Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D1041–D1046***
doi:10.1093/nar/gkr1142

Dr.VIS: a database of human disease-related viral integration sites

Xin Zhao[1], Qi Liu[1], Qingqing Cai[1], Yanyun Li[2], Congjian Xu[2], Yixue Li[3], Zuofeng Li[3,*] and Xiaoyan Zhang[1,*]

[1]School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai, [2]Obstetrics and Gynecology Hospital of Fudan University, Shanghai, 200011 and [3]Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai, China

Received October 24, 2011; Revised and Accepted November 9, 2011

ABSTRACT

Viral integration plays an important role in the development of malignant diseases. Viruses differ in preferred integration site and flanking sequence. Viral integration sites (VIS) have been found next to oncogenes and common fragile sites. Understanding the typical DNA features near VIS is useful for the identification of potential oncogenes, prediction of malignant disease development and assessing the probability of malignant transformation in gene therapy. Therefore, we have built a database of human disease-related VIS (Dr.VIS, http://www.scbit.org/dbmi/drvis) to collect and maintain human disease-related VIS data, including characteristics of the malignant disease, chromosome region, genomic position and viral–host junction sequence. The current build of Dr.VIS covers about 600 natural VIS of 5 oncogenic viruses representing 11 diseases. Among them, about 200 VIS have viral–host junction sequence.

INTRODUCTION

The contribution of infectious agents to the development of serious human diseases, especially tumors, is increasingly understood (1). It is estimated that viral infections contribute to 15–20% of all human cancers (2). Research has revealed that integration of viral genomes into human chromosomes is necessary for most viral induction of tumor development, which can activate or inactivate host genes by means of provirus insertion (2,3). This holds not only for retroviruses such as human T-cell leukemia virus (4), but also for a number of non-retroviruses such as human papillomavirus (5) and hepatitis B virus (2,6). Finally, integration events can cause rearrangements of viral and host sequences (7), expression of fused transcripts, deletions of chromosomal sequences and transpositions of viral sequences from one chromosome to another (8–10). Viral integration is site-specific in many cases (11). Moreover, viruses differ in their preferred insertion site (12). Viral integration sites (VIS) have become a key to associating viral infection and human malignant disease.

Up to date, at least seven viruses have been compellingly associated with human malignant diseases, including:

(1) HTLV-1 (adult T-cell leukemia and tropical spastic paraparesis) (13);
(2) HPV (cervical cancer, head and neck cancer and anogenital cancer) (14,15);
(3) HHV-8 (Kaposi's sarcoma) (16);
(4) EBV (Burkitt's lymphoma) (17);
(5) HBV (hepatocellular carcinoma) (18);
(6) MCV, Merkel cell polyomavirus (Merkel cell carcinoma) (19); and
(7) HIV (AIDS and B-cell lymphoma) (1).

There are many viruses that are potentially associated with human malignant diseases such as Simian virus 40 (brain cancer, bone cancer and mesothelioma), BK virus (prostate cancer) and so on (1–3). Some are still under study, such as xenotropic murine leukemia virus-related virus whose relationship with prostate cancer is still controversial (20–22). Most of those viruses have a significant integration step in viral infection and disease development.

Viral integration can activate gene expression to cause malignant disease if the VIS is close to an oncogene. This process known as insertional mutagenesis (23), has allowed identification of potential cellular oncogenes through mapping of retroviral integration sites (23,24). This work has also led to the development of a database of cancer-associated genes (23,25).

*To whom correspondence should be addressed. Tel: +86 021 65980233; Fax: +86 021 65981041; Email: xyzhang@tongji.edu.cn
Correspondence may also be addressed to Zuofeng Li. Tel: +86 021 54065268; Fax: +86 021 54065057; Email: lizuofeng@gmail.com

Gene therapy holds promise for curing many malignant diseases. However, current gene therapy methods have limited control over where a therapeutic virus inserts into the human genome. It was reported that several patients developed T-cell leukemia during treatment of X-linked severe combined immunodeficiency (SCID-X1), because of viral integration near the proto-oncogenes LMO2, BMI1 and CCND2 (23,26).

