UNIVERSITY OF CALIFORNIA, SAN DIEGO

Retroviral Integration: Mechanism and Consequences

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Mary Kathleen Lewinski

Committee in charge:

Theodore Friedmann, Chair
Frederic D. Bushman, Co-chair
John C. Guatelli
Victor Nizet
Douglas D. Richman
John A. T. Young

2005

The dissertation of Mary Kathleen Lewinski is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____
Co-chair

_____
Chair

University of California, San Diego

2005

DEDICATION

To my family, for your patience, love and support. Thank you for instilling in me an appreciation for higher education.

To my friends, who put up with me even when I cancelled our plans so that I could get some lab work done.

With much love,

M.K.L.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter II

Chapter III

LIST OF TABLES

ACKNOWLEDGEMENTS

The text of Chapter Two, in full, has been submitted for publication as:

Lewinski, M. K., Yamashita, M., Emerman, M., Shinn, P., Leipzig, J., Hannenhalli, S., Berry, C. C., Ecker, J. R., and Bushman, F. D. "Retroviral Gag and integrase act synergistically to determine integration target specificity," 2005.

The dissertation author was the primary researcher and author.

The text of Chapter Three, in full, is a reprint of the material as it appears in the

Journal of Virology:

Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C. C., Ecker, J. R., and Bushman, F. D. "Genome-wide analysis of chromosomal features repressing HIV transcription". *J Virol* **79**, 6610-9, 2005.

The dissertation author was the primary researcher and author.

VITA

2000	B.A., Philosophy, University of Southern California
2000	B.A., Political Science, University of Southern California
2005	Ph.D., Biomedical Sciences, University of California, San Diego
In progress	M.D., University of California, San Diego School of Medicine

PUBLICATIONS

M. K. Lewinski, D. Bisgrove, P. Shinn, H. Chen, C. Hoffmann, S. Hannenhalli, E. Verdin, C. C. Berry, J. R. Ecker, and F. D. Bushman. "Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription," Journal of Virology 79 (11): 6610-6619, June, 2005.

RESEARCH EXPERIENCE

1998-1999	Department of Political Science, University of Southern California Advisor: Sunhyuk Kim, Ph.D. Undergraduate Honors Thesis: *The Prospect of Democracy in Central Asia: A Critical Analysis of the Political Climates of Kazakhstan and Uzbekistan*
1998-1999	Department of Neuroscience, University of Southern California Advisor: William O. McClure, Ph.D.
2001-2002	Department of Pathology, University of California, San Diego School of Medicine Advisor: Celsa Spina, Ph.D.
2002	Department of Pathology, University of California, San Diego School of Medicine Advisor: Steffan Ho, M.D., Ph.D.
2002-2005	Infectious Disease Laboratory, The Salk Institute for Biological Studies Advisor: Frederic D. Bushman, Ph.D.

ABSTRACT OF THE DISSERTATION

Retroviral Integration: Mechanism and Consequences

by

Mary Kathleen Lewinski

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2005

Professor Theodore Friedmann, Chair

Professor Frederic D. Bushman, Co-chair

Integration of retroviral cDNA into the host cell genome is a process central to the replication cycle of retroviruses and is mediated by the virally-encoded integrase protein. While any DNA sequence can be a target for integration, retroviruses do not integrate randomly into the host cell chromosomes. Recent studies have found that different retroviruses have distinct target site selection preferences for various genomic features. We have sequenced integration sites from human immunodeficiency virus (HIV), murine leukemia virus (MLV) and three HIV-MLV chimeras and determined that both integrase and the viral Gag proteins act together to

determine virus-specific integration site selection preferences.

Once integrated, the provirus is transcribed by the host cell machinery into messenger RNA and the viral RNA genome. A number of factors are thought to contribute to the level of proviral gene expression. For HIV, these include the activation state of the host cell, CpG methylation, nucleosome organization, and mutations in the viral transactivator, *tat*, or transcription factor binding sites in the viral promoter and enhancer. Factors that negatively influence HIV gene expression are of interest because of the phenomenon of viral latency, where HIV persists in the genome of host cells undetected due to a lack of expression. We set out to determine the extent to which integration site influences expression of the HIV provirus. In this study, we infected Jurkat T cells with an HIV-based vector transducing GFP and separated cells into GFP-expressing and non-expressing populations. We then sequenced integration sites from these two populations. Low proviral expression correlated with integration into (1) gene deserts, (2) centromeric heterochromatin, and (3) very highly expressed host cell genes. These data suggest that particular genomic features influence the expression of HIV proviruses and provide models for postintegration latency in cells from HIV-infected patients.

I. INTRODUCTION

Retroviruses are a group of RNA viruses that replicate by integrating a DNA copy of their genome into the host cell chromosome and relying on the cellular machinery to produce copies of the viral genomic RNA and viral proteins. Many retroviruses are vertebrate pathogens, causing the formation of tumors, as with the oncoretroviruses, or crippling the host's immune system, as with human immunodeficiency virus (HIV), the cause of the AIDS pandemic.

Central to the retroviral replication cycle is integration of the viral cDNA into the host cell genome. Integration is essential for the production of new virions. Beyond being indispensable for the virus, integration can have significant consequences for the host—occasionally resulting in insertional mutagenesis and contributing to transcriptional latency of HIV.

THE RETROVIRAL REPLICATION CYCLE

Following binding of the retroviral envelope glycoproteins to its cellular receptor(s), the virus fuses with the cell membrane, releasing its core into the host cell cytoplasm. The viral genomic RNA is then reverse transcribed to form double-stranded DNA. The viral DNA—in a complex with integrase and other viral and cellular proteins—enters the nucleus. There, the viral integrase protein covalently joins the viral DNA to the host cell DNA. Once integrated, the viral DNA—called the "provirus"—acts as a transcription template for efficient synthesis of viral mRNA and genomic RNA. Viral proteins are translated and assemble with the viral genomic

RNA. These new virions then bud from the host cell membrane, mature, and infect new host cells. The provirus persists indefinitely in the host cell chromosome and is inherited by daughter cells like any other gene during cell division.

Integrase is essential to viral replication. For retroviral DNA to efficiently direct the production of progeny virions it must become covalently integrated into the host cell chromosome (reviewed in (Coffin *et al.*, 1997; Hansen *et al.*, 1998)). Some expression from unintegrated viral DNA can be detected (Panganiban and Temin, 1983), but this is not sufficient to sustain a spreading infection (Engelman *et al.*, 1995; Englund *et al.*, 1995). Analyses of mutants have identified the viral integrase coding region of the retroviral *pol* gene as essential for the integration process (Donehower, 1988; Donehower and Varmus, 1984; Panganiban and Temin, 1984; Quinn and Grandgenett, 1988; Schwartzberg *et al.*, 1984). Also essential are regions at the ends of the viral long terminal repeats (LTRs) that serve as recognition sites for the integrase protein (Colicelli and Goff, 1985; Colicelli and Goff, 1988; Panganiban and Temin, 1983).

Phenotypes of integrase mutants. Extensive work has shown that integrase mutants can have a variety of effects on viral replication. Integrase mutants containing substitutions in the enzyme active site (considered below) generally have effects only at the integration step in the viral life cycle. However, integrase may play additional roles in viral replication, perhaps as a structural component of replication intermediates. Integrase is present as part of the retroviral *gag-pol* polyprotein during assembly and budding and is present in reverse transcription complexes after infection

of new host cells (Fassati and Goff, 1999; Nermut and Fassati, 2003). Many mutants of integrase, including deletion mutants, can have pleiotropic effects on the viral life cycle, including effects on particle budding, infectivity and reverse transcription (Engelman*, et al.*, 1995).

MECHANISM OF INTEGRATION

Integration of viral DNA into the host cell chromosome involves several coordinated steps: processing of the viral DNA ends, coordinated joining of those ends to target DNA, and repairing of the gaps. The first two reactions are catalyzed by the viral integrase protein while the last is mediated by as-yet-undefined factors.

DNA breaking and joining reactions catalyzed by integrase. The viral genomic RNA is reverse transcribed to form a linear double-stranded DNA molecule, the precursor to the integrated provirus (Brown *et al.*, 1987; Brown *et al.*, 1989; Fujiwara and Mizuuchi, 1988). The provirus is colinear with unintegrated linear viral DNA (Dhar *et al.*, 1980; Hughes *et al.*, 1978) but differs from the reverse transcription product in that it is missing two (or for some retroviruses, three) bases from each end (Hughes *et al.*, 1981). Flanking the integrated provirus are direct repeats of the cellular DNA that are usually 4-6 base pairs in length, depending on the viral integrase (Hughes*, et al.*, 1981; Vincent *et al.*, 1990). This duplication of cellular sequences flanking the viral DNA is generated as a consequence of the integration mechanism (Coffin*, et al.*, 1997).

Linear viral DNA is found in a complex with proteins in the cytoplasm of infected cells. These complexes (termed “preintegration complexes”) can be isolated

and have been shown to mediate integration of viral DNA into target DNA *in vitro* (Bowerman *et al.*, 1989; Brown*, et al.*, 1987; Ellison *et al.*, 1990; Farnet and Haseltine, 1990; Farnet and Haseltine, 1991).

The development of *in vitro* assays with purified integrase has allowed its enzymatic functions to be elucidated. The provirus is the result of two reactions catalyzed by the viral integrase: terminal cleavage and strand transfer. Studies with purified integrase have shown that it is sufficient for both 3′ end cleavage (Bushman and Craigie, 1991; Craigie *et al.*, 1990; Katzman *et al.*, 1989; Sherman and Fyfe, 1990) and joining of the viral DNA to the cellular chromosome or naked target DNA (Bushman *et al.*, 1990; Craigie*, et al.*, 1990; Katz *et al.*, 1990). Most integrase proteins catalyze the removal of two bases from the 3′ end of each viral DNA strand, leaving recessed 3′ hydroxyl groups (Brown*, et al.*, 1989; Fujiwara and Mizuuchi, 1988; Roth *et al.*, 1989; Sherman and Fyfe, 1990). This terminal cleavage reaction is required for proper integration. It may allow the virus to create a standard end from viral DNA termini that can be heterogeneous due to the terminal transferase activity of reverse transcriptase (Miller *et al.*, 1997; Patel and Preston, 1994). In addition, the terminal cleavage step is coupled to the formation of a stable integrase-DNA complex (Ellison and Brown, 1994; Vink *et al.*, 1994). A recent study has suggested that this 3′ end processing facilitates the formation of a complex that is capable of directing concerted integration of the viral DNA ends (Li and Craigie, 2005). Following terminal cleavage, a recessed hydroxyl is exposed that immediately follows a CA dinucleotide. This CA is conserved among retroviruses and many related transposons.

Evidence suggests that more internal LTR sites are also important for integration (Balakrishnan and Jonsson, 1997; Bushman and Craigie, 1990; Leavitt *et al.*, 1992). After end processing, integrase catalyzes the covalent attachment of hydroxyl groups at the viral DNA termini to protruding 5′ phosphoryl ends of the host cell DNA (Brown*, et al.*, 1987; Brown*, et al.*, 1989; Fujiwara and Mizuuchi, 1988). Both the viral DNA 3′ end cleavage and strand transfer reactions are likely mediated by single-step transesterification chemistry as shown by stereochemical analysis of the reaction course (Engelman *et al.*, 1991).

Purified integrase can also catalyze the "reverse" of the strand transfer reaction, termed disintegration (Chow *et al.*, 1992). Assays for disintegration activity have been useful in the analysis of defective integrase mutants because the requirements for disintegration seem to be more lenient than those for integration.

Biochemical analysis of purified integrase revealed that it requires a divalent metal—either Mg^{2+} or Mn^{2+}—to carry out reactions with model substrates (Chow*, et al.*, 1992). As is discussed below, several structures of integrase show a divalent metal bound at the active site. Modeling suggests that two cations at the active site are important, the second of which is likely carried to the active site by the DNA substrate (Bujacz *et al.*, 1997; Lins *et al.*, 2000). A more recent report detailing Cys substitutions at HIV-1 integrase active site residues D64 and D116 suggested that these resides act by binding divalent metal (Gao *et al.*, 2004). Divalent metal is also involved in assembly and stabilization of integrase-DNA complexes (Bujacz*, et al.*, 1997; Gao*, et al.*, 2004; Hazuda *et al.*, 1997; Lee *et al.*, 1995; Yi *et al.*, 1999).

Host factors involved in repair of gaps in integration intermediates.

Integrase carries out the terminal cleavage and strand transfer steps that initiate viral DNA integration. Concerted integration of both ends of the viral DNA, followed by melting of the target DNA segments between the points of joining, yields single-stranded gaps at each host-virus DNA junction, and a two base overhang derived from the viral DNA. The manner by which this intermediate is subsequently repaired to yield the fully integrated provirus is unclear. For many parasitic DNA replication reactions, the parasite carries out reaction steps only up to a point that the host cannot easily reverse, forcing the host to complete the job (Bushman, 2001; Craig *et al.*, 2002). For retroviral integration, it is reasonable to infer that host DNA repair enzymes complete provirus formation. DNA gap repair enzymes are known to be involved in a variety of DNA repair pathways, so their recruitment to gaps at host-virus DNA junctions is readily envisioned. Consistent with this, known gap repair enzymes have been shown to act on model host-virus DNA junctions *in vitro* (Yoder and Bushman, 2000).

STRUCTURE OF THE INTEGRASE PROTEIN AND MULTIMERS

The integrase protein is composed of three separate domains—the N-terminal zinc-binding domain, the catalytic core and the C-terminal DNA-binding domain. The three-domain structure was initially suggested by partial proteolysis studies (Engelman *et al.*, 1993). Later their structures were solved by NMR and x-ray crystallography. The crystal and NMR structures of each domain indicate that each dimerizes (Cai *et al.*, 1997; Chen *et al.*, 2000a; Chen *et al.*, 2000b; Eijkelenboom *et al.*, 1999; Goldgur

et al., 1999; Goldgur *et al.*, 1998; Lodi *et al.*, 1995; Maignan *et al.*, 1998; Wang *et al.*, 2001; Yang *et al.*, 2000), but the relevance of these structures to integrase function *in vivo* remains under investigation. It is known that all three domains are essential for the full catalytic activity of integrase (Drelich *et al.*, 1992; Schauer and Billich, 1992; Vink *et al.*, 1993). The structure and function of each domain, along with what is known about how they are assembled in the full-length protein and in integrase oligomers, are discussed in turn below.

N-terminal domain of integrase. The N-terminal domain (approximately the first 50 amino acids) of integrase is thought to promote DNA binding and multimerization. It has a conserved HHCC zinc-binding motif with an overall fold resembling that of the helix-turn-helix bacterial repressors (Cai*, et al.*, 1997; Eijkelenboom *et al.*, 1997) that is conserved in all retroviral and retrotransposon integrases. Evidence indicates that this domain must bind Zn^{2+} to function (Bushman *et al.*, 1993; Coffin*, et al.*, 1997; Eijkelenboom*, et al.*, 1997).

Integrase mutants with the N-terminal domain deleted or with substitutions in the conserved His or Cys residues are significantly impaired in their ability to catalyze 3′ end cleavage and strand transfer reactions but still maintain disintegration activity (Bushman*, et al.*, 1993; Bushman and Wang, 1994; Engelman and Craigie, 1992; Vincent *et al.*, 1993). Other mutants of less highly conserved amino acids in the N-terminal domain have weak end cleavage and strand transfer activities (Vincent*, et al.*, 1993). Adding Zn^{2+} *in vitro* was found to enhance the Mg^{2+}-dependent terminal cleavage reaction by HIV-1 integrase (Lee and Han, 1996). This suggests that the N-

terminal domain, while having no direct role in catalysis, might play some role in viral DNA recognition.

Another possible role for the N-terminal domain of integrase is in multimerization (Heuer and Brown, 1998; Lee *et al.*, 1997; Zheng *et al.*, 1996) (discussed in more detail below). Studies of the zinc-binding properties of integrase found that the Zn^{2+}-bound N-termini dimerized (Yang *et al.*, 1999) and Zn^{2+}-bound integrase tetramerized more easily than integrase without Zn^{2+} or with mutations in the HHCC motif (Zheng*, et al.*, 1996). Binding of Zn^{2+} to the N-terminal domain of integrase likely stabilizes the enzyme, allowing for proper multimerization and efficient enzymatic activity. Cross-linking studies have also implicated the N-domain in binding of target DNA (Heuer and Brown, 1997).

Catalytic core of integrase. The central domain of integrase (e.g. residues 50-212 of HIV-1 integrase) functions primarily in catalysis and DNA binding. The catalytic core is comprised of mixed alpha helix and beta sheets folded such that three acidic residues of the $D,DX_{35}E$ motif are in close proximity. This three-dimensional structure is an RNaseH-type fold that is conserved among members of the $D,DX_{35}E$ phosphotransferase enzyme family that includes retroviral and retrotransposon integrases and bacterial transposases (Dyda *et al.*, 1994; Kulkosky *et al.*, 1992; Rowland and Dyke, 1990; Yang and Steitz, 1995).

Site-directed mutagenesis of conserved amino acids in this catalytic core resulted in integrase proteins that were inactive in 3′ end cleavage, DNA strand transfer and disintegration assays, suggesting that this domain is essential for catalysis

(Engelman and Craigie, 1992; Hazuda *et al.*, 1994; Leavitt *et al.*, 1996). In fact, the catalytic domain alone is sufficient to catalyze disintegration (Bushman*, et al.*, 1993; Bushman and Wang, 1994; Kulkosky *et al.*, 1995; Vink*, et al.*, 1993), although efficient 3′ end cleavage and strand transfer also require the N-terminal and C-terminal domains (Bushman and Wang, 1994; Drelich*, et al.*, 1992; Schauer and Billich, 1992; Vink*, et al.*, 1993).

Each residue of the D,DX_{35}E motif catalytic triad is required for catalysis of integration (Engelman and Craigie, 1992; Kulkosky*, et al.*, 1992; van Gent *et al.*, 1993a). D,DX_{35}E motif residues D64 and D116 of HIV-1 integrase are thought to act by coordinating at least one divalent metal ion and probably two (Gao*, et al.*, 2004). While initial crystal structures of the catalytic domain did not include a bound cation (Bujacz *et al.*, 1996a; Bujacz *et al.*, 1995; Dyda*, et al.*, 1994), later structures (Bujacz *et al.*, 1996b; Goldgur*, et al.*, 1998; Maignan*, et al.*, 1998) and models (Lins *et al.*, 1999) showed that the aspartic acid residues of the catalytic triad can coordinate Mn^{2+} and/or Mg^{2+}. One structure of the avian sarcoma virus (ASV) integrase catalytic domain has been visualized with two bound metal atoms (although the Zn^{2+} and Cd^{2+} atoms bound are not biological ligands) (Bujacz*, et al.*, 1997) and the catalytic domain of HIV-1 integrase with two bound cations at the active site was subsequently modeled (Lins*, et al.*, 2000). Although integrase bound to two metal atoms has not yet been proven capable of catalyzing integration *in vitro*, these crystal and model structures suggest that the mechanism involves two bound cations.

In addition to catalysis of terminal cleavage and strand transfer reactions, the core domain functions in binding to viral DNA. Studies with chimeric integrases have shown that the core domain is responsible for recognition of the viral DNA substrate (Katzman and Sudol, 1995; Katzman and Sudol, 1998; Pahl and Flugel, 1995) and cross-linking studies with HIV-1 integrase found that core domain residues Q148 and Y143 bind to the viral DNA ends (Esposito and Craigie, 1998). Cross-linking data suggest that the conserved residues K156 and K159 of HIV-1 integrase (near the active site in the catalytic core domain) are essential for the interaction between integrase and viral DNA, specifically the conserved deoxyadenosine (Jenkins *et al.*, 1997). Further, the core domain is thought to be responsible for target site selection *in vitro* (Appa *et al.*, 2001; Harper *et al.*, 2003; Shibagaki and Chow, 1997).

C-terminal domain of integrase. The final 75-100 amino acids of integrase comprise the C-terminal domain—the least conserved of the three domains. Structural analysis has found that it has an SH3-type fold, and may form dimers (Eijkelenboom *et al.*, 1995; Eijkelenboom*, et al.*, 1999; Lodi*, et al.*, 1995). The C-terminal domain has strong but nonspecific DNA-binding activity and thus has been called the DNA-binding domain (Engelman *et al.*, 1994; Khan *et al.*, 1991; Lutzke and Plasterk, 1998; Lutzke *et al.*, 1994; Mumm and Grandgenett, 1991; Vink*, et al.*, 1993; Woerner and Marcus-Sekura, 1993). Its ability to dimerize in solution has led some to suggest that the C-terminal domain plays a role in multimerization (Andrake and Skalka, 1995; Lutzke and Plasterk, 1998). Mutagenesis data support a role for the C-terminal domain in proper folding of the integrase protein (Moreau *et al.*, 2003).

Structures containing two domains of integrase. Although there are NMR and crystal structures for the individual domains of integrase, these are not sufficient to determine the structural arrangement of domains in full-length integrase protein. Full-length integrase has not been crystallized. However, several structures of two-domain integrase fragments have been solved. These two-domain structures provide insight into the mechanism of host and viral DNA binding by and multimerization of integrase.

Two-domain structures with the catalytic core and C-terminal domains have been solved for Rous sarcoma virus (RSV) (Yang*, et al.*, 2000), HIV-1 (Chen*, et al.*, 2000a) and simian immunodeficiency virus (SIV) (Chen*, et al.*, 2000b) integrases. Additionally, the structure of a two-domain HIV-1 integrase fragment with the catalytic and N-terminal domains has been determined (Wang*, et al.*, 2001). In each of these structures the catalytic core domains are associated as dimers, as they are in structures of the catalytic domain alone (Bujacz*, et al.*, 1995; Goldgur*, et al.*, 1999; Goldgur*, et al.*, 1998; Lubkowski *et al.*, 1999; Maignan*, et al.*, 1998). However, the position of the C-terminal domain varies considerably among these two-domain structures. The two-domain structure of RSV integrase shows the C-terminal domains associated as a dimer in a canted conformation such that one C-terminal domain contacts its catalytic domain (Yang*, et al.*, 2000). In the catalytic/C-terminal two-domain structure for HIV-1 integrase, the catalytic cores exist as dimers, but the C-terminal domains are monomeric and at the ends of extended alpha-helical linkers such that the structure is in a Y conformation (Chen*, et al.*, 2000a). In the two-domain

structure of SIV integrase, only one of the four C-terminal domains associated with two dimers of the catalytic domain can be visualized (Chen*, et al.*, 2000b). The C-terminal domain is poorly conserved among different retroviral integrases so it is not entirely unexpected that its conformations differ in these two-domain structures. However, it is unclear whether any of these structures is similar to the actual conformation of these domains *in vivo*. The variation in C-terminal domain position relative to the catalytic domain can be attributed to the flexibility of the linker and/or the lack of the stabilizing N-terminal domain or DNA.

An HIV-1 integrase fragment that includes the catalytic core and N-terminal domains also crystallized as a dimer (Wang*, et al.*, 2001). In this structure, the N-terminal domains are arranged differently than seen in dimers of the individual N-terminal domain (Cai*, et al.*, 1997). This two-domain structure can accommodate the C-terminal domain in the same orientation observed in the catalytic/C-terminal two-domain structure of HIV-1 integrase (Chen*, et al.*, 2000a). This suggests that the N-terminal domain could stabilize the structure of the C-terminal and catalytic domains of HIV-1 integrase.

The two-domain structures of integrase allow for modeling of integrase bound to viral and target DNA. Using time-resolved fluorescence anisotropy (TFA) (Deprez *et al.*, 2000; Leh *et al.*, 2000), protein footprinting (Dirac and Kjems, 2001), and cross-linking data (Esposito and Craigie, 1998; Gao *et al.*, 2001; Heuer and Brown, 1997; Heuer and Brown, 1998; Jenkins*, et al.*, 1997) in addition to the structural data reviewed above, Podtelezhnikov and colleagues modeled HIV-1 integrase dimers

bound to DNA (Podtelezhnikov *et al.*, 2003). Their model differs from the full-length integrase structure suggested by Wang and colleagues (Wang*, et al.*, 2001) in that the domains are tightly compacted together. This conflicts with the catalytic core/C-terminal two-domain HIV-1 integrase structure (Chen*, et al.*, 2000a), which has the two domains linked by an extended alpha-helix. Such a structure was not compatible with the TFA data. In their model of this compacted integrase dimer bound to DNA, the terminal three bases of viral DNA interact only with the catalytic core domain while host target DNA binds to all three domains (Podtelezhnikov*, et al.*, 2003). This model is able to accommodate both structural and experimental data. The C-terminal and catalytic core domains are known to bind DNA nonspecifically (Engelman*, et al.*, 1994). Also, in this model the zinc finger of the N-terminal domain contacts host DNA as seen with cross-linking data (Heuer and Brown, 1997).

The structures of these two-domain integrases and the subsequent models of integrase-DNA complexes lend further support to the idea that integrase acts as a tetramer. Dimers in the two-domain structures have the catalytic core active sites on opposite sides of the complex—too far apart to account for the spacing between sites of integration of the viral DNA ends. This suggests that integration *in vivo* proceeds with each viral DNA end associated with an integrase dimer assembled as a tetramer.

Multimerization of integrase. As mentioned above, structural analysis of integrase and its domains has determined that integrase can self-associate to form dimers and tetramers *in vitro*. Studies have shown that pairs of integrase mutants that are inactive alone can complement each other and function to near-wild-type levels *in*

vitro (Engelman*, et al.*, 1993; Fletcher *et al.*, 1997; van Gent *et al.*, 1993b). This suggests that integrase acts as a multimer. Other studies have found that multimerization is required for integrase end cleavage and joining reactions (Jones *et al.*, 1992), with the smallest functional integrase unit being a dimer (Bao *et al.*, 2003; Jones*, et al.*, 1992). Cross-linking studies with preintegration complexes indicate that integrase molecules associate as tetramers *in vivo* (Gao*, et al.*, 2001).

Podtelezhnikov and colleagues modeled the structure of an HIV-1 integrase tetramer bound to viral and host DNA using TFA data (Deprez*, et al.*, 2000; Leh*, et al.*, 2000) and computer simulations of the hydrodynamic properties of integrase oligomers. They also incorporated data from crystal structures, cross-linking and other biochemical data on integrase-DNA interactions. They reasoned that their model dimer of HIV-1 integrase (discussed above) is not sufficient to catalyze concerted integration because the active sites are too far apart to account for the five base pairs that separate the points of joining of HIV-1 DNA ends to the target DNA. Thus a tetramer, with a dimer catalyzing integration of each viral DNA end, is the likely functional oligomer *in vivo*. The model tetramer is composed of monomers with the same structure. One monomer from each dimer catalyzes the integration of one end of the viral DNA while the other monomer serves a structural role. The viral DNA is bound to the catalytic core domain of the active monomer as described above and also contacts the C-terminal domain of the monomer that catalyzes integration of the other viral DNA end, as suggested by experimental data (Esposito and Craigie, 1998; Heuer and Brown, 1997). The host DNA binds the catalytic core domain near

the site of integration and contacts both C-terminal domains of the catalytically active monomers at the five base pairs between the points of joining of the viral and target DNA. The N-terminal domains bind host DNA outside of this region according to the model. The dimer-dimer interface involves N-terminal domains of the structural, catalytically inactive monomers, explaining why zinc-binding facilitates tetramerization (Deprez*, et al.*, 2000; Zheng*, et al.*, 1996). This model tetramer is structurally similar to the Tn5 transposase-DNA complex (Rice and Baker, 2001).

A consequence of higher-order assembly of nucleoprotein complexes containing integrase is coupled joining (also termed concerted integration). Coupled joining is the integration of both viral DNA ends into opposite strands of the target DNA. Correct integration *in vivo* requires joining of both ends of viral DNA with two points in target DNA that are a specific number of base pairs apart (five for HIV-1, four for murine leukemia virus), depending on the retrovirus. Such coupled joining reactions can be reproduced under carefully controlled conditions *in vitro* (Aiyar *et al.*, 1996; Carteau *et al.*, 1999; Goodarzi *et al.*, 1995; Sinha *et al.*, 2002), though as yet complex assembly is somewhat inefficient. Coupled joining can be detected as a DNA product of a distinctive length in gels and by sequencing of viral-host DNA junctions to ensure that target site duplication of the correct length is formed after gap repair. The DNA forms detected as a result of the reactions *in vitro* are frequently a mixture of coupled and uncoupled products. While progress has been made, efficient reconstitution of integration complexes from purified components has not been fully

achieved. One possibility is that additional proteins have a role in assembly of fully functional integrase complexes.

COMPOSITION OF INTEGRASE COMPLEXES IN VIVO

Integration *in vivo* is carried out by a nucleoprotein complex that includes the viral DNA and integrase (Bowerman*, et al.*, 1989; Brown*, et al.*, 1987; Ellison*, et al.*, 1990; Farnet and Haseltine, 1990; Farnet and Haseltine, 1991; Li *et al.*, 2001; Miller*, et al.*, 1997). With the development of assays involving preintegration complexes (PICs) purified from virally infected cells (Brown*, et al.*, 1987; Ellison*, et al.*, 1990; Farnet and Haseltine, 1990), it has become possible to study the organization and function of authentic replication intermediates. PIC preparations have been generated for cells infected with HIV-1 and murine leukemia virus (MLV). For avian sarcoma-leukosis virus (ASLV), complexes did not efficiently complete reverse transcription, suggesting a late block in replication (Lee and Coffin, 1991). A limitation on studies of PICs has been the difficulty of obtaining large amounts of material. Even if cells are infected at high multiplicity, only several PICs per cell can be purified. Therefore, only small quantities of PICs can be studied in the background of a complex mixture of cellular proteins. So far PICs have not been purified to homogeneity. Nevertheless it has been possible to infer a number of their features using sensitive biochemical approaches.

PICs can be shown to have proteins tightly bound at the viral DNA ends. The ends are protected from attack by exonucleases (Miller*, et al.*, 1997) or recombination complexes (Wei *et al.*, 1997; Wei *et al.*, 1998b). In addition, it can be shown that the

ends of the viral DNA are held together by a protein-DNA complex because the viral DNA can be cut internally with restriction endonucleases and integration can still occur with the viral DNA ends (Miller*, et al.*, 1997).

Viral proteins in the preintegration complex. Given the difficulty of purifying PICs, it has been challenging to get precise information on their protein composition. PICs of HIV-1 have been shown to contain the viral integrase, matrix and reverse transcriptase (Miller*, et al.*, 1997) but very little capsid (Farnet and Haseltine, 1991). A study using fluorescent microscopy to track the transit and composition of viral complexes in the host cell suggested that some capsid remains associated with most but not all viral particles through the initiation of reverse transcription (McDonald *et al.*, 2002). The point at which HIV capsid dissociates from the reverse transcription complex or the PIC has not been clarified. The HIV Vpr and nucleocapsid proteins are detectable in early fractions and probably remain associated with the PIC but this has been difficult to demonstrate with more purified preparations (Miller*, et al.*, 1997). For MLV, integrase and capsid are readily detected in the PIC, suggesting that more capsid remains associated with MLV than HIV-1 PICs (Bowerman*, et al.*, 1989; Li*, et al.*, 2001).

Several viral proteins have been shown to stimulate reactions with purified integrase *in vitro*, notably the nucleocapsid protein (NC) (Carteau*, et al.*, 1999; Gao *et al.*, 2003). Under specific conditions *in vitro*, the magnitude of the stimulation by NC can be 1000-fold or more (Carteau*, et al.*, 1999). The effects of NC mutants *in vivo* have been difficult to study because NC is required for multiple steps in the viral life

cycle, including RNA dimerization, packaging, and reverse transcription. Studies from Gorelick and coworkers using carefully selected NC mutants have provided some support for the idea that NC is important for integration as well (Buckman *et al.*, 2003; Carteau*, et al.*, 1999). They found that viral DNAs captured in junctions between 2-LTR circles tended to be predominantly uncleaved by integrase in the presence of the zinc-finger residue substitution CCCC/CCHC NC mutant, suggesting a requirement for NC in integrase-catalyzed terminal cleavage of viral DNA. This readout is indirect but does support the notion that NC is involved in integrase function.

Host proteins in the preintegration complex. Several host cell proteins have been suggested to be important for retroviral DNA integration. None have yet been shown to be strictly required for integration *in vivo*, however, leaving the importance of each proposed protein uncertain.

The functions of many DNA-binding proteins and DNA-modifying enzymes are assisted by architectural DNA-binding proteins. These proteins act by changing the direction of the long axis of the DNA helix and/or neutralizing negative charges in the DNA phosphate backbone, assisting in the formation of precise three-dimensional nucleoprotein structures. Many such examples have been reported (Bushman, 2001; Craig*, et al.*, 2002) to the point where it would be surprising if architectural DNA-binding proteins were not involved in integration. A complication, however, is that in many cases multiple small basic proteins can satisfy the requirement for architectural

DNA-binding proteins, so that redundancy greatly complicates the assessment of *in vivo* importance of any single protein.

In one experimental paradigm, PICs were subject to gel filtration in the presence of high salt, resulting in a loss of integrase activity (Chen and Engelman, 1998; Farnet and Bushman, 1997; Harris and Engelman, 2000; Lee and Craigie, 1994; Li *et al.*, 1998). Adding back extracts from uninfected cells was found to restore activity. Fractionation of such extracts has led to the identification of several cellular proteins that can support reconstitution, two of which have been identified as HMGA and BAF. HMGA was identified through studies of HIV-1 PICs (Farnet and Bushman, 1997) while BAF was identified with studies of MLV (Cai *et al.*, 1998; Lee and Craigie, 1994). MLV PICs exposed to high salt tend to use their own DNA as an integration target, a process called autointegration. BAF succeeded in blocking autointegration, hence the name: barrier-to-autointegration factor. The importance of both of these proteins *in vivo* is uncertain. Cells knocked-out for the two HMGA family proteins nevertheless supported wild-type levels of integration (Beitzel and Bushman, 2003), indicating that either HMGA is not important *in vivo* or it is redundant with other factors. Mutation of BAF is lethal to cells and so cells lacking this factor cannot be studied. However, MLV autointegration is very efficient *in vitro*, suggesting that there may be a mechanism—such as BAF binding to viral DNA—that blocks this *in vivo* (Lee and Craigie, 1994; Lee and Coffin, 1990). The viral NC shows some activity in reconstitution after salt-stripping (Farnet and Bushman, 1997), raising the possibility that this viral protein is a contributor during normal infection.

Assays *in vitro* using purified integrase can also be used to assess the function of candidate cofactors. Both NC and HMGA have been shown to stimulate reactions with purified HIV-1 integrase (Carteau *et al.*, 1997; Gao*, et al.*, 2003; Hindmarsh *et al.*, 1999) while BAF appears to inhibit the activity of purified integrase (unpublished results). In contrast, a recent study suggests that NC, HMGA1 and BAF have no effect on the concerted integration of viral DNA ends by integrase (Li and Craigie, 2005). The relationship of these results to integration *in vivo* has not been clarified.

Another route to identifying candidate cellular proteins has involved searching for proteins that bind tightly to HIV-1 integrase. The first such protein to be identified using the yeast two-hybrid assay was Ini1 (Kalpana *et al.*, 1994), a cellular protein that is a member of the SWI/SNF chromatin remodeling complex. Purified Ini1 is able to stimulate integration *in vitro* under certain conditions. Data suggestive of its *in vivo* importance comes from overexpression of Ini1 fragments. Overexpression of Ini1 fragments that contain the integrase-interaction domain has shown very strong dominant-negative effects, though unexpectedly, these inhibited HIV late in the viral replication cycle—after integration (Yung *et al.*, 2001). Though these data are provocative, it is still uncertain what role, if any, Ini1 plays in normal HIV replication.

Yet another cellular protein identified by binding to HIV-1 integrase is LEDGF/p75 (Cherepanov *et al.*, 2003). LEDGF was first identified as a transcriptional mediator protein using biochemical assays (Ge *et al.*, 1998). LEDGF was also identified as a stress-responsive transcription factor in ocular tissues, hence the name: lens epithelium-derived growth factor. The name notwithstanding, LEDGF

appears to be expressed in most tissues assayed. LEDGF can affect the location of integrase inside cells (Llano *et al.*, 2004b). In the presence of LEDGF, HIV-1 integrase can be detected bound to cellular chromatin, suggesting that LEDGF may help bring integrase to target DNA (Maertens *et al.*, 2003). Binding to LEDGF also protects integrase from proteolysis (Llano *et al.*, 2004a). Despite this data supporting its potential importance *in vivo*, so far functional studies indicate that knock-down of LEDGF does not diminish viral replication. LEDGF can stimulate the function of purified integrase *in vitro* (Cherepanov*, et al.*, 2003), but this is a somewhat permissive assay. While provocative, these data fail to establish a definitive role for LEDGF in integration.

RETROVIRAL INTEGRATION TARGETING

While most sequences tested *in vitro* can serve as a targets for integration (Bor *et al.*, 1996; Brown*, et al.*, 1987; Craigie*, et al.*, 1990), all retroviruses tested exhibit nonrandom selection of integration target sites in cells (Mitchell *et al.*, 2004; Schroder *et al.*, 2002; Wu *et al.*, 2003). Possible explanations for integration target site specificity include the variable accessibility of certain regions of chromosomal DNA or tethering of the PIC to genomic sites through its interaction with specific cellular DNA-binding proteins.

Target DNA sequence and structure preferred for integration. *In vitro* studies of integration target site selection with naked DNA found that retroviruses exhibit weak primary sequence preferences (Bor*, et al.*, 1996; Carteau *et al.*, 1998; Fitzgerald and Grandgenett, 1994; Goodarzi *et al.*, 1997; Pryciak and Varmus, 1992).

There is some evidence that proteins complexed with integrase might affect target selection at the primary sequence level, as the sequence preferences of purified HIV-1 integrase differ somewhat from those of PICs (Bor*, et al.*, 1996; Kitamura *et al.*, 1992). Genome-wide studies (considered below) have shown that different retroviruses have weak but distinguishable primary sequence preferences *in vivo* (Carteau*, et al.*, 1998; Holman and Coffin, 2005; Stevens and Griffith, 1996; Wu *et al.*, 2005), but they play only a minor role in integration target site selection.

Proteins bound to target DNA can influence integration positively or negatively. Steric hindrance prevents integration from occurring in chromosomal areas occupied by DNA-binding proteins as assayed *in vitro* (Bushman, 1994; Pryciak and Varmus, 1992) and observed *in vivo* (Maxfield *et al.*, 2005; Weidhaas *et al.*, 2000). However, not all protein-bound target DNA is unfavorable for integration. DNA assembled into nucleosomes, for instance, has been shown to be more favorable for integration than naked DNA (Pruss *et al.*, 1994a; Pruss *et al.*, 1994b; Pryciak and Varmus, 1992). Close examination of sites preferred in nucleosomal DNA indicates that the most severely bent regions of DNA on the nucleosomes are hotspots for integration (Pruss*, et al.*, 1994a; Pryciak *et al.*, 1992), suggesting that distortion of DNA itself facilitates integration. In fact, distortion of DNA in non-nucleosomal protein complexes has been shown to favor integration (Bor *et al.*, 1995; Muller and Varmus, 1994). Distortion of viral and target DNA is likely to be an essential step in the process of integration (Bushman and Craigie, 1992; Scottoline *et al.*, 1997) so targeting of DNA that is already distorted could facilitate the reaction.

Genome-wide studies of integration targeting. Large studies of integration site selection across the human genome have been performed for three retroviruses: HIV-1, MLV and ASLV. The integration target site selection preferences were shown to differ among these retroviruses. Transcription units were strongly favored targets of HIV-1 integration (Schroder*, et al.*, 2002) regardless of host cell type studied (Mitchell*, et al.*, 2004; Wu*, et al.*, 2003). MLV integrase favored transcription units to a lesser extent, but exhibited a strong bias for areas within five kilobases of transcription start sites, with twenty percent of integration sites found in these regions (Wu*, et al.*, 2003). No bias was found in the location of HIV integration sites within transcription units—that is, the frequency of integration was the same across the length of the transcription unit (Mitchell*, et al.*, 2004; Wu*, et al.*, 2003). ASLV shows the most random distribution of integration sites with only a weak preference for genes (Mitchell*, et al.*, 2004; Narezkina *et al.*, 2004).

To study the influence of host cell gene expression on integration targeting, microarrays were used for transcriptional profiling of the target cells. The median expression level of genes targeted for integration by HIV was found to be much higher than the median expression of all genes assayed, indicating that HIV has a preference for integration into active genes (Schroder*, et al.*, 2002). Studies of HIV integration site selection and gene transcription in two other human cell types revealed that in these cell types as well, transcription units were favored integration targets (Mitchell*, et al.*, 2004). The bias for active genes was tissue-specific in that genes targeted for integration in a specific host tissue were more likely to be highly expressed in that cell

type than in the others tested. MLV and ASLV were both shown to have a weak preference for active genes (Mitchell, *et al.*, 2004; Narezkina, *et al.*, 2004). Surprisingly, studies of ASLV integration into two genes in quail cells suggested that high level transcription disfavored integration (Maxfield, *et al.*, 2005; Weidhaas, *et al.*, 2000). Why the results of these studies differ from the genome-wide studies is unknown.

Several chromosomal features were found to influence integration site selection in the genome-wide studies. HIV integration is biased towards GC-rich regions and cytogenetic R bands (Elleder *et al.*, 2002; Mitchell, *et al.*, 2004; Schroder, *et al.*, 2002). This might be explained by HIV's preference for gene-rich regions of chromosomes, which are correlated with high gene expression, high GC content and R banding.

CpG islands are chromosomal regions enriched in the CpG dinucleotide corresponding to gene regulatory regions. Studies have found that CpG islands are favored integration targets of MLV (Laufs *et al.*, 2004; Wu, *et al.*, 2003). However, while CpG islands are found in regions of high gene density—regions that are favored for HIV-1 integration—CpG islands themselves are disfavored for HIV-1 integration (Mitchell, *et al.*, 2004).

Early studies of integration targeting suggested that MLV integration may be biased towards DNase I hypersensitive sites (Rohdewohld *et al.*, 1987; Vijaya *et al.*, 1986). This bias was suggested to be a consequence of favored integration into areas of open chromatin, which are more accessible to the integration machinery (Panet and

Cedar, 1977). Accessibility may also be the explanation for the opposite trend seen with heterochromatin. HIV-1 disfavors integration into alpha satellite DNA, a marker for centromeric heterochromatin (Carteau*, et al.*, 1998; Schroder*, et al.*, 2002). Centromeric heterochromatin is tightly packed and presumably less accessible to the retroviral PIC.

There appears to be a bias in the selection of whole chromosomes for HIV integration that cannot be entirely accounted for by the variations in gene density among the chromosomes (Laufs *et al.*, 2003; Mitchell*, et al.*, 2004). If found to be reproducible, this may point to additional factors involved in integration targeting, such as the intranuclear positions of chromosomes.

The initial genome-wide studies of integration targeting mentioned above were all done with human cells. Human cells were chosen because of their relevance to medicine and the feasibility of such studies following the completion of the human genome sequence. However, the biological relevance of studies in cell types that are not the natural hosts of the viruses studied is unclear. It is possible that cellular factors responsible for integration site selection are not well conserved among different species. One study that considered integration targeting in nonhuman cells (Hematti *et al.*, 2004) surveyed integration site selection by SIV- and MLV-based vectors in rhesus macaque hematopoietic stem cells. Hematti and colleagues found that SIV integration preferences are similar to those of HIV—with a strong bias towards transcription units. MLV targeting was the same as that in human cells—with a preference for regions near transcription start sites. A recent study of ASLV and HIV

integration site selection in chicken cells found that the targeting preferences were the same as observed in human cells (Barr *et al.*, 2005). These data suggest that there is conservation of the cellular determinants of integration site selection among vertebrates.

The most striking result from these genome-wide studies of integration targeting is that different retroviruses have distinct integration target site preferences. This suggests that virus-specific factors—not simply the accessibility of genomic targets—determine integration site selection.

Targeting of integration by tethering factors. All retroviruses studied thus far exhibit at least a weak preference for integration into transcription units, as discussed above. This could be explained, in part, by the accessibility of open chromatin to the PIC. However, accessibility alone cannot account for the distinct target site preferences of HIV-1, MLV and ASLV. Virus-specific factors likely play a role. An attractive model based on studies of the yeast retrotransposons is that the retroviral PICs interact with tethering factors bound to specific regions of host cell chromosomes that direct integration to nearby sites.

The yeast retrotransposons, such as the Ty elements, are very similar to retroviruses in genome organization and replication. The major difference is that, unlike retroviruses, retrotransposons lack *env* genes and thus do not have an extracellular stage in their replication cycle. Because they cannot produce progeny that leave the host cell, retrotransposons must avoid killing their host during replication. Replication without disruption of host cell transcription is particularly

difficult for Ty elements because their host—*Saccharomyces cerevisiae*—has a very gene-dense genome. Ty1, Ty3 and Ty5 have each developed their own strategy for targeting their integration to benign regions of the yeast genome (reviewed in (Boeke and Devine, 1998; Bushman, 2003; Sandmeyer, 2003)). Both Ty1 and Ty3 integrate upstream of Pol III-transcribed genes. Ty3 does this through integrase binding to the Pol III transcription complex and directing insertion of its DNA nearby (Kirchner *et al.*, 1995). Ty5 integrase targets telomeres or the silent mating loci by interacting with the heterochromatin protein Sir4p (Zhu *et al.*, 2003; Zhu *et al.*, 1999).

As with retrotransposons, tethering of the retroviral PIC to target DNA might play a role in retroviral integration site selection. *In vitro* studies with artificial tethering of retroviral integrases have confirmed the feasibility of such a mechanism for integration targeting. In these studies, integrase fusions to sequence-specific DNA-binding domains were able to direct site-specific integration *in vitro* (Bushman, 1994; Bushman and Miller, 1997; Goulaouic and Chow, 1996; Holmes-Son and Chow, 2000; Katz *et al.*, 1996; Tan *et al.*, 2004).

Several cellular factors are known to bind PICs and/or facilitate integration *in vitro*, suggesting they might influence targeting of retroviral integration to cellular chromosomes. They are BAF, HMGA1, Ini-1, Ku and LEDGF/p75, among others (Bushman, 2001; Bushman, 2003; Coffin*, et al.*, 1997; Engelman, 2005; Sandmeyer, 2003). Of these, an attractive candidate tethering factor for HIV-1 integrase is LEDGF/p75. LEDGF binds to HIV integrase and is found in HIV PICs but does not bind MLV integrase (Cherepanov*, et al.*, 2003; Llano*, et al.*, 2004b; Maertens*, et al.*,

2003). A recent study has found that HIV integration in genes, particularly LEDGF-responsive genes, is modestly but significantly reduced in LEDGF-knock down cells (A. Ciuffi, F. D. Bushman and colleagues, submitted). This suggests that LEDGF may act as one of several tethering factors for the HIV PIC.

CONSEQUENCES OF INTEGRATION INTO HOST CHROMOSOMES

The fates of the provirus and its host cell are intimately intertwined. The provirus can influence transcription of host genes in its vicinity and the chromosomal environment exerts its effects on proviral transcription. This reciprocal relationship is at least in part responsible for two phenomena associated with retroviral integration—insertional mutagenesis and viral latency.

Insertional mutagenesis. Defining the determinants of integration targeting has become topical recently due to setbacks faced in gene therapy trials using retroviral vectors. In these trials, two of nine children successfully treated for X-linked severe combined immunodeficiency (X-SCID) with an MLV-based vector delivering the *IL2RG* gene developed T cell leukemia (Hacein-Bey-Abina *et al.*, 2003a; Hacein-Bey-Abina *et al.*, 2003b). In both of these children, the leukemic cells harbored vector DNA integrated in or near the *LMO-2* gene—a candidate proto-oncogene—that resulted in an increase in *LMO-2* expression.

Retroviruses have long been implicated in tumorigenesis in animals (reviewed in (Coffin*, et al.*, 1997)). They can exert oncogenic effects by transducing an oncogene (v-*onc*) that they encode, as in oncogenesis by acute transforming viruses. As another means of oncogenesis, the retrovirus can integrate near and affect the

transcription of a cellular proto-oncogene or tumor suppressor. This phenomenon is called proviral insertional mutagenesis. In the case of promoter insertion, the provirus integrates in the same orientation upstream of a proto-oncogene, thus increasing its expression via the promoter and enhancer elements of one LTR. If inserted upstream in the opposite orientation of the gene or downstream in either orientation, the proviral enhancer sequences can boost expression of the proto-oncogene. A provirus can be inserted within a gene where it disrupts the formation of normal transcripts. This can result in proteins that lack regulatory domains, abolishing negative regulation, or it can increase mRNA stability. A retroviral insertion can also contribute to oncogenesis by inactivating one copy of a tumor suppressor gene. In order for cancer to result in this case, however, there must be inactivation of the other genomic copy of the tumor suppressor as well.

The propensity of retroviruses to cause cancer in model vertebrates by the above mechanisms has made them invaluable to the discovery of proto-oncogenes. Numerous proto-oncogenes have been identified through the sequencing of integration sites in endogenous tumors (Coffin*, et al.*, 1997). Also, high-throughput methods for proviral tagging of proto-oncogenes have recently been developed (Li *et al.*, 1999; Suzuki *et al.*, 2002).

For the adverse events in the X-SCID gene therapy trial, the *IL2RG*-transducing MLV-based vector integrated near the transcription start site of the *LMO-2* gene. In one case the vector was in the 5′ promoter region in the same orientation as *LMO-2* and in the other case it was in the first intron in the opposite orientation

(Hacein-Bey-Abina*, et al.*, 2003a; Hacein-Bey-Abina*, et al.*, 2003b). Insertions in *Lmo-2* and *Il2rg* have been associated with leukemia in mice by retroviral tagging (Dave *et al.*, 2004). It is likely that the growth-promoting effects of the *IL2RG* transgene and the increased expression of *LMO-2* acted synergistically in the development of leukemia in these children.

The retroviral vector employed in the X-SCID gene therapy trial was based on MLV. MLV-based vectors are widely used for gene transduction in animals and in human gene therapy. As discussed above, studies of MLV integration targeting have determined that MLV exhibits a preference for integration near transcription start sites (Wu*, et al.*, 2003). Considering the potential for activation of the host gene via promoter insertion or proviral enhancer activity and MLV's integration targeting preference for transcription start sites underscores the potential dangers of using MLV as a gene therapy vector. Its weak bias for integration in genes (Mitchell*, et al.*, 2004) might make ASLV a preferable vector for gene therapy applications. Further studies of integration targeting such as the one presented in Chapter Two are necessary in order to elucidate the determinants of site selection so that less toxic gene therapy vectors can be engineered.

Integration and viral latency. The primary obstacle to the eradication of HIV from the body, and thus, a cure for AIDS, is the existence of latent viral reservoirs. These are cells that harbor replication-competent but unexpressed virus that is invulnerable to antiretroviral therapy and immune surveillance. Upon cessation of drug therapy, the latent provirus can reseed the body with virus.

Several factors contribute to this phenomenon of latency, many of which are consequences of the life cycle of the primary host cells of HIV—$CD4^+$ T lymphocytes. $CD4^+$ T lymphocytes vary widely in how permissive they are for HIV replication depending on their activation status. For instance, HIV-1 infection of resting $CD4^+$ T cells does not result in integrated provirus due to a number of blocks in replication—a phenomenon called preintegration latency (Pierson *et al.*, 2002). Only if the resting host cell is activated before the viral PIC is degraded can productive infection of these cells occur.

Stable latent reservoirs are the result of postintegration latency (reviewed in (Lassen *et al.*, 2004)). Sometimes, an activated $CD4^+$ T cell is infected with HIV and then reverts to a quiescent memory phenotype. The integrated provirus is thought to remain unexpressed because the quiescent T cell lacks sufficient nuclear levels of the necessary transcription factors and its chromatin is condensed and inactive (Brown *et al.*, 1999; Lassen*, et al.*, 2004; Setterfield *et al.*, 1983).

Recent studies have considered additional factors determining postintegration latency *in vivo* (Finzi *et al.*, 1997; Wong *et al.*, 1997), including the proviral integration site (Han *et al.*, 2004). The site of integration in the genome has long been known to influence the expression of genes within proviruses. Nevertheless, the extent to which integration site plays a role in transcriptional repression of HIV proviruses has been hard to study *in vivo*. This is, in part, due to the scarcity of latently infected cells in infected individuals. Han and colleagues characterized 74 integration sites from T cells of patients on prolonged antiretroviral therapy and found

that the distribution of sites was similar to that of HIV in cell culture—active genes were favored targets (Han*, et al.*, 2004). However, because these proviruses were not sequenced or otherwise tested, it is impossible to know whether they were truly latent—that is, replication competent but silenced—or inactivated by mutation. In fact, only one percent of inactive HIV proviruses are thought to be authentically latent (Chun *et al.*, 1997a; Chun *et al.*, 1997b).

Because of the challenges of studying HIV latency *in vivo*, cell culture models of this phenomenon have been developed and studied. Jordan and colleagues developed a model system in which the human $CD4^+$ T cell line, Jurkat, was infected with an HIV-based vector expressing green fluorescent protein (GFP) from the viral promoter (Jordan *et al.*, 2001). Following infection, cells were sorted into GFP-expressing and non-expressing populations. Cells from the GFP-negative population were stimulated with a cytokine or mitogen to activate expression of the silenced proviruses. Those cells harboring inducible proviruses were analyzed. Initial studies with this model found that inducible proviruses were frequently found integrated into heterochromatic regions (Jordan *et al.*, 2003). The results of a genome-wide study using this model comparing integration sites of productively infected and latent proviruses are presented in Chapter Three.

SUMMARY

While much is known about the biochemistry of retroviral integration, we are only beginning to elucidate the varied and complex virus-host cell interactions at this essential step in the retroviral replication cycle. The following investigations of the

viral determinants of integration target site selection (Chapter Two) and the effects of genomic location on subsequent proviral gene expression (Chapter Three) provide insight into the host-virus relationship, with implications for human health.

A study of the viral determinants of integration target site selection involving the analysis of integration preferences of HIV-MLV chimeras, presented in Chapter Two, found that viral Gag and integrase proteins together determine the different targeting preferences of HIV-1 and MLV. These results enable us to refine models for the mechanism of integration targeting and bring us closer to identifying cellular factors that interact with the retroviral PIC to direct site-specific integration. These viral and cellular factors that mediate the process of integration *in vivo* are potential drug targets in the case of HIV. Further, this study was the first to demonstrate the transfer of integration targeting preferences of one virus to another through alteration of the viral genome. This holds promise for the gene therapy field, suggesting that safer retroviral vectors could be engineered by substituting *gag* and *integrase* gene fragments from retroviruses that prefer to integrate into more benign regions of the human genome.

A study comparing genomic features of integration sites from well-expressed and transcriptionally silenced proviruses, presented in Chapter Three, confirms that integration site does influence proviral expression and suggests that it can contribute to transcriptional silencing of HIV. These results advance our understanding of HIV latency, the primary obstacle to a cure for AIDS.

