey the main mechanistic themes that were demonstrated through studies of specific examples to date. As all three mechanisms can affect the same type of process (e.g. epigenetic regulation), I will discuss these examples by the main processes lncRNAs were demonstrated to affect. For a more comprehensive surveys please refer to [106, 129-131].

lncRNAs as epigenetic modulators

The first mechanistic themes of lncRNAs function were discovered through early studies of XIST that was shown to coat the inactive X allele from which it was also transcribed, as well as affect its chromatin state [131, 132]. In addition, early work demonstrated that single stranded RNA (ssRNA) is required for heterochromatin formation mediated by the repressive chromatin modifying complex HP1 [133]. Members of the Polycomb family of chromatin modifiers were shown to bind ssRNA, and ssRNA but not ssDNA was required for the maintenance of the repressive histone modification H3K27me3 and H3K9me3 controlled by these modifiers [134]. This and the lack of a definitive consensus Polycomb binding sequence in mammals, akin to the Polycomb response elements in fruit flies [135], supported the hypothesis that Polycomb recruitment in mammals might be directed by RNA [136, 137]. Taken together these studies suggested that other lncRNAs are involved in epigenetic regulation, and perhaps guided the next early studies to investigate lncRNA function from that perspective [129].

HOTAIR from the HOXC cluster, which represses HOXD genes and affects their repressed chromatin state, was the first lncRNA shown to bind the Polycmb repressive complex PRC2 [42]. This pioneering work introduced the idea that a lncRNA can be involved in epigenetic regulation while acting in *trans*. Soon after, the mechanistic aspects of an RNA's effect on PRC2 activity started to be uncovered, when Xist and its short isoform repeat A (RepA) were also shown to bind PRC2 [138].

A current working model suggests that in the initiation of XIC, RepA is first transcribed from the Xist first exon, recruits PRC2 to the Xist promoter and induces its transcription [132, 138]. Full length Xist then binds PRC2 through its repeat A and recruits it to the rest of the inactive X chromosome. The RepA/Xist PRC2 interaction can be competitively inhibited by the binding of antisense transcript Tsix to PRC2, as happens on the active X. The spreading of the Xist-PRC2 complex across the XI involves the binding of Xist through its repetitive binding domain RepC to another polycomb protein YY1 that tethers Xist co-transcriptionally to a 'nucleation center' in the XI [139]. The following spreading to the distal regions is then midiated by the 3 dimensional conformation of the genome that brings secondary nucleation centers in proximity to the primary one and enable the deposition of PRC2 along the XI [140].

Notably, several steps of this model are still debated, as some of studies *in-vivo* and in human systems describe different behaviors of Xist, RepA and their recruitment of PRC2 [131, 141]. For example, using mouse genetics it was shown that while not being able to repress the XI, XIST can still associate with chromatin and spread when the RepA repeat is deleted (which is not consistent with the necessity of RepA for loading) [142]. Indeed, the large body of work that investigates "the mother of all lncRNAs" over two decades tells that deciphering the mechanism of one lncRNA may be very puzzling.

lncRNAs were reported to bind many other types of chromatin modifiers, including activating modifiers, and were shown to affect their activity based on the fact that a knockdown of the lncRNA had a negative effect on the binding of a chromatin modulator or the deposition of the corresponding chromatin mark on a target locus. These lncRNA-chromatin modulator pairs include: Air and G9a [143], ANRIL and CBX7 [144], Anril and Suz12 [145], Mistral and MLL1 [146], HOTAIR and the LSD1/CoREST/REST complex [147], HOTTIP and WDR5 [148], and a few activating lncRNAs with Mediator [149]. Other interactions were reported between Kcnq1ot1 and G9a and PRC2 [150], SRA RNA with CTCF [151] as well as interactions of 30% of expressed mouse embryonic stem cells (mES) lincRNA with at least one of 11 chromatin modifiers [63]. Large scale and genome wide studies further showed that numerous lncRNAs bind PRC2 and other repressive modifiers [62, 152, 153]. However, the extent to which many of these RNA-protein interactions are direct and specific is still controversial [141, 154].

One suggestive model for the involvement of lncRNAs in chromatin state regulation is that lncRNAs provide target specificity to the unspecific and ubiquitous chromatin modifiers by binding them and recruiting them to their target genomic loci (**Figure 1.4**) [129]. This target recognition may either be primarily encoded in the primary RNA sequence, that can base pair or form a triplex with the targeted DNA, or on features of the RNA secondary or tertiary structure. The chromatin state regulation model is specifically appealing if lncRNAs are acting in proximity to their site of transcription as it provides a possible explanation for how the target DNA is being recognized by the non-specific chromatin modulators (for example by the nascent transcript tethering the modifying complex as happens in fission yeast [155]).

In principle, lncRNAs can also guide chromatin modifiers to distal targets through indirect interactions such as base pairing with nascent transcripts or binding to DNA-binding proteins (**Figure 1.4**) [107]. Indeed, these are concordant with a broader model that suggests that lncRNA can act as molecular scaffolds that bring into proximity several components to form a unique functional complex by their binding to distinct domains on the RNA (**Figure 1.4**) [156, 157]. This model is also consistent with a patchy conservation pattern of lncRNAs [83, 110] (discussed in **Section 1.4**), as domains across the molecules can in principle be swapped, while preserving function. This model was proposed to describe yeast telomerase RNA, when it was shown that protein binding domains can be moved to three distinct locations on the RNA with retention of telomerase function *in vivo* [157]. More recent studies propose that the *cis* and *trans* model can be reconciled, by showing that some lincRNAs use the three dimensional structure of the genome to bring distal sites and on other chromosomes into proximity [140, 158]. These studies show that the *trans* sites (i.e. not in the locus transcribing the lncRNA) bound by the lncRNA are in fact proximal to the transcribing locus as demonstrated by genome-wide chromosome conformation capture (Hi-C) [140, 159] or by RNA-FISH [158].

lincRNAs as transcriptional regulators

In addition to regulating gene expression by interacting with epigenetic factors, lncRNAs can influence transcription directly. Models have been proposed for how lncRNAs can prevent transcription factors (TFs) from binding to their DNA targets, including: (1) by acting as decoys of the DNA binding element and binding the DNA binding domain of the TF directly thus sequestering it from its targets (**Figure 1.4**), or (2) competitively inhibiting the TF by binding the targeted DNA elements [106]. The first model was proposed to explain how the binding of lncRNA GAS5 to Glucocorticoid Receptors prevents them from binding their responsive genes [160], and how the lncRNA PANDA which binds the TF NF-YA sequesters it away of its pro-apoptotic targets [161]. This model is also demonstrated in the case of the cytoplasmic lncRNA NRON that binds the TF NFAT and retains it in a cytoplasmic ribonucleic complex. This complex in turn induces NFAT's phosphorylated form thus preventing it from shuttling into the nucleus (as only the un-phosphorylated form is imported to the nucleus) [40, 162].

lncRNAs can also directly interfere with the activity of RNA-Polymerase II (Pol II). This was demonstrated by studies showing that RNA transcribed from a SINE family of transposable element, and in specific Alu elements in humans and B2 elements in mouse, can act in *trans* during a heat shock response and bind Pol II. This prevents the formation of the pre initiation complex at promoters thus making it transcriptionally inert [163, 164] [165]. lncRNAs can also act as co-activators [166] or co-repressors [167]. For example, the SRA RNA is a bi-functional transcript (i.e. has a ncRNA isoform and a protein coding mRNA) that serves as a co-activator of a number of steroid hormone receptors and is present in co-regulator ribonucleoprotein complexes [166, 168].

lncRNAs as post transcriptional regulators

Being a key and dynamic organelle the nucleus includes several membraneless compartments, commonly referred to as nuclear bodies. These nuclear-bodies were implicated in a variety of functions, including: sequestration and modification of proteins, RNA processing, ribonucleoprotein complex assembly, as well as epigenetic regulation [169]. Localization to such nuclear bodies is perhaps the most prominent feature of the highly abundant and early discovered lncRNAs NEAT1 and NEAT2 (MALAT1) [170].

NEAT1 interacts with paraspeckles proteins such as p54/NONO and PSP and is required for the formation and stability of paraspeckles [171-174], nuclear bodies involved in the retention of mRNAs that undergo Adenosine-to-Inosine editing, possibly by relying on NEAT1's continued transcription [175]. MALAT1, neighboring NEAT1, localizes to nuclear speckles, and is involved in regulation of context specific alternative splicing by binding and distributing serine/arginine (SR) splicing factors to transcription sites and by regulating their phosphorylation [176].

Malat1 and another lncRNA Tug1 were also shown to participate in trafficking genes between silencing (Polycomb bodies) and activating (interchromatin granules) nuclear compartments by binding to the PRC1 component Cbx4/Pc2 in its un-methylated or methylated forms, respectively [177]. In addition, Xist and Kcnq1ot1 were shown to localize their targets, the XI and Kcnq1ot1 imprinted domain to the perinucleolar compartment to maintain their silencing [150, 178]. Taken together, lncRNAs can regulate the localization of genes and proteins to specific compartments. Yet, it should be noted that the given examples are of lncRNAs that are an order of magnitude more expressed than the average lncRNA and therefore involvement in such global processes may be less common among lowly expressed lncRNAs.

As with other RNA such as small RNAs, lncRNA can regulate mRNA processing and stability by base pairing recognition. Alternative splicing may be regulated by transcription of antisense transcripts such as NATs that bind to a complementary sense strand to form a duplex and mask a splice site [91]. lncRNAs can also affect splicing by binding components of the splicing machinery and hindering spliceosome formation, as in the case of MIAT that sequesters the splicing factor SF1 [179] .

lncRNA can also be involved in the destabilization of mRNAs in the cytoplasm. One such example are the half-STAU1-binding site RNAs (1/2 sbsRNA) that are lncRNAs containing complementary Alu elements that base pair with Alu elements on the 3' UTRs of mRNAs targets of a decay process mediated by STAU1 and are necessary for the process [180]. lncRNAs were also described as repressing microRNA function, either by binding to the mRNA target and masking the microRNA binding site [91], or by binding the microRNA itself and sequestering it from its targets, as has been proposed for a number of expressed pseudogenes and competing endogenous RNAs [181, 182]. While the expression levels of pseudogenes is relatively low in order to have a fine tuning regulatory effect on microRNA regulation [183], such a role of a 'microRNA sponge' was suggested as more plausible for the thousands of recently discovered circular mRNAs which are highly expressed in the cytoplasm and are more resistant to degradation due to their circular form [184, 185].

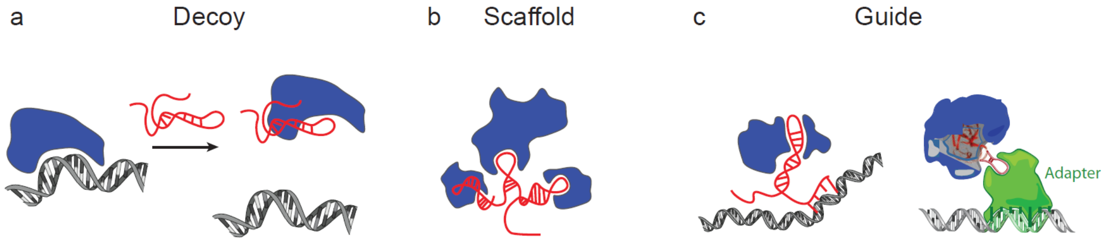


Figure 1.4| Models for lncRNA mechanisms. Adapted with permission from *Rinn and Chang*[129].(a) lncRNAs can act as decoys that titer a protein (e.g. a transcription factor) away from binding its target (e.g. DNA binding site). (b) lncRNAs can act as scaffolds that bring several protein complexes into proximity. (c) lncRNA can act as guides that bring a protein complex to its target DNA site. The lncRNA can bund the DNA directly or bind an adaptor that specifically binds the target site. lncRNA (red), DNA (gray helix), proteins (blue/green).

1.5 Challenges in studying lncRNAs

After providing an historical view of the “birth” of the lncRNA field, how were they cataloged initially and some of their suggested mechanisms of their action, I will now summarize the key areas of discussion and investigation with respect to lncRNAs as a functional class. While some points may seem obscure, these have been guiding the focus on how lncRNAs’ functions are being investigated by the wide community.

Many lncRNAs have low abundance. While the most well-studied lncRNAs are substantially expressed, the vast majority of lncRNAs have low abundance (on average approximately an order of magnitude less than coding genes, [100]). This poses a challenge to models suggesting that lncRNAs can act as *trans* guides of chromatin modifiers to their targets or as competitors of microRNA targets, as it is unclear how their low concentration will be sufficient to find their binding partner and impose a real effect from a stoichiometric perspective. A reconciling hypothesis suggests that lincRNAs expression only appears low when measured in bulk and that there is actually substantial expression in a sub population of cells [186]. Alternatively, a model in which many of the low abundance lncRNAs work in *cis* near their site of transcription may explain how a transcript can pursue its function and find its interacting partners, while having only one or few copies per cell. One such example is the lincRNA HOTTIP that transcribed at 1-5 copies per cell and regulates the chromatin state of its nearby loci [148]. Moreover, a seemingly *trans* effect may be in fact activity proximal to the transcription site if three dimensional chromosome architecture is taken into consideration. This model was suggested to explain previous observations on XIST [140] and the lncRNA FIRRE [158].

Evolutionary conservation. The second common concern that dismisses lncRNAs as a functional class is that unlike other types of functional RNA (e.g. tRNA, rRNA, snoRNA, microRNA) they demonstrate low rate of cross-species conservation. Yet, multiple studies on the subsets of lncRAs that resemble mRNA (i.e. capped, poly adenylated and spliced) found evidence of purifying selection acting on lncRNAs exons and promoters relative to a neutral substitution rate [58, 70, 117]. The extent of this conservation was debated for some time [71], but overall it suggested these subsets included functional molecules. Later analysis of lincRNAs' exons showed that their overall conservation rates mostly relied on short patches of conserved sequence [83].

With the growing availability of lncRNA catalogs across species this focus shifted with the realization that a very small fraction of the lncRNA transcripts in one species have a transcribed orthologous transcript in another species [100, 110, 187]. Since BLAST is not sensitive enough to detect homology between lncRNAs [107], we and others have used syntenic mapping to map corresponding loci between genomes of two species based on conserved gene order (synteny). While the level of sequence homology was relatively low, the existence of syntenic orthologs has been demonstrated to be a useful criterion to prioritize candidates for functional studies [158, 188] and was demonstrated as a successful criterion in discovering functional lincRNAs with conserved function in-vivo from zebrafish to mammals [110] .

One possible explanation of the overall low conservation of lncRNAs, is that the studied sets are burdened by a large fraction of non-functional transcripts. Alternatively, lncRNA might be under pressure to conserve structure rather than sequence and current methods used to evaluate their conservation are not tailored to detect such constraints as they focus on primary sequence similarity. Even the functional and conserved repeats within XIST are interspersed within poorly conserved sequence, resulting in overall low conservation scores [189]. Overall, future understanding of the selective pressures acting on lncRNAs will enable the development of methods to estimate conservation levels that are tailored for this gene type, and those in turn will provide a better understanding on how lncRNAs evolve.

Cis vs. trans activity. As noted above, an early debate was on whether most functional lncRNA are acting in *cis* or *trans* [190]. While most other class of RNA (e.g. TERC, snoRNAs, RnaseP, microRNA etc') act in *trans,* the early examples of XIST and other lncRNAs involved in imprinting repeatedly demonstrated activity in *cis,* that is, in proximal to their transcription site. This and other characteristics noted above supported the idea that many other lincRNAs are also acting in *cis*. While several studies tried to address this [117, 191] to conclusively distinguish between the two, it is necessary to apply both loss-of-function and rescue gain-of-function perturbations. Other approaches such as single molecule imaging [192] and chromatin localization of RNA [140, 193, 194] can also be revealing as they can provide information for where the transcript is localized relative to its site of transcription and therefore guide following studies (as in [139, 158, 195, 196]).

Act of transcription or functional molecules? Another debate that relates to lncRNAs being low abundant and lowly conserved is the possibility that the act of transcription rather than the transcript itself serves as the functional unit. This is also consistent with the high number of isoforms observed for some of these transcripts. To distinguish between the two possibilities one may show that hindering the mature transcript (such as inserting transcription termination signal downstream to the promoter as in [197]) while maintaining transcription has a phenotypic outcome, or by showing that a genetic perturbation can be rescued in *trans* by a transgene [198].

Sub-cellular localization of lncRNAs. It is currently not settled whether lncRNAs are primarily localized to the nucleus or cytoplasm [107, 118]. Localization of their majority to the nucleus may relax the concerns that many of these encode peptides or newly evolving protein coding genes [199, 200] . RNA-Seq from specific cellular fractions (e.g., nuclear, cytoplasmic, chromatin) was previously used to approach this question, however, since this data provide enrichments relative to different population of RNA in each fraction it cannot be interpreted conclusively [107, 118]. One possibility to resolve this is by using single molecule RNA-FISH that provides absolute rather than relative estimates of counts.

Newly evolving protein coding gene? Recent application of ribosome profiling, an assay in which the RNA sequences bound and protected by ribosomes is isolated and sequenced by RNA-Seq, suggest that some of the currently annotated lncRNAs may encode small peptides and newly evolving proteins [199, 200] . Specifically, one study suggested that a large fraction on mES lincRNAs are bound by the ribosome and are perhaps coding. However, this should be interpreted with care as only ~20% of mES were abundant enough to be evaluated in this study, and since both H19 and MALAT1, well studied lnRNAs that function in the nucleus, were also detected on the ribosome. lncRNAs may bind ribosomes either purposely to regulate translation activity, or accidently as they contain all the recognition features of mRNAs. They can even encode a peptide that may not be functional. Alternatively, some may encode small functional peptides [128]. Yet, in a collaboration with the Saghatelian lab we evaluated proteomics data enriched for small peptides from K562 cells, which suggested that less than 0.5% of lincRNAs expressed in that cell may encode peptides [201]. Overall, it seems that ribosome profiling can be useful for the task of noncoding classification only when coupled to the proper computational analysis [202, 203] or when accounting for codon periodicity [204].

Are lncRNA essential genes? Only recently, the function of several lncRNAs was evaluated by classic genetic knockout studies using animal models. Surprisingly, neither knockout of MALAT1 nor NEAT1, two of the most highly abundant and functionally characterized lncRNAs, appeared to have a distinguishing robust phenotype in mice [205, 206]. However, other studies showed that the loss of each of 15 different lncRNAs can cause developmental defects in either zebrafish, chicken or mouse (reviewed at [130]), including a recent study of 18 lincRNA knockout mice, three of which (~13%) were required for proper development or life [207].

1.6 Single molecule imaging as a tool for studying lncRNAs' function

Single molecule RNA-FISH

Our current understanding of lncRNAs expression mostly relies on high-throughput bulk measurements using microarrays technologies or cDNA sequencing. More recently, single cell polymerase chain reaction (PCR) based measurements and single cell RNA-Seq provide the opportunity to explore the population heterogeneity of individual transcripts [208]. These methods, however, still suffer from low sensitivity in targeting low abundance transcripts such as lncRNAs [209]. Single molecule RNA fluorescent in situ hybridization (FISH) provides an alternative in which one can simultaneously obtain single cell quantitative and spatial measurements of a transcript [192].

RNA FISH is an imaging based technique to detect and localize single RNA transcripts in fixed cells based on the Watson-Crick hybridization of complementary nucleic acid labeled probes [210]. First, cells are fixed and permeabilized, then soaked with an excess of labeled probe(s) targeting the transcript of interest, unbound probes are then washed away, and cells are imaged. RNA-FISH protocols mostly differ in the nucleic acid used to generate probes and the type of labeling scheme used to detect the probe via microscopy. The development of florescence based techniques for probe labeling [211, 212] was a significant advancement to the original in situ hybridization (ISH) methods that targeted DNA [213] and RNA [214] using radioactively labeled probes. In traditional FISH, fluorophores are coupled to oligonucleotides by enzymatic means such as nick translation or *in-vitro* transcription [215] which result in a random distribution of the fluorescent dye along the probe. While eliminating the challenges in the handling and the stability of radioactively labeled probes, traditional FISH methods still suffer from low spatial resolution due to cellular auto fluorescence as well as low consistency of labeling across experiments and do not provide single molecule resolution [192, 210].

Improvements in probe designs, imaging technology and image processing software enabled the development of single molecule RNA-FISH (smRNA-FISH) [216, 217]. The key improvement was the replacement of a single randomly labeled long probe with a set of consistently labeled short probes (first with a few 50 pb probes labeled with 3-5 fluorophores [216], and later with 32-48 single label 20bp probes [217]). The binding of multiple probes localizes a sufficient number of fluorophores to the target RNA such that a single RNA molecule is reliably visible as a fluorescent spot via fluorescent microscopy [216, 217]. This provides higher sensitivity as it is unlikely that all of the oligonucleotides target inaccessible regions of the transcript. Using multiple short probes also provide higher specificity since off-target binding of a single oligonucleotide in the probe pool will either be undetectable or clearly distinguishable relative to brighter spots corresponding to the true RNA [217].

Current methods for RNA-FISH can be broadly divided to two main classes: those which measure the signal directly and those which rely on signal amplification. In the direct methods the probe itself is labeled [216, 217]. These direct methods face specificity challenges when the target is too short as well as a limitation of low signal when the total number of targeting fluorophores is small. Later methods try to improve low signal by using more stable oligonucleotides or signal amplification strategies [218-220].

To increase target specificity, a recent approach uses branched DNA (bDNA) technology to enhance sm-FISH [221]. In bDNA smRNA-FISH, signal is detected and amplified only when a pair of consecutive oligonucleotides binds the target of interest thus eliminating the possibility of signal detected from a single or few non-paired oligonucleotides binding to an off-target. In brief, primary pairs of unlabeled oligonucleotides bind the target, a long preamplifier DNA molecule binds such pairs and serves as a trunk to which several fluorescently labeled amplifier probes bind. The bDNA smRNA-FISH multi-step protocol was recently compared [221] to the traditional smRNA-FISH [216] and was shown to significantly enhance the signal-to-noise ratio resulting with clearer signal with similar detection accuracy [221]. While offering significant advancement in automated high-throughput and image analysis in the same levels of accuracy, sm-bDNA FISH still suffers from limited detection of nuclear transcripts (due to low yield in entrance of the long (>100bp) trunk probes to the nucleus) and its current state offers only simultaneous imaging of four different RNA molecules [221].

Other recent variations of smRNA-FISH use combinatorial labeling of the probe directly to detect multiple transcripts at a time. These include iceFISH which detects up to 20 transcripts at a time and was used to measure gene expression and chromosome structure at once [222] as well as the coupling of smRNA-FISH with optical super-resolution microscopy (SRM) to detect and measure up to 32 genes simultaneously [223]. Another finer resolution approach is SNP-FISH which can detect a specific allele by coupling simultaneous imaging of multiple fluorescent dyes with image analysis of co-localized spots [224]. Taken together, these advances are poising smRNA-FISH to become a high-throughput tool to study gene expression in a single cell with spatial resolution.

RNA-FISH: a tool for understanding lncRNA functions

One area of research that is less explored is the application of smRNA-FISH to detect less abundant transcripts such lncRNAs. Indeed, RNA-FISH was instrumental in detecting the more abundant and well-studied lncRNAs, including XIST, MALAT1, NEAT1, MIAT, and GAS5, uncovering their unique cellular localization patterns and deciphering their interaction with other molecules and its effect on their localization [160, 170, 225, 226]. Offering higher resolution and sensitivity smRNA-FISH was also applied to detect low abundance lncRNAs [62, 158, 195, 196, 227-230] as well as to estimate the population abundance of lncRNAs that are expressed on average at one or less copies per cell (as in HOTTIP [148]).

Besides the detection of individual molecules, sm-FISH can be used to address broader open questions in a systems level. Specifically, if applied to a large and representative set of lncRNAs it can in principle: (1) uncover their cellular localization patterns and reveal common themes, (2) provide an absolute measure to whether lncRNAs are predominantly in the nucleus, an open question that was so far addressed only by relative measurements of RNA-Seq on cellular compartments [118], (3) reveal localization patterns during specific cell states such as mitosis, or in response to perturbation, (4) illuminate on the interaction of lncRNAs with their neighboring gene by detecting them simultaneously.

Simultaneous analysis of lincRNAs and other molecules may help address the on-going debate of whether lncRNAs that appear to be involved in epigenetic regulation (based on their interaction with chromatin modifiers) are acting in *cis* or in *trans*. Indeed, using multi-color smRNA- FISH, perturbations and single cell correlation analysis, lincHox-a1 was recently shown to repress its neighbor Hoxa1 in *cis* [196]. Moreover, using perturbation and RNA-FISH, Jpx, a neighbor of Xist, was shown to activate Xist expression, surprisingly, in *trans* [195].

Taken together sm-FISH has a great potential to advance studies of lncRNAs biology, especially in the early steps of investigating a new gene. Yet, it is still not routinely applied to study new lncRNA and a comprehensive survey of the cellular patterns of lncRNAs is still not available.

1.7 Research objectives

Application of emerging technologies has a great impact on the discovery and understanding of lncRNAs biology. In this thesis I present two studies in which I applied an emerging technology for the first time to systematically study human lncRNAs and addressed a key need by our community at a given time. I have the following specific objectives:

1. Annotate, characterize and catalog human lincRNAs using a large compendium of RNA-Seq. In chapter 2 I describe my work to develop and apply an integrative approach to annotate over 8000 lincRNAs using ab initio transcript reconstruction from RNA-Seq of 20 different tissue and cell types. I then cataloged and characterized these lincRNAs by over 30 properties, revealing their striking tissue specificity and additional traits.

2. Characterize the localization patterns and cell-to-cell variability of human lncRNAs in a single cell and single molecule resolution using smRNA-FISH. In chapter 3 I describe my work in which I used smRNA-FISH to survey 61 lncRNAs and catalog their abundance and cellular localization patterns in three human cell types. This reveals diverse localization patterns of lncRNAs as well as a bias toward nuclear localization. In addition, this highlighted the challenges with applying this technology on the specific class of lncRNA and presented a strategy to approach these challenges.

1.8 References

1. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
2. Brenner, S., F. Jacob, and M. Meselson, *An unstable intermediate carrying information from genes to ribosomes for protein synthesis.* Nature, 1961. **190**: p. 576-581.
3. Holley, R.W., et al., *Structure of a Ribonucleic Acid.* Science, 1965. **147**(3664): p. 1462-5.
4. Holley, R.W., et al., *Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid.* The Journal of biological chemistry, 1965. **240**: p. 2122-8.
5. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.
6. Nissen, P., et al., *The structural basis of ribosome activity in peptide bond synthesis.* Science, 2000. **289**(5481): p. 920-30.
7. Muth, G.W., L. Ortoleva-Donnelly, and S.A. Strobel, *A single adenosine with a neutral pKa in the ribosomal peptidyl transferase center.* Science, 2000. **289**(5481): p. 947-50.
8. Warner, J.R., et al., *Rapidly labeled HeLa cell nuclear RNA. I. Identification by zone sedimentation of a heterogeneous fraction separate from ribosomal precursor RNA.* Journal of Molecular Biology, 1966. **19**(2): p. 349-61.
9. Soeiro, R., H.C. Birnboim, and J.E. Darnell, *Rapidly labeled HeLa cell nuclear RNA. II. Base composition and cellular localization of a heterogeneous RNA fraction.* Journal of Molecular Biology, 1966. **19**(2): p. 362-72.
10. Weinberg, R.A. and S. Penman, *Small molecular weight monodisperse nuclear RNA.* Journal of Molecular Biology, 1968. **38**(3): p. 289-304.
11. Maxwell, E.S. and M.J. Fournier, *The small nucleolar RNAs.* Annual review of biochemistry, 1995. **64**: p. 897-934.
12. Berget, S.M., et al., *Spliced segments at the 5' termini of adenovirus-2 late mRNA: a role for heterogeneous nuclear RNA in mammalian cells.* Cold Spring Harbor symposia on quantitative biology, 1978. **42 Pt 1**: p. 523-9.
13. Matera, A.G., R.M. Terns, and M.P. Terns, *Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs.* Nature reviews. Molecular cell biology, 2007. **8**(3): p. 209-20.
14. Guerrier-Takada, C., et al., *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.* Cell, 1983. **35**(3 Pt 2): p. 849-57.
15. Kruger, K., et al., *Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena.* Cell, 1982. **31**(1): p. 147-57.
16. Zaug, A.J., P.J. Grabowski, and T.R. Cech, *Autocatalytic cyclization of an excised intervening sequence RNA is a cleavage-ligation reaction.* Nature, 1983. **301**(5901): p. 578-83.
17. Breaker, R.R., *Riboswitches and the RNA World.* Cold Spring Harbor Perspectives in Biology, 2010.
18. Gilbert, W., *The RNA World.* Nature, 1986(319): p. 618.
19. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
20. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.* Nature, 2000. **403**(6772): p. 901-6.
21. Lee, R.C. and V. Ambros, *An extensive class of small RNAs in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 862-4.
22. Lau, N.C., et al., *An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 858-62.
23. Lagos-Quintana, M., et al., *Identification of novel genes coding for small expressed RNAs.* Science, 2001. **294**(5543): p. 853-8.
24. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
25. Aravin, A.A., G.J. Hannon, and J. Brennecke, *The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race.* Science, 2007. **318**(5851): p. 761-4.
26. Brannan, C.I., et al., *The product of the H19 gene may function as an RNA.* Molecular and Cellular Biology, 1990. **10**(1): p. 28-36.
27. Bartolomei, M.S., S. Zemel, and S.M. Tilghman, *Parental imprinting of the mouse H19 gene.* Nature, 1991. **351**(6322): p. 153-5.

28. Gabory, A., H. Jammes, and L. Dandolo, *The H19 locus: role of an imprinted non-coding RNA in growth and development.* BioEssays : news and reviews in molecular, cellular and developmental biology, 2010. **32**(6): p. 473-80.
29. Cai, X. and B.R. Cullen, *The imprinted H19 noncoding RNA is a primary microRNA precursor.* RNA, 2007. **13**(3): p. 313-6.
30. Keniry, A., et al., *The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r.* Nature cell biology, 2012. **14**(7): p. 659-65.
31. Brown, C.J., et al., *A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome.* Nature, 1991. **349**(6304): p. 38-44.
32. 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, 1992. **71**(3): p. 527-42.
33. Brockdorff, N., et al., *The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus.* Cell, 1992. **71**(3): p. 515-26.
34. Penny, G.D., et al., *Requirement for Xist in X chromosome inactivation.* Nature, 1996. **379**(6561): p. 131-7.
35. Wutz, A. and R. Jaenisch, *A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation.* Molecular Cell, 2000. **5**(4): p. 695-705.
36. Lee, J.T., *Epigenetic regulation by long noncoding RNAs.* Science, 2012. **338**(6113): p. 1435-9.
37. Lee, J.T., L.S. Davidow, and D. Warshawsky, *Tsix, a gene antisense to Xist at the X-inactivation centre.* Nature Genetics, 1999. **21**(4): p. 400-4.
38. Lyle, R., et al., *The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1.* Nature Genetics, 2000. **25**(1): p. 19-21.
39. Young, T.L., T. Matsuda, and C.L. Cepko, *The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina.* Curr Biol, 2005. **15**(6): p. 501-12.
40. Willingham, A.T., et al., *A strategy for probing the function of noncoding RNAs finds a repressor of NFAT.* Science, 2005. **309**(5740): p. 1570-3.
41. Kanduri, C., N. Thakur, and R.R. Pandey, *The length of the transcript encoded from the Kcnq1ot1 antisense promoter determines the degree of silencing.* The EMBO journal, 2006. **25**(10): p. 2096-106.
42. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.
43. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
44. Pheasant, M. and J.S. Mattick, *Raising the estimate of functional human sequences.* Genome Research, 2007. **17**(9): p. 1245-53.
45. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
46. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
47. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.
48. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.
49. Rinn, J.L., et al., *The transcriptional activity of human Chromosome 22.* Genes Dev, 2003. **17**(4): p. 529-40.
50. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
51. Struhl, K., *Transcriptional noise and the fidelity of initiation by RNA polymerase II.* Nature Structural & Molecular Biology, 2007. **14**(2): p. 103-5.
52. Wang, J., et al., *Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs.* Nature, 2004. **431**(7010): p. 1 p following 757; discussion following 757.
53. Eddy, S.R., *Non-coding RNA genes and the modern RNA world.* Nature reviews. Genetics, 2001. **2**(12): p. 919-29.
54. Shoemaker, D.D., et al., *Experimental annotation of the human genome using microarray technology.* Nature, 2001. **409**(6822): p. 922-7.
55. Kapranov, P., et al., *Large-scale transcriptional activity in chromosomes 21 and 22.* Science, 2002. **296**(5569): p. 916-9.

56. Agarwal, A., et al., *Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays.* BMC genomics, 2010. **11**: p. 383.
57. van Bakel, H., et al., *Most Dark Matter Transcripts Are Associated With Known Genes.* PLoS Biol, 2010. **8**(5): p. e1000371.
58. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
59. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
60. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
61. Bernstein, B.E., A. Meissner, and E.S. Lander, *The mammalian epigenome.* Cell, 2007. **128**(4): p. 669-81.
62. 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, 2009. **106**(28): p. 11667-72.
63. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011. **477**(7364): p. 295-300.
64. Huarte, M., et al., *A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response.* Cell, 2010. **142**(3): p. 409-19.
65. Loewer, S., et al., *Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells.* Nat Genet, 2010. **42**(12): p. 1113-7.
66. Sun, L., et al., *Long noncoding RNAs regulate adipogenesis.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(9): p. 3387-92.
67. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.
68. Ravasi, T., et al., *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.* Genome Res, 2006. **16**(1): p. 11-9.
69. Maeda, N., et al., *Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs.* PLoS Genet, 2006. **2**(4): p. e62.
70. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.
71. Marques, A.C. and C.P. Ponting, *Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness.* Genome Biol, 2009. **10**(11): p. R124.
72. Mercer, T.R., et al., *Specific expression of long noncoding RNAs in the mouse brain.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 716-21.
73. Dinger, M.E., et al., *Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation.* Genome Res, 2008. **18**(9): p. 1433-45.
74. Pang, K.C., et al., *Genome-wide identification of long noncoding RNAs in CD8+ T cells.* J Immunol, 2009. **182**(12): p. 7738-48.
75. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell. **143**(1): p. 46-58.
76. Jia, H., et al., *Genome-wide computational identification and manual annotation of human long noncoding RNA genes.* RNA, 2010. **16**(8): p. 1478-87.
77. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.
78. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nat Methods, 2011. **8**(6): p. 469-77.
79. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
80. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
81. Burrows, M. and D.J. Wheeler, *A block sorting lossless data compression algorithm*, in *Technical Report*1994, DEC, Digital Systems Research Center: Palo Alto, California.
82. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature Biotechnology, 2011. **29**(7): p. 644-52.
83. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.

84. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotech, 2010. **28**(5): p. 511-515.
85. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.
86. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature Biotechnology, 2013. **31**(1): p. 46-53.
87. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nature methods, 2010. **7**(12): p. 1009-15.
88. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
89. Katayama, S., et al., *Antisense transcription in the mammalian transcriptome.* Science, 2005. **309**(5740): p. 1564-6.
90. He, Y., et al., *The antisense transcriptomes of human cells.* Science, 2008. **322**(5909): p. 1855-7.
91. Faghihi, M.A. and C. Wahlestedt, *Regulatory roles of natural antisense transcripts.* Nature reviews. Molecular cell biology, 2009. **10**(9): p. 637-43.
92. Louro, R., A.S. Smirnova, and S. Verjovski-Almeida, *Long intronic noncoding RNA transcription: expression noise or expression choice?* Genomics, 2009. **93**(4): p. 291-8.
93. Heo, J.B. and S. Sung, *Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA.* Science, 2011. **331**(6013): p. 76-9.
94. Pink, R.C., et al., *Pseudogenes: pseudo-functional or key regulators in health and disease?* RNA, 2011. **17**(5): p. 792-8.
95. Tay, Y., J. Rinn, and P.P. Pandolfi, *The multilayered complexity of ceRNA crosstalk and competition.* Nature, 2014. **505**(7483): p. 344-52.
96. Seila, A.C., et al., *Divergent transcription from active promoters.* Science, 2008. **322**(5909): p. 1849-51.
97. Seila, A.C., et al., *Divergent transcription: a new feature of active promoters.* Cell Cycle, 2009. **8**(16): p. 2557-64.
98. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
99. Preker, P., et al., *RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.* Science, 2008. **322**(5909): p. 1851-1854.
100. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & development, 2011. **25**(18): p. 1915-27.
101. Sigova, A.A., et al., *Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(8): p. 2876-81.
102. Kim, T.-K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-187.
103. Wang, D., et al., *Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA.* Nature, 2011. **474**(7351): p. 390-4.
104. Lam, M.T., et al., *Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription.* Nature, 2013. **498**(7455): p. 511-5.
105. Li, W., et al., *Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation.* Nature, 2013. **498**(7455): p. 516-20.
106. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.
107. Ulitsky, I. and D.P. Bartel, *lincRNAs: genomics, evolution, and mechanisms.* Cell, 2013. **154**(1): p. 26-46.
108. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nature methods, 2006. **3**(3): p. 211-22.
109. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.* Nature, 2011. **469**(7328): p. 97-101.
110. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.
111. Okazaki, Y., et al., *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.* Nature, 2002. **420**(6915): p. 563-73.
112. Pruitt, K., T. Tatusova, and D. Maglott, *Chapter 18, The Reference Sequence (RefSeq) Project.* , in *The NCBI handbook [Internet]*.2002, National Library of Medicine (US), National Center for Biotechnology Information: Bethesda (MD).

113. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Research, 2012. **22**(9): p. 1760-74.

114. Dinger, M.E., et al., *Differentiating protein-coding and noncoding RNA: challenges and ambiguities.* PLoS computational biology, 2008. **4**(11): p. e1000176.

115. Sharp, P.M., et al., *Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity.* Nucleic Acids Research, 1988. **16**(17): p. 8207-11.

116. Badger, J.H. and G.J. Olsen, *CRITICA: coding region identification tool invoking comparative analysis.* Molecular biology and evolution, 1999. **16**(4): p. 512-24.

117. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.

118. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Research, 2012. **22**(9): p. 1775-89.

119. Liu, J., et al., *Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis.* The Plant cell, 2012. **24**(11): p. 4333-45.

120. Boerner, S. and K.M. McGinnis, *Computational identification and functional predictions of long noncoding RNA in Zea mays.* PLoS One, 2012. **7**(8): p. e43047.

121. Broadbent, K.M., et al., *A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs.* Genome Biology, 2011. **12**(6): p. R56.

122. Nam, J.W. and D.P. Bartel, *Long noncoding RNAs in C. elegans.* Genome Research, 2012. **22**(12): p. 2529-40.

123. Kong, L., et al., *CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.* Nucleic Acids Research, 2007. **35**(Web Server issue): p. W345-9.

124. Wang, L., et al., *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.* Nucleic Acids Research, 2013. **41**(6): p. e74.

125. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.* Bioinformatics, 2011. **27**(13): p. i275-i282.

126. Pauli, A., et al., *Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.* Genome Research, 2012. **22**(3): p. 577-91.

127. Washietl, S., et al., *RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data.* RNA, 2011. **17**(4): p. 578-94.

128. Pauli, A., et al., *Toddler: an embryonic signal that promotes cell movement via Apelin receptors.* Science, 2014. **343**(6172): p. 1248636.

129. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annual review of biochemistry, 2012. **81**: p. 145-66.

130. Fatica, A. and I. Bozzoni, *Long non-coding RNAs: new players in cell differentiation and development.* Nature reviews. Genetics, 2014. **15**(1): p. 7-21.

131. Augui, S., E.P. Nora, and E. Heard, *Regulation of X-chromosome inactivation by the X-inactivation centre.* Nature reviews. Genetics, 2011. **12**(6): p. 429-42.

132. Froberg, J.E., L. Yang, and J.T. Lee, *Guided by RNAs: X-inactivation as a model for lncRNA function.* Journal of Molecular Biology, 2013. **425**(19): p. 3698-706.

133. Maison, C., et al., *Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component.* Nature Genetics, 2002. **30**(3): p. 329-34.

134. Bernstein, E., et al., *Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin.* Mol Cell Biol, 2006. **26**(7): p. 2560-9.

135. Schwartz, Y.B. and V. Pirrotta, *Polycomb silencing mechanisms and the management of genomic programmes.* Nature reviews. Genetics, 2007. **8**(1): p. 9-22.

136. Bernstein, E. and C.D. Allis, *RNA meets chromatin.* Genes Dev, 2005. **19**(14): p. 1635-55.

137. Simon, J.A. and R.E. Kingston, *Mechanisms of polycomb gene silencing: knowns and unknowns.* Nat Rev Mol Cell Biol, 2009. **10**(10): p. 697-708.

138. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.

139. Jeon, Y. and J.T. Lee, *YY1 tethers Xist RNA to the inactive X nucleation center.* Cell, 2011. **146**(1): p. 119-33.

140. Engreitz, J.M., et al., *The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.* Science, 2013. **341**(6147): p. 1237973.

141. Brockdorff, N., *Noncoding RNA and Polycomb recruitment.* RNA, 2013. **19**(4): p. 429-42.

142. Wutz, A., T.P. Rasmussen, and R. Jaenisch, *Chromosomal silencing and localization are mediated by different domains of Xist RNA.* Nature Genetics, 2002. **30**(2): p. 167-74.

143. Nagano, T., et al., *The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin.* Science, 2008. **322**(5908): p. 1717-20.

144. Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.* Molecular Cell, 2010. **38**(5): p. 662-74.

145. Kotake, Y., et al., *Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene.* Oncogene, 2011. **30**(16): p. 1956-62.

146. Bertani, S., et al., *The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin.* Molecular Cell, 2011. **43**(6): p. 1040-6.

147. Tsai, M.C., et al., *Long noncoding RNA as modular scaffold of histone modification complexes.* Science, 2010. **329**(5992): p. 689-93.

148. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.* Nature, 2011. **472**(7341): p. 120-4.

149. Lai, F., et al., *Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.* Nature, 2013. **494**(7438): p. 497-501.

150. Pandey, R.R., et al., *Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation.* Molecular Cell, 2008. **32**(2): p. 232-46.

151. Yao, H., et al., *Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA.* Genes & development, 2010. **24**(22): p. 2543-55.

152. Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.

153. Kanhere, A., et al., *Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2.* Mol Cell, 2010. **38**(5): p. 675-88.

154. Davidovich, C., et al., *Promiscuous RNA binding by Polycomb repressive complex 2.* Nature Structural & Molecular Biology, 2013. **20**(11): p. 1250-7.

155. Moazed, D., *Small RNAs in transcriptional gene silencing and genome defence.* Nature, 2009. **457**(7228): p. 413-20.

156. Guttman, M. and J.L. Rinn, *Modular regulatory principles of large non-coding RNAs.* Nature, 2012. **482**(7385): p. 339-46.

157. Zappulla, D.C. and T.R. Cech, *Yeast telomerase RNA: a flexible scaffold for protein subunits.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(27): p. 10024-9.

158. Hacisuleyman, E., et al., *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.* Nature Structural & Molecular Biology, 2014. **21**(2): p. 198-206.

159. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

160. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Science signaling, 2010. **3**(107): p. ra8.

161. Hung, T., et al., *Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters.* Nat Genet, 2011. **43**(7): p. 621-9.

162. Sharma, S., et al., *Dephosphorylation of the nuclear factor of activated T cells (NFAT) transcription factor is regulated by an RNA-protein scaffold complex.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(28): p. 11381-6.

163. Espinoza, C.A., et al., *B2 RNA binds directly to RNA polymerase II to repress transcript synthesis.* Nature Structural & Molecular Biology, 2004. **11**(9): p. 822-9.

164. Mariner, P.D., et al., *Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock.* Molecular Cell, 2008. **29**(4): p. 499-509.

165. Yakovchuk, P., J.A. Goodrich, and J.F. Kugel, *B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(14): p. 5569-74.

166. Lanz, R.B., et al., *A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex.* Cell, 1999. **97**(1): p. 17-27.

167. Wang, X., et al., *Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription.* Nature, 2008. **454**(7200): p. 126-30.

168. Lanz, R.B., et al., *Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA).* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(25): p. 16081-6.

169. Mao, Y.S., B. Zhang, and D.L. Spector, *Biogenesis and function of nuclear bodies.* Trends in genetics : TIG, 2011. **27**(8): p. 295-306.

170. Hutchinson, J.N., et al., *A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains.* BMC genomics, 2007. **8**: p. 39.

171. Clemson, C.M., et al., *An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.* Molecular Cell, 2009. **33**(6): p. 717-26.

172. Sasaki, Y.T., et al., *MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(8): p. 2525-30.

173. Sunwoo, H., et al., *MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles.* Genome Research, 2009. **19**(3): p. 347-59.

174. Chen, L.L. and G.G. Carmichael, *Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA.* Molecular Cell, 2009. **35**(4): p. 467-78.

175. Mao, Y.S., et al., *Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs.* Nature cell biology, 2011. **13**(1): p. 95-101.

176. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Molecular Cell, 2010. **39**(6): p. 925-38.

177. Yang, L., et al., *ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs.* Cell, 2011. **147**(4): p. 773-88.

178. Zhang, L.F., K.D. Huynh, and J.T. Lee, *Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing.* Cell, 2007. **129**(4): p. 693-706.

179. Tsuiji, H., et al., *Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1.* Genes to cells : devoted to molecular & cellular mechanisms, 2011. **16**(5): p. 479-90.

180. Gong, C. and L.E. Maquat, *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.* Nature, 2011. **470**(7333): p. 284-8.

181. Poliseno, L., et al., *A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.* Nature, 2010. **465**(7301): p. 1033-8.

182. Cesana, M., et al., *A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA.* Cell, 2011. **147**(2): p. 358-69.

183. Wilusz, J.E. and P.A. Sharp, *Molecular biology. A circuitous route to noncoding RNA.* Science, 2013. **340**(6131): p. 440-1.

184. Memczak, S., et al., *Circular RNAs are a large class of animal RNAs with regulatory potency.* Nature, 2013. **495**(7441): p. 333-8.

185. Hansen, T.B., J. Kjems, and C.K. Damgaard, *Circular RNA and miR-7 in cancer.* Cancer research, 2013. **73**(18): p. 5609-12.

186. Dinger, M.E., et al., *Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.* Briefings in functional genomics & proteomics, 2009. **8**(6): p. 407-23.

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

188. Marin-Bejar, O., et al., *Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2.* Genome Biology, 2013. **14**(9): p. R104.

189. Duret, L., et al., *The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.* Science, 2006. **312**(5780): p. 1653-5.

190. Orom, U.A. and R. Shiekhattar, *Long non-coding RNAs and enhancers.* Curr Opin Genet Dev, 2011. **21**(2): p. 194-8.

191. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011.

192. Itzkovitz, S. and A. van Oudenaarden, *Validating transcripts with probes and imaging technology.* Nature methods, 2011. **8**(4 Suppl): p. S12-9.

193. Chu, C., et al., *Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.* Molecular Cell, 2011. **44**(4): p. 667-78.

194. Simon, M.D., et al., *The genomic binding sites of a noncoding RNA.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(51): p. 20497-502.
195. Tian, D., S. Sun, and J.T. Lee, *The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation.* Cell, 2010. **143**(3): p. 390-403.
196. Maamar, H., et al., *linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis.* Genes & development, 2013. **27**(11): p. 1260-71.
197. Martens, J.A., L. Laprade, and F. Winston, *Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene.* Nature, 2004. **429**(6991): p. 571-4.
198. Grote, P., et al., *The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse.* Developmental cell, 2013. **24**(2): p. 206-14.
199. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
200. Carvunis, A.R., et al., *Proto-genes and de novo gene birth.* Nature, 2012. **487**(7407): p. 370-4.
201. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nature chemical biology, 2013. **9**(1): p. 59-64.
202. Guttman, M., et al., *Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.* Cell, 2013. **154**(1): p. 240-51.
203. Chew, G.L., et al., *Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs.* Development, 2013. **140**(13): p. 2828-34.
204. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.
205. Zhang, B., et al., *The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult.* Cell reports, 2012. **2**(1): p. 111-23.
206. Nakagawa, S., et al., *Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice.* The Journal of cell biology, 2011. **193**(1): p. 31-9.
207. Sauvageau, M., et al., *Multiple knockout mouse models reveal lincRNAs are required for life and brain development.* eLife, 2013. **2**: p. e01749.
208. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nature reviews. Genetics, 2013. **14**(9): p. 618-30.
209. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-40.
210. Raj, A. and S. Tyagi, *Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes.* Methods in enzymology, 2010. **472**: p. 365-86.
211. Bauman, J.G., et al., *A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA.* Experimental cell research, 1980. **128**(2): p. 485-90.
212. Singer, R.H. and D.C. Ward, *Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog.* Proceedings of the National Academy of Sciences of the United States of America, 1982. **79**(23): p. 7331-5.
213. Gall, J.G. and M.L. Pardue, *Formation and detection of RNA-DNA hybrid molecules in cytological preparations.* Proceedings of the National Academy of Sciences of the United States of America, 1969. **63**(2): p. 378-83.
214. Harrison, P.R., et al., *Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA.* FEBS letters, 1973. **32**(1): p. 109-12.
215. Langer, P.R., A.A. Waldrop, and D.C. Ward, *Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes.* Proceedings of the National Academy of Sciences of the United States of America, 1981. **78**(11): p. 6633-7.
216. Femino, A.M., et al., *Visualization of single RNA transcripts in situ.* Science, 1998. **280**(5363): p. 585-90.
217. Raj, A., et al., *Imaging individual mRNA molecules using multiple singly labeled probes.* Nature methods, 2008. **5**(10): p. 877-9.
218. Pare, A., et al., *Visualization of individual Scr mRNAs during Drosophila embryogenesis yields evidence for transcriptional bursting.* Current biology : CB, 2009. **19**(23): p. 2037-42.
219. Kloosterman, W.P., et al., *In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes.* Nature methods, 2006. **3**(1): p. 27-9.
220. Larsson, C., et al., *In situ detection and genotyping of individual mRNA molecules.* Nature methods, 2010. **7**(5): p. 395-7.