Therefore, understanding the genes and DNA features near disease-related VIS will abet the identification of potential oncogenes, prediction of malignant disease development and assessment of the probability of malignant transformation in gene therapy. However, numerous identified VIS are still widely scattered in published papers. In this study, we developed a database of human disease-related VIS (Dr.VIS) to collect and maintain those data from the literature (PubMed) and public databases (GenBank) (27). Furthermore, each VIS is linked to the UCSC Genome Browser (28) and Ensembl Genome Browser (29) for more detailed viewing of genomic traits.

MATERIALS AND METHODS

Data model of VIS and clusters

The following characteristics are listed for each human disease-related VIS: virus name, chromosome region, locus, genomic position, viral–host junction sequence and corresponding human disease. The chromosome region is denoted as cytogenetic band. The locus must have been approved by HGNC (30) and can be a microRNA or an interrupted gene with specific coordinates of subcomponents (exons or introns). Genomic position is the position of the insertion point in the genome as represented in the Human Genome Assembly 2009 (hg19) (31). Viral–host junction sequence is always recorded as 'human genome–viral genome–human genome'.

In Dr.VIS, VIS representing the same virus name, chromosome region and human disease, are clustered to generate a unique data entry called a viral integration cluster (or VIS cluster) for convenient data organization. Genomic traits of a VIS cluster include common fragile site (32), microRNA, gene distribution and son on. More detailed traits are crosslinked to HGNC (30), UCSC (33) and Ensembl (29), through their chromosome coordinates. Furthermore, each VIS cluster is assigned a confidence code (Table 1) to indicate its frequency.

Table 1. Confidence codes

Code	Description	Integration sites count
WK	Well known	$f \geq 5$
SS	Strongly supported	$1 < f < 5$
SO	Single observation	$f = 1$

Collection of VIS associated with human diseases

VIS related to human disease were collected from PubMed and GenBank (Figure 1). All VIS deposited in Dr.VIS are sequenced or detected from natural samples of patients. A Perl script extracted viral–host junction sequences from GenBank by matching keywords (i.e. integration site) and annotation of both host and virus (i.e. Homo sapiens and a virus) as regular expressions. The script extracted PMIDs from the original literature reporting junction sequences, for subsequent manual retrieval and processing curation from PubMed.

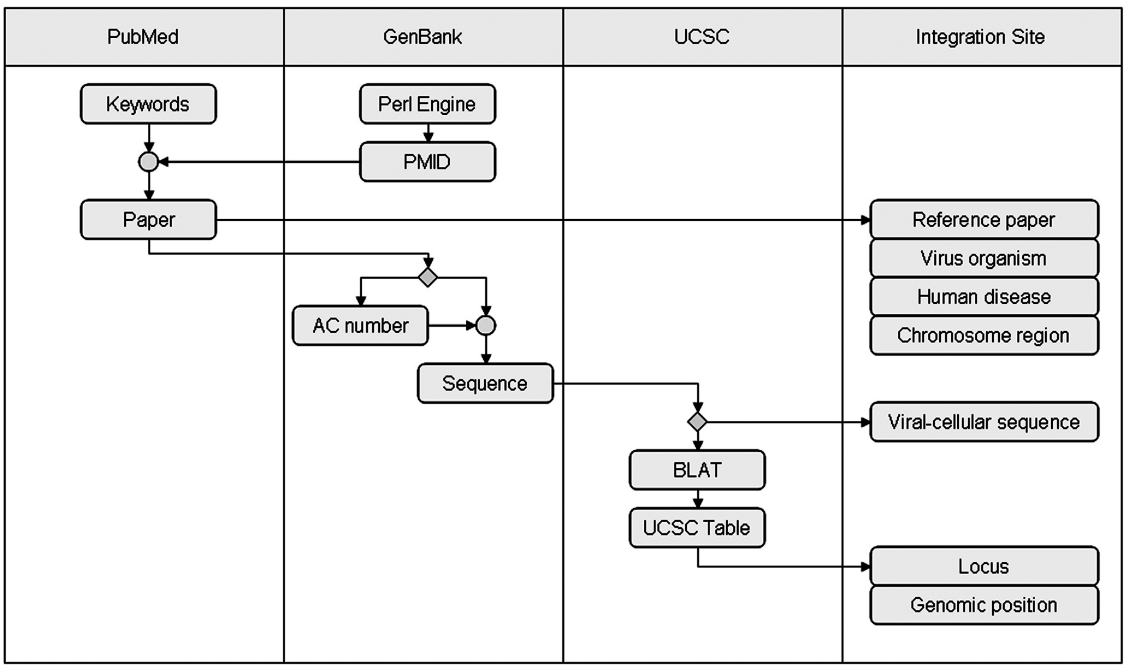

Figure 1. Work flow of data collection and re-mapping.

