CHAPTER 1

Introduction and scope of the thesis

Melanie Winkle
1 Introduction

1.1 Composition of the human genome

For a long time it has been a puzzling fact why the vast majority of the human genome does not encode for proteins; the main functional entities of cells. As shown in Figure 1A, coding exons make up less than two percent of the genome. A much larger part is occupied by noncoding sequences, either located within genes (i.e. intronic) and regulatory elements (ca. 24%) or within intergenic spaces (ca. 18%). The largest part of the genome (ca. 45%) is assigned to variable classes of transposable elements (TEs), mobile DNA sequences inserted into and moving within the genome, while around 11% contains simple repetitive sequence and duplications (estimations based on1, 2). However, the line between noncoding sequences and TEs or repeats is rather thin, as noncoding loci often contain these elements3. In addition, both repeats4, 5 and TEs3, 6-8 may act as functional entities in noncoding genes. Thus, although human cells go through great lengths to replicate their entire genome as fault-free as possible, a very minor part serves the production of proteins. This clearly suggests that there is a substantial piece of information missing in the jigsaw puzzle of the human genome.

1.2 The dawn of long noncoding RNA research

Major advances in sequencing technologies now provide a much more intricate view of the human transcriptome, showing that about two-thirds of the genome is actively transcribed in at least one cell type or tissue9. The presence of tightly regulated, independent transcriptional units lacking coding potential (Box 1)9-11 was followed by a plethora of publications discussing the significance of the noncoding genome for human life12-14. The identification and mapping of ten-thousands of long noncoding (Inc)RNA genes2, 15-18 eventually lead to the emergence of an entire research field. Accordingly, the amount of studies addressing IncRNA function has risen from a single publication in 200619 to over 7,000 in 2017 (Figure 1B). Of note, the highly abundant IncRNA XIST involved in X chromosome inactivation was identified as early as 199120. The experimental support required for the hypothesis that IncRNAs have important functions was delivered by the first murine knockout study in 2013. For five of 18 IncRNAs knockout induced developmental errors, peri- or postnatal lethality21. Thus, the central dogma of RNA functioning as a mere intermediate between DNA and protein has been rattled since the discovery of IncRNAs as functional entities.

Box 1 The CODING POTENTIAL of a transcript can be calculated based on the quality and length of putative open reading frames (ORFs) as well as the similarity of the putative protein product to known proteins. To exclude the production of peptides from small ORFs that may be present within noncoding genes, ribosome sequencing data as well as mass-spectrometry data (peptide sequences) can be considered.
INTRODUCTION AND SCOPE OF THE THESIS

1.3 LncRNA (R)evolution

Doubts about the functionality of lncRNAs are mainly based on the limited evolutionary conservation present in lncRNA as compared to coding genes. However, purifying selection does occur in approx. 5% of the genome, a much larger fraction than occupied by coding sequence (<2%). Closer inspection of lncRNA loci indeed identified evolutionary constraint, albeit mostly restricted to lncRNA promoters, splice sites and shorter stretches of primary sequence. The definition of lncRNA orthologs accordingly applies less stringent sequence alignment cutoffs and may additionally depend on positional conservation.

Box 2 For assessment of evolutionary conservation of lncRNAs the following distinctions are made:

- **LncRNA orthologs** are based on sequence similarity (~30% of exonic sequence can be aligned).
- **Syntenic lncRNA orthologs** are based on genomic association (synteny) and thus contain conserved lncRNA-mRNA pairs.
- **Transcriptional conservation** is present when the orthologous sequence also produces an RNA transcript. Orthologs that do not show transcriptional conservation are also referred to as ‘pseudoconserved’.
- **Regulatory conservation** is present when the lncRNA ortholog shows the same tissue specificity.
conservation (i.e. syntenic IncRNA orthologs or IncRNA-mRNA pairs). Furthermore, transcriptional conservation needs to be considered, i.e. whether the orthologous sequence is also actively transcribed. Considering these factors, protein coding genes are highly conserved: Comparison of human coding genes with chimpanzee and rat, showed high sequence (99 and 93%, respectively) as well as transcriptional conservation (92 and 90%, respectively)\(^24\). This confirms that the coding genomic content is very static across large eras of evolution (also known as the G-value paradox). Interestingly, the amount of noncoding sequence (i.e. the ratio of noncoding sequence over total genome size) shows a linear correlation with the biological complexity of organisms\(^12, 25\). Accordingly, the conservation of IncRNAs decreases more rapidly with evolutionary distance: Of a set of 1,900 human IncRNAs 98 and 54% had orthologs in chimpanzee and rat, respectively, of which 80 and 35% also show transcriptional conservation\(^24\). In other words, an estimated 70 to 80% of IncRNA genes have developed at a late point in evolution and are largely specific to the primate lineage\(^26-28\). Interestingly, of these primate-IncRNAs approx. 20% are brain-specific\(^26\). Novel IncRNA genes are thought to originate from protein coding gene copies or pseudogenes that have lost their coding potential\(^28, 29\), or from the insertion and adaption of TEs\(^30\). TEs are able to acquire functional significance, e.g. as regulatory elements or by shaping splicing patterns, thereby creating novel transcriptional units with tissue specific expression\(^6, 28, 30\). These observations hint at a rapid turnover in IncRNA evolution and raise the question when these transcripts have undergone their ‘revolution’, i.e. at which point they have gained actual functional significance. The majority of IncRNAs that are conserved across multiple mammalian species, i.e. (syntenic) IncRNA orthologs, also show regulatory conservation (see Box 2), i.e. expression in the same tissues\(^24\). Of all IncRNAs initially mapped across 20 human tissues\(^16\), nearly 80% showed high tissue specificity. More than half of these were specifically expressed in testis\(^16, 26\), a tissue with a highly permissive transcriptional landscape due to an open chromatin conformation\(^31\). Evolving IncRNA genes might respond to weak stimuli in such a permissive environment, while functional IncRNAs are more likely to be under tight regulation. In accordance with this, ‘younger’ IncRNA genes are more often testis specific than ‘older’ IncRNA genes, and brain specific IncRNAs show the highest grade of (regulatory) conservation\(^26\). Together this shows that the noncoding part of the human genome is much more dynamic and subject to continuative evolutionary turnover.

1.4 LncRNA classes

LncRNAs are classified according to their size (>200 nucleotides) and their genomic position in relation to protein coding genes or regulatory elements. The major classes defined to date (Figure 1C) are intergenic, antisense, bidirectional, intronic and overlapping IncRNAs, as well as RNAs transcribed from enhancer (eRNAs) or promoter regions (pRNAs). In addition, a growing number of circular (circ)RNA transcripts generated by back-splicing reactions have been identified\(^32\). The majority of IncRNAs identified so far are intergenic or bidirectional IncRNAs (>10,000 transcripts), antisense
INTRODUCTION AND SCOPE OF THE THESIS

(−5,000) and intronic transcripts (−600), with just a few known overlapping lncRNAs (<200) (estimations based on2). However, these numbers likely also reflect the difficulty of distinguishing functional intronic, overlapping and promoter-derived lncRNAs from primary mRNA transcripts. Active transcription is often observed at enhancers producing enhancer (e)RNAs. Accordingly a significant number of lncRNA loci are marked by the typical enhancer chromatin mark H3K4me1. Furthermore, about 4% of the lncRNA loci encode small RNAs such as micro (mi)RNAs or small nucleolar (sno) RNAs within their introns (compared to 7% of coding genes)2.

