HIV-1 in the RNA world: Transcription regulation, miRNAs and antiviral RNAs

Harwig, A.

Citation for published version (APA):
Chapter One

General Introduction
Scope of the thesis
Introduction

RNA

All organisms, from bacteria to human, use three biological molecules that each serve critical functions in the expression of genes in the cell. These are deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. RNA is synthesized from DNA in a process called transcription (1, 2). RNA differs from DNA in that it is composed of ribose instead of deoxyribose sugars, it contains Uracil instead of Thymine and it is mostly single-stranded (ss) instead of double-stranded (ds). The flow of genetic information in a cell is from DNA through RNA to proteins, thus DNA makes RNA makes protein. The genetic information in the form of DNA in the nucleus is connected by the messenger RNA (mRNA) to the protein-producing ribosomes in the cytoplasm (3). But RNA molecules play many more critical functions in the cell, e.g. it forms the enzymatic and structural component of ribosomes (rRNA) and the translational building blocks (tRNA), and is the main effector of post-transcriptional gene regulation (via microRNAs or miRNAs in the RNA interference (RNAi) mechanism). RNA also serves as the genetic material for many clinically relevant viruses like Ebola virus, hepatitis C virus and the human immunodeficiency virus type I (HIV-1). In this thesis, some known and novel RNA functions are investigated with HIV-1 as model system. The main focus of this work is on the mechanism of HIV-1 transcription, the RNAi pathway and the interplay between both. Therefore, to understand this thesis, the general biology of HIV-1 will first be covered, after which the mechanism of transcription and RNAi will be explained in more detail.

The retrovirus HIV-1

HIV-1 was identified as the causative agent of acquired immunodeficiency syndrome (AIDS) in 1983 (4). This virus is mainly transmitted through sexual intercourse and traverses epithelial barriers at mucosal surfaces of the genital and anorectal tracts (reviewed in (5)). Since its discovery many advances have been made in the treatment of this chronic infection. At present, HIV-1 infection does not have to lead to an unavoidable death as there are several drug treatments that can inhibit viral replication and consequently prolong the life of the infected individual. However, a complete viral cure is not possible because, like all members of the retrovirus family, HIV-1 will settle itself in the human genome during virus replication. HIV-1 belongs to the lentivirus genus, a subfamily of the Retroviridae (6). Retroviruses carry two copies of the positive-stranded RNA genome in each virion (7). Upon entry of the cell, this RNA genome is reverse transcribed into double-stranded DNA that is subsequently integrated in the host cell genome (8). The viral genome will thus become an integral part of the host cell genome, which is one of the reasons why HIV-1 is so difficult to cure. Here the virus can remain dormant to establish a latent infection, but new RNA and proteins can be produced later upon activation to make infectious virus particles and reignite the spreading infection (9, 10).
Replication cycle of HIV-1

HIV-1 replicates primarily in CD4-positive T lymphocytes, but dendritic cells, monocytes and macrophages can also be infected by this virus (11, 12). All these target cells are part of the human immune system. By replicating in these immune cells – which may result in cell-death – HIV-1 will trigger a gradual collapse of the human defense system, resulting in opportunistic infections and ultimately death of the infected individual. The HIV-1 virion particle will recognize the CD4 receptor on target cells via the gp120 subunit of the viral Envelope (Env) protein present on the viral membrane (Fig. 1) (13). This interaction induces a conformational change in the trimeric Env protein structure, allowing subsequent binding to the cellular co-receptor CXCR4 or CCR5 (14-17). The viral membrane is then fused with the cellular membrane, releasing the viral core into the cytoplasm (18). The core is partially uncoated to form the pre-integration complex (19-21). Within the viral core, the RNA genome is converted into dsDNA by the viral Reverse Transcriptase (RT) enzyme that is co-packaged in the viral core (22-24). The core is transported from the cytoplasm to the nucleus where the viral DNA is inserted into the host genome by the viral Integrase (IN) enzyme and cellular co-factors (25-27). Viral integration appears to be random with respect to the chromosomal site, but there is a preference for integration into active genes (28-30). Once integrated, the so-called HIV-1 provirus will hijack the transcription machinery of the host cell to start the production of unspliced and singly-spliced viral RNAs (31, 32). The unspliced RNAs have a bivalent role: they can be translated into viral proteins or packaged as viral genome in virions. The spliced mRNAs are used for translation of the other viral proteins necessary for virus replication. The viral structural proteins and two full-length RNA genomes assemble at the cell membrane to form a nascent virion particle, which is released from the cell by budding (33-38). The virions mature to become infectious, such that new host cells can be infected to start the replication cycle over again (39-42).