221. Battich, N., T. Stoeger, and L. Pelkmans, *Image-based transcriptomics in thousands of single human cells at single-molecule resolution.* Nature methods, 2013. **10**(11): p. 1127-33.

222. Levesque, M.J. and A. Raj, *Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation.* Nature methods, 2013. **10**(3): p. 246-8.

223. Lubeck, E. and L. Cai, *Single-cell systems biology by super-resolution imaging and combinatorial labeling.* Nature methods, 2012. **9**(7): p. 743-8.

224. Levesque, M.J., et al., *Visualizing SNVs to quantify allele-specific expression in single cells.* Nature methods, 2013. **10**(9): p. 865-7.

225. Clemson, C.M., et al., *XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure.* The Journal of cell biology, 1996. **132**(3): p. 259-75.

226. Sone, M., et al., *The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons.* Journal of cell science, 2007. **120**(Pt 15): p. 2498-506.

227. Mohammad, F., et al., *Kcnq1ot1/Lit1 noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region.* Molecular and Cellular Biology, 2008. **28**(11): p. 3713-28.

228. Carpenter, S., et al., *A long noncoding RNA mediates both activation and repression of immune response genes.* Science, 2013. **341**(6147): p. 789-92.

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

230. Bumgarner, S.L., et al., *Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment.* Molecular Cell, 2012. **45**(4): p. 470-82.

Chapter 2 | Integrative Annotation of Human Large Intergenic Non-Coding RNAs Reveals Global Properties and Specific Subclasses

Contributions: Moran N. Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev†, John L. Rinn†. (†) equal contribution.

This chapter originally appeared as: Cabili et al. Genes and Development, 2011[1].

Abstract

Large intergenic non-coding RNAs (lincRNAs) are emerging as key regulators of diverse cellular processes. Determining the function of individual lincRNAs remains a challenge. Recent advances in RNA sequencing (RNA-Seq) and computational methods allow for an unprecedented analysis of such transcripts. Here, we present an integrative approach to define a reference catalogue of over 8,000 human lincRNAs. Our catalogue unifies previously existing annotation sources with transcripts we assembled from RNA-Seq data collected from ~4 billion RNA-Seq reads across 24 tissues and cell types. We characterize each lincRNA by a panorama of more than 30 properties, including sequence, structural, transcriptional, and orthology features. We find that lincRNA expression is strikingly tissue specific compared to coding genes, and that they are typically co-expressed with their neighboring genes, albeit to a similar extent to that of pairs of neighboring protein-coding genes. We distinguish an additional sub-set of transcripts that have high evolutionary conservation but may include short open reading frames, and may serve either as lincRNAs or as small peptides. Our integrated, comprehensive, yet conservative reference catalogue of human lincRNAs reveals the global properties of lincRNAs and will facilitate experimental studies and further functional classification of these genes.

Authors' contributions

MNC, AR and JLR designed the study. MNC performed the analysis with the help of CT and LG and with the guidance of AR and JLR. CT and LG introduced the tissue specificity score. LG, MK and BTV performed experiments. MNC wrote the manuscript and all authors reviewed and revised it.

2.1 Introduction

A few dozen long non-coding RNAs (lncRNA) are known to play important regulatory roles in diverse processes, such as X-inactivation (*XIST*) [2], imprinting (*H19*; *Kcnq1ot1*) [3, 4], and development (*HOTAIR*, *COLDAIR*) [5, 6]. Recent genomic studies have shown that a substantial portion of the mammalian genome may be transcribed [7], suggesting the presence of many more non-coding transcripts, and spurring efforts to catalogue them [7, 8], using data collected with tiling microarrays [9, 10], shotgun sequencing of expressed sequence tags (ESTs) and cloned cDNA [7, 11], and maps of histone modification patterns [12]. In particular, recent studies have focused on large intergenic non-coding RNAs (lincRNAs) [12-15], which do not overlap annotated protein-coding regions, as this facilitates experimental manipulation and computational analysis.

Recent work has suggested various functions and molecular mechanisms for lincRNAs [16, 17], including the regulation of epigenetic marks and gene expression [2, 4, 5, 14, 18-20]. Other studies have inferred and tested the functional role of lincRNAs in processes such as pluripotency and p53 response pathways, by associating the expression of lincRNAs to those of protein-coding genes [12, 21-23] . More globally, a recent comprehensive screen identified dozens of lincRNAs required to maintain pluripotency and suggested that these lincRNAs work in *trans* [24]. Another class of 'enhancer RNAs' may be either by-products of transcription [25, 26] or serve to activate gene expression in *cis* [15, 27]. Despite these intriguing studies of individual lincRNAs, generalizing these findings to thousands of lincRNAs remains a substantial challenge. Collectively, lincRNAs are likely to reflect different families with distinct roles.

A first requirement towards functional categorization is a systematic catalogue of lincRNAs transcripts and their expression across tissues. In practice, however, researchers studying human lincRNAs are faced with an excessive set of non-coding transcripts of varying or unknown reliability, which may be not-well-defined [14] and have little or no expression data [8], or with very small sets of experimentally-validated ones [28]. Transcripts in current annotations of the human transcriptome from the GENCODE/ HAVANA [8] or UCSC Genome Browser [29] are valuable resources, but, it is hard to evaluate their biological characteristics in the absence of expression levels and further processing.

Recent advances in RNA-Seq [30] and computational methods for transcriptome reconstruction [31-33] now provide an opportunity to comprehensively annotate and characterize lincRNA transcripts. Indeed, an initial application of this approach in three mouse cell types characterized the gene-structure of over a thousand mouse lincRNAs, most of which were not previously identified [31].

Here, we present an integrative approach to define a reference set of lincRNAs that unifies existing annotation sources with transcripts reconstructed from over 4 billion RNA-Seq reads collected across 24 human tissues and cell types. We developed a conservative, broadly-applicable, pipeline to identify transcripts that are sufficiently expressed and have a negligible potential to encode proteins. We identified 8,195 putative lincRNAs, of which 4,662 (57%) form a ‘stringent’ set. We characterized each lincRNA in the catalogue by a panorama of structural, sequence and expression features, as an initial step towards fine categorization.

We used these features to test some of the proposed roles and characteristics of lincRNAs in a global and systematic way. For example, we find that lincRNAs – at all expression levels – are expressed in a highly tissue specific manner, much more so than protein coding genes. We observe no significant enrichment of correlated co-expression between lincRNAs and their neighboring genes beyond that expected for any two neighboring protein-coding genes. We identified expressed orthologous transcripts in another vertebrate species for 993 (12%) of human lincRNAs. An additional set of 2,305 other transcripts with high evolutionary conservation but ambiguous coding potential may function as non-coding RNAs or as small peptides. Finally, we highlight 414 lincRNAs that reside within intergenic regions previously associated with specific diseases/traits by genome wide association studies, as candidates for future disease-focused studies. Our reference catalogue will facilitate future experimental and computational studies to uncover lincRNA functions.

2.2 Results

A computational approach for comprehensive annotation of lincRNAs

To comprehensively identify human lincRNAs we developed a computational approach that integrates RNA-Seq data with available annotation resources (**Figure 2.1a**) and consists of 4 key steps (**2.4 Methods**). (**1**) transcriptome reconstruction of each sample from RNA-Seq data using two transcript assemblers, Cufflinks [32] and Scripture [31]; (**2**) compilation of all non-coding and unclassified transcripts previously annotated; (**3**) integration of RNA-Seq reconstructions with all annotation resources, using Cuffcompare [32], to determine a unique set of isoforms for each transcript locus; and (**4**) processing of the collected transcripts to identify lincRNAs, defined as transcripts that are reliably expressed, large, multi-exonic, non-coding and intergenic.

There are two main challenges in applying this integrative approach to annotate lincRNA gene loci: (**i)** distinguishing lowly expressed lincRNAs [31] from the tens of thousands of lowly expressed, single-exon, unreliable fragments assembled from RNA-Seq; and (**ii)** distinguishing novel transcripts encoding proteins or short peptides from bona fide non-coding ones. To address the first challenge, we remove unreliable lowly expressed transcripts using a learned read coverage threshold (**2.4 Methods**) and focus only on multi-exonic transcripts. To address the second challenge, we evaluated the coding potential of each of the remaining putative lincRNAs using two methods. First, we removed any putative open reading frames (ORFs) that are evolutionarily constrained to preserve synonymous amino-acid content, as reflected by a positive Phylogenetic Codon Substitution Frequency (PhyloCSF) metric [34], calculated for each locus across 29 mammals (**2.4 Methods**). Second, we scanned each transcript, in all three reading frames, to exclude transcripts that encode any of the 31,912 protein domains cataloged in the protein family database Pfam [35].

An annotated human lincRNA catalogue

To generate a human lincRNA catalogue, we applied our pipeline to poly adenylated RNA-Seq data collected from 24 human tissues and cell lines. These included both single- and paired- end reads that are 50 or 75 bases long, sequenced on Illumina platforms (~4 billion reads total; ~175 million reads per sample on average; **2.4 Methods**). We integrated those with annotations from RefSeq [36], the UCSC Genome Browser [29], and GENCODE (version 4) [8] that were processed through our pipeline. We eliminated all annotated non-lincRNA transcripts (*e.g.*, annotated protein-coding genes, microRNAs, tRNAs, pseudogenes).

The initial catalogue consists of a provisional set of 8,195 intergenic transcripts (**Figure 2.1b**). Although many of the previously annotated transcripts are also captured by the ones assembled from the sequencing data (**Figure 2.1b-c**, 1,864 lincRNAs identified by both), most (4,816) novel lincRNAs were only identified using RNA-Seq. Based on the three samples for which we had 2 biological replicates (brain, testes and lung fibroblasts), the reconstructed transcripts are highly reproducible: 70%-80% of assembled transcripts in the lower-coverage replicate are also assembled in the higher-coverage replicate (**Supplementary Table 1.1**; **2.4 Methods**).

Despite the high correspondence between protein-coding transcripts reconstructed by Cufflinks and Scripture (~85% of coding genes; **2.4 Methods**; **Supplementary Figure 1.1a**), there were larger differences between the non-coding transcripts assembled by the two methods, due to the differences in how each assembler reconstructs low-abundance transcripts (**Supplementary Figure 1.1.1b**, ~43% of the putative lincRNAs were identified by only one source). This is comparable to previously observed discrepancies in reconstruction of lowly-expressed protein coding genes (Garber et al. 2011) and is handled below.

We annotated each putative lincRNA in the provisional catalogue with a comprehensive ‘profile’ listing dozens of traits such as its chromatin state, maximal expression level, proximity to coding genes, and evolutionary conservation (**2.4 Methods**, **Supplementary Dataset 1-2**). Below, we use these features to define particular criteria by which we focus our analysis. Future users may leverage the annotated catalogue through criteria of their choosing.

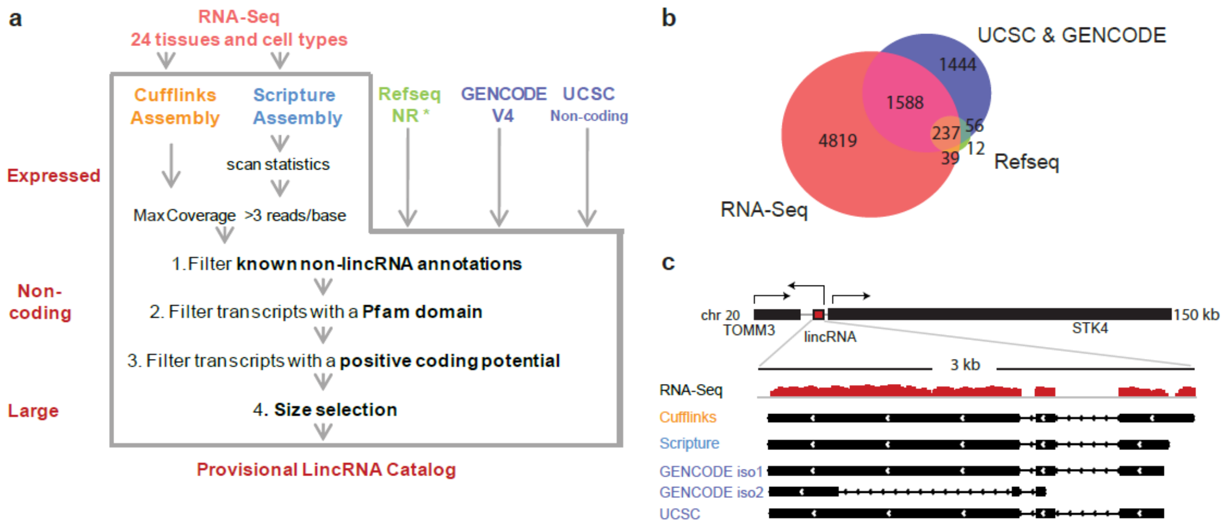


Figure 2.1| LincRNA catalog generation. **(a)** An integrative computational pipeline to map, reconstruct and determine the coding potential of lincRNAs based on known annotations and computational methods and its application to human lincRNAs. The pipeline takes as input RNA seq data (top, red) and existing annotation sources (top, RefSeq NR, Gencode and UCSC annotation for human). RNA-Seq data is assembled by two assemblers – Cufflinks (gold) and Scripture (blue). Transcripts from all inputs are filtered by known annotations, presence of a Pfam domain, and positive coding potential. Transcripts annotated by RefSeq NR (*) where **not** filtered by the Pfam domain scan and the coding potential score. Finally, only multi-exonic transcripts larger than 200 base pairs (bp) are retained. **(b)** The number of lincRNA loci identified and their overlap with other annotation sources. The Venn diagram shows the overlap between transcripts from RNA-Seq assembly (red), GENCODE and UCSC (purple) and Refseq (green). **(c)** A representative example of a noncoding transcript that was reconstructed by Cufflinks and Scripture and was also curated in GENCODE and UCSC. Top: the human genomic locus of the human lincRNAs (red) and its protein-coding neighbors (black, arrowhead: direction of transcription). Bottom: magnified view of the lincRNA locus showing the coverage of RNA-Seq reads from the testes (red) and the transcripts identified by each source (black). Abbreviations: *iso* = isoform.

A stringent set of 4,662 human lincRNAs

We defined a *stringent lincRNA set* that includes those loci for which at least one lincRNA isoform was reconstructed in at least two different tissues, or by two assemblers in the same tissue (**2.4 Methods**). This leverages the unique benefits of each assembler, while in principle removing transcripts with insufficient coverage. The stringent set includes 4,662 lincRNA loci (14,353 transcripts), 2,798 of which (~60%) were not identified by RefSeq, UCSC and GENCODE. We focused on the characteristics of this stringent set.

LincRNAs are alternatively spliced and preferentially proximal to developmental regulators

We characterized the basic features of lincRNAs, comparing them to protein-coding genes, when appropriate. **First**, the **size** of lincRNAs is smaller than that of protein-coding transcripts and they have fewer exons (on average, 2.9 exons and a transcript length of ~1 kilobase (kb) for lincRNAs vs. 10.7 exons and ~2.9 kb for protein-coding transcripts; **Supplementary Figure 1.2a-b**). Notably, we may under-estimate the length and exons of lincRNAs, since their lower abundance may result in incomplete assembly. **Second**, lincRNAs are **alternatively spliced** (on average, ~2.3 isoforms per lincRNA locus, **Supplementary Figure 1.2c**). **Third**, lincRNA **loci** are located from a few bases to over 3 megabase from a protein-coding locus, with 28% within 10 kb of their coding neighbor (median = ~40 kb; **Supplementary Figure 1.2d**). **Finally**, protein coding genes proximal (<= 10 kb) to lincRNAs are enriched for those associated with development and transcriptional regulation (*e.g.*, *GATA2*, *GZF1*, and *NEUROG* all have lincRNA neighbors, **Supplementary Figure 1.3**), consistent with previous reports [12, 37].

Many lincRNAs are characterized by K4-K36 domains

We next explored the chromatin features of lincRNA loci as reflected in chromatin state maps from the 9 ENCODE cell lines and other cells [14, 38]. We examined each locus for the presence of a ‘K4-K36 domain’, a chromatin signature of actively transcribed genes, which we previously used to identify lincRNAs [12]. This domain consists of histone 3 lysine 4 tri-methylation (H3K4me3) at the promoter followed by histone 3 lysine 36 tri-methylation (H3K36me3) along the transcribed region. Despite the lack of paired matched samples of histone modifications and RNA-Seq, 24% of the lincRNAs in our catalogue have previously-defined chromatin K4-K36 domains, and ~ 40% have such domains when using less stringent criteria (with the remaining exhibiting partial signatures; **Supplementary Figure 1.4, 2.4 Methods**).

lincRNA genes are no more likely to overlap enhancers than protein coding genes

Recent studies reported short transcripts derived from enhancer elements, termed eRNAs, that are most likely not poly-adenylated [25]. While this suggests that eRNAs and lincRNAs come from different classes, it is possible that longer poly adenylated transcripts may also be derived from enhancer elements, and hence be related to eRNAs. To test this possibility, we examined the overlap between lincRNAs' exons and two recent annotations of human enhancers based on genome-wide chromatin state maps. 27% of our lincRNAs and 44% of coding genes overlap 111,362 genomic regions previously suggested to function as enhancers [38] in 9 ENCODE cell lines (each overlap P <0.001, permutation test, **2.4 Methods**). When considering a more stringent subset of regions that are more likely to function only as enhancers (**2.4 Methods**), ~10% and 14% of lincRNAs and coding genes overlap such regions (both P <0.001), respectively. Both lincRNAs and protein-coding genes have even lower overlaps (both <3%, P<0.001) to an enhancer set from human embryonic stem (ES) cells [39], possibly due to the lack of biological correspondence between the cell types and the tissue-specific nature of both lincRNAs and enhancers. Notably, this low overlap persists even when comparing more closely matched samples. Thus, only 15% of lincRNAs defined in mouse ES cells [31] overlap enhancers defined in mouse ES cells ([40]; **2.4 Methods**), and less than 1% of lincRNA defined in mouse neuronal progenitor cells [31] overlap enhancer elements that express eRNAs in mouse cortical neurons ([25] ; **2.4 Methods**). Taken together, this data suggests that lincRNAs and eRNAs represent different subtypes of lncRNAs.

LincRNAs are expressed in a more tissue specific manner than protein coding genes

The maximal expression levels of lincRNAs are lower than those of protein coding genes, across the 24 samples (**Figure 2.2a**), with a ~10-fold lower median maximal expression level (**Figure 2.2b**, expression estimated with Cufflinks [32], **2.4 Methods**). Importantly, lincRNAs identified by RefSeq annotations were similarly lowly expressed relative to coding genes (~10 fold lower; **Supplementary Figure 1.5**). These lower expression levels are consistent with previous reports [31, 41], suggesting a general property of lincRNAs.

The vast majority of lincRNAs exhibit tissue specific expression patterns, more so than protein coding genes, based on unsupervised clustering of expression profiles (**Figure 2.2a**). We further calculated a Tissue Specificity Score for each transcript, using an entropy-based metric, which relies on Jansen-Shannon (JS) divergence (**2.4 Methods**). This specificity metric (ranging from 0 to 1) quantifies the similarity between a transcript's expression pattern across tissues and another predefined pattern that represents the extreme case in which a transcript is expressed only in one tissue.

Based on this measure, the majority of lincRNAs (78%) are tissue specific, relative to only ~19% of coding genes ($P < 10^{-300}$, Fisher exact test, **Figure 2.2c** and **Supplementary Figure 1.6**). These differences are **not** the result of the low expression levels of lincRNAs and hold true for lincRNAs and protein coding genes expressed at similar levels (**Figure 2.2b-c**, **Supplementary Figure 1.6**). This was particularly true for the 35% of more highly expressed lincRNAs (and comparably expressed protein coding genes, each with a maximal expression level of 3 to 20 FPKM (fragments per kilobase of exons per million fragments mapped)). Thus, lincRNAs exhibit more tissue specificity than protein coding genes at different expression ranges.

Approximately a third of our lincRNAs are specific to testes. Very few (<2%) of those overlap with a previously defined set of testes-specific small piRNAs (~30 nucleotides long, [42]. Thus, testes-specific lincRNAs may define a new class of RNAs in this organ. Testes-specific lincRNAs do not bias the global transcriptional characteristics above: lincRNAs that are **not** testes-specific are also lowly expressed and tissue specific (presenting a qualitatively similar distribution with only moderately reduced tissue specificity scores; **Supplementary Figure 1.5 and 6a**).

Finally, we predicted putative functions for our lincRNAs based on the known functions of protein-coding genes with similar expression patterns. We clustered lincRNAs and protein-coding genes using k-means clustering of the tissue specificity distance measure (**2.4 Methods),** and annotated each cluster with enriched functions of the protein-coding gene members. Clusters of tissue-specific lincRNAs and protein-coding genes are enriched for processes specific to that tissue or its differentiation (e.g: a liver-specific cluster is enriched with functional terms such as cholesterol and lipid transport and homeostasis**, Supplementary Figure 1.7**, **Supplementary Dataset 2-3**).

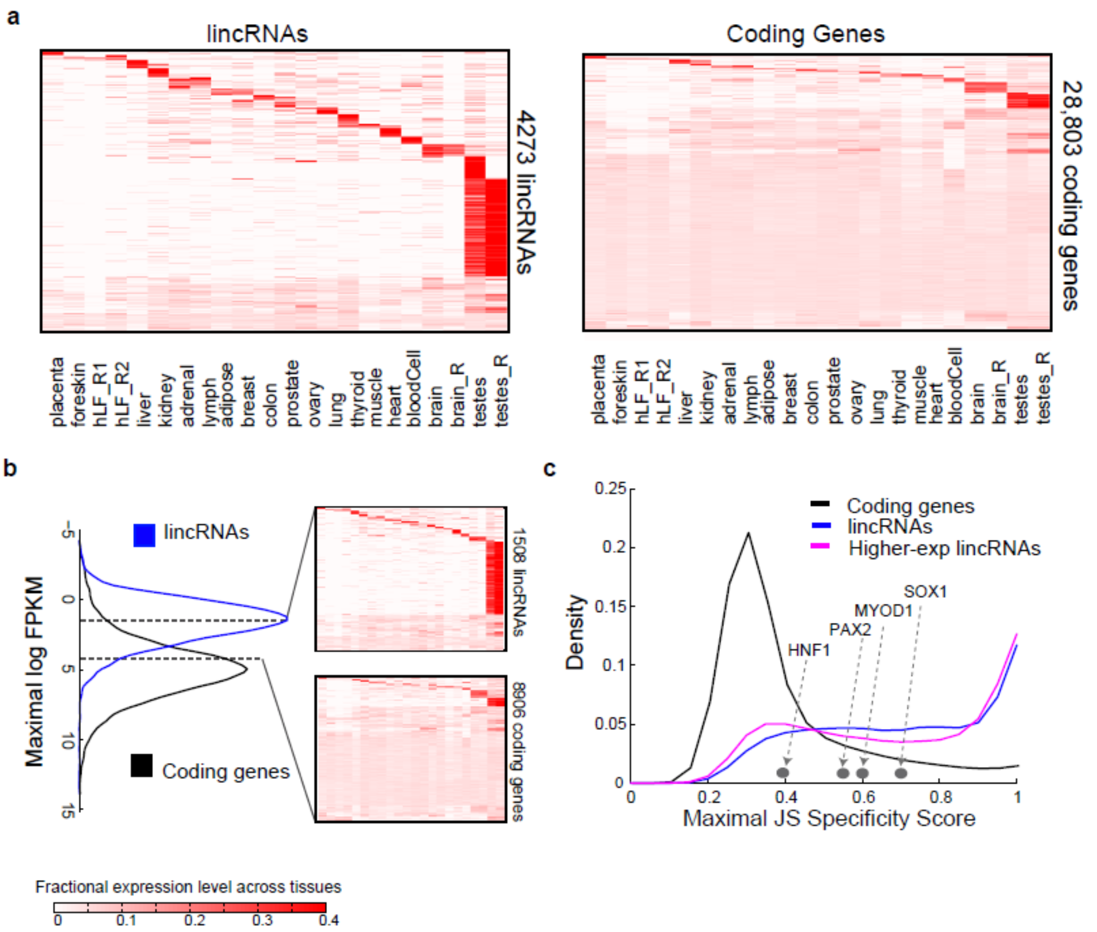


Figure 2.2| Tissue specificity of lincRNAs and coding genes. **(a)** Abundance of 4273 lincRNA (rows, left panel) and 28,803 protein-coding genes (rows, right panel) across tissues (columns). Rows and columns are ordered based on a k-means clustering of lincRNAs and protein coding genes. Color intensity represents the fractional density across the row of log normalized FPKM counts as estimated by Cufflinks (saturating less than 4% of the top normalized expression values, **2.4 Methods**). **(b)** lincRNAs are more lowly expressed than protein coding genes. Maximal expression abundance (log2 normalized FPKM counts as estimated by Cufflinks) of each lincRNA (left panel, blue) and coding (left panel, black) transcripts across tissues, Right panel shows the expression levels of 1508 lincRNAs (top right) and 8906 coding genes (bottom right) that have a maximal expression level within the range bounded by the dashed segments on the left panel ([1.6-4.3] log2 FPKM; see **2.4 Methods**). Heat maps are clustered and visualized as in (a). **(c)** Tissue specific expression. Shown are distributions of maximal tissue specificity scores calculated for each transcript across the tissues from the data in panel *a*, for coding genes (black), lincRNAs (blue) and the 1508 highly expressed lincRNAs (pink; as in panel *b*). Examples of the tissue specificity score of coding genes with known tissue specific patterns are marked by gray dots.

LincRNAs are co-expressed with neighboring coding genes at levels similar to those expected for any pair of chromosomal neighbors

The enrichment of specific gene functions in protein-coding genes neighboring lincRNAs and the presence of some pairs of neighboring lincRNA:protein-coding genes within expression clusters raise the hypothesis that such organization may be important for the regulatory function of lincRNAs. In particular, recent studies suggested that some lincRNAs may act in *cis* and affect the gene expression of their chromosomal neighborhood [15, 37].

One expectation from this hypothesis is that the expression of lincRNAs and their neighboring gene loci would be correlated across our samples. To test this hypothesis, we focused on the expression patterns of 1,361 (28%) of our stringent lincRNAs that are located within 10kb from a coding gene. Indeed, these lincRNAs and their coding neighbors were more correlated to each other than random gene pairs ($P < 5*10^{-46}$, KS-test; $P < 10^{-307}$, Student t-test, effect size = 0.86; **Figure 2.3a, 2.4 Methods**).

We must interpret this result with care, since the co-expression between a lincRNA and its protein-coding neighbor may result from either (**i**) a true *cis* effect of lincRNAs on its neighbor, or (**ii**) proximal transcriptional activity in the surrounding open chromatin [43], since co-expression of chromosomal protein-coding gene neighbors was previously shown across eukaryotes [44, 45]. Supporting the second possibility, pairs of neighboring protein-coding genes were also more correlated to each other than random pairs ($P < 3.4*10^{-159}$, Kolomogorv-Smirnov (KS) test ; **Figure 2.3a**). Furthermore, the correlation between lincRNA:Protein-coding gene neighbors was only modestly higher than between Protein-coding gene:Protein-coding gene neighbors of a similar distance (effect size = 0.23, $P < 4.3*10^{-7}$, KS-test; $P < 6.9*10^{-7}$, Student t-test; **Figure 2.3a**).

To further distinguish between these two possibilities, we focused on those protein-coding genes that had a lincRNA neighbor on one side and a coding neighbor on the other side, and used a **paired** test to compare the correlation between each protein-coding gene and its lincRNA neighbor, to that between the same protein-coding gene and its protein-coding gene neighbor.

This paired comparison showed a weak **opposite** trend, where pairs of coding gene neighbors are slightly more correlated to each other than neighboring lincRNA:protein-coding gene pairs ($P < 0.001$ paired Student t-test; effect size = 0.23), thus favoring option (ii) an effect of gene proximity.

Taken together, this analysis suggests that overall lincRNAs are **not** more correlated to their protein-coding gene neighbors than expected for a pair of neighboring protein coding gene loci. Yet, the ultimate test of *cis* or *trans* regulatory mechanisms for lincRNAs requires experimental gain- or loss- of function data.

Divergently Transcribed lincRNAs

Unstable, likely non-coding, transcripts can also be derived from divergent (bi-directional) transcription, in both yeast and mammals [46-48] . These may be either byproducts of chromatin remodeling and recruitment of the transcription machinery to the neighboring gene's promoter or functional transcripts. Due to limited read length and computational methods, previous studies did not determine whether these transcripts are spliced. Interestingly, several functionally studied lincRNAs, including *Tug1* [49], *HOTAIR* [5], and *HOTTIP* [27], are divergent transcripts. We therefore hypothesized that other divergently transcribed transcripts may be spliced and polyadenylated lincRNAs.

Indeed, 588 (~13 %) of our stringent lincRNAs are spliced transcripts divergently transcribed within 10 kb of a coding gene promoter, with a majority (~65%) that initiate within 1 kb of a coding gene's annotated transcription start site (**Supplementary Figure 1.8**). Furthermore, ~35% of the 588 pairs share a H3K4me3 domain (a hallmark of active promoters) based on the ENCODE chromatin state maps (**2.4 Methods**), although we cannot definitively determine if these divergently encoded pairs are also divergently transcribed from the same promoter. These divergent coding genes neighbors are enriched for developmental and metabolic processes (**Supplementary Figure 1.3b**). Focusing on the 68% that are spliced in the tissue where they are maximally transcribed (**2.4 Methods**: "*Estimating expression abundance*"), there is only a slightly higher correlation between divergent lincRNAs and neighboring coding genes than for divergent coding gene pairs (effect size =0.27, P < 0.008 KS-test; P < 0.009, Student t-test; **Figure 2.3b**). Furthermore, while ~49% of the divergently transcribed lincRNAs are tissue specific, for approximately half of those, the neighboring gene is ubiquitously expressed (**Figure 2.3c**). Thus, although there are clearly bidirectionally transcribed, spliced lincRNAs in our catalogue, we found no clear additional distinguishing features for this set.

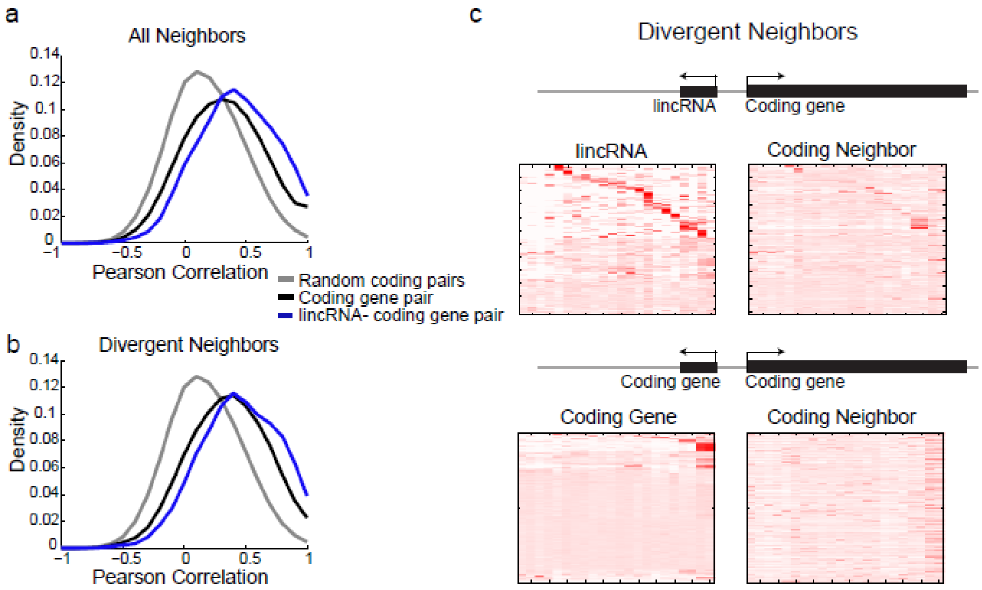


Figure 2.3| Chromosomal domains of gene expression. (a) Correlation of expression patterns between pairs of neighboring genes. Shown are distributions of Pearson correlation coefficient in expression levels across the tissues in Figure *2a* between either 6524 pairs of coding gene neighbors (black), 497 pairs of lincRNAs and their neighboring coding gene (blue), or 10,000 random pairs of protein coding genes (gray; null model) *. **(b)** Shown are distributions of Pearson correlation coefficients calculated as in (a) but only for 223 pairs of divergently transcribed pairs of lincRNA and protein-coding gene (blue) or 1575 pairs of divergently transcribed protein-coding genes *. **(c)** Expression patterns of pairs of divergently expressed genes. Shown are expression patterns (presented as in **Figure 2a**) for pairs of divergently transcribed lincRNA (rows, top left) and protein coding genes (rows, top right), or pairs of divergently transcribed protein-coding genes (rows, bottom left and right panels) *. (*) Only lincRNAs that have spliced read support when maximally expressed and that are not testes-specific are presented (refer to Supplementary Note "*Estimating expression abundance*" for further detail).

Expressed syntenic orthologs of human lincRNAs in mammals and vertebrates

We and others have previously reported evidence for purifying selection at different sets of mammalian lincRNAs [12, 13, 15]. A recent study has also identified expressed orthologs of a few highly-conserved and brain-expressed mouse lncRNAs in species as distant as opossum and chicken [50]. However, the number of human lincRNAs that have an orthologous, actively expressed, **transcript** in other species remains unknown.

To identify human lincRNAs with orthologous expressed transcripts in other species (supported by experimental evidence), we surveyed a catalogue of mammalian and non-mammalian vertebrate transcripts that were syntenicaly mapped to the human genome by TransMap [51], a cross-species mRNA alignment method. TransMap maps all known transcripts (*e.g.*, full-length cDNAs and others in RefSeq or UCSC) and ESTs across vertebrate species using syntenic BLASTZ alignments [52] that use conserved gene order (synteny). Since EST coverage varies between species (**Supplementary Table 1.2**), TransMap can only provide a lower-bound estimate of orthologous transcripts.

Of the 8,195 lincRNAs, 993 are syntenically paired with an orthologous transcript (**Figure 2.4a-d**), comprising a Trans-mapped lincRNA set (~135 expected by random permutations; **2.4 Methods; 2.4 Methods**). 702 of the Trans-mapped lincRNAs are in the stringent lincRNA set (~15% of stringent lincRNAs). The majority (53%) of the Trans-mapped lincRNAs were not previously annotated in the human transcriptome (GENCODE, RefSeq or UCSC; **Supplementary Figure 1.9a**). Trans-mapped lincRNAs have tissue specificity and low expression comparable to that of all other lincRNAs (**Supplementary Figures 1.6a and 1.9b-c**). 59% of the Trans-mapped lincRNAs were mapped to annotated transcripts that had evidence beyond ESTs. Supporting our non-coding classification scheme, only 18% of the 641 lincRNAs with Trans-mapped orthologous transcripts in mouse were classified as coding in mouse and only ~11% have a positive PhyloCSF score (**Supplementary Figure 1.9a**; **2.4 Methods**). Trans-mapped lincRNAs have orthologs in species from mouse to fish, with closer species with more transcriptome data showing more orthologs than distant ones (**Figure 2.4d**).

Orthologous lincRNAs exhibit modest sequence homology

We evaluated the degree of sequence similarity between the Trans-mapped transcripts. We measured the portion of each lincRNA transcript's length that is aligned to the orthologous transcript. The majority of trans-mapped lincRNAs are only moderately spanned by an orthologous mapped transcript (a median of 21% and 56% of their transcript or genomic locus aligned, respectively; **Figure 2.4e**). In loci where lincRNAs are Trans-mapped to mouse coding transcripts, a larger portion of the human *locus* but a smaller portion of the mouse *transcript* aligns between the species (**Figure 2.4e**, **Supplementary Figures 1.10a-b**). This may be due to either cryptic small peptides in the human transcript or the evolution of a non-coding transcript from a coding one. The available data is insufficient to distinguish between these hypotheses, which can be tested as paired cross species RNA-Seq samples are collected.

We next compared the fraction of identical bases aligned between the lincRNAs and their orthologs to that of random sequence pairs, randomly selected syntenic blocks, or orthologous coding genes. Trans-mapped lincRNAs and their orthologous transcripts show similar sequence identity to that of randomly selected syntenic blocks, which is lower than pairs of orthologous protein coding genes and higher than for random pairs of genomic regions of similar size (**Figure 2.4f**, **Supplementary Figures 1.10c-d**; **2.4 Methods**). With only 34% of the human genome syntenically mapped to the mouse genome [53], the resemblance of Trans-mapped lincRNAs to random syntenic blocks still implies evolutionary constraint to preserve sequence elements.

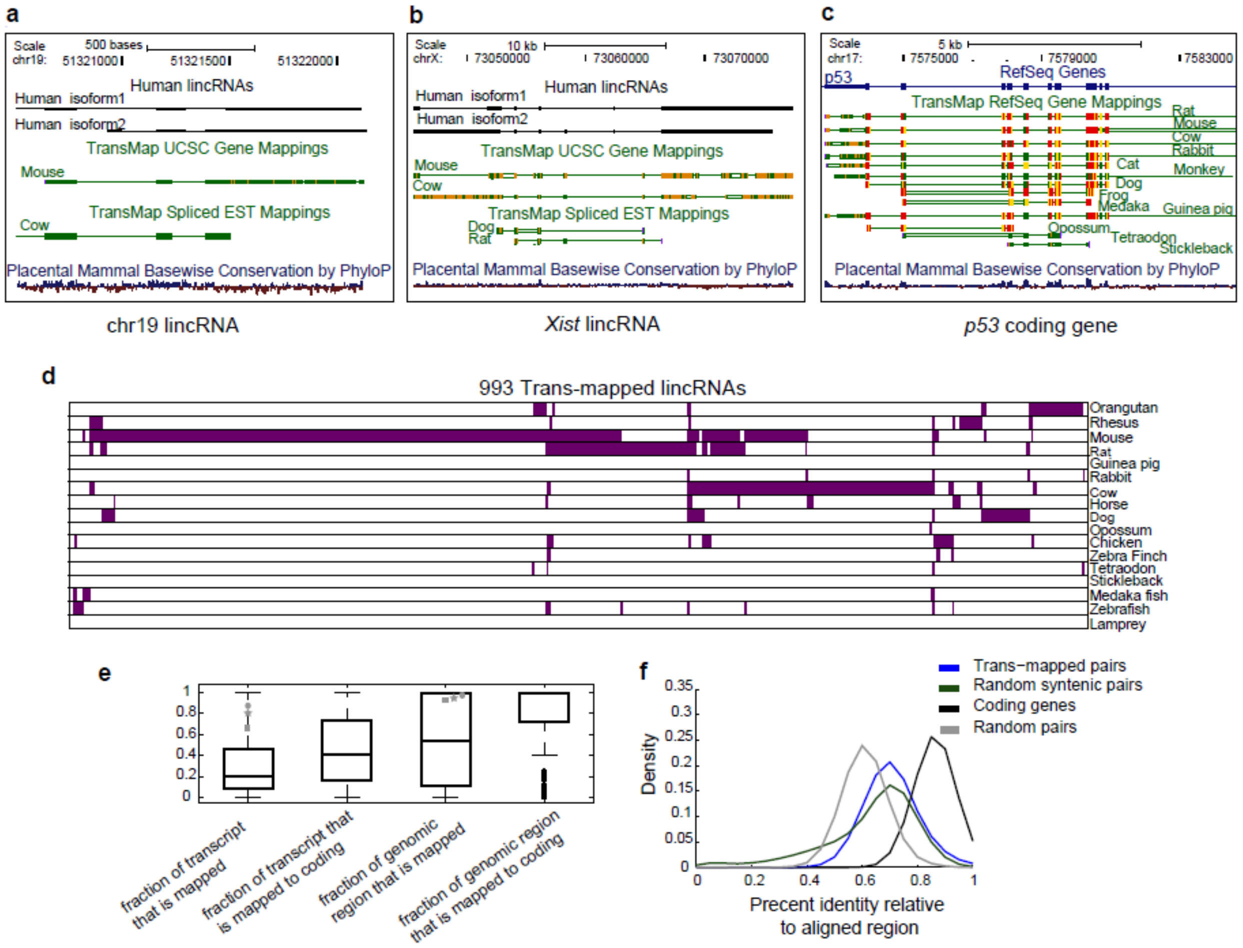


Figure 2.4| Orthologous transcripts of human lincRNAs in mammals and other vertebrates. (a) A human lincRNA with syntenic TransMap mappings to mouse and cow. Shown are UCSC browser [54] tracks showing two isoforms of the human lincRNA (black, top tracks), the mouse and cow transcripts (green, middle tracks) that were Trans-mapped to its human locus, and the base-wise conservation calculated by PhyloP at this locus (red-blue, bottom track). **(b)** Syntenic Trans-Mapping to *XIST***.** Tracks presented as in (a). **(c)** Syntenic Trans-Mapping to *p53*. **(d)** Species distribution of 993 human lincRNAs with trans-mapped orthologs (columns) and the species in which the trans-mapped transcripts were found (rows, purple). **(e)** Characteristics of trans-mapping to human lincRNAs. Boxplots of the fraction of the human lincRNA transcript that is aligned to an ortholog (1st and 2nd boxes), and the fraction of the lincRNAs genomic locus covered by the syntenic mapping of the ortholog (3rd and 4th boxes), for all trans-mapped lincRNAs (1st and 3rd boxes) or only for those lincRNAs that were mapped to mouse coding transcripts (2nd and 4th boxes). Gray square, star and circle represent *XIST*, *HOTAIR* and the lincRNA shown in panel 4a, respectively. **(f)** Distribution of the percentage of identical bases across the FSA [55] pair-wise alignments between human and mouse trans-mapped transcript pairs. Blue: lincRNAs and their mouse orthologs, Black: human coding genes and their mouse orthologs, Green: randomly selected 1 kilobase human and mouse syntenic blocks. Gray: random pairing of human lincRNAs and mouse transcripts (from the set marked in blue). All statistics presented in this figure were calculated at the locus level (i.e each lincRNA loci was accounted once, rather than accounting for all of its isoforms).

Novel transcripts with potential coding capacity

While our stringent lincRNA classification strategy focused on non-coding transcripts, we also characterized 2,305 transcripts that were excluded by our coding potential criteria (a Pfam domain, a positive PhyloCSF score, or previously annotated as pseudogenes) and termed them the *transcripts of uncertain coding potential* (*TUCP*) set (**2.4 Methods**). These may include lincRNAs as well as other transcripts. The majority (1,533; ~66%) were previously annotated as pseudogenes which, due to our focus on multi-exonic transcripts, are probably not retro-transposed spliced mRNAs that were integrated back to the genome (**Figure 2.5a**). Similar to the stringent set, TUCP transcripts are expressed at lower and more tissue-specific patterns than protein-coding genes (**Figures 2.5b-c**).

The coding potential of most of these transcripts was very low compared to known coding genes, and only 32% (757) exceeded our PhyloCSF score criteria (**Figures 2.5a,d**; **2.4 Methods**). The evolutionary-constrained ORFs in these transcripts are mostly short (51% are less than 70 amino acids long) and cover a small portion of the transcript (53% cover less than 25%, **Figures 2.5e-f**). Thus, some of these transcripts may encode small functional peptides [56], whereas others may function as non-coding RNA.

TUCP transcripts are under stronger purifying selection than stringent lincRNAs. First, the exonic sequence in TUCP transcripts is more highly conserved than that of stringent lincRNAs ($P < 10^{-116}$, effect size = 0.77; **Supplementary Figure 1.11**; **2.4 Methods**), even when excluding pseudogenes (**Supplementary Figure 1.11)**. Second, a larger fraction of them have a Trans-mapped syntenic ortholog (~36% (838) or ~34% when excluding pseudogenes, compared with ~15% (702) of stringent lincRNAs), and the syntenic alignments cover a slightly larger portion of the transcript (**Supplementary Figure 1.12**). Third, 74% of the Trans-mapped transcripts have an ortholog in a species more distant than mouse (vs. 37% of the Trans-mapped lincRNAs; ~67% when excluding TUCP pseudogenes; **Figure 2.5g**).

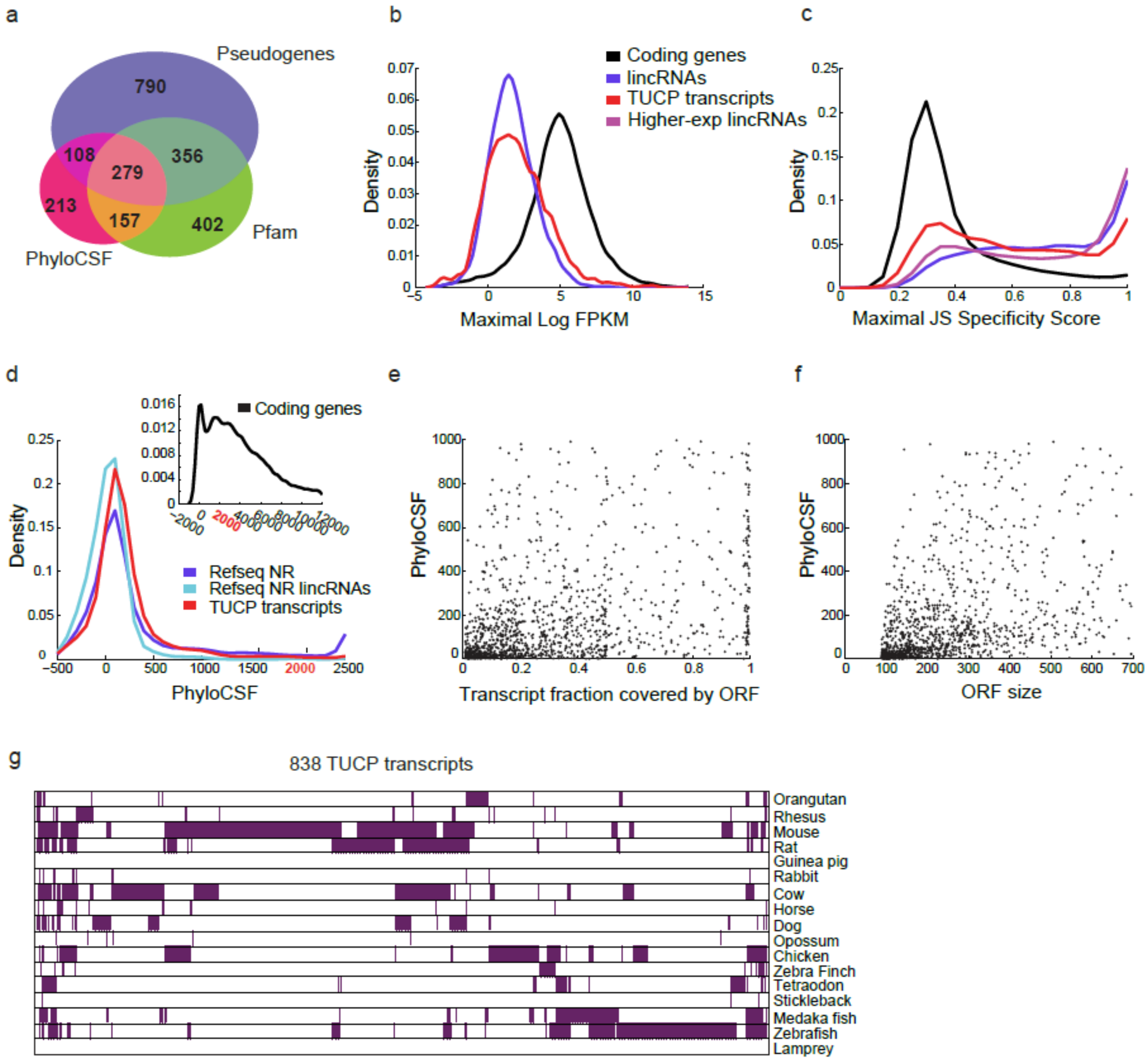


Figure 2.5| Novel transcripts with potential coding capacity. (a) Characteristics of TUCP transcripts. Shown is a Venn diagram of the 2305 TUCP set transcripts annotated as pseudogenes (purple), containing a Pfam domain (green), having a PhyloCSF score higher than the pipelines set criteria (pink), or combinations thereof. **(b)** Expression levels of TUCP transcripts. Shown are distributions of maximal expression abundance (log normalized FPKM counts as estimated by Cufflinks) in TUCP (red), stringent set lincRNA (blue), and coding (black) transcripts. **(c)** Tissue specificity of TUCP transcripts. Shown are distributions of maximal tissue specificity scores calculated for each transcript in the TUCP set (red), stringent lincRNA set (blue), coding (black), and higher expressed lincRNAs (magenta, transcripts defined as in Figure 2c). **(d)** PhyloCSF scores of TUCP transcripts. Shown is the distribution of PhyloCSF scores of the TUCP transcripts (red), all non coding genes in Refseq (blue) or the subset of Refseq classified as lincRNA by our pipeline (light blue). Inset: the corresponding distribution for

protein coding genes, which spans a much wider range of positive scores. **(e,f)** Putative ORFs in TUCP transcripts. Shown are scatter plots of the fraction of each transcript spanned by an ORF (e, X- axis) or of the ORF size (f, in nucleotides, X axis) vs. the PhyloCSF score of that ORF (Y axis), for the 1404 TUCP transcripts that had a PhyloCSF score greater than 0. **(g)** Orthologs for TUCP transcripts. Shown are 838 TUCP transcripts (columns) with trans-mapped orthologs and the species in which the trans-mapped transcripts were found (rows, purple).

lincRNAs in disease-associated regions

Although genome wide association studies (GWAS) have identified thousands of common genetic variants related to specific traits or disease phenotypes, many of these variants (~43% [57]) lie in intergenic regions and hence remain largely unexplained. We identified 414 lincRNAs from our comprehensive catalogue (215 of the stringent set) that are located within 1,146 disease- or trait- associated regions from the Published GWAS Catalogue [57] that do not contain annotated coding genes (**2.4 Methods; Supplementary Dataset 2**). Notably, 30 and 81 of those lincRNAs overlap a common variant that was associated with a disease phenotype within their exon or their intron, respectively (both tag and imputed SNPs). Another 76 intergenic disease/trait regions overlap 84 TUCP transcripts (**Supplementary Dataset 6**).