II. RETROVIRAL GAG AND INTEGRASE ACT SYNERGISTICALLY TO DETERMINE INTEGRATION TARGET SPECIFICITY

A. ABSTRACT

Retroviruses differ in their preferences for sites for viral DNA integration in the human genome. HIV integrates preferentially within active transcription units, whereas murine leukemia virus (MLV) integrates preferentially near transcription start sites and CpG islands. We have investigated the viral determinants of integration site selection using chimeric viruses with MLV genes substituted for their HIV counterparts. Chimeras containing MLV structural proteins (*gag*) or MLV integrase (*IN*) showed only slight differences compared to HIV. However, an HIV derivative with both MLV *gag* and *IN* (HIVmGagmIN) was fully switched to the MLV integration specificity. We found that MLV but not HIV targeted DNase I hypersensitive sites, and HIVmGagmIN also targeted these sites. Fourteen transcription factor binding motifs were enriched near MLV and HIVmGagmIN integration sites, specifying potential cellular factors mediating integration targeting. These findings disclose an unexpected function of Gag proteins and point to new models for retroviral integration targeting.

B. INTRODUCTION

The selection of target sites for integration of retroviral DNA is central to the biology of retroviruses and the application of retroviral vectors to gene therapy. Retroviral integration site selection is not strongly sequence-specific with respect to

target DNA (Carteau, *et al.*, 1998; Holman and Coffin, 2005; Stevens and Griffith, 1996; Wu, *et al.*, 2005), but integration *in vivo* shows pronounced favored and disfavored chromosomal regions. Early studies of MLV suggested that integration may be favored in open chromatin (Panet and Cedar, 1977), since a positive correlation was detected between integration frequency and DNase I hypersensitive sites (Rohdewohld, *et al.*, 1987; Vijaya, *et al.*, 1986). More recently, the completion of the draft human genome sequence has allowed systematic studies of integration targeting by high-throughput sequencing of integration acceptor sites (Mitchell, *et al.*, 2004; Narezkina, *et al.*, 2004; Schroder, *et al.*, 2002; Wu, *et al.*, 2003), revealing that integration site selection differs among retroviruses. HIV integration sites are found predominantly in active transcription units (Mitchell, *et al.*, 2004; Schroder, *et al.*, 2002). For MLV, in contrast, over 20% of integration events are near transcription start sites and associated CpG islands, while integration within transcription units is only slightly favored (Wu, *et al.*, 2003). ASLV shows the most random pattern of integration site selection, favoring transcription units only weakly and not favoring transcription start sites (Mitchell, *et al.*, 2004; Narezkina, *et al.*, 2004). Here we investigate the mechanisms dictating these different integration site preferences.

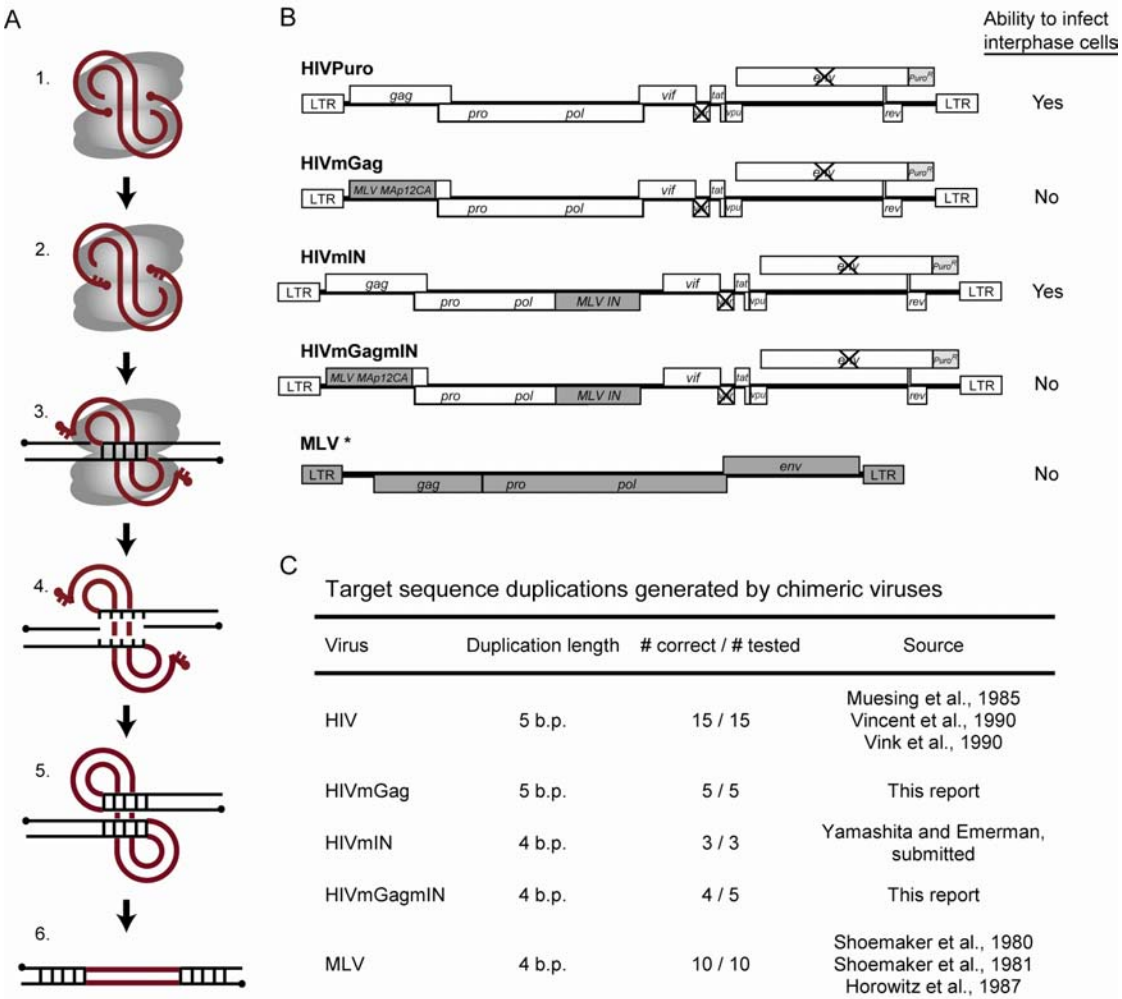
The DNA breaking and joining reactions that mediate retroviral integration are well worked out (Figure 1A). Prior to integration, two nucleotides are removed from each 3' end of the unintegrated linear viral DNA by the virus-encoded integrase (IN) protein, exposing recessed 3' hydroxyl groups (Brown, *et al.*, 1989; Fujiwara and Mizuuchi, 1988; Hughes, *et al.*, 1981; Roth, *et al.*, 1989; Sherman and Fyfe, 1990)

(See Figure 1A). IN then joins the recessed 3' hydroxyl groups to protruding 5' ends in the target DNA (Bushman*, et al.*, 1990; Craigie*, et al.*, 1990; Katz*, et al.*, 1990). Unpairing of the DNA between the points of joining results in formation of DNA gaps, which are then filled in and sealed, probably by host cell gap repair enzymes (Yoder and Bushman, 2000). A consequence of the gap repair step is the creation of a short duplication of the target site DNA at each host-virus DNA junction. The length of this duplication is characteristic for each virus—5 bp for HIV (Muesing *et al.*, 1985; Vincent*, et al.*, 1990; Vink *et al.*, 1990) and 4 bp for MLV (Horowitz *et al.*, 1987; Shoemaker *et al.*, 1980; Shoemaker *et al.*, 1981).

Here we investigate the requirements for integration targeting *in vivo* using chimeric viruses in which gene segments of MLV were substituted for the corresponding segments of the HIV genome (Figure 1B). The chimeras contained MLV *gag* gene segments substituted for HIV *gag* (HIVmGag) or MLV *IN* substituted for HIV *IN* (HIVmIN) (Yamashita and Emerman, 2004; Yamashita and Emerman, submitted). Guided by our initial studies, we constructed and analyzed an additional HIV-based virus containing both MLV *gag* and MLV *IN* (HIVmGagmIN). Previous characterization has shown that these viruses differ in their ability to infect interphase cells, and this property maps to the *gag* gene (Yamashita and Emerman, 2004; Yamashita and Emerman, submitted). That is, MLV integrates only after mitosis, while HIV can integrate any time during the cell cycle (Lewis *et al.*, 1992; Lewis and Emerman, 1994; Roe *et al.*, 1993; Weinberg *et al.*, 1991) although integration during mitosis appears to be disfavored (Katz *et al.*, 2003; Mannioui *et al.*, 2004). The

chimeric viruses HIVmGag and HIVmGagmIN have the same cell cycle requirements as MLV (Yamashita and Emerman, 2004; Yamashita and Emerman, submitted), while HIVmIN has the same cell cycle requirements as HIV (Yamashita and Emerman, submitted). Integration target site selection was assayed by cloning and sequencing 2582 junctions between human DNA and proviral DNA generated by infection of human cells with the chimeric and control viruses.

We found that HIVmGagmIN favored integration near transcription start sites and CpG islands, matching the preferences of MLV. In contrast, HIVmGag and HIVmIN exhibited much more modest differences in integration targeting compared to wild-type HIV. We used new genome-wide data on preferential DNase I cleavage sites (Crawford *et al.,* 2004; Crawford *et al.,* submitted) to analyze the relationship to integration, and found that MLV but not HIV favored integration near DNase I cleavage sites. Like MLV, the HIVmGagmIN virus favored integration near preferential DNase I cleavage sites as well. We also examined the association of transcription factor binding motifs with integration site sequences from each of the data sets and found fourteen motifs that were enriched near both MLV and HIVmGagmIN sites, thereby identifying possible cellular proteins guiding integration by MLV and HIVmGagmIN. In contrast, no single motif was common among HIV, HIVmGag and HIVmIN. These data indicate that Gag and IN work synergistically to direct integration site selection, and suggest models where either i) the cell cycle entry point specified by Gag and tethering through IN direct target site selection, or else ii) both Gag and IN bind co-operatively to tethering factors that guide integration.



Virus	Duplication length	# correct / # tested	Source
HIV	5 b.p.	15 / 15	Muesing et al., 1985 Vincent et al., 1990 Vink et al., 1990
HIVmGag	5 b.p.	5 / 5	This report
HIVmIN	4 b.p.	3 / 3	Yamashita and Emerman, submitted
HIVmGagmIN	4 b.p.	4 / 5	This report
MLV	4 b.p.	10 / 10	Shoemaker et al., 1980 Shoemaker et al., 1981 Horowitz et al., 1987

Figure 1: Retroviral DNA integration and the chimeric viruses used in this study. A) The DNA breaking and joining reactions mediating integration. Gray ovals represent IN monomers, thick red lines are viral DNA, black lines are target DNA, and dots are 5' ends. (1) Linear blunt-ended viral cDNA is bound by IN (gray) as part of the preintegration complex. (2) IN removes two nucleotides from the 3' ends of the viral DNA, exposing recessed 3' hydroxyl groups. (3) IN joins the recessed 3' ends of viral DNA to the target DNA. (4) Unpairing of the target DNA between the joined ends of the viral DNA yields gaps in the target DNA. (5) DNA repair enzymes fill in the gaps. (6) The provirus is flanked by repeated segments of the target DNA. B) Chimeric HIV derivatives containing segments of MLV. At the top is the HIV parent virus, with *vpr* and *env* inactivated and the puromycin resistance gene in place of *nef*. Following that are the chimeras, with substitutions of MLV *gag* gene segments (*MA*, *p12* and *CA*-coding regions) for HIV *MA* and *CA* or substitution of MLV *IN* for HIV *IN*, or both. *The MLV genome is shown for comparison. The MLV used in this study (MLVPuro) was an MLV-based vector (LPCX) encoding the puromycin resistance gene with Gag, Pol and amphotropic Env provided in trans. Construction and characterization of these viruses and chimeras, including an analysis of their ability to infect interphase cells, are described in (Yamashita and Emerman, 2004; and Yamashita and Emerman, submitted). C) Target sequence duplication lengths made by HIV, MLV and the chimeric viruses.

C. RESULTS

Cloning and analysis of integration sites. The chimeric viruses used in this study were deleted for the *env* gene and complemented with the envelope of vesicular stomatitis virus (VSV-G) to boost titer and restrict infection to a single round. These chimeras were less infectious than the wild-type virus (Yamashita and Emerman, 2004; Yamashita and Emerman, submitted), so the puromycin resistance gene was cloned in place of *nef* and infected cells were selected with puromycin to enrich for provirus-containing cells. (Some effects of Puromycin selection on integration site recovery are examined in Appendix 1.) *Vpr* was also deleted because of its cellular toxicity (Rogel *et al.*, 1995). In order to control for possible biases in integration site recovery due to puromycin selection, control infections were carried out with an HIV derivative transducing the puromycin resistance gene (termed "HIVPuro") and an MLV vector (LPCX) also transducing the puromycin resistance gene (termed "MLVPuro"). HeLa cells were chosen as infection target cells because they are highly susceptible to infection and they were used in a previous study comparing MLV and HIV integration targeting (Wu*, et al.*, 2003).

To clone integration sites, genomic DNA from infected cells was extracted, digested with MseI and ligated to adapters. The junctions between proviral DNA and genomic DNA were amplified by nested PCR using primers complementary to proviral and adapter sequences, cloned, sequenced, and mapped to the human genome as described (Lewinski *et al.*, 2005; Mitchell*, et al.*, 2004; Schroder*, et al.*, 2002; Wu*, et al.*, 2003). Newly determined sets of integration sites (a total of 2582 sites for the

Table 1: Integration site data sets used in this study

Data Set	Cell Type	No. of integration sites	Source
HIVPuro	HeLa	525	This report
HIVmGag	HeLa	493	This report
HIVmIN	HeLa	494	This report
HIVmGagmIN	HeLa	526	This report
MLVPuro	HeLa	544	This report
MLV-Burgess	HeLa	917	Wu et al., 2003
HIV-pooled	various*	2055	Carteau et al., 1998 Schroder et al., 2002 Wu et al., 2003 Mitchell et al., 2004
ASLV	293T-TVA, HeLa	834	Mitchell et al., 2004 Narezkina et al., 2004
L1 LINE	HeLa	127	Gilbert et al., 2002 Symer et al., 2002

* SupT1, HeLa, H9, IMR-90, PBMC

five viruses studied) were compared to each other and to previously reported data sets (Table 1). The distribution of integration sites was also compared to random sites in the human genome generated computationally.

As a test for correct integration by the chimeric viruses, we determined the target site duplication lengths for a few integration events of each (Figure 1C). Each chimeric virus showed mostly the duplication length characteristic of the virus donating the *IN* segment, which is as expected because IN is known to dictate the length of the duplication (Bushman*, et al.*, 1990; Craigie*, et al.*, 1990; Katz*, et al.*, 1990). For unknown reasons one duplication out of five for the HIVmGagmIN chimera was 5 bp instead of the expected 4 bp; all others were as expected. In addition, all integration events showed evidence of correct cleavage at the viral DNA 3' end by IN. These data support the idea that the IN-DNA complexes of the chimeras generally assembled and functioned normally.

Integration frequency near transcription start sites and CpG islands. Approximately 500 unique sequences for each of the five viruses were mapped to the human genome and nearby features were assessed (Figure 2A). Figure 2B shows the distribution of integration sites in three selected chromosomal regions.

We first evaluated the frequency of integration near transcription start sites and CpG islands (Figure 3A and B and Table 2). The MLVPuro control exhibited a strong preference for integration near transcription start sites—26.1% of MLVPuro sites were within plus or minus 5kb of a RefSeq gene transcription start site compared to 5.6% of random control sites. For the HIVPuro virus, 6.9% were near transcription start sites,

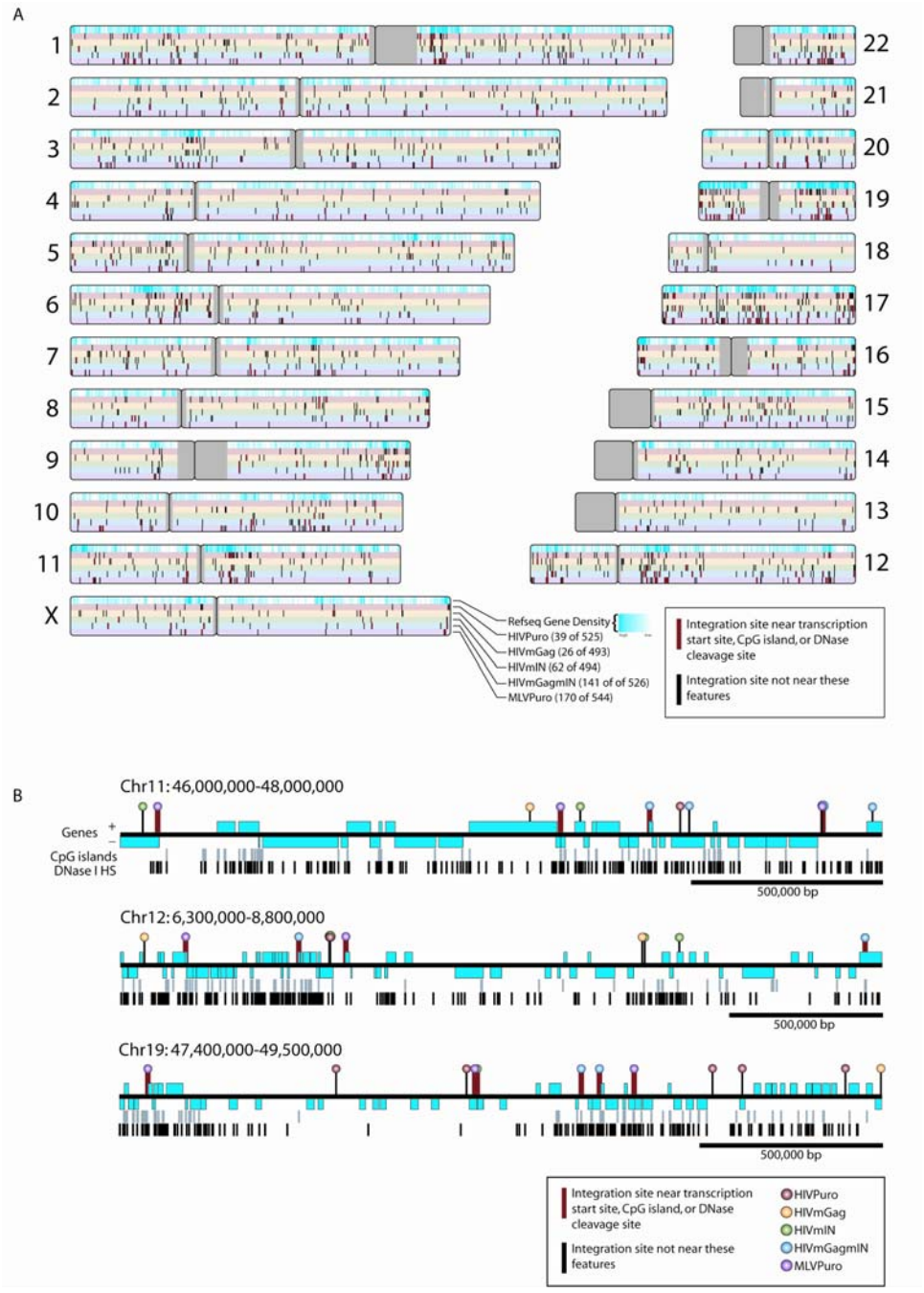


Figure 2: Sites of retroviral integration in the human genome. A) Positions of integration sites on the human chromosomes. The human chromosomes are shown numbered. Centromeric regions (which are mostly unsequenced) are shown in gray. Relative gene density is indicated in the top bar on each chromosome by the intensity of the cyan coloration. Integration site data sets (lower bars) are color coded as indicated. Sites of integration near transcription start sites, CpG islands, or multiple DNase I cleavage sites are shown as red dashes (the number of these in each data set is indicated in parentheses), other sites are black. B) Close-up view of selected chromosomal regions. See the figure for legend.

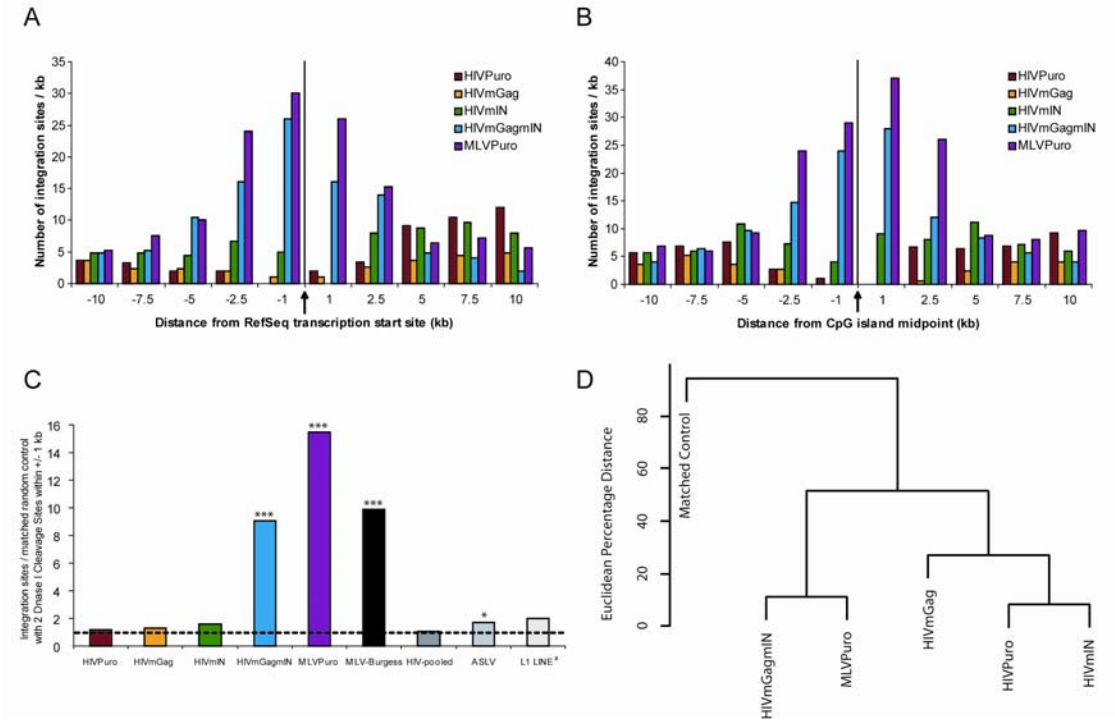


Figure 3: Frequency of integration near genomic features, and clustering based on these results. Features examined included A) transcription start sites and B) CpG islands. The number of integration sites within each interval was divided by the number of kilobases of that interval yielding the number of sites per kb. C) Integration near DNase I hypersensitive sites. For each data set the proportion of integration sites found within 1 kb of two DNase I hypersensitive sites was divided by the proportion in the matched random control set. The dotted line represents the expected bar height if the observed data did not differ from random. # L1 was analyzed with respect to an unmatched random set. *** P-value < 0.0001 by Chi-square comparison to random. * 0.05 > P-value >0.01. D) Clustering of integration site data sets using a machine learning algorithm. One hundred and nine types of genomic features were used to categorize the data sets. See Appendix 3 for details.

Table 2: Integration near genomic features

	Percentage of integration sites (P-values for Chi-square comparison to random)					
	Human Genome (random sites)	HIVPuro	HIVmGag	HIVmIN	HIVmGagmIN	MLVPuro
Within ± 5kb of a RefSeq transcription start site	5.6%	6.9% (0.2017)	3.9% (0.1013)	10.9% (<0.0001)	22.4% (<0.0001)	26.1% (<0.0001)
Within ± 1kb of a CpG island midpoint	1.7%	0.2% (0.0081)	0.0% (0.0038)	2.2% (0.3471)	9.9% (<0.0001)	11.8% (<0.0001)
Within RefSeq genes	32.2%	77.9% (<0.0001)	66.7% (<0.0001)	71.5% (<0.0001)	42.4% (<0.0001)	44.3% (<0.0001)
With 2 DNase I hypersensitive sites in a window ± 1kb	1.2%	1.0% (0.6327)	1.6% (0.3706)	1.4% (0.6307)	8.9% (<0.0001)	11.4% (<0.0001)
Within 500 bp of a multispecies conserved sequence midpoint	28.9%	36.4% (0.0003)	39.4% (<0.0001)	35.6% (0.0016)	43.5% (<0.0001)	46.3% (<0.0001)
Within 500 bp of a MCS midpoint in intergenic regions*	24.2%	24.1% (0.9869)	29.3% (0.1400)	29.8% (0.1301)	42.6% (<0.0001)	41.9% (<0.0001)

* Defined as integration sites outside of RefSeq genes.

which was not significantly greater than random. Thus the preferential integration near transcription start sites by MLV but not HIV reported previously (Schroder*, et al.*, 2002; Wu*, et al.*, 2003) was reproduced here.

The HIVmIN and HIVmGag chimeras differed from MLV, exhibiting 3.9% (HIVmGag) or 10.9% (HIVmIN) of integration events near transcription start sites. In this and other features, the HIVmIN chimera did display a somewhat more MLV-like pattern of integration site selection than HIV or the HIVmGag chimera, suggesting that MLV IN may be in part responsible for MLV integration targeting.

However, the doubly substituted HIVmGagmIN chimera integrated with high frequency near transcription start sites (22.4% of sites), and was indistinguishable from MLV. Thus both the determinants in *gag* and *IN* were required to transfer the preference for integration near transcription start sites from MLV to HIV.

The integration frequency near CpG islands was then compared. CpG islands are regions rich in the CpG dinucleotide that are undermethylated and frequently associated with gene regulatory regions (Bird, 1986; Larsen *et al.*, 1992). MLV favors integration near CpG islands while HIV disfavors these sites (Mitchell*, et al.*, 2004; Wu*, et al.*, 2003). We quantified integration frequency near CpG islands and found that both the MLVPuro and HIVmGagmIN viruses favored integration near these sites—11.8% and 9.9% of sites, respectively, were within 1 kb of a CpG island midpoint, compared to 1.7% of random sites. HIVPuro and HIVmGag viruses significantly disfavored regions within 1 kb of a CpG island midpoint (0.2% and 0%, respectively). The HIVmIN chimera, which had 2.2% of sites within 1 kb of a CpG

island midpoint, did not differ significantly from random but did favor these sequences to a greater degree than the HIVPuro virus (p-value = 0.0026 by Chi-square).

In summary, the HIVmGagmIN chimera resembled MLV in its strong preference for integration near transcription start sites and CpG islands. Neither MLV *gag* nor *IN* alone could transform targeting of HIV chimeras to the MLV pattern. The HIVmIN chimera exhibited an intermediate preference for integration near transcription start sites and CpG islands, suggesting that MLV IN does play some role in targeting to these regions.

Another difference between HIV and MLV is the different frequency of integration within transcription units. The HIVPuro virus favored integration in these sequences (77.9% in RefSeq genes), while the MLVPuro virus showed a much weaker trend (44.3% in RefSeq genes), which is only slightly above the frequency for random sites (32.2%). The double chimera HIVmGagmIN did not differ significantly from the MLVPuro virus, again indicating the similarity between the two. The HIVmGag and HIVmIN chimeras showed intermediate phenotypes, being down 11% and 6%, respectively, in the frequency of targeting transcription units compared to HIVPuro, but still significantly greater than the MLVPuro or HIVmGagmIN viruses. Thus analysis of integration in transcription units also indicated that both MLV *gag* and *IN* were needed for MLV-like specificity, while also indicating that transfer of either MLV *gag* and *IN* alone had modest but discernable effects.

Integration frequency near favored DNase I cleavage sites in chromatin. Early studies of MLV integration targeting suggested that MLV favors DNase I

hypersensitive sites for integration (Panet and Cedar, 1977; Rohdewohld*, et al.*, 1987; Vijaya*, et al.*, 1986). DNase I hypersensitive sites are believed to be nucleosome-depleted chromosomal regions associated with regulatory elements (Gross and Garrard, 1988). Genome-wide mapping of DNase I cleavage sites in chromatin has revealed that they are enriched near the boundaries of transcription units and near CpG islands, reinforcing the idea that they are markers for regulatory regions (Crawford *et al.*, 2004).

To assess the correlation between retroviral integration and DNase I cleavage frequency genome-wide, we quantified integration sites within 1 kb of two DNase I cleavage sites. We chose to use two cleavage sites in the analysis instead of a single site to better match the experimental definition of DNase I hypersensitive sites, which relies on multiple cleavage events. The conclusions were similar whether one, two, or three DNase sites were used for the analysis (similarly, the segment lengths used for comparison did not strongly affect the conclusions (data not shown)). For technical reasons, Crawford *et al.* analyzed cleavage sites in resting T cells, but many DNase I sites are expected to be present in cells of from diverse tissues (Sabo *et al.*, 2004), so we have extrapolated these data to the HeLa cells used in our study.

Figure 3C shows the proportion of integration sites that were in intervals (plus or minus one kb of the integration sites) containing two or more DNase I cleavage sites compared to random controls. The percentages are listed in Table 2. We also analyzed previously published data sets from MLV (Wu*, et al.*, 2003), HIV (Carteau*, et al.*, 1998; Mitchell*, et al.*, 2004; Schroder*, et al.*, 2002; Wu*, et al.*, 2003), ASLV

(Mitchell, *et al.*, 2004; Narezkina, *et al.*, 2004), and the L1 retrotransposon (Gilbert *et al.*, 2002; Symer *et al.*, 2002) and plotted these in Figure 3C for comparison.

Of all the elements analyzed, MLV showed by far the strongest preference for integration near DNase I cleavage sites. HIV and L1 elements showed no preference for integration near DNase I sites, while ASLV showed a weak preference that barely achieved statistical significance. Thus, contrary to the expectation that open chromatin at DNase I cleavage sites is globally favorable for integration, we find that favored integration near these sites is specific to MLV.

The double chimera HIVmGagmIN was similar to the MLVPuro virus in that it strongly favored DNase I hypersensitive sites for integration. Like the HIVPuro virus, the HIVmGag and HIVmIN chimeras did not favor these sites for integration above random. Thus substituting both MLV *gag* and *IN* into HIV was required to transfer the tendency to favor integration near DNase I cleavage sites.

Integration frequency near multispecies conserved sequences. We also investigated the relationship between retroviral integration sites and multispecies conserved sequences (MCS), which are defined as genomic regions that have been highly conserved among diverse vertebrates (Siepel *et al.*, 2005). Although the role of many of these sequences is unclear, at least some appear to be conserved regulatory elements and others conserved exons. HIVPuro, MLVPuro and the chimeric viruses each exhibited a modest preference for integration into the MCSs (Table 2). Because MCSs are in part exons, this tendency can be partially attributed to favored integration in transcription units.

When MCSs within and outside of genes were considered separately, however, differences in integration preferences were observed. The most striking result was that the MLVPuro and HIVmGagmIN viruses exhibited clear preferences for integration near intergenic MCSs, while the HIVPuro, HIVmGag, and HIVmIN viruses had no preference for these regions. Although the nature of MCSs is not fully clarified, these findings do provide another indication of the parallels between integration by the MLVPuro and HIVmGagmIN viruses.

Integration frequency and transcriptional activity. We next assessed the effects of transcriptional activity on integration frequency using transcriptional profiling data for the HeLa target cells. All viruses tested favored active transcription units for integration (Figure 4A). The median expression level of genes targeted for integration was highest for the HIVPuro, HIVmIN, and HIVmGag viruses, followed by the MLVPuro and HIVmGagmIN viruses. All were higher than randomly selected transcription units.

Figure 4B shows the frequency of integration for each virus in genes which have been classified by their expression levels. All viruses differed significantly from random in their distribution across expression-level bins (p-value < 0.0001 by Chi-square). The MLVPuro and HIVmGagmIN data sets showed slightly weaker trends than the other data sets. Thus the MLVPuro and HIVmGagmIN viruses were similar by this measure as well.

Global comparison of trends in integration targeting. To assess the similarities among integration site data sets, a machine learning algorithm based on

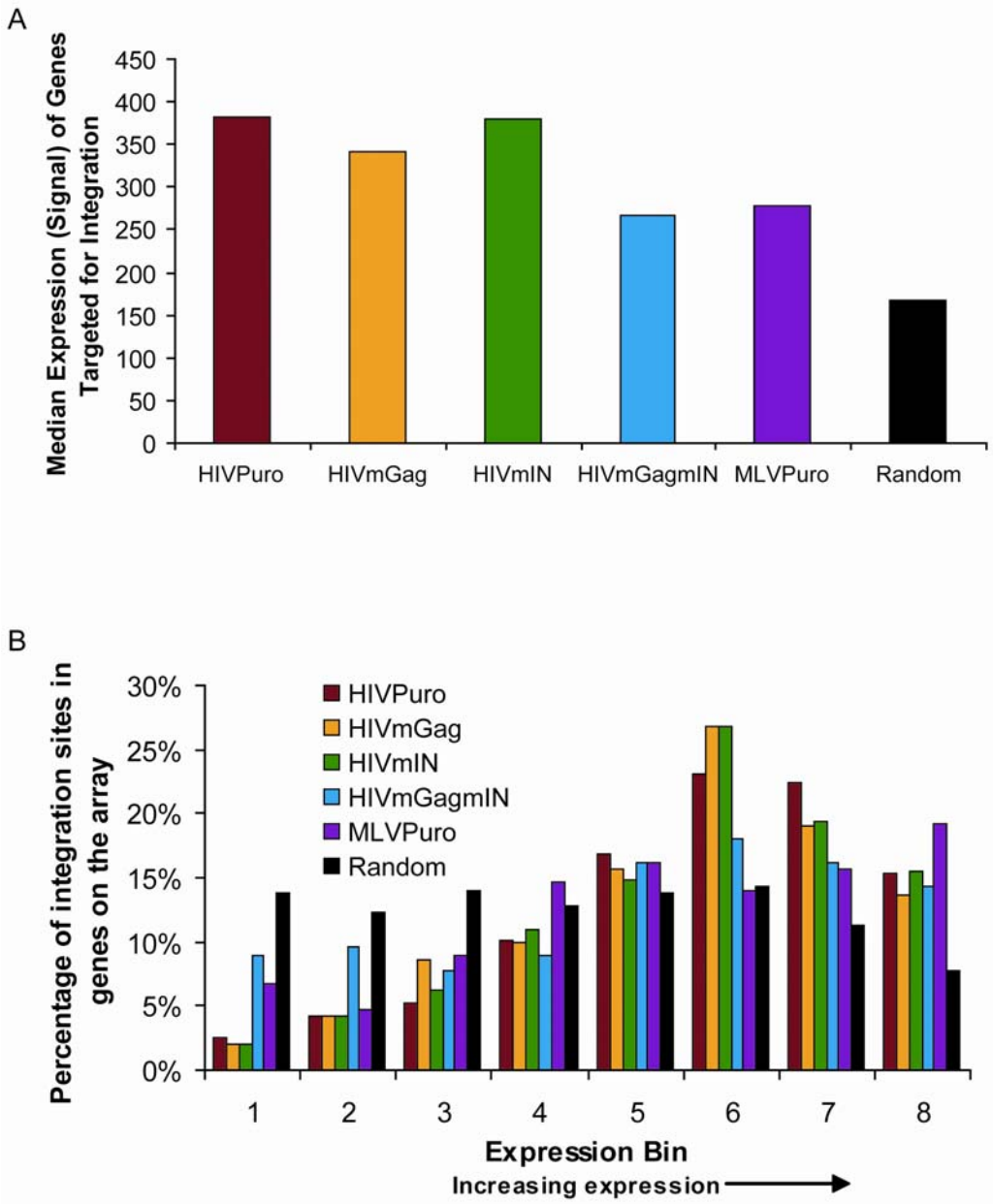


Figure 4: Effects of transcriptional activity on integration. A) Median expression levels of genes targeted for integration by the different viruses. The units are "signal" as defined by Affymetrix Microarray Suite 5.0 software. B) Frequency of integration in transcription units as a function of the level of expression. To classify the expression level of transcription units targeted for integration we used HeLa cell transcriptional profiling data assayed with Affymetrix HG-U133A microarrays. Probes on the array were ranked by expression level and divided into eight expression bins of equal size, with the 1/8 lowest expressing genes in bin 1 and the 1/8 highest expressing genes in bin 8. Integration sites in genes were distributed in the appropriate bins by expression level, summed, and expressed as a percentage of the total.

RandomForest was developed to cluster the data sets, taking into account 109 different types of genomic features (Figure 3D and Appendix 3). The MLVPuro and HIVmGagmIN integration site data sets were clustered together by this means, as were the HIVPuro and HIVmIN data sets. The HIVmGag data set was the most distinct, though it was closer to the HIVPuro and HIVmIN data sets than to MLVPuro and HIVmGagmIN. An analysis of targeting in the HIVmGag data set indicated that it showed much less preference for integration in gene rich regions than did HIVPuro or HIVmIN, largely accounting for the difference (data not shown).

Sequence motifs at integration sites. To investigate possible cellular factors directing integration site selection, we asked whether any known transcription factor binding motifs were significantly enriched in genomic sequences near integration sites. It has not so far been possible to associate binding sites for specific cellular proteins with integration sites, but if cellular sequence-specific DNA-binding proteins tether integration complexes to favored sites, then such interactions might be detectable in large data sets.

We evaluated possible enrichment of 347 transcription factor binding motifs from the TRANSFAC databases within plus or minus 1 kb of integration sites compared to 5000 randomly chosen 2 kb intergenic regions. Also included in this study is a previously published set of MLV integration sites in HeLa cells (termed MLV-Burgess; (Wu*, et al.*, 2003)). The MLVPuro, MLV-Burgess, and HIVmGagmIN data sets showed by far the highest numbers of significantly enriched binding site motifs (42, 35, and 23, respectively). The HIVPuro, HIVmGag, and

HIVmIN data sets returned far fewer (3, 1, and 2). Strikingly, for the MLV group of motifs, many were common to all three data sets, or shared between two of the three (Figure 5). Elimination of overlapping motifs from the raw data yielded 14 significantly enriched factors common to all three, thus specifying a set of cellular factors that may guide MLV (and HIVmGagmIN) integration. Varying the parameters used in the bioinformatic analysis showed that repeating the analysis under more permissive conditions returned even larger numbers of significantly enriched motifs (data not shown). No single motif was common to the HIVPuro, HIVmGag, and HIVmIN data sets taken together.

The location of MLVPuro, MLV-Burgess, and HIVmGagmIN integration sites could then be compared to the positions of enriched transcription factor binding motifs. The peak frequency of enriched motifs was not at the point of integration, but offset by at least 200 bp (p-value = e^{-21}). Thus any favorable interactions between MLV integration complexes and these transcription factors must extend over this distance along the integration target DNA.

D. DISCUSSION

We report a study of integration target site selection by hybrid viruses containing segments of MLV substituted for their HIV counterparts. Surprisingly, we found that it was necessary to transfer both MLV *gag* and *IN* to HIV to confer the MLV integration target site preferences on a chimeric virus. These data reveal a new function for retroviral Gag proteins in integration targeting and suggest that the

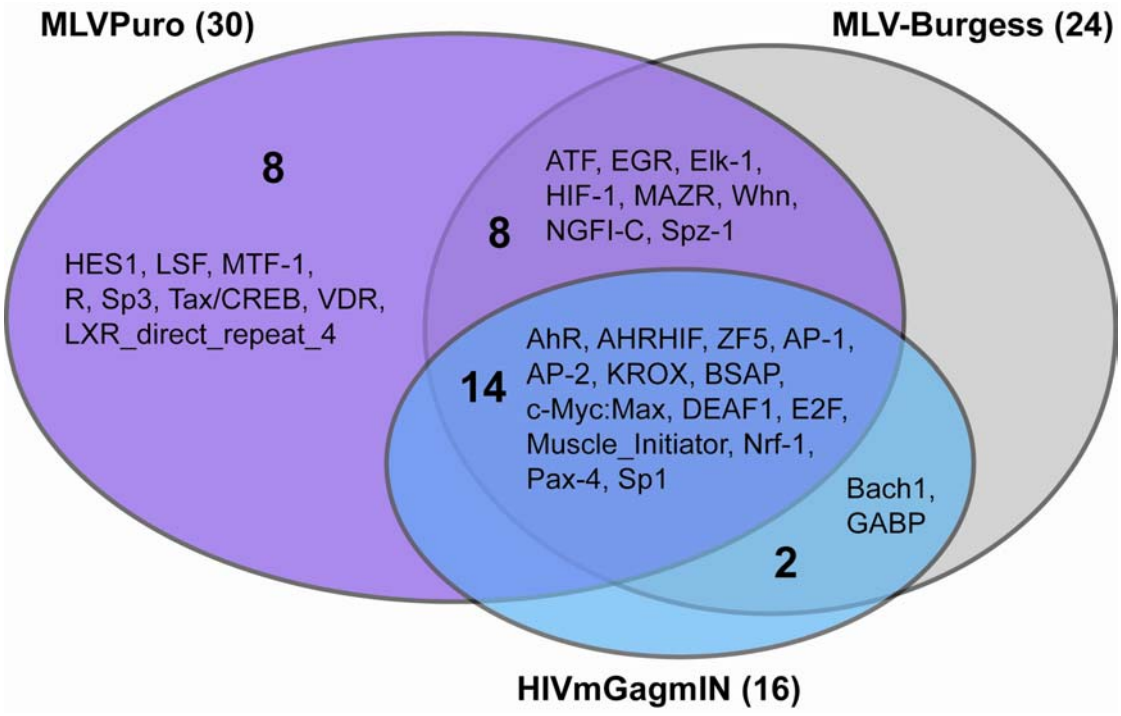


Figure 5: Diagram of the relationship of transcription factor binding sites enriched in the MLVPuro, MLV-Burgess, and HIVmGagmIN integration site data sets. The genomic sequences within one kilobase of each integration site were compared to 5000 randomly selected 2 kb intergenic regions. The indicated sequences were enriched greater than or equal to 1.65-fold. All comparisons achieved p-value of less than or equal to 0.001.

simplest versions of several previous models for integration site selection are unlikely to be correct.

Integration and open chromatin. The early observations that MLV favored integration near DNase I hypersensitive sites (Panet and Cedar, 1977; Rohdewohld*, et al.*, 1987; Vijaya*, et al.*, 1986) led to the proposal that open chromatin was generally favorable for retroviral DNA integration. However, our data indicate that DNase I sensitive regions are not universally favorable—only MLV, and not HIV, ASLV or L1, strongly favored integration near these sites. Analysis of those MLV integration sites found near DNase I cleavage sites revealed that they show a strong tendency to be near transcription start sites, CpG islands, and clustered transcription factor binding sites (data not shown). These data support a model in which the presence of DNase I cleavage sites is a marker for binding of specific cellular proteins, probably associated with gene control regions. It is unclear whether relatively greater exposure of DNA at these sites is involved at all—binding of integration complexes to specific factors at these sites may fully explain the observations.

Other measurements from the genome-wide data are consistent with a role for accessibility in integration targeting, but here too other explanations are possible. Integration of all the elements studied is favored at least weakly in transcription units (Mitchell*, et al.*, 2004; Narezkina*, et al.*, 2004; Schroder*, et al.*, 2002; Wu*, et al.*, 2003), consistent with greater accessibility of these sequences, but it is also possible that transcription units have specifically bound proteins that account for favored integration. Centromeres are disfavored integration targets (Carteau*, et al.*, 1998;

Schroder, *et al.*, 2002), and this could be because they are tightly wrapped in centromere-specific proteins and therefore inaccessible, but other candidate mechanisms include a lack of positive-acting cellular factors at centromeres or unfavorable intranuclear positions of centromeres. At present, none of the available data strictly requires models based on DNA accessibility to explain integration targeting.

Integration targeting via tethering. Another mechanism for directing integration to specific locations invokes interactions between integration complexes and cellular proteins bound at favored sites. Such a model has been strongly supported for the retrovirus-related Ty elements in yeast, where interactions between Ty integrase proteins and cellular DNA-binding proteins appear to account for selective integration targeting (Boeke and Devine, 1998; Bushman, 2003; Sandmeyer, 2003; Zhu, *et al.*, 1999). For retroviral INs, model *in vitro* studies have confirmed that tethering integration complexes to target DNA artificially can result in selective integration nearby (Bushman, 1994; Bushman and Miller, 1997; Goulaouic and Chow, 1996; Holmes-Son and Chow, 2000; Katz, *et al.*, 1996; Tan, *et al.*, 2004). A simple model for retroviral integration targeting invokes tethering interactions between chromatin-associated cellular factors and IN proteins. Different retroviral INs would interact with different DNA-bound factors, accounting for the differences in target site selection among the retroviruses.

However, the integration preferences of the HIV derivative containing MLV *IN* (HIVmIN) were closer to HIV than to MLV. This argues against a determinant in

IN serving as the sole mediator of integration specificity. The HIVmIN chimera did show slightly increased frequencies of integration near transcription start sites and CpG islands, like MLV, but did not show elevated frequency near MLV-favored transcription factor binding motifs, DNase I cleavage sites or intergenic MCS sequences. Thus the evidence suggests that IN influences integration targeting, but binding of IN alone to cellular factors is not sufficient to explain retroviral targeting preferences. Similarly, tethering through Gag proteins cannot explain the data, because transfer of MLV *gag* alone to HIV (to make the HIVmGag chimera) did not confer the MLV targeting phenotype.

For the case of HIV, the cellular LEDGF protein is a candidate HIV tethering factor, since this protein has been found to bind tightly to HIV IN but not to MLV IN (Cherepanov*, et al.*, 2003; Llano*, et al.*, 2004a; Llano*, et al.*, 2004b; Maertens*, et al.*, 2003). LEDGF has been suggested to be a component of transcription complexes, which could distribute the protein across transcription units, which are the favored targets for HIV integration (Ge*, et al.*, 1998). When LEDGF was depleted from cells (Llano*, et al.*, 2004b), the frequency of HIV integration in transcription units was diminished (A Ciuffi, M. Llano, E. Poeschla, P. S., J. L., C. B., J. E., and F. D. B, submitted). However, a tethering interaction between LEDGF and HIV IN is not the full explanation for HIV integration targeting, because 1) the LEDGF knockdown did not fully eliminate favored HIV integration in transcription units, 2) substituting MLV IN for HIV IN (to make HIVmIN) reduced integration in transcription units only

modestly, and 3) comparison of the HIVmIN chimera to HIVmGagmIN indicated that HIV Gag plays a role as well.

Effects of cell cycle on integration targeting. HIV and MLV differ in the cell cycle dependence of infection, which also has the potential to influence integration targeting. HIV can infect cells regardless of cell cycle phase (Lewis*, et al.*, 1992; Weinberg*, et al.*, 1991) while MLV infection requires host cells to pass through mitosis (Lewis and Emerman, 1994; Roe*, et al.*, 1993). The transcriptional state of a cell is known to vary with the cell cycle, so the organization of chromosomal DNA encountered by the MLV and HIV integration complexes should differ. The HIVmGag chimera exhibited cell cycle-restricted infectivity, like that of MLV (Yamashita and Emerman, 2004)—thus HIVmGag would encounter the chromosomal DNA in the same state as would MLV. However, the targeting preferences of the HIVmGag chimera, while different from those of HIV, do not resemble the integration site selection preferences of MLV. Thus cell cycle-associated changes in chromatin structure, combined with the differential cell cycle dependence of HIV and MLV infection, cannot fully account for the different integration site preferences.

Combined models for the mechanism of integration target site selection. Models for the mechanism of integration targeting must take into account the involvement of both IN and Gag proteins. One simple possibility is that IN and Gag both act as required tethering factors by binding to cellular proteins, though to explain the data, the tethering interaction must be strongly dependent on simultaneous binding by both IN and Gag. Another possibility would combine the role of Gag in specifying

the cell-cycle stage of infection together with tethering interactions through IN. According to this idea, it might be necessary for an integration complex to enter the nucleus at the proper stage of the cell cycle for IN to have the opportunity to encounter its binding partner(s) on chromosomes. This idea is particularly attractive to explain the MLV integration preference, since MLV enters the nucleus at a restricted point in the cell cycle. As one test of this idea, it should be possible to map the MLV Gag determinants of targeting within HIVmGagmIN, which is of interest because the p12-CA portion is known confer the dependence of infection on mitosis (Yamashita and Emerman, 2004).

Cellular factors directing MLV integration. The identification of enriched sequence motifs at integration sites of MLV and HIVmGagmIN allows more specific models of MLV integration to be proposed. Transcription factors that bind to these motifs are strong candidates for tethering MLV integration complexes near favored sites, perhaps via contacts with MLV IN and/or Gag. However, the bioinformatic analysis indicated that fully 14 binding motifs were enriched near MLV or HIVmGagmIN integration sites, and relaxing the criteria used in the analysis returned even more enriched motifs (unpublished data). Thus it appears unlikely that one or a few transcription factors bound to these motifs are solely responsible for targeting. One possibility is that there are many surfaces in MLV integration complexes that bind transcription factors, possibly involving both IN and Gag, with different surfaces docking with different transcription factors. Another possibility is that MLV integration complexes do not bind directly to these transcription factors, but rather to

additional proteins recruited by them, such as transcriptional mediator proteins or basal transcription factors. One or a few such proteins might be recruited by many different transcription factors, explaining how so many transcription factor binding sites could be associated with integration sites. The transcription factor binding motifs were mostly present at a distance of at least 200 bp from the site of MLV integration, consistent with the idea that large multi-protein transcription complexes bind across the intervening region. Thus all of the genomic features correlating with MLV integration (transcription start sites, CpG islands, DNase I cleavage sites, MCSs, and enriched transcription factor binding sites) may be markers for a class of multi-protein transcription complexes. Further experiments will be needed to determine the composition of these potential complexes and the specific protein-protein interactions mediating favored MLV integration at these sites.

E. EXPERIMENTAL PROCEDURES

DNA constructions. To generate the MLVPuro data set, we used LPCX (Clontech), which is an MLV-based vector that expresses the puromycin resistance gene from the MLV LTR. All other vectors used were based on the full-length HIV clone pLAI (Peden *et al.*, 1991). *Vpr* has been mutated by the insertion of 4 bases at the NcoI site at 5207 and *env* has a deletion between the BglII sites at 6634 and 7214 (Rogel*, et al.*, 1995). The puromycin resistance gene was cloned in place of *nef*. The MLV *gag* gene segment encoding MA, p12 and CA from pAMS (Miller *et al.*, 1985) was cloned in place of HIV MA and CA for MHIV-mMA12CA-ΔenvΔvprΔnef-puromycin^{R} (for the HIVmGag data set) and MHIV-mMA12CA-mIN-ΔenvΔvprΔnef-

puromycin^R (for HIVmGagmIN) as described previously (Yamashita and Emerman, 2004). For MHIV-mIN-ΔenvΔvprΔnef-puromycin^R (HIVmIN) and MHIV-mMA12CA-mIN-ΔenvΔvprΔnef-puromycin^R (HIVmGagmIN), the MLV IN-encoding portion of the pAMS *pol* gene was cloned in place of HIV IN, starting at the same position of the 5′ end of the HIV IN gene segment. The 3′ end of the HIV IN-encoding region with the cPPT remains and is separated from the end of MLV IN by 2 stop codons. (The junction sequence is CGTGGAAGCCCTTAATAGTCTgaattc.)

Infections. Vesicular stomatitis virus G protein (VSV-G)-pseudotyped virus was prepared as described previously (Yamashita and Emerman, 2004). HeLa cells were infected by spinoculation (O'Doherty *et al.*, 2000) with concentrated viral supernatant and 20 micrograms/ml DEAE-dextran. Infected cells were selected with 0.7 micrograms/ml puromycin for two weeks. Genomic DNA was extracted from pooled colonies.

Cloning integration sites. Genomic DNA was digested with MseI and ligated to a linker as described previously (Wu*, et al.*, 2003). The ligase was heat-inactivated at 65°C for 15 minutes and the genomic DNA was digested with a second restriction enzyme to limit the amplification of an internal viral fragment. SpeI was used for the MLVPuro virus and SacI was used for the HIV-based viruses. Viral-host DNA junctions were amplified by nested PCR using primers specific for the proviral LTR (reading out from the 3' end) and the linker essentially as described (GeneWalker Kit, Clontech). Nested PCR products were cloned using the TOPO TA cloning system

(Invitrogen). Clones were sequenced and mapped to the human genome with BLAT (University of California, Santa Cruz).

For analysis of the length of target site duplications, integration site clones were randomly chosen and genomic sequence-specific primers were designed. The viral-host DNA junction from the 5' LTR of the provirus was amplified from undigested genomic DNA and cloned using the TOPO TA cloning system (Invitrogen). Oligonucleotides used in this study are listed in Appendix 4.

Bioinformatic analysis. A detailed statistical analysis is presented in Appendices 2 and 3. In order to control for possible biases in the data sets due to the choice of restriction endonuclease used in cloning integration sites, each experimental integration site was paired with ten randomly selected sites in the genome that were exactly the same distance from an MseI site. These matched random control sites were generated *in silico* and used for comparison to the integration site data sets as previously described (Mitchell*, et al.*, 2004).

The statistical analysis of favored binding motifs (Figure 5) was carried out as follows. Let X and Y denote sets of significant factors around the integration sites in two independent experiments, with c factors in common. Assuming a random sampling of |X| and |Y| distinct factors from a pool of 347 transcription factors, the hypergeometric p-value estimates the probability of sampling c or more common factors.

For the analysis of the effects of host cell transcription on integration, we

acquired a set of HeLa transcriptional profiling data (assayed with Affymetrix HG-U133A microarrays) from NCBI Gene Expression Omnibus (GSM23372, GSM23373, GSM23377 and GSM23378 (Carson *et al.*, 2004)). For the analysis in Figure 4B, the signal values for each probe across the four arrays were averaged and ranked according to expression level.

The text of Chapter Two, in full, has been submitted for publication as:

Lewinski, M. K., Yamashita, M., Emerman, M., Shinn, P., Leipzig, J., Hannenhalli, S., Berry, C. C., Ecker, J. R., and Bushman, F. D. “Retroviral Gag and integrase act synergistically to determine integration target specificity,” 2005.

The dissertation author was the primary researcher and author.

III. GENOME-WIDE ANALYSIS OF CHROMOSOMAL FEATURES REPRESSING HUMAN IMMUNODEFICIENCY VIRUS TRANSCRIPTION

A. ABSTRACT

We have investigated regulatory sequences in noncoding human DNA that are associated with repression of an integrated human immunodeficiency virus type 1 (HIV-1) promoter. HIV-1 integration results in the formation of precise and homogeneous junctions between viral and host DNA, but integration takes place at many locations. Thus, the variation in HIV-1 gene expression at different integration sites reports the activity of regulatory sequences at nearby chromosomal positions. Negative regulation of HIV transcription is of particular interest because of its association with maintaining HIV in a latent state in cells from infected patients. To identify chromosomal regulators of HIV transcription, we infected Jurkat T cells with an HIV-based vector transducing green fluorescent protein (GFP) and separated cells into populations containing well-expressed (GFP-positive) or poorly expressed (GFP-negative) proviruses. We then determined the chromosomal locations of the two classes by sequencing 971 junctions between viral and cellular DNA. Possible effects of endogenous cellular transcription were characterized by transcriptional profiling. Low-level GFP expression correlated with integration in (i) gene deserts, (ii) centromeric heterochromatin, and (iii) very highly expressed cellular genes. These data provide a genome-wide picture of chromosomal features that repress transcription and suggest models for transcriptional latency in cells from HIV-infected patients.

