UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

Characterization of sncRNAs in response to HIV

Miguel Tavares Paiva Santos Pereira

Mestrado em Biologia Molecular e Genética

Dissertação orientada por:
Doutora Margarida Gama Carvalho

2019

Acknowledgments

Foram quase dois anos de trabalho árduo e de momentos altos e baixos. Uma experiência extremamente enriquecedora a todos os níveis. Conheci muita gente, que por mais pequeno que tenha sido o contributo, tornaram esta experiência espetacular.

Quero agradecer imenso à Doutora Margarida Gama Carvalho pela oportunidade de trabalhar neste projeto, que me permitiu um crescimento científico enorme, a todos os níveis. A todas as correções, conselhos e apoios que deu ao longo destes quase dois anos. Vou continuar a tentar melhorar, pois sei que há muito para evoluir, mas sei que cresci bastante desde o primeiro dia.

Em seguida, gostaria de agradecer a todos os elementos do" RNA Systems Biology -Gama & Pinto lab", foi um prazer ter-vos conhecido e ter privado convosco, em contexto dos "Lab meetings" e momentos de descontração. Um obrigado especial à Tânia pela paciência enorme e ajuda que me deu na parte bioinformática, que foi deveras essencial e ajudou-me a aprender sobre a mesma. À Inês, Afonso, Nuno e João pelos ensinamentos. Um obrigado à Inês também pelas magníficas conversas e momentos de descontração, uma amizade que ficará para sempre no meu coração. Um obrigado enorme à Iga pelas conversas e oportunidade de ajudar e ensinar, foi um prazer enorme e uma experiência especial. Um enorme obrigado também à Cláudia, uma pessoa que valeu imensa a pena conhecer. Uma companheira de curso, que se conhece anos depois, mas como diz o provérbio "mais vale tarde do que nunca". Uma amizade para recordar e manter.

Gostaria também de agradecer aos elementos do laboratório da Doutora Margarida Amaral pela ajuda e ensinamentos necessários durante etapas deste longo caminho, que enriqueceram sem dúvida alguma esta viagem maravilhosa.

Um agradecimento ao Vicente e à Andreia, aos apoios que demos uns aos outros em momentos de dificuldade e aos momentos de alegria e diversão, que tornaram o caminho bem mais fácil. A todos os meus amigos pelas várias frases de apoio, que por mais simples que parecessem foram sempre boas de ouvir, em conjunto com os bons momentos que ajudavam a abstrair do Mestrado.

Um agradecimento aos meus avós por se preocuparem em saber sempre como estava a correr o Mestrado e acreditarem que ia conseguir. À Beatriz e Abraão, por todos os dias de total descontração, diversão e relaxamento, que ajudaram imenso a descansar do trabalho. Ao meu pai, pela motivação, preocupação e crença em mim. Todos os momentos contigo foram uma ajuda preciosa para que eu chegasse ao fim desta aventura. Um obrigado infinito à minha querida mãe. Sem ela, esta aventura nunca teria sido possível. Sempre acreditaste em mim e apoiaste-me mesmo quando as coisas foram complicadas de começar. Há muito na vida que te devo a ti e esta é uma delas.

Por último, ao meu amor e tesouro, à minha adorável namorada. Por seres a minha melhor amiga, companheira, por acreditares e mim, por me amares, por todos os momentos que partilhamos juntos, por tudo. Sem ti, esta viagem não seria a mesma.

Resumo extenso

Em estudos anteriormente feitos no laboratório, têm se testado possíveis associações de pequenos RNAs não codificantes (sncRNAs) à infeção pelo vírus da imunodeficiência humana (HIV). Os sncRNAs têm entre 18 até 200 nucleótidos de comprimento, não codificam para proteínas, sendo responsáveis pela regulação génica ao nível transcricional e pós-transcricional, processamento e modificação de RNA. Existem diversos sncRNAs, sendo os mais curtos em extensão, os pequenos RNAs de interferência (siRNAs) e os microRNAs (miRNAs). São sncRNAs com uma origem diferente, mas têm em comum a utilização do complexo de silenciamento induzido por RNA (RISC) para a sua função de silenciamento. O HIV tem como alvo principal os linfócitos T CD4+, que são parte crucial da resposta imune humana. As células T CD4 podem ser *naïve*, efetoras ou de memória. Por sua vez, o HIV pode estar em estado latente ou ativo durante a infeção, mediante do estado das células T CD4, podendo mesmo regular esse estado. Quando as células estão no estado quiescente, um estado de repouso, ou são células que ainda não encontraram o antigénio apropriado, células naïve, a infeção é latente e não há a produção de transcritos virais. A ativação das células *naïve* procede através do recetor de células T (TCR) em conjunto com outros fatores, podendo-se assim ativar células T CD4 naïve em laboratório.

Durante os estudos anteriores no laboratório de acolhimento, novos pequenos RNAs (sRNAs), propostos como associados a retrotransposões, foram identificados a partir de dados de sequenciação. Esses dados foram obtidos a partir da análise de bibliotecas de sequências de RNA de células T CD4 naïve não estimuladas, ou estimuladas in vitro via TCR, sem infeção, infetadas com HIV-1, ou HIV-2. Estes novos sRNAs apresentam potencial para ser um novo mecanismo de defesa contra o HIV por parte do hospedeiro, pois aparentam ter novas categorias funcionais, diferentes das dos miRNAs. Estes sRNAs são mais curtos em extensão e com função ainda não definida.

O objetivo deste trabalho consistiu em confirmar a hipótese de alterações na expressão dos novos três sRNAs e caracterizar a sua origem e potencial impacto na interação entre hospedeiro e o vírus do HIV. Para tal, recorreu-se a dois modelos experimentais diferentes, uma linha celular J-Lat e células primárias usadas para a produção das bibliotecas de sequenciação. A linha celular J-Lat é uma linha celular derivada de Jurkat com um provirus de HIV integrado, mas transcricionalmente latente, com algumas modificações para eliminar o potencial infeccioso. O gene da proteína da fluorescência verde (GFP) substitui a sequência codificante da proteína nef e existe uma mutação frameshift da env. A transcrição latente pode ser ativada pelo factor de necrose tumoral-α (TNF-α) através da ativação da via do NF-κB, que é responsável pela eficiente transcrição do promotor long terminal repeat (LTR) do HIV. Desse modo, a utilização da linha celular J-Lat serve como um modelo biológico, onde é possível testar a hipótese colocada em relação às possíveis alterações de expressão dos sRNAs alvo e os impactos na interação entre hospedeiro, J-Lat, e vírus, HIV. Em relação às células primárias, a utilização de amostras infetadas de células humanas primárias em conjunto com as amostras controlo também permite o seu uso como modelo biológico.

Para a deteção dos sRNAs alvo, uma vez que as sequências são muito curtas entre 15 e 17 nucleótidos, o método de deteção foi transcrição reversa com reação de polimerase em cadeia (RT-PCR), mas com o uso do método de cauda de Poly A para produção de DNA complementar (cDNA) na transcrição reversa (RT). O método por cauda de Poly A permitiria criar um cDNA a partir de sequências muito curtas e com especificidade. Começou-se por estabelecer as condições de cultura das células e de reativação do vírus latente. Fez-se cultura de J-Lat e estimulou-se com TNF-α para ativação da

transcrição de transcritos virais. A confirmação da reativação do vírus latente fez-se por visualização da sequência codificante da proteína GFP no genoma viral por microscopia de fluorescência e deteção dos transcritos virais por RT-PCR. O RNA extraído pelo método de TRIzol de células J-Lat tratadas com TNF-α e de células controlo, foram então utilizados para os ensaios de deteção de alterações de expressão dos sRNAs alvo em resposta à presença de transcritos virais.

Os vários ensaios tiveram resultados inconclusivos. Os sRNAs alvos foram detetados, mas não se conseguiu obter uma expressão diferencial ao longo das amostras. Tendo em conta tais resultados, decidiu-se usar as amostras de células primárias, pois haveria a possibilidade de o fenómeno de expressão diferencial só se poder observar em células primárias. No entanto, os mesmo resultados foram obtidos, alvos sRNAs detetados, mas com ausência de expressão diferencial ao longo das amostras. Mediante de tais resultados, procedeu-se a uma análise bioinformática para explorar de forma mais detalhada a presença destes sRNAs nas bibliotecas de sequenciação e determinar a sua potencial origem.

Os dados utilizados consistiram nos dados originais de sequenciação anteriormente publicados e que suportaram a identificação dos sRNAs alvo. Contagens das sequências alvo nos dados de sequenciação de amostras de células estimuladas e infetadas com HIV-1, confirmaram a observação anterior de aumento de expressão destas sequências em comparação com amostras de células não infetadas. Havendo um aumento mais acentuado nas contagens das sequências exatas dos sRNAs alvo do que de sequências longas contendo os sRNAs alvo, concluiu-se que as sequências alvo são independentes de sequências longas que as contenham. Em seguida, procedeu-se ao mapeamento dos sRNAs alvo no genoma humano de forma a identificar a sua potencial origem. Dois sRNAs alvo foram associados a um RNA de transferência (tRNA), tRNA$^{Lys}{}_3$, que é o *primer* para transcrição reversa do HIV-1. O PBS sncRNA do tRNA$^{Lys}{}_3$ alinha com os dois sRNAs alvo, mas esse alinhamento consiste em 12 nucleótidos dos 18 totais. Os restantes 5 nucleótidos dos sRNAs alvo alinham com os nucleótidos a montante da sequência do PBS sncRNA, mas com um "mismatch". A posição em questão no tRNA$^{Lys}{}_3$ corresponde a uma posição com uma modificação, A58, em que podem potencialmente ocorrer vários "mismatches" durante a transcrição reversa de RNAs para posterior sequenciação. Este nucleótido é também aquele que distingue os dois sRNAs entre si, o que pode significar que correspondem a apenas um sRNA novo.

A enorme similaridade entre os dois sRNAs alvo e o PBS sncRNA levou a que fosse procurada a sequência do PBS sncRNA e a sequência genómica do sRNA sem o "mismatch" A58. A comparação de abundância das quatro sequências levou à conclusão de que os valores do PBS sncRNA são de 7 até 32 vezes inferior nas bibliotecas não infetadas e 32 até 156 nas bibliotecas infetadas em comparação com as outras três sequências. Portanto, a sequência genómica do sRNA encaixa como a nova sequência do PBS sncRNA.

A existência de sRNAs derivados de tRNAs não é novidade, mas ainda é um campo com muitos pormenores desconhecidos. sRNAs derivados de tRNAs podem ser classificados como tRFs, fragmentos derivados de tRNAs. Estes sRNAs são metades dos tRNAs, quer por uma das extremidades, 5' ou 3', ou, a meio do tRNA, numa das partes do "loop". Os comprimentos das sequências também variam, entre 17 até 19, 18 até 22, ou, 30 até 33 nucleótidos. As funções dos tRFs têm sido associadas a inúmeros processos desde interferirem com a transcrição reversa, expressão de transposões, inibição de tradução, cancro, ou, infeção viral. Portanto, o sRNA alvo pode muito bem pertencer a uma nova classe de tRFs, pois não encaixa totalmente nas classes já existentes e que não são totalmente conhecidas em termos de funções, sendo o nosso sRNA alvo a nova sequência do PBS sncRNA com função já conhecida.

O sRNA alvo mais curto de 15 nucleótidos foi associado com um RNA ribossomal (rRNA), 18S rRNA muito utilizado em estudos filogenéticos e a dois retrotransposões THE1C e MSTA, que fazem parte de um “Endogenous Retrovirus-Mammalian apparent LTR retrotransposon” (ERV-MaLR). A ‘Repeat 3’ no 18S rRNA é localizada na base de um “stem-loop” compatível com o caminho de processamento de um miRNA canónico. Alguns estudos já realizados reportam a existência de miRNA derivados de rRNA em humano. Tendo já sido associados a piRNAs podendo funcionar como “small guide RNAs”.

tRNAs e rRNAs são dos RNAs mais abundantes nas células. Os vários níveis de estrutura dos rRNAs e tRNAs, as diferentes espécies de tRNAs e as modificações de tRNAs dificultam a deteção de fragmentos provenientes destes RNAs por métodos de PCR, podendo explicar a dificuldade na deteção da expressão diferencial dos alvos sRNAs.

Este trabalho levou à descoberta de uma nova sequência do PBS sncRNA, que poderá ajudar a perceber de uma forma diferente a relação entre HIV-1 e o hospedeiro, e por último, um miRNA derivado de rRNA com possível atuação durante a infeção por HIV-1. Haverá então a necessidade de execução de futuros ensaios para um melhor entendimento de funções e confirmação do modo de atuação destas sequências e o seu impacto na relação entre o HIV-1 e o hospedeiro.

Palavras chave

tRNA$^{Lys}{}_{3}$, tRFs, miRNA derivados de rRNA, PBS sncRNA, HIV.

Resumo

Os pequenos RNAs não codificantes (sncRNAs) são RNAs que não codificam para proteínas que variam em tamanho de 18 até 200 nucleótidos e são responsáveis pela regulação génica ao nível transcricional e pós transcricional, processamento e modificação de RNA. Em estudos anteriormente feitos no laboratório., novos pequenos RNAs (sRNAs) associados a retrotransposões, foram identificados de dados de sequenciação obtidos de RNA de células T naïve humanas não estimuladas e estimuladas via recetor de células T (TCR) e não infetadas e infetadas com HIV-1 ou HIV-2. Estes sRNAs aparentavam ter novas categorias funcionais, distintas dos micro RNAs (miRNAs), com características que sugerem o potencial para representar um novo mecanismo de defesa do hospedeiro contra a infeção por HIV. Distinta dos miRNAs, sncRNAs com 22 nucleótidos em extensão que são responsáveis por impedir a produção de proteína através de ligação e depois por clivagem do RNA mensageiro (mRNA) ou repressão de tradução, estes novos pequenos sRNAs são mais curtos e com função ou funções desconhecidas.

O objetivo era testar as alterações de expressão nestes novos sRNAs, três sRNAs, e caracterizar o seu impacto funcional na interação entre células hospedeiras e o vírus do HIV. Para esse propósito, modelos biológicos foram usados, uma linha celular com o nome J-Lat e células primárias, usadas para a construção de bibliotecas de sequenciação. A linha celular J-Lat tem integrada um proviorus HIV, mas transcricionalmente latente que pode ser ativado. A indução do NF-κB através da estimulação do fator de necrose tumoral alfa (TNF-α) leva à produção de transcritos virais, tornando a linha celular J-Lat adequada para ser um modelo biológico para este estudo. Tendo isso em conta, as células primárias de células não infetadas e infetadas, também servem como modelo biológico.

As sequências dos sRNAs alvo com 15 e 17 nucleótidos em extensão provou ser difícil para deteção por métodos de PCR padrão, pois o método tem de ser específico na deteção de sequências curtas em extensão. A transcrição reversa com um método RT-PCR com uma transcrição reversa conhecida por cauda de Poly A, provou ser confiável para a deteção de alvos curtos em extensão enquanto ao mesmo tempo ser específica.

