Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism

Zheng Hu[1,12], Da Zhu[1,12], Wei Wang[2,12], Weiyang Li[3,4,12], Wenlong Jia[3,5,12], Xi Zeng[3,12], Wencheng Ding[1], Lan Yu[1], Xiaoli Wang[1], Liming Wang[1], Hui Shen[1], Changlin Zhang[1], Hongjie Liu[3], Xiao Liu[3], Yi Zhao[3], Xiaodong Fang[3], Shuaicheng Li[5], Wei Chen[3], Tang Tang[6], Aisi Fu[7], Zou Wang[7], Gang Chen[1], Qinglei Gao[1], Shuang Li[1], Ling Xi[1], Changyu Wang[1], Shujie Liao[1], Xiangyi Ma[1], Peng Wu[1], Kezhen Li[1], Shixuan Wang[1], Jianfeng Zhou[8], Jun Wang[3,4,9–11], Xun Xu[3], Hui Wang[1] & Ding Ma[1]

Human papillomavirus (HPV) integration is a key genetic event in cervical carcinogenesis[1]. By conducting whole-genome sequencing and high-throughput viral integration detection, we identified 3,667 HPV integration breakpoints in 26 cervical intraepithelial neoplasias, 104 cervical carcinomas and five cell lines. Beyond recalculating frequencies for the previously reported frequent integration sites *POU5F1B* (9.7%), *FHIT* (8.7%), *KLF12* (7.8%), *KLF5* (6.8%), *LRP1B* (5.8%) and *LEPREL1* (4.9%), we discovered new hot spots *HMGA2* (7.8%), *DLG2* (4.9%) and *SEMA3D* (4.9%). Protein expression from *FHIT* and *LRP1B* was downregulated when HPV integrated in their introns. Protein expression from *MYC* and *HMGA2* was elevated when HPV integrated into flanking regions. Moreover, microhomologous sequence between the human and HPV genomes was significantly enriched near integration breakpoints, indicating that fusion between viral and human DNA may have occurred by microhomology-mediated DNA repair pathways[2]. Our data provide insights into HPV integration-driven cervical carcinogenesis.

HPV infection is recognized as the main cause of cervical carcinomas[3]. While most infections are cleared spontaneously by the immune system, a few persist for years and eventually cause cancer[4]. It may therefore be that, in addition to HPV infection, other viral risk factors and host susceptibilities are required to initiate the transformation process[5,6]. HPV DNA integration into the host genome is considered one of the most important risk factors for cervical carcinoma development[1,7]. The level of HPV integration was reported to be positively correlated with cervical intraepithelial neoplasia (CIN) grades and has even been proposed as a marker for disease progression[8,9]. Thus, identifying HPV integration hot spots in the human genome and elucidating the mechanisms of integration will yield insight into HPV-induced cervical carcinogenesis.

The first study to identify the exact HPV integration breakpoints in the human genome was carried out in 1987, when a single integrated copy of the virus was detected between *KLF5* and *KLF12* in SiHa cells[10]. Since then several studies to detect genomic integration loci using polymerase chain reaction (PCR) sequencing of host–viral DNA[10–14] or RNA[15] junctions have been reported. To our knowledge, 417 HPV breakpoints affecting approximately 389 genes have been identified by these methods (**Supplementary Table 1**). However, these studies suffered from relatively small sample sizes, and their breakpoints were either biased toward restriction enzyme sites in the human genome or limited to HPV early genes. We therefore performed high-throughput viral integration detection (HIVID), a next-generation sequencing and computational method developed by our group[16] (**Supplementary Figs. 1–3**), in 26 CINs, 104 cervical carcinomas and five cell lines. Our data provide, to our knowledge, the first genome-wide, unbiased, single-nucleotide-resolution HPV integration map in CINs and cervical carcinomas, revealing new hot spots and potential mechanisms.

In this study, we first conducted whole-genome sequencing (WGS; >30× coverage) and HIVID in parallel on the HPV-positive cell lines SiHa and HeLa, as well as two cervical carcinomas (**Supplementary**

[1]Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. [2]Department of Obstetrics and Gynecology, Nanfang Hospital, Southern Medical University, Guangzhou, China. [3]Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China. [4]School of Bioscience and Bioengineering, South China University of Technology, Guangzhou, China. [5]Department of Computer Science, City University of Hong Kong, Hong Kong, China. [6]WuHan Frasergen Bioinformatics Co., Ltd , Wuhan, China. [7]Wuhan Institute of Biotechnology, Wuhan, China. [8]Department of Hematology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China. [9]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia. [10]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark. [11]Centre for iSequencing, Aarhus University, Aarhus, Denmark. [12]These authors contributed equally to this work. Correspondence should be addressed to D.M. (dingma424@yahoo.com), H.W. (huit71@sohu.com) or X.X. (xuxun@genomics.cn).

Received 15 April; accepted 5 December; published online 12 January 2015; doi:10.1038/ng.3178

