Correspondence

Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence

Daniel Elleder[a], Adam Pavlíček[a], Jan Pačes[a,b], Jiří Hejnar[a,*]

First published online 5 March 2002

After entering a cell during infection, the human immunodeficiency virus type 1 (HIV-1) undergoes a series of steps including reverse transcription of its genome and culminating in integration of proviral DNA into the host chromosomes. The further fate of the individual provirus is to a great extent influenced by the efficiency of provirus transcription, dependent upon the site of its integration [1]. How HIV-1 and other retroviruses choose their integration sites is still far from completely understood. It seems that the integration is not strictly specific, because most or all genomic regions are potential targets, but neither is it a random event. Locally, there are up to several hundred-fold differences in the usage of target sites due to the local DNA structure, bending, distortion and wrapping around nucleosomes (reviewed in [2]).

The non-randomness on the scale of genomic regions has been much less addressed [3]. Methods employed to study in vivo retrovirus integration sites include restriction enzyme digestions and blotting, fluorescence in situ hybridization, PCR-based assays, and most importantly cloning and sequencing the virus–host integration junctions. Most studies analyze only a small number of integration sites, or focus on selected genomic regions. To date, the most representative study is provided by Carteau et al. [4]. It lists a set of 61 HIV-1 integration site sequences obtained after short experimental infection of the human T-cell line SupT1. Of these, 59 sequences are available in GenBank, together with 104 control genomic sequences for comparison. The authors analyzed the sequences using the nr, dbEST and MONTH databases as of November 1997. They concluded that there is no significant difference between integration sites and controls, except that centromeric alphoid repeats are selectively absent at integration sites. The availability of the human genome sequence [5] creates a great opportunity for a new genome-wide analysis of these data. By mapping the exact positions of the integration sites we can analyze large DNA regions flanking the proviruses and describe the genomic features present.

We used the BLAT program to map the genomic positions of integration sites in the most recent GoldenPath assembly of 6 August 2001 (http://genome.ucsc.edu). Of 59 sequences available in GenBank, we succeeded in mapping 48, where the level of homology was satisfactory (Table 1 in supplementary material on the web; http://www.elsevier.com/PII/S0014579302026121). For each mapped integration we collected several genomic features available in the GoldenPath assembly. The first was the presence of transcribed sequences, either as the 'known protein coding genes' category (from the RefSeq project) or as the 'human mRNAs from GenBank' category. In addition, 800 cytogenetic band resolution is available, light or dark according to Giemsa staining. Next we calculated the GC level of 100 kb regions surrounding the integration sites symmetrically and the gene densities along these flanking regions. We compared these data with the whole genome summary statistics that we calculated for the GoldenPath assembly and looked for any differences indicating possible integration preference.

In our analysis, 54.2% (26 of 48) of the mapped integration sites fall in genes, which is significantly higher compared to the genome average calculated as 22.2% ($P < 0.00001$, χ^2 test). For the broader category of mRNAs this comparison is 68.7% to 30.7% ($P < 0.00001$, χ^2 test). This implies that potentially transcribed regions represent strongly preferred targets for HIV-1 integration. Out of the 33 integrations in transcription units, 18 and 15 are in sense and antisense orientation, respectively, 28 map to introns, two to exons, two to 5′ untranslated regions (UTR) and one to a 3′ UTR. Giemsa light (R) and dark (G) bands were targeted in 68.7% and 31.2%, compared to genome averages of 44.0% and 48.5%, respectively, estimated from the GoldenPath assembly ($P < 0.003$, χ^2 test). The average GC content of 100 kb regions flanking the integration sites was calculated to be 44.4%, higher than the whole genome average of 41.0%. The distribution of integrations is clearly biased, with more hits belonging to the GC-richer genomic regions (Fig. 1).

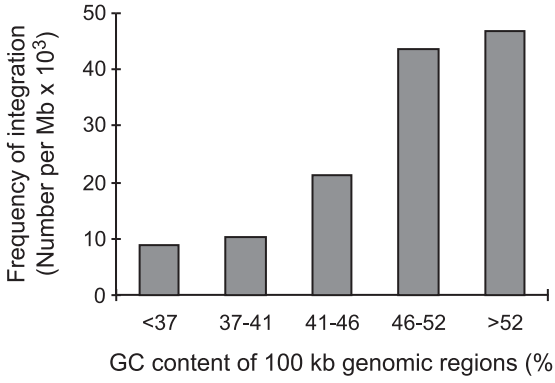

Fig. 1. Distribution of 48 HIV-1 integration sites in the human genome with respect to the GC content. On the horizontal axis, there are 100 kb genomic regions arranged by their GC content, the GC% intervals are consistent with the proposed division of the genome into five categories (isochores) [6]: L1 (< 37%), L2 (37–41%), H1 (41–46%), H2 (46–52%) and H3 (> 52%). The vertical axis shows the probability of targeting the individual classes with a specified GC content. This is calculated as the number of integrations whose 100 kb surrounding belongs to the corresponding GC% interval divided by the extent occupied by the 100 kb genomic regions also belonging to the same GC% interval. The size of the genome belonging to individual GC% categories in 100 kb windows is 571 Mb for L1, 977 Mb for L2, 706 Mb for H1, 321 Mb for H2 and 85 Mb for H3. The graph clearly shows an increasing tendency, with higher probability that regions with higher GC content will be targeted.

PII: S0014-5793(02)02612-1

To characterize the gene density, we calculated the average number of entries of the 'known genes' category (5′ ends) in 1 Mb regions surrounding the integration sites as 12.0. This is several times higher than the genome average of 4.4 (estimated as 14 200 entries in the 'known genes' category per 2.85 Gb sequenced genome size). The means are significantly different according to the Mann–Whitney *U*-test ($P < 0.0001$). We obtained similar results for 100 kb regions and for the mRNA category (data not shown).

