UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

UNDERSTANDING THE INTERACTIONS BETWEEN RETROVIRAL VECTORS AND THE GENOME OF HUMAN KERATINOCYTES

ANA RITA MICAELO MENDES

MASTER IN APPLIED MICROBIOLOGY

2009/2010

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

UNDERSTANDING THE INTERACTIONS BETWEEN RETROVIRAL VECTORS AND THE GENOME OF HUMAN KERATINOCYTES

ANA RITA MICAELO MENDES

THESIS COORDINATORS:

FULVIO MAVILIO

CENTRO DI MEDICINA RIGENERATIVA STEFANO FERRARI, UNIVERSITY OF MODENA AND REGGIO EMILIA

MARGARIDA GAMA-CARVALHO

CENTER FOR BIODIVERSITY, FUNCTIONAL AND INTEGRATIVE GENOMICS, FACULTY OF SCIENCES, UNIVERSITY OF LISBON AND FACULTY OF MEDICINE, UNIVERSITY OF LISBON

MASTER IN APPLIED MICROBIOLOGY

2009/2010

INDEX

“Retroviruses are unique among infectious agents, both in the way they interact with the host cell and organism and in the consequences of this interaction—not only to the life of the infected host, but also in some cases to the host's descendants. No other infectious agent of higher eukaryotes regularly integrates its genetic information into the host genome; no other regularly acquires host genes into its genome, no other can infect the germ line of its host; and no other has played such an important part in so many aspects of modern biology.” (in Retroviruses)

ABSTRACT

Retroviruses are a unique family of infectious agents, its diverse life cycle and survival strategies have been an important object of study to modern biology. The integration pattern of gamma-retroviruses (RV) and lentiviruses (LV) in the human genome follows a non-random very distinct distribution, but the precise factors involved in choice of the integration sites are still unknown.

To further study this question, RV (MLV) and LV (HIV) integration sites were mapped after infection of primary human keratinocytes with vectors derived from these viruses. 3020 and 1601 unique sites for MLV and HIV, respectively, were retrieved and analysed for correlation with several genomic sites. Moreover, a comparative analysis between the new dataset and previously obtained datasets on lymphocytes T and haematopoietic stem cells Cd34$^+$ was performed to understand the involvement of the cellular program in viral integration choices.

The outcome of this project suggests substantial differences in the molecular mechanisms tethering RV and LV PICs to human chromatin. Both vectors' integration locates preferably inside genes and target mostly active genes with a higher frequency in lymphocytes, these characteristics showed a higher frequency in HIV's integrations. MLV is attracted to TSS proximal regions with a higher frequency in lymphocytes, while HIV is repelled by these regions independently of the cell-type. MLV positively associates with epigenetic markers of promoters and enhancers and has a higher integration in evolutionary conserved non-coding sequences in mammals (CNCs). Additionally it targets regions rich in transcription factor binding sites (TFBS) with a cell-specific pattern, revealing a possible preference for functional TFBS rather then a general feature preference. HIV associates with epigenetic markers for the body of transcribed genes and has no preference for CNCs. HIV showed a more widely distributed integration, with clusters that spread through much larger regions of the genome and, therefore, target more genes then MLV's clusters which have a more packed distribution. Analysis of gene function revealed some cell-specificity in MLV's targets although common genes probably related to the transcription machinery are also highly targeted. HIV targets genes involved essentially in general cell-functions. Furthermore we predict that retroviruses choose early replicating regions of the genome to integrate.

Since primary human keratinocytes are clinically relevant in gene therapy for correction of inherited skin diseases, safety parameters related to both vectors were also analysed. The results of this study predict that HIV has a lower probability of generating insertional gene activation events that could lead to oncogene activation. Furthermore, since MLV has a more cell-specific integration pattern, there's a different probability of integrating near specific oncogenes according to the cell-type.

Findings of this work reveal cellular specific patterns of integration as well as common integration preferences, clarifying the cellular machinery's role in integration site-selection mechanism. Overall these results strengthen models for integration targeting based on tethering of viral components to host proteins and suggest a higher involvement of the cellular transcriptional machinery in MLV’s integration in comparison to HIV’s.

KEY WORDS: retrovirus, lentivirus, integration, viral vectors.

RESUMO

Os retrovírus fazem parte de uma família única de agentes infecciosos; as particularidades do seu ciclo de vida, a forma como interagem com o hospedeiro e as consequências desta interacção têm sido objecto de estudo fulcral da biologia moderna. O impacto dos retrovírus na sociedade actual é inegável e a busca pela sua compreensão tem direccionado esforço e recursos da comunidade científica.

O processo de integração retroviral no genoma humano é de primordial importância uma vez que, dele depende a sobrevivência do vírus no hospedeiro. A integração dos retrovírus segue um padrão não-aleatório que varia entre as diversas famílias virais, podendo-se facilmente identificar uma família retroviral pelas suas preferências na escolha do local de integração (Bushman, 2005). Apesar dos mecanismos envolvidos nas reacções de corte e ligação do genoma viral ao do hospedeiro serem já bastante conhecidos, os factores envolvidos na escolha do local de integração são ainda uma questão em aberto.

Para tentar compreender melhor esta questão, mapeei 3020 locais únicos de integração de *moloney leukaemia virus* (MLV) e 1601 locais de integração de *human immunodeficiency virus* (HIV) no genoma de queratinócitos humanos primários. Sendo MLV da família *oncoretroviridae* e HIV da família *lentiviridae*, procurei estabelecer as semelhanças e diferenças no processo de escolha do local de integração de cada um. A par de analisar as preferências integrativas neste tipo celular, realizei uma análise comparativa com dados anteriores obtidos em células estaminais hematopoiéticas (HSC) Cd34^{+} e linfócitos T de forma a tentar compreender o envolvimento do programa celular nas escolhas integrativas.

Os resultados deste trabalho demonstram que MLV, membro da família *gamma-retroviridae*, e HIV, da família *lentiviridae*, apresentam padrões de integração não aleatórios muito distintos. Os vírus em estudo apresentam diferenças substanciais nos mecanismos moleculares que atraem os complexos de pré-integração viral (PICs) para a cromatina humana revelando estratégias evolutivas muito diversas. Ambos os vectores se integram preferencialmente dentro de genes e maioritariamente em genes activos, com uma frequência mais elevada em linfócitos. Estas características demonstraram-se mais exacerbadas nas integrações de HIV.

MLV é atraído por regiões próximas de *transcription start sites* (TSS), com uma frequência mais elevada em linfócitos, enquanto HIV é repelido por estas mesmas zonas do genoma com igual frequência em todos os tipos celulares. Ambos os vírus se integram preferencialmente dentro de genes em comparação com o controlo, no entanto, esta preferência é bastante mais acentuada em HIV, em particular em linfócitos, apresentando mais de 80% das integrações em genes em comparação com 50% das integrações de MLV. Genes com expressão activa são candidatos preferenciais para a integração viral, incluindo ≈70% dos *hits* únicos. Esta preferência é privilegiada nas integrações dentro de genes de HIV, sendo mais elevada em Cd34+HSC

quando comparada com queratinócitos. MLV, mas não HIV, revela alguma preferência por zonas do genoma conservadas não-codificantes (CNCs) entre mamíferos e relaciona-se positivamente com regiões ricas em *transcription factor binding sites,* com um padrão particular para cada tipo celular. Esta particularidade sugere uma preferência por TFBS activos mais do que uma preferência geral pela vizinhança destas regiões genómicas.

Foram também analisados diversos marcadores epigenéticos para os diversos tipos celulares. Esta análise revelou uma forte associação de MLV a regiões com promotores e *enhancers* e uma associação negativa a zonas com marcadores geralmente associados a cromatina inactiva ou repressão da expressão génica. Por outro lado, HIV associa-se positivamente a marcadores epigenéticos para o corpo de genes transcritos e negativamente para marcadores de regiões com promotores e *enhancers,* assim como zonas de cromatina inactiva.

Clusters de integração viral foram definidos estatisticamente, tendo em conta o tamanho da amostra e a distância entre uma integração e a integração consecutiva. Verificou-se que MLV apresenta clusters com maior densidade de integração, enquanto os clusters de HIV se expandem através de uma área mais vasta da cromatina. Consequentemente, os clusters de HIV incluem um maior número de genes relativamente a MLV.

Através da utilização do programa Ingenuity analisei os genes alvo de ambos os vírus relativamente à sua função em todos os tipos celulares. Verifica-se que MLV tem como alvo genes específicos do tipo celular assim como genes relacionados com funções mais gerais. HIV integra-se essencialmente em genes com funções de regulação da expressão génica.

Seguidamente, coloquei a hipótese de que as integrações retrovirais se localizariam em zonas dos cromossomas que se replicam cedo durante o ciclo replicativo, uma vez que, estas zonas se relacionam com zonas de cromatina activa assim como com genes de elevada expressão em determinado tipo celular. Apesar de considerar os resultados com alguma cautela, tendo em conta a reduzida informação disponível em relação ao tempo de replicação, é previsível uma forte correlação com zonas dos cromossomas que replicam cedo, por parte de ambos os vírus.

Uma vez que os vírus estudados são aplicáveis ao nível da terapia génica de doenças epiteliais hereditárias, verifiquei alguns parâmetros de segurança relativamente a ambos os vectores virais. Após confronto de uma lista de *common insertion sites* e oncogenes com os genes alvo de MLV e HIV, prevê-se que os lentivírus sejam vectores mais seguros no entanto, dadas as limitações das análises, não foi possível concluir com bastante fiabilidade sobre esta questão.

A partir do seu padrão de integração, é possível que os gamma-retrovirus tenham desenvolvido um mecanismo que tira proveito da maquinaria celular transcripcional para promover a sua própria expressão, ligando integração do provírus com regiões de elevada actividade no genoma, ricas em TFBS, TSS, promotores, *enhancers* e com uma conformação epigenética característica de uma cromatina activa. Isto sugere um modelo no qual factores de transcrição

(TFs) ubíquos ligados aos complexos de pré-integração (PICs) de retrovírus, interagem com componentes gerais dos complexos de ligação a *enhancers*, por exemplo co-reguladores, complexos de remodelação da cromatina ou complexos mediadores, em vez de TFs ou famílias de TFs específicas. A ligação dos PICs a fábricas transcripcionais, onde promotores e regiões regulatórias são relocalizadas através de mecanismos específicos do tipo celular, pode ser a causa dos clusters específicos com alta frequência de integração de MLV assim como do targeting preferencial de genes associados a redes regulatórias específicas do tipo celular. Trabalhos recentes dão suporte a esta hipótese, propondo uma interacção directa entre integrase e remodelação da cromatina, reparação de DNA e factores transcripcionais. De um ponto de vista evolutivo, esta cooperação pode ser interpretada como o desenvolvimento dos mecanismos através dos quais os retrotransposões escolhem regiões alvo específicas através da ligação a proteínas da célula hospedeira. Um mecanismo ligando selecção de locais alvo a regulação génica pode ter evoluído para maximizar a probabilidade de que os gammaretrovírus sejam transcritos no genoma da célula-alvo e, possivelmente para induzir a expansão das células infectadas através de desregulação insercional de reguladores de crescimento específicos do tipo celular.

Por outro lado, os lentivírus desenvolveram uma estratégia totalmente diferente, interagindo com muito menos interferência com a cromatina e maquinaria celular do hospedeiro, isto é, não provocando situações de desregulação insercional uma vez que não escolhem a proximidade de regiões regulatórias para se integrar. Esta estratégia prolonga o tempo de vida das células hospedeiras e consequentemente do hospedeiro, aumentando a probabilidade de disseminação viral para novos hospedeiros.

Este trabalho contribui assim significativamente para aprofundar as questões da integração retroviral no genoma humano, assim como para inferir sobre a segurança da utilização de vectores virais em doenças hereditárias da pele.

PALAVRAS-CHAVE: retrovírus, lentivírus, integração, vectores

INTRODUCTION

Retroviruses are obligate intracellular parasites whose replication depends absolutely on their hosts. In its integrated form, as provirus, the retroviral genome mimics a cellular gene, the interaction of the viral genome with the host machinery involved in gene regulation and expression is, therefore, extensive. This interplay has important consequences, resulting in an adaptation of different retroviruses to exploit the complex cellular host machinery in different ways and in the co-evolution of hosts to acquire new resistance mechanisms. From the virus point of view, it must choose strategies that allow the host to optimize viral production and its transmission to fresh hosts, avoiding undesirable side effects like blocking the normal function of a receptor but, not necessarily having a benign effect from the host's perspective.

Choice of the integration site is an essential step in retroviral life cycle once it influences the following replication process and consequently the virus' fate. In fact, clustering of different viruses on the basis of their integration preferences shows a high degree of overlaping with phylogenetic trees based on the sequence similarity of their integrases, which, in turn, are in good agreement with traditional trees based on genomic sequences (Derse, Crise et al. 2007). This emphasises that the survival of a retrovirus cannot but depend on the site in the host cell genome it integrates in.

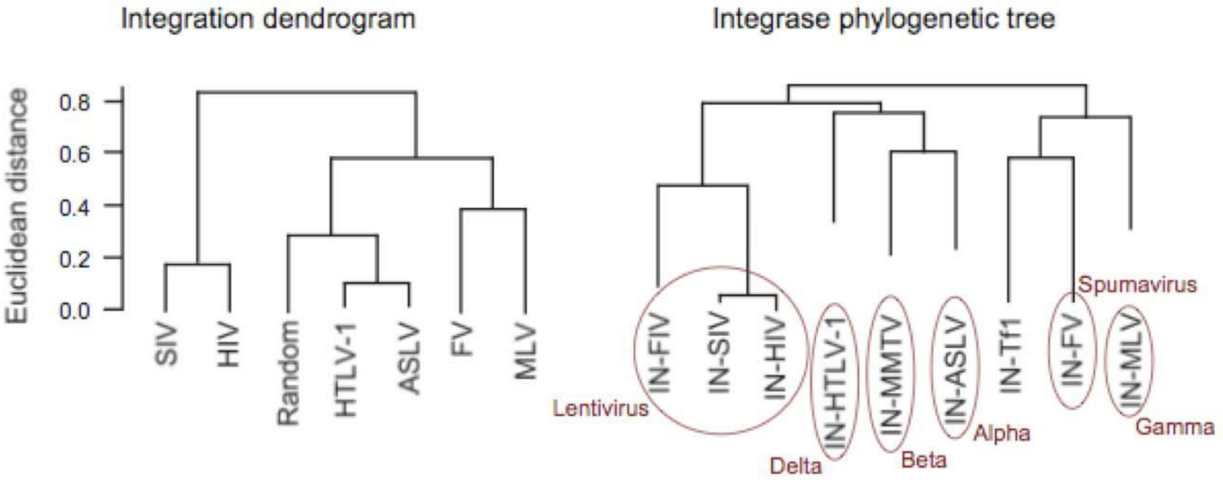


Figure 1. Retroviruses' phylogenetic trees based on integrase sequence and integration preferences. (A) Dendrogram based on location of integration in relation to 69 genomic features. Unsupervised hierarchical clustering, with euclidean distance and average linkage was used to generate the dendrogram. (B) Phylogenetic tree with integrase sequences, including P31822 (FIV), P03365 (MMTV), and AAA35339 (Tf1) integrases, showing that the three integrases have been placed into different clusters. (Derse, Crise et al.2007)

While the DNA breaking and joining reactions mediating integration are well understood, integration site selection is still a largely unknown molecular mechanism. A deeper understanding of the subject would have a double impact both on basic retrovirology as well as on its clinical applications (new targets for antiretroviral drugs and site directed gene-therapy).

Early studies considered integration a random process since there were no obvious patterns in the integration sequences, and disruption or activation of genes was thought to be a rare event. Nevertheless, some factors presumably involved in the enhancement or reduction of integration were identified, such as DNA bending induced by nucleosomal assembly, steric hidrance of DNA binding proteins and DNA physical structure. However, the assumption of random integration was reconsidered after reported high frequency events of oncogene activation in a gene therapy clinical trial of X-linked severe combined immunodeficiency (X-SCID) using retroviral vectors. 5 out of 8 patients developed leukaemia as the result of insertional activation of the oncogene LMO2 in 2 independent clinical trials, giving proof that the integration site-selection process is far from being random (Hacein-Bey-Abina, Le Deist et al. 2002; Hacein-Bey-Abina, von Kalle et al. 2003; McCormack and Rabbitts 2004).

Since complete human genome sequencing has become available, it has been possible to study in statistically rigorous manner retroviral integration by sequencing the junctions of provirus and the host's genome in genome-wide studies. Several large scale, high-throughput methods were designed and, more recently, massive parallel sequencing techniques were adapted to increase the output of integration studies (Margulies, Egholm et al. 2005; Wang, Ciuffi et al. 2007; Wang, Garrigue et al. 2008). Once acquired, integration data can be correlated with several genomic features in an attempt to find patterns and preferences of site-selection. Since no significant differences in terms of integration distribution were found between wild-type viruses and the vectors derived from them, the latter, for practical and safety matters, are most commonly used in insertion studies. The results of these large-scale surveys have been so far very revealing, not only uncovering genomic features systematically and specifically associated with retroviral insertions, but also pointing out that each retrovirus has a unique, characteristic pattern of integration within the human genome. Nevertheless, these results still fail on providing an explanation to all viral integrations and a more absolute answer as to which mechanisms are in fact involved in the virus choice to integrate. Therefore, retroviruses still remain largely unpredictable and more studies are needed to uncover its biology.

- A brief review of retroviruses and retroviral vectors

Retroviruses comprise a large family of enveloped RNA virus with virions of 80-100 nm in diameter and an outer lipid envelope that incorporates and displays the viral glicoproteins. Their RNA (7-12 kb) is linear, single-stranded, nonsegmented, and of positive polarity. The hallmark of the family is its replicative strategy, which includes as essential steps after cell entry, reverse transcription of the virion RNA into linear double-stranded DNA and the subsequent integration of viral DNA into the cell's chromatin.