The 215 stringent lincRNAs in these regions are typically expressed in a tissue–specific manner, which in a few cases directly corresponds to the tissue relevant to the associated disease (**Supplementary Table 1.3**). For example, a lincRNA positioned ~3 kb downstream of a thyroid cancer associated SNP in chromosome 14q13.3 (rs944289, OR = 1.37; P = $2.0*10^{-9}$;[58] is strongly expressed specifically in the thyroid (~5.4 log2 FPKM). The 'tag SNP' and the proximal lincRNA are within a 249 kb linkage disequilibrium (LD) region that does not contain any known genes. rs944289 is ~3.5 kb upstream to the transcription start site of the thyroid specific lincRNA. rs944289[T] is predicted to be part of a binding motif for *C/EBP-alpha* (**2.4 Methods; Supplementary Figure 1.13**), and may affect the lincRNA's expression. The LD region is ~250 kb upstream to the gene *NKX2-1* (*TTF1*), a transcription factor with a prominent role in thyroid development, and a previously suggested candidate gene for this SNP association. The lincRNA may be an additional candidate playing a role in thyroid-specific processes (possibly in coordination with the neighboring *NKX2-1*) and in thyroid cancer.

2.3 Discussion

We have generated a reference catalogue of 8,195 human lincRNAs based on integrating RNA-Seq data from 24 tissues and cell types with publically available transcript annotations. 58% of the transcripts in our catalogue are novel and are now identified for the first time using RNA-Seq. We have annotated each lincRNA with a broad range of structural, expression and evolutionary features, shedding new light on their global properties, and testing or generalizing previous hypotheses.

lincRNAs are remarkably tissue specific compared to protein coding genes. This possibility was previously raised [16, 17] based on differential expression patterns in specific biological systems and has several implications. **First**, researchers studying a particular system may benefit from RNA-Seq profiling followed by *de-novo* assembly in that system. **Second**, it is consistent with the hypothesis that some lincRNAs interact with chromatin modulators and provide their target specificity. **Third**, it may indicate that lincRNAs could serve as specific fine tuners. **Fourth**, the low level of lincRNAs expression in a complex tissue, such as brain, may in fact be a by-product of their expression only in few specific cells. Future targeted perturbations of tissue specific lincRNAs defined in our study may elucidate their role in tissue specific processes.

Could many lincRNAs act as enhancer elements promoting the transcription of their neighboring coding genes? Recent studies have demonstrated that several lincRNAs have enhancer like functions [15, 27]. While our co-expression analysis is consistent with this notion, it is insufficient to suggest a global trend in which lincRNAs act as enhancers of their neighbors since neighboring coding genes exhibit similar co-expression patterns. Further systematic perturbation studies in individual systems (as in [15]) may help assess the scope of this function. Notably, a very recent study that systematically perturbed 150 lincRNAs expressed in mouse embryonic stem cells suggested that lincRNAs primarily affect gene expression in *trans* [24]. Collectively, this suggests that some lincRNAs can work in *cis* while others in *trans*.

993 lincRNAs have an orthologous transcript expressed from a syntenic region in another species, ~50% of which were identified for the first time in this study. These lincRNAs had only moderate sequence identity and alignment to their orthologs. This moderate conservation may indicate the importance of transcription from a specific genomic location, or the reduced selective pressure on the primary sequence of non-coding RNAs [2, 59], or the rapid evolution of new functions. It may also be due to alignment to orthologous ESTs that are incomplete transcripts. Our analysis was limited by available transcript data in other species; and will be enhanced as more transcriptomes are sequenced in other organisms.

TUCP intergenic transcripts did not pass our stringent classification criteria as lincRNA, due to evidence of possible protein-coding potential. These transcripts have similar expression levels and tissue specificity as the stringent lincRNA set, but significantly higher level of sequence conservation. Many could encode small peptides, similar to those that function in *D. Melanogaster* embryogenesis [56]. Another 1,533 TUCP transcripts are classified as pseudogenes, and may represent pseudogenes that have evolved to function as non-coding regulatory agents. Ribosome profiling [60] and mass spectrometry of small peptides will help to resolve which of the TUCP transcripts are more likely to be coding.

Substantial progress has been recently made towards the essential goal of annotating long non-coding RNA loci. Our work presents an integrative yet conservative computational approach to mapping lincRNA transcripts that can be used for mapping new transcripts in other species. This is critical to overcome major barriers for future experiments (e.g. cloning, expression profiling and gain- and loss- of function), as well as for the interpretation of genetic association studies. Indeed, 414 lincRNAs in our catalogue stand out as located within intergenic regions associated with common disease. Future work will be necessary to identify RNA sequence domains that relate to function [2], and to further classify lincRNAs into families. Our panorama of lincRNA properties will greatly advance these goals.

2.4 Methods

RNA-Seq data sets

We used two datasets of RNA-Seq for transcriptome reconstruction. The first includes poly-adenylated RNA samples from 16 tissues that were sequenced using Illumina HiSeq 2000 as part of the Human Body Map 2 project (235 million reads per sample on average**; Supplementary Table 1.4**). The second dataset included 8 additional tissues and cell-lines each sequenced the Illumina Genome Analyzer II (GAII) (54 million reads per sample on average; **Supplementary Table 1.4**). The Human Body Map 2 data is accessible from ArrayExpress, ArrayExpress accession: E-MTAB-513:http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-MTAB-513&expandefo=on. The 8 additional tissues and cell-lines are available at GEO accession GSE30554 (See **Supplementary Table 1.4**).

Publically available annotations

All known annotations that were used for the analysis of this paper are specified in **Supplementary Table 1.5**.

Library preparation protocol.

RNA was extracted from in-house cultured cell lines (lung and foreskin fibroblasts) using TRIzol (Invitrogen) following the manufacturer instructions. After extraction a DNAse treatment was applied using Turbo DNAse (Ambion) and a second RNA extraction with TRIzol was performed. Human tissue total RNA samples were obtained from Ambion.

All samples were sequenced using a standard Illumina mRNA-Seq Protocol. The standard mRNA-seq library preparations for both the 16 individual tissues and the 8 additional samples, was done using the “Illumina mRNA-Seq Sample Prep Kit” (Part #RS-100-0801). Briefly, poly A+ RNA was purified from 100ng of total RNA with oligo-dT beads. Purified mRNA was then fragmented with divalent cations under elevated temperature. First strand cDNA synthesis was performed with random hexamers and reverse transcriptase. Second strand cDNA synthesis was performed using RNAseH and DNA PolI. Following cDNA synthesis, the double stranded products were end repaired, a single “A” base was added, and Illumina PE adaptors were ligated onto the cDNA products. The ligation products were purified using gel electrophoresis. The target size range on the gel for these libraries was ~300 bp for the Human Body Map samples and ~350-500 for the 8 additional samples, such that the final library for sequencing would have cDNA inserts with sizes of ~200 bp or 160-380 bp long, respectively . Following gel purification, the adapter ligated cDNA was then amplified with 15 cycles of PCR for the Human Body Map samples and 14 cycles for the 8 additional samples.

RNA-Seq read mapping

All sequenced reads were aligned to the human genome (NCBI 37, Hg19) using the spliced read aligner TopHat version V1.1.4 [62]. Briefly, using a two step mapping processes, TopHat first uses Bowtie [63] to align reads that are directly mapped to the genome (with no gaps). It then determines the possible location of gaps in the alignment based on canonical and non-canonical splice sites flanking the aligned reads. Finally, it uses gapped alignments to align the reads that were not aligned by Bowtie in the first step.

For this analysis, we used 2 iterations of TopHat alignments to maximize the use of splice site information derived across all samples. To this end, the reads from each sample were first aligned using the paired-end alignment option when possible (default parameters and 'min-anchor=5', 'min-isoform-fraction =0'). Next, we generated a pooled splice-sites (or "junctions") file by combining all predicted splice sites across all alignments. We then re-aligned each sample using the pooled splice-sites file (using 'raw-juncs' and 'no-novel-juncs' parameters).

RNA-Seq transcriptome assembly

The transcriptome of each sample was assembled from the mapped reads separately by both Scripture [31] and Cufflinks [32]. Briefly, Cufflinks and Scripture are *ab-initio* transcriptome assemblers that reconstruct the transcriptome based on RNA-Seq reads that were aligned to the genome using a spliced read aligner (e.g. TopHat) [64]. Both methods use spliced read information to determine exons connectivity, but with two different approaches. Scripture uses a statistical segmentation model to distinguish expressed loci from experimental noise and uses spliced reads to assemble expressed segments. It reports all statistically significantly expressed isoforms in a given locus. Cufflinks uses a probabilistic model to simultaneously assemble and quantify the expression level of a minimal set of isoforms that provides a maximum likelihood explanation of the expression data in a given locus.

The main difference between the output of the two methods results from the different approach each assembler uses when reconstructing low abundant transcripts [33]. Whereas Scripture employs a significance threshold on expression levels to filter out transcripts and reports all significantly expressed isoforms [31], Cufflinks will report a minimal set of isoforms that explains the expression of a given locus without employing such a threshold [32]. The observed discrepancy across lincRNA loci is comparable to that previously observed for low expressed protein-coding genes reconstructed by these assemblers [33].

To obtain transcriptome assemblies from the read alignments, Cufflinks version V1.0.0 was run using default parameters (and ‘min-frags-per-transfrag=0’) and Scripture version 1.0 was run with default parameters, but without using paired-end information (to avoid conflicts that occurred while running Cufflinks abundance estimation mode in later steps). In case several lanes were available per sample the corresponding aligned reads were pooled to a single source file prior to running the assemblers.

lincRNA classification pipeline

After obtaining a unique set of assembled isoforms from all processed tissue assemblies and known annotations we ran the set through the following filters:

(1)Size selection. We excluded single exon transcripts and ones smaller than 200 bases.

(2)Minimal read coverage threshold. We ran Cufflinks with its transcript abundance calculation mode to estimate the read coverage of each transcript across the 24 tissues and cell types. We eliminated transcripts with a maximal coverage below 3 reads per base. This coverage threshold was set by optimizing the sensitivity and specificity of identifying full length vs. partial length transcripts of protein coding genes annotated in Refseq or non-coding genes annotated in UCSC. To this end, we calculated the number of full length and partial length transcripts identified at each coverage threshold (considering the maximal coverage threshold in which a transcript was identified across all tissues). We used area under the curve (AUC) calculations to determine the optimal threshold for the coding and non-coding sets and took their average as the final threshold.

(3)Filter of known non-lincRNAs annotations. We eliminated all transcripts that had an exon overlapping a transcript from any of the following sets: (a) coding genes annotated in RefSeq, UCSC or GENCODE 4, (b) microRNA, tRNAs, snoRNAs, rRNAs annotated in Ensembl, (c) pseudogenes. See **Supplementary Table 1.5** for specific details on each annotation set.

(4)Positive coding potential threshold. We estimated for each transcripts the degree of evolutionary pressure on sequence substitutions acting to preserve an open reading frame. To this end, we scored the coding potential of all remaining transcripts using PhyloCSF (phylogenetic codon substitution frequency) [34]. Briefly, PhyloCSF determines whether a multi-species nucleotide sequence alignment in a specific locus is more likely to represent a protein coding than a non-coding transcript. To do so it applies a probabilistic model that examines the over-representation of evolutionary signatures characteristic of alignments of conserved coding regions, such as the high frequencies of synonymous codon substitutions and conservative amino acid substitutions. We ran PhyloCSF using a multiple sequence alignment of 29 mammalian genomes (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/) to obtain the best scoring ORF greater than 29 amino acids across all three reading frames. We excluded from our catalog all transcripts that contained such an open reading frame with a PhyloCSF score greater than 100. The PhyloCSF score threshold was determined to optimize the specificity and sensitivity in correctly classifying coding and noncoding transcripts annotated in RefSeq (RefSeq coding and RefSeq lincRNAs; see below and in **Supplementary Table 1.5**). PhyloCSF =100 corresponds to a false negative rate of 6% for coding genes (6% of coding genes are classified as non-coding) and a false positive rate of ~10% (9.5% of noncoding transcripts are classified as coding).

(5)Known protein domain filter. We evaluated which of the remaining transcripts contained a known protein coding domain. To this end, we translated each transcripts in all three possible frames and used HMMER-3 [65] to identify occurrence of any of the 31,912 known protein family domains documented in the Pfam database (release 24; used both PfamA and PfamB ; [35]). Briefly, HMMER-3 uses hidden Markov models (HMMs) to scan each amino acid sequence and classify whether it resembles any of the known domains in the database. We have excluded transcripts with a Pfam hit that was declared significant.

(6)Intergenic classification. We classified all the remaining transcripts that did not overlap the genomic region of a known non-lincRNA annotation as potential lincRNAs. All transcripts that passed all the filters above are available in **Supplementary Dataset 1**.

To derive a unique set of lincRNAs that includes previous annotations we used Cuffcompare to integrate the RNA-Seq derived lincRNAs with the predetermined set of lincRNAs previously annotated by RefSeq, UCSC or GENCODE 4. The publically available lincRNA sets were derived by running specific steps of our lincRNA classification pipeline on the transcripts annotated in the public datasets (**Figure 2.1a**, See **Supplementary Table 1.5** for specific details).

Notably, there were 364 loci (4 %) for which the number of isoforms identified by Scripture was over 10 (and often over few dozens or hundreds). Such complexity of isoform makes it impossible to estimate abundance in our downstream analysis. We therefore considered only the isoforms identified by Cufflinks and public annotations for these loci and provide the additional possible isoforms at http://www.broadinstitute.org/genome_bio/human_lincrnas.

Stringent set classification

We defined a *stringent lincRNA set* that includes transcripts loci for which at least one lincRNA isoform: (i) was reconstructed in at least two tissues, or (ii) was reconstructed by two assemblers in the same tissue. An isoform was considered as reconstructed by two sources in case an exact sub-sequence of exon-intron boundaries was identified in the isoforms from the two sources. That is, at least two neighboring exons were independently assembled by two sources. This criterion can still be sensitive to errors in the alignments as both assembly approaches use the same input alignments.

Classification of *Transcripts of Uncertain Coding Potential* (TUCP)

We defined the *Transcripts of Uncertain Coding Potential* (TUCP) set as the set of transcripts that passed the size and abundance filters of our pipeline but were either: (i) classified as pseudogenes by Vega [66], (ii) had a PhyloCSF score that exceeded our determined threshold, or

(iii) had a Pfam domain hit. We included only the maximum likelihood isoforms reconstructed by cufflinks for this set to decrease the isoform complexity in each locus.

While our abundance estimation for pseudogenes may be biased by incorrect read mapping to corresponding paralogous loci, our results estimating global characteristics of expression levels and tissue specificity were similar regardless of whether pseudogene transcripts were included or excluded from the TUCP set (data not shown).

Assembly performance estimation on protein coding genes

We estimated the number of protein coding transcripts annotated in RefSeq that were correctly assembled by Cufflinks, Scripture or both by counting how many of the known transcripts had a partially compatible or fully compatible transcript (**Supplementary Figure 1.1a**) assembled from any of our 24 tissues and cell-lines. A transcript was considered as having a fully compatible assembled isoform if the exact exon-intron chain was recovered in the assembled transcript, and partially compatible if at least 2 exact exons and their connecting intron were recovered in the assembled transcript (**Supplementary Figure 1.1a**).

lincRNAs catalog and annotation

The complete lincRNA catalog (including the TUCP transcripts) as well as all RNA-Seq alignments and transcriptome reconstruction are available at: http://www.broadinstitute.org/genome_bio/human_lincrnas. Specific description of all characterization fields are provided on the site. All **Supplementary Datasets** are available through the site.

Chromatin state of actively transcribed genes

We screened for evidence of a chromatin signature of histone 3 lysine 4 tri methylation (H3K4me3) across the promoter followed by a H3K36me3 along the transcribed region (K4-K36 domains) across lincRNAs loci. To this end, we used publically available segmentations of K4 and K36 chromatin domains detected from ChIP-Seq (chromatin immunoprecipitation followed by sequencing) data across 9 ENCODE cell lines [38]. These cell lines (with the one exception of lung fibroblast), are not included in our 24 tissue/cell line compendium. We identified all lincRNAs that had an enrichment of H3K4me3 within +/- 2 KB of the transcript start site and enrichment of H3K36me3 somewhere along the transcripts genomic region in the same cell type. We defined lincRNAs with such overlap as K4-K36 lincRNAs whenever neither the K4 nor the K36 chromatin domains were also overlapping a coding gene. We also included in the K4K36 lincRNA set those lincRNAs that overlapped genomic regions previously classified as transcribing human lincRNAs based on similar chromatin state maps [14]. Details of the random model generation are specified below.

Chromatin state of enhancers

To evaluate the number of lincRNAs that coincide with enhancers we examined genomic regions previously classified as containing a chromatin domain characterizing enhancers [38]. In brief, Ernest et al. applied a multivariate hidden Markov model to classify the chromatin state of 200 base windows along the genome across 9 encode cell lines based on the combinatorial patterns of 9 chromatin marks. We enumerated the number of lincRNAs that had an exon overlap with a region that was classified as a strong enhancer (state #4) at any of the 9 cell types. Since the chromatin state classification were assigned to just a sub region of the transcript and such sub region may have been classified as having different states in different cell type, we also applied a more stringent criteria to define potential enhancer regions. In this criterion, we now focused on the 200 bases windows in which an enhancer state was classified in the majority of cells in which the chromatin state classification was different than heterochromatin (state #13). This includes all regions that were classified as enhancer in only one or few tissues and not by other alternative states (such as: promoters, transcription elongation etc').

Enhancers sets used for the comparison with mouse lincRNAs from a same/similar cell types were obtained as following. Enhancers that were defined in mouse embryonic stem cells based on regions bound by *CHD7* and *p300* and also enriched with H3K4me1 and H3K27ac were obtained from [40]. Intergenic enhancer elements that produce eRNA in mouse cortical neurons based on binding to *p300/CBP*, enrichment in H3K4me1 and RNA-Seq support (> 7 reads) were obtained from [25].

Estimating expression abundance and normalization

We estimated the expression abundance of all lincRNAs and protein coding genes by running Cufflinks in its expression abundance estimation mode across our 24 samples [32]. We have used the complete non-coding transcripts catalog and all coding transcripts annotated in UCSC for a comprehensive representation of transcripts along the genome while performing abundance estimation. FPKM calls were log2 normalized (after addition of ε=0.05). The HeLa and liver samples from the 8 samples set were eliminated from further expression analysis due to low coverage and a lower expression range in comparison to other samples.

Estimating expression abundance: note

In some cases, the abundance estimation does not represent a specific isoform of interest but rather a different transcript. This is due to cases when there is read coverage across the exons but no spliced reads supporting the specific transcript of interest (as displayed using the Integrative Genome Viewer [69] in **Supplementary Figure 1.14**). This problem will affect any transcript abundance estimation method that doesn't explicitly constrain requirements for spliced read support of a specific isoform. To the best of our knowledge such method is not yet available. To address this issue, we flagged each transcript in the catalog that did not have spliced read support in the tissue where it is maximally expressed and excluded these transcripts from the neighbor correlation analysis (below).

Normalization of expression vectors for tissue specificity calculation

To calculate the tissue specificity scores of a transcript we needed to convert the transcript's expression vector to an abundance density (as the JS metric is applied on discrete probability distributions). To this end, we added a pseudo-count of 1 to the raw FPKM expression vector of each transcript and applied a log 2 normalization to obtain a non-negative expression vector. We then normalized this expression vector to a density vector by dividing by the total expression counts. Formally:

(1) $V' = \frac{\log_2(V+1)}{\sum_{i=1}^{n} \log_2(v_i+1)}$, where $V = (v_1, \dots, v_n)$ is the original raw FPKM abundance estimation of the transcript and V' is the new normalized density vector.

Tissue Specificity Score

To evaluate the tissue specificity of a transcript we relied on [32] and devised an entropy-based measure that quantifies the similarity between a transcript's expression pattern and another predefined pattern that represent an extreme case in which a transcript is expressed only in one tissue . This specificity measure relies on the Jensen-Shannon (JS) divergence. The JS divergence of 2 discrete probability distributions p^1, p^2 is defined to be:

$$JS(p^1, p^2) = H\left(\frac{p^1+p^2}{2}\right) - \frac{H(p^1)+H(p^2)}{2} \tag{2}$$

where H is the entropy of a discrete probability distribution

$$p = (p_1, p_2 .., p_n), 0 \leq p_i \leq 1 \; and \; \sum_{i=1}^{n} p_i = 1$$

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{3}$$

Relying on the theorem that the square root of the JS divergence is a metric [61], we define the distance between two tissue expression patterns $e^1 and\ e^2$, $e^i = (e_1^i, .., e_n^i)$ as

$$JS_{dist}(e^1, e^2) = \sqrt{JS(e^1, e^2)} \tag{4}$$

The tissue specificity of a transcript's expression pattern e across *n* tissues with respect to tissue *t* can then be defined as:

$$JS_{sp}\,(e|t) = 1 - JS_{dist}(e, e^t) \tag{5}$$

where e^t is a predefined expression pattern that represents the extreme case in which a transcript is expressed in only one tissue. Formally: $e^t = (e_1^t, .., e_n^t)\ , s.t\ e_i^t = \{ \begin{matrix} 1 & if\ i = t \\ 0 & otherwise \end{matrix} \}$.

Finally, we define the tissue specificity score of a transcript as the maximal tissue specificity score across all *n* tissues of the transcripts expression pattern *e*:

$$JS_{sp}(e) = argmax_t\ JS_{sp}(e|t),\ t = 1 \dots n \tag{6}$$

K- means clustering of tissue expression patterns

We applied K-means clustering with random seeds to obtain clusters of lincRNAs and protein coding genes using the JS_{dist} (**2.4 Methods**) as a distance metric. Applying the JS distance metric allowed better separation of heterogeneous clusters to tissue specific clusters as compared to using a Euclidean distance or Pearson correlation as a distance metric. For the clusters presented in this paper we used K=30 after optimizing the selection of K to minimize the distances of data within clusters while maximizing the distance between clusters using a Silhouette function [70] (**Supplementary Figure 1.15**). Briefly, we used K-means clustering with 31 values of K (k=20,25,30,…,80) . For each run, we calculated the Sillhouette function on each transcript's expression pattern e^i:

(7) $$Si(e^i) = \frac{b(e^i) - a(e^i)}{\max(a(e^i), b(e^i))}$$

where

(8) $a(e^i) = E(Dist(e^i, e^j)\ |e^i \in c^x \ and\ e^j \in c^x)$, where c^x is the cluster to which e^i was assigned.

(9) $b(e^i) = \min_{c^x} E(Dist(e^i, e^j)\ |e^i\ not \in c^x \ and\ e^j \in c^x)$

That is, *a(e^i)* is the average distance of a sample to all other samples in its cluster and *b(e^i)* is the minimal average distance of a sample from all other clusters.

We then calculated two summary statistics:

(i) the mean silhouette value across all transcripts: $A = E\left(Si(e^i)\right),\ i \in (1 \dots n)$

(ii) the mean of mean cluster Silhouette scores across *m* clusters :

(10) $$B = E(Si_{cluster}(c^x)), x \in (1 \dots m)$$

(11) where, $Si_{cluster}(c^x) = E(Si(e^i)|\ e^i \in c^x)$

K=30 obtained a maximal Sillhouette score according to both statistics and was chosen for our analysis.

Random permutation model of intergenic transcripts

We used the following random permutation model to calculate the enrichment of our lincRNA set with different genomic features (e.g. K4K36 domains, enhancer regions). For each calculation we generated a 1,000 random sets that preserved the same transcript structures distribution of our original lincRNA set but were uniformly sampled from the un-annotated fraction of the genome. Thus, given our transcript set we have randomly re-positioned each transcript in the un-annotated fraction of the genome. An enrichment P-vale was estimated by the number of times the number of overlaps between a random set and the genomic feature of interest exceeded that of our lincRNA set. We defined the un-annotated fraction of the genome by removing centromer regions and all regions that were annotated with a transcript by GENCODE 4.

Functional enrichment analysis of coding gene sets

We estimated the enrichment of different coding gene sets with Gene Ontology (GO) [67] functional terms using the David Bioinformatics Tool [68] and reported the results for GO-FAT biological process terms. GO-FAT are a sub-set of the GO annotation set derived by David by eliminating broad GO-terms that are high in the GO-term tree hierarchy. This is designed to avoid redundancy of annotation sets and overshadowing of the broad terms when applying multiple testing corrections.

Expression correlation of lincRNAs and their neighboring genes

We estimated the expression correlation between the expression pattern of a lincRNA and its closest coding gene neighbor by calculating the Pearson correlation coefficient between their density-normalized expression vectors (as described above). Density normalization was applied in order to compare vectors of similar magnitudes (as the expression levels of coding genes is 10 orders of magnitude higher than that of lincRNAs).

Identification of Trans-mapped syntenic orthologs of human lincRNAs

We have downloaded all available TransMap mappings of expressed transcripts to the human genome (NCBI39/Hg19) from the UCSC Genome Browser (http://genome.ucsc.edu/ ;[51]. The TransMap methodology maps all annotated transcripts of one species to the other by using the syntenic BLASTZ alignments between two species [52]. First, it aligns all mRNA sequences of species *a* to its own genome. Then it uses the syntenic alignment between species *a* and *b* to project the mRNA sequence of *a* to the genome of *b* and finally refines this mapping. We have crossed all UCSC, Refseq, mRNAs and ESTs transcripts Trans-mapped to human with our lincRNA set and included every lincRNA that had an exon overlap with a Trans-mapped transcript in the Trans-mapped lincRNA set. We have used the UCSC classification of coding and non-coding transcripts applied on human and mouse transcripts known to UCSC (and downloaded from the UCSC genome browser ; [29]).

Refined alignment of human lincRNAs and their mouse orthologs

To assess the alignment quality of the TransMap lincRNAs and their syntenic orthologs we realigned the transcript sequence of all human lincRNAs and their mouse orthologs using the Fast Statistical Alignment algorithm with default parameters [55]. We compared the fraction of aligned bases and fraction of identical bases with respect the human reference sequence across 4 sets: (a) human lincRNAs and their mouse orthologs, (b) random sequence pairs, (c) randomly selected syntenic blocks and (d) orthologous coding genes known to Refseq (**Figure 4f** and **Supplementary Figure 1.10c-d**). The random sequence pairs were obtained by shuffling the human lincRNAs and mouse ortholog pairs. Randomly selected syntenic block were obtained by uniformly sampling 1 KB blocks from the un-annotated fraction of the genome that is also syntenicaly mapped to mouse.

Sequence conservation level estimation in novel transcripts with potential coding capacity

We used the SiPhy algorithm and software package (http://www.broadinstitute.org/genome_bio/siphy/ ;[71]) to estimate ω, which quantifies how well a sequence substitution pattern across a multiple sequence alignment fits a neutral selection model, or is constrained by purifying selection. Specifically, ω represents the deviation ('contraction' or 'extension') of the phylogenetic tree's branch length in a specific position in the genome sequence compared to the neutral tree, based on the total number of substitutions estimated from the alignment of the region of interest across 20 placental mammals (build Hg18, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/). We estimated ω across the exons of: (1) coding genes annotated in RefSeq, (2) TUCP transcripts after excluding pseudogenes, (3) Trans-Mapped lincRNAs, (4) the stringent set lincRNAs, (5) intronic segments that were size-matched to their neighboring coding exon and were randomly selected from the intron, and (6) ancestral repeats of human and mouse. The human-mouse ancestral repeats were downloaded from the UCSC Genome Browser [54] and a subset of 3000 repeats was uniformly sampled and used for this analysis.

lincRNAs in disease associated regions

We downloaded a list of disease-associated SNPs ($P < 5*10^{-5}$) from the National Human Genome Research Institute (NHGRI) catalog of published genome wide association studies [57] . We then extracted regions which are in linkage disequilibrium (LD) with each of the SNPs by first finding the left-most and right-most SNPs that are in LD ($R^2 > 0.5$) with the aforementioned SNP and then finding the closet recombination hotspots as described in [72]. We than crossed our lincRNAs and TUCP sets with disease/trait associated LD regions. The results are reported in **Supplementary Dataset 2 and 6**.

Transcription factor binding motif were first identified using AliBaba 2.1 through TRANSFAC [73] and then scanned to determine whether the motif is conserved by running SiPhy [71] in comparison to 5000 randomly chosen coding gene promoter regions (+/- 2KB from transcription start site).

2.5 References

1. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & development, 2011. **25**(18): p. 1915-27.
2. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.
3. Leighton, P.A., et al., *Disruption of imprinting caused by deletion of the H19 gene region in mice.* Nature, 1995. **375**(6526): p. 34-39.
4. Pandey, R.R., et al., *Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation.* Mol Cell, 2008. **32**(2): p. 232-46.
5. Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.
6. Heo, J.B. and S. Sung, *Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA.* Science, 2011. **331**(6013): p. 76-9.
7. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
8. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.
9. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.
10. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.
11. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
12. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
13. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.
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, 2009. **106**(28): p. 11667-72.
15. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.
16. Ponting, C.P., P.L. Oliver, and W. Reik, *Evolution and functions of long noncoding RNAs.* Cell, 2009. **136**(4): p. 629-41.
17. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nat Rev Genet, 2009. **10**(3): p. 155-9.
18. Nagano, T., et al., *The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin.* Science, 2008. **322**(5908): p. 1717-20.
19. Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.
20. Koziol, M.J. and J.L. Rinn, *RNA traffic control of chromatin complexes.* Curr Opin Genet Dev, 2010. **20**(2): p. 142-8.
21. Huarte, M., et al., *A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response.* Cell, 2010. **142**(3): p. 409-19.
22. Loewer, S., et al., *Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells.* Nat Genet, 2010. **42**(12): p. 1113-7.
23. Hung, T., et al., *Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters.* Nat Genet, 2011. **43**(7): p. 621-9.
24. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011.
25. Kim, T.-K., et al., *Widespread transcription at neuronal activity-regulated enhancers.* Nature, 2010. **465**(7295): p. 182-187.
26. De Santa, F., et al., *A large fraction of extragenic RNA pol II transcription sites overlap enhancers.* PLoS Biol, 2010. **8**(5): p. e1000384.

27. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.* Nature, 2011. **472**(7341): p. 120-4.
28. Amaral, P.P., et al., *lncRNAdb: a reference database for long noncoding RNAs.* Nucleic Acids Res, 2010. **39**(Database issue): p. D146-51.
29. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-46.
30. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.
31. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.
32. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotech, 2010. **28**(5): p. 511-515.
33. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nat Methods, 2011. **8**(6): p. 469-77.
34. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.* Bioinformatics, 2011. **27**(13): p. i275-i282.
35. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **38**(Database issue): p. D211-22.
36. Pruitt, K., T. Tatusova, and D. Maglott, *Chapter 18, The Reference Sequence (RefSeq) Project.* , in *The NCBI handbook [Internet]*.2002, National Library of Medicine (US), National Center for Biotechnology Information: Bethesda (MD).
37. Ponjavic, J., et al., *Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain.* PLoS Genet, 2009. **5**(8): p. e1000617.
38. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
39. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2010. **470**(7333): p. 279-83.
40. Zentner, G.E., P.J. Tesar, and P.C. Scacheri, *Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions.* Genome Res, 2011.
41. Ravasi, T., et al., *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.* Genome Res, 2006. **16**(1): p. 11-9.
42. Girard, A.l., et al., *A germline-specific class of small RNAs binds mammalian Piwi proteins.* Nature, 2006. **442**(7099): p. 199-202.
43. Ebisuya, M., et al., *Ripples from neighbouring transcription.* Nat Cell Biol, 2008. **10**(9): p. 1106-13.
44. Cohen, B.A., et al., *A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.* Nat Genet, 2000. **26**(2): p. 183-6.
45. Hurst, L.D., C. Pal, and M.J. Lercher, *The evolutionary dynamics of eukaryotic gene order.* Nat Rev Genet, 2004. **5**(4): p. 299-310.
46. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
47. Seila, A.C., et al., *Divergent transcription from active promoters.* Science, 2008. **322**(5909): p. 1849-51.
48. Preker, P., et al., *RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.* Science, 2008. **322**(5909): p. 1851-1854.
49. Young, T.L., T. Matsuda, and C.L. Cepko, *The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina.* Curr Biol, 2005. **15**(6): p. 501-12.
50. Chodroff, R.A., et al., *Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes.* Genome Biol, 2010. **11**(7): p. R72.
51. Zhu, J., et al., *Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage.* PLoS Comput Biol, 2007. **3**(12): p. e247.
52. Schwartz, S., et al., *Human-mouse alignments with BLASTZ.* Genome Res, 2003. **13**(1): p. 103-7.
53. Kent, W.J., et al., *Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11484-9.
54. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
55. Bradley, R.K., et al., *Fast Statistical Alignment.* PLoS Comput Biol, 2009. **5**(5): p. e1000392.
56. Kondo, T., et al., *Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis.* Science, 2010. **329**(5989): p. 336-9.

57. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

58. Gudmundsson, J., et al., *Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations.* Nat Genet, 2009. **41**(4): p. 460-464.

59. 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, 1992. **71**(3): p. 527-42.

60. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.

61. Fuglede, B. and F. Topsoe, *Jensen-Shannon divergence and Hilbert space embedding.* Proceedings of the IEEE International Symposium on Information Theory, 2004: p. 31.

62. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

63. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

64. Haas, B.J. and M.C. Zody, *Advancing RNA-Seq analysis.* Nat Biotech, 2010. **28**(5): p. 421-423.

65. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Research, 2011.

66. Wilming, L.G., et al., *The vertebrate genome annotation (Vega) database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D753-60.

67. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

68. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat. Protocols, 2008. **4**(1): p. 44-57.

69. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotech, 2011. **29**(1): p. 24-26.

70. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.

71. Garber, M., et al., *Identifying novel constrained elements by exploiting biased substitution patterns.* Bioinformatics, 2009. **25**(12): p. i54-62.

72. Raychaudhuri, S., et al., *Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions.* PLoS Genet, 2009. **5**(6): p. e1000534.

73. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.

Chapter 3 | Localization and Abundance Analysis of Human lncRNAs at Single Cell and Single Molecule Resolution

Contributions: Moran N. Cabili, Margaret Dunagin , Patrick McCalahan, Andrew Biaesch, Olivia Padovan-Merhar, Aviv Regev†, John L. Rinn† and Arjun Raj†. (†) Equal contribution.

Abstract

Long noncoding RNAs (lncRNAs) have emerged as key players in different cellular processes and are required for diverse functions *in vivo*. However, fundamental aspects of lncRNA biology remain poorly characterized, including their subcellular localization, abundance and expression variability from cell to cell. Here, we used single molecule, single-cell RNA fluorescence *in situ* hybridization (RNA FISH) to survey 61 lncRNAs, chosen by properties such as conservation, tissue specific expression, and expression abundance, cataloging their abundance and cellular localization patterns in three human cell types. These lncRNAs displayed diverse sub-cellular localization patterns ranging from strictly nuclear localization to almost exclusively cytoplasmic localization, with the majority localized primarily in the nucleus. The low abundance of these lncRNAs as measured in bulk cell populations cannot be explained by high expression in a small subset of 'jackpot' cells. Simultaneous analysis of lncRNAs and mRNAs from corresponding divergently transcribed loci showed that divergent lncRNAs do not present a distinct localization pattern and are not always co-regulated with their neighbor. Overall, our study highlights important differences and similarities between lncRNAs and mRNAs in terms of abundance, localization and cell-to-cell variability. The rich set of localization patterns we observe are consistent with a broad range of potential functions for lncRNA, and will guide the formulation of hypotheses for further mechanistic studies.

Author Contributions

AR^1, JLR and NMC designed the study. MD performed the experiments with guidance by AR^1. PM and AB performed the pilot experiments with guidance by AR^1. MD and MNC performed the image processing. MNC performed the analysis with guidance by AR^1, JLR and AR^2. AR^1, MNC, MD, JLR and AR^2 discussed and interpreted the results. MNC wrote the manuscript and all authors reviewed and revised it. OPM performed experiments and contributed mRNA control data. Arjun Raj (AR^1), Aviv Regev (AR^2).

3.1 Introduction

Thousands of lncRNAs expressed from mammalian genomes await functional characterization [1]. A plethora of studies now demonstrate that dozens of these lncRNAs can regulate diverse cellular processes involved in development and disease onset and progression [2, 3]. While the regulatory roles and mechanism of action of several individual lncRNAs were elucidated, it is still unclear how general these roles and mechanisms are [4].

Many fundamental aspects of lncRNA biology that could provide important insights into their potential mechanism of action are yet to be characterized, including their subcellular localization patterns and even their abundance within single cells. For instance, if a lncRNA resides primarily in the nucleus near its site of transcription it may suggest a transcriptional regulatory function in *cis* [5-11]. Moreover, quantifying lncRNA abundance in single cells can help resolve whether the generally low abundance of lncRNAs in bulk assays is a result of uniformly low abundance per cell or due to a small proportion of cells that highly express the lncRNA. Such questions cannot be addressed by current lncRNA profiles using bulk population assays.

RNA fluorescence *in situ* hybridization (RNA FISH) [12, 13] can be powerfully deployed to address these questions to study lncRNA mechanism of action. For example, early studies of the lncRNA XIST [14, 15], a key regulator of X inactivation and a paradigm for studying chromatin regulation by lncRNAs [16], used RNA FISH to demonstrate that XIST is located on the inactive X-chromosome, suggesting that it may be involved in chromosome architecture regulation [8, 11]. More recently, RNA FISH in combination with immunofluorescence revealed that the lncRNAs MALAT1, NEAT 1 and MIAT co-localize with different speckle-like structures in the nucleus [5, 17-20]. RNA FISH also demonstrated how the lncRNA GAS5 changes its localization in response to stimulus [6]. These examples are, however, amongst the mostly highly abundant RNAs in the cell, whereas the vast majority of lncRNAs are considerably less abundant, precluding the use of conventional RNA FISH techniques that have relatively low sensitivity.

More recently, single molecule RNA FISH techniques, based on hybridization of multiple short, fluorescently labeled, oligonucleotide probes [21, 22], were used to estimate the absolute level and subcellular localization of even low abundance lncRNAs [7, 9, 10, 23-28]. Some studies assessed the correlation between a lncRNA and its putative mRNA target (simultaneously monitored with two differently colored fluorescent dyes) to reveal potential regulatory interactions [10, 25]. For example, combining multicolor correlations with subcellular analysis revealed that lincHOXA1 represses the neighboring Hoxa1 in *cis* in a subpopulation of cells [9]. Apart from an early study that surveyed *in situ* hybridization data to brain tissue and at low resolution [29], no study has systematically applied single molecule RNA FISH to explore lncRNAs from cDNA and RNA-seq catalogs [30-36], where thousands of transcripts await functional characterization. Furthermore, no study has tackled the unique challenges posed by lncRNAs, which are shorter, lower abundance and more likely to contain repeats than mRNA [30, 37].

Here, we used single molecule, single-cell RNA-FISH to characterize the abundance and sub-cellular localization patterns of 61 lncRNAs across three human cell types. We focused on the subclass of intergenic lncRNAs (lincRNAs) [38] from our well-annotated Human lincRNA Catalog [30], and selected a subset spanning a wide range of tissue specificity and expression level and encompassing both syntenically orthologous lincRNAs [30, 35] and divergently transcribed lincRNAs [30, 33, 39-41]. We show that lncRNA FISH is more prone to artifacts and established a pipeline that controls for these by rigorously validating single molecule RNA FISH probe sets. Using validated probe sets we demonstrate that sub-cellular localization patterns vary across lncRNAs, but are consistent across cell types, typically with most exhibiting a much stronger nuclear localization than mRNA. In mitotic cells, lncRNAs do not associate with chromatin, showing that (at least for the examined cases) retention to specific regulatory regions through mitosis is likely not a mechanism of mitotic inheritance. We also measured cell-to-cell variability in abundance, showing that many lncRNAs are in fact quite homogenous across cells, and overall no different than mRNA in their heterogeneity levels. Finally, simultaneous analysis of matching pairs of divergently transcribed lncRNAs and mRNAs shows that these are not always co-regulated and that the localization patterns of divergently transcribed lncRNA do not differ from those of other lncRNAs.

3.2 Results

A single molecule, single cell RNA FISH survey of lncRNAs in 3 human cell types

To characterize the abundance and localization patterns of lncRNAs in the three different cell types, we studied 61 lncRNAs selected to span a range of parameters (**Figure 3.1a**) using single molecule RNA-FISH. We used a two-step approach to validate the single molecule RNA-FISH signals (**Figure 3.1a**): first, we imaged each lncRNA using a complete probe set to identify the lncRNAs for which we obtained an RNA FISH signal, and second, we relabeled the individual oligonucleotides comprising our probe sets to validate the signals we observed as being specific to the target (a 'two-color co-localization' step, below). For those probe sets that did validate, we analyzed the characteristics of the associated lncRNAs.

Specifically, we manually curated a candidate set of 61 lncRNA for screening (**Figure 3.1b-c; 3.4 Methods; Supplementary Dataset 1; Supplementary Table 2.1**), such that: (**1**) the lncRNAs in our set are significantly expressed in at least one of human foreskin fibroblasts (hFF), human lung fibroblasts (hLF) or HeLa cells, the target cell lines for our study; **(2)** the lncRNAs span a wide range of expression levels and tissue specificity (**Supplementary Figure 2.2.1**); (**3**) the set includes a subset of 43 lncRNAs that have an expressed syntenic ortholog in mouse; (**4**) the set includes a subset of 16 lincRNAs that are transcribed divergently to a neighboring mRNA (within 10 KB). These criteria and subsets are not mutually exclusive (**Figure 3.1b**). Finally, we included 16 "classic" previously studied lncRNAs as a point of reference (**Supplementary Table 2.2**). These "classic" lncRNAs are generally far more abundant than the other lincRNAs in our survey. We used three types of mRNA controls (**Supplementary dataset 1.4-1.5;** 34 in total): (**1**) 9 mRNAs transcribed divergently to those 'divergent lncRNAs' in this study; (**2**) CCNA2, a cyclin simultaneously imaged with all other lncRNAs; and (**3**) 24 mRNAs that span a wide range of expression levels in hFF (Padovan and Raj, personal communication).

To visualize single lncRNA molecules, we used an established protocol for single molecule RNA FISH [22], where we design 10-48 complementary DNA oligonucleotides, each 20 bases long and labeled with a single fluorophore at its 3' end (**Figure 3.1a**). When these probes hybridize to a single RNA molecule, the packing of so many fluorophores to a single location renders the RNA molecule detectable by fluorescence microscopy. This method is typically also highly specific as only hybridization of a large fraction of the probe set yields a detectable signal [22]. We successfully designed probe sets and screened 61 lncRNAs in hFF, hLF and HeLa cells (**3.4 Methods; Supplementary Dataset 1.6**), 53 of which presented a detectable signal in at least one cell type.

Since lncRNAs are, on average, more lowly expressed and have higher repeat content [37] than protein coding mRNAs (for which single molecule RNA-FISH is commonly applied; **Supplementary Figure 2.1**), they are more prone to off-target signals. In particular, a single oligonucleotide binding to a highly abundant, well localized off-target transcript (such as those generated from repetitive sequences) can create a spurious signal. For mRNA targets, such signals are readily distinguishable from the dispersed, diffraction-limited spots corresponding to individual mRNA, but several lncRNAs, such as the *cis*-bound lncRNAs XIST or KCNQ1OT1 [8, 26], may actually display such signals legitimately. Indeed, we detected a few intriguing staining patterns in our pilot imaging screen that we nevertheless suspected to reflect off-target hybridization of one or a few oligonucleotides (**Supplementary Figure 2.2**). For instance, by aligning the probe sets' oligonucleotides to the transcriptome, we were able to identify for one such probe set (XLOC_010514) one specific oligonucleotide that was hybridizing to a highly abundant and highly localized RNA (MALAT1) by only 15 nucleotides, yet resulting in a significant detectable signal (**Supplementary Figure 2.2a; 3.4 Methods**).

To control for these and other possible off-target effects, we used a two-color co-localization approach [21, 22] in which we analyze each lncRNA after partitioning each probe set into two subsets ('even' and 'odd' oligonucleotides), each labeled with a differently colored fluorophore (**Figure 3.1a, Supplementary Figure 2.2b-c**). For a valid probe set, the signals from these two subsets should largely colocalize with each other, with the number of colocalized spots roughly equaling those obtained with the full probe set ('quantitative consistency'). Should a single oligonucleotide hybridize to a highly abundant incorrect 'off' target, this would manifest itself as a non-colocalized signal visible in only the odd or even channel ('qualitative inconsistency').

Alternatively, if some other form of non-specific background is present, it can result in a large difference between the number of spots identified with the full probe set and the number of colocalized spots (quantitative inconsistency). Using this validation step, we eliminated 19 probe sets from further analysis, as they had major qualitative or quantitative differences in the two color co-localization assay, underscoring the importance of testing for off target effects for lncRNA-FISH (**Supplementary Figure 2.2d-e**; **Supplementary Dataset 2; Supplementary Table 2.3**). Another 8 probe sets had no discernible signal. We were unable to attribute the cases of no detectable signal or co-localization inconsistencies to low number of oligonucleotides and observed a very slight bias toward lower abundance lncRNAs (Kruskal-Wallis one way analysis of variance $P< 8.4X10^{-3}$; **Supplementary Figure 2.3**). Overall, we proceeded with 34 lncRNAs for which we had a valid detectable signal.

We further evaluated each of the remaining 34 probe sets in each of the screened cell types to ensure the molecule count distribution obtained from the two color co-localization assay and the corresponding single color full probe set assay were quantitatively similar (**3.4 Methods**; **Figure 3.1a**; **Supplementary Figure 2.2e** ; **Supplementary Figure 2.18** ; **Supplementary Dataset 2**). This resulted in a set of 70 lncRNA-cell type pairs with valid signal, corresponding to 34 unique lncRNAs; **Supplementary Dataset 2**; **Supplementary Figure 2.19**). Note that probe validation in each cell type is essential as some probes were valid in one cell type but not in another **(Supplementary Figure 2.4)**. Altogether, we acquired over 2,000 images overall in 3-5 separate fluorescence channels, with 2-3 biological replicates per gene-cell pair (the final analysis included 80, 24 and 28 cells per gene on average, for HeLa, hLF and hFF, respectively).