Condições apropriadas para a reativação do provirus HIV latente foram estabelecidas. As J-Lat foram colocadas em cultura e estimuladas com TNF-α para levar à produção de transcritos virais. Visualização por microscopia de fluorescência da sequência codificante da GFP no genoma viral, mais a deteção de transcritos virais por RT-PCR, confirmam a reativação do vírus latente. O testar de alterações de expressão por cauda de Poly A e PCR de RNA de células tratadas com TNF-α e células controlo, produziram resultados inconsistentes. Alvos sRNAs foram detetados, mas sem a observação ao longo das amostras de expressão diferencial. Foi então usado RNA das células primárias dos dados de sequenciação para verificar, se o fenómeno poderia ser só observado em células T primárias. Infelizmente, os mesmos resultados foram obtidos com os alvos sRNAs detetados, mas outra vez sem expressão diferencial.

Análise bioinformática foi o próximo passo executado, não apenas para tentar verificar as alterações de expressão, mas também para tentar determinar a origem dos alvos sRNAs. Os mesmos dados de sequenciação usados para a identificação dos novos sRNAs foram usados para a análise. Os resultados mostraram que houve um aumento de expressão das sequências nos dados de RNA de células não infetadas e células infetadas, confirmando as suposições iniciais. O aumento em expressão foi mais alto em correspondências exatas para os alvos sRNAs do que em sequências mais longas que as contenham, o que indica que os alvos sRNAs são independentes dessas sequências mais longas. Na determinação da origem, duas das sequências foram associadas a tRNA$^{\text{Lys}}_{3}$, o *primer* para a transcrição reversa do HIV-1 e a outra sequência a um RNA ribossomal, 18S rRNA, e dois retrotransposões, THE1C e MSTA.

'Repeat 1' e 'Repeat 2' alinham com o tRNA$^{\text{Lys}}_{3}$, com um "mismatch" e com o PBS sncRNA em 12 nucleótidos dos seus 18 totais. A posição "mismatch" é localizada na posição 58, onde existe uma modificação, A58. Na transcrição reversa do RNA para sequenciação, essa posição é propensa a levar a "mismatches". Essa mesma localização é a onde o nucleótido é diferente na 'Repeat 1' e 'Repeat 2'. Desse modo, isso leva-nos a crer que as duas sequências são uma. A sequencia genómica sRNA foi procurada em conjunto com o PBS sncRNA. Os resultados revelaram que a quantidade de PBS sncRNA presente nos dados de sequenciação são 7 até 32 vezes menos nas bibliotecas de células não infetadas e 32 até 156 vezes menos nas bibliotecas de células infetadas, do que das sequências da 'Repeat 1', 'Repeat 2' e sequência genómica sRNA. Portanto, a sequência genómica sRNA é uma nova sequência PBS sncRNA.

Investigação adicional sugere que o sRNA podem bem fazer parte de uma classe recente de sRNA, fragmentos derivados de tRNA (tRFs). São sncRNAs que são derivados de tRNAs e são usualmente metades de extremidades 5' ou 3' de um tRNA, mas podem também existir metades do meio do tRNA. A nossa sequência parece fazer parte de uma nova subcategoria destas metades. The tRFs têm sido associados a muitos processos como silenciamento de RNA, cancro, ou, resposta a infeções virais.

Em relação à 'Repeat 3', nós determinamos a posição relativa da sequência no contexto da estrutura secundária reportada do 18S rRNA. A posição é localizada na base de um "stem-loop" que é compatível com o caminho de processamento do miRNA canónico. Estudos reportam a existência de miRNAs derivados de rRNA em humanos e ratos. Alguns destes miRNA derivados de rRNA têm sido associados a piRNAs e podem servir de pequenas RNAs guia.

As dificuldades na deteção da expressão diferencial dos alvos sRNAs podem ser devidas a especificidade dos tRNAs e rRNAs. Eles estão entre os mais abundantes nas células. Ambos têm vários níveis de estrutura e os tRNAs têm diversas espécies e modificações. Todos estes aspetos levam a uma grande dificuldade na deteção de fragmentos de RNAs por métodos de PCR padrão.

Esta dissertação trouxe à luz uma nova sequência de PBS sncRNA e um miRNA derivado de rRNA, que poderá ter um impacto num entendimento diferente da infeção por HIV-1 e a relação entre vírus e hospedeiro. Ensaios futuros são necessários para determinar quanto a descoberta de uma nova sequência do PBS sncRNA muda o nosso entendimento da infeção por HIV-1 e qual é a ligação entre o miRNA derivado de rRNA e o HIV-1, para determinar qual é a função exata e como isso afeta o hospedeiro e o HIV-1.

Palavras chave

tRNA$^{\text{Lys}}_{3}$, tRFs, miRNA derivados de rRNA, PBS sncRNA, HIV.

Summary

Small non-coding RNAs (sncRNAs) are RNAs that do not encode for proteins and range in size from 18 to 200 nucleotides and are responsible for gene regulation at the transcriptional and post-transcriptional level, RNA processing and modification. In previous studies performed in the laboratory, novel small RNAs (sRNAs) associated with retrotransposons, were identified from sequencing data obtained from RNA of CD4 human naïve T cells non-stimulated and stimulated via T cell receptor (TCR) and non-infected and infected with either HIV-1 or HIV-2. These sRNAs appeared to have new functional categories, distinct from micro RNAs (miRNAs), with characteristics that suggested the potential to represent a new defense mechanism of the host against HIV infection. Distinct from the miRNAs, sncRNAs with about 22 nucleotides in length that are responsible for preventing protein production through binding and then whether by cleavage of messenger RNA (mRNA) or translation repression, these novel sRNAs are shorter and have unknown function or functions.

The aim was to test expression alterations in these novel sRNAs, three sRNAs, and characterized their functional impact on the interaction between host cells and HIV virus. For that purpose, biological models were used, a cell line named J-Lat and primary cells, used to build the sequencing libraries. The J-Lat cell line has an integrated, but transcriptionally latent HIV provirus that can be activated. The induction of the NF-κB through the tumour necrosis factor alpha (TNF-α) stimulation leads to production of viral transcripts, making the J-Lat cell line suitable to be a biological model for this study. Given that, non-infected cells and infected primary cells, also suit as a biological model.

Our target sRNAs sequences with 15 and 17 nucleotides in length proved difficult for detection by standard PCR methods, because the method needs to be specific while detecting short sequences in length. A reverse transcription with polymerase chain reaction (RT-PCR) method with a specific reverse transcription know as Poly A tailing, proved to be reliable for the detection of short targets in length while also being specific.

Appropriate conditions for the reactivation of the latent HIV provirus were established. The J-Lat were cultured and stimulated with TNF-α to lead to the production of viral transcripts. Visualization by florescence microscopy of the coding sequence of the green fluorescence protein (GFP) in the viral genome, plus the detection of viral transcripts by RT-PCR, confirmed the reactivation of the latent virus. Testing of the expression alterations by Poly A tailing and PCR from RNA of TNF-α treated cells and control cells, produced inconsistent results. The targets sRNAs were detected, but there was no differential expression observed along the tested samples. Thus, RNA from the primary cells of the sequencing data was used to verify, if the phenomena could only be observed in primary T cell. Alas, the same results were obtained with target sRNAs detected, but again without differential expression.

Bioinformatic analysis was the next step taken, not only to try to check those expression alterations, but also to try to determine the target sRNAs origin. The same sequencing data used for the identification of those novel sRNAs was used for the analysis. The results showed that there was an increase in expression of the sequences in the data of RNA from non-infected cells to infected cells, confirming the earlier assumptions. The fold increase in expression was higher in exact matches for the target sRNAs than in longer reads containing them, which indicates that the target sRNAs are independent from longer reads. In the origin determination, two of the sequences were associated with transfer RNA Lysine 3 (tRNA^{Lys_3}), the *primer* for reverse transcription of HIV-1 and the other sequence to a ribosomal RNA, 18S rRNA, and two retrotransposons, THE1C and MSTA.

‘Repeat 1’ and ‘Repeat 2’ align with the $tRNA^{Lys}_3$ with one mismatch and the PBS sncRNA on 12 nucleotides of its total 18. The mismatch position is located on position 58, where there is a modification, A58. On reverse transcription of RNA for sequencing, that position is prone to lead to mismatches. That same location is the one where the nucleotide is different in ‘Repeat 1’ and ‘Repeat 2’. Thus, that lead to believe that those two sequences are one. That sRNA genomic sequence was searched alongside with the PBS sncRNA. The results showed that the amount of PBS sncRNA sequence present in the sequence data is 7 to 32 times lower in non-infected libraries and 32 to 156 times lower in infected libraries, than the sequences for ‘Repeat 1’, ‘Repeat 2’ and sRNA genomic sequence. Therefore, the sRNA genomic sequence is a new PBS sncRNA sequence.

Additional research suggests that the sRNA might well be part of a recent class of sRNAs, tRNA-derived fragments (tRFs). They are sncRNAs that are derived from tRNAs and are usually halves of the 5’ end or 3’ end of a tRNA, but there can also exist halves from the middle of the tRNA. Our sequence seems to be part of a new subcategory of these halves. The tRFs have been associated with a lot of processes like RNA silencing, cancer, or, response to viral infections.

Regarding the ‘Repeat 3, we determined the relative position of its sequence in the context of the reported secondary structure of the 18S rRNA. The position is located at the base of a stem-loop that is compatible with the canonical miRNA processing pathway. Studies report the existence of rRNA-derived miRNAs in humans and mice. Some of these rRNA-derived miRNAs have been linked to piRNAs and might work as a small guide RNAs.

The difficulties in the detection of differential expression of the target sRNAs could be due to the specificities of the tRNAs and rRNAs. They are among the most abundant in the cells. Both have several levels of structure and the tRNAs have several species and modifications. All these aspects lead to a greater difficulty in detecting fragments from these RNAs by standard PCR methods.

This dissertation has brought to light a new PBS sncRNA sequence and a rRNA-derived miRNA, which can have an impact in a different understanding of the HIV-1 infection and the relation between virus and host. Future assays are necessary to determine how much the discovery of a new PBS sncRNA sequence changes our understanding of the HIV-1 infection and what is the link between the rRNA-derived miRNA and the HIV-1, to determine what is the exact function and how it affects the host and the HIV-1.

Keywords

$tRNA^{Lys}_3$, tRFs, rRNA derived-miRNA, PBS sncRNA, HIV.

Index

Acknowledgments II

Resumo extenso III

Resumo VI

Summary VIII

Figures XII

Tables XIII

Abbreviations XIV

1. Introduction 1

1.1. CD4+ T cells and HIV-1 1

1.2. Small non-coding RNAs 3

1.3. Detection of sRNAs 4

1.4. Novel sRNAs from sequencing data 5

1.5. Aims 7

2. Methods 8

2.1. Cell culture (with TNF-α assay) and RNA extraction – J-Lat 8

2.1.1 Culturing cells 8

2.1.2 TNF-α assay 8

2.1.3 RNA extraction 8

2.2. Primers design 8

2.3. RT-PCR 8

2.4. Bioinformatic analysis 9

2.4.1 Repeats counting 9

2.4.2 Repeats locations 10

2.4.3 target sRNAs alignments 10

3. Results 11

3.1. Establishing J-Lat as a model for regulated expression of HIV-1 transcripts 11

3.1.1 Assessment of reactivation efficiency using fluorescence microscopy 11

3.1.2 Detection of HIV-1 transcripts by RT-PCR 12

3.2. Detection of novel sRNAs in response to HIV-1 reactivation in J-Lat cells 13

3.3. Detection of expression levels of novel sRNAs in response to HIV-1 infection in primary cells . 14

3.4. Bioinformatic analysis of target sRNAs 15

3.4.1 Presence of target sRNAs in small RNA-Seq libraries of human naïve T CD4+ cells 15

3.4.2 Mapping of possible genomic origins for the target sRNAs 17

4. Discussion 23

5.1. References 27
6. Supplementary Information... 31
6.1. Oligonucleotide sequences 31

Figures

Fig.1.1 - HIV provirus integrated genome in J-Lat. 3

Fig.1.2 – Stem loop RT-qPCR method. 5

Fig.1.3 - Poly A tailing method for cDNA synthesis and PCR. 5

Fig.3.1 - Evaluation of the response of J-Lat cells to TNF-α treatment using fluorescence microscopy. 12

Fig.3.2 - Detection of HIV-encoded transcripts in J-Lat cells 13

Fig.3.3 - Detection of target sRNAs in J-Lat cells. 14

Fig.3.4 - Detection of target sRNAs in primary CD4+ T cells. 15

Fig.3.5 - Alignment of “Repeat 1”, “Repeat 2” and “Repeat 3” sequences with HIV-1 vector pNL4-3. 19

Fig.3.6 - Secondary structure of the human tRNA Lys3, primer of the HIV-1 reverse transcription. 20

Fig.3.7 - Exact matches of “Repeat 1”, “Repeat 2”, Genomic sequence and PBS in the sequencing data. 21

Fig.3.8 - Secondary structure model for region V5 between human gene bp 677-910. 22

Tables

Table.1.1 – Small RNA libraries of stimulated human naïve CD4 T cells. 6

Table.1.2 – The target small RNA sequences. 6

Table.3.1 - Assessment of RNA quantity and quality from J-Lat TNF-α treated and control cells using UV spectrophotometry. 12

Table.3.2 - Description of the small- RNA-seq libraries used for analysis. 16

Table.3.3 - Count of target sRNA sequences in the sequencing data. 16

Table.3.4 - Locations of the sRNA target sequences in the Human Genome. 18

Table.3.5 - Count of the sRNA genomic sequence from the tRNA Lys 3 and PBS sncRNA sequence in the sequencing data. 20

Abbreviations

Ag – antigen.

APC – antigen presenting cell.

ARH12 – ras homolog family member A.

bp – base pairs.

CCA – cytosine-cytosine-adenine.

CCR5 – C-C chemokine receptor type 5.

CD4 – cluster of differentiation 4.

CD28 – cluster of differentiation 28.

CD80 – cluster of differentiation 80.

CD86 – cluster of differentiation 86.

CDS – Oligo-dT adaptor.

CPM – Counts per million.

CXCR4 – C-X-C chemokine receptor type 4.

DNA – deoxyribonucleic acid.

dsRNA – double-stranded RNA.

EDTA – Ethylenediaminetetraacetic acid.

eGFP – enhanced GFP.

Env – Short for envelope. Protein for viral envelope.

ERV – Endogenous retrovirus.

ERV-MaLR – Endogenous Retrovirus-Mammalian apparent LTR retrotransposon.

FBS – Fetal bovine serum.

Fw – Forward.

Gag – Group-specific antigen.

GC content – guanine-cytosine content.

GFP – green fluorescence protein.

gRNA – guide RNA.

HERV – human endogenous retrovirus.

HERV-K – human endogenous retrovirus K.

HIV – human immunodeficiency viruses.

IL-2 – interleukine -2.

LINE – long interspersed elements.

LTR – long terminal repeats.

m_1 A – 1-methyladenosine.