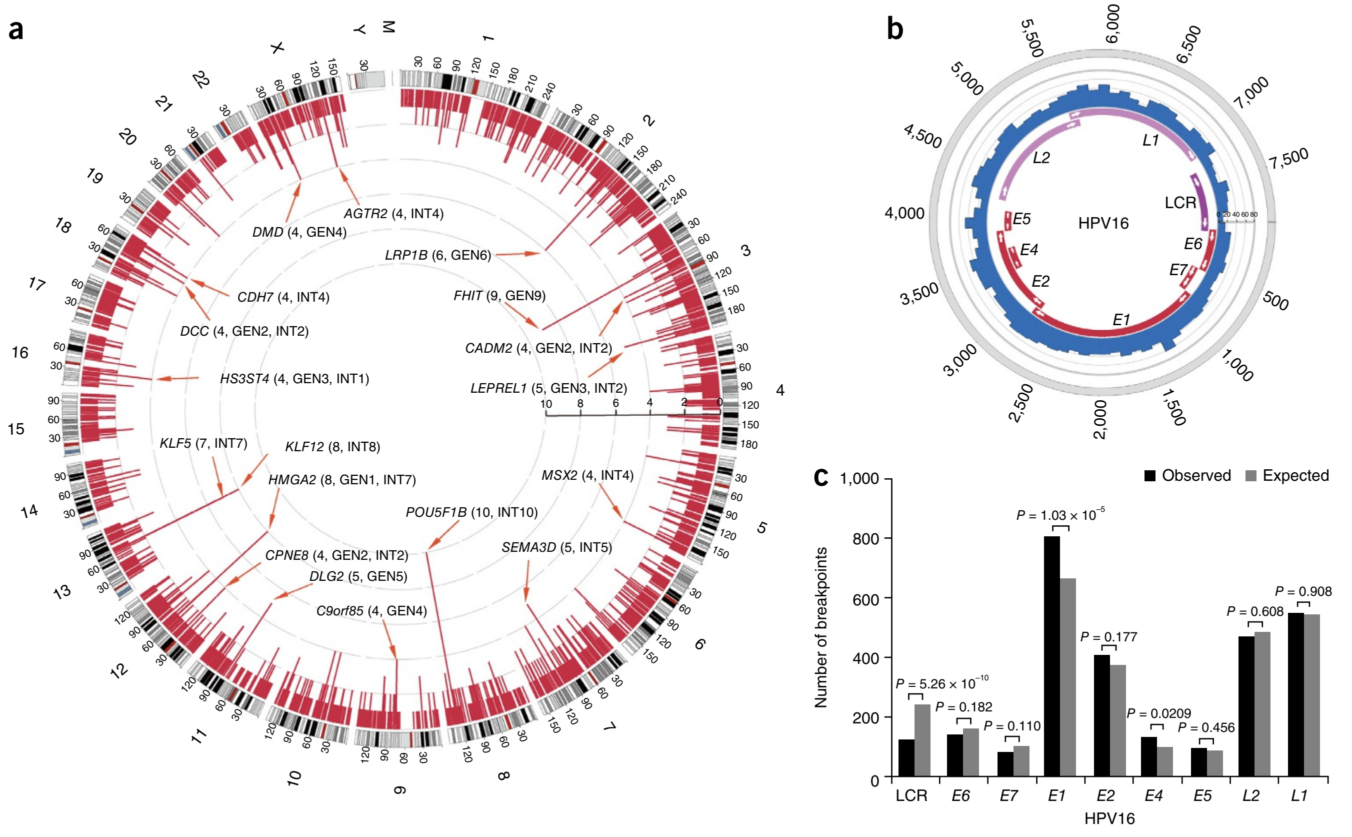

Figure 1 Distribution of breakpoints in the human and HPV genomes in 135 samples. (**a**) Distribution of integration breakpoints in the human genome. In the outer circle, each bar denotes the location of HPV integration into the 24 human chromosomes and the mitochondrial genome. Histogram axis units represent number of samples, and outer DNA numbering is given in millions of bases. In the inner circle, each red bar depicts the frequency of HPV integration. Some loci with high rates of integration are marked. GEN, breakpoints located within genes. INT, breakpoints located <500 kb from genes. (**b**) Distribution of integration breakpoints in the HPV16 genome. Histograms (black) of the frequency of breakpoints in the samples were constructed for 100-bp intervals. Histogram axis units represent number of breakpoints, and outer DNA numbering is given in bases. HPV genes with different functions are colored. (**c**) Comparison of the observed (blue) and expected (gray) numbers of breakpoints in the HPV16 genome. *P* values were calculated by chi-squared tests.

Tables 2–5). HIVID not only detected 10 of 11 HPV integration breakpoints identified via WGS but also discovered 135 other breakpoints not identified by WGS. These data prove that, by comparison to WGS, HIVID is a sensitive method for genome-wide survey of HPV integration breakpoints.

We next performed HIVID in more samples and detected a total of 3,666 HPV integration breakpoints in 103 of 135 samples, including 14 of 26 CINs, 85 of 104 cervical carcinomas and 4 of 5 HPV-positive cell lines (**Supplementary Tables 4** and **5** and **Supplementary Note**). Sanger sequencing and RNA sequencing (RNA-seq) were applied to validate HIVID breakpoints (**Supplementary Figs. 4** and **5**, **Supplementary Tables 6–9** and **Supplementary Note**). We calculated the frequency of affected genes (**Supplementary Table 10; for** detailed frequency distribution by type, see **Supplementary Fig. 6**, **Supplementary Tables 11–13** and Online Methods) and plotted

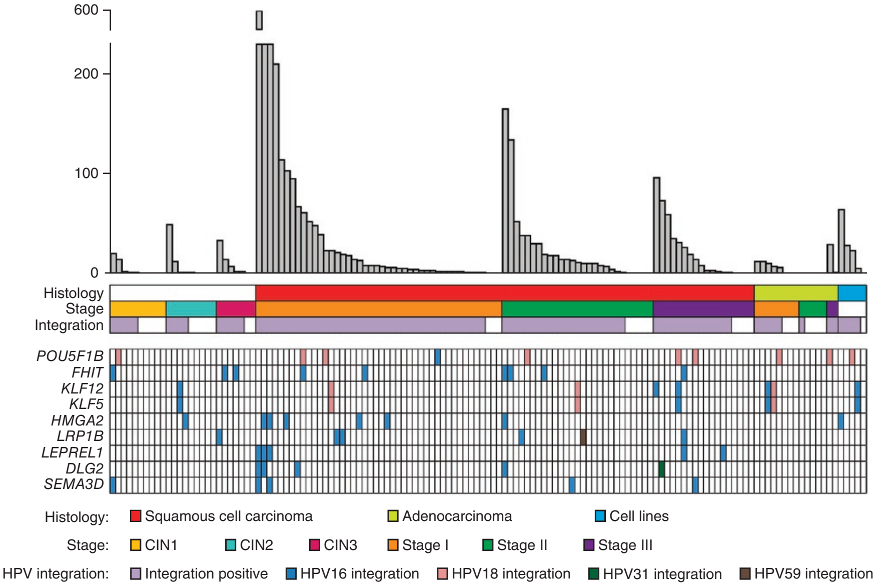

Figure 2 Clinical annotation of HPV integration sites in 104 cervical carcinomas, 26 CINs and five cell lines. All panels are aligned with vertical tracks representing 135 individuals. The data are sorted by histology, stage, integration (middle) and number of HPV integrations (top). The bottom heat map shows the distribution of HPV integrations into the nine genes that most frequently underwent integration events (≥5 events) in cervical carcinomas, CINs and cell lines.

Table 1 Demographic and clinicopathologic characteristics of cervical cancer and CIN patients and their associations with HPV integrations

		HPV integrations[a]						
	No. of	Negative		Positive			Median	
Variable	patients	No.	%	No.	%	*P* value[b]	integrations[a]	*P* value[c]
Cervical cancer patients								
Age								
<40	28	5	17.9	23	82.1	0.791	9.0	0.662
40–49	57	12	21.1	45	78.9		6.0	
≥50	17	2	11.8	15	88.2		8.0	
Stage								
Stage 1	52	6	11.5	46	88.5	0.076	7.0	0.445
Stage 2+	52	13	25.0	39	75.0		9.0	
Metastasis								
Negative	82	15	18.3	67	81.7	1.000	7.5	0.761
Positive	22	4	18.2	18	81.8		11.0	
Pathology								
Squamous cell carcinoma	89	12	13.5	77	86.5	**0.006**	8.0	**0.004**
Adenocarcinoma	15	7	46.7	8	53.3		1.0	
Differentiation								
Well or moderately	74	14	18.9	60	81.1	0.788	7.0	0.679
Poorly	30	5	16.7	25	83.3		9.0	
Neoadjuvant chemotherapy								
Negative	68	10	14.7	58	85.3	0.196	8.0	0.310
Positive	36	9	25.0	27	75.0		7.5	
Number of pregnancies								
<3	25	7	28.0	18	72.0	0.235	2.0	0.228
≥3	77	12	15.6	65	84.4		10.0	
Number of births								
<3	76	15	19.7	61	80.3	0.774	7.0	0.787
≥3	26	4	15.4	22	84.6		9.5	
Number of abortions								
<2	51	13	25.5	38	74.5	0.075	6.0	0.082
≥2	51	6	11.8	45	88.2		10.0	
Tubal ligation								
Negative	75	13	17.3	62	82.7	0.557	7.0	0.882
Positive	25	6	24.0	19	76.0		12.0	
CIN patients								
Grade								
CIN1	10	5	50	5	50	0.615	0.5	0.404
CIN2	9	5	55.6	4	44.4		0.0	
CIN3	7	2	28.6	5	71.4		2.0	
Age								
<40	12	7	58.3	5	41.7	0.666	0.0	0.600
40–49	10	4	40.0	6	60.0		1.0	
≥50	4	1	25.0	3	75.0		2.0	
CIN versus cancer								
CIN	26	12	46.2	14	53.8	**0.003**	1.0	**0.001**
Cancer	104	19	18.3	85	81.7		8.0	