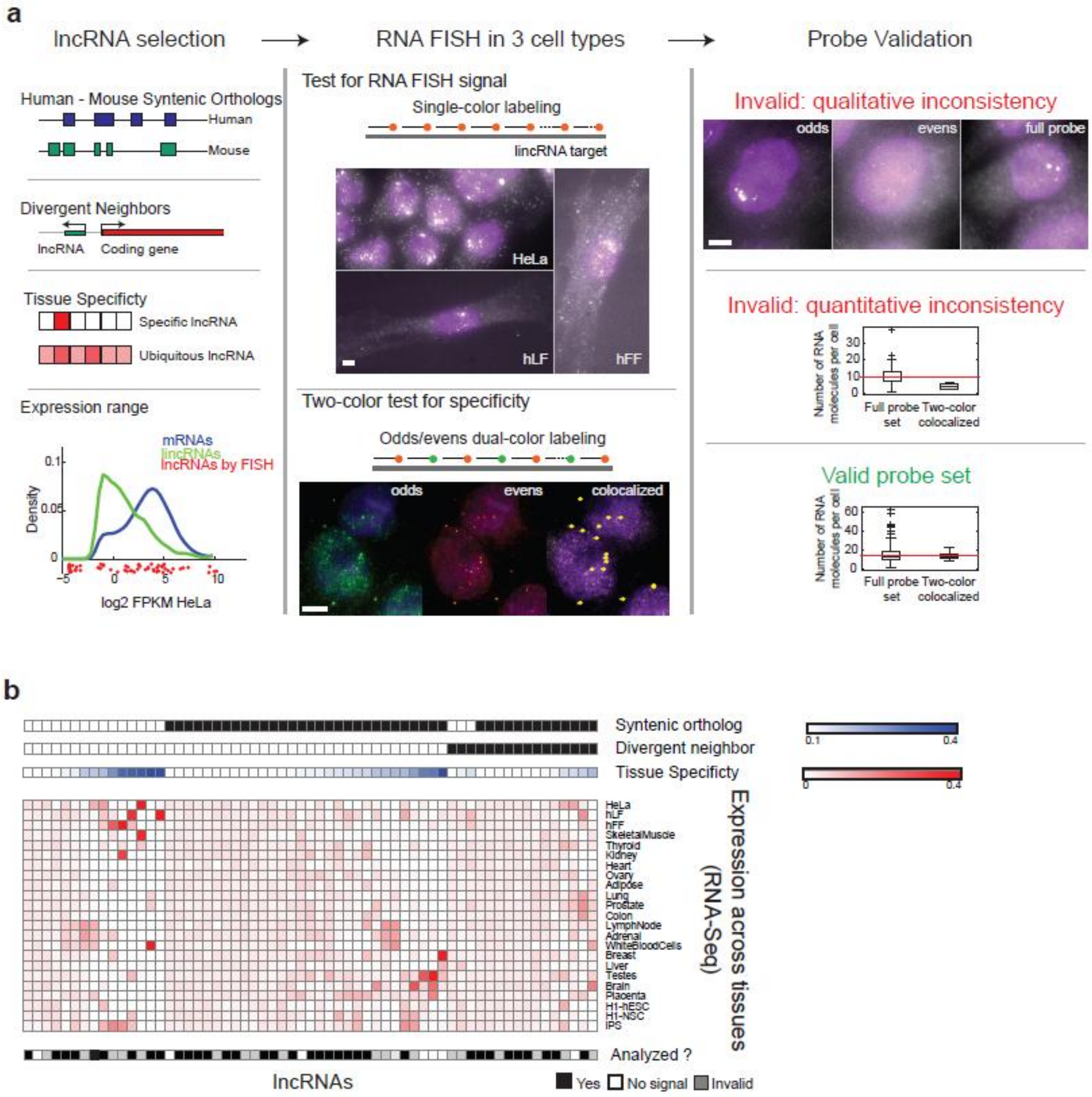


Figure 3.1| An RNA-FISH survey of lncRNAs
(a) Study workflow. **(b)** Key features of 61 lincRNAs for which probe sets were successfully designed and were imaged in the study. Shown are for each of 61 lincRNA (column) the following features from top to bottom: whether it has a syntenic ortholog (black: has ortholog) or a divergently transcribed mRNA neighbor (black: has such neighbor), the extent of tissue specificity across 23 tissues (blue color intensity: maximal tissue specificity score as in [30] across the tissues presented in the figure; white to blue color bar), its expression level as measured by RNA-Seq (red intensity: the fractional density across the row of $\log_2$(FPKM) as estimated by Cufflinks; white to red color bar) in each of 23 tissues (heatmap rows; **Supplementary Dataset 4** ;), and the extent of analysis performed (black: lncRNAs with valid probe set that were included in the final analysis; white: lncRNAs showing no signal; gray: lncRNAs with an invalid probe based on the two-color co-localization assay).

Most lncRNAs are located primarily in the nucleus

The 70 lncRNA-cell type pairs examined (**Figure 3.2**; **Supplementary Figure 2.5**) span a range of localization patterns which we classified as type: (**I**) 1-2 large foci in the nucleus (9 pairs); (**II**) large nuclear foci and single molecules scattered through the nucleus (11 pairs); (**III**) predominantly nuclear, without foci (18 pairs); (**IV**) cytoplasmic and nuclear (28 pairs); and (**V**) predominantly cytoplasmic (4 pairs). Validating our approach, 11 of the 12 lncRNA that were previously imaged by RNA-FISH [6, 8, 19, 23, 42-49] showed patterns that are consistent with previous reports (**Supplementary Dataset 1.3).** These include the large nuclear foci previously observed for XIST and Kcnq1ot1 [8, 11, 44], localization of GAS5 to both the nucleus and cytoplasm [6] and the speckle and para-speckle structures of MALAT1 and NEAT1, respectively [19, 42].

The majority of lncRNAs (55% classified as class I-III; 38 lncRNA-cell type pairs) are predominantly in the nucleus (**Figure 3.3a-b**; **3.4 Methods;** compared to 1/49 of mRNAs using the I-III classification criteria of more than 65% of molecules in the nucleus), with ~13% of lncRNA-cell type pairs mainly located in 1-2 large foci (type I). This bias toward nuclear localization was significant compared to localization of mRNAs (67% of lncRNAs vs. 10% of mRNAs have more than 50% of their RNA in the nucleus; Kolmogorov Smirnov (KS) $P < 13X10^{-11}$; **Figure 3.3a-b**). Within the lncRNA set, divergent lncRNAs presented a slightly higher bias toward nuclear localization (KS $P < 2.12X10^{-2}$; **Figure 3.3c**) while syntenic orthologs did not present such bias over the lncRNA background distribution. The latter set does, however, exhibit a slight bias toward higher expression (KS $P < 3.25X10^{-3}$; **Figure 3.3d**).

In the vast majority (85%) of cases, the lncRNA localization pattern was consistent across the cell types where data was available. The notable exceptions were 5 lncRNAs (lincFOXF1, TERC, XLOC_005764, GAS5, XLOC_002746) that displayed distinct patterns in at least two cell types. These differences, however, appeared mostly to result from changes in overall expression leading to the appearance of additional bright foci in the nucleus (**Figure 3.2**, magenta stars, **Supplementary Figure 2.6-8**; **Supplementary Dataset 3.2**). For example, we identified large lncRNA foci for TERC and XLOC_005764 in HeLa (type II), where they are more abundant (~81 and ~22 molecules per cell, respectively) than in hFF (type III, ~17 and ~4 molecule per cell, respectively), where these foci are missing. Similarly, GAS has dominant nuclear foci in HeLa (type II, ~195 molecules per cell), and less frequent foci in fibroblasts, where its expression is lower (type VI, ~75 molecules per cell). In other cases, higher abundance was associated with the appearance in the cytoplasm as well. For example, lincFOXF1 is more abundant in fibroblasts, where it more frequently appears in the cytoplasm (type VI in fibroblasts *vs*. type II in HeLa).

Focal lncRNAs do not persist in their nuclear foci during mitosis

The appearance of bright nuclear foci of specific lncRNAs raised the question of whether these foci persist through mitosis; if they are involved in transcription regulation, then persistence at the target locus through mitosis could suggest a potential mechanism for the maintenance of epigenetic states through cell division. To address this question we examined the staining in mitotic cells of six such lncRNA (~50% of such cases), including XIST. None of the examined cases presented nuclear foci in cell undergoing mitosis. (The potential foci we observed in ~1/3 of ANRIL mitotic cells were not validated when using 2-color co-localization; **Figure 3.3e; Supplementary Dataset 3.3; Supplementary Table 2.4**). Notably, for five of the lncRNAs, including XIST, we observed some molecules spread throughout the cytoplasm during mitosis (consistent with previous observations for XIST [8]). In the case of XLOC_001515 we did not observe any lncRNA molecules whatsoever during mitosis.

We also checked whether any of our lncRNAs had evidence for G1 or S/G2 dependent expression by simultaneously measuring the cyclin CCNA2 transcript in every image we obtained. CCNA2 expresses exclusively in the S, G2 and M phases of the cell cycle, making its mRNA abundance a marker for cell cycle phase [50, 51]. We identified two lncRNAs that positively correlated in their expression with CCNA2 (lincSFPQ and XLOC_001226), and one negatively correlated (XLOC_011185), irrespective of cell volume (Padovan and Raj, personal communication; **Supplementary Dataset 3.4**; **Supplementary Table 2.5; Supplementary**

Figure 2.9; **3.4 Methods**). This suggests that expression of these lncRNAs is regulated through the cell cycle.

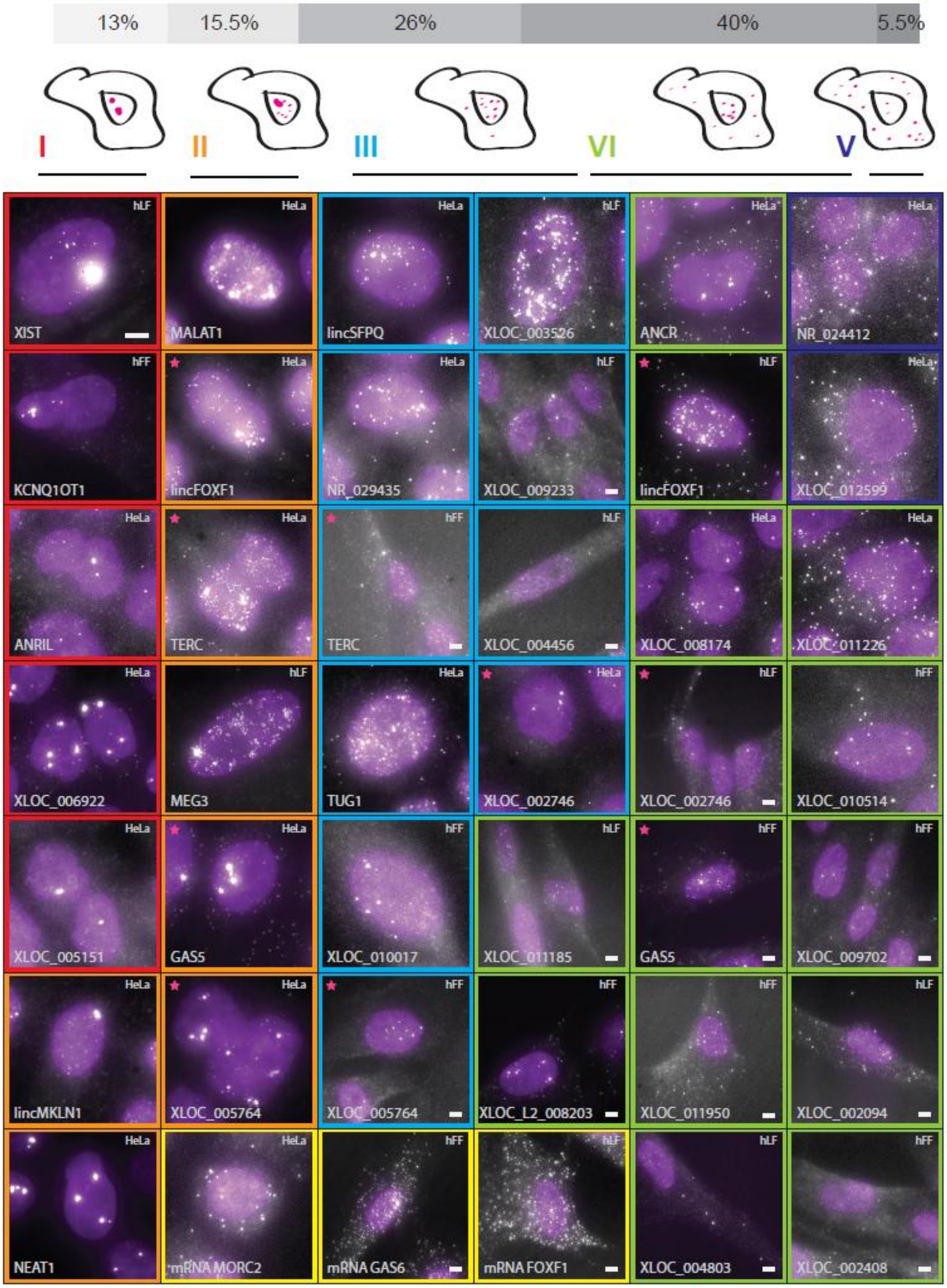


Figure 3.2| lncRNAs exhibit a variety of cellular localization patterns. See legend below.

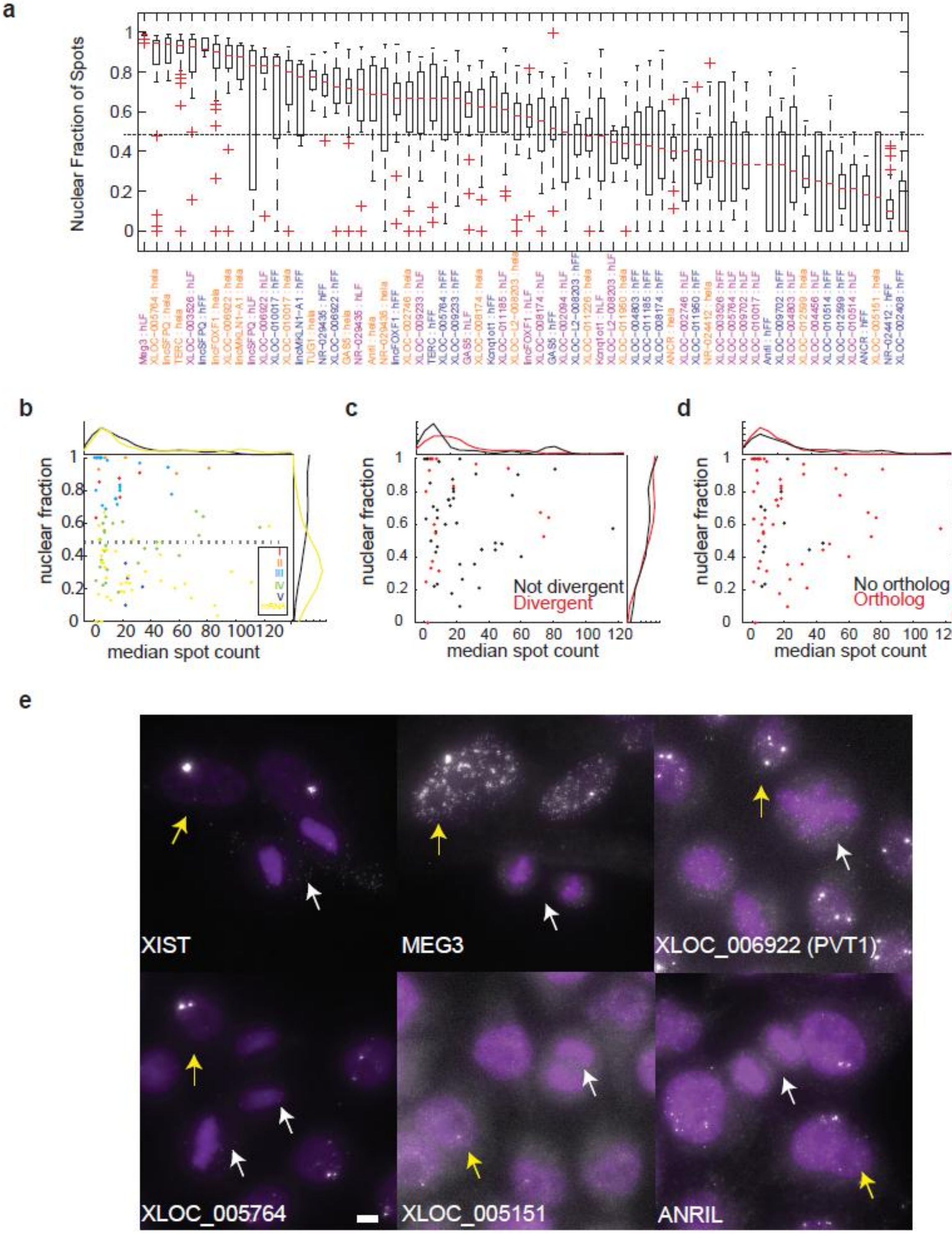


Figure 3.3| Nuclear localization of lncRNAs. See legend below.

Figure 3.2| lncRNAs exhibit a variety of cellular localization patterns. Florescence micrographs of representative expressing cells for each of 34 lncRNAs with a validated probe set. LncRNA-cell pairs are classified to cellular localization types I-V as described in the **3.4 Methods** (marked by their border color). Magenta stars mark five lncRNAs that are presented in two different cell types and two different classes (see same row for comparison). Scale bar, 5 µm; when a scale bar is not specified, reference the scale bar within the top left image. Top panel: fraction of each classification for each type across the full set of 70 valid lncRNA-cell pairs imaged.

Figure 3.3| Most lincRNAs are predominantly localized to the nucleus
(a) Boxplots describing the distribution of the fraction of molecules localized to the nucleus (Y axis) for each validated lncRNA-cell pair (X axis, orange: HeLa, blue: hFF, purple: hLF). Red bar: medians. Whiskers are is 1.5* the inner quartile range. **(b)** Scatter plot of the relationship between expression level (X axis; median number of molecules per cell) and nuclear localization (Y axis, median fraction of nuclear spots across all expressing cells). Each data point is one gene-cell pair and is colored by its classification to the localization classes I-V (**3.4 Methods**) of Figure 2. mRNA sets 1-2 (yellow) serve as a reference. Histograms on top and right are the distribution of all lncRNAs- (black) and mRNA- (yellow) cell pairs. **(c)** Scatter and histograms as in (b) but for lncRNA with (red) or without (black) a divergently transcribed mRNA counterpart. **(d)** Scatter and histograms as (b) but for lncRNA with (red) or without (black) a syntenic ortholog. **(e)** Representative image of mitotic cells (marked with white arrows) lacking foci that are seen in interphase cells (marked with yellow arrows). Scale bar, 5 µm.

The extent of cell-to-cell variability in lncRNA expression is not markedly greater than that of mRNAs

When measured in bulk cell populations, lncRNAs are typically expressed at low levels compared to mRNAs [1, 30]. Several studies have hypothesized that these bulk measurements may obscure an extreme cell-to-cell heterogeneity, such that lncRNA are expressed very highly in a small fraction of cells, but lowly or not at all in most others cells, resulting in average low expression [52, 53]. We tested this hypothesis by quantifying the cell-to-cell variability of the lncRNAs in our panel.

We first confirmed that the average (cell population) expression level estimates for our lncRNAs are generally consistent between RNA-FISH and RNA-Seq (Pearson ρ = 0.52; P-value < $4.3X10^{-6}$; **3.4 Methods**), with the discrepancies possibly due to the high variability in RNA-Seq abundance estimates for some of the examined transcripts (**Supplementary Figure 2.10**). The distribution of single cell counts demonstrated the relatively low expression of lncRNAs across cells, with 43% of lncRNA-cell pairs having 10 or less molecules per cell on average and with a median of 14 molecules across all gene-cell pair distribution medians (*vs*. 36 for the 49 mRNA-cell pairs we examined; **Figure 3.4a**).

In most cases, cell-to-cell variability in lncRNA levels did not reveal the presence of low frequency, highly expressing cells, and is similar to that of protein coding mRNAs expressed at comparable average levels (**3.4 Methods**; **Figure 3.4c**). In particular, with one notable exception (described below), the mean and the median molecule counts were similar, highlighting the lack of outlier cells in the single cell distributions (**3.4 Methods**; **Figure 3.4b**; **Supplementary Figure 2.15**; Pearson ρ = 0.98, P-value <2.5X10^{-39}).

Since we only obtained a few dozen cells for most of the lncRNA-cell line pairs examined (due to limited imaging throughput), we cannot rule out the possibility of a particularly rare cell with extraordinarily high expression levels. To increase our power, we imaged 500-700 cells for each of 4 lncRNA in HeLa cells (**Supplementary Figure 2.12**), including XLOC_004456 that displayed no signal in HeLa in our initial assessment (**Supplementary Figure 2.3**). None of these images revealed the presence of any highly expressing outlier cells. With a sample size of n=500 cells, we can place an upper bound of 0.6% of cells that may be outliers with a statistical power of 0.95 (**3.4 Methods**).

The one case we found of clear 'jackpot' outlier cells is the tissue specific lncRNA XLOC_003526 (**Figure 3.4d-e**): it is lowly expressed on average (FPKM <1 in a population of hLF RNA-Seq, with few, if any, spliced reads; **Supplementary Figure 2.11**), but in RNA-FISH ~25% of the cells express it highly (107 +/- 26 molecules on average), whereas the other cells express it very lowly (9 +/- 1.2 molecules on average). Moreover, cells highly expressing XLOC_003526 presented a distinct non-random localization pattern reminiscent of MEG3. XLOC_003526 has no known expressed orthologs and is encoded in a poorly conserved 900 Kb gene desert. Expression of XLOC_003526 does not correlate with a specific phase of the cell cycle (as defined by CCNA2 expression in the same cells; above). Indeed, the distinct localization pattern of XLOC_003526 was not observed in any of the 12 mitotic hLF cells we examined (**Supplementary dataset 3.3**).

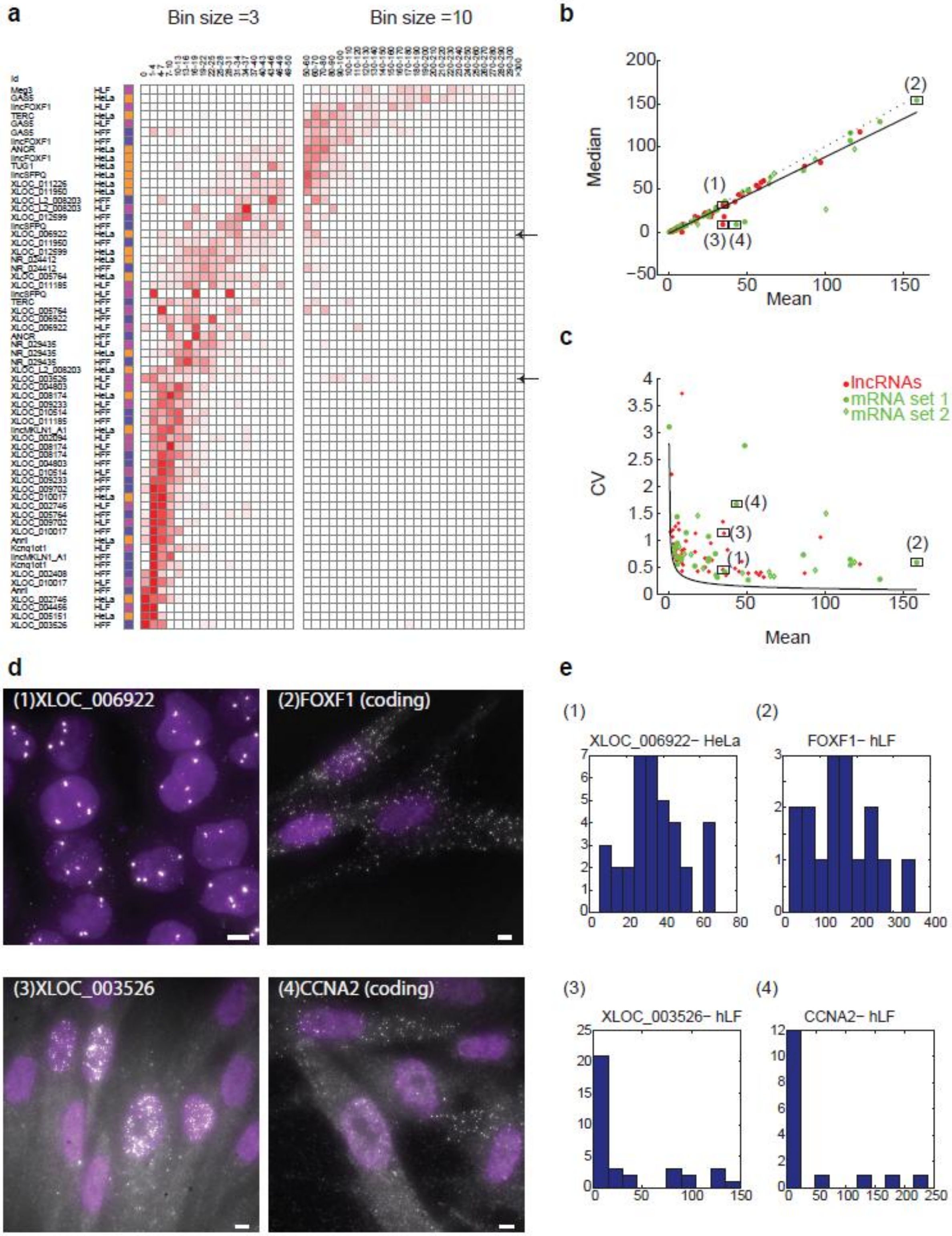


Figure 3.4 | Cell-to-cell variability does not appear to explain the low abundance of the lncRNAs in our survey. See legend below.

Figure 3.4 | Cell-to-cell variability does not appear to explain the low abundance of the lncRNAs in our survey

(a) Distribution of RNA single molecules counts (bins, columns; Red color intensity: fractional density of molecule counts across the population) for the 64 lncRNA-cell pair in the validated set that are quantitative (rows, **3.4 Methods**). Cell type is color coded in left column (orange: HeLa orange; blue: hFF; purple: hLF). Values from 0-50 molecules are in bins sized 3 (left); values 50-300 are in bins sized 10 (right) with the last bin for all cells with more than 300 molecules. One lncRNA (XLOC_003526) has heterogeneous expression across cells, and one with homogenous expression (XLOC_006922), are pointed by black arrows and referenced in figures b-c. **(b,c)** The relationship between the mean molecule count (X axis) vs. median molecule count (Y axis, b) or vs. variability in molecule counts (Y axis, coefficient of variation, c) for the 64 lncRNA-cell pairs in the quantitative validated set (red), mRNA set 1 (green circles; **3.4 Methods**) and mRNA set2 (green diamonds; **3.4 Methods**). A linear regression line in b (black) supports the consistency of the majority of transcript-cell pairs with a unimodal distribution (Y=0.87X-1.25, Pearson ρ = 0.96). Dotted line is Y=X. Black curve in (c) is the theoretic Poisson distribution. Four transcripts marked (1-4) are analyzed further in d-e. LncRNA pairs with mean > 170 (less than 10% of all pairs) are not presented, but show a similar pattern on a log scale. **(d)** Fluorescence micrographs of single molecule RNA FISH of a homogenously expressed lncRNA (1-XLOC_006922; top left) and mRNA (2-FOXF1; top right) and of a heterogeneously expressed lncRNA (3-XLOC_003526; bottom left) and mRNA (4 – CCNA2; bottom right). XLOC_003526 and CCNA2 are both heterogeneous but do not correlate with each other based on co-staining in two colors. Scale bar, 5 µm. **(e)** Molecule count distributions for each of the example transcript 1-4.

Cellular localization of divergently transcribed lncRNA-mRNA transcript pairs

We have previously distinguished a subset of lincRNAs that are transcribed divergently from protein coding genes' promoters (~500, ~13% of human lincRNAs [30, 33]; **Figure 3.5b**). These may be related to the widespread phenomena of divergent transcription [39-41, 54, 55], but are distinguished by the fact that the resulting transcripts are stable, processed and spliced. While not much is known about the function of this subset and their site of action, one hypothesis is that divergent lncRNAs are co-regulated with their neighbors and possibly have a regulatory effect on their neighbor at the transcription site [33, 56]. Several studies using bulk assays observed that divergent transcripts are usually co-expressed [33, 40, 41, 56]; To look for such correlations at the single cell level, we co-stained cells for divergent lncRNA and their mRNA neighbor, looking for any correlations in single cell abundances as well as any distinctive spatial localization patterns that might indicate local regulation (such as localization to one or few foci such as for the cis acting XIST/KCNQ1OT1 (**Figure 3.2**).

We therefore used two-color RNA-FISH to analyze eight of the nine candidate divergent lncRNAs for which we had valid probe sets along with their neighboring mRNA (**Figure 3.5; Supplementary dataset 3.6**). We observed that in most cases (7/8) the bi-directionally promoted lncRNAs were not simply localized at one or few foci (characteristics of type I; likely to be the site of transcription), but rather were localized throughout the cell (**Figure 3.5a-b; Supplementary Figure 2.13**). For example, XLOC_011950 and XLOC_010514 have a substantial cytoplasmic fraction and no nuclear foci (type VI). NR_029435, TUG1 and XLOC_009233 are mostly nuclear but with no apparent foci (type III). Lastly, while lincMKLN1 (type II), lincFOXF1 and GAS5 (type II and VI) all present nuclear foci in some cell types, lincFOXF1 and GAS5 are also substantially expressed outside these foci and in the cytoplasm. Together, the different subcellular localizations displayed by divergent lncRNAs were distinct from each other, and were not qualitatively different from those of the other lncRNAs in our survey.

We also observed a spectrum of correlation and expression levels of the lncRNA and its neighboring protein coding gene (**Figure 3.5b**). lincFOXF1 and its neighbor are both tightly correlated (Pearson ρ =0.91) in hLF and expressed at similar expression levels (~121 and 158 molecules per cell on average, respectively). XLOC_010514 and its neighbor are also tightly correlated in hLF (Pearson ρ =0.84), but the lncRNA is an order of magnitude more lowly expressed (~256 and 7 molecules per cell on average, respectively). XLOC_011950 and its neighbor are positively correlated and expressed in a similar range in HeLa, but do not correlate in hFF, where they are still expressed to the same extent on average (**Figure 3.5c**; **Supplementary Figure 2.14**). NR_029435 and GAS5 are positively correlated with their neighbors in HeLa (Pearson ρ= 0.4 and 0.43, respectively) and expressed to a higher extent than their coding neighbor (2 fold and 30 fold increase, respectively), although it is unclear whether these relatively mild correlations result from a generic correlation with cellular volume (Padovan and Raj, personal communication). We note that there was no correspondence between the existence of an expression correlation between the lncRNA and its neighbor and a particular subcellular localization pattern. Taken together, while divergent lncRNA in this study share a common genomic layout, they do not have the same localization pattern, expression levels nor co-expression levels with their neighboring coding gene.

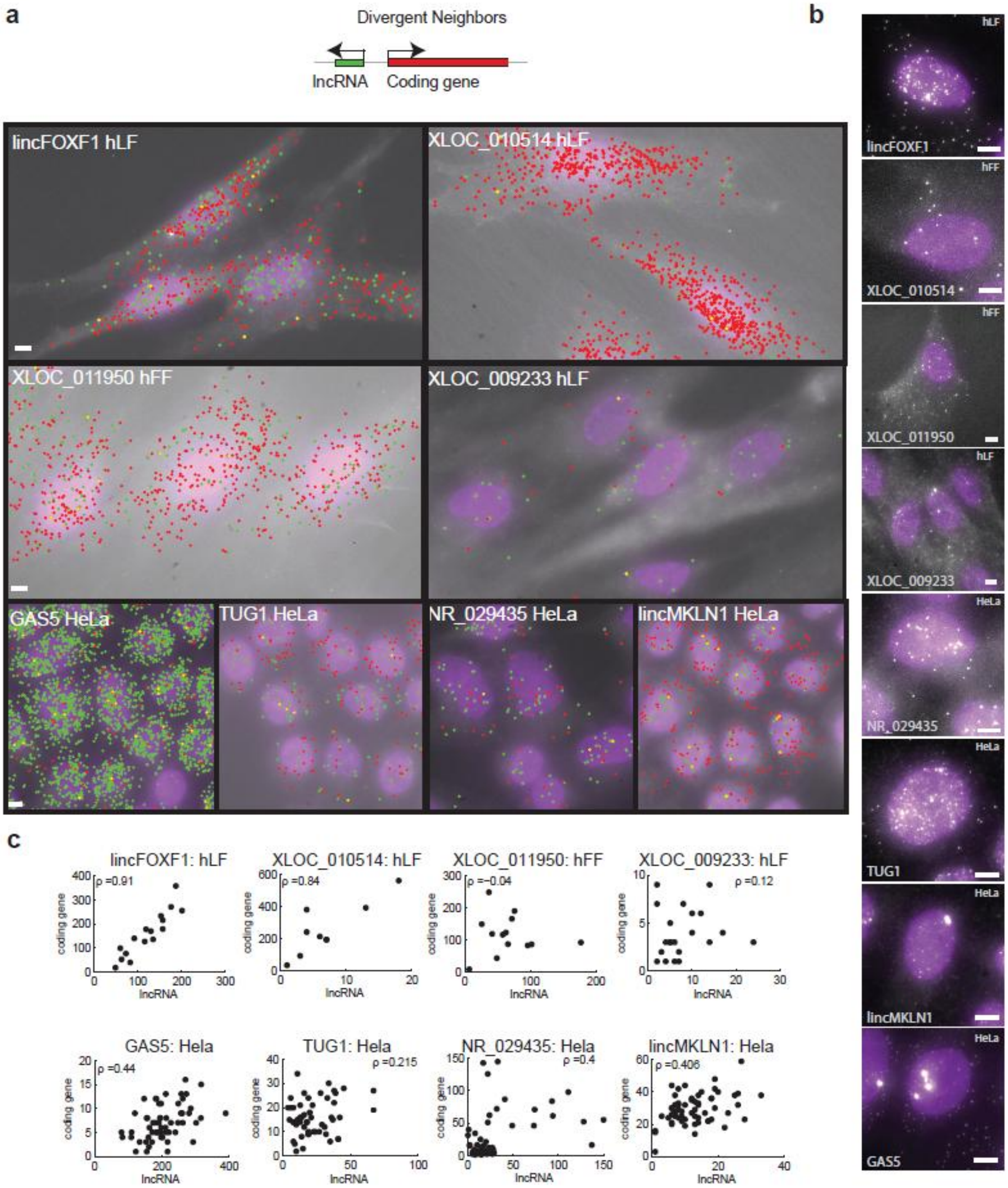


Figure 3.5 | Cellular localization of divergent lincRNAs and their neighbors
(a) Two-color overlay micrograph presenting florescence probes targeting the lincRNA (green) and coding neighbor (red). Co-localized spots are marked yellow. The lincRNA and cell type are marked on the image. Scale bar, 5 µm; marked on the left most image. Top: illustration of the positional genomic orientation of a divergent lincRNA and its coding gene neighbor. **(b)** Representative fluorescence micrographs as shown in *Figure 2* for the lincRNAs in *a*. Scale bar, 5 µm. **(c)** Scatter plots of the relationship in each cell between the expression level of the lincRNA (X axis, molecule count) and that of

its neighboring coding gene (Y axis). Pearson correlation coefficients (ρ) after removal of outliers (**3.4 Methods**) are denoted on top. Data in **(a-c)** is presented for 8 of the 9 lincRNA-gene neighbor pairs for which a valid probe set and fluorescence signal is available.

3.3 Discussion

In this study we have applied single molecule RNA-FISH to quantitatively characterize the expression and localization of 34 lncRNAs, chosen to span diverse characteristics, in individual cells in three human cell types (overall, 70 gene-cell pairs). Our systematic analysis optimized new experimental and computational techniques for the unique challenges posed by lncRNAs, addressed several open questions in lncRNA biology, and shed new light on their putative function.

While single molecule RNA-FISH has a high success rate when applied to mRNAs [21, 22], our screen highlighted specific challenges in the context of lncRNAs. **First**, lncRNAs are shorter than mRNAs and have a higher repeat content: ~75% of our lncRNA candidate set contains a repeat element *vs*. only 1 of the mRNAs in this study [37]. Since probes targeting repetitive sequences are associated with a higher background, this greatly constrained our probe design, typically resulting in smaller probe sets. Thus, while we originally attempted to design probes for 87 lncRNAs, only 61 yielded designs with greater than 10 oligonucleotides. **Second**, lncRNAs appear hybridize to off-targets at a higher frequency than mRNAs, likely due to smaller probe sets, lower sequence specificity of individual oligonucleotides, and higher abundance of off target hybridization sites relative to the target. Our two-color co-localization step was essential to control for these, by revealing quantitative and qualitative inconsistencies for 20 lncRNAs (~33% of lncRNAs *vs*. 1 of 11 mRNAs (9%) that had no signal in this screen). **Third**, unlike mRNAs, legitimate lncRNA signals can appear as bright nuclear foci, whereas such signals are rarely if ever observed for mRNAs, where foci are readily discarded as false positives. Yet, such type of false positives are common for lncRNAs but cannot be automatically dismissed. For example, there are 8 cases of large foci in HeLa appearing in only one channel (*e.g.*, Figure 3.1a) that might represent a highly expressed and localized off target (presumably of repetitive sequence or perhaps the transcription site of another gene). **Finally**, it is essential to separately validate probes in every cell type under investigation, because a probe may show discrepancies only in some cell types (**Supplementary Figure 2.4**). Such confounding probes can be detected experimentally and specific oligonucleotides can be eliminated as we applied for XLOC_010514

(**3.4 Methods**; **Supplementary Figure 2.2a**). However, we did not apply this detection and elimination process to all invalid probe sets due to scalability constrains.

LncRNAs have diverse sub-cellular localization patterns, with a bias toward primary localization to the nucleus (55%), followed with localization both to the cytoplasm and the nucleus (40%), whereas only few lncRNAs were predominantly cytoplasmic. In contrast, mRNAs were primarily localized to the cytoplasm, with only a minority of the transcripts of any mRNA species (on average 23 %) localized in the nucleus (equivalent to our classification to type V which includes ~ 5% of our set). Overall, fully 95% of our lncRNAs display a nuclear bias, as compared to mRNA. This bias toward nuclear localization is consistent with previous analysis that relied on relative RNA-Seq measurements of subcellular fractions [31, 57], but suggests more conclusive evidence as it relies on absolute measurements (see [1] for discussion).

We observe bright nuclear foci for ~30% of the lncRNA we examined, but no bright cytoplasmic foci for any of them. These foci may be consistent with a role for these lncRNAs in chromatin regulation [4], such as previously demonstrated for XIST [8], KCNQOT1[44], AIR [58] and other lncRNA involved in imprinting [59]. The absence of nuclear foci in mitotic cells for five of six lncRNAs that present such foci during interphase strongly suggests that a putative association with specific regulatory regions is likely not a mechanism of mitotic inheritance in these cases. We distinguish two classes of bright nuclear foci. Tight spatially localized foci are in a diffraction limit spot, as for single RNA molecules, such as ANRIL and XLOC_006922. These are likely localized to the transcription site itself, potentially during transcriptional bursts [60]. Conversely, foci with spatially extended localization, such as MALAT1 and XIST, could either represent spreading of the lncRNA over a broad area of chromatin (*e.g.*, XIST) or localization to a particular nuclear region, such as speckles for MALAT1. Of note, we also observed distinct non-random localization pattern throughout the nucleus for MEG3 and XLOC_003526 reminiscent of MALAT1 (**Figure 3.2**) and possibly consistent with localization to a specific nuclear body [61].

Our single cell analysis suggests that – at least for our set – the low abundance of lncRNAs in bulk cell population is most likely not a result of high expression in a small subset of 'jackpot' cells as previously hypothesized [52, 53]. Overall, the extent of cell-to-cell variability of lncRNAs resembled that of mRNA expressed at similar levels. Although in some cases the number of imaged cells is low, we nevertheless observe a relatively homogenous expression of few molecules per cell, and the average expression from RNA-FISH and (bulk) RNA-Seq are fairly correlated (**Supplementary Figure 2.8**). Furthermore, this conclusion is strongly supported when we analyzed hundreds of cells for a few representative lncRNAs. We cannot rule out the possibility that a few very rare "jackpot" cells exist (which could be discovered by even deeper imaging studies), but those are not necessary to explain the average expression in bulk assays.

There are a few important exceptions to this general rule. First, a striking exception is XLOC_003562, a lncRNA that is very lowly expressed based on bulk estimates (<1 FPKM), but is expressed at ~110 molecules per cell in ~25% of the cells in our analysis, and has intriguing non-random localization pattern reminiscent of MEG3 and MALAT1 (**Figure 3.4d**). Second, two rare daughter cells, which probably result from asymmetric division of HeLa cells, express NR_029435 very highly and uniquely (**Supplementary Figure 2.16**). Finally, three of the lncRNA in our set may oscillate with the cell cycle, based on correlation analysis with CCNA2, one of which (XLOC_001185) possibly down regulated in proliferating cells. Thus, some lncRNAs show more substantial cell-to-cell variability. Nevertheless, this is within the range of phenomena also observed for mRNAs, and the frequency of positive cells is not particularly low.

While the vast of divergent transcription results with short unstable transcripts [39-41, 54, 55] , we and others reported over 500 lincRNAs that are transcribed divergently to protein coding genes [30, 33, 56]. It was recently suggested that these more stable transcripts evolve by "strengthening the U1-PAS axis", such that a gain in U1 binding sites coupled with loss of poly-adenylation(A) sites (PAS) promotes transcription of elongated and spliced transcripts upstream of promoters through evolution [55].

We selected 16 pairs of divergently transcribed lncRNAs to analyze the relationship between their localization or expression levels to that of their neighboring gene mRNA. We detected a valid signal for 9 of them (a success rate similar to other types of lncRNAs in this study) but this did not significantly correlate with properties associated with higher stability (transcript length, number of exons or maximal expression). Most divergently transcribed lncRNAs correlated in their expression with their neighbor across individual cells, with varying correlations possibly owing to different stability of the lncRNA and the mRNA. Our divergent lncRNAs did not share a common localization pattern. Two of the divergently transcribed lncRNAs (GAS5 and lincFOXF1) were highly expressed (~120-200 molecule per cell on average), and two were substantially elevated compared to their neighbor (GAS5 and NR_029435). Only 3/8 divergently transcribed lncRNAs had one or few strong foci, a pattern expected if the transcribed lncRNA is directly or indirectly involved in transcription regulation around its site of transcription. One of these 3 linRNAs is lncMKLN1 (PINT), a potential tumor suppressor lncRNA, regulated by P53, whose mouse syntenic ortholog was previously shown to bind PRC2 [62]. The vast majority of the examined cases (7/8) spread-out through the nucleus and also appeared in the cytoplasm. Overall, while sharing a genomic layout, divergent lncRNAs in this study did not share a common localization and expression pattern.

In summary, our study highlights important differences and similarities between lncRNAs and mRNAs, and provides a workflow for applying single molecule FISH to study lncRNA. The rich set of localization patterns we observe suggest a broad range of potential functions for lncRNA and highlights specific lncRNAs for future hypothesis driven mechanistic studies.

3.4 Methods

Selection of candidate lncRNA set

We manually selected the lncRNAs candidate set from our human lincRNA catalog [30] to represent different features: (**1**) tissue specificity, (**2**) evolutionary conservation by an expressed mouse syntenic ortholog, and (**3**) divergent transcription from coding genes' promoters. We selected the set such that it can be divided to a positive and a negative subset with respect to each trait so we can study teach feature with respect to other features (*e.g.*, expression level, nuclear localization, etc.). These subsets were not mutually exclusive across these three features (**Figure 3.1b**). We also ensured that the set includes transcripts expressed in at least one of the three cell types we studied (hLF, hFF or HeLa), and represent a wide range of expression levels (as we did not have an estimate for the minimal expression level required for detection by RNA-FISH).

Features of each lncRNA were extracted from our human lincRNA catalog and were evaluated as previously described [30]. Briefly, to evaluate tissue specificity we use a tissue specificity score (ranging from 0 to 1) that is calculated by an entropy based metric. We consider transcripts with a score greater than 0.4 as tissue specific. Divergent lncRNAs are those that are bi-directionally transcribed within 10 KB from a protein-coding gene promoter (illustrated in **Figure 3.1a**). Orthologous lncRNAs in this study are those for which there is an annotated expressed transcript in their syntenically mapped genomic region in mouse (mapped by the TransMap algorithm [63]).

The selection process was performed by the following steps. **(1)** we selected only the lincRNAs that were is the stringent set and were significantly expressed in foreskin or lung fibroblasts. We determined significant expression based on a scan statistic [30, 38]. **(2)** We divided the set from (1) to two subsets according to the presence of a mouse syntenic ortholog. **(3)** We sorted each of the two subsets by *(a)* expression level, *(b)* tissue specificity, *(c)* transcript length, *(d)* the presence of a chromatin signature of actively transcribed genes [30, 38], and *(e)* being a divergent transcript. **(3)** We then screened the sorted subsets top-down manually selecting those lincRNAs that had an isoform that was clearly supported by RNA-Seq spliced read based on visual inspection of the RNA-Seq read alignments and the reconstructed transcripts. **(4)** We partitioned these candidates to three expression bins and balanced the subsets sizes by eliminating transcripts from the top two bins, giving preference to divergent transcripts, tissue specific transcripts, and then to longer transcripts that have a chromatin signature of actively transcribed genes and a low number of isoforms. **(5)** We repeated steps 1-4 using the transcripts expressed in HeLa cells. All selection criteria were based on previously estimated features that are available through the human lincRNA catalog table ([30] ; http://www.broadinstitute.org/genome_bio/human_lincrnas/). Finally, we included 16 previously studied lncRNAs curated from the literature (**Supplementary dataset 1.3; Supplementary Table 2.2**).

Selection of control mRNA sets

We selected two sets of mRNA controls (**Supplementary dataset 1.4-1.5)**. Set 1 is comprised of mRNAs transcribed divergently to our 'divergent lncRNAs' and CCNA2, a cyclin simultaneously imaged in all images. Set two consists of mRNAs selected to span a wide range of expression levels in human foreskin fibroblasts (Padovan and Raj, personal communication).

Design and synthesis of RNA FISH probe sets

We designed oligonucleotides sets using software available through Stellaris Probe Designer (http://www.biosearchtech.com/stellarisdesigner/). Since the software avoids sequence elements deemed to cause high levels of background, it can sometimes result in only a limited number of potential oligonucleotides targeting a particular RNA. As a conservative choice, we only included in the actual screen those lncRNAs for which we had at least 10 designed oligonucleotides. **Supplementary dataset 1.6** contains all the oligonucleotide sequences used in this study.

We ordered all Stellaris™-type oligonucleotides from Biosearch Technologies, but instead of a dye on the 3' end of the oligonucleotide, we ordered oligonucleotides with an amine group on the 3' end, to which we coupled either Alexa Fluor 594 (Life Technologies), Cy3 (GE Healthcare) or Atto 647N (Atto-Tec). After coupling, we removed the unlabeled oligonucleotides via HPLC purification. For the data using full probe sets, we labeled the lncRNA oligonucleotides with Alexa Fluor 594, the coding neighbor mRNA oligonucleotides (when applicable) with Cy3, and Cyclin A2 mRNA oligonucleotides with Atto 647N. When validating the lncRNA oligonucleotides via co-localization, we labeled the even numbered oligonucleotides in Alexa Fluor 594 and the odd numbered oligonucleotides with Cy3.

Cell culture and RNA FISH

We cultured human foreskin fibroblasts (CRL-2097, ATCC), human lung fibroblasts (IMR-90, ATCC), and HeLa cells (gift from the lab of Phillip Sharp, MIT) in Dulbecco's modified Eagle's medium with Glutamax (DMEM, Life Technologies), supplemented with 10% fetal bovine serum, Penicillin and Streptomycin. We grew the cells in 2-well chambered coverglass (Lab Tek). We washed cells with 1x phosphate buffered saline (PBS) and then fixed them in 3.7% formaldehyde in 1x PBS for 10 minutes at room temperature. After fixation, we washed the cells twice with 1x PBS and then permeabilized them in 70% ethanol at 4°C at least overnight or until we performed FISH staining.

We performed RNA FISH staining as previously described [22, 64]. Briefly, we washed cells with a solution of 10% formamide in 2x sodium citrate buffer (SSC), then applied the appropriate amount of probe in a hybridization solution containing 10% formamide, 2x SSC, and 10% dextran sulfate (w/v). Hybridization was allowed to occur overnight in a humid chamber at 37°C. Cells were then washed twice for 30 minutes at 37°C with 10% formamide in 2x SSC. DAPI was applied during the second wash. Cells were then rinsed twice with 2x SSC before imaging.

Imaging

After performing RNA FISH, we imaged the cells on a Nikon Ti-E inverted fluorescence microscope using a Plan Apochromat 100x objective and a cooled CCD camera. We acquired around 25-30 optical slices at 0.3µm intervals, thereby covering the entire vertical extent of the cell. As described previously, we used bandpass filters specifically for these channels that have essentially no signal crossover [51], and acquired successive image stacks for DAPI (nuclear stain), each fluorescence channel targeted with an RNA FISH probe. We also acquired images in a fluorescence channel with a 488nm excitation (similar to fluorescein/Alexa 488); this channel has no probe in it, and thus reveals the degree of autofluorescent background in the sample.

Image analysis

Image analysis was performed using custom software written in Matlab (The Mathworks, Natick, MA) as previously described [22]. Briefly, images were first manually segmented to define cellular boundaries by using a custom user interface. Images were then processed with a linear filter akin to a Laplacian-of-Gaussian to remove non-uniform background and to enhance particulate signals. RNA particles in each channel were then identified in a semi-automated manner by selecting an intensity threshold above which a spot is considered an RNA particle. Specifically, the threshold was computationally estimated (and then manually confirmed or adjusted) by identifying a plateau in the graph comparing the intensity threshold (X axis) and total particles above that threshold (Y axis; **Supplementary Figure 2.16**). We then determined each spot's intensity by fitting a two-dimensional Gaussian to the spot signal and obtaining amplitude. Finally, we determined which spots co-localize across channels following the methods outlined in Levesque et al. [65] in a two stage process: first, we find spots that colocalize within a relatively large spatial window, then we use those colocalized spots to register the two images (correcting for any shifts between channels) and run the colocalization again, but this time with a smaller window. We ignored spots that co-localized with spots identified in the GFP channel (which represent auto-fluorescent background). Details regarding subsequent analysis steps are described in the following sections.