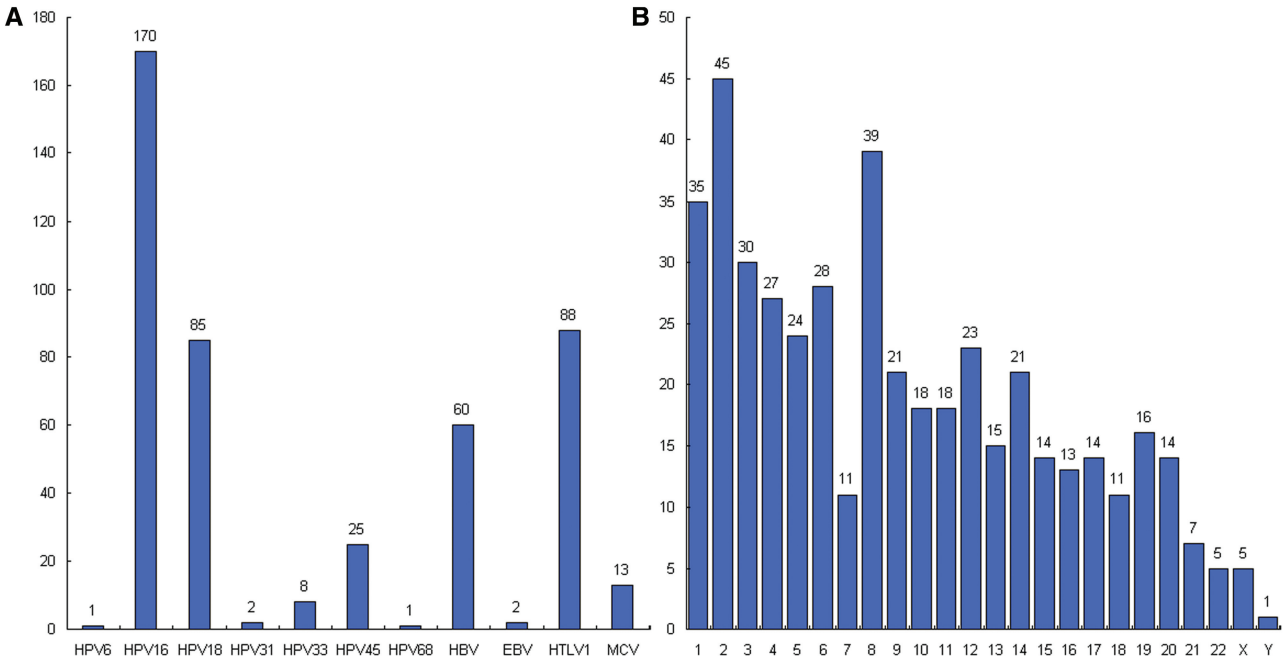

Figure 2. Distribution of VIS clusters associated with human malignant diseases. (**A**) Frequency of VIS clusters by virus type, (**B**) frequency of VIS clusters versus chromosome.

Papers reporting disease-related viral integration into the human genome were collected from PubMed in two ways, by script as described immediately above, and by manual search of the keywords virus, integration site, cancer, tumor, malignancy and disease. About 200 initially selected papers were obtained and filtered for relevance; curators read nearly 80 finally selected papers in full to extract the VIS characteristics required in the data model. In some cases, exact junctions were transcribed from illustrations in the papers. Sequences denoted with accession numbers are downloaded directly from GenBank.

Re-mapping of VIS

Three fields of a VIS (genomic position, chromosome region and locus) are updated by re-mapping according to the viral–host junction sequence obtained (Figure 1).

Mapping of genomic position. The genomic position of a VIS in the Human Genome Assembly 2009 (hg19) (31) is identified using BLAT from UCSC (33), provided that the identity of the BLAT result exceeds 80%. When there are two or more positive alignments, a manual check helps to choose the correct one.

