SCIENTIFIC REPORTS

OPEN

TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration

Received: 29 May 2018
Accepted: 26 November 2018
Published online: 24 January 2019

Sonja Lagström[1,2], Sinan Uğur Umu[2], Maija Lepistö[3], Pekka Ellonen[3], Roger Meisal[1], Irene Kraus Christiansen[1,4], Ole Herman Ambur[5] & Trine B. Rounge[2]

HPV genomic variability and chromosomal integration are important in the HPV-induced carcinogenic process. To uncover these genomic events in an HPV infection, we have developed an innovative and cost-effective sequencing approach named TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing). TaME-seq combines tagmentation and multiplex PCR enrichment for simultaneous analysis of HPV variation and chromosomal integration, and it can also be adapted to other viruses. For method validation, cell lines (n = 4), plasmids (n = 3), and HPV16, 18, 31, 33 and 45 positive clinical samples (n = 21) were analysed. Our results showed deep HPV genome-wide sequencing coverage. Chromosomal integration breakpoints and large deletions were identified in HPV positive cell lines and in one clinical sample. HPV genomic variability was observed in all samples allowing identification of low frequency variants. In contrast to other approaches, TaME-seq proved to be highly efficient in HPV target enrichment, leading to reduced sequencing costs. Comprehensive studies on HPV intra-host variability generated during a persistent infection will improve our understanding of viral carcinogenesis. Efficient identification of both HPV variability and integration sites will be important for the study of HPV evolution and adaptability and may be an important tool for use in cervical cancer diagnostics.

Human papillomavirus (HPV) is the main cause of cervical cancer[1], one of the most common cancers in women worldwide, causing more than 200,000 deaths each year[2,3]. A persistent infection with HPV high-risk genotypes is recognised as a necessary cause of cancer development[4]. Of the 13 carcinogenic high-risk types, HPV16 and 18 are associated with about 70% of all cervical cancers[5,6]. HPV infection is also associated with cancer in penis, vulva, vagina, anus, and head and neck[7]. However, only a small fraction of HPV infections at any site will progress to cancer[8]. This indicates that in addition to HPV infection, additional factors such as HPV genomic variability and integration, could contribute to the HPV-induced carcinogenic process. An appropriate sequencing approach is needed to uncover these genomic events during a persistent HPV infection.

HPV contains an approximately 7.9 kb circular double-stranded DNA genome, consisting of early region (E1, E2, E4-7) genes, late region (L1, L2) genes and an upstream regulatory region (URR)[9]. To date, more than 200 HPV types have been identified[10]. Each individual HPV type shares at least 90% sequence identity in the conserved L1 open reading frame (ORF) nucleotide sequence. Isolates of the same HPV types that differ by 1–10% or 0.5–1% across the genome are referred to as variant lineages or sublineages, respectively[11,12].

Despite phylogenetic relatedness, HPV variant lineages can differ in their carcinogenic potential[13–16]. Traditionally, studies have focused on cancer risk of main variants. However, recent studies have revealed variability below the level of variant lineages that may be evidence of intra-host viral evolution and adaptation[17–20]. In contrast to a limited number of studies on HPV variability, HPV integration into the host genome has been more

[1]Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway. [2]Department of Research, Cancer Registry of Norway, Oslo, Norway. [3]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [4]Clinical Molecular Biology (EpiGen), Medical Division, Akershus University Hospital and Institute of Clinical Medicine, University of Oslo, Norway. [5]Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway. Correspondence and requests for materials should be addressed to T.B.R. (email: trine.rounge@kreftregisteret.no)

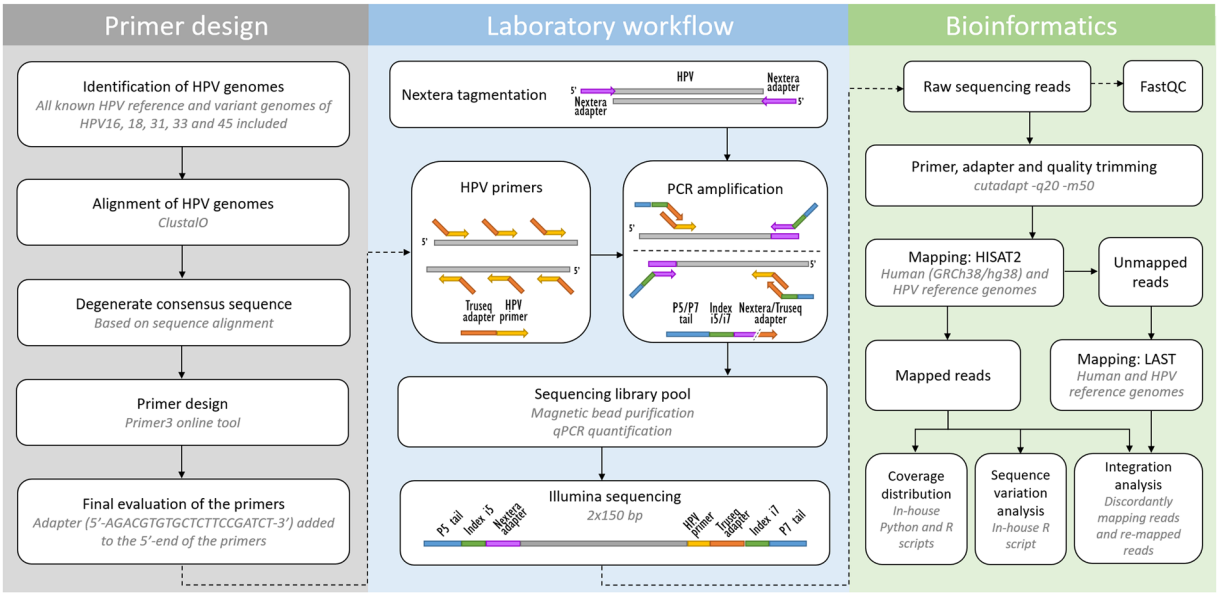


Figure 1. Primer design, laboratory and bioinformatics workflows of the TaME-seq method.

widely studied and is regarded as a determining event in cervical carcinogenesis[21–23]. Upon integration, disruption or complete deletion of the E1 or E2 gene is often observed, resulting in constitutive expression of the E6 and E7 oncogenes[24–26], inactivation of cell cycle checkpoints and genetic instability[23]. Viral integration may also lead to modified expression of cellular genes nearby, disruption of genes, as well as genomic amplifications that may promote oncogenesis[23,27]. The finding of certain chromosomal clusters of integration in precancerous lesions and cancers[28] also suggests a selective advantage of specific HPV integrations. Still, several important questions remain for HPV integration and more comprehensive analyses of integration sites are needed in order to expand our understanding of HPV pathogenesis.

The development of next generation sequencing (NGS) technologies has provided new tools for viral genomic research. During the recent years, a few studies have described different NGS based approaches to study HPV variability and integration in the human genome. The most common approaches used in HPV genomic analyses are based on target enrichment using highly multiplexed degenerate primers[29], enrichment by multiplex PCR using HPV16 forward primers[30], bead-based target capture[31–33], and rolling circle amplification[34] followed by NGS. These methods are however designed to detect either HPV integration or HPV variability. In addition, target capture methods poorly enrich HPV and remain expensive due to high probe cost and off-target sequencing.

In order to contribute to the understanding of the role of intra-host HPV genomic variability and chromosomal integration in carcinogenesis, we have developed an innovative library preparation strategy followed by an in-house bioinformatics pipeline named TaME-seq (tagmentation-assisted multiplex PCR enrichment sequencing). TaME-seq combines tagmentation and multiplex PCR enrichment, allowing simultaneous HPV genomic variability and integration analysis (Fig. 1). TaME-seq, with highly efficient target enrichment and reduced sequencing cost, enables deep sequencing analysis in order to find low frequency variants and rare integration events. Here, we present the results of HPV integration and genomic variability analysis in HPV16, 18, 31, 33 and 45 positive clinical samples and cell lines. The method described here provides an important tool for comprehensive studies of HPV genomic variability and chromosomal integration, and it can also be adapted to studies on other viruses such as retroviruses, adeno-associated viruses and integrating human herpesviruses.

Results

Read mapping analysis and genome coverage. Table 1 summarises liquid-based cytology (LBC) samples (n = 21), cell lines (n = 4) and plasmid samples (n = 3) included in the analysis. The samples generated 154.8 million raw reads of which 72.5 million reads (47%) mapped to the target HPV reference genomes. Only a small fraction (0.08%) of the reads mapped to other HPV types than those reported positive by HPV genotyping. The mean coverage ranged from 303 to 273898, while the fraction of the genome covered by minimum 10 × ranged from 0.35 to 1, and the fraction of the genome covered by minimum 100 × ranged from 0.33 to 1 (Table 1). HPV genome sequencing coverage aligned to the target HPV genomes with the location of HPV genomic regions and primers is visualised for CaSki, HeLa, LBC34, LBC11 and MS751 (Fig. 2). Overall, the samples showed varying HPV genome coverage profiles (Supplementary Figs S1–S5). Totally, 10 HPV positive samples were excluded from further analysis due to poor sequencing coverage (Supplementary Table S1). Sequencing of the HPV negative control samples resulted in no or negligible amount (<500) of reads mapped to target HPV genomes (Supplementary Table S2). The MS751 cell line was confirmed not to contain HPV18 sequences (Supplementary Table S1)[35].

Deletions in HPV genomes. The method enables identification of regions covered with very few or no sequencing reads, interpreted as large HPV genomic deletions. Cell lines HeLa and MS751 are known to contain partial HPV genomes due to deletions of 2.5 kb and 5 kb, respectively[35,36], which was confirmed by our method

Sample	Sample type	Raw reads	Trimmed reads	Reads mapped to target HPV	% Reads mapped to target HPV	Mean coverage	Fraction of genome covered by minimum	
							10×	100×
HPV16								
CaSki	Cell line	16138790[b]	12944262	12634651	78%	184716	1.00	1.00
SiHa	Cell line	151168[b]	133360	67496	45%	1018	0.96	0.83
SiHa-1	Cell line	5948008[c]	3735936	1249594	21%	17561	0.93	0.90
SiHa-1	Cell line	844178[b]	532874	181199	21%	2554	0.92	0.78
SiHa-2	Cell line	1405886[c]	789664	420774	30%	5609	0.91	0.85
SiHa-2	Cell line	158672[b]	90150	48412	31%	646	0.84	0.52
WHO std HPV16	Plasmid	359638[b]	304002	278987	78%	4104	0.99	0.96
LBC1[a]	LBC	128008[b]	108756	75323	59%	1124	0.96	0.88
LBC7[a]	LBC	62246[b]	51590	25567	41%	384	0.94	0.66
HPV18								
HeLa	Cell line	1433248[b]	1120824	394420	28%	5897	0.68	0.62
WHO std HPV18	Plasmid	2021206[b]	1358182	1098783	54%	15447	0.99	0.96
LBC103[a]	LBC	1477706[b]	1209564	74358	5%	1056	0.93	0.83
LBC105[a]	LBC	190664[b]	160450	32695	17%	484	0.51	0.34
LBC107	LBC	2180284[b]	1881868	978435	45%	14663	1.00	0.99
LBC108[a]	LBC	5407154[b]	3773986	3360463	62%	46691	1.00	0.98
LBC48[a]	LBC	641378[b]	433884	72589	11%	988	0.95	0.83
HPV31								
LBC16	LBC	276994[b]	191290	74465	27%	1065	0.94	0.80
LBC24[a]	LBC	471666[b]	348416	24197	5%	355	0.96	0.69
LBC32	LBC	2446832[b]	1523572	1319939	54%	18983	0.99	0.98
LBC34	LBC	3285680[b]	1841812	1723631	52%	23790	0.99	0.96
HPV33								
HPV33 plasmid	Plasmid	13824396[b]	5202718	5230090	38%	61527	1.00	1.00
LBC11	LBC	2852262[b]	1052512	986936	35%	12038	0.99	0.98
LBC30	LBC	77128[b]	51682	21431	28%	303	0.93	0.63
LBC31[a]	LBC	4276740[c]	2831408	44917	1.1%	544	0.76	0.60
LBC52	LBC	154936[b]	86990	34390	22%	439	0.95	0.62
LBC65[a]	LBC	368260[b]	248142	144022	39%	1993	1.00	0.91
HPV45								
MS751	Cell line	1221694[b]	1047286	56291	5%	845	0.35	0.33
LBC13[a]	LBC	496370[b]	389306	58293	12%	849	0.96	0.78
LBC29	LBC	211052[b]	122502	45925	22%	614	0.91	0.69
LBC36[a]	LBC	2412532[b]	1822912	1579570	65%	22093	1.00	0.97
LBC54	LBC	50169422[c]	26385910	20570184	41%	256857	1.00	1.00
LBC64[a]	LBC	5121416[c]	3040714	307476	6%	3943	0.95	0.88

Table 1. Read counts and sequencing coverage of HPV positive cell lines, plasmids and LBC samples. [a]Sample has multiple HPV infections. [b]Sequenced on MiSeq sequencing platform. [c]Sequenced on HiSeq 2500 sequencing platform.

(Fig. 2). A large deletion of 4.8 kb was revealed in the clinical sample LBC105, indicating partial or complete deletion of HPV18 genes E1, E2, E4, E5, L1 and L2 (Supplementary Fig. S2).