Retroviruses are broadly divided into simple and complex, distinguishable by the organization of their genomes. All retroviruses contain 4 major coding domains with information for virion proteins: *gag*, for the matrix, capsid and nucleoprotein structures; *pol*, for the reverse transcriptase and integrase enzymes; *pro*, for the protease enzyme; and *env*, for the viral envelope protein. In addition to the basic coding domains, complex retroviruses also encode several accessory genes, derived from multiple splicing. Both ends of all viral genomes contain terminal noncoding sequences, composed of 5' and 3' unique sequences (U5 and U3 regions), and of two direct repeats (R) where the transcription start site and the polyadenylation signal are located (Coffin, 1997).

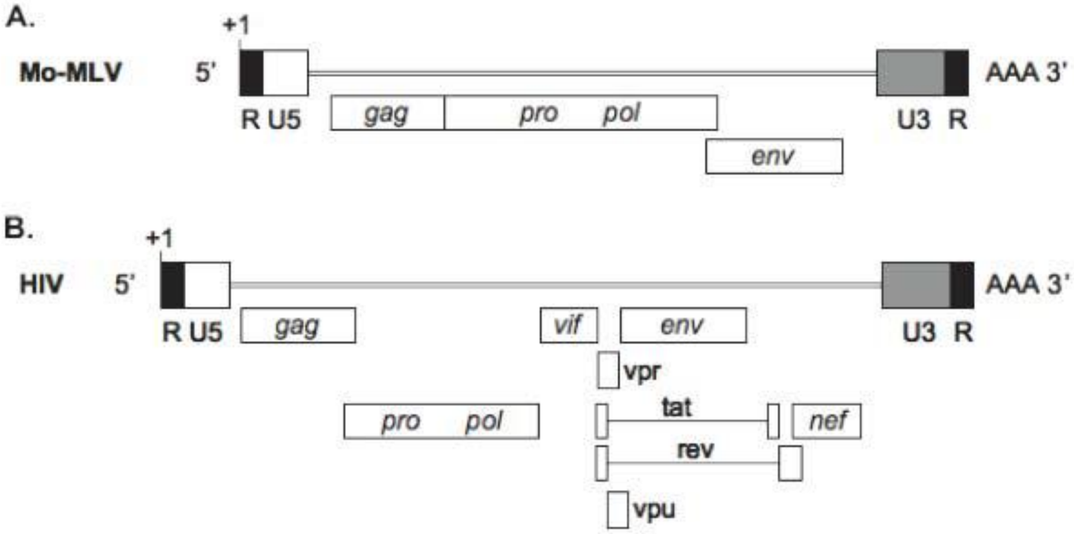


Figure 2. Simple and complex retroviral genomes. (A) Moloney murine leukaemia virus genomic RNA is only made up of four elementary coding regions, *gag, pro, pol, and env*. Terminal noncoding R, U5 and U3 regions are depicted; transcription start site (+1) and polyadenylation signal (AAA) are specified. **(B)** Human immunodeficiency virus is a complex retrovirus, with six accessory, partially overlapping genes (*vif, vpr, tat, rev, vpu* and *nef*) in addition to the four basic coding domains (Cattoglio, PhD dissertation).

Based on evolutionary relatedness, retroviruses are further classified into seven *genera*. Moloney murine leukaemia virus (Mo-MLV) and human immunodeficiency virus-type 1 (HIV-1), object of this thesis, belong to the *genus* of gamma-retroviruses (also known as oncoretroviruses) and of lentiviruses, respectively. Murine leukaemia viruses (Mo-MLV), are known to induce tumours in murine hosts and other vertebrates. Members of this family can either be endogenous or exogenous. In its replicative form, MLV has a single-stranded positive sense RNA genome that replicates via a DNA intermediate by the process of reverse transcription. HIV is a member of the Lentiviridae family, the name lentivirus means slow virus, so called because these viruses take a long time to cause exposed disease. Most lentiviruses target cells of the immune system and thus disease is often manifested as immunodeficiency. Primary targets of HIV are activated CD4+ T4 helper lymphocytes but the virus can also infect several other cell types including macrophages. Its genome, like MLV's, is made of positive sensed RNA but, as a complex retrovirus, it is also composed of accessory genes namely vif, vpr, tat, rev, vpu and nef, in addition to the four basic coding domains (gag, pro, pol and env). These genes encode small proteins and overlap with the structural genes (especially ENV) but are in different reading frames. Mutants in the TAT and REV genes show that both proteins are necessary for virus production (Meredith, 2009).

Life cycle of retroviruses and its essential interactions with the host can be summarized in five main steps. First, the host cell must express a specific receptor on its surface to provide a site for the virus to bind and trigger the entry process, mediated by the viral Env protein. Second, the cell must supply deoxynucleotides in adequate concentration for the virion reverse transcriptase to transform the RNA genome into DNA. Third, there must be a means for the viral DNA to access host chromosomes as targets for viral integration; integration may also require the aid of host repair enzymes. Fourth, host machinery and components are necessary to express viral RNA and carry out the processing (polyadenylation and splicing) and transport of both viral genomic and messenger RNA to the cytoplasm. Finally, host cell machinery is necessary for the synthesis, folding, modification, and transport of viral proteins to the membrane assembly sites (Coffin, Hughes and Varmus 1997).

The integration process, catalyzed by the viral integrase protein, takes place in the context of the preintegration complex (PIC), a nucleoprotein agglomerate formed after double stranded DNA synthesis that contains, in addition to the virus genetic material, the viral proteins integrase, nucleocapsid, virion protein R and matrix, as well as specific cellular proteins (Coffin, 1997). After completion of viral DNA synthesis, the integrase removes two nucleotides from the 3' end of both strands of viral DNA, adjacent to a conserved *CA* dinucleotide, generating recessed 3'-hydroxyl groups; in the subsequent cleavage-ligation reaction, the processed 3'-hydroxyl ends are joined to protruding 5' ends of the target DNA. Complete integration is achieved when cellular enzymes repair gaps at each host-virus DNA junction, resulting in a 4 to 6 base pair repeat in the host DNA, flanking each proviral end.

Retroviruses used in this study are retroviral vectors also applied for therapeutic purposes in the correction of genetic diseases, therefore, even though they keep the same integration pattern as wild type viruses, they differ in that they lack their basic coding elements. In this case, retroviral proteins are provided in *trans* by packaging cells so that the vectors are replication-incompetent, only the minimal viral elements required for high efficiency transfer are retained. The *cis*-acting elements essential for retroviral replication, together with the viral proteins (retrotranscriptase, integrase and protease) packaged within the virions make viral replication cycle possible. Essential elements still maintained in the viral genome include a promoter and a polyadenylation signal for viral genome production in the packaging cell, a packaging signal for incorporation of vector RNA into virions, signals required for reverse transcription and short repeats at the termini of viral LTRs necessary for integration. All the intervening genomic material can be replaced with the sequence of interest, to accommodate up to 10 kb of heterologous DNA, in our case a reporter gene.

- **Retroviral integration features: state of the art**

Several features have been investigated for a role in integration site-selection, one of the first and most intuitive was the **sequence at target DNA sites**.

Although the viral sequence required by integrase is already known (a dinucleotide CA is invariably positioned exactly 2 bp from both ends of the viral termini and the significant roles of the sequences internal to the CA dinucleotide extending for up to 15 bp also play a significant role), the sequences present at target DNA sites seem very diverse. In a recent study, Wu et al. discovered a statistically weak palindromic consensus at integration target sites while examining a large number of sequences from several retroviruses, including HIV-1, SIV, MLV, and ASLV. This consensus is weakly conserved but distinguishable between different retroviruses. The most probable hypothesis is that this conserved sequence is needed to meet the spatial or energy requirements of the integration complex, rather than the most favourite sequence at each base.

Although the host cell DNA sequences hosting integration events show some detectable similarity to one another, this similarity is very modest thus, retroviral DNA integration is not tightly sequence-specific (Stevens and Griffith 1996; Carteau, Hoffmann et al. 1998; Wu, Li et al. 2005; Berry, Hannenhalli et al. 2006).

Taking advantage of the large number of genome browsers and databases available, integration sites can be correlated positively or negatively with any genomic feature annotated, by comparison of the viral integration frequency near that feature to the frequency expected from random integration. Several genomic features were shown to have a correlation with integration site-selection: **coding and non-coding transcription units**, **CpG islands, centromeric regions, repetitive elements, fragile sites, transcription factor binding sites (TFBS)**.

Another interesting approach is to correlate viral integration with cell-specific features, like **gene expression, epigenetic modifications**, and others that reflect cell-specific transcriptional regulatory pathways. Unfortunately, only a limited number of datasets for some cell-types is already available, therefore, one should be cautious in choosing the appropriate dataset to use for correlation with integration data.

For their clinical relevance, Mo-MLV and HIV-1 are the most studied retroviruses, having also the most largely and extensively studied integration pattern.

Mo-MLV showed to have a preference for active genes and particularly for regions with a role in transcription regulation like transcription start sites, CpG islands, DNase I hypersensitive sites and transcription factor binding sites. With a distinctively different pattern from Mo-MLV, HIV-

1, seems to have a negative correlation with regions involved in transcriptional regulation. It shows to have a preference to target genes in highly active regions of chromatin. These distinct patterns of integration imply a very different viral strategy for survival. One way to interpret these data is that, Mo-MLV chooses to integrate in regions that assure its transcription and further processing. The deregulation of some classes of genes resultant of Mo-MLV integration confers some growth and/or survival advantage for the virus, leading to their *in vivo* amplification. On the other hand, HIV-1 prefers to integrate in chromatin regions that, even though assure its survival and proliferation, since these are highly conserved and active, are not involved in transcriptional regulation and thus not prone to gene deregulation. This strategy implies a longer period of survival for the host and, consequently, a higher chance of viral propagation to a large number of hosts.

- Proposed mechanisms for integration site selection

It has been proposed that integration is favoured in regions of **open chromatin** once MLV tends to integrate near DNase I hypersensitive-sites, maybe because these regions might be more accessible to the pre-integration complex (PIC) (**Figure 3**). This hypothesis has been supported by evidence of preferential integration in transcriptionally active chromatin and disfavoured integration in the centromeric heterochromatic region. However, only MLV strongly favoured integration near to these sites and the diversity of genomic targets both in MLV and among other retroviruses implies more complex and distinct mechanisms involved in integration site-selection for each family.

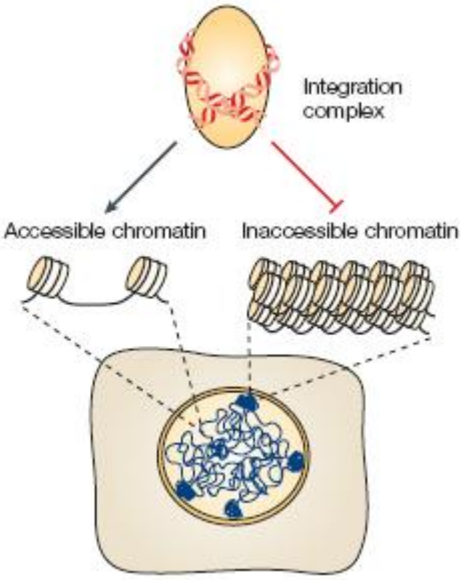


Figure 3. Proposed mechanisms that direct integration site selection by retroviruses- accessibility of target DNA. According to this model, chromosomal DNA is relatively inaccessible to integration complexes when packed in nucleosomes and other proteins. Exposure of target sequences promotes integration. (Bushman, Lewinski et al. 2005)

A different model invoked the effect of **cell cycle** in the mechanism for integration targeting (**Figure 4**). HIV can infect cells regardless of the cell-cycle phase, while MLV requires the host cell to pass through mitosis. The transcriptional state of a cell is known to vary with the cell cycle, so the organization of chromosomal DNA encountered by HIV and MLV complexes should differ. However, HIV constructs with cell-cycle restricted infectivity showed a very different integration pattern in relation to MLV, even though the pattern was also different from the wild type HIV. These results suggested that cell cycle does play some role in retroviral targeting but other factors seem to dominate. Additionally, studies of HIV integration targeting in non-dividing cells and in dividing cells didn't show significant differences, providing more evidence against this targeting model.

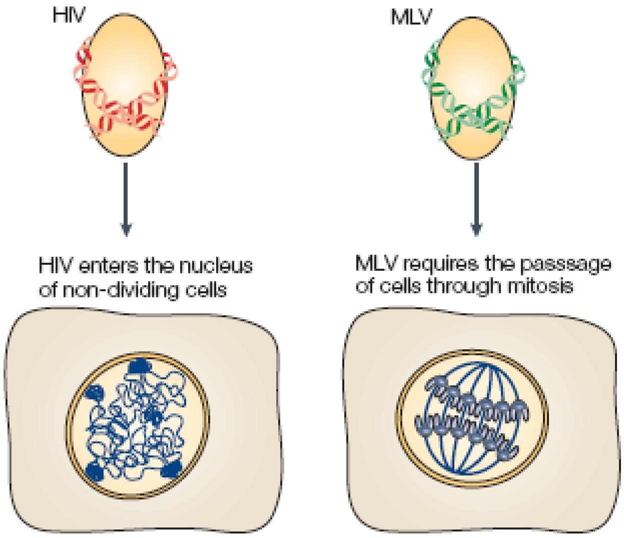

HIV
MLV
HIV enters the nucleus of non-dividing cells
MLV requires the passage of cells through mitosis

Figure 4. Proposed mechanisms that direct integration site selection by retroviruses - timing of nuclear entry during the cell cycle. To enter the nucleus, MLV requires the passage of cells through mitosis, whereas HIV can enter cells at different stages of the cell cycle. If the state of chromosomes differs at different points in the cell cycle, this could influence integration targeting (Bushman, Lewinski et al. 2005).

Another candidate mechanism for integration is the **tethering of PIC** to the genome by cellular or viral proteins (**Figure 5**). Through binding of the PIC to this (or these) supposed protein(s), the complex would be directed to a specific location in the genome where it would eventually integrate. Several cellular proteins have been isolated as physically bound to viral PICs, some common to both HIV and MLV and others specific, also for some of them, association occurs via direct interaction with the viral integrase. Some of these proteins studied so far will be described in the next paragraphs.

Kalpana et al. 1994 used the HIV-1 integrase protein as "bait" to screen for cellular proteins that might participate in viral DNA integration. Their screen yielded a cDNA clone encoding a protein that binds specifically to integrase both *in vivo* and *in vitro.* The integrase-binding protein, designated **integrase-interacting protein 1 (Ini1)**, displays a high degree of sequence similarity to the yeast protein Snf5, implicated in the transcriptional activation of a number of genes. They also found that, at certain concentrations, Ini1 increased the efficiency of integration (Kalpana, Marmon et al. 1994).

Farnet et al. 1997, presented data indicating that the **high mobility group protein HMGA1** (previously known by HMGI(Y)) is required for function of HIV-1 preintegration complexes (PICs) isolated from infected cells. Integration activity was lost from PICs following treatment with 600mM KCl. This activity could be restored, however, by addition of extracts from uninfected SupT1 cells, suggesting that a host activity might be required. Screening for the complementing protein yielded

HMGA1. Analysis of protein composition by western blotting revealed that HMGA1 was present in PICs, but depleted from PICs by high salt treatment. Purified HMGA1 alone was not sufficient to carry out integration when mixed with purified HIV-1 cDNA, supporting a model in which HMGA1 is required as an accessory factor for the function of HIV-1 PICs. This work shows that the function of HMGA1 in integration seems to be related to one of the steps of integration process, possibly the covalent strand transfer step, rather than to PIC targeting in the genome. Lin et al. 2003, also studied the HMG 1 family proteins in Mo-MLV, finding a similar role to the one found in HIV-1. The outcome of this study also suggested that binding of multivalent HMGA1 monomers to multiple cDNA sites compacts retroviral cDNA, thereby promoting formation of active integrase-cDNA complexes. HMGA1 has not been found to bind to integrase directly, consistent with models in which HMGA1 acts by binding to the cDNA (Farnet and Bushman 1997).

Barrier-to-autointegration factor (BAF) is a homodimeric protein with 89 amino acid residues and is highly conserved among species. This protein was found to protect viral DNA from autointegration both in Mo-MLV and HIV. Furthermore, it was found that BAF can promote the association of PICs with target DNA. In vitro studies have revealed that BAF bridges double-stranded DNA with no detectable sequence specificity. DNA bridging results in intramolecular compactation at low DNA concentrations and intermolecular aggregation at high DNA concentrations. Therefore BAF's activity may have two different outcomes: intramolecular bridging may compact the viral DNA into a rigid structure, making it less accessible as a target for autointegration and anchoring of PICs to other DNA may promote intermolecular integration in target DNA (Suzuki and Craigie 2002; Lin and Engelman 2003).

The human protein **Rad18** is also known to interact with integrase. This interaction results in an increased stabilization of integrase, which in its natural form is a particularly unstable protein. The re-localization of integrase and its co-localization with Rad18 in a subset of cells suggests an additional function for this association. Human Rad18 contains a putative SAP-box, a domain recently recognized to mediate the binding of certain proteins to specific A/T-rich DNA regions known as the scaffold attachment regions (SAR). Interestingly, PARP-1, Ku antigens, and HMGA1, which are involved in retroviral integration, have all been found to be SARbinding proteins. An intriguing possibility is that the molecules relevant for HIV-1 integration cluster together, perhaps in the vicinity of SARs, achieving in this way the coordination required for these complex reactions (Mulder, Chakrabarti et al. 2002).

Human **embryonic ectoderm development protein (EED)**, a member of the superfamily of WD-40 repeat proteins, and of the Polycomb group proteins, has been identified as a cellular partner of the matrix (MA) protein of HIV-1. EED was also found to interact with the integrase both *in vitro* and *in vivo* in yeast. EED-binding site(s) are located in the C-terminal domain of the integrase, between residues 212 and 264. In EED, two putative discrete integrase-binding sites were mapped to its N-terminal moiety, at a distance from the MA-binding site, but EED-integrase interaction also required the integrity of the EED last two WD repeats. EED showed an apparent positive effect on integrase-mediated DNA integration reaction in vitro, in a dose-dependent manner (Violot, Hong et al. 2003).