HIV-1 genome organization

The HIV-1 proviral DNA genome is approximately 9.8 kilobasepair (kbp) in length and contains nine genes (Fig. 1) (43, 44). Common to all retroviruses, HIV-1 encodes the Gag, Pol and Env proteins (45). Gag and Pol are translated from the unspliced mRNA as a polyprotein precursor and subsequently cleaved by the viral Protease (PR) (46, 47). Gag is cleaved into the structural subunits Matrix (MA), Capsid (CA) and Nucleocapsid (NC), while Pol encodes for the viral enzymes RT, IN and PR. The Env open reading frame encodes for the surface (gp120) and trans-membrane (gp41) subunits of the Envelope glycoprotein. The regulatory proteins Tat (48) and Rev (49) are translated from multiply-spliced mRNAs. The transactivator protein Tat enhances transcription from the long-terminal repeat (LTR) promoter and the Rev protein is required for nuclear export of unspliced and singly-spliced viral transcripts. In addition, the HIV-1 genome encodes four accessory proteins Vif, Vpr, Vpu and Nef. These proteins are not strictly required for virus replication and exhibit various functions, e.g. disrupting the antiviral function of restriction factors like APOBEC (50), suppression of potential antigen-specific T cells (51), degradation of the CD4 receptor to avoid super-infection (52) and modulation of MHC class 1 cell surface molecules to avoid immune recognition (53).
The HIV-1 coding region is flanked by non-coding LTRs, which are subdivided in the U3 (unique 3’), R (repeat) and U5 (unique 5’) domains. Multiple critical sequence motifs are encoded by the LTRs. The 5′ LTR is active as the transcriptional promoter and 3′ LTR-encoded sequences encode the polyadenylation signal (54). Viral transcription starts at the first residue of R in the 5′ LTR and the transcript is polyadenylated at the last residue of R in the 3′ LTR.
Transcription of HIV-1

HIV-1 transcription is a critical step in the viral replication cycle in which the integrated provirus uses the cellular transcription machinery to produce the viral RNAs. The HIV-1 provirus encodes a strong promoter in the U3 region of the 5' LTR (Fig. 1) (55, 56). This U3 promoter comprises a core promoter and an enhancer region (57, 58). The core promoter is a powerful promoter with three tandem Sp1 binding sites (59, 60) and a TATA box (CATATAA) element (56, 61). The enhancer region contains binding sites for NF-kB (62, 63), NFAT (64), AP1 (56, 65) and variable other transcription factors dependent on the HIV-1 subtype (66). This enhancer upregulates HIV-1 transcription, but removal of its binding sites is not deleterious in certain cell lines (67). In contrast, the core promoter is essential for transcription. HIV-1 transcription is a quite unique process that not only requires DNA promoter elements, but also an RNA element within the newly formed transcript, the trans-acting region (TAR) hairpin structure (Fig. 2A) (61, 68). TAR binds the viral Tat protein to fully activate HIV-1 transcription (Fig. 2B). Below we will discuss in more detail the function of this Tat-TAR axis.