B. INTRODUCTION

The position of genes within chromosomes is known to modulate their rate of transcription (Wolffe, 1998), but relatively few studies have systematically compared regulation at multiple chromosomal sites. Of these, most have focused on identifying positively acting promoters and enhancers by "enhancer trapping" or related approaches (Friddle *et al.*, 2003; Lukacsovich and Yamamoto, 2001). Here we have used human immunodeficiency virus (HIV) integration to identify negatively acting chromosomal features, an issue of interest both in understanding global control of transcription and in assessing HIV transcriptional latency in patients.

Retroviral model systems provide a tractable means of studying the influence of chromosomal context on transcription. Each integrated provirus is joined to flanking cellular DNA at exactly the same points at the ends of the viral DNA, but integration takes place at many different sites in the host cell chromosomes. Thus, the viral genome provides a homogeneous transcription template that can be analyzed at different chromosomal locations, allowing the influence of flanking chromosomal features to be assessed.

Early during HIV gene expression, transcription is initiated by polymerase II from the viral long terminal repeat (LTR) under the control of cellular factors, including NF-κB, SP1, NFAT, and others (Emerman and Malim, 1998; Freed, 2004). Most of the resulting transcripts terminate within 100 nucleotides of the transcription initiation site (Kao *et al.*, 1987). A low level of full-length transcripts is nevertheless synthesized, and a portion of these are spliced to yield the mRNA encoding Tat. In

the late phase of viral transcription, Tat accumulates in the host cell and binds to the TAR site on the viral RNA, recruiting the cyclin T-CDK9 complex and facilitating transcriptional elongation (Garber and Jones, 1999; Wei *et al.*, 1998a).

HIV transcription is known to be sensitive to the chromosomal environment at the site of integration (Jordan*, et al.*, 2003; Jordan*, et al.*, 2001). In one example of such regulation, Jordan *et al.* found that proviruses integrated into centromeric heterochromatin had undetectable levels of basal transcription. However, activation of transcription by treatment with tumor necrosis factor alpha (TNF-α) or 12-O-tetradecanoylphorbol 13-acetate (TPA), both of which induce the NF-κB pathway, allowed activation of such proviruses (Jordan*, et al.*, 2003; Jordan*, et al.*, 2001). Additional factors proposed to affect HIV transcription are reviewed in references (Freed, 2004) and (Garber and Jones, 1999).

Chromosomal features repressing HIV gene expression are of particular interest due to their possible influence on clinical latency in HIV infection. For many HIV-infected patients, treatment with highly active antiretroviral therapy can reduce viral loads to undetectable levels but, unfortunately, cells persist long term that harbor integrated proviruses capable of reseeding virus production after cessation of therapy. One well-characterized reservoir is in resting CD4-positive T cells (Chun*, et al.*, 1997b; Finzi*, et al.*, 1997; Wong*, et al.*, 1997). A low percentage of these cells harbor transcriptionally inactive HIV proviruses which may be induced to produce HIV upon T-cell activation. The finding that centromeric heterochromatin represses HIV gene expression, along with other known mechanisms for down-modulating HIV gene

expression (Blankson *et al.*, 2002; Freed, 2004; Garber and Jones, 1999; Sheridan *et al.*, 1997; Verdin, 1991), provides candidate explanations connecting transcriptional repression to clinical latency.

To study how expression from the HIV type 1 (HIV-1) promoter is affected by the integration site of the provirus, we isolated cells containing stably expressed and inducible proviruses, determined integration sites by sequencing 971 host-virus DNA junctions, and then asked what identifiable features were enriched in each population. Several notable biases were found, suggesting potential mechanisms by which the chromosomal environment may modulate HIV transcription.

C. MATERIALS AND METHODS

Vector preparation and infections. To produce the Tat and green fluorescent protein (GFP)-transducing HIV-based vector, 293T cells were cotransfected with pEV731 (LTR-Tat-IRES-GFP) (Jordan*, et al.*, 2001), the packaging construct pCMVdeltaR8.91, and the vesicular stomatitis virus G protein-producing pMD.G construct (Naldini *et al.*, 1996). Viral supernatant was harvested 48 h later and filtered through a 0.45-μm filter unit. Vector titer was determined by infection of 6×10^5 Jurkat cells with various amounts of vector supernatant and 4 μg/ml Polybrene (hexadimethrine bromide; Sigma). Cells were harvested 96 h after infection and analyzed by fluorescence-activated cell sorting for GFP expression.

Jurkat cells were cultured at a density of 3×10^5 to 1×10^6 cells/ml in RPMI 1640 medium with 10% fetal bovine serum, 100 U/ml penicillin, 100 μg/ml streptomycin, and 2 mM L-glutamine at 37°C. Cells were infected at a multiplicity of

infection of 0.1 with 4 μg/ml Polybrene for cloning integration sites and at 1.0 for analysis by transcriptional profiling. To date, comparisons between integration site data sets made with HIV-based vectors (Schroder*, et al.*, 2002; Wu*, et al.*, 2003) have not shown any differences with integration sites made with authentic HIV (Carteau*, et al.*, 1998; Wu*, et al.*, 2003).

Acquisition of stably bright and inducible cell populations. Jurkat cells were fluorescence-activated cell sorter (FACS) analyzed into GFP-positive and GFP-negative populations 2 to 4 days postinfection as described elsewhere (Jordan*, et al.*, 2003; Jordan*, et al.*, 2001). At this stage, about 7% of cells were GFP positive. The GFP-positive cells were sorted for GFP expression a second time 2 weeks postinfection, and DNA was extracted (QIAgen DNeasy tissue kit), yielding stably expressed proviruses. At this stage, about 90% of cells were GFP positive (geometric mean of GFP fluorescence measured in FL1 from a representative experiment was 215). GFP-negative Jurkat cells were sorted twice more for lack of GFP expression and then cultured with TNF-α for 17 h prior to sorting. After induction, approximately 0.25% of cells became GFP positive (geometric mean, 63.3, when analyzed 4 days after sorting). Note that the absolute level of the fluorescent signal measured in FL1 varied depending on the instrument used and the gate drawn compared to the uninfected control. The cells that were inducibly GFP positive were collected and DNA was extracted, yielding the inducible sample. The inducible cells became dark upon withdrawal of TNF-α (over 90% became dim 2 weeks after removal of TNF),

indicating dependence of expression on the inducing agent. The fraction of inducible cells seen in this study was similar to that reported previously (Jordan*, et al.*, 2003).

Integration site cloning and mapping to the genome. DNA from stably expressed and inducible populations was digested with three restriction endonucleases with six-base recognition sites (NheI, SpeI, and XbaI, essentially as described in (Schroder*, et al.*, 2002)) or with MseI (which has a four-base recognition site, as described in (Wu*, et al.*, 2003)). Digested DNA was then ligated to the appropriate adapter and amplified by nested PCR as described previously (Schroder*, et al.*, 2002). Oligonucleotides used are listed in the supplemental material in Appendix 4 (Table S2). Integration site sequences were determined to be authentic if they began at the junction with the HIV LTR, had a sequence identity of >98%, and yielded a unique best hit when mapped to the human genome using BLAT (UCSC).

A small data set (20 sites) was also generated using TPA as an inducing agent and analyzed. This set was biased in favor of integration in genes, and 2/20 were in alphoid repeats, paralleling sites analyzed after induction with TNF-α (data not shown).

Expression analysis. A total of 3 x 10^6 Jurkat cells (in triplicate per treatment group) were plated and either left untreated in culture, infected with the vesicular stomatitis virus G protein-pseudotyped LTR-Tat-IRES-GFP HIV-based vector (with 4 μg/ml Polybrene) at a multiplicity of infection of 1 for 24 h, or treated with 10 ng/ml TNF-α for 17 h. Cells were harvested, and total RNA was extracted using the QIAgen RNeasy kit. Labeling and hybridization of RNA to Affymetrix HG-U133A arrays was

performed using the Affymetrix protocol. Analysis used Affymetrix Microarray Analysis Suite 5.1 software. Changes in transcriptional activity were quantified using EASE and significance analysis of microarrays (SAM) to determine the false discovery rate. For the comparison of untreated Jurkat cells to HIV-infected cells, 575 genes were found to change at least twofold in activity (accepting a 1% false discovery rate). For the comparison of untreated cells to TNF-α-treated cells, 10 genes were found to be upregulated and 32 were downregulated under the same criteria.

Statistical analysis. A detailed statistical analysis is presented in the supplemental material (Appendix 5). An analysis of the randomly selected genes yielded a surprising result which suggested that the bias for favored integration in active genes (see Figure 9, below) is stronger than the figure may suggest. Randomly selected sites that were mapped to genes were distributed into classes by expression level as in Figure 9, below, and analyzed. The random sites did not yield a uniform distribution in each expression class, but instead revealed a bias in favor of the least-well expressed genes (values were as follows: class 1, 15.1 to 16.1%; class 2, 14.6 to 15.7%; class 3, 15.1 to 15.3%; class 4, 12.8 to 13.4%; class 5, 11.4 to 11.6%; class 6, 11.7 to 12.1%; class 7, 10.8 to 11.2%; class 8, 6.2 to 6.7%; $P < 0.0001$ by Chi-square; the range is for all three data sets in Figure 9A to C, below). This is probably explained by the finding that highly expressed genes tend to have shorter introns (Castillo-Davis *et al.*, 2002) and so are smaller targets for integration. This emphasizes that the tendency to integrate in active genes is likely stronger than

previously appreciated, because active genes are typically smaller than poorly expressed genes.

For the Mann-Whitney test to compare expression signals for the stably expressed and inducible proviruses, the data were filtered to remove noise by analyzing only genes that were called "present" on at least two out of three arrays.

Nucleotide sequence accession numbers. The sequences for the integration sites newly determined in this study have been deposited at NCBI and assigned accession numbers CZ442176 to CZ443146. Microarray data have been deposited at the NCBI GEO repository under accession number GSE2504.

D. RESULTS

Isolation of integration sites from cells containing stably expressed and inducible proviruses. To acquire cells containing stably expressed or weakly expressed proviruses, Jurkat cells (a $CD4^+$ T-cell line) were infected with an HIV-based vector that encoded the HIV transcriptional activator Tat and GFP (LTR-Tat-IRES-GFP) (Jordan*, et al.*, 2001) (Figure 6A). Cells were infected at a low multiplicity of infection (0.1) to minimize the fraction harboring more than one provirus. Cells were then separated several times by FACS into GFP-expressing and nonexpressing populations (Figure 6B). The GFP-negative population was treated with TNF-α, an agent that is known to activate LTR transcription (Schmid *et al.*, 1991) and thereby to activate transcription from silent proviruses. Cells were then sorted to obtain the induced GFP-positive population. Previous studies using this model have shown that most of these inducible proviruses are silent due to integration

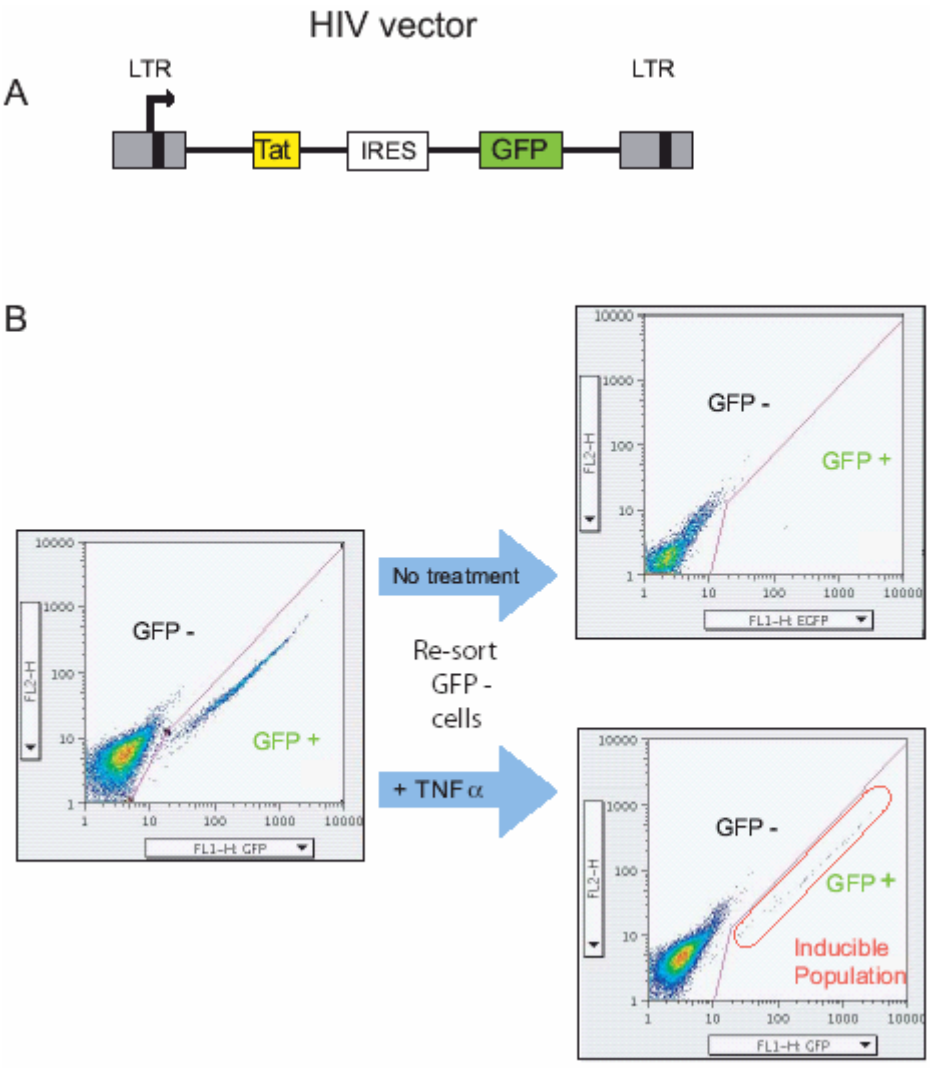


Figure 6: Acquisition of cells containing stably expressed and inducible proviruses. A) Tat-transducing HIV-based vector used in this study. Tat, HIV-encoded transcriptional activator; IRES, internal ribosome entry site. Transcription initiates within the left LTR. B) Acquisition of cells containing stably expressed and inducible proviruses by FACS. Cells were infected at a multiplicity of about 0.1 and sorted for GFP-positive and -negative cells (left side). GFP-positive cells were collected and then sorted a second time to isolate a stably bright fraction. The GFP-negative (dark) population was sorted twice, and the dark cells were collected each time. The stably dark cells were then treated with TNF-α, and the resulting bright cells were collected (right side).

in chromosomal sites unfavorable for gene expression (Jordan, *et al.*, 2003; Jordan, *et al.*, 2001). In addition, focusing on the inducible fraction minimizes possible complications resulting from the inactivation of viral genomes by mutation. Integrated proviruses that were not expressed and were uninducible were not studied.

Chromosomal integration sites from cells in the stably expressed and inducible populations were then cloned using ligation-mediated PCR and sequenced (Schroder, *et al.*, 2002; Wu, *et al.*, 2003). The chromosomal distributions of these sites were compared to two data sets generated by infection of lymphoid cells (SupT1 cells or primary peripheral blood mononuclear cells) with HIV-based vectors (Mitchell, *et al.*, 2004; Schroder, *et al.*, 2002). The cells in these studies were not fractionated by the level of proviral gene expression, and so these data sets provide an overview of integration site selection by HIV. A set of 10,000 random sites in the human genome generated *in silico* was also included for comparison (Table 3).

Frequency of integration in genes. Since the complement of human genes has not been fully clarified, we used four different gene catalogs to analyze the frequency of integration in transcription units (Table 4). For all sets of HIV integration sites and all types of gene calls, integration was strongly biased in favor of transcription units (Mitchell, *et al.*, 2004; Schroder, *et al.*, 2002; Wu, *et al.*, 2003). For example, using the well-characterized RefSeq genes for comparison, the human genome contains 31.1% genes, while HIV integration site data sets showed frequencies of integration in genes from 66.1% (SupT1 cells) to 73.4% (Jurkat cells, inducible integration sites). The stably expressed and inducible populations of

Table 3: Integration site data sets used in this study

Data set	Vector	Cell type	No. of integration sites	Source or reference
Stably expressed	HIV: LTR-Tat-IRES-GFP	Jurkat	587	This report
Inducible	HIV: LTR-Tat-IRES-GFP	Jurkat	384	This report
HIV/SupT1	HIV p 156 (CMV-GFP)	SupT1	493	Schroder et al., 2002
HIV/PBMC	HIV p 156 (CMV-GFP)	PBMC[a]	550	Mitchell et al., 2004
Random			10,000	This report

[a] PBMC, peripheral blood mononuclear cells

Table 4: Integration in transcription units[b]

Chromosomal feature	Frequency (%) of transcription units at integration sites in:				
	Human genome (random sites)	Stably expressed sites, HIV/Jurkat	Inducible sites, HIV/Jurkat	HIV/SupT1	HIV/PBMC
Acembly	49.2	87.6	89.1	83.2	87.8
GenScan	64.3	78.4	78.6	76.1	79.5
RefSeq	31.1	71.2	73.4	66.1	69.1
UniGene	50.8	79.2	80.7	72.6	75.1

[b] All comparisons to random show $P < 0.0001$.

proviruses both showed similar high frequencies of integration in genes (see the statistical information provided in the supplemental material, Appendix 5).

Primary sequences at integration sites. The primary sequences that served as integration targets were analyzed separately for the stably expressed and inducible proviruses (Figure 7). The sequences from both data sets showed inverted repeat symmetry centered on the sequence 5′GT(A/T)AC3′ as previously reported (Bor*, et al.*, 1996; Carteau*, et al.*, 1998; Stevens and Griffith, 1996). The more detailed analysis reported here also shows the presence of a longer consensus, with notable conservation about one turn of the helix in either direction out from the conserved sequences. No binding sites for known transcription factors were significantly enriched in either data set (data not shown). Thus, we could not detect any clear differences between the two data sets in the local sequences at integration sites.

Integration in repeated sequences: inducible proviruses are more frequently found in alphoid repeats. Despite these similarities between the stably expressed and inducible integration sites, three features were found to differ. Each suggests a chromosomal feature disfavoring HIV transcription. The first involved the frequency of integration in repeated sequences (Table 5).

The frequency of integration in alphoid repeats was 4.3% in the inducible Jurkat sites but only 0% to 0.5% in the other HIV data sets. Alphoid repeats are mostly found in centromeres, and packaging of DNA in centromeric heterochromatin is known to repress transcription of many genes (She *et al.*, 2004; Wallrath, 1998). These data support the idea that HIV DNA embedded in centromeric heterochromatin

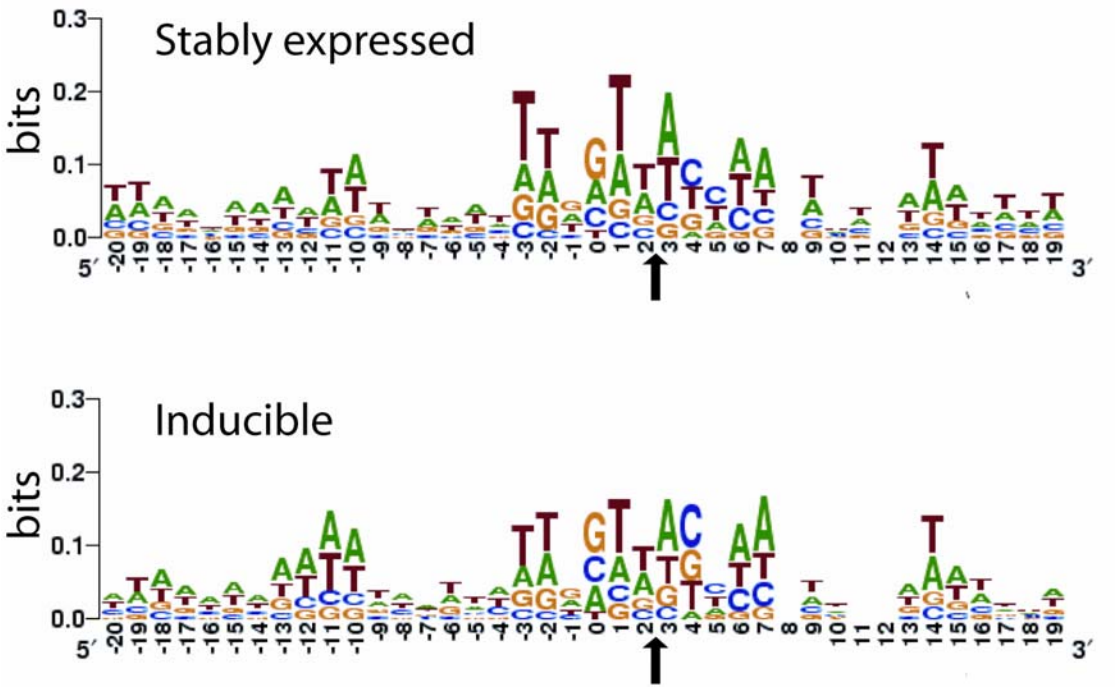


Figure 7: Primary sequences surrounding the stably expressed and inducible proviruses. The weak consensus sequence seen at the stably expressed (top) and inducible (bottom) proviruses was rendered so that the degree of conservation is proportional to the height of each letter, using LOGO (http://weblogo.Berkeley.edu/logo.cgi). The y axis reflects the information content at each base, so that perfect conservation would have a score of 2 bits. The points of joining between the HIV and human DNA lie between -1 and 0 (for the sequenced HIV DNA end) and between 4 and 5 on the other strand for the other end of the HIV DNA. Thus, the points of joining, and the integration consensus sequence, are symmetric around position 2 (arrow).

Table 5: Integration in repeated sequences[a]

Chromosomal feature	Frequency (%) of repeated sequences at integration sites in:				
	Human genome (random sites)	Stably expressed sites, HIV/Jurkat	Inducible sites, HIV/Jurkat	HIV/SupT1	HIV/PBMC
SINES					
Alu	9.4	9.1 (0.8325)	9.5 (0.9002)	17.6 (<0.0001)	10.1 (0.5246)
MIR	2.5	3.0 (0.4186)	1.7 (0.3087)	1.5 (0.107)	3.2 (0.2713)
DNA elements	2.7	2.1 (0.3491)	3.9 (0.1207)	2.4 (0.6898)	3.9 (0.0844)
LTR elements (HERV)	7.7	5.1 (0.0124)	3.5 (0.0007)	4.5 (0.0035)	2.5 (<0.0001)
LINE	18.0	21.2 (0.0368)	15.2 (0.1207)	19.2 (0.4347)	15.5 (0.132)
Alpha satellite	0.3	0.1 (0.5807)	4.3 (<0.0001)	0.5 (0.2987)	0.0 (0.2142)

[a] The percentages are relative to all sites in the data set; values in parentheses are p-values (Chi-square) compared to random sites.

is poorly expressed, so that enriching for poorly expressed proviruses enriched for those in alphoid repeats (Jordan*, et al.*, 2003; Jordan*, et al.*, 2001).

A small number of integration sites (20 total) were isolated from cells after induction with TPA instead of TNF-α. Of these, two were in alphoid repeats, paralleling results with TNF-α induction (data not shown).

All HIV integration site data sets showed that human endogenous retroviruses (HERVs) are significantly disfavored targets ($P < 0.013$), as reported previously for the SupT1 data set (Schroder*, et al.*, 2002). HERVs are enriched outside transcription units, while HIV integration is favored within transcription units, accounting for the observed bias.

Inducible proviruses are more frequently found in gene deserts. A second difference was found in an analysis of the positions of stably expressed and inducible proviruses in intergenic regions. The stably expressed proviruses were more frequently found in short intergenic regions, indicative of favored integration in gene-rich chromosomal domains, as seen previously (Mitchell*, et al.*, 2004; Schroder*, et al.*, 2002). In contrast, the inducible proviruses were much more frequently found in long intergenic regions or "gene deserts" (Figure 8) ($P < 0.0007$, regardless of gene call used for the analysis) (see the statistical information provided in Appendix 5).

This finding was reinforced by an analysis of the density of integration events compared to the density of CpG islands, which are more common in gene-dense regions. The stably expressed proviruses were found more commonly in regions of

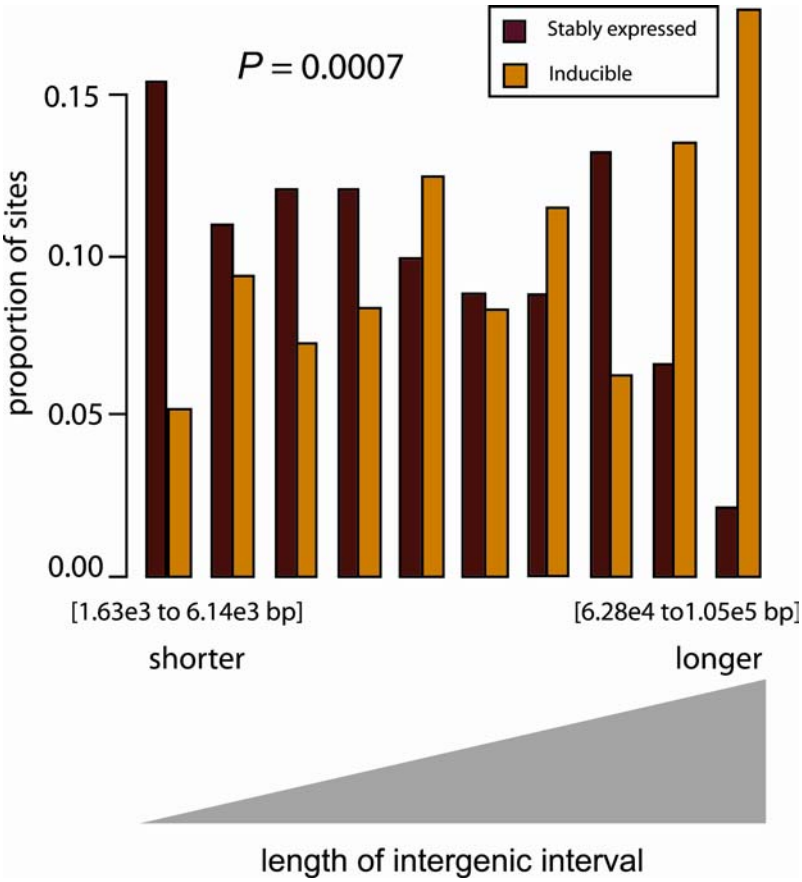


Figure 8: Frequency of stably expressed or inducible proviruses in intergenic regions of different lengths. Shorter intergenic regions are shown to the left, and longer ones are to the right. GenScan genes were used for this analysis, though the conclusions were similar for other gene sets as well (see the statistical information provided in Appendix 5). The p-value is obtained from the logistic regression of event type (stable or inducible) on a cubic B-spline basis (i.e., a third-order polynomial) for intergenic distance. The units on the x axis indicate lengths of intergenic regions, in base pairs. Lengths of intergenic regions for each category were defined by the following boundaries (from left to right, in bp): 1,627, 6,135, 10,506, 14,900, 21,907, 28,989, 36,333, 43,531, 62,837, 104,802, and 3,182,720. The inducible proviruses in the rightmost five bins accounted for 14% of all inducible proviruses.

high CpG island density, whereas the inducible sites were enriched in regions of low density (P = 0.002) (see the statistical information provided in Appendix 5). This indicates that the inducible proviruses are enriched in long intergenic regions that are depleted of both genes and CpG islands.

Inducible proviruses are more frequently found in very highly expressed cellular genes. A third chromosomal feature correlating with inducible HIV gene expression was identified by transcriptional profiling analysis of the Jurkat target cells. The expression signals of cellular genes hosting integration events were tabulated for the stably expressed and inducible proviruses. The median for both groups of genes was found to be higher than the median of all the probe sets on the HU133A microarrays used (stably expressed = 152, inducible = 177, all genes on the array = 66; units are "signal," as defined by Affymetrix MAS 5.1). Genes in both the stably expressed and inducible populations were also more active than genes from the random control population in Table 3 (random = 57; $P < 0.0001$ for comparison to either the stably expressed or inducible populations; Mann-Whitney test). This broadly parallels previous studies of HIV, which revealed that active genes were favored as integration targets (Mitchell*, et al.*, 2004; Schroder*, et al.*, 2002; Wu*, et al.*, 2003).

Thus, it was unexpected that the stably expressed and inducible data sets differ from each other. The median expression value for genes hosting inducible proviruses was found to be significantly higher than the median of genes hosting stably expressed proviruses (P = 0.0004; Mann-Whitney test).

To analyze this issue in more detail, expression signals of genes hosting integration events were divided into classes by their signal values and the distribution was examined (Figure 9A). As with previous studies, genes hosting integration events were found more commonly in the more highly expressed genes. The inducible proviruses were more frequently found in the highest expression class: 24% of inducible integration sites (in genes represented on the array) compared to 14% for the stably expressed set ($P = 0.003$; Chi-square test). In previous studies, genes in the highest expression class (eighth bin) were consistently found to be less favorable for integration (Mitchell*, et al.*, 2004; Schroder*, et al.*, 2002); here, this is seen as well for the stably bright population but not the inducible population. Thus, we infer that integration in the very highly expressed genes was associated with the inducible phenotype and, specifically, that the transcription level in bin 8 is unfavorable for HIV transcription. Inducible proviruses in highly expressed genes were found in both orientations relative to the direction of host gene transcription (data not shown). An analysis of the placement of integration sites within genes showed no obvious bias; for example, the inducible sites in the most highly transcribed genes (eighth bin) were not clustered near the start site of transcription (data not shown).

The relationship between integration targeting and host cell transcription was probed further by repeating the transcriptional profiling measurements under two additional conditions. Jurkat cells were infected with the HIV-Tat-GFP vector prior to RNA isolation, or cells were treated with 10 ng/ml TNF and RNA was isolated subsequently. These manipulations caused clearly detectable changes in transcription.

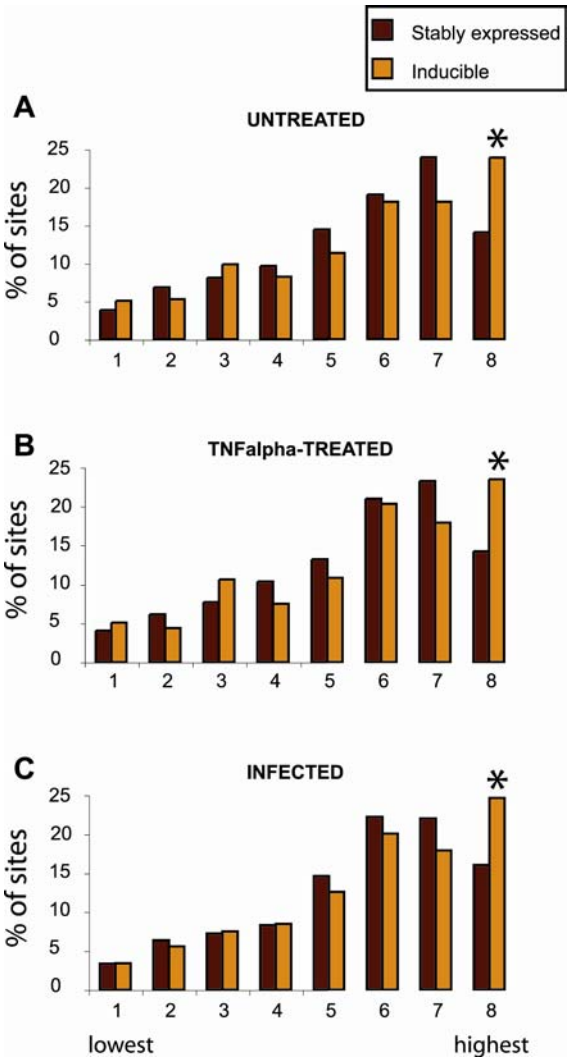


Figure 9: Inducible proviruses are found more commonly in very highly active genes. Expression levels were assayed in Jurkat cells (three independent Affymetrix HU133A microarrays for each condition) and scored using the Affymetrix Microarray Suite 5.1 software package. To classify the expression levels of genes hosting integration events, class boundaries were first generated by dividing all the genes on the array into eight classes according to their relative level of expression. Genes that hosted integration events were then distributed into the classes defined by these boundaries, summed, and expressed as a percentage of the total number of integration sites in genes on the array. The leftmost class in each panel contains the 1/8 most weakly expressed genes, and the rightmost class contains the 1/8 most highly expressed. The highest signal value represented in each expression bin (for untreated Jurkat cells) was as follows: bin 1, 9.2; bin 2, 20.6; bin 3, 38.6; bin 4, 66; bin 5, 117; bin 6, 227; bin 7, 488; bin 8, 12050. Integration sites were analyzed using data from untreated Jurkat cells (A), TNF-treated Jurkat cells (B), or HIV-Tat-GFP-infected Jurkat cells (C) ($P < 0.003$; Chi-square test). Inducible proviruses in the eighth class (most highly expressed) accounted for about 17% of the total.

Notably, infection with the Tat-transducing vector caused down-modulation of a large family of genes involved in signal transduction and immune responses, potentially a biologically significant activity of Tat involved in evasion of the host immune response (de la Fuente *et al.*, 2002; Izmailova *et al.*, 2003; Kanazawa *et al.*, 2000). In Figure 10, signal intensities from Affymetrix HU133A microarrays were analyzed by SAM (http://www-stat.Stanford.EDU/tibs/SAM/) to identify significantly affected genes and then clustered according to gene ontology using EASE (http://david.niaid.nih.gov/david/ease.htm). A large set of Tat-repressed genes (115 probe sets corresponding to 108 different genes) was identified as overrepresented compared to all genes queried by the microarray in the "signal transducer activity" category ($P = 1.16 \times 10^{-5}$; Fisher exact test with Bonferroni correction for multiple comparisons). Expression values were normalized by dividing by the mean. In cases where multiple probe sets queried the activities of a single gene, the values were found to be closely similar and a single representative probe set was used for the figure.

Treatment with TNF resulted in induction of a number of previously characterized TNF-inducible genes. Though these changes were readily detectable, overall transcription in the cell types studied was still quite similar (correlation coefficients for pair-wise comparisons of any two microarrays showed $R > 0.98$). Analysis of genes hosting integration events using these transcriptional profiling data sets also indicated that very highly transcribed cellular genes were more common targets in the inducible data set (Figure 9B and C).

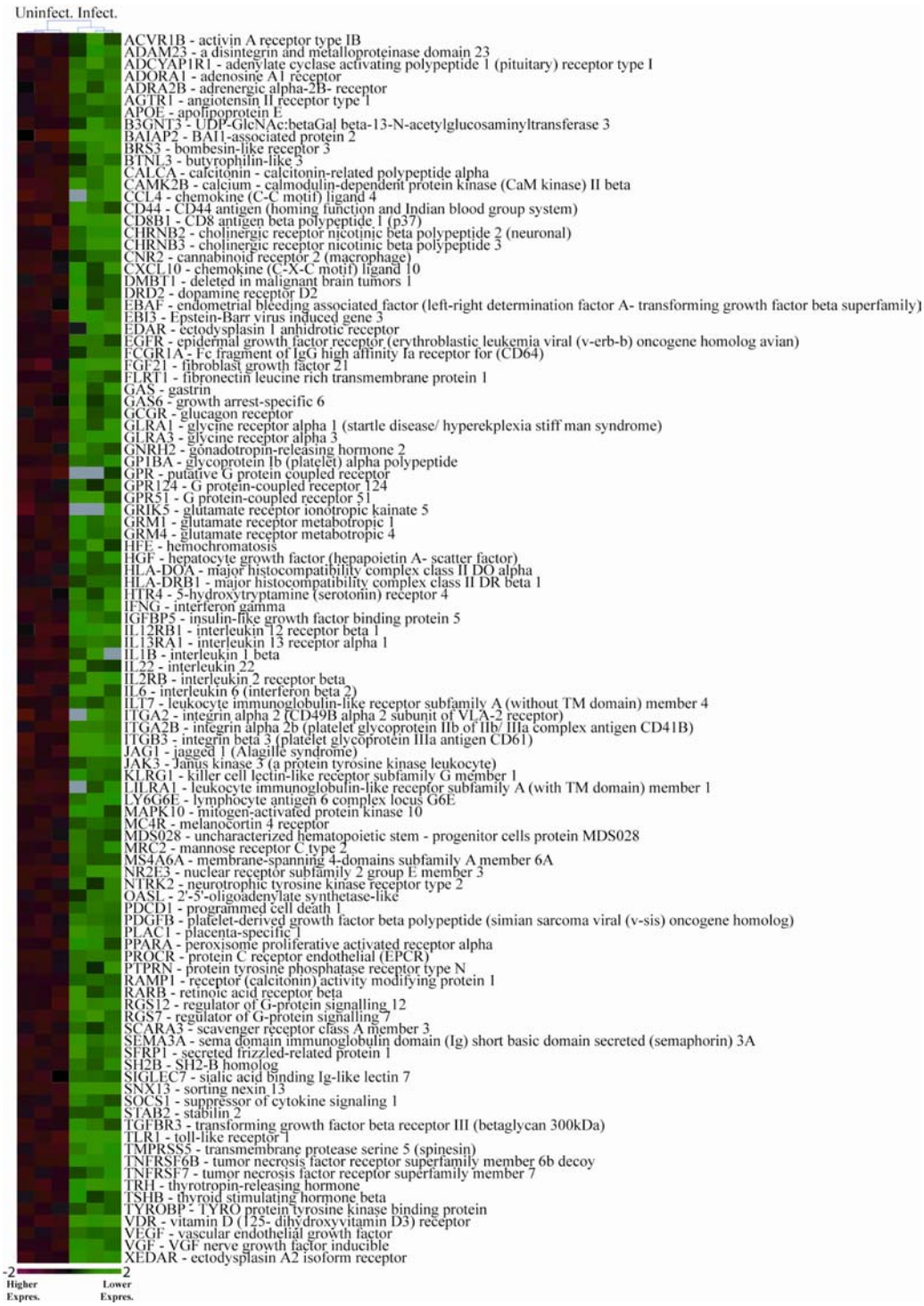


Figure 10: Tat down-modulates host cell genes important in signal transduction and immune responses. The three left columns show results from uninfected cells, and the three right columns show results from cells infected with the Tat-transducing HIV-based vector. Gray tiles indicate negative values. All significantly affected genes called by EASE in the "signal transducer activity" category are shown, except for six olfactory receptors and one taste receptor.

Jurkat cells as model HIV target cells: assessment using transcriptional profiling. The transcriptional profiling data on Jurkat cells could be used to investigate how closely the Jurkat cell line models the primary cells normally targeted by HIV infection *in vivo*. Transcriptional profiles of uninfected Jurkat cells were compared to 79 transcriptional profiles of human cells and tissues (data from (Su *et al.*, 2004)). A cluster analysis is shown in Figure 11. Transcriptional profiles of Jurkat cells clustered with profiles of a collection of leukocytes, including $CD4^+$ T cells. Jurkat cell transcription did differ somewhat from $CD4^+$ T cells, however, which could be due to the transformed state of Jurkat cells or to differences in the execution of the microarray experiments. Inspection of the Jurkat transcriptional profiles indicates that many of the genes expected to be active in $CD4^+$ T cells are indeed robustly expressed (Figure 10 and data not shown), consistent with previous studies in which Jurkat cells were shown to be active in assays of T-cell function (e.g., references (Frumento *et al.*, 1997) and (Manger *et al.*, 1986)). In summary, transcription in the Jurkat cell clusters with authentic $CD4^+$ T cells, helping to validate the use of Jurkat cells as a model of infection *in vivo*.

E. DISCUSSION

Here we compared the chromosomal placement of HIV proviruses that were stably expressed after integration to proviruses that were poorly expressed but inducible upon treatment of cells with TNF-α. Three chromosomal features correlated with inducible expression: centromeric heterochromatin, gene deserts, and highly active host transcription units. Each of these is discussed below. However, only about

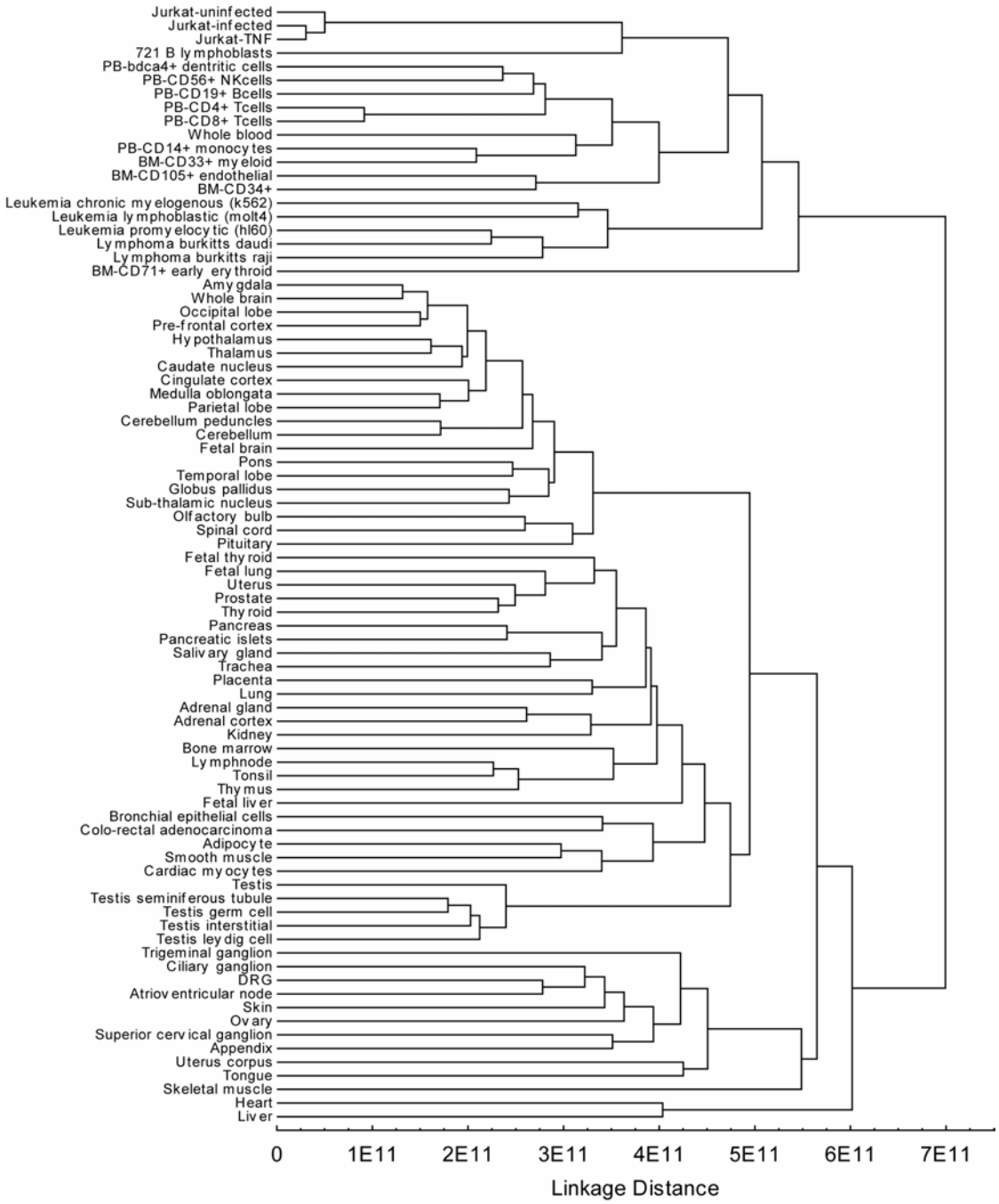


Figure 11: Clustering of transcriptional profiles from Jurkat cells with human leukocytes. Data for human tissues are from reference (Su, *et al.*, 2004). All analyses used Affymetrix HU133A microarrays. Transcription signal values were averaged between replicates and ranked prior to clustering. Squared Euclidean distance and unweighted pair-group average linkage (also know as UPGMA) cluster analysis of the transcriptional profiles was carried out using Statistica 7.0.

40% of the inducible proviruses were associated with one of these three features, and so further chromosomal environments unfavorable for expression may yet be found. In addition, studies from others using this model suggest that low-level GFP expression may also result from stochastic fluctuations in Tat levels. For cells expressing low levels of Tat protein, fluctuations in Tat concentration may extinguish LTR-driven transcription, and this may become "locked in" because Tat protein is required to activate its own expression (D. Schaffer and coworkers, personal communication).

Silencing HIV proviruses by transcriptional interference. A significantly greater proportion of the inducible proviruses were found in the most highly expressed fraction of host genes (Figure 9), suggesting that very-high-level host gene transcription interferes with transcription of an integrated provirus. Many studies have established that transcriptional interference can repress gene expression (Callen *et al.*, 2004; Cullen *et al.*, 1984; Greger *et al.*, 2000; Greger *et al.*, 1998; Hausler and Somerville, 1979; Martens *et al.*, 2004), and a model HIV promoter has previously been shown to be sensitive to transcriptional interference in HeLa cells (Greger*, et al.*, 1998). For a provirus in the same orientation as the host cell gene, read-through transcription may repress by blocking access of factors to the downstream promoter or by actively dislodging bound proteins (Callen*, et al.*, 2004; Greger*, et al.*, 2000; Greger*, et al.*, 1998; Hausler and Somerville, 1979; Martens*, et al.*, 2004). In the HeLa cell model, read-through transcription was found to repress HIV transcription by dislodging bound Sp1 (Greger*, et al.*, 1998). A provirus in an orientation opposite that

of the host gene may be silenced by the above mechanisms, or by transcriptional "trainwrecking" whereby two RNA polymerase complexes collide during convergent elongation. Convergent transcription could also result in transcription of both DNA strands and formation of double-stranded RNA, which might silence proviral transcription via RNA interference (reviewed in references (Hu *et al.*, 2004; Plasterk, 2002)), RNA-directed DNA methylation (Morris *et al.*, 2004), induction of the interferon response (Fields and Kinpe, 1996), or generation of antisense RNA (Scherer and Rossi, 2003).

Inducible proviruses are integrated more commonly in gene deserts. A strong trend was seen involving integration sites outside genes, in which long intergenic regions or gene deserts more frequently hosted inducible proviruses. Short intergenic regions more commonly hosted stably expressed proviruses. A similar trend was also seen comparing the frequency of integration in CpG islands, which are known to be associated with genes. A variety of mechanisms could account for this bias, none mutually exclusive. Gene deserts may be heterochromatic, and so packaged in proteins unfavorable for efficient transcription (Jenuwein, 2001; Jenuwein and Allis, 2001; Wallrath, 1998). Gene deserts may be enriched in binding sites for transcriptional silencer proteins, though no candidate binding sites emerged from our analysis of primary sequences at integration sites. Intranuclear positioning of gene deserts could also be a factor (Boyle *et al.*, 2001; Casolari *et al.*, 2004; Chubb and Bickmore, 2003). A recent study suggested that activation of genes in yeast can be accompanied by translocation of the genes to a nuclear pore complex (Casolari*, et al.*,

2004). Thus, proviruses integrated into gene-sparse regions may be localized within nuclear domains that are unfavorable for transcription.

Integration in centromeric heterochromatin disfavors HIV gene expression. Repression of HIV expression after integration in alphoid repeats was previously observed by Eric Verdin and colleagues using the Jurkat model (Jordan*, et al.*, 2003; Jordan*, et al.*, 2001). Heterochromatin adopts a condensed structure that blocks access of the transcriptional machinery (She*, et al.*, 2004; Wallrath, 1998). Thus, a simple model to explain our results is that wrapping of the proviral DNA in heterochromatin blocks access of the transcriptional machinery and thereby represses transcription.

Models for the mechanism of transcriptional latency in patients. HIV-infected patients on successful long-term antiretroviral therapy nevertheless harbor cells containing latent proviruses, and after cessation of treatment HIV from these cells can reinitiate active replication (Chun*, et al.*, 1997b; Finzi*, et al.*, 1997; Han*, et al.*, 2004; Wong*, et al.*, 1997). Our findings reveal mechanisms by which the surrounding chromosomal environment may silence some integrated proviruses while leaving them inducible by TNF-α treatment. The data presented here suggest that proviruses integrated in centromeric heterochromatin, gene deserts, and highly transcribed genes may contribute to the latent population.

Direct studies of integration sites from latently infected cells in patients have been challenging. One report investigated the distribution of HIV integration sites in resting $CD4^+$ lymphocytes of patients on effective highly active antiretroviral therapy

(Han*, et al.*, 2004). However, this work was complicated by the fact that defective proviruses greatly outnumber latent proviruses in patient cells (Chun*, et al.*, 1997b; Finzi*, et al.*, 1997; Wong*, et al.*, 1997). Han *et al.* cloned 74 integration sites and found that 93% of the proviruses were integrated within active transcription units (Han*, et al.*, 2004). If these sites are representative of latent integration sites in patients, then the transcriptional interference model may be the most attractive based on our data.

The text of Chapter Three, in full, is a reprint of the material as it appears in the

Journal of Virology:

Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C. C., Ecker, J. R., and Bushman, F. D. "Genome-wide analysis of chromosomal features repressing HIV transcription". *J Virol* **79**, 6610-9, 2005.

The dissertation author was the primary researcher and author.

IV. CONCLUSIONS

Genome-wide studies of integration targeting have provided substantial insight into the virus-host cell interaction. The differential integration target site selection preferences of retroviruses could reflect subtle differences in their replication strategies, analogous to the pressures driving Ty retrotransposon targeting of integration to benign regions of the yeast genome. For instance, HIV-1 has a small window in which to replicate because productively infected cells are quickly eliminated by cytotoxic T lymphocytes and the cytopathic effects of the virus. In order to maximize progeny production, HIV may have evolved to target integration to regions of the host genome most conducive to high proviral gene expression, such as gene-rich regions of chromosomes and active cellular genes. The preferred target sites of MLV (transcription start sites, CpG islands and DNase I hypersensitive sites, among others) might be near binding sites for transcription factors that aid in MLV gene expression or are genomic regions where the provirus might escape silencing by CpG methylation.

The studies presented in the previous chapters have contributed to our understanding of the mechanism and consequences of retroviral integration. Evidence that the retroviral Gag proteins as well as integrase determine integration target site selection preferences suggests modifications to the simplest models of integration targeting, i.e., that regions of open chromatin are preferentially targeted for integration because they are accessible, that binding of integrase to specific tethering factors

directs integration to sites nearby, or that variations in transcriptional state of the host cell at different phases of the cell cycle account for differences between cell-cycle restricted (MLV) and unrestricted (HIV) viruses. The differential preferences of these viruses for DNase I hypersensitive sites and active genes argue against the idea that open chromatin is the primary determinant of integration targeting. The observation that an HIV-based chimera with MLV *integrase* (HIVmIN) does not have target site selection preferences similar to MLV while a chimera with both MLV *IN* and *gag* (HIVmGagmIN) does argues against the direct interaction between integrase and tethering factors being the only determinant of target site selection. This data suggests that Gag plays a role, either directly, by binding in a highly co-operative fashion with IN to tethering factors, or indirectly, by restricting nuclear entry of MLV PICs to a specific point in the cell cycle. Cell-cycle related changes in chromatin conformation and nuclear organization alone cannot account for the differences in targeting between HIV and MLV because an MLV *gag*-substituted HIV chimera (HIVmGag) that is cell-cycle restricted like MLV does not exhibit integration site selection preferences like those of MLV. In a refined model of integration targeting, the phase of the cell cycle determines whether tethering factors for the PIC are bound to preferred sites in the cellular DNA. MLV capsid-p12, by remaining associated with integration complexes of MLV and the HIVmGagmIN chimera, restricts access of the PIC to the cellular DNA until after mitosis. At this point in the cell cycle, tethering factors that interact with integrase and/or other elements of the PIC could be bound near transcription start

sites, CpG islands and DNase I hypersensitive sites, directing integration near these features.

This modification of integration targeting by swapping elements of the retroviral genome suggests a strategy for engineering safer retroviral gene therapy vectors. While MLV-based gene therapy vectors have been successfully employed to treat X-linked severe combined immunodeficiency, their insertion near the transcription start site of the *LMO-2* proto-oncogene has contributed to the development of leukemia in at least two patients (Hacein-Bey-Abina*, et al.*, 2003a; Hacein-Bey-Abina*, et al.*, 2003b). Such insertional mutagenesis is a significant risk with these vectors considering the preference MLV has for integration in and near promoters. By substituting *gag* and *IN* coding regions from a virus (such as ASLV) that prefers to integrate in regions of the genome less likely to disrupt host gene expression, a safer hybrid vector could be produced.

A genome-wide comparison of integration sites from well expressed and poorly expressed HIV-1 proviruses suggested that integration site does play a role in HIV expression and is a candidate contributor to the phenomenon of postintegration latency. Three genomic features were significantly enriched at integration sites of reversibly silenced proviruses: gene deserts, centromeric heterochromatin and very highly expressed host genes. Gene deserts (long intergenic regions) likely have an intranuclear position that is unfavorable for proviral gene expression. Centromeric heterochromatin has a condensed conformation that blocks access of transcriptional machinery to proviruses in these regions. High levels of host gene transcription could

silence proviral expression by transcriptional interference. These genomic features are not favored targets of HIV integration and this is consistent with the idea that HIV has evolved to preferentially target integration to genomic regions favorable for its expression. In the rare instances where viral cDNA does integrate into gene deserts, heterochromatin or very highly expressed host genes, the level of proviral gene transcription may be low. A model for the contribution of integration site to viral latency could be that HIV infects an activated $CD4^+$ T cell, completes the process of reverse transcription, and integrates into a chromosomal region that represses proviral transcription. Expression of viral proteins is suppressed long enough for the host cell to survive and revert to a quiescent memory T cell. In the memory T cell, the virus remains latent for years until the host cell encounters its antigen, is activated and produces progeny virions. Determining how relevant this model is to the clinical phenomenon of HIV latency will require further study.

The publication of the human genome sequence has allowed for these large-scale studies of genomic features associated with integration target sites of chimeras and viruses sorted by expression level. Careful analysis of this data has allowed us to identify the viral determinants of integration site selection and to elucidate the influence of integration site on proviral expression, thus contributing to our understanding of the mechanisms and some of the consequences of retroviral integration.