Poly(ADP-ribose)polymerase-1 (PARP-1) is a nuclear protein mainly associated with chromatin, and is proposed to be involved in the process of DNA repair, including the repair of gapped intermediates generated during retroviral integration. This protein accumulates at the active mammalian centromere on metaphase chromosomes, and is associated with centromeric DNA and proteins, suggesting that PARP-1 and poly(ADP-ribosyl)ation reaction may be involved in the regulation of centromere function. Kameoka's et al. results suggested that poly(ADP-ribosyl)ation, although it's not required for efficient HIV-1 integration, seems to be necessary during integration near the centromere region. Even though a low frequency of integration has been shown in centromeric DNA, these rare events represent an important retroviral mechanism of survival since integration in these sites leads to the establishment of a latent infection that can be reactivated (Kameoka, Nukuzuma et al. 2005).

The most studied cellular tethering factor so far, is the **lens epithelium derived growth factor (LEDGF) or p75**, an ubiquitously expressed nuclear protein, tightly associated with chromatin throughout the cell cycle. It is a transcriptional coactivator involved in stress response, autoimmune disease, cancer and HIV replication. This protein was shown to be a strong interactor of the HIV-1 integrase and to stimulate its catalytic activity in vitro. It binds at its C-terminus to lentiviral IN protein dimers while the N-terminal half binds to chromatin (Llano, Vanegas et al. 2006; Engelman and Cherepanov 2008). When LEDGF/p75 was depleted from cells using RNA interference, integration in transcription units was diminished, documenting a role in integration targeting (Ciuffi, Mitchell et al. 2006). LEDGF/p75-responsive genes were identified by transcriptional profiling and found to be favored integration targets for both HIV (Ciuffi, Llano et al. 2005) and another lentivirus, feline immunodeficiency virus (Kang, Moressi et al. 2006). Correlation analysis with genomic features revealed an association with active chromatin markers, such as H3 and H4 acetylation, H3K4 monomethylation and RNA polymerase II binding. Interestingly, some associations did not correlate with HIV-1 integration indicating that not all LEDGF/p75 complexes on the chromosome are prone to HIV-1 integration (De Rijck, Bartholomeeusen et al.). Moreover, in cells depleted for LEDGF/p75, HIV integration was still favoured in transcription units. Thus additional factors may be involved in guiding HIV integration (Ciuffi, Llano et al. 2005).

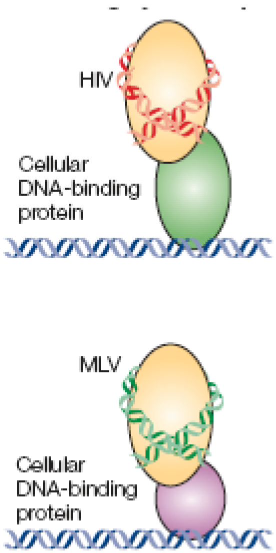


Figure 5. Proposed mechanisms that direct integration site selection by retroviruses - tethering by cellular proteins. This model proposes that specific interactions between integration complexes and cellular proteins bound locally on target DNA promote integration at nearby sites. (Bushman, Lewinski et al. 2005)

Although essential for PIC integration activity, interaction with cellular proteins doesn't explain by itself the integration characteristics of Mo-MLV. Integrase has an essential role targeting sites for MLV integration. Proof of this principle comes from experiments using an HIV vector packed with a MLV integrase, which gains preference for MLV integration sites (regulatory regions with transcription start sites (TSSs), CpG islands and TFBSs). Additionaly, the high integration near TFBSs is dramatically reduced when the U3 transcriptional enhancer is deleted from the Mo-MLV LTRs. These results indicate that the integrase and the U3 enhancer are the major viral determinants of Mo-MLV selection of regulatory regions in the genome. A plausible explanatory model is that **cellular** transcription factors binding the Mo-MLV U3 enhancer cooperate with the integrase in directing PICs towards regions actively engaged by the transcriptional machinery. Accordingly, some of the TFBSs enriched around Mo-MLV integrations are *consensus* motifs for transcription factors already known to bind the U3 enhancer and drive proviral expression after integration (*e.g.,* members of the ETS family and the bivalent YY1 transcription factor).

This thesis analyzes a collection of retroviral integration sites retrieved from primary human keratinocytes at an early time point after infection, when clonal selection in culture is very unlikely to have occurred. Epithelial keratinocytes are exposed to retroviral infection, for example during oral-sexual contact and breast-feeding, making the epithelium a potential site of primary retroviral infection and dissemination. Even though these cells do not express the common receptor and co-receptors found on permissive cells, integration is shown to occur and persist in daughter cells after keratinocytes' division (Vacharaksa, Asrani et al. 2008). After integration the life cycle of the virus aborts and no newly assembled virus particles are detectable, although, captured infectious viruses are harboured for at least 48h and transferred to blood mononuclear cells (Vacharaksa, Asrani et al. 2008). Epithelial cells can, therefore, be actively providing a route to systemic retroviral infection, making them an interesting target to study viral integration mechanisms.

The general goal of this work was to describe the integration preferences of retroviruses in the genome of these clinically relevant cells, perform a comparative analyses with previously obtained results for other cell-types and possibly provide new insights into the molecular mechanisms responsible for differential retroviral integration targeting. Furthermore, we intended to evaluate the safety parameters related to the use of these viral vectors in protocols of gene therapy for genetic skin diseases, the main research line followed in the laboratory.

METHODS AND MATERIALS

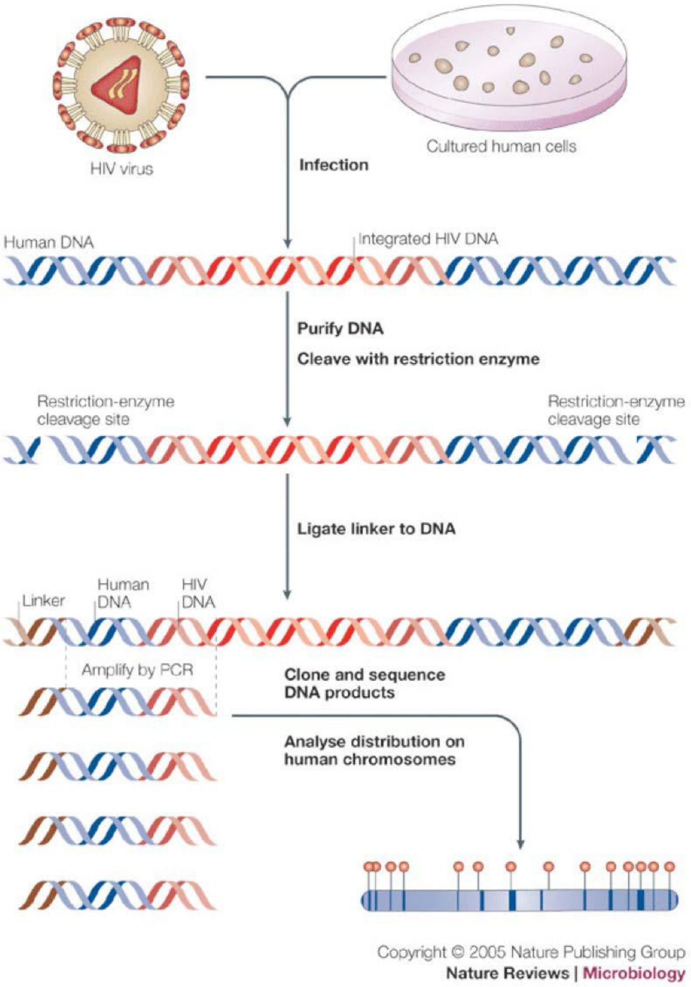

HIV virus
Cultured human cells
Infection
Human DNA
Integrated HIV DNA
Purify DNA
Cleave with restriction enzyme
Restriction-enzyme cleavage site
Restriction-enzyme cleavage site
Ligate linker to DNA
Linker
Human DNA
HIV DNA
Amplify by PCR
Clone and sequence DNA products
Analyse distribution on human chromosomes
Copyright © 2005 Nature Publishing Group
Nature Reviews | Microbiology

Figure 6. General overview of the methods used to analyze retroviral integration sites in the human genome (Bushman, Lewinski et al. 2005)

1. Virus production and cell culture

The vectors used were gamma-retroviral MFG.GFP and lentiviral derivated pRRLsin-18.ppt.K14.GFP.WPRE. Briefly, MFG.GFP vector expressing the EGFP cDNA under the control of the Moloney leukaemia virus LTR was constructed in the MFG backbone and integrated in the amphotropic Gp+envAM12 packaging cell line. The viral supernatant used for keratinocytes infection was collected from one clone of Am12MFG GFP packaging cell line. The VSV-G-pseudotyped pRRLsin-18.ppt.K14.GFP.WPRE lentiviral vector, containing a K14 promoter-driven GFP expression cassette, was produced by transient co-transfection of 293T cells with a second generation packaging system. Viral supernantant collected from transfected 293T cells was then concentrated by ultracentrifugation in order to increase the viral titer. Transfection efficiency was evaluated by analysis of GFP expression by flow cytometry 48 hours after infection.

2. Transduction of cells

For retroviral transduction subconfluent primary keratinocytes (taken from male newborns foreskin) were trypsinized and plated (2x10^5 cells) in multi-6 wells plates, without feeder layer, in the presence of KBM (Lonza) medium supplemented with glutamine (2%) insulin (5 µg/ml) adenite (0.18mM) hydrocortisone (0.4 µg/ml) and EGF. Retroviral transduction was performed by spinoculation (3 rounds at 1,800 rpm for 45min) in the presence of 8 µg/ml polybrene.

For lentiviral transduction subconfluent primary skin keratinocytes were trypsinized and 10^5 cells resuspended into 2 ml of Kno medium containing the retroviral virus at MOI of 8-10 in the presence of 8ug/ml polybrene. The transduction mixture containing keratinocytes, virus and polybrene was then plated on lethally irradiated 3T3-J2 cells. The medium was replaced after 5 hours. Transduced keratinocytes were grown to confluence, trypsinized and re-plated onto new feeder-layers for further analysis. Transduction efficiency was evaluated by analysis of GFP expression by flow cytometry 48 hours and two weeks after infection.

3. Cloning and analysis of viral insertion sites

Genomic DNA was extracted and digested with *Mse*I. 3' LTR vector genome junctions were amplified by linker-mediated PCR (LM-PCR) (Cattoglio et al., 2007), adapted to the GS-FLX Genome Sequencer (Roche/454 Life Sciences, Branford, CT) pyrosequencing platform. For each transduction, we performed 2 restriction digestions, 6 linker ligations and 18 nested PCRs, with nested primers specific for the linker and the 3' LTR containing a bead-capture tag and a sequencing tag. A 4-nucleotide multiplex tag was also added to the 3' LTR nested primer downstream of the sequencing tag to discriminate between different samples (see Supplemental methods). Pooled LM-PCR amplicons were quantified (NanoDrop Technologies, Wilmington, DE), checked by an Agilent Bioanalyzer (Agilent Technologies, Palo Alto, CA), size-fractionated by SPRI beads (Agencourt Bioscience Corporation, Beverly, MA), and sequenced according to the GS-FLX manufacturer's instructions.

4. Sequence analysis

All UCSC Known Genes having their transcription start site (TSS) at ±50 kb from an integration or random site were annotated as targets, as done on previous works (Cattoglio et.al, 2007). In case of multiple transcript variants, we arbitrarily chose the isoform with the nearest TSS to an integration or random site. For each site, we annotated the genomic features (CpG islands, conserved non-coding sequences, conserved TFBSs) whose hg18 coordinates overlapped for at least 1 nucleotide with the ±50 kb interval around the insertion site. We used UCSC tracks (http://genome.ucsc.edu) for both CpG islands (27,639 items) and conserved TFBSs (3,807,783 items). Genomic coordinates of 82,335 mammalian CNCs were described (Kim and Pritchard 2007). To generate a matched control dataset, we randomly extracted 20,000,000 sites from the human genome and discarded sites with the nearest *Mse*I recognition site (TTAA) at <20 bp (the minimum requirement for a blast search) or >500 bp (the maximum estimated length for efficient

454 bead loading). The resulting 14,260,000 sequences were processed through the same bioinformatic pipeline used for integration sequences, ending with a library of 11,655,601 unique sites.

5. Gene expression profiling

The expression profile of primary keratinocytes was determined by microarray analysis. RNA was isolated from 1-2 x 10^6 primary human keratinocytes, transcribed into biotinylated cRNA and hybridized to Affymetrix HG-U133A Gene Chip arrays according to the Affymetrix instructions. Scanned images were processed by the Affymetrix GCOS suite, and transcript levels determined with the GCOS absolute analysis algorithm. To correlate retroviral integration and gene activity, expression values from the keratinocytes microarrays were divided into four classes, i.e. absent, low (below the 25th percentile in a normalized distribution), intermediate (between the 25th and the 75th percentile) and high (above the 75th percentile).

6. Functional gene analysis

Genes were analysed by the network-based Ingenuity pathways analysis tool (Ingenuity Systems, www.ingenuity.com). Gene identifiers were uploaded into the application, and mapped to their corresponding Focus Gene in the Ingenuity Pathways Knowledge Base. Networks were algorithmically generated based on the direct or indirect interaction between Focus Genes. The functional analysis of each network identified the biological functions and/or diseases that were most significant to the genes in the network (Bonferroni correction).

Results

The retroviral vector used for this study is an MFG-based retroviral vector expressing the eGFP cDNA under the control of Moloney leukemia virus LTRs; the lentiviral vector is a SIN-18 vector in which the expression of the same cDNA is under the control of the internal full-length K14 promoter (for maps of retroviral vectors see Appendix).

Human primary keratinocytes from healthy donors were transduced with the gamma-retroviral and lentiviral vector with an efficiency of 60% and 80% EGFP positive cells, respectively. Transduced cells with lentiviral vector were cultivated onto a feeder layer while transduced cells with gamma-retroviral vector were cultivated directly on plastic dishes to avoid murine feeder contamination (since murine cells are the original host of these viruses). Genomic DNA was extracted and vector-genome junctions were cloned and sequenced by a LM-PCR approach adapted to the different vector types (Cattoglio et al., 2007) and mapped onto the human genome. Cumulatively, we mapped 3020 MLV and 1601 HIV independent integrations.

Analysis of the distribution of viral integrations along the genome was performed on my dataset originated from primary human keratinocytes. Furthermore, my data was compared with previous datasets of Cd34$^+$ hematopoietic stem cells (32631 MLV hits and 28382 HIV hits) from Cattoglio *et al.* Blood in press and lymphocytes (8277 MLV hits) from Recchia *et al.* unpublished and (7780 HIV hits) from Wang *et al.*, 2007.

Retrovirus integrate preferably inside genes

Viral and random integrations along the genome of the three cell-types, were classified as being inside genes or outside genes. **Figure 7** represents the distribution as a % of the total number of integrations, showing that HIV has a high tendency to integrate inside genes ≈75% in keratinocytes and Cd34$^+$HSC and a even higher preference in lymphocytes with 83% of integrations inside genes ($p < 2.2e-16$).

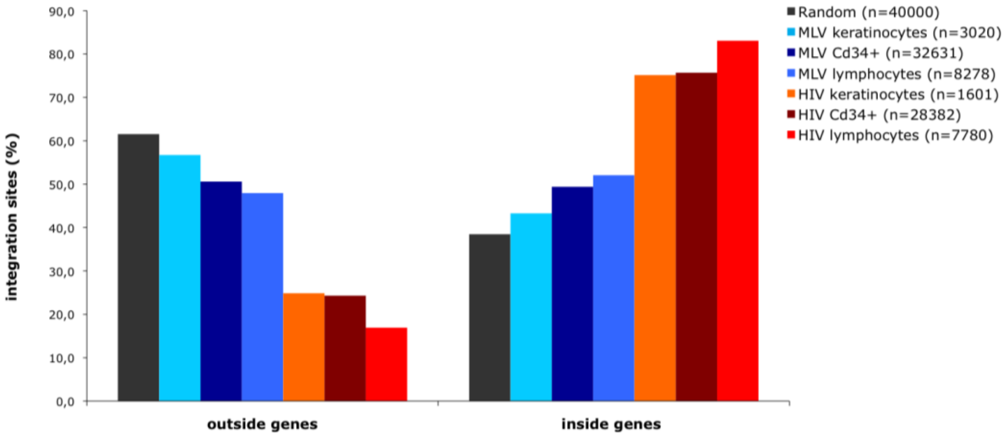


Figure 7. Distribution of MLV, HIV and random integration sites with respect to Known Genes (UCSC definition) on primary human keratinocytes, lymphocytes and Cd34$^+$HSC.

MLV shows a much lower but still, statistically relevant (two-sample test for equality of proportions with continuity correction, Rweb 1.03) preference to integrate inside genes than would be expected from a random distribution through the genome in all cell-types, although, with a significantly higher freuency in lymphocytes and Cd34$^+$HSC in relation to keratinocytes with 52%, 49% and 43% of integrations inside genes, respectively ($p < 0,0001815$). Random integrations were 38,5% inside genes and 61,5% outside, reflecting mostly the gene content of the human genome.

Following the previous results, we hypothesized that chromosomes with an overrepresentation of viral integration would be the ones with higher gene density. To verify the proposed hypothesis, integration frequency along chromosomes of keratinocytes was analysed by comparison with the random distribution and then correlated with values of gene density for each human chromosome. In fact, we could confirm an association between viral integration and gene density, since the 4 more gene dense chromosomes (17, 19, 20, 22) (data on chromosome gene density taken from NCBI http://www.ncbi.nlm.nih.gov/genome/guide/human/) perfectly overlapped with the 4 more targeted ones (**Figure 8**), however, the variation observed in all the chromosomes could not be explained solely by the gene density parameter.