Mapping of locus. The locus of integration is always interrupted, and potentially inactivated, by viral insertion. Loci were identified using the Genes and Gene Tracks Table from UCSC (34), and VIS were mapped to the gene component (exon, intron, 3′-untranslated region, promoter) on the basis of BLAT hit. All recognized loci were required to have been approved by the HGNC (30).

Mapping of chromosome region. The chromosome region (cytogenetic band) was subsequently calculated based on the insertion site's genomic position and the Chromosome Band Table from UCSC (34).

Clustering of VIS

As described in the data model, VIS are conditionally clustered as a unique data entry termed viral integration cluster (VIS cluster). A confidence code is assigned to each VIS cluster indicating its frequency, according to the number of insertion sites that it contains (Table 1). Statistics of integration clusters compellingly associated with human malignant disease are illustrated for the current build in Figure 2.

Web interfaces

Data browser. The data browser presents a catalog of links to chromosome, virus and disease. Currently, there are 24 chromosomes, 12 viruses and 12 diseases, which can be browsed for VIS.

Data search. Three search engines (keywords, position and the jQuery search engine) are implemented in the data interface. Users can search Dr.VIS with keywords of disease, virus, chromosome region, and so on, using the keyword search engine. VIS clusters can also be selected on the basis of genomic position or chromosome region (cytogenetic band). Users can filter the search result through the jQuery search engine, which is embed in the table list and is powered by jQuery.

Data visualization. For each VIS cluster, Dr.VIS provides an interface (Figure 3) with details and links to the UCSC Genome Browser and the Ensembl Genome Browser. The graphic view (Figure 4) summarizes the distribution of

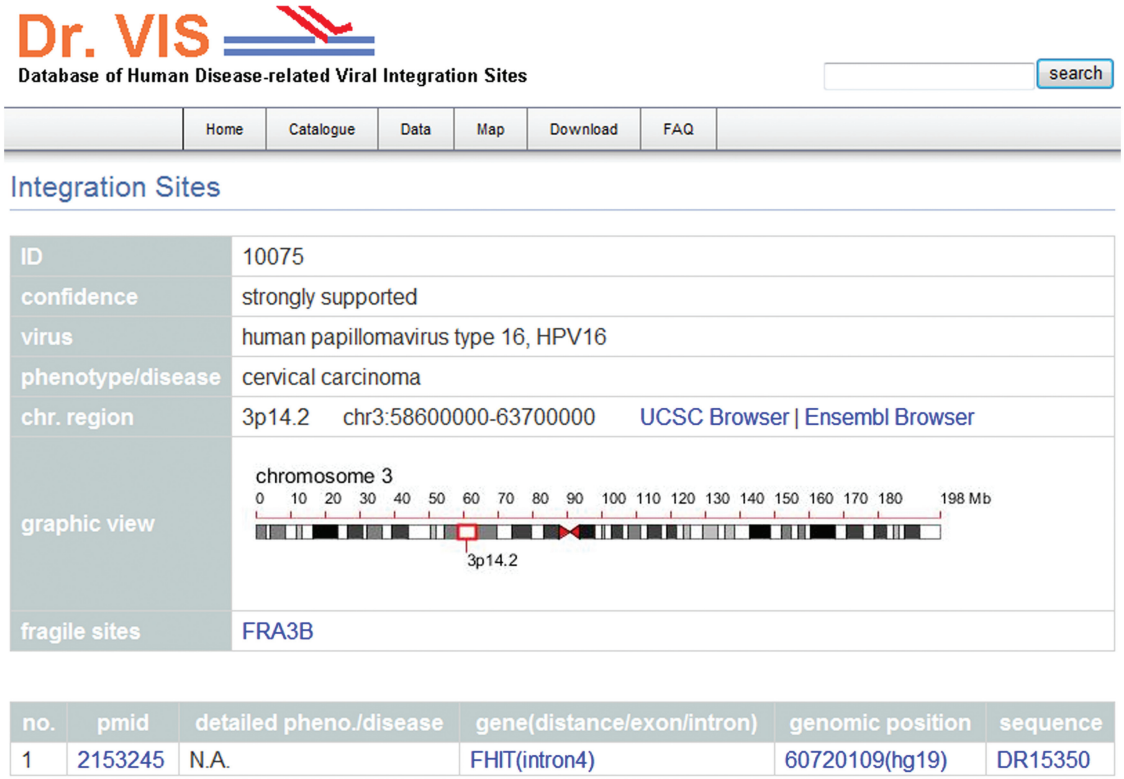

Figure 3. Screenshot of the VIS details interface.