[a]All integration breakpoints were included in this analysis. For analysis of integration breakpoints restricted to those located less than 500 kilobases from annotated genes, see **Supplementary Table 16**. [b]*P* values were calculated by chi-squared test. [c]*P* values were calculated by Mann-Whitney *U* test or Kruskal-Wallis test. Bold *P* values indicate $P < 0.05$.

the data against the human genome. Breakpoints were distributed throughout the whole genome, but with a handful of hot spots (**Fig. 1a**, 33 genes with ≥4 events and 9 genes with ≥5 events; **Supplementary Figs. 7** and **8** and **Supplementary Note**). Notably, 3,320 breakpoints (~90%) were located <500 kb from annotated genes and 1,546 breakpoints (~42%) were located within genes (**Supplementary Fig 9**), many of which involved tumor-associated pathways (**Supplementary Tables 14** and **15**). The strong tendency of HPV integration toward functional genes and the coexistence of hot spots with scattered breakpoints suggest that HPV may, from the beginning, randomly integrate into the host genome on the basis of genome accessibility[7,17,18]. However, over the long-term course of carcinogenesis, integrations at recurrent loci may provide selection advantages to host cells[19,20], leading to the recurrence of hot-spot genes in different samples.

To profile the integration pattern on the virus, we also annotated breakpoints on the HPV16 genome (**Fig. 1b**; for breakpoints of other HPV types, see **Supplementary Fig. 10**). Unexpectedly, in contrast with reports that integrated HPV16 should retain intact oncogenes *E6* and *E7* with the long control region (LCR)[1], we found that breakpoints could occur in any part of the viral genome, perhaps enabling the virus to adapt to the changing environment during carcinogenesis. Therefore, previous methods[9,21,22] to identify HPV16 integration status on the basis of the *E2*/*E6* ratio may be inaccurate because the *E6* gene may be disrupted in some events. We calculated the frequency in each viral gene or region and found that breakpoints were prone to occur in *E1* instead of *E2* ($P = 1.03 \times 10^{-5}$; **Fig. 1c**). In addition, we determined that breakpoints were less prone to occur in the LCR ($P = 5.26 \times 10^{-10}$ for the region containing the viral promoter; **Supplementary Fig. 11**), which was probably preserved because it acted as a strong *cis*-activator of nearby oncogene expression, promoting the malignant transformation of host cells[23,24].

As HPV integrations may be specific biomarkers for prediction of clinical outcomes, we analyzed the association between viral integration and clinicopathological parameters in CINs and cervical carcinomas (**Fig. 2**). The overall rate of HPV integration in all samples was 76.3% (103 of 135). Both integration rate and integration number in squamous cell carcinomas (86.5%, median eight integrations, range 0–599) were significantly higher than in adenocarcinomas (53.3%, median 1 integration, range 0–29, $P = 6 \times 10^{-3}$ for rate, $P = 4 \times 10^{-3}$ for number). During the progression of cervical lesions, the rate of HPV integration rose markedly from 53.8% (14 of 26) of CINs to 81.7% (85 of 104) of cervical carcinomas ($P = 3 \times 10^{-3}$; **Table 1**, **Supplementary Table 16** and **Supplementary Figs. 12** and **13**). And the number of integrations increased from a median of 1 integration per case in CINs (range 0–49) to a median of 8 integrations per case in cervical carcinomas (range 0–599, $P = 1 \times 10^{-3}$). Our data reveal that HPV integrations occurred in the initiating stage (for example, CIN1) of cervical carcinogenesis. The increase of both integration rate and number from CINs to cancer highlights their potential values as predictors of disease progression[22].

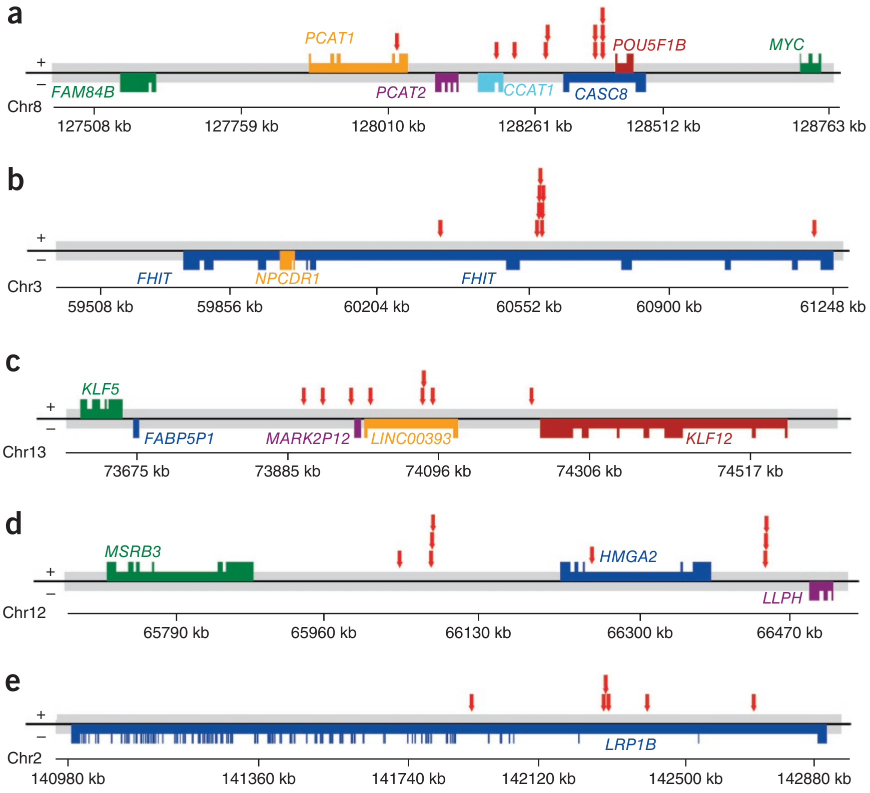

Figure 3 Mapping five HPV integration hot spots in 135 samples. (**a**–**e**) The HPV breakpoint sites on the recurrently affected genes *POU5F1B* (**a**), *FHIT* (**b**), *KLF5*–*KLF12* (**c**), *HMGA2* (**d**) and *LRP1B* (**e**) were mapped to the human hg19 reference sequence. Red arrows, locations of HPV integration breakpoints in a given sample. Chr, chromosome; boxes, gene regions; bulges, exons.