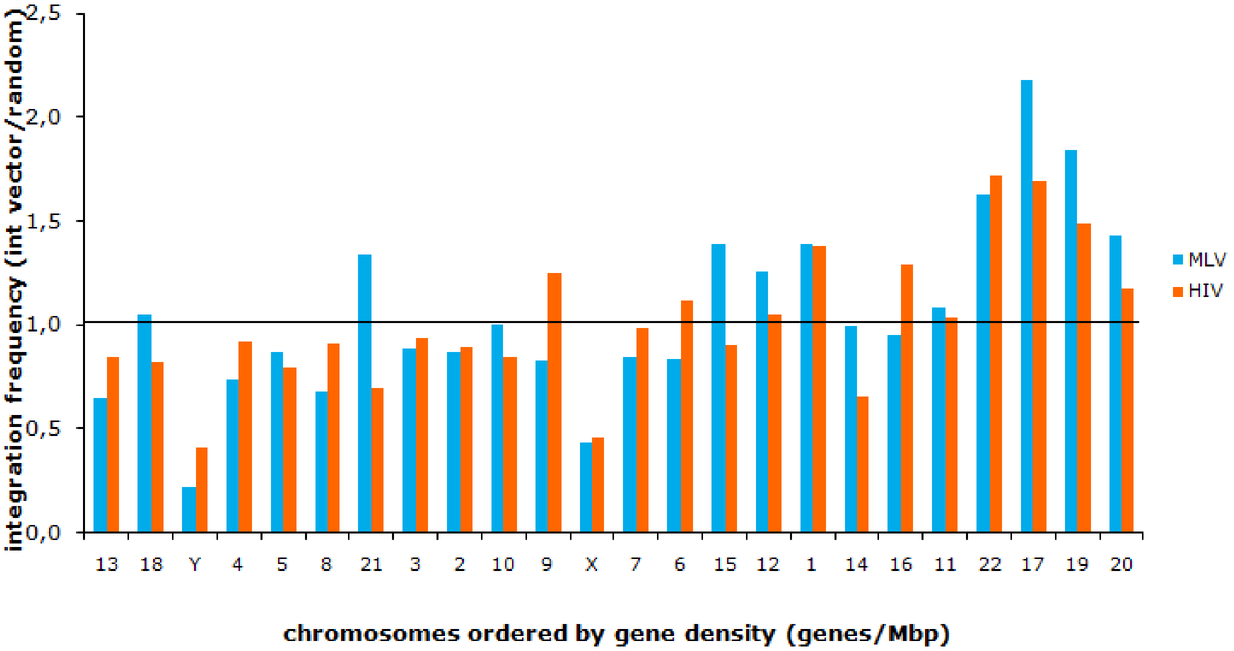


Figure 8. Frequency of integration of MLV (blue bars) and HIV (orange bars), in relation to the control, according to the gene density of chromosomes in primary human keratinocytes.

MLV is attracted to TSS proximal regions while HIV is repelled

Single integrations were then classified into three groups (**Figure 9**), TSS proximal when located at a distance of +/-2,5 kb from the TSS of any gene (UCSC Known Gene track), intragenic when located inside a gene but not near a TSS and intergenic when outside any gene and at a distance higher then 2,5 kb from any TSS.

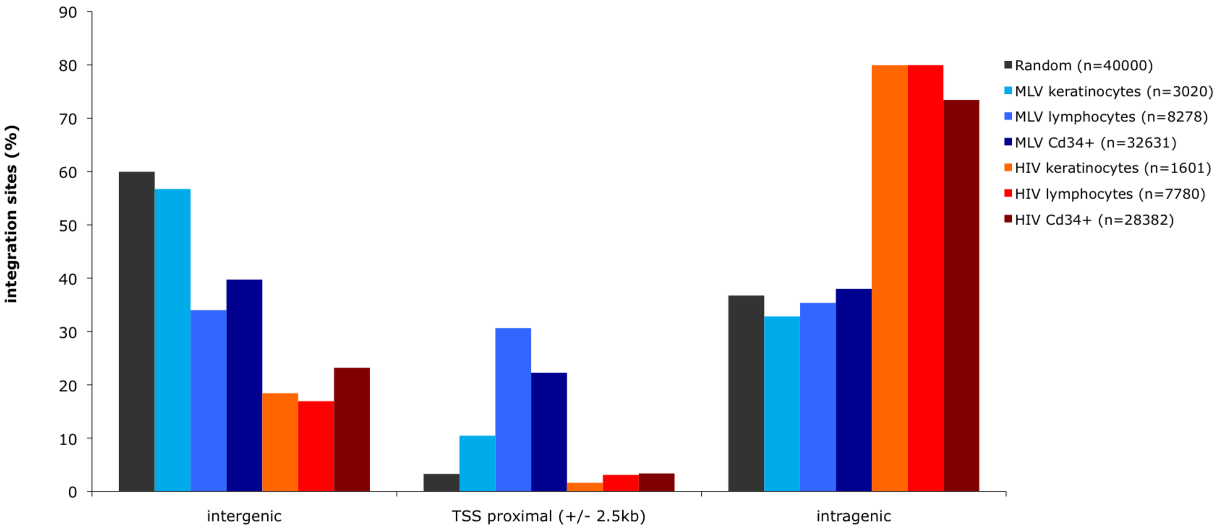


Figure 9. Distribution of MLV, HIV and random integration sites with respect to Known Genes (UCSC definition) on primary human keratinocytes, lymphocytes and Cd34+HSC.

MLV showed a higher integration in TSS proximal regions in comparison to the control ($p < 2.2e-16$) in all cell-types. Lymphocytes showed the highest bias towards TSS regions accounting for 30,6% of all integrations, followed by Cd34^{+}HSC with 22,3% and the lowest frequency for keratinocytes with 10,5% of integrations in this region. In contrast, HIV integrations were disfavoured by TSS proximal regions in keratinocytes with only 1,6% ($p = 0.000291$) of total integrations and had a similar distribution to the control in the other two cell-types, ≈3%.

To further look at this particular feature, all integration events occurring within 50kb from the TSS of any Known Gene were plotted in 2,5 kb intervals in either direction of the TSS (**Figure 10 A**). MLV showed a high incidence of integrations within 5 kb from the TSS while HIV showed a moderate but statistically relevant higher incidence of integration downstream of the TSS confirming it's favouritism for intragenic integrations. Mapping at a closer distance from the TSS at 5kb in both directions (**Figure 10 B**), revealed that MLV is attracted to TSS regions while HIV seems to be repelled by the same regions. Both MLV and HIV have a mirror plot with MLV increasing the number of integrations near the TSS while a decrease of integration is shown by HIV. In the region close to the TSS MLV shows a sudden drop of integration, in lymphocytes and Cd34^{+}HSC this occurs in the 0bp to 100bp region while in keratinocytes the same drop is localized in the region 100bp downstream from the previous ones. A second drop is seen in the region from 300bp to 400bp from the TSS in lymphocytes and, once again, 100bp downstream in keratinocytes, but is not observable in Cd34^{+}HSC. These drops in integration frequency are possibly due to occupancy from the basal transcription machinery. Both of these distinct characteristic patterns of integration represent a viral integration distribution non-dependent on cell specificities once it was observed in all the studied cell-types. Although, a significant higher frequency of integration by MLV was observed in lymphocytes, mainly at a distance from -1,7 to

1,7 kb from the TSS of these cells. HIV's distribution showed no differences between all the studied cell-types. Random integrations were evenly distributed throughout the analysed region.

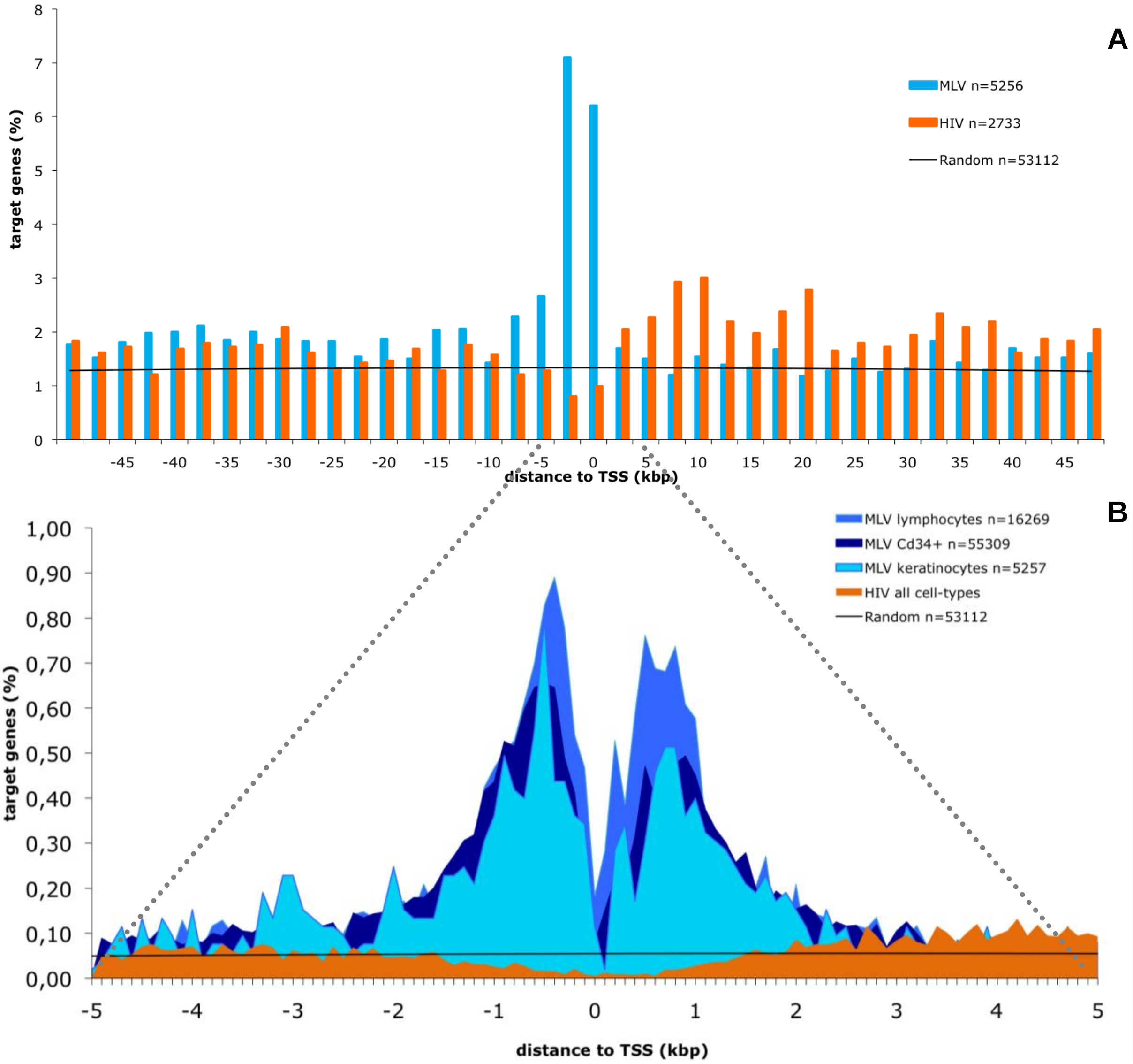


Figure 10. Distribution of retroviral integrations around transcription start sites. The % of the total number of targeted genes (n) is plotted on the Y axis. The black line indicates the distribution of control random sites. **(A)** Distribution of the distance of MLV (blue bars) and HIV (orange bars) integration sites in primary human keratinocytes from the transcription start site (TSS) of targeted genes at 50kp resolution. **(B)** Comparative analysis of distribution of retroviral integrations around transcription start sites on primary human keratinocytes, lymphocytes and Cd34^{+}HSC at 5kbp resolution.

Retroviruses target mostly active genes

To understand if the genes targeted by retrovirus, in particular those with TSS proximal integrations, were active, an Affymetrix microarrays run (HG-U133 plus 2.0) was performed in triplicate to determine the expression profile of over 18,900 genes in keratinocytes activated in culture in the same conditions used for retroviral transduction. Affymetrix probe sets were re-annotated with custom CDF files according to the Bioconductor indications (Dai, Wang et al. 2005), to obtain a single expression value for each gene. Expression levels were divided into four classes: absent (black portion of histogram bars), low (below the 25th percentile of the normalized distribution, yellow), intermediate (between the 25th and the 75th percentile, orange) and high (above the 75th percentile, red). The percentage distribution of the expression values of genes targeted by all integration/random sites (all ISs), TSS-proximal sites (TSS-proximal ISs) and intragenic sites (intragenic ISs) are shown by the left, middle or right group of bars, respectively.

In keratinocytes, both MLV and HIV showed a significant preference for active genes, once 69% and 74% (p-value < 2.2e-16) respectively of targeted genes were active in all insertion sequences, compared to 58% of the random. A even bigger difference was revealed by MLV's integration in TSS proximal regions with 85% (p-value < 2.2e-16) of active genes and 87% (p-value = 9.421e-10) active in intragenic integrations as opposed to 60% of activity both in TSS and intragenic regions with the random. HIV had no significant preference for TSS proximal active genes but showed an evident preference for active genes part of intragenic integrations (p-value < 2.2e-16).

Results previously obtained for Cd34$^+$HSC (Cattoglio *et al*., Blood in press) revealed the same overall preference for active genes. In comparison to the actual data, we can suggest that there's a even higher preference for HIV to integrate in active genes among the intragenic integrations in Cd34$^+$HSC than in keratinocytes with 88,5% and 73% of targeted expressed genes, respectively.

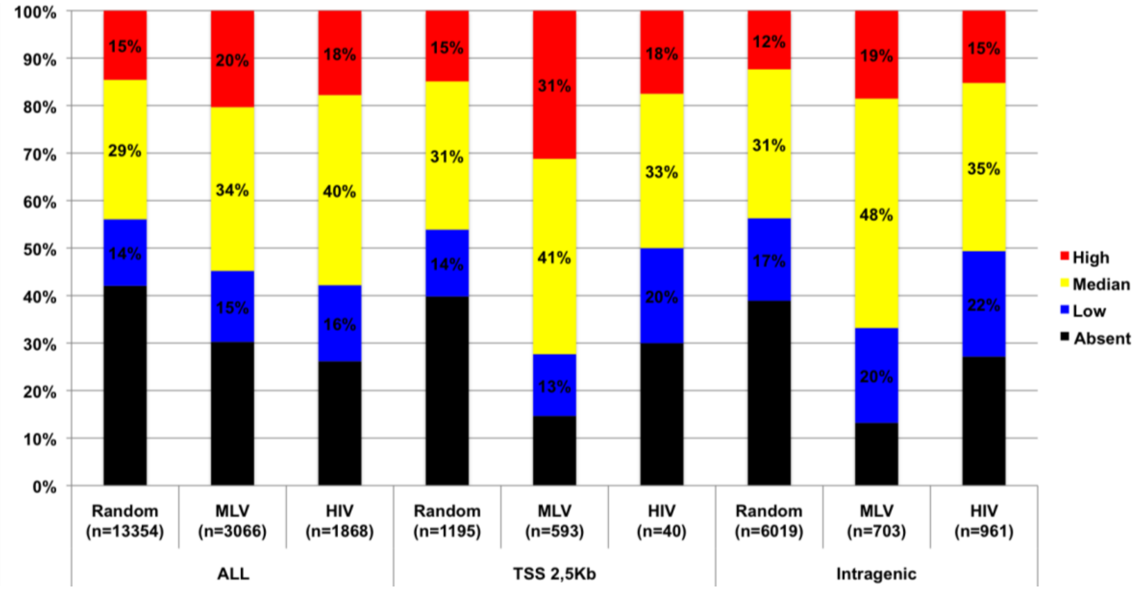


Figure 11. Association between retroviral integration and gene activity in primary human keratinocytes. Histogram distribution of expression values from an Affymetrix microarray analysis of RNA obtained from primary human keratinocytes. The number of genes belonging to each category is indicated in parenthesis under the correspondent bar. Genes targeted more than once are considered more than once for the analysis.

Moreover, with the Affymetrix data for gene expression in Cd34^{+}HSC cells we analysed the proportion of genes hit (intronic and exonic integrations) by both viruses and the random dataset in comparison to the total number of genes in the genome. A significantly lower number of non-expressed genes were targeted by both MLV and HIV, corresponding to 6% and 5%, respectively, of the genome in comparison to 14% of the random distribution. No differences were seen in the proportion of expressed target genes by MLV and the random which corresponds to 20% of genes in the genome, however a significant difference was observed for HIV which targeted 26% of expressed genes (p-value < 2.2e-16). Globally MLV targeted less genes with a total of 24% in comparison to the random and HIV which targeted 34% and 31%, respectively.

Regions highly targeted by retroviruses, defined as clusters of integration, were also analysed, comprising ≈ 10% of the genes in the genome in both viruses. Expressed genes hit by clusters accounted for 8% and 10% of the genes in Cd34+ cells for MLV and HIV, respectively, while non-expressed genes hit by clusters represented only 2% and 1% of the total number of genes in the genome.

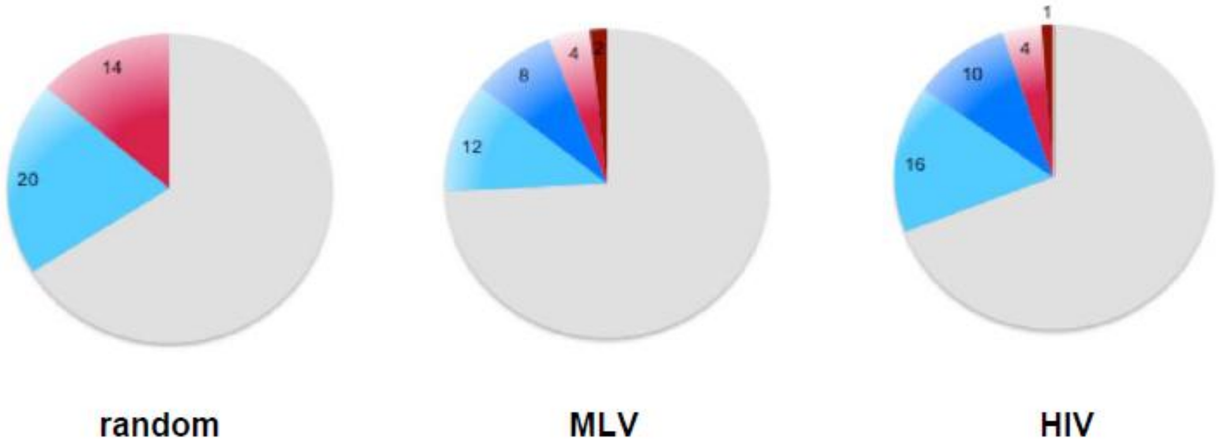


Figure 12. Proportion of genes targeted by random, MLV and HIV distribution, in the genome of human Cd34+ cells. Blue represents targeted expressed genes, red represents targeted non-expressed genes and grey non-targeted genes. Dark blue and red show the proportion of genes targeted in clusters.