APPENDIX 1

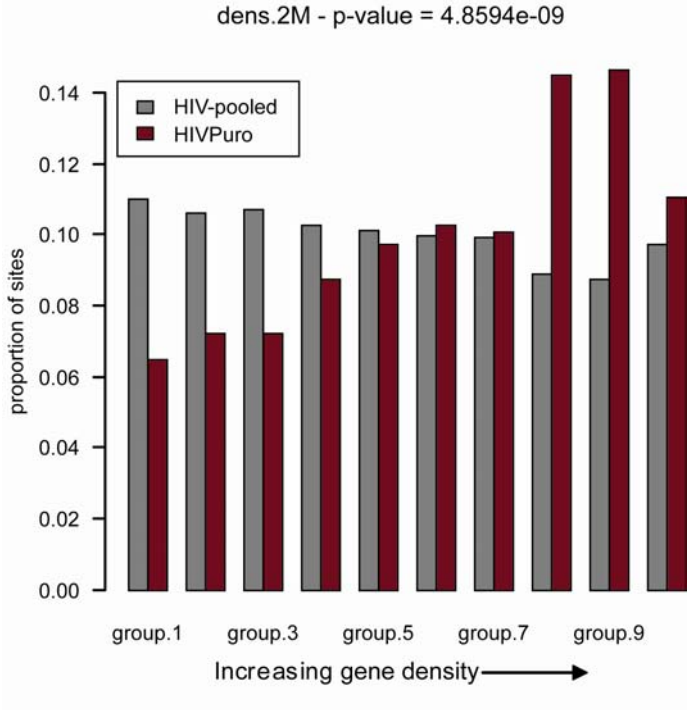
The Effects of Puromycin Selection on Integration Site Recovery

HIVPuro vs. unselected pooled HIV data sets

HIVPuro favors gene-rich regions over unselected HIV-pooled. The following plot examines the association of integration sites with gene density in a 2 megabase window surrounding each locus. The data is divided into deciles of gene density, with the most gene-poor decile on the left (group.1) and the most gene-rich decile on the right (group.10). We plot the proportion of integration sites from Puromycin-selected HIVPuro and unselected HIV-pooled data sets that fall in 2 megabase windows with the indicated gene density. The boundaries of each gene density group are as follows (as genes/bp):

	lower category		upper
1	0.000000e+00	group.1	1.309524e-06
2	1.309524e-06	group.2	2.000000e-06
3	2.000000e-06	group.3	2.828333e-06
4	2.828333e-06	group.4	3.808333e-06
5	3.808333e-06	group.5	4.916667e-06
6	4.916667e-06	group.6	6.333333e-06
7	6.333333e-06	group.7	9.183333e-06
8	9.183333e-06	group.8	1.330060e-05
9	1.330060e-05	group.9	1.849594e-05
10	1.849594e-05	group.10	4.108333e-05

The p-value given is the result of fitting a cubic polynomial to the gene density values.

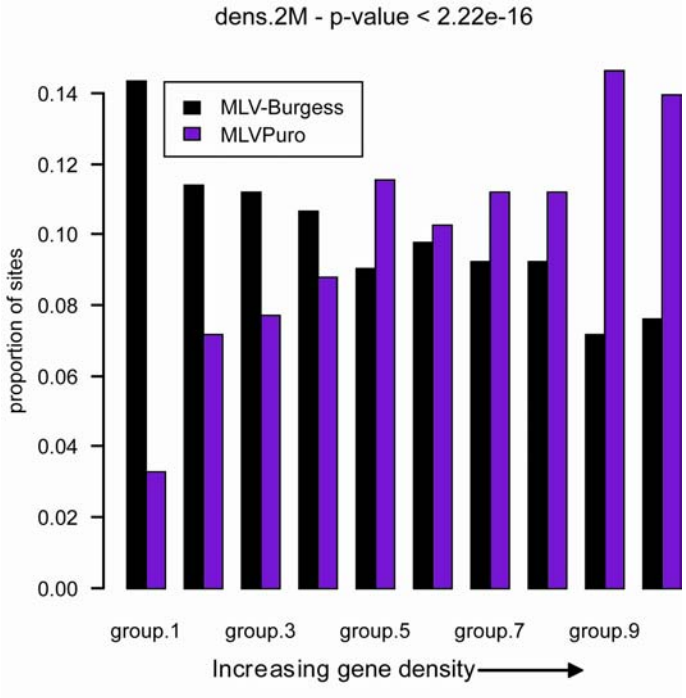


MLVPuro vs. unselected MLV data set

MLVPuro favors gene-rich regions over unselected MLV-Burgess. This plot examines the association of integration sites with gene density in a 2 megabase window surrounding each locus. The data is divided into deciles of gene density, with the most gene-poor decile on the left (group.1) and the most gene-rich decile on the right (group.10). We plot the proportion of integration sites from Puromycin-selected MLVPuro and unselected MLV-Burgess data sets that fall in 2 megabase windows with the indicated gene density. The boundaries of each gene density group are as follows (as genes/bp):

	lower category		upper
1	0.000000e+00	group.1	1.250000e-06
2	1.250000e-06	group.2	1.916667e-06
3	1.916667e-06	group.3	2.736111e-06
4	2.736111e-06	group.4	3.620833e-06
5	3.620833e-06	group.5	4.645833e-06
6	4.645833e-06	group.6	5.800000e-06
7	5.800000e-06	group.7	7.583333e-06
8	7.583333e-06	group.8	1.006667e-05
9	1.006667e-05	group.9	1.565000e-05
10	1.565000e-05	group.10	3.950000e-05

The p-value given is the result of fitting a cubic polynomial to the gene density values.



These data represent the converse of the finding presented in Chapter Three that gene-poor regions or “gene deserts” repress HIV transcription. Together, these results suggest that on average integration in gene-rich regions is more favorable for subsequent proviral gene expression and that this is true for MLV as well as HIV.

APPENDIX 2

Association of Genomic Features with Integration

Charles C. Berry

Contents

1 Introduction

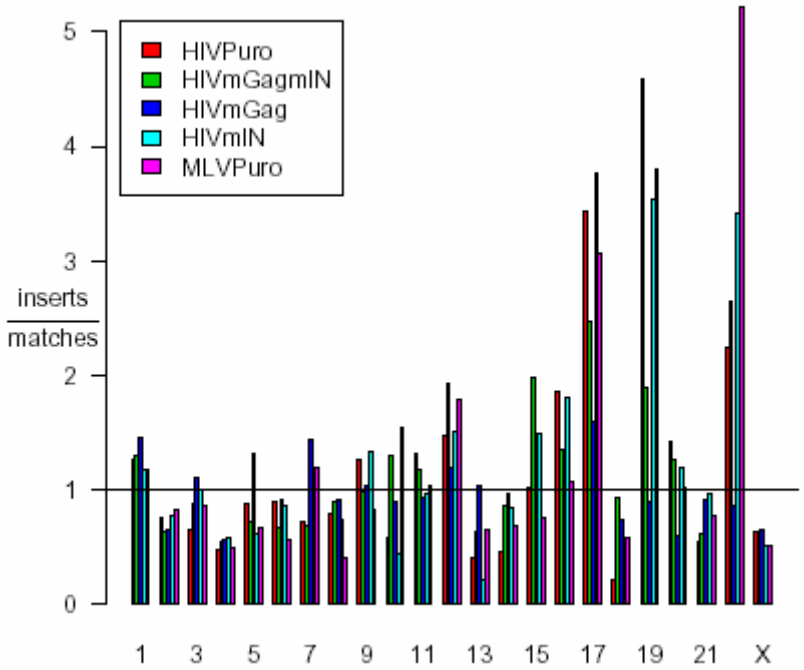
In this document, I examine the association of integration sites with various genomic features.

The data consist of both actual integration sites and sets of control sites, each set chosen to match the spacing (in bases) from the nearest restriction site (according to the direction in which the sequence was read) to an integration site. The numbers of insertion and matching sites for several data sets are shown below:

	type	
Origin.of.data.set	insertion	match
HIVPuro	525	5240
HIVmGagmIN	526	5260
HIVmGag	493	4930
HIVmIN	494	4920
MLVPuro	544	5430

The advantage of choosing 'control' sites that match the spacing from the nearest restriction site is that biases due to location and density of restriction sites are eliminated by applying the classical multinomial logit model (reviewed in (McCullagh and Nelder, 1999)). This model allows regression procedures to be applied to the study of integration intensity as a function of genomic features. The clogit function of the R survival library implements estimation and fitting for such models along with the usual likelihood ratio and Wald tests.

The distribution of relative frequency of insertions across the chromosomes is given in this barplot:

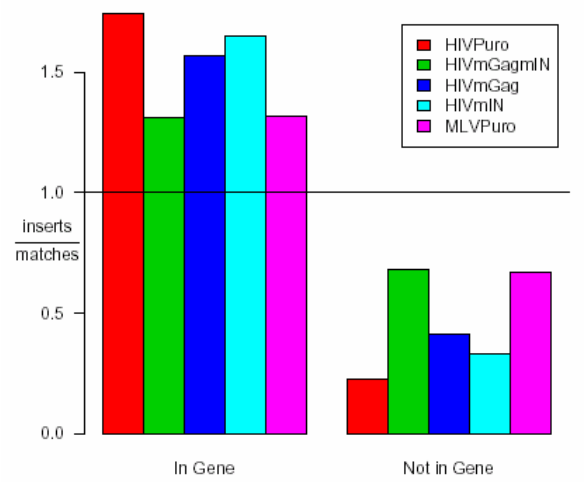


It seems evident that there are some chromosomes that are particularly favored for integration. This is reinforced by a test of statistical significance. The test performed used the likelihood ratio statistic for the multinomial logit model (reviewed in (McCullagh and Nelder, 1999)) as implemented by the clogit function of the R survival library. The null hypothesis tested is that the ratio of true integration events to matched control sites is constant across all chromosomes. This test attains a p-value of $< 2.22e{-}16$.

2 Preference for Genes

2.1 Acembly Genes

Here we examine the preference that integration events have for genes. In the following plot we show the relative frequency of integrations in genes according to the 'Acembly' annotation. The bars grouped over the label "In Gene" give the relative frequency of integration events (compared to control sites) between bases located within Acembly gene annotations, while the bars over the label "Not in Gene" give the relative frequency of integration events (compared to control sites) between bases not located within Acembly gene annotations.

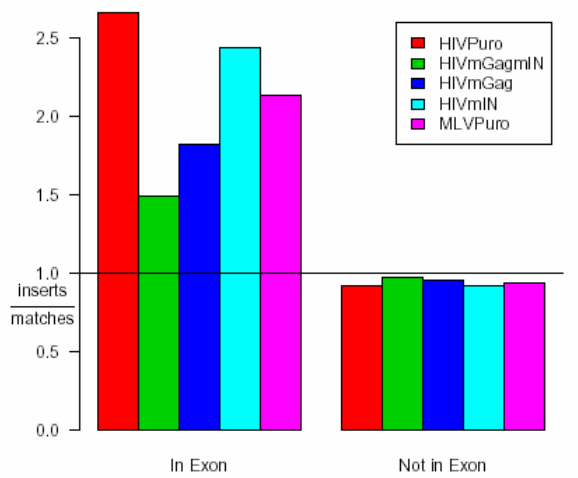


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e{-}16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e{-}16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	2.040	0.1430	14.30	3.24e-46
HIVmGagmIN	0.659	0.0966	6.82	8.93e-12
HIVmGag	1.350	0.1160	11.60	2.97e-31
HIVmIN	1.610	0.1260	12.80	2.04e-37
MLVPuro	0.688	0.0957	7.19	6.34e-13

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the HIVPuro data set, while the smallest is seen in the HIVmGagmIN data set.

In the following plot we show the relative frequency of insertions in exons according to the 'Acembly' annotation. The bars grouped over the label "In Exon" give the relative frequency of integration events (compared to control sites) between bases located in exons according to the Acembly annotation, while the bars over the label "Not in Exon" give the relative frequency of integration events (compared to control sites) between bases not located in exons according to the Acembly gene annotation.



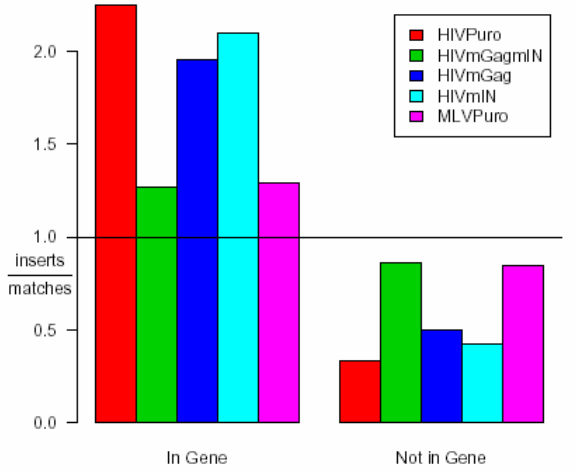
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.479	0.149	3.210	0.001310
HIVmGagmIN	0.150	0.177	0.845	0.398000
HIVmGag	0.177	0.173	1.020	0.308000
HIVmIN	0.433	0.153	2.830	0.004720
MLVPuro	0.559	0.153	3.640	0.000269

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as ”In Exon” is net of that due to being in a gene. Note that in the barplot above the ’Not in Exon’ bars include both the introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

2.2 RefGenes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the ’refGene’ annotation.



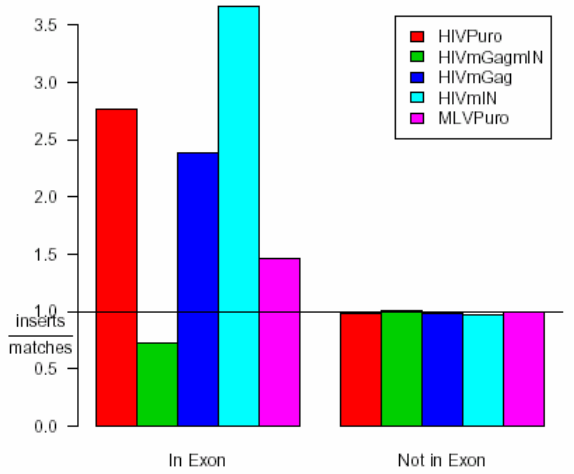
It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e{-}16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e{-}16$. Here is the table

of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	1.910	0.1100	17.30	5.44e-67
HIVmGagmIN	0.387	0.0930	4.16	3.12e-05
HIVmGag	1.380	0.1010	13.60	2.34e-42
HIVmIN	1.590	0.1060	15.10	2.12e-51
MLVPuro	0.425	0.0912	4.66	3.12e-06

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the HIVPuro data set, while the smallest is seen in the HIVmGagmIN data set.

In the following plot we show the relative frequency of insertions in exons according to the 'refGene' annotation.



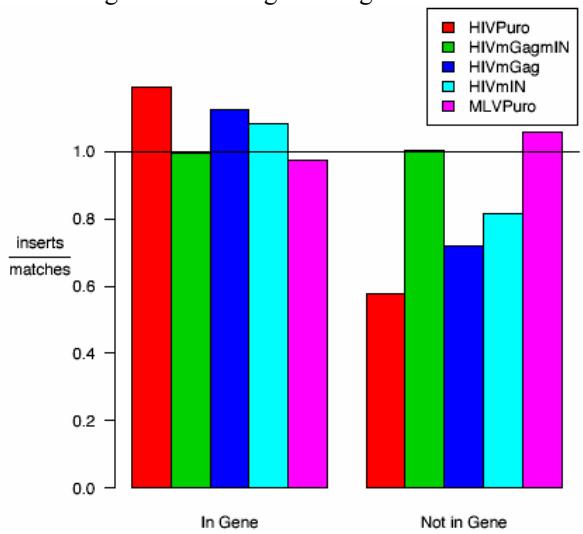
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.251	0.274	0.917	0.3590
HIVmGagmIN	-0.580	0.470	-1.230	0.2170
HIVmGag	0.212	0.286	0.741	0.4580
HIVmIN	0.604	0.244	2.470	0.0134
MLVPuro	0.130	0.316	0.410	0.6820

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as "In Exon" is net of that due to being in a gene.

2.3 GenScan Genes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'genScan' annotation.

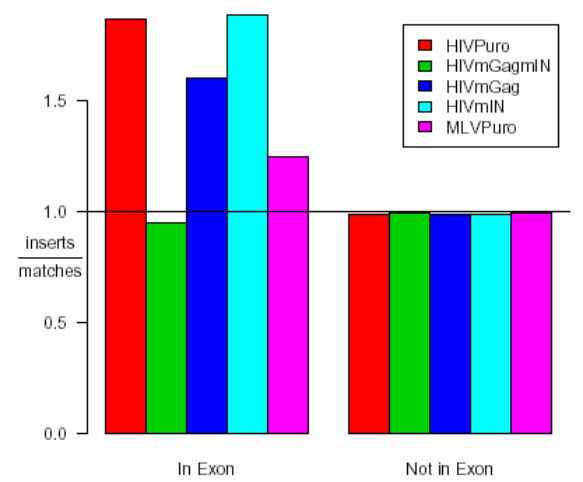


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of 1.3431e−07. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains 7.1446e − 08. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.72500	0.1180	6.1700	7.02e-10
HIVmGagmIN	-0.00615	0.0983	-0.0626	9.50e-01
HIVmGag	0.43600	0.1120	3.9100	9.34e-05
HIVmIN	0.27600	0.1090	2.5400	1.12e-02
MLVPuro	-0.08220	0.0955	-0.8610	3.89e-01

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the HIVPuro data set, while the smallest is seen in the MLVPuro data set.

In the following plot we show the relative frequency of insertions in exons according to the 'genScan' annotation.



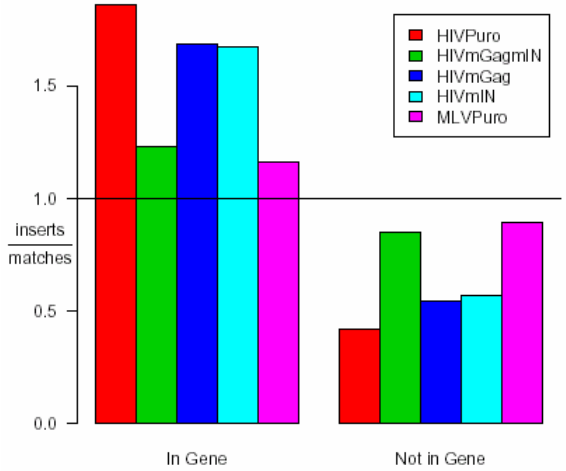
Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.4620	0.319	1.450	0.1470
HIVmGagmIN	-0.0477	0.431	-0.111	0.9120
HIVmGag	0.3630	0.364	0.998	0.3180
HIVmIN	0.5680	0.295	1.920	0.0545
MLVPuro	0.2530	0.379	0.667	0.5050

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as ”In Exon” is net of that due to being in a gene.

2.4 UniGenes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the ’uniGene’ annotation.

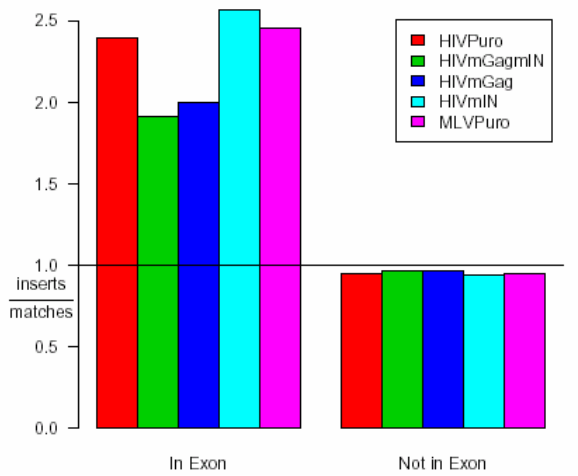


It seems evident that there is a strong tendency for insertions to occur in genes. A formal test of significance bears this out with a p-value of $< 2.22e{-}16$. Also, it appears that the tendency of different viruses to integrate into genes varies, and a test for this hypothesis attains $< 2.22e{-}16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	1.480	0.1050	14.20	1.77e-45
HIVmGagmIN	0.369	0.0919	4.02	5.94e-05
HIVmGag	1.140	0.1010	11.30	1.84e-29
HIVmIN	1.090	0.1000	10.90	1.44e-27
MLVPuro	0.266	0.0901	2.95	3.14e-03

As is evident, there are some differences in the coefficients. The largest coefficient is seen in the HIVPuro data set, while the smallest is seen in the MLVPuro data set.

In the following plot we show the relative frequency of insertions in exons according to the ’uniGene’ annotation.



Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.270	0.184	1.46	1.44e-01
HIVmGagmIN	0.499	0.210	2.37	1.76e-02
HIVmGag	0.230	0.196	1.17	2.42e-01
HIVmIN	0.515	0.178	2.89	3.88e-03
MLVPuro	0.853	0.183	4.65	3.31e-06

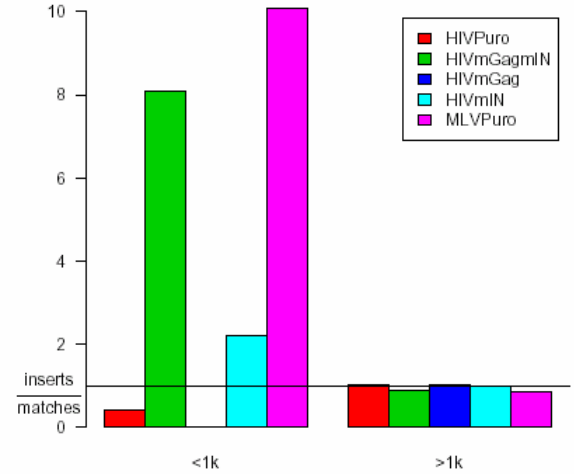
The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as ”In Exon” is net of that due to being in a gene.

3 CpG Island Neighborhoods

Here we study the effect of being in the neighborhood of CpG Islands. Following Wu and colleagues (Wu, *et al.*, 2003), who found that the neighborhoods within ±1kb of CpG islands are enriched for MLV insertions, we study such neighborhoods.

3.1 1 kilobase neighborhoods

The following plot shows the effect of being in or within ±1kb of a CpG island:



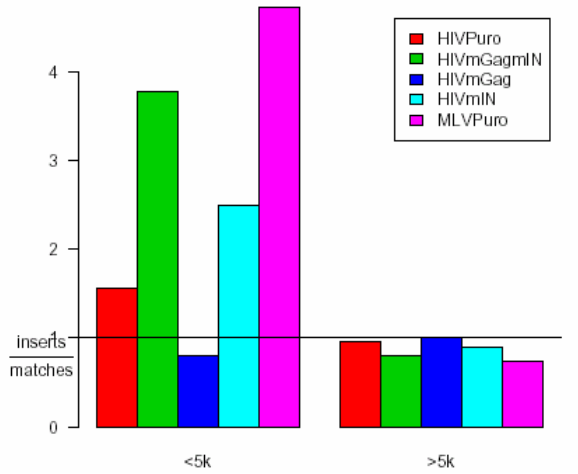
A formal test of significance comparing the difference attains a p-value of < 2.22e−16. A test for differences between viruses attains < 2.22e−16. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	-0.896	0.591	-1.5200	1.29e-01
HIVmGagmIN	2.200	0.172	12.8000	1.14e-37
HIVmGag	-14.700	555.000	-0.0264	9.79e-01
HIVmIN	0.823	0.260	3.1700	1.54e-03
MLVPuro	2.460	0.156	15.8000	4.01e-56

The largest coefficient is seen in the MLVPuro data set, while the smallest is seen in the HIVmGag data set.

3.2 5 kilobase neighborhoods

The following plot shows the effect of being in or within ±5kb of a CpG island:



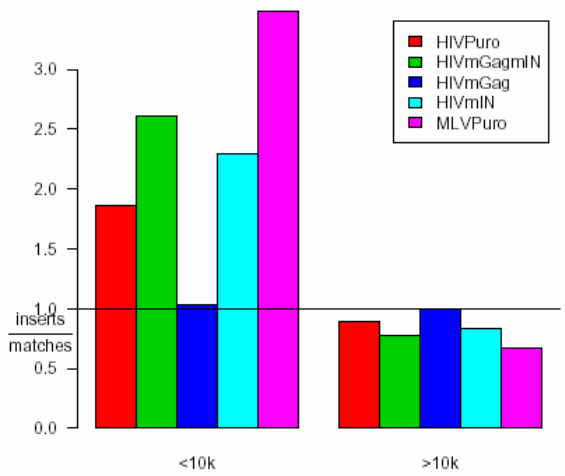
A formal test of significance comparing the difference attains a p-value of < 2.22e−16. A test for differences between viruses attains < 2.22e−16. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.477	0.164	2.91	3.63e-03
HIVmGagmIN	1.560	0.116	13.40	4.45e-41
HIVmGag	-0.233	0.209	-1.12	2.65e-01
HIVmIN	1.020	0.135	7.58	3.33e-14
MLVPuro	1.870	0.110	16.90	2.53e-64

The largest coefficient is seen in the MLVPuro data set, while the smallest is seen in the HIVmGag data set.

3.3 10 kilobase neighborhoods

The following plot shows the effect of being in or within ±10kb of a CpG island:



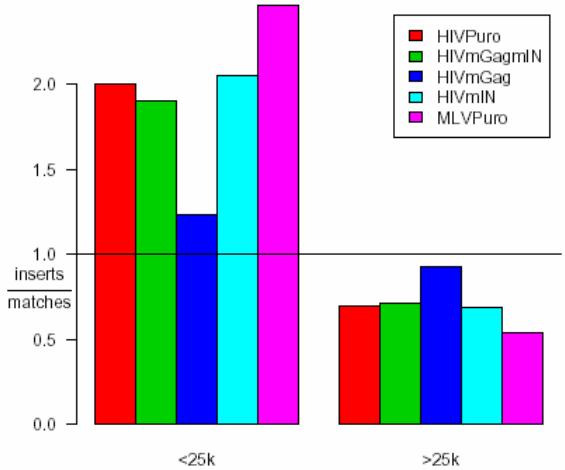
A formal test of significance comparing the difference attains a p-value of < 2.22e−16. A test for differences between viruses attains < 2.22e−16. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	0.7290	0.1160	6.28 3.	46e-10
HIVmGagmIN	1.2100	0.1020	11.80	3.26e-32
HIVmGag	0.0338	0.1470	0.23	8.18e-01
HIVmIN	1.0000	0.1130	8.87	7.34e-19
MLVPuro	1.6700	0.0994	16.80	4.78e-63

The largest coefficient is seen in the MLVPuro data set, while the smallest is seen in the HIVmGag data set.

3.4 25 kilobase neighborhoods

The following plot shows the effect of being in or within ±25kb of a CpG island:



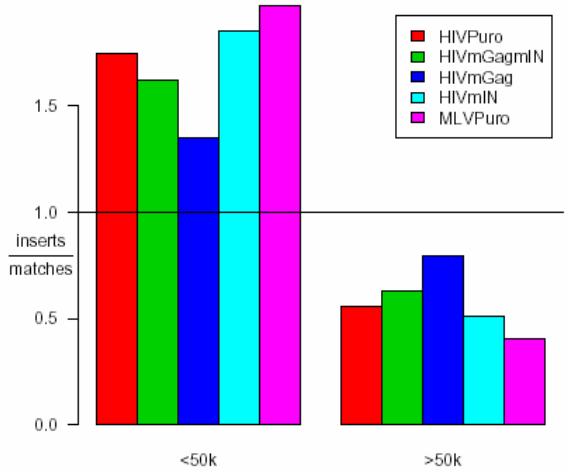
A formal test of significance comparing the difference attains a p-value of < 2.22e−16. A test for differences between viruses attains < 2.22e−16. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	1.050	0.0941	11.2	3.95e-29
HIVmGagmIN	0.985	0.0933	10.6	4.25e-26
HIVmGag	0.283	0.1050	2.7	6.85e-03
HIVmIN	1.090	0.0968	11.2	3.18e-29
MLVPuro	1.540	0.0948	16.3	1.33e-59

The largest coefficient is seen in the MLVPuro data set, while the smallest is seen in the HIVmGag data set.

3.5 50 kilobase neighborhoods

The following plot shows the effect of being in or within ±50kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of $< 2.22e{-}16$. A test for differences between viruses attains $8.0065e{-}14$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites along with their standard errors, z statistics, and p-values for each data set:

	coef	se	z	p
HIVPuro	1.160	0.0970	12.00	5.56e-33
HIVmGagmIN	0.946	0.0937	10.10	5.97e-24
HIVmGag	0.532	0.0948	5.62	1.95e-08
HIVmIN	1.290	0.1020	12.70	3.45e-37
MLVPuro	1.610	0.1050	15.40	2.44e-53

The largest coefficient is seen in the MLVPuro data set, while the smallest is seen in the HIVmGag data set.

4 Gene Density, Expression 'Density', and CpG Island Density

In this section the association with gene density is examined. The 'genes' that are counted are the genes represented on the Affymetrix HU133A microarray. In addition, we count the number of such genes expressed at various levels. The levels are:

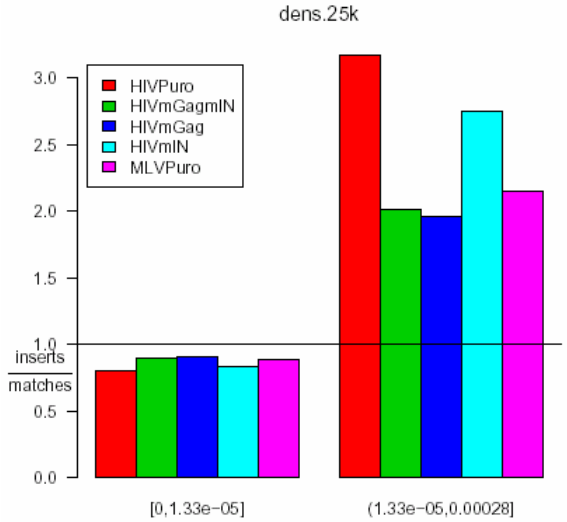
low.ex Count genes whose expression is in the upper half and divide by number of bases
med.ex Count genes whose expression is in the upper 1/8th and divide by number of bases
high.ex Count genes whose expression is in the upper 1/16th and divide by number of bases

The bolded terms are used as abbreviations in what follows. The abbreviation **dens** is used to indicate gene density as number of genes per base.

4.1 25 kilobase window

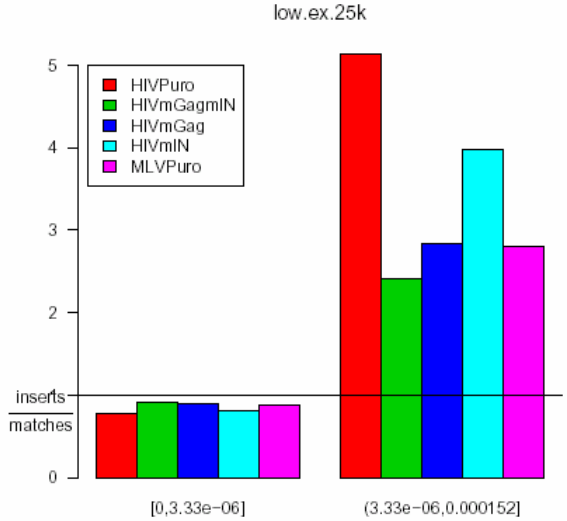
In the barplot that follows we examine the association of insertion sites with gene density in a 25 kilobase window surrounding each locus. More such plots will follow and the method of their construction is always to try to divide the data according to the deciles of density. However, it often happens that there is a very skewed distribution of density and even the 90th percentile is zero. In that case, the barplots simply show the sites for which the density is zero and those for which it is non-zero. If there are fewer than ten groups of bars, the groupings contain ten percent of the sites each except for the leftmost grouping which will contain all of the remaining sites.

Also note that the title of the plot contains clues as to its content; the prefix indicates the type of variable studied while the suffix indicates the window width in the number of bases. The p-value given is the result of fitting a cubic polynomial to the gene density values.



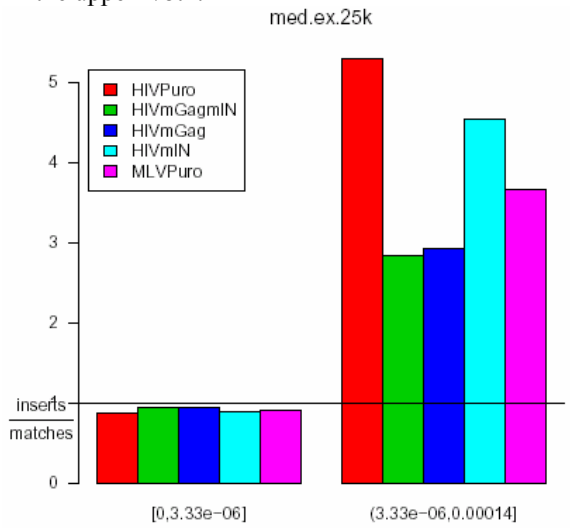
	coef	se	z	p
HIVPuro	1.460	0.103	14.10	6.11e-45
HIVmGagmIN	0.729	0.115	6.36	2.01e-10
HIVmGag	0.870	0.118	7.39	1.49e-13
HIVmIN	1.230	0.109	11.20	2.56e-29
MLVPuro	0.974	0.107	9.12	7.77e-20

Here are the results for expression density. First, we count just genes that are in the upper half.



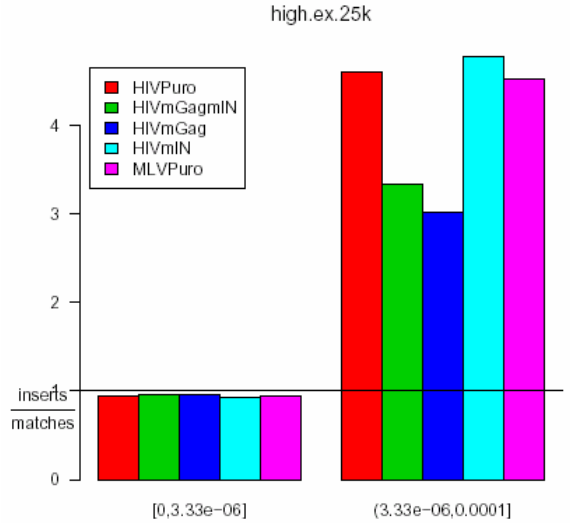
	coef	se	z	p
HIVPuro	1.910	0.119	16.00	8.42e-58
HIVmGagmIN	0.968	0.136	7.14	9.36e-13
HIVmGag	1.160	0.137	8.52	1.59e-17
HIVmIN	1.570	0.123	12.80	1.28e-37
MLVPuro	1.150	0.125	9.21	3.20e-20

Now we count genes in the upper 1/8th:



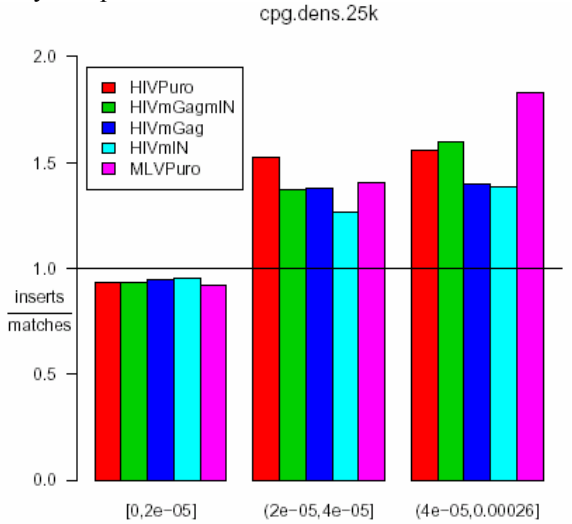
	coef	se	z	p
HIVPuro	1.80	0.145	12.40	3.01e-35
HIVmGagmIN	1.10	0.168	6.53	6.42e-11
HIVmGag	1.15	0.176	6.52	6.95e-11
HIVmIN	1.65	0.152	10.90	1.78e-27
MLVPuro	1.40	0.157	8.93	4.42e-19

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.58	0.197	8.03	1.00e-15
HIVmGagmIN	1.26	0.218	5.77	8.11e-09
HIVmGag	1.14	0.234	4.88	1.07e-06
HIVmIN	1.63	0.198	8.25	1.62e-16
MLVPuro	1.57	0.205	7.66	1.83e-14

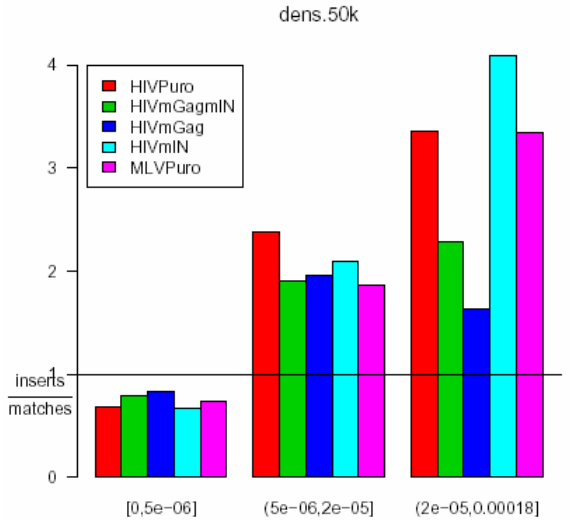
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.469	0.0945	4.96	6.95e-07
HIVmGagmIN	0.649	0.0936	6.93	4.15e-12
HIVmGag	0.448	0.0965	4.64	3.44e-06
HIVmIN	0.370	0.0981	3.77	1.61e-04
MLVPuro	0.634	0.0914	6.94	3.99e-12

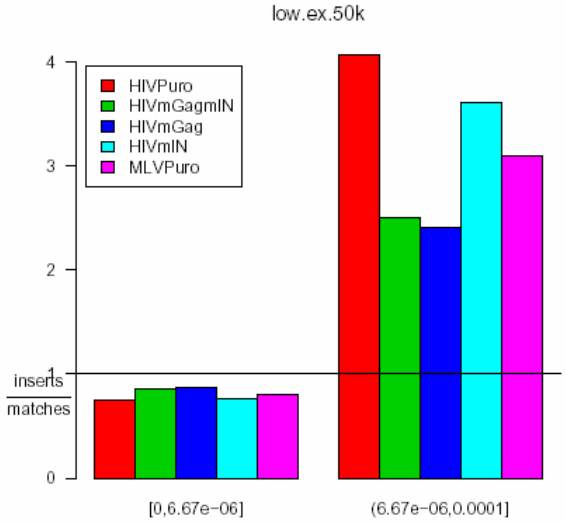
4.2 50 kilobase window

In the barplot that follows we examine the association of insertion sites with expression density in a 50 kilobase window surrounding each locus. First, we count just the number of genes represented on the chip.



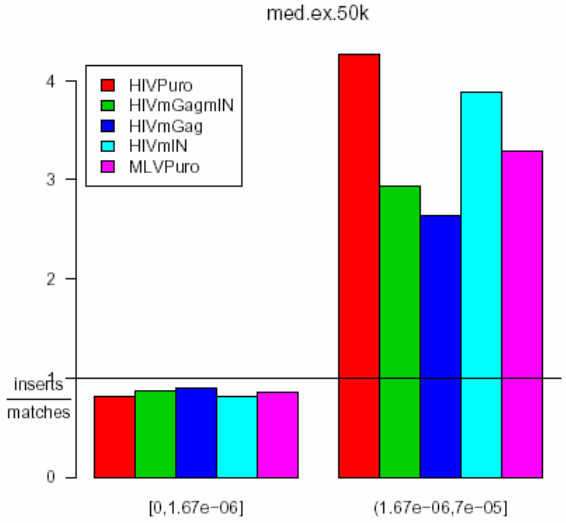
	coef	se	z	p
HIVPuro	1.500	0.0947	15.90	1.02e-56
HIVmGagmIN	0.941	0.0962	9.78	1.40e-22
HIVmGag	0.900	0.1010	8.89	6.35e-19
HIVmIN	1.370	0.0978	14.00	2.11e-44
MLVPuro	1.150	0.0932	12.30	8.07e-35

Here are the results for expression density. First, we count just genes that are in the upper half.



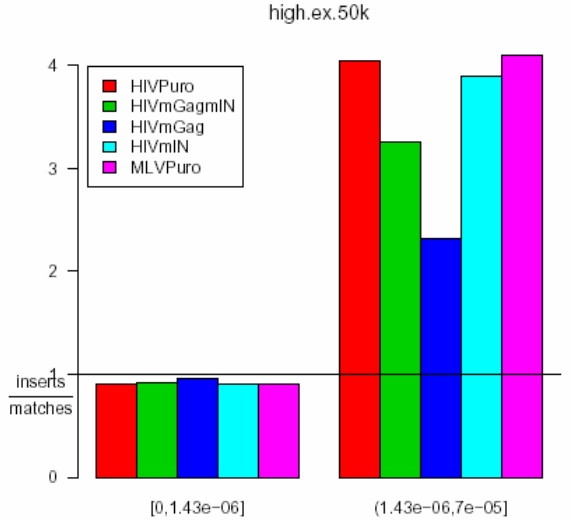
	coef	se	z	p
HIVPuro	1.83	0.102	18.00	2.52e-72
HIVmGagmIN	1.07	0.108	9.89	4.72e-23
HIVmGag	1.16	0.112	10.40	3.43e-25
HIVmIN	1.60	0.105	15.30	1.56e-52
MLVPuro	1.30	0.102	12.70	3.88e-37

Now we count genes in the upper 1/8th:



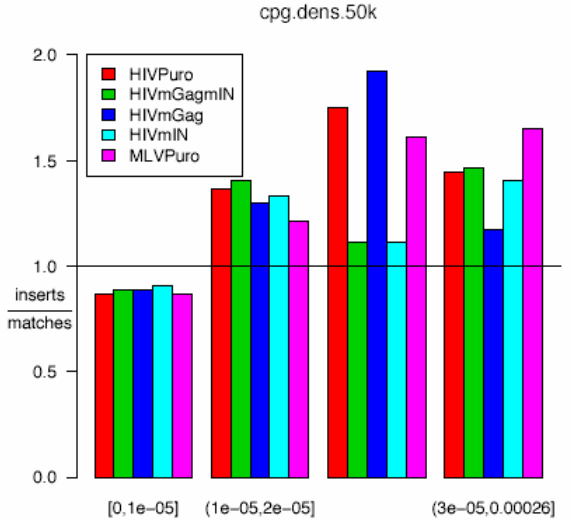
	coef	se	z	p
HIVPuro	1.63	0.118	13.80	3.93e-43
HIVmGagmIN	1.20	0.127	9.50	2.14e-21
HIVmGag	1.08	0.139	7.75	9.52e-15
HIVmIN	1.55	0.123	12.60	2.61e-36
MLVPuro	1.35	0.122	11.00	2.96e-28

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.490	0.154	9.70	3.09e-22
HIVmGagmIN	1.270	0.163	7.81	5.73e-15
HIVmGag	0.887	0.192	4.63	3.70e-06
HIVmIN	1.450	0.154	9.41	4.78e-21
MLVPuro	1.510	0.152	9.89	4.64e-23

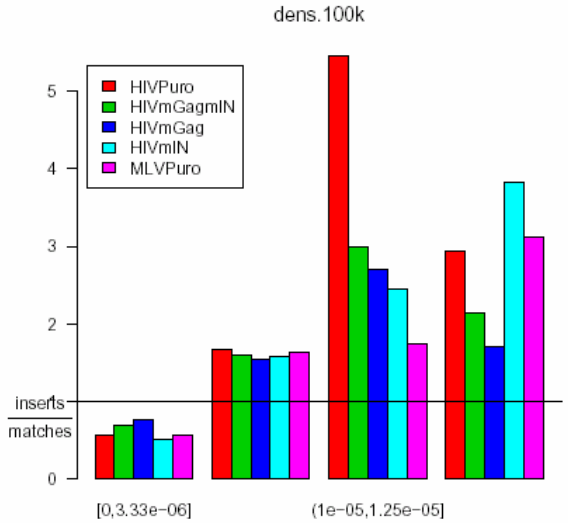
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.580	0.0934	6.22	5.04e-10
HIVmGagmIN	0.636	0.0933	6.82	9.29e-12
HIVmGag	0.581	0.0958	6.06	1.33e-09
HIVmIN	0.499	0.0956	5.22	1.81e-07
MLVPuro	0.678	0.0913	7.43	1.10e-13

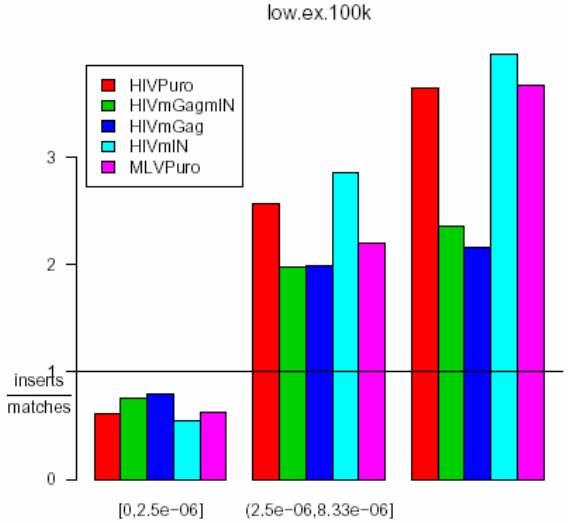
4.3 100 kilobase window

In the barplot that follows we examine the association of insertion sites with expression density in a 100 kilobase window surrounding each locus. First, we count just the number of genes represented on the chip.



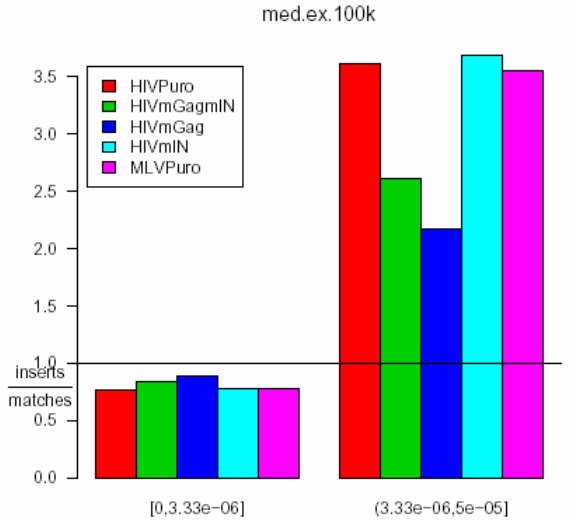
	coef	se	z	p
HIVPuro	1.530	0.0988	15.50	3.72e-54
HIVmGagmIN	0.995	0.0928	10.70	7.63e-27
HIVmGag	0.893	0.0955	9.35	8.82e-21
HIVmIN	1.560	0.1020	15.30	1.51e-52
MLVPuro	1.400	0.0963	14.50	9.06e-48

Here are the results for expression density. First, we count just genes that are in the upper half.



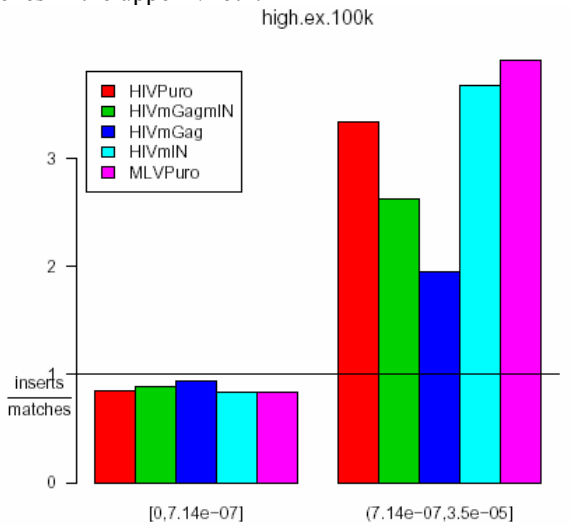
	coef	se	z	p
HIVPuro	1.78	0.0967	18.4	1.07e-75
HIVmGagmIN	1.08	0.0947	11.4	2.67e-30
HIVmGag	1.08	0.0991	10.9	1.41e-27
HIVmIN	1.79	0.0994	18.0	1.03e-72
MLVPuro	1.61	0.0939	17.1	8.91e-66

Now we count genes in the upper 1/8th:



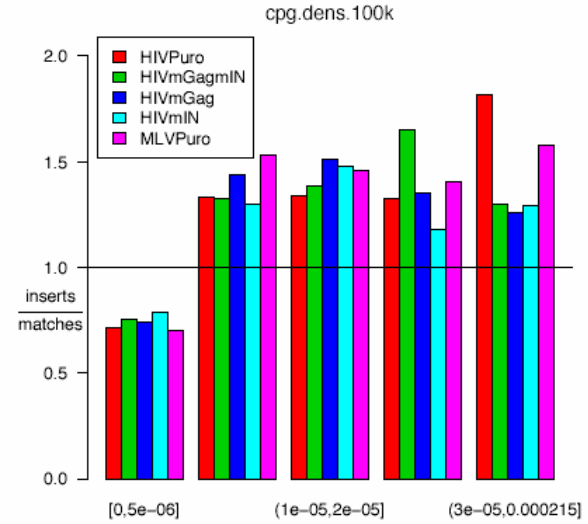
	coef	se	z	p
HIVPuro	1.600	0.102	15.60	5.47e-55
HIVmGagmIN	1.120	0.107	10.50	9.88e-26
HIVmGag	0.948	0.118	8.05	7.98e-16
HIVmIN	1.570	0.106	14.80	7.00e-50
MLVPuro	1.500	0.102	14.70	3.89e-49

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.360	0.124	11.00	3.07e-28
HIVmGagmIN	1.080	0.132	8.17	3.12e-16
HIVmGag	0.737	0.155	4.77	1.84e-06
HIVmIN	1.460	0.127	11.50	1.47e-30
MLVPuro	1.550	0.121	12.80	2.15e-37

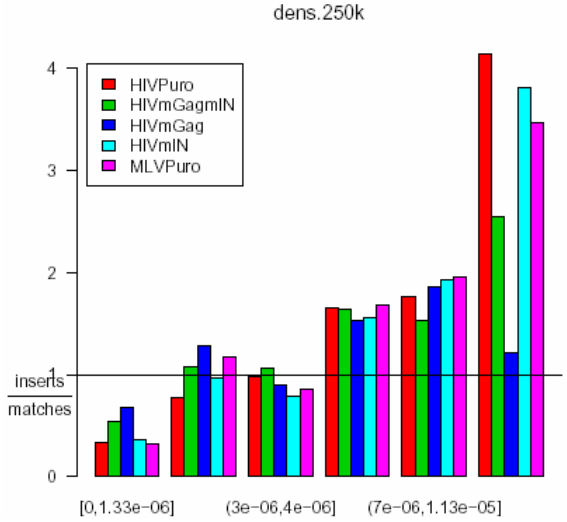
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.710	0.0929	7.64	2.21e-14
HIVmGagmIN	0.609	0.0923	6.60	4.17e-11
HIVmGag	0.637	0.0948	6.71	1.90e-11
HIVmIN	0.527	0.0946	5.58	2.45e-08
MLVPuro	0.762	0.0907	8.40	4.65e-17

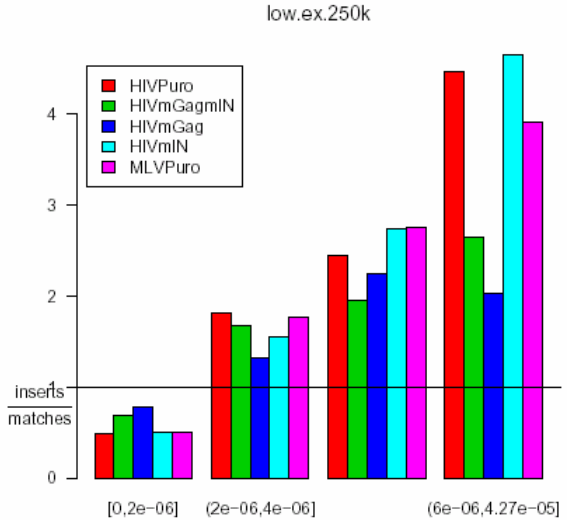
4.4 250 kilobase window

In the barplot that follows we examine the association of insertion sites with expression density in a 250 kilobase window surrounding each locus. First, we count just the number of genes represented on the chip.



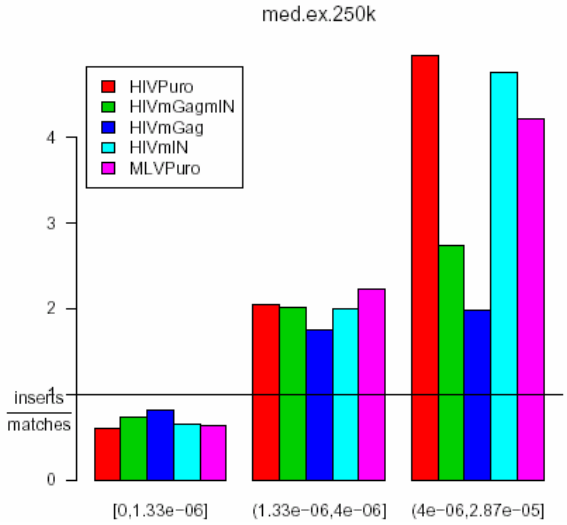
	coef	se	z	p
HIVPuro	1.630	0.1180	13.80	2.24e-43
HIVmGagmIN	1.060	0.1020	10.40	2.31e-25
HIVmGag	0.672	0.0986	6.82	9.39e-12
HIVmIN	1.610	0.1180	13.70	1.99e-42
MLVPuro	1.750	0.1210	14.50	1.86e-47

Here are the results for expression density. First, we count just genes that are in the upper half.



