



## Rearrangements of viral and human genomes at human papillomavirus integration events and their allele-specific impacts on cancer genome regulation

Vanessa L. Porter, Michelle Ng, Kieran O'Neill, et al.

*Genome Res.* published online December 5, 2024

Access the most recent version at doi:[10.1101/gr.279041.124](https://doi.org/10.1101/gr.279041.124)

---

**P<P** Published online December 5, 2024 in advance of the print journal.

**Accepted Manuscript** Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

**Open Access** Freely available online through the *Genome Research* Open Access option.

**Creative Commons License** This manuscript is Open Access. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at <http://creativecommons.org/licenses/by/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Published by Cold Spring Harbor Laboratory Press

1 **Title:**

2 Rearrangements of viral and human genomes at human papillomavirus integration events  
3 and their allele-specific impacts on cancer genome regulation

4  
5 **Author List:**

6 Vanessa L. Porter<sup>1,2,3</sup>, Michelle Ng<sup>1,3</sup>, Kieran O'Neill<sup>1</sup>, Signe MacLennan<sup>1,2,3</sup>, Richard D.  
7 Corbett<sup>1</sup>, Luka Culibrk<sup>1,4</sup>, Zeid Hamadeh<sup>5,6</sup>, Marissa Iden<sup>7,8</sup>, Rachel Schmidt<sup>7,8</sup>, Shirng-Wern  
8 Tsaih<sup>7,8</sup>, Carolyn Nakisige<sup>9</sup>, Martin Origa<sup>9</sup>, Jackson Orem<sup>9</sup>, Glenn Chang<sup>1,10</sup>, Jeremy Fan<sup>1,4</sup>,  
9 Ka Ming Nip<sup>1,4</sup>, Vahid Akbari<sup>1,2</sup>, Simon K. Chan<sup>1</sup>, James Hopkins<sup>1</sup>, Richard A. Moore<sup>1</sup>, Eric  
10 Chuah<sup>1</sup>, Karen L. Mungall<sup>1</sup>, Andrew J. Mungall<sup>1</sup>, Inanc Birol<sup>1,2</sup>, Steven J. M. Jones<sup>1,2</sup>, Janet S.  
11 Rader<sup>7,8</sup>, Marco A. Marra<sup>1,2,3</sup>

12

13 **Author Affiliations:**

14 <sup>1</sup>Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada

15 <sup>2</sup>Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

16 <sup>3</sup>Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

17 <sup>4</sup>Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

18 <sup>5</sup>Cytogenomics Laboratory, Vancouver General Hospital, Vancouver, BC, Canada

19 <sup>6</sup>Department of Pathology and Laboratory Medicine, University of British Columbia,  
20 Vancouver, BC, Canada

21 <sup>7</sup>Department of Obstetrics and Gynecology, Medical College of Wisconsin, Milwaukee, WI,  
22 53226, USA.

23 <sup>8</sup>Medical College of Wisconsin Cancer Center, Milwaukee, WI, 53226, USA.

24 <sup>9</sup>Uganda Cancer Institute, Kampala, Uganda

25 <sup>10</sup>Genome Science and Technology Graduate Program, University of British Columbia,  
26 Vancouver, BC, Canada

## 27 Abstract

28 Human papillomavirus (HPV) integration has been implicated in transforming HPV  
29 infection into cancer. To resolve genome dysregulation associated with HPV integration, we  
30 performed Oxford Nanopore long-read sequencing on 72 cervical cancer genomes from a  
31 Ugandan dataset that was previously characterized using short-read sequencing. We found  
32 recurrent structural rearrangement patterns at HPV integration events, which we categorized  
33 as: del(letion)-like, dup(lication)-like, translocation, multi-breakpoint, or repeat region  
34 integrations. Integrations involving amplified HPV-human concatemers, particularly multi-  
35 breakpoint events, frequently harbored heterogeneous forms and copy numbers of the viral  
36 genome. Transcriptionally active integrants were characterized by unmethylated regions in  
37 both the viral and human genomes downstream from the viral transcription start site, resulting  
38 in HPV-human fusion transcripts. In contrast, integrants without evidence of expression lacked  
39 consistent methylation patterns. Furthermore, whereas transcriptional dysregulation was  
40 limited to genes within 200 kilobases of an HPV integrant, dysregulation of the human  
41 epigenome in the form of allelic differentially methylated regions affected megabase expanses  
42 of the genome, irrespective of the integrant's transcriptional status. By elucidating the  
43 structural, epigenetic, and allele-specific impacts of HPV integration, we provide insight into  
44 the role of integrated HPV in cervical cancer.

45

46

47

48

49

50

## 51 Introduction

52 Human papillomavirus (HPV), an 8-kilobase (kb), double-stranded, circular DNA virus,  
53 drives nearly all cervical cancers and a subset of head and neck cancers and anogenital  
54 cancers (de Sanjosé et al. 2018). Vaccination against high-risk HPV types has effectively  
55 reduced cervical cancer rates (Falcaro et al. 2021). However, cervical cancer incidence  
56 remains high in countries lacking vaccine availability (Bruni et al. 2016). The HPV genome  
57 normally exists as episomes in infected cells; however, during HPV-driven oncogenesis, the  
58 viral DNA often becomes integrated into the host cell genome (Huang et al. 2008). HPV  
59 integration is associated with changes in the sequence content of HPV, viral gene copy  
60 number, and epigenetic regulation of the viral genome (Warburton et al. 2018; Groves et al.  
61 2016; Warburton et al. 2021). Genomic and epigenomic changes can also occur in adjacent  
62 human genomic regions (Tian et al. 2023; Adey et al. 2013; Mima et al. 2023; Groves et al.  
63 2021; Symer et al. 2022), and may have cancer-promoting consequences, as they often affect  
64 oncogenes within or near the site of HPV integration (Iden et al. 2021).

65

66 HPV integration is not a normal part of the HPV life cycle and occurs when double-  
67 stranded breaks in non-contiguous regions of the human genome are bridged using surrogate  
68 double-stranded DNA from the virus (Akagi et al. 2014). In cancers, HPV integration sites  
69 occur throughout the genome, but several loci are recurrently affected, including regions near  
70 *MYC*, *TP63*, *FHIT*, and *KLF5* (Hu et al. 2015; Gagliardi et al. 2020; Warburton et al. 2021; Fan  
71 et al. 2023; Symer et al. 2022; Bodelon et al. 2016). Different reports have indicated that  
72 integration may occur more frequently at regions with HPV microhomology, at fragile sites, or  
73 at sites that are actively transcribed (Hu et al. 2015; Warburton et al. 2021). Some of these  
74 apparent discrepancies may be attributed to differences in the sequencing approaches used  
75 (i.e. whole genome, HPV capture, RNA) (Fan et al. 2023). HPV integration sites are also  
76 associated with dense structural alterations, perhaps as a result of unstable HPV-human  
77 concatemers and intermediates that can form during replication and cause multiple HPV-

78 human breakpoints across a genomic locus, which is referred to as an integration event  
79 (Kadaja et al. 2009; Akagi et al. 2023). The genomic consequences of HPV integration have  
80 thus far been difficult to study due to the limitations of short reads in capturing the complexity  
81 of human-HPV structures. However, recent advancements in long-read DNA sequencing  
82 technology have enhanced our ability to interpret repetitive sequences, structural alterations,  
83 and DNA methylation signals (Kovaka et al. 2023). Combined with enhanced haplotype  
84 phasing capabilities (Akbari et al. 2021), long-read sequencing may therefore be used to  
85 resolve complex HPV-integrated genomes and methylomes.

86

87         To investigate the structural changes resulting from HPV integration and their impacts  
88 on human and viral genome regulation, we used Oxford Nanopore long-read sequencing to  
89 characterize HPV-positive cervical cancer tumor genomes from the HIV Tumor Molecular  
90 Characterization Project (HTMCP). We found 438 HPV-human integration breakpoints  
91 mapping to 129 integration events across 69 HPV-integrated cervical cancer samples, and  
92 characterized the genomic structures, methylation patterns, and transcriptional regulation  
93 associated with these loci. Our analysis revealed the extent of *cis*-linked genomic  
94 dysregulation resulting from HPV integration, including structural rearrangements and  
95 modulation of virus and host gene expression and epigenetic regulation.

96

## 97 Results

### 98 *Long-read whole genome sequencing of cervical tumors*

99         We sequenced 72 cervical cancer samples from the HTMCP cohort using whole  
100 genome Oxford Nanopore Technologies (ONT) long-read sequencing (Methods). The  
101 samples yielded an average of 102 gigabase pairs (Gb) of data (range 52.6 - 153 Gb; median  
102 coverage = 34×; Supplemental Fig. S1). We achieved long read lengths in our samples  
103 (median N50 = 17.5 kb; range 9.0 - 34.1 kb; Supplemental Fig. S1) that were of high read  
104 quality, as measured by the base error rate (median = 4.8%; range 2.2% - 8.4%) and the

105 proportion of artifactual chimeric DNA fusions (median = 5.0%; range 1.0% - 12.4%). These  
106 data are supplemented by high-quality, short-read whole genome and RNA sequencing (RNA-  
107 seq), as previously described (Gagliardi et al. 2020).