We also gained the opportunity to investigate the distribution of integration hot spots (**Fig. 2**) in different cancer stages and histological types from a genome-wide perspective. We first recalculated the frequencies of previously reported integrations at *POU5F1B* (9.7%), *FHIT* (8.7%), *KLF12* (7.8%), *KLF5* (6.8%), *LRP1B* (5.8%) and *LEPREL1* (4.9%) in our data set (**Supplementary Table 10**; for detailed gene functions, see **Supplementary Note**). More importantly, we identified new hot spots, including *HMGA2* (7.8%), *DLG2* (4.9%) and *SEMA3D* (4.9%). Notably, integrations at *POU5F1B*, *KLF5*–*KLF12*, *FHIT*, *HMGA2*, *LRP1B* and *SEMA3D* were detected in both CINs and cancer, and integrations at *POU5F1B* and *KLF5*–*KLF12* were further shared by squamous cell carcinomas and adenocarcinomas. The involvements of these hot spots in CINs indicate their roles in the early stage of cervical carcinogenesis. Intriguingly, we noted that HPV integrations exerted different effects on these important targets, which were related with their integration positions (**Figs. 3** and **4** and **Supplementary Fig. 14**). *FHIT* and *LRP1B*, whose HPV breakpoints were in their introns, generally displayed decreased protein expression in neoplastic tissues compared to that in adjacent normal tissues (**Fig. 4a**). And loss of *FHIT* and *LRP1B* protein expression was associated with HPV integration in their introns (**Fig. 4b**). By contrast, *MYC* (near *POU5F1B*) and *HMGA2*, with HPV integrations occurring in their flanking regions, showed increased protein expression in neoplastic tissues relative to that in adjacent normal tissues (**Fig. 4a**). And elevated protein expression of *MYC* and *HMGA2* were related to HPV integration in their flanking regions (**Fig. 4b**). Our data support the hypothesis that HPV may efficiently survey the human genome, activating or inactivating genes that favor

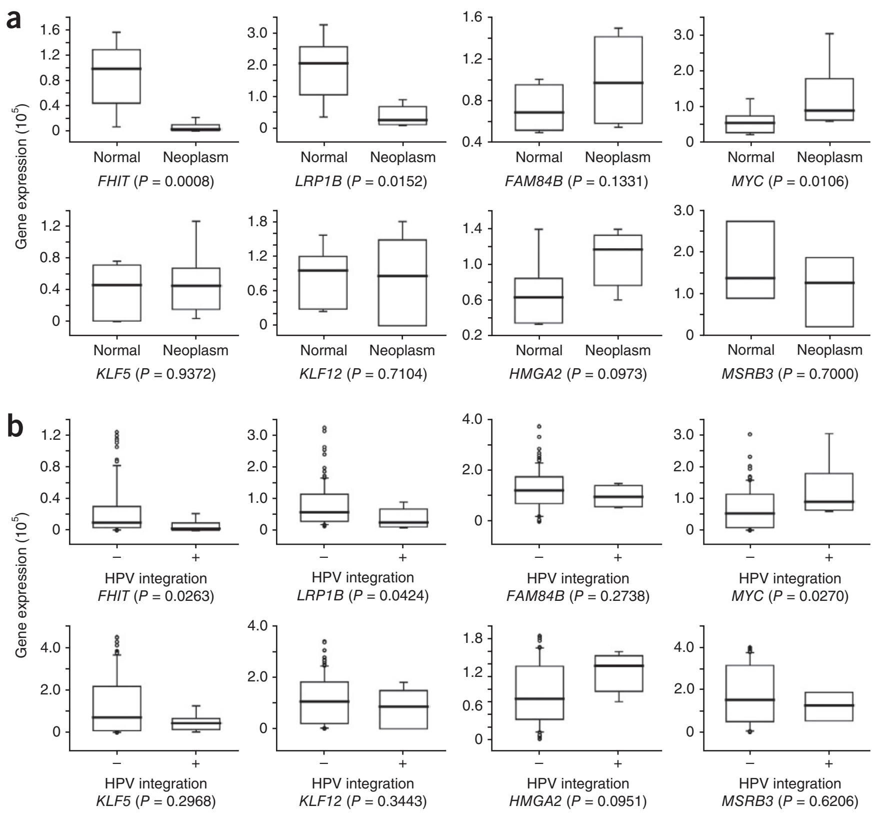

Figure 4 Effects of HPV integration on gene expression in samples assessed by immunohistochemistry. (**a**) Protein expression levels (arbitrary units) from genes that frequently harbored HPV integrations in cervical neoplasm tissues versus adjacent normal tissues. (**b**) Comparison of protein expression levels from genes that frequently underwent integration events in samples with or without HPV integration. Staining optical intensities and *P* values (Mann-Whitney *U* test) are shown in both panels. For each box plotting, the central mark represents the median, the edges of the box represent the 25th and 75th percentiles, and the whiskers are the most extreme data points not considered outliers. For numbers of samples analyzed per group, see **Supplementary Table 22**; information on the samples subjected to immunohistochemical staining is summarized in **Supplementary Table 23**.