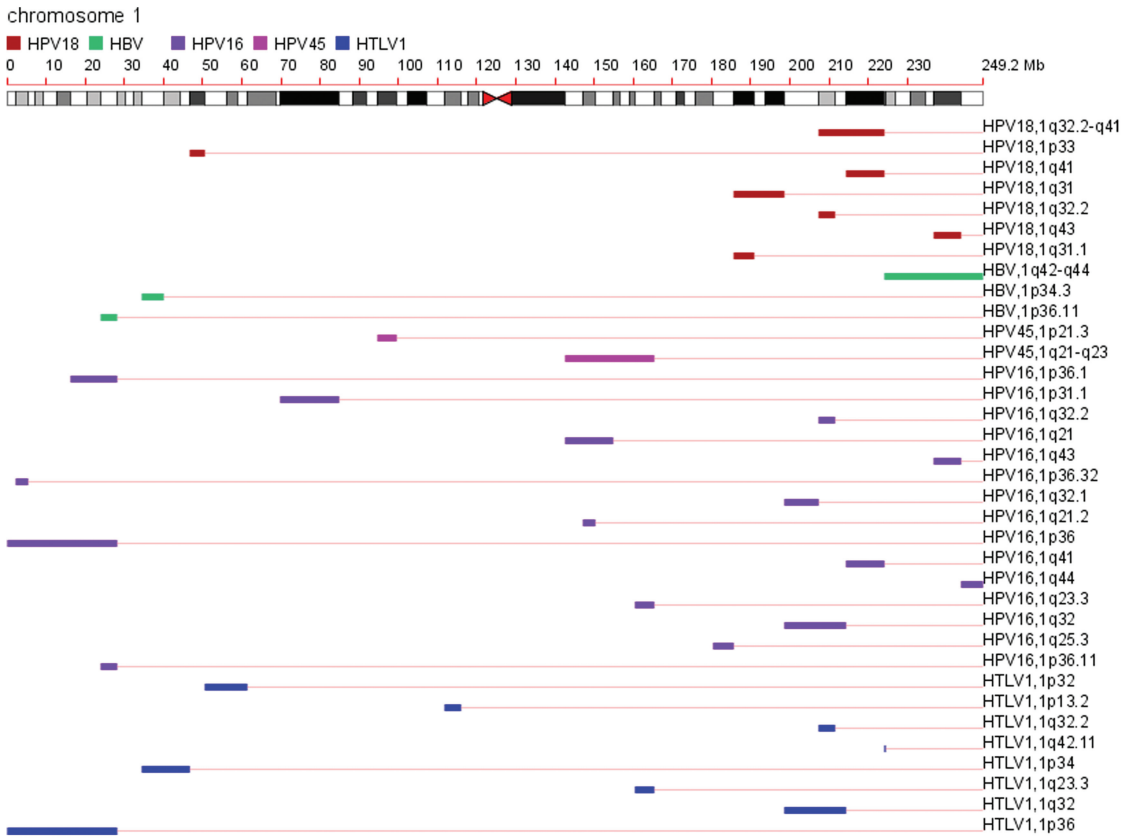

Figure 4. Screenshot of the graphic view of VIS located in human chromosome 1.

VIS clusters over any human chromosome. Any or all of the viruses can be selected for display.