108

### 109 *Detection of HPV integration using long-read sequencing*

110 The molecular and clinical properties of the 72 samples are shown in Fig. 1A  
111 (Supplemental Table S1). From a sequence perspective, we define HPV integration as the  
112 recombination of HPV with human DNA, which can be detected from chimeric reads that align  
113 to both the HPV and human genomes. We observed that HPV integration events involved at  
114 least two double-stranded human genome breaks and two double-stranded HPV genome  
115 breaks (Fig. 1B). Breakpoints were grouped into an “integration event” if either of the following  
116 conditions was met: (1) the HPV breakpoints co-occurred on one or more of the same reads,  
117 or (2) the breakpoints mapped within 500 kb of each other (Cancer Genome Atlas Research  
118 Network et al. 2017; Gagliardi et al. 2020) (Methods). Applying these conditions, we detected  
119 438 integration breakpoints belonging to 129 integration events across 69 samples with HPV  
120 integration (Fig. 1A; Supplemental Table S2). To determine whether an event was transcribed,  
121 we searched for HPV-human fusion transcripts occurring in the vicinity of the integration event  
122 using existing RNA-seq data (Methods). At most, two events per sample produced fusion  
123 transcripts, but even samples without a recurrent RNA fusion site near an integration event  
124 expressed HPV (Fig. 1A; Supplemental Table S2). In three samples with deep and evenly  
125 distributed read coverage across the HPV genome, we did not detect HPV integration  
126 breakpoints, indicative of highly amplified episomal HPV. One sample harbored three  
127 integration events from two different HPV types, HPV16 (two events) and HPV59 (one event),  
128 although only the HPV59 event was expressed (Supplemental Table S2).

129

130 Next, we compared the number and genomic locations of HPV integration breakpoints  
131 identified in our long-read sequencing data to previous short-read sequencing data (Gagliardi

132 et al. 2020) (Methods). Across the 69 HTMCP samples with HPV integration, 424 and 438  
133 integration breakpoints were called in the short-read and long-read sequencing datasets,  
134 respectively (Supplemental Table S2 and Supplemental Table S3). Integration breakpoint  
135 calls using short- and long-reads were identical for 33 samples. Eighteen samples contained  
136 more calls in the long-read data, and 18 contained more calls in the short-read data  
137 (Supplemental Fig. S2A-B). Two samples where HPV integration was not detected with short  
138 reads contained one integration breakpoint each in the long-read data. One of these samples  
139 had even and deep read coverage indicative of episomal HPV, but an integration event was  
140 detected at a low variant allele frequency (VAF = 0.00025), suggesting that integration may  
141 have occurred sub-clonally (Supplemental Table S2). The second sample contained an  
142 integration site within a repetitive region of Chromosome 21p11 that was detected with long  
143 reads, but the region contained no short reads with adequate mapping quality.

144

145 Two samples had highly discordant integration breakpoint calls when detecting HPV-  
146 human junction sites, with an HPV58-integrated sample having 59 more calls in the long-read  
147 data (82 vs 23) and an HPV16-integrated sample having 53 more calls in the short-read data  
148 (56 vs 3; Supplemental Fig. S2A-B). We confirmed overlap between integration breakpoint  
149 calls in each technology for these two samples. In the HPV58-integrated sample, 67/105  
150 (64%) integration breakpoints called in either the long-read or short-read data overlapped  
151 repetitive regions, and 30/59 (51%) breakpoints resided in repetitive regions in the HPV16-  
152 integrated sample, suggesting that read alignments may be confounded by repetitive  
153 sequences.

154

155 As expected, the number of integration breakpoint calls per sample was correlated  
156 between technologies (Supplemental Fig. S2C; Spearman's correlation,  $R=0.78$ ,  $p=3.8\times 10^{-15}$ ). Both technologies detected similar proportions of breakpoint calls within CpG islands and  
157 exonic, intronic, and repetitive regions. (Supplemental Fig. S2D). Long reads detected more  
158 integration breakpoints per sample, and discordance between calls made using long- and  
159

160 short-read technologies could be due to low-confidence mapping of short reads in repetitive  
161 regions.

162

### 163 *Structural alterations associated with HPV integration events*

164         Based on the sequence alignment patterns of HPV-containing reads and the HPV  
165 integration breakpoints, we defined five categories of integration structures in the 69 integrated  
166 samples. These were (del)etion-like, (dup)lication-like, translocation, multi-breakpoint, and  
167 repeat region integrations, which we categorized using a custom workflow (Fig. 1B-C;  
168 Methods). Multi-breakpoint integrations were the most prevalent, comprising 41/129 (32%) of  
169 integration events in our dataset (Fig. 1D; Supplemental Table S2). Next were dup-like  
170 (37/129) and del-like integrations (29/129), which differed by the presence or absence of the  
171 intervening human sequence between the two HPV breakpoints (Fig. 1B-C). We also  
172 investigated the potential of integration events existing as extrachromosomal circular DNA  
173 (ecDNA) using the events' *de novo* assemblies. We identified eight events as potential  
174 ecDNAs, most of which (6/8) were categorized as dup-like integrations (Supplemental Table  
175 S4). Translocation (3/129), and repeat region (3/129) integrations were rare (Fig. 1D;  
176 Supplemental Table S2). The remainder of the events (16/129) were unmatched single  
177 breakpoints that we were unable to categorize (Supplemental Table S2). While multi-  
178 breakpoint events were the most prevalent, only 49% of these events produced HPV fusion  
179 transcripts, in contrast to 76% of dup-like events. In fact, dup-like events were more often  
180 expressed when compared to all other types of events (Fisher's exact test,  $p=0.00096$ , Fig.  
181 1D). The distributions of the integration categories also differed between HPV types (Fig. 1E).

182

183         Consistent with previous reports, we observed regions of recurrent HPV integration  
184 (>2 events within 1 Mb bins) in the genome (Symer et al. 2022; Gagliardi et al. 2020; Bodelon  
185 et al. 2016; Hu et al. 2015; Warburton et al. 2021; Fan et al. 2023; Cancer Genome Atlas  
186 Research Network et al. 2017), most of which contained events that produced fusion

187 transcripts (Fig. 1F). The most frequently integrated loci in our dataset were near *KLF5/KLF12*  
188 (13q22; Supplemental Fig. S3A), *MYC* (8q24; Supplemental Fig. S3B), and *TP63* (3q28;  
189 Supplemental Fig. S3C). All of these loci contained four integration events within 1 Mb of the  
190 respective gene and all integration events produced HPV-human fusion transcripts. All events  
191 in the 8q24 locus were multi-breakpoint events, whereas the 13q22 and 3q28 loci exclusively  
192 contained dup-like events (Supplemental Table S2). Another recurrently integrated locus was  
193 detected near the repeat-containing ribosomal gene *RNA28SN2* (21p11) in four samples (Fig.  
194 1F). Three of the four events were not expressed, but we suspect this was due to a limited  
195 ability to map short RNA-seq reads in this low complexity locus. The one expressed event was  
196 a multi-chromosomal event with a fusion transcript detected in a more mappable region  
197 overlapping *FAM83B* on Chromosome 6p12. We compared the expression of *MYC*, *TP63*,  
198 and *KLF5* in samples with and without HPV integration within 1 Mb of the respective genes  
199 (Fig. 1G; Supplemental Table S5) and observed that *MYC* and *TP63* expression was  
200 significantly increased in the integrated samples, whereas *KLF5* expression was not affected  
201 (Wilcoxon; *MYC*  $p=0.026$ ; *TP63*  $p=0.024$ ; *KLF5*  $p=0.61$ ; Fig. 1G).

202

### 203 *HPV structure and copy number heterogeneity within integration events*

204 We defined an HPV integrant as an uninterrupted segment of the integrated virus  
205 genome situated between two HPV-human breakpoints. Examples of possible integrant  
206 structures are shown in Fig. 2A. An HPV breakpoint can be unambiguously linked to another  
207 HPV breakpoint when the entirety of the integrant is contained on a single read, and we refer  
208 to each unique pairing as a breakpoint pair (Supplemental Table S6). Using such reads, we  
209 found evidence for different configurations of viral genome segments – some of which were  
210 rearranged and/or concatemered – between a single breakpoint pair (Fig. 2B). We refer to  
211 these as heterologous integrants, and hypothesize that they may arise as a result of unequal  
212 amplification of the HPV genome during the replication of HPV-human concatemers, akin to  
213 “heterocateny” described by Akagi et al. (Akagi et al. 2023) and the “onion-skin” structures

214 described by Kadaja et al. (Kadaja et al. 2009) (Fig. 2B). Heterologous integrants were found  
215 in 21% (45/212) of breakpoint pairs in our cohort (Fig. 2C; Supplemental Table S6), and up to  
216 15 unique integrant configurations were observed within a single breakpoint pair (Fig. 2C).  
217 Fig. 2D summarizes the different configurations of HPV integrants present in an example multi-  
218 breakpoint event. This sample harbored four connected breakpoints within the HPV genome,  
219 splitting it into four segments, which were variably rearranged across the integration event  
220 (Fig. 2D). We used read lengths and alignments to distinguish the HPV integrant sizes and  
221 structures (Methods), and detected various combinations – some of them periodic – of the  
222 four HPV segments between the different HPV-human breakpoint pairs within the event (Fig.  
223 2D).

224

225 Heterogeneity in HPV genome structure has been reported in the episome prior to  
226 integration (Rossi et al. 2023). We therefore investigated HPV multimers, i.e. concatenated  
227 HPV genomes, and other HPV genome structural variants (SVs) in our four predominantly  
228 episomal samples (including one sample with sub-clonal integration, VAF = 0.00025). HPV  
229 genome multimers were found to be the dominant configuration in two samples, while the  
230 other two contained predominantly single copy episomes, although we cannot exclude the  
231 possibility that shorter reads may bias these two samples towards detection of smaller  
232 configurations (Fig. 2E). Sample HTMCP-03-06-02156 showed an accumulation of read  
233 lengths between one and two HPV copies, indicating episomal structural variation, and we  
234 indeed found 23 viral SVs in this sample (Supplemental Fig. S4). We thus hypothesize that  
235 integrant rearrangement and copy number heterogeneity may partially originate within the  
236 episome prior to HPV integration.

237

238 Most heterologous integrants (91%) originated from multi-breakpoint events and were  
239 not detected in del-like or translocation integration events, compatible with the notion that  
240 amplified copy number is associated with their formation (Fig. 2F; Supplemental Table S6).  
241 HPV18 integrants were more frequently heterologous than other HPV types (35% vs. 21%;

242 Fisher's exact test,  $p=0.0071$ ; Fig. 2G) and contained significantly more viral genome copies  
243 per integrant than HPV16 (Wilcoxon,  $p=0.040$ ; Fig. 2H).