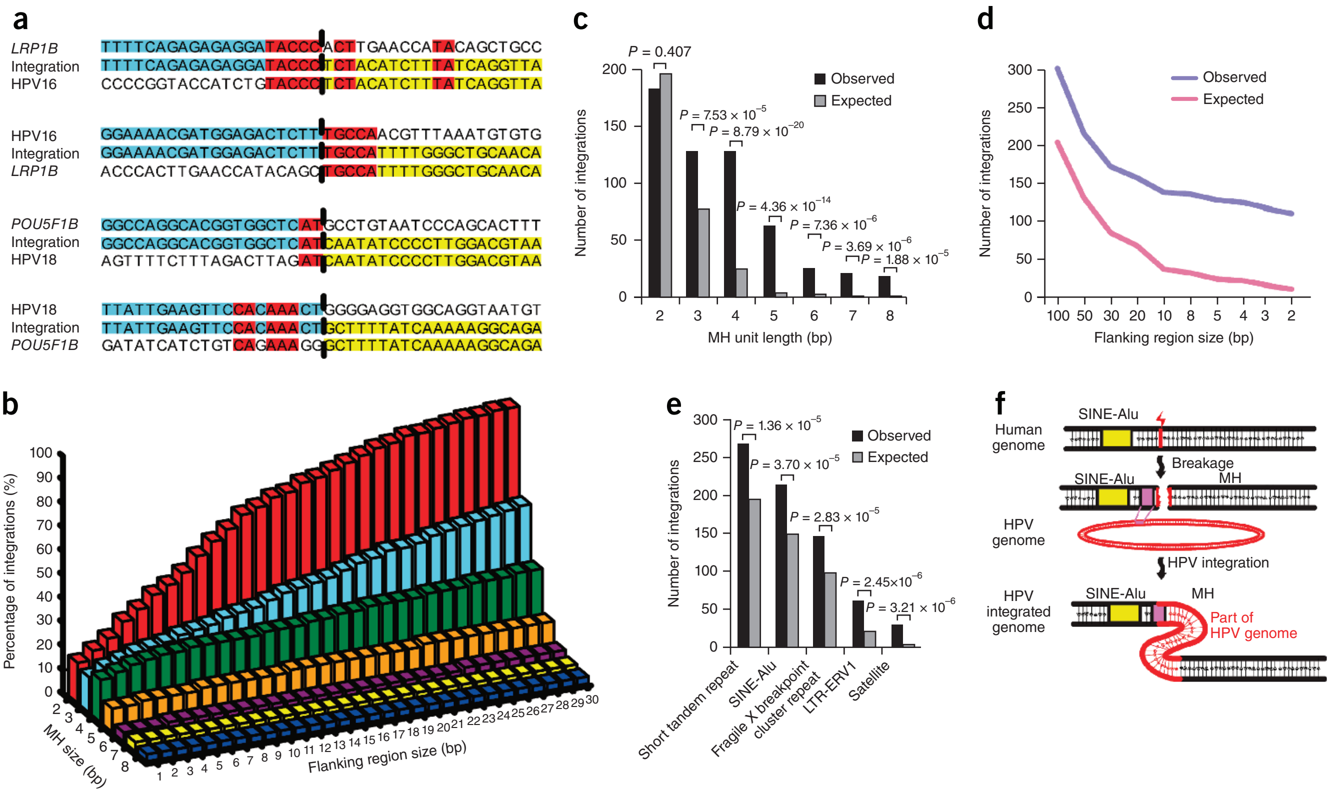

Figure 5 MH sequences are significantly enriched in the regions flanking integration sites. In the 135 samples, there were 600 major integration sites used for MH analysis (see Online Methods). (**a**) Alignment of the sequence around the integration site between the human genome and the HPV genome. Junction boundaries are shown as vertical dashed lines. *LRP1B* and *POU5F1B* are hot spots for HPV16 and HPV18, respectively. All HPV sequences are from the reference strand. Blue, upstream partner; yellow, downstream partner; red, nucleotides that align to both reference sequences (MHs). (**b**) Lego plots of MHs of different sizes in flanking regions around integration sites. (**c**) Comparison between observed and expected integrations harboring MHs of different sizes in 5-bp flanking regions. *P* values are noted at the top of bars. (**d**) Comparison between observed and expected integrations with 4-bp MHs in flanking regions of different sizes. (**e**) Genomic elements significantly enriched with integration sites, compared with expected occurrences. *P* values are noted at the top of bars. (**f**) Schematic of a mechanism connecting breakage of the human genome around SINE-Alu motifs with integration of the HPV genome. Yellow, SINE-Alu; purple, MH; red, HPV sequence. *P* values were calculated by chi-squared tests.

positive clonal selection and thereby providing more opportunities for the malignant transformation of host cells[25].

Another important unanswered question is the mechanism of HPV integration. In our data set, we observed a significant enrichment of microhomologies (MHs) between the human genome and the HPV genome at or near integration breakpoints (**Fig. 5a–d**, **Supplementary Fig. 15**, **Supplementary Tables 17** and **18**, and **Supplementary Note**), particularly when MH length was ≥4 bp ($P < 0.001$; **Fig. 5c**). When the radius of the flanking region was extended, the enrichment of 4-bp MHs remained significant ($P < 0.001$; **Fig. 5d**). This phenomenon indicates that MH-mediated DNA repair pathways, such as fork stalling and template switching (FoSTeS)[26,27] and microhomology-mediated break-induced replication (MMBIR)[27,28], may be the main mechanisms mediating the integration process.

Because FoSTeS and MMBIR pathways were usually triggered by local genomic instabilities (**Supplementary Fig. 16**), we analyzed the genomic instability–related genomic elements in the flanking regions of integrated sites[2,28] and surveyed their relations to copy number variations (CNVs) by high-resolution array-based comparative genomic hybridization (array-CGH; **Supplementary Figs. 17** and **18**). Several genomic instability–related genomic elements were determined to be significantly enriched ($P < 0.001$; **Fig. 5e**), including fragile X breakpoint cluster repeats, short tandem repeats, long terminal repeat–endogenous retroviral sequence 1 (LTR-ERV1), satellite and short interspersed nuclear element (SINE)-Alu repeats. When only integrations with MHs (4 bp length in ± 5-bp flanking region) were considered, the enrichments of satellite and SINE-Alu DNA remained evident ($P = 1.25 \times 10^{-4}$ and $P = 4.4 \times 10^{-3}$, respectively; **Supplementary Fig. 19a**). By contrast, array-CGH showed that CNVs were significantly enriched not only at all HPV integration sites (**Supplementary Fig. 20a**) but also at the ones with MHs (**Supplementary Fig. 20b**) and specific genomic elements (**Supplementary Figs. 21** and **22**). Thus we propose a new viral integration model: under certain circumstances (for example, HPV infection), local genomic elements (satellite and SINE-Alu DNA) are unstable and tend to form DNA breaks or stalled replication forks, resulting in activation of the FoSTeS or MMBIR pathways. Then, facilitated by the MHs flanking the breakpoint, HPV may hijack MH-mediated DNA repair pathways to fuse itself to the broken host genome and complete the integration process (**Fig. 5f**, **Supplementary Fig. 23** and **Supplementary Table 19**). Similar evidence supporting our theory can be found in a previously reported data set (**Supplementary Figs. 15**, **19b** and **24**, and **Supplementary Tables 18** and **20**).

In summary, we report a genome-wide unbiased analysis of HPV integration in CINs and cervical carcinomas. The HIVID strategy of HPV integration breakpoint identification represents a new tool for cervical screening programs. The integration hot spots identified in our study could be used as biomarkers for early personalized treatment and prognosis assessment of patients with cervical carcinoma[29,30] (**Supplementary Table 21**). We also discovered a significant enrichment of MHs at HPV integration sites, suggesting integration events are facilitated by MH-mediated DNA repair pathways. Our data shed light on HPV integration hot spots in cervical carcinomas and their underlying mechanisms.

URLs. Power and sample size calculation, http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize; repetitive region list, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/rmsk.txt.gz; non-B region list, http://nonb.abcc.ncifcrf.gov/apps/site/default.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. SRA: SRA180295, SRA189004 and SRA189003.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by funds from the National Development Program (973) for the Key Basic Research of China (2015CB553903 and 2013CB911304) and the National Natural Science Funding of China (81230038, 81372805, 81172466, 81272859, 81372801, 81372804, 81402158, 81370469, 81372806, 81370469, 81172468, 81101964, 81230052 and 81302266). The study is also sponsored by the Chinese 863 Program (2012AA02A507, 2012AA02A201), Guangdong Enterprise Key Laboratory of Human Disease Genomics, ShenZhen Engineering Laboratory for Clinical Molecular Diagnostic and China National GeneBank–Shenzhen. We thank all participants recruited for this study. We thank our colleagues from BGI, H. Huang, L. Huang, X. Zhuang, L. Lin, H. Cao, X. Fang, X. Zhang, Y. Shuang and H. Yang for sequencing and analysis.