HPV-human integration sites. A two-step strategy was applied to detect possible integration sites (Fig. 3). A total of 27 integration sites were detected in cell lines CaSki, SiHa, HeLa and MS571 (Table 2). For CaSki, 16 previously reported integration sites[30,32,37] were confirmed. In addition, three novel sites were identified. These mapped to HPV16 E6, E2 and L1 genes. One was located in an intronic region of the gene *BRSK1*; two were located more than 50 kb from annotated genes (Table 2). Three sites, including one previously reported site as a control[30,37], were subjected to Sanger sequencing to confirm the integration sites (Supplementary Table S3). Integration sites identified in SiHa, HeLa and MS751 were consistent with previous studies[31,35–39] and were not subjected to validation by Sanger sequencing. Additionally, two integration sites were detected in the clinical sample LBC105 (Table 2). The integration breakpoints were mapped to the HPV E1 and L1 genes flanking the deleted region (Supplementary Fig. S2) and they were located in intronic regions of the gene *GTF2IRD1* (Table 2). Both integration sites were confirmed by Sanger sequencing (Supplementary Table S3).

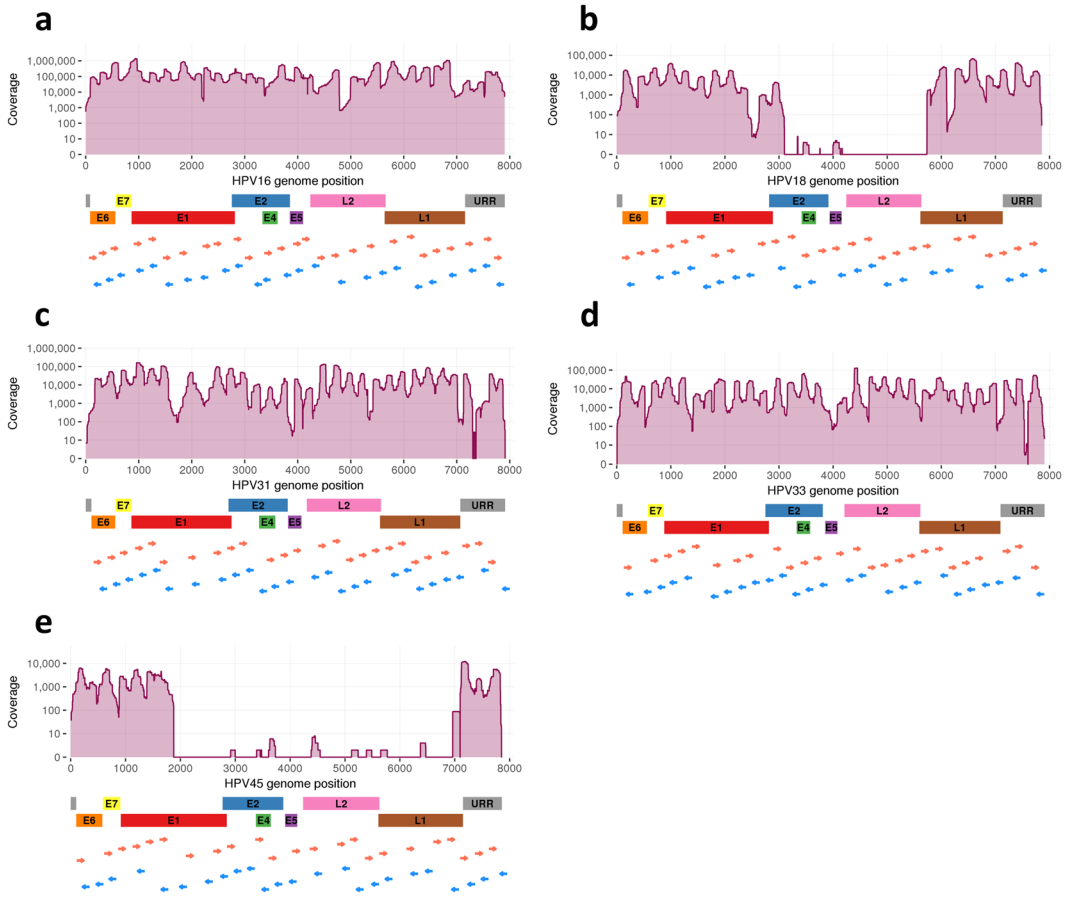


Figure 2. HPV genome sequencing coverage in HPV positive samples. The coverage plots of (**a**) CaSki, (**b**) HeLa, (**c**) LBC34, (**d**) LBC11, and (**e**) MS751 are aligned to the respective target HPV genomes. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV primers is indicated below the genomic positions.

Evaluation of variant calling using SiHa technical replicates. Sequencing libraries of the SiHa cell line served as technical replicates to assess the variant calling performance. In both SiHa-1 and SiHa-2, more variable sites were detected with higher mean coverage (Fig. 4). Number of variable sites in SiHa-1 ranged from 477 to 809 and mean coverage ranged from 2554 to 17561. Number of variable sites in SiHa-2 ranged from 257 to 522 and mean coverage ranged from 646 to 5609 (Fig. 4; Supplementary Table S4). First, reproducibility of variant calling was assessed within the same SiHa sequencing library. Concordance rate of variable sites was calculated using HiSeq 2500 result as the reference value. The concordance rates varied from 92% (HiSeq down-sampled 90%) to 45% (MiSeq) in SiHa-1 and from 89% (HiSeq downsampled 90%) to 27% (MiSeq) in SiHa-2 (Supplementary Table S4). Concordance rates of variants, including low frequency variation, between replicates (different library, same sequencing platform) were calculated to evaluate the effect of library preparation steps on the number of variable sites found in each sample. Concordance rates were 21% and 19% in SiHa-1 and SiHa-2, respectively (Supplementary Table S5).

HPV genomic variability. Variability was analysed in cell lines and LBC samples. Samples had variable sites (variant allele frequency $>0.2\%$ and coverage $\geq 100\times$) in all genes with the exception of regions that were deleted or had low sequencing coverage. The number of variable sites was normalised by the length of each HPV genomic region. Genomic regions had varying percentages of variable sites (0–28%) in each of the samples. Overall, there were samples within each HPV type that had $>15\%$ variable sites in at least one HPV gene (Fig. 5). Principally, samples with higher mean coverage had more variable sites (Supplementary Table S6), which is in line with the results from the variant analysis done on SiHa replicates (Fig. 4). CaSki had most variable sites (1017) of the cell lines and LBC54 had most variable sites (1641) of the clinical samples (Supplementary Table S6). A variant profile with variable site positions and variant allele frequency (VAF) is shown for CaSki and LBC54 (Fig. 6). Overall, the results show considerable variability in the samples throughout the HPV genome (Fig. 5, Supplementary Figs S6–S10).

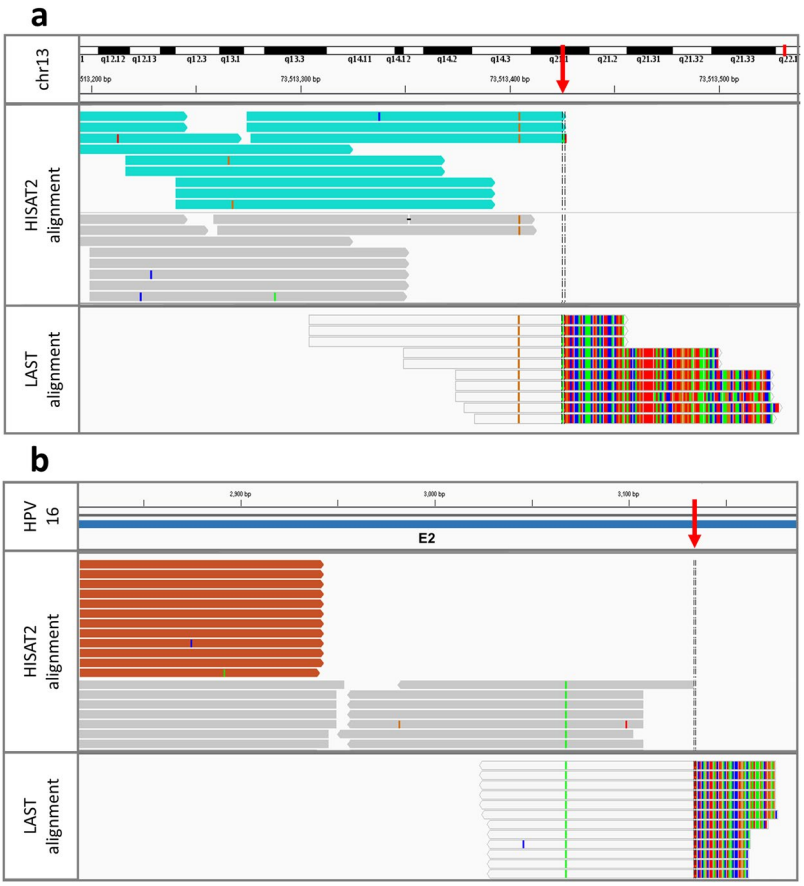


Figure 3. An IGV visualisation of HISAT2 and LAST alignments to find HPV-human integration breakpoints. All the reads were first mapped with HISAT2 and then the unmapped reads were remapped with LAST. (**a**) SiHa reads mapping to chromosome 13 (GRCh38/hg38). Light blue HISAT2 reads have pairs mapping to HPV16 reference genome. Multi-coloured parts of the LAST reads are mismatched bases that map to HPV16 (not visualised). (**b**) SiHa reads mapping to HPV16 reference genome. Orange HISAT2 reads have pairs mapping to chromosome 13 (GRCh38/hg38). Multi-coloured parts of the LAST reads are mismatched bases that map to chromosome 13 (not visualised). Red arrows point to the exact breakpoint positions.

Discussion

Here, we present a novel cost-efficient approach, TaME-seq, for the simultaneous analysis of HPV variation and chromosomal integration. Previous methods have been less effective and/or limited to either one of the two analyses[29–34]. To demonstrate the performance of TaME-seq, we employed HPV16, 18, 31, 33 and 45 positive clinical samples, HPV positive cell lines and HPV plasmids. With 47% of the total of 154.8 million raw reads mapped on the target HPV reference genomes, TaME-seq proved to be highly efficient in HPV target enrichment. Other approaches for HPV target enrichment have reported much lower HPV mapping ratios[32,40], requiring more sequencing and therefore at a higher sequencing cost. TaME-seq currently covers HPV16, 18, 31, 33 and 45, being the most common HPV genotypes in cervical cancer[5]. TaME-seq can be extended to cover additional HPV types, as well as other viruses, by implementing new primers to the method.

The ability of TaME-seq to detect chromosomal integration sites has been shown for the HPV positive cervical cancer cell lines CaSki, SiHa, HeLa and MS751. CaSki cells contain a high copy number (~600 copies/cell) of integrated full-length HPV16 arranged in concatemers[41,42]. SiHa (1–2 HPV16 copies/cell)[39,41] and HeLa (10–50 HPV18 copies/cell)[43] cells harbour integrated HPV genomes. MS751 cells contains integrated HPV45[35], but in contrast to the product specification sheet (ATCC, Manassas, VA) no HPV18, which was verified in our analyses. For CaSki, 16 previously reported integration sites[30,32,37] were detected by our method. In addition, three novel integration sites were identified. Known integration sites in SiHa[31,37,39], HeLa[31,36] and MS751[35], as well as large deletions demonstrated in HeLa[36] and MS751[35], were confirmed by the TaME-seq method. Of the 21 LBC samples, HPV integration sites could only be detected in one sample, being in line with previous studies reporting no or few HPV integration events in LSIL/ASC-US samples[44,45]. However, other studies report integration events also in LSIL samples[32,46]. The detection of integrated forms of the virus is also dependent on the amount of episomes in the sample; low copy integration sites may remain undetected against a high background of episomal HPV.

The high sequencing coverage throughout the HPV genome enables detection of low frequency variants. Variant calling was evaluated using SiHa replicates to set the variant calling threshold. Previous studies have used variant calling thresholds of 0.5% or 1%[17,34]. With the high coverage provided by the TaME-seq method there is

Sample	HPV		Human (GRCh38/hg38)		# Unique discordant read pairs	# Unique junction reads
	Breakpoint	ORF	Chromosomal locus	Breakpoint		
HPV16						
CaSki	273	E6	20p11.1	chr20:26276796	19	0[e]
	494[a]	E6	20p11.1	chr20:26341342[b]	7	0[e]
	582	E7	19q13.42	chr19:55310208	0	15
	975	E1	Xq27.3	chrX:145696778	0	7
	1398	E1	2p23.3	chr2:27135968	6	0[e]
	1793	E1	10p14	chr10:11700197	4	0[e]
	2987	E2	Xq27.3	chrX:145708231	3	8
	3239	E2	7p22.1	chr7:6925283	5	0[e]
	3631[a]	E2	19q13.42	chr19:55310043[c]	3	0[e]
	3729	E2	6p21.1	chr6:45691388	0	11
	4654	L2	11p15.4	chr11:6741077	11	0[e]
	5432	L2	11q22.1	chr11:100766632	2	0[e]
	5698	L1	10p14	chr10:11700617	20	0[e]
	5698	L1	5p11	chr5:46292081	2	0[e]
	5762	L1	11q22.1	chr11:100771699	4	0[e]
	6572	L1	19q13.42	chr19:55307445	3	0[e]
	7123[a]	L1	20p11.1	chr20:26357640[b]	20	0[e]
	7733	URR	11p15.4	chr11:6740842	2	0[e]
	7733	URR	2p23.3	chr2:27137265	6	0[e]
SiHa	3133	E2	13q22.1	chr13:73513425	7	7
	3385	E2/E4	13q22.1	chr13:73214729	3	0[e]
HPV18						
HeLa	2066	E1	8q24.21	chr8:127229053	2	0[e]
	2887	E2	8q24.21	chr8:127221122	13	0[e]
	5730	L1	8q24.21	chr8:127218384	11	89
	7655	URR	8q24.21	chr8:127221804	3	0[e]
LBC105	1561	E1	7q11.23	chr7:74525628[d]	0	10
LBC105	6528	L1	7q11.23	chr7:74515883[d]	2	0[e]
HPV45						
MS751	1646	E1	18q11.2	chr18:23024744	10	0[e]
	7120	L1	18q11.2	chr18:23021388	15	0[e]

Table 2. Chromosomal integration sites detected by TaME-seq. [a]Novel breakpoint in CaSki cell line. [b]No annotated genes within 50 kb from the breakpoint. [c]Intronic region in gene *BRSK1*. [d]Intronic region in gene *GTF2IRD1*. [e]When number of unique junction reads is 0, the breakpoint coordinates are not exact.