Transcription can be divided into distinct phases: RNA polymerase II (RNAPII) is recruited to the HIV-1 promoter, initiates transcription, escapes from the promoter, possibly pauses at promoter-proximal sites, elongates the RNA transcript and eventually terminates transcription (69, 70). Recruitment of RNAPII occurs at the core promoter (71, 72). The non-processive RNAPII complex will initiate transcription of the HIV-1 RNA genome, starting at the transcription start site (TSS). However, soon after initiation most RNAPII molecules will pause, yielding short transcripts (73). This is caused by the folding of these short nascent RNA transcripts into the stable TAR hairpin structure (Fig. 2C; left panel) (68). Folding of the stable TAR hairpin structure on the nascent transcript will force the non-processive RNAPII complex to backtrack (74-80), such that the 3' end of the nascent transcript is displaced from the RNAPII catalytic center. As a result, the polymerase is paused until it is reactivated (Fig. 2C; left panel). Several other factors have been reported to suppress transcription elongation, such as negative elongation factors (NELFs), DRB sensitivity inducing factor (DSIF) (81-84), and microprocessor (85). RNAPII pausing on the TAR hairpin provides a time window in which transcription factors can be recruited that activate RNAPII to its full potential (Fig. 2C; middle panel) (86). One of these factors is the viral trans-activator protein Tat (Fig. 2B). Tat can recruit the positive transcription elongation factor-b (P-TEFb) from the 7SK small nuclear ribonucleoprotein (snRNP) complex that is bound at the HIV-1 promoter to direct it to the short TAR transcript (Fig. 2C; middle panel) (87-89). Once released from 7SK snRNP and bound to TAR, P-TEFb will phosphorylate the C-terminal domain (CTD) of RNAPII, and the NELF and DSIF proteins. This phosphorylation results in a processive RNAPII complex that is able to overcome the TAR pause, and phosphorylated NELF and DSIF proteins that cannot bind tightly to the RNA (81). This will allow RNAPII to produce full-length HIV-1 transcripts (Fig. 2C; right panel). The HIV-1 RNA transcripts are usually polyadenylated at the polyadenylation signal encoded by the 3' LTR and can be spliced or remain unspliced (see chapter 2).
Figure 2. The players involved in HIV-1 transcription initiation and elongation. (A) Structure of the TAR hairpin. (B) Domain organization of the viral Tat protein and P-TEFb component CycT1. Amino acid (aa) positions are indicated and the domain functions are described. (C) Model of RNAPII pausing (II) and reactivation (►) upon TAR transcription. Phosphorylation of the CTD domain of RNAPII is indicated (P). The catalytic center of RNAPII is positioned at the black arrow. See the text for more details.
**The Tat protein**

Tat is a small regulatory protein translated from a multiply-spliced HIV-1 transcript (Fig. 1). The protein consists of 86 to 101 amino acids, depending on the virus isolate, and is vital for HIV-1 replication (90, 91). The Tat protein can be subdivided in different domains (Fig. 2B). The first 58 amino acids that include the activation and RNA binding domains are important for its trans-activation function (92). The RNA binding domain, consisting of multiple lysine and arginine residues (the RKKR-motif) (93), is involved in binding to the bulge region of TAR. The activation domain interacts with the CyclinT1 (CycT1) component of P-TEFb (Fig. 2B) (94, 95). P-TEFb consists of CycT1 and a cyclin-dependent kinase (CDK9) and has a key role in activation of the elongation phase of transcription. P-TEFb is present at the U3 promoter in an inactive state when bound to the 7SK snRNP. Tat binds tightly to the P-TEFb complex by forming a Zinc-mediated bridge with the Cys261 residue of CycT1 (96, 97). When in complex with CycT1, the affinity of Tat for TAR is increased 9-fold (94, 98). This high affinity will lead to recruitment of Tat-P-TEFb to TAR, releasing the 7SK snRNP and allowing P-TEFb to activate transcription (87-89). Many alternative functions have been attributed to Tat. For instance, Tat has been reported to function as a suppressor of the Dicer nuclease of the RNA interference mechanism (99, 100), although this result could not be reproduced by others (101).