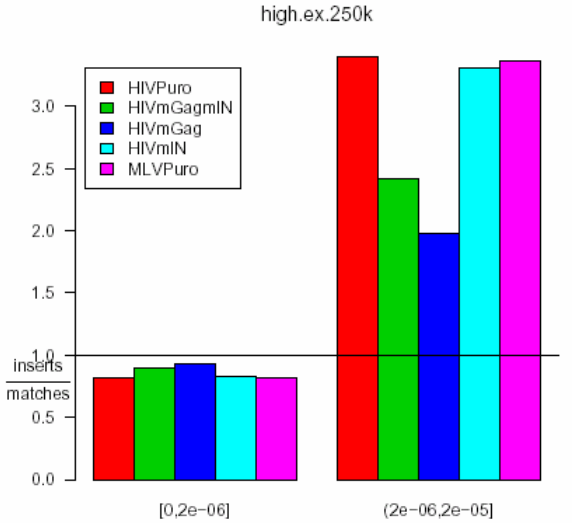
	coef	se	z	p
HIVPuro	1.850	0.1090	17.0	1.17e-64
HIVmGagmIN	1.270	0.0968	13.1	2.53e-39
HIVmGag	0.983	0.0961	10.2	1.48e-24
HIVmIN	1.910	0.1130	17.0	8.29e-65
MLVPuro	1.910	0.1090	17.5	2.93e-68

Now we count genes in the upper 1/8th:



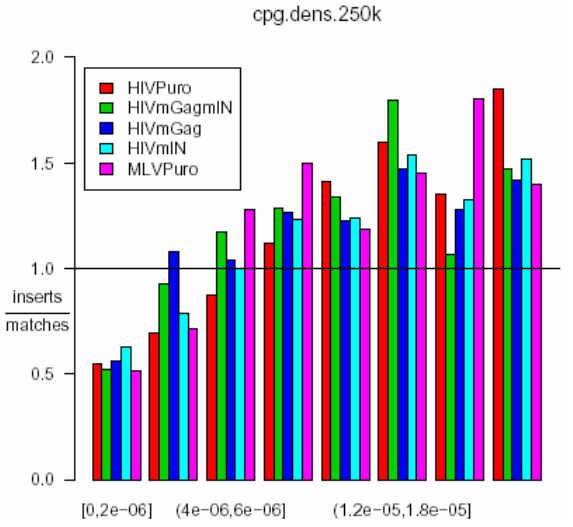
	coef	se	z	p
HIVPuro	1.650	0.0952	17.30	5.31e-67
HIVmGagmIN	1.170	0.0939	12.50	1.30e-35
HIVmGag	0.845	0.0975	8.66	4.53e-18
HIVmIN	1.630	0.1000	16.30	9.82e-60
MLVPuro	1.670	0.0957	17.40	7.23e-68

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.360	0.0991	13.70	8.40e-43
HIVmGagmIN	1.130	0.1020	11.00	3.66e-28
HIVmGag	0.733	0.1130	6.47	9.64e-11
HIVmIN	1.390	0.1030	13.50	2.10e-41
MLVPuro	1.470	0.0973	15.20	7.28e-52

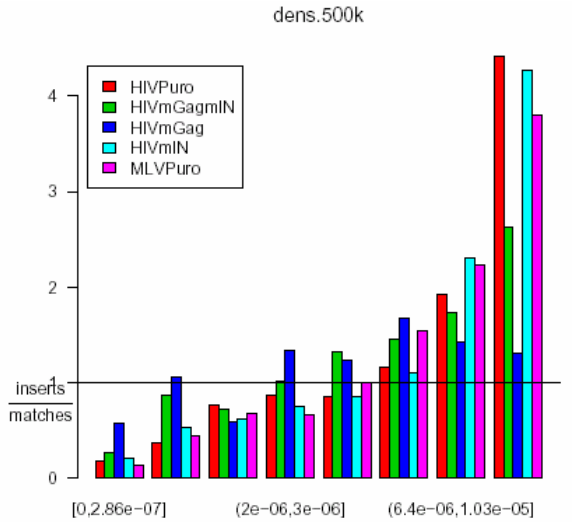
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.859	0.0959	8.96	3.37e-19
HIVmGagmIN	0.690	0.0933	7.40	1.39e-13
HIVmGag	0.594	0.0954	6.23	4.77e-10
HIVmIN	0.646	0.0962	6.71	1.92e-11
MLVPuro	0.771	0.0922	8.36	6.21e-17

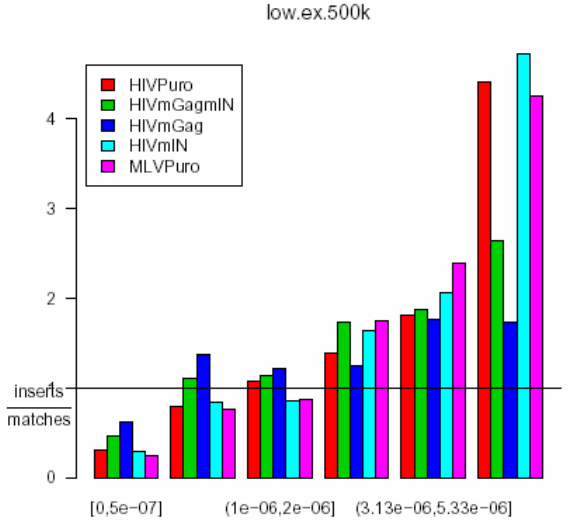
4.5 500 kilobase window

In the barplot that follows we examine the association of insertion sites with expression density in a 500 kilobase window surrounding each locus. First, we count just the number of genes represented on the chip.



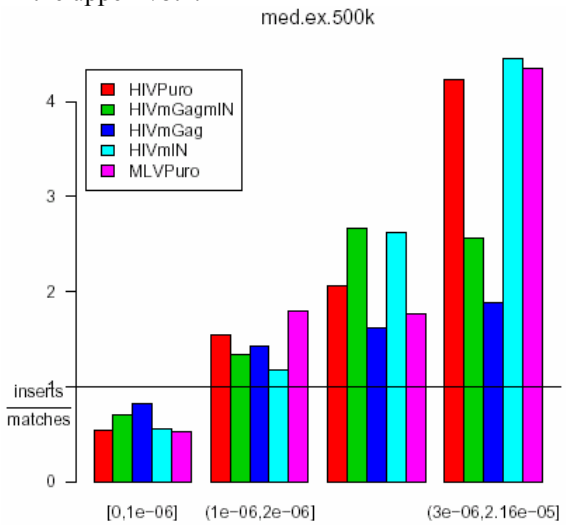
	coef	se	z	p
HIVPuro	1.630	0.1150	14.20	1.39e-45
HIVmGagmIN	1.190	0.1020	11.60	3.04e-31
HIVmGag	0.742	0.0979	7.58	3.58e-14
HIVmIN	1.610	0.1160	14.00	3.11e-44
MLVPuro	1.780	0.1180	15.10	3.02e-51

Here are the results for expression density. First, we count just genes that are in the upper half.



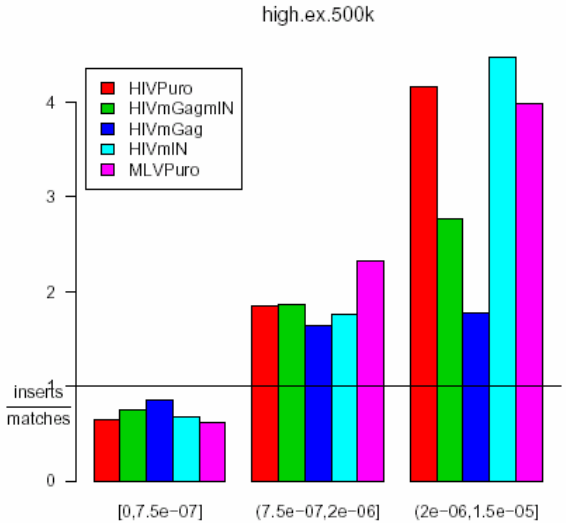
	coef	se	z	p
HIVPuro	1.820	0.1220	14.90	2.27e-50
HIVmGagmIN	1.290	0.1060	12.20	1.71e-34
HIVmGag	0.848	0.0993	8.54	1.31e-17
HIVmIN	1.840	0.1250	14.70	4.40e-49
MLVPuro	2.030	0.1300	15.60	4.02e-55

Now we count genes in the upper 1/8th:



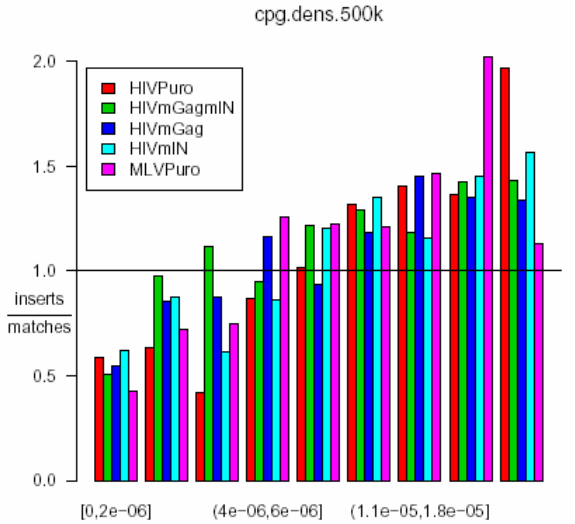
	coef	se	z	p
HIVPuro	1.580	0.1020	15.50	3.46e-54
HIVmGagmIN	1.300	0.0975	13.30	1.25e-40
HIVmGag	0.788	0.0952	8.28	1.24e-16
HIVmIN	1.610	0.1060	15.20	4.90e-52
MLVPuro	1.880	0.1080	17.40	1.15e-67

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.360	0.0937	14.50	2.22e-47
HIVmGagmIN	1.130	0.0938	12.00	2.04e-33
HIVmGag	0.655	0.0993	6.59	4.45e-11
HIVmIN	1.380	0.0976	14.10	3.30e-45
MLVPuro	1.580	0.0933	17.00	1.75e-64

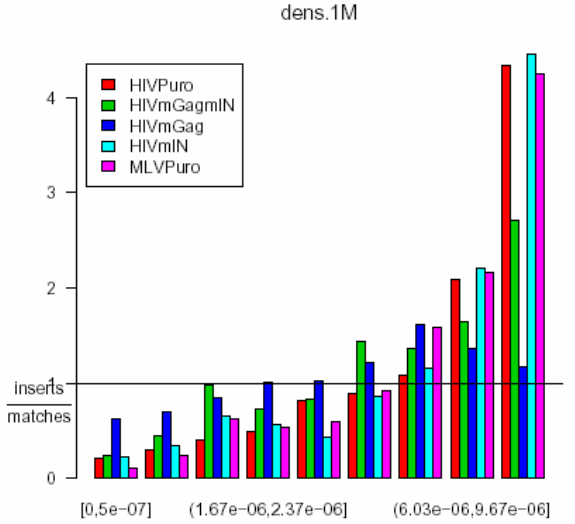
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.798	0.0968	8.24	1.71e-16
HIVmGagmIN	0.581	0.0937	6.20	5.69e-10
HIVmGag	0.506	0.0957	5.28	1.26e-07
HIVmIN	0.643	0.0971	6.62	3.63e-11
MLVPuro	0.737	0.0934	7.89	3.08e-15

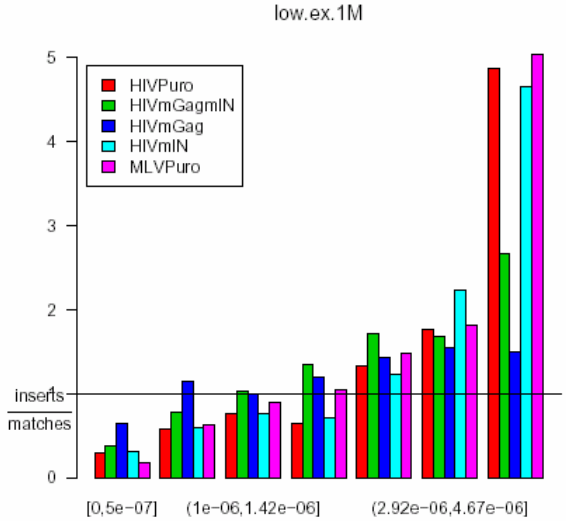
4.6 1 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in a 1 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



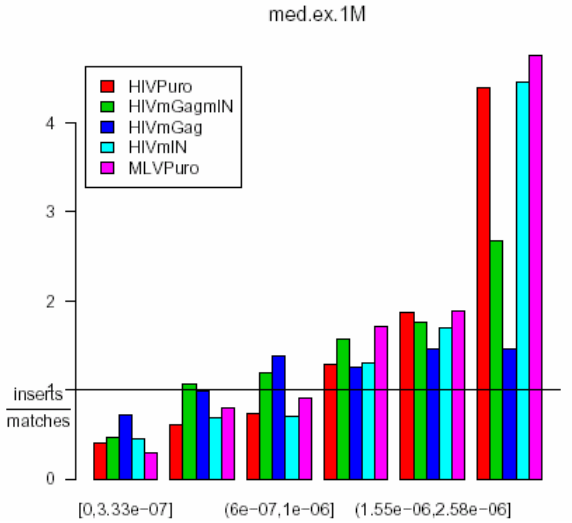
	coef	se	z	p
HIVPuro	1.72	0.1210	14.2	1.56e-45
HIVmGagmIN	1.14	0.1040	11.0	2.94e-28
HIVmGag	0.53	0.0963	5.5	3.75e-08
HIVmIN	1.50	0.1150	13.0	1.13e-38
MLVPuro	1.77	0.1210	14.6	2.57e-48

Here are the results for expression density. First, we count just genes that are in the upper half.



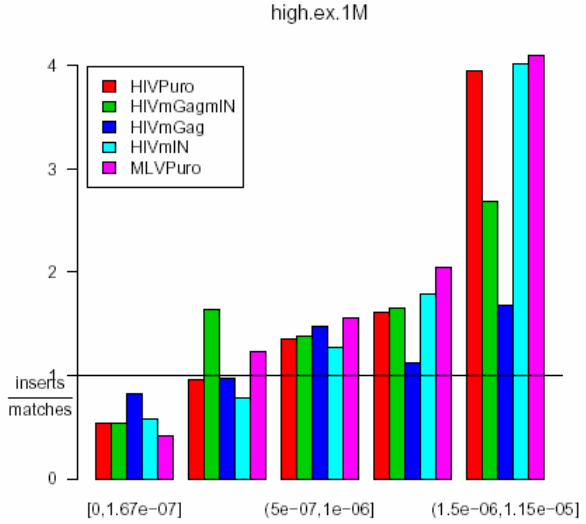
	coef	se	z	p
HIVPuro	1.660	0.1130	14.60	1.86e-48
HIVmGagmIN	1.310	0.1030	12.70	1.10e-36
HIVmGag	0.601	0.0956	6.29	3.18e-10
HIVmIN	1.630	0.1150	14.20	1.39e-45
MLVPuro	1.990	0.1230	16.20	9.09e-59

Now we count genes in the upper 1/8th:



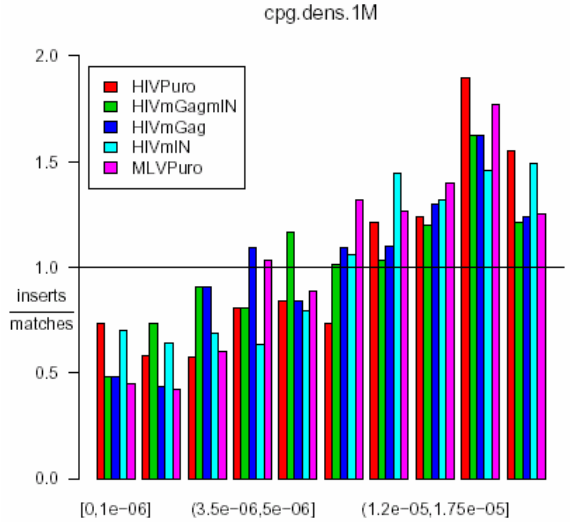
	coef	se	z	p
HIVPuro	1.410	0.110	12.70	4.15e-37
HIVmGagmIN	1.250	0.105	11.90	1.73e-32
HIVmGag	0.595	0.097	6.13	8.82e-10
HIVmIN	1.290	0.109	11.80	5.02e-32
MLVPuro	1.840	0.123	14.90	2.30e-50

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.190	0.0969	12.20	1.69e-34
HIVmGagmIN	1.190	0.0973	12.30	1.43e-34
HIVmGag	0.426	0.0951	4.48	7.49e-06
HIVmIN	1.120	0.0998	11.20	5.70e-29
MLVPuro	1.570	0.1030	15.30	9.83e-53

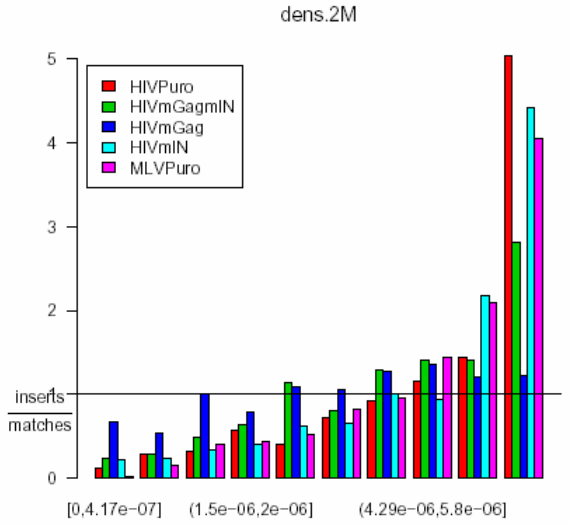
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.626	0.0943	6.63	3.31e-11
HIVmGagmIN	0.417	0.0923	4.51	6.40e-06
HIVmGag	0.511	0.0957	5.34	9.35e-08
HIVmIN	0.682	0.0982	6.95	3.67e-12
MLVPuro	0.734	0.0930	7.90	2.86e-15

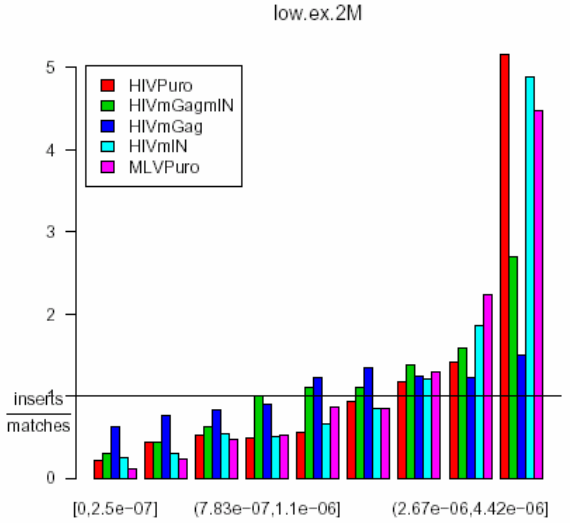
4.7 2 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in a 2 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



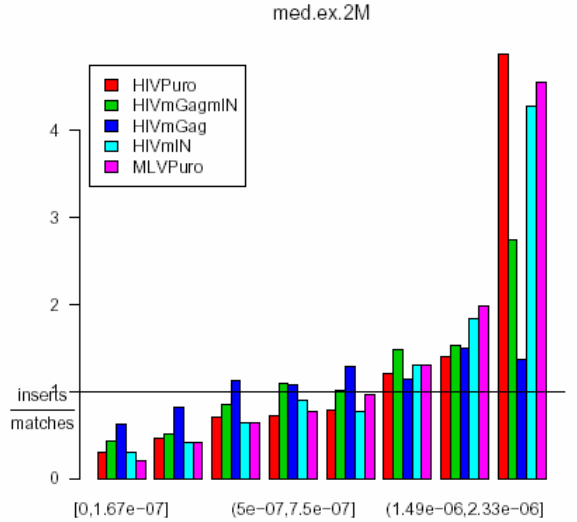
	coef	se	z	p
HIVPuro	1.640	0.1180	13.90	9.90e-44
HIVmGagmIN	1.000	0.1000	10.00	1.41e-23
HIVmGag	0.431	0.0956	4.51	6.62e-06
HIVmIN	1.580	0.1190	13.30	1.81e-40
MLVPuro	1.780	0.1210	14.70	1.10e-48

Here are the results for expression density. First, we count just genes that are in the upper half.



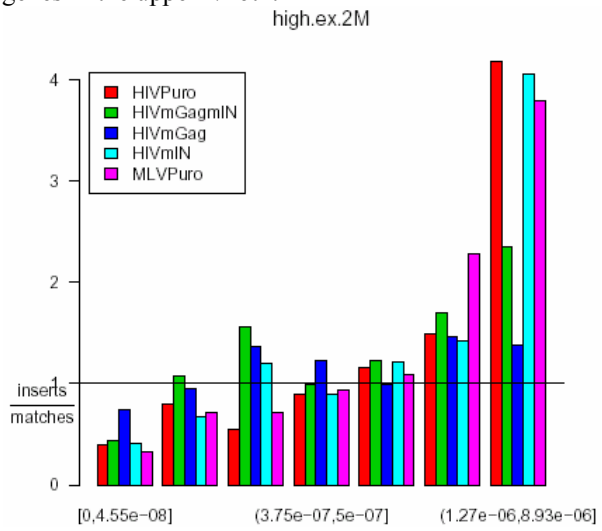
	coef	se	z	p
HIVPuro	1.560	0.1160	13.50	1.27e-41
HIVmGagmIN	1.100	0.1020	10.80	5.02e-27
HIVmGag	0.566	0.0961	5.89	3.82e-09
HIVmIN	1.590	0.1200	13.30	3.38e-40
MLVPuro	1.850	0.1240	14.90	6.15e-50

Now we count genes in the upper 1/8th:



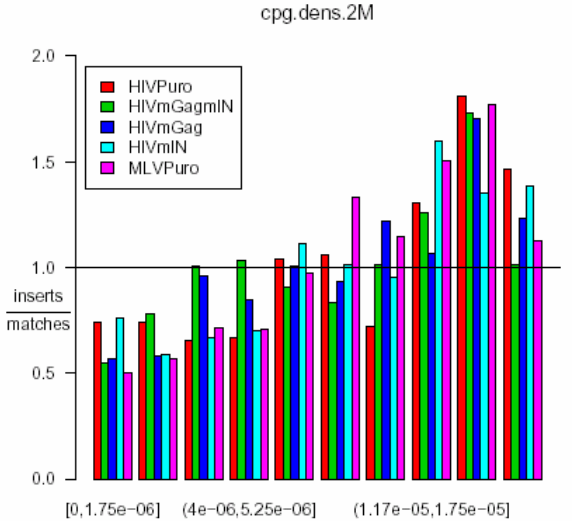
	coef	se	z	p
HIVPuro	1.410	0.1080	13.10	4.10e-39
HIVmGagmIN	1.070	0.0998	10.70	1.06e-26
HIVmGag	0.496	0.0958	5.18	2.24e-07
HIVmIN	1.480	0.1130	13.10	2.43e-39
MLVPuro	1.640	0.1140	14.40	4.35e-47

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.190	0.0993	12.00	2.47e-33
HIVmGagmIN	0.933	0.0949	9.83	8.45e-23
HIVmGag	0.458	0.0948	4.84	1.33e-06
HIVmIN	1.270	0.1040	12.20	1.95e-34
MLVPuro	1.400	0.1010	13.80	1.33e-43

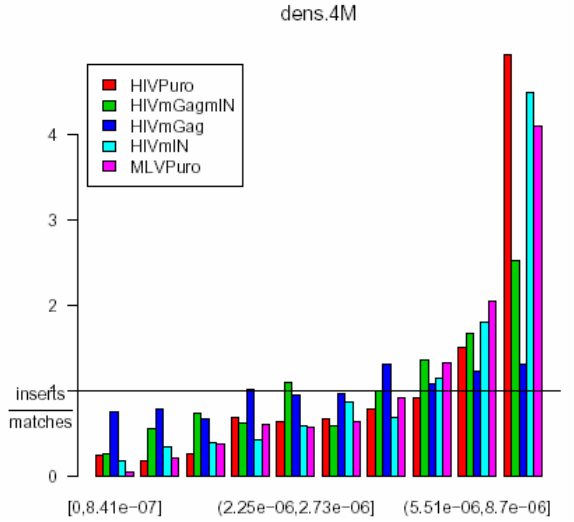
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.503	0.0944	5.33	9.77e-08
HIVmGagmIN	0.309	0.0921	3.35	8.00e-04
HIVmGag	0.449	0.0962	4.67	3.04e-06
HIVmIN	0.505	0.0973	5.19	2.12e-07
MLVPuro	0.699	0.0935	7.48	7.52e-14

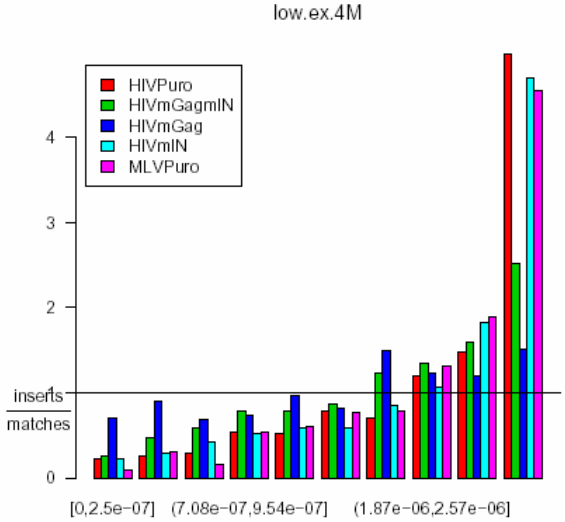
4.8 4 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in a 4 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



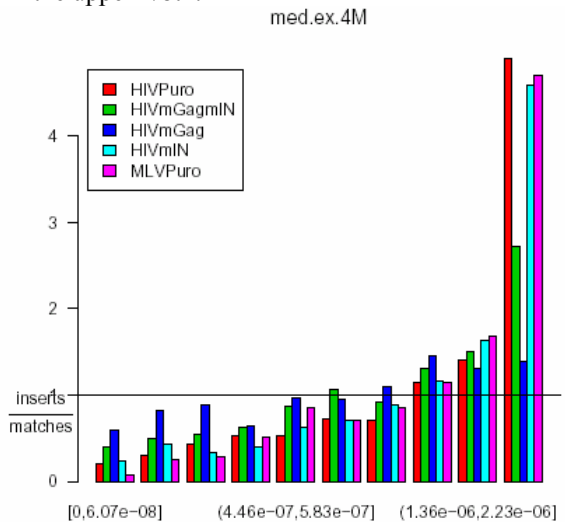
	coef	se	z	p
HIVPuro	1.400	0.1120	12.50	1.11e-35
HIVmGagmIN	0.760	0.0961	7.91	2.59e-15
HIVmGag	0.336	0.0947	3.55	3.86e-04
HIVmIN	1.480	0.1160	12.80	2.13e-37
MLVPuro	1.510	0.1120	13.40	3.51e-41

Here are the results for expression density. First, we count just genes that are in the upper half.



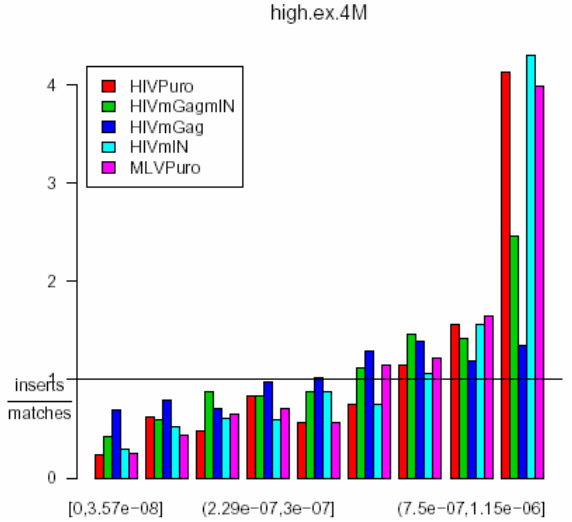
	coef	se	z	p
HIVPuro	1.530	0.1160	13.20	8.35e-40
HIVmGagmIN	0.964	0.0996	9.68	3.68e-22
HIVmGag	0.445	0.0957	4.65	3.25e-06
HIVmIN	1.420	0.1140	12.50	1.01e-35
MLVPuro	1.630	0.1170	14.00	1.98e-44

Now we count genes in the upper 1/8th:



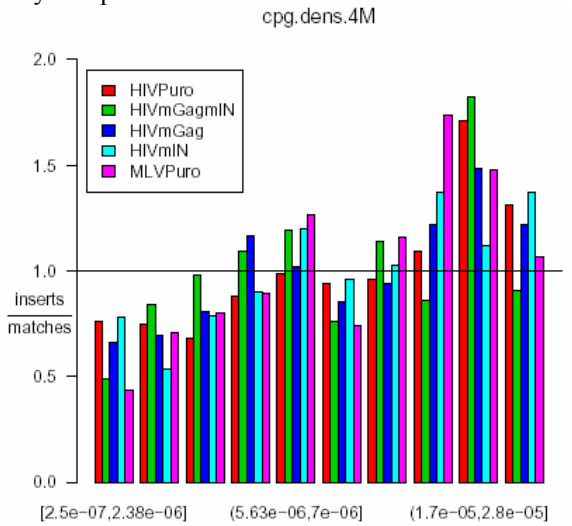
	coef	se	z	p
HIVPuro	1.400	0.1110	12.60	2.03e-36
HIVmGagmIN	0.932	0.0985	9.46	3.17e-21
HIVmGag	0.458	0.0962	4.77	1.87e-06
HIVmIN	1.430	0.1140	12.50	7.38e-36
MLVPuro	1.430	0.1090	13.10	5.03e-39

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.140	0.1040	11.00	4.55e-28
HIVmGagmIN	0.866	0.0977	8.87	7.61e-19
HIVmGag	0.468	0.0962	4.87	1.11e-06
HIVmIN	1.270	0.1100	11.60	7.05e-31
MLVPuro	1.290	0.1060	12.20	2.40e-34

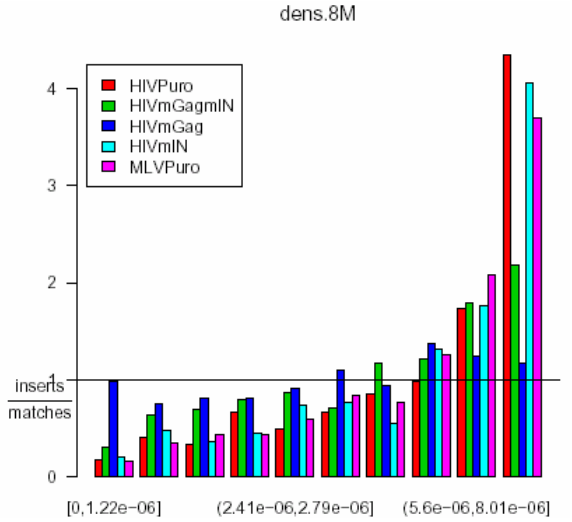
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.385	0.0930	4.14	3.44e-05
HIVmGagmIN	0.178	0.0915	1.95	5.15e-02
HIVmGag	0.273	0.0946	2.88	3.95e-03
HIVmIN	0.335	0.0954	3.51	4.55e-04
MLVPuro	0.409	0.0910	4.49	7.00e-06

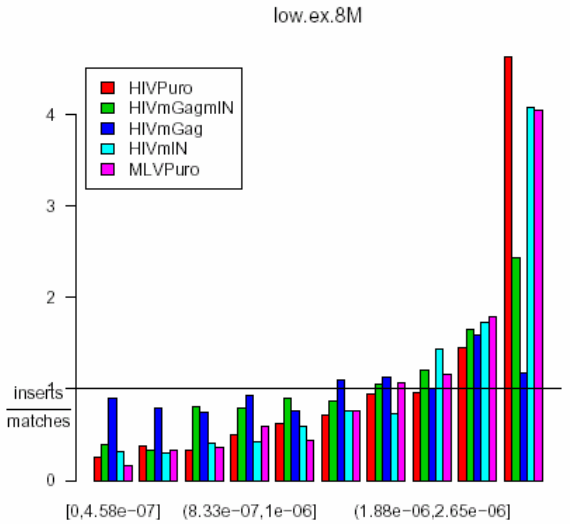
4.9 8 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in an 8 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



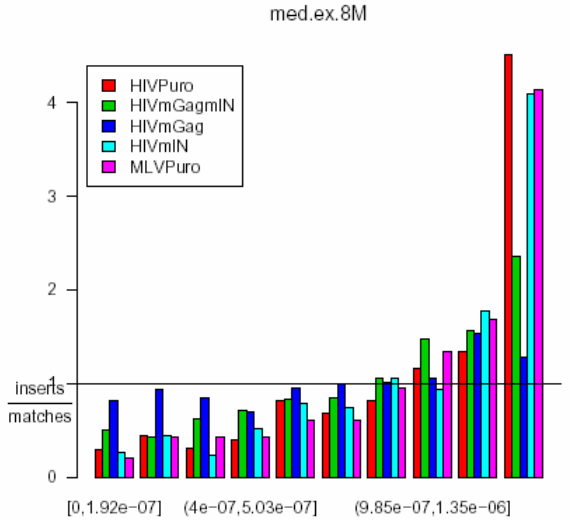
	coef	se	z	p
HIVPuro	1.350	0.1100	12.20	2.74e-34
HIVmGagmIN	0.754	0.0959	7.86	3.80e-15
HIVmGag	0.300	0.0948	3.16	1.56e-03
HIVmIN	1.310	0.1110	11.80	3.88e-32
MLVPuro	1.410	0.1090	12.90	6.94e-38

Here are the results for expression density. First, we count just genes that are in the upper half.



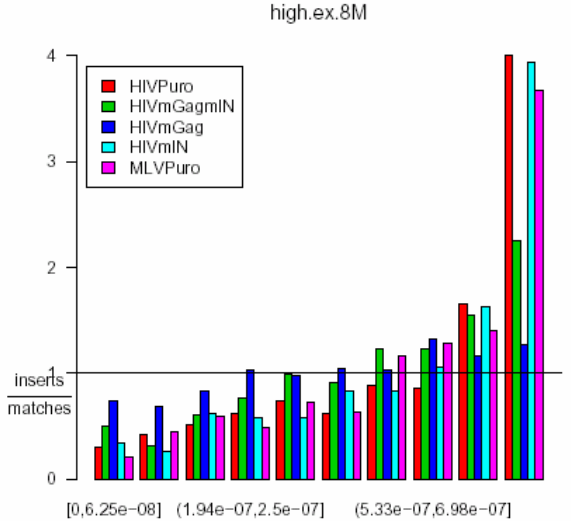
	coef	se	z	p
HIVPuro	1.340	0.1100	12.30	1.59e-34
HIVmGagmIN	0.791	0.0963	8.21	2.19e-16
HIVmGag	0.369	0.0953	3.87	1.07e-04
HIVmIN	1.430	0.1150	12.40	1.46e-35
MLVPuro	1.480	0.1120	13.20	8.83e-40

Now we count genes in the upper 1/8th:



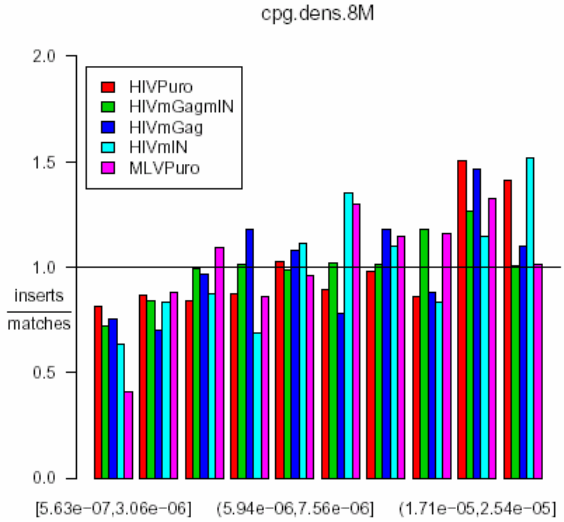
	coef	se	z	p
HIVPuro	1.260	0.1080	11.70	1.23e-31
HIVmGagmIN	0.832	0.0965	8.62	6.62e-18
HIVmGag	0.316	0.0956	3.30	9.54e-04
HIVmIN	1.310	0.1100	11.90	8.01e-33
MLVPuro	1.370	0.1080	12.70	5.30e-37

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.080	0.1040	10.50	1.21e-25
HIVmGagmIN	0.792	0.0964	8.22	2.07e-16
HIVmGag	0.311	0.0955	3.25	1.15e-03
HIVmIN	1.220	0.1080	11.30	1.25e-29
MLVPuro	1.180	0.1030	11.50	1.18e-30

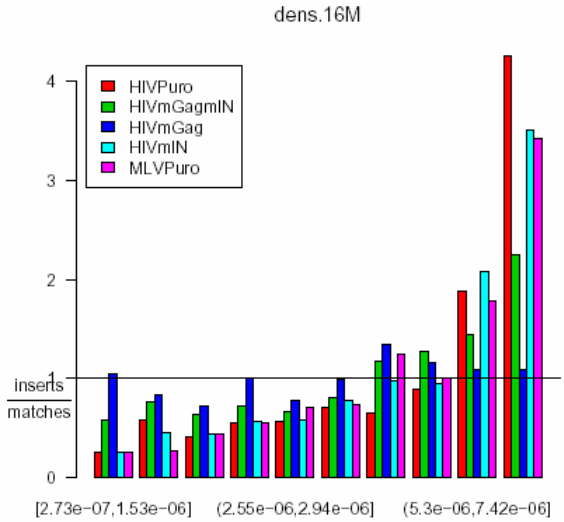
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.231	0.0924	2.50	0.012400
HIVmGagmIN	0.189	0.0914	2.07	0.038800
HIVmGag	0.134	0.0943	1.42	0.156000
HIVmIN	0.349	0.0955	3.65	0.000264
MLVPuro	0.352	0.0906	3.88	0.000103

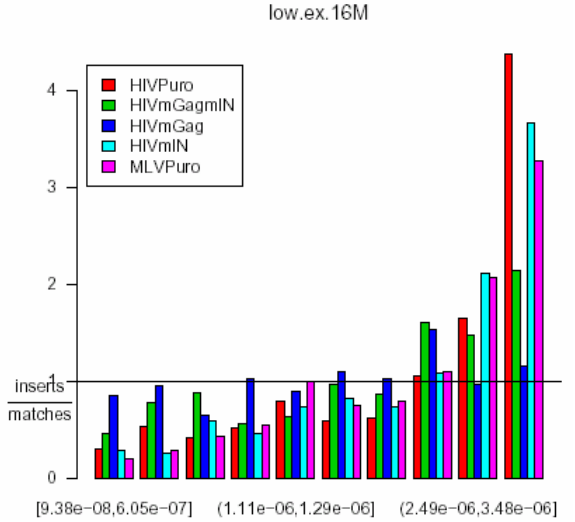
4.10 16 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in a 16 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



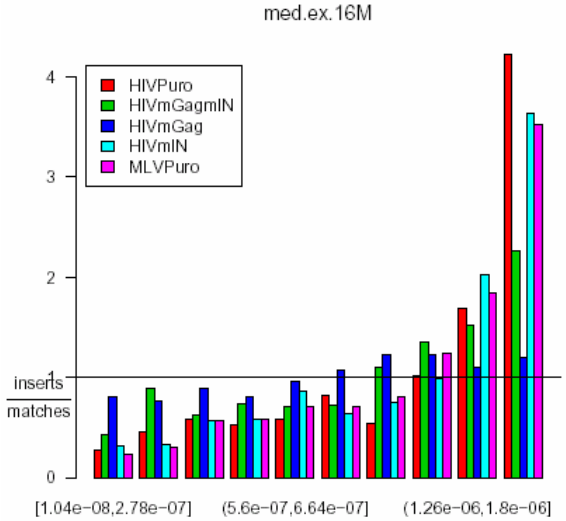
	coef	se	z	p
HIVPuro	1.160	0.1060	11.00	5.28e-28
HIVmGagmIN	0.702	0.0950	7.39	1.47e-13
HIVmGag	0.254	0.0946	2.68	7.36e-03
HIVmIN	1.260	0.1100	11.40	3.79e-30
MLVPuro	1.240	0.1040	11.90	2.00e-32

Here are the results for expression density. First, we count just genes that are in the upper half.



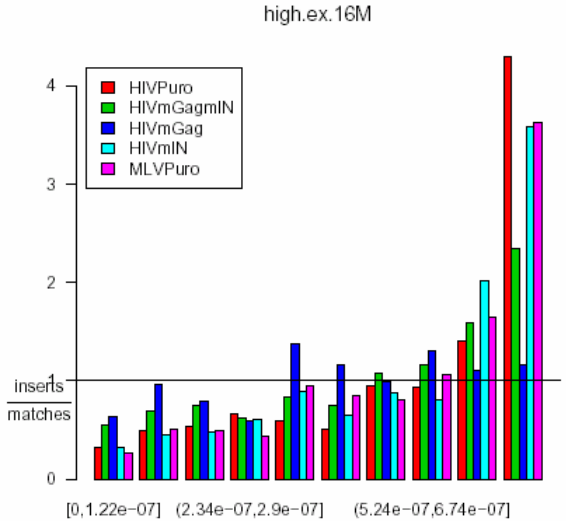
	coef	se	z	p
HIVPuro	1.080	0.1040	10.40	2.67e-25
HIVmGagmIN	0.746	0.0953	7.83	4.73e-15
HIVmGag	0.281	0.0948	2.97	2.98e-03
HIVmIN	1.260	0.1100	11.50	2.06e-30
MLVPuro	1.160	0.1030	11.20	2.49e-29

Now we count genes in the upper 1/8th:



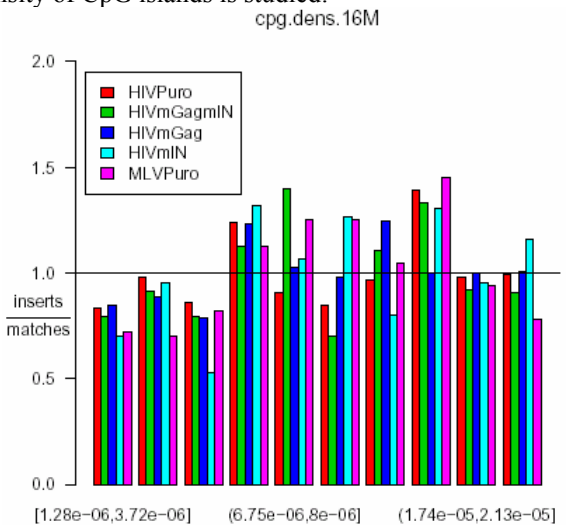
	coef	se	z	p
HIVPuro	1.140	0.1050	10.80	2.73e-27
HIVmGagmIN	0.708	0.0952	7.43	1.09e-13
HIVmGag	0.322	0.0953	3.38	7.29e-04
HIVmIN	1.070	0.1050	10.10	4.73e-24
MLVPuro	1.180	0.1030	11.40	2.53e-30

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.050	0.1030	10.20	1.62e-24
HIVmGagmIN	0.675	0.0949	7.11	1.15e-12
HIVmGag	0.273	0.0951	2.87	4.14e-03
HIVmIN	1.040	0.1050	9.92	3.39e-23
MLVPuro	1.060	0.1000	10.50	5.90e-26

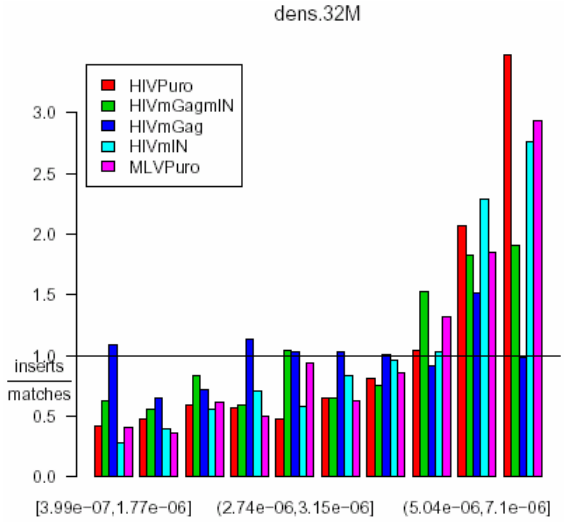
Here the effect of density of CpG islands is studied:



	coef	se	z	p
HIVPuro	0.0785	0.0921	0.853	0.3940
HIVmGagmIN	-0.0152	0.0913	-0.166	0.8680
HIVmGag	0.0922	0.0944	0.977	0.3290
HIVmIN	0.1940	0.0947	2.050	0.0402
MLVPuro	0.1650	0.0899	1.830	0.0666

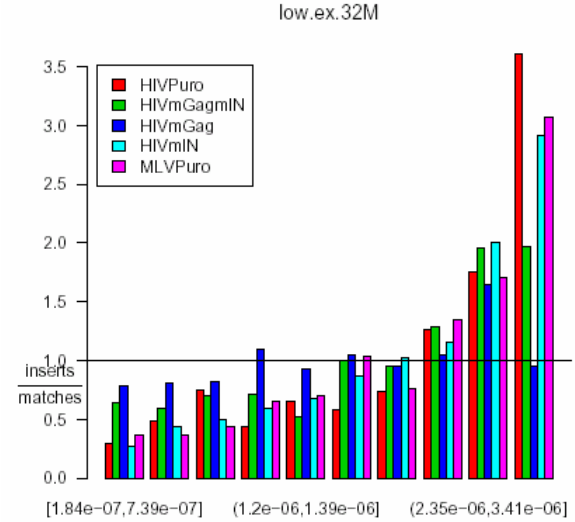
4.11 32 megabase window

In the barplot that follows we examine the association of insertion sites with expression density in a 32 megabase window surrounding each locus. First, we count just the number of genes represented on the chip.



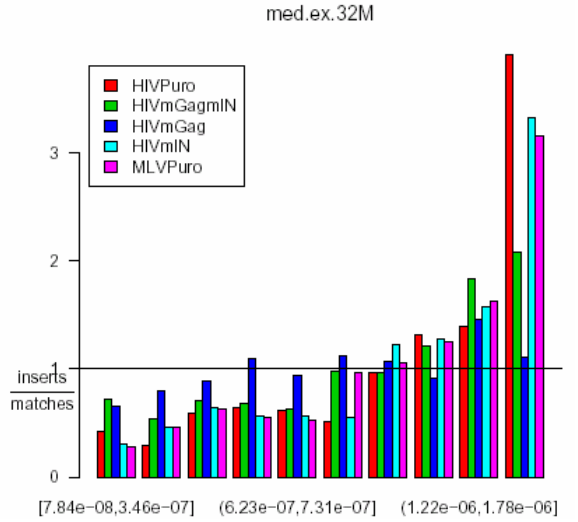
	coef	se	z	p
HIVPuro	1.100	0.1030	10.70	1.69e-26
HIVmGagmIN	0.593	0.0942	6.30	3.03e-10
HIVmGag	0.159	0.0941	1.69	9.13e-02
HIVmIN	1.120	0.1070	10.50	1.25e-25
MLVPuro	0.958	0.0982	9.76	1.72e-22

Here are the results for expression density. First, we count just genes that are in the upper half.



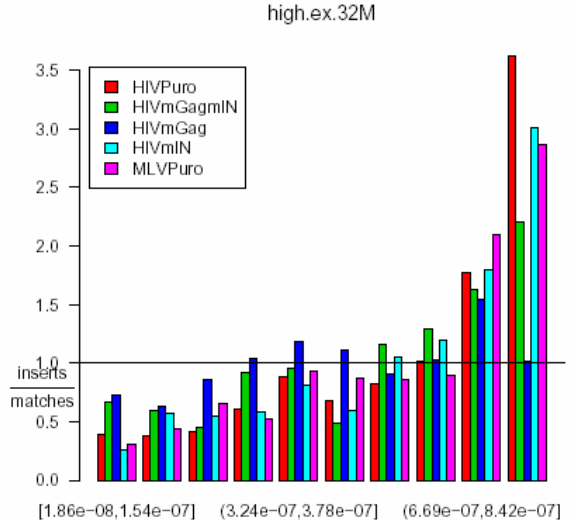
	coef	se	z	p
HIVPuro	1.050	0.1030	10.20	2.52e-24
HIVmGagmIN	0.801	0.0966	8.29	1.18e-16
HIVmGag	0.232	0.0949	2.44	1.45e-02
HIVmIN	1.150	0.1070	10.80	5.35e-27
MLVPuro	1.100	0.1010	10.90	1.47e-27

Now we count genes in the upper 1/8th:



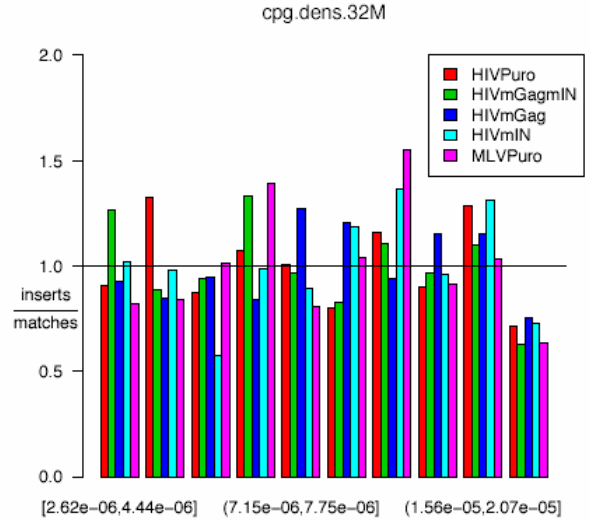
	coef	se	z	p
HIVPuro	1.080	0.1030	10.40	2.14e-25
HIVmGagmIN	0.764	0.0962	7.94	2.01e-15
HIVmGag	0.251	0.0948	2.65	8.01e-03
HIVmIN	1.120	0.1060	10.50	1.02e-25
MLVPuro	1.150	0.1020	11.30	1.23e-29

And here we count genes in the upper 1/16th:



	coef	se	z	p
HIVPuro	1.020	0.1020	10.10	6.39e-24
HIVmGagmIN	0.624	0.0946	6.59	4.36e-11
HIVmGag	0.228	0.0947	2.41	1.59e-02
HIVmIN	1.000	0.1050	9.60	8.29e-22
MLVPuro	0.950	0.0976	9.73	2.25e-22

Here the effect of density of CpG islands is studied:

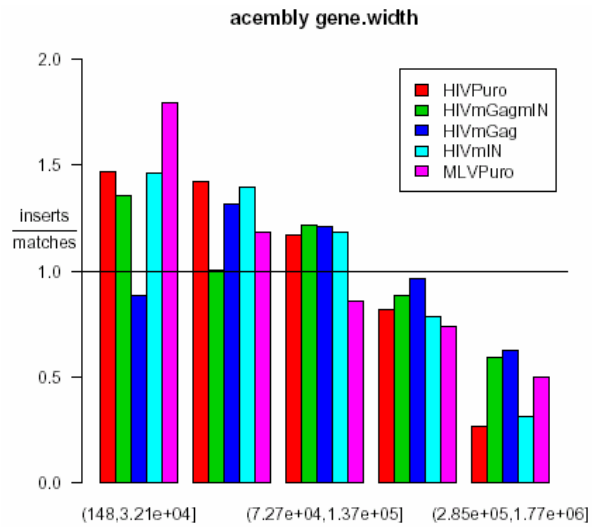


	coef	se	z	p
HIVPuro	-0.0774	0.0918	-0.843	0.3990
HIVmGagmIN	-0.1530	0.0921	-1.660	0.0966
HIVmGag	0.0777	0.0943	0.823	0.4100
HIVmIN	0.2080	0.0947	2.200	0.0281
MLVPuro	0.0573	0.0899	0.638	0.5240

5 Juxtaposition with Gene Start and End Positions

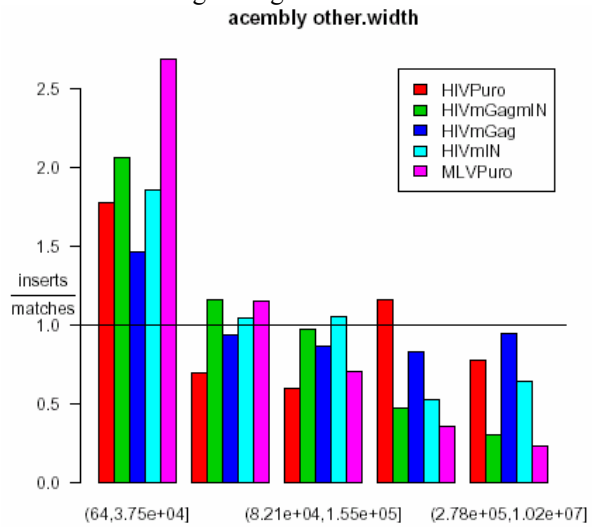
5.1 Acembly Annotations

In this section we study the effect of juxtaposition in terms of gene start and end positions. The first barplot shows the effect of gene width for those insertions that are located within an Acembly gene. The table following the barplot shows the p-values for a test of the hypothesis that the proportions in each of the categories that define the bars are equal in the insertions and their matches. This p-value is obtained from the $5 \times 2 \times k$ table of counts defined by gene width category, insertion/match status, and stratum (consisting of an insertion and its matched sites) using a likelihood ratio test for the hypothesis of no association between gene width category and insertion/match status. The test used compared the log-linear model (Bishop *et al.*, 1975) with all two-way configurations to that with no gene width category and insertion/match status configuration.



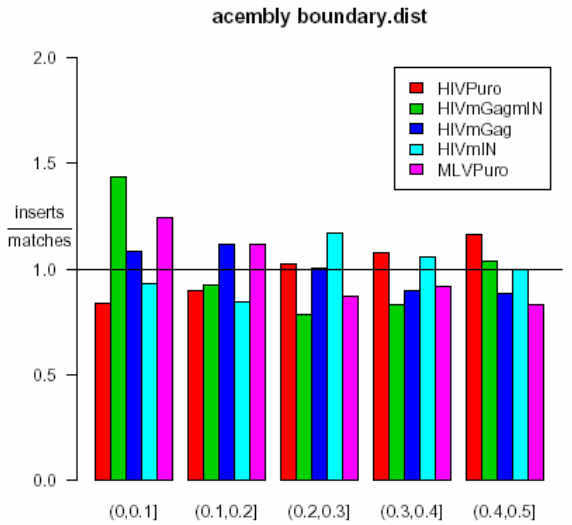
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
2.77e-24	7.78e-05	3.47e-04	1.30e-20	4.19e-14

The next plot uses the width of a non-gene region for insertions that fall into such regions.



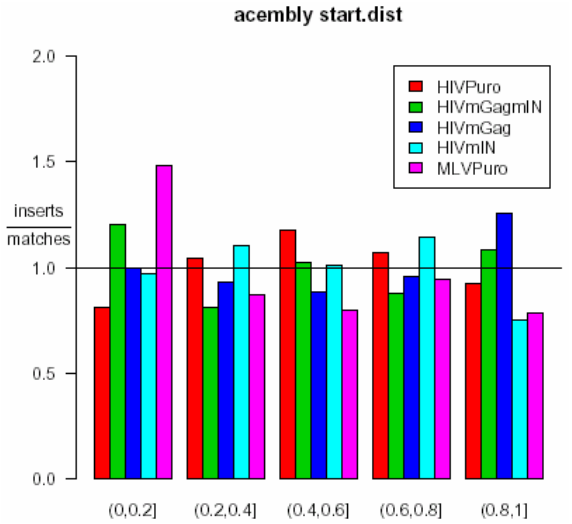
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
6.95e-02	4.95e-16	2.76e-01	1.12e-03	9.18e-24

The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, '0.25' refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.



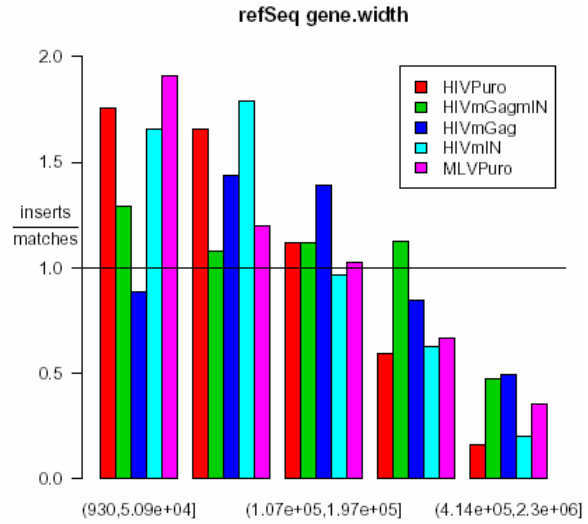
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
1.16e-01	2.21e-05	3.64e-01	2.16e-01	1.00e-02

This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point. Locations near '0' are relatively near the beginning of transcription, while those near '1' are near the termination of the transcript.