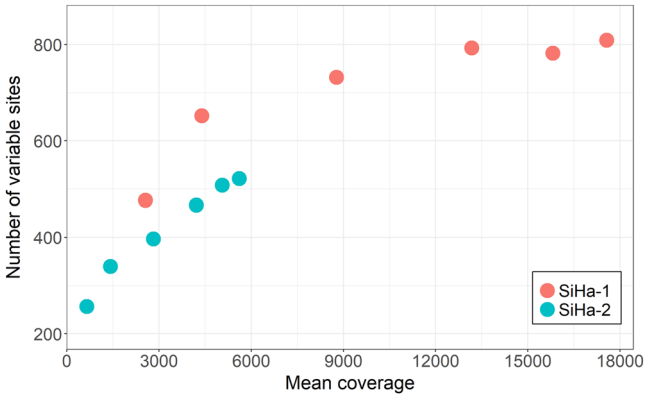


Figure 4. Number of variable sites in SiHa replicates. SiHa-1 (red dots) and SiHa-2 (blue dots) served as technical replicates to assess the variant calling performance. In SiHa libraries, sequenced on MiSeq and HiSeq 2500 platforms, increasing number of variable sites were detected with higher mean coverage.

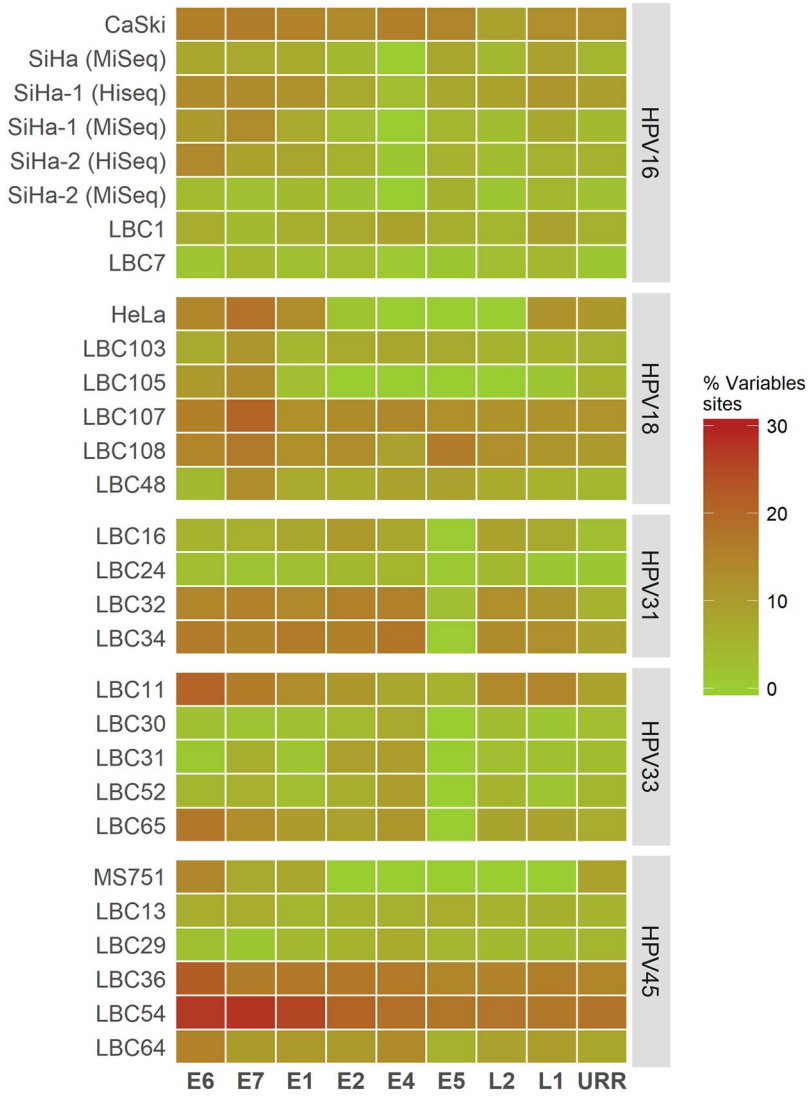


Figure 5. Proportion of variable sites in HPV genes in HPV positive samples. The number of variable sites was normalised by the length of each HPV gene. Gradient green (0% variable sites) to red (30% variable sites) color-coding of the results is shown to present the considerable variability in the samples throughout the HPV genome.

potential for detecting very low frequency variation. We have therefore analysed the variation using 0.2% as the variant calling threshold. Multiple and stringent filtering steps was included to filter out non-reliable variants, as we are approaching the inherent error rate profile of the PCR amplification and Illumina sequencing[47]. However, the threshold for variant calling is dependent on experimental and analytical basis and must be set according to the study aims.

The results from the SiHa analysis indicate that calling ultra-low frequency variants is dependent on the sequencing coverage. Lower sequencing coverage results in the detection of fewer variants and less concordance between sample replicates. In order to find ultra-low frequency variants, high sequencing coverage is required. Figure 4 shows that at the mean coverage of 12000×, the number of variants in SiHa-1 is approaching saturation. This indicates that more variants are not likely to be found even with higher sequencing coverage. Finally, differences in sequencing coverage affect the number of variable sites found, but also experimental approaches due to stochastic sampling and variant calling can fail to reveal low frequency variants. Overall, our results uncover low frequency variants in the samples, potentially introduced by DNA repair mechanisms and APOBEC enzyme mediated DNA editing[48–50], although some bias may be introduced by PCR and sequencing. Variable sites are present in all genes of the studied HPV types. Traditionally, studies have focused on sequence variation on a viral sublineage level[13–16] or the high variability has been interpreted as HPV variant co-infections[29]. The development of NGS technologies has provided comprehensive tools for the study of HPV genomic variability. Recent studies have reported high HPV variability that may be evidence of intra-host viral evolution and adaptation generated during a chronic HPV infection[17–20].

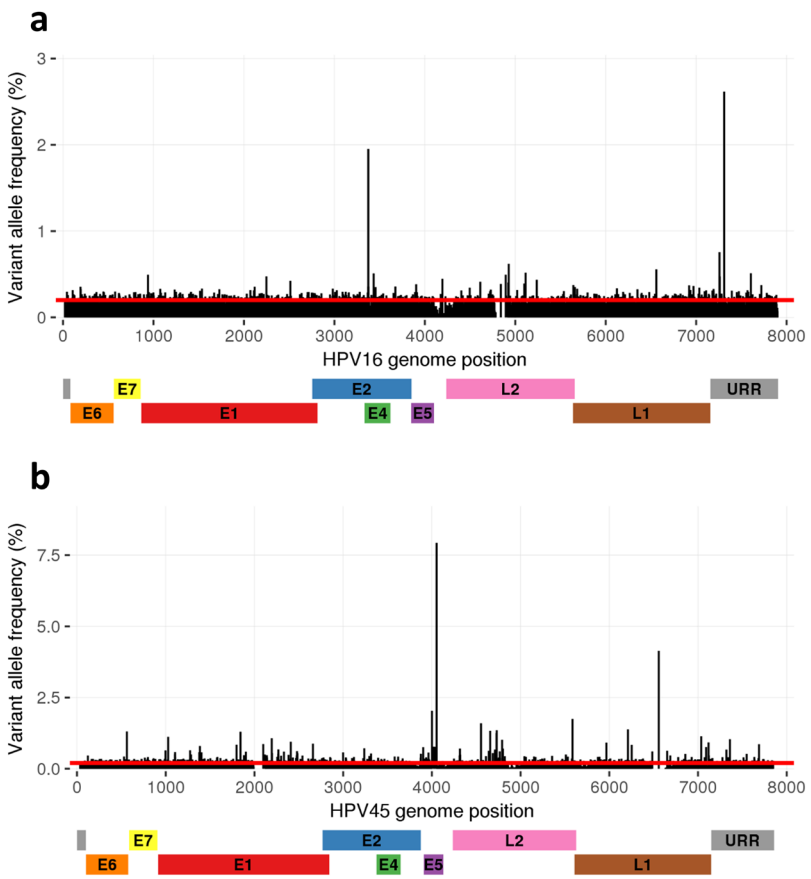


Figure 6. HPV nucleotide variation observed in two samples. The plots showing variable sites and variant allele frequency (%) in (**a**) CaSki, and (**b**) LBC54 are aligned to the respective target HPV genomes. The location of genes and URR is indicated below the genomic positions. The red line indicates the variant calling threshold value of 0.2%.

Our study has some limitations. Firstly, TaME-seq is not intended for determining HPV genotypes and we recommend it for analyses of HPV variability and integration events in samples with known HPV status. Secondly, due to variation in amplification efficacy, an uneven coverage is seen for different genomic regions. Sudden drops in the coverage, that are not genomic deletions, may be due to suboptimal primer performance or poor alignment against the reference genomes. This issue can be solved partly by designing new primers covering these regions and optimising the primer performance. Also, the read alignment step can be further optimised. Alternatively, alignment could be performed by *de novo* assembly to create consensus sequences for the alignment. Thirdly, enough viral DNA and good dsDNA quality are important for achieving consistent tagmentation results in the Nextera protocol[51]. Sample preparation of the excluded LBC samples failed likely due to very low viral load in the samples, which was not quantified separately.

In summary, we have developed a NGS approach that allows the simultaneous study of HPV genomic variability and chromosomal integration. TaME-seq is applicable to large sample cohorts due to its highly efficient target enrichment, leading to less off-target sequences and therefore reduced sequencing cost. Comprehensive studies on HPV intra-host variability generated during a persistent infection will improve our understanding of viral carcinogenesis. Efficient identification of HPV genomic variability and integration sites will be important both for the study of HPV evolution, adaptability and may be a useful tool for cervical cancer diagnostics.

Methods

Samples. Anonymised LBC samples from routine cervical cancer screening were included in the study, comprising cases of atypical squamous cells of undetermined significance (ASC-US) and low-grade squamous intraepithelial lesions (LSIL). HPV positive samples with the cobas 4800 HPV test (Roche Molecular Diagnostics, Pleasanton, CA) were extracted for DNA using the automated system NucliSENS easyMAG (BioMerieux Inc., France) with off-board lysis. The samples were HPV genotyped using the modified GP5+/6+ PCR protocol (MGP)[52], followed by HPV type-specific hybridisation using Luminex suspension array technology[53] or the Anyplex™ II HPV28 assay (Seegene, Inc., Seoul, Korea). LBC samples (n = 31) were positive for HPV16, 18, 31, 33 or 45 alone, or had multiple infections including at least one of the five types. DNA extracted from the HPV positive cervical carcinoma cell lines CaSki, SiHa, HeLa and MS751 (ATCC, Manassas, VA) served as positive controls. WHO international standards for HPV 16 (1st WHO International Standard for Human Papillomavirus Type 16 DNA, NIBSC code: 06/202) and 18 (1st WHO International Standard for Human Papillomavirus Type 18

DNA, NIBSC code: 06/206)(NIBSC, Potters Bar, Hertfordshire, UK) and a plasmid containing the strain HPV33[54] were used as additional positive controls. Laboratory-grade water and DNA from an HPV negative human sample were included as negative controls. DNA was quantified by the fluorescence-based Qubit dsDNA HS assay (Thermo Fisher Scientific Inc.,Waltham, MA, USA).

Primer design. HPV16, 18, 31, 33, and 45 whole genome reference and variant sequences were obtained from the PapillomaVirus Episteme (PaVE) database[55]. All the available reference and variant sequences within an HPV type were aligned using the multiple sequence alignment tool ClustalO[56]. The sequence alignment was converted to a consensus sequence for each HPV type in CLC Sequence viewer version 7.7.1 (QIAGEN Aarhus A/S). TaME-seq HPV primers were designed using Primer3[57] and HPV consensus sequences as the source sequence. Finally, primers were modified by adding an Illumina TruSeq-compatible adapter tail (5′-AGACGTGTGCTCTTCCGATCT-3′) to the 5′-end and then synthesised by Thermo Fisher Scientific, Inc. (Waltham, MA).