**The TAR RNA hairpin**

TAR is a structured RNA element present at both the 5’ and 3’ ends of the HIV-1 RNA genome (68). TAR folds a 57-nt hairpin with 23 bp and a 3-nt bulge that divides the stem into two helical domains (Fig. 2A). Due to the presence of the triple-U bulge, the top of the stem is bend with respect to the bottom in an angle of ~50° (102). Analysis of the interaction between Tat and TAR RNA by NMR revealed that Tat binding forces a conformational change in the TAR RNA structure, which results in straightening of the TAR hairpin (Fig. 2C) (103, 104). TAR has been reported to act as transcriptional ‘pause-hairpin’ (75). Indeed, when RNAP II is not activated, it will pause directly after transcribing the TAR hairpin, which results in production of short TAR-containing transcripts (73). This pausing of RNAPII can lead to the establishment of a latent reservoir of integrated viruses (105). It has been reported that microprocessor, the other nuclease of the RNAi mechanism, can cleave the TAR hairpin and hereby induce termination of transcription (85). TAR may also act as substrate in the RNAi mechanism to produce miRNA regulators of gene expression (99, 106-108), but this possibility has been disputed by others (101, 109, 110).
Figure 3. The RNAi pathway. Schematic of pri-miRNA processing into an active miRNA that executes mRNA silencing. See the text for further details.
The RNAi mechanism

RNAi is an evolutionarily conserved post-transcriptional gene silencing (PTGS) mechanism that is triggered by dsRNA. The dsRNA precursors are cleaved into small microRNAs (miRNAs), ~21 nt in size, that target complementary mRNAs for inactivation (Fig. 3). I will briefly discuss the canonical RNAi pathway, followed by a survey of the core players and a description of recently described non-canonical pathways.

In the cell, miRNAs are encoded by long primary transcripts (pri-miRNA) (111). The pri-miRNA folds a characteristic hairpin structure that is recognized and cleaved by the nuclear microprocessor complex into a precursor miRNA hairpin (pre-miRNA) with a 2-nt overhang. Microprocessor consists of at least two proteins: the ribonuclease III (RNAse III)-like enzyme Drosha and the RNA-binding DGCR8 (DiGeorge syndrome critical region 8) (112, 113). The pre-miRNA is transported to the cytoplasm by Exportin-5 (Exp5) (114) and subsequently cropped by the RNAseIII-like Dicer enzyme into a miRNA duplex (115, 116). This miRNA duplex is loaded into the RNA-induced silencing complex (RISC) (117, 118). The strand with the thermodynamically least stable 5’ end becomes the guide strand, while the passenger strand is degraded (119, 120). Argonaute (Ago) protein performs the nuclease activity of RISC and uses the guide strand to recognize a complementary mRNA target (121). Depending on the extent of basepairing within the miRNA:mRNA duplex, the mRNA is silenced by cleavage or translationally repressed (118, 122).

Dicer can also recognize and cleave small hairpin RNAs (shRNAs), which results in the production of small interfering RNAs (siRNAs). The shRNAs can be produced from man-made transgene constructs using cellular or viral promoter elements, usually RNAPIII promoters that drive the expression of small transcripts. In these constructs, the ends of the shRNAs are defined by RNAP initiation and termination signals, in such a way that an RNA hairpin is formed with a 2-nt overhang that is recognized by Dicer. Dicer cleavage removes the loop domain of the shRNA, resulting in a duplex small RNA molecule that can be loaded into Ago2 to activate the RISC complex. Unlike miRNAs, this shRNA path does not require microprocessor cleavage.
**Microprocessor**

Processing of the pri-miRNA is critical in miRNA biogenesis because microprocessor and Dicer determine the ends of the mature miRNA sequence, and thus its target specificity (112, 123, 124). DGCR8 (~86kDa) serves as the molecular anchor of microprocessor and contains two RNA binding domains (RBD) (125) (Fig. 4A). Drosha (~160 kDa) is the endonuclease responsible for pri-miRNA cleavage and has a single dsRBD and two RNase III domains (112, 126). The ability of DGCR8 to bind RNA is modulated by acetylation of lysine residues within the dsRBDs (127). DGCR8 recognizes the ss-dsRNA junction at the base of the short hairpin (128, 129). The position and size of the pre-miRNA terminal loop region also affects cleavage (128, 130-132). Microprocessor cleaves optimally at sites that are positioned ~11 bp from the base and ~21 bp from the top of the hairpin (Fig. 4). With sub-optimal distances, microprocessor will cleave at multiple sites, yielding multiple pre-miRNA species. The dsRBD of Drosha is necessary but not sufficient for interaction with the pri-miRNA (123). Each of the two RNase III domains of Drosha will cleave one side of the RNA stem and leave a 5’ phosphate and a 3’ 2-nt staggered overhang on the pre-miRNA that are necessary for recognition by Exp5 and Dicer (123, 133, 134). Exp5 recognizes pre-miRNAs with a stem of more than 14 bp containing a short 3’ overhang and exports the pre-miRNA out of the nucleus (135, 136).