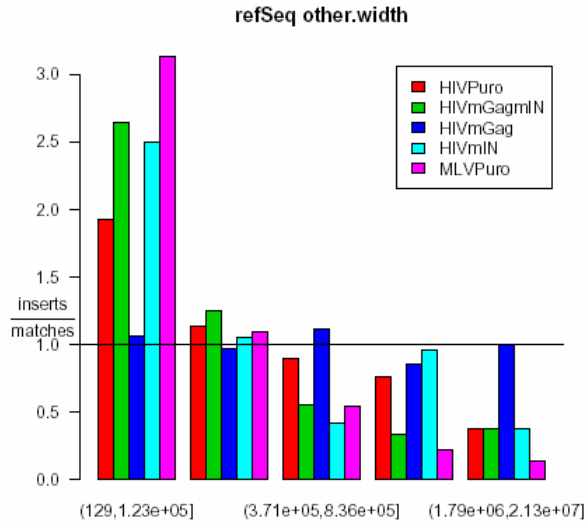


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
4.47e-02	5.51e-02	1.79e-01	8.81e-02	1.07e-06

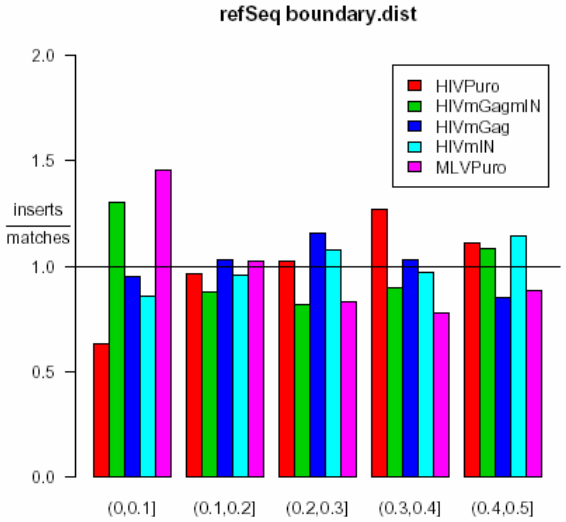
5.2 RefSeq Annotations



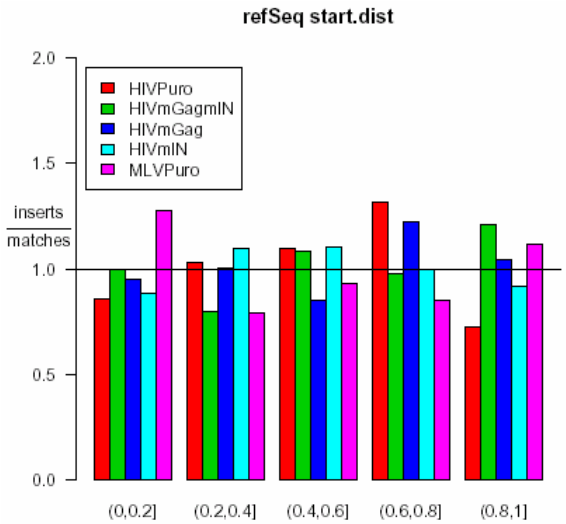
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
1.63e-37	1.00e-03	6.13e-08	4.00e-31	7.75e-13



HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
3.39e-05	6.23e-41	5.64e-01	2.61e-13	9.69e-62

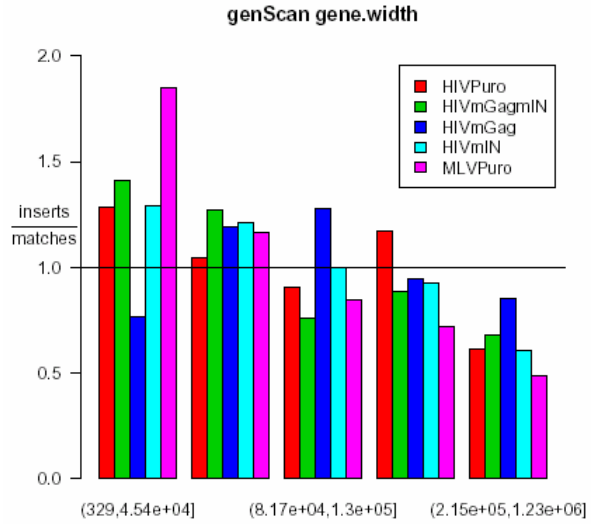


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
7.88e-05	2.35e-03	3.27e-01	3.15e-01	3.34e-06

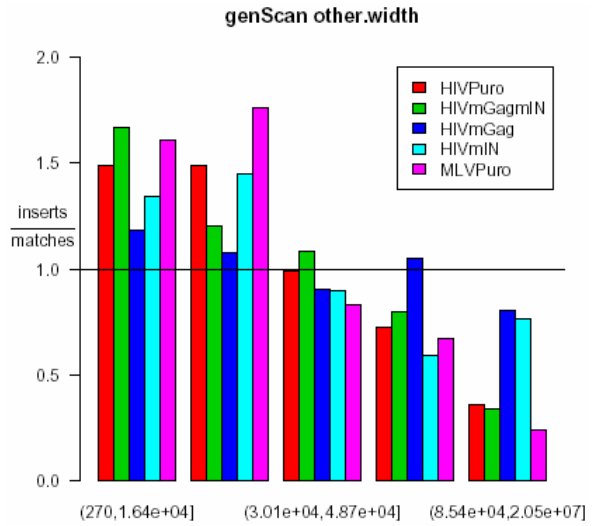


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
0.00159	0.12800	0.27700	0.42400	0.00164

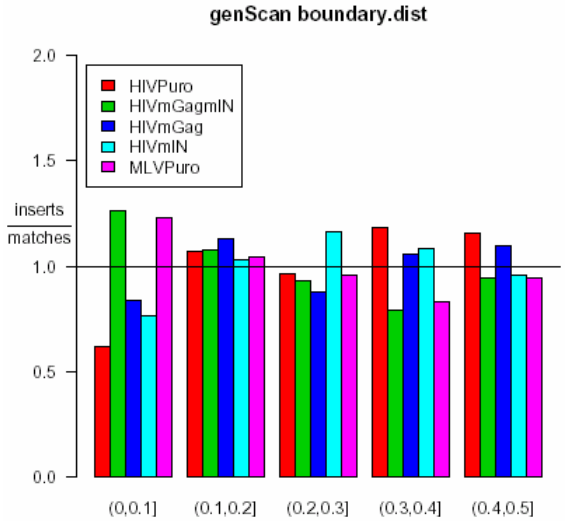
5.3 GenScan Annotations



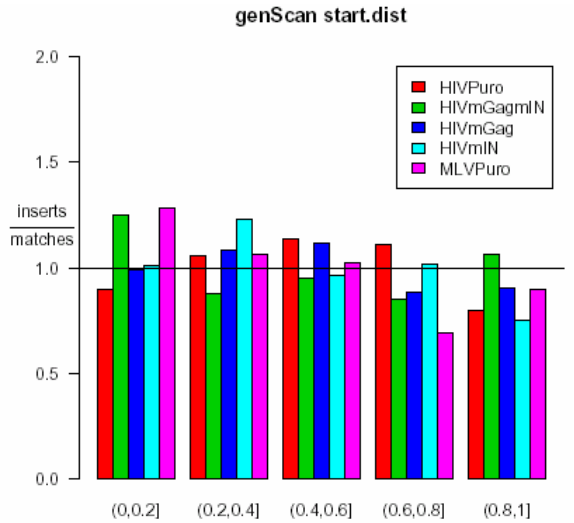
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
4.22e-05	3.72e-06	1.72e-02	4.78e-05	2.64e-14



HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
2.44e-03	2.11e-06	8.67e-01	8.51e-06	1.12e-12

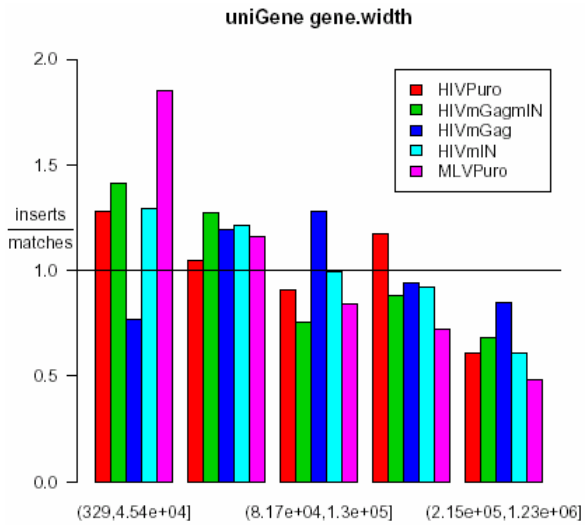


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
7.39e-05	1.14e-02	1.24e-01	5.21e-02	4.80e-02

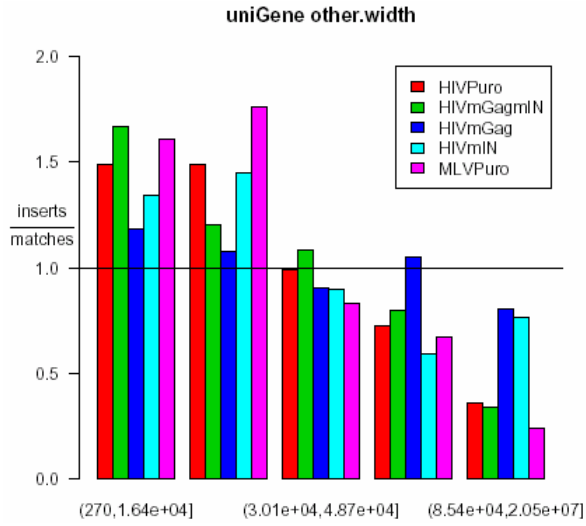


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
0.08980	0.05220	0.50700	0.04610	0.00192

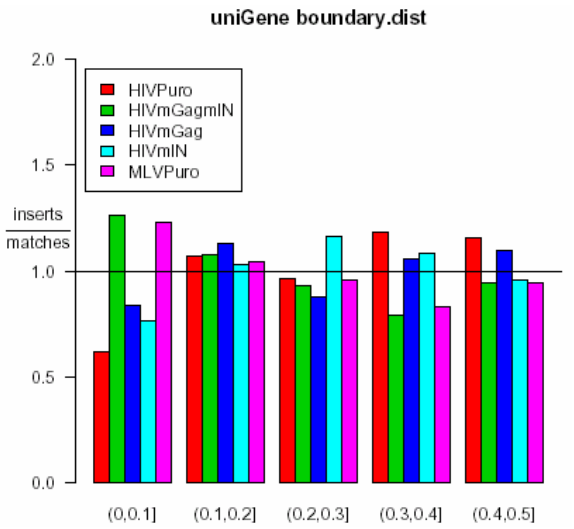
5.4 UniGene Annotations



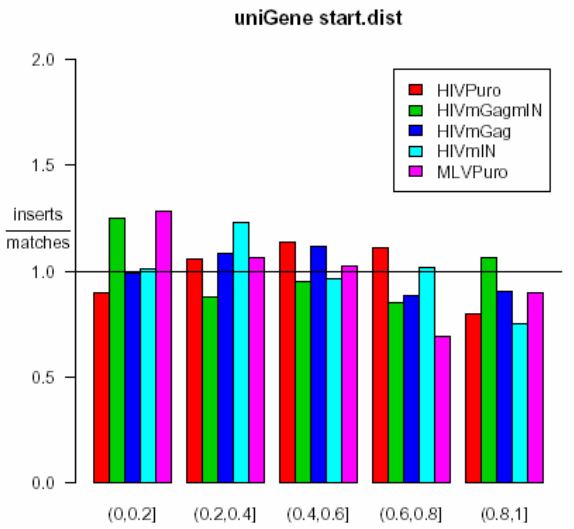
HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
4.22e-05	3.72e-06	1.72e-02	4.78e-05	2.64e-14



HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
2.44e-03	2.11e-06	8.67e-01	8.51e-06	1.12e-12



HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
7.39e-05	1.14e-02	1.24e-01	5.21e-02	4.80e-02

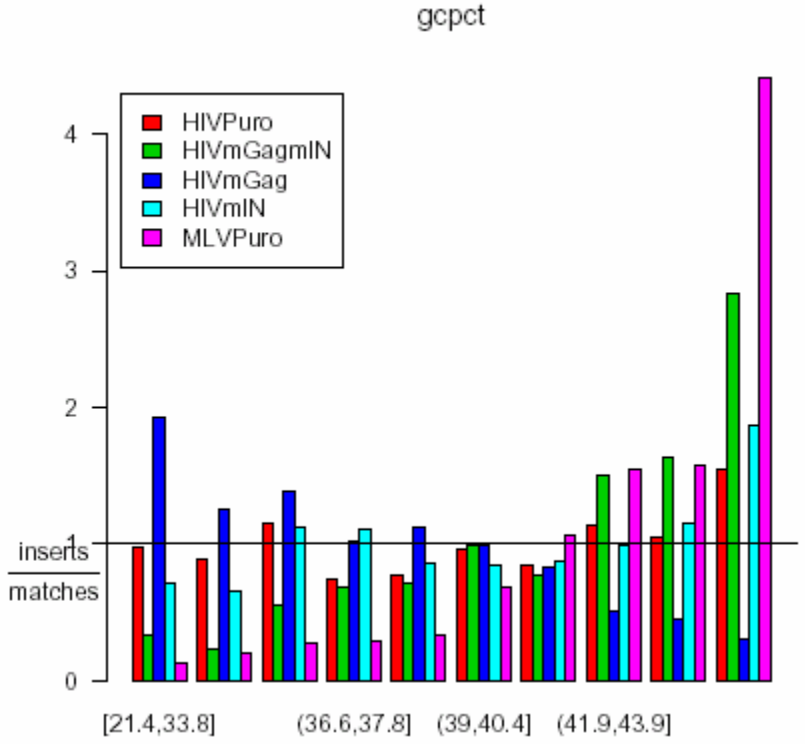


HIVPuro	HIVmGagmIN	HIVmGag	HIVmIN	MLVPuro
0.08980	0.05220	0.50700	0.04610	0.00192

6 GC content

Here we study the effect of GC content on insertion. The GC content is taken from the Human Genome Draft at GoldenPath from the table http://genome.ucsc.edu/goldenPath/14nov2002/database/gcPercent.txt.gz.

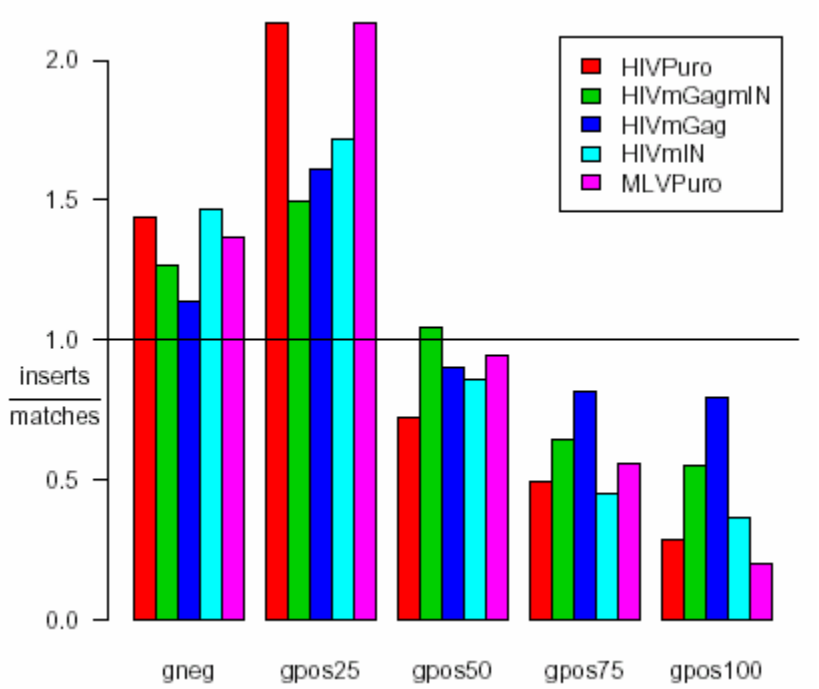
Following the plot is a table of fitted coefficients based on splitting the GC percent data at the median.



	coef	se	z	p
HIVPuro	0.185	0.0924	2.00	4.57e-02
HIVmGagmIN	1.120	0.1050	10.70	9.91e-27
HIVmGag	-0.772	0.1030	-7.50	6.46e-14
HIVmIN	0.248	0.0960	2.59	9.71e-03
MLVPuro	2.030	0.1330	15.20	2.01e-52

7 Cytobands

Here we study the association of cytoband with insertion intensity. The data are obtained from http://genome.ucsc.edu/goldenPath/14nov2002/database/cytoBand.txt.gz.



A formal test of significance attains a p-value of $< 2.22e{-}16$. Here is the table of coefficients of the log ratio of intensities for true insertion sites versus control insertion sites (comparing each category of Giemsa staining to 'gneg') along with their standard errors, z statistics, and p-values:

	coef	se	z	p
cyto.typegpos100	-1.120	0.0737	-15.10	8.11e-52
cyto.typegpos25	0.317	0.0673	4.71	2.47e-06
cyto.typegpos50	-0.403	0.0641	-6.28	3.40e-10
cyto.typegpos75	-0.805	0.0723	-11.10	8.27e-29

APPENDIX 3

Similarity of Integration Sites of Different Integration Complexes

Charles C. Berry

1 Introduction

The aim of this report is to assess the tendency of different integration complexes—in this case retroviral vectors composed of HIV, MLV, or elements of both—to select particular genomic loci as favored integration targets. Previously, it has been shown that HIV and MLV favor different sites for integration. It is of particular interest to characterize the degree to which different integration complexes favor the same or different sites.

With a very large number of integration events (of the order of 10 per base by complex or 150,000,000,000 for this study), this could be done directly by counting the number of events at each genomic locus for each integration complex, then comparing the counts. Integration complexes that tend to share high counts at some loci and share low counts at other loci presumably share features that govern integration targeting. On the other hand, integration complexes whose counts do not correlate in this fashion presumably do not share features relevant to integration targeting.

Practically, it is not now possible to collect such large samples of integration events, so another strategy is needed. A number of genomic features (e.g. local GC percentage, exons, actively transcribing genes) have been identified that correlate with integration of HIV and/or MLV. By applying a machine learning algorithm to a sample of integration events, a function can be created that maps the local genomic features to a vector of probabilities of integration of different types.

The overall strategy used here is to characterize the integration intensity for different integration complexes at particular genomic positions according to a collection of features associated with each position. This will be done by using a supervised machine learning algorithm to form a classification rule. After this, a sample of genomic positions will be studied to determine which complexes share features that govern integration targeting.

2 Data Used

The number of integration sites used for each integration complex used summarized here:

	count
HIVPuro	524
HIVmGAGmIN	526
HIVmGAG	493
HIVmIN	492
MLVPuro	543
matchedControl	500

The 'matchedControl' sites are randomly sampled *in silico* from the genome (according to Chromosome, Position on the chromosome, and Strand), but at a similar distance from the restriction site used in these experiments as one of the actual insertion sites. A second set of randomly sampled sites is later used to compare the predicted targets of the different integration complexes.

The features used are as follows:

In Gene The position is or is not in a gene according to each of these annotation schemes: Acembly, RefSeq, UniGene, and GenScan. (4 features)

In Exon The position is or is not in an exon according to each of these annotation schemes: Acembly, RefSeq, UniGene, and GenScan. (4 features)

Gene Density The density of genes according to each of the annotation schemes and within windows with widths of ±50,000 bases, ±100,000 bases, ±250,000 bases, ±500,000 bases, and ±1,000,000 bases. Each density is the number of genes counted divided by the number of bases. (20 features)

Density of Expressed Genes Using the genes on the Affymetrix HU133A GeneChip, the number of such genes, the numbers whose 'average difference score' were characterized as at least 'low' (above the median), at least 'medium' (above the 75th percentile), and at least 'high' (above the 87.5th percentile) were counted in windows of widths ±12,500 bases, ±25,000 bases, ±50,000 bases, ±125,000 bases, ±250,000 bases, ±500,000 bases, ±1,000,000 bases, ±2,000,000 bases, ±4,000,000 bases, ±8,000,000 bases and ±16,000,000 bases. (44 features)

GC percentage In running windows of width 5120 bases. Derived from the file gc5Base.txt.gz from the GoldenPath website (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/). (1 feature)

In CpG Island In or not in a CpG island according to the cpgIsland.txt.gz from the GoldenPath website. (1 feature)

CpG Island Neighborhoods Whether site is within ±500, ±2,500, ±5,000, ±12,500, or ±25,000 bases of a CpG island. (5 features)

CpG Island Density The density of CpG islands in windows of widths ±12,500, ±25,000, ±50,000, ±125,000, ±250,000, ±500,000, ±1,000,000, ±2,000,000, ±4,000,000, ±8,000,000, and ±16,000,000 bases. Each density is the number of island counted divided by the number of bases. (11 features)

DNase I Site Density The number of DNase I sites in windows of widths ±500, ±1000, and ±5000 bases. Each density is the number of sites counted divided by the number of bases. (3 features)

Juxtaposition of Transcription Start/Stop Sites Various measures are used: the width of the gene if the insertion site is in one or else the width of the intergenic region, the fraction of that distance from/to the nearest gene boundary, the absolute distance to the nearest transcription start site, and the signed distance to the nearest start site (negative values precede start sites). These are calculated for each of these annotation schemes: Acembly, RefSeq, UniGene, and GenScan. (16 features)

3 Training the Predictive Algorithm

The algorithm used in this report is the *randomForest* algorithm of Breiman (Breiman, 2001). It was chosen for its proven ability to perform classification (including estimation of posterior probabilities) on data sets with modest numbers of observations but with many variables. In addition, accurate estimates of classification error and measures of the marginal importance of classifying variables are obtained as a by-product of the *bagging* algorithm used by the procedure.

Roughly speaking, the algorithm grows a collection of binary trees—splitting the data recursively to create branches in which the one or a few classes dominate. The use of resampling procedures for selecting the objects to be classified and the predictor variables for which candidate splits are allowed generates a collection of trees. These sampling procedures counter the tendency to overfit the training data and are responsible for the excellent performance of the randomForest algorithm. Each tree in the collection will produce a predicted class for a vector of predictor variables, and the 'votes' of the collection of trees is used to assign the ultimate prediction.

The implementation used is that of Liaw and Wiener (Liaw and Wiener, 2002) ('randomForest' version 4.5-12—an R package (R Development Core Team, 2005)) and is based on the Fortran code of Leo Breiman and Adele Cutler.

The default values for options in the randomForest function that govern the approach to training the classifier were used with the exception of the prior probabilities which were set equal to one another. In subsequent runs, the number of variables screen for each candidate split was varied to half and twice the default number; there was little effect on the classification results.

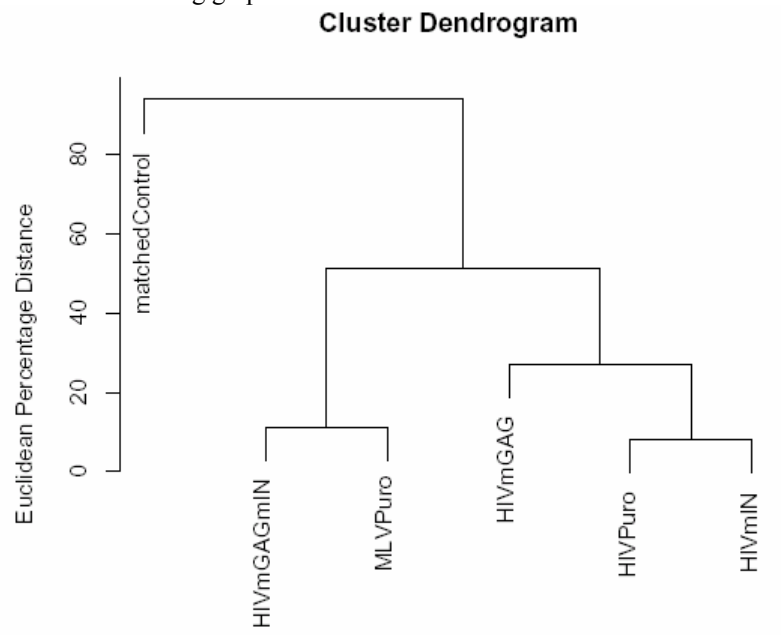
4 Results

4.1 Classification of the Training Data

The classifications made on the training dataset are summarized in the following table:

	predicted					
actual	HIVPuro	HIVmGAGmIN	HIVmGAG	HIVmIN	MLVPuro	matchedControl
HIVPuro	167	32	125	108	52	40
HIVmGAGmIN	46	140	67	37	173	63
HIVmGAG	83	37	192	68	37	76
HIVmIN	138	42	96	123	62	31
MLVPuro	52	147	46	40	225	33
matchedControl	8	17	18	2	16	439

As is evident from inspection of the table, matched control sites are rarely mistaken as bona fide integration sites (439 of 500 were correctly classified). Close inspection also shows that several rows have roughly similar patterns of counts. HIVPuro and HIVmIN are similar as are MLVPuro and HIVmGAGmIN. Hierarchical clustering was performed by percentaging each row in the table above (using Euclidean distance between the resulting rows of percents and the 'complete' clustering method). The results appear in the following graph:



As one might expect the HIVPuro - HIVmIN pair merge first, then the MLVPuro - HIVmGAGmIN pair merge, and the matched controls are last to merge.

The marginal importance of variables can be judged by randomly permuting its values among those available for splitting a tree at a given point. The values in the table below reflect the decrease in accuracy (i.e. the fraction correctly classified) for each of a collection of variables (each variable is among the top 5 for at least one of the classes of integration complex or matched control).

	HIVPuro	HIVmGAGmIN	HIVmGAG	HIVmIN	MLVPuro	matchedControl
signed.dx.ref	0.021	0.022	0.021	0.017	0.027	0.006
refGene.genes	0.018	0.013	0.017	0.017	0.010	0.000
acembly.genes	0.012	0.006	0.009	0.009	0.006	0.009
signed.dx.ace	0.011	0.008	0.011	0.011	0.013	0.008

gcpct	0.011	0.014	0.036	0.012	0.042	-0.002
signed.dx.uni	0.008	0.011	0.008	0.006	0.017	0.003
general.wd.ref	0.009	0.010	0.012	0.011	0.010	0.007
uni.500k	0.004	0.000	0.002	0.006	0.002	0.023
ace.500k	0.007	0.002	0.002	0.004	0.004	0.022
ace.200k	0.007	0.001	0.004	0.006	0.003	0.021
ace.1M	0.007	0.001	0.006	0.005	0.004	0.014
ace.100k	0.005	0.002	0.001	0.007	0.000	0.014

The variable names describe the following features:

signed.dx.ref	Distance from(+)/to(-) nearest refGene gene start
refGene.genes	In refGene gene
acembly.genes	In acembly gene
signed.dx.ace	Distance from(+)/to(-) nearest acembly gene start
gcpct	GC percent in 5120 base window
signed.dx.uni	Distance from(+)/to(-) nearest uniGene gene start
general.wd.ref	Width of refGene intergenic region
uni.500k	uniGene density in 500 kilobase window
ace.500k	acembly gene density in 500 kilobase window
ace.200k	acembly gene density in 200 kilobase window
ace.1M	acembly gene density in 1megabase window
ace.100k	acembly gene density in 100 kilobase window

As is evident the more important variables for distinguishing between integration sites and matched control sites tend not to be so important for distinguishing among the different integration events (and vice versa). In particular, the GC percentage and the juxtaposition of transcription start sties are at or near the top of each list for the integration sites but not for the matched control sites. Measures of gene density are most important for classifying matched controls, but not for discriminating among integration complexes.

It is interesting to consider whether these 12 variables classify the integration complexes as well as the full collection of 109 variables used earlier. Here is the table of classification results.

	predicted					
actual	HIVPuro	HIVmGAGmIN	HIVmGAG	HIVmIN	MLVPuro	matchedControl
HIVPuro	151	40	139	119	51	24
HIVmGAGmIN	44	159	63	28	182	50
HIVmGAG	103	39	201	62	39	49
HIVmIN	136	41	101	124	67	23
MLVPuro	57	136	38	45	231	36
matchedControl	11	23	24	5	23	414

These results differ slightly from those seen above—on the whole the classification accuracy hardly differs. It is probably worth taking another look at the variable importance measure now that many redundant variables have been eliminated. This table shows the revised variable importance measures:

	HIVPuro	HIVmGAGmIN	HIVmGAG	HIVmIN	MLVPuro	matchedControl
signed.dx.ref	0.040	0.062	0.052	0.041	0.069	0.018
refGene.genes	0.046	0.040	0.053	0.046	0.031	0.001
acembly.genes	0.023	0.018	0.023	0.018	0.013	0.013
signed.dx.ace	0.012	0.028	0.021	0.023	0.024	0.023
gcpct	0.018	0.036	0.069	0.026	0.091	-0.002
signed.dx.uni	0.007	0.020	0.009	0.009	0.026	0.022
general.wd.ref	0.014	0.033	0.020	0.020	0.013	0.014
uni.500k	0.003	0.001	0.017	0.019	0.013	0.056

ace.500k	0.016	0.017	0.012	0.013	0.011	0.055
ace.200k	0.020	0.010	0.011	0.011	0.008	0.051
ace.1M	0.031	0.007	0.023	0.021	0.014	0.036
ace.100k	0.013	0.007	0.007	0.011	0.000	0.037

Again, different variables tend to register as important for discriminating among integration complexes as opposed to discriminating between them and the matched control sites. Now that many redundant variables have been eliminated the marginal importance of most variables has increased.

4.2 Classification of Genomic Locations

To get a better sense of the relation between the attractiveness of a genomic location to different integration complexes the predicted probabilities for a set of random matched controls are computed using the original classification trees.

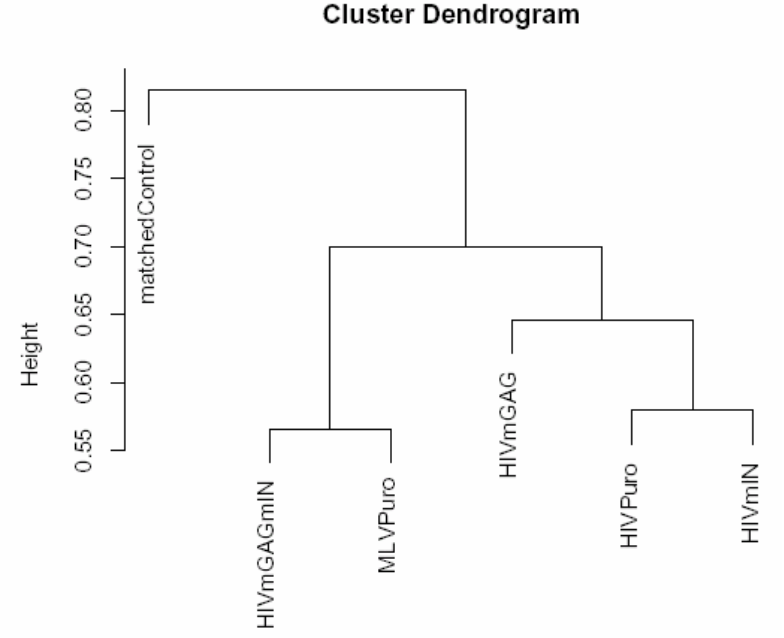
Taking p_{ij} as the posterior probability of integration complex category j for genomic location i, a normalization is performed:

$$\tilde{p}_{ij} = p_{ij} / \sum_i p_{ij}$$

Then the symmetrized Kullback-Leibler distance is calculated as

$$KL(j,k) = \sum_i \tilde{p}_{ij} \log(\tilde{p}_{ij}/\tilde{p}_{ik}) + \tilde{p}_{ik} \log(\tilde{p}_{ik}/\tilde{p}_{ij})$$

Hierarchical clustering of these distances is presented in the following graph:



The same nodes are merged to form this tree as the earlier one, although the order is slightly different than in the tree formed from the classification of the training set.

APPENDIX 4

Oligonucleotides used in the studies

Table S1: Oligonucleotides used for the study in Chapter Two

Oligonucleotide name	Sequence 5′-3′	Comments
MseI linker+	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC	adapter top strand
MseI linker-	P-TAGTCCCTTAAGCGGAG-N	adapter bottom strand; P=phosphate, N=amine blocking group
MseI linker primer 1	GTAATACGACTCACTATAGGGC	adapter primer, PCR 1
MseI linker nested primer	AGGGCTCCGCTTAAGGGAC	adapter primer, nested PCR
MKL-3	CTTAAGCCTCAATAAAGCTTGCCTTGAG	HIV primer, PCR 1
MKL-5	TGACTCTGGTAACTAGAGATCCCTCAG	HIV primer, nested PCR
MLVPuro-primer1	GAATCGTGGTCTCGCTGTTCCTTGG	MLVPuro primer, PCR 1
MLVPuro-nested	GGTCTCCTCTGAGTGATTGACTACC	MLVPuro primer, nested PCR
HIVU3rev	GCCTCTTCTACCTTATCTGGCTCAACTG	HIV-specific primer to clone 5' LTR-genomic DNA junction
mGag9B10pr	CTGAAACTGGAACTTCTCTCCCATT	integration site sequence-specific primer to clone 5' LTR-genomic DNA junction
mGag1C10pr	GAACATGTCACTTAACTTTTCCATTCCA	
mGag10E01pr	GAGAAGCTGTAGGAGTGTTCAGAGTCA	
mGag11C03pr	CGCCACATTCTTACTGCACATTAAAG	
mGag6F04pr	GTCTCTTAAACATCTGAATGTGCATCTT	
mGagmIN5G11pr	AGTAGCCCTTTTCTTAATTGCCAGTG	
mGagmIN3E03pr	GATGATAATGATGATGATTACAGATGGGA	
mGagmIN5H02pr	ATAGTGATGTCTGCTTTCTAGATGCTGC	
mGagmIN2A05pr	TCCACTGCTCACTTTATAGGCCCTG	
mGagmIN6B02pr	GTCTCCACTTCTTCCTTTAGGTCAAC	

Table S2: Oligonucleotides used for the study in Chapter Three

Oligonucleotide name	Sequence 5′-3′	Comments
HincII	GTAATACGACTCACTATAGGGCACGCGTGGTCGACG GCCCGGGCTGC	adapter 1 top strand
mNheIAvrIISpeII	P-CTAGGCAGCCCG-N	adapter 1 bottom strand; P=phosphate, N=amine blocking group
ASB 9	GACTCACTATAGGGCACGCGT	adapter 1 primer, PCR 1
ASB 16	GTCGACGGCCCGGGCTGCCTA	adapter 1 primer, nested PCR
MseI linker+	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC	adapter 2 top strand
MseI linker-	P-TAGTCCCTTAAGCGGAG-N	adapter 2 bottom strand; P=phosphate, N=amine blocking group
MseI linker primer 1	GTAATACGACTCACTATAGGGC	adapter 2 primer, PCR 1
MseI linker nested primer	AGGGCTCCGCTTAAGGGAC	adapter 2 primer, nested PCR
MKL-3	CTTAAGCCTCAATAAAGCTTGCCTTGAG	HIV primer, PCR 1
MKL-5	TGACTCTGGTAACTAGAGATCCCTCAG	HIV primer, nested PCR

APPENDIX 5

Association of Genomic Features with Integration in Stably Expressed and Inducible Cell Lines

Charles C. Berry

Contents

1 Introduction

In this document, I examine the association of integration sites in cells selected as stably expressed (labeled 'IBB' hereafter) or inducible (labeled 'ID') with various genomic features.

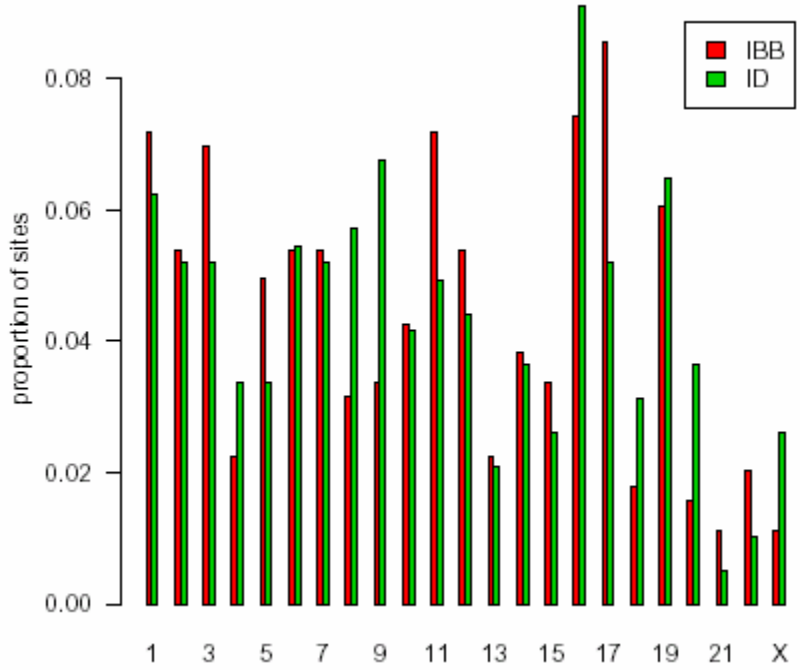
The numbers are shown below:

exp.group

IBB ID

447 388

The distribution of relative frequency of insertions across the chromosomes is given in this barplot:

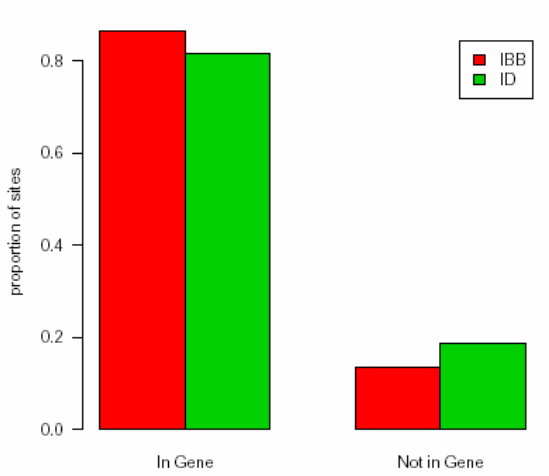


Are there chromosomes that are particularly favored for integration by one group over the other? This was tested for statistical significance. The test performed used the likelihood ratio statistic for the logistic regression model (reviewed in (McCullagh and Nelder, 1999)) as implemented by the glm function of R using the binomial family. The null hypothesis tested is the ratio of true integration events in the two groups is constant across all chromosomes. This test attains a p-value of 0.17674.

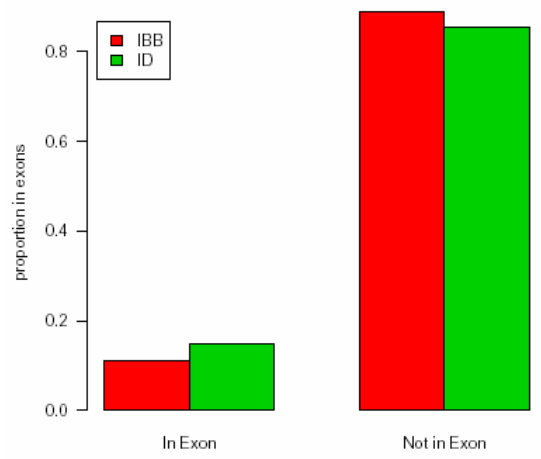
2 Preference for Genes

2.1 Acembly Genes

Here we examine the relative preference that integration events in the two groups have for genes. In the following plot we show the relative frequency of integrations in genes according to the 'Acembly' annotation. The bars grouped over the label "In Gene" give the relative frequency of integration events (compared to control sites) between bases located within Acembly gene annotations, while the bars over the label "Not in Gene" give the relative frequency of integration events (compared to control sites) between bases not located within Acembly gene annotations.



Is there is a difference in the tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.053439. In the following plot we show the relative frequency of insertions in exons according to the 'Acembly' annotation. The bars grouped over the label "In Exon" give the relative frequency of integration events (compared to control sites) between bases located in exons according to the Acembly annotation, while the bars over the label "Not in Exon" give the relative frequency of integration events (compared to control sites) between bases not located in exons according to the Acembly gene annotation.



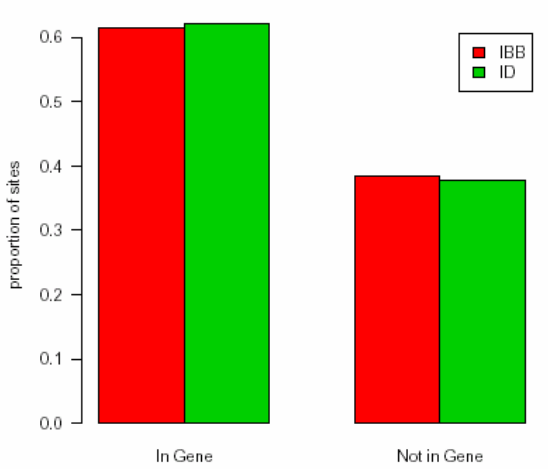
Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.166	0.174	0.953	0.3410
in.gene	-0.426	0.193	-2.210	0.0270
in.exon	0.391	0.211	1.860	0.0632

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot the bars above 'Not in Exon' include both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

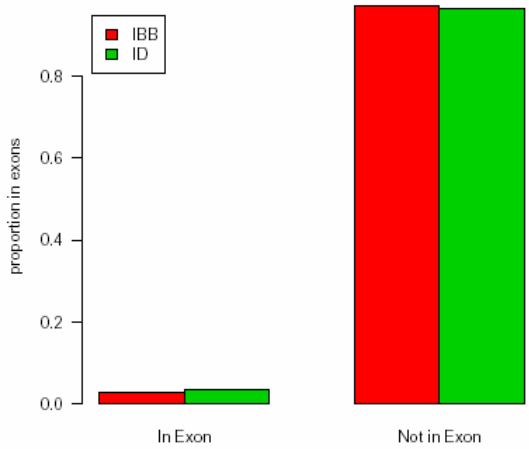
2.2 RefGenes

Here we examine the relative preference that insertions of the two types have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'refGene' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.86057.

In the following plot we show the relative frequency of insertions in exons according to the 'refGene' annotation.



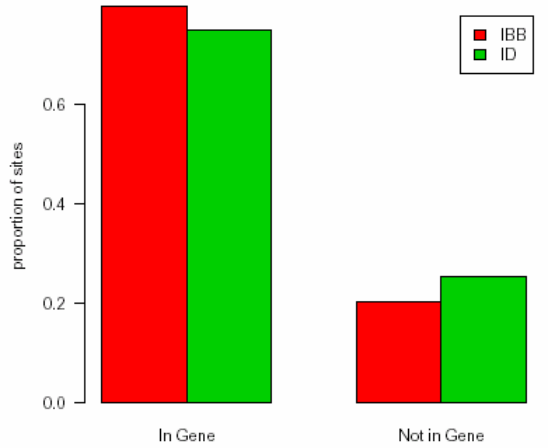
Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-0.1570	0.112	-1.4000	0.162
in.gene	0.0137	0.144	0.0947	0.925
in.exon	0.2180	0.396	0.5500	0.583

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot the bars above 'Not in Exon' include both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

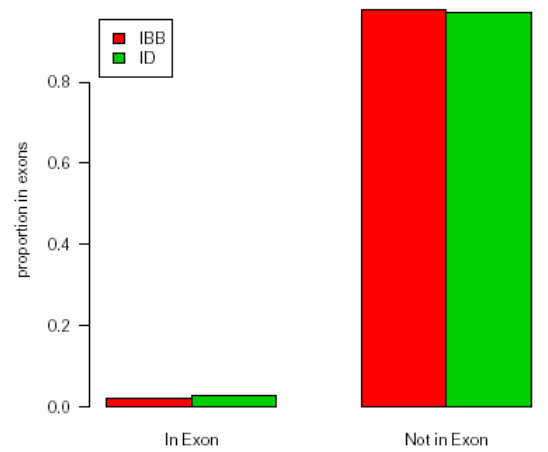
2.3 GenScan Genes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'genScan' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.091842.

In the following plot we show the relative frequency of insertions in exons according to the 'genScan' annotation.



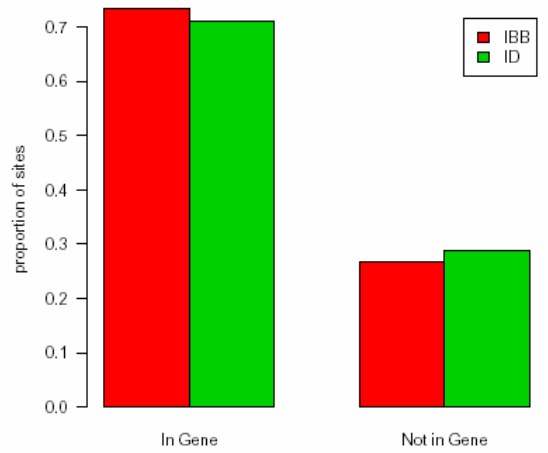
Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	0.0741	0.146	0.509	0.611
in.gene	-0.2890	0.166	-1.740	0.082
in.exon	0.3110	0.444	0.699	0.485

The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot the bars above 'Not in Exon' include both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

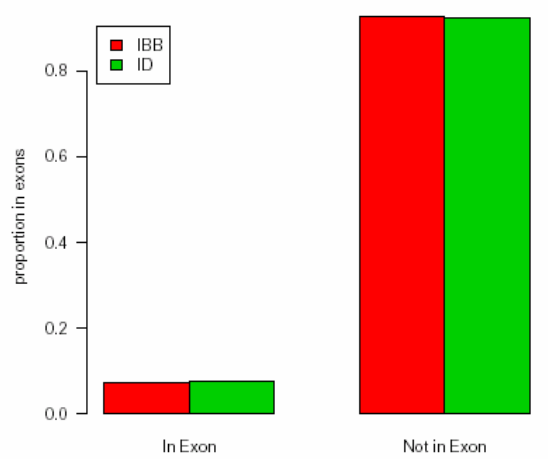
2.4 UniGenes

Here we examine the preference that insertions have for genes. In the following plot we show the relative frequency of insertions in genes according to the 'uniGene' annotation.



Is there is a tendency for insertions to occur in genes? A formal test of significance yields a p-value of 0.46991.

In the following plot we show the relative frequency of insertions in exons according to the 'uniGene' annotation.



Here is the table of coefficients of the log ratio of intensities for ID sites versus IBB sites along with their standard errors, z statistics, and p-values:

	coef	se	z	p
(Intercept)	-0.0606	0.132	-0.460	0.645
in.gene	-0.1210	0.157	-0.769	0.442
in.exon	0.0863	0.267	0.324	0.746

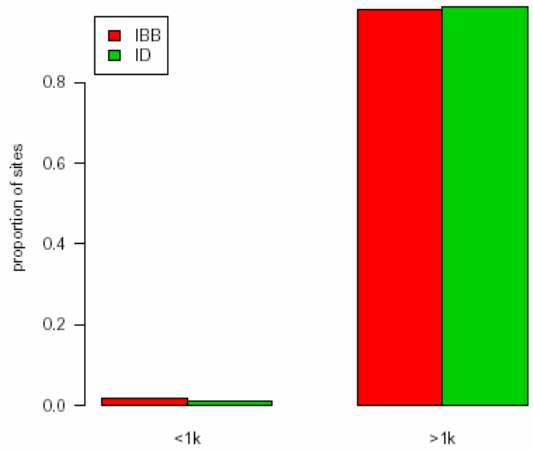
The model on which these coefficients are based include terms for whether the site is in a gene or not. Thus, the effect shown as 'in.exon' is net of that due to being in a gene. Note that in the barplot the bars above 'Not in Exon' include both introns and intergenic regions, so the impression given by the table may differ from that for the barplot.

3 CpG Island Neighborhoods

Here we study the effect of being in the neighborhood of CpG Islands. Following Wu and colleagues (Wu, *et al.*, 2003), who found that the neighborhoods within ±1kb of CpG islands are enriched for MLV insertions, we study such neighborhoods.

3.1 1 kilobase neighborhoods

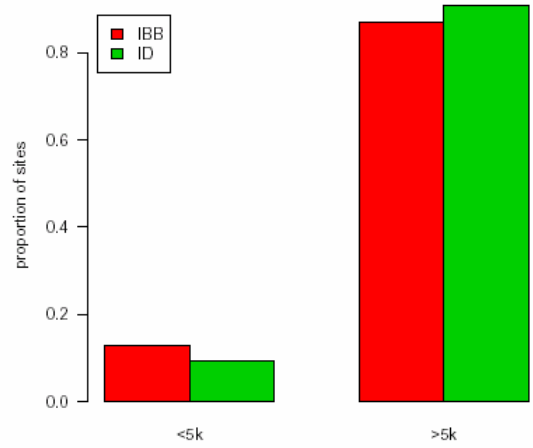
The following plot shows the effect of being in or within ±1kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.55736.

3.2 5 kilobase neighborhoods

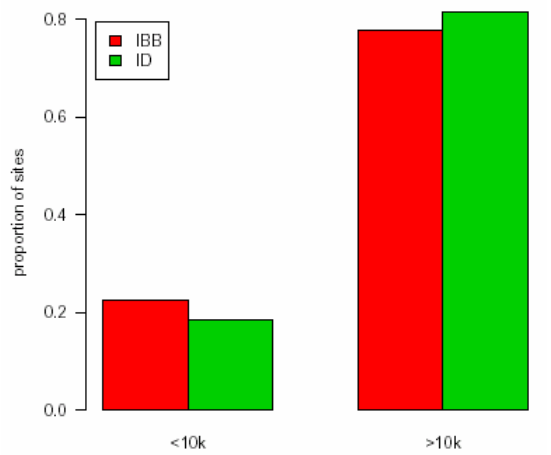
The following plot shows the effect of being in or within ±5kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.09008.

3.3 10 kilobase neighborhoods

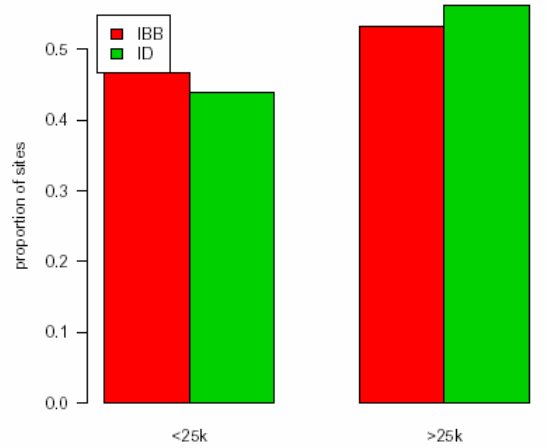
The following plot shows the effect of being in or within ±10kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.17307.

3.4 25 kilobase neighborhoods

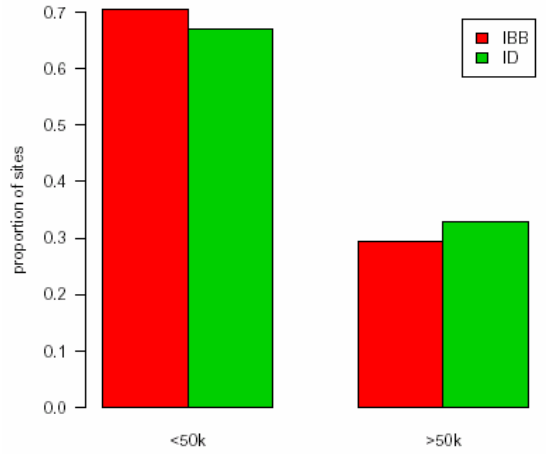
The following plot shows the effect of being in or within ±25kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.39436.

3.5 50 kilobase neighborhoods

The following plot shows the effect of being in or within ±50kb of a CpG island:



A formal test of significance comparing the difference attains a p-value of 0.28185.

4 Gene Density, Expression 'Density', and CpG Island Density

In this section the association with gene density is examined. The 'genes' that are counted are the Ensembl genes. In addition, we study various functions of the EST counts for the Ensembl genes using data described by Versteeg and colleagues (Versteeg *et al.*, 2003) and CpG Island density. Based on preliminary observations, it was decided to determine the density of ESTs found in a region in the following ways:

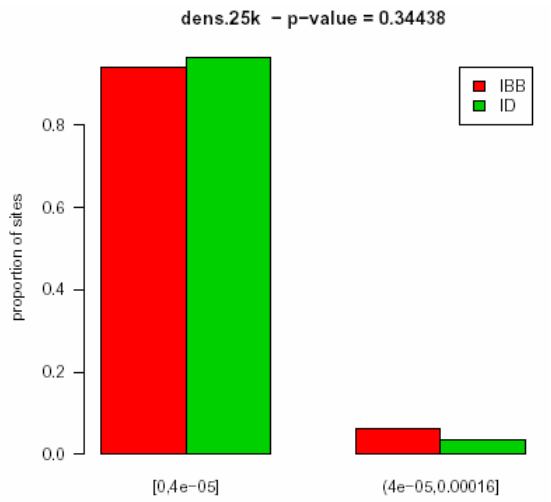
count.exprs Count only one EST per gene and divide by number of bases
exprs Count up to 200 ESTs per gene and divide by number of bases
big.exprs Count only the ESTs in excess of 200 per gene and divide by number of bases

The bolded terms are used as abbreviations in what follows. The abbreviation **dens** is used to indicate gene density as number of genes per base.

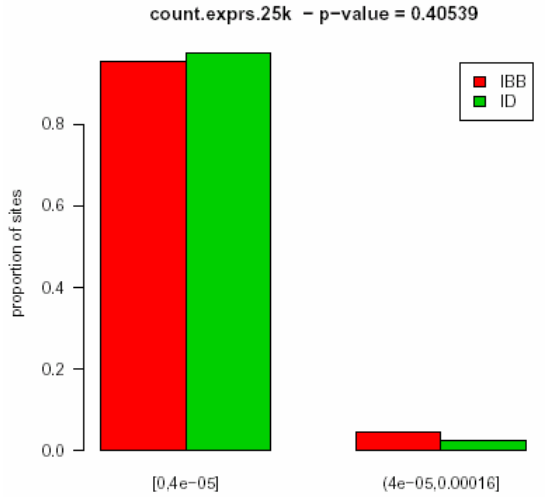
4.1 25 kilobase window

In the barplot that follows we examine the association of insertion sites with gene density in a 25 kilobase window surrounding each locus. More such plots will follow and the method of their construction is always to try to divide the data according to the deciles of density. However, it often happens that there is a very skewed distribution of density and even the 90th percentile is zero. In that case, the barplots simply show the sites for which the density is zero and those for which it is non-zero. If there are fewer than ten groups of bars, then the groupings contain ten percent of the sites each except for the leftmost grouping which will contain all of the remaining sites.