Library preparation and sequencing. Primer pools for each HPV type were prepared by combining primers separately in equal volumes. Samples were subjected to tagmentation using Nextera DNA library prep kit (Illumina, Inc., San Diego, CA). Tagmented DNA was purified using DNA Clean & Concentrator™-5 columns (Zymo Research, Irvine, CA) according the manufacturer's instructions or ZR-96 DNA Clean & Concentrator™-5 plates (Zymo Research, Irvine, CA) according to the Nextera® DNA Library Prep Reference Guide (15027987 v01) before PCR amplification for target enrichment. Amplification was performed using Qiagen Multiplex PCR Master mix (Qiagen, Hilden, Germany) according to the manufacturer's instructions. For each sample, two PCR reactions were performed separately with 0.75 µM of HPV primer pools, 0.5 µM of i7 index primers (adapted from Kozich *et al.*[58]) and 1 µl of i5 index primers from the Nextera index kit (Illumina, Inc., San Diego, CA). The cycling conditions were as follows: initial denaturation and hot start at 95 °C for 5 minutes; 30 cycles at 95 °C for 30 seconds, at 58 °C for 90 seconds and at 72 °C for 20 seconds; final extension at 68 °C for 10 minutes. Following amplification, libraries were pooled in equal volumes and the final sample pool was purified with Agencourt® AMPure® XP beads (Beckman Coulter, Brea, CA). The quality and quantity of the pooled libraries were assessed on Agilent 2100 Bioanalyzer using Agilent High Sensitivity DNA Kit (Agilent Technologies Inc., Santa Clara, CA) and by qPCR using KAPA DNA library quantification kit (Kapa Biosystems, Wilmington, MA). Sequencing was performed on the MiSeq platform (Illumina, Inc., San Diego, CA) or on the HiSeq 2500 platform (Illumina, Inc., San Diego, CA). Samples were sequenced as 151 bp paired-end reads and two 8 bp index reads.

Sequence alignment. Raw paired-end reads were trimmed for adapters, HPV primers, quality (-q 20) and finally for minimum length (-m 50) using cutadapt (v1.10)[59]. Trimmed reads were mapped to human (GRCh38/hg38) and HPV16, 18, 31, 33 and 45 reference genomes obtained from the PaVE database[55] using HISAT2 (v2.1.0)[60]. Mapping statistics and sequencing coverage were calculated using the Pysam package[61] with an in-house Python (v3.5.4) script. Downstream analysis was performed using an in-house R (v3.4.4) script. Results from both reactions of the same sample were combined and method performance was then evaluated based on the percentage of obtained reads mapped to the HPV reference genome, mean sequencing coverage and percentage of HPV reference genome coverage for each sample. Further analysis was performed when a sample had >20000 reads mapped to the target HPV reference genome. The target HPV genomes correspond to the HPV types for which the samples were reported positive by HPV genotyping.

Detecting HPV-human integration sites. The paired-end reads that mapped (HISAT2) with one end to a human chromosome and the other end to the target HPV reference genome were identified as discordant read pairs. If a specific position had ≥2 read pairs with unique start or end coordinates, it was considered as a potential integration site. To determine the exact position of HPV-human integration breakpoints, previously unmapped reads were remapped to human and HPV reference genomes (as above) using the LAST (v876) aligner (options -M -C2)[62]. Positions covered by ≥3 junction reads, with unique start or end coordinates, were considered as potential integration breakpoints. Integration site detection was not based on reads sharing the same start and end coordinates as these reads were considered as potential PCR duplicates. Selected HPV integration breakpoints were confirmed by PCR amplification and Sanger sequencing.

Sequence variation analysis. Mapped nucleotide counts over HPV reference genomes and average mapping quality values of each nucleotide were retrieved from BAM files and variant calling was performed using an in-house R script. To reduce the effects of PCR amplification and sequencing artefacts in the variation analysis, filtering was applied before the variant calling. Nucleotides seen ≤2 times in each position and nucleotides with mean Phred quality score of <20 were filtered out. Nucleotide counts from both reactions of the same sample were combined and variant allele frequencies (VAF) of the three minor alleles in each position were calculated. If results from either of the reaction showed >5 times larger VAF with <20% of the total coverage, it was discarded from variant calling. Finally, variants were called if VAF was >0.2% and coverage was ≥100×.

Two sequencing libraries of SiHa cell line served as technical replicates to assess the variant calling performance. The technical replicates were sequenced on the MiSeq platform or on the HiSeq 2500 platform. In addition, HiSeq raw sequencing data was downsampled randomly and defined portions (90%, 75%, 50% and 25%) of the original reads were further analysed. Reproducibility of calling variants in the replicates was assessed by calculating concordance rate. The concordance rate (R_c) between duplicates was defined as follows:

$$R_c = \frac{N_c}{mean(N_1, N_2)}$$

where N_c was the number of concordant variants between a pair of replicate samples, and N_1 and N_2 were the total number of variants detected in each of the duplicated sample.

Ethical approval. This study was approved by the regional committee for medical and health research ethics, Oslo, Norway [2017/447] and we confirm that all experiments were performed in accordance with the committee's guidelines and regulations.

Data Availability

Sequence data from cell lines will be available at European Nucleotide Archive (ENA) accession number ERP111061. Plasmids are third party property and requests must be made to International Human Papillomavirus Reference Center and Institut Pasteur. Sequencing data from clinical samples will be available from the authors upon request with obtained ethical approval. Clinical sequence data may be deposited at the European Genome-phenome Archive (EGA) (ethical and legal assessments are on-going).

References

1. Walboomers, J. M. *et al*. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **189**, 12–19, 10.1002/(sici)1096-9896(199909)189:1<12::aid-path431>3.0.co;2-f (1999).
2. Ferlay, J. *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–386, https://doi.org/10.1002/ijc.29210 (2015).
3. Fitzmaurice, C. *et al*. The Global Burden of Cancer 2013. *JAMA Oncol* **1**, 505–527, https://doi.org/10.1001/jamaoncol.2015.0735 (2015).
4. Bosch, F. X., Lorincz, A., Munoz, N., Meijer, C. J. & Shah, K. V. The causal relation between human papillomavirus and cervical cancer. *J. Clin. Pathol.* **55**, 244–265 (2002).
5. de Sanjose, S. *et al*. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *The Lancet Oncology* **11**, 1048–1056, https://doi.org/10.1016/s1470-2045(10)70230-8 (2010).
6. Crosbie, E. J., Einstein, M. H., Franceschi, S. & Kitchener, H. C. Human papillomavirus and cervical cancer. *The Lancet* **382**, 889–899, https://doi.org/10.1016/s0140-6736(13)60022-7 (2013).
7. Forman, D. *et al*. Global burden of human papillomavirus and related diseases. *Vaccine* **30**(Suppl 5), F12–23, https://doi.org/10.1016/j.vaccine.2012.07.055 (2012).
8. Moscicki, A. B. *et al*. Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine* **30**(Suppl 5), F24–33, https://doi.org/10.1016/j.vaccine.2012.05.089 (2012).
9. Bernard, H. U. Taxonomy and phylogeny of papillomaviruses: an overview and recent developments. *Infect. Genet. Evol.* **18**, 357–361, https://doi.org/10.1016/j.meegid.2013.03.011 (2013).
10. Bzhalava, D., Eklund, C. & Dillner, J. International standardization and classification of human papillomavirus types. *Virology* **476**, 341–344, https://doi.org/10.1016/j.virol.2014.12.028 (2015).
11. Bernard, H. U. *et al*. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology* **401**, 70–79, https://doi.org/10.1016/j.virol.2010.02.002 (2010).
12. Burk, R. D., Harari, A. & Chen, Z. Human papillomavirus genome variants. *Virology* **445**, 232–243, https://doi.org/10.1016/j.virol.2013.07.018 (2013).
13. Cornet, I. *et al*. HPV16 genetic variation and the development of cervical cancer worldwide. *Br. J. Cancer* **108**, 240–244, https://doi.org/10.1038/bjc.2012.508 (2013).
14. Mirabello, L. *et al*. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. *J. Natl. Cancer Inst.* **108**, https://doi.org/10.1093/jnci/djw100 (2016).
15. Chan, P. K. *et al*. Geographical distribution and oncogenic risk association of human papillomavirus type 58 E6 and E7 sequence variations. *Int. J. Cancer* **132**, 2528–2536, https://doi.org/10.1002/ijc.27932 (2013).
16. Chen, A. A., Gheit, T., Franceschi, S., Tommasino, M. & Clifford, G. M. Human Papillomavirus 18 Genetic Variation and Cervical Cancer Risk Worldwide. *J. Virol.* **89**, 10680–10687, https://doi.org/10.1128/jvi.01747-15 (2015).
17. de Oliveira, C. M. *et al*. High-level of viral genomic diversity in cervical cancers: A Brazilian study on human papillomavirus type 16. *Infect. Genet. Evol.* **34**, 44–51, https://doi.org/10.1016/j.meegid.2015.07.002 (2015).
18. Mirabello, L. *et al*. HPV16 E7 Genetic Conservation Is Critical to Carcinogenesis. *Cell* **170**, 1164–1174 e1166, https://doi.org/10.1016/j.cell.2017.08.001 (2017).
19. Hirose, Y. *et al*. Within-Host Variations of Human Papillomavirus Reveal APOBEC-Signature Mutagenesis in the Viral Genome. *J. Virol.* https://doi.org/10.1128/jvi.00017-18 (2018).
20. Dube Mandishora, R. S. *et al*. Intra-host sequence variability in human papillomavirus. *Papillomavirus Res*, https://doi.org/10.1016/j.pvr.2018.04.006 (2018).
21. Zur Hausen, H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat. Rev. Cancer* **2**, 342–350, https://doi.org/10.1038/nrc798 (2002).
22. Pett, M. & Coleman, N. Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J. Pathol.* **212**, 356–367, https://doi.org/10.1002/path.2192 (2007).
23. McBride, A. A. & Warburton, A. The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog.* **13**, e1006211, https://doi.org/10.1371/journal.ppat.1006211 (2017).
24. Jeon, S., Allen-Hoffmann, B. L. & Lambert, P. F. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J. Virol.* **69**, 2989–2997 (1995).
25. Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. & Murakami, I. Human papillomavirus molecular biology and disease association. *Rev. Med. Virol.* **25**(Suppl 1), 2–23, https://doi.org/10.1002/rmv.1822 (2015).
26. Ziegert, C. *et al*. A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene* **22**, 3977–3984, https://doi.org/10.1038/sj.onc.1206629 (2003).
27. Peter, M. *et al*. Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J. Pathol.* **221**, 320–330, https://doi.org/10.1002/path.2713 (2010).
28. Kraus, I. *et al*. The Majority of Viral-Cellular Fusion Transcripts in Cervical Carcinomas Cotranscribe Cellular Sequences of Known or Predicted Genes. *Cancer Res.* **68**, 2514–2522, https://doi.org/10.1158/0008-5472.Can-07-2776 (2008).
29. Cullen, M. *et al*. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res* **1**, 3–11, https://doi.org/10.1016/j.pvr.2015.05.004 (2015).
30. Xu, B. *et al*. Multiplex Identification of Human Papillomavirus 16 DNA Integration Sites in Cervical Carcinomas. *PLoS One* **8**, e66693, https://doi.org/10.1371/journal.pone.0066693 (2013).
31. Liu, Y., Lu, Z., Xu, R. & Ke, Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget* **7**, 5852–5864, https://doi.org/10.18632/oncotarget.6809 (2016).

32. Hu, Z. *et al*. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat. Genet.* **47**, 158–163, https://doi.org/10.1038/ng.3178 (2015).
33. Holmes, A. *et al*. Mechanistic signatures of HPV insertions in cervical carcinomas. *npj Genomic Medicine* **1**, https://doi.org/10.1038/npjgenmed.2016.4 (2016).
34. Kukimoto, I. *et al*. Genetic variation of human papillomavirus type 16 in individual clinical specimens revealed by deep sequencing. *PLoS One* **8**, e80583, https://doi.org/10.1371/journal.pone.0080583 (2013).
35. Geisbill, J., Osmers, U. & Durst, M. Detection and characterization of human papillomavirus type 45 DNA in the cervical carcinoma cell line MS751. *J. Gen. Virol.* **78**(Pt 3), 655–658, https://doi.org/10.1099/0022-1317-78-3-655 (1997).
36. Adey, A. *et al*. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211, https://doi.org/10.1038/nature12064 (2013).
37. Akagi, K. *et al*. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* **24**, 185–199, https://doi.org/10.1101/gr.164806.113 (2014).
38. Mincheva, A., Gissmann, L. & zur Hausen, H. Chromosomal integration sites of human papillomavirus DNA in three cervical cancer cell lines mapped by *in situ* hybridization. *Med. Microbiol. Immunol.* **176**, 245–256 (1987).
39. 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).
40. Li, T. *et al*. Universal Human Papillomavirus Typing Assay: Whole-Genome Sequencing following Target Enrichment. *J. Clin. Microbiol.* **55**, 811–823, https://doi.org/10.1128/JCM.02132-16 (2017).
41. Baker, C. C. *et al*. Structural and transcriptional analysis of human papillomavirus type 16 sequences in cervical carcinoma cell lines. *J. Virol.* **61**, 962–971 (1987).
42. Yee, C., Krishnan-Hewlett, I., Baker, C. C., Schlegel, R. & Howley, P. M. Presence and expression of human papillomavirus sequences in human cervical carcinoma cell lines. *Am. J. Pathol.* **119**, 361–366 (1985).
43. Meissner, J. D. Nucleotide sequences and further characterization of human papillomavirus DNA present in the CaSki, SiHa and HeLa cervical carcinoma cell lines. *J. Gen. Virol.* **80**(Pt 7), 1725–1733, https://doi.org/10.1099/0022-1317-80-7-1725 (1999).
44. 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, https://doi.org/10.1016/j.ygyno.2003.11.035 (2004).
45. Liu, Y. *et al*. Genome-wide profiling of the human papillomavirus DNA integration in cervical intraepithelial neoplasia and normal cervical epithelium by HPV capture technology. *Sci. Rep.* **6**, 35427, https://doi.org/10.1038/srep35427 (2016).
46. Li, H. *et al*. Preferential sites for the integration and disruption of human papillomavirus 16 in cervical lesions. *J. Clin. Virol.* **56**, 342–347, https://doi.org/10.1016/j.jcv.2012.12.014 (2013).
47. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 125, https://doi.org/10.1186/s12859-016-0976-y (2016).
48. Warren, C. J. *et al*. APOBEC3A functions as a restriction factor of human papillomavirus. *J. Virol.* **89**, 688–702, https://doi.org/10.1128/JVI.02383-14 (2015).
49. Kukimoto, I. *et al*. Hypermutation in the E2 gene of human papillomavirus type 16 in cervical intraepithelial neoplasia. *J. Med. Virol.* **87**, 1754–1760, https://doi.org/10.1002/jmv.24215 (2015).
50. Chen, J. & Furano, A. V. Breaking bad: The mutagenic effect of DNA repair. *DNA Repair (Amst)* **32**, 43–51, https://doi.org/10.1016/j.dnarep.2015.04.012 (2015).
51. Lamble, S. *et al*. Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* **13**, 104, https://doi.org/10.1186/1472-6750-13-104 (2013).
52. Soderlund-Strand, A., Carlson, J. & Dillner, J. Modified general primer PCR system for sensitive detection of multiple types of oncogenic human papillomavirus. *J. Clin. Microbiol.* **47**, 541–546, https://doi.org/10.1128/JCM.02007-08 (2009).
53. Schmitt, M. *et al*. Bead-based multiplex genotyping of human papillomaviruses. *J. Clin. Microbiol.* **44**, 504–512, https://doi.org/10.1128/JCM.44.2.504-512.2006 (2006).
54. Beaudenon, S. *et al*. A novel type of human papillomavirus associated with genital neoplasias. *Nature* **321**, 246–249, https://doi.org/10.1038/321246a0 (1986).
55. Van Doorslaer, K. *et al*. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.* **41**, D571–578, https://doi.org/10.1093/nar/gks984 (2013).
56. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).
57. Untergasser, A. *et al*. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115, https://doi.org/10.1093/nar/gks596 (2012).
58. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120, https://doi.org/10.1128/AEM.01043-13 (2013).
59. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **2011** 17, https://doi.org/10.14806/ej.17.1.200 (2011).
60. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
61. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
62. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493, https://doi.org/10.1101/gr.113985.110 (2011).