Single molecule count correction

We applied the following heuristic to estimate the number of molecules within a large focus. For each lncRNA, we divided the total integrated signal in a single detectable spot (focus) by the median signal of the spots within the same cell. The final spot count x_i for a spot p_i was $x_i = max(1, floor(s_i/c_i))$, where s_i is the total intensity of spot p_i, and c_i is the median intensity of across all the spots within the same cell as p_i. c_i was used to estimate the signal intensity from hybridization to a single molecule. Application of this correction affected on average 5% +/- 0.5% STDV of the originally-detected spots, causing on average an increase of 12% +/- 1.7% in the mean spot counts (**Supplementary Figure 2.17**). Following this calculation, the spot count estimations of XIST, MALAT1, and NEAT1 were outliers in that the corrected counts resulted in over an order of magnitude higher estimate than the other cases. Moreover, shorter exposure time was required in these three cases in order to obtain a clear image, as molecules in these cases are clumped together posing a challenge to resolve single spots. Therefore, we consider the spot count estimation of XIST, MALAT1, and NEAT1 as not quantitative and eliminate these from the quantitative parts of the analysis. Five other lncRNAs (Anril, GAS5, NR_029435, lincFOXF1, and TERC) included outlier cells (<4 cells per lncRNA) for which the fitting algorithm yielded an implausibly large value for a specific spot, as verified by eye (resulting in a total spot count a 15 IQR (inter quartile range) higher than the median count across all imaged cells for that gene). These specific cells were eliminated from the analysis (visual inspection confirmed these were not "jackpot cells"; **Supplementary dataset 2.4**).

Identification of off-target hybridization of an XLOC_010514 probe to MALAT1

To identify candidate oligonucleotides of the XLOC_010514 lncRNA that potentially hybridize to other RNAs in the cell we aligned the sequences against the Refseq [66] transcriptome using BLAST [67] (with default parameters for short read alignments; word size =7, match score =1 mismatch score = -3). We then ranked the next best hits based on their expression levels by RNA-Seq. Elimination of oligonucleotide #5 that is predicted to have 15 exact matches with MALAT1 and reimaging resulted with elimination of the localization pattern that was similar to MALAT1.

Validation of probe sets by two-color co-localization

To validate each probe set we used a two-color co-localization approach similar to that previously described [21, 22]. Briefly, we partitioned each probe set to the even- and odd-numbered oligonucleotides and coupled each subset with a different fluorophore (evens with Alexa594, odds with Cy3). We then hybridized the two probe sets and imaged each color.

To determine the total number of RNA particles above background signal in each color we pursued the following procedure. First, we determined the total number of particles imaged in each cell using the full probe set coupled to Alexa 594 (termed the 'single-colored probe set'), using the previously described, semi-automated procedure [22] employed in Image Analysis, above (**Supplementary Figure 2.16**). We also estimated the distribution of particle counts for the single-colored probe set and its mean m_i. Next, for every cell in the two-color co-localization dataset we selected the x_i particles with the highest signal for each of the even-numbered and odd-numbered probe subsets, where $x_i = max\ (50, 5*m_i)$. We then calculated the number of co-localized spots among these x_i spots from each color in every cell. Finally, we determined the distribution of the number of co-localized spots for each probe set across cells. We only consider the co-localized spots as representing a true mRNA particle in each channel when we analyze images acquired in the two-color assay.

We applied this analysis to every probe set in each of the three cell types (HeLa, hLF, hFF) in which it displayed a signal. A probe set was considered invalid in a specific cell type if there was either (**Figure 3.1a**, **Supplementary Figure 2.2d**): (**1**) a *qualitative difference* between the localization pattern obtained using one color channel *vs.* the other; or (**2**) a *quantitative difference* defined as a statistically significant difference in the distribution of the number of co-localized particles and the single-color probe set particles ($P < 0.05$, Mann-Whitney U ranksum test). The remaining cell-probe set pairs were considered valid and images acquired with the full-single-colored probe set were used for all subsequent analyses. Manual examination recovered 14 additional borderline cases in which the clear pattern seen in one cell type was similar to that in a different cell type for which the two color and single color assays were consistent. The specific classifications and distribution comparisons are specified in **Supplementary dataset 2.1, Supplementary Table 2.3,** and **Supplementary Figure 2.18**.

For many of the two-color experiments it was impossible to robustly determine the total number of mRNA particles in each channel using the plateau method [22] used for the single-colored probe set (**Supplementary Figure 2.16b**). This is likely due to the smaller number of oligonucleotides that actually hybridize to the target when using only half the probe set, resulting in a lower contrast between the real signal and background [22] . The approach we used above to evaluate the number of co-localized spots does not rely on the plateau method and is not sensitive to the selection of an intensity threshold.

Localization to the nucleus

Nuclear localization of a spot was heuristically determined based on co-localization with DAPI after considering the maximal signal across all z-stacks. We determined nuclear localization by two approaches that yielded similar results: (**1**) the percent of spots across the entire cell population localized to the nucleus ('molecule level'); or (**2**) the percent of cells in which more than 50% of the spots were localized to the nucleus ('cell level'). Classification of a gene as predominantly nuclear was estimated based on the 'cell level' approach by calculating the fraction of nuclear spots for each cell, and then taking the median across this distribution.

Each lncRNA:cell-type pair was assigned to one of the following classes: (I) 1-2 large foci, (II) both large foci and single molecules scattered through the nucleus, (III) predominantly nuclear (without foci), (VI) cytoplasmic and nuclear, and (V) predominantly cytoplasmic.

Assignment was performed with the following steps: **(1)** For each lncRNA-cell pair we calculated the fraction of nuclear spots for each cell, and then determined the median of that distribution. **(2)** LncRNA-cell pairs with a median fraction of nuclear spots > 0.65 were then manually assigned to classes I, II, or III, by manual inspection of the images and visual recognition of large foci. **(3)** LncRNA-cell pairs with a median fraction of nuclear spots < 0.35 and an average spot count > 20 were classified as V. The selection of a spot count threshold was made in order to be conservative when classifying to V. **(4)** All other cases were classified as IV. **(5)** Finally, we reassigned two borderline cases to IV (lincFOXf1-hFF and XLOC_011950-hFF, median nuclear fraction of 0.67, 0.35 respectively), since we were unable to manually identify specific cells that support a predominant localization to either compartment. Assignments to localization patterns are specified at **Supplementary dataset 3.2**.

Estimation of population expression abundance based on RNA-Seq

We estimated the expression abundance of all lncRNAs and protein-coding genes by running Cuffdiff2 (non-diff mode) [68] across a set of HeLa, hLF and hFF samples as well as a second set that included these samples in addition to a previously published RNA-Seq human tissue compendium ([30, 37]; **Supplementary Dataset 4**). RNA-Seq libraries were aligned to Hg19 using Tophat [69]. We used our entire human noncoding transcripts catalog [30] complemented with additional lncRNAs added in this screen and all coding transcripts annotated in the UCSC Browser [70] for a comprehensive representation of transcripts along the genome, while performing abundance estimation (**Supplementary Dataset 4**). FPKM calls were $\log_2$-transformed (after addition of $\varepsilon = 0.05$). Expression matrices and other information related to the RNA-Seq analysis are provided in **Supplementary dataset 4.**

Identification of lncRNAs with extreme heterogeneity

XLOC_003526 was first detected as having a subpopulation of highly expressing cells based on visual inspection of the data. We have attempted to identify other such cases systematically by two other approaches: (**1**) detecting outliers in the regression line fitting the samples' mean and median. This detects only XLOC_003526. (**2**) Fitting a Poisson to the mean *vs*. CV plot and detecting outliers with a mean value outside the confidence intervals around the estimated mean. None of the candidates identified by the latter approach appeared to present a distinguished subset of expressing cells based on visual examination. Other attempts to fit a Gaussian or a mixture of Gaussians to model bimodality in the data failed due to small sample size.

Selection of lncRNAs for a large sample size imaging survey

We examined the lncRNAs that have a low average molecule count relative to their estimated expression based on RNA-Seq (by looking at off-diagonal data points in **Supplementary Figure 2.8**) in HeLa cells (to enable higher throughput). We selected XLOC_L2_008203, XLOC_008174 and XLOC_004456, as they matched these criteria. We also included NR_029435 since we initially observed one rare case of a highly expressing cell and wanted to further estimate the frequency such case occurs.

Rare cell power calculation

We performed the following power calculation to determine the probability of not observing eve one 'jackpot' cell (defined informally as *max (100, 10*IQR(v)+mean(v))* , where *v* is the count vector and *IQR* is the inner quartile range) among *n* samples cells. Let *p* be the probability of a jackpot cell. Then the probability of not finding any jackpot cell among *n* sampled cell is $(1\text{-}p)^n$. Therefore, the probability of finding at least one jackpot cell is $1\text{-}(1\text{-}p)^{\mathrm{n}}$. In our studies, we measured at least n=500 cells without observing a jackpot cell. Setting our desired statistical power to 0.95, we achieve $p < 0.006$.

Correlation with cell cycle phase

To determine whether the expression of any of the lncRNAs in our study is correlated with phases of the cell cycle, we simultaneously measured in every image we acquired the cyclin CCNA2, which is exclusively expressed in the S, G2 and M phases of the cell cycle. We then applied several approaches to determine if a lncRNA's expression is cell cycle dependent. (**1**) For every lncRNA – cell type pair we calculated the Pearson correlation coefficient (r) between CCNA2 and the lncRNA molecule counts. We consider $|r|>0.4$ as indication of cell cycle associated expression. For those associated cases, we estimate if the observed correlation was dependent on cell volume by fitting a linear regression model to predict the lncRNA molecule counts by using the cell area and CCNA2 levels as predictors (using the Matlab function LinearModel.fit). We then evaluated if CCNA2 was a significant predictor ($P<0.05$). (**2**) We estimated a threshold on CCNA2 levels for each cell type that distinguished cells in G1 from all other phases. We derived this threshold by plotting the distributions of levels of CCNA2 in all the cells we imaged from a specific cell type. The resulting distributions are bimodal, allowing us to determine a threshold t=0 for HeLa cells and t=20 for fibroblasts. For each lncRNA in every cell type we then split the population of cells by these thresholds and compared differences in the distributions of molecule counts using a KS test. Finally, we followed by a visual examination of the candidates that were significant based on these two criteria as a final conformation step.

Divergent neighbor correlation analysis

The presented Pearson correlation values between lncRNA-mRNA divergent pairs were calculated after removal of outliers. Outliers were defined as those that are three standard deviations over the mean of any one of the two variables. Correlation values before and after the removal of outliers are specified in **Supplementary Dataset 3.5**.

Catalog access

Our linc-FISH catalog can be accessed at http://www.broadinstitute.org/genome_bio/human_lincrnas/ (select lincRNA-FISH catalog on the left menu). All **Supplementary Datasets** as well as raw image data can be downloaded from the website. Individual images can be viewed through an image database linked to the website. (please login with the following credentials : username-reviewers password –p@55w0rd).

3.5 References

1. Ulitsky, I. and D.P. Bartel, *lincRNAs: genomics, evolution, and mechanisms.* Cell, 2013. **154**(1): p. 26-46.
2. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annual review of biochemistry, 2012. **81**: p. 145-66.
3. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nature reviews. Genetics, 2009. **10**(3): p. 155-9.
4. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.
5. Hutchinson, J.N., et al., *A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains.* BMC genomics, 2007. **8**: p. 39.
6. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Science signaling, 2010. **3**(107): p. ra8.
7. Hacisuleyman, E., et al., *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.* Nature Structural & Molecular Biology, 2014. **21**(2): p. 198-206.
8. Clemson, C.M., et al., *XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure.* The Journal of cell biology, 1996. **132**(3): p. 259-75.
9. Maamar, H., et al., *linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis.* Genes & development, 2013. **27**(11): p. 1260-71.
10. Tian, D., S. Sun, and J.T. Lee, *The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation.* Cell, 2010. **143**(3): p. 390-403.
11. 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, 1992. **71**(3): p. 527-42.
12. Singer, R.H. and D.C. Ward, *Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog.* Proceedings of the National Academy of Sciences of the United States of America, 1982. **79**(23): p. 7331-5.
13. Harrison, P.R., et al., *Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA.* FEBS letters, 1973. **32**(1): p. 109-12.
14. Brown, C.J., et al., *Localization of the X inactivation centre on the human X chromosome in Xq13.* Nature, 1991. **349**(6304): p. 82-4.
15. Brockdorff, N., et al., *The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus.* Cell, 1992. **71**(3): p. 515-26.
16. Froberg, J.E., L. Yang, and J.T. Lee, *Guided by RNAs: X-inactivation as a model for lncRNA function.* Journal of Molecular Biology, 2013. **425**(19): p. 3698-706.
17. Ip, J.Y. and S. Nakagawa, *Long non-coding RNAs in nuclear bodies.* Development, Growth & Differentiation, 2012. **54**(1): p. 44-54.
18. Sone, M., et al., *The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons.* Journal of cell science, 2007. **120**(Pt 15): p. 2498-506.
19. Clemson, C.M., et al., *An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.* Molecular Cell, 2009. **33**(6): p. 717-26.
20. Sasaki, Y.T., et al., *MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(8): p. 2525-30.
21. Femino, A.M., et al., *Visualization of single RNA transcripts in situ.* Science, 1998. **280**(5363): p. 585-90.
22. Raj, A., et al., *Imaging individual mRNA molecules using multiple singly labeled probes.* Nature methods, 2008. **5**(10): p. 877-9.
23. 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, 2009. **106**(28): p. 11667-72.
24. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.* Nature, 2011. **472**(7341): p. 120-4.
25. Bumgarner, S.L., et al., *Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment.* Molecular Cell, 2012. **45**(4): p. 470-82.
26. Mohammad, F., et al., *Kcnq1ot1/Lit1 noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region.* Molecular and Cellular Biology, 2008. **28**(11): p. 3713-28.

27. Carpenter, S., et al., *A long noncoding RNA mediates both activation and repression of immune response genes.* Science, 2013. **341**(6147): p. 789-92.
28. Kretz, M., et al., *Control of somatic tissue differentiation by the long non-coding RNA TINCR.* Nature, 2013. **493**(7431): p. 231-5.
29. Mercer, T.R., et al., *Specific expression of long noncoding RNAs in the mouse brain.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 716-21.
30. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & development, 2011. **25**(18): p. 1915-27.
31. Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Research, 2012. **22**(9): p. 1775-89.
32. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.
33. Sigova, A.A., et al., *Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(8): p. 2876-81.
34. Pauli, A., et al., *Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.* Genome Research, 2012. **22**(3): p. 577-91.
35. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.
36. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.
37. Kelley, D. and J. Rinn, *Transposable elements reveal a stem cell-specific class of long noncoding RNAs.* Genome Biology, 2012. **13**(11): p. R107.
38. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
39. Seila, A.C., et al., *Divergent transcription: a new feature of active promoters.* Cell Cycle, 2009. **8**(16): p. 2557-64.
40. Seila, A.C., et al., *Divergent transcription from active promoters.* Science, 2008. **322**(5909): p. 1849-51.
41. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
42. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Molecular Cell, 2010. **39**(6): p. 925-38.
43. Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.* Molecular Cell, 2010. **38**(5): p. 662-74.
44. Terranova, R., et al., *Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos.* Developmental cell, 2008. **15**(5): p. 668-79.
45. Tsuiji, H., et al., *Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1.* Genes to cells : devoted to molecular & cellular mechanisms, 2011. **16**(5): p. 479-90.
46. Zhong, F., et al., *Disruption of telomerase trafficking by TCAB1 mutation causes dyskeratosis congenita.* Genes & development, 2011. **25**(1): p. 11-6.
47. Zhu, Y., et al., *Telomerase RNA accumulates in Cajal bodies in human cancer cells.* Molecular biology of the cell, 2004. **15**(1): p. 81-90.
48. Yang, L., et al., *ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs.* Cell, 2011. **147**(4): p. 773-88.
49. Tomlinson, R.L., et al., *Telomerase reverse transcriptase is required for the localization of telomerase RNA to cajal bodies and telomeres in human cancer cells.* Molecular biology of the cell, 2008. **19**(9): p. 3793-800.
50. Eward, K.L., et al., *Cyclin mRNA stability does not vary during the cell cycle.* Cell Cycle, 2004. **3**(8): p. 1057-61.
51. Levesque, M.J. and A. Raj, *Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation.* Nature methods, 2013. **10**(3): p. 246-8.
52. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-40.
53. Dinger, M.E., et al., *Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.* Briefings in functional genomics & proteomics, 2009. **8**(6): p. 407-23.

54. Preker, P., et al., *RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.* Science, 2008. **322**(5909): p. 1851-1854.

55. Almada, A.E., et al., *Promoter directionality is controlled by U1 snRNP and polyadenylation signals.* Nature, 2013. **499**(7458): p. 360-3.

56. Lepoivre, C., et al., *Divergent transcription is associated with promoters of transcriptional regulators.* BMC genomics, 2013. **14**: p. 914.

57. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

58. Nagano, T., et al., *The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin.* Science, 2008. **322**(5908): p. 1717-20.

59. Barlow, D.P., *Genomic imprinting: a mammalian epigenetic discovery model.* Annual Review of Genetics, 2011. **45**: p. 379-403.

60. Raj, A., et al., *Stochastic mRNA synthesis in mammalian cells.* PLoS Biology, 2006. **4**(10): p. e309.

61. Mao, Y.S., B. Zhang, and D.L. Spector, *Biogenesis and function of nuclear bodies.* Trends in genetics : TIG, 2011. **27**(8): p. 295-306.

62. Marin-Bejar, O., et al., *Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2.* Genome Biology, 2013. **14**(9): p. R104.

63. Zhu, J., et al., *Comparative Genomics Search for Losses of Long-Established Genes on the Human Lineage.* PLoS Comput Biol, 2007. **3**(12): p. e247.

64. Raj, A., et al., *Variability in gene expression underlies incomplete penetrance.* Nature, 2010. **463**(7283): p. 913-8.

65. Levesque, M.J., et al., *Visualizing SNVs to quantify allele-specific expression in single cells.* Nature methods, 2013. **10**(9): p. 865-7.

66. Pruitt, K., T. Tatusova, and D. Maglott, *Chapter 18, The Reference Sequence (RefSeq) Project.* , in *The NCBI handbook [Internet]*.2002, National Library of Medicine (US), National Center for Biotechnology Information: Bethesda (MD).

67. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-10.

68. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature Biotechnology, 2013. **31**(1): p. 46-53.

69. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

70. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-46.

Chapter 4 | Discussion

4.1 Contributions

Integrative annotation of human lincRNAs reveals global properties and specific subclasses

When I initiated my studies, our understanding of lncRNA biology was very rudimentary. Yet, several groups had described evidence for evolutionary constrains acting on lncRNA sequences, which implied potential function [1-4]. This was ground breaking as it suggested that, similar to the microRNA field, a whole new layer of regulatory processes was about to be discovered. At the time, while evolutionary conservation, as an indicator of function was a guiding theme, we did not have the means to explore how many of these newly discovered mouse lincRNAs actually have an expressed human ortholog and what was the meaning of these partially conserved sequences.

At that time the main annotation groups were curating growing numbers of cDNA clones classified as unknown transcripts in human and mouse [2, 5, 6]. However, these were yet to be assembled into a characterized collection of lncRNAs, and even when the first efforts started, it was unclear in which tissues and context the transcripts were expressed and how comprehensive were these sets. As early evidence suggested that lncRNAs were tissue specific [7-9], a comprehensive catalog detailing where lncRNAs are expressed and what their characteristics are was needed, as it could serve as a guiding map for the emerging community that had a growing interest in lincRNA biology.

The emergence of RNA-Seq and the development of ab initio transcriptome assemblers presented a unique opportunity to generate such a catalog at single nucleotide resolution in a cost effective manner and unprecedented scale. Our goal was to generate, for the first time, a detailed lincRNA catalog describing a large set of properties such as conservation, expression patterns, tissue specificity and chromatin state, to facilitate the generation of testable functional hypotheses. For example, we identified a set of 215 lincRNAs located in gene deserts genetically associated with disease in genome wide association studies (GWAS). This included a thyroid specific lincRNA located in a gene desert associated with thyroid cancer, which was subsequently shown by an independent study to be differentially expressed in thyroid cancer patient samples. As we predicted, this differential expression was influenced by a single nucleotide polymorphism (SNP) in a transcription factor binding site at the promoter [10].

One of the main advantages of our tissue compendium is that it allowed us to conclusively show that lincRNAs are more tissue specific compared to protein coding genes across a broad range of expression levels. This had both theoretical and practical implications. First, it supported the hypothesis that lincRNAs may be involved in processes that determine cell identity. Second, it made it clear that when studying lincRNAs in a new system, one greatly benefits from reconstructing its unique transcriptome from RNA-Seq data as it likely includes important transcripts not documented in available catalogs. Indeed, when using similar approaches tissue specificity was later demonstrated during zebrafish development [11] and across mammals [12, 13]. Interestingly, tissue specificity was shown to be conserved, such that expression of a lincRNA that has an ortholog is likely to be specific to the same tissue in different species [12].

Another key point introduced in our study was that evolutionary conservation was not evaluated based on cross species sequence homology, but rather by looking for an expressed syntenic ortholog. This was important since weak evidence of sequence homology is meaningless with respect to the transcript’s function if an ortholog is not being expressed. We identified ~1,000 human lincRNAs that had an expressed syntenic ortholog, ~700 of which were part of our stringent set. Since at the time of our study a parallel dataset of mouse tissues was not available, we could only provide a lower bound estimate for the number of lincRNAs that have a syntenic expressed ortholog based on publically available catalogs from other species. Surprisingly, our initial estimate of ~13-15% did not significantly increase when we later compared lincRNAs expressed from similar tissues in human and mouse (e.g: human and mouse lung fibroblasts and embryonic stem cells; MNC unpublished data). Moreover, a recent study comparing human lincRNAs to expressed orthologs in 5 different mammals report ~700 syntenic orthologs in the mouse, similar to our initial report [12]. Overall, the moderate but existing conservation of a subset of lincRNA expressing loci is consistent with the view of RNA as a more evolvable molecule, in which structure rather than sequence has a primary effect on function [14].

Our study also highlighted a subset of ~500 divergent promoters that produce stable and long divergent spliced transcripts. These were recently suggested to evolve by a process termed "strengthening of the U1-PAS axis" in which a gain in U1 binding sites coupled with loss of poly-adenylation(A) sites (PAS) through evolution promotes transcription of elongated and spliced transcripts upstream of promoters [15]. Interestingly, we noticed that for 30% of divergent lincRNA loci, the loci was expressing a spliced divergent transcript only in tissues that are different than where divergent-non-spliced transcripts were most abundant. In addition, while overall we observed co-expression of divergent lincRNAs and their neighbors (which was only slightly higher than for divergent coding gene pairs), about 25% of these divergent lincRNAs are tissue specific whereas their coding neighbor is ubiquitously expressed.

While the vast majority of divergent transcripts are co-expressed with their neighbors [16-19], a few studies using RNA-FISH, including ours, highlighted exceptions to this rule [20, 21]. These provide examples by which one can explore a recently proposed model suggesting that divergent transcription can serve as a mean to evolve new genes with a function that is independent of their neighbor [22].

We also distinguished a subset of transcripts of uncertain coding potential (TUCPs), which are transcripts that we could not definitively classify as non-coding, as they included some evidence for conserved open reading frames (ORFs) or protein domains. This set of transcripts might be of particular interest as it includes specific genes with higher sequence conservation, which are more likely to be functional. In a following collaboration, we demonstrated that less than 0.5% of lincRNAs and TUCPs have evidence for potentially encoding small peptides based on mass-spectrometry data that was enriched for small peptides [23]. This suggests that some of the lncRNAs found to be enriched in poly-ribosome fractions might be doing something different than encoding small peptides [24, 25]. Alternatively, some of these transcripts may be 'proto-genes', non-genic sequences that are randomly translated to generate a pool that may result in *de-novo* gene birth over time, a model supported by studies in yeast [26]. Indeed, Pri and more recently Toddler/Elabela are examples of unknown genes with short ORFs that do encode short peptides essential for development [27-30].

Localization and abundance analysis of human lncRNAs at single cell and single molecule resolution

Our first study provided a map that enabled us to understand some of the global properties of lincRNAs. Still, a key step toward understanding function is to track the subcellular localization of these RNAs. Importantly, the low abundance and cell/tissue specificity of lincRNAs raise the possibility that they may only function in a small fraction of high expressing cells amongst the bulk population. Currently, neither single molecule imaging nor single cell based analyses are routinely used to study lncRNAs functions. These would greatly benefit the community and serve as an early guide for mechanistic studies targeting a novel gene.

With this in mind, we have systematically applied single-cell single-molecule RNA-FISH to dozens of lncRNAs and in this process tackled specific challenges in applying this technique to study this class of molecules. One such challenge involves target specificity, since we found in a number of cases that signal resulting from hybridization to highly concentrated off-target appears similar to the patterns we found legitimate for other lncRNA. To control this, we used a two color co-localization step to validate each probe set and introduced a computational approach to assess probe set validity from this data (i.e. comparison of the distribution of co-localized spots rather than the fraction of co-localized spots). Based on our survey of 61 lncRNAs, we show that this strategy is essential for FISH analyses of lincRNAs, since inconsistencies were found for approximately a third of our probe sets and specific types of invalid signals were distinguished.

While we were only able to survey a limited set, our study demonstrates the great benefit of applying RNA-FISH at an early step when studying the function of novel lncRNAs. We demonstrated that lncRNAs have a range of localization patterns and are biased toward the nucleus. To date, small interfering RNA (siRNA) perturbation and RNA affinity purifications are routinely used as initial steps to determine which genes may be targeted by a lincRNA and with which proteins it interacts. However, using RNA-FISH *Maamar et al.* demonstrated that targeting a nuclear lncRNA using siRNA vs. perturbation through locked nucleic acid (LNA; which specifically act in the nucleus) has different phenotypic outcomes [20]. In addition, the interpretation of recently-introduced chromatin localization assays [31-33], will be influenced by the RNA's localization shown in *situ*. For example, single molecule RNA-FISH sowed that the lincRNA Firre is localized in one main foci around its transcription start site, but a chromatin localization assay also indicated it binds *trans* sites [34]. This latter finding guided subsequent RNA-FISH experiments that corroborated a proposed model suggesting that the *trans* sites are in fact proximal to the focus in three dimensions. Thus, together with our study, this highlights the importance to determine whether a lincRNA is predominantly nuclear and what type of localization pattern it has prior to deciding which assays to use.

Our single cell data suggests that in the majority of examined cases, it is unlikely that a rare subset of high expressing cells explain the overall low abundance measured across the population. Yet, we do demonstrate this possibility with XLOC_003526, a lincRNA that has extremely low abundance based on RNA-Seq, but is in fact expressed highly and in a very striking pattern in only ~22% of fibroblast cells. This localization pattern is reminiscent of MALAT1, a lncRNA localized to nuclear speckles and involved in alternative splicing regulation by distributing serine/arginine (SR) splicing factors to transcription sites [35]. The discovery of MALAT1 striking localization pattern very early on was what guided following studies leading to its function [36]. We also demonstrated a similar pattern for MEG3, an imprinted lncRNA with an unknown mechanism of action that was shown to act as a tumor suppressor in several cancers [37]. Our findings suggest the testable hypothesis that MEG3 and XLOC_003526 may localize to specific nuclear bodies and are involved in the processes specific to these compartments (e.g. involved in alternative splicing in similar to MALAT1).

Overall, using systematic RNA-Seq and FISH-based surveys we uncovered key properties of lincRNAs including their tissue specificity, levels of syntenic orthology as well as their diverse cellular localization patterns. We applied new technologies, which were timely to our understanding of novel genes, highlighted challenges when applying them to lncRNAs and proposed strategies to overcome these challenges. Finally, we generated publically available catalogs that provide detailed information for each studied lncRNA. These will benefit the scientific community when generating testable mechanistic hypotheses to elucidate the function of lncRNAs.

4.2 Future directions

Our lncRNA-FISH survey revealed interesting non-random nuclear localization patterns for several novel lincRNAs (e.g. XLOC_005764, MEG3, XLOC_006922, XLOC_005151). These are promising candidates for chromatin localization assays in which one can detect the genomic regions where an RNA is localized by affinity capture of a target RNA from fixed cells and sequencing bound DNA [32-33]. Another interesting candidate for follow-up experiments is the lincRNA XLOC_003526, which is highly expressed in only 22% of lung fibroblast cells. One possible approach to learn more about this gene is to perform single cell RNA-Seq on ~100 human lung fibroblasts followed by differential gene expression analysis between the sub-population that expresses the lincRNA (expected to be ~20%) and the one which does not. One can then use a guilt-by-association strategy to learn which other pathways are specifically expressed in the same sub-population of cells and generate testable hypotheses on possible functions for this lincRNA. In parallel, one can apply simultaneously RNA-FISH and immunoflorescence targeting proteins markers of specific nuclear bodies, to learn if XLOC_003526, or MEG3 localize to specific nuclear compartments similarly to MALAT1 and NEAT1 (as suggested by their localization patterns).

One future avenue that I find particularly interesting is understanding to what extent syntenic orthologous lincRNAs have similar functions. We found that while their position is conserved, syntenic orthologous lincRNAs in human and mouse share moderate sequence identity (lower than coding genes, higher than random regions and similar to randomly chosen syntenic regions). This suggests that while having similar transcriptional regulation across species, some syntenic orthologs, such as those transcribed divergently from a coding gene promoter for example, may have evolved to perform different functions. An important study in zebrafish demonstrated that two syntenically orthologous lincRNAs from human and mouse were able to rescue a developmental phenotype caused by knocking down their zebrafish ortholog [38]. Yet, we are still waiting with great anticipation to learn whether these mammalian orthologs have similar developmental roles in mouse and human.

A possible approach to screen for additional candidate orthologs is to systematically select orthologous lincRNAs that are substantially expressed in similar cell systems in human and mouse (or other species pairs that can be more easily manipulated *in vivo*) and perform single molecule RNA-FISH to learn if these have similar expression and localization patterns. One can then follow with perturbation, RNA affinity purification assays or other approaches to learn the function in one species and then use genetics to learn which sequence is essential for these interactions. Similar experiments in the other species' cell system and a comparison of the affected/interacting genes/proteins can provide a mean to estimate the similarity across species. If these are similar, one should then test if the ortholog from one species can complement the lack of its counterpart in the other species. For example, would a knockout of a *trans* acting mouse lincRNA be rescued by overexpressing its human ortholog. Another approach to gain insight on conserved function is to apply recent tools which combine structure-dependent chemical probing and RNA-Seq to predict the secondary structure of each species' molecule and then evaluate how conserved are these predicted structures [39-41].

Another intriguing avenue is to learn whether divergent lincRNAs have an independent function away from their transcription site. We have tried to approach this systematically by selecting all divergent lincRNAs expressed in foreskin fibroblasts (spliced, long and significantly expressed transcripts) and imaging them simultaneously with their neighbors. While we did find a few that are localized away from their transcription site or that were not correlated with their neighbors, our set was too small to suggest any general properties.

One of the difficulties in studying divergent transcripts is that localization away from the transcription site can only suggest a potential for *trans* action, but does not eliminate the possibility of *cis* activity. In addition, co-expression doesn't imply a regulatory effect between the neighbors. Most importantly, to study the independent function of the lincRNA one would ideally want to perturb the nascent divergent lincRNA transcript at the transcription site and would preferably avoid genetic manipulation of the genomic loci, as this might have an effect on the neighbors' promoter (as these are shared). Such a tool is currently unavailable and currently this question can only be addressed by combining different approaches of genome editing (e.g.

examining outcome after: (1) gene deletion, (2) an insertion of early termination site and (3) replacement of the gene sequence with a reporter).

One possibility to screen for such targets would be to apply a single-cell variation of native elongating transcript (NET)-Seq [42] or global run on (GRO)-Seq [18] that would provide an estimate for how many divergent lincRNAs are not co-expressed with their neighbor in a given cell (as in [20]). Single molecule RNA-FISH can then be used to learn the localization patterns of such divergent lincRNA/coding genes pairs. One can then select specific lincRNA targets that localize to the transcription site and ones that do not, and use perturbation to start understanding their functions.

4.3 Perspective and concluding remark

After the initial discovery of lncRNAs and their recognition by a wider community, the field is now “marching toward mechanism”. The emergence of RNA-Seq and the plethora of assays based on its power enabled the discovery of lncRNAs, the characterization of their global properties and their detection across biological systems. The early attempts to understand mechanism suffered from difficulties in effectively applying available molecular assays on lncRNAs. For example, it is very hard to obtain 80% knockdown of nuclear lncRNAs and while different approaches can be applied to detect protein binding partners it is not always clear what the proper controls are. Yet, systematic studies carried in the past few years provided the community with a better understanding of the strengths and weaknesses of available tools.

The recent and rapid developments of the CRISPR-Cas9 system to edit, regulate and target specific genomic loci seems like a promising avenue to efficiently and effectively target lncRNAs, a much needed tool by our community [43]. This approach is especially appealing as variations of the CRISPR-Cas9 system enables genomic editing as well as perturbing transcription (epigenetically or post-transcriptionally) while leaving the genome intact. This may be adapted in the future to generate an ideal tool that will target a nascent RNA for degradation while tethered to the transcription site. Such a tool will enable the distinction between the lncRNA causing an effect, the missing DNA element or the act of transcription. Still, we should adopt this new tool with caution, as off-target effects caused by this system are not fully appreciated at this early time point and strategies to improve its specificity are still evolving [43]. It would also be important to systematically understand if there are any additional challenges specific to particular types of lncRNA genes (due to higher prevalence of transposable element sequence, for example [44]).

In the near future, our field will mostly benefit from detailed mechanistic understanding of specific examples by combining classic genetics and biochemistry with new emerging tools. The current literature suggests that lncRNAs have different mechanisms of action and do not only act on chromatin [45]. Yet, XIST remains one of the only well characterized models, which acts on chromatin in *cis*. Moreover, applying RNA-FISH as an integral part of every study would be key to understanding the function of unknown lncRNAs. As it’s been done routinely in studies of XIST, MALAT1 and NEAT1, this should be integrated early in studies to understand where the lncRNA is localized and how consistent it is across cells and after perturbations. Another general approach that will greatly advance our understanding of lincRNA mechanisms would be to gain knowledge on RNA secondary or tertiary structures by combining structure-dependent chemical probing and RNA-Seq to better predict the shape of the molecule and which sequence elements may be important for its function [39-41].

Having a better mechanistic understanding of more lincRNAs will provide us with models on which we can test novel high throughput molecular tools. We will then be able to apply those systematically and understand lincRNA characteristics on a class level. Evidently, this is an iterative process and the first round of systematic studies provided a much better perspective going into the next round of mechanistic studies. It has been an exciting time to witness the birth of the lincRNA field and to be part of a growing community that is trying to decipher the function of a new and challenging class of genes. I am looking forward to see all the new discoveries on lincRNAs in the coming years. The stage is now set, lncRNAs are marching toward mechanism.

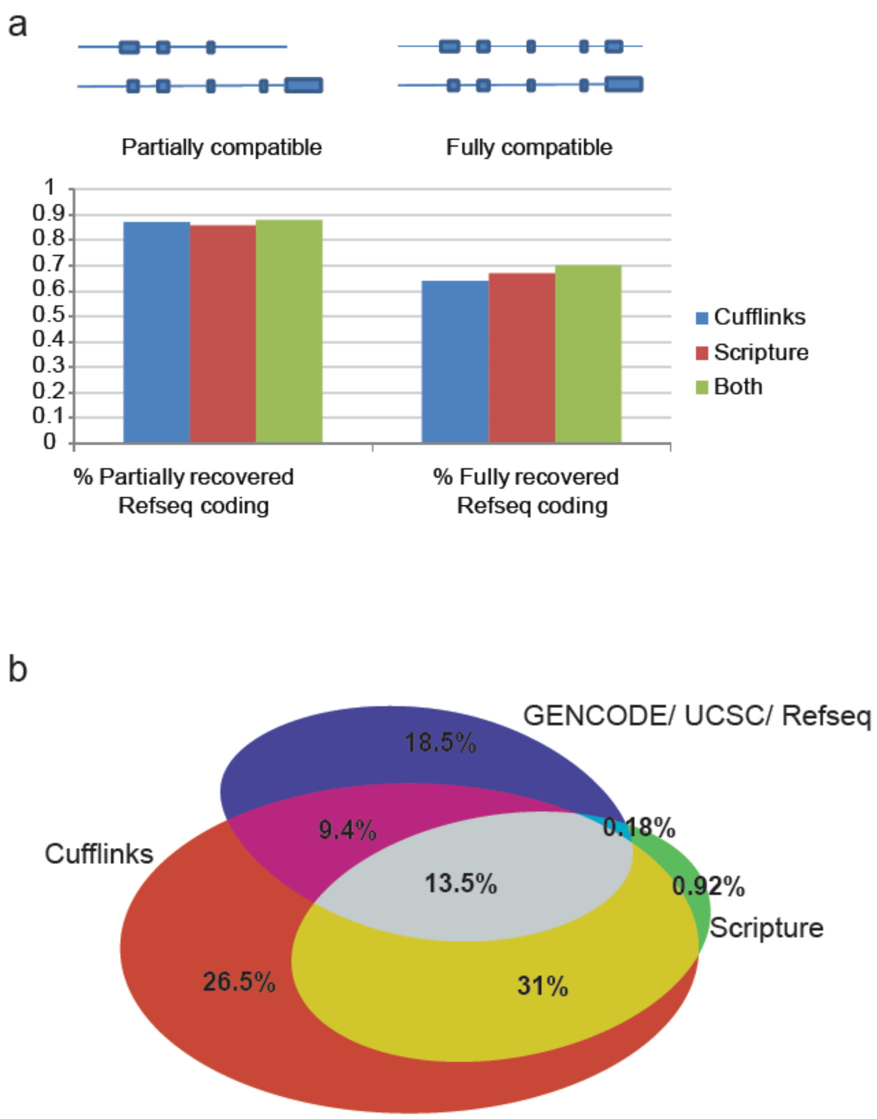
4.4 References

1. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
2. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.
3. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.
4. 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, 2009. **106**(28): p. 11667-72.
5. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
6. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.
7. Dinger, M.E., et al., *Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation.* Genome Res, 2008. **18**(9): p. 1433-45.
8. Mercer, T.R., et al., *Specific expression of long noncoding RNAs in the mouse brain.* Proc Natl Acad Sci U S A, 2008. **105**(2): p. 716-21.
9. Pang, K.C., et al., *Genome-wide identification of long noncoding RNAs in CD8+ T cells.* J Immunol, 2009. **182**(12): p. 7738-48.
10. Jendrzejewski, J., et al., *The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(22): p. 8646-51.
11. Pauli, A., et al., *Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.* Genome Research, 2012. **22**(3): p. 577-91.
12. Washietl, S., M. Kellis, and M. Garber, *Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals.* Genome Research, 2014.
13. Kutter, C., et al., *Rapid turnover of long noncoding RNAs and the evolution of gene expression.* PLoS Genet, 2012. **8**(7): p. e1002841.
14. Ulitsky, I. and D.P. Bartel, *lincRNAs: genomics, evolution, and mechanisms.* Cell, 2013. **154**(1): p. 26-46.
15. Almada, A.E., et al., *Promoter directionality is controlled by U1 snRNP and polyadenylation signals.* Nature, 2013. **499**(7458): p. 360-3.
16. Sigova, A.A., et al., *Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(8): p. 2876-81.
17. Seila, A.C., et al., *Divergent transcription from active promoters.* Science, 2008. **322**(5909): p. 1849-51.
18. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.* Science, 2008. **322**(5909): p. 1845-8.
19. Lepoivre, C., et al., *Divergent transcription is associated with promoters of transcriptional regulators.* BMC genomics, 2013. **14**: p. 914.
20. Maamar, H., et al., *linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription in cis.* Genes & development, 2013. **27**(11): p. 1260-71.
21. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Science signaling, 2010. **3**(107): p. ra8.
22. Wu, X. and P.A. Sharp, *Divergent transcription: a driving force for new gene origination?* Cell, 2013. **155**(5): p. 990-6.
23. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nature chemical biology, 2013. **9**(1): p. 59-64.
24. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
25. Chew, G.L., et al., *Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs.* Development, 2013. **140**(13): p. 2828-34.
26. Carvunis, A.R., et al., *Proto-genes and de novo gene birth.* Nature, 2012. **487**(7407): p. 370-4.
27. Chng, S.C., et al., *ELABELA: a hormone essential for heart development signals via the apelin receptor.* Developmental cell, 2013. **27**(6): p. 672-80.

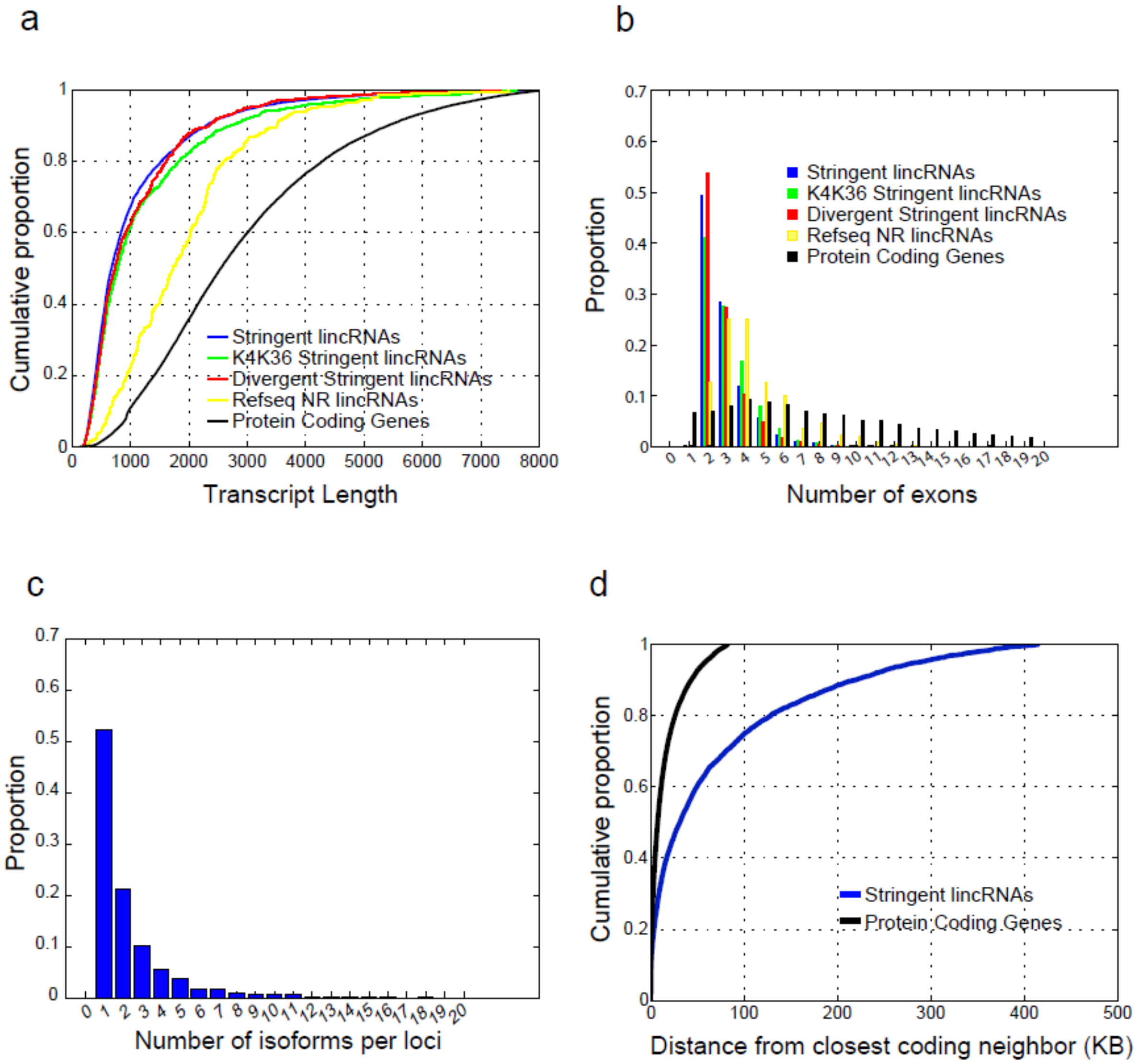
28. Pauli, A., et al., *Toddler: an embryonic signal that promotes cell movement via Apelin receptors.* Science, 2014. **343**(6172): p. 1248636.
29. Kondo, T., et al., *Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA.* Nat Cell Biol, 2007. **9**(6): p. 660-5.
30. Kondo, T., et al., *Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis.* Science, 2010. **329**(5989): p. 336-9.
31. Chu, C., et al., *Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.* Molecular Cell, 2011. **44**(4): p. 667-78.
32. Simon, M.D., et al., *The genomic binding sites of a noncoding RNA.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(51): p. 20497-502.
33. Engreitz, J.M., et al., *The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.* Science, 2013. **341**(6147): p. 1237973.
34. Hacisuleyman, E., et al., *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.* Nature Structural & Molecular Biology, 2014. **21**(2): p. 198-206.
35. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Molecular Cell, 2010. **39**(6): p. 925-38.
36. Hutchinson, J.N., et al., *A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains.* BMC genomics, 2007. **8**: p. 39.
37. Zhou, Y., X. Zhang, and A. Klibanski, *MEG3 noncoding RNA: a tumor suppressor.* Journal of molecular endocrinology, 2012. **48**(3): p. R45-53.
38. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.
39. Mortimer, S.A., et al., *SHAPE-Seq: High-Throughput RNA Structure Analysis.* Current protocols in chemical biology, 2012. **4**(4): p. 275-97.
40. Kertesz, M., et al., *Genome-wide measurement of RNA secondary structure in yeast.* Nature, 2010. **467**(7311): p. 103-7.
41. Weeks, K.M., *Advances in RNA structure analysis by chemical probing.* Current opinion in structural biology, 2010. **20**(3): p. 295-304.
42. Churchman, L.S. and J.S. Weissman, *Nascent transcript sequencing visualizes transcription at nucleotide resolution.* Nature, 2011. **469**(7330): p. 368-373.
43. Sander, J.D. and J.K. Joung, *CRISPR-Cas systems for editing, regulating and targeting genomes.* Nature Biotechnology, 2014.
44. Kelley, D. and J. Rinn, *Transposable elements reveal a stem cell-specific class of long noncoding RNAs.* Genome Biology, 2012. **13**(11): p. R107.
45. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.

Appendix

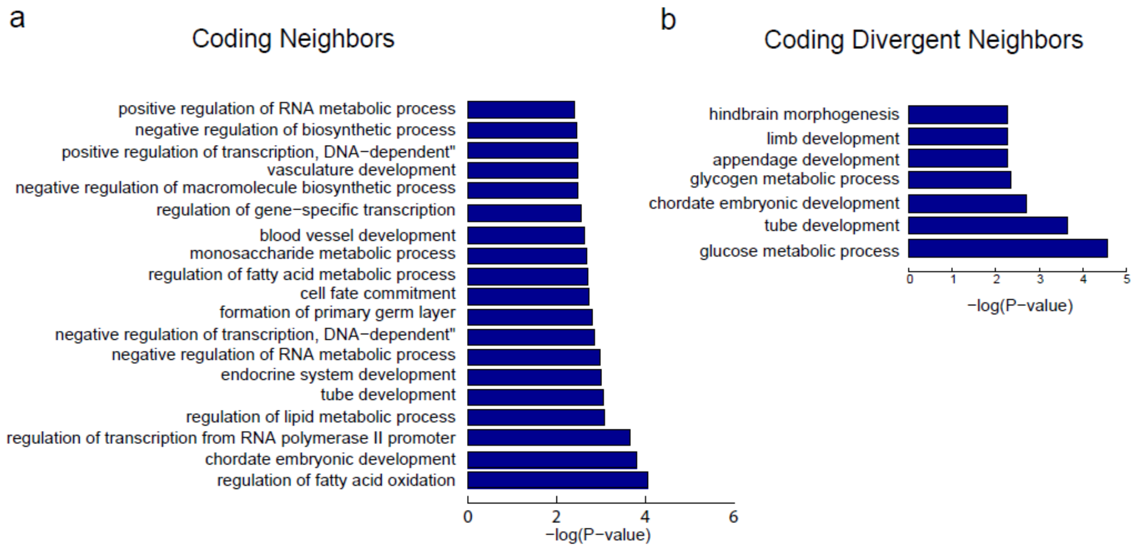
Supplementary Figures and Tables of Chapter 2



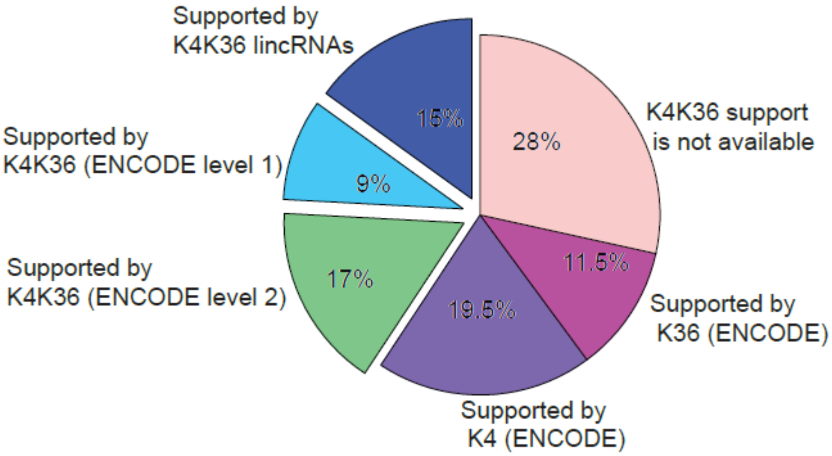
Supplementary Figure 1.1| Evaluation of lincRNA Transcript Structures. (a) Performance of *de-novo* assemblers on protein coding genes. Top**:** Illustration of compatibility between transcript isoforms. In fully compatible transcript (right) both isoforms have the same exact set of introns and at most differ only in the 5' and 3' boundaries of their 5' and 3' exons, respectively. Partially compatible transcripts (left) have a subset (at least 2 sequential exons) of their exon-intron chain that is identical in both transcripts. Bottom: Shown are the fractions of protein coding transcripts annotated in Refseq (24,717 transcripts) that were partially recovered or fully recovered by Cufflinks (blue) Scripture (red) or by the unique union of both assemblers (green), after assembling the 24 tissues and cell-lines used for our analysis. Partial recovery corresponds to all Refseq transcripts for which a transcript with at least 2 correctly connected exons was assembled (87%, 86%, 88% for Cufflinks, Scripture and the union, respectively). Fully recovered corresponds to all Refseq transcripts for which an exact gene structure was assembled (the same exact exon-intron chain without accounting for the exact 5' and 3' boundaries of the 5' and 3' exons, respectively; 64%, 67%, 70% for Cufflinks, Scripture and the union, respectively). (b) A schematic Venn diagram describing the partition of the provisional lincRNA loci with respect to the annotation resources that identified a transcript as a lincRNA.