MHC II – Major Histocompatibility Complex II.

miRNA – micro RNA.

mRNA – messenger RNA.

Nef – Negative regulatory factor.

NF-κB – nuclear factor kappa-light-chain-enhancer of activated B cells.

NGS – Next-generation sequencing.

No-RT – No reverse transcription.

NTC – Non-Template Control.

ORF – open reading frame.

PBS – primer binding site.

PBS – Phosphate-buffered saline.

PCR – polymerase chain reaction.

piRNA – piwi-interacting RNA.

PIWI – P-element induced wimpy testis.

Poly A – poly adenylated.

ppt – polypurine tract.

pre-miRNA – precursor miRNA.

pri-miRNA – primary miRNA.

p-TEBb – human positive transcription elongator factor b.

qPCR – quantitative PCR.

Rev – regulator of expression of viral proteins.

RISC – RNA induced silencing complex.

RNA – ribonucleic acid.

RNase H – ribonuclease H.

RPMI medium – Roswell Park Memorial Institute medium.

rRNA – ribosomal RNA.

R sequence – direct repeats of ends of viral RNA.

RT – reverse transcription.

RT-PCR – Reverse Transcription-Polymerase Chain Reaction.

Rv – Reverse.

SAM file – Sequence Alignment Map file.

SINE – short interspersed element.

siRNA – short or small interfering RNA.

sncRNA – small non-coding RNA.

snoRNA – small nucleolar RNA.

snRNA – small nuclear RNA.

sRNA – small RNA.

SVA elements – SINE-R, VNTR and Alu elements.

TAE – Tris based, acetic acid and EDTA.

Tat – transactivator protein.

TCR – T cell receptor.

Th – T helper.

Tm – Melting temperature.

TNF-α – tumour necrosis factor alpha.

Treg – regulatory T.

tRNA – transfer RNA.

$tRNA^{Lys}_3$ – transfer RNA lysine 3.

UCSC genome browser – University of California, Santa Cruz genome browser.

UV – Ultraviolet.

Vif – Viral infectivity factor.

VNTR – variable number of tandem repeats.

Vpr – Viral protein R.

Vpu – Viral protein U.

1. Introduction

1.1. CD4+ T cells and HIV-1

CD4+ T cells are a crucial element in the human immune response. They are crucial due to their capacity to help B cells make antibodies, to induce macrophages to develop enhanced microbicidal activity, to recruit neutrophils, eosinophils and basophils to sites of infection and inflammation[1]. Mature T cells are produced in the thymus and released into the bloodstream in low numbers. These cells are considered immunological naïve, because they have not yet encountered the appropriate antigen[2]. Naïve CD4+ T cells can differentiate into Th1, Th2, Th17 and Treg (regulatory T) cells, as a response to a pattern of signals received during their initial interaction with an antigen presented by an APC (Antigen Presenting Cell) using its MHC II (Major Histocompatibility Complex) to the CD4+ T cell TCR (T cells receptor), becoming effector cells[2,3]. The TCR combined with CD3 forms the TCR complex necessary for the activation of the naïve T cells, because the TCR does not contain signalling domains and therefore needs the multisubunit signalling apparatus, that is the CD3[4]. There can also exist costimulation to turn naïve cells into effector cells. CD28 and its ligands CD80 and CD86, are expressed on professional antigen-presenting cells[5]. Costimulation by CD28, synergizes with signalling through the TCR, leading to T cell activation by enhancing gene expression, increasing proliferation and interleukine-2 (IL-2) production, giving protection from signal-1-induced apoptosis to effectively promote the progression of T cells from naïve to effector cells and memory populations of Th1 and Th2 phenotypes[5]. Th1 cells are critical for immunity to intracellular microorganisms and Th2 cells for immunity to a great number of extracellular pathogens, including helminths[1,6]. Immunity against microbes, like extracellular bacteria and some fungi is performed by Th17 cells[7]. The last one, Treg cells, are responsible for suppressing potentially deleterious activities of Th cells[8].

HIV main target is the CD4+ T cells. The interaction of the HIV with the chemokine coreceptors, CCR5 or CXCR4 allows its entry[9]. That entry is characterized by the fusion of membranes of the host cell and the virion. With the fusion of the contents the virion can enter the cell and set the stage for the reverse transcription[10,11].

Reverse transcription in HIV-1 takes place in newly infected cells and the reverse transcriptase contains the enzymatic activities required for the reverse transcription to occur. These are a DNA polymerase than can copy an RNA or DNA template, and an RNase H that is able to degrade RNA, if it is part of an RNA-DNA duplex. Genomic RNA is plus-stranded and it will serve as a template for the synthesis of the first DNA strand, the minus strand, by using as a *primer* an host $\text{tRNA}^{\text{Lys}}{}_3$, whose 3' end is base paired to a complementary sequence near the 5' end of the viral RNA called the primer binding site (PBS), that as about 180 nucleotides from the 5' end of genomic RNA. DNA synthesis creates an RNA-DNA duplex that will serve as a substrate for the RNase H. The RNase degradation removes the 5' end of the viral RNA, exposing the newly synthesized minus strand of DNA. The ends of the viral RNA are direct repeats with the designation, R. They act as a bridge that permits the newly synthesized minus-strand DNA to be transferred to the 3' end of the viral RNA. Two copies of the viral RNA genome, one of the strands, the first or the minus-strand DNA transfer can involve the R sequence at the 3' ends of either of the two RNAs[11].

With the transfer complete, minus-strand synthesis can continue along the length of the genome with the RNase H degradation also happening. A purine rich sequence called polypurine tract (ppt), that is resistant to the RNase H degradation, serves as a *primer* for the initiation of the plus strand of DNA.

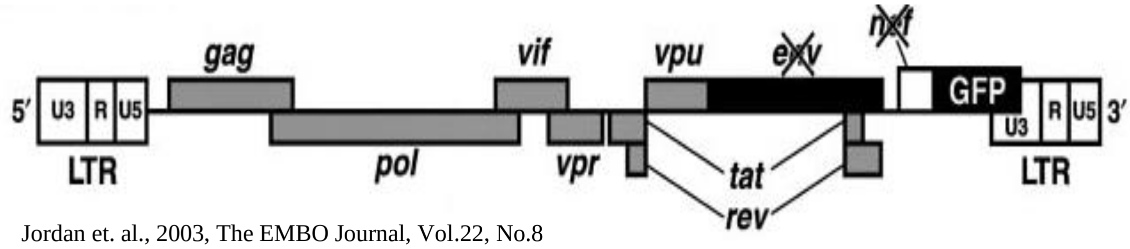
The HIV-1 has two ppt, one located near the 3' end of the RNA and the other one near the middle of the genome. The first one is essential for viral replication and the other one increases the ability of the virus to complete the synthesis of the plus-strand of DNA, even tough is not essential[11]. On the generation of the plus-strand the minus-strand of DNA is copied, but also the first 18 nucleotides of the tRNA$^{Lys}{}_3$ *primer*. After the tRNA is copied into DNA, it becomes a substrate for RNase H. HIV-1 cleaves the tRNA one nucleotide from the 3' end, leaving a single A ribonucleotide at the 5' end of the minus strand[11].

Minus-strand DNA synthesis could proceed along the entire length of the RNA genome, but genomic RNAs found in virions are usually nicked. The existence of a second copy of the RNA genome allow minus-strand DNA synthesis to transfer to the second RNA template, thus bypassing the nick in the original template, contributing to an efficient recombination. When the minus-strand DNA synthesis nears the 5' end of the genomic RNA, the PBS is copied, setting the stage for the plus-stranded transfer. The 3' end of the plus strand DNA contains 18 nucleotides that were copied from the tRNA *primer*, which are complementary to the 18 nucleotides at the 3' end of the minus-strand DNA that were copied from the PBS. These two complementary sequences anneal, and DNA synthesis extends both minus and plus strands to the ends of both templates. HIV-1 plus-strand DNA can be also synthesized from multiple initiation sites[11].

The reverse transcription creates a DNA product that is longer than the derived RNA genome. Both ends of the DNA contain sequences from each end of the RNA, U3 from the 3' end and U5 from the 5' end. Thus, each end of the viral DNA has the same sequence, U3-R-U5, that are the long terminal repeats (LTRs), that after integration, will be the ends of the provirus. The full-length linear viral DNA ends are defined at the U5, by the RNase H cleavage that removes the tRNA *primer*, and on the U3 end, by the cleavages that that generate and remove the ppt *primer*. The specificity of the RNase H cleavage is important, because the ends of the linear viral DNA are the substrates for integration[12].

HIV latency can come from different ways, one of those is a consequence of the normal physiology of CD4+ T lymphocytes, because these lymphocytes can at any given time be in a resting G_0 state. These resting cells are deeply quiescent cells with a very low metabolic rate and a unique morphology characterized by a small cytoplasmatic volume[13]. HIV gene expression is dependent on inducible host transcription factors that are only transiently activated after exposure to an antigen[12]. In adults, about half of the resting cells are naïve, having yet to encounter an appropriate antigen (Ag). Antigen-driven responses involve a burst of cellular proliferation and differentiation that gives rise to effector cells. The effector cells usually die quickly, but the ones that survive return to a resting state of G_0 and are designated as memory cells, because they already have encountered an antigen. Apart from that, they have an altered pattern of gene expression enabling long-term survival and rapid response to the appropriate antigen[13]. Therefore, the result with the HIV infection is a stably integrated, but transcriptionally silent form of the virus in the lymphocyte cells, in a cell whose function is to survive for long periods of time[14]. So, there can be several pathways for HIV latency. Preactivation latency that involves the direct infection of resting CD4+T cells. Postactivation latency involving either infection of activated cells and reversion to a resting state, or, direct infection of T cells transitioning from effector memory to resting central memory cells after T cell receptor (TCR) stimulation[15].

Regarding the passage from HIV latency to active HIV infection with expression of viral transcripts, one of the options is by inducing the NF-κB through the tumour necrosis factor alpha (TNF-α) stimulation. The NF-κB is essential for efficient transcription from the HIV long terminal repeat (LTR)[15]. The HIV latency is present in cells lines for possible HIV study as biological models. One example is the J-Lat cell line that is derived from the Jurkat cell line. The J-Lat has an integrated, but transcriptionally latent HIV provirus with green fluorescence protein (GFP) gene replacing nef coding sequence and env with a frameshift mutation (Fig.1.1). The latent state can be viral re-activated by using the TNF-α to allow the expression of viral transcripts[16].



Jordan et. al., 2003, The EMBO Journal, Vol.22, No.8

Fig.1.1 - HIV provirus integrated genome in J-Lat.

After the viral genome is transcribed the HIV provirus is integrated into the host chromatin where it becomes subject to transcription by the host RNA polymerase II. However, the RNA polymerase fails to travel far on the viral template. The short viral transcripts are unable to support viral replication. The HIV encodes an essential accessory protein to overcome this restriction, the transcriptional transactivator protein (Tat), along with a cofactor, the human positive transcription elongation factor b (p-TEBb)[17]. This way transcription from the HIV-1 long terminal repeat (LTR) can occur by interaction with the viral transactivation response (TAR) element[18]. With the retroviral pre-mRNA transcribed, it is spliced into viral mRNAs that exhibit all the characteristics of cellular mRNAs, since they bear a 5’ cap structure and a 3’ poly(A) tail. Alternative splicing in HIV gives rise to over 30 different mRNAs species that are then exported to the cytoplasm by different pathways. These species have been grouped in mRNAs that do not undergo splicing and are full-length transcripts that encode for Gag and Gag-Pol polyproteins, singly spliced transcripts that generate viral proteins Env, Vif and Vpu and finally fully spliced transcripts that express Rev, Tat , Vpr and Nef. The HIV-1 protein synthesis depends on the host cell translation machinery for ribosomes, tRNAs, amino acids and all the necessary initiation, elongation and termination factors. Cleavage of polypeptides and their assembly as an HIV-1 particle along with the budding, are the last steps that will then lead to the release of the HIV-1 particle from the infected cell, that is then able to infect other susceptible cells around[9,19,20].

1.2. Small non-coding RNAs

The number of known RNAs has increased a lot over the years, discovering different types of RNA such as small RNAs, thanks to cloning and sequencing of size fractionated RNAs[21]. The functions range a lot, and small RNAs are linked to almost every biological process, developmental timing, cell differentiation, cell proliferation, cell death, metabolic control, transposon silencing and antiviral defence. RNAs are divided in coding RNA and non-coding RNA, which then are divided into other sub-categories[21].

There are a lot of non-coding RNAs, like ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) related to translation, which were the first to be discovered. Besides these first two, many others have been discovered, small nuclear RNAs (snRNAs) responsible for splicing, small nucleolar RNAs (snoRNAs) for methylation of rRNAs and guide RNAs (gRNAs) regarding RNA editing. From these non-coding RNAs there are the small non-coding RNAs of which the snRNAs, snoRNAs and gRNAs are part of, due to their size. The gRNAs are around 50-70 nucleotides while the other two, less than 200 nucleotides. The small non-coding RNAs range in size from micro RNAs (miRNAs) of 21 to 25 nucleotides to the sncRNAs that are no larger than 200 nucleotides in length[22]. Small non-coding RNAs are responsible for gene regulation at the transcriptional and post-transcriptional level, RNA processing and modification[23].

From the sncRNAs with the smallest length some of the examples are, piwi-interacting RNA (piRNA), short interfering RNA (siRNA) and micro RNA (miRNA). piRNA are made from a precursor of 25 to 27 nucleotides in length, a single stranded precursor transcript unlike the double strand precursors of the miRNA and the siRNA. They are small silencing RNAs and bear 2'-O-methyl-modified 3' termini and they guide PIWI-clade Argonautes (PIWI proteins) instead of AGO-clade proteins that are used for the miRNA and siRNA pathways. The piRNA are about 21 to 35 nucleotides in length. piRNAs silence transposons by repressing their transcription or by cleaving their mRNAs. Other functions have been associated with piRNAs like viral defence[24]. The siRNAs are created by the cleavage of a longer double-stranded RNA and tend to have a 2-nucleotide overhang on the 3' end of each strand. The cleavage is performed by Dicer, a RNase III-like enzyme. The silencing by the siRNA happens with the assistance of a multiprotein component complex named RISC (RNA induced silencing complex). The siRNA strands are then separated and the 5' end that is more stable is most likely to be integrated in the active RISC complex. The mRNA is cleaved by the action of catalytic RISC protein through the assistance of the antisense single-stranded siRNA that guides and aligns the RISC complex to the target mRNA[25,26]. miRNAs function consists mainly on binding to mRNA to prevent protein production by translation repression or through mRNA cleavage. The choice of translation inhibition or mRNA cleavage depends on the level of complementarity with the miRNA[25,26]. Like siRNAs they are about 19 to 20 nucleotides long with 5'-phosphate and 3'-hydroxyl ends. Mature miRNAs are formed by a two-step cleavage of primary miRNA (pri-miRNA), first cleavage to precursor miRNA (pre-miRNA) by Drosha, followed by another cleavage to miRNA, this time by Dicer, duplex that unwinds and assembles with the effector RISC. Drosha and Dicer are two ribonuclease III nucleases. The main difference between the miRNAs and the siRNAs is that the miRNA is derived from a 60 to 70 nucleotide RNA hairpin and the siRNAs from a long double-stranded RNA (dsRNA)[25,26].