Acknowledgements

We thank Mona Hansen and Hanne Kristiansen-Haugland for DNA sample extraction and HPV genotyping, and Tobias Neidel for primer design for HPV31, 33 and 45. This work was funded by a grant from South-Eastern Norway Regional Health Authority (project number 2016020).

Author Contributions

S.L. designed primers, performed the experiments, analysed the results and drafted the manuscript text. S.U.U. contributed to the data analysis. M.L. and P.E. performed the pilot experiments and P.E. designed the initial TaME-seq assay concept. R.M. contributed to the primer design process and designed primers. I.K.C. and O.H.A. contributed to study design and result interpretation. T.B.R. contributed to the study design, data analysis and result interpretation. All authors contributed to writing, reading and approving the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36669-6.

Competing Interests: S.L., M.L., P.E., R.M., I.K.C., O.H.A. and T.B.R. and their corresponding institutions have filed a patent application at the technology transfer company Inven2, Oslo, Norway on the protocol described here.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

TaME-seq: An efficient sequencing approach for characterisation of HPV genomic variability and chromosomal integration

Sonja Lagström[1,2], Sinan Uğur Umu[2], Maija Lepistö[3], Pekka Ellonen[3], Roger Meisal[1], Irene Kraus Christiansen[1,4], Ole Herman Ambur[5], Trine B. Rounge[2,*]

Author affiliations:
[1]Department of Microbiology and Infection Control, Akershus University Hospital, Lørenskog, Norway
[2]Department of Research, Cancer Registry of Norway, Oslo, Norway
[3]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
[4]Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital and University of Oslo, Lørenskog, Norway
[5]Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway

*Corresponding author:
E-mail: trine.rounge@kreftregisteret.no

Supplementary Table S1. Read counts and sequencing coverage of HPV positive samples that were excluded from the analysis.

Sample	Sample type	Raw reads	Trimmed reads	Reads mapped to target HPV	% Reads mapped to target HPV	Mean coverage	Fraction of genome covered by minimum	
							10×	100×
HPV16								
LBC43[a]	LBC	47788[c]	34778	15283	32%	205	0.86	0.44
HPV18								
MS751[a]	Cell line	890142[b]	837366	220	0.0%	3	0.08	0.00
LBC110[a]	LBC	1673786[b]	1464308	7958	0.5%	118	0.82	0.36
LBC10[a]	LBC	144822[b]	107538	1133	0.8%	17	0.28	0.04
LBC18[a]	LBC	700120[b]	107538	160	0.0%	2	0.05	0.00
LBC41[a]	LBC	2839890[c]	1996944	16982	0.6%	212	0.62	0.40
LBC56	LBC	508874[b]	406150	647	0.1%	9	0.32	0.00
HPV31								
LBC8	LBC	120100[b]	84434	3220	0.4%	47	0.60	0.14
LBC17[a]	LBC	330244[b]	228388	712	0.2%	10	0.18	0.03
LBC18[a]	LBC	214800[b]	163400	439	0.2%	6	0.17	0.00
HPV45								
LBC40	LBC	205342[b]	166784	107	0.1%	2	0.05	0.00

[a] Sample has multiple HPV infections.

[b] Sequenced on MiSeq sequencing platform.

[c] Sequenced on HiSeq 2500 sequencing platform.

Supplementary Table S2. Read counts and sequencing coverage of HPV negative control samples.

Sample	Raw reads	Trimmed reads	Reads mapped to target HPV	% Reads mapped to target HPV	Mean coverage	Fraction of genome covered by minimum	
						10×	100×
HPV16							
H_2O	1060[b]	794	482	45%	7	0.24	0.00
Human	38710[b]	33928	54	0.1%	1	0.00	0.00
HPV18							
H_2O	214[b]	146	0	0.0%	0	0.00	0.00
Human	496112[b]	412056	51	0.0%	1	0.00	0.00
HPV31							
H_2O	810[b]	594	0	0.0%	0	0.00	0.00
Human	340858[b]	285822	0	0.0%	0	0.00	0.00
HPV33							
H_2O	4828[b]	3406	297	6.2%	4	0.14	0.00
Human	3010522[c]	1707226	22	0.0%	0	0.00	0.00
HPV45							
H_2O	178[b]	144	16	9.0%	0	0.00	0.00
Human	1237502[b]	1075344	72	0.0%	1	0.01	0.00

[b] Sequenced on MiSeq sequencing platform.

[c] Sequenced on HiSeq 2500 sequencing platform.

Supplementary Table S3. Integration breakpoints confirmed by PCR amplification and Sanger sequencing.

Sample	Junction sequence[a]	HPV				Human		
		Type	Start	End	ORF	Chromosomal locus	Start	End
CaSki[b]	GCCAAATATATATATATATAC ACACACACATATATATGTATA CTATATACTATAGTATATACA GTATATATAGTATATATGTAA ACTATAGCCAAATATATATAT AGCCATTAGTTGCAGTTCAAT TGCTTGTAATGCTTTATTCTTT GATACAGCCAGCGTTGGCACC ACCTGGTGGTTAATATGTTTA AATCCCATTTCTCTGGCCTTGT AATAAATAGCACATTCTAGGC GCATGTGTTTCCAATAG	HPV16	2987	2848	E2	Xq27.3	145708341	145708231
CaSki	TAATATAAGGGGTCGGTGGAC CGGTCGATGTATGTCTTGTTG CAGATCATCAAGAACACGTAA AGAAACCCAGCTGTAATCATG CATGGAGATACACCTACATTG CATGAATATATGTGCCACATT TTCTTAATCCAGTCTATCATTG TTGGACATTTGGGTTGGTTCC	HPV16	481	598	E6	20p11.1	26341312	26341369
CaSki	TTCCCTTTCAGAGAGCACGTT TAAAACACCCTTTTTGTAGTA TCTGGAAGTGCACATTTGGAG GGCTTTGATGCATATGGTGAA AAAGGAAATGACTCATATGAT ACAACTGCTAAACGCAAAAA ACGTAAGCTGTAAGTATTGTA TGTATGTTGAATTAGTGTTGTT TGTTGTTTATATGTTTGTATGT GCTTGTATGTGCT	HPV16	7123	7221	L1	20p11.1	26357830	26357922
LBC105	TACAAGTGACAATAGCAATAT AGAAAATGTAAATCCACAATG TACCATAGCACAATTAAAAGA CTTGTTAAAAGTAAACAATAA ACAAGGAGCTATGTTAGCATG CCACCATGCCACGCCCAGTTA ATTTTTGTATTTTTGTAGAGAC GGGGTTTCACCATGTTGGCCA GGCTGGTCTCGAACTCTCGAG CTCAAGTGATTCACCTGCCTC GGCCTCCCAAAGTGCA	HPV18	1459	1561	E1	7q11.23	74525628	74525503
LBC105	CAAGCCTGGGGCTATTTCTAG GCGAGAGGTGGCAGTGACTTG AGCCAGGGCAGGGACAGTAG GGGTGGAGGTTTGGAAAACG GCTGGAGCAGAACTTCTGTGT CACTGTGAGGTACCATTGGAT ATTTGTCAGTCTATTTGTAAAT ATCCTGATTATTTACAAATGT CTGCAGATCCTTATGGGGATT CCATGTTTTTTTGCTTACGGCG TGAGCAGCTTTTTGCTAGGCA TTTTTGGAATAGGGCAGGT	HPV18	6264	6407	L1	7q11.23	74515867	74515764

[a] Black letters: human sequence (aligned to GRCh38/hg38); red letters: HPV sequence (aligned to HPV reference sequences obtained from the PaVE database); green letters: nucleotides shared between HPV and human genomes; blue letters: nucleotides that did not align.

[b] Previously reported integration site that was used as a control.

Supplementary Table S4. Reproducibility of variant calling was assessed within same SiHa sequencing libraries. Concordance rate of variable sites was calculated using HiSeq 2500 results as reference.

Sample	Sequence platform (downsampled)	Mean coverage	Variable sites	% Concordance
SiHa-1	HiSeq 2500	17561	809	-
	HiSeq 2500 (90%)	15809	782	92
	HiSeq 2500 (75%)	13176	793	84
	HiSeq 2500 (50%)	8773	732	73
	HiSeq 2500 (25%)	4386	652	61
	MiSeq	2554	477	45
SiHa-2	HiSeq 2500	5609	522	-
	HiSeq 2500 (90%)	5057	508	89
	HiSeq 2500 (75%)	4212	467	78
	HiSeq 2500 (50%)	2811	397	64
	HiSeq 2500 (25%)	1415	340	48
	MiSeq	646	257	27

Supplementary Table S5. Concordance rate of variable sites in two technical replicates of SiHa sequenced on a same platform.

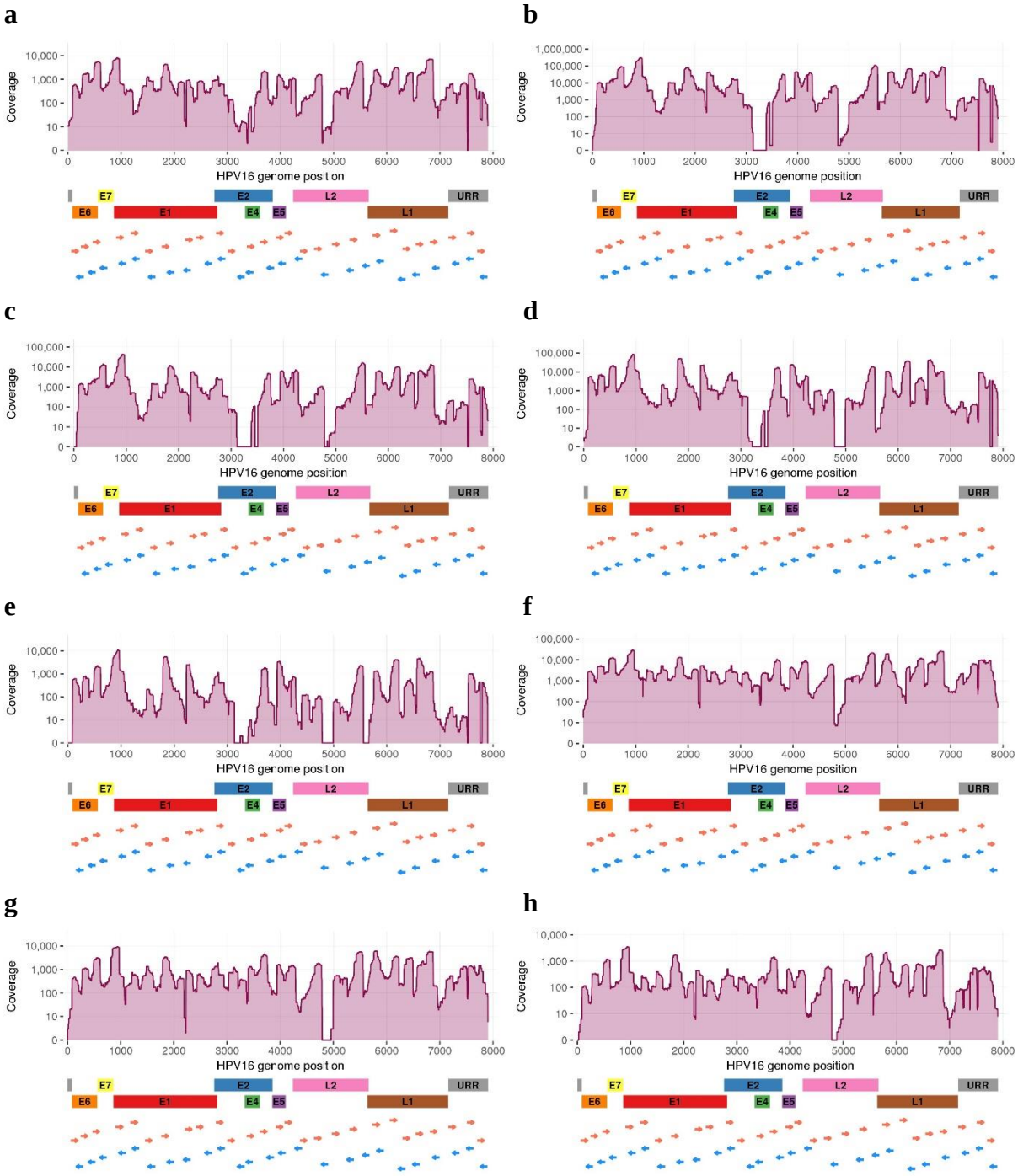
Sample	Sequence platform	Mean coverage	Variable sites	% Concordance
SiHa-1	HiSeq 2500	17561	809	21
SiHa-2		5609	522	
SiHa-1	MiSeq	2554	477	19
SiHa-2		646	257	

Supplementary Table S6. Mean coverage, total number of variable sites and percentage of variable sites in each HPV genes in the HPV positive cell lines and LBC samples.