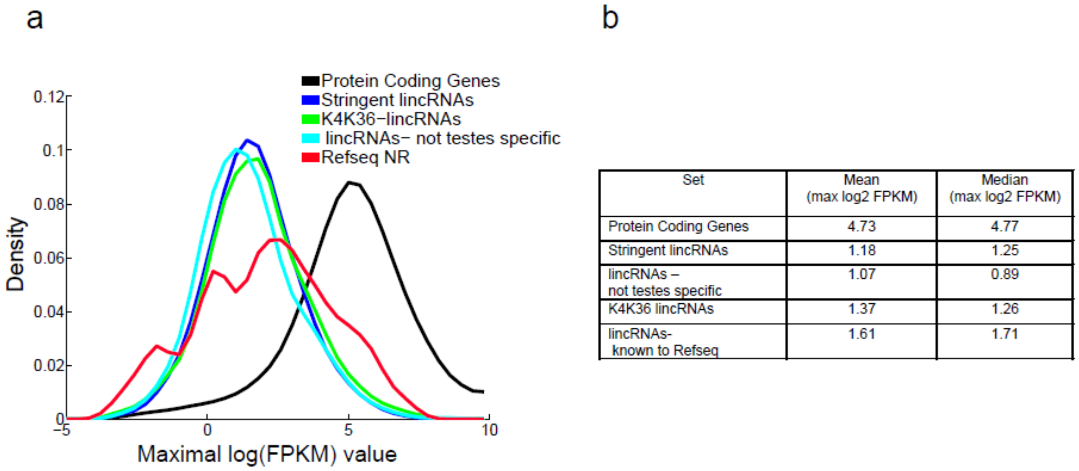
Supplementary Figure 1.2| Structural characteristics of lincRNAs in comparison to protein-coding transcripts. (a) Transcript length. Shown are cumulative distributions of transcript length for the 'conservative lincRNAs' (blue), conservative lincRNAs with K4-K36 support (green), conservative lincRNAs that are divergently expressed (red), lincRNAs annotated in RefSeq NR (yellow) and protein-coding genes (black). Data is shown after eliminating the top 5% of coding genes which are larger than 8KB. To control for the varying number isoforms across a transcript locus, only a single isoform with the maximal number of exons was selected to represent each lincRNA or coding gene locus. **(b)** Analysis of lincRNA Exon number. Shown is the distribution of exon number for different sets of lincRNAs and coding genes, defined and color coded as in (a). Only a single isoform with the maximal exon number was selected in each locus. Data is shown after eliminating the top 15% of coding genes that have more than 20 exons. **(c)** Number of isoforms. Shown is the distribution of the number of isoforms identified per lincRNA locus. **(d)** Distance from neighboring coding locus. Shown is the cumulative distribution of the distance (in kilobase) of a lincRNA from its closest coding neighbor.



Supplementary Figure 1.3| Functional enrichment in neighbors of lincRNAs. Shown are enrichments (-log$_{10}$ P value, Fisher exact test, X axis) of Gene Ontology (GO) [1] terms that are significantly enriched among lincRNA neighboring coding genes that are within 10 kilobase **(a)** or the subset of those that are divergent **(b)**. Only terms that passed a false discovery rate (FDR) correction smaller than 10% are shown (calculated by the David functional annotation tool [2]). Of the 959 neighboring coding genes that were also annotated by GO-FAT (a more focused sub-set of GO [2]), 284 (~28%) were included in at least one of the enriched functional sets. A full list of enriched functional sets is provided in **Supplementary Dataset 4**.

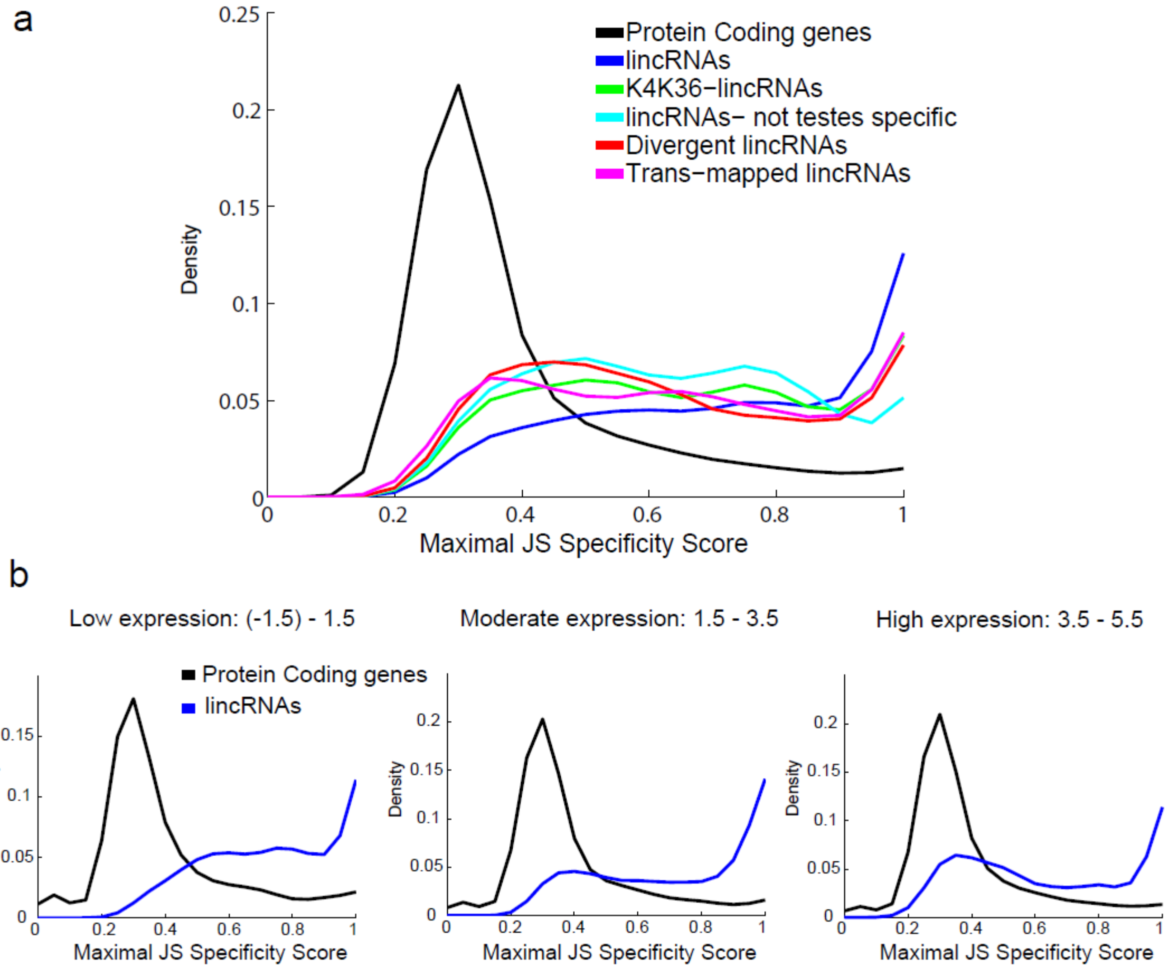


Supplementary Figure 1.4| Evidence for K4-K36 chromatin domains across the conservative set lincRNAs loci. Distribution of the 4662 conservative lincRNA loci with respect to evidence of a chromatin signature of actively transcribed genes (histone 3 lysine 4 tri-methylation (H3K4me3) across the promoter region , followed by H3K36me3 along the transcribed region; K4K36 domain) across the transcripts loci. 15% (701, green) of the lincRNAs loci overlap with a lincRNAs region identified by the K4K36 domain in a previous study [3]. 9% (425, blue) were supported by a K4-K36 domain identified in at least one of 9 cell lines analyzed as part of the Encode project (Encode level 1; **2.4 Methods**) and that is not shared with a coding gene. An additional 17% (772) were also supported by a K4-K36 domain from the Encode cell-lines, however, this time the K4-K36 domain may also overlap a known coding gene (purple ; **2.4 Methods**). A lincRNA was considered as having a K4-K436 domain in the Encode set if it overlapped a H3K4me3 peak within +/- 2 KB from its start site, and also overlapped a H3K36me3 pick across the transcribed region.

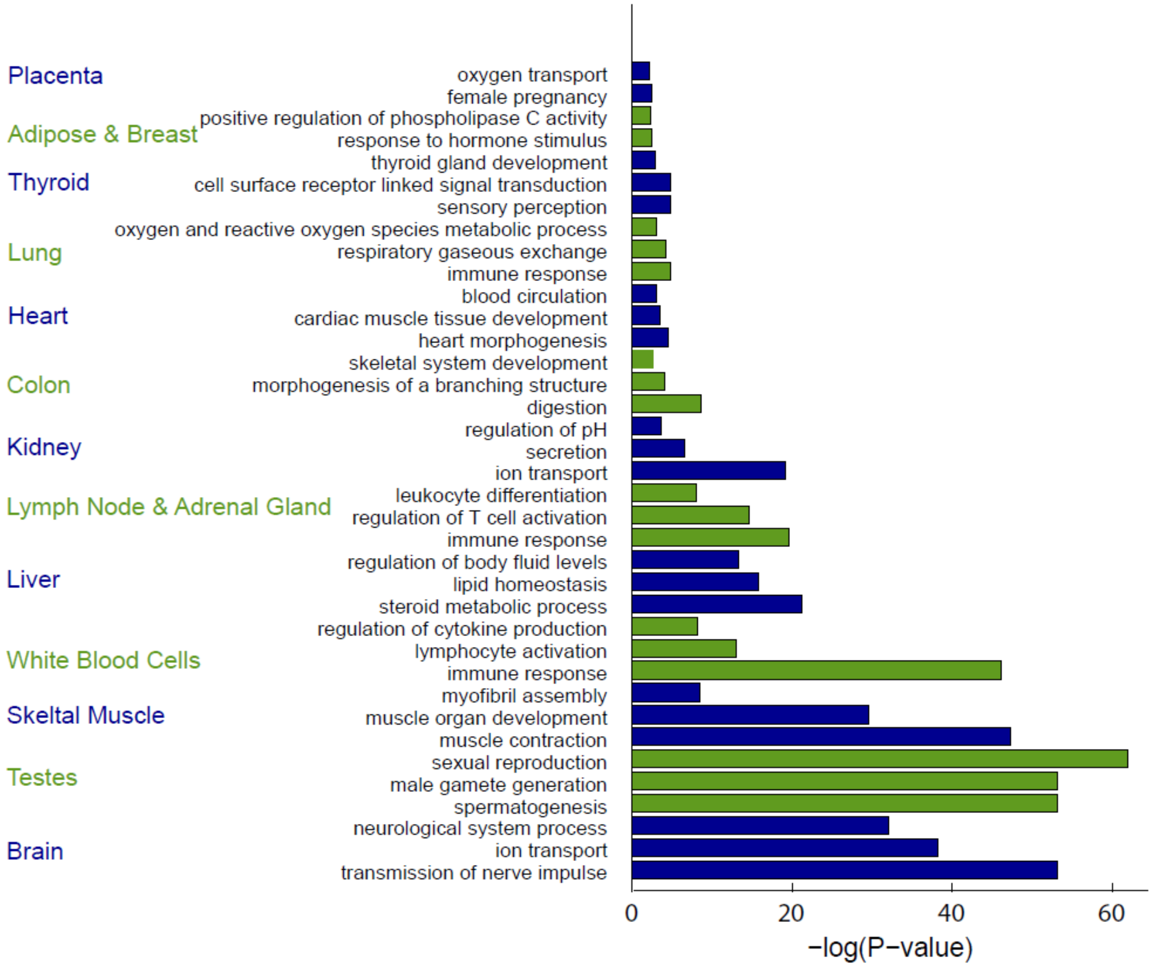


Set	Mean (max log2 FPKM)	Median (max log2 FPKM)
Protein Coding Genes	4.73	4.77
Stringent lincRNAs	1.18	1.25
lincRNAs – not testes specific	1.07	0.89
K4K36 lincRNAs	1.37	1.26
lincRNAs- known to Refseq	1.61	1.71

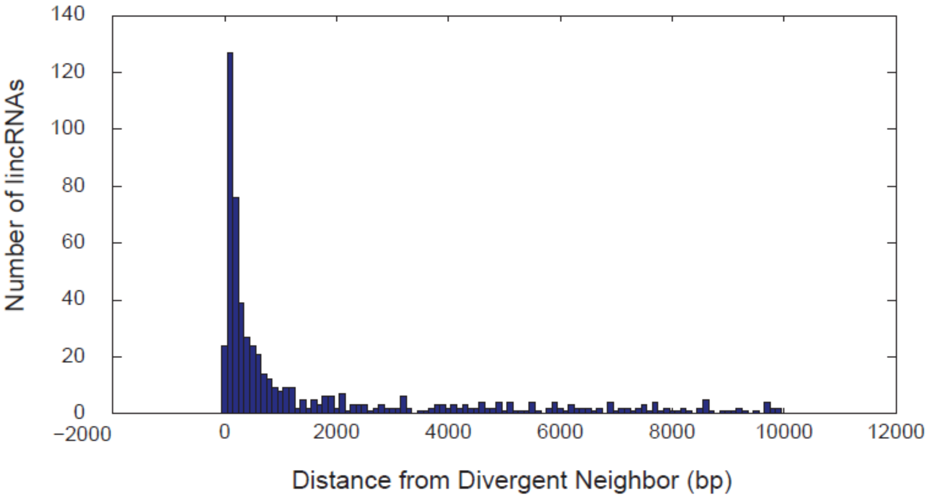
Supplementary Figure 1.5| Maximal expression levels of lincRNAs. (a) Shown are distributions of maximal expression abundance across all tissues (log2 normalized FPKM counts as estimated by Cufflinks) of conservative lincRNAs (blue), conservative lincRNAs with K4-K36 support (green, **2.4 Methods**), lincRNAs which are not testes-specific (cyan), lincRNAs known to RefSeq (red; RefSeq NR) and coding genes (black). **(b)** Mean and median of maximal expression scores for each of the transcript sets presented in (a).



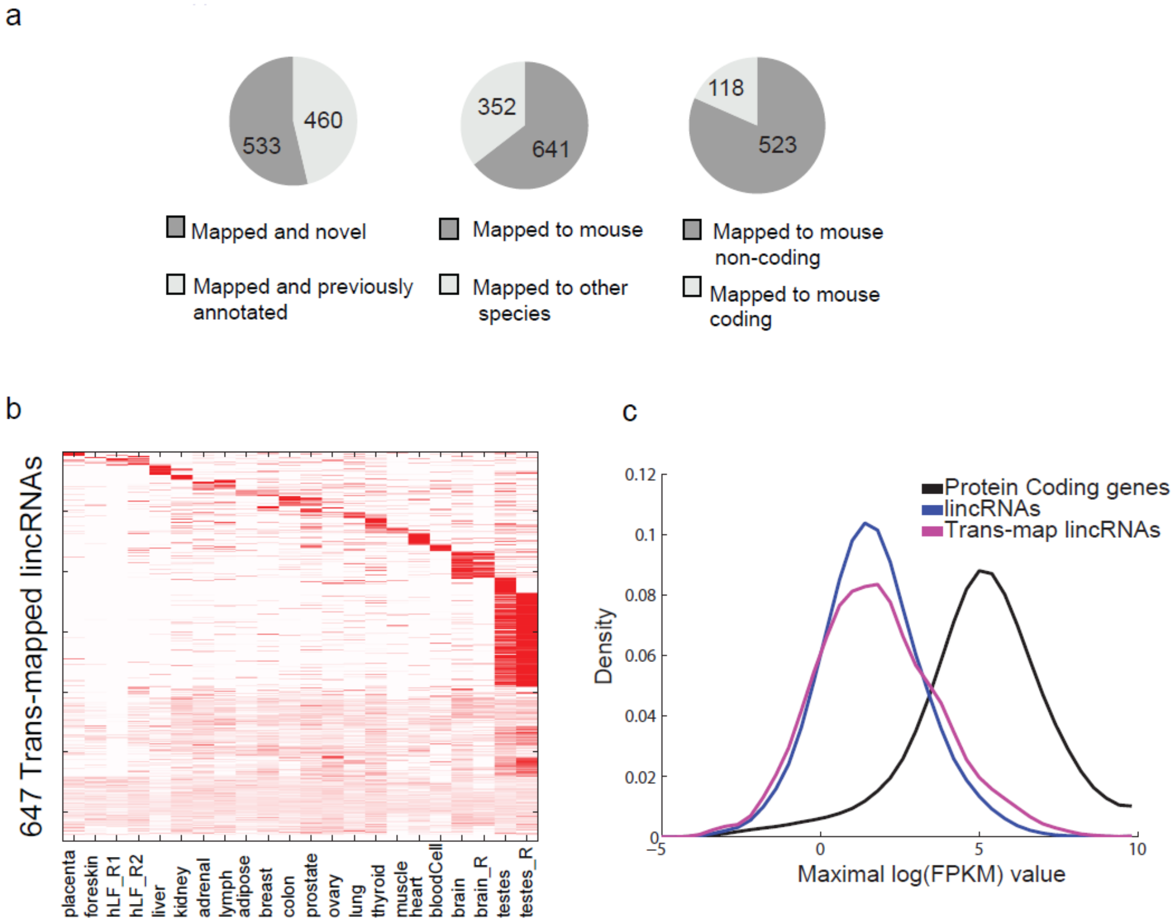
Supplementary Figure 1.6| Tissue specificity of lincRNAs expression. (a) Shown are distributions of maximal tissue specificity scores calculated for each transcript across all tissues, for coding genes (black), different sets of lincRNAs (blue, green, and cyan, defined as in **Supplementary Figure 1.5**), as well as lincRNAs that are divergently transcribed (red) and trans-mapped lincRNAs (magenta). **(b)** Comparison of distribution of tissue specificity scores between lincRNAs and coding genes at three levels of maximal expression ranges. Transcripts in each panel have a maximal expression level within the range specified on top of the panel: low expression (left panel) ranges between (-1.5) -1.5 log2 FPKM, moderate expression (middle panel) ranges from 1.5-3.5 log2 FPKM and high expression (right panel) ranges from 3.5-5.5 log2 FPKM.



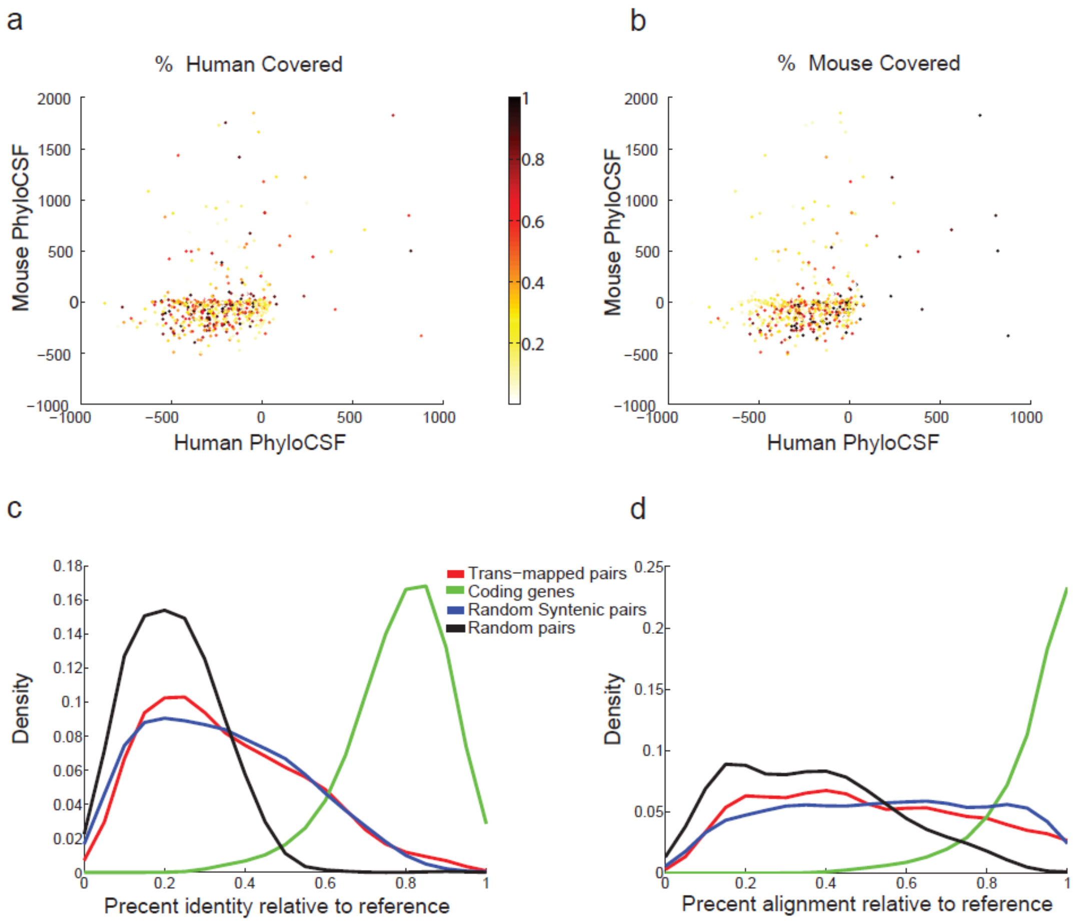
Supplementary Figure 1.7| Functional enrichment of tissue specific expression clusters. Representative Gene Ontology (GO) terms that are significantly enriched among protein coding genes assigned to tissue specific clusters as was calculated by David functional annotation tool [2]. Shown are the nominal P-values calculated by a Fisher exact test (FDR < 10%). A full list of enriched functional sets is provided in **Supplementary Dataset 3**.



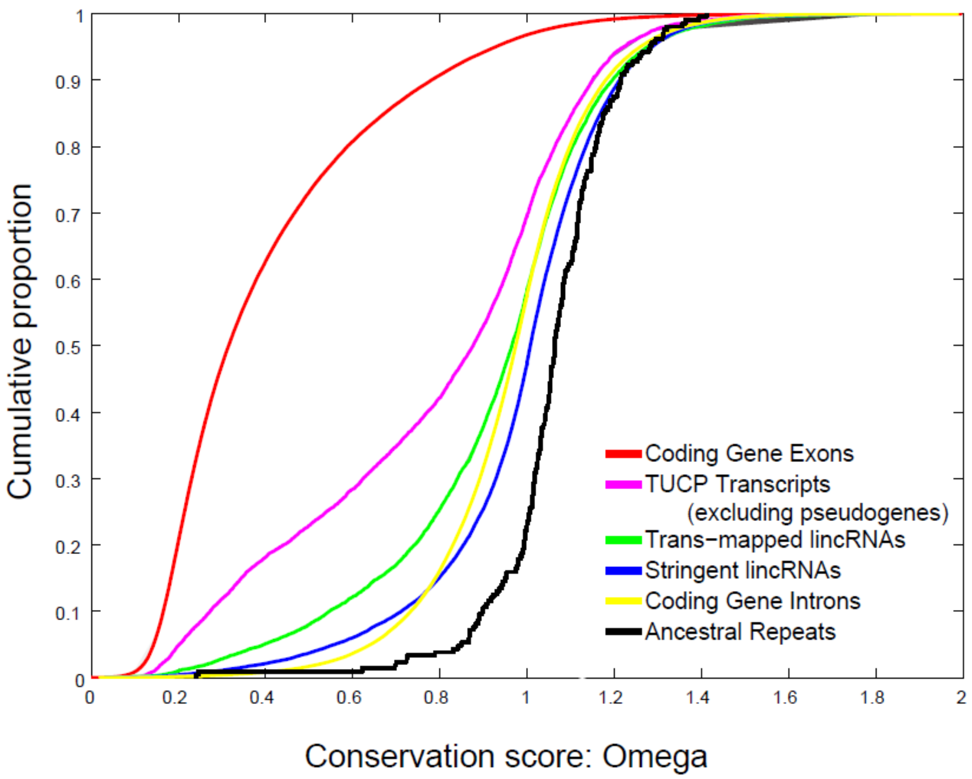
Supplementary Figure 1.8| Distance between divergent lincRNAs and their coding neighbors. Shown is the distribution of base pair distance between the transcription start site of 588 lincRNAs and coding genes that are divergently transcribed.



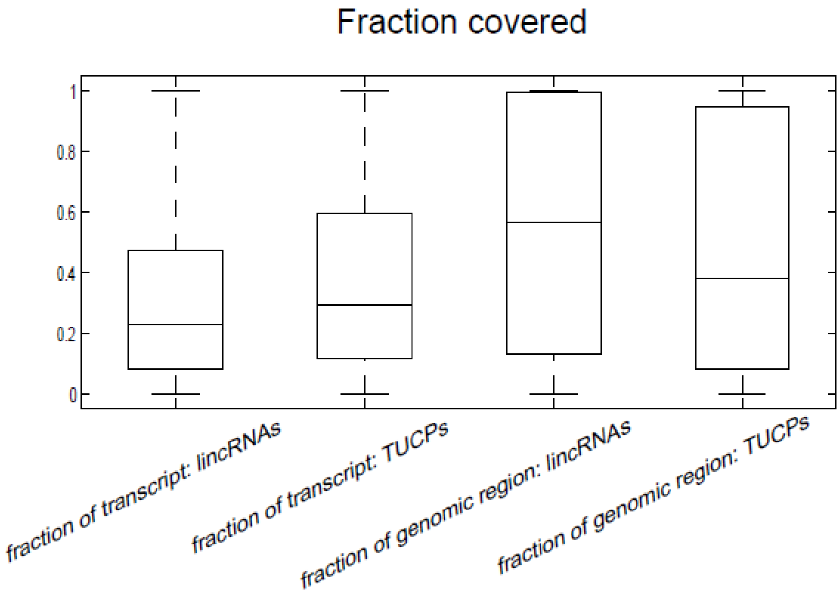
Supplementary Figure 1.9| Trans-mapped lincRNAs. (a) 533 of the 993 Trans-mapped lincRNAs (53%, dark grey) are designated as novel, such that they are not included in the public annotation resources used for this study and were only identified using RNA-Seq. **(b)** 641 (64%, dark grey) of the trans-mapped lincRNAs are mapped to mouse (and possibly to other species), while the rest are mapped only to species other than mouse (light grey). **(c)** Of the 641 lincRNAs that have a mouse syntenic ortholog, only 118 (18%, light grey) of their orthologs are classified as protein coding genes in the UCSC classification. **(d)** Expression patterns of 647 lincRNA that are in the conservative set and are trans-mapped. Shown are the expression patterns of the 647 lincRNAs (rows) across the tissues (columns, data presented as in **Figure 2a**). Color intensity is the fractional density across the row of log normalized FPKM counts as estimated by Cufflinks (saturating less than 4% of the top normalized expression values). **(e)** Distributions of maximal expression abundance of each lincRNA (blue), Trans-mapped lincRNAs (pink) and coding (black) transcripts across tissues. Expression levels are log2 normalized FPKM counts as estimated by Cufflinks.



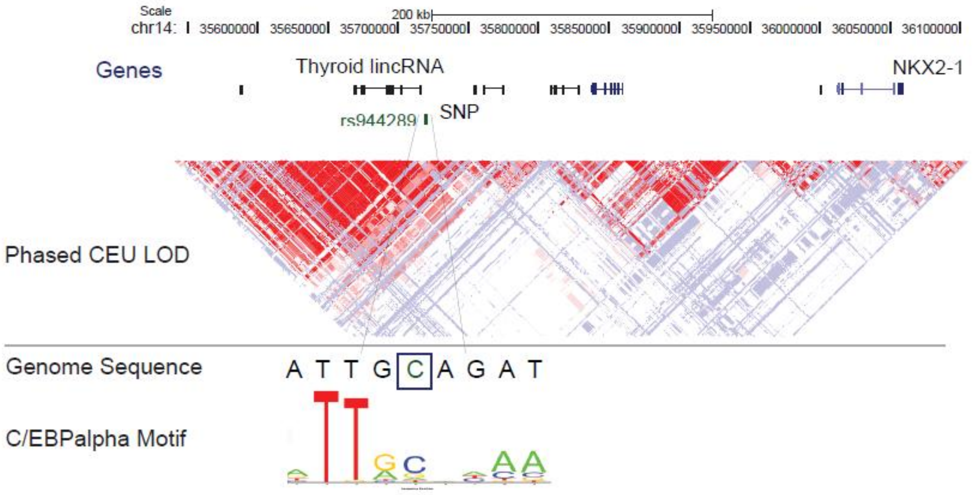
Supplementary Figure 1.10| Alignments of trans-mapped lincRNAs to orthologous transcripts. (a,b) Shown are scatter plots comparing the PhyloCSF score (coding potential) in 614 pairs of human Trans-mapped lincRNAs (X axis) and their syntenic mouse orthologs (Y axis). Dots are colored either by the fraction of the human lincRNA transcript that is covered by the mapping to a mouse syntenic ortholog (a) or by the fraction of the mouse orthologous transcript that is covered by the mapping (b). Mouse transcripts with a positive PhyloCSF have a significantly smaller region covered by a human ortholog (P < 0.04, K-S test). **(c-d)** Distribution of % identity (c) and % alignment (including gaps and mismatches, d) of the human transcript in either trans-mapped pairs of lincRNAs and its orthologous transcript (red), random sequence pairs (black), randomly selected pairs of syntenic blocks (blue), and pairs of orthologous coding genes (green) (**2.4 Methods**).



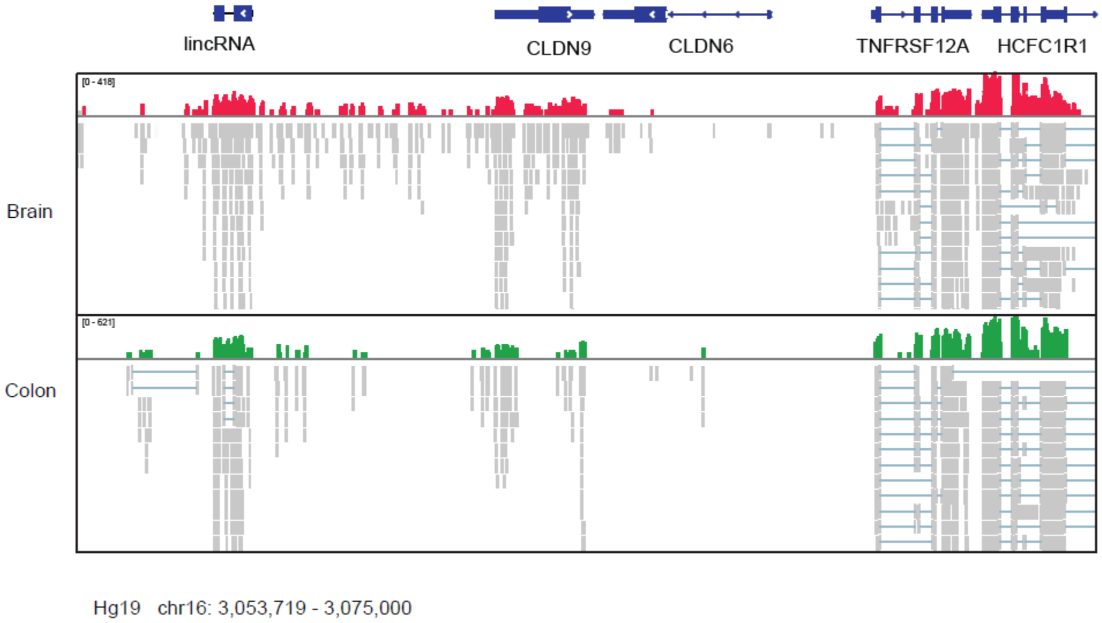
Supplementary Figure 1.11| Exon sequences of TUCP transcripts are under stronger evolutionary conservation. Shown is the cumulative distribution of conservation levels across 29 mammals (omega [4]) in the exons of protein coding genes (red), TUCP transcripts (magenta), Trans-mapped lincRNAs(green), conservative lincRNAs (blue) , introns of coding genes (yellow) and ancestral repeats (black) . Only transcripts with a sufficient cross-species alignment support (branch length > 0.5) are included in the plot, corresponding to over than 80% of each set, except for the ancestral repeats for which only 40% had sufficient alignment support. The ancestral repeats serve as a null model of neutral selection. The intron set was created by uniformly sampling a size matched intronic fragment from the intron neighboring each coding exon. Pseudogenes were excluded from the TUCP transcripts to avoid bias in the conservation estimate caused by alignments to the paralog gene of the pseudogenes.



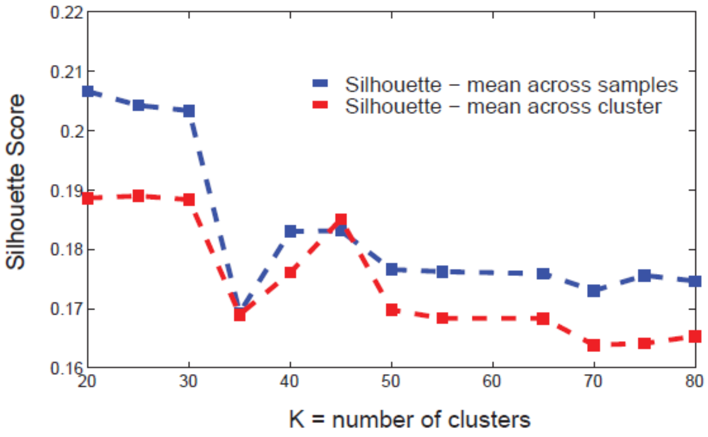
Supplementary Figure 1.12| Evaluation of Trans-mapped TUCP transcripts. Shown are boxplots representing the fraction of the human transcript that is syntenicaly aligned to an ortholog (1st and 2nd boxes) and the fraction of the lincRNAs genomic locus covered by the syntenic mapping of the ortholog (3rd and 4th boxes) for lincRNAs (1st and 3rd boxes) or TUCP transcripts (2nd and 4th boxes). Both fraction of transcript mapped and fraction of genomic region mapped are significantly different between conservative lincRNAs and TUCP transcripts ($P<3.3*10^{-4}$ and $P<1.8*10^{-4}$, respectively, KS test), but with small effect sizes (24% and 18%, respectively).



Supplementary Figure 1.13| A thyroid specific lincRNA within an intergenic region associated with thyroid cancer. Top: Shown are UCSC genome browser (Kent et al. 2002) tracks showing the log odds (LOD, blue (low) to red (high)) scores between SNPs from the CEU HapMap population and the relative location of coding genes (blue) and lincRNAs (black) in a region spanning a thyroid cancer associated SNP (rs944289, green), based on the UCSC Genome Browser, build 36. Bottom: a magnified view of the 9 nucleotide surrounding the SNP (rs944289; marked by blue box) and the Logo view of the C/EBPalpha binding motif that overlaps the SNP.



Supplementary Figure 1.14| Caveats of transcript abundance estimation approach. Shown is a snapshot from the Integrative Genome Viewer (Robinson et al. 2011) describing a lincRNA that is maximally expressed in brain according to Cufflinks' abundance information, although there are no spliced RNA-Seq reads that support the expression of this specific spliced isoform (in comparison to the colon where there are spliced reads supporting the exon). Upper panel: transcripts structure (blue). Middle panel: RNA-Seq coverage (red) and reads (gray*) in brain. Bottom panel: RNA-Seq coverage (green) and reads (gray*) in colon. * RNA-Seq read are represented as gray rectangles while light blue lines mark gaps across the read alignment.



Supplementary Figure 1.15| Optimizing the selection of the number of clusters (K) for K-means clustering. Shown is the distribution of Silhouette (Si) scores (y-axis) across different values of K (x-axis), the number of clusters used to obtain tissue specific clusters of lincRNAs and coding genes. Average Silhouette score was calculated by the mean Si across all data samples (blue), or by taking the average of each cluster's average Si (red). Higher Si implies a better clustering of the data.

Supplementary Table 1.1| Transcript reconstruction comparison between biological replicates.

	Human Lung Fibroblast	Brain	Testes
% of stringent set transcript loci from the low coverage sample that were identified by a matching multi-exon or single-exon isoform in the high coverage sample	69%	74.5%	81.6%

Supplementary Table 1.2| Number of transcripts across species

Species	Species name	Number of transcripts trans-mapped across species	Number of ESTs	Number of RefSeq genes
Mouse	*Mus musculus*	2313560	4367822	28539
Zebra fish	*Danio rerio*	792572	1646898	16158
Cow	*Bos Taurus*	685264	1621863	13964
Rat	*Rattus norvegicus*	424072	1121086	17172
Madka fish	*Oryzias latipes*	389820	703843	584
Chicken	*Gallus gallus*	258332	618531	5468
Dog	*Canis familiaris*	121789	399902	1212
Tetraodon	*Tetraodon nigroviridis*	103162		
Rhesus	*Macaca mulatta*	34270	68320	2566
Orangutan	*Pongo pygmaeus abelii*	29657	46027	3555
Zebra finch	*Taeniopygia guttata*	27017	92700	170
Horse	*Equus caballus*	15803	37539	996
Rabbit	*Oryctolagus cuniculus*	13561	36332	1334
Stickleback fish	*Gasterosteus aculeatus*	9042	301622	
Opossum	*Monodelphis domestica*	1712	459	467
Platypus	*Ornithorhynchus anatinus*	486	47782	383

Supplementary Table 1.3| Disease associated intergenic regions containing a lincRNA that is expressed in a disease related tissue.

Disease/trait	SNP	P value; OR/ beta	lincRNAs	Tissue	Reference	Note
Thyroid cancer	rs944289	2E-9 ;1.37	XLOC_010996	Thyroid	[5]	Strikingly high expression in the thyroid
Waist-hip ratio	rs1055144	9.9E-25 ; 0.04	XLOC_006016	Adipose	[6]	Ortholog transcript in cow, some EST support
LDL cholesterol	rs2126259	7E-12; 0.02	XLOC_006706	Liver	[7]	Transcript identified by GENCODE with support by RNA-Seq.
Bipolar disorder	rs472913	2E-7; 1.18	XLOC_000856	Brain	[8]	Partial isoform identified by GENCODE
Prostate cancer	rs9284813	5E-9	XLOC_002726	Prostate	[9]	Partial isoform identified by GENCODE

Supplementary Table 1.4| RNA-Seq datasets.

Tissue/ Cell type	Dataset	Sequencing platform	Read length	Number of aligned reads
Human lung fibroblasts (hLF)	Rinn lab	Illumina genome analyzer II (GAII)	75 base pairs (bp), paired ends (PE)	33849576
hLF2	Rinn lab	GAII	75 pb, PE (2 lanes)	136310398
Foreskin fibroblasts	Rinn lab	GAII	75 pb, PE	58787488
Brain	Rinn lab	GAII	75 pb, PE	50661137
HeLa	Rinn lab	GAII	75 pb, PE	30134104
Liver	Rinn lab	GAII	75 pb, PE	32132725
Placenta	Rinn lab	GAII	75 pb, PE	47384953
Testes	Rinn lab	GAII	75 pb, PE	49171324
Adipose	Illumina Human Body Map2	High-Seq	75 pb single end (SE), 50 bp , PE	221715491
Adrenal gland	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	218793183
Brain	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	204772783
Breast	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	219398164
Colon	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	244916925
Heart	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	246232044
Kidney	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	229699125
Liver	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	239123970
Lung	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	240042489
Lymph Node	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	249015289
Ovary	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	241958761
Prostate	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	251395450
Skeletal Muscle	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	242833112
White Blood Cells	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	248288643
Testes	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	241527185
Thyroid	Illumina Human Body Map2	High-Seq	75 pb SE, 50 bp PE	235187522

Supplementary Table 1.5. Publically available annotations used for this study.

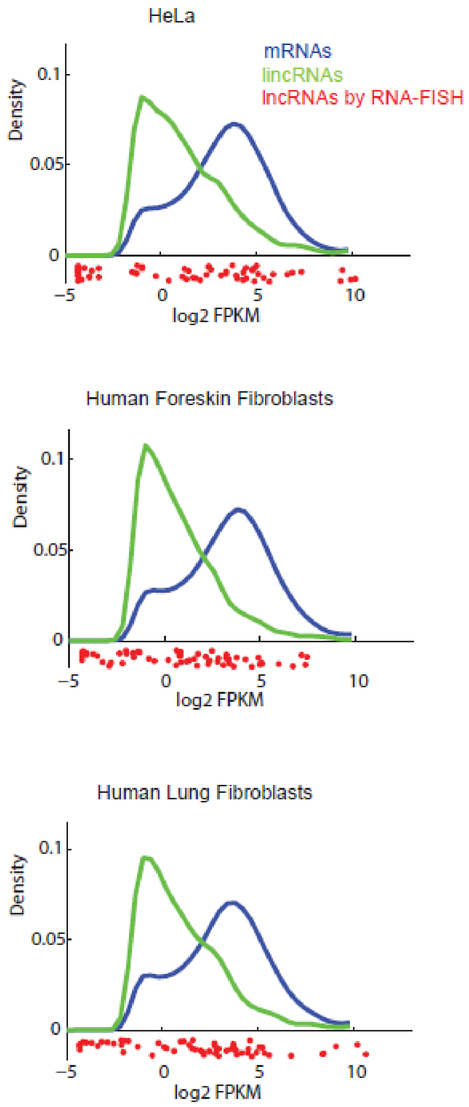
Dataset name	Downloaded from (date)	refs	# of transcripts	Note
Protein coding transcripts known to RefSeq	The UCSC genome browser ; NCBI37/Hg19 (July 2010)	[10]	31911	Extracted only transcripts with an identifier beginning with NM_
microRNA and snoRNAs set1	The UCSC genome browser; NCBI37/Hg19 (November 2010)	[11, 12]	1341	Data was collected from miRBASE and snoRNABase databases.
miRNA, miRNA_pseudogene, rRNA, scoRNA, snoRna, tRNAs	Biomart Ensembl ; NCBI37/Hg19 (November 2010; Ensembl Release 60)	[13]	6219	Based on Ensembl annotation
Pseudogenes set1 – Vertebrate Genome Annotation (Vega) pseudogenes	The UCSC genome browser NCBI37/Hg19; (July 2010; Vega Release 38)	[14]	9534	Classification scheme: http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html Eliminated the ones that overlapped a Refseq NR_
Pseudogenes set2 - Yale Gerstein Group Pseudogenes	From : http://pseudogenes.org/ (July 2010 ; build 37)	[15]	16363	
UCSC coding genes	The UCSC genome browser NCBI37/Hg19 (November 2010); The kgTxInfo primary table	[16, 17]	69201	Extracted transcripts annotated as coding (ref classification 2008 ; http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_kgMethod=1, http://genome.ucsc.edu/cgi-bin/hgGene?hgg_do_txInfoDescription=1)
GENCODE coding genes	GENCODE Version 4, from : ftp://ftp.sanger.ac.uk/pub/gencode/ Nov 2010	[18]	29782	Downloaded file : GENCODE_v4.annotation.GRCh37.gtf.gz Extracted genes that were manually annotated (level 1-2) with the types : gene_type = "protein_coding" , transcript_status = "KNOWN"
Noncoding transcripts known to Refseq	The UCSC genome browser ; NCBI37/Hg19 (July 2010)	[10]	4883	Extracted only transcripts with an identifier beginning with NR_

Non coding transcripts known to UCSC	The UCSC genome browser NCBI37/Hg19 (November 2010); The kgTxInfo primary table	[16, 17]	7673	Extracted transcripts annotated as non-coding (ref classification 2008)

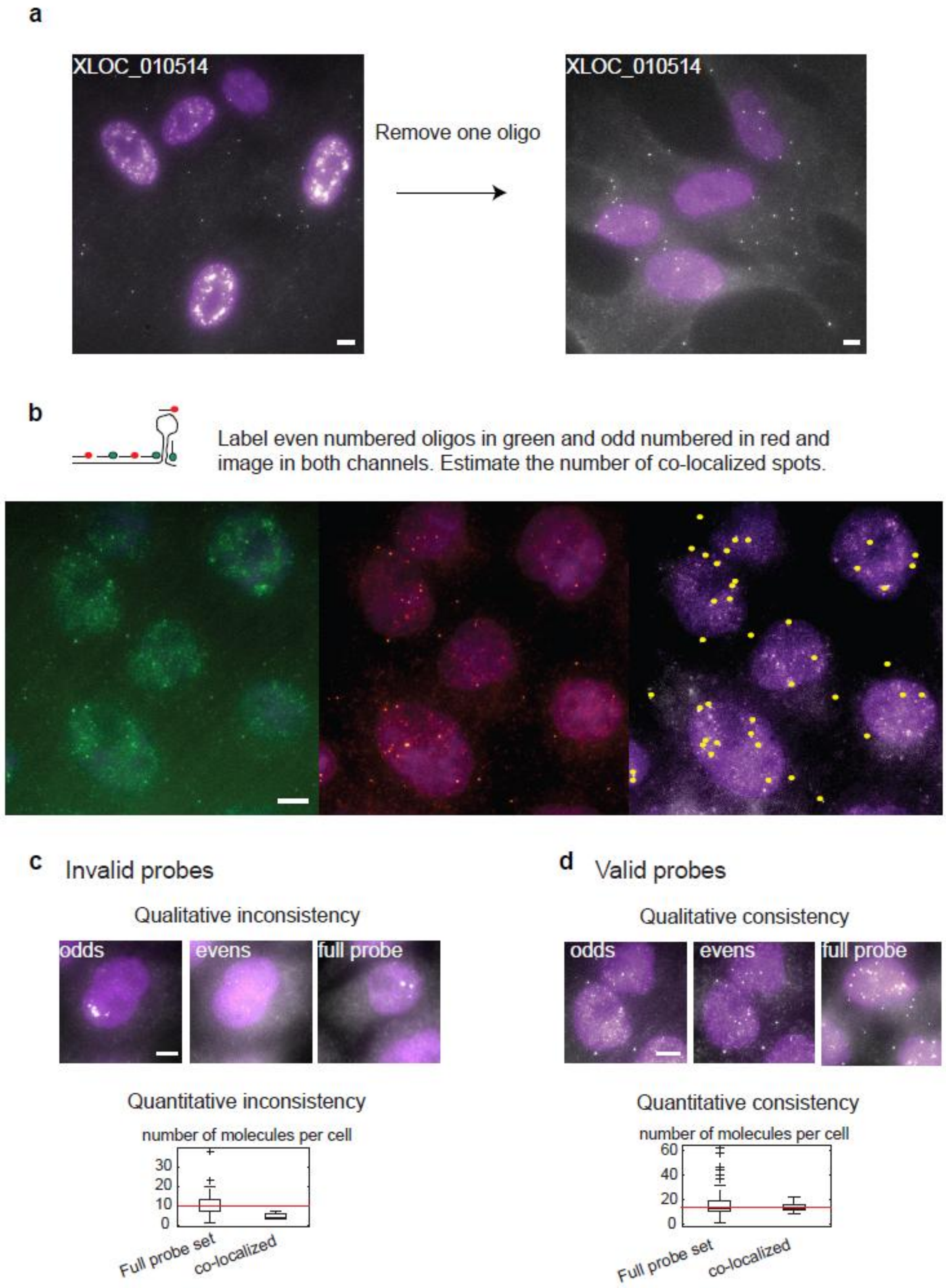
Non coding transcripts known to GENCODE	GENCODE Relese 4
	From : ftp://ftp.sanger.ac.uk/pub/GENCODE/
	Nov 2010
		[18]	14598	Extracted genes that were annotated with the types :
				gene_type = "processed_transcript", or "lincRNA", or "non_coding";
				transcript_status = "processed_transcript", or "lincRNA" , or "non_coding", or "misc_RNA";
				Based on the Havana annotation guidelines : http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/assets/guidelines.pdf

RefSeq derived lincRNAs				Run "non coding transcripts known to Refseq" set through steps 1,3,6 of our lincRNA classification pipeline
GENCODE derived lincRNAs				Run "non coding transcripts known to GENCODE" set through steps 1,3,4,5,6 of our lincRNA classification pipeline
UCSC derived lincRNAs				Run "non coding transcripts known to UCSC" set through steps 1,3,4,5,6 of our lincRNA classification pipeline

Supplementary Figures and Tables of Chapter 3



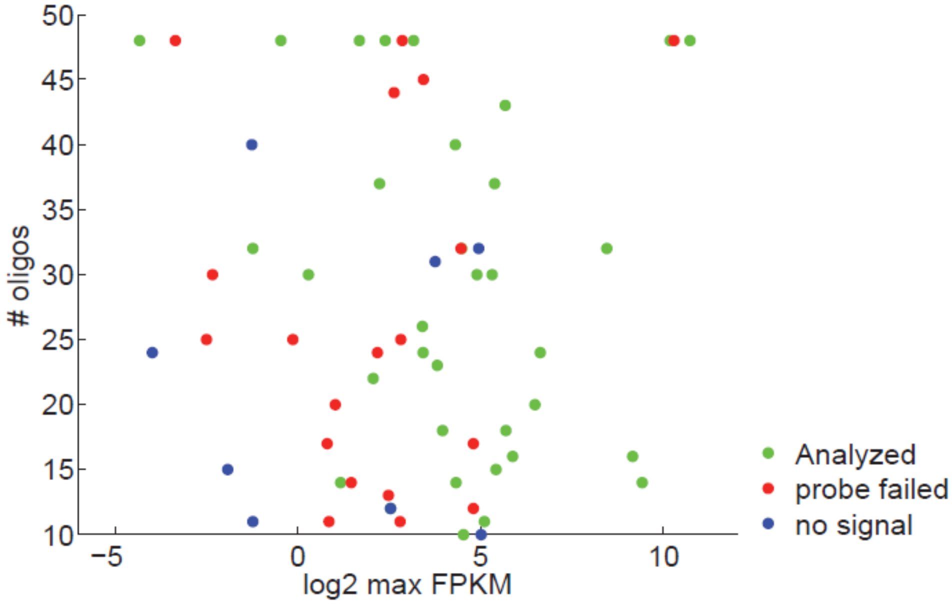
Supplementary Figure 2.1| LncRNA expression. Shown are the distribution of (bulk) expression abundance estimates ($\log_2$(FPKM) as estimated by Cuffdiff2[19]) of all expressed lncRNAs (green), coding genes (blue) and the 61 selected lncRNAs for this study (red) in HeLa (top), human foreskin fibroblasts (middle) and human lung fibroblasts (bottom).