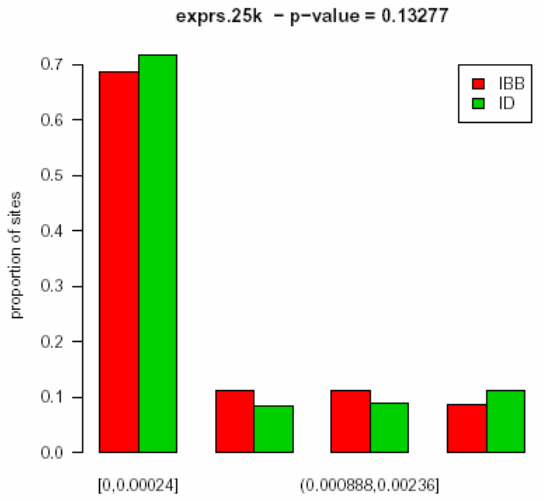
Also note that the title of the plot contains clues as to its content; the prefix indicates the type of variable studied while the suffix indicates the window width in the number of bases. The p-value given is the result of fitting a quadratic polynomial to the gene density values.



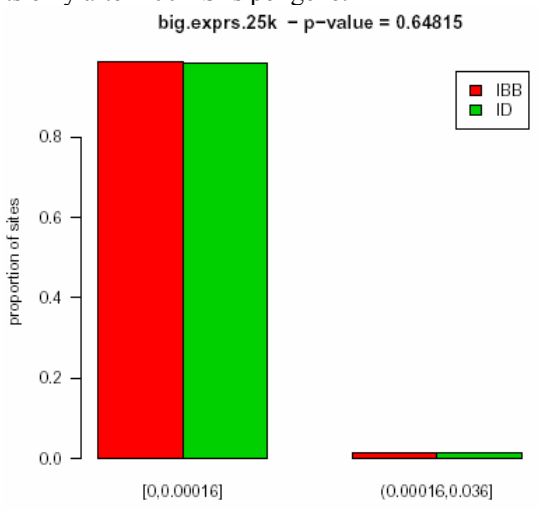
In the barplot that follows we examine the association of insertion sites with expression density in a 25 kilobase window surrounding each locus. First, we count just one EST per gene:



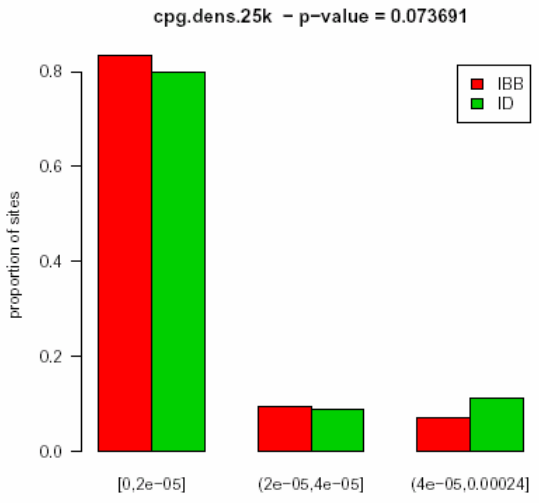
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

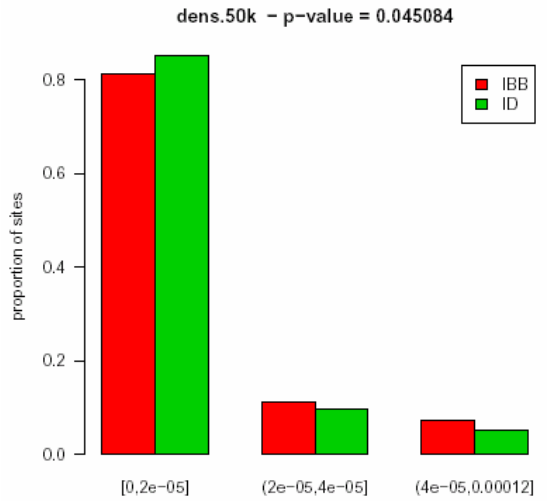


Here the effect of density of CpG islands is studied:

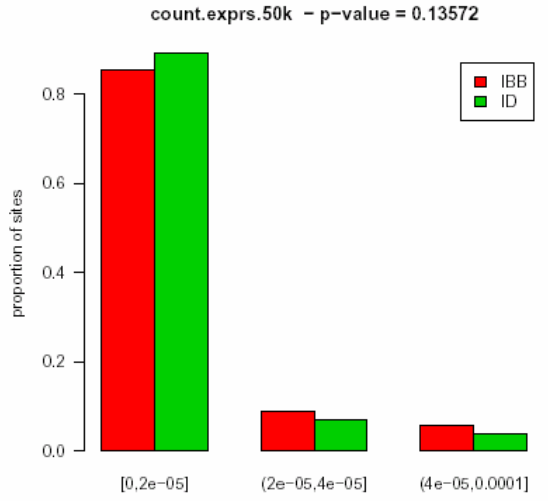


4.2 50 kilobase window

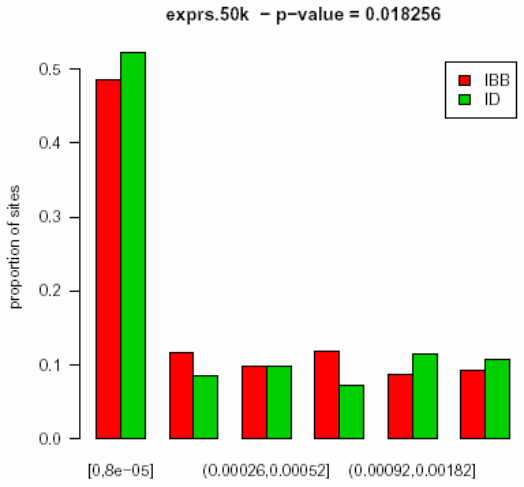
First, we see gene density:



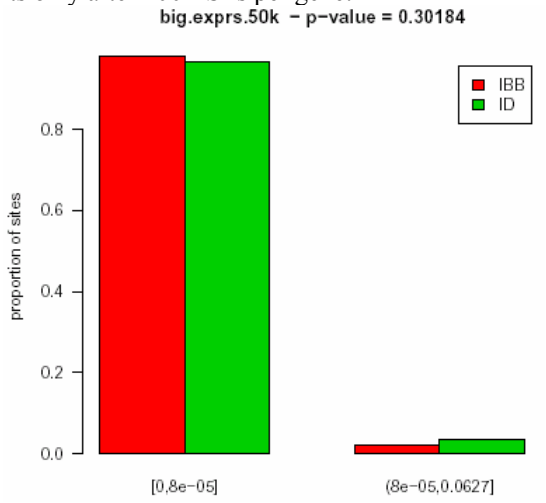
Here are the results for EST density. First, we count just one EST per gene:



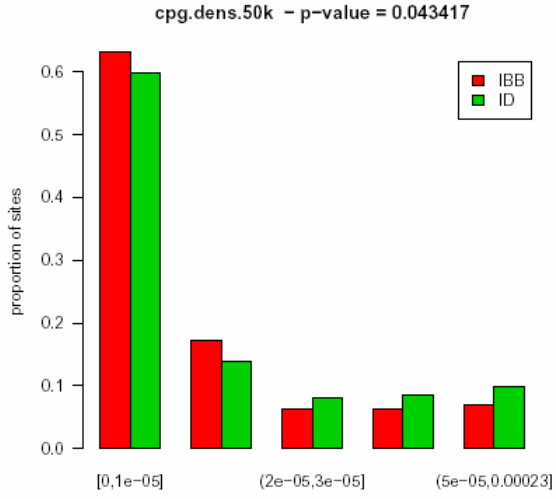
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

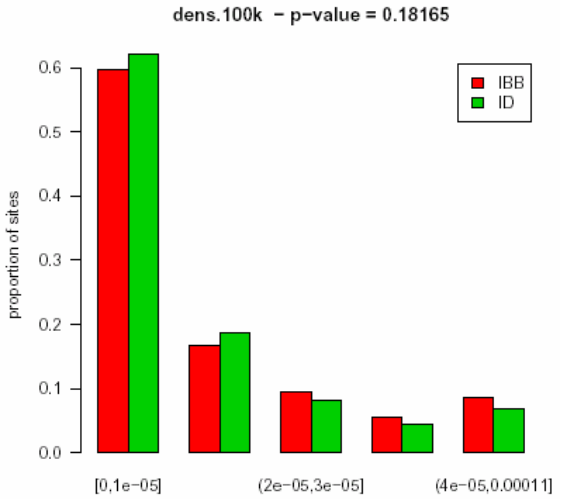


Here the effect of density of CpG islands is studied:

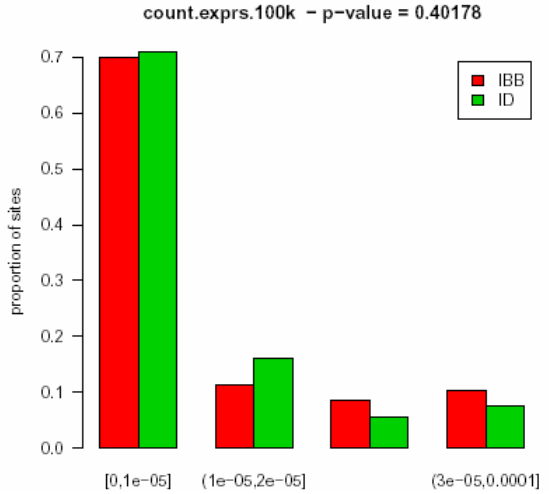


4.3 100 kilobase window

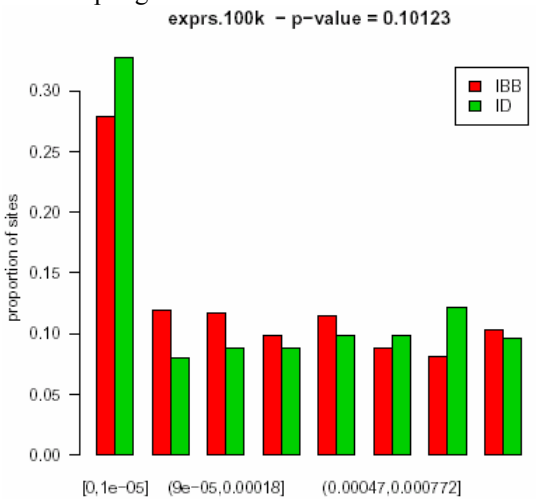
First, we see gene density:



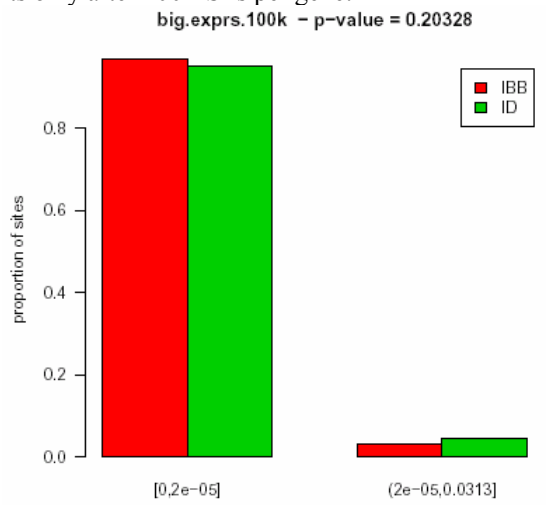
Here are the results for EST density. First, we count just one EST per gene.



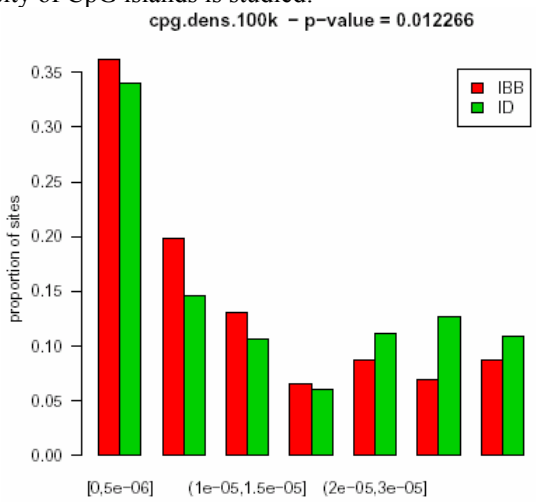
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

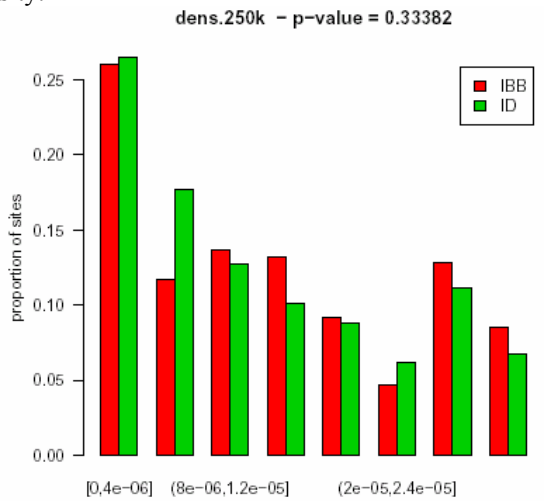


Here the effect of density of CpG islands is studied:

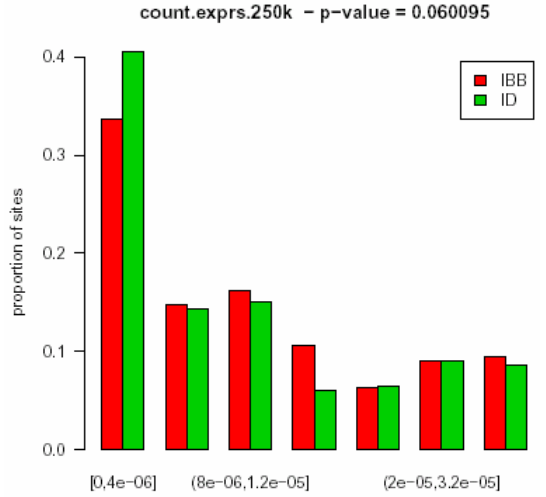


4.4 250 kilobase window

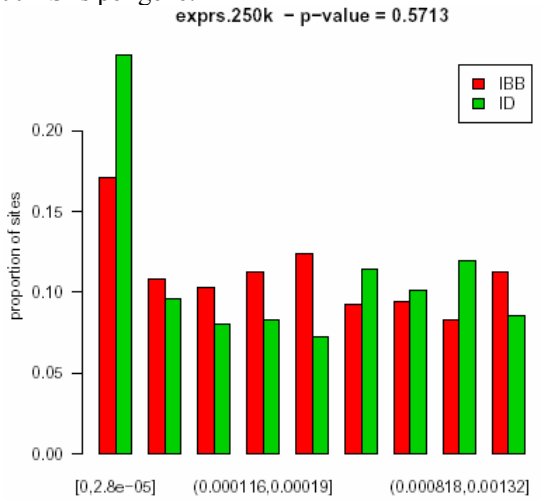
First, we see gene density:



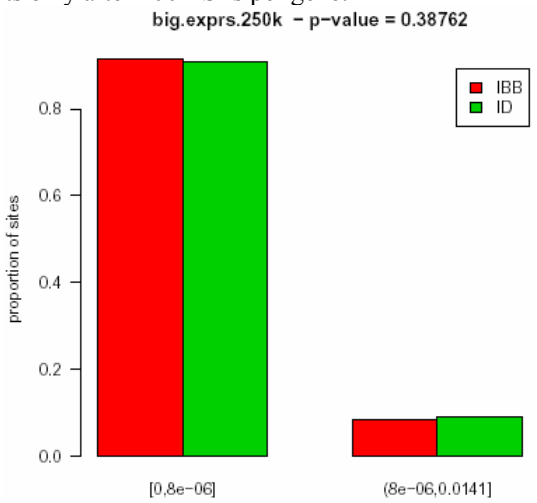
Here are the results for EST density. First, we count just one EST per gene:



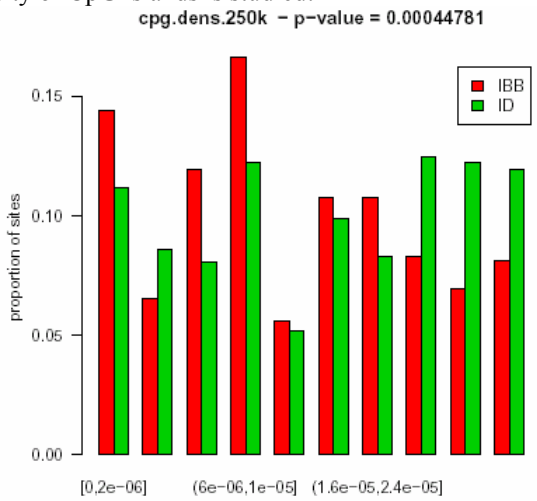
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

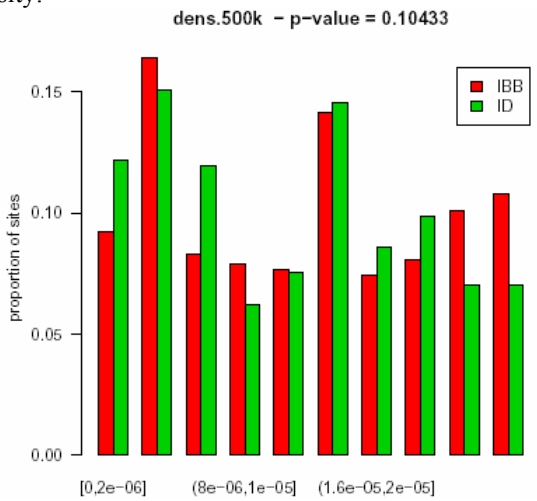


Here the effect of density of CpG islands is studied:

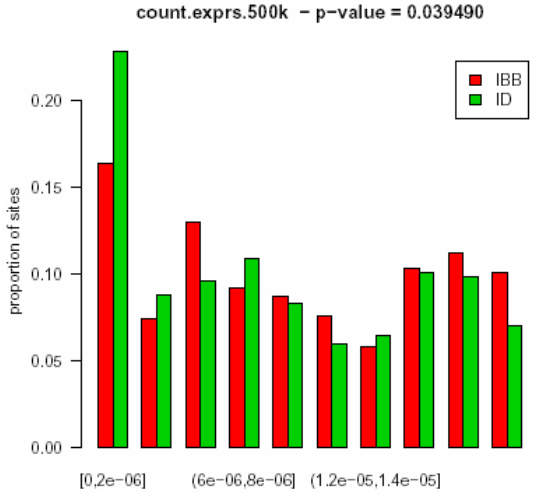


4.5 500 kilobase window

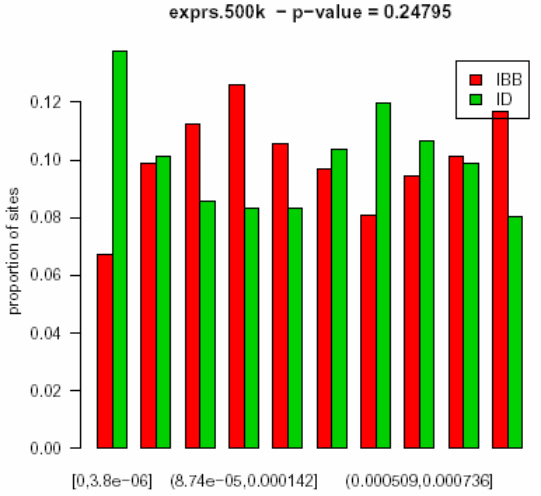
First, we see gene density:



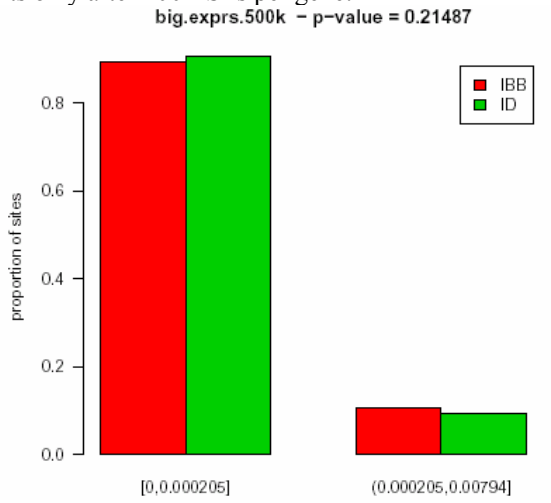
Here are the results for EST density. First, we count just one EST per gene:



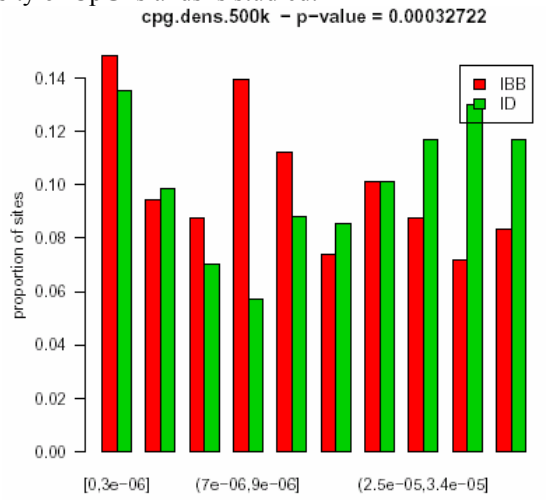
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

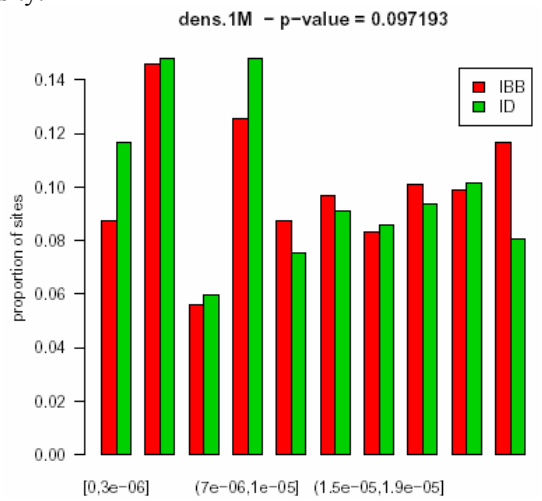


Here the effect of density of CpG islands is studied:

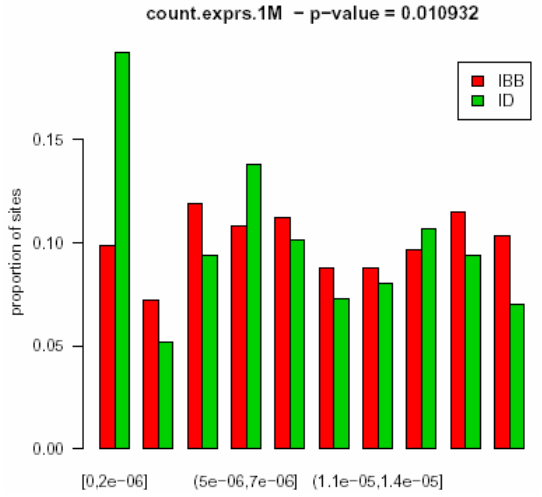


4.6 1 megabase window

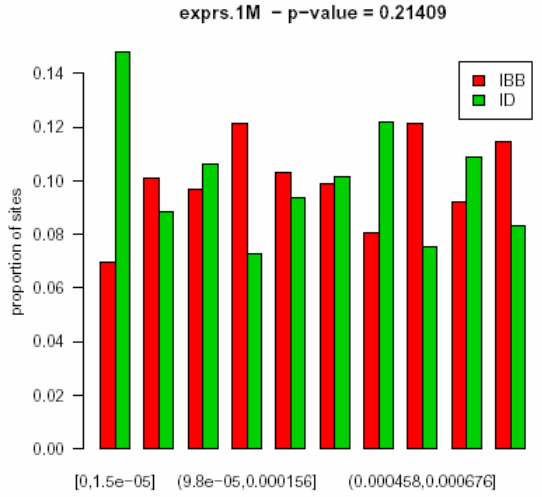
First, we see gene density:



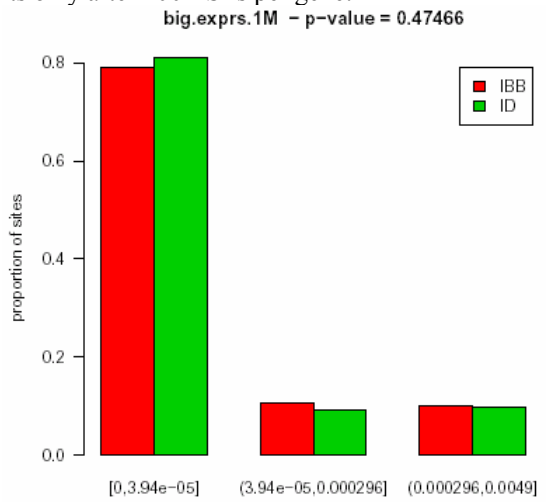
Here are the results for EST density. First, we count just one EST per gene:



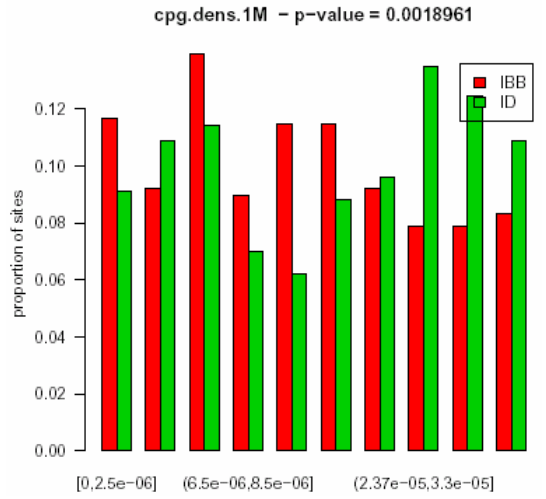
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

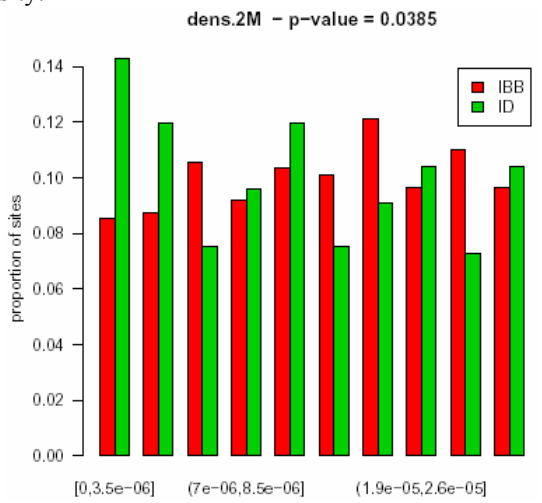


Here the effect of density of CpG islands is studied:

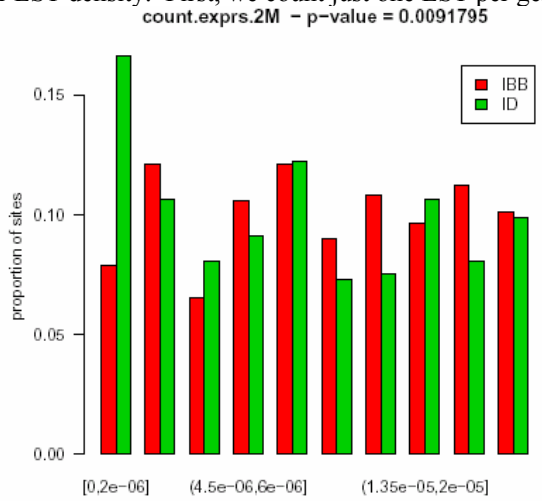


4.7 2 megabase window

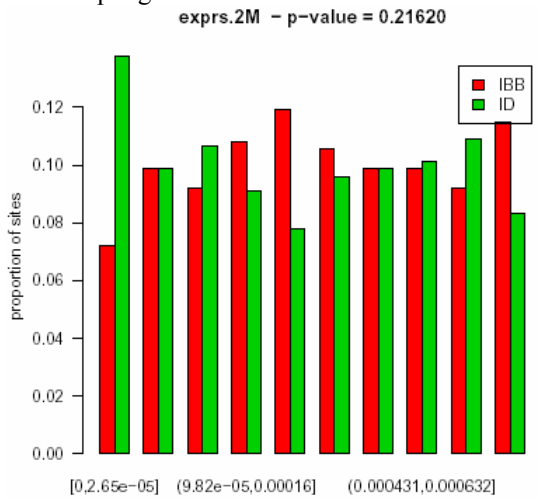
First, we see gene density:



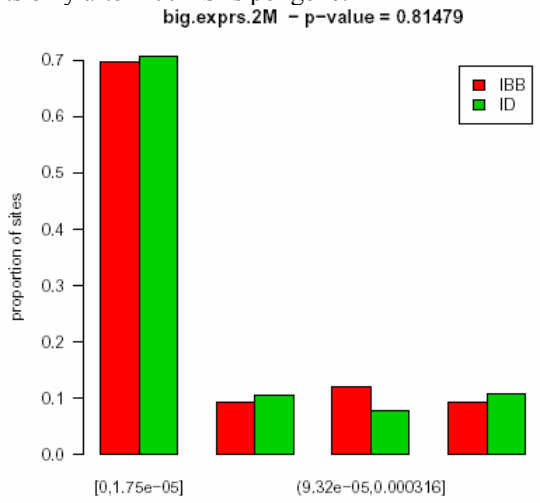
Here are the results for EST density. First, we count just one EST per gene:



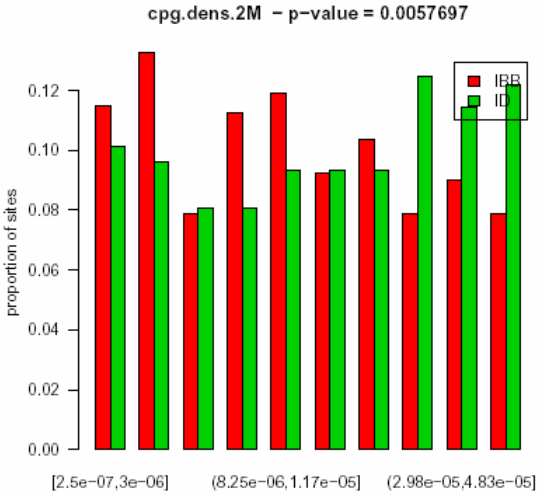
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

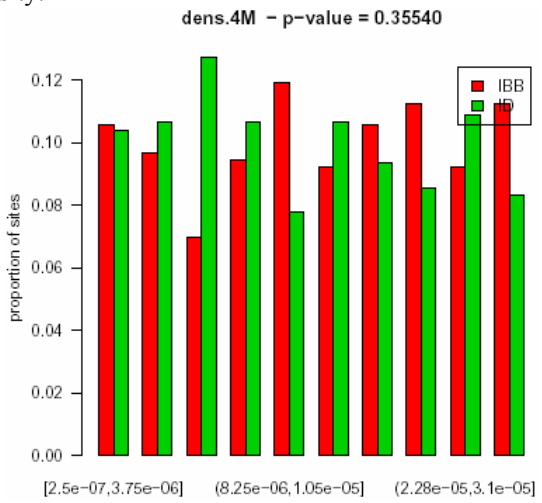


Here the effect of density of CpG islands is studied:

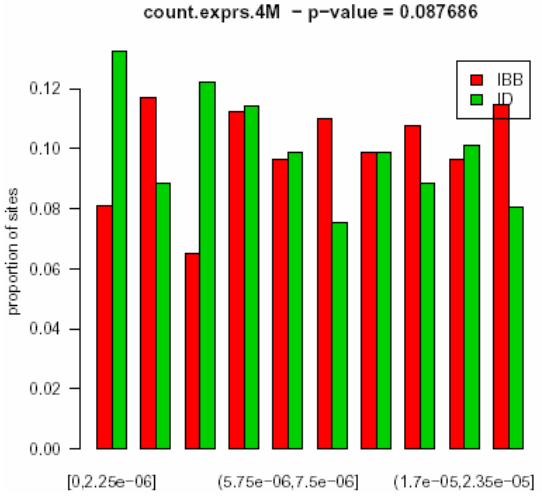


4.8 4 megabase window

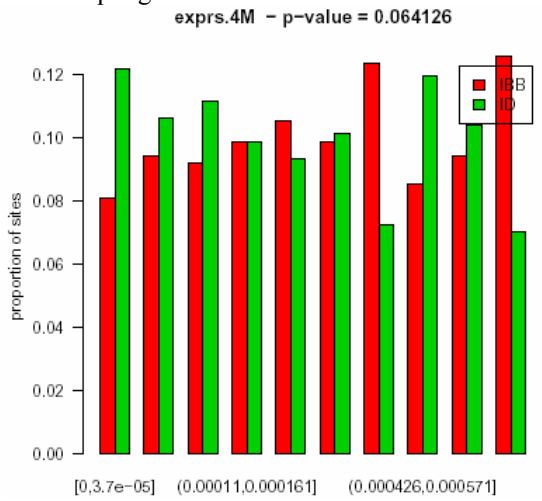
First, we see gene density:



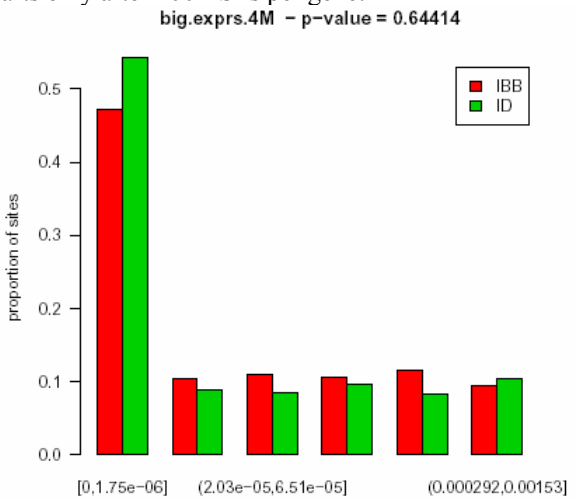
Here are the results for EST density. First, we count just one EST per gene:



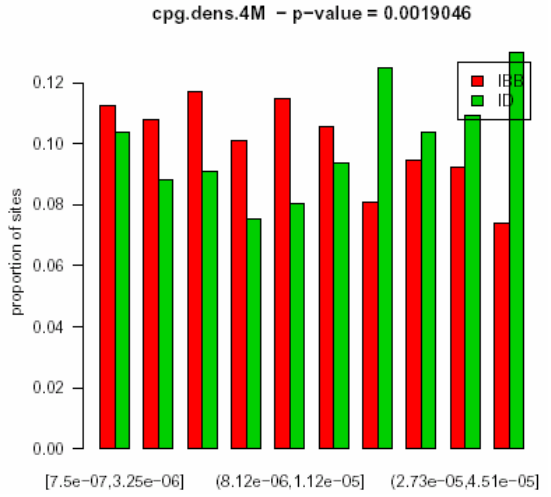
Now we count up to 200 ESTs per gene:



And here counting starts only after 200 ESTs per gene:

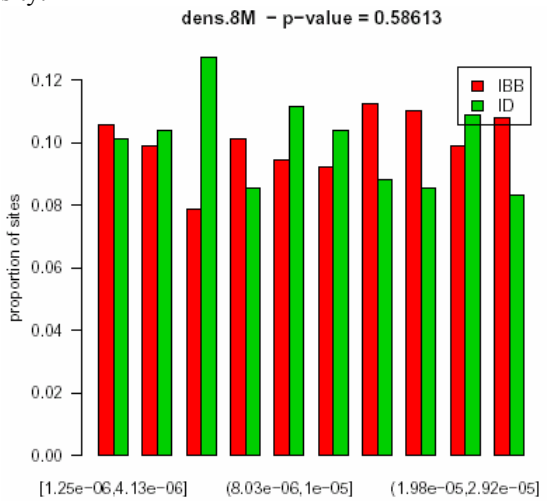


Here the effect of density of CpG islands is studied:

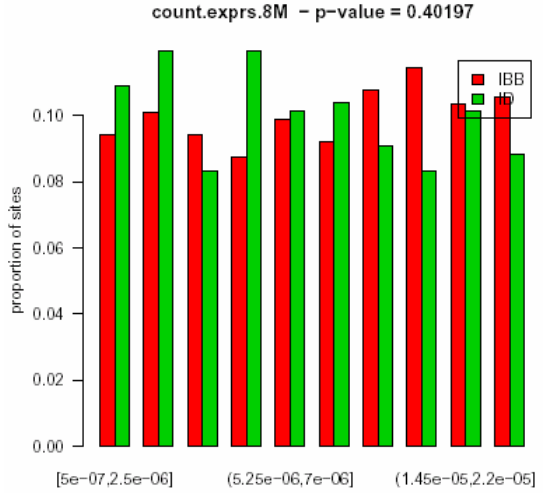


4.9 8 megabase window

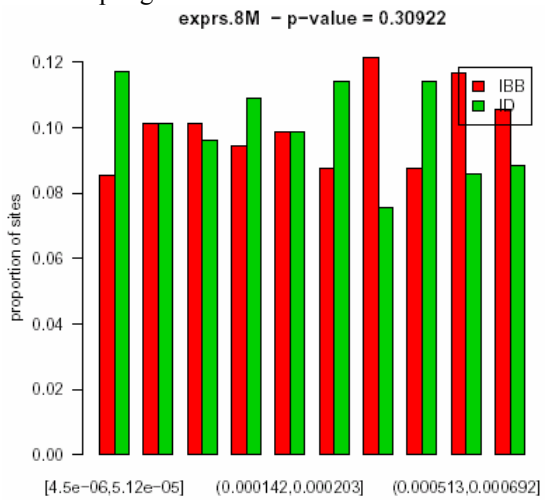
First, we see gene density:



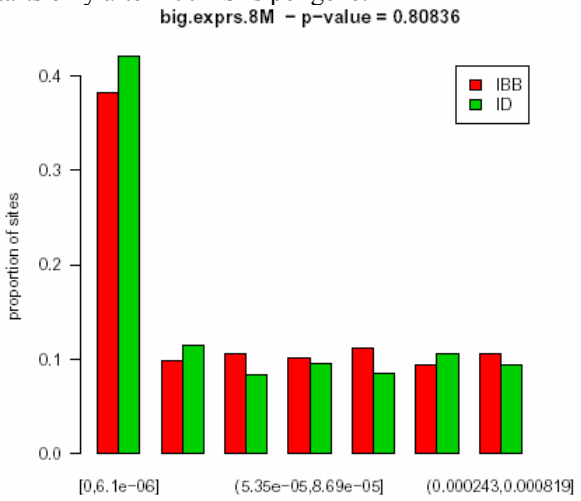
Here are the results for EST density. First, we count just one EST per gene:



Now we count up to 200 ESTs per gene:

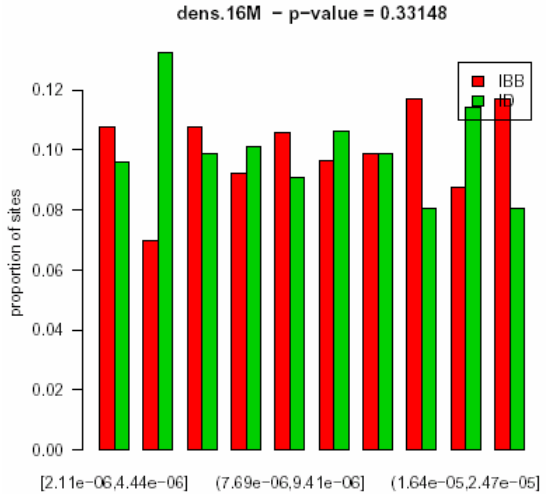


And here counting starts only after 200 ESTs per gene:

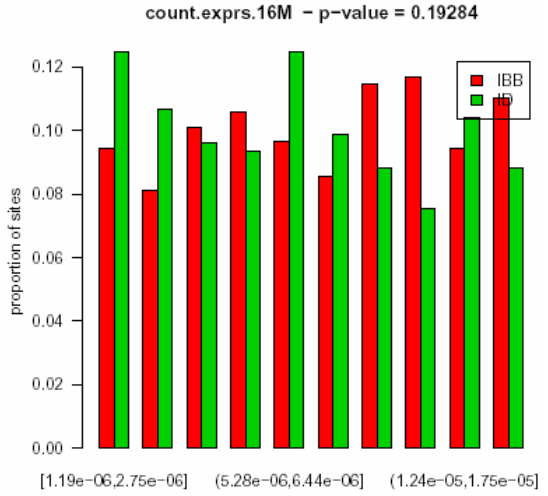


4.10 16 megabase window

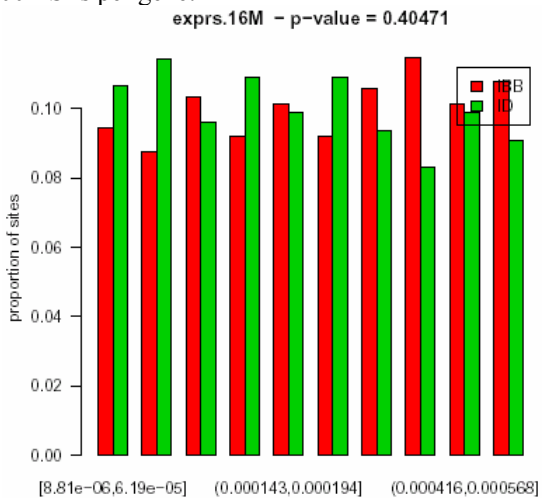
First, we see gene density:



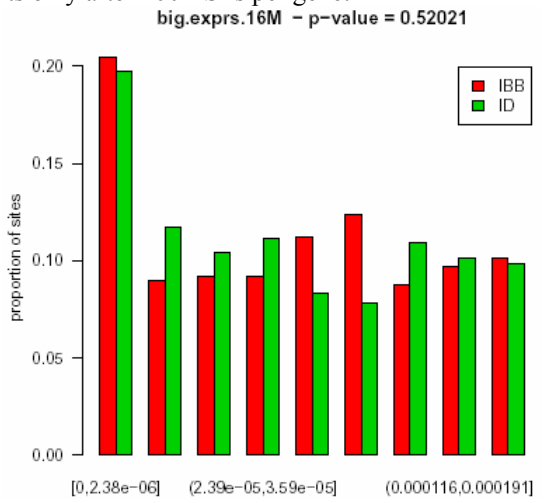
Here are the results for EST density. First, we count just one EST per gene:



Now we count up to 200 ESTs per gene:

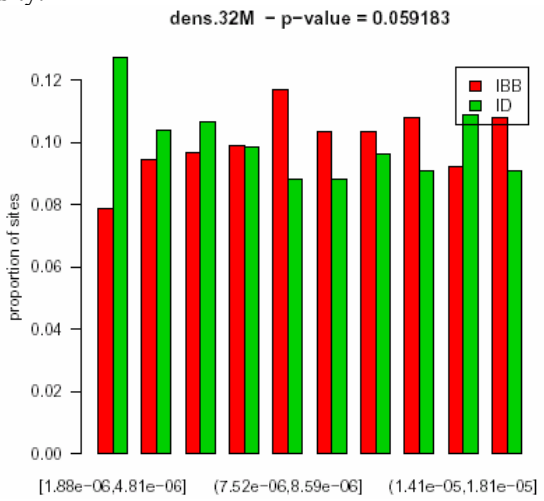


And here counting starts only after 200 ESTs per gene:

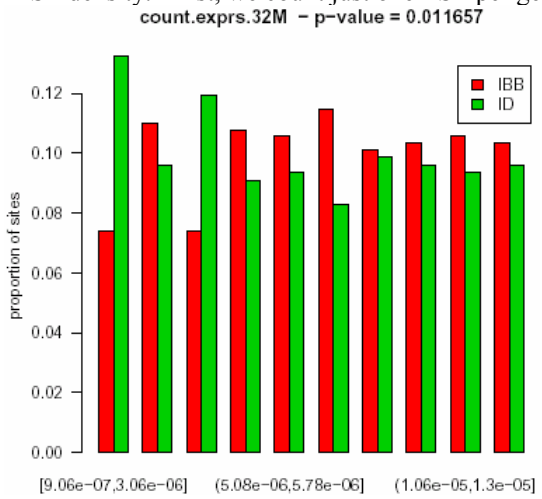


4.11 32 megabase window

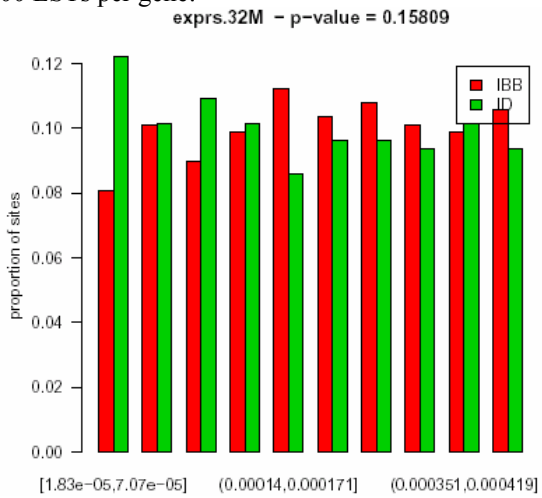
First, we see gene density:



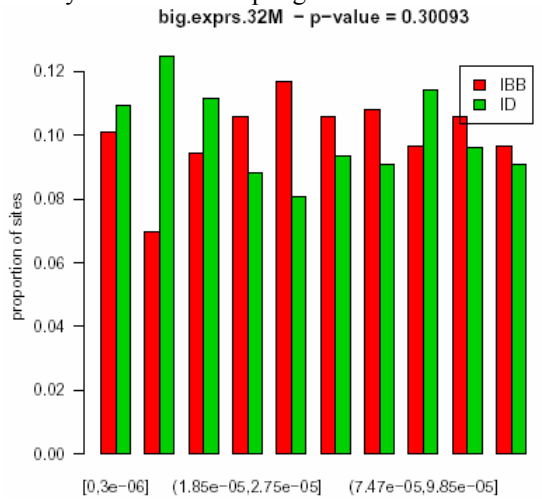
Here are the results for EST density. First, we count just one EST per gene:



Now we count up to 200 ESTs per gene:



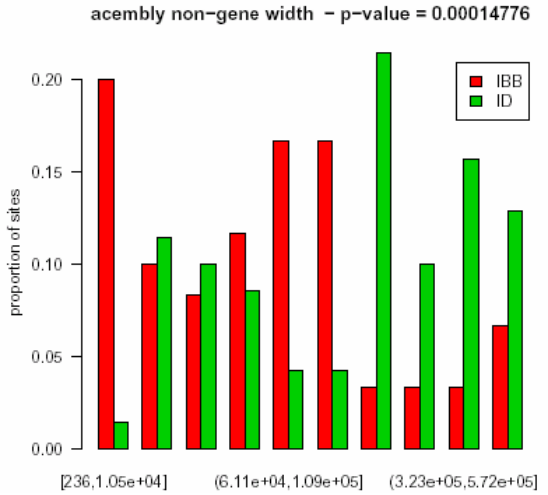
And here counting starts only after 200 ESTs per gene:



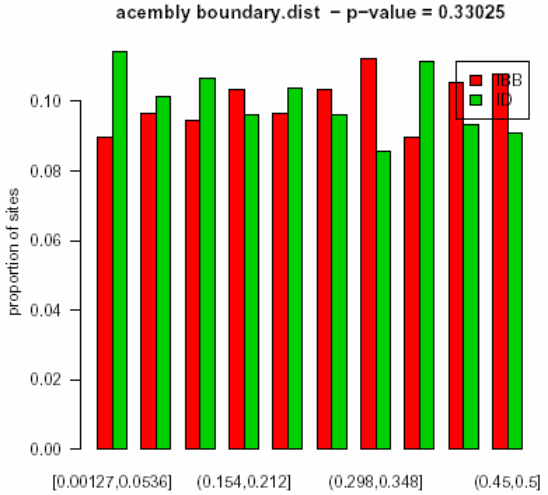
5 Juxtaposition with Gene Start and End Positions

5.1 Acembly Annotations

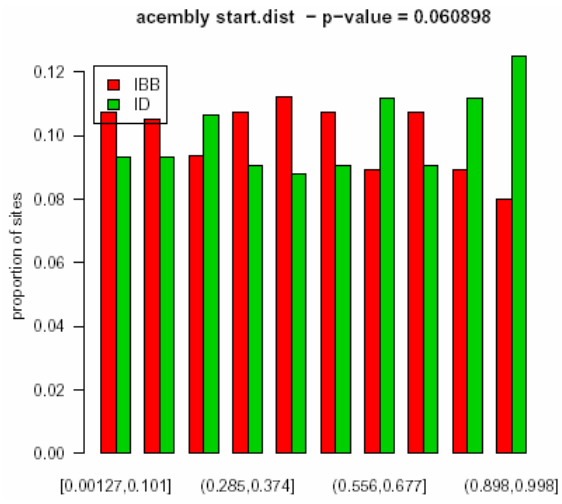
In this section we study the effect of juxtaposition in terms of gene start and end positions. The first barplot shows the effect of gene width for those insertions that are located within an Acembly gene. The next plot uses the width of a non-gene region for insertions that fall into such regions.



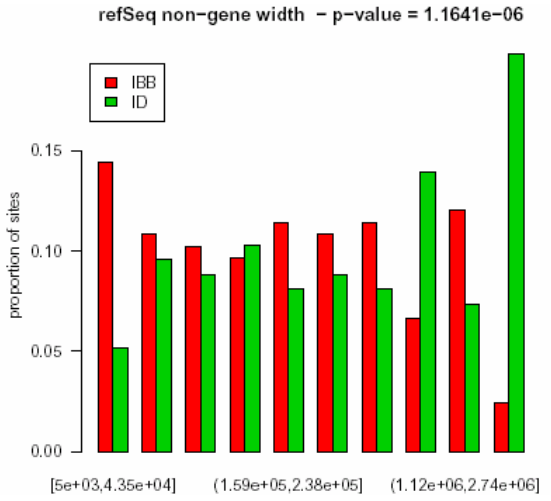
The next plot studies the distance to the nearest boundary between a gene and a non-gene region. The distance is expressed as a fraction of the length of the region. Thus, '0.25' refers to one quarter of the distance from the site to nearest boundary divided by the total width of the region.

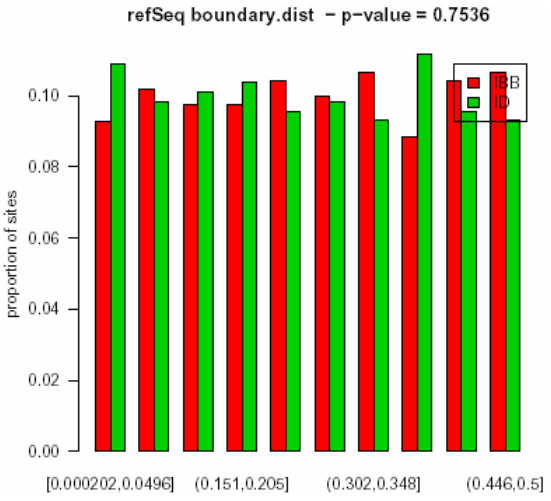


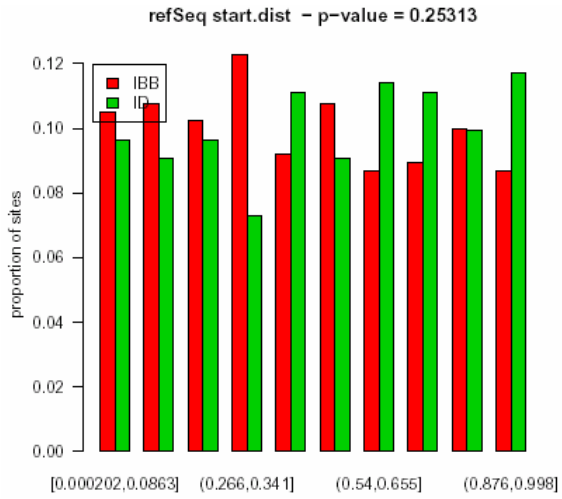
This plot studies the effect of nearness to the beginning of a transcript. For sites in genes, it is the distance to the start of the gene divided by the width of the gene. For other sites it is the distance from the site to the nearer gene if that gene boundary is also a transcription starting point. Locations near '0' are relatively near the beginning of transcription, while those near '1' are near the termination of the transcript.