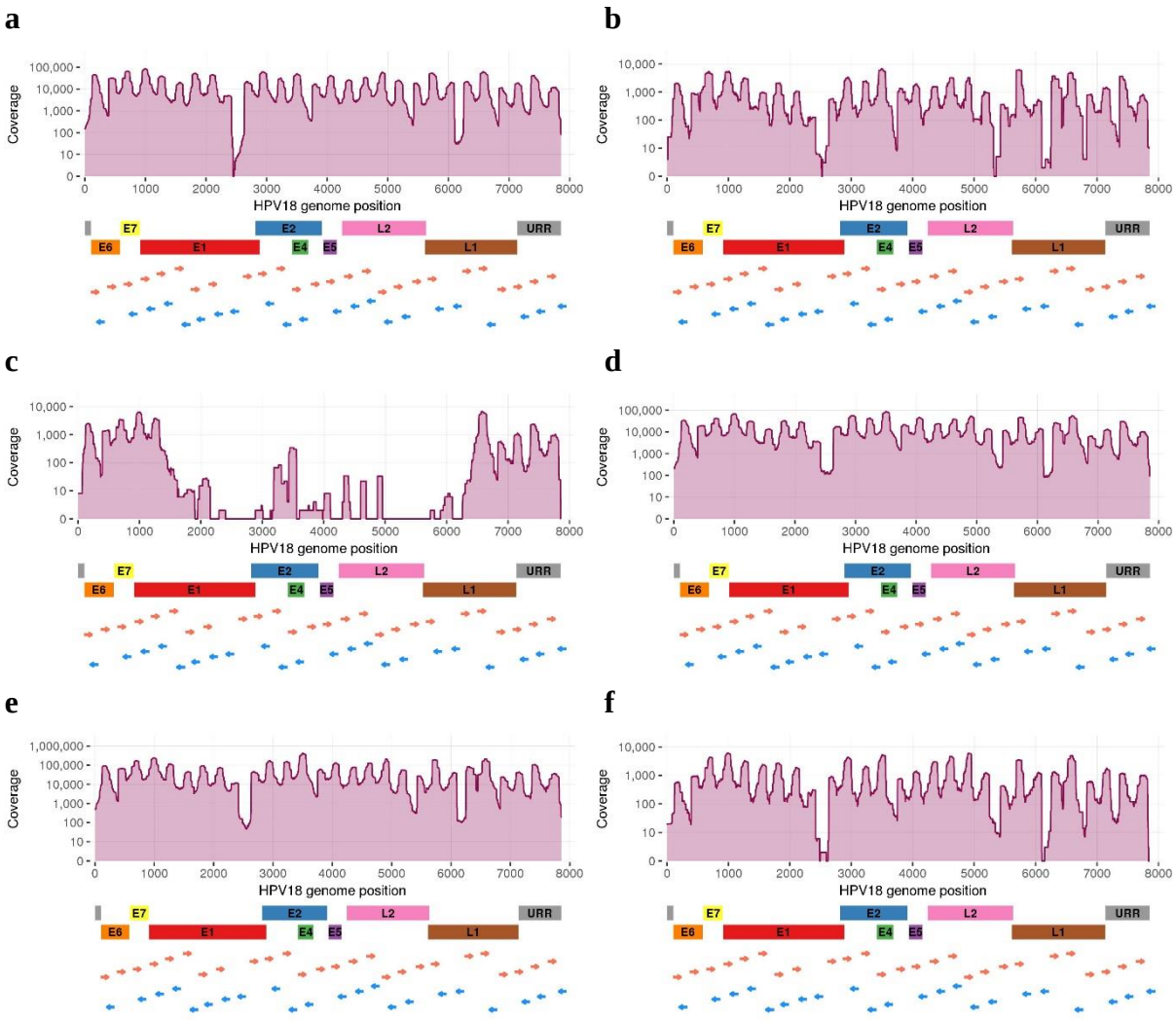
Sample	Mean coverage	Variable sites	% Variable sites								
			E6	E7	E1	E2	E4	E5	L2	L1	URR
HPV16											
CaSki	184716	1017	15.7	16.2	15.1	13.6	15.6	14.7	8.9	12.8	12.7
SiHa[b]	1018	500	7.8	7.4	7.1	4.0	0.3	7.9	4.4	8.6	4.8
SiHa-1[c]	17561	809	13.4	13.5	12.0	7.3	3.1	7.9	8.8	11.5	9.4
SiHa-1[b]	2554	477	10.5	13.1	7.5	3.2	0.3	4.8	3.3	7.6	4.0
SiHa-2[c]	5609	522	13.8	8.8	8.3	5.8	1.4	5.6	3.6	5.9	5.9
SiHa-2[b]	646	257	3.8	2.4	3.9	2.3	0.0	6	1.6	4.7	2.5
LBC1	1124	525	6.7	4.0	6.2	7.5	8.7	6.3	5.0	9.1	5.8
LBC7	384	244	1.9	4.7	2.7	3.0	0.7	1.2	3.1	4.5	1.4
HPV18											
HeLa	5897	679	14.5	18.2	13.1	1.8	0.0	0.0	0.0	12.1	11.0
LBC103	1056	456	7.3	11.3	4.8	7.2	7.9	7.2	5.2	5.4	5.7
LBC105	484	222	10.5	13.5	3.0	0.0	0.0	0.0	0.0	1.6	5.3
LBC107	14663	999	15.5	20.8	12.6	13.3	14.2	12.6	12.0	11.9	11.8
LBC108	46691	967	14.5	16.7	12.5	13.0	9.0	16.2	12.7	11.2	10.4
LBC48	988	521	4.4	12.9	7.5	7.0	8.6	9.0	7.0	5.6	4.6
HPV31											
LBC16	1065	561	5.3	6.1	7.8	10.7	7.8	0.4	8.9	7.3	3.1
LBC24	355	206	3.1	2.0	2.6	4.3	4.9	0.8	4.4	1.3	1.0
LBC32	18983	932	14.4	15.2	13.8	15.2	15.2	2.4	12.5	11.2	5.5
LBC34	23790	1060	16.7	14.8	16.6	15.4	17.5	0.4	13.4	12.8	8.6
HPV33											
LBC11	12038	1032	20.9	16.3	13.2	11.0	7.9	5.3	13.8	14.7	8.9
LBC30	303	202	2.4	2.0	2.5	3.7	7.5	0.0	3.3	1.5	3.1
LBC31	544	272	0.9	6.5	2.1	9.2	9.9	0.0	2.9	2.6	3.5
LBC52	439	313	4.9	6.1	3.2	6.2	9.9	0.0	5.1	1.9	5.0
LBC65	1993	732	17.1	12.9	10.3	9.0	11.1	0.0	8.4	8.7	6.9
HPV45											
MS751	845	321	14.3	7.5	7.9	0.0	0.0	0.0	0.0	0.1	8.9
LBC13	849	432	6.7	6.9	4.7	5.4	5.9	6.8	5.5	6.0	5.2
LBC29	614	329	2.7	1.2	4.4	5.2	7.0	4.5	3.9	4.2	4.9
LBC36	22093	1279	22.2	16.2	17.0	17.1	16.8	14.4	15.0	16.0	14.4
LBC54	256857	1641	27.5	28.0	25.8	20.8	18.7	17.6	17.7	17.2	18.0
LBC64	3943	782	15.1	10.0	10.7	10.4	13.6	5.9	9.0	9.8	7.9

[b] Sequenced on MiSeq sequencing platform.

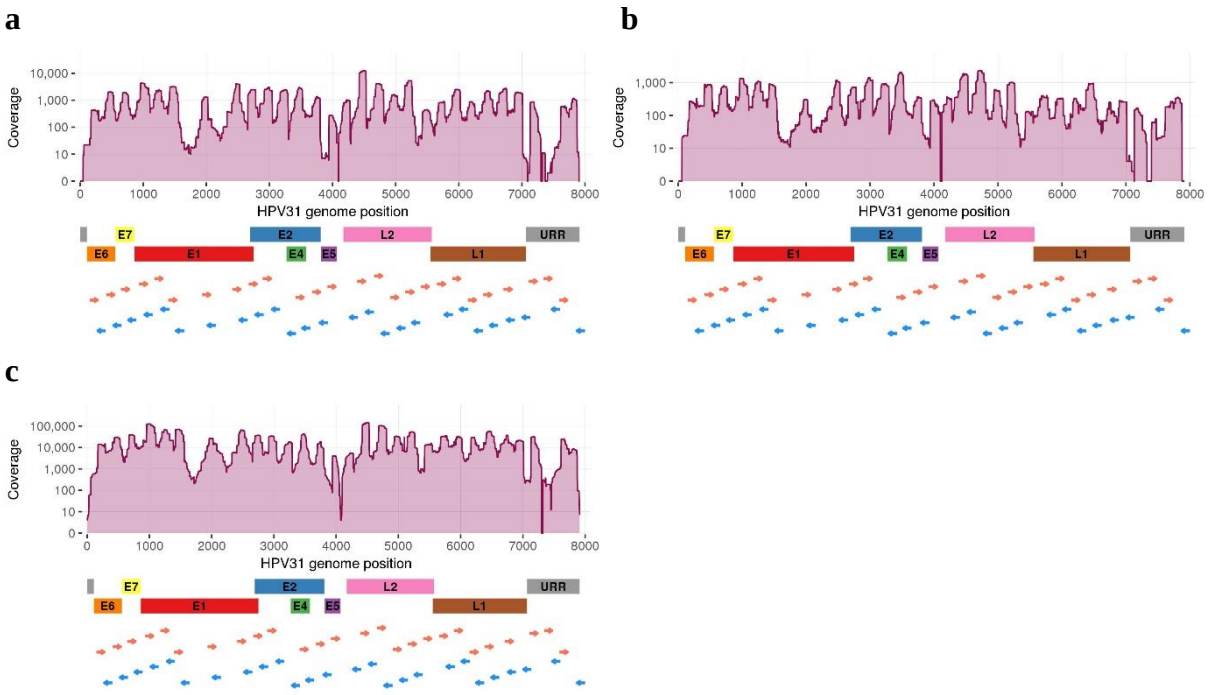
[c] Sequenced on HiSeq 2500 sequencing platform.