1.3. Detection of sRNAs

Since sRNAs with a very short length have been discovered, methods have been designed to better detect them. The stem loop RT-qPCR method is a method used for detection of miRNAs and its quantification. A specific stem-loop RT *primer* and forward *primer* are designed for each unique sRNA[27]. The stem-loop RT *primer* has to have nucleotides of the 3' end complementary to the 5' end of the sRNA for cDNA synthesis and the forward *primer* for PCR has 80-90% of its sequence equal to the sRNA sequence (Fig.1.2). Given that, if a sRNA is too short in sequence length (Less than 18 nucleotides), this two *primers* have a very high chance of hybridizing and form heterodimers and give a positive signal, even when there is not one.

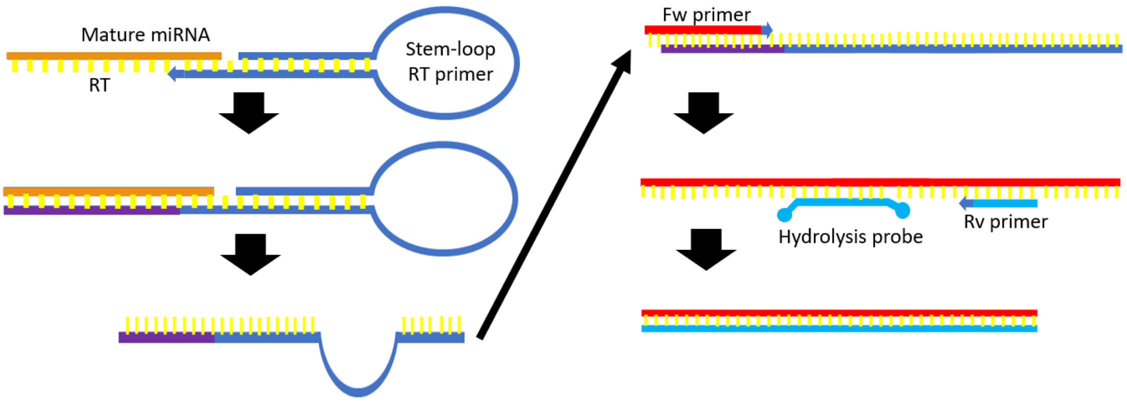


Fig.1.2 – Stem loop RT-qPCR method.

The traditional reverse transcription with PCR would also have the problem with sRNAs with a very short sequence in length (Less than 18 nucleotides), because the necessary *primers* for the PCR amplification would contain almost the entire sequence of the sRNA. Thus, having a very high chance of hybridizing amongst themselves, create heterodimers, but most importantly give a positive signal when there could not be one.

The Poly A tailing for cDNA synthesis and posterior PCR amplification is a method that might well solve problems of rightly detecting very short (Less than 18 nucleotides in length) sRNAs. The Poly A tailing method also relies, like in the traditional reverse transcription, of an anneal of an oligo-dT to the Poly A tail that is added to the target sRNA 3' end. The difference in this method is that the Poly A tailing oligo-dT (CDS) is different, because it does not only have the several thymines to anneal to the Poly A tail, but also has more nucleotides that form a sequence on which on the PCR, an Universal Reverse (ARH12) *primer* will anneal to a specific part of that sequence Fig.1.3). Therefore, the generated cDNA is the only one that is most likely to anneal to the specific Universal Reverse *primer* and produce a signal, making these method particularly better than the others for the correct detection of very short (Less than 18 nucleotides) sRNAs.

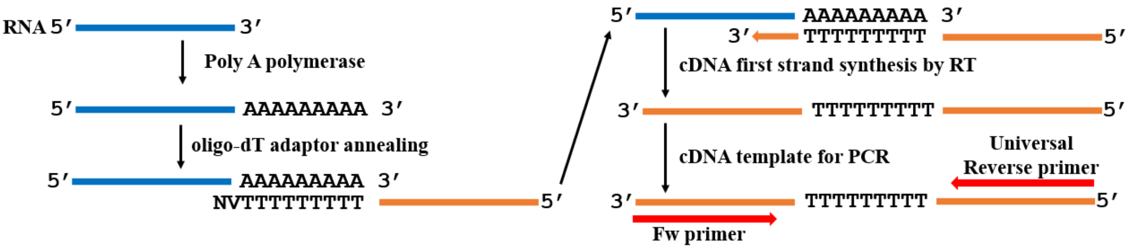


Fig.1.3 - Poly A tailing method for cDNA synthesis and PCR.

1.4. Novel sRNAs from sequencing data

In previous studies in the laboratory, human naïve CD4 T cells were purified from the blood of seronegative individuals and stimulated via T cell receptor (TCR), infected stimulated and non-stimulated cells with either HIV-1 or HIV-2 and produced small RNA libraries for expression profiling by next-generation sequencing (NGS) (Table.1.1). The infected cells consisted of infection with clones, HIV-1_{NL4-3} regarding an infection of HIV-1 and for the infection of HIV-2, the clone HIV-2_{ROD}. The infection was performed for 24 hours in the naïve CD4 T cells. Therefore, stimulation only occurred after. Stimulation of the TCR on the CD4 T cells was performed *in vitro* for 72 hours using immobilized anti-CD3 and soluble anti-CD28 antibodies[28].

Table.1.1 – Small RNA libraries of stimulated human naïve CD4 T cells.

Treatment		Library ID
Stimulated	Non-infected	GHM-17 GHM-19
	HIV-1 infected	GHM-20 GHM-22
	HIV-2 infected	GHM-23 GHM-24 GHM-25

From these libraries, small RNAs (Table.1.2) were identified, from viral and human origin, whose expression is altered in activated CD4+ T cells and/or infected by HIV-1 or HIV-2. Among these small RNAs, some appeared to have new functional categories, distinct from the miRNAs, with characteristics that suggest the potential to represent a new defense mechanism of the host against HIV infection.

Table.1.2 – The target small RNA sequences.

sRNA	Sequence (5'-3')	Size (bp)
'Repeat 1'	TTCAGGTCCCTGTTCGG	17
'Repeat 2'	TTCATGTCCCTGTTCGG	17
'Repeat 3'	TTGGATCTTGGGAGC	15

These small RNAs have been associated with retrotransposons. Retrotransposons can move from place to place in a genome by reverse transcription of an RNA transposition element. There are a few classes of retrotransposons. Long terminal repeats (LTR) retrotransposons that have a direct few hundred base pairs long at each end. LTR retrotransposons have a gag and pol gene and in some a gene like a retroviral env gene. Each one of the genes codes for a primary translation product that is processed into viral proteins needed for virion formation, infection of cells and reverse transcription of new DNA. Apart from the LTR retrotransposons there are the non-LTR retrotransposons, that do not have terminal repeats either direct or indirect. Long interspersed elements (LINEs) are the bigger non-LTR retrotransposons and have two open reading frames (ORFs). One encodes for an RNA-binding protein and the second for a nuclease, usually related to an apurinic-apyrimidinic repair endonuclease, a reverse transcriptase and in some cases and RNase H domain. Since the reverse transcription usually ends before the first strand of DNA is complete, the promotor necessary for transcription of the RNA transposition intermediate will have been lost, making the new LINE copies unable to transpose. If a mutation affects a complete LINE that influences that promoter it would still made the LINEs unable to transpose, therefore only very few LINEs are able to transpose. Short interspersed elements (SINEs) are also non-LTRs and have a high copy numbers even though they do not have coding capacity, so they usually use LINE machinery for their own ends. In most cases, they are hybrid elements with the 5' sequence derived from a tRNA and the 3' region related to the 3' end of a LINE. The most predominant SINEs in humans are the Alu elements, named due to site for the restriction enzyme AluI. SVA elements are non-LTRs and non-autonomous retrotransposons and they are so named, because they have a composite structure, with a 5' Alu-like sequence , a variable number of tandem repeats (VNTRs) region , a sequence derived from the

3’end of the human endogenous retrovirus HERV-K, including the LTR and a 3’ poly(A). Evidence suggests that they are mobilized by the LINEs[29].

1.5. Aims

The aim of the present work is to test the expression alterations of the novel classes of small RNAs identified in the laboratory and characterize their functional impact on the interaction between host cells and HIV virus. The J-Lat cell line provides a biological model capable of studying the expression of viral transcripts in cultured cells. The J-Lat cell line, as previously stated, is a Jurkat-derived cell line with an integrated, but transcriptionally latent HIV provirus, on which green fluorescence protein (GFP) gene replaces the nef coding sequence and env has a frameshift mutation (Fig.1.1)[16].The latent HIV provirus could be re-activated by TNF-α to express the viral transcripts. Thus, control cells and TNF-α treated cells would serve as the biological model to analyze if the expression of viral transcripts would lead to an increase or any kind of expression alterations of the target sRNAs. As it was already referred, the target sRNAs have been associated with retrotransposons, furthermore, the sample sequences of the target sRNAs have been more associated with the human genome with a better alignment than with the HIV genome. The search for an origin and number of those origin locations, could provide a better understanding of what these sRNAs are and what their functions might be. Finally, it could also provide insight on why there is accumulation of the target sRNAs in response to HIV infection and if their function could be some sort of mechanism against the HIV infection or rather a mechanism that assists the HIV infection.

2. Methods

2.1. Cell culture (with TNF-α assay) and RNA extraction – J-Lat

2.1.1 Culturing cells

Immortalized human T lymphocyte J-Lat cells, a Jurkat derived cell-line[30], were cultured at 1×10^7 cells/mL in RPMI-1640 (1x) medium + GlutaMAX (Gibco) supplemented with 10% Fetal bovine serum (FBS) (Gibco).

2.1.2 TNF-α assay

Latent HIV provirus in J-Lat cells was activated by stimulation with TNF-α (Sigma-Aldrich) with a concentration of 10 ng/µL for 24 hours in a P35 Petri dish. J-Lat cells without the TNF-α treatment were also prepared in a P35 Petri dish to serve as a control.

The cells were observed on the fluorescence microscope, DMI4000 coupled with a Leica DFC365 FX using an 20x NA 0.7 objective and the Leica LAS X software for photo acquisition, excitation at 450-490 nm, for evaluation of the response of J-Lat cells to TNF-α treatment.

2.1.3 RNA extraction

TRIzol reagent (Invitrogen) was used for isolation of total RNA from the TNF-α treated cells and control cells, according to the manufacturer's instructions. Finally, the pellet was resuspended in 40 µL of RNase free water + EDTA 0.1 mM and incubated at 57,5 ºC for 10 minutes, according to the manufacturer's instructions, in a Labnet AccuBlock Digital Dry Bath. RNA concentration and absorbances were then measured on the NanoDrop, Spectrophotometer ND-1000 (NanoDrop Technologies) with the help of the software, NanoDrop 1000 3.8.1 (NanoDrop Technologies). After that, RNA samples were stored at -80ºC.

2.2. Primers design

Oligonucleotide sequences (STAB VIDA) were used in Reverse Transcription (RT), Poly A tailing and PCR (Supplemental Table.1 and Supplemental Table.2). Only the last two oligonucleotide sequences (Supplemental Table.2) were required to be design, since the rest already existed in the lab.

For the design of the *primers* two tools were used, one from Thermo Fisher Scientific, the Tm calculator and the other from Integrated DNA Technologies, the Oligo Analyzer 3.1.

2.3. RT-PCR

For the detection of viral transcripts cDNA synthesis was done using NZY M-MuLV Reverse Transcriptase (NZYTech) according to the manufacturer's instructions. A control with no RT was performed.

For the detection of target sRNAs cDNA synthesis was done using Poly (A) Polymerase, Yeast (Thermo Scientific) according to the manufacturer's instructions, plus the NZY M-MuLV Reverse Transcriptase (NZYTech) according to the manufacturer's instructions. Of note, the Oligo-dT adaptor (CDS) (STAB VIDA) was added before the reaction termination step of the Poly A tailing, so that it could ligate to the Poly A tailed sequence, before doing the Reverse Transcription. A control with no RT was performed.

PCR was performed in a C100 Touch Thermal Cycler (BIO-RAD) with NZYTaq II DNA Polymerase (NZYTech) according to the manufacturer's instructions. A control with no template was performed for each target.

PCR products were separated by a 2% agarose gel. 1/6 volume of 6x Loading Dye to the PCR product sample of each tube. Total volume of each sample used in the electrophoresis run. NZYDNA Ladder VI (NZY Tech) used according manufacturer's instructions. The electrophoresis run was done at 100 V for 90 minutes. The bands were visualized using ChemiDoc XRS+ System (BIO-RAD) with Image Lab Software (BIO-RAD) that also allowed to capture images.

2.4. Bioinformatic analysis

Bioinformatic analysis was performed using simple Linux commands.

2.4.1 Repeats counting

From sequencing data of uninfected stimulated CD4+ T cells (GHM-17 and GHM-19) and HIV-1 infected stimulated CD4+ T cells (GHM-20 and GHM-22), the filtered data was analyzed using the following commands for different types of counts regarding the target sRNAs on fastq. files.

Number of reads with "x" length

```
awk -F "" 'NR%4==2 {print NF}' FICHEIRO | sort -nk 1 | uniq -c | awk 'BEGIN{print "Size,   Frequency"} {print $2,$1}'
```

Exact match

```
grep -w "REPEAT" <GHM fastq. file name> | wc -l
```

Contained

```
grep "REPEAT" <GHM fastq. file name> | wc -l
```

Unique in contained

```
grep "REPEAT" <GHM fastq. file name> | uniq | wc -l
```

Number of reads in a fastq. file (in total):

```
cat <GHM fastq file name> | wc -l   - Gives the “x”.

expr x / 4
```

2.4.2 Repeats locations

From the SAM files of each GHM obtained from aligning the three target sRNAs (‘Repeat 1’, ‘Repeat 2’ and ‘Repeat 3’) to the human genome, version Feb. 2009 (GRCh37/hg19).

Genomic locations of target sRNAs

```
samtools view -F 4 <GHM SAM file name> | grep "REPEAT" | awk '{print $3, $4, $10}' | sort | uniq -c
```

The positions of the exact matches for each target sRNA were then searched on the UCSC Genome Browser. The version of the Human Genome used, was the same that was used for the alignment to create the SAM files, version Feb. 2009 (GRCh37/hg19). Given that, each SAM file only gave the beginning of each position, each position was the beginning of the interval to put on the UCSC Genome Browser so that the entire sequence was browsed. Therefore, each interval began with the position from the SAM file and then the final position of the interval was 16 nucleotides ahead for the ‘Repeat 1’ and ‘Repeat 2’ and 14 nucleotides ahead for the ‘Repeat 3’.

2.4.3 target sRNAs alignments

Alignments were made by using the National Center for Biotechnology Information (NCBI) tool Basic Local Alignment Search Tool (BLAST), Nucleotide BLAST. Human genome used, version Feb. 2009 (GRCh37/hg19). HIV-1 vector used, pNL4-3 (AF324493).