MLV's clusters are more packed while HIV's expand through a wider chromatin region

To analyse features present in highly targeted regions, clusters of viral integration in the genome were defined taking in consideration the size of the dataset and the distance from one integration to the consecutive one. A maximal distance between consecutive integrations corresponding to a false discovery rate of 0.01 was statistically established to define if an integration is inside or outside a cluster. For the keratinocytes dataset a distance of less or equal to 9,905 bp and 17,318 bp between two consecutive integrations was considered for MLV's and HIV's integrations, respectively. For the Cd34^{+}HSC dataset we considered a window of three integrations within 12,587 bp for MLV and three integrations within 14,460 bp for HIV.

With this definition, 320 clusters for MLV and 102 for HIV in keratinocytes were identified, containing 23.6% (714) and 13.6% (218) of the total integration sites. For the Cd34^{+}HSC dataset 3,497 MLV clusters were identified comprising 65.3% of all integrations and 2,446 HIV clusters with 50.6% of the total number of integrations.

The size of clusters was very different between the two viruses, MLV's clusters being more compact and HIV's spreading through a larger region of the chromatin. Differences in clustering pattern were even more evident in the larger dataset for Cd34^{+}HSC, where cluster size varied in length from 3,199 bp to 78,534 bp for MLV and from 7,271 bp to 200,508 bp for HIV's clusters. Furthermore, HIV's clusters have a maximum of 110 integrations by cluster while MLV's maximum is less than half, with 42 integrations by cluster in Cd34+HSC.

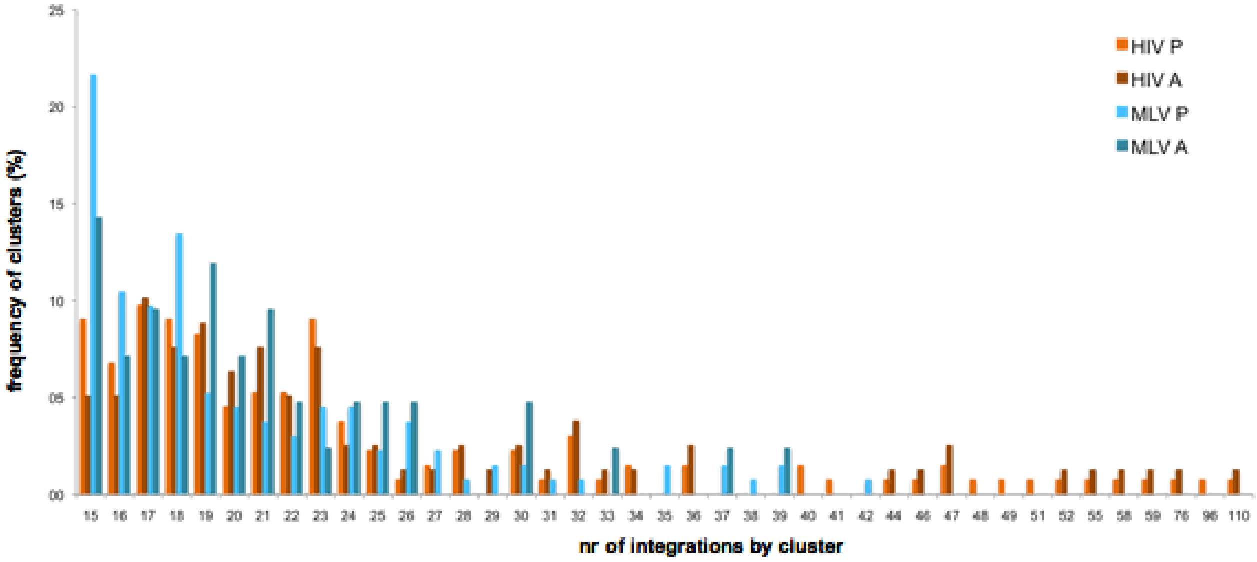


Figure 13. Cd34+ cells cluster analysis. Graph shows the frequency of clusters in relation to the number of integrations by clusters in both MLV (blue bars) and HIV (orange bars). Clusters were classified as having present expression (lighter bars) or absent expression (darker bars) according to its gene constitution. Clusters with both expressed and non-expressed genes were considered twice for analysis purposes.

MLV integration positively associates with epigenetic markers of promoters and enhancers

Using UCSC track for epigenetic markers in keratinocytes, histone modifications present in clusters were analysed. MLV's clusters were positively correlated with histone modifications associated with enhancers and promoters of active genes, namely H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac and also with polymerase 2. Histone modifications associated with inactive genes and heterochromatic regions like H3K9me1 and H3K27me3, and H4K20me1 or to the body of actively transcribed genes like H3K36me3 were negatively correlated with MLV's clusters in keratinocytes. For HIV's clusters no obvious association with epigenetic markers was seen, therefore, we would have to perform a global analysis on all the integrations to have a better notion of its preferences.

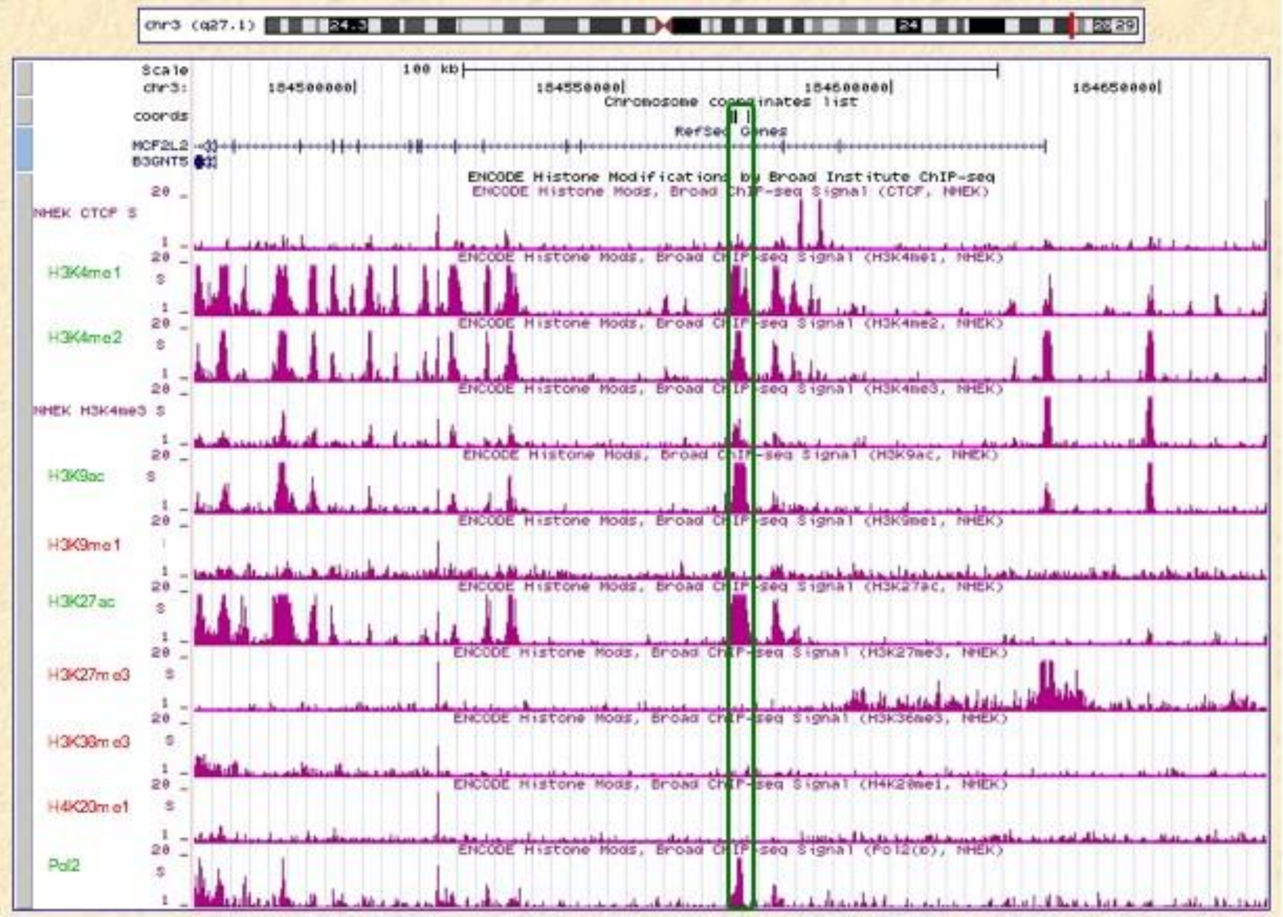

chr3 (q27.1)
Scale
chr3:
100 kb
184500000
184550000
184600000
184650000
Chromosome coordinates list
coords
RefSeq Genes
MCF2L2
B3GNT5
ENCODE Histone Modifications by Broad Institute ChIP-seq
ENCODE Histone Mods, Broad ChIP-seq Signal (CTCF, NHEK)
NHEK CTCF S
H3K4me1
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me1, NHEK)
H3K4me2
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me2, NHEK)
NHEK H3K4me3 S
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me3, NHEK)
H3K9ac
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K9ac, NHEK)
H3K9me1
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K9me1, NHEK)
H3K27ac
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K27ac, NHEK)
H3K27me3
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K27me3, NHEK)
H3K36me3
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K36me3, NHEK)
H4K20me1
ENCODE Histone Mods, Broad ChIP-seq Signal (H4K20me1, NHEK)
Pol2
ENCODE Histone Mods, Broad ChIP-seq Signal (Pol2(b), NHEK)
20
1

Figure 14. MLV's cluster of 5 integrations (marked green) associated with epigenetic markers for primary human keratinocytes available in the UCSC genome browser. Cluster's position overlaps with picks relative to several epigenetic markers and polymerase 2.

Furthermore, possible associations with other genomic features annotated by UCSC were verified, namely CpG islands, phylogenetically conserved sequences and transcription factor binding sites that could be involved in the viruses choice to integrate.

The influence of CpG islands on viral integration was analysed by plotting integrations +/- 10kb around the CpG islands' midpoint. Mostly the same distribution observed before for TSS proximal regions, with a clearly higher incidence of MLV's integrations and a negative correlation by HIV, was verified. However, as shown on previous works, most of the CpG islands chosen by MLV overlap with TSS proximal regions (Cattoglio, Facchini et al. 2007), therefore, this feature shouldn't be considered as independent and instead as being correlated with the previous. That is, the higher integration near CpG islands should be a consequence of selection for TSS proximal regions. This feature showed the same pattern of integration in all studied cell-types (data not shown).

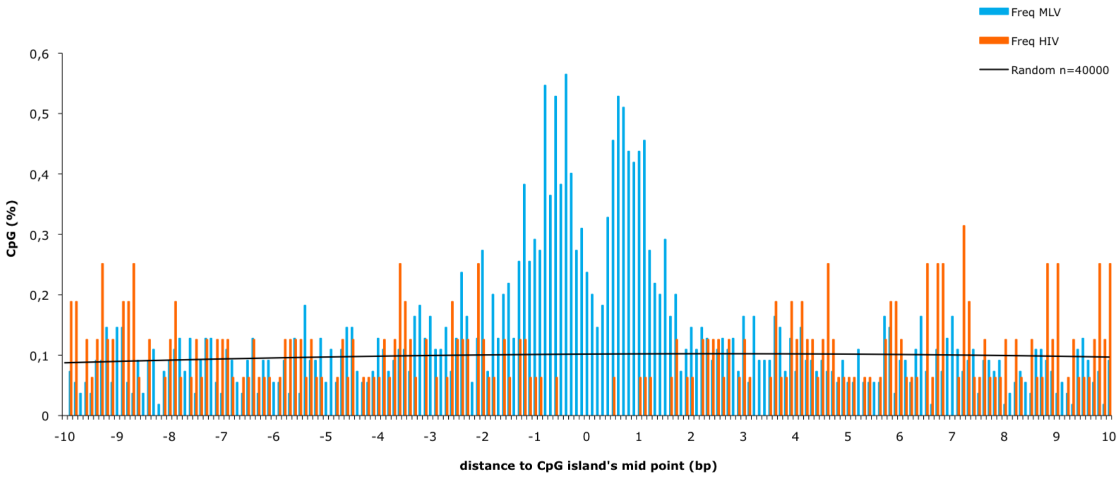


Figure 15. Distribution of retroviral integrations around CpG islands on keratinocytes. Distribution of the distance of MLV (blue bars) and HIV (orange bars) integration sites from the midpoint of CpG islands at 100-bp resolution. The % of the total number of CpG islands (n) is plotted on the Y axis. The black line indicates the distribution of control random sites.

MLV has a superior integration in evolutionary conserved elements

Correlation of viral integration with phylogenetically conserved non-coding sequences (CNCs) among mammals was done using the database available on UCSC and the dataset on keratinocytes. Integration positions in the genome were plotted against the midpoint of conserved sequences along a +/-10kb distance (**Figure 16**). MLV revealed an enrichment of CNCs around integration sites. Looking at the -1kb to 1kb window around the midpoint of the CNC, 10,4% of MLV's integrations locate in this region in comparison to 5,4% of the random distribution (p < 2.2e-16). HIV showed no significant difference in the distribution around this feature in comparison to the control. This particular pattern was similar between all analysed cell-types.

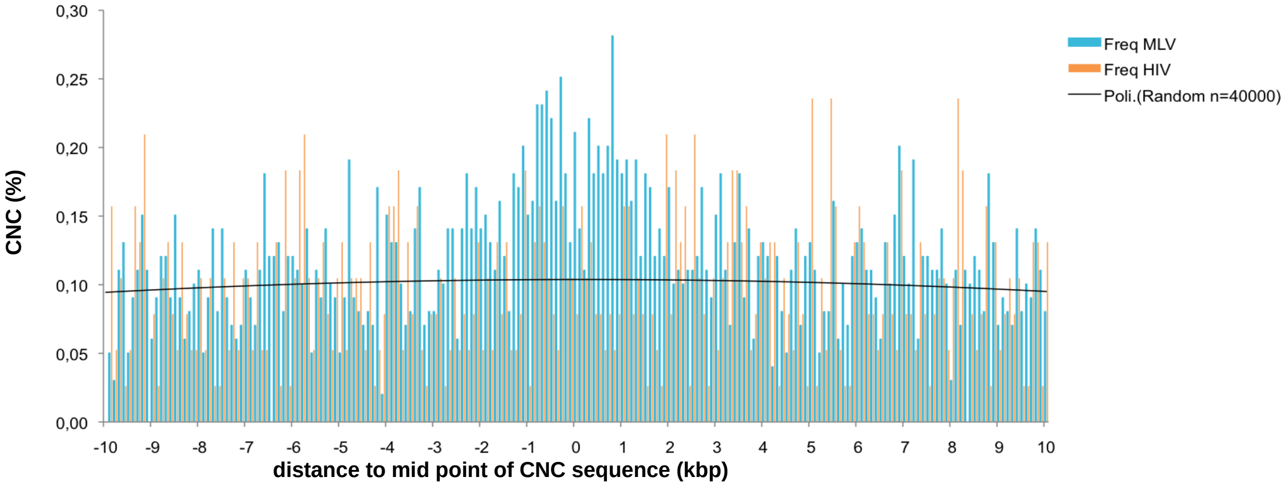


Figure 16. Distribution of retroviral integrations around conserved non-coding sequences on keratinocytes. Distribution of the distance of MLV (blue bars) and HIV (red bars) integration sites from the midpoint of mammalian evolutionarily conserved non-coding sequences (CNCs) at 100bp resolution. The % of the CNCs (n) is plotted on the Y axis. The black line indicates the distribution of control random sites.

MLV has a higher integration in oncogenes and common insertion sites

To evaluate the safety parameters of MLV and HIV derived vectors, the number of oncogenes and common insertion sites (CIS) targeted by both vectors was determined in keratinocytes and compared with the previous values established for lymphocytes and Cd34$^+$HSC. CIS are regions in the genome that have been hit by viral insertions in multiple independent tumors significantly more than expected by chance (de Ridder, Uren et al. 2006), this database is drawn from MLV's integrations. An oncogene is a normal gene that can become an oncogene due to mutations or increased expression. Usually they code for proteins that help to regulate cell growth and differentiation and are often involved in signal transduction and execution of mitogenic signals. Therefore, till its deregulation an oncogene only has the potential to become tumorigenic. Cancer associated genes were annotated using libraries of proto-oncogenes (http://microb230.med.upenn.edu/protocols/cancergenes.html) and genes associated with common insertion sites in murine tumors (http://rtcgd.ncifcrf.gov).

Together oncogenes and CIS represented around 11% of all targeted genes for MLV in all cell-types, a number that is significantly higher in comparison to the 7,8% of the control (p-value = 7.916 e^{-9}). HIV didn't target a significantly relevant number of oncogenes or CIS. Once the CIS dataset was drawn from MLV integrations, it has a different weight in terms of significance when compared with MLV or HIV data. If we consider only the oncogene dataset, which should be less biased for MLV integrations, there's 7% of integrations in oncogenes by MLV versus 6,9% of HIV and 5,6% of the control; this difference is still statistically relevant for MLV and not for HIV (p-value = 0.001763). These results must be looked at with some caution once genes considered as oncogenes, as said above, are also involved in functions like cell growth and development, and being highly expressed genes are also preferential targets for retroviruses. It shouldn't be assumed that gene deregulation always follows viral integration in genes.