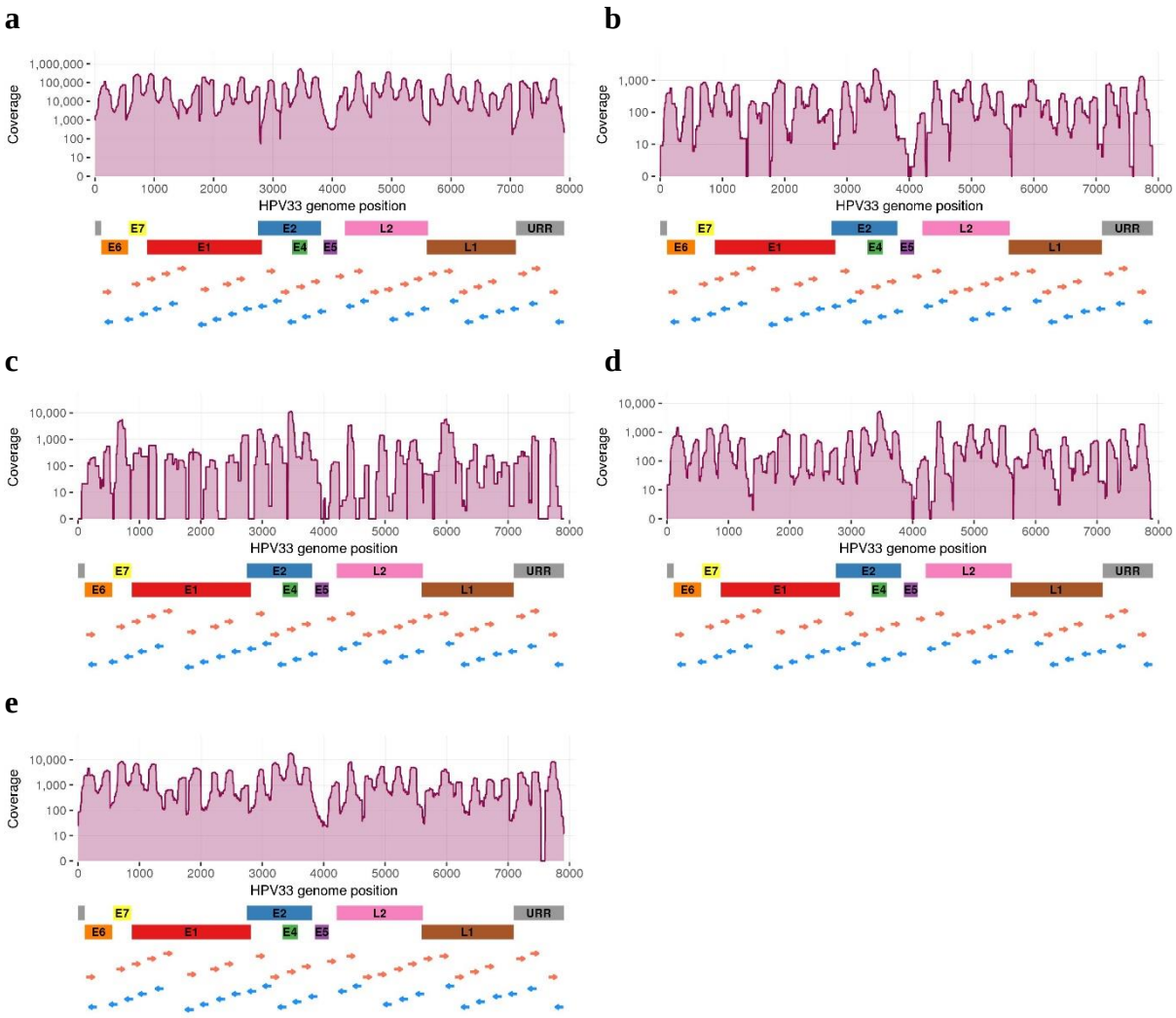
Supplementary Figure S1. HPV genome sequencing coverage of HPV16 positive samples a) SiHa (sequenced on MiSeq), b) SiHa-1 (sequenced on HiSeq), c) SiHa-1 (sequenced on MiSeq), d) SiHa-2 (sequenced on HiSeq), e) SiHa-2 (sequenced on MiSeq), f) WHO standard for HPV16, g) LBC1, and h) LBC7. The coverage plots are aligned to the HPV16 genome. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV16 primers is indicated below the genomic positions.



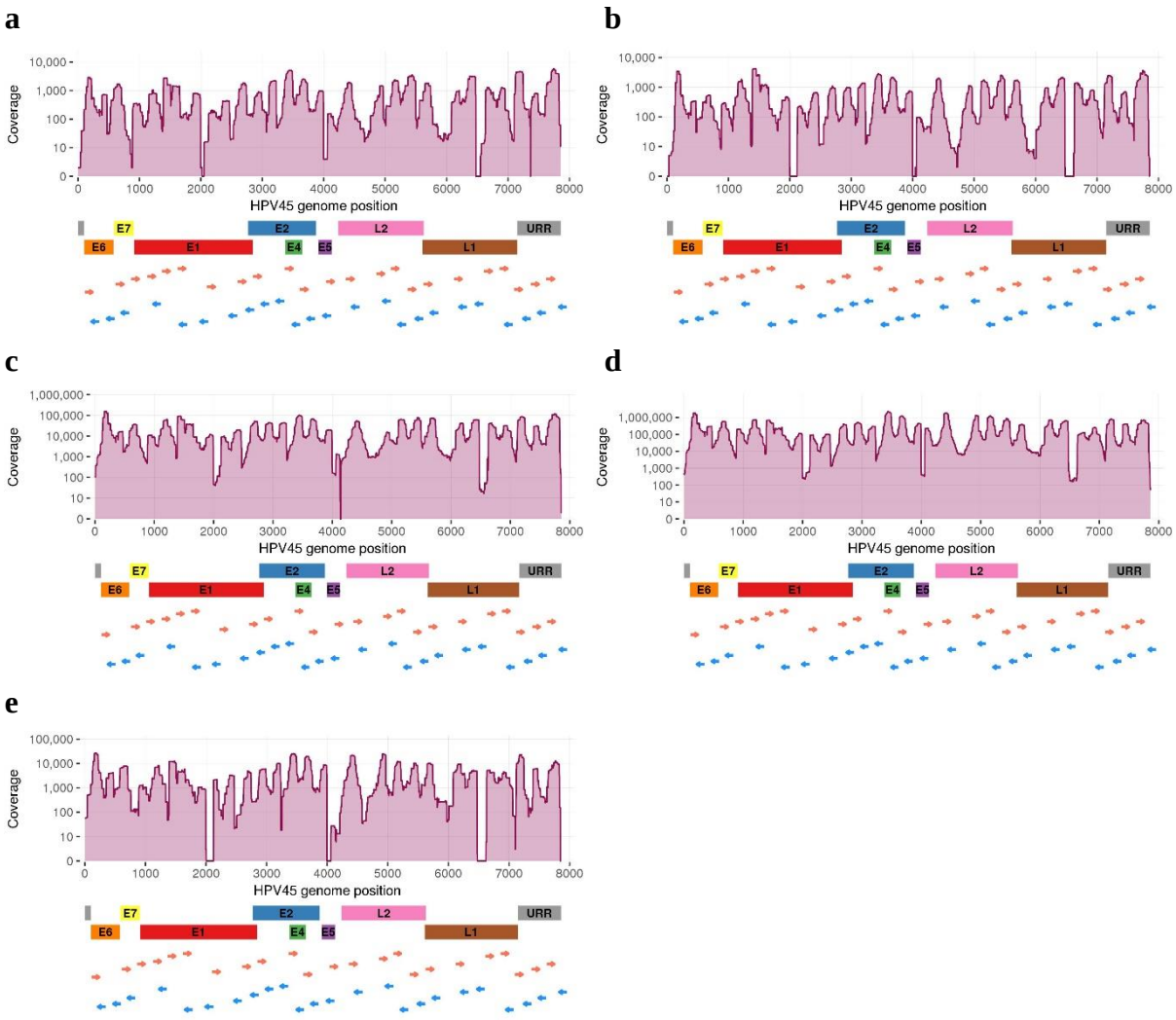
Supplementary Figure S2. HPV genome sequencing coverage of HPV18 positive samples a) WHO standard for HPV18, b) LBC103, c) LBC105, d) LBC107, e) LBC108, and f) LBC48. The coverage plots are aligned to the HPV18 genome. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV18 primers is indicated below the genomic positions.



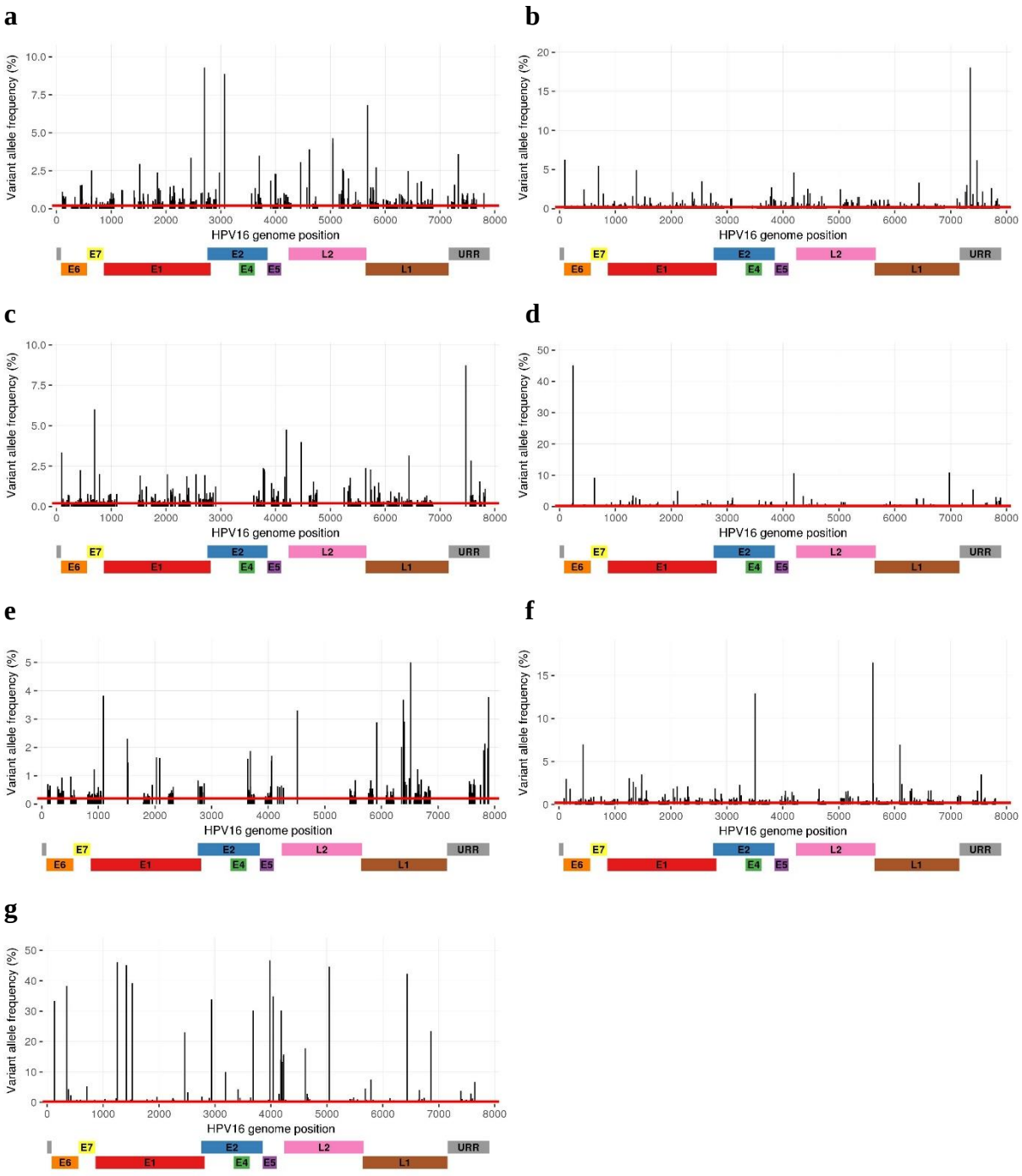
Supplementary Figure S3. HPV genome sequencing coverage of HPV31 positive samples a) LBC16, b) LBC24, and c) LBC32. The coverage plots are aligned to the HPV31 genome. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV31 primers is indicated below the genomic positions.



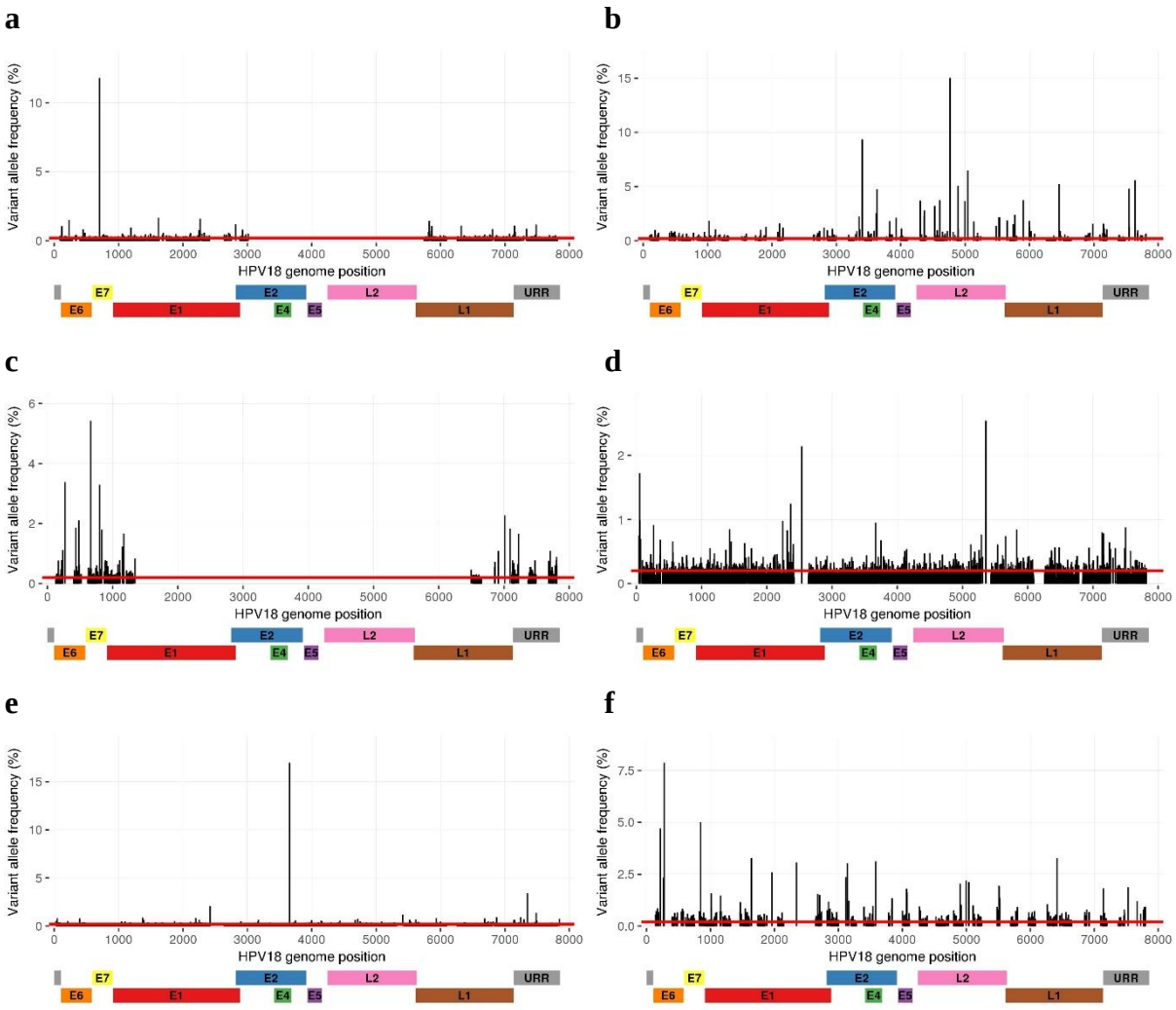
Supplementary Figure S4. HPV genome sequencing coverage of HPV33 positive samples a) HPV33 plasmid, b) LBC30, c) LBC31, d) LBC52, and e) LBC65. The coverage plots are aligned to the HPV33 genome. Location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV33 primers is indicated below the genomic positions.