244

245 We were unable to determine the complete integrant for 123 breakpoints because no  
246 reads linked the detected breakpoint to any other. For each of these incomplete integrants,  
247 we used the longest HPV sequence contained on a read to determine the minimum size of  
248 the integrant (Fig. 2I; Supplemental Table S6). Incomplete integrants often contained at least  
249 two HPV copies (mean = 2.1) and the longest incomplete HPV integrant was supported by a  
250 read that spanned 48 kb (Fig. 2I). In contrast, the largest complete integrant between a  
251 breakpoint pair was 37 kb. Thus, even with long-read sequencing, we did not achieve  
252 complete resolution of HPV integrant lengths in 123/335 (37%) of HPV breakpoints.

253

#### 254 *Comparison of two-breakpoint events*

255 Two-breakpoint integration events involving one chromosome comprised 66/129  
256 (51%) of events in our cohort and occurred when one HPV integrant was inserted between  
257 two breaks in the human genome. These were associated with either del-like insertions or  
258 dup-like expansions (Fig. 1B), which were differentiated based on the alignment patterns of  
259 the HPV-containing chimeric reads. Most predicted ecDNAs (7/8) were also two-breakpoint  
260 events (Supplemental Table S4). Del-like events were characterized by a copy number loss  
261 between the integration breakpoints, in some cases accompanied by amplification of the  
262 regions flanking the integration (Fig. 3A). In dup-like events, including six potential ecDNAs,  
263 the region between the breakpoints was amplified; in predicted ecDNAs, read alignments were  
264 fully contained between and did not extend beyond the breakpoints (Fig. 3A).

265

266 We predict that candidate ecDNA integration events exist outside the chromosome in  
267 a self-contained circular structure, where the human region between the integration  
268 breakpoints is bridged by HPV (examples shown in Fig. 3B). The first example shows a near-

269 complete copy of HPV18 connecting two breakpoints from an intergenic region on  
270 Chromosome 2 (Fig 3b, left). The second example shows a candidate ecDNA that contains a  
271 portion of *TP63* along with a portion of HPV45 (Fig. 3B, middle). The third example shows an  
272 atypical ecDNA integration event, in which the ecDNA contains part of HPV52 along with a  
273 rearranged intergenic segment of Chromosome 13 (Fig. 3B, right). The eight candidate  
274 ecDNAs had an average assembly size of 48.8 kb (range 22.4 - 78.0 kb; Fig 3c) and  
275 predominantly involved clade A7 HPV types (4 HPV18, 3 HPV45; Supplemental Table S4).  
276 We predicted no HPV16 ecDNAs.

277

278 The human-HPV breakpoints in dup-like and del-like integration events were  
279 distributed differently across the human genome. Firstly, there was greater spacing between  
280 the two breakpoints in dup-like events, with a median distance of 59.1 kb compared to 938 bp  
281 (Wilcoxon,  $p=7.6\times 10^{-6}$ ; Fig. 3D; Supplemental Table S7). The dup-like events were also less  
282 likely to be completely contained within genic regions and more likely to have a breakpoint  
283 within the 3' or 5' end of a gene (Fisher's Exact,  $p=0.040$ ; Fig. 3E; Supplemental Table S7).

284

### 285 *HPV integration is associated with full chromosome arm translocations*

286 Translocation integrations, in which two breakpoints occurred on different  
287 chromosomes (Akagi et al. 2023), were rare in our study. There were only three instances,  
288 involving three different HPV types (HPV18, 52, and 31). The read alignments from the HPV52  
289 translocation integration event are shown in Supplemental Fig. S5. This event involved a focal  
290 amplification around the site of HPV integration and a segment of HPV containing *E6* and *E7*  
291 (Supplemental Fig. S5A). The Chromosome 8 breakpoint, although within a gene, occurred  
292 near the pericentromeric region and was adjacent to a one copy loss upstream of the  
293 breakpoint. The Chromosome 1 breakpoint occurred in a genic region on the p-arm and  
294 showed copy gain across a ~70 Mb segment upstream of the integration breakpoint  
295 (Supplemental Fig. S5B). This indicates that the Chromosome 1 region was duplicated and

296 recombined with the q-arm of Chromosome 8, the p-arm of Chromosome 8 was lost, and the  
297 HPV52 segment was focally amplified around the translocation junction (Supplemental Fig.  
298 S5C).

299

### 300 *Multi-breakpoint HPV integration events are structurally complex*

301 Multi-breakpoint events, comprising 41/129 (32%) of all events, were the most  
302 common type of event in our study. Fig. 4A summarizes the number of HPV-human  
303 breakpoints per multi-breakpoint event, which ranged from 3 to 32 (Supplemental Table S8).  
304 On average, transcribed integration events contained more breakpoints per event than non-  
305 transcribed events (Wilcoxon,  $p=0.0084$ ; Fig. 4B). Integrations with higher numbers of HPV  
306 breakpoints were also associated with more human-only structural variants (SVs) overlapping  
307 the integration event (Spearman's correlation,  $R=0.69$ ,  $p=6.0\times 10^{-7}$ ) indicating general  
308 instability within these regions (Fig. 4C; Supplemental Table S8).

309

310 We explored how HPV breakpoints were connected within multi-breakpoint events.  
311 Five examples are shown in Fig. 4D, three of which originate from the recurrently integrated  
312 region on Chromosome 8 near *MYC* (Fig. 1F). The presence of heterologous integrants  
313 between breakpoint pairs is indicated, as is the degree of connectivity of each breakpoint (Fig.  
314 4D). HTMCP-03-06-02058 and HTMCP-03-06-02267 both feature one highly connected  
315 breakpoint that pairs with all other breakpoints in the event. This contrasts with the HTMCP-  
316 03-06-02149 event, in which each breakpoint connects to two others. The two multi-  
317 chromosome events shown in Fig. 4D were the two most rearranged events. The HTMCP-03-  
318 06-02109 event involved a heavily rearranged Chromosome 3, while Chromosomes 7 and 17  
319 contained one and two breakpoints, respectively. The single breakpoint on Chromosome 7  
320 connected to almost every other breakpoint on the other two chromosomes. In contrast,  
321 HTMCP-03-06-02175 contained three highly-connected breakpoints but none connected with  
322 all three other chromosomes.

323

324 To resolve the copy number patterns that can occur at multi-breakpoint events, we  
325 delved deeper into the structure of HTMCP-03-06-02149 event one (Fig. 4E). This event was  
326 of particular interest because it encompassed, but did not disrupt, the oncogene *MYC*. Read  
327 depth conversions were used to predict absolute copy number ratios for each segment bridged  
328 by a HPV breakpoint pair in the event. The three segments that were connected by non-  
329 heterologous integrants, including the region containing *MYC*, each showed a one copy  
330 increase above the baseline adjacent to the event (Fig. 4E). The segment between the  
331 heterologous integrant breakpoint pair showed a five-copy increase and this breakpoint pair  
332 was associated with heterologous integration, compatible with the notion that amplification  
333 may be involved in the generation of heterologous integrants (Fig. 4E). Thus, each of the five  
334 copies in the concatenated HPV-human amplification have altered HPV integrant structures.  
335 The resulting proposed structure therefore contained seven HPV integrants linking the human  
336 genome segments including two uninterrupted copies of *MYC*, as illustrated in Fig. 4E.

337

### 338 *HPV integration events show consistent orientation-dependent regulatory patterns*

339 ONT sequencing simultaneously yields DNA methylation information along with DNA  
340 sequence data (Simpson et al. 2017). We therefore investigated methylation patterns at HPV  
341 integration events, including within the integrated HPV genome and the flanking human  
342 regions, and related these to the transcriptional status of the event. Methylation of CpGs within  
343 HPV integrants and up to 5 kb on either side of the breakpoints is shown in Fig. 5A-B for dup-  
344 like and del-like integration events (Supplemental Table S9 and S10). The integration events  
345 are oriented according to the direction of HPV transcription. In transcribed events, we  
346 observed that human regions upstream of HPV transcription were methylated, while human  
347 regions downstream of HPV transcription were unmethylated, and that the HPV late genes  
348 (*L1* and *L2*) were methylated, while the early genes (specifically *E6* and *E7*) were  
349 unmethylated after the long control region (LCR) up to the 3' breakpoint (Fig. 5A-B). This

350 pattern was especially pronounced in dup-like events (Fig. 5A). In contrast, events without  
351 evidence of HPV expression tended to lack consistent methylation patterns adjacent to HPV  
352 and contained partially methylated CpGs across the HPV genome (Fig. 5A-B). In all integrants,  
353 the LCR, a non-coding regulatory region that activates HPV early gene transcription (Maki et  
354 al. 1996), was invariably hypomethylated compared to the genic region of HPV (Fig. 5A-B).

355

356 We next investigated how the production of HPV fusion transcripts correlated with  
357 methylation patterns across HPV integration events. Integration often truncates or disrupts the  
358 HPV genome in such a way that the transcription termination signal for the early genes is lost  
359 (normally positioned after *E5*). We would thus expect transcripts produced from such  
360 integrants to read through downstream into the human genome, where transcription might  
361 terminate. We show the epigenetic and transcriptional landscapes of HPV integrants using the  
362 assemblies of three representative events (Fig. 5C). One event occurred in an intron of  
363 *NEGR1* upstream from its 3' exon, and produced a *E7<sup>^</sup>NEGR1* fusion transcript containing  
364 the poly(A) tail of the human gene (Fig. 5C). However, the other two events did not overlap  
365 any genic elements on the same strand. We observed that, across all samples, HPV-human  
366 RNA fusions consistently occurred downstream from the nearest HPV integration breakpoint,  
367 especially when focusing on the most abundantly expressed fusion sites (Fig. 5D), and they  
368 occurred within the demethylated region adjacent to the HPV integrant (Fig 5C). In fact, events  
369 with a demethylated downstream region showed significantly higher normalized expression in  
370 the 5 kb bin downstream from HPV compared to events that were methylated (Wilcoxon,  
371  $p=2.3\times 10^{-5}$ , Fig. 5E). Events that generated HPV fusion transcripts had lower downstream  
372 methylation (Wilcoxon,  $p=1.9\times 10^{-5}$ , Fig. 5F) and higher downstream expression (Wilcoxon,  
373  $p=7.3\times 10^{-9}$ , Fig. 5G) than non-transcribed events, and expression and methylation in the  
374 downstream region were negatively correlated (Pearson's correlation,  $R=-0.42$ ,  $p=0.00086$ ,  
375 Fig. 5H). However, the increased expression adjacent to the HPV integrant was unique to the  
376 downstream region: in events with a demethylated downstream region, there was a significant  
377 imbalance in expression between the upstream and downstream regions, relative to HPV,

378 whereas no such difference was observed in the methylated events (Wilcoxon,  $p=1.6\times 10^{-5}$ ,  
379 Fig. 5I).