![A](image1)

**Figure 4. Microprocessor structure and function.**

(A) Domain structure of the microprocessor components DGCR8 and Drosha. (B) Model of microprocessor-mediated processing of a pri-miRNA into pre-miRNA. See the text for further details.
**Dicer**

Once arrived in the cytoplasm, Dicer will recognize the pre-miRNA and crop this molecule to a mature miRNA duplex (137). As described above, shRNAs can be introduced at this point into the RNAi mechanism and follow a very similar path. Dicer is a large enzyme (~200kDa) that contains three helicase domains, a pair of tandem RNase III domains, a dsRBD, a ruler and PAZ (Piwi Argonaut and Zwille) domain (Fig. 5A). The Dicer enzyme adopts an L-shaped form with the helicases at the base and PAZ near the top. The ruler domain folds back, bringing the RNase III domains in the vicinity of the Dicer base (Fig. 5B) (138-140). The three N-terminal helicase domains of Dicer are arranged in a hand-shaped form at the base of the Dicer protein and facilitate pre-miRNA recognition by interacting with the terminal loop (141). The PAZ domain contains two adjacently positioned RNA binding pockets of which one will recognize the 3’ overhang of 2 nt and the other will dock the 5’ phosphate (142-144). The Dicer protein will subsequently undergo a conformational change that links the dsRBD and RNase III domains to the top of the pre-miRNA (Fig. 5B) (140). Once in proximity of the RNA hairpin, the RNase III domains will cleave in a 2-nt staggered fashion on a fixed distance from the terminal end of the pre-miRNA. This cropping of the pre-miRNA into a miRNA duplex occurs at a distance of 21 to 25 bp, or ~70 Å, which is ~2 helical turns of RNA as determined by the ruler domain (134, 138, 144). Internal bulges and loops may influence the cleavage specificity of Dicer (145). Dicer uses the TAR RNA-binding protein (TRBP) and protein activator of PKR (PACT) as co-factors, but their precise roles has not yet been established (146-148).

Figure 5. Dicer structure and function. (A) Domain organization of the Dicer nuclease. (B) Model of Dicer-mediated processing of a pre-miRNA into a miRNA duplex. See the text for further details.
**Ago/RISC**

The processed RNA duplex is subsequently released by Dicer and loaded onto the Ago component of RISC (118, 122, 149). Several Ago variants have been identified and all Ago proteins contain an N-terminal domain, PAZ domain, middle domain (MID) and PIWI (P-element induced wimpy testes) domain (Fig. 6A) (150, 151). These domains are separated by two linkers and arranged around a central binding groove (Fig. 6B). The strand with its 5’ end at the most unstable side of the small RNA duplex – the guide strand – will be loaded into RISC, while the passenger strand is degraded (119, 120). Removal of the passenger strand occurs either by unwinding (Ago1, 3 and 4) or by cleavage, i.e. slicing, of the passenger strand (Ago2). Ago2 is the only Ago variant with a PIWI domain capable of slicing activity (121, 152, 153).

The 5’ end of the guide strand will dock in a pocket of the MID domain (154-156) and induce a conformational change that locks this otherwise flexible protein and the small RNA in a stable structure (151, 157-159). The short RNA is threaded through the central binding groove and exposes the seed region (nucleotides 2 to 5) of the guide strand for target site recognition (Fig. 6B) (150, 160, 161). Once the target RNA sequence is basepaired with the guide strand, Ago2 will undergo stepwise conformational changes, resulting in a widening of the central groove and allowing more of the guide strand to pair with the target mRNA (160). Depending on the level of complementarity between the guide strand and the target mRNA, the latter will be translationally repressed (partial complementarity) or sliced by the PIWI domain (full complementarity) (162-164).