AUTHOR CONTRIBUTIONS

D.M. took full responsibility for the study, especially in conceiving, designing and supervising the research together with Z.H. and H.W. W.W., J.Z., J.W. and X.X. participated in the study design. G.C., Q.G., Shuang Li., L.X., C.W., S. Liao, X.M., P.W., K.L. and S.W. supervised the diagnosis of patients and subject recruitment. D.Z., H.S., W.J., W.L., X.Z., H.L., X.F., Shuaicheng Li and X.L. performed statistical analysis. W.L., X.Z. and Y.Z. performed HIVID and RNA-seq analysis. W.J. performed FuseSV (in-house software for MH analysis) and WGS analysis. W.D., L.Y., L.W., H.S., X.W., C.Z., W.C., T.T., A.F. and Z.W. performed the experiments. The manuscript was drafted by Z.H., D.Z., W.J., W.L. and X.Z. under the supervision of D.M., H.W. and X.X. All authors critically reviewed the article and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Pett, M. & Coleman, N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.* **212**, 356–367 (2007).
2. Lawson, A.R. *et al.* RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res.* **21**, 505–514 (2011).
3. zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2**, 342–350 (2002).
4. Crosbie, E.J., Einstein, M.H., Franceschi, S. & Kitchener, H.C. Human papillomavirus and cervical cancer. *Lancet* **382**, 889–899 (2013).
5. Shi, Y. *et al.* A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. *Nat. Genet.* **45**, 918–922 (2013).
6. Stanley, M.A., Pett, M.R. & Coleman, N. HPV: from infection to cancer. *Biochem. Soc. Trans.* **35**, 1456–1460 (2007).
7. Wentzensen, N., Vinokurova, S. & von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **64**, 3878–3884 (2004).
8. Hudelist, G. *et al.* Physical state and expression of HPV DNA in benign and dysplastic cervical tissue: different levels of viral integration are correlated with lesion grade. *Gynecol. Oncol.* **92**, 873–880 (2004).
9. Arias-Pulido, H., Peyton, C.L., Joste, N.E., Vargas, H. & Wheeler, C.M. Human papillomavirus type 16 integration in cervical carcinoma in situ and in invasive cervical cancer. *J. Clin. Microbiol.* **44**, 1755–1762 (2006).
10. el Awady, M.K., Kaplan, J.B., O'Brien, S.J. & Burk, R.D. Molecular analysis of integrated human papillomavirus 16 sequences in the cervical cancer cell line SiHa. *Virology* **159**, 389–398 (1987).
11. Thorland, E.C. *et al.* Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. *Cancer Res.* **60**, 5916–5921 (2000).
12. Luft, F. *et al.* Detection of integrated papillomavirus sequences by ligation-mediated PCR (DIPS-PCR) and molecular characterization in cervical cancer cells. *Int. J. Cancer* **92**, 9–17 (2001).
13. Kalantari, M., Blennow, E., Hagmar, B. & Johansson, B. Physical state of HPV16 and chromosomal mapping of the integrated form in cervical carcinomas. *Diagn. Mol. Pathol.* **10**, 46–54 (2001).
14. Xu, B. *et al.* Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PLoS ONE* **8**, e66693 (2013).
15. Klaes, R. *et al.* Detection of high-risk cervical intraepithelial neoplasia and cervical cancer by amplification of transcripts derived from integrated papillomavirus oncogenes. *Cancer Res.* **59**, 6132–6136 (1999).
16. Li, W. *et al.* HIVID: an efficient method to detect HBV integration using low coverage sequencing. *Genomics* **102**, 338–344 (2013).
17. Dall, K.L. *et al.* Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Res.* **68**, 8249–8259 (2008).
18. Ferber, M.J. *et al.* Preferential integration of human papillomavirus type 18 near the c-myc locus in cervical carcinoma. *Oncogene* **22**, 7233–7242 (2003).
19. Schmitz, M. *et al.* Loss of gene function as a consequence of human papillomavirus DNA integration. *Int. J. Cancer* **131**, E593–E602 (2012).
20. Peter, M. *et al.* MYC activation associated with the integration of HPV DNA at the MYC locus in genital tumors. *Oncogene* **25**, 5985–5993 (2006).
21. Boulet, G.A. *et al.* Human papillomavirus 16 load and E2/E6 ratio in HPV16-positive women: biomarkers for cervical intraepithelial neoplasia >or=2 in a liquid-based cytology setting? *Cancer Epidemiol. Biomarkers Prev.* **18**, 2992–2999 (2009).
22. Gradíssimo Oliveira, A., Delgado, C., Verdasca, N. & Pista, A. Prognostic value of human papillomavirus types 16 and 18 DNA physical status in cervical intraepithelial neoplasia. *Clin. Microbiol. Infect.* **19**, E447–E450 (2013).
23. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).
24. Ojesina, A.I. *et al.* Landscape of genomic alterations in cervical carcinomas. *Nature* **506**, 371–375 (2014).
25. Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I.B. & Durst, M. Non-random integration of the HPV genome in cervical cancer. *PLoS ONE* **7**, e39632 (2012).
26. Lee, J.A., Carvalho, C.M. & Lupski, J.R.A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
27. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849–853 (2009).
28. Verdin, H. *et al.* Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLoS Genet.* **9**, e1003358 (2013).
29. Campitelli, M. *et al.* Human papillomavirus mutational insertion: specific marker of circulating tumor DNA in cervical cancer patients. *PLoS ONE* **7**, e43393 (2012).
30. Das, P. *et al.* HPV genotyping and site of viral integration in cervical cancers in Indian women. *PLoS ONE* **7**, e41012 (2012).

ONLINE METHODS

HPV typing. The two most commonly used consensus primer sets, MY09/MY11 and GP5+/GP6+, were used to amplify a broad spectrum of HPV genotypes in a single reaction. Samples were also subjected to PCR using specific primers for HPV types 11, 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66 and 68 as previously described[31–33] (**Supplementary Fig. 25** and **Supplementary Tables 24**–**26**). The PCR products were Sanger sequenced, and the sequencing results were aligned by the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST). All HPV-positive specimens were referred to subsequent high-throughput sequencing.

Sample collections. All samples were taken from Tongji hospital, Wuhan and Jingmen No. 2 People's Hospital, Jingmen, Hubei, China, between 2007 and 2014. Exfoliated cervical epithelial cells were obtained with cervical brushes. Patients with abnormal cytological results were referred to subsequent cervical biopsy. The exfoliated cervical epithelial cell samples from patients diagnosed with CIN were subjected to DNA extraction. Cancer tissues were obtained by experienced surgeons and were frozen in liquid nitrogen rapidly to await DNA extraction. Typical tumor sites were determined by two independent pathologists. Typical cancer tissues were cut from the specimens with new scalpels, and all operations are carried out carefully to avoid contamination between samples. Written informed consent was obtained from each patient, and the study was approved by the institution's ethics committee. The inclusion criteria of this study were (i) HPV-positive CINs and HPV-positive cervical carcinomas (ii) obtained from consenting patients and (iii) having adequate and sufficiently high quality DNA (quantity ≥ 3 μg, concentration ≥ 20 ng/μL and an apparent main band upon electrophoresis) for next-generation sequencing. Because sample size calculation for adequate power required assumptions based on previous observation, we first selected a different integration rate between CINs and cervical carcinomas as a specific effect. Prior data indicated that the integration rate among controls (CINs) was 0.4 (ref. 34). If the true integration rate for experimental subjects (cervical carcinomas) was 0.9 (ref. 34), then at least 17 cervical carcinomas and 17 CINs were to be needed to be able to reject the null hypothesis that the failure rates for experimental and control subjects were equal with probability (power) 0.8. The type I error probability associated with this test of this null hypothesis is 0.05. We used a continuity-corrected chi-squared statistic or Fisher's exact test to evaluate this null hypothesis. These power and sample size calculations were done using the software Power and Sample Size Calculation version 3.0. We initially screened 126 CINs (60 CIN1s, 34 CIN2s and 32 CIN3s) and 131 cervical carcinomas (101 squamous cell carcinomas and 30 adenocarcinomas, of which 26 CINs (10 CIN1s, 9 CIN2s and 7 CIN3s) and 104 cervical carcinomas (89 squamous cell carcinomas and 15 adenocarcinomas) were positive for HPV infection and had sufficient and qualified DNA.