380

381 The methylation patterns in the integrated HPV genomes contrasted with the four  
382 samples harboring episomal HPV (Supplemental Fig. S6A), in which viral sequences were  
383 hypomethylated compared to the integrated HPV genomes, particularly in the late genic  
384 region. The hypermethylation of select HPV genes therefore is a common feature of integrated  
385 HPV genomes in our cohort and was not seen in samples containing HPV episomes. In the  
386 three translocation events, HPV methylation did not follow the standard pattern and we  
387 suspect they instead matched the methylation landscape of the human genome  
388 (Supplemental Fig. S6B).

389

#### 390 *Allele-specific expression and methylation patterns in regions of HPV integration*

391 HPV integrates into only one allele, leaving the other unaffected (Karimzadeh et al.  
392 2023). We leveraged this to explore how HPV integration may have allele-specific impacts on  
393 the epigenome and human gene targets on a broader scale. We phased reads into haplotypes  
394 and determined the allele from which the HPV-containing reads originated, and then identified  
395 differentially methylated regions (DMRs) between the two haplotypes across the tumor  
396 genome (Methods). DMR density was compared to a null distribution across the human  
397 chromosomes to identify regions with significantly dense clusters of DMRs, referred to as DMR  
398 hotspots (Methods). HPV integration sites were then intersected with the DMR hotspots to  
399 identify events that potentially affected DNA methylation in an allele-specific manner. Across  
400 the samples in our dataset, 33 integration event loci overlapped autosomal DMR hotspots  
401 (Fig. 6A; Supplemental Fig. S7; Supplemental Table S11). HPV-overlapping DMR hotspots  
402 had a median size of 6.1 megabases (Mb; range 1.8 - 26 Mb) compared to the overall genome-  
403 wide median of 3.9 Mb (range 18 kb - 56 Mb) and mostly overlapped multi-breakpoint events  
404 (Supplemental Fig. S8A-B).

405

406           There was variation in DMR distribution at the integration-associated hotspots, in  
407 regards to their position relative to HPV and in size and density (Fig. 6A). We therefore tested,  
408 in each integration region, if high DMR density was unique to the sample harboring HPV  
409 integration (Methods; Supplemental Fig. S9). Three and six events were significantly enriched  
410 for DMRs on one (i.e. uni- directional) and both sides of HPV (i.e. bi-directional), respectively,  
411 in the integrated sample ( $p_{adj}<0.05$ ; Benjamini-Hochberg corrected permutation test; Fig. 6A;  
412 Supplemental Fig. S10 and S11). Four of the six bi-directional DMR-enriched events produced  
413 fusion transcripts (Fig. 6A; Supplemental Table S11), whereas none of the uni-directional  
414 DMR-enriched events had HPV-human transcripts detected (Fig. 6A; Supplemental Table  
415 S11). The density of DMRs appeared focal around uni-directional DMR-enriched integration  
416 events, affecting <1 Mb adjacent to the HPV integrant, while bi-directional enrichment was  
417 more distributed, spanning 1-8 Mb from the HPV integrant (Fig. 6A; Supplemental Fig. S12).  
418 DMR densities in the remaining 24 events were not statistically different from the same regions  
419 in other cervical cancer samples, suggesting the high density of DMRs in these regions may  
420 occur independently of HPV integration (Fig. 6A).

421

422           We also investigated the direction of DNA methylation changes on the integrated  
423 haplotype. All three uni-directional DMR-enriched integration events (events 1-3) were  
424 hypomethylated on the HPV-integrated allele, while bi-directional DMR-enriched events  
425 showed both hypermethylation (events 31-33) and hypomethylation (events 28-29) (Fig. 6B;  
426 Supplemental Table S11). We note that, though methylation changes generally occurred in a  
427 uniform direction over a broad region of the human genome in the significantly DMR-enriched  
428 events, transcribed events still followed the pattern described in Fig. 5 in the CpGs adjacent  
429 to the HPV-human breakpoints, suggesting that the local and distal methylation changes may  
430 be independent phenomena (Supplemental Fig. S11B).

431

432 We also analyzed the association between HPV integration and DMR density across  
433 our cohort and observed that regions surrounding HPV integration exhibited a significantly  
434 higher density of DMRs compared to 10,000 randomly selected control regions with  
435 comparable GC-content (Supplemental Fig. S13A; Methods). This effect was more  
436 pronounced closer to the integration event, but was significant up to a 2 Mb ( $\pm 1$  Mb) window  
437 (Wilcoxon,  $p_{adj} < 0.05$ , Fig. 6C; Supplemental Fig. S13B), indicating that HPV's impact on DNA  
438 methylation was strongest near the integration event and gradually diminished around  $\pm 1$  Mb  
439 away from the event (Fig. 6C).

440

441 We next searched for nearby genes whose transcription was potentially affected by  
442 HPV integration. We limited our analysis to loci within 1 Mb of HPV integration sites, which is  
443 approximately the window within which DMR enrichment was observed. We identified 2,066  
444 genes near HPV integrants across 68 samples. Of these, 101 genes had outlier expression  
445 levels in 31 integrated samples and 43 integration event loci (Fig. 6D-E; Supplemental Table  
446 S12; Supplemental Fig. S14). We examined these genes for allele-specific expression (ASE),  
447 which would be consistent with a *cis*-acting element (e.g. HPV integration) altering gene  
448 expression on one allele. Of the expression outliers that we could test for ASE, 69% (44/64)  
449 showed ASE (Methods; Supplemental Table S12). 50% of the outlier genes (50/101) were  
450 within 200 kb of the HPV integration site (in 26 samples and 31 events), and these tended to  
451 be overexpressed (48/50 genes; 96%) and have ASE (32/38 tested genes; 84%). Outlier  
452 genes between 200 kb and 1 Mb away (51/101) contained significantly fewer that were  
453 overexpressed (30/51 genes; 59%; Fisher's exact test,  $p = 7.4 \times 10^{-6}$ ), indicating that HPV  
454 integration may have greater influence on gene expression within 200 kb of the event.

455

456 When comparing gene expression in the integrated sample versus the rest of the  
457 cohort, genes overlapping expressed HPV integration events (i.e. HPV-human fusion  
458 transcripts were detected) showed significantly higher upregulation than those overlapping  
459 non-expressed HPV integration events (Fig. 6F). We also observed greater upregulation of

460 non-overlapping genes within 200 kb of expressed HPV integration events (Fig. 6F). Genes  
461 200-500 kb away from HPV integration showed no significant difference between expressed  
462 and non-expressed events (Fig. 6F).

463

464 The nuclear receptor genes *NR4A1* and *NR4A3* (Wenzl et al. 2015) were the genes  
465 most highly overexpressed in an allele-specific manner from HPV-containing haplotypes  
466 (Supplemental Fig. 14; Supplemental Fig. 15A-B). *NR4A1* (Chromosome 12) and *NR4A3*  
467 (Chromosome 9) were detected from two independent integration events in two different  
468 samples. Non-integrated samples showed ASE of *NR4A1* and *NR4A3* (11 and 2,  
469 respectively), but in both cases, the major allele frequency observed in the integrated sample  
470 was much higher than the majority of samples (Supplemental Fig. 15A-B). Two outlier genes,  
471 *NR4A3* and *CBARP*, overlapped significant DMR-enriched events (event 31 and 29; Fig. 6A-  
472 B; Supplemental Table S12), and the remaining seven significant DMR-enriched events were  
473 not associated with changes in gene expression within 200 kb, suggesting that allelic  
474 methylation and gene expression changes may be two independent consequences of HPV  
475 integration.

476

477 To determine how HPV may contribute to the upregulation of oncogenes, we explored  
478 the DNA methylation landscape around HPV integration in the *NR4A3*-activated sample. This  
479 event was a multi-breakpoint event, with one breakpoint upstream of *NR4A3* on Chromosome  
480 9 that connected to a Chromosome 14 locus harboring twelve breakpoints (Supplemental Fig.  
481 S16). The Chromosome 9 locus overlapped a DMR hotspot that was hypermethylated on the  
482 integrated allele (Fig. 6G, top). However, several DMRs near the *NR4A3* promoter were  
483 unmethylated on the integrated allele, including an 821-bp DMR in exon two of *NR4A3* at the  
484 3' end of an unmethylated promoter element (Fig. 6G, bottom), distinguishing it from other  
485 cervical cancer samples we profiled (Supplemental Fig. S17; bottom). The broad methylation  
486 on the HPV-containing allele was also unique to the integrated sample (Supplemental Fig.  
487 S17; top). We thus observed, in the *NR4A3*-activated sample, both HPV-associated large-

488 scale allelic methylation as well as focal demethylation of potential *NR4A3* regulatory  
489 elements, and high expression of *NR4A3* transcripts from the integrated allele.