Label even numbered oligos in green and odd numbered in red and image in both channels. Estimate the number of co-localized spots.

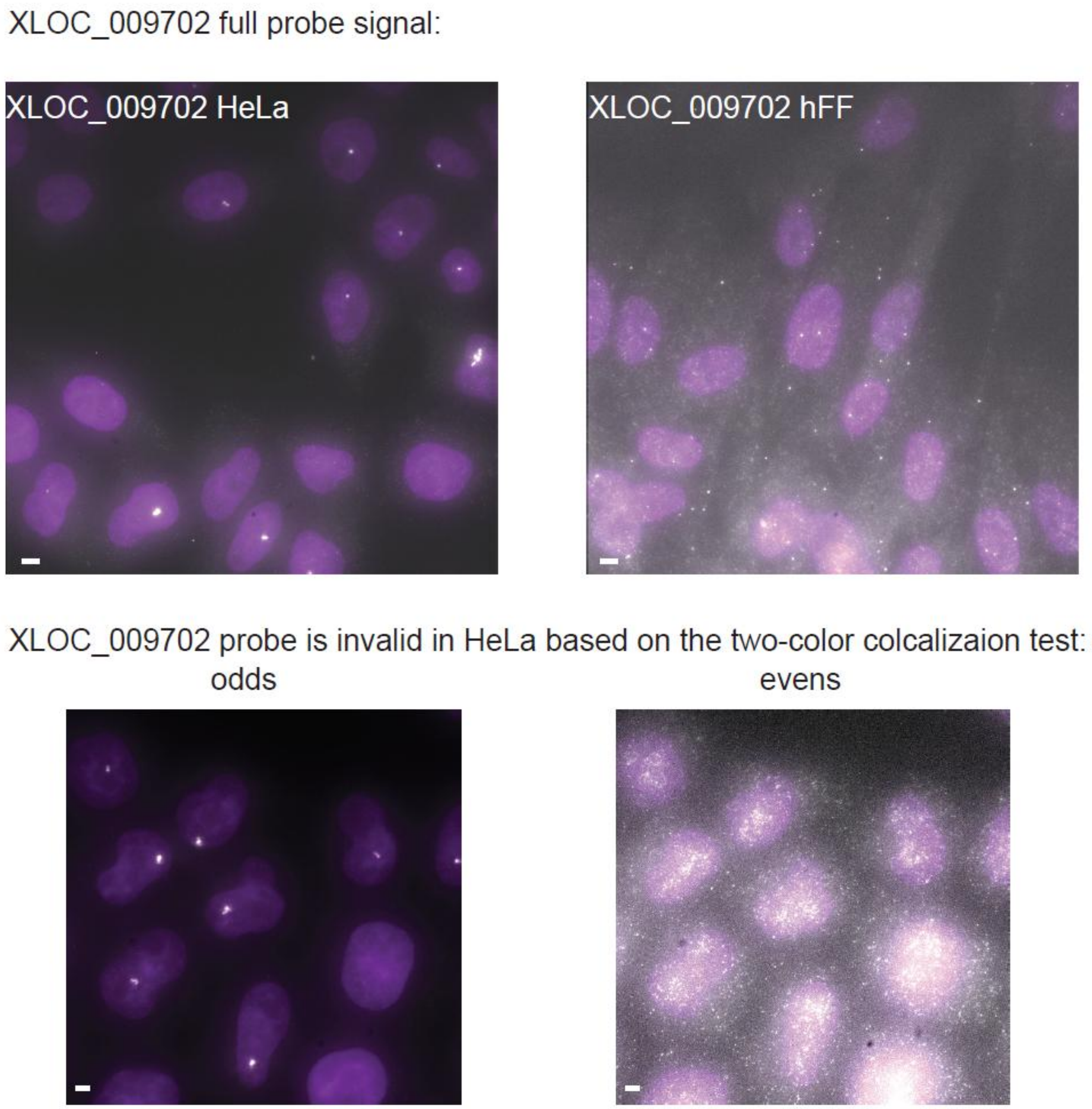
Supplementary Figure 2.2 | Two color co-localization approach for RNA-FISH probe set validation.

Supplementary Figure 2.2 | Two color co-localization approach for RNA-FISH probe set validation **(a)** Florescence micrograph of the original probe set (left) targeting XLOC_010514 and the validated probe set after eliminating one oligonucleotide in hFF (right; **3.4 Methods**). **(b)** Top: Illustration of the two color labeling of even and odd numbered oligonucleotides within a probe set for the co-localization probe validation assay. Bottom: Florescence micrographs of the NR_029435 probe set in HeLa of the even numbered oligonucleotides (left, green; Alexa 594), odd numbered oligonucleotides (middle, orange; Cy3) and co-localized spots over the even numbered set micrograph (right, yellow over white; Alexa 594). **(c)** Examples of qualitative and quantitative inconsistencies. Top: qualitative inconsistency between the localization pattern of MIAT in HeLa cells as determined by odd numbered oligonucleotides (left), even numbered (middle) and full probe set (right). The pattern matches between odd and full set, but not evens. Bottom: quantitative inconsistency between the distribution of RNA molecule counts (Box plots) based on the full probe set (left) and based on the number of co-localized spots when imaging in two colors (right). Red bar: medians. Whiskers are is 1.5* the inner quartile range. **(d)** Examples of qualitative and quantitative consistencies. As in *c* but for the valid probe set for NR_029435 in HeLa cells, which is both qualitatively (top) and quantitatively (bottom) consistent.

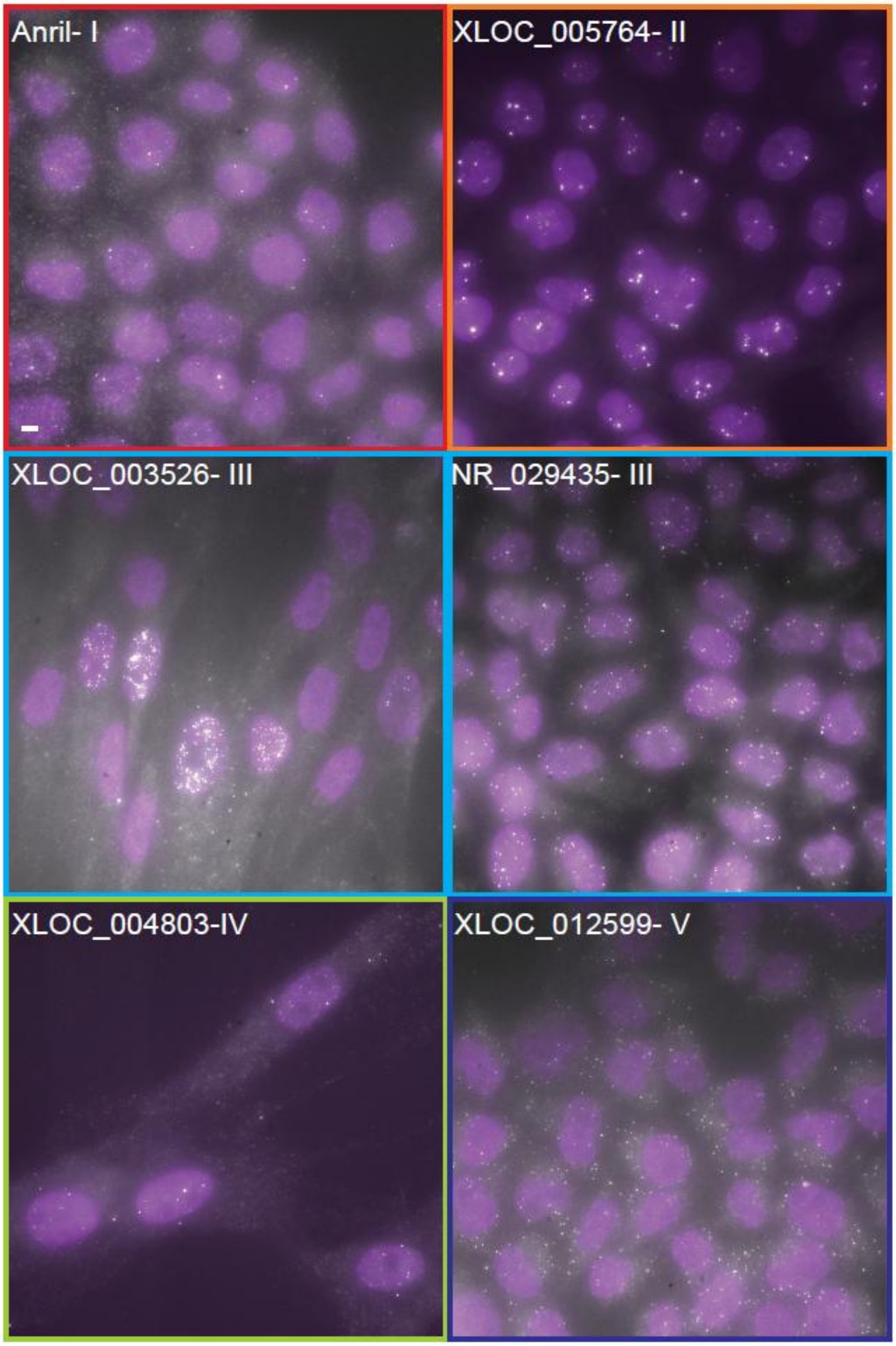


Supplementary Figure 2.3 | Dependency of probe set success on the number of oligonucleotides and lncRNA expression

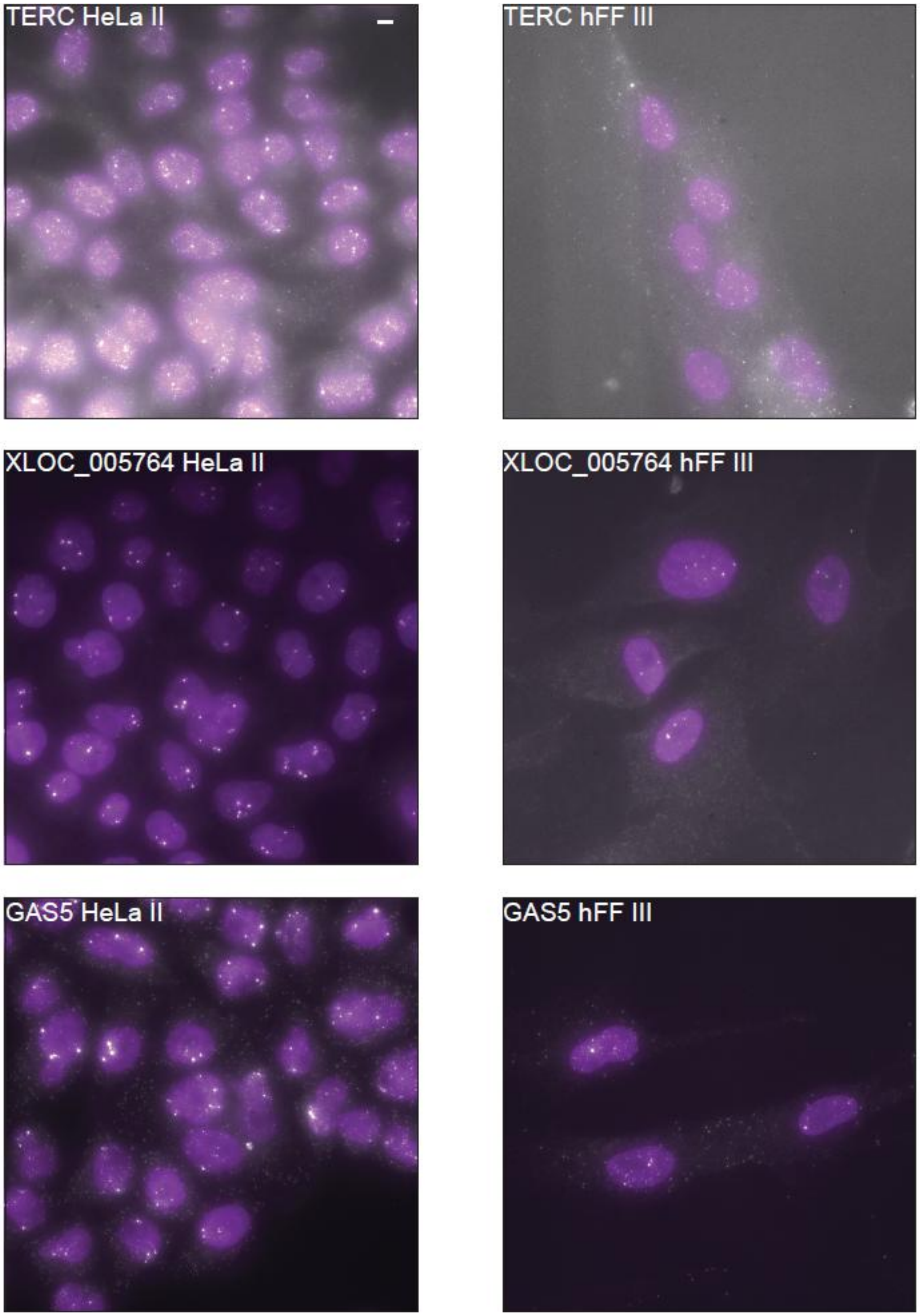
Scatter plot showing the relationship between the population average expression level as measured by RNA-Seq (X-axis, maximal $\log_2$(FPKM) across HeLa/hLF/hFF) and the number of oligonucleotides designed to target each lncRNA (Y-axis). Data points are colored by their classification to valid probe set included in the analysis (green), probe failed in the co-localization assay (red) or hybridization resulted with no signal (blue).


XLOC_009702 full probe signal:
XLOC_009702 HeLa
XLOC_009702 hFF
XLOC_009702 probe is invalid in HeLa based on the two-color colcalizaion test:
odds
evens

Supplementary Figure 2.4 | Qualitative inconsistency of the XLOC_009702 probe set in one cell type but not another
Shown are micrographs for XLOC_009702 imaged with the full probe set (top) in either HeLa (top left) or hFF (top right), or with the odds and even probe sets in HeLa (bottom left and right, respectively). The probe set is valid in hFF but fails the two-color co-localization assay in HeLa cells.

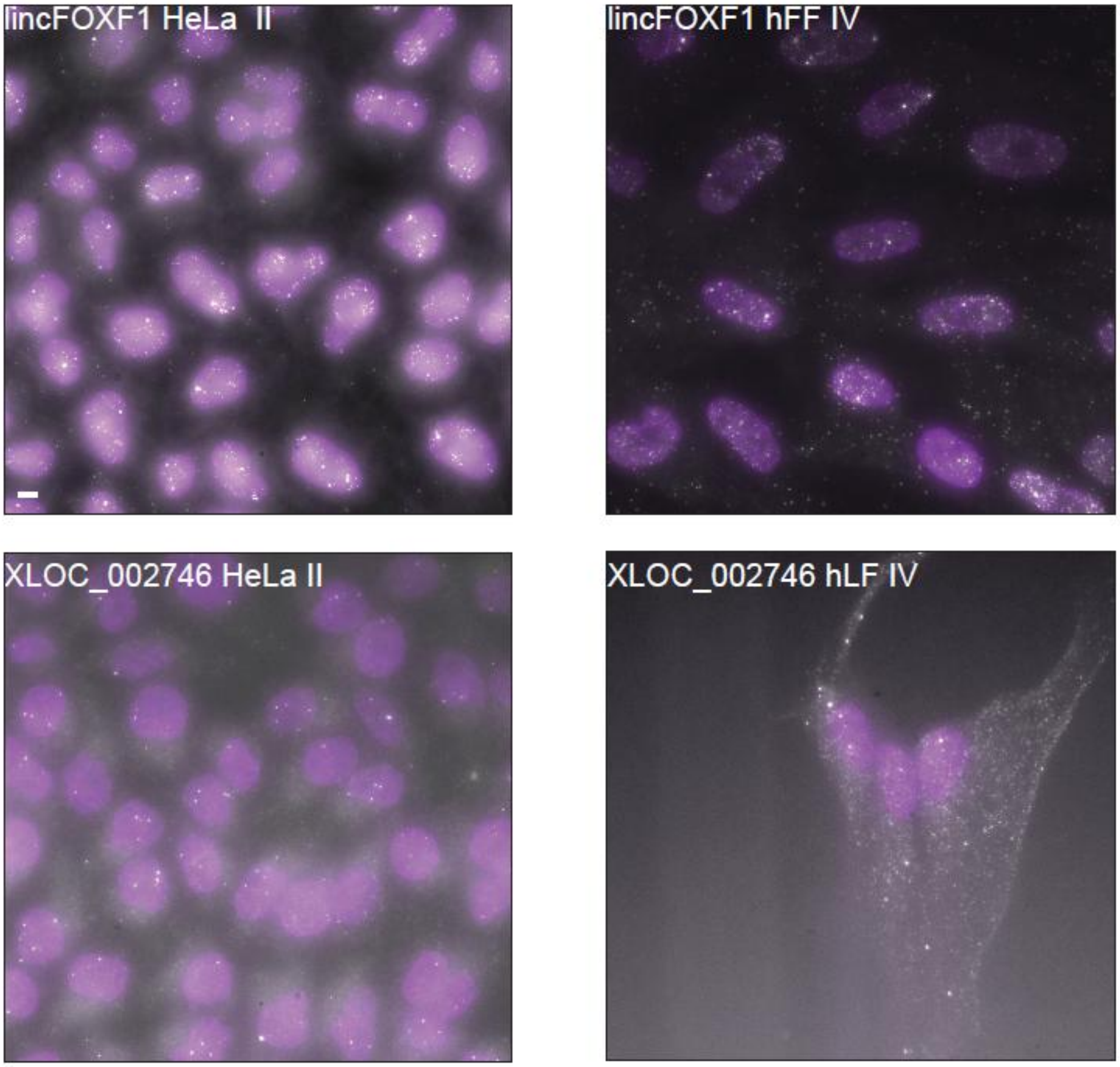


Supplementary Figure 2.5 | Cellular localization patterns
Shown are representative whole image fields for selected examples from each pattern type I-V, as defined in *Figure 2*. Scale bar, 5 µm.



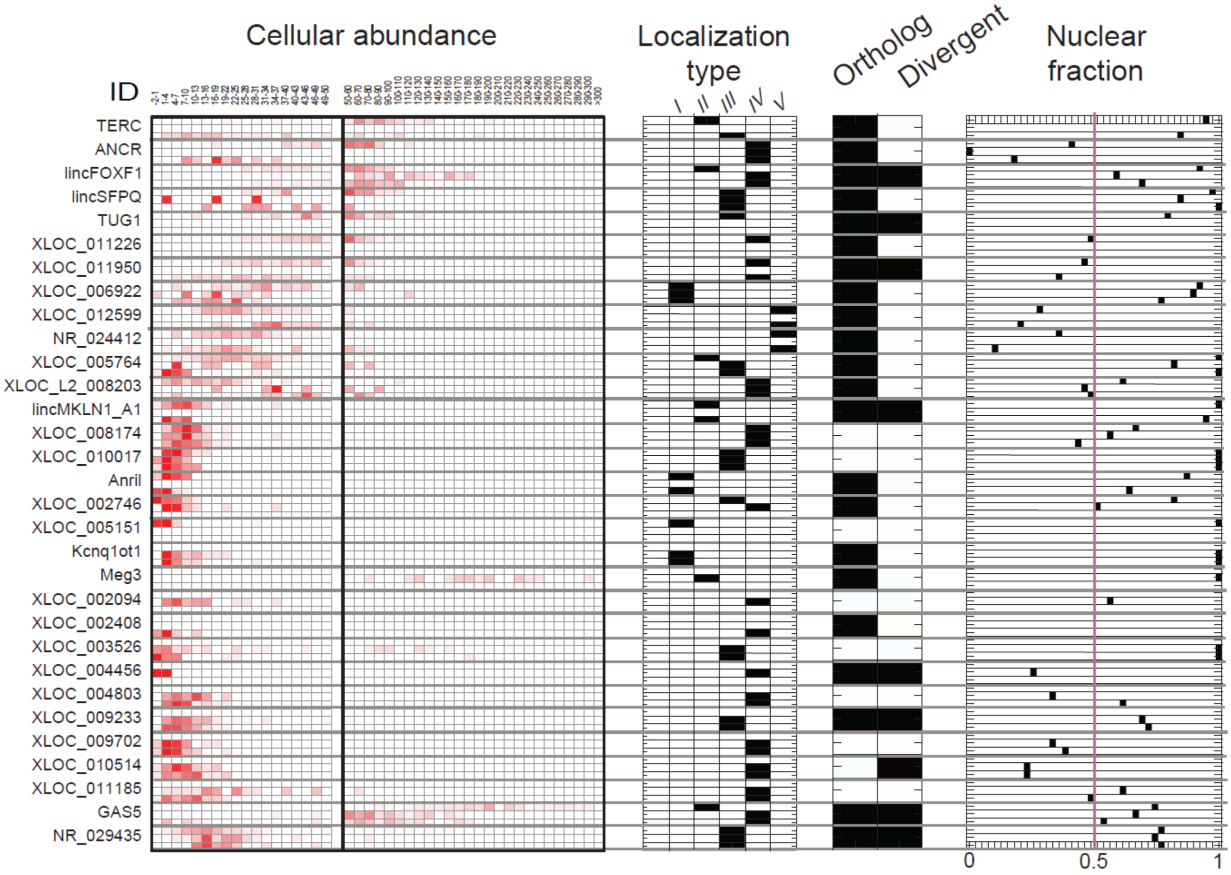
Supplementary Figure 2.6 | Presence of nuclear foci for the same lincRNA in one cell type but not another

Images demonstrating a pattern switch for TERC (top), XLOC_005764 (middle) and GAS5 (bottom) between HeLa cells (left), where expression is higher and nuclear foci are identified, and hFF (right), where foci are absent. Scale bar, 5 µm.



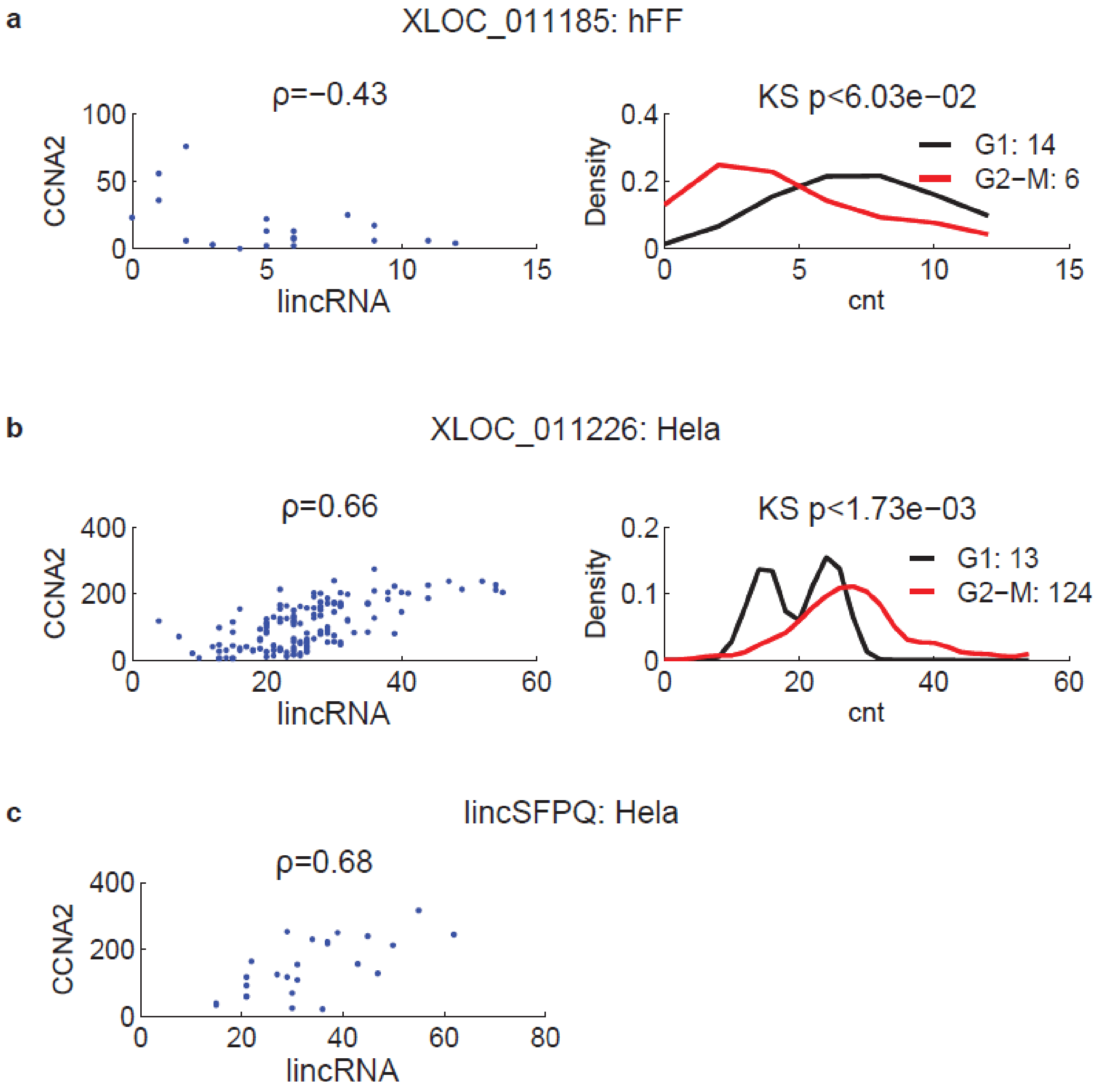
Supplementary Figure 2.7 | Presence of cytoplasmic transcripts for the same lincRNA in one cell type but not another

Images demonstrating a pattern switch for lincFOXF1 (top) and XLOC_002746 (bottom) between fibroblasts (right), where expression is higher and cytoplasmic transcripts are identified, and HeLa cells (left) where only nuclear expression is detected. Scale bar, 5 µm.

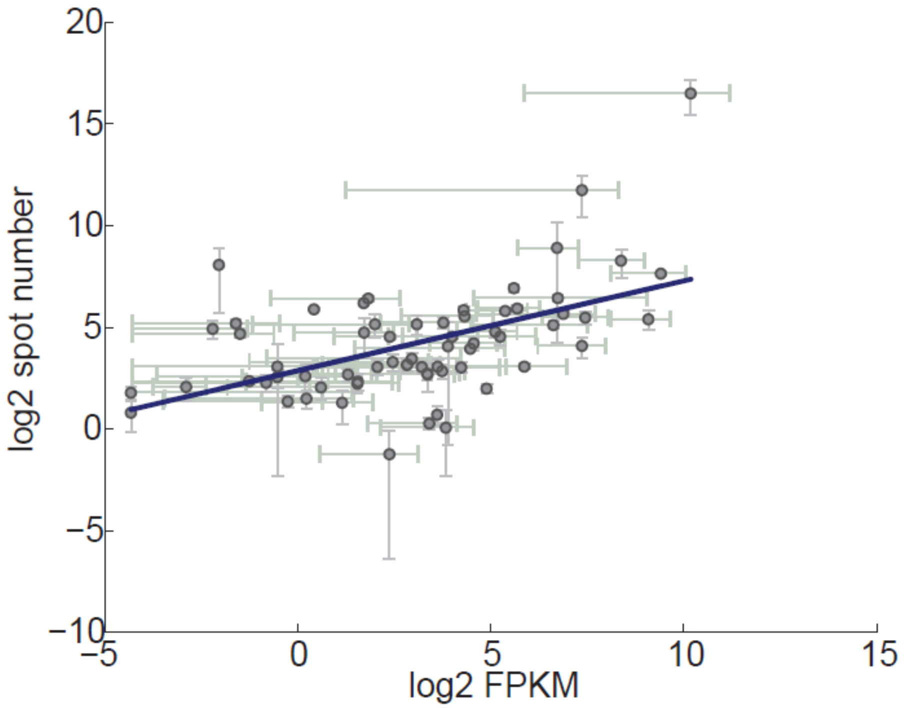


Supplementary Figure 2.8 | Summary of features for each lncRNA in the study

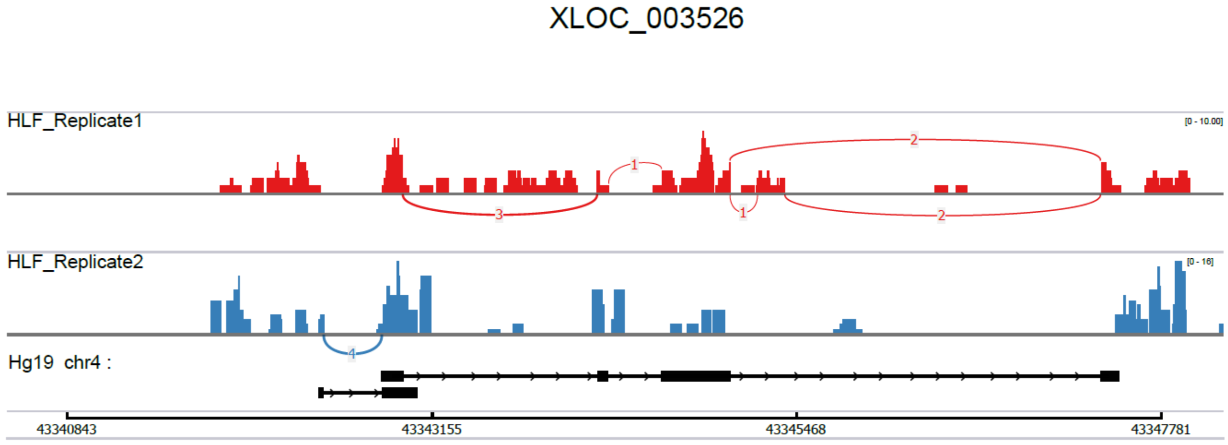
For each lncRNA in the validated and quantitative probe set (rows) shown are from left to right: abundance in HeLa, hLF and hFF (top to bottom sub rows, respectively; presented as fractional density of single molecule counts as in *Figure 4a*); classification to I-V localization type as in *Figure 2*; having a mouse orthologs or being divergently transcribed; and the median fraction of spots localized to the nucleus in expressing cells. Cell types for which the probe set is not valid are blanked. Positive classifications are marked black in the three right most panels.



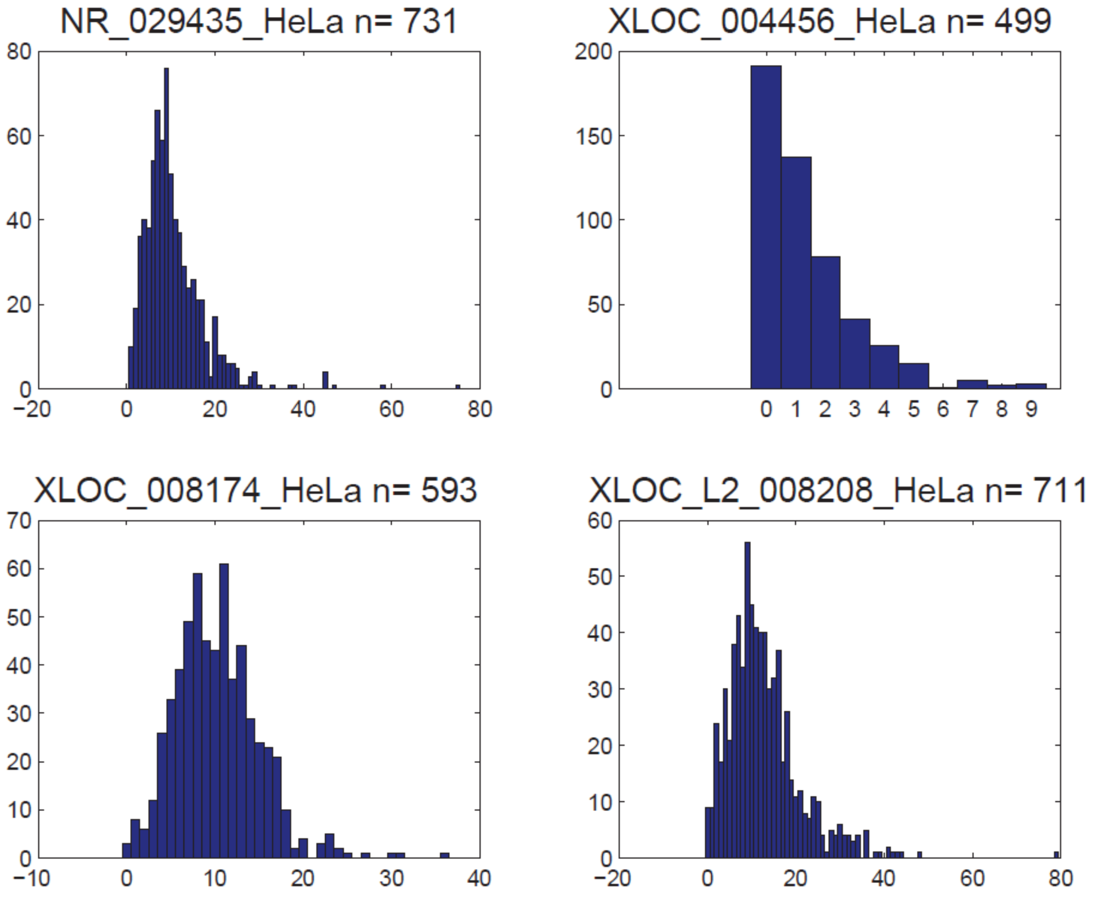
Supplementary Figure 2.9 | lncRNAs with cell-cycle correlated expression
For each of three lncRNAs found to be significantly correlated with CCNA2, irrespective of cell volume, shown are a scatter plot (left) comparing the estimated molecule counts of the lncRNA (X-axis) and CCNA2 (Y-axis) in each (Pearson correlation ρ is noted on top) and histograms (right) of the molecule count distribution of the lncRNA in cells in G1 (black) or G2/S/M (red) (number of cells is in top right; P-value of testing whether the distributions differ using a Kolmogorov-Smirnov test is on top). Cells were classified as G1 or G2/S/M based on a threshold on CCNA2 levels as specified in the **3.4 Methods**. (a) XLOC_011185 in hLF. (b) XLOC_011226 in HeLa. (c) lincSFPQ in HeLa. lincSFPQ did not have any cells in G1.



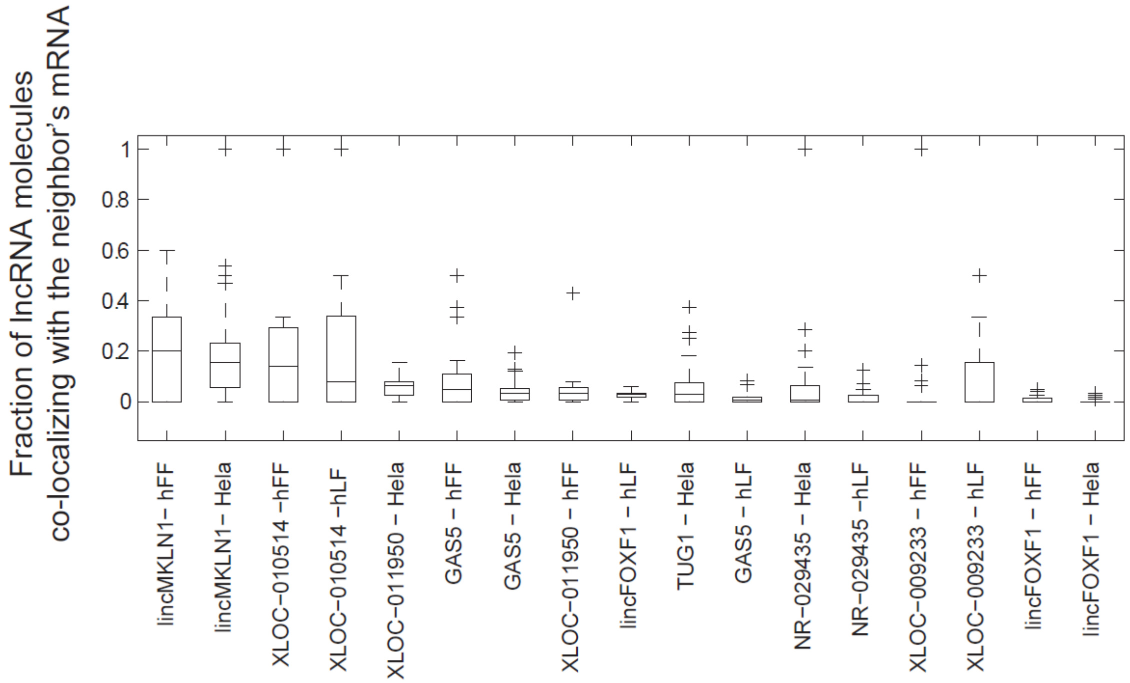
Supplementary Figure 2.10 | Correlation between mean (bulk) expression and single cell average Shown is a scatter plot comparing for each of the lncRNAs its expression level estimate in a bulk population from RNA-Seq (X-axis, $\log_2$(FPKM) from Cuffdiff2) and the mean number of mRNA molecule identified by single molecule RNA-FISH (Y-axis, $\log_2$ scale). Error bars correspond to a 95% confidence interval. Regression line Y = 4.42X+2.86 (blue line; Pearson ρ=0.52).



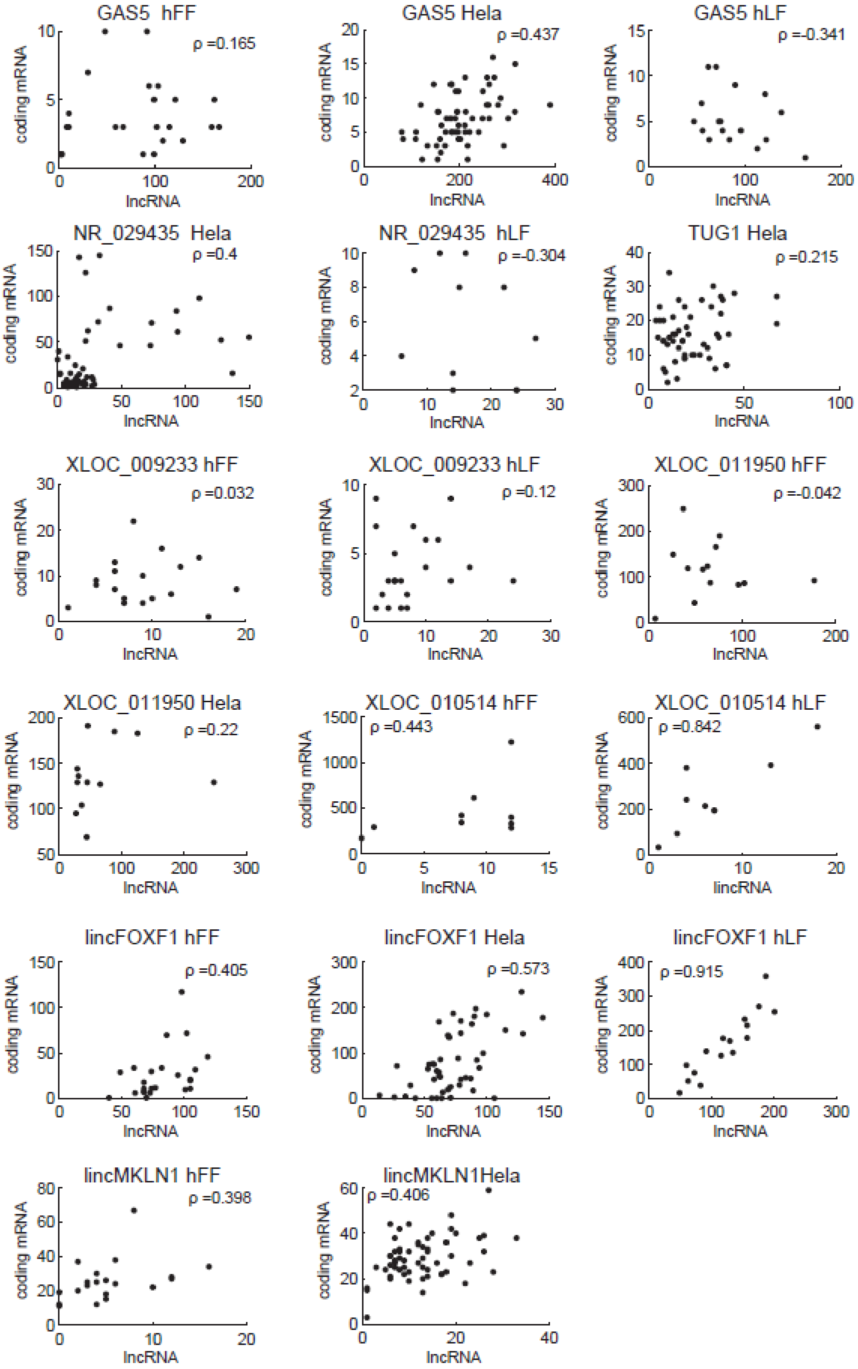
Supplementary Figure 2.11 | low RNA-Seq read coverage across the XLOC_003526 locus
A Sashimi plot [20] (as displayed by IGV [21]), describing XLOC_003526 transcript reconstruction by Cufflinks [22]. hLF RNA-Seq read alignment across the XLOC_003526 locus (top) from two replicates (red and blue coverage tracks) show the aggregate coverage of both spliced and un-spliced reads. Exon junctions spanned by the spliced reads are marked by bow curves. The number on top of the bow curve specifies how many spliced reads support this specific junction. There are only very few spliced reads in this data and these are all essential for reconstruction of the correct transcript (bottom).



Supplementary Figure 2.12 | Transcript abundance distributions across large numbers of cells Shown are the distributions of number of molecules per cell detected when imaging many (n>498) cells for each of four different lncRNAs. The lncRNA ID, cell type and sample size (n) are specified on top.

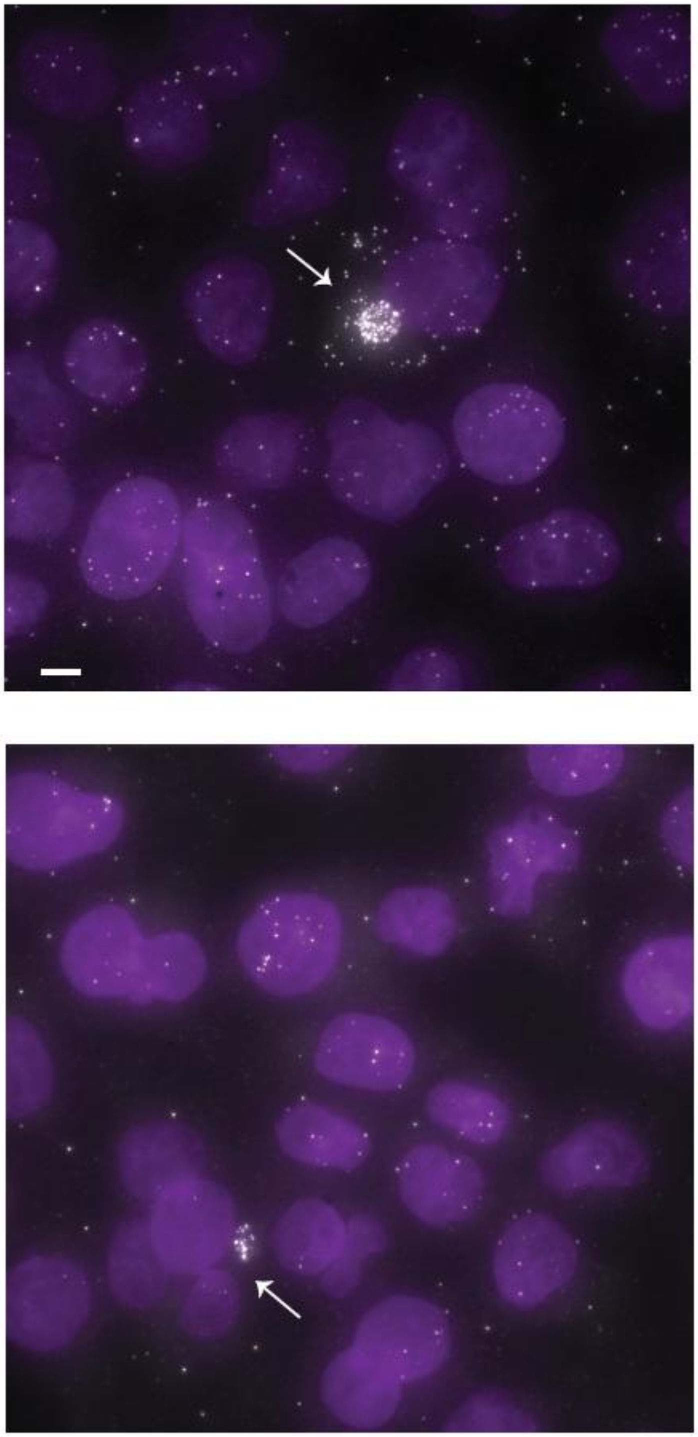


Supplementary Figure 2.13 | Divergently transcribed lncRNAs do not strongly co-localize with their neighboring mRNAs
For each lncRNA-cell pair (X axis), shown is a Box plot of the distribution of the fraction of divergent lncRNA molecules that co-localize with their neighboring mRNA molecule across the imaged cells.

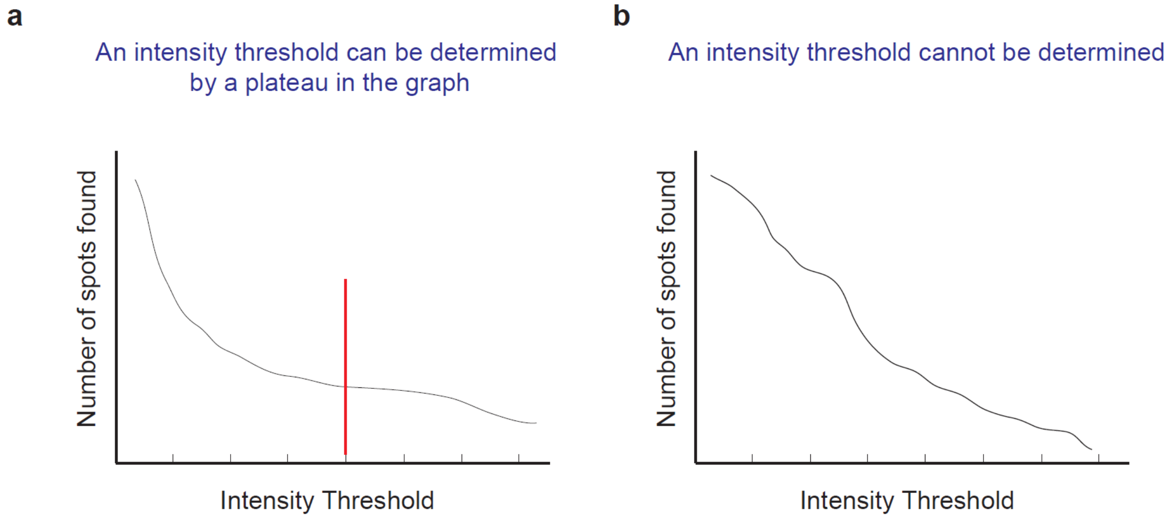


Supplementary Figure 2.14 | Correlation of expression between lncRNAs and their divergently transcribed mRNA neighbors

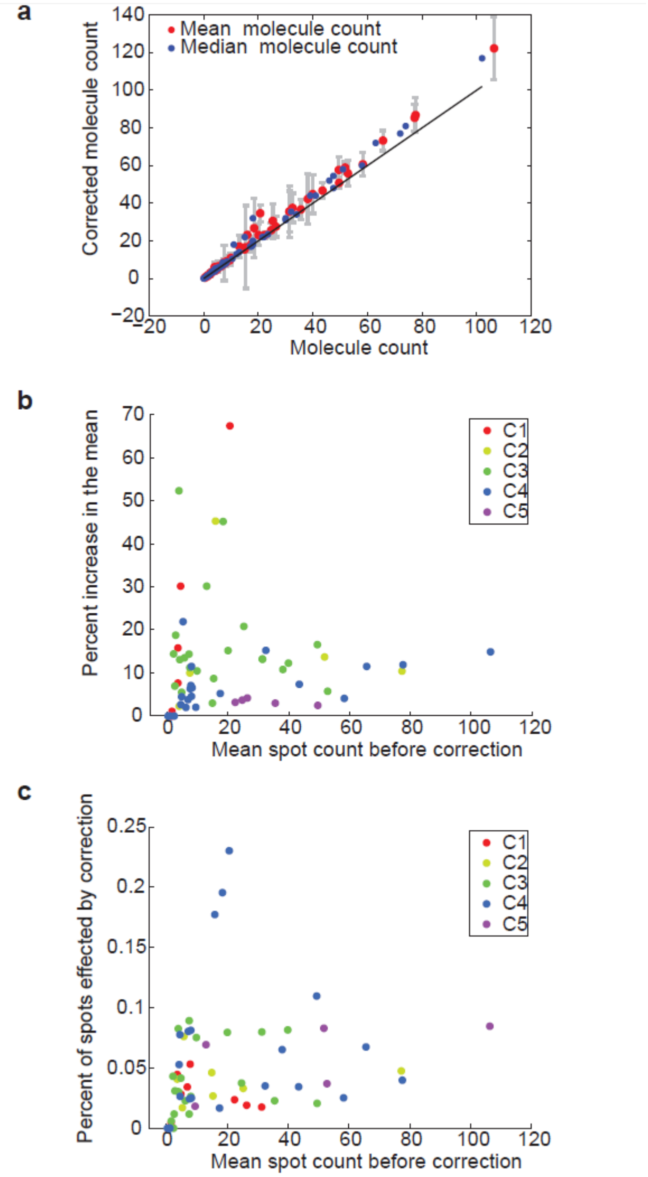
For each lncRNA:mRNA:cell-type combination included in the analysis for divergent transcripts shown are scatter plots of the relation between the expression level in each cell of a lncRNA (X-axis, molecule count) and its cognate, divergently transcribed, mRNA (Y axis, molecule count). Pearson correlation coefficient (ρ) after removal of outliers (**3.4 Methods**) is denoted on top.