Additionally, targeted oncogenes are different among cell-types, for example LMO2 gene responsible for the leukaemia cases in the clinical trial for X-SCID wasn't targeted neither in keratinocytes nor lymphocytes.

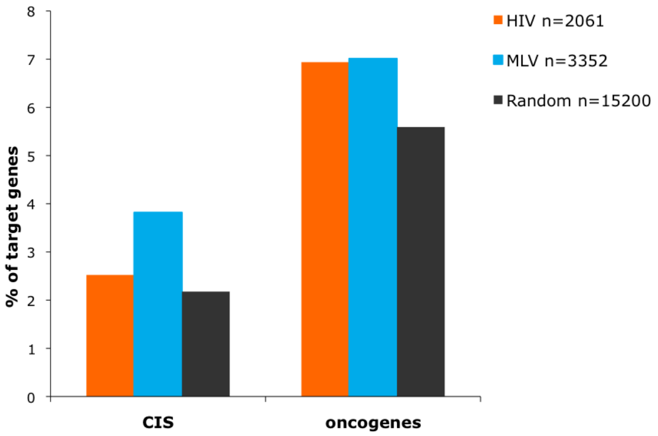


Figure 17. Percentage of oncogenes and common insertion sites hit by MLV and HIV vectors in keratinocytes. All genes hit by HIV and MLV were plotted against available lists of experimentally determined oncogenes and CIS to determine the % of total integrations occurring in these specific genes.

MLV targets regions rich in transcription factor binding sites with a cell-specific pattern

It has been recently shown, in works done with Cd34$^+$ cells, that MLV integration is favoured by proximity to transcription factor binding sites (TFBS) (Felice, Cattoglio et al. 2009). To check if an equal pattern occurs and if the same transcription factor families are targeted in different cell-types, TFBS distribution around integration sites in keratinocytes and lymphocytes was analysed and the final results were compared with the previous ones obtained for Cd34+ cells. The dataset used on conserved TFBS was the one available on UCSC, which lists the evolutionary conserved TFBS between human, murine and rat genomes, a robust definition that increases the significance of motif discovery. Conserved TFBS were considered in a window of 2,000bp around integration sites in all the genome. MLV showed a high overrepresentation of TFBS, almost 70% of MLV sequences in all cell-types contained at least one TFBS versus 44,6% of the random sequences ($p<10^{-15}$). Moreover, 143 out of 258 binding motifs were enriched in comparison to the control over 2 folds on keratinocytes within the analysed window, whereas 155 out of the same 258 were enriched on lymphocytes. These values contrast with the ones previously obtained for Cd34$^+$HSC, which had a similar enrichment in only 43 out of 258 binding motifs. HIV had only 4 motifs out of 258 enriched over 2 folds in keratinocytes and only 1 in Cd34$^+$HSC, therefore, TFBS location doesn't seem to be a preferred feature for integration of this virus. The most targeted motifs are common among all cell-types and consist of binding sites for the transcription factor Sp1, Ets, NFY, Elk, NRF2, Myc, AP2, E2F, MAZR, Egr3 and CREB. In comparison to the previous results on Cd34$^+$HSC, retroviruses seem to integrate with a higher frequency near TFBS in keratinocytes and lymphocytes, in particular V$SP1_Q6 which is enriched ≈3 folds more both in keratinocytes and lymphocytes in comparison to Cd34$^+$HCS. Additionally, a much higher diversity of TFBS families is highly targeted both in keratinocytes and lymphocytes when compared to targeting in Cd34$^+$HSC.

Table 1. Conserved transcription factor binding sites (TFBSs) over-represented around MLV integration sites (±1,000 bp). For each TFBS, the TRANSFAC matrix ID and accession number are specified, together with the putative binding transcription factor(s) (rightmost column) as well as the number of folds each TFBS is enriched on each cell-type in relation to the control. The table shows only TFBSs over-represented at least 2 folds with respect to the frequency around random sites.

acession number	matrix ID	lymphocytes	Keratinocytes	CD34+	binding factors
V$SP1_Q6	M00196	40,4	31,4	12,4	Sp1
V$AP2ALPHA_01	M00469	33,8	3,3	8,7	AP-2alphaA
V$SP1_01	M00008	27,1	15,8	8,3	Sp1
V$CETS1P54_01	M00032	20,8	12,2	6,7	c-Ets-1
V$ELK1_02	M00025	20,2	11,9	6,2	Elk-1
V$NMYC_01	M00055	18,2	11,1	5,5	N-Myc
V$NRF2_01	M00108	18,0	11,2	6,6	NRF-2
V$NFY_01	M00287	16,5	12,3	5,8	NF-Y
V$AP2_Q6	M00189	14,9	10,2	5,2	AP-2alphaA, AP-2gamma
V$MAZR_01	M00491	14,0	9,1	5,0	MAZR
V$E2F_03	M00516	13,7	9,3	5,0	E2F, E2F-1
V$CREB_01	M00039	13,2	7,6	3,4	CREB, deltaCREB
V$ELK1_01	M00007	12,1	6,1	4,7	Elk-1
V$AP2GAMMA_01	M00470	12,0	4,2	4,0	AP-2gamma
V$EGR3_01	M00245	11,4	8,5	3,3	Egr-3
V$PAX5_01	M00143	11,0	6,1	3,5	Pax-5
V$PAX4_01	M00373	11,0	7,2	3,5	Pax-4a
V$USF_Q6	M00187	10,3	7,1	3,5	USF1
V$USF_C	M00217	10,2	6,3	3,8	USF1
V$MYCMAX_03	M00615	10,0	6,5	3,3	c-Myc, Max
V$ARNT_01	M00236	9,9	6,8	3,4	Arnt
V$ATF_01	M00017	9,5	7,3	3,0	ATF
V$NGFIC_01	M00244	9,4	6,8	2,8	Egr-4
V$PAX5_02	M00144	9,3	5,5	2,9	Pax-5
V$EGR1_01	M00243	9,1	6,9	3,0	Egr-1
V$MAX_01	M00119	8,9	4,3	2,8	Max1
V$E2F_02	M00050	8,8	3,8	2,7	E2F, E2F-1, E2F-2, E2F-3a, E2F-4, E2F-5
V$AHRARNT_02	M00237	8,4	5,4	2,3	AhR, Arnt
V$USF_01	M00121	8,3	4,4	2,9	USF1
V$AHR_01	M00139	8,2	5,1	2,6	AhR
V$TAXCREB_01	M00114	8,0	4,9	2,7	CREB, deltaCREB
V$CREBP1_Q2	M00179	7,9	6,4	2,7	ATF-2
V$EGR2_01	M00246	7,9	5,8	2,2	Egr-2
V$PAX4_03	M00378	7,8	6,9	2,6	Pax-4a
V$MZF1_02	M00084	7,7	5,6	2,9	MZF-1
V$ARNT_02	M00539	7,3	4,8	2,5	Arnt
V$CREBP1CJUN_01	M00041	7,3	4,1	2,2	ATF-2, c-Jun
V$SPZ1_01	M00446	7,1	5,2	2,8	Spz1
V$STAT3_02	M00497	7,0	6,3	2,3	STAT3
V$GATA2_01	M00076	6,8	5,2		GATA-2
V$NFY_Q6	M00185	6,7	4,4	2,4	CP1A, CP1C, NF-Y, NF-YA
V$STAT1_01	M00224	6,7	4,0	2,3	STAT1alpha, STAT1beta
V$CMYB_01	M00004	6,6	4,1	2,3	c-Myb
V$MYCMAX_01	M00118	6,6	3,4	2,4	c-Myc, Max1
V$GATA1_01	M00075	6,6	3,9		GATA-1
V$RREB1_01	M00257	6,6	3,4		RREB-1
V$YY1_02	M00069	6,4	4,1	2,0	YY1
V$E2F_01	M00024	6,4	3,9		E2F
V$CREB_02	M00113	6,3	5,3	2,2	CREB, deltaCREB
V$NFKAPPAB_01	M00054	6,2	5,6		NF-kappaB, NF-kappaB1, RelA
V$AHRARNT_01	M00235	6,2	5,8	2,2	AhR, Arnt
V$TAXCREB_02	M00115	5,8	5,1		CREB, deltaCREB
V$SREBP1_01	M00220	5,8	5,1	2,3	SREBP-1a, SREBP-1b, SREBP-1c
V$NFKAPPAB50_01	M00051	5,8	4,5		NF-kappaB1
V$CREL_01	M00053	5,8	4,5		c-Rel
V$MZF1_01	M00083	5,6	2,9		MZF-1
V$NFKAPPAB65_01	M00052	5,4	3,7		RelA
V$P300_01	M00033	5,4	4,5	2,1	p300
V$P53_01	M00034	5,1	5,2		p53, p53
V$HEN1_02	M00058	5,0	3,6		HEN1
V$CREB_Q4	M00178	4,9	4,1		CREB
V$NFKB_C	M00208	4,8	4,1		NF-kappaB, NF-kappaB1, NF-kappaB2
V$AP4_01	M00005	4,7	3,4		AP-4

V$ZIC2_01	M00449	4,6	3,9	ZIC2
V$ATF6_01	M00483	4,6	3,7	ATF6
V$MIF1_01	M00279	4,5	3,6	MIF-1
V$CP2_01	M00072	4,5	3,7	CP2
V$CREB_Q2	M00177	4,5	4,0	CREB
V$BACH2_01	M00490	4,4	7,0	Bach2
V$XBP1_01	M00251	4,4	2,9	XBP-1
V$NRSF_01	M00256	4,3	3,3	NRSF form 1, NRSF form 2
V$OLF1_01	M00261	4,3	3,0	Olf-1
V$HEN1_01	M00068	4,2	4,2	HEN1
V$STAT3_01	M00225	4,1	2,7	STAT3
V$AP1_01	M00517	4,1	6,1	AP-1, c-Fos, c-Jun, Fra-1, JunB, JunD, FosB
V$MYCMAX_02	M00123	4,1	3,1	c-Myc, Max1
V$ZIC3_01	M00450	4,0	3,1	Zic3
V$ZIC1_01	M00448	3,9	3,0	Zic1
V$PAX3_01	M00360	3,9	2,7	Pax-3
V$E47_01	M00002	3,9	2,6	E47
V$PAX2_01	M00098	3,8	3,8	Pax-2
V$ROAZ_01	M00467	3,7	2,5	Roaz
V$MYOGNF1_01	M00056	3,7	2,8	NF-1
V$HMX1_01	M00433	3,6	3,3	Nkx5-1
V$ZID_01	M00085	3,5	2,7	ZID
V$NFKB_Q6	M00194	3,4	3,4	NF-kappaB, NF-kappaB1
V$HNF4_01	M00134	3,4	3,3	HNF-4alpha2
V$STAT_01	M00223	3,4	3,0	STAT1alpha, STAT1beta, STAT2,STAT3,STAT4, STAT6
V$ISRE_01	M00258	3,4		ISGF-3
V$NFE2_01	M00037	3,4	5,1	NF-E2, NF-E2 p45
V$HOX13_01	M00023	3,4	2,5	HOXA5
V$SREBP1_02	M00221	3,4	3,5	SREBP-1a, SREBP-1b, SREBP-1c
V$LMO2COM_01	M00277	3,4	3,4	Lmo2
V$AP1_Q6	M00174	3,3	3,1	AP-1
V$RFX1_02	M00281	3,2		RFX1
V$AREB6_03	M00414	3,2	3,8	AREB6
V$BACH1_01	M00495	3,2	4,9	Bach1
V$AP1FJ_Q2	M00172	3,2	3,3	AP-1, c-Fos, c-Jun
V$HNF4_01_B	M00411	3,2	3,0	HNF-4alpha1
V$STAT1_03	M00496	3,2	3,3	STAT1
V$NCX_01	M00484	3,1	2,3	NCX
V$STAT5A_02	M00460	3,1		STAT5A
V$PAX2_02	M00486	3,1	2,2	Pax-2
V$YY1_01	M00059	3,0	2,0	YY1
V$PAX4_04	M00380	3,0	3,0	Pax-4a
V$GATA3_01	M00077	3,0	2,4	GATA-3
V$SRF_Q6	M00186	2,9	2,2	SRF
V$AP1_Q2	M00173	2,8	3,7	AP-1
V$MYOD_01	M00001	2,8	2,8	MyoD
V$ER_Q6	M00191	2,8	2,5	ER-alpha
V$NFY_C	M00209	2,8	2,8	CP1A, NF-Y, NF-YA
V$RFX1_01	M00280	2,8	2,7	RFX1
V$AREB6_04	M00415	2,7		AREB6
V$USF_02	M00122	2,7	2,5	USF1
V$AP1_Q4	M00188	2,7	3,2	AP-1
V$IRF2_01	M00063	2,7	2,0	IRF-2
V$COMP1_01	M00057	2,7		COMP1
V$IRF7_01	M00453	2,7		IRF-7A
V$AML1_01	M00271	2,7		AML1a
V$MYB_Q6	M00183	2,7	2,1	c-Myb
V$MYOD_Q6	M00184	2,6		MyoD
V$LUN1_01	M00480	2,6	2,1	LUN-1
V$AP1_C	M00199	2,6	5,2	AP-1
V$HTF_01	M00538	2,6	2,0	HTF
V$STAT5A_01	M00457	2,6	2,1	STAT5A
V$SEF1_C	M00214	2,5	2,3	SEF-1 (1)
V$ARP1_01	M00155	2,5	2,6	ARP-1
V$SRF_C	M00215	2,5		SRF
V$IRF1_01	M00062	2,5		IRF-1
V$CEBP_C	M00201	2,5		C/EBPalpha
V$AP4_Q6	M00176	2,5	2,6	AP-4
V$COUP_01	M00158	2,5		COUP-TF1, HNF-4alpha2
V$MEIS1_01	M00419	2,4	2,1	Meis-1
V$AREB6_01	M00412	2,4	3,0	AREB6
V$CHOP_01	M00249	2,4		C/EBPalpha, CHOP-10
V$PPARG_03	M00528	2,4	2,0	PPAR-gamma1, PPAR-gamma2

V$AREB6_02	M00413	2,4	2,2	AREB6
V$NF1_Q6	M00193	2,3	3,1	NF-1
V$TCF11MAFG_01	M00284	2,3	3,0	LCR-F1
V$IK1_01	M00086	2,3	2,4	Ik-1
V$HSF2_01	M00147	2,3		HSF2
V$STAT5B_01	M00459	2,3		STAT5B
V$AP4_Q5	M00175	2,3	2,4	AP-4
V$TCF11_01	M00285	2,3	2,6	LCR-F1
V$PPARA_01	M00242	2,2	2,4	PPAR-alpha, PPAR-alpha
V$PPARG_01	M00512	2,2		PPAR-gamma1, PPAR-gamma2
V$HSF1_01	M00146	2,2		HSF1 (long)
V$GATA1_03	M00127	2,2		GATA-1
V$E47_02	M00071	2,1	2,3	E47
V$LYF1_01	M00141	2,1		LyF-1
V$IK2_01	M00087	2,1	2,1	Ik-2
V$PPARG_02	M00515	2,1	2,2	PPAR-gamma1, PPAR-gamma2
V$GRE_C	M00205	2,1		GR-alpha
V$CEBPB_02	M00117	2,1	2,3	C/EBPbeta
V$IK3_01	M00088	2,1	2,2	Ik-3
V$HOXA3_01	M00395		2,41	HOXA3
V$NFAT_Q6	M00302		2,36	NF-AT1, NF-AT1, NF-AT2, NF-AT3, NF-AT4
V$GR_Q6	M00192		2,15	GR-alpha, GR-beta
V$SOX9_B1	M00410		2,13	Sox9
V$RP58_01	M00532		2,06	RP58
V$SRF_01	M00152		2,04	SRF
V$PAX6_01	M00097		2,02	Pax-6

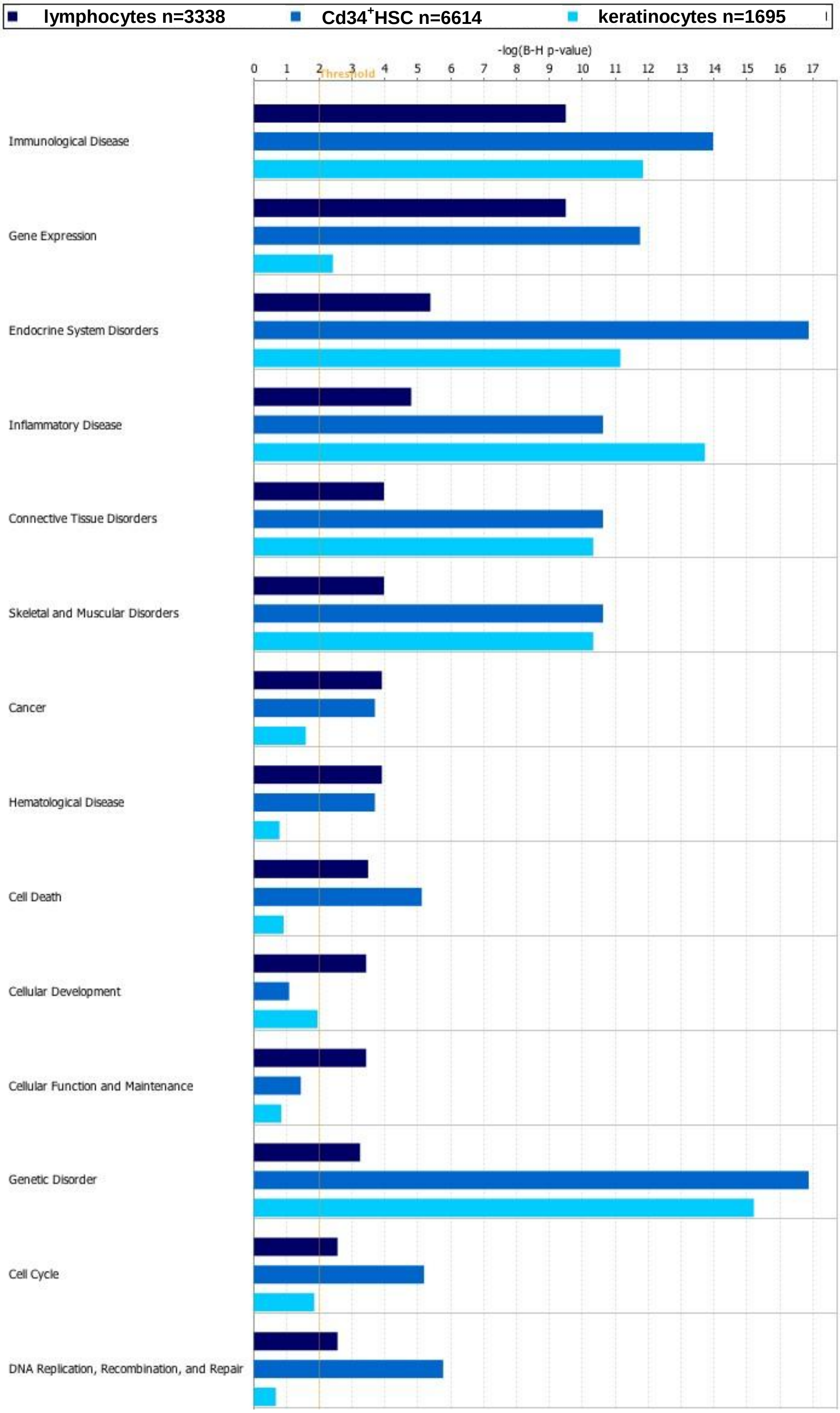
Ingenuity Pathways Analysis suggests functions for retroviruses' target genes

Using Ingenuity Pathways Analysis software, statistically relevant functions of genes targeted by retrovirus (threshold value of 0.01, using the Bonferroni correction for multiple testing) in keratinocytes, lymphocytes and Cd34$^+$HSC was investigated (**Figure 18**). A different definition of target gene was used, taking in consideration both of the viruses' preferences. For HIV, genes with intronic and exonic integrations were analysed while, for MLV, genes with integrations in the -2,5 till the TSS region were also included.