490

## 491 Discussion

492 HPV integration plays a crucial role in the development of HPV-driven cancers and is  
493 associated with complex dysregulation of the host genome, the nature and effects of which  
494 remain incompletely understood. Our study aimed to provide a detailed overview of genomic  
495 structures and epigenomic and transcriptional changes associated with HPV integration  
496 events in cervical tumors harboring various HPV types. We used ONT long-read sequencing  
497 to generate reads that could span the distances between HPV-human breakpoints, thereby  
498 enabling the reconstruction of complex events. Based on the HPV-containing reads, we  
499 categorized integration events, and related HPV-associated transcription and structural  
500 alterations to local ( $\pm 5$  kb) and distal ( $\pm 1$  Mb) regulatory changes in the tumor.

501

502 The simplest HPV integration structures we observed involved two virus-human  
503 breakpoints. Categories of two-breakpoint events included dup-like and del-like, and, less  
504 frequently, translocation events. Dup-like events, and to a lesser extent, del-like events,  
505 showed consistent methylation patterns on the HPV integrant and adjacent human DNA,  
506 where methylation in the human region upstream of the first breakpoint persisted along HPV  
507 sequences until the LCR, where a hypomethylated state began and extended through the  
508 second breakpoint into the downstream human DNA. A demethylated downstream region was  
509 associated with the production of HPV-human fusion transcripts, which in some cases  
510 contained part of a human gene. *De novo* assembly indicated a circular topology for some  
511 events, which were categorized as putative ecDNAs. Some HPV-human chimeric candidate  
512 ecDNAs contained no human coding sequences, consistent with hypotheses that co-amplified  
513 HPV-human regions can act as cellular super-enhancers (Warburton et al. 2018) and that  
514 ecDNAs can act as mobile enhancer elements (Zhu et al. 2021).

515

516 In 21% of HPV integrants, we observed different rearrangements and copy numbers  
517 of the viral genome between the same two HPV-human breakpoints. The phenomenon of  
518 heterologous integration was first suggested by Kadaja et al. (Kadaja et al. 2009), who showed  
519 that, upon integration, re-replication of the HPV genome generates intermediate structures  
520 whose resolution by DNA damage repair machinery leaves heterologous structures. In our  
521 cohort, heterologous integrants were only detected in regions with amplification between the  
522 breakpoints. This supports Kadaja et al.'s observations and potentially extends them by raising  
523 the possibility that the heterologous integrants may represent different copies within an  
524 amplification rather than different cell populations in a heterogeneous tumor, although we note  
525 that our bulk sequencing approach has limited ability to distinguish between these two  
526 possibilities.

527

528 Of the integration types, multi-breakpoint events were the most prevalent and complex.  
529 These events contained 3 to 33 HPV-human breakpoints across one or more loci, and were  
530 associated with amplified virus-human concatemers and other SVs in the tumor genome, akin  
531 to "heterocateny" described by Akagi et al. (Akagi et al. 2023). We reconstructed the two most  
532 complex multi-breakpoint integration events in our dataset: One contained a single breakpoint  
533 that connected to most other breakpoints in the event, while the other involved three highly-  
534 connected breakpoints that joined different regions of the event, suggesting different  
535 mechanisms. The single highly connected breakpoint suggests that the event arose at a single  
536 point in time within a highly fragmented region, which was reassembled using multiple HPV  
537 integrants to link the fragments, akin to a local chromothripsis event (Li et al. 2020). In contrast,  
538 the event with multiple highly-connected segments is compatible with the notion that the  
539 integration gained complexity during the tumor's evolution (Akagi et al. 2023). When new  
540 double-stranded breaks occur, HPV integrants scattered across the genome may enable  
541 recombination between these regions through shared homology.

542

543 HPV integration has been reported to affect nearby gene expression, both through the  
544 production of viral-human fusion transcripts and by introducing and amplifying enhancers that  
545 can affect host genes (Gagliardi et al. 2020; Brant et al. 2019; Singh et al. 2024; Tian et al.  
546 2023; Adey et al. 2013; Fan et al. 2023; Liu et al. 2023; Groves et al. 2021). We confirmed  
547 local and distal changes to gene regulation and extended these observations into the  
548 methylome. Adjacent to HPV integrants, we uncovered a novel association between the  
549 transcriptional status of the event and its downstream methylation state. High expression  
550 levels downstream of integrants overlapped demethylated regions on the human genome,  
551 which we hypothesize may play a role in termination of HPV-human fusion transcripts. We  
552 also identified allelic changes in methylation and gene expression. HPV integrated regions  
553 were associated with allelic DMRs up to  $\pm 1$  Mb away from the integration event, and in nine  
554 events, we could specifically associate DMR density with the presence of HPV. Both  
555 transcribed and non-transcribed events were associated with DMR-enriched regions.  
556 Meanwhile, gene expression changes tended to occur within 200 kb of a transcribed event.  
557 We hypothesize that the local and distal effects of HPV integration on the methylome may  
558 occur through independent mechanisms. Longer-range methylation changes on the HPV  
559 integrated allele could arise as a host defense mechanism in response to viral integration, as  
560 reported for other viruses and endogenous retroviruses (Watanabe et al. 2015; Jähner and  
561 Jaenisch 1985; Blazkova et al. 2009).

562

563 The methylation patterns within and adjacent to HPV integrants indicate that certain  
564 epigenetic states in the viral and host genomes are necessary to link the two and perhaps  
565 enable the production of HPV-human fusion transcripts. In contrast, the unintegrated HPV  
566 episomes in our study were comparatively hypomethylated, most notably in the late genic  
567 region. It has long been reported that the rate of HPV integration increases with cervical  
568 dysplasia severity (Cheung et al. 2008; Vinokurova et al. 2008). This is consistent with clinical  
569 studies reporting a positive correlation between HPV genome methylation and cervical  
570 precancer severity (Clarke et al. 2012; Wentzensen et al. 2012; Mirabello et al. 2013; Vasiljević

571 et al. 2014; Mirabello et al. 2015; Noyez and van de Wal 1989; Clarke et al. 2018; Bowden et  
572 al. 2019; Liu et al. 2017). Methylation of the viral late genes (*L1/L2*) in particular exhibited high  
573 diagnostic sensitivity for detecting high-grade neoplasms in HPV16-positive women (Bowden  
574 et al. 2019), and we speculate this may be indicative of viral integration as infection persists  
575 and progresses into neoplasia.

576

577 In our cohort, the *MYC* locus on Chromosome 8 (Symer et al. 2022; Gagliardi et al.  
578 2020; Bodelon et al. 2016; Hu et al. 2015) harbored recurrent multi-breakpoint events, which  
579 were associated with increased *MYC* expression in the integrated samples. Chromosome  
580 loops connecting the integration region to *MYC* in a haplotype-specific manner have been  
581 described in HeLa cells (Adey et al. 2013), providing a mechanism by which *MYC* upregulation  
582 may be achieved. Recurrent dup-like integration events on Chromosome 3 within *TP63*  
583 (Kamal et al. 2021; Gagliardi et al. 2020; Bodelon et al. 2016; Liu et al. 2016) were also  
584 observed in our dataset, and were associated with elevated *TP63* expression. A mechanism  
585 has yet to be described, but three of the four integrants were transcribed on the opposite  
586 strand, suggesting an orientation-independent mechanism such as enhancer activation. We  
587 also observed HPV-associated allele-specific activation of *NR4A3*, which was situated 170 kb  
588 downstream of an HPV breakpoint belonging to a multi-breakpoint event spanning two  
589 chromosomes. In salivary gland acinic cell carcinoma, the cancer is driven by translocations  
590 that juxtapose super-enhancers from other chromosomes upstream of *NR4A3* (Haller et al.  
591 2019b, 2019a). We speculate that HPV-driven activation of *NR4A3* may occur through a  
592 similar mechanism. The ASE of *NR4A3* and its distance from the HPV breakpoint also suggest  
593 a physical interaction between enhancer elements involved in the HPV integration event and  
594 *NR4A3*. These enhancers may come from the HPV LCR or from the chromosome segments  
595 involved in the multi-breakpoint event. Further studies are required to determine how HPV  
596 integration leads to the activation of certain genes while driving megabase-scale  
597 hypermethylation on affected haplotypes.

598

599           Although we have explored new aspects of HPV integration biology, our study has  
600 several limitations. First, we did not examine how variations in HPV integrant sequences,  
601 whether from sub-lineages of HPV types or somatic mutations, might influence the regulation  
602 of HPV integration events and their effects on the surrounding genome. Second, our methods  
603 for reconstructing HPV integration events used two conditions to group HPV breakpoints  
604 (Methods), the second of which assumes that closely juxtaposed breakpoints belong to the  
605 same event in the absence of reads spanning the breakpoints. For two-breakpoint events, we  
606 assume that each breakpoint represents opposite ends of the integrant, and from this, we infer  
607 orientation and the direction of transcription. However, for multi-breakpoint events, we do not  
608 assume any specific linkages or orientations unless supported by the reads. Third, we do not  
609 provide orthogonal experimental evidence of predicted ecDNA events and rely on the topology  
610 of the supporting reads. However, several groups have reported circular structures at HPV  
611 integration events and validated these results using specialized techniques (Akagi et al. 2023;  
612 Tian et al. 2023). We note that our informatics methods are limited by read lengths, so larger  
613 ecDNAs are likely missed by our current approach, which relies on reads fully spanning the  
614 event. Read lengths remain a general limitation of our study, particularly as relates to resolving  
615 heterologous integrants and multi-breakpoint events. Lastly, we used bulk long-read WGS in  
616 this study, making it difficult to distinguish whether observations exist within individual cells or  
617 across different cell populations in the tumors. To confidently discern these two possibilities,  
618 a single-cell approach would be required.