3. Results

3.1. Establishing J-Lat as a model for regulated expression of HIV-1 transcripts

To answer the question of whether the accumulation of the target sRNAs is induced in response to the expression of HIV-1 transcripts, the J-Lat cell line was chosen as a biological model that could permit the safe expression of viral transcripts, since it has an integrated, but transcriptionally latent HIV provirus. Thus, when viral reactivation occurs it will express viral transcripts and permit the analysis to check if there was an increase in expression of those viral transcripts. Therefore, J-Lat cells existent in the lab were thawed and cultured to perform reactivation assays using TNF-α to induce the transcription of the latent HIV provirus integrated in the cells, to mimic the effect of an active HIV infection.

3.1.1 Assessment of reactivation efficiency using fluorescence microscopy

We began by establishing the appropriate conditions for the reactivation of latent HIV-1.

24h after splitting, cells were treated with 10 µg/mL of TNF-α, or, left in normal medium for control. A sample was taken for TNF-α treatment and another sample to serve as a control, cells without TNF-α. 24 hours later, the cells were observed in the fluorescence microscope to check if the TNF-α had been able to activate the cells to express the viral transcripts. The confirmation of that fact would be the presence of green fluorescence, due to the presence of the GFP coding sequence in the HIV provirus. The sample with TNF-α had about 75% of cells expressing the GFP (Fig.3.1), while the sample without the TNF-α had about 1% of cells expressing the GFP (Fig.3.1), corresponding to the basal HIV-1 reactivation levels previously described, which is generally of lower intensity than the one observed upon TNF- α treatment, as can be observed in the comparison between panels C and D[16].

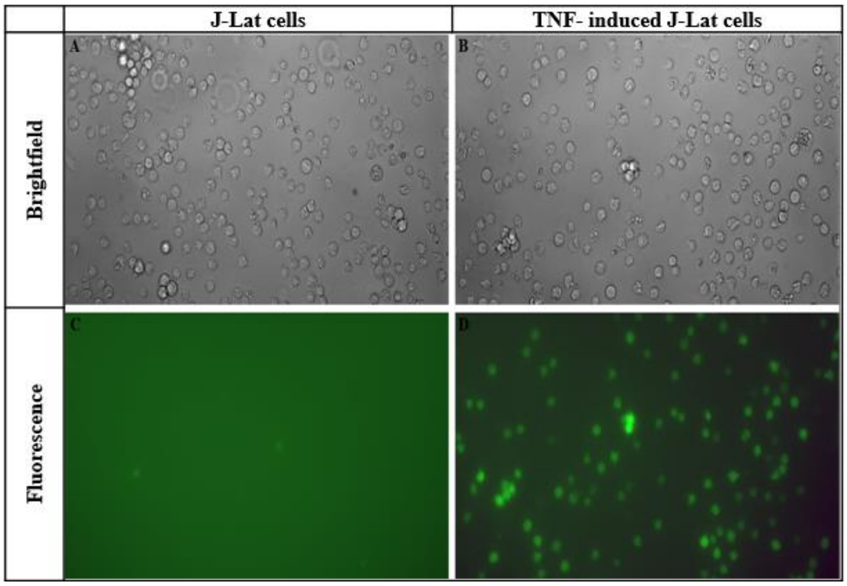


Fig.3.1 - Evaluation of the response of J-Lat cells to TNF-α treatment using fluorescence microscopy. Legend: Subclonfluent J-Lat cells were incubated for 24h with medium (A, C) or 10 µg/mL of TNF-α (B, D) and bright field microscopy (A, B) or fluorescence microscopy (C, D) images were acquired.

3.1.2 Detection of HIV-1 transcripts by RT-PCR

Having confirmed the efficient induction of GFP expression by fluorescence microscopy, we next sought to confirm the presence of HIV-1 transcripts by RT-PCR. For this purpose, TNF-α treated and control cells were used for RNA extraction using the TRIzol method. Three independent reactivation experiments were performed, and the extracted RNA samples were quantified by UV spectrophotometry (Table.3.1). The results showed that the RNA samples were of good quality to be used for cDNA synthesis.

Table.3.1 - Assessment of RNA quantity and quality from J-Lat TNF-α treated and control cells using UV spectrophotometry.

		Absorbance ratios	
Samples	ng/µL	260/280	260/230
J-Lat TNF-α (1)	215,4	1,97	1,8
J-Lat TNF Ø (1)	177	2,11	1,06
J-Lat TNF-α (2)	172,4	2	2,1
J-Lat TNF Ø (2)	259	2,01	2,1
J-Lat TNF-α (3)	134,6	2,02	1,75
J-Lat TNF Ø (3)	212,1	2,01	1,78

To confirm the expression of viral transcripts in the J-Lat cells, a PCR was performed on cDNA synthesized from the extracted RNA. mRNA of eGFP and Tat protein (an HIV protein) were detected alongside an endogenous housekeeping control, the U6 snRNA. The results showed that the Tat band and eGFP was present in TNF-treated J-Lat cells and in control cells, confirming that there is always a basal expression of the HIV provirus, but the intensity of the band was higher in the former (Fig.3.2).

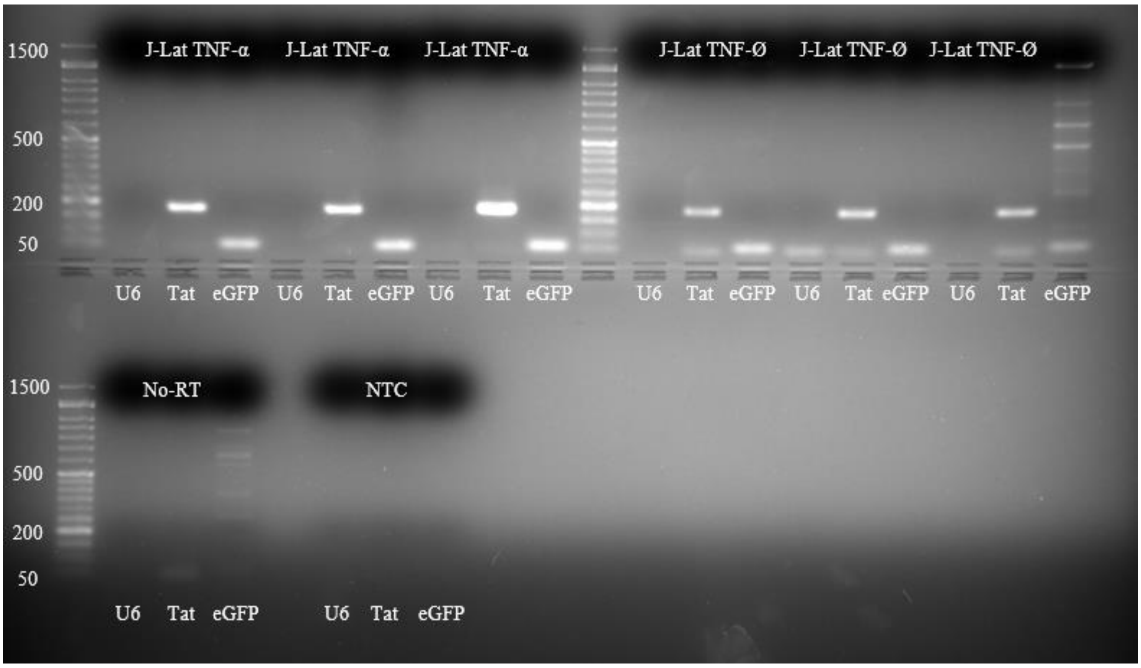


Fig.3.2 - Detection of HIV-encoded transcripts in J-Lat cells. Legend: Agarose gel electrophoresis of the PCR reactions of replicate cDNA samples from J-Lat cells treated with TNF-α and control samples. The expected amplicon sizes are: U6snRNA – 99 bp; Tat mRNA – 183 bp; eGFP mRNA – 55 bp.

In this assay, we had a problem with the detection of the U6snRNA, probably due to the *primers* or another technical aspect in performing the assay. However, since this was not a quantitative approach, even though no U6 band appeared, the presence of all the other bands was enough to confirm the presence of amplified cDNA and the expression of viral transcripts, thus satisfying our aims.

As the presence of viral transcripts in the J-Lat cells in response to TNF-α treatment was confirmed, it is possible to use these RNA samples to try to detect the expression of the target sRNAs and verify if there was an increase in their abundance in response to the presence of HIV-1 transcripts.

3.2. Detection of novel sRNAs in response to HIV-1 reactivation in J-Lat cells

Having RNA with viral transcripts from TNF-α treated J-Lat cells and control samples, we sought to assess the accumulation levels of target sRNAs on the presence of HIV-1 transcripts by a specific RT-PCR method. The cDNA synthesis was performed using the Poly A tailing method, which allowed more specificity and the ability to amplify the three specific sequences through PCR. Since the sequences length are very short (15, 17 nucleotides in length), other methods could lead to the production of unbiased results. mRNA of Tat protein (an HIV protein), the three target sRNAs were detected alongside an endogenous housekeeping control, the U6 snRNA.

The outcome of several assays did not produce the expected results. No significant difference in the specific target sequences (‘Repeat 1’, ‘Repeat 2’ and ‘Repeat 3’) was observed, between control samples and TNF-α J-Lat treated cell samples (Fig.3.3).The band intensity seemed constant between the two types of samples regarding the target sRNAs, with only one assay showing a slightly higher band intensity on the TNF-α treated J-Lat cell sample. The mRNA of the Tat protein only appeared in one of the assays, on the TNF-α treated J-Lat cell sample, even though the presence of viral transcripts had been confirmed in the former assay (Fig.3.2).

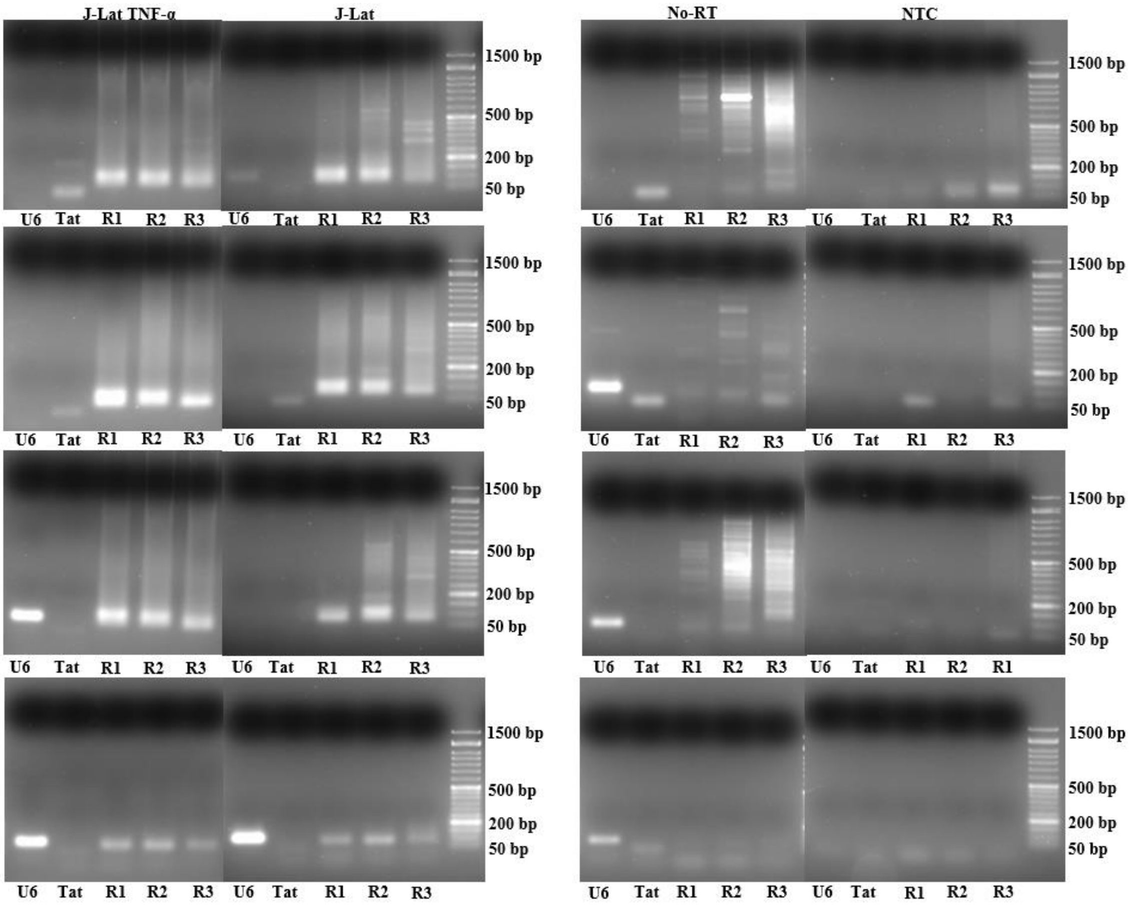


Fig.3.3 - Detection of target sRNAs in J-Lat cells. Legend: Agarose gel electrophoresis of the PCR reactions of replicate cDNA samples from J-Lat cells treated with TNF-α and control samples. The expected amplicon sizes are: U6snRNA – 99 bp; Tat mRNA – 183 bp; R1 target sRNA – 80 bp; R2 target sRNA – 81 bp; R3 target sRNA – 79 bp.

Thus, we were unable to obtain any clear evidence supporting an increased expression of our target sRNAs in J-Lat cells, even in the presence of HIV-1 transcripts. Given that this phenomena may only be observed in primary T cells, we next sought to detect the sRNAs in the original samples that were used to generate the sequencing libraries, which were still available in the lab.

3.3. Detection of expression levels of novel sRNAs in response to HIV-1 infection in primary cells

Given that RNA from primary cells used in earlier studies to generate the sequencing data libraries was still available, three different pairs of samples derived from individual buffy coats (the fraction of anticoagulated blood that contains most of the leucocytes and platelets obtained from density gradient centrifugation of the blood, in this case from blood donors) were used to try to detect the accumulation of novel sRNAs in response to HIV-1 infection. The pairs of buffy coats consisted on in vitro stimulated human naïve CD4+ T cells infected with the HIV-$1_{\mathbf{NL4-3}}$ clone and control samples.

The same specific RT-PCR method with Poly A tailing for cDNA synthesis was used. Alas, the results on the detection of mRNA of Tat protein (an HIV protein), the three target sRNAs alongside an endogenous housekeeping control, the U6 snRNA, were the same as the TNF-α treated J-Lat cell samples and control samples (Fig.3.4). In this case, the results besides being inconsistent, were also contrary to the hypothesis in two of the four assays. The control samples were the ones with a slightly higher band intensity of the target sRNAs. The mRNA of the Tat protein was not detected. The results were also inconsistent throughout the four assays in all the targets, from U6snRNA, mRNA of the Tat protein and the three novel sRNAs, regarding the primary cell samples.