This analysis showed that, in all cell-types, MLV significantly targeted genes involved broadly in gene expression, cell death, cellular development, function and maintenance, cell cycle and DNA replication, recombination and repair. Moreover, genes targeted by MLV are involved in immunological and inflammatory diseases, cancer, genetic disorders, haematological disease and endocrine system disorders as well as in cell-specific disorders related to skeletal, muscular and connective tissues. One main observable difference between cell-types is that cancer and haematological disease are not significant categories among integrations in keratinocytes. From this analysis we could also see that HIV targets mostly genes involved in general cell functions like gene expression, cell cycle, post-translational modification and molecular transport. HIV's target genes are also involved in a number of disorders like endocrine system and genetic disorders, as well as, metabolic, cardiovascular, gastrointestinal and immunological diseases.

Associations between genes targeted by MLV and HIV in all cell-types were then analysed, starting from 1692 and 1213 genes for MLV and HIV, respectively, in keratinocytes; 3338 and 3098 for MLV and HIV in lymphocytes; 6607 and 3097 for MLV and HIV in Cd34$^+$HSC. Several gene networks with direct and indirect connections between the analysed genes were uncovered. Networks with the highest score, comprising from 28 to 35 focus molecules, were analysed since these would have no doubt of statistical relevancy.

Functions linked to each of the networks were very diverse, therefore, it wasn’t possible to draw any conclusion from the obtained results. Possibly, there’s no bias towards any functional network or the size of the datasets is insufficient to put in evidence any highly targeted network.


■ lymphocytes n=3338
■ Cd34^{+}HSC n=6614
■ keratinocytes n=1695
-log(B-H p-value)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Threshold
Immunological Disease
Gene Expression
Endocrine System Disorders
Inflammatory Disease
Connective Tissue Disorders
Skeletal and Muscular Disorders
Cancer
Hematological Disease
Cell Death
Cellular Development
Cellular Function and Maintenance
Genetic Disorder
Cell Cycle
DNA Replication, Recombination, and Repair

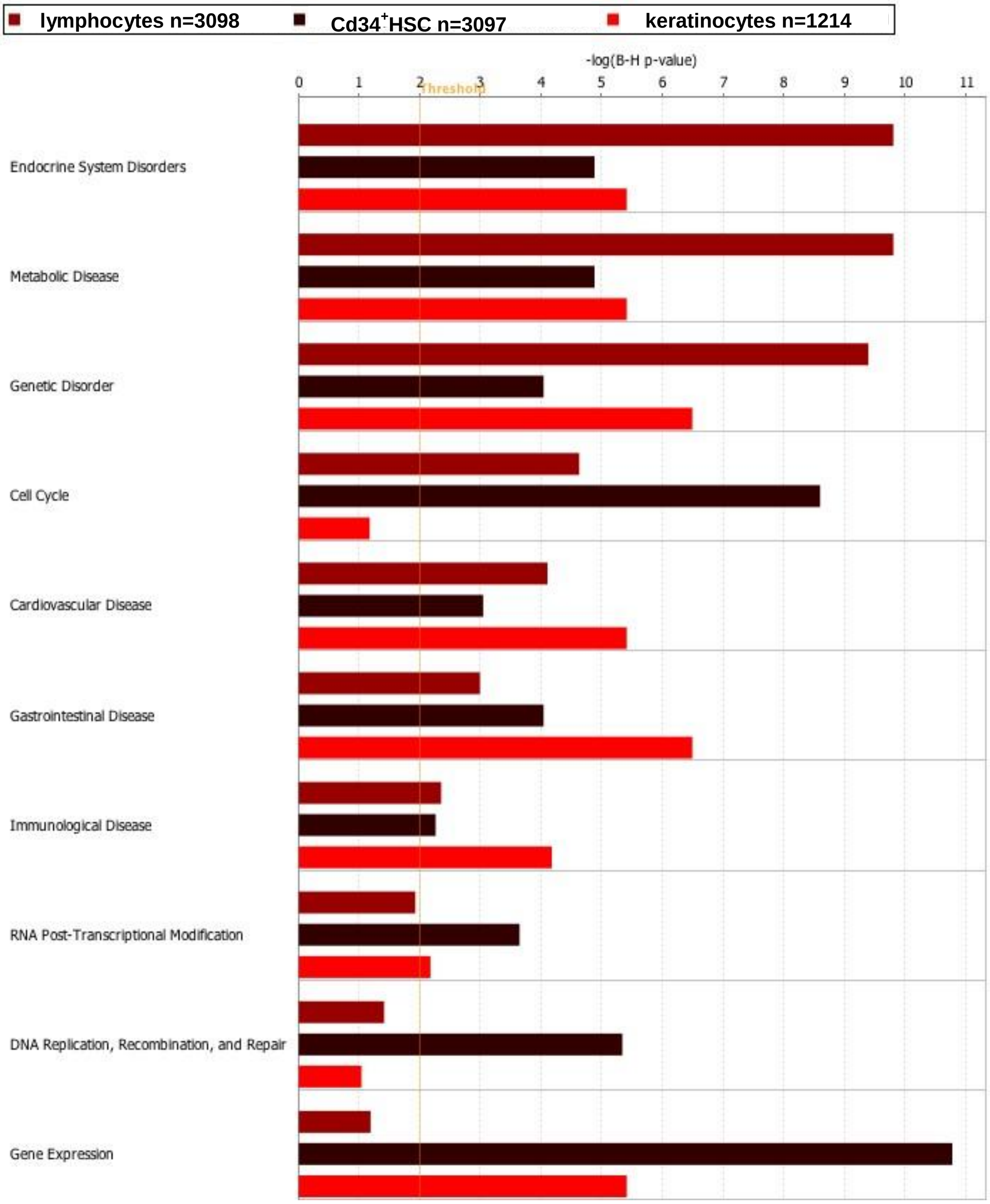


Figure 18. Ingenuity® functional analysis of genes targeted by MLV and HIV in lymphocytes, Cd34^{+}HSC and keratinocytes. Shown are those function/disease categories significantly over-represented among MLV (blue bars) or HIV (red bars) target genes with respect to the Ingenuity® Pathways Knowledge Base, used as background population for the analysis. P values were corrected for multiple testing by the Bonferroni method (B-H), Statistical significance was set at a p value of 0.01 (2 in log scale, yellow vertical line). n represents the number of target genes analysed for each cell-type. Cd34+HSC dataset was sampled for correction of disparities due to large differences in target gene number in HIV's analysis.

Retroviruses choose early replicating regions of the genome to integrate

Replication timing (RT) along human chromosomes has been shown to correlate with gene expression and/or chromatin modifications (Hansen, Thomas et al. ; Woodfine, Fiegler et al. 2004), being that early replication is associated with expression and gene dense regions while late replication is associated with repression. Therefore, we hypothesized that retrovirus would integrate mainly in early replicating zones of the genome. To test this hypothesis we correlated data on RT in lymphocytes (Woodfine, Fiegler et al. 2004) with a dataset of retroviral integration in the same cell-type (Wang, Ciuffi et al. 2007). The dataset available online quantifies RT using human genomic microarrays covering the entire genome with clones (mean insert size 150kb) spaced at 1Mb intervals. Since replication foci seem to have an average size of 60-100 kb and to be synchronously activated in clusters of more than 10 replicons (Tomilin, Solovjeva et al. 1995), we considered the viral integrations in a window of 1Mb from the annotated position for RT. Therefore, we correlated the positions of viral integration with the closest annotated position for RT in all the genome and attributed the correspondent RT value, expressed as the ratio S:G1 phase DNA. Genomic regions with values from 1 to 1,4 were considered late replicating, whereas regions with values from 1,8 to 2,2 were considered early replicating.

Both viruses showed a significant higher integration in early replicating regions and a much lower integration in late replicating zones in comparison to the random (p-value< 2,2 e-16). The same analysis was also done on MLV integrations localized inside genes but no significant difference was shown.

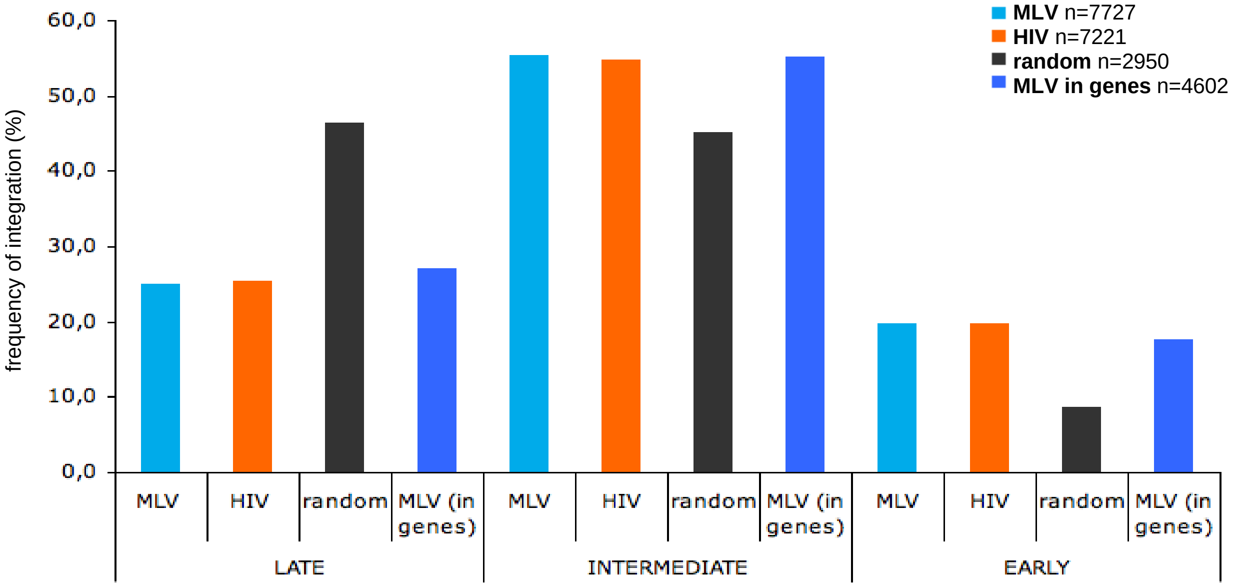


Figure 19. Distribution of viral integrations among early, intermediate and late replication zones in the genome of human lymphocytes. Graph shows the frequency of integration (%), considering all MLV's and HIV's integrations and MLV's intragenic integrations, correlated with the closest annotated replication timing in all human chromosomes on lymphocytes.

Discussion

Retroviruses have clear biases in the way they integrate the human genome, one can quite easily distinguish between retroviral families by looking at their integration pattern. However, these differences are not fully understood and only recently, with the possibility of large scale sequencing of integration sites, we could begin to have a better notion of what might drive the mechanism of viral integration.

The main aim of this work was to create and analyse an extended library of integration sequences of MLV and HIV after infection of primary human keratinocytes. By creating a new large-scale survey, based on clinically relevant cells that are often involved in retroviral lifecycle, we intended to widen our boundaries in terms of analysis feedback. Large-scale surveys of viral integration in different biological sets make possible to perform in a statistically rigorous manner more diverse types of analysis with a wider fidelity of results. Other studies on different genomic features of the human genome are also instrumental in widening our possibilities in terms of analysis. Furthermore, we intended to understand cell specific differences by comparing our dataset with previous ones obtained for lymphocytes and Cd34$^+$HSC.

I first tried to understand retroviruses' integration preferences in relation to the position of annotated genes in the UCSC database, starting from the new dataset on keratinocytes and then comparing with the ones on lymphocytes and Cd34$^+$HSC. It's noticeable that both MLV and HIV share a preference to integrate inside genes, although, HIV's integration is much more biased towards intragenic locations than MLV's. Furthermore, when comparing the pattern among the three different cell-types, it's evident that integration inside genes is much more exacerbated in lymphocytes for both viruses.

The next step was to evaluate distribution around the TSS region, once this feature had been reported to strongly correlate with retroviral integration (Wu, Li et al. 2003). MLV showed a high frequency of integration in these regions in comparison to the control and to HIV. Additionally, it's clear that MLV had an increasingly higher percentage of integration near TSS regions depending on the cell-type, being that the lowest observed frequency was seen in keratinocytes, followed by Cd34$^+$HSC and the highest in lymphocytes. When looking at higher resolutions (50kb and 5kb), plotting integration frequencies against distance to the TSS, it's possible to observe once again HIV's preference to integrate in the body of genes but not in the TSS region. In fact, HIV seems to be repelled by TSS regions with an equal pattern in all cell-types. MLV shows a mirror like distribution around TSS in all cell-types with a very high frequency of integration in the regions close to the TSS (-2,5 to 2,5kb) and a drop of integration in the 0 to 300bp region downstream the TSS. Notably, once again in lymphocytes, there's a higher frequency of integration by MLV in the -1,7 to 1,7kb region around TSS in comparison to the other two cell-types.

HIV's integration was proposed to be correlated with gene expression in previous works, even with relatively small datasets (Schroder, Shinn et al. 2002). To check if integrations correlated with gene activity in both viruses on all the cell-types, Affymetrix, a microarray assay, was performed on keratinocytes and confronted with the results obtained from the same experiment done on Cd34$^+$HSC (Cattoglio *et al*., Blood in press) and just with MLV on lymphocytes (Recchia *et. al.*, unpublished). A high frequency of integration was confirmed in expressed genes by both viruses, in particular HIV. Although, in keratinocytes there was a lower frequency of expressed genes targeted by HIV in intragenic hits in comparison to Cd34$^+$HSC.

These first analysis showed that, overall, the pattern of integration of both viruses is maintained among different cell-types, although, relevant differences were seen in relation to the frequency of events around features linked to viral integration. Summed up, these differences suggest there's a more efficient viral exploitation of primary target cells, in this case lymphocytes, reinforcing the idea that the cell's machinery is largely involved in directing PICs to the genome.

Both viruses seem to target the same amount of genes, with a high preference for expressed ones. However, even though there's a significant preference for expressed genes, more than ¼ of non-expressed genes are targeted, therefore, gene expression cannot explain viral preference by itself. Major differences are seen when looking at viral integration clusters: HIV targets a higher number of genes, approximately twice as much in comparison to MLV which can be explained by the significantly larger size of HIV's clusters. Furthermore, MLV's clusters are more compact and HIV's are much larger reaching a higher number of integrations. By analysing the distance between consecutive integrations and cluster's size we realize that HIV seems to integrate in a wider portion of the genome in comparison to MLV, suggesting that the first is attracted by a more generally available genetic feature. This assumption, in addition to the results of the gene function analysis, are in agreement with the available evidence that LV PICs are tethered to the human genome by LEDGF and possibly other chromatin remodelling or DNA-repair complexes widely distributed in the chromatin.

There are clearly some biases on integration distribution in whole chromosomes by both viruses, while a part might be explained by gene density, this factor doesn't seem to be the only one playing a part in viral distribution. Possibly, another important factor to look at in future experiments might be nuclear positioning of the chromosomes, since this was proposed to be a cell-specific feature, maintained among the same cell-type but differing in cells of distinct tissues. Spatial distribution of genes within the nucleus contributes to transcriptional control, allowing optimal gene expression as well as constitutive or regulated gene repression. Previous works reported that PICs preferentially distribute in decondensed areas of the chromatin with a striking positioning in the nuclear periphery, while heterochromatin regions are largely disfavored. These observations provide an indication of how the nuclear architecture may initially orient the selection of retroviral integration sites (Albanese, Arosio et al. 2008).