619  
620           Overall, our study leveraged long-read sequencing to identify the genomic and  
621 epigenetic impacts of HPV integration in cervical cancer. Recurrent integration-related  
622 structural rearrangements were observed, which had varied effects on the viral and human  
623 genomes. Multi-breakpoint events tended to be structurally complex and harbored  
624 heterologous configurations of the HPV genome linking its integration breakpoints. In dup-like  
625 and del-like integrations, demethylation downstream of the viral LCR was associated with the  
626 production of HPV-human fusion transcripts. We also observed widespread methylation and

627 expression changes in *cis* with HPV integration, both proximal and distal to the integration site.  
628 The methylome in particular showed allelic differences megabases away from HPV  
629 integration. Our findings illustrate the multi-omic dysregulation associated with HPV  
630 integration in cervical cancer and how this may result in oncogenic functions.

631

## 632 **Methods**

### 633 *Ethics approval and consent to participate*

634 The HTMCP patient cohort design was approved by the Fred Hutchinson Cancer  
635 Research Center Institutional Review Board (7662) and complied with ethical regulation;  
636 patient accrual received institutional and governmental approval, and informed consent was  
637 obtained from all patients (Gagliardi et al. 2020). The molecular characterization performed in  
638 this study had approval from the BC Cancer Research Ethics Board (UBC BC Cancer REB  
639 H19-03010) and the Medical College of Wisconsin Institutional Review Board.

640

### 641 *Sample selection*

642 Our sample selection balanced HPV types, homologous recombination deficiency  
643 (HRD) scores (which may impact SV generation) (Telli et al. 2016), and HPV integration  
644 statuses, to what was previously identified across the whole cohort using short-read  
645 sequencing (Gagliardi et al. 2020). Our final dataset included 72 samples from the HTMCP  
646 (Gagliardi et al. 2020), which consisted of 25 HPV16 tumors, 21 HPV18 tumors, 12 HPV45  
647 tumors, and 14 tumors with less common HPV types, approximately reflecting the proportions  
648 of HPV types present in the larger HTMCP cohort. Visual inspection of the HPV genome  
649 alignments was performed to confirm the HPV-containing reads were of high-quality, both in  
650 mapping quality and in length and contiguity. Three samples that were initially sequenced  
651 were excluded in the final dataset due to an insufficient number of HPV-aligning reads,  
652 ambiguous or low-quality alignments, or highly fragmented and uninterpretable reads across

653 the HPV genome due to low N50s and high chimeric rates in the sample. None of these  
654 samples had HPV integration detected by our workflow, but their low-quality alignments  
655 disqualified them from accurately representing episomal HPV.

656

#### 657 *ONT library construction and whole-genome sequencing*

658 HTMCP samples either had sufficient DNA left over from an earlier extraction  
659 (Gagliardi et al. 2020) or required fresh extraction. For archival (frozen) tissues, nucleic acids  
660 were extracted using bead-based or column purification methods. Blue Pippin size-selection  
661 (Sage Science) was performed on 5 µg DNA to deplete shorter DNA molecules (<15 kb) from  
662 the final library to achieve 2 µg of final input DNA at a concentration of 42 ng/µL. Blue Pippin  
663 was not performed on samples with insufficient DNA yields. 41 samples were sequenced on  
664 R9.4 flow cells and 31 samples used R10.4.1 flow cells, depending on when they were  
665 sequenced (Supplemental Table S13). The ONT ligation-based library preparation kit (SQK-  
666 LSK110 for R9 samples and SQK-LSK114 for R10 flow cells) was implemented on a NIMBUS  
667 liquid handling robot (Hamilton), followed by the whole genome PCR-free library construction  
668 for Oxford Nanopore sequencing, as described (Dixon et al. 2023). The libraries were loaded  
669 onto quality-controlled flow cells exhibiting >5000 active pores within 5 days of construction to  
670 preserve the pore-targeting adapter. The library recovery after ligation and bead purifications  
671 was expected to be >40% of input (i.e. >800 ng). For yields 300-500 ng the entire library was  
672 loaded onto a PromethION flow cell. For libraries with >500 ng yield, 2/3 of the library was  
673 loaded initially followed by a nuclease (DNase I) flush of the flow cell to restore the activity of  
674 the pores that had become clogged with DNA. The remaining 1/3 of the library was then loaded  
675 onto the restored flow cell for a maximum total of 750 ng. PromethION flow cells were typically  
676 run for 72 hours for maximal sequencing yield.

677

## 678 *ONT Primary Data Analysis*

679           Raw signal from the PromethION sequencer was basecalled using ONT's Guppy 5  
680 with the "super-accurate" model for the 41 R9 samples, and Dorado (v. 0.6.1)  
681 (<https://github.com/nanoporetech/dorado>) for the 31 R10 samples (Supplemental Table S13).  
682 These sequence data were aligned using minimap (v. 2.15) (Li 2018) to a custom reference  
683 containing GRCh38 with no ALT contigs and 17 HPV types from the HTMCP cohort,  
684 downloaded from the PaVE database (<https://pave.niaid.nih.gov/>) (Van Doorslaer et al. 2017).  
685 Subsequent primary analysis was carried out via a NextFlow workflow (Supplemental Table  
686 S13). Structural variants were called from the aligned BAM using Sniffles (v. 1.0.12b and v.  
687 2.0.7) (Sedlazeck et al. 2018; Smolka et al. 2024) and CuteSV (v. 1.0.12) (Jiang et al. 2020);  
688 however, downstream HPV integration analyses used custom SV detection parameters using  
689 Sniffles, as described later (v. 1.0.12) (Sedlazeck et al. 2018). Small variants were called using  
690 Clair3 (v. 0.1-r8) (Luo et al.) and phased using WhatsHap (v. 1.0) (Martin et al. 2016), with  
691 phase blocks retained for later analyses. For the 41 samples run on R9 flow cells, DNA  
692 methylation (5mC) was called at the read level using Nanopolish (v. 0.13.2) (Simpson 2018)  
693 and phased into haplotypes using NanoMethPhase (v. 1.0) (Akbari et al. 2021). For the 31  
694 samples run on R10 flow cells, methylation calls were extracted from the BAM file using  
695 modbam2bed (v. 0.9.1; <https://github.com/epi2me-labs/modbam2bed>) and phasing  
696 information from WhatsHap (v. 1.0) (Martin et al. 2016) was added as an additional tag on the  
697 BAM (Supplemental Table S13). Statistical testing for DMRs was performed using DSS (v.  
698 2.47.1) (Feng and Wu 2019) in both workflows. No batch effects were observed when  
699 comparing the methylation frequencies at 463 UCSC Genome Browser  
700 (<http://genome.ucsc.edu>) CpG islands (Supplemental Fig. S18).

701

## 702 *Identifying HPV integration breakpoints, integration events, and integration event loci*

703           Sniffles (v. 1.0.12) (Sedlazeck et al. 2018) was used to call translocations on the ONT  
704 long-read WGS using the following specifications to maximize the accuracy for detected HPV

705 integration breakpoints: `max_distance = 50, max_num_splits = -1, report_BND,`  
706 `num_reads_report = -1, min_support = 5, and min_seq_size = 500.` A minimum of 5 consensus  
707 reads was required to detect an HPV integration breakpoint. An R script (v. R 4.0; [http://www.r-](http://www.r-project.org)  
708 [project.org](http://www.r-project.org)) was developed that then iteratively grouped HPV breakpoints together if one or  
709 more reads overlapped between the breakpoints as indicated in the VCF. A distance threshold  
710 of 500 kb was also implemented using BEDTools (v. 2.30.0) (Quinlan 2014) to further group  
711 breakpoints that mapped near to each other but lacked reads long enough to intersect. The  
712 first condition ensured breakpoints that appeared distant to each other on the reference, but  
713 that were physically linked through fusion rearrangements, could be paired together. The  
714 second condition linked breakpoints that mapped near each other but that lacked reads long  
715 enough to span between them. The collection of HPV breakpoints that were grouped together  
716 through these two methods were referred to as an integration event. All read names belonging  
717 to an integration event were retained for later analyses. Integration event loci were defined as  
718 the integration breakpoints within an event that map within 500 kb of each other, as determined  
719 using BEDTools (v. 2.30.0) (Quinlan 2014). Integration events spanning multiple  
720 chromosomes or large genomic distances would have multiple integration event loci. The  
721 integration event loci were used for regional analyses such as finding recurrently integrated  
722 loci, determining expression changes of neighboring genes, and overlapping DMR hotspots.

723

#### 724 *Detection of HPV integration event expression using HPV-human fusion transcripts*

725 Previously described short-read RNA-seq data (Gagliardi et al. 2020) were used to  
726 detect HPV-human breakpoints in fusion transcripts. These data are available for download  
727 through dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>, phs000528), as part of the NCI Cancer  
728 Genome Characterization Initiative (CGCI; phs000235). The data were realigned to the  
729 hg38/HPVs reference genome using STAR (v. 2.7.11) (Dobin et al. 2013) with the following  
730 parameters, as described by Fan et al. (Fan et al. 2023): `--outSAMstrandField intronMotif, --`  
731 `chimSegmentMin 12, --chimJunctionOverhangMin 8, --chimOutJunctionFormat 1, --`

732 alignSJDBoverhangMin 10, --alignMatesGapMax 100000, --alignIntronMax 100000, --  
 733 limitSjdbInsertNsj 1500000, --alignSJstitchMismatchNmax 5 -1 5 5, --outSAMattrRGline  
 734 ID:GRPundef, --chimMultimapScoreRange 3, --chimScoreJunctionNonGTAG 4, --  
 735 chimMultimapNmax 20, --chimNonchimScoreDropMin 10, --peOverlapNbasesMin 12, --  
 736 peOverlapMMp 0.1, --alignInsertionFlush Right, --alignSplicedMateMapLminOverLmate 0 --  
 737 alignSplicedMateMapLmin 30, --chimOutType Junctions WithinBAM SoftClip. RNA junctions  
 738 between the HPV genome and human chromosomes were filtered to those with at least 10  
 739 congruent reads and within 100 kb of an ONT-called integration event using BEDTools (v.  
 740 2.30.0) (Quinlan 2014) and a custom python script. Thus, expressed HPV integration events  
 741 had at least one HPV-human RNA junction with at least 10 reads within 100 kb of an HPV  
 742 breakpoint. The number of supporting reads for each HPV-human junction was used to  
 743 determine the relative expression level of each junction per sample.