Supplementary Figure 2.15 | Two rare daughter HeLa cells express NR_029436 highly
Shown are micrographs of NR_029436 with two rare, highly expressing, daughter HeLa cells (157 and 41 molecules per cell for the top and bottom images, respectively). Scale bar, 5 µm.

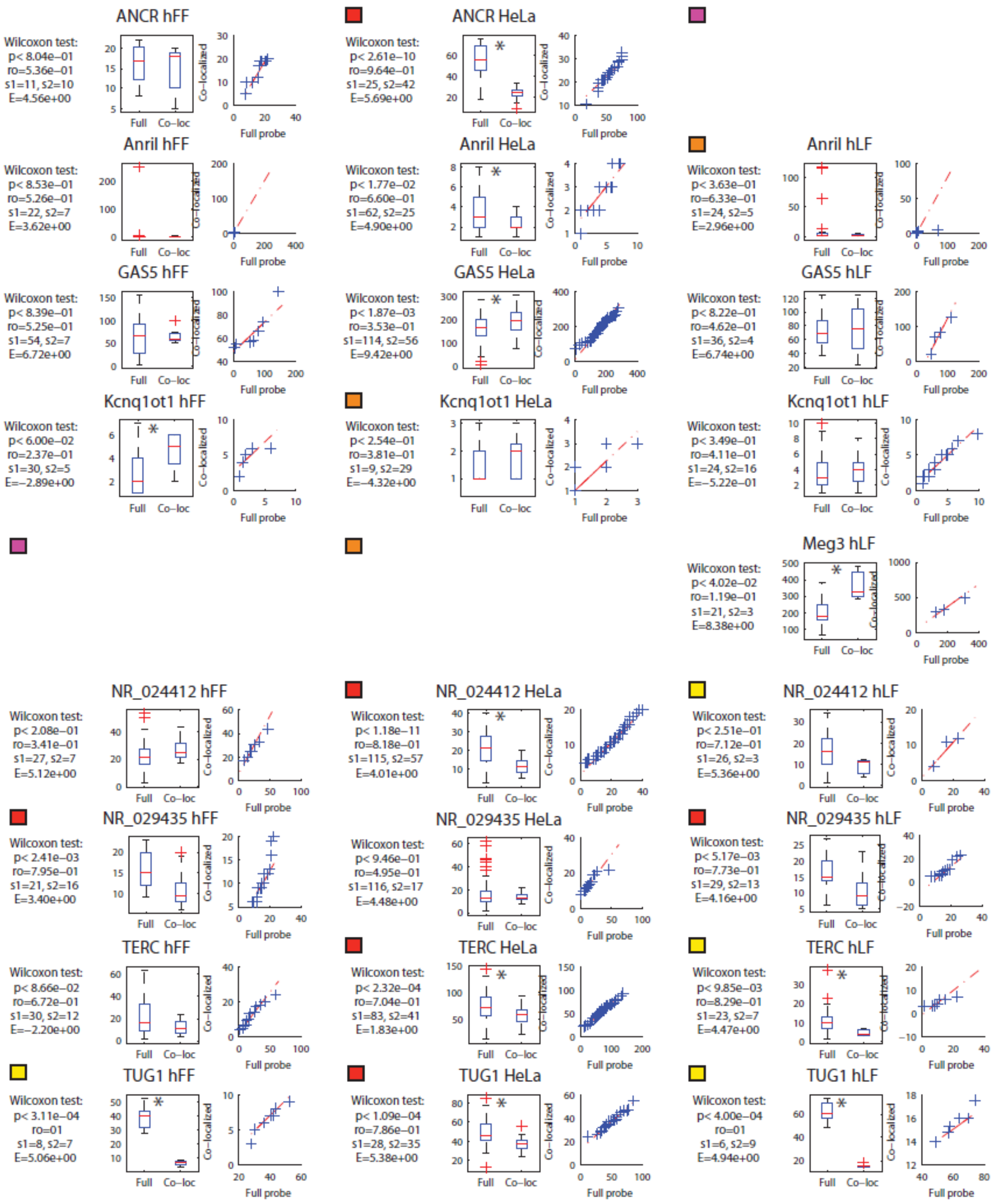


Supplementary Figure 2.16 | Determining a spot intensity threshold (X) to detect valid RNA spots Shown are illustration plots of the number of detected spots (Y axis) above each detection threshold (X-axis log2(intensity)) in **(a)** a case where a detection threshold can be determined by identifying a plateau in the graph; and **(b)** a case where a plateau is not identifiable. The latter is common for lncRNA images acquired with half of the probe set in the two-color co-localization assay.

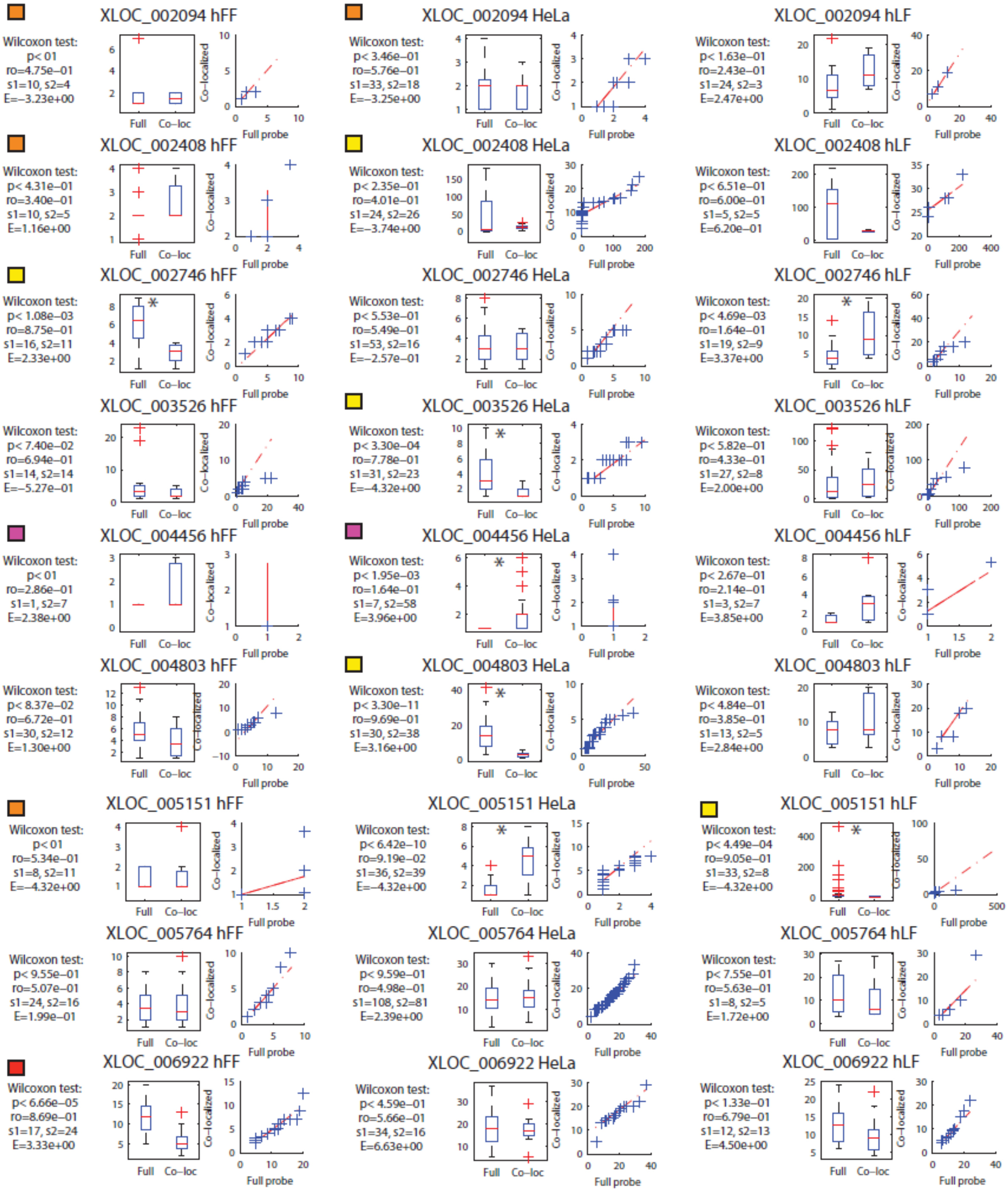


Supplementary Figure 2.17 | Single molecule count correction
Scatter plot of the mean (red) and median (blue) total spot count before (X-axis) and after (Y-axis) spot count correction. Error bars (gray) are the standard error of the mean. Black line Y=X. **(b-c)** Scatter plots of the mean spot count before correction (X axis) *vs*. either the percent increase in the mean after applying the correction (*b*) or the mean percent of spots that were affected by the correction across the cells (*c*). Data points are colored by classification to I-V class types, as in *Figure 2 (*noted here as C1-C5, respectively*)*. Top 10% (n=6) highest data points exhibit similar behavior and are excluded from this figure for resolution purposes. Each data point describes the population of cells expressing a specific lncRNA in a specific cell type.

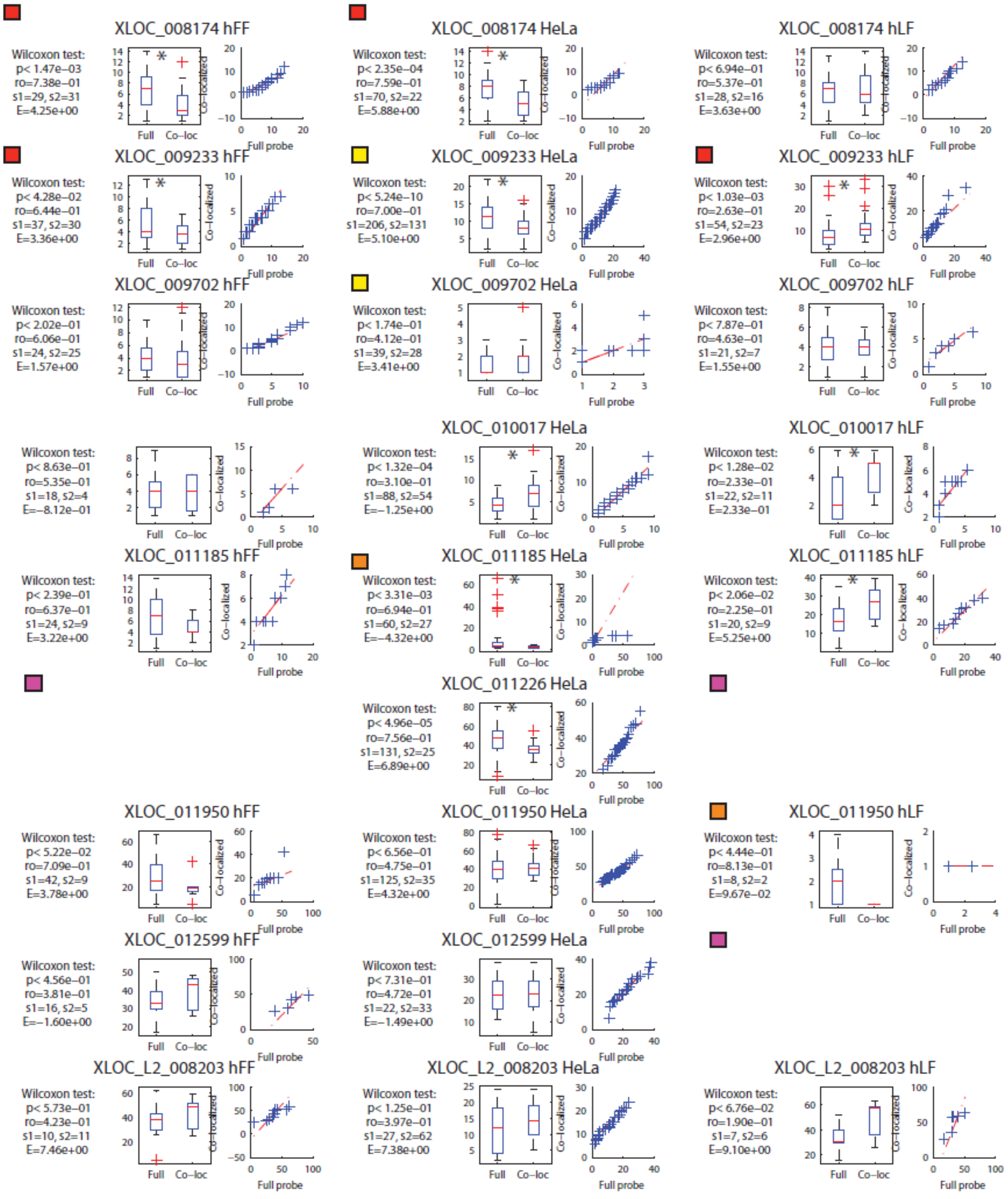
Supplementary Figure 2.18 | Probe set validation by two color co-localization. See following pages. Each row present three panels describing the same lncRNA in the three cell types assayed in this study. Each panel compares for that lincRNA:cell type pair, the count distribution as estimated by the full probe (full, s1) or by the spots that co-localize in the two-color assay (Co-loc, s2) by their: descriptive statistics (left), Box plots (middle), and quantile-quantile (QQ) plots (right). The descriptive statistics report from top to bottom: P-value of a Mann Whitney U test when comparing the two count distributions; r_o=U/s1*s2, where U is the ranksum statistic, and s_i is the sample size in each set, provides an estimate from 0 to 1 of the extent of overlap between the two distributions, where r_o=0.5 represents perfect similarity; sample set sizes for s1 and s2; and E, the log_2FPKM expression estimate from RNA-Seq. Some panels are marked by a colored box on the top left with the following color code: yellow- quantitatively or qualitatively inconsistent signal (classification 2.1-2.3); purple – insufficient data due to technical reasons (classification 2.4-2.7); orange – no signal in either the full probe set or the two-color assay (classification 3); red - valid after manual recovery. A borderline case with a pattern that is similar to a valid pattern of the gene in another cell type (classification 1.5). Panels where a box is absent are valid sets. Classifications are specified in **Supplementary Dataset 2.1**.



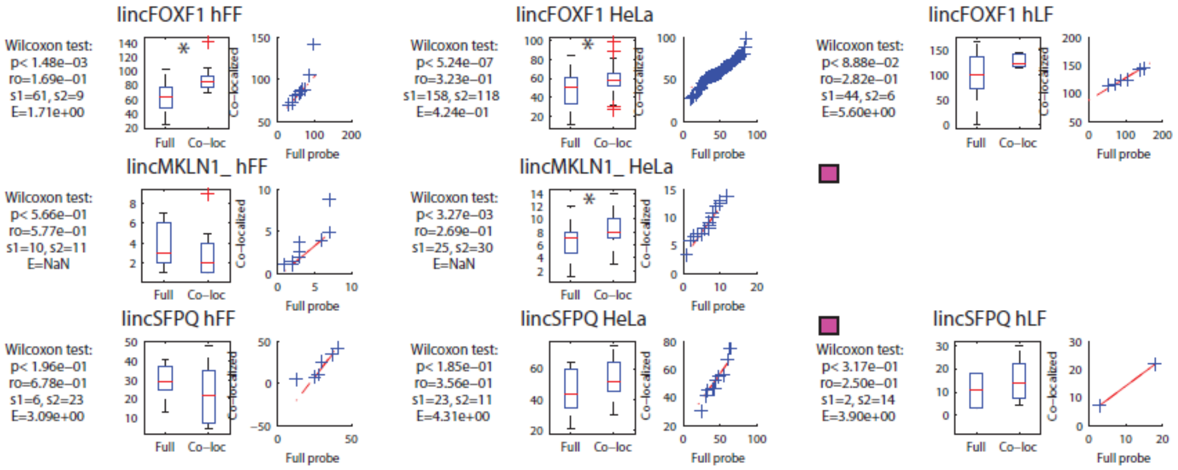
Supplementary Figure 2.18a | Probe set validation by two color co-localization. See legend above.



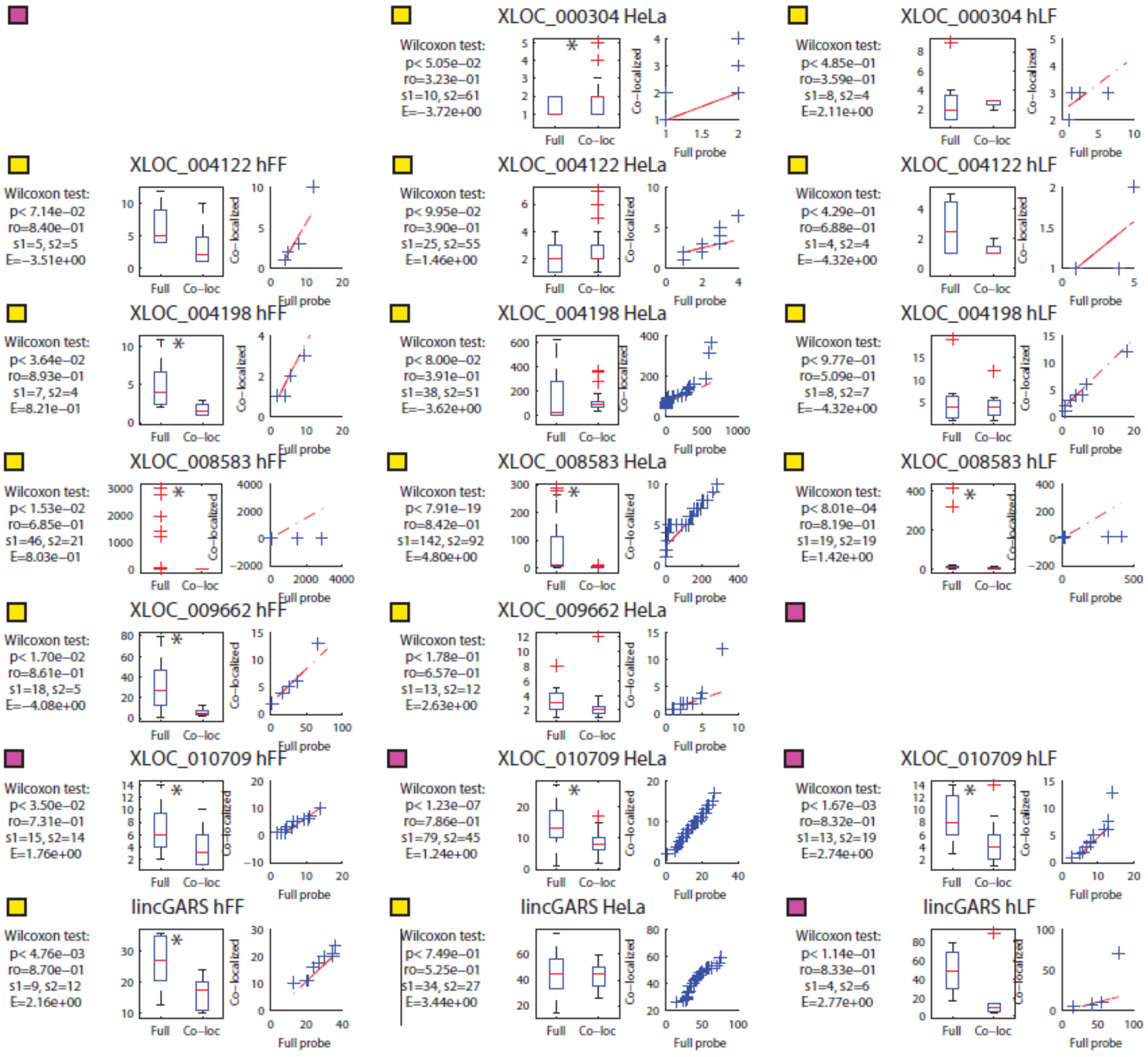
Supplementary Figure 2.18b | Probe set validation by two color co-localization. See legend above.



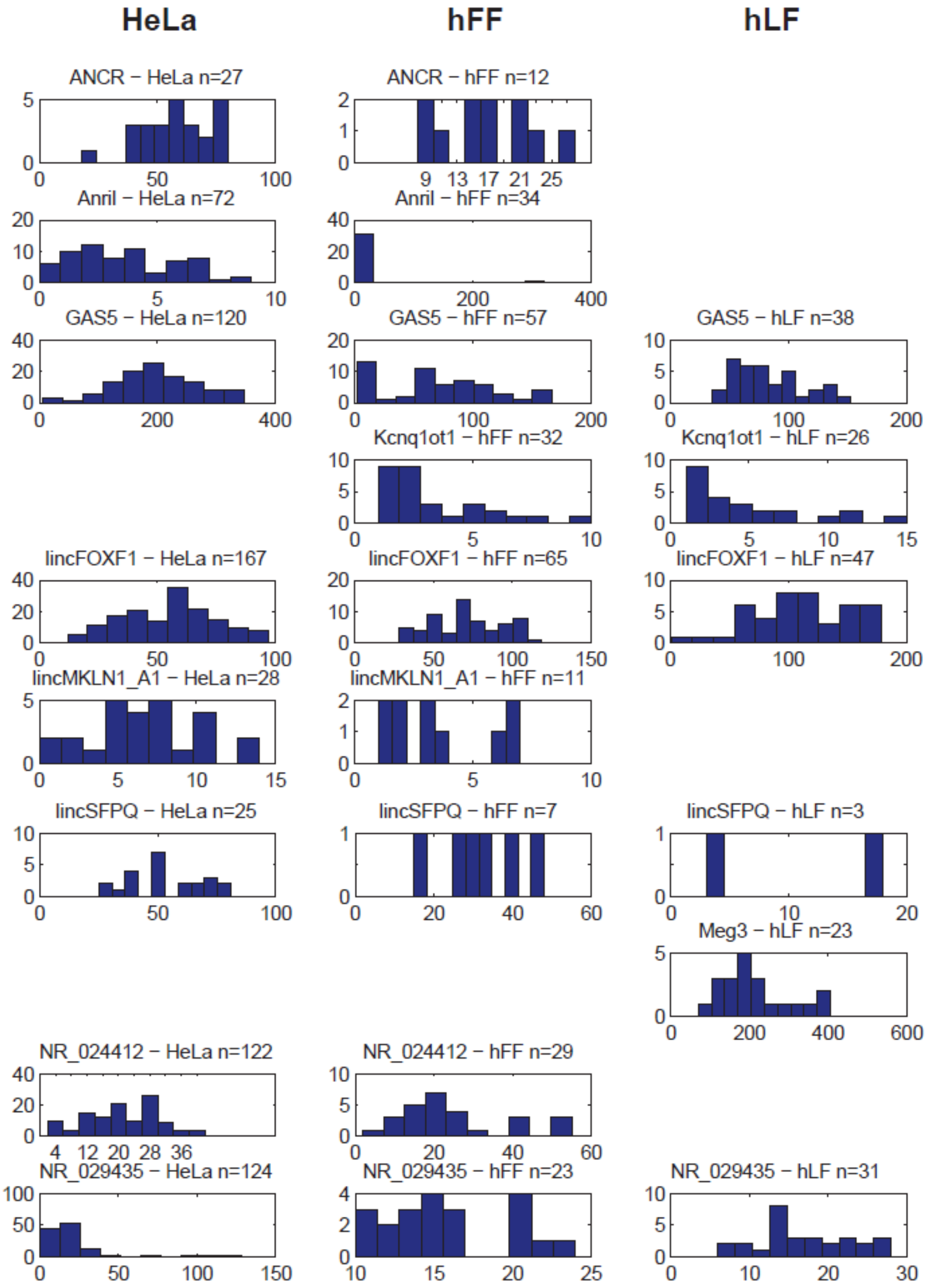
Supplementary Figure 2.18c | Probe set validation by two color co-localization. See legend above.



Supplementary Figure 2.18d | Probe set validation by two color co-localization. See legend above.

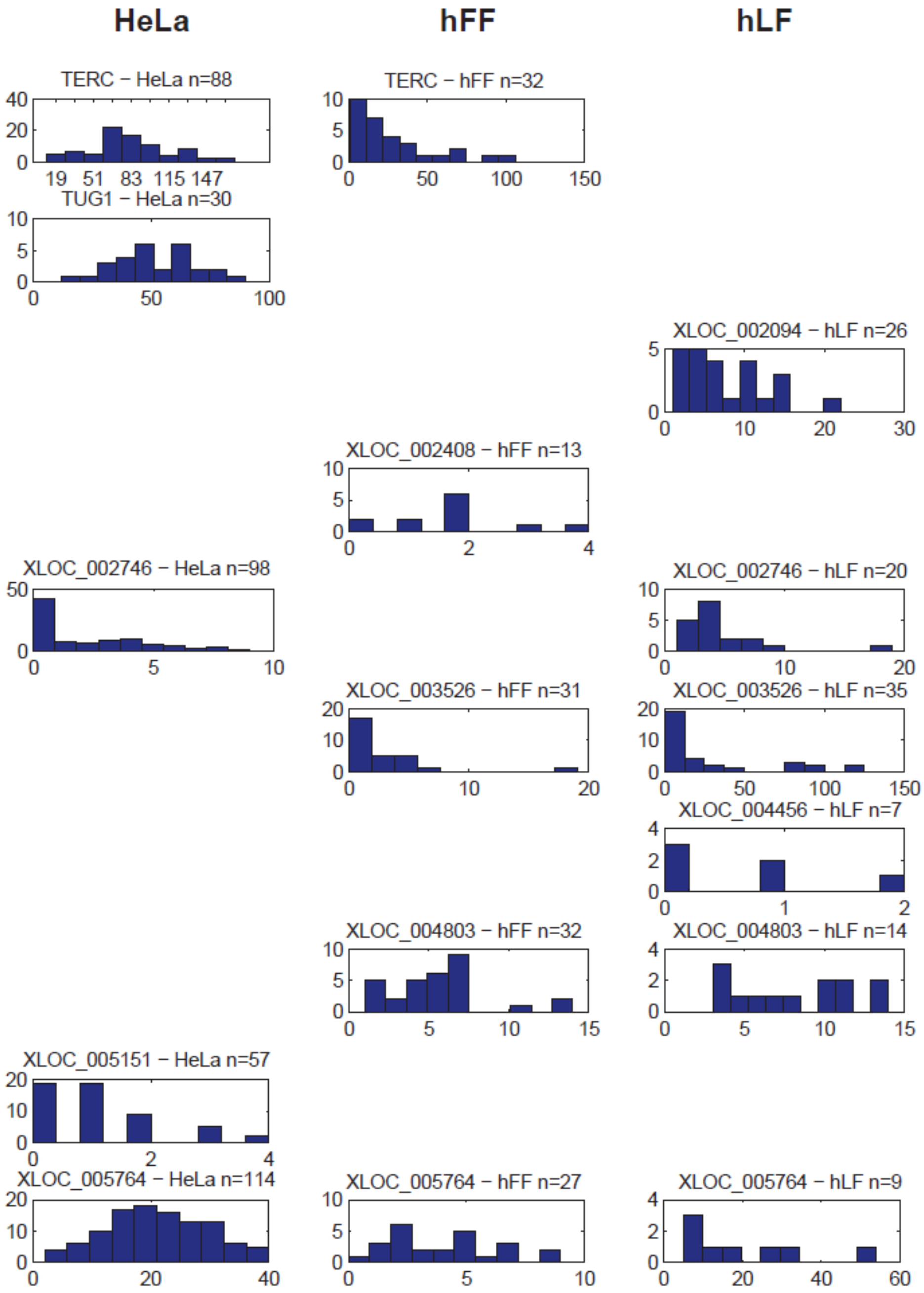


Supplementary Figure 2.18e | Probe set validation by two color co-localization. See legend above.

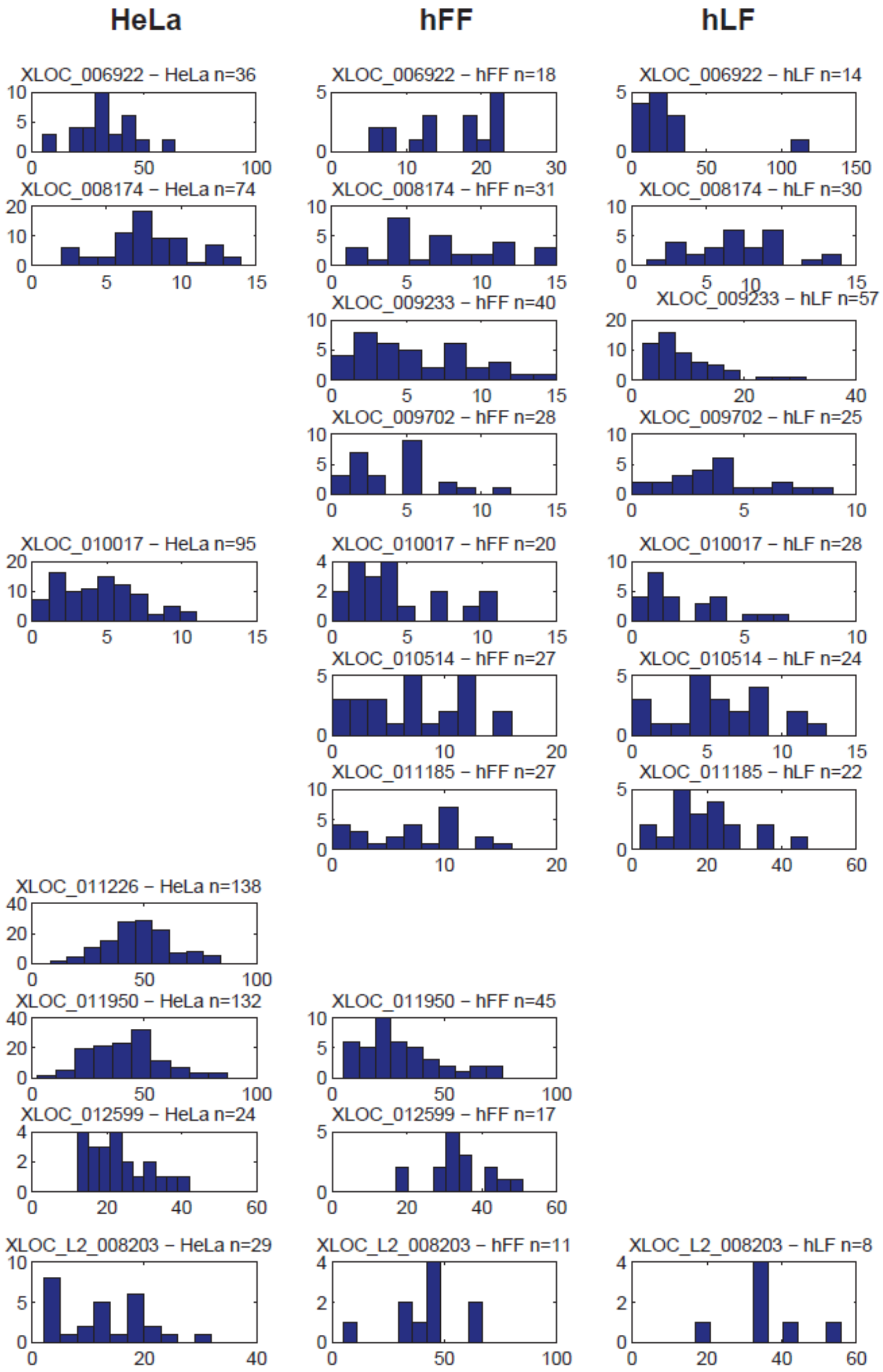


Supplementary Figure 2.19a | Molecule count distributions

For each lincRNA (row) and each cell type (column) from the validated lincRNA:cell type set shown is a distribution of the molecule counts . The lncRNA and cell type as well as the total number of cells are noted on top of each distribution.



Supplementary Figure 2.19b | Molecule count distributions
For each lincRNA (row) and each cell type (column) from the validated lincRNA:cell type set shown is a distribution of the molecule counts . The lncRNA and cell type as well as the total number of cells are noted on top of each distribution.



Supplementary Figure 2.19c | Molecule count distributions

For each lincRNA (row) and each cell type (column) from the validated lincRNA:cell type set shown is a distribution of the molecule counts . The lncRNA and cell type as well as the total number of cells are noted on top of each distribution.

Supplementary Table 2.1| lncRNA-FISH candidate set characteristics. “ValidSignal?” Indices : 1-yes 2-failed co-loc 3-nosignal.

Screen ID	Valid Signal?	#Oligos	Ortholog	Divergent	Number Of Exons	Transcript Len (bp)	Tissue Specificity
ANCR	1	18	1	0	3	855	0.087
ANRIL	1	30	1	0	19	3834	0.098
GAS5	1	14	1	1	12	632	0.097
h19	2	32	1	0	5	2308	0.129
KCNQ1OT1	1	48	1	0	1	91671	0.175
lincFOXF1	1	43	1	1	6	6662	0.195
lincGARS	2	45	0	1	7	3871	0.087
lincMKLN1	1	37	1	1	4	2948	0.107
lincSFPQ	1	40	1	0	3	2670	0.087
MALAT1	1	48	1	0	1	8707	0.086
MEG3	1	32	1	0	8	1722	0.162
MIAT	2	48	1	0	4	9982	0.105
NEAT1	1	48	1	0	1	3735	0.088
NR_024412	1	15	1	0	3	1127	0.089
NR_029435	1	32	1	1	4	2442	0.079
NRON	2	48	1	0	1	2730	0.287
TERC	1	10	1	0	1	451	0.236
TUG1	1	37	1	1	3	7104	0.079
XIST	1	48	1	0	6	19553	0.179
XLOC_000304	2	24	1	1	4	952	0.098
XLOC_002094	1	12	0	0	4	717	0.834
XLOC_002263	2	25	1	0	8	1619	0.096
XLOC_002408	1	14	1	0	3	1058	0.131
XLOC_002746	1	24	1	0	7	9656	0.112
XLOC_003526	1	22	0	0	5	990	0.417
XLOC_004122	2	14	1	1	3	1194	0.166
XLOC_004198	2	11	0	0	4	1771	0.394
XLOC_004456	1	18	1	1	4	730	0.091
XLOC_004803	1	48	0	0	2	3120	0.133
XLOC_005151	1	48	0	0	3	3047	0.654
XLOC_005764	1	48	1	0	4	2933	0.155
XLOC_006198	2	17	0	0	5	1091	0.317
XLOC_006922	1	24	1	0	3	1218	0.131
XLOC_008174	1	16	0	0	4	2886	0.083
XLOC_008583	2	12	0	0	3	988	0.102
XLOC_009233	1	11	1	1	3	673	0.091
XLOC_009447	3	24	1	0	2	3206	0.234
XLOC_009474	2	48	1	0	2	3710	0.087

XLOC_009662	2	44	0	0	3	2716	0.443
XLOC_009702	1	26	0	0	5	1895	0.228
XLOC_010017	1	30	0	0	7	2207	0.135
XLOC_010202	2	17	1	1	2	1454	0.084
XLOC_010263	3	32	1	1	3	6207	0.172
XLOC_010514	1	23	0	1	2	1985	0.162
XLOC_010556	2	25	1	0	5	1589	0.224
XLOC_010709	2	11	0	1	2	1083	0.133
XLOC_010853	3	40	1	0	4	863	0.360
XLOC_011185	1	30	0	0	3	4496	0.114
XLOC_011226	1	20	1	0	5	8974	0.084
XLOC_011264	2	25	0	0	3	2516	0.208
XLOC_011950	1	14	1	1	5	2768	0.123
XLOC_012187	2	20	1	0	2	3690	0.218
XLOC_012192	2	13	1	0	4	3169	0.090
XLOC_012197	3	12	0	0	4	1304	0.209
XLOC_012564	3	15	1	0	5	7748	0.499
XLOC_012599	1	32	1	0	5	2455	0.149
XLOC_012980	3	10	0	0	2	342	0.083
XLOC_013841	2	30	1	1	2	2794	0.234
XLOC_014160	3	11	1	0	8	1958	0.353
XLOC_L2_008203	1	16	1	0	2	793	0.099
XLOC_L2_010926	3	31	1	0	3	2603	0.083

Supplementary Table 2.2| "classic" lncRNAs imaged in this study. A previous reverence that included RNA-FISH of the corresponding gene, is specified if available.

Name	Previous publication with RNAFISH	Pubmed ID	Cell type in previous study	Consistent?
DANCR	NA			NA
ANRIL	Yap et al. Mol Cell. 2010 [23]	PMID: 20541999	prostate cancer cell	Signal in previous paper is not punctate. Single molecules cannot be defined
GAS5	Kino et al. Sci. Signal. 2010 [24]	PMID: 20124551	HeLa	yes
KCNQ1OT1	Terranova et al. Mol. Cell 2008 [25]	PMID: 18848501	mouse trophectodermal stem (TS) cells	yes
lincFOXF1	Khalil et al. PNAS 2009 [3]	PMID: 19571010	hFF	yes
lincMKLN1	Khalil et al. PNAS 2009[3]	PMID: 19571010	hFF	yes
lincSFPQ	Khalil et al. PNAS 2009[3]	PMID: 19571010	hFF	yes
MALAT1	Tripathi et al. Cell 2010 [26]	PMID: 20797886	HeLa	yes
MEG33	NA			NA
MIAT	Tsuiji et al. Genes to Cells 2011 [27]	PMID: 21463453	DF1 HeLa and chicken spinal cord cells	failed probe in our screen
NEAT1	Clemson et al. Mol Cell 2009 [28]	PMID: 19217333	HeLa	yes
TERC	Zhu et al. 2004 [29]	PMID: 14528011	HeLa HFF-Bjs	Overall yes
TUG1	Yang et al. cell 2011 [30]; Khalil et al. PNAS 2009 [3]	PMID: 22078878 ;PMID: 19571010	HeLa;hFF	yes; yes
XIST	Clemson et al. JCB 1996 [31]; Brown et al. Cell 1992 [32]	PMID: 8636206; PMID: 1423611	normal human fibrblasts	yes
lincGARS	NA			no signal in our study
NRON	NA			no signal in our study
H19	Jouvenot et al. Curr. Bio. 2009 [33]	PMID: 10531031	mouse :E13.5 fetal liver	no signal in our study

Supplementary Table 2.3| Probe set two-color co-localization validation results. **(a)** Validation score index. **(b)** Validation score results.

(a) Validation score index.

Subclass number	Class type
1	Valid
1.5	Valid after manual recovery: a borderline case with a pattern similar to a valid pattern of the gene in another cell type
2.1	Invalid: qualitative due to foci found in one channel and not the other
2.2	Invalid : quantitative
2.3	Invalid: qualitative; a conservative classification of a borderline case in manual examination identifies clear spots in one channel in the two color assay.
2.4	Not enough data
2.5	Single color image not available – technical reasons
2.6	2 color image not available but signal is very consistent with the gene's signal in a valid cell type
2.7	2 color image not available due to technical reasons
3	No signal in single color image

(b) Validation score results.

screenID	**FinalStatus**	**HeLa**	**hFF**	**hLF**
ANCR	1.0	1.5	1	2.6
Anril	1.0	1.5	1	3
GAS5	1.0	1	1	1
KCNQIOT1	1.0	3	1	1
lincFOXF1	1.0	1	1	1
lincMKLN1_A1	1.0	1	1	2.7
lincSFPQ	1.0	1	1	2.4
MALAT1	1.0	1	1	1
MEG3	1.0	3	2.6	1
NEAT1	1.0	1	2.6	2.6
NR_024412	1.0	1.5	1	2.3
NR_029435	1.0	1	1.5	1.5
TERC	1.0	1.5	1	2.2
TUG1	1.0	1.5	2.2	2.2
XIST	1.0	3	2.5	1
XLOC_002094	1.0	3	3	1
XLOC_002408	1.0	2.3	3	1.5
XLOC_002746	1.0	1	2.2	1

XLOC_003526	1.0	2.2	1	1
XLOC_004456	1.0	2.4	2.4	1
XLOC_004803	1.0	2.2	1	1
XLOC_005151	1.0	1	3	2.2
XLOC_005764	1.0	1	1	1
XLOC_006922	1.0	1	1.5	1.5
XLOC_008174	1.0	1.5	1.5	1
XLOC_009233	1.0	2.2	1.5	1
XLOC_009702	1.0	2.1	1	1
XLOC_010017	1.0	1	1	1
XLOC_010514	1.0	3	1	1
XLOC_011185	1.0	3	1	1
XLOC_011226	1.0	1.5	2.6	2.6
XLOC_011950	1.0	1	1	3
XLOC_012599	1.0	1	1	2.6
XLOC_L2_008203	1.0	1	1	1
h19	2.0	2	3	3
lincGARS	2.0	2.3	2.2	2.4
MIAT	2.0	2.2	2.2	2.4
NRON	2.0	2	3	3
XLOC_000304	2.0	2.3	2.5	2.3
XLOC_002263	2.0	2.1	3	3
XLOC_004122	2.0	2.3	2.3	2.3
XLOC_004198	2.0	2.3	2.2	2.3
XLOC_006198	2.0	2.7	2.7	2.7
XLOC_008583	2.0	2.2	2.2	2.2
XLOC_009474	2.0	2.1	3	3
XLOC_009662	2.0	2.3	2.2	2.5
XLOC_010202	2.0	2.1	3	3
XLOC_010556	2.0	2.1	3	3
XLOC_011264	2.0	2.1	2.7	2.7
XLOC_012187	2.0	3	3	2.1
XLOC_012192	2.0	2.1	3	3
XLOC_013841	2.0	2.1	3	3
XLOC_009447	3.0	3	3	3
XLOC_010263	3.0	3	3	3
XLOC_010853	3.0	3	3	3
XLOC_012197	3.0	3	3	3
XLOC_012564	3.0	3	3	3
XLOC_012980	3.0	3	3	3
XLOC_014160	3.0	3	3	3
XLOC_l2_010926	3.0	3	3	3
XLOC_010709	2.0	2.7	2.7	2.7

Supplementary Table 2.4| Mitotic cells analysis.

lincRNA name	Localization pattern	Cell type	# of mitotic cells	# of mitotic cells with foci on chromosomes
XLOC_005764	II	HeLa	7	0
XLOC_005151	I	HeLa	12	0
XLOC_006922	I	HeLa	6	0
XIST	I	hLF	18	0
ANRIL	I	HeLa	15	5
MEG3	II	HeLa	10	0
XLOC_003526	III	hLF	12	0

Supplementary Table 2.5| Correlation analysis of lncRNAs with CCNA2. Presented are lncRNA that had a significant score in at least one of the test described in 3.4 Methods. Pearson ρ specifies the correlation between the lncRNA and CCNA2 that are imaged simultaneously. The mean number of spots per cell is specified for the lncRNA and CCNA2. Kolmogorov-Smirnov (KS) test p-value is calculated between the distributions of lncRNA counts after splitting G1 and G2/S/M by a threshold on CCNA2 counts. (#) refers to the numbers of lncRNA cells in each sub population (G1 or G2/S/M). Regression P-value is the significance of CCNA2 as a predictor in a linear regression to predict the lncRNA level using CCNA2 and cell area as predictors. Cases that were finally considered as significant are marked red.

Name	cell	#Cellls	Pearson ρ	lncRNA spots	CCNA2 spots	KS P_val	# G1	# G2/S/M	Regression Pval
lincSFPQ	Hela	25	0.68	33.10	147.00	NA	0	25	2.31E-03
XLOC_011226	Hela	137	0.66	25.90	105.00	1.73E-03	13	124	1.00E-10
TUG1	Hela	30	0.58	29.70	82.90	4.18E-01	8	22	
XLOC_003526	hLF	24	0.55	6.88	6.50	1.36E-01	23	1	
XLOC_006922	hLF	11	0.53	10.00	8.91	8.70E-01	10	1	
GAS5	Hela	117	0.46	118.00	79.80	4.10E-03	20	97	7.09E-01
XLOC_011950	Hela	126	0.45	23.00	80.60	4.62E-01	35	91	
lincSFPQ	hFF	7	0.40	20.10	72.30	1.58E-01	2	5	
NR_024412	Hela	122	0.37	12.20	66.50	1.40E-01	54	68	
MEG3	hLF	20	0.35	136.00	5.75	NA	20	0	
NR_029435	Imr90	23	-0.36	9.35	15.00	9.40E-02	18	5	
XLOC_011185	hFF	20	-0.44	5.20	16.40	6.30E-02	14	6	5.02E-02

References for Supplementary Material

1. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
2. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat. Protocols, 2008. **4**(1): p. 44-57.
3. 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, 2009. **106**(28): p. 11667-72.
4. Garber, M., et al., *Identifying novel constrained elements by exploiting biased substitution patterns.* Bioinformatics, 2009. **25**(12): p. i54-62.
5. Gudmundsson, J., et al., *Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations.* Nat Genet, 2009. **41**(4): p. 460-464.
6. Heid, I.M., et al., *Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.* Nat Genet, 2010. **42**(11): p. 949-960.
7. Waterworth, D.M., et al., *Genetic variants influencing circulating lipid levels and risk of coronary artery disease.* Arterioscler Thromb Vasc Biol, 2010. **30**(11): p. 2264-76.
8. Scott, L.J., et al., *Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci U S A*, 2009. **106**(18): p. 7501-7506.
9. Takata, R., et al., *Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population.* Nat Genet, 2010. **42**(9): p. 751-4.
10. Pruitt, K., T. Tatusova, and D. Maglott, *Chapter 18, The Reference Sequence (RefSeq) Project.* , in *The NCBI handbook [Internet]*.2002, National Library of Medicine (US), National Center for Biotechnology Information: Bethesda (MD).
11. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.
12. Lestrade, L. and M.J. Weber, *snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.* Nucleic Acids Res, 2006. **34**(Database issue): p. D158-62.
13. Flicek, P., et al., *Ensembl 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D800-6.
14. Wilming, L.G., et al., *The vertebrate genome annotation (Vega) database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D753-60.
15. Zhang, Z., et al., *PseudoPipe: an automated pseudogene identification pipeline.* Bioinformatics, 2006. **22**(12): p. 1437-9.
16. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-46.
17. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Research, 2008. **36**(suppl 1): p. D773-D779.
18. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.
19. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature Biotechnology, 2013. **31**(1): p. 46-53.
20. Katz, Y.W., Eric T.; Silterra, Jacob; Schwartz, Schraga; Wong, Bang; Mesirov, Jill P.; Airoldi, Edoardo M.; Burge, Christopher B., *Sashimi plots: Quantitative visualization of RNA sequencing read alignments*, 2013, ARXIV: eprint arXiv:1306.3466. p. 2013arXiv1306.3466K.
21. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotech, 2011. **29**(1): p. 24-26.
22. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotech, 2010. **28**(5): p. 511-515.
23. Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.* Molecular Cell, 2010. **38**(5): p. 662-74.
24. Kino, T., et al., *Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor.* Science signaling, 2010. **3**(107): p. ra8.
25. Terranova, R., et al., *Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos.* Developmental cell, 2008. **15**(5): p. 668-79.
26. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Molecular Cell, 2010. **39**(6): p. 925-38.

27. Tsuiji, H., et al., *Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1.* Genes to cells : devoted to molecular & cellular mechanisms, 2011. **16**(5): p. 479-90.
28. Clemson, C.M., et al., *An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.* Molecular Cell, 2009. **33**(6): p. 717-26.
29. Zhu, Y., et al., *Telomerase RNA accumulates in Cajal bodies in human cancer cells.* Molecular biology of the cell, 2004. **15**(1): p. 81-90.
30. Yang, L., et al., *ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs.* Cell, 2011. **147**(4): p. 773-88.
31. Clemson, C.M., et al., *XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure.* The Journal of cell biology, 1996. **132**(3): p. 259-75.
32. 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, 1992. **71**(3): p. 527-42.
33. Jouvenot, Y., et al., *Biallelic transcription of Igf2 and H19 in individual cells suggests a post-transcriptional contribution to genomic imprinting.* Current biology : CB, 1999. **9**(20): p. 1199-202.