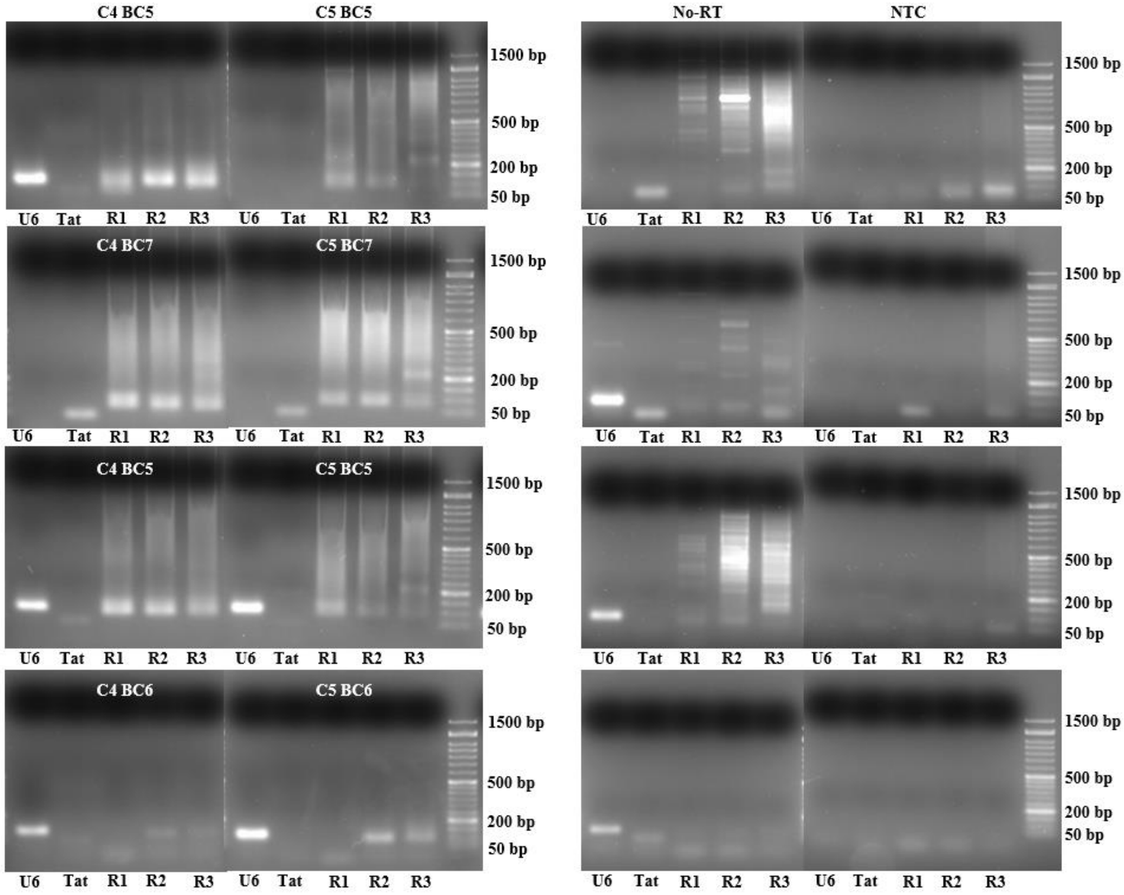


Fig.3.4 - Detection of target sRNAs in primary CD4+ T cells. Legend: Agarose gel electrophoresis of the PCR reactions of replicate cDNA samples from several pools (BC5, BC6 and BC7) of primary stimulated CD4+ T cells infected with an HIV-1 clone (C5), HIV-$1_{\mathbf{NL4-3}}$ and control samples (C4). The expected amplicon sizes are: U6snRNA – 99 bp; Tat mRNA – 183 bp; R1 target sRNA – 80 bp; R2 target sRNA – 81 bp; R3 target sRNA – 79 bp.

Given the inconsistent results in the RT-PCR assays performed to detect the target sRNAs, we decided to perform a more detailed bioinformatic analysis to confirm their expression pattern and to learn more about their potential origin.

3.4. Bioinformatic analysis of target sRNAs

3.4.1 Presence of target sRNAs in small RNA-Seq libraries of human naïve T CD4+ cells

The previous bioinformatic analysis performed in the lab identified a significant increase in the average expression levels of the target sRNA sequences in response to HIV-1 infection. However, this was a preliminary analysis that did not look deeply into the variability of results between different samples and did not explore in great detail the potential genomic point of origin of these sequences, nor their presence in the context of longer sRNA reads. Thus, we set out to reanalyze this data in more detail.

The sequencing data used for the analysis was comprised of libraries generated from small RNA isolated from naïve, CD4+ stimulated non-infected cells (GHM-17 and GHM-19) and HIV-1 infected cells (GHM-20 and GHM-22) (Table.3.2). Of note, the GMH-17/GMH-20 and GMH-19/GMH-22 libraries were produced from the same sample of cells, and thus may be considered control/experiment pairs.

Table.3.2 - Description of the small- RNA-seq libraries used for analysis.

Library ID	Donor ID	Sample ID	Treatment	Library Size	Average read length
GHM-17	A;D;E	C4 Pool 1	Stimulated	5479509	20
GHM-19	G;H;K	C4 Pool 3	Stimulated	4272410	20
GHM-20	A;D;E	C5 Pool 1	Stimulated-HIV-1	7506592	20
GHM-22	G;H;K	C5 Pool 3	Stimulated-HIV-1	11991015	20

Simple Linux commands (see methods) where used to search the libraries for exact matches to the identified target sRNAs. Since these sequences might also be found within larger sequencing reads, we widened our search to consider this. These cases are of particular interest as they may help to identify the genomic regions that are contributing to generate the target sRNA sequences. Since a large number of such sequences was found, we further determined the number of unique reads present in the data. The results obtained are presented in Table.3.3.

Table.3.3 - Count of target sRNA sequences in the sequencing data. Legend: 'exact match'- number of reads identical to the target sRNA sequence; 'contained' - number of occurrences of the target sRNA sequence within a longer read; 'unique in contained' - number of unique reads containing the target sRNA sequence.

Library ID	sRNA	Exact match	Fold	Contained	Fold	Unique in cont.	CPM
GHM-17	Repeat 1	14	-------	326	-------	170	2.55
	Repeat 2	17	-------	305	-------	143	3.10
	Repeat 3	154	-------	4317	-------	3141	28.10
GHM-19	Repeat 1	13	-------	66	-------	20	3.04
	Repeat 2	46	-------	213	-------	61	10.76
	Repeat 3	174	-------	2741	-------	1933	4.07
GHM-20	Repeat 1	1663	118.7	2116	6.4	259	221.53
	Repeat 2	1487	87.4	1810	5.9	139	198.09
	Repeat 3	947	6.1	15871	3.6	11350	126.15
GHM-22	Repeat 1	677	52.0	1128	17.0	242	56.45
	Repeat 2	535	11.6	916	4.3	159	44.61
	Repeat 3	282	1.6	10257	3.7	6843	23.51

The results (Table.3.3) of the counting analysis confirmed the previous observation that exact matches to 'Repeat 1' and 'Repeat 2' sRNAs have an increase in expression in response to HIV-1 infection, especially in the pair GHM-17 and GHM-20. 'Repeat 3', showed a smaller increase in expression in infected libraries.

Longer reads containing the ’Repeat 1’ and Repeat 2’ sRNAs have also an increase in expression in response to HIV-1 infection, but in this case the pair GHM-19 and GHM-22 has the greater fold increase. ‘Repeat 3’ has also a smaller increase in expression in infected libraries. The amount of these unique longer reads increases in response to HIV-1 infection. The CPM indicates that the increase in expression is not masked by a larger library size in the infected ones, because the CPM is higher in the infected libraries.

The fold increase in the exact matches to the target sRNAs is in average 7 times higher for ‘Repeat 1’, 10 times for ‘Repeat 2’ and 2 times for ‘Repeat 3’ in the pair GHM-17 and GHM-20, than the one observed in the longer reads. This indicates that the target sRNAs are independent from these longer reads. Therefore, the HIV-1 infection triggers a mechanism that would highly increase the expression of these target sRNAs specifically and not a unique set of longer reads that contains them.

3.4.2 Mapping of possible genomic origins for the target sRNAs

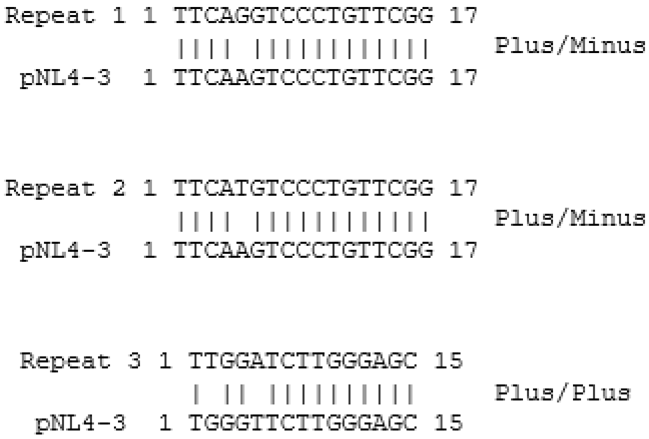
The counting analysis indicated that the infection or replication of HIV-1 could trigger a mechanism that would highly increase the expression or accumulation of these target sRNA sequences. To further investigate the matter, SAM files from the sequencing data aligned with the human genome were analyzed regarding the position of the target sRNAs within the human genome. Several locations for the ‘Exact matches’ were found in the human genome (Table.3.4), although they suggest that the sequences are not repeats, because the number of possible genomic locations were no greater than eight.

The locations where browsed in the UCSC Genome browser with results showing new findings and others in some agreement with earlier assumptions. Locations of the ’Repeat 1’ and ‘Repeat 2’ matched with the tRNA $^{\mathbf{Lys}}\mathbf{{}_3}$ gene, which serves as a *primer* for the reverse transcription of HIV-1. This explains the greater increase of these sequences in comparison with the ‘Repeat 3’. ‘Repeat 3’ matched some retrotransposons, THE1C and MSTA, Long Terminal Repeats (LTR) and the 18S rRNA gene, in agreement with the earlier assumptions about the association of the sequences with retrotransposons, which was the case for this one.

Table.3.4 - Locations of the sRNA target sequences in the Human Genome. Legend: Version Feb. 2009 (GRCh37/hg19) of the Human Genome.

<table>
<tr><th>sRNA</th><th>Genomic location</th><th>Annotation</th><th>Alignment</th><th>Strand</th></tr>
<tr><td rowspan="7">Repeat 1</td><td>chr11:59,323,955-59,323,971</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr1:204,475,708-204,475,724</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr17:74,513,667-74,513,683</td><td>--------------------</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCAGGTCCCTGTTTGG 17</td><td>+/+</td></tr>
<tr><td>chr17:8,022,526-8,022,542</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr19:18,740,364-18,740,380</td><td>L2b</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCAGGTCCCTGTTGGG 17</td><td>+/+</td></tr>
<tr><td>chr5:159,716,110-159,716,126</td><td>L2b</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||||||||||||| |
Gen. 1 TTCAGGTCCCTGTTCAG 17</td><td>+/+</td></tr>
<tr><td>chr6:28,918,859-28,918,875</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCAGGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td rowspan="8">Repeat 2</td><td>chr11:59,323,955-59,323,971</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr1:204,475,708-204,475,724</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr1:230,396,323-230,396,339</td><td>MIR3</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCATGTCCCTGTTGGG 17</td><td>+/+</td></tr>
<tr><td>chr16:35,108,056-35,108,072</td><td>--------------------</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCATGTCCCTGTTGGG 17</td><td>+/+</td></tr>
<tr><td>chr17:8,022,526-8,022,542</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td>chr22:18,777,327-18,777,343</td><td>--------------------</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCATGTCCCTGTTTGG 17</td><td>+/+</td></tr>
<tr><td>chr22:21,578,438-21,578,454</td><td>--------------------</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||||||||||||| ||
Gen. 1 TTCATGTCCCTGTTTGG 17</td><td>+/+</td></tr>
<tr><td>chr6:28,918,859-28,918,875</td><td>tRNA-Lys-AAA</td><td>Rep. 1 TTCATGTCCCTGTTCGG 17
|||| ||||||||||||
Gen. 1 TTCAAGTCCCTGTTCGG 17</td><td>+/+</td></tr>
<tr><td rowspan="6">Repeat 3</td><td>chr11:110,902,196-110,902,210</td><td>THE1C</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
<tr><td>chr16:33,963,179-33,963,193</td><td>SSU-rRNA_Hsa</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
<tr><td>chr2:19,130,736-19,130,750</td><td>MSTA</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
<tr><td>chr2:76,763,319-76,763,333</td><td>--------------------</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
<tr><td>chr8:63,566,215-63,566,229</td><td>--------------------</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
<tr><td>chrX:8,938,899-8,938,913</td><td>MSTA</td><td>Rep.1 TTGGATCTTGGGAGC 15
|||||||||||||||
Gen.1 TTGGATCTTGGGAGC 15</td><td>+/+</td></tr>
</table>

Of note, we confirmed that, as expected from their previous identification, sequences ‘Repeat 1’ and ‘Repeat 2’ also align to the NL4-3 HIV-1 (AF324493) genomic sequence, but in an antisense direction and with one mismatch, thus reducing the likelihood that they are the result of natural viral transcription (Fig.3.5).’Repeat 3’ aligned to the NL4-3 HIV-1 genomic sequence, in a sense direction, but with two mismatches, which also reduces the likelihood of being the result of natural viral transcription (Fig.3.5).

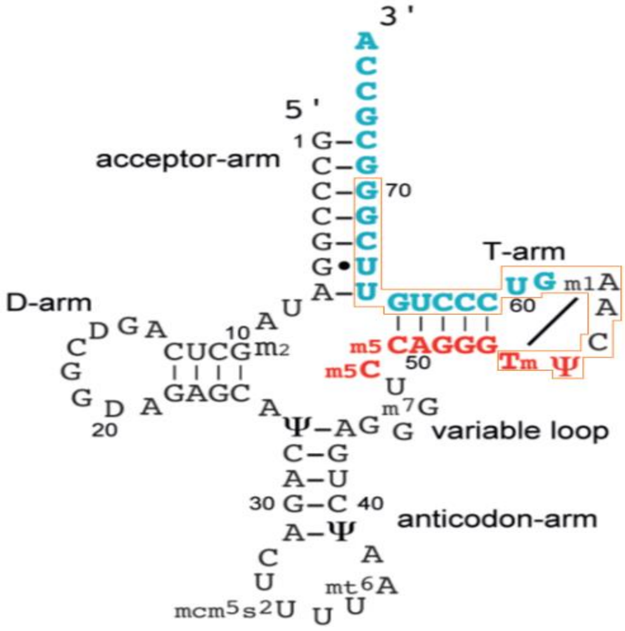


```
Repeat 1 1 TTCAGGTCCCTGTTCGG 17
           |||| ||||||||||||    Plus/Minus
 pNL4-3  1 TTCAAGTCCCTGTTCGG 17


Repeat 2 1 TTCATGTCCCTGTTCGG 17
           |||| ||||||||||||    Plus/Minus
 pNL4-3  1 TTCAAGTCCCTGTTCGG 17


 Repeat 3 1 TTGGATCTTGGGAGC 15
            | || ||||||||||     Plus/Plus
  pNL4-3  1 TGGGTTCTTGGGAGC 15
```

Fig.3.5 - Alignment of “Repeat 1”, “Repeat 2” and “Repeat 3” sequences with HIV-1 vector pNL4-3.