DISCUSSION

VIS associated with malignant disease were always detected in samples from patients. Many useful approaches have been applied or newly developed to identify VIS such as fluorescence in situ hybridization (FISH), linear amplification mediated PCR (LAM-PCR) (35), amplification of papillomavirus oncogene transcripts assay (APOT), detection of integrated papilloma sequences PCR (DIPS-PCR) and next-generation sequencing (36–38). In addition to VIS, directly detected in naturally infected samples, many integration sites have been indentified in artificial experiments or in silico (39), as with SeqMap (23). Dr.VIS was developed as a comprehensive database of VIS associated with human malignant diseases. Dr.VIS is intended to facilitate biomedical applications or systematic researches into molecular causation and anomalies. The current build focuses on, oncogenic viruses demonstrably associated with human cancers. Viruses potentially resulting in anomalies are also of great interest. Updates of Dr.VIS will be continuously supported, since causative viruses continue to be identified and the number of documented VIS is rapidly increasing.

ACKNOWLEDGEMENTS

The authors thank Dr Charles F. Crane, USDA-ARS, West Lafayette, IN, USA for the great support during manuscript preparation. The authors are grateful to Ms Xinyi Liu from Shanghai Jiao Tong University for comments on the earlier versions of the manuscript. The authors also thank Dr Michael Galperin for the helpful suggestions.

FUNDING

Funding for open access charge: State Key Basic Research Program (973) (2011CB910204); National Natural Science Foundation of China (81101955); Major State Basic Research Development Program (2010CB945501); the 863 Hi-Tech Program of China (2009AA02Z308); National Key Technology R&D Program in the 11th Five Year Plan of China (2008BAI64B01) and Major State Basic Research Development Program of China (2010CB529200).

Conflict of interest statement. None declared.

REFERENCES

1. Talbot,S.J. and Crawford,D.H. (2004) Viruses and tumours – an update. *Eur. J. Cancer*, **40**, 1998–2005.
2. Morissette,G. and Flamand,L. (2010) Herpesviruses and chromosomal integration. *J. Virol.*, **84**, 12100–12109.
3. McLaughlin-Drubin,M.E. and Munger,K. (2008) Viruses associated with human cancer. *Biochim. Biophys. Acta*, **1782**, 127–150.
4. Giam,C.Z. and Jeang,K.T. (2007) HTLV-1 Tax and adult T-Cell leukemia. *Front. Biosci.*, **12**, 1496–1507.
5. zur Hausen,H. (1996) Papillomavirus infections–a major cause of human cancers. *Biochim. Biophys. Acta*, **1288**, F55–F78.
6. Kremsdorf,D., Soussan,P., Paterlini-Brechot,P. and Brechot,C. (2006) Hepatitis B virus-related hepatocellular carcinoma: paradigms for viral-related human carcinogenesis. *Oncogene*, **25**, 3823–3833.
7. Brechot,C. (2004) Pathogenesis of hepatitis B virus-related hepatocellular carcinoma: old and new paradigms. *Gastroenterology*, **127**, S56–S61.
8. Dandri,M., Burda,M.R., Bürkle,A., Zuckerman,D.M., Will,H., Rogler,C.E., Greten,H. and Petersen,J. (2002) Increase in de novo HBV DNA integrations in response to oxidative DNA damage or inhibition of poly(ADP-ribosyl)ation. *Hepatology*, **35**, 217–223.
9. Wang,H.P., Zhang,L., Dandri,M. and Rogler,C.E. (1998) Antisense downregulation of N-myc1 in woodchuck hepatoma cells reverses the malignant phenotype. *J. Virol.*, **72**, 2192–2198.
10. Klimov,E., Vinokourova,S., Moisjak,E., Rakhmanaliev,E., Kobseva,V., Laimins,L., Kisseljov,F. and Sulimova,G. (2002) Human papilloma viruses and cervical tumours: mapping of integration sites and analysis of adjacent cellular sequences. *BMC Cancer*, **2**, 24.
11. Kotin,R.M., Siniscalco,M., Samulski,R.J., Zhu,X.D., Hunter,L., Laughlin,C.A., McLaughlin,S., Muzyczka,N., Rocchi,M. and Berns,K.I. (1990) Site-specific integration by adeno-associated virus. *Proc. Natl Acad. Sci. USA*, **87**, 2211–2215.
12. Lewinski,M.K., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., Crawford,G., Collins,F., Shinn,P., Leipzig,J., Hannenhalli,S. *et al.* (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog.*, **2**, e60.
13. Poiesz,B.J., Ruscetti,F.W., Gazdar,A.F., Bunn,P.A., Minna,J.D. and Gallo,R.C. (1980) Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc. Natl Acad. Sci. USA*, **77**, 7415–7419.
14. Dürst,M., Gissmann,L., Ikenberg,H. and zur Hausen,H. (1983) A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc. Natl Acad. Sci. USA*, **80**, 3812–3815.
15. Boshart,M., Gissmann,L., Ikenberg,H., Kleinheinz,A., Scheurlen,W. and zur Hausen,H. (1984) A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. *EMBO J.*, **3**, 1151–1157.
16. Chang,Y., Cesarman,E., Pessin,M.S., Lee,F., Culpepper,J., Knowles,D.M. and Moore,P.S. (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, **266**, 1865–1869.
17. Epstein,M.A., Achong,B.G. and Barr,Y.M. (1964) Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet*, **1**, 702–703.
18. Dane,D.S., Cameron,C.H. and Briggs,M. (1970) Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. *Lancet*, **1**, 695–698.
19. Laude,H.C., Jonchère,B., Maubec,E., Carlotti,A., Marinho,E., Couturaud,B., Peter,M., Sastre-Garau,X., Avril,M.F., Dupin,N. *et al.* (2010) Distinct merkel cell polyomavirus molecular features in tumour and non tumour specimens from patients with merkel cell carcinoma. *PLoS Pathog.*, **6**, e1001076.
20. Kim,S., Kim,N., Dong,B., Boren,D., Lee,S.A., Das Gupta,J., Gaughan,C., Klein,E.A., Lee,C., Silverman,R.H. *et al.* (2008) Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. *J. Virol.*, **82**, 9964–9977.
21. Switzer,W.M., Jia,H., Zheng,H., Tang,S. and Heneine,W. (2011) No association of xenotropic murine leukemia virus-related viruses with prostate cancer. *PLoS One*, **6**, e19065.
22. Sakuma,T., Hué,S., Squillace,K.A., Tonne,J.M., Blackburn,P.R., Ohmine,S., Thatava,T., Towers,G.J. and Ikeda,Y. (2011) No evidence of XMRV in prostate cancer cohorts in the Midwestern United States. *Retrovirology*, **8**, 23.
23. Hawkins,T.B., Dantzer,J., Peters,B., Dinauer,M., Mockaitis,K., Mooney,S. and Cornetta,K. (2011) Identifying viral integration sites using SeqMap 2.0. *Bioinformatics*, **27**, 720–722.