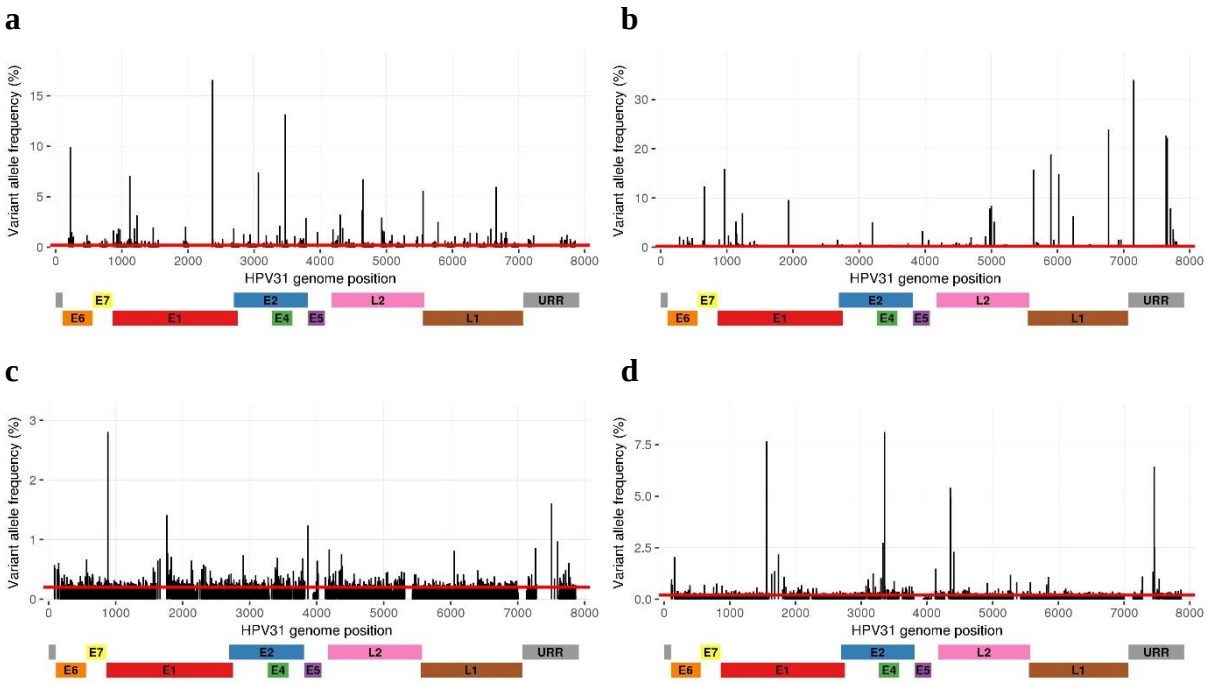
Supplementary Figure S5. HPV genome sequencing coverage of HPV45 positive samples a) LBC13, b) LBC29, c) LBC36, d) LBC54, and e) LBC64. The coverage plots are aligned to the HPV45 genome. The location of early (E1, E2, E4-7), late (L1, L2) genes, URR, and forward (red arrows) and reverse (blue arrows) HPV45 primers is indicated below the genomic positions.



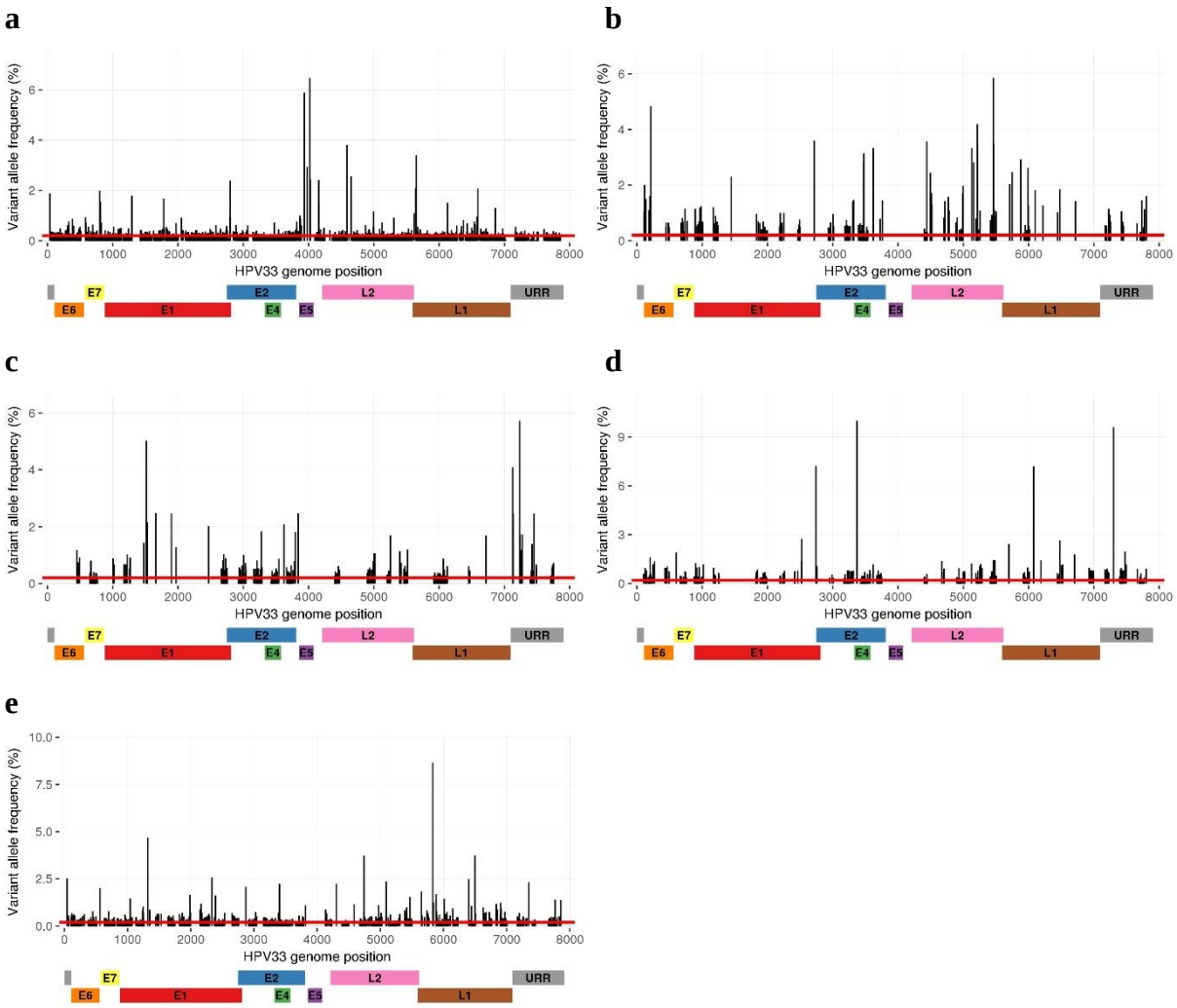
Supplementary Figure S6. Variable sites and variant allele frequency (%) in HPV16 positive samples a) SiHa (sequenced on MiSeq), b) SiHa-1 (sequenced on HiSeq), c) SiHa-1 (sequenced on MiSeq), d) SiHa-2 (sequenced on HiSeq), e) SiHa-2 (sequenced on MiSeq), f) LBC1, and g) LBC7. The variant plots are aligned to the HPV16 genome with the location of genes and URR. The red line indicates the variant calling threshold value of 0.2%.



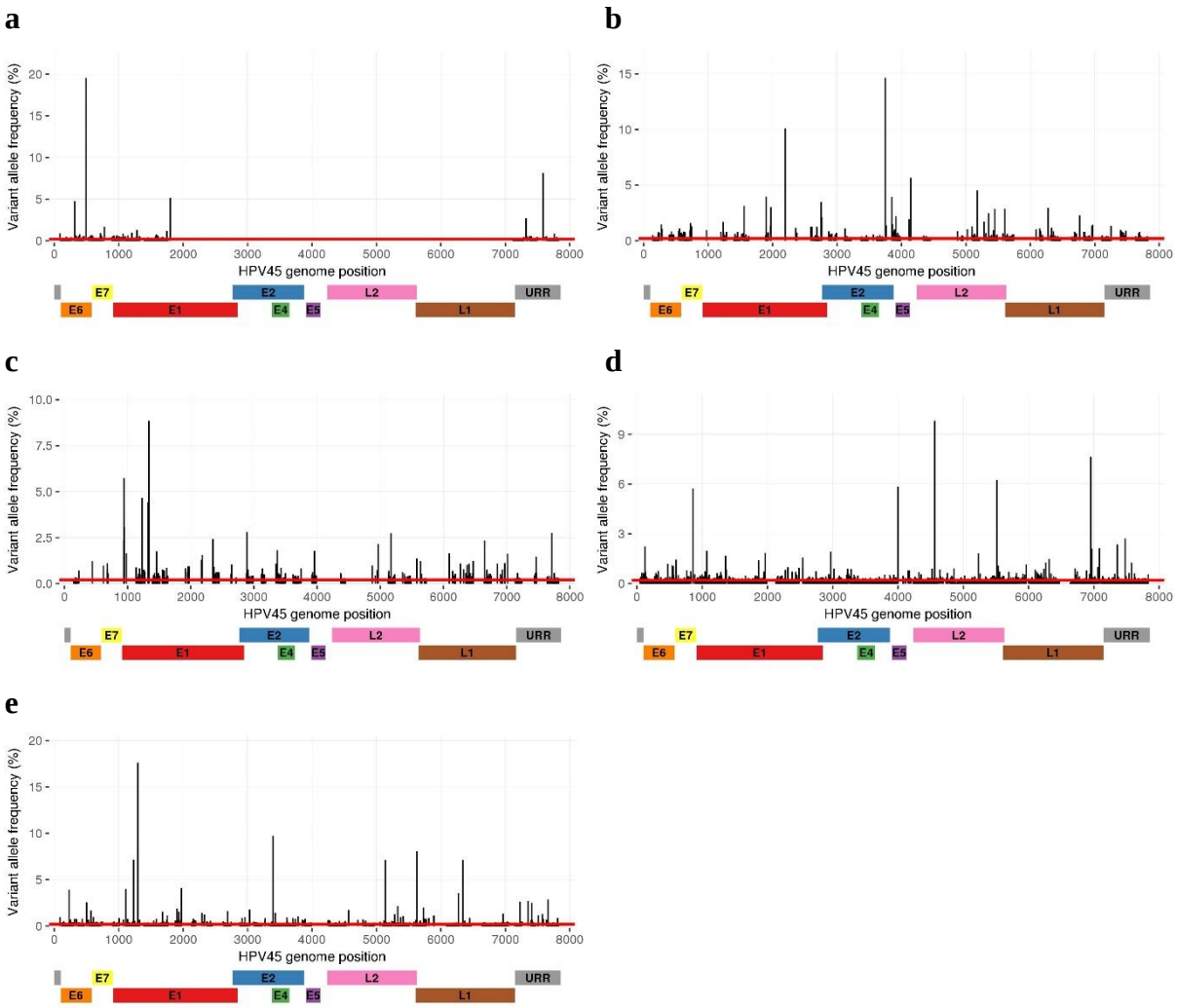
Supplementary Figure S7. Variable sites and variant allele frequency (%) in HPV18 positive samples a) HeLa, b) LBC103, c) LBC105, d) LBC107, e) LBC108, and f) LBC48. The variant plots are aligned to the HPV18 genome with the location of genes and URR. The red line indicates the variant calling threshold value of 0.2%.



Supplementary Figure S8. Variable sites and variant allele frequency (%) in HPV31 positive samples a) LBC16, b) LBC24, and c) LBC32, and d) LBC34. The variant plots are aligned to the HPV31 genome with the location of genes and URR. The red line indicates the variant calling threshold value of 0.2%.



Supplementary Figure S9. Variable sites and variant allele frequency (%) in HPV33 positive samples a) LBC11, b) LBC30, c) LBC31, d) LBC52, and e) LBC65. The variant plots are aligned to the HPV33 genome with the location of genes and URR. The red line indicates the variant calling threshold value of 0.2%.



Supplementary Figure S10. Variable sites and variant allele frequency (%) in HPV45 positive samples a) MS751, b) LBC13, c) LBC29, d) LBC36, and e) LBC64. The variant plots are aligned to the HPV45 genome with the location of genes and URR. The red line indicates the variant calling threshold value of 0.2%.