The accumulation of a sRNA derived from the tRNA$^{Lys}{}_3$ Primer Binding Region (PBS) has been previously reported[31] and was also detected during the initial analysis of this dataset in our laboratory. Thus, we were extremely intrigued why the ‘Repeat 1’ and ‘Repeat 2’ sequences were not originally detected as fragments of this tRNA. To understand this, we performed a careful comparison between these sRNA sequences and the structure of the tRNA$^{Lys}{}_3$ and the PBS sRNA (Fig. 3.6)

‘Repeat 1’ and ‘Repeat 2’ align with the sequence of the tRNA$^{Lys}{}_3$ with one mismatch. That aligned part represents twelve nucleotides of the of the PBS sequence that binds to the HIV-1 for reverse transcription (Fig.3.6). The A58 mismatch position where the sequences do not align with the tRNA is the position of the nucleotide that distinguishes the ‘Repeat 1’ from the ‘Repeat 2’. The PBS sncRNA sequence does not contain the A58 mismatch position, being one nucleotide upstream.



Sleiman et. al., 2013, PLOS One, Vol.8, Issue 6

Fig.3.6 - Secondary structure of the human tRNA Lys3, primer of the HIV-1 reverse transcription. Legend: In blue the sequence of the PBS, in red the anti-PAS sequence and the delimited area in orange corresponding to the sequence of 'Repeat 1' and 'Repeat 2'.

'Repeat 1' and 'Repeat 2' origin is the $\text{tRNA}^{\text{Lys}}{}_3$ with the possibility of being a single new sRNA, being that the mismatch position is the same that differs them. There is high similarity of the sRNAs sequences with the PBS sncRNA sequence, thus suggesting an association.

Given our discoveries regarding the origin of the 'Repeat 1' and 'Repeat 2' sequences, we decided to analyze the presence of similar sequences without the A58 mismatch, as well as of the PBS sncRNA sequence in our dataset (Table.3.5).

Table.3.5 - Count of the sRNA genomic sequence from the tRNA Lys 3 and PBS sncRNA sequence in the sequencing data. Legend: 'exact match'- number of reads identical to the target sRNA sequence; 'contained' - number of occurrences of the target sRNA sequence within a longer read; 'unique in contained' - number of unique reads containing the target sRNA sequence.

Library ID	sRNA	Exact match	Fold	Contained	Fold	Unique in cont.	CPM
GHM-17	Genomic sequence	1	-------	64	-------	29	0.18
	PBS	1	-------	21	-------	13	0.18
GHM-19	Genomic sequence	6	-------	73	-------	35	1.40
	PBS	0	-------	11	-------	3	0.0
GHM-20	Genomic sequence	335	335	438	6.8	52	47.29
	PBS	3	3	48	2.2	21	0.39
GHM-22	Genomic sequence	149	24.8	244	3.0	49	12.42
	PBS	12	N/A	51	4.6	21	1.00

The results (Table.3.5) of the counting analysis showed the exact matches to PBS sncRNA sequence and sRNA genomic sequence with an increase in expression in response to HIV-1 infection. The same was observed for longer reads containing either sequence. The amount of these unique longer reads also increases in response to HIV-1 infection in both sequences. The CPM indicates that the increase in expression is not masked by a larger library size in the infected ones, because the CPM is higher in the infected libraries.

Increase in expression in response to HIV-1 infection is confirmed in either sequence, but the exact matches to PBS sncRNA sequence in the dataset are just 1 for the non-infected libraries and 15 for the infected ones. Exact matches for sRNA genomic sequence are 7 times higher than the PBS sncRNA sequence and 32 times higher in infected libraries, even though PBS sncRNA is necessary for HIV-1 reverse transcription and therefore, should have a much higher read count. Fold increase in sRNA genomic sequence is 37 times higher in exact matches than the one observed in the longer reads, thus confirming that the sRNA genomic sequence is independent from the longer reads like the target sRNAs.

With the findings regarding the amount of PBS sncRNA in our dataset, we decided to compare the amount of PBS sncRNA sequence, sRNA genomic sequence, 'Repeat 1' and 'Repeat 2' sequences in our dataset (Fig.3.7).

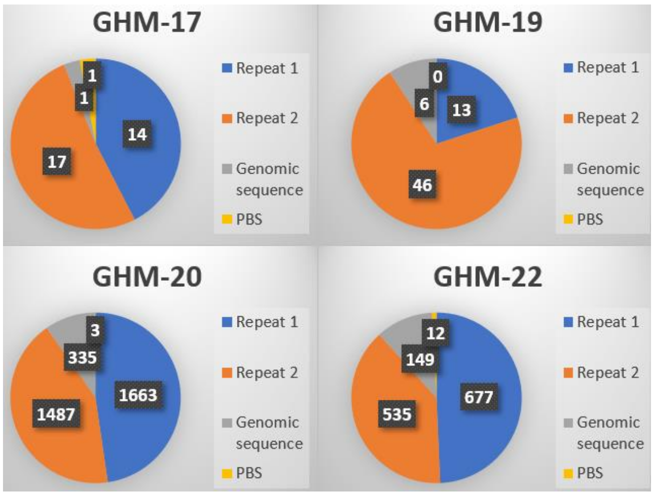
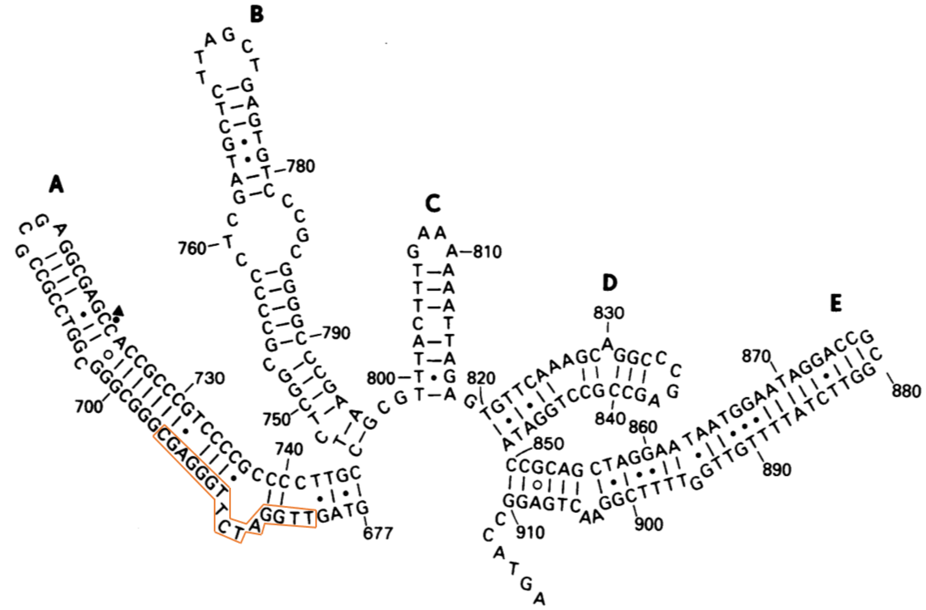


Fig.3.7 - Exact matches of "Repeat 1", "Repeat 2", Genomic sequence and PBS in the sequencing data.

The results (Fig.3.7) showed that the amount of the PBS sncRNA sequence is much lower than 'Repeat 1' sequence, 'Repeat 2' sequence and sRNA genomic sequence in non-infected and infected libraries. In non-infected libraries the PBS sncRNA sequence is 7 to 32 times lower than the other three sequences and 32 to 156 times lower in the infected libraries.

Given the difference in our dataset between the amount of PBS sncRNA sequence and the sRNA 'Repeat 1', or, sRNA 'Repeat 2', or sRNA genomic sequence, there is the possibility that the known PBS sncRNA sequence is wrong and the sRNA genomic sequence is the right PBS sncRNA sequence.

Finally, we decided to investigate in greater detail the possible origin of the 'Repeat 3' sequence from the 18s rRNA. For this purpose, we determined the relative position of the sequence in the context of the reported secondary structure of this molecule[32]. The results, presented in Fig.3.8, reveal that the 'Repeat 3' is located at the base of a stem-loop that is compatible with the canonical miRNA processing pathway[33–35].



Gonzalez and Schmickel, 1986, American Society of Human Genetics, Vol.38

Fig.3.8 - Secondary structure model for region V5 between human gene bp 677-910. Legend: Delimited area from bp 681-695 corresponding to the 'Repeat 3' sequence.

Therefore, sRNA 'Repeat 3' has the potential to be a miRNA-like fragment associated with the RISC silencing complex with similar biogenesis and function to miRNA. In terms of function, since there is an increase in expression in response to an HIV-1 infection, its silencing target must be associated with it.

4. Discussion

Processing and analysis of sequencing data from small libraries derived from RNA pools of uninfected stimulated CD4+ T cells (GHM-17 and GHM-19) and from HIV-1 infected stimulated CD4+ T cells (GHM-20 and GHM-22) gave six potential sRNAs[28]. From those six, three appeared to have a large increase in raw read counts between RNA from stimulated non-infected cells and RNA from stimulated infected cells that suggested the accumulation of these three sRNAs in response to HIV-1 infection. Although the function of the sequences was not clear, there was the possibility of being a sort of defense mechanism against the HIV- infection.

A biological model was required to find out, if the accumulation of target sRNAs is induced in response to the expression of HIV-1 transcripts. The J-Lat cell line with an integrated, but transcriptionally latent HIV provirus, allows for a safe expression of viral transcripts, and was therefore chosen as the biological model.

Appropriate conditions for the reactivation of latent HIV-1 were established. The reactivation with TNF-α in J-Lat cells was performed and the efficient induction of GFP expression by fluorescence microscopy was confirmed (Fig.3.1). The following step was confirmation of the presence of HIV-1 transcripts, which was done by RT-PCR from RNA extracted from TNF-α treated and control cells. There was detection of mRNA of eGFP and an HIV-1 viral transcript, mRNA of the Tat protein[36] in TNF-α treated J-Lat cells and in control cells, confirming the existence of a basal expression of the HIV provirus, but the intensity of the band was higher in the former (Fig.3.2).

The extracted RNA was then used to assess the accumulation levels of target sRNAs on the presence of HIV-1 transcripts by a specific RT-PCR method. The cDNA synthesis was performed using the Poly A tailing, which allowed the ability to amplify the target sRNAs sequences without producing unbiased results that others method could produce. Alas, no significant difference in the target sRNAs sequences ('Repeat 1', 'Repeat 2' and 'Repeat 3') was observed, between control samples and TNF-α J-Lat treated cell samples (Fig.3.3). Therefore, we sought to detect them in the primary cells samples that were used to generate the sequencing libraries, on the basis that the phenomena may only be observed in primary T cells.

Once again, the results were inconsistent throughout the assays, on the target sRNAs and controls (Fig.3.4). Given that the assays did not produced any clear evidence supporting an increased expression of our target sRNAs in response to HIV-1 transcripts, we decided to perform a more detailed bioinformatic analysis to confirm their expression pattern and to learn more about their potential origin.

The bioinformatic analysis focused in the variability of results between different samples, the potential genomic point of origin of the target sRNAs and their presence in the context of longer sRNA reads. For this purpose, it was used sequencing data from libraries generated from small RNA isolated from naïve, CD4+ stimulated non-infected cells (GHM-17 and GHM-19) and HIV-1 infected cells (GHM-20 and GHM-22) (Table.3.2)[28]. The fold increase from non-infected to infected libraries regarding the exact matches to the target sRNAs is in average 7 times higher for 'Repeat 1', 10 times for 'Repeat 2' and 2 times for 'Repeat 3'in the pair GHM-17 and GHM-20, than the one observed in longer reads (Table.3.3). This, alongside with the CPM values being higher in the infected libraries, indicates that the target sRNAs are independent from the longer reads , but also that there is an increase in expression of the target sRNAs in response to HIV-1 infection. We can therefore say that the HIV-1 infection triggers a mechanism that would highly increase the expression of these target sRNAs specifically and not a unique set of longer reads that contains them.

Of note, the sequences ‘Repeat 1’, ‘Repeat 2’ and ‘Repeat 3’ align to the NL4-3 HIV-1 genomic sequence (Fig.3.5), as expected from previous identification. However, the first two align in an antisense direction with one mismatch and the last in a sense direction, but with two mismatches, which reduces the likelihood of the three being the result of natural viral transcription.

For further investigation, SAM files from the sequencing data aligned with the human genome were analyzed regarding the position of the target sRNAs within the human genome. An immediate aspect that popped up was the fact that the sequences were not repeats, since the number of possible genomic locations were no greater than eight (Table.3.4).

Through our analysis, we confirmed that ‘Repeat 1’ and ‘Repeat 2’ are actually fragments derived from the $tRNA^{Lys}{}_3$ that represents the major *primer* used by HIV-1 for reverse transcription, and which gives rise to the accumulation of a well know cleavage product previously termed PBS sncRNA[31]. However, they differ from the PBS sequence in two main aspects. The target sRNAs sequences have one nucleotide less than the PBS sncRNA in terms of length, aligning with twelve nucleotides of the PBS sncRNA (Fig.3.6). The position 58 in the tRNA $^{\mathbf{Lys}}\mathbf{{}_3}$ is an A, but in our sRNA sequences we find a G and T for ‘Repeat 1’ and ‘Repeat 2’, respectively[31,37]. The sequence with the adenine as also been found in the sequencing data and the same increase in expression in response to an HIV-1 infection was observed (Table.3.5).

An interesting aspect of this position is that the adenine is a 1-methyladenosine (m_1 A). This is a frequent modification that is found on position 58 of eukaryotic tRNAs. Such modification can happen through catalyzation by enzymes or a reaction between RNA and certain alkylating agents. The sites where m_1 A is known to be or postulated to be, are sites were mismatch signals occurred in different RNA-Seq protocols and our case could well be another one. RNA-Seq protocols are based on reverse transcription followed by the sequencing of the cDNA and the numerous parameters of the reverse transcription with an RNA modification can easily lead to a misincorporation by the RT enzyme, especially when encountering an RNA modification[38].

The PBS sncRNA in our sequencing data was detected with 7 to 32 times lower values in non-infected libraries and 32 to 156 times lower values in infected libraries than ‘Repeat 1’, ‘Repeat 2’ and the sRNA genomic sequence (Fig.3.7).If the PBS sncRNA is the major *primer* used by HIV-1 for reverse transcription, than it would lead to a major increase in expression in response to HIV-1 infection, but the values that match that are the ones of the ‘Repeat 1’, ‘Repeat 2’ and sRNA genomic sequence. This indicates that the sRNA genomic sequence is the correct PBS sncRNA sequence.