744

#### 745 *Comparison of HPV integration calls between ONT and Illumina*

746 Previously described short-read WGS data (Gagliardi et al. 2020), available for  
 747 download through dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>, phs000528) as part of the NCI  
 748 Cancer Genome Characterization Initiative (CGCI; phs000235), were realigned to the  
 749 hg38/HPVs reference genome using minimap2 (v. 2.15) (Li 2018). The short-read SV caller  
 750 Manta (v. 1.6.0) (Chen et al. 2016) was used to call translocations between the human  
 751 chromosomes and HPV genomes, and a minimum congruent read pair threshold of five reads  
 752 was applied with BCFtools (v. 1.15.1) (Danecek et al. 2021). Integration breakpoints within  
 753 500 kb of each other were combined into integration events using BEDTools (v. 2.30.0)  
 754 (Quinlan 2014). Subsequently, integration breakpoint calls for 69 HTMCP samples (those with  
 755 integration detected using one or both technologies) were summed and compared by sample  
 756 and by event. The genomic overlap of integration breakpoint calls for Illumina and ONT was  
 757 determined using the BEDTools (v. 2.27.1) (Quinlan 2014) intersect function against hg38  
 758 annotations obtained from the Table Browser page on the University of California, Santa Cruz

759 website (<http://genome.ucsc.edu>) (Karolchik et al. 2004). Specifically, the GENCODE V43  
760 knownGene annotation (Harrow et al. 2012) filtered to exons or introns, cpplandExt, and  
761 RepeatMasker (Chen 2004) annotations were downloaded as BED files (accessed May 2,  
762 2023).

763

#### 764 *Assembling HPV integration event contigs*

765 Each HPV integration event was assembled into an integration contig for  
766 characterization. The events were first subsetted into event-only BAM files with Picard tools'  
767 (v. 2.26.6; <https://broadinstitute.github.io/picard/>) FilterSamReads function using the read  
768 name text files created when grouping the integration events. The reads were then converted  
769 into FASTQ files using SAMtools (v. 1.12) (Li et al. 2009). Each set of event reads were then  
770 run through the assembler Flye (v. 2.9) (Kolmogorov et al. 2019) with three rounds of polishing.  
771 The assembly was mapped back to the reference chromosome for assembly annotation and  
772 the reads were mapped back to the assembly to check assembly quality using minimap2 (v.  
773 2.23) (Li 2018). Sniffles (v. 1.0.12) (Sedlazeck et al. 2018) was run on the reads aligned to the  
774 assembly to check for rearrangements that may not be assembled correctly, such as  
775 insertions, deletions, and duplications.

776

#### 777 *Detection of potential ecDNAs*

778 Circularity was tested on all events to discover putative ecDNA events that were  
779 predicted to be extrachromosomal. The assemblies of events that successfully assembled  
780 were aligned to the reference genome using minimap2 (v. 2.23) (Li 2018). The alignments of  
781 the event reads were then subtracted using BEDTools (v. 2.30.0) (Quinlan 2014) from the  
782 assembly alignment. If there was minimal to no coverage outside the assembly region and  
783 there was adequate coverage (>2 reads) on each border of the assembly, and the assemblies

784 that contained a contig that was predicted as circular by Flye (v. 2.9) (Kolmogorov et al. 2019),  
785 then circularity was predicted.

786

### 787 *Categorizing HPV integration events*

788 The categorization of the integration events followed the decision chart in Fig. 1c.  
789 Repeats were detected on the FASTA files of the reads using RepeatMasker (v. 4.2.1) (Chen  
790 2004). The output GFF file was then checked for the following repeat sequences: TAR1, ALR,  
791 HSAT4, HSAT5, HSAT6, and SST1. If over 50% of the reads contained repeat sequences,  
792 then the integration event was categorized as a repeat integration. If no repeat was detected,  
793 then the number of breakpoints in the event was counted. Events with more than two  
794 breakpoints were categorized as multi-breakpoint events. Two-breakpoint events were then  
795 categorized as translocation integration if the two breakpoints were on different chromosomes,  
796 del-like integration if the region in between the two breakpoints had no read coverage in the  
797 HPV-containing reads, and dup-like integration if there was read coverage between the  
798 breakpoints. Circularity was tested separately from this categorization based on assembly as  
799 outlined above.

800

### 801 *HPV integrant copy number calling*

802 The HPV-containing read alignments in PAF format were analyzed using an R script  
803 (v. R 4.0; <http://www.r-project.org>) script to determine HPV integrant lengths in a read-specific  
804 manner. Reads containing the same human and HPV breakpoints were grouped together,  
805 allowing a leniency in the exact breakpoints of  $\pm 30$  bp. The HPV alignments that occurred  
806 between two SV-detected breakpoints were extracted from each read and the total length of  
807 the HPV integrant was determined by its cumulative length on the read. If only one breakpoint  
808 was detected, the length of the HPV alignments before or after the break was calculated for  
809 analyses of incomplete HPV integrants. All unique breakpoint pairs (or incomplete  
810 breakpoints) were determined for each sample. HPV integrant sizes for each breakpoint pair

811 were separated into read groups if there was a difference of >300 bp to the next closest size.  
812 Reads with integrant sizes within <300 bp of each other were therefore grouped together.  
813 Each unique breakpoint pair or breakpoint was categorized: “heterologous” if >1 integrant  
814 size group was detected, “partial” if the HPV integrant was <1 HPV genome equivalent in size,  
815 “full” if the HPV integrant was >1 HPV genome equivalent in size, and “incomplete” if the  
816 breakpoint was not paired. The maximum integrant size of heterologous integrants was  
817 determined as the mean size of the largest read group. The maximum size of incomplete  
818 integrants was determined by the read containing the longest HPV alignment before or after  
819 the breakpoint.

820

### 821 *Comparison of two-breakpoint events*

822 The genomic distance between breakpoints was calculated for all two-breakpoint  
823 events by subtracting the reference position of the most 5' breakpoint from the most 3'  
824 breakpoint. A Wilcoxon ranked-sum test was used to calculate the significance of the  
825 difference between dup-like and del-like events. The position of the two-breakpoint events  
826 relative to genic regions was determined by intersecting the region between the two-  
827 breakpoint (event region) with the gene regions (Ensembl 100 GRCh38) using BEDTools (v.  
828 2.30.0) (Quinlan 2014). The events were categorized as “genic” if >90% of the event region  
829 fell within a gene, “partially genic” if the event was between 10-90% within a gene, and  
830 “intergenic” if <10% of the event region was within a gene. The proportions of events that were  
831 genic, partially genic, and intergenic were compared using a Fisher’s Exact test.

832

### 833 *Copy number analysis*

834 The copy number profiles (determined using Illumina WGS) were obtained for HTMCPC-  
835 03-06-02054 and other samples containing translocation events. The Ploidetect (v. 3.0)  
836 (Culibrk et al. 2021) pipeline was run using the “short” sequence type (see

837 <https://github.com/lculibrk/Ploidetect-pipeline>, section 1.1.4). The results were selected for the  
838 first predicted model that did not have a ploidy of 1.

839

#### 840 *Multi-breakpoint event reconstruction and visualization*

841 The multi-breakpoint event loci were visualized using a custom R script (v. R 4.0;  
842 <http://www.r-project.org>) that positioned the HPV integrant connections. The HPV breakpoint  
843 pairs were grouped into their respective events, and the selected multi-breakpoint events in  
844 Fig. 4D were visualized by connecting the breakpoints within a breakpoint pair using an arched  
845 dotted line. The heterologous integrants were colored differently from integrants with a single  
846 integrant structure. The number of connections each breakpoint participated in (i.e. the  
847 number of breakpoint pairs it belonged to) was also indicated by the color of the dot. Read  
848 depth of the example event in Fig. 4E was calculated using SAMtools (v. 1.12) (Li et al. 2009)  
849 in the regions between each breakpoint pair. SV breakpoints were mapped to the HPV  
850 integration events by using overlapping read names, as done for assembling HPV breakpoints  
851 into events, using a custom python script.

852

#### 853 *HPV genome methylation, flanking methylation, and flanking expression*

854 The HPV genome methylation frequencies per CpG were calculated using the HPV-  
855 containing reads from each group of reads belonging to a single HPV integrant. These reads  
856 were used to calculate the average methylation flanking the integrant for del-like and dup-like  
857 integration events. For the flanking regions, methylation frequencies were averaged across  
858 500 bp bins up to 5000 bp upstream and downstream of the breakpoints on the human DNA.  
859 The upstream/downstream directions were oriented according to the strandedness of the HPV  
860 integrant such that the directions were relative to that of HPV transcription. For visualization,  
861 the HPV genome positions were shifted to begin at the *E6* start position and end at the *L1*  
862 stop position, with all LCR positions represented as negative values so they could be  
863 visualized together.

864

865           The stranded position of HPV-human RNA fusion sites was compared to the relative  
866 position of the nearest DNA integration breakpoint. For assembly, we re-aligned RNA-seq data  
867 to the HPV integration event contig using minimap (v. 2.15) (Li 2018) and re-mapped the  
868 methylation and transcript fusion sites for visualization. For del-like and dup-like integration  
869 events, we calculated the RPKM of the RNA-seq reads in the 5 kb region upstream and  
870 downstream from HPV and used these normalized expression values to compare events. We  
871 analyzed a 0-2.5 kb window around HPV for upstream and downstream methylation due to  
872 read length limitations in some events. The average methylation value was calculated by  
873 averaging the methylation frequencies across all reads, and an average methylation frequency  
874 of <25% in the downstream region was considered “unmethylated”.