24. Buchberg,A.M., Bedigian,H.G., Jenkins,N.A. and Copeland,N.G. (1990) Evi-2, a common integration site involved in murine myeloid leukemogenesis. *Mol. Cell. Biol.*, **10**, 4658–4666.
25. Akagi,K., Suzuki,T., Stephens,R.M., Jenkins,N.A. and Copeland,N.G. (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.*, **32**, D523–D527.
26. 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. *et al.* (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*, **302**, 415–419.
27. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
28. Sanborn,J.Z., Benz,S.C., Craft,B., Szeto,C., Kober,K.M., Meyer,L., Vaske,C.J., Goldman,M., Smith,K.E., Kuhn,R.M. *et al.* (2011) The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Res.*, **39**, D951–D959.
29. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
30. Seal,R.L., Gordon,S.M., Lush,M.J., Wright,M.W. and Bruford,E.A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.
31. The Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
32. Thorland,E.C., Myers,S.L., Gostout,B.S. and Smith,D.I. (2003) Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene*, **22**, 1225–1237.
33. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
34. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2002) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
35. Hüser,D., Gogol-Doring,A., Lutter,T., Weger,S., Winter,K., Hammer,E.M., Cathomen,T., Reinert,K. and Heilbronn,R. (2010) Integration preferences of wildtype AAV-2 for consensus rep-binding sites at numerous loci in the human genome. *PLoS Pathog.*, **6**, e1000985.
36. Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.
37. Mardis,E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.
38. Cronin,M. and Ross,J.S. (2011) Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology. *Biomark. Med.*, **5**, 293–305.
39. Li,Z., Liu,X., Wen,J., Xu,Y., Zhao,X., Li,X., Liu,L. and Zhang,X. (2011) DRUMS: a human disease related unique gene mutation search engine. *Hum. Mutat.*, **32**, E2259–E2265.