A small sequence of 17 bp that comes from a tRNA, matches a type of sRNAs that are classified as tRNA-derived fragments (tRFs). tRFs are abundant non-coding RNAs that are widespread in most organisms. They have been linked to stress responses, cancer, cell-cell signaling via exosomes, response to viral infection and neurological disorders. Already, there are sub-classes of the tRFs, 5’ and 3’ halves of 30-33 nucleotides and short 18 to 22 nucleotides 3’ fragments[39]. The new thing about our sequence is that it is not from the beginning of a 5’ end, or, 3’ end. It starts from the 3’ end side, but upstream of the sequence without containing the CCA trinucleotide necessary for maturation being the site where there is the attachment for the ester-linked amino acid[40] and for the interaction with the ribosome during protein synthesis[41]. In any case, although it does not contain the CCA trinucleotide, the fact that the

RNA modification in position 58 is the position where the mismatched occurred, suggests that the fragment still came from matured tRNA, since these modifications are post-transcriptional. Apart from the first functions mentioned earlier, there are other functions such as translation inhibition, inhibition of apoptosis, suppression of breast cancer progression. The latest function linked to tRFs is gene silencing, since the RNAi silencing machinery has been implicated in the biogenesis of tRFs species, but regarding the 5' tRFs of 30-33 nucleotides. In this case, Dicer would be responsible for the cleavage of the fragments from some human tRNA, which is not our case, since they are more related to the 3' tRFs. Regarding the 3' tRFs ones, they are associated with endogenous retroviruses (ERV) because ERVs are used as *primers* for reverse transcription. Recently, tRFs of sizes ranging from 17-19 nucleotides have been studied and implicated in interfering with reverse transcription, while larger ones from 22 nucleotides interfere with transposons expression[39].

Regarding the other sRNA sequence, it has been related either to a ribosomal RNA, 18S rRNA and/or transposons related to a specific HERV. It aligns to a segment of the transcripts of THE1C and MSTA, who are both Long Terminal Repeats (LTRs) from a Mammalian apparent LTR retrotransposon derived from an endogenous retrovirus. It is not clear what is the significance or if there is any, regarding the location inside the transcripts. In the case of the 18S rRNA it is found from position bp 681 to 695 (Fig.3.8) and is therefore part of the Variable region 5 (V5) of the human 18S rRNA in a region that only exists in eukaryotes[32] with 9 variable regions V1- to V9[42]. The V5 does not have long highly variable regions and is considered a short region, probably because a great part is only present in eukaryotes as stated earlier[42]. The 18S rRNA is frequently used for phylogenetic studies[32] and is present in all eukaryotic cells being the structural RNA of the small component of the cytoplasmic ribosomes of eukaryotes[42].

'Repeat 3' in 18S rRNA is located at the base of a stem-loop that is compatible with the canonical miRNA processing pathway (Fig.3.8)[33–35]. There have been a few studies that report the existence, biogenesis and functions of rRNA-derived miRNAs or miRNA-like fragments, in mice and human [33,43]. Some of these rRNA-derived miRNAs have been related to piRNAs, which form a hook structure and potentially work as a small guide RNA[43]. Therefore, sRNA 'Repeat 3' has the potential to be a miRNA-like fragment and associated with the RISC silencing complex with similar biogenesis and function to miRNAs. In terms of function, since there is an increase in expression in response to an HIV-1 infection, its silencing target must be associated with it.

Cellular transcriptome is composed of the following RNA species of tRNAs, three tRNA gene-derived RNA species, precursor tRNAs (pre-tRNAs), mature tRNAs and tRNA-derived small RNA fragments[44]. Having identical sequences, these RNA species cannot be distinguished by normal PCR methods. tRNA modifications are also one of the issues, and tRNAs harbor the highest density of nucleoside with over 100 post-transcriptional modifications associated with tRNA folding and function, like codon recognition. With many of these modifications inhibiting Watson-Crick base pairing and thus arresting reverse transcription, making several biased results and that could lead to mismatches[44]. The several levels of structure that both rRNA and tRNA can have, is also one aspect that could be a problem for the detection of the sequences by PCR methods, because the cDNA synthesis for posterior PCR is more difficult in more complex structures with loops and hairpins[32,42,44].

The 'Repeat 1' and 'Repeat 2' are in fact just one sequence, a sRNA genomic sequence from the $tRNA^{Lys}_3$. 'Repeat 1' and 'Repeat 2' correspond to the sRNA, but with a common mismatch in position A58, that occurs during reverse transcription. Increase in expression of the sRNA genomic sequence in response to HIV-1 infection is explained by being the PBS sncRNA responsible for HIV-1 reverse transcription. ‘Repeat 3’ can be a rRNA-derived miRNA with a function in the RISC silencing complex associated with HIV-1 infection. An association with the LTRs is also possible, since there is a correlation of increase in expression of ERVs with HIV-1 infection during some stages of infection[45]. For the detection of the sequences, use of approaches like Northern blot to detect RNA, or, kits for cDNA synthesis might work better than the ones used[46], since detection by PCR methods is difficult due to specificity and the existence of tRNA modifications, as we have observed in our case regarding the 1-methyladenosine (m_1 A).

Our work as brought to light a new PBS sncRNA sequence and a rRNA-derived miRNA, which can have an impact in a different understanding of the HIV-1 infection in terms of the reverse transcription of the HIV-1 and the host’s defense against the virus. In the future, the detection of these two sequences will prove essential to then perform additional assays. Assays to evaluate how this affects our knowledge and study of HIV-1, since we have been assuming the PBS sncRNA with a different sequence. And finally, assays to permit the understanding of the link between the rRNA-derived miRNA and the HIV-1, to determine what is the exact function and how it affects the host and the HIV-1.

5.1. References

1. Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* **112**, 1557–1569 (2008).
2. Berard, M., Tough, D. F. & Rg, B. Qualitative differences between na¨ıve and memory T cells. *Immunology* **106**, 127–138 (2002).
3. Leung, S. *et al.* The cytokine milieu in the interplay of pathogenic Th1/Th17 cells and regulatory T cells in autoimmune disease. *Cell. Mol. Immunol.* **7**, 182–189 (2010).
4. Birnbaum, M. E. *et al.* Molecular architecture of the αβ T cell receptor–CD3 complex. *Proc. Natl. Acad. Sci.* **111**, 17576–17581 (2014).
5. Kohlmeier, J. E., Chan, M. A. & Benedict, S. H. Costimulation of naive human CD4+ T cells through intercellular adhesion molecule-1 promotes differentiation to a memory phenotype that is not strictly the result of multiple rounds of cell division. *Immunology* **118**, 549–588 (2006).
6. Mosmann, T. R. & Coffman, R. L. TH1 and TH2 Cells: Different Patterns of Lymphokine Secretion Lead to Different Functional Properties. *Ann Rev Immunol* **7**, 145–173 (1989).
7. Tesmer, L. A., Lundy, S. K., Sarkar, S. & Fox, D. A. Th17 cells in human disease. *Immunol. Rev.* **223**, 87–113 (2008).
8. Corthay, A. How do Regulatory T Cells Work? *Scand. J. Immunol.* **70**, 326–336 (2009).
9. Maartens, G., Celum, C. & Lewin, S. R. HIV infection: epidemiology, pathogenesis, treatment, and prevention. *The Lancet* **384**, 258–271 (2014).
10. Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: Cell Binding and Entry. *Cold Spring Harb. Perspect. Med.* **2**, a006866–a006866 (2012).
11. Hu, W.-S. & Hughes, S. H. HIV-1 Reverse Transcription. *Cold Spring Harb. Perspect. Med.* **2**, a006882–a006882 (2012).
12. Persaud, D., Zhou, Y., Siliciano, J. M. & Siliciano, R. F. Latency in Human Immunodeficiency Virus Type 1 Infection: No Easy Answers. *J. Virol.* **77**, 1659–1665 (2003).
13. Siliciano, R. F. & Greene, W. C. HIV Latency. *Cold Spring Harb. Perspect. Med.* **1**, a007096–a007096 (2011).

14. Rezaei, S. D. *et al.* The Pathway To Establishing HIV Latency Is Critical to How Latency Is Maintained and Reversed. *J. Virol.* **92**, e02225-17 (2018).

15. Painter, M. M., Zaikos, T. D. & Collins, K. L. Quiescence Promotes Latent HIV Infection and Resistance to Reactivation from Latency with Histone Deacetylase Inhibitors. *J. Virol.* **91**, e01080-17 (2017).

16. Jordan, A. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J.* **22**, 1868–1877 (2003).

17. Ott, M., Geyer, M. & Zhou, Q. The Control of HIV Transcription: Keeping RNA Polymerase II on Track. *Cell Host Microbe* **10**, 426–435 (2011).

18. Modai, S. *et al.* HIV-1 infection increases microRNAs that inhibit Dicer1, HRB and HIV-EP2, thereby reducing viral replication. *PLOS ONE* **14**, e0211111 (2019).

19. Ohlmann, T., Mengardi, C. & López-Lastra, M. Translation initiation of the HIV-1 mRNA. *Translation* **2**, e960242 (2014).

20. Bieniasz, P. D. The Cell Biology of HIV-1 Virion Genesis. *Cell Host Microbe* **5**, 550–558 (2009).

21. Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.* **10**, 126–139 (2009).

22. Choudhuri, S. Small noncoding RNAs: Biogenesis, function, and emerging significance in toxicology. *J. Biochem. Mol. Toxicol.* **24**, 195–216 (2010).

23. Klimenko, O. V. Small non-coding RNAs as regulators of structural evolution and carcinogenesis. *Non-Coding RNA Res.* **2**, 88–92 (2017).

24. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.* **20**, 89–108 (2019).

25. Dana, H. *et al.* Molecular Mechanisms and Biological Functions of siRNA. *Int. J. Biomed. Sci.* **13**, 48–57 (2017).

26. MacFarlane, L.-A. & R. Murphy, P. MicroRNA: Biogenesis, Function and Role in Cancer. *Curr. Genomics* **11**, 537–561 (2010).

27. Chen, C. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.* **33**, e179–e179 (2005).

28. Amaral, A. J. *et al.* miRNA profiling of human naive CD4 T cells links miR-34c-5p to cell activation and HIV replication. *EMBO J.* **36**, 346–360 (2017).

29. Finnegan, D. J. Retrotransposons. *Curr. Biol.* **22**, R432–R437 (2012).

30. Schneider, U., Schwenk, H.-U. & Bornkamm, G. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int. J. Cancer* **19**, 621–626 (1977).

31. Yeung, M. L. *et al.* Pyrosequencing of small non-coding RNAs in HIV-1 infected cells: evidence for the processing of a viral-cellular double-stranded RNA hybrid. *Nucleic Acids Res.* **37**, 6575–6586 (2009).

32. Gonzalez, I. L. & Schmickel, R. D. The Human 18S Ribosomal RNA Gene: Evolution and Stability. *Alm J Hum Genet* **38**, 419–427 (1986).

33. Yoshikawa, M. & Fujii, Y. R. Human Ribosomal RNA-Derived Resident MicroRNAs as the Transmitter of Information upon the Cytoplasmic Cancer Stress. *BioMed Res. Int.* **2016**, 1–14 (2016).

34. Shukla, G. C., Singh, J. & Barik, S. MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Mol Cell Pharmacol* **3**, 83–92 (2012).

35. Bikard, D., Loot, C., Baharoglu, Z. & Mazel, D. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. *Microbiol. Mol. Biol. Rev.* **74**, 570–588 (2010).

36. Bagashev, A. & Sawaya, B. E. Roles and functions of HIV-1 Tat protein in the CNS: an overview. *Virol. J.* **10**, 1–20 (2013).

37. Sleiman, D., Barraud, P., Brachet, F. & Tisne, C. The Interaction between tRNALys3 and the Primer Activation Signal Deciphered by NMR Spectroscopy. *PLoS ONE* **8**, e64700 (2013).

38. Hauenschild, R. *et al.* The reverse transcription signature of *N* -1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* **43**, 9950–9964 (2015).

39. Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* **170**, 61-71.e11 (2017).

40. Jackman, J. E. tRNA Biogenesis. in *Encyclopedia of Life Sciences* (ed. John Wiley & Sons, Ltd) a0020894 (John Wiley & Sons, Ltd, 2010). doi:10.1002/9780470015902.a0020894

41. Hori, H. *et al.* Transfer RNA Synthesis and Regulation. in *eLS* (ed. John Wiley & Sons Ltd) a0000529.pub3 (John Wiley & Sons, Ltd, 2014). doi:10.1002/9780470015902.a0000529.pub3

42. Hadziavdic, K. *et al.* Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. *PLoS ONE* **9**, e87624 (2014).

43. Wei, H. *et al.* Profiling and Identification of Small rDNA-Derived RNAs and Their Potential Biological Functions. *PLoS ONE* **8**, e56842 (2013).

44. Honda, S., Shigematsu, M., Morichika, K., Telonis, A. G. & Kirino, Y. Four-leaf clover qRT-PCR: A convenient method for selective quantification of mature tRNA. *RNA Biol.* **12**, 501–508 (2015).

45. Nazar, R. Ribosomal RNA Processing and Ribosome Biogenesis in Eukaryotes. *IUBMB Life Int. Union Biochem. Mol. Biol. Life* **56**, 457–465 (2004).

46. Ormsby, C. E. *et al.* Human Endogenous Retrovirus Expression Is Inversely Associated with Chronic Immune Activation in HIV-1 Infection. *PLoS ONE* **7**, e41021 (2012).

6. Supplementary Information

6.1. Oligonucleotide sequences

Supplemental Table.1 - Oligonucleotide sequences used in Reverse transcription (RT), Poly A tailing or PCR.

Designation	Method	Sequence (5'-3')
Oligo-dT	Reverse transcription (RT)	TTTTTTTTTTTTTTTTTT
Oligo-dT adaptor (CDS)	Poly A tailing	AAGCAGTGGTAACAACGCAGAGTACCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Universal Reverse (ARH12)	PCR	AAGCAGTGGTAACAACGCAGAGT

Supplemental Table.2 - Oligonucleotide sequences use in PCR.

Gene	Forward Primer (5'-3')	Reverse Primer (5'-3')	Size (bp)
Tat	GCATCTCCTATGGCAGGAAG	CCGTTCACTAATCGAATGGA	183
eGFP	AGTCCGCCCTGAGCAAAGA	TCCAGCAGGACCATGTGATC	55
snRNA	**Forward Primer (5'-3')**	**Reverse Primer (5'-3')**	**Size (bp)**
U6	GCTTCGGCAGCACATATACTA	AAATATGGAACGCTTCACGA	99
sRNA	**sRNA sequence**	**Forward Primer (5'-3')**	**Reverse Primer (5'-3')**
'Repeat 1'	TTCAGGTCCCTGTTCGG	GACGGTTCAGGTCCCTGTTC	AAGCAGTGGTAACAACGCAGAGT
'Repeat 2'	TTCATGTCCCTGTTCGG	GACGGGTTCATGTCCCTGTTC	AAGCAGTGGTAACAACGCAGAGT
'Repeat 3'	TTGGATCTTGGGAGC	GCAGGGTTGGATCTTGGGAG	AAGCAGTGGTAACAACGCAGAGT