Cell culture. Human cervical cancer cell lines CaSki, HeLa and SiHa were purchased from the American Type Culture Collection (ATCC). The immortalized human cervical keratinocyte cell line HaCaT came from the China Center for Type Culture Collection (CCTCC, Wuhan, China) and were transfected with HPV16 plasmid. The immortalized human cervical keratinocyte cell line S12 were obtained as a gift from K. Raj (Centre for Radiation, Chemical and Environmental Hazards, Health Protection Agency, Chilton, Didcot, UK) and M. Stanley (Department of Pathology, University of Cambridge)[35]. CaSki, HeLa and SiHa were grown in Dulbecco's modified Eagle's medium (Gibco) supplemented with 10% fetal bovine serum and antibiotics. S12 and HaCaT cells were maintained in DF12 medium, which consists of a mix of Dulbecco's modified Eagle's medium (Gibco) and Ham F-12 medium (Gibco) containing 5% fetal bovine serum (Gibco), penicillin, streptomycin and supplements: 8.4 ng/ml cholera toxin (Sigma), 5 μg/ml insulin (Sigma), 24.3 μg/ml adenine (Sigma), 0.5 μg/ml hydrocortisone (Sigma) and 10 ng/ml epithelial growth factor (PeproTech). All cells used were grown at 37 °C with 5% CO_2.

HPV fragments enrichment and sequencing. Full-length HPV genomes of 17 types (6, 11, 16, 18, 31, 33, 35, 39, 45, 52, 56, 58, 59, 66, 68, 69 and 82) were used to design the HPV probes by MyGenostics (**Supplementary Fig. 26** and **Supplementary Table 27**). Sequencing libraries were constructed following the instructions from Illumina. Genomic DNA was sheared to around 150–200 bp DNA fragments using a Covaris E-210 ultrasonicator (Covaris, Inc., Woburn, MA). These fragments were purified, end blunted, A-tailed and adaptor-ligated. The concentration of libraries was quantified by Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). The hybridization process was carried according to MyGenostics GenCap Target Enrichment Protocol (GenCap Enrichment, MyGenostics, USA). Libraries were hybridized with HPV probes including 17 types of HPV at 65 °C for 24 h and then washed to remove uncaptured fragments. The eluted fragments were amplified by 18 cycles of PCR to generate libraries for sequencing. Libraries were quantified and proceeded to 101 cycles of paired-end index sequencing in the Illumina HiSeq 2000 sequencer according to manufacturer's instructions (Illumina Inc., San Diego, CA; **Supplementary Figs. 1** and **2**).

Breakpoints detection and annotation of HPV integration sites. We carried out the analysis according to a previous algorithm[16]. Briefly, low quality reads and duplicate reads, as well as adaptor contamination reads, were removed to obtain clean reads for subsequent analysis. Clean reads were mapped to human (NCBI build 37, hg19) and HPV genome (HPV6: FR751337.1, HPV82: AF293961.1, HPV69: AB027020.1, HPV68: FR751039.1, HPV66: EF177191.1, HPV59: EU918767.1, HPV58: HQ537777.1, HPV56: EF177181.1, HPV52: HQ537751.1, HPV45: EF202167.1, HPV39: M62849.1, HPV35: HQ537730.1, HPV33: HQ537688.1, HPV31: HQ537687.1, HPV18: AY262282.1, HPV16: NC_001526.2, HPV11: HE574705.1). Those reads that paired-end aligned perfectly to the human or HPV genome were removed and the chimeric paired-end reads (partial read sequence aligned to human genome and partial aligned to HPV genome) were reserved. The reserved reads underwent paired-end reads assembly. The paired-end reads assembly was used to reconstruct fragment sequences and was able to increase the efficacy of locating the exact position of breakpoints (**Supplementary Fig. 3a**). The paired-end assembled reads were remapped to the human and HPV genomes using BWA[36]. The joint position of human and HPV sequence was the breakpoint for HPV integration (**Supplementary Fig. 3b**). The normalized number of support paired-end assembled reads (NNSS) were used to minimize the impact of total sequencing data. NNSS can be considered the number of supporting reads of each HPV integration breakpoint in every million read pairs. Integration breakpoints with NNSS >1 were selected. With this standard, the Sanger sequencing validation rate for HIVID was 83.1% in our study and 82.7% in a previous study[16]. The annotation of integrated breakpoints was performed by ANNOVAR[37]. Because integrated HPV was considered a strong *cis*-activator of flanking genes and *cis*-acting enhancers can influence their target genes over long distances (up to 1 Mb for upstream enhancers[38] and 850 kb for downstream enhancers[39]), breakpoints located <500 kb from annotated genes were included to calculate the affected gene frequency in HPV-integrated samples.

Whole genome sequencing and analysis. Genomic DNA from four samples was purified for ≥30×-coverage paired-end sequencing. DNA was fragmented with a Covaris E-210 ultrasonicator. By optimizing the shearing parameters, DNA fragments were concentrated at a peak of 500 bp for the relevant libraries. These fragments were purified, blunted, polyadenylated and ligated with adaptors. After size selection by gel electrophoresis, 10 to 12 cycles of PCR were performed. Paired-end 90-bp read length sequencing was performed on the HiSeq 2000 sequencer according to the manufacturer's instructions (Illumina). We carried out the bioinformatics analysis of HPV integration as in a previous study[40].

Detecting integration breakpoints by RNA-seq. We selected 11 samples (7 cervical cancers and 4 cell lines) with hot-spot genes detected by HIVID sequencing and sufficiently high quality RNA (quantity ≥ 400 ng, concentration ≥ 5 ng/μL, RNA integrity number (RIN) ≥ 7.0, 28S/18S ≥ 1.0, a smooth baseline and normal 5S peak in the electropherogram). RNA-seq libraries were sequenced as paired-end 90-bp sequence tags using the standard Solexa pipeline. We carried out the analysis of integration sites using the transcriptome data according to a previous method[16]. We removed reads that perfectly aligned to human or HPV genome and reserved chimeric paired-end reads. These chimeric reads comprised partial human genome sequence

doi:10.1038/ng.3178

and partial HPV genome sequence, and were used to identify HPV integration breakpoints in the transcriptome.

The definition of consistent breakpoints with RNA. The breakpoints in RNA should be located in the same gene as the breakpoints of DNA. The breakpoints in DNA and corresponding RNA should be less than 600 kb from each other if the breakpoints were detected in the intergenic region.

Validation of HPV integration positions. PCR and Sanger sequencing were used to verify the selected HPV integration breakpoints from HIVID. PCR primers were designed on the basis of the paired-end assembled fragment, in which one primer was located in the human genome and the other in the HPV genome. PCR was performed using a GeneAmp PCR System 9700 thermal cycler and sequenced on an Applied Biosystems 3730x DNA analyzer (Life Technologies, Inc).