Following this rational, the epigenetic background surrounding integrations is an important factor to look at and, as shown in previous works (Wang, Ciuffi et al. 2007), it might give important leads to further understand the mechanism of retroviral integration. Recent data available in the UCSC database on epigenetic modifications in primary human keratinocytes, made possible to perform an analysis of the epigenetic markers in the vicinity of retroviral integrations in this cell-type. The results obtained from the epigenetic markers analysis, even though are very diverse between both viruses, are in agreement with one of the first hypothesis which proposed that condensed, inactive chromatin is unfavourable to viral integration. Furthermore, this analysis demonstrated that MLV integrates with a high frequency in regions marked with H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, histone modifications associated with promoter and enhancer locations (Heintzman, Stuart et al. 2007) and also positively associates with polymerase 2. This characteristic may be instrumental in future works, as a way to uncover new promoter and enhancer regions in different cell-types using the virus as a reporter, for example, to unveil the transcriptional program of human stem cells. As to HIV, it seems to disfavour these regions but instead strongly associates with the H3K36me3 marker, associated with the body of actively transcribed genes (Sims and Reinberg 2009), therefore, showing a more discreet integration pattern. Histone modifications associated with inactive genes and heterochromatic regions like H3K9me1 and H3K27me3, and H4K20me1 (Schotta, Sengupta et al. 2008) were negatively associated with both viruses confirming the viral preference for active chromatin regions.

Furthermore we looked at the possible association between other genomic features like CpG islands and CNCs with viral integration. While the positive association observed between MLV's integrations and CpG islands is questionable in terms of relevance (once it overlaps promoter regions), it is interesting to note that the positive correlation with phylogenetically conserved sequences is maintained among different cell-types with the same frequency. Most CNCs are segments of around 100–300 bp that are widely distributed across the human genome. They are not preferentially located near genes as previously hypothesized (Dermitzakis, Kirkness et al. 2004). In some cases, clusters of CNCs are found in gene deserts and a subset of these CNCs have been shown to play functional roles as enhancers (Nobrega, Ovcharenko et al. 2003; de la Calle-Mustienes, Feijoo et al. 2005). Even though its exact function is still unknown, it has been shown repeatedly that screening for CNCs is an effective method for identifying cis-regulatory modules of gene expression, and that these are likely to form part of the genomic circuitry that uniquely defines mammal development (Kim and Pritchard 2007). Therefore, in terms of retroviral strategy, integrating preferably around these sequences would increase the virus' probability of survival once CNCs seem to be fundamental regions that are unlikely subject to disruption and in which mutation occurs only in occasional short bursts of evolution.

TFBS conserved among mammals in the vicinity of viral integrations were analysed in all cell-types once it was recently reported to be strongly correlated with MLV's integration (Felice, Cattoglio *et al*. 2009). This analysis, besides confirming a high enrichment of this feature near

MLV's integration sites, revealed that a different diversity of TFBS is highly targeted by different cell-types, additionally some TFBS are commonly enriched in all cell-types and some only in particular cells. These results lead to the hypothesis that the differences observed in the different cell-types might be explained by TFBS activity on each cell-type. Therefore, lymphocytes and keratinocytes should have a higher number of active TFBS in comparison to Cd34^{+}HSC that, as progenitor cells, still might have less functional TFBS. Hence, besides preference for some specific families of TFBS (some common and some cell-specific) there seems to be a high number of integrations near active TFBS.

Ingenuity Pathways Analysis software was used to verify the function of our generated target gene lists. We verified that, genes targeted by MLV are involved in general cell-functions as well as in more cell-specific functions, while HIV targets significantly genes mainly involved in regulation of gene expression. The magnitude of MLV's tissue-specific biases is quite low probably because most of the cellular transcriptional program is common among all cell-types. The outcome of this analysis is in agreement with the hypothesis that MLV targets genomic elements involved in transcriptional regulation while HIV is correlated with more widely distributed genes, related to functions like regulation of gene expression, chromatin remodelling and DNA repair. Nevertheless this software showed some limitations once results varied accordingly to the number of genes analysed, therefore, its results should be seen as suggestions and not absolute results. Since the datasets available were very different in terms of number of integrations, it wasn't possible to make an accurate comparison between common genes hit in all cell-types and genes hit in a particular cell-type.

Correlation of the replication timing of a chromosomal region with viral integration might also be an interesting feature to investigate, once the chromatin state found by the virus at the time of infection could play an important role in retroviral integration. Unfortunately there are still few studies with large enough data to correlate with viral integrations. Data on the whole genome already available was done on lymphocytes (Woodfine, Fiegler et al. 2004) for its obvious clinical relevance, therefore, we used the dataset on retroviral integrations in this cell-type to correlate with replication timing on all the human genome. Once the annotated positions along the chromosomes for replication timing are limited, a correlation between proviral positions and this data is still very preliminary and can only be seen as a suggestion of viral preferences. Results of this association proposed a high correlation between early replicating regions and retroviral integration in both viruses analysed. This suggestion is in agreement with several findings that associate early replication regions with features also previously correlated to retroviral integration: Hansen et. al reported that DNA replication typically initiates within foci of accessible chromatin (Hansen, Thomas et al.); Karnani et. al, observed that a delay of replication time was accompanied by a decrease in level of gene expression and appearance of repressive chromatin marks (Karnani, Taylor et al. 2007); Woodfine and his team, reported significant positive correlations between the mean replication timing and mean gene density, additionally they associated regions of high-level expression with early replication but, rather than the association with the extent of expression they strongly correlate replication timing with the probability of gene expression (Woodfine, Fiegler et al. 2004). While it seems unlikely that this characteristic is responsible for integration site selection by itself, whether it plays a part together with the other features observed or is just a consequence of the preference for an overlapping feature, is still to be determined in future experiments.

Although there are some similarities between HIV's and MLV's pattern of integration, for instance the preference to integrate inside active genes and probably in early replicating regions of the genome, notably, different viral families have distinct strategies when integrating the genome. Gamma-retroviruses and lentiviruses analysed in this work have mainly different, some times opposite, preferences of integration in the human genome revealing completely distinct evolutionary paths.

From its integration pattern it seems that gamma-retroviruses might have evolved a mechanism that takes advantage of the cellular transcriptional machinery to promote its own expression, coupling proviral integration with highly active regions of the genome rich in TFBS, TSS, promoters, enhancers and with a epigenetic conformation characteristic of an active chromatin. This suggests an indirect tethering model in which ubiquitous TFs bound within RV PICs interact with general components of the enhancer-binding complexes, such as co-regulators, chromatin remodelling or mediator complexes, rather than with specific TFs or TF families. Tethering of PICs to transcription factories, where promoters and regulatory regions are relocated by cell-specific mechanisms, may in turn be the cause of the RV-specific, high frequency of integration clusters and preferred targeting of genes associated to cell-specific regulatory networks described above. Recent works provide support for this hypothesis, proposing a direct interaction between integrase and chromatin remodelling, DNA repair and transcription factors (Studamire and Goff 2008). From a evolutionary point of view, this cooperation may be interpreted as the development of the mechanisms by which retrotransposons target their integration to specific genomic regions, tethered by host cell proteins. A mechanism coupling target site selection to gene regulation may have evolved to maximize the probability for gammaretroviruses to be transcribed in the target cell genome, and possibly to induce expansion of infected cells by insertional deregulation of cell-specific growth regulators.

On the other hand, lentiviruses developed a totally different strategy interacting with much less interference with the host's chromatin and cellular machinery. This strategy prolongs the lifespan of infected cells and consequently of the host, increasing the probability of viral dissemination to new hosts.

There are still a lot of questions that remain unanswered mainly regarding HIV's integration mechanism. Through the bioinformatic analysis done so far, features attracting HIV are not very obvious and, although some preferences and disfavoured regions are already known, further works in different directions are needed to fully understand its choices. Since HIV can infect non-dividing cells, while MLV only infects cells if they undergo mitosis, viral integration complexes might access the genome at different stages in the cell cycle. It's possible that chromatin's conformation found by viruses at the time of infection, might also play an important part in the virus' integration choice. Additional studies following virus as they integrate cells in different stages of the cell-cycle might further elucidate these questions.

This thesis also shows the importance of the cell context in determining the frequency of integration into certain genomic regions. Once gene therapy for inherited skin diseases was the main goal of the laboratory, this work was also instrumental in understanding the safety parameters of using viral vectors in this type of cells. Regarding these issues, it is possible that the probability of targeting dominantly acting proto-oncogenes is different in distinct cell-types. For instance, LMO2 locus, which gave rise to the leukaemia cases in the X-SCID clinical trials after viral insertional activation, wasn't hit by neither of the virus in epithelial cells and T-cells where it is not expressed. Moreover, genes functionally related by Ingenuity software to cancer and haematological disease were not relevant MLV targets in keratinocytes, although, there was a significant preference for oncogenes when comparing integrations with experimentally determined proto-oncogenes. These data suggest that HIV-based vectors might be safer in gene therapy clinical trials than MLV-based vectors.

References

Albanese, A., D. Arosio, et al. (2008). "HIV-1 pre-integration complexes selectively target decondensed chromatin in the nuclear periphery." PLoS One 3(6): e2413.

Berry, C., S. Hannenhalli, et al. (2006). "Selection of target sites for mobile DNA integration in the human genome." PLoS Comput Biol 2(11): e157.

Bushman, F., M. Lewinski, et al. (2005). "Genome-wide analysis of retroviral DNA integration." Nat Rev Microbiol 3(11): 848-58.

Carteau, S., C. Hoffmann, et al. (1998). "Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target." J Virol 72(5): 4005-14.

Cattoglio, C., G. Facchini, et al. (2007). "Hot spots of retroviral integration in human CD34+ hematopoietic cells." Blood 110(6): 1770-8.

Ciuffi, A., M. Llano, et al. (2005). "A role for LEDGF/p75 in targeting HIV DNA integration." Nat Med 11(12): 1287-9.

Ciuffi, A., R. S. Mitchell, et al. (2006). "Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts." Mol Ther 13(2): 366-73.

Coffin, JM. (1997) Retroviruses: Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Res 33(20): e175.

de la Calle-Mustienes, E., C. G. Feijoo, et al. (2005). "A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts." Genome Res 15(8): 1061-72.

de Ridder, J., A. Uren, et al. (2006). "Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens." PLoS Comput Biol 2(12): e166.

De Rijck, J., K. Bartholomeeusen, et al. "High-resolution profiling of the LEDGF/p75 chromatin interaction in the ENCODE region." Nucleic Acids Res.

Dermitzakis, E. T., E. Kirkness, et al. (2004). "Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment." Genome Res 14(5): 852-9.

Derse, D., B. Crise, et al. (2007). "Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses." J Virol 81(12): 6731-41.

Engelman, A. and P. Cherepanov (2008). "The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication." PLoS Pathog 4(3): e1000046.

Farnet, C. M. and F. D. Bushman (1997). "HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro." Cell 88(4): 483-92.

Felice, B., C. Cattoglio, et al. (2009). "Transcription factor binding sites are genetic determinants of retroviral integration in the human genome." PLoS One 4(2): e4571.

Hacein-Bey-Abina, S., F. Le Deist, et al. (2002). "Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy." N Engl J Med 346(16): 1185-93.

Hacein-Bey-Abina, S., C. von Kalle, et al. (2003). "A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency." N Engl J Med 348(3): 255-6.

Hansen, R. S., S. Thomas, et al. "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing." Proc Natl Acad Sci U S A 107(1): 139-44.

Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet 39(3): 311-8.

Kalpana, G. V., S. Marmon, et al. (1994). "Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5." Science 266(5193): 2002-6.

Kameoka, M., S. Nukuzuma, et al. (2005). "Poly(ADP-ribose)polymerase-1 is required for integration of the human immunodeficiency virus type 1 genome near centromeric alphoid DNA in human and murine cells." Biochem Biophys Res Commun 334(2): 412-7.

Kang, Y., C. J. Moressi, et al. (2006). "Integration site choice of a feline immunodeficiency virus vector." J Virol 80(17): 8820-3.

Karnani, N., C. Taylor, et al. (2007). "Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas." Genome Res 17(6): 865-76.

Kim, S. Y. and J. K. Pritchard (2007). "Adaptive evolution of conserved noncoding elements in mammals." PLoS Genet 3(9): 1572-86.

Lin, C. W. and A. Engelman (2003). "The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes." J Virol 77(8): 5030-6.

Llano, M., M. Vanegas, et al. (2006). "Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75." J Mol Biol 360(4): 760-73.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature 437(7057): 376-80.

McCormack, M. P. and T. H. Rabbitts (2004). "Activation of the T-cell oncogene LMO2 after gene therapy for X-linked severe combined immunodeficiency." N Engl J Med 350(9): 913-22.

Meredith LW, Sivakumaran H, Major L, Suhrbier A, Harrich D (2009) Potent Inhibition of HIV-1 Replication by a Tat Mutant. PLoS ONE 4(11): e7769. doi:10.1371/journal.pone.0007769

Mulder, L. C., L. A. Chakrabarti, et al. (2002). "Interaction of HIV-1 integrase with DNA repair protein hRad18." J Biol Chem 277(30): 27489-93.

Nobrega, M. A., I. Ovcharenko, et al. (2003). "Scanning human gene deserts for long-range enhancers." Science 302(5644): 413.

Schotta, G., R. Sengupta, et al. (2008). "A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse." Genes Dev 22(15): 2048-61.

Schroder, A. R., P. Shinn, et al. (2002). "HIV-1 integration in the human genome favors active genes and local hotspots." Cell 110(4): 521-9.

Sims, R. J., 3rd and D. Reinberg (2009). "Processing the H3K36me3 signature." Nat Genet 41(3): 270-1.

Stevens, S. W. and J. D. Griffith (1996). "Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration." J Virol 70(9): 6459-62.

Studamire, B. and S. P. Goff (2008). "Host proteins interacting with the Moloney murine leukemia virus integrase: multiple transcriptional regulators and chromatin binding factors." Retrovirology 5: 48.

Suzuki, Y. and R. Craigie (2002). "Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes." J Virol 76(23): 12376-80.

Tomilin, N., L. Solovjeva, et al. (1995). "Visualization of elementary DNA replication units in human nuclei corresponding in size to DNA loop domains." Chromosome Res 3(1): 32-40.

Vacharaksa, A., A. C. Asrani, et al. (2008). "Oral keratinocytes support non-replicative infection and transfer of harbored HIV-1 to permissive cells." Retrovirology 5: 66.

Violot, S., S. S. Hong, et al. (2003). "The human polycomb group EED protein interacts with the integrase of human immunodeficiency virus type 1." J Virol 77(23): 12507-22.

Wang, G. P., A. Ciuffi, et al. (2007). "HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications." Genome Res 17(8): 1186-94.

Wang, G. P., A. Garrigue, et al. (2008). "DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer." Nucleic Acids Res 36(9): e49.

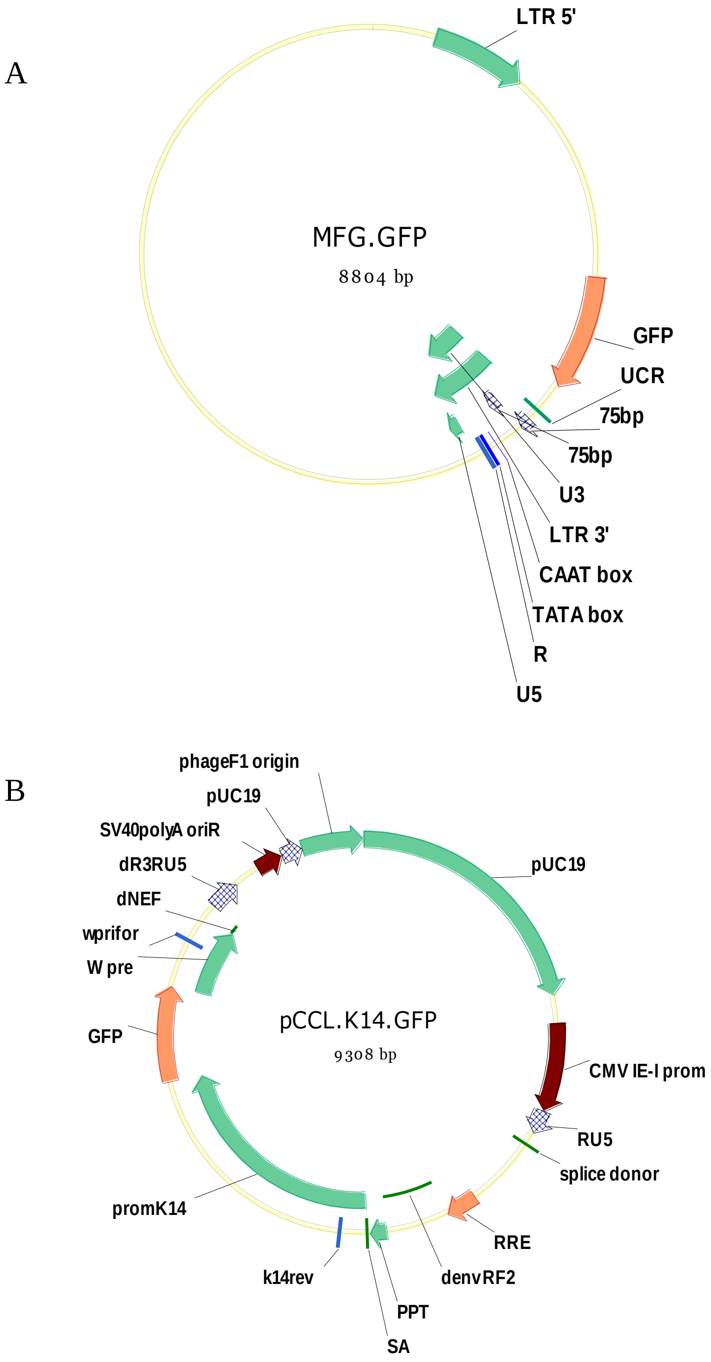
Woodfine, K., H. Fiegler, et al. (2004). "Replication timing of the human genome." Hum Mol Genet 13(2): 191-202.

Wu, X., Y. Li, et al. (2003). "Transcription start regions in the human genome are favored targets for MLV integration." Science 300(5626): 1749-51.

Wu, X., Y. Li, et al. (2005). "Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses." J Virol 79(8): 5211-4.

Wu X, L. Y., Crise B, Burgess SM. (2003). "Transcription start regions in the human genome are favored targets for MLV integration." Science 300: 1749-1751.

Appendix



Map of the retroviral vectors used to infect primary human keratinocytes.
A) oncoretrovirus (MLV) B) lentivirus (HIV)