1.5 LncRNA characteristics

Some main characteristics distinguish lncRNAs from mRNA transcripts: (1) a generally lower expression level2, 16, (2) a lower number of exons (mostly two to three)2, 16, (3) an expression pattern that is more often restricted to a single cell type or tissue2, 33, and (4) more frequent nuclear localization with a low amount or complete lack of transcripts in the cytoplasm34. Similar to mRNAs, lncRNAs are (alternatively) spliced and stabilized through 5’cap structures and polyadenylation16. Exceptions to this are most eRNA transcripts (very short lived) and circRNAs (extremely stable structure). Moreover, novel triple-helix and snoRNA-mediated stabilization mechanisms of non-polyadenylated lncRNAs have been described35, indicating that more yet unidentified mechanisms may be at hand. Chemical modification of lncRNA transcripts can potentially add flexibility to the system, as is observed in protein coding transcripts. The most prevalent types of RNA modifications, 5-methylcytosine (m5C), N6-methyladenosine (m6A), pseudouridine (Ψ), and adenosine-to-inosine editing (A-to-I) indeed occur in thousands of lncRNA transcripts16, 37. Though few studies have addressed the functional significance of these modifications, involvements in e.g. secondary lncRNA structure and/or binding interactions with proteins have been suggested (reviewed in38). In addition, the functions of these modifications in lncRNAs may be similar to those described for mRNAs, e.g. regulating stability, nuclear export or miRNA targeting (FIGURE 1D)37, 39.

1.6 LncRNA functions

RNA molecules have the capacity to create complex molecular interaction networks. Linear lncRNAs can bind to DNA or RNA based on sequence homology thereby forming triplex structures40, while lncRNAs folded into secondary structures may form various complex interactions with proteins and/or RNA/DNA41. Together, these factors contribute to a complicated jigsaw puzzle of possible interactions and functionalities (FIGURE 1E), including transcriptional and post-transcriptional control mechanisms (reviewed in42, 43). Interaction with DNA can occur not only through base-pairing with DNA, but also through duplex formation with nascent RNA transcripts, by interaction with DNA-bound proteins, chromatin marks or by transcriptional tethering (i.e. the lncRNA transcript being functional during its transcription)44-46. This can lead to gene expression regulation in cis or trans through e.g. steric hindrance or recruitment of
transcription factors and/or co-factors or via recruitment of epigenetic activators and repressors. Furthermore, various post-transcriptional mechanisms can be at hand, affecting mRNAs (splicing, stability, localization, availability), proteins (stability, localization, availability, activation/repression) and protein complexes (scaffolding) or microRNA availability. Thus, the flexibility and interaction capacity that is intrinsic to lncRNAs creates a broad functional repertoire. Furthermore, RNA secondary structures and chemical modifications can add additional diversity and fine-tuning to their functionality.

**FIGURE 2** Schematic of normal B-cell maturation and lymphomagenesis. (*left to right*) Immature naïve B cells from the bone marrow enter the secondary lymphoid organs. B cells are activated upon encountering antigens and receiving activating stimuli from T helper cells, resulting in the formation of germinal centers (GCs). In the dark zone of the GC, B cells undergo clonal expansion and somatic hypermutation to diversify the B cell receptor repertoire. In the light zone, B cells are selected for antigen affinity. Apoptosis is induced in cells with sub-optimal antigen affinity, while high affinity B cells may differentiate into effector cells (i.e. memory B or antibody secreting plasma cells) or re-enter the dark zone for further expansion and germinal center maintenance. (*grey arrows*) B-cell lymphomas most commonly arise from GC B or post-GC B cells. ABC DLBCL – activated B cell-like diffuse large B-cell lymphoma; BL – Burkitt lymphoma; CLL – chronic lymphocytic leukemia; FL – Follicular lymphoma; GC B DLBCL – germinal center B cell-like diffuse large B-cell lymphoma; HL – Hodgkin lymphoma; PMBL – primary mediastinal B-cell lymphoma. (*bottom*) Major molecular components involved in the B cell maturation process with special emphasis on Myc. Myc (*asterisk*) is expressed upon T cell-activation of B cells, during the early stages of germinal center formation, and in subsets of light zone B cells, while otherwise being repressed by BCL6 or BLIMP1.
1.7 B-cell lymphoma and Myc

B-cell lymphoma malignancies show an annual incidence of approx. 20 new cases per 100,000 persons. The most common subtypes include diffuse large B-cell lymphoma (DLBCL), Hodgkin lymphoma (HL), follicular lymphoma (FL) and chronic lymphocytic leukemia (CLL). Less common B-cell lymphomas include mantle cell lymphoma (MCL), marginal zone lymphoma (MZL), primary mediastinal B-cell lymphoma (PMBL) and Burkitt lymphoma (BL). Despite significant advancements in treatment strategies, lymphoma-associated deaths are still high with 5-year survival rates between 30-90% depending on lymphoma subtype and stage at diagnosis. Furthermore, successfully treated patients may suffer from secondary treatment effects including cardiac disease, secondary malignancies and infertility.