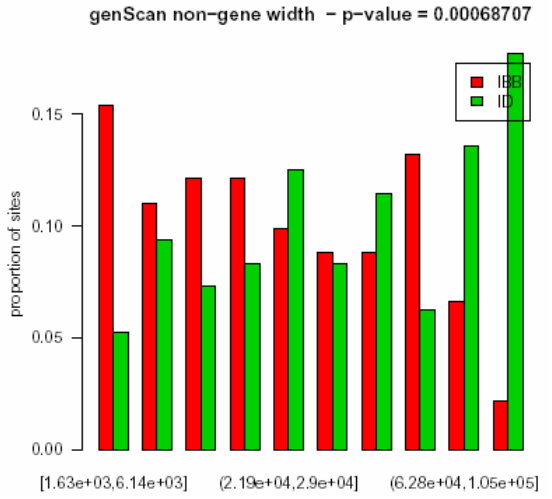
5.2 RefSeq Annotations

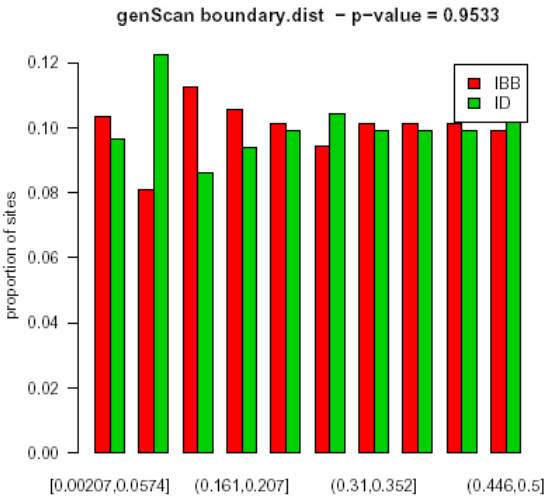


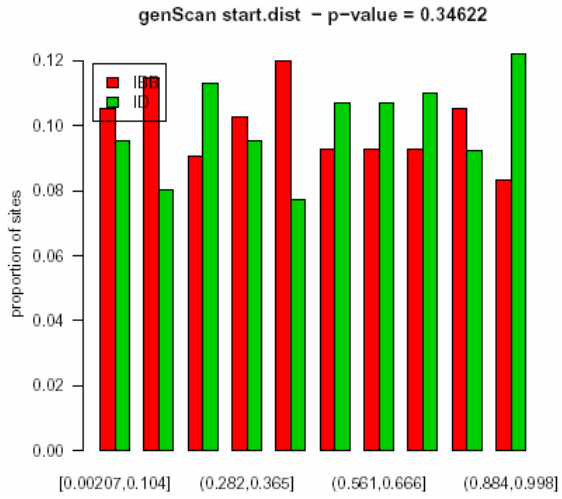




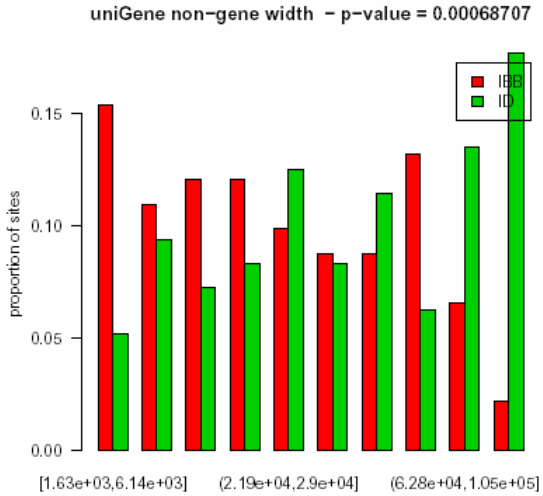
5.3 GenScan Annotations

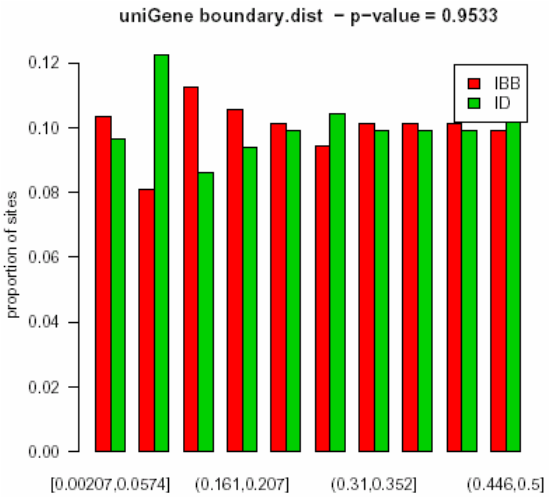


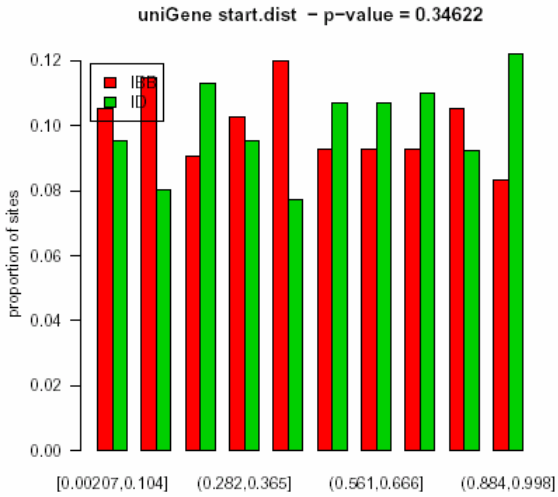




5.4 UniGene Annotations



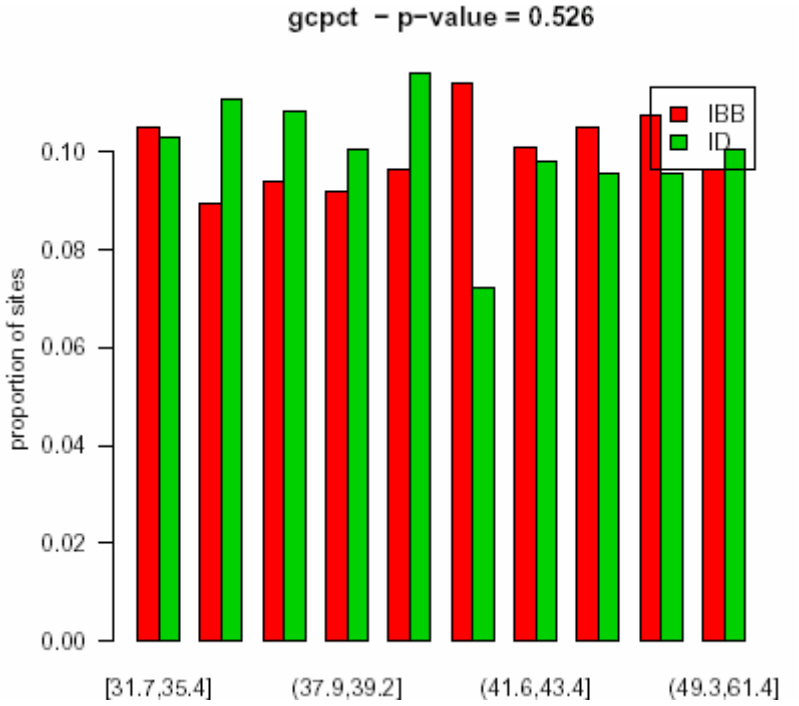




6 GC content

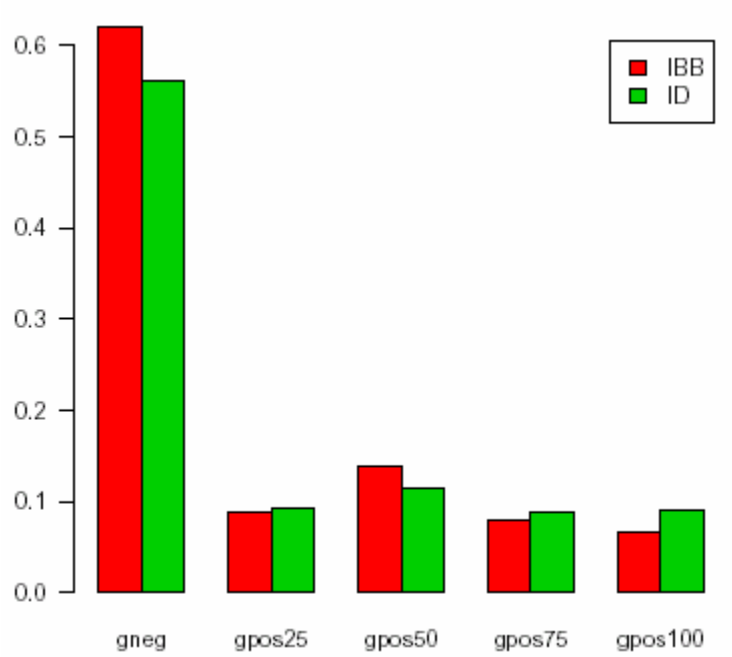
Here we study the effect of GC content on insertion. The GC content is taken from the Human Genome Draft at GoldenPath from the table http://genome.ucsc.edu/goldenPath/14nov2002/database/gcPercent.txt.gz.

Following the plot is a table of fitted coefficients based on splitting the GC percent data at the median.



7 Cytobands

Here we study the association of cytoband with insertion intensity. The data are obtained from http://genome.ucsc.edu/goldenPath/14nov2002/database/cytoBand.txt.gz.



A formal test of significance attains a p-value of 0.41588.

REFERENCES

Aiyar, A., Hindmarsh, P., Skalka, A. M., and Leis, J. (1996). Concerted integration of linear retroviral DNA by the avian sarcoma virus integrase in vitro: dependence on both long terminal repeat termini. *J Virol* **70,** 3571-80.

Andrake, M. D., and Skalka, A. M. (1995). Multimerization determinants reside in both the catalytic core and C terminus of avian sarcoma virus integrase. *J Biol Chem* **270,** 29299-306.

Appa, R. S., Shin, C. G., Lee, P., and Chow, S. A. (2001). Role of the nonspecific DNA-binding region and alpha helices within the core domain of retroviral integrase in selecting target DNA sites for integration. *J Biol Chem* **276,** 45848-55.

Balakrishnan, M., and Jonsson, C. B. (1997). Functional identification of nucleotides conferring substrate specificity to retroviral integrase reactions. *J Virol* **71,** 1025-35.

Bao, K. K., Wang, H., Miller, J. K., Erie, D. A., Skalka, A. M., and Wong, I. (2003). Functional oligomeric state of avian sarcoma virus integrase. *J Biol Chem* **278,** 1323-7.

Barr, S. D., Leipzig, J., Shinn, P., Ecker, J. R., and Bushman, F. D. (2005). Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J Virol* **79,** 12035-44.

Beitzel, B., and Bushman, F. (2003). Construction and analysis of cells lacking the HMGA gene family. *Nucleic Acids Res* **31,** 5025-32.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321,** 209-13.

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). "Discrete multivariate analyses: Theory and practice." MIT Press.

Blankson, J. N., Persaud, D., and Siliciano, R. F. (2002). The challenge of viral reservoirs in HIV-1 infection. *Annu Rev Med* **53,** 557-93.

Boeke, J. D., and Devine, S. E. (1998). Yeast retrotransposons: finding a nice quiet neighborhood. *Cell* **93,** 1087-9.

Bor, Y. C., Bushman, F. D., and Orgel, L. E. (1995). In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc Natl Acad Sci U S A* **92,** 10334-8.

Bor, Y. C., Miller, M. D., Bushman, F. D., and Orgel, L. E. (1996). Target-sequence preferences of HIV-1 integration complexes in vitro. *Virology* **222,** 283-8.

Bowerman, B., Brown, P. O., Bishop, J. M., and Varmus, H. E. (1989). A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev* **3,** 469-78.

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., and Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet* **10,** 211-9.

Breiman, L. (2001). Random forests. *Machine Learning* **45**.

Brown, K. E., Baxter, J., Graf, D., Merkenschlager, M., and Fisher, A. G. (1999). Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. *Mol Cell* **3,** 207-17.

Brown, P. O., Bowerman, B., Varmus, H. E., and Bishop, J. M. (1987). Correct integration of retroviral DNA in vitro. *Cell* **49,** 347-56.

Brown, P. O., Bowerman, B., Varmus, H. E., and Bishop, J. M. (1989). Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc Natl Acad Sci U S A* **86,** 2525-9.

Buckman, J. S., Bosche, W. J., and Gorelick, R. J. (2003). Human immunodeficiency virus type 1 nucleocapsid zn(2+) fingers are required for efficient reverse transcription, initial integration processes, and protection of newly synthesized viral DNA. *J Virol* **77,** 1469-80.

Bujacz, G., Alexandratos, J., Qing, Z. L., Clement-Mella, C., and Wlodawer, A. (1996a). The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant. *FEBS Lett* **398,** 175-8.

Bujacz, G., Alexandratos, J., Wlodawer, A., Merkel, G., Andrake, M., Katz, R. A., and Skalka, A. M. (1997). Binding of different divalent cations to the active site of avian sarcoma virus integrase and their effects on enzymatic activity. *J Biol Chem* **272,** 18161-8.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R. A., and Skalka, A. M. (1995). High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J Mol Biol* **253,** 333-46.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R. A., and Skalka, A. M. (1996b). The catalytic domain of avian sarcoma virus integrase: conformation of the active-site residues in the presence of divalent cations. *Structure* **4,** 89-96.

Bushman, F. D. (1994). Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc Natl Acad Sci U S A* **91,** 9233-7.

Bushman, F. D. (2001). "Lateral DNA Transfer: Mechanisms and Consequences." Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bushman, F. D. (2003). Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115,** 135-8.

Bushman, F. D., and Craigie, R. (1990). Sequence requirements for integration of Moloney murine leukemia virus DNA in vitro. *J Virol* **64,** 5645-8.

Bushman, F. D., and Craigie, R. (1991). Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proc Natl Acad Sci U S A* **88,** 1339-43.

Bushman, F. D., and Craigie, R. (1992). Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites. *Proc Natl Acad Sci U S A* **89,** 3458-62.

Bushman, F. D., Engelman, A., Palmer, I., Wingfield, P., and Craigie, R. (1993). Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding. *Proc Natl Acad Sci U S A* **90,** 3428-32.

Bushman, F. D., Fujiwara, T., and Craigie, R. (1990). Retroviral DNA integration directed by HIV integration protein in vitro. *Science* **249,** 1555-8.

Bushman, F. D., and Miller, M. D. (1997). Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J Virol* **71,** 458-64.

Bushman, F. D., and Wang, B. (1994). Rous sarcoma virus integrase protein: mapping functions for catalysis and substrate binding. *J Virol* **68,** 2215-23.

Cai, M., Huang, Y., Zheng, R., Wei, S. Q., Ghirlando, R., Lee, M. S., Craigie, R., Gronenborn, A. M., and Clore, G. M. (1998). Solution structure of the cellular factor BAF responsible for protecting retroviral DNA from autointegration. *Nat Struct Biol* **5,** 903-9.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G. M., and Gronenborn, A. M. (1997). Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat Struct Biol* **4,** 567-77.

Callen, B. P., Shearwin, K. E., and Egan, J. B. (2004). Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol Cell* **14,** 647-56.

Carson, J. P., Zhang, N., Frampton, G. M., Gerry, N. P., Lenburg, M. E., and Christman, M. F. (2004). Pharmacogenomic identification of targets for adjuvant therapy with the topoisomerase poison camptothecin. *Cancer Res* **64,** 2096-104.

Carteau, S., Batson, S. C., Poljak, L., Mouscadet, J. F., de Rocquigny, H., Darlix, J. L., Roques, B. P., Kas, E., and Auclair, C. (1997). Human immunodeficiency virus type 1 nucleocapsid protein specifically stimulates Mg2+-dependent DNA integration in vitro. *J Virol* **71,** 6225-9.

Carteau, S., Gorelick, R. J., and Bushman, F. D. (1999). Coupled integration of human immunodeficiency virus type 1 cDNA ends by purified integrase in vitro: stimulation by the viral nucleocapsid protein. *J Virol* **73,** 6670-9.

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

Casolari, J. M., Brown, C. R., Komili, S., West, J., Hieronymus, H., and Silver, P. A. (2004). Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell* **117,** 427-39.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nat Genet* **31,** 415-8.

Chen, H., and Engelman, A. (1998). The barrier-to-autointegration protein is a host factor for HIV type 1 integration. *Proc Natl Acad Sci U S A* **95,** 15270-4.

Chen, J. C., Krucinski, J., Miercke, L. J., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., and Stroud, R. M. (2000a). Crystal structure of the HIV-1 integrase

catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci U S A* **97,** 8233-8.

Chen, Z., Yan, Y., Munshi, S., Li, Y., Zugay-Murphy, J., Xu, B., Witmer, M., Felock, P., Wolfe, A., Sardana, V., Emini, E. A., Hazuda, D., and Kuo, L. C. (2000b). X-ray structure of simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50-293)--an initial glance of the viral DNA binding platform. *J Mol Biol* **296,** 521-33.

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., and Debyser, Z. (2003). HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem* **278,** 372-81.

Chow, S. A., Vincent, K. A., Ellison, V., and Brown, P. O. (1992). Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science* **255,** 723-6.

Chubb, J. R., and Bickmore, W. A. (2003). Considering nuclear compartmentalization in the light of nuclear dynamics. *Cell* **112,** 403-6.

Chun, T. W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J. A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T. C., Kuo, Y. H., Brookmeyer, R., Zeiger, M. A., Barditch-Crovo, P., and Siliciano, R. F. (1997a). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183-8.

Chun, T. W., Stuyver, L., Mizell, S. B., Ehler, L. A., Mican, J. A., Baseler, M., Lloyd, A. L., Nowak, M. A., and Fauci, A. S. (1997b). Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc Natl Acad Sci U S A* **94,** 13193-7.

Coffin, J. M., Hughes, S. H., and Varmus, H. E. (1997). "Retroviruses." Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Colicelli, J., and Goff, S. P. (1985). Mutants and pseudorevertants of Moloney murine leukemia virus with alterations at the integration site. *Cell* **42,** 573-80.

Colicelli, J., and Goff, S. P. (1988). Sequence and spacing requirements of a retrovirus integration site. *J Mol Biol* **199,** 47-59.

Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. (eds.) (2002). "Mobile DNA II." American Society Microbiology.

Craigie, R., Fujiwara, T., and Bushman, F. (1990). The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration in vitro. *Cell* **62,** 829-37.

Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E. D., Wolfsberg, T. G., and Collins, F. S. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci U S A* **101,** 992-7.

Cullen, B. R., Lomedico, P. T., and Ju, G. (1984). Transcriptional interference in avian retroviruses--implications for the promoter insertion model of leukaemogenesis. *Nature* **307,** 241-5.

Dave, U. P., Jenkins, N. A., and Copeland, N. G. (2004). Gene therapy insertional mutagenesis insights. *Science* **303,** 333.

de la Fuente, C., Santiago, F., Deng, L., Eadie, C., Zilberman, I., Kehn, K., Maddukuri, A., Baylor, S., Wu, K., Lee, C. G., Pumfery, A., and Kashanchi, F. (2002). Gene expression profile of HIV-1 Tat expressing cells: a close interplay between proliferative and differentiation signals. *BMC Biochem* **3,** 14.

Deprez, E., Tauc, P., Leh, H., Mouscadet, J. F., Auclair, C., and Brochon, J. C. (2000). Oligomeric states of the HIV-1 integrase as measured by time-resolved fluorescence anisotropy. *Biochemistry* **39,** 9275-84.

Dhar, R., McClements, W. L., Enquist, L. W., and Vande Woude, G. F. (1980). Nucleotide sequences of integrated Moloney sarcoma provirus long terminal repeats and their host and viral junctions. *Proc Natl Acad Sci U S A* **77,** 3937-41.

Dirac, A. M., and Kjems, J. (2001). Mapping DNA-binding sites of HIV-1 integrase by protein footprinting. *Eur J Biochem* **268,** 743-51.

Donehower, L. A. (1988). Analysis of mutant Moloney murine leukemia viruses containing linker insertion mutations in the 3' region of pol. *J Virol* **62,** 3958-64.

Donehower, L. A., and Varmus, H. E. (1984). A mutant murine leukemia virus with a single missense codon in pol is defective in a function affecting integration. *Proc Natl Acad Sci U S A* **81,** 6461-5.

Drelich, M., Wilhelm, R., and Mous, J. (1992). Identification of amino acid residues critical for endonuclease and integration activities of HIV-1 IN protein in vitro. *Virology* **188,** 459-68.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R., and Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* **266,** 1981-6.

Eijkelenboom, A. P., Lutzke, R. A., Boelens, R., Plasterk, R. H., Kaptein, R., and Hard, K. (1995). The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nat Struct Biol* **2,** 807-10.

Eijkelenboom, A. P., Sprangers, R., Hard, K., Puras Lutzke, R. A., Plasterk, R. H., Boelens, R., and Kaptein, R. (1999). Refined solution structure of the C-terminal DNA-binding domain of human immunovirus-1 integrase. *Proteins* **36,** 556-64.

Eijkelenboom, A. P., van den Ent, F. M., Vos, A., Doreleijers, J. F., Hard, K., Tullius, T. D., Plasterk, R. H., Kaptein, R., and Boelens, R. (1997). The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Curr Biol* **7,** 739-46.

Elleder, D., Pavlicek, A., Paces, J., and Hejnar, J. (2002). Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence. *FEBS Lett* **517,** 285-6.

Ellison, V., Abrams, H., Roe, T., Lifson, J., and Brown, P. (1990). Human immunodeficiency virus integration in a cell-free system. *J Virol* **64,** 2711-5.

Ellison, V., and Brown, P. O. (1994). A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration in vitro. *Proc Natl Acad Sci U S A* **91,** 7316-20.

Emerman, M., and Malim, M. H. (1998). HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science* **280,** 1880-4.

Engelman, A. (2005). The ups and downs of gene expression and retroviral DNA integration. *Proc Natl Acad Sci U S A* **102,** 1275-6.

Engelman, A., Bushman, F. D., and Craigie, R. (1993). Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *Embo J* **12,** 3269-75.

Engelman, A., and Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *J Virol* **66,** 6361-9.

Engelman, A., Englund, G., Orenstein, J. M., Martin, M. A., and Craigie, R. (1995). Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. *J Virol* **69,** 2729-36.

Engelman, A., Hickman, A. B., and Craigie, R. (1994). The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *J Virol* **68,** 5911-7.

Engelman, A., Mizuuchi, K., and Craigie, R. (1991). HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67,** 1211-21.

Englund, G., Theodore, T. S., Freed, E. O., Engelman, A., and Martin, M. A. (1995). Integration is required for productive infection of monocyte-derived macrophages by human immunodeficiency virus type 1. *J Virol* **69,** 3216-9.

Esposito, D., and Craigie, R. (1998). Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. *Embo J* **17,** 5832-43.

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

Farnet, C. M., and Haseltine, W. A. (1990). Integration of human immunodeficiency virus type 1 DNA in vitro. *Proc Natl Acad Sci U S A* **87,** 4164-8.

Farnet, C. M., and Haseltine, W. A. (1991). Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. *J Virol* **65,** 1910-5.

Fassati, A., and Goff, S. P. (1999). Characterization of intracellular reverse transcription complexes of Moloney murine leukemia virus. *J Virol* **73,** 8919-25.

Fields, B. N., and Kinpe, D. M. (1996). "Virology." Raven Press, New York, N.Y.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D., and Siliciano, R. F. (1997).

Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278,** 1295-300.

Fitzgerald, M. L., and Grandgenett, D. P. (1994). Retroviral integration: in vitro host site selection by avian integrase. *J Virol* **68,** 4314-21.

Fletcher, T. M., 3rd, Soares, M. A., McPhearson, S., Hui, H., Wiskerchen, M., Muesing, M. A., Shaw, G. M., Leavitt, A. D., Boeke, J. D., and Hahn, B. H. (1997). Complementation of integrase function in HIV-1 virions. *Embo J* **16,** 5123-38.

Freed, E. O. (2004). HIV-1 and the host cell: an intimate association. *Trends Microbiol* **12,** 170-7.

Friddle, C. J., Abuin, A., Ramirez-Solis, R., Richter, L. J., Buxton, E. C., Edwards, J., Finch, R. A., Gupta, A., Hansen, G., Holt, K. H., Hu, Y., Huang, W., Jaing, C., Key, B. W., Jr., Kipp, P., Kohlhauff, B., Ma, Z. Q., Markesich, D., Newhouse, M., Perry, T., Platt, K. A., Potter, D. G., Qian, N., Shaw, J., Schrick, J., Shi, Z. Z., Sparks, M. J., Tran, D., Wann, E. R., Walke, W., Wallace, J. D., Xu, N., Zhu, Q., Person, C., Sands, A. T., and Zambrowicz, B. P. (2003). High-throughput mouse knockouts provide a functional analysis of the genome. *Cold Spring Harb Symp Quant Biol* **68,** 311-5.

Frumento, G., Corradi, A., Ferrara, G. B., and Rubartelli, A. (1997). Activation-related differences in HLA class I-bound peptides: presentation of an IL-1 receptor antagonist-derived peptide by activated, but not resting, CD4+ T lymphocytes. *J Immunol* **159,** 5993-9.

Fujiwara, T., and Mizuuchi, K. (1988). Retroviral DNA integration: structure of an integration intermediate. *Cell* **54,** 497-504.

Gao, K., Butler, S. L., and Bushman, F. (2001). Human immunodeficiency virus type 1 integrase: arrangement of protein domains in active cDNA complexes. *Embo J* **20,** 3565-76.

Gao, K., Gorelick, R. J., Johnson, D. G., and Bushman, F. (2003). Cofactors for human immunodeficiency virus type 1 cDNA integration in vitro. *J Virol* **77,** 1598-603.

Gao, K., Wong, S., and Bushman, F. (2004). Metal binding by the D,DX35E motif of human immunodeficiency virus type 1 integrase: selective rescue of Cys substitutions by Mn2+ in vitro. *J Virol* **78,** 6715-22.

Garber, M. E., and Jones, K. A. (1999). HIV-1 Tat: coping with negative elongation factors. *Curr Opin Immunol* **11,** 460-5.

Ge, H., Si, Y., and Roeder, R. G. (1998). Isolation of cDNAs encoding novel transcription coactivators p52 and p75 reveals an alternate regulatory mechanism of transcriptional activation. *Embo J* **17,** 6723-9.

Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110,** 315-25.

Goldgur, Y., Craigie, R., Cohen, G. H., Fujiwara, T., Yoshinaga, T., Fujishita, T., Sugimoto, H., Endo, T., Murai, H., and Davies, D. R. (1999). Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: a platform for antiviral drug design. *Proc Natl Acad Sci U S A* **96,** 13040-3.

Goldgur, Y., Dyda, F., Hickman, A. B., Jenkins, T. M., Craigie, R., and Davies, D. R. (1998). Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc Natl Acad Sci U S A* **95,** 9150-4.

Goodarzi, G., Chiu, R., Brackmann, K., Kohn, K., Pommier, Y., and Grandgenett, D. P. (1997). Host site selection for concerted integration by human immunodeficiency virus type-1 virions in vitro. *Virology* **231,** 210-7.

Goodarzi, G., Im, G. J., Brackmann, K., and Grandgenett, D. (1995). Concerted integration of retrovirus-like DNA by human immunodeficiency virus type 1 integrase. *J Virol* **69,** 6090-7.

Goulaouic, H., and Chow, S. A. (1996). Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. *J Virol* **70,** 37-46.

Greger, I. H., Aranda, A., and Proudfoot, N. (2000). Balancing transcriptional interference and initiation on the GAL7 promoter of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **97,** 8415-20.

Greger, I. H., Demarchi, F., Giacca, M., and Proudfoot, N. J. (1998). Transcriptional interference perturbs the binding of Sp1 to the HIV-1 promoter. *Nucleic Acids Res* **26,** 1294-301.

Gross, D. S., and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57,** 159-97.

Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J. L., Fraser, C. C., Cavazzana-Calvo, M.,

and Fischer, A. (2003a). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* **348,** 255-6.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C. S., Pawliuk, R., Morillon, E., Sorensen, R., Forster, A., Fraser, P., Cohen, J. I., de Saint Basile, G., Alexander, I., Wintergerst, U., Frebourg, T., Aurias, A., Stoppa-Lyonnet, D., Romana, S., Radford-Weiss, I., Gross, F., Valensi, F., Delabesse, E., Macintyre, E., Sigaux, F., Soulier, J., Leiva, L. E., Wissler, M., Prinz, C., Rabbitts, T. H., Le Deist, F., Fischer, A., and Cavazzana-Calvo, M. (2003b). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302,** 415-9.

Han, Y., Lassen, K., Monie, D., Sedaghat, A. R., Shimoji, S., Liu, X., Pierson, T. C., Margolick, J. B., Siliciano, R. F., and Siliciano, J. D. (2004). Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol* **78,** 6122-33.

Hansen, M. S., Carteau, S., Hoffmann, C., Li, L., and Bushman, F. (1998). Retroviral cDNA integration: mechanism, applications and inhibition. *Genet Eng (N Y)* **20,** 41-61.

Harper, A. L., Sudol, M., and Katzman, M. (2003). An amino acid in the central catalytic domain of three retroviral integrases that affects target site selection in nonviral DNA. *J Virol* **77,** 3838-45.

Harris, D., and Engelman, A. (2000). Both the structure and DNA binding function of the barrier-to-autointegration factor contribute to reconstitution of HIV type 1 integration in vitro. *J Biol Chem* **275,** 39671-7.

Hausler, B., and Somerville, R. L. (1979). Interaction in vivo between strong closely spaced constitutive promoters. *J Mol Biol* **127,** 353-6.

Hazuda, D. J., Felock, P. J., Hastings, J. C., Pramanik, B., and Wolfe, A. L. (1997). Differential divalent cation requirements uncouple the assembly and catalytic reactions of human immunodeficiency virus type 1 integrase. *J Virol* **71,** 7005-11.

Hazuda, D. J., Wolfe, A. L., Hastings, J. C., Robbins, H. L., Graham, P. L., LaFemina, R. L., and Emini, E. A. (1994). Viral long terminal repeat substrate binding characteristics of the human immunodeficiency virus type 1 integrase. *J Biol Chem* **269,** 3999-4004.

Hematti, P., Hong, B. K., Ferguson, C., Adler, R., Hanawa, H., Sellers, S., Holt, I. E., Eckfeldt, C. E., Sharma, Y., Schmidt, M., von Kalle, C., Persons, D. A., Billings, E. M., Verfaillie, C. M., Nienhuis, A. W., Wolfsberg, T. G., Dunbar, C. E., and Calmels, B. (2004). Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol* **2,** e423.

Heuer, T. S., and Brown, P. O. (1997). Mapping features of HIV-1 integrase near selected sites on viral and target DNA molecules in an active enzyme-DNA complex by photo-cross-linking. *Biochemistry* **36,** 10655-65.

Heuer, T. S., and Brown, P. O. (1998). Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry* **37,** 6667-78.

Hindmarsh, P., Ridky, T., Reeves, R., Andrake, M., Skalka, A. M., and Leis, J. (1999). HMG protein family members stimulate human immunodeficiency virus type 1 and avian sarcoma virus concerted DNA integration in vitro. *J Virol* **73,** 2994-3003.

Holman, A. G., and Coffin, J. M. (2005). Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci U S A* **102,** 6103-7.

Holmes-Son, M. L., and Chow, S. A. (2000). Integrase-lexA fusion proteins incorporated into human immunodeficiency virus type 1 that contains a catalytically inactive integrase gene are functional to mediate integration. *J Virol* **74,** 11548-56.

Horowitz, J. M., Holland, G. D., King, S. R., and Risser, R. (1987). Germ line integration of a murine leukemia provirus into a retroviruslike sequence. *J Virol* **61,** 701-7.

Hu, W. Y., Bushman, F. D., and Siva, A. C. (2004). RNA interference against retroviruses. *Virus Res* **102,** 59-64.

Hughes, S. H., Mutschler, A., Bishop, J. M., and Varmus, H. E. (1981). A Rous sarcoma virus provirus is flanked by short direct repeats of a cellular DNA sequence present in only one copy prior to integration. *Proc Natl Acad Sci U S A* **78,** 4299-303.

Hughes, S. H., Shank, P. R., Spector, D. H., Kung, H. J., Bishop, J. M., Varmus, H. E., Vogt, P. K., and Breitman, M. L. (1978). Proviruses of avian sarcoma virus are terminally redundant, co-extensive with unintegrated linear DNA and integrated at many sites. *Cell* **15,** 1397-410.

Izmailova, E., Bertley, F. M., Huang, Q., Makori, N., Miller, C. J., Young, R. A., and Aldovini, A. (2003). HIV-1 Tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages. *Nat Med* **9,** 191-7.

Jenkins, T. M., Esposito, D., Engelman, A., and Craigie, R. (1997). Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking. *Embo J* **16,** 6849-59.

Jenuwein, T. (2001). Re-SET-ting heterochromatin by histone methyltransferases. *Trends Cell Biol* **11,** 266-73.

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science* **293,** 1074-80.

Jones, K. S., Coleman, J., Merkel, G. W., Laue, T. M., and Skalka, A. M. (1992). Retroviral integrase functions as a multimer and can turn over catalytically. *J Biol Chem* **267,** 16037-40.

Jordan, A., Bisgrove, D., and Verdin, E. (2003). HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *Embo J* **22,** 1868-77.

Jordan, A., Defechereux, P., and Verdin, E. (2001). The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *Embo J* **20,** 1726-38.

Kalpana, G. V., Marmon, S., Wang, W., Crabtree, G. R., and Goff, S. P. (1994). Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science* **266,** 2002-6.

Kanazawa, S., Okamoto, T., and Peterlin, B. M. (2000). Tat competes with CIITA for the binding to P-TEFb and blocks the expression of MHC class II genes in HIV infection. *Immunity* **12,** 61-70.

Kao, S. Y., Calman, A. F., Luciw, P. A., and Peterlin, B. M. (1987). Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330,** 489-93.

Katz, R. A., Greger, J. G., Boimel, P., and Skalka, A. M. (2003). Human immunodeficiency virus type 1 DNA nuclear import and integration are mitosis independent in cycling cells. *J Virol* **77,** 13412-7.

Katz, R. A., Merkel, G., Kulkosky, J., Leis, J., and Skalka, A. M. (1990). The avian retroviral IN protein is both necessary and sufficient for integrative recombination in vitro. *Cell* **63,** 87-95.

Katz, R. A., Merkel, G., and Skalka, A. M. (1996). Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. *Virology* **217,** 178-90.

Katzman, M., Katz, R. A., Skalka, A. M., and Leis, J. (1989). The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the in vivo sites of integration. *J Virol* **63,** 5319-27.

Katzman, M., and Sudol, M. (1995). Mapping domains of retroviral integrase responsible for viral DNA specificity and target site selection by analysis of chimeras between human immunodeficiency virus type 1 and visna virus integrases. *J Virol* **69,** 5687-96.

Katzman, M., and Sudol, M. (1998). Mapping viral DNA specificity to the central region of integrase by using functional human immunodeficiency virus type 1/visna virus chimeric proteins. *J Virol* **72,** 1744-53.

Khan, E., Mack, J. P., Katz, R. A., Kulkosky, J., and Skalka, A. M. (1991). Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Res* **19,** 851-60.

Kirchner, J., Connolly, C. M., and Sandmeyer, S. B. (1995). Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science* **267,** 1488-91.

Kitamura, Y., Lee, Y. M., and Coffin, J. M. (1992). Nonrandom integration of retroviral DNA in vitro: effect of CpG methylation. *Proc Natl Acad Sci U S A* **89,** 5532-6.

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P., and Skalka, A. M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Mol Cell Biol* **12,** 2331-8.

Kulkosky, J., Katz, R. A., Merkel, G., and Skalka, A. M. (1995). Activities and substrate specificity of the evolutionarily conserved central domain of retroviral integrase. *Virology* **206,** 448-56.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13,** 1095-107.

Lassen, K., Han, Y., Zhou, Y., Siliciano, J., and Siliciano, R. F. (2004). The multifactorial nature of HIV-1 latency. *Trends Mol Med* **10,** 525-31.

Laufs, S., Gentner, B., Nagy, K. Z., Jauch, A., Benner, A., Naundorf, S., Kuehlcke, K., Schiedlmeier, B., Ho, A. D., Zeller, W. J., and Fruehauf, S. (2003). Retroviral vector integration occurs in preferred genomic targets of human bone marrow-repopulating cells. *Blood* **101,** 2191-8.

Laufs, S., Nagy, K. Z., Giordano, F. A., Hotz-Wagenblatt, A., Zeller, W. J., and Fruehauf, S. (2004). Insertion of retroviral vectors in NOD/SCID repopulating human peripheral blood progenitor cells occurs preferentially in the vicinity of transcription start regions and in introns. *Mol Ther* **10,** 874-81.

Leavitt, A. D., Robles, G., Alesandro, N., and Varmus, H. E. (1996). Human immunodeficiency virus type 1 integrase mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection. *J Virol* **70,** 721-8.

Leavitt, A. D., Rose, R. B., and Varmus, H. E. (1992). Both substrate and target oligonucleotide sequences affect in vitro integration mediated by human immunodeficiency virus type 1 integrase protein produced in Saccharomyces cerevisiae. *J Virol* **66,** 2359-68.

Lee, M. S., and Craigie, R. (1994). Protection of retroviral DNA from autointegration: involvement of a cellular factor. *Proc Natl Acad Sci U S A* **91,** 9823-7.

Lee, S. P., and Han, M. K. (1996). Zinc stimulates Mg2+-dependent 3'-processing activity of human immunodeficiency virus type 1 integrase in vitro. *Biochemistry* **35,** 3837-44.

Lee, S. P., Kim, H. G., Censullo, M. L., and Han, M. K. (1995). Characterization of Mg(2+)-dependent 3'-processing activity for human immunodeficiency virus type 1 integrase in vitro: real-time kinetic studies using fluorescence resonance energy transfer. *Biochemistry* **34,** 10205-14.

Lee, S. P., Xiao, J., Knutson, J. R., Lewis, M. S., and Han, M. K. (1997). Zn2+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro. *Biochemistry* **36,** 173-80.

Lee, Y. M., and Coffin, J. M. (1990). Efficient autointegration of avian retrovirus DNA in vitro. *J Virol* **64,** 5958-65.

Lee, Y. M., and Coffin, J. M. (1991). Relationship of avian retrovirus DNA synthesis to integration in vitro. *Mol Cell Biol* **11,** 1419-30.

Leh, H., Brodin, P., Bischerour, J., Deprez, E., Tauc, P., Brochon, J. C., LeCam, E., Coulaud, D., Auclair, C., and Mouscadet, J. F. (2000). Determinants of Mg2+-

dependent activities of recombinant human immunodeficiency virus type 1 integrase. *Biochemistry* **39,** 9285-94.

Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2005). Genome-wide analysis of chromosomal features repressing HIV transcription. *J Virol* **79,** 6610-9.

Lewis, P., Hensel, M., and Emerman, M. (1992). Human immunodeficiency virus infection of cells arrested in the cell cycle. *Embo J* **11,** 3053-8.

Lewis, P. F., and Emerman, M. (1994). Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus. *J Virol* **68,** 510-6.

Li, J., Shen, H., Himmel, K. L., Dupuy, A. J., Largaespada, D. A., Nakamura, T., Shaughnessy, J. D., Jr., Jenkins, N. A., and Copeland, N. G. (1999). Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat Genet* **23,** 348-53.

Li, L., Farnet, C. M., Anderson, W. F., and Bushman, F. D. (1998). Modulation of activity of Moloney murine leukemia virus preintegration complexes by host factors in vitro. *J Virol* **72,** 2125-31.

Li, L., Olvera, J. M., Yoder, K. E., Mitchell, R. S., Butler, S. L., Lieber, M., Martin, S. L., and Bushman, F. D. (2001). Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *Embo J* **20,** 3272-81.

Li, M., and Craigie, R. (2005). Processing of Viral DNA Ends Channels the HIV-1 Integration Reaction to Concerted Integration. *J Biol Chem* **280,** 29334-9.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* **2,** 18-22.

Lins, R. D., Briggs, J. M., Straatsma, T. P., Carlson, H. A., Greenwald, J., Choe, S., and McCammon, J. A. (1999). Molecular dynamics studies on the HIV-1 integrase catalytic domain. *Biophys J* **76,** 2999-3011.

Lins, R. D., Straatsma, T. P., and Briggs, J. M. (2000). Similarities in the HIV-1 and ASV integrase active sites upon metal cofactor binding. *Biopolymers* **53,** 308-15.

Llano, M., Delgado, S., Vanegas, M., and Poeschla, E. M. (2004a). Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. *J Biol Chem* **279,** 55570-7.

Llano, M., Vanegas, M., Fregoso, O., Saenz, D., Chung, S., Peretz, M., and Poeschla, E. M. (2004b). LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J Virol* **78,** 9524-37.

Lodi, P. J., Ernst, J. A., Kuszewski, J., Hickman, A. B., Engelman, A., Craigie, R., Clore, G. M., and Gronenborn, A. M. (1995). Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry* **34,** 9826-33.

Lubkowski, J., Dauter, Z., Yang, F., Alexandratos, J., Merkel, G., Skalka, A. M., and Wlodawer, A. (1999). Atomic resolution structures of the core domain of avian sarcoma virus integrase and its D64N mutant. *Biochemistry* **38,** 13512-22.

Lukacsovich, T., and Yamamoto, D. (2001). Trap a gene and find out its function: toward functional genomics in Drosophila. *J Neurogenet* **15,** 147-68.

Lutzke, R. A., and Plasterk, R. H. (1998). Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding. *J Virol* **72,** 4841-8.

Lutzke, R. A., Vink, C., and Plasterk, R. H. (1994). Characterization of the minimal DNA-binding domain of the HIV integrase protein. *Nucleic Acids Res* **22,** 4125-31.

Maertens, G., Cherepanov, P., Pluymers, W., Busschots, K., De Clercq, E., Debyser, Z., and Engelborghs, Y. (2003). LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J Biol Chem* **278,** 33528-39.

Maignan, S., Guilloteau, J. P., Zhou-Liu, Q., Clement-Mella, C., and Mikol, V. (1998). Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. *J Mol Biol* **282,** 359-68.

Manger, B., Weiss, A., Hardy, K. J., and Stobo, J. D. (1986). A transferrin receptor antibody represents one signal for the induction of IL 2 production by a human T cell line. *J Immunol* **136,** 532-8.

Mannioui, A., Schiffer, C., Felix, N., Nelson, E., Brussel, A., Sonigo, P., Gluckman, J. C., and Canque, B. (2004). Cell cycle regulation of human immunodeficiency virus type 1 integration in T cells: antagonistic effects of nuclear envelope breakdown and chromatin condensation. *Virology* **329,** 77-88.

Martens, J. A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature* **429,** 571-4.

Maxfield, L. F., Fraize, C. D., and Coffin, J. M. (2005). Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc Natl Acad Sci U S A*. **102,** 1436-41.

McCullagh, P., and Nelder, J. A. (1999). "Generalized linear models." Chapman & Hall Ltd.

McDonald, D., Vodicka, M. A., Lucero, G., Svitkina, T. M., Borisy, G. G., Emerman, M., and Hope, T. J. (2002). Visualization of the intracellular behavior of HIV in living cells. *J Cell Biol* **159,** 441-52.

Miller, A. D., Law, M. F., and Verma, I. M. (1985). Generation of helper-free amphotropic retroviruses that transduce a dominant-acting, methotrexate-resistant dihydrofolate reductase gene. *Mol Cell Biol* **5,** 431-7.

Miller, M. D., Farnet, C. M., and Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J Virol* **71,** 5382-90.

Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2,** E234.

Moreau, K., Faure, C., Violot, S., Verdier, G., and Ronfort, C. (2003). Mutations in the C-terminal domain of ALSV (Avian Leukemia and Sarcoma Viruses) integrase alter the concerted DNA integration process in vitro. *Eur J Biochem* **270,** 4426-38.

Morris, K. V., Chan, S. W., Jacobsen, S. E., and Looney, D. J. (2004). Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305,** 1289-92.

Muesing, M. A., Smith, D. H., Cabradilla, C. D., Benton, C. V., Lasky, L. A., and Capon, D. J. (1985). Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus. *Nature* **313,** 450-8.

Muller, H. P., and Varmus, H. E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *Embo J* **13,** 4704-14.

Mumm, S. R., and Grandgenett, D. P. (1991). Defining nucleic acid-binding properties of avian retrovirus integrase by deletion analysis. *J Virol* **65,** 1160-7.

Naldini, L., Blomer, U., Gallay, P., Ory, D., Mulligan, R., Gage, F. H., Verma, I. M., and Trono, D. (1996). In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272,** 263-7.

Narezkina, A., Taganov, K. D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A. M., and Katz, R. A. (2004). Genome-wide analyses of avian sarcoma virus integration sites. *J Virol* **78,** 11656-63.

Nermut, M. V., and Fassati, A. (2003). Structural analyses of purified human immunodeficiency virus type 1 intracellular reverse transcription complexes. *J Virol* **77,** 8196-206.

O'Doherty, U., Swiggard, W. J., and Malim, M. H. (2000). Human immunodeficiency virus type 1 spinoculation enhances infection through virus binding. *J Virol* **74,** 10074-80.

Pahl, A., and Flugel, R. M. (1995). Characterization of the human spuma retrovirus integrase by site-directed mutagenesis, by complementation analysis, and by swapping the zinc finger domain of HIV-1. *J Biol Chem* **270,** 2957-66.

Panet, A., and Cedar, H. (1977). Selective degradation of integrated murine leukemia proviral DNA by deoxyribonucleases. *Cell* **11,** 933-40.

Panganiban, A. T., and Temin, H. M. (1983). The terminal nucleotides of retrovirus DNA are required for integration but not virus production. *Nature* **306,** 155-60.

Panganiban, A. T., and Temin, H. M. (1984). The retrovirus pol gene encodes a product required for DNA integration: identification of a retrovirus int locus. *Proc Natl Acad Sci U S A* **81,** 7885-9.

Patel, P. H., and Preston, B. D. (1994). Marked infidelity of human immunodeficiency virus type 1 reverse transcriptase at RNA and DNA template ends. *Proc Natl Acad Sci U S A* **91,** 549-53.

Peden, K., Emerman, M., and Montagnier, L. (1991). Changes in growth properties on passage in tissue culture of viruses derived from infectious molecular clones of HIV-1LAI, HIV-1MAL, and HIV-1ELI. *Virology* **185,** 661-72.

Pierson, T. C., Zhou, Y., Kieffer, T. L., Ruff, C. T., Buck, C., and Siliciano, R. F. (2002). Molecular characterization of preintegration latency in human immunodeficiency virus type 1 infection. *J Virol* **76,** 8518-31.

Plasterk, R. H. (2002). RNA silencing: the genome's immune system. *Science* **296,** 1263-5.

Podtelezhnikov, A. A., Gao, K., Bushman, F. D., and McCammon, J. A. (2003). Modeling HIV-1 integrase complexes based on their hydrodynamic properties. *Biopolymers* **68,** 110-20.

Pruss, D., Bushman, F. D., and Wolffe, A. P. (1994a). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci U S A* **91,** 5913-7.

Pruss, D., Reeves, R., Bushman, F. D., and Wolffe, A. P. (1994b). The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem* **269,** 25031-41.

Pryciak, P. M., Sil, A., and Varmus, H. E. (1992). Retroviral integration into minichromosomes in vitro. *Embo J* **11,** 291-303.

Pryciak, P. M., and Varmus, H. E. (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69,** 769-80.

Quinn, T. P., and Grandgenett, D. P. (1988). Genetic evidence that the avian retrovirus DNA endonuclease domain of pol is necessary for viral integration. *J Virol* **62,** 2307-12.

R Development Core Team. (2005). *In* "R Foundation for Statistical Computing", Vienna, Austria.

Rice, P. A., and Baker, T. A. (2001). Comparative architecture of transposase and integrase complexes. *Nat Struct Biol* **8,** 302-7.

Roe, T., Reynolds, T. C., Yu, G., and Brown, P. O. (1993). Integration of murine leukemia virus DNA depends on mitosis. *Embo J* **12,** 2099-108.

Rogel, M. E., Wu, L. I., and Emerman, M. (1995). The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *J Virol* **69,** 882-8.

Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R., and Breindl, M. (1987). Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. *J Virol* **61,** 336-43.

Roth, M. J., Schwartzberg, P. L., and Goff, S. P. (1989). Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell* **58,** 47-54.

Rowland, S. J., and Dyke, K. G. (1990). Tn552, a novel transposable element from Staphylococcus aureus. *Mol Microbiol* **4,** 961-75.

Sabo, P. J., Humbert, R., Hawrylycz, M., Wallace, J. C., Dorschner, M. O., McArthur, M., and Stamatoyannopoulos, J. A. (2004). Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A* **101,** 4537-42.

Sandmeyer, S. (2003). Integration by design. *Proc Natl Acad Sci U S A* **100,** 5586-8.

Schauer, M., and Billich, A. (1992). The N-terminal region of HIV-1 integrase is required for integration activity, but not for DNA-binding. *Biochem Biophys Res Commun* **185,** 874-80.

Scherer, L. J., and Rossi, J. J. (2003). Approaches for the sequence-specific knockdown of mRNA. *Nat Biotechnol* **21,** 1457-65.

Schmid, R. M., Perkins, N. D., Duckett, C. S., Andrews, P. C., and Nabel, G. J. (1991). Cloning of an NF-kappa B subunit which stimulates HIV transcription in synergy with p65. *Nature* **352,** 733-6.

Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110,** 521-9.

Schwartzberg, P., Colicelli, J., and Goff, S. P. (1984). Construction and analysis of deletion mutations in the pol gene of Moloney murine leukemia virus: a new viral function required for productive infection. *Cell* **37,** 1043-52.

Scottoline, B. P., Chow, S., Ellison, V., and Brown, P. O. (1997). Disruption of the terminal base pairs of retroviral DNA during integration. *Genes Dev* **11,** 371-82.

Setterfield, G., Hall, R., Bladon, T., Little, J., and Kaplan, J. G. (1983). Changes in structure and composition of lymphocyte nuclei during mitogenic stimulation. *J Ultrastruct Res* **82,** 264-82.

She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., Clark, R., Graves, T., Gulden, C. L., Alkan, C., Bailey, J. A., Sahinalp, C., Rocchi, M., Haussler, D., Wilson, R. K., Miller, W., Schwartz, S., and Eichler, E. E. (2004). The structure and evolution of centromeric transition regions within the human genome. *Nature* **430,** 857-64.

Sheridan, P. L., Mayall, T. P., Verdin, E., and Jones, K. A. (1997). Histone acetyltransferases regulate HIV-1 enhancer activity in vitro. *Genes Dev* **11,** 3327-40.

Sherman, P. A., and Fyfe, J. A. (1990). Human immunodeficiency virus integration protein expressed in Escherichia coli possesses selective DNA cleaving activity. *Proc Natl Acad Sci U S A* **87,** 5119-23.

Shibagaki, Y., and Chow, S. A. (1997). Central core domain of retroviral integrase is responsible for target site selection. *J Biol Chem* **272,** 8361-9.

Shoemaker, C., Goff, S., Gilboa, E., Paskind, M., Mitra, S. W., and Baltimore, D. (1980). Structure of a cloned circular Moloney murine leukemia virus DNA molecule containing an inverted segment: implications for retrovirus integration. *Proc Natl Acad Sci U S A* **77,** 3932-6.

Shoemaker, C., Hoffman, J., Goff, S. P., and Baltimore, D. (1981). Intramolecular integration within Moloney murine leukemia virus DNA. *J Virol* **40,** 164-72.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15,** 1034-50.

Sinha, S., Pursley, M. H., and Grandgenett, D. P. (2002). Efficient concerted integration by recombinant human immunodeficiency virus type 1 integrase without cellular or viral cofactors. *J Virol* **76,** 3105-13.

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

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101,** 6062-7.

Suzuki, T., Shen, H., Akagi, K., Morse, H. C., Malley, J. D., Naiman, D. Q., Jenkins, N. A., and Copeland, N. G. (2002). New genes involved in cancer identified by retroviral tagging. *Nat Genet* **32,** 166-74.

Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G., and Boeke, J. D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110,** 327-38.

Tan, W., Zhu, K., Segal, D. J., Barbas, C. F., 3rd, and Chow, S. A. (2004). Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. *J Virol* **78,** 1301-13.

van Gent, D. C., Oude Groeneger, A. A., and Plasterk, R. H. (1993a). Identification of amino acids in HIV-2 integrase involved in site-specific hydrolysis and alcoholysis of viral DNA termini. *Nucleic Acids Res* **21,** 3373-7.

van Gent, D. C., Vink, C., Groeneger, A. A., and Plasterk, R. H. (1993b). Complementation between HIV integrase proteins mutated in different domains. *Embo J* **12,** 3261-7.

Verdin, E. (1991). DNase I-hypersensitive sites are associated with both long terminal repeats and with the intragenic enhancer of integrated human immunodeficiency virus type 1. *J Virol* **65,** 6790-9.

Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., and van Kampen, A. H. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13,** 1998-2004.

Vijaya, S., Steffen, D. L., and Robinson, H. L. (1986). Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J Virol* **60,** 683-92.

Vincent, K. A., Ellison, V., Chow, S. A., and Brown, P. O. (1993). Characterization of human immunodeficiency virus type 1 integrase expressed in Escherichia coli and analysis of variants with amino-terminal mutations. *J Virol* **67,** 425-37.

Vincent, K. A., York-Higgins, D., Quiroga, M., and Brown, P. O. (1990). Host sequences flanking the HIV provirus. *Nucleic Acids Res* **18,** 6045-7.

Vink, C., Groenink, M., Elgersma, Y., Fouchier, R. A., Tersmette, M., and Plasterk, R. H. (1990). Analysis of the junctions between human immunodeficiency virus type 1 proviral DNA and human DNA. *J Virol* **64,** 5626-7.

Vink, C., Lutzke, R. A., and Plasterk, R. H. (1994). Formation of a stable complex between the human immunodeficiency virus integrase protein and viral DNA. *Nucleic Acids Res* **22,** 4103-10.

Vink, C., Oude Groeneger, A. M., and Plasterk, R. H. (1993). Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein. *Nucleic Acids Res* **21,** 1419-25.

Wallrath, L. L. (1998). Unfolding the mysteries of heterochromatin. *Curr Opin Genet Dev* **8,** 147-53.

Wang, J. Y., Ling, H., Yang, W., and Craigie, R. (2001). Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *Embo J* **20,** 7333-43.

Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H., and Jones, K. A. (1998a). A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92,** 451-62.

Wei, S. Q., Mizuuchi, K., and Craigie, R. (1997). A large nucleoprotein assembly at the ends of the viral DNA mediates retroviral DNA integration. *Embo J* **16,** 7511-20.

Wei, S. Q., Mizuuchi, K., and Craigie, R. (1998b). Footprints on the viral DNA ends in moloney murine leukemia virus preintegration complexes reflect a specific association with integrase. *Proc Natl Acad Sci U S A* **95,** 10535-40.

Weidhaas, J. B., Angelichio, E. L., Fenner, S., and Coffin, J. M. (2000). Relationship between retroviral DNA integration and gene expression. *J Virol* **74,** 8382-9.

Weinberg, J. B., Matthews, T. J., Cullen, B. R., and Malim, M. H. (1991). Productive human immunodeficiency virus type 1 (HIV-1) infection of nonproliferating human monocytes. *J Exp Med* **174,** 1477-82.

Woerner, A. M., and Marcus-Sekura, C. J. (1993). Characterization of a DNA binding domain in the C-terminus of HIV-1 integrase by deletion mutagenesis. *Nucleic Acids Res* **21,** 3507-11.

Wolffe, A. P. (1998). "Chromatin." Academic Press, San Diego, CA.

Wong, J. K., Hezareh, M., Gunthard, H. F., Havlir, D. V., Ignacio, C. C., Spina, C. A., and Richman, D. D. (1997). Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278,** 1291-5.

Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300,** 1749-51.

Wu, X., Li, Y., Crise, B., Burgess, S. M., and Munroe, D. J. (2005). Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol* **79,** 5211-4.

Yamashita, M., and Emerman, M. (2004). Capsid is a dominant determinant of retrovirus infectivity in nondividing cells. *J Virol* **78,** 5670-8.

Yang, F., Leon, O., Greenfield, N. J., and Roth, M. J. (1999). Functional interactions of the HHCC domain of moloney murine leukemia virus integrase revealed by nonoverlapping complementation and zinc-dependent dimerization. *J Virol* **73,** 1809-17.

Yang, W., and Steitz, T. A. (1995). Recombining the structures of HIV integrase, RuvC and RNase H. *Structure* **3,** 131-4.

Yang, Z. N., Mueser, T. C., Bushman, F. D., and Hyde, C. C. (2000). Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. *J Mol Biol* **296,** 535-48.

Yi, J., Asante-Appiah, E., and Skalka, A. M. (1999). Divalent cations stimulate preferential recognition of a viral DNA end by HIV-1 integrase. *Biochemistry* **38,** 8458-68.

Yoder, K. E., and Bushman, F. D. (2000). Repair of gaps in retroviral DNA integration intermediates. *J Virol* **74,** 11191-200.

Yung, E., Sorin, M., Pal, A., Craig, E., Morozov, A., Delattre, O., Kappes, J., Ott, D., and Kalpana, G. V. (2001). Inhibition of HIV-1 virion production by a transdominant mutant of integrase interactor 1. *Nat Med* **7,** 920-6.

Zheng, R., Jenkins, T. M., and Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proc Natl Acad Sci U S A* **93,** 13659-64.

Zhu, Y., Dai, J., Fuerst, P. G., and Voytas, D. F. (2003). Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A* **100,** 5891-5.

Zhu, Y., Zou, S., Wright, D. A., and Voytas, D. F. (1999). Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p. *Genes Dev* **13,** 2738-49.