Thus, it appears that HIV-1 preferably integrates into genes, genome regions with increased gene density, cytogenetic light bands, and GC-rich regions. These features obviously are interdependent, genes being more frequent in GC-rich regions and in Giemsa light bands, and light bands having a higher average GC content than dark bands [6]. It is possible that the preference for gene integrations causes the bias observed for chromosomal bands and GC distribution. The specifications of genes targeted are available in the supplementary material. It would be interesting to quantify their expression in the SupT1 cell line used. The observed preferences support the model of favored integration into regions with an open chromatin structure.

The share of gene integrations may further increase, because we consider only genes identified with high confidence in the GoldenPath assembly, not genes predicted by computer programs. On the other hand, we could not identify genomic positions of 11 integration site sequences, not even in the public Celera genome or in the draft HTGS sequences. These presumably belong to the unsequenced or heterochromatin portion of the genome and could weaken the statistical significance of gene targeting.

This is, to our knowledge, the first study that maps the exact genomic positions of HIV-1 integrations. Previous approaches relied just on the short sequences obtained by PCR or cloning and could not analyze larger flanking regions and more distant genomic features. The reason why Carteau et al. found only 18% integrations into transcription units was that most introns targeted were not in the database at that time.

The features previously reported in HIV-1 integration sites were described in a small number of cases, for example, the presence of topoisomerase II cleavage sites was described based on a single integration [7]. Broader analyses did not reveal any significant preferences, except for the increased number of Alu and L1 repeats, but this was not confirmed later [4]. Even the most recent studies do not report any significant preferences of HIV-1 integration [8]. Also, most studies analyzed integrations in long-term cell lines or in patient material. In such cases, the repeated cell divisions can select for some infected clones with growth advantage or greater provirus stability and the initial integration site distribution can be disturbed. Carteau et al. harvested cells just 14 h post-infection, thus avoiding this effect. Most studies publish only the sequence analysis, but Carteau et al. deposited all published sequences in GenBank. All this prompted us to limit our analysis to this homogeneous and most representative set of data.

As to other retrovirus species, Moloney murine leukemia virus was reported to integrate into transcriptionally active genome regions in five of nine cases analyzed [9]. For avian sarcoma leukosis virus RAV-1, the transcriptional activity of one locus was found to be associated with a decrease in integration frequency [10]. Surveys of HTLV-I integrations did not characterize any specific preferred targets [11].

The preference of HIV-1 for transcriptionally active regions with open chromatin conformation could be advantageous, allowing higher transcription of the provirus and thus more efficient continuation of its replication cycle. However, disruption of some genes may be harmful to the host cell and this may select against such integration events during subsequent cell divisions. The mechanisms of the proposed integration preference may involve just a greater accessibility of open chromatin regions to the viral preintegration complex, or its specific nuclear localization. The integrase protein of HIV-1 also interacts with Ini1/hSNF5, a component of the chromatin remodeling complex [12]. This interaction could actively target the integrations to a subset of genomic locations favorable for transcription.

Our approach demonstrates the advantage of using the human genome sequence for analyzing the pattern of retroviral integration. Of further interest should be the analysis of more integration events, including other retroviral genera, other cell types, and, particularly, quiescent HIV-1-infected cells.

Acknowledgements: This study was supported by Grants 204/01/0632 and 524/01/0866 from the Grant Agency of the Czech Republic. Many thanks are due to Jan Svoboda for his encouragement and critical reading of the manuscript.

References

[1] Jordan, A., Defechereux, P. and Verdin, E. (2001) EMBO J. 20, 1726–1738.
[2] Holmes-Son, M.L., Appa, R.S. and Chow, S.A. (2001) Adv. Genet. 43, 33–69.
[3] Withers-Ward, E.S., Kitamura, Y., Barnes, J.P. and Coffin, J.M. (1994) Genes Dev. 8, 1473–1487.
[4] Carteau, S., Hoffmann, C. and Bushman, F. (1998) J. Virol. 72, 4005–4014.
[5] IHGSC (International Human Genome Sequencing Consortium) (2001) Nature 409, 860–921
[6] Saccone, S., DeSario, A., Wiegant, J., Raap, A.K., Della Valle, G. and Bernardi, G. (1993) Proc. Natl. Acad. Sci. USA 90, 11929–11933.
[7] Howard, M.T. and Griffith, J.D. (1993) J. Mol. Biol. 232, 1060–1080.
[8] Lyn, D., Bennett, N.A., Shiramizu, B.T., Herndier, B.G. and Igietseme, J.U. (2001) Cell. Mol. Biol. 47, 981–986.
[9] Scherdin, U., Rhodes, K. and Breindl, M. (1990) J. Virol. 64, 907–912.
[10] Weidhaas, J.B., Angelichio, E.L., Fenner, S. and Coffin, J.M. (2000) J. Virol. 74, 8382–8389.
[11] Leclercq, I., Mortreux, F., Cavrois, M., Leroy, A., Gessain, A., Wain-Hobson, S. and Wattel, E. (2000) J. Virol. 74, 2305–2312.
[12] Kalpana, G.V., Marmon, S., Wang, W., Crabtree, G.R. and Goff, S.P. (1994) Science 266, 2002–2006.

*Corresponding author. Fax: (42)-2-24310955.
E-mail address: hejnar@img.cas.cz (J. Hejnar)

[a] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, CZ-16637 Prague, Czech Republic*
[b] *Center for Integrated Genomics, Flemingovo 2, CZ-16637 Prague, Czech Republic*