The majority of B-cell lymphomas arise during or after the germinal center reaction (FIGURE 2). GC B cells are especially susceptible to malignant transformation due to their high proliferative rate and the genomic hypermutation machinery active during B-cell receptor diversification. Somatic hypermutation and class-switch recombination require the generation of double strand breaks within the immunoglobulin heavy/light chain gene (IgH/IgL) loci, which are induced by AID (encoded by AICDA). AID off-targeting may result in translocation of proto-oncogenes to the active Ig loci, resulting in constitutive oncogene expression. Such translocations are commonly observed in BL (MYC-IgH/IgL), DLBCL (BCL6-IgH, BCL2-IgH, MYC-IgH/IgL), FL (BCL2-IgH) and MCL (CCND1-IgH). In DLBCL, MYC rearrangements are seen in approximately 10% of the cases and a lower percentage is seen in FL, MCL and HL. Myc rearrangements have been associated with more aggressive lymphoma subtypes and a poor clinical outcome. A direct causative role for Myc-regulated genes and cancer development has been established for several protein coding genes and miRNAs. During normal B cell maturation, Myc is only expressed in specific subsets of cells (FIGURE 2): Myc is induced upon T cell activation and remains present during the initial stages of GC formation, to then be repressed by BCL6 in the active GC. A subset of light zone GC B cells with high affinity B-cell receptors express Myc upon interaction with T cells. These Myc+ light zone GC B cells (Myc+, BCL6-, NF-κB+) are thought to re-enter the dark zone for further proliferation and are essential for GC maintenance. As MYC is frequently translocated in B-cell lymphomas and translocations generally require the active transcription of its partners, these Myc+ GC B cell populations may be especially prone to malignant transformation.

2 Aim and scope of the thesis

LncRNA research does not only greatly expand our understanding of human cell biology and its intricate, cell-type specific regulatory mechanisms, it also provides
an additional point-of-view on the deregulation of these cellular systems in cancer cells. In the future, fundamental research on lncRNA expression deregulation and functionality will likely provide a solid basis for a vast expansion of drugable targets in cancer therapy. This study aims at the characterization of lncRNA expression patterns in normal B cell subsets and different types of B-cell lymphoma and a further functional characterization of deregulated lncRNAs.

In CHAPTER 2, the topic is introduced more thoroughly in a review addressing the current state of lncRNA research in normal B-cell development and B-cell malignancies. CHAPTER 3 addresses lncRNA expression patterns in normal sorted B-cell populations (i.e. naïve, GC and memory B cells) and assess lncRNA deregulation in HL cell lines compared to their presumed cell of origin. The study of HL is especially challenging due to a minor contribution of actual tumor cells (<1%) within an extensive inflammatory background. Studying lncRNAs deregulated in HL cell lines provided us with HL-specific candidate lncRNAs whose expression was further studied in primary HL tissue samples by RNA FISH. In CHAPTER 4, we focused on lncRNAs regulated by the oncogenic transcription factor Myc, using genome wide expression analyses in a B-cell model with a repressible MYC allele and in primary lymphoma samples characterized by high or low Myc expression. With these experiments we will for the first time establish the relevance of lncRNAs for the Myc transcriotional network. In CHAPTER 5 we further deepen the understanding of Myc-regulated lncRNAs by using an additional in vitro model, i.e. Myc knockdown in BL cell lines. A reliable set of Myc regulated lncRNAs relevant to BL pathogenesis will be established by a further selection of candidates based on an early response to MYC induction and proven Myc occupancy near the transcription start site. Further studies on the top Myc-induced candidate KTN1-AS1 will be performed to establish its role in guiding the downstream effects of Myc in BL cells. In CHAPTER 6 our findings are summarized, discussed and future perspectives are elaborated.
3 References