HPV variant assignment. The genome sequences of HPV variants were downloaded from NCBI nucleotide database via GenBank[41]. Mutations of each HPV variant were identified by comparison with the reference sequence of its related HPV type. Reads that failed to map to the human reference were retrieved and aligned against the reference of each HPV type. Mutations of all HPV variants were used as a database to check their existence in a single sample. Unique mutations of each HPV variant were also confirmed. The HPV variants of each sample were determined on the basis of the mutations present (**Supplementary Figs. 27** and **28**).

Microhomology (MH) in the flanking region of the integration position. Clusters were found to harbor integration positions located near each other, among which one major integration case with more supporting reads than others could always be found. We applied in-house software, named FuseSV (not published), to detect the major case in each 2-Mb window, then used them to investigate the MHs around the integration breakpoints (see **Supplementary Fig. 29** for hot-spot genes). All major cases were required to have at least 5 supporting reads. We extracted the bilateral sequence from the human and HPV genomes and compared the paired sequences base by base to confirm the presence of the same sequences. We defined the contiguous identical bases that could not be extended as one single MH unit. After obtaining the MHs, we calculated the MHs' presence in a given flanking region (**Fig. 5b** and **Supplementary Table 18**). We defined one MH as located in the flanking region when it overlapped with this region. In the calculation, one MH was used only for the MH of its size, and not for any other, shorter MHs. To determine the expected MH distribution, viral integration breakpoints were randomly produced in the non-N region of the human genome reference. The number of random breakpoints was equal to that of the integration cases we reported. These breakpoints within the human genome were randomly and nonredundantly paired with the HPV breakpoints detected from the sequencing data, generating the data set of expected and observed integration cases. Next we extracted the bilateral sequences flanking the breakpoints from the human and HPV reference at a defined radius—for example, 20 bp long. The number of MHs of different sizes was counted in the defined range. Given the flanking region size and the MH size, the number of integration cases between expected and observed was compared via a chi-squared test.

Genomic elements around the integration position. The repetitive region list was downloaded from the UCSC genome browser. The non-B region list was from Non-B DB[42]. A given integration is confirmed to be adjacent to a genomic element when the element overlaps with the flanking region (200 bp). We confirmed 51 common motifs previously reported[2,43] and analyzed their presence, on either strand, within ±25 bp of the integration site. Randomly selected breakpoints across the entire human genome were used to calculate the expected presence.

Immunohistochemical staining. Formalin fixed and paraffin-embedded sections (4 μm) were subjected to immunohistochemical staining. The slides were incubated overnight at 4 °C with the primary antibody. We used rabbit anti-FHIT (1:30, ab3074, Abcam), mouse anti-LRP1B (1:50, SC-49229, Santa Cruz), rabbit anti-FAM84B (1:50, 18421-1-AP, ProteinTech), mouse anti-MYC (1:100, GTX80249, GeneTex), rabbit anti-KLF5 (1:200, GTX103289, GeneTex), rabbit anti-KLF12 (1:100, sc-84347, Santa Cruz), rabbit anti-HMGA2 (1:100, SAB2701959, Sigma) and rabbit anti-MSRB3 (1:100, A89705, Sigma). Antibody detection was performed using diaminobenzidine. Images were photographed from three randomly chosen fields using cellSens Dimension (version 1.8.1, Olympus), and the staining intensity was measured using ImagePro Plus (Version 6.0, Media Cybernetics). Information of the samples subjected to immunohistochemical staining is summarized in detail in **Supplementary Tables 22** and **23**.

Comparative genomic hybridization. For each sample, 250 ng genomic DNA was digested by NspI nuclease for 2 h at 37 °C. Digested DNA and adaptors were ligated by T4 DNA ligase for 3 h at 16 °C. Ligated DNA was amplified, fragmented, end-labeled with biotin and then hybridized to an Affymetrix CytoScan HD Array. Arrays were incubated at 50 °C for 16 h in a Hybridization Oven 645 with rotary motion (60 r.p.m.). Arrays were washed and stained in a Fluidics Station 450 with protocol "CytoScanHD_Array_450" and scanned with scanner 3000 7 G controlled by Affymetrix GeneChip Command Console Software (AGCC v4.0.0). Raw data were analyzed by Chromosome Analysis Suite (ChAS) Software v2.1 and copy number was determined by the Affymetrix CytoScanHD REF model. All microarray experiments were carried out by the CytoScan HD Array Kit and Reagent Kit Bundle (Cat No. 901835) and following the manufacturer's protocol. Data from the samples subjected to array-CGH are summarized in detail in **Supplementary Tables 4** and **28**.

Fluorescence *in situ* hybridization. Whole-genome plasmids of HPV16 and HPV18 were labeled by standard nick translation with digoxigenin–dUTP; the bacterial artificial chromosome (BAC) RP11-1145O20 plasmid was labeled with biotin–dUTP. Probes and target DNA were denatured simultaneously for 10 min at 90 °C before hybridization overnight at 37 °C. The HPV probe was detected by peroxidase-conjugated sheep anti-digoxigenin Fab fragments (1:200; 11207733910, Roche), then Cy3-labeled tyramide (1:50, NEL704A001KT, Perkin Elmer). The first peroxidase was inactivated by H_2O_2/sodium acetate/sodium azide buffer. The RP11-1145O20 probe was detected by streptavidin–HRP (1:100, NEL701A001KT, PerkinElmer), then FITC-labeled tyramide (1:50, NEL701A001KT, PerkinElmer). Slides were mounted in Vectashield (Vector Laboratories) containing 4′,6-diamidino-2-phenylindole (DAPI).

31. Karlsen, F. *et al.* Use of multiple PCR primer sets for optimal detection of human papillomavirus. *J. Clin. Microbiol.* **34**, 2095–2100 (1996).
32. Baay, M.F. *et al.* Comprehensive study of several general and type-specific primer pairs for detection of human papillomavirus DNA by PCR in paraffin-embedded cervical carcinomas. *J. Clin. Microbiol.* **34**, 745–747 (1996).
33. Walboomers, J.M. *et al.* Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **189**, 12–19 (1999).
34. Woodman, C.B., Collins, S.I. & Young, L.S. The natural history of cervical HPV infection: unresolved issues. *Nat. Rev. Cancer* **7**, 11–22 (2007).
35. Stanley, M.A., Browne, H.M., Appleby, M. & Minson, A.C. Properties of a non-tumorigenic human cervical keratinocyte cell line. *Int. J. Cancer* **43**, 672–676 (1989).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
38. Lettice, L.A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
39. Li, L. *et al.* A far downstream enhancer for murine Bcl11b controls its T-cell specific expression. *Blood* **122**, 902–911 (2013).
40. Sung, W.K. *et al.* Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44**, 765–769 (2012).
41. Burk, R.D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232–243 (2013).
42. Cer, R.Z. *et al.* Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41**, D94–D100 (2013).
43. Abeysinghe, S.S., Chuzhanova, N., Krawczak, M., Ball, E.V. & Cooper, D.N. Translocation and gross deletion breakpoints in human inherited disease and cancer I: nucleotide composition and recombination-associated motifs. *Hum. Mutat.* **22**, 229–244 (2003).

doi:10.1038/ng.3178