---

**Figure 6. Ago2 structure and function.** (A) Domain organization of the Ago2 enzyme. Linkers separate the N and PAZ domains (L1) and the PAZ and MID domains (L2). (B) Model of miRNA loading in Ago2 and subsequent target mRNA recognition and cleavage. See the text for further details.
Non-canonical miRNA processing

Next to the canonical pathway described above, alternative processing routes for miRNA-like small RNAs have recently been described (165, 166). These non-canonical routes can be divided in Drosha-independent and Dicer-independent pathways. In the Drosha-independent pathways, alternative processing is required to create a short hairpin RNA that can function as substrate for Dicer. This can be achieved by splicing out of a small stem-loop structure (mirtrons) (167-169), transcription of a short hairpin transcript such that RNAP initiation and termination determines the 5’ and 3’ ends (170-172), or processing of small non-coding RNAs like tRNAs (173) and small nucleolar RNAs (snoRNAs) (174).

Dicer-independent pathways require an alternative way to process a small RNA hairpin, such that it can be loaded into RISC to exert its function. Ago2 is able to process the unique pre-miR-451 that is too small (stem of 17 bp) to be recognized by Dicer (175, 176). The same path was described for small shRNA variants with a stem of 17 to 19 bp that were therefore called AgoshRNAs (177). Ago2 mediates the processing and the produced guide strand subsequently directs RISC-mediated target RNA cleavage (176-178). Processing by Ago2 differs much from regular Dicer processing. The Ago2 PIWI domain cleaves the 3’ arm of the miR451 and AgoshRNA hairpins between bp 10 and 11, creating a molecule that remains partially basepaired. For miR-451, it has been described that poly(A)-specific ribonuclease (PARN) trims the 3’ end to create the mature species (179). Dicer-independent AgoshRNAs are promising therapeutics because no passenger strand is generated, which reduces off-target effects (180). In addition, such AgoshRNAs can be used in cells like monocytes that do not express the Dicer nuclease (181).
Scope of the thesis

The research described in this thesis employs HIV-1 as model system to investigate the molecular processes of RNA transcription and small RNA processing in the RNAi mechanism. In particular, we address the interplay between these two processes. In chapter 2-4 we study aspects of HIV-1 transcription. Chapters 5 and 6 deal with retrovirus-encoded miRNAs and their biogenesis during RNA transcription. Chapters 7-10 focus on the optimization of RNAi reagents as therapeutic tools to target HIV-1.

In chapter 2 we review recent findings related to HIV-1 RNA biosynthesis. Special attention is paid to the viral TAR RNA hairpin and we discuss TAR-encoded miRNAs. In chapter 3, we investigate the contribution of the Tat protein to HIV-1 replication and possible functions in addition to the well-known transcription function. To do so, we use a modified virus in which the Tat-TAR axis is functionally replaced by the Tet-On system for inducible gene expression. In chapter 4, we investigate the mechanism of TAR-mediated transcriptional pausing and premature termination. We present evidence for a novel role of microprocessor cooperating with the co-factors Setx, Xrn2 and Rrp6.

A long-standing controversy whether or not retroviruses can produce miRNAs is reviewed in chapter 5. The main argument against this possibility is the reasoning that miRNA production would lead to cleavage and inactivation of the viral RNA genome, but some recent exceptions were described. In chapter 6, we study whether a miRNA is encoded by the HIV-1 TAR hairpin and a novel non-canonical biogenesis route is described. The role of the TAR-binding Tat protein in this process is also addressed.

The next chapters focus on the recently described shRNA variants that skip Dicer-processing and are alternatively processed by Ago2. The molecules that use such a non-canonical processing route are called AgoshRNAs. Chapter 7 focuses on the critical AgoshRNA properties (stem length, top G-U basepair) needed to activate Ago2 processing. Details of the processing of these AgoshRNA variants are subsequently addressed by deep sequence analysis in chapter 8. In chapter 9, we perform a subsequent AgoshRNA mutational analysis, thereby focusing on the top and bottom base pairs. Chapter 10 probes the mechanism of non-canonical AgoshRNA processing by Ago2. The use of these AgoshRNA molecules as anti-HIV therapeutics is discussed.
References


150. **Schirle NT, MacRae IJ.** 2012. The crystal structure of human Argonaute2. Science 336:1037-1040.