875

#### 876 *Hotspots of differentially methylated regions*

877           Differentially methylated regions (DMRs) were identified using the DSS (Feng and Wu  
878 2019) 2-sample analysis. DMR hotspots were identified using an R script (v. R 4.0;  
879 <http://www.r-project.org>) that compared the actual density of DMRs across each chromosome  
880 to a Gaussian null distribution of the same number of DMRs. DMR hotspots were defined as  
881 regions where the actual density of DMRs was significantly higher than the null distribution.  
882 These hotspots were intersected with the HPV integration event locations using BEDTools (v.  
883 2.30.0) (Quinlan 2014). Events and hotspots on Chromosome X were excluded. The phase  
884 block borders of the HPV integration events were also identified by intersecting with BEDTools  
885 (v. 2.30.0) (Quinlan 2014), and the haplotype of the integration event was determined as the  
886 haplotype that contained the majority of HPV-containing reads at that integration event locus.  
887 The *P* values for the enrichment of DMRs upstream and downstream from the HPV integration  
888 locus were calculated by generating 1000 random genomic regions of the same size as the  
889 test region (500 kb), calculating the DMR density within these regions for the test sample and  
890 all other samples, calculating the fold change in DMR density between the test sample vs. the

891 mean of all other samples in each region, and determining the percentile of the test region's  
892 fold change compared to the null distribution of the 1000 random regions.

893

#### 894 *Association of DMRs with HPV integration across tumors*

895 The DMR density at five window sizes (100 kb, 500 kb, 1 Mb, 2 Mb, and 5 Mb) was  
896 calculated by dividing the number of bases within a DMR by the window size. We compared  
897 the HPV-integrated loci ( $n=147$ ) to a control set ( $n=10,000$ ) to determine significance. The  
898 control set was generated by first calculating the GC-content distribution of the HPV-integrated  
899 test regions HPV-integrated at the five window sizes. We then measured the background  
900 distribution of GC-content in sliding windows of the same window sizes (100 kb, 500 kb, 1 Mb,  
901 2 Mb, and 5 Mb) across the genome (sliding by 100 kb for each. We selected 10,000 control  
902 regions to create a GC-content distribution that mimics the test regions' distribution and  
903 calculated the DMR density in randomly selected samples for each. We then compared the  
904 DMR density at the test regions to the control regions using a Benjamini-Hochberg-adjusted  
905 Wilcoxon rank-sum test.

906

#### 907 *Accession of gene expression data*

908 The gene expression data for the HTMCP cohort (Gagliardi et al. 2020) was  
909 downloaded from the GDC data portal on August 31st 2023 (<https://gdc.cancer.gov/>). The  
910 transcripts per million (TPM) normalized values were used to create an expression matrix of  
911 the 72 samples used in this study.

912

#### 913 *Gene expression outliers and allele-specific expression analysis*

914 The distance between protein-coding genes (Ensembl 100 GRCh38) and integration  
915 event loci was calculated using BEDTools (v. 2.30.0) (Quinlan 2014). Genes that were within  
916  $\pm 1$  Mb of the integration events were tested to see if they were expression outliers, i.e. more

917 than 1.5 IQR below Q1 or 1.5 IQR above Q3. The fold change of the outliers was calculated  
918 as  $\log_2(\text{TPM of the tested sample}/\text{TPM of the median sample})$ . ASE was determined using  
919 IMPALA using the phased VCF (Chang et al. 2023). Genes with ASE were defined as genes  
920 that had a major allele frequency threshold greater than 0.65 and an adjusted  $P$  value less  
921 than 0.05. Genes that were unable to be tested for ASE did not contain a phased single  
922 nucleotide variant (SNV) within the genic region.

923

## 924 *Statistical analyses*

925 No sample sizes other than previously specified molecular features of the initial dataset  
926 were predetermined. Unless otherwise stated, all statistical tests correspond to two-sided  
927 tests.  $P$  value methods and multiple-test correction are reported in the text. Wilcoxon in the  
928 text refers to the Wilcoxon rank-sum test.

929

## 930 Data Access

931 The whole genome long-read sequencing data generated in this study have been  
932 submitted to the dbGaP database (<https://www.ncbi.nlm.nih.gov/gap/>) under accession  
933 number phs003780 ([https://www.ncbi.nlm.nih.gov/projects/gap/cgi-  
934 bin/study.cgi?study\\_id=phs003780.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003780.v1.p1)). Data analysis methods and source code can be  
935 found as Supplementary Code and as open-source workflows on GitHub  
936 ([https://github.com/MarraLab/VPorter HPV integration](https://github.com/MarraLab/VPorter_HP_V_integration)).

937

## 938 Competing Interests Statement

939 V.L.P., K.O., and S.J.M.J. received payment in the form of travel and accommodation from  
940 Oxford Nanopore Technologies (ONT) to attend and speak at ONT's annual meeting.

941

## 942 Acknowledgements

943 We are grateful for contributions from the other members of various groups at  
944 Canada's Michael Smith Genome Sciences Centre, including those from the Biospecimen,  
945 Library Construction, Sequencing, Bioinformatics, Technology Development, Quality  
946 Assurance, LIMS, Purchasing and Project Management teams. We are grateful to Dr. Adi Steif  
947 for constructive analytical suggestions. The results published here are in whole or part based  
948 upon data generated by the Cancer Genome Characterization Initiative (phs000235), the HIV+  
949 Tumor Molecular Characterization Project - Cervical Cancer (HTMCP - CC) project, developed  
950 by the NCI. The data used for this analysis are available at <https://www.ncbi.nlm.nih.gov/gap/>  
951 and <https://gdc.cancer.gov/>. Information about CGCI projects can be found at  
952 <https://ocg.cancer.gov/programs/cgci>. V.L.P., L.C., and S.M. were the recipients of CIHR  
953 Frederick Banting and Charles Best Canada Graduate Scholarships GSD-152374, GSD-  
954 164207, and FBD-187583, respectively. M.N. was supported by postdoctoral fellowship  
955 awards from the Canadian Institutes of Health Research (FRN-188098) and Michael Smith  
956 Health Research BC (RT-2023-3168). S.J.M.J. is the recipient of the Canada Research Chair  
957 in Computational Genomics. M.A.M. is the UBC Canada Research Chair in Genome Science.  
958 This work was supported in part by funding provided by the Canadian Institutes for Health  
959 Research (CIHR award FDN-143288 and PJT-180410) to M.A.M. and NCI R21CA241013 and  
960 R01CA262198 to J.S.R. and MCW OBGYN WHRP fund. M.A.M. is The Terry Fox Leader in  
961 Cancer Genome Science. This study was conducted with the financial support of The Terry  
962 Fox Research Institute and the Terry Fox Foundation. The views expressed in the publication  
963 are the views of the authors and do not necessarily reflect those of The Terry Fox Research  
964 Institute or the Terry Fox Foundation.

965

## 966 Author Contributions

967 This project was conceived by V.L.P. and M.A.M. Data were generated by Canada's  
 968 Michael Smith Genome Sciences Centre at BC Cancer and analyses were performed by  
 969 V.L.P., M.N., K.O., and S.M. M.A.M. supervised this work. Contribution to methods used in  
 970 data analyses: V.L.P., M.N., K.O., R.C., L.C., Z.H., G.C., J.F., K.M.N., V.A., I.B., and S.J.M.J.  
 971 Contribution to data generation: K.O., M.N., R.C., M.I., R.M., S-W.T., S.K.C., J.H., R.M., E.C.,  
 972 K.L.M., A.J.M., S.J.M.J., and J.S.R. Cohort and clinical data collection: C.N., J.O, and M.O.  
 973 V.L.P., M.A.M, and M.N. wrote the manuscript. All authors reviewed and edited the  
 974 manuscript.

975

## 976 References

- 977 Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J.  
 978 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer  
 979 cell line. *Nature* **500**: 207–211.
- 980 Akagi K, Li J, Broutian TR, Padilla-Nash H, Xiao W, Jiang B, Rocco JW, Teknos TN, Kumar  
 981 B, Wangsa D, et al. 2014. Genome-wide analysis of HPV integration in human  
 982 cancers reveals recurrent, focal genomic instability. *Genome Res* **24**: 185–199.
- 983 Akagi K, Symer DE, Mahmoud M, Jiang B, Goodwin S, Wangsa D, Li Z, Xiao W, Dunn JD,  
 984 Ried T, et al. 2023. Intratumoral Heterogeneity and Clonal Evolution Induced by HPV  
 985 Integration. *Cancer Discov* **13**: 910–927.
- 986 Akbari V, Garant J-M, O'Neill K, Pandoh P, Moore R, Marra MA, Hirst M, Jones SJM. 2021.  
 987 Megabase-scale methylation phasing using nanopore long reads and  
 988 NanoMethPhase. *Genome Biol* **22**: 68.
- 989 Blazkova J, Trejbalova K, Gondois-Rey F, Halfon P, Philibert P, Guiguen A, Verdin E, Olive  
 990 D, Van Lint C, Hejnar J, et al. 2009. CpG methylation controls reactivation of HIV  
 991 from latency. *PLoS Pathog* **5**: e1000554.
- 992 Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. 2016. Genomic  
 993 characterization of viral integration sites in HPV-related cancers. *Int J Cancer* **139**:  
 994 2001–2011.
- 995 Bowden SJ, Kalliala I, Veroniki AA, Arbyn M, Mitra A, Lathouras K, Mirabello L, Chadeau-  
 996 Hyam M, Paraskevidis E, Flanagan JM, et al. 2019. The use of human  
 997 papillomavirus DNA methylation in cervical intraepithelial neoplasia: A systematic  
 998 review and meta-analysis. *EBioMedicine* **50**: 246–259.
- 999 Brant AC, Menezes AN, Felix SP, de Almeida LM, Sammeth M, Moreira MAM. 2019.  
 1000 Characterization of HPV integration, viral gene expression and E6E7 alternative  
 1001 transcripts by RNA-Seq: A descriptive study in invasive cervical cancer. *Genomics*  
 1002 **111**: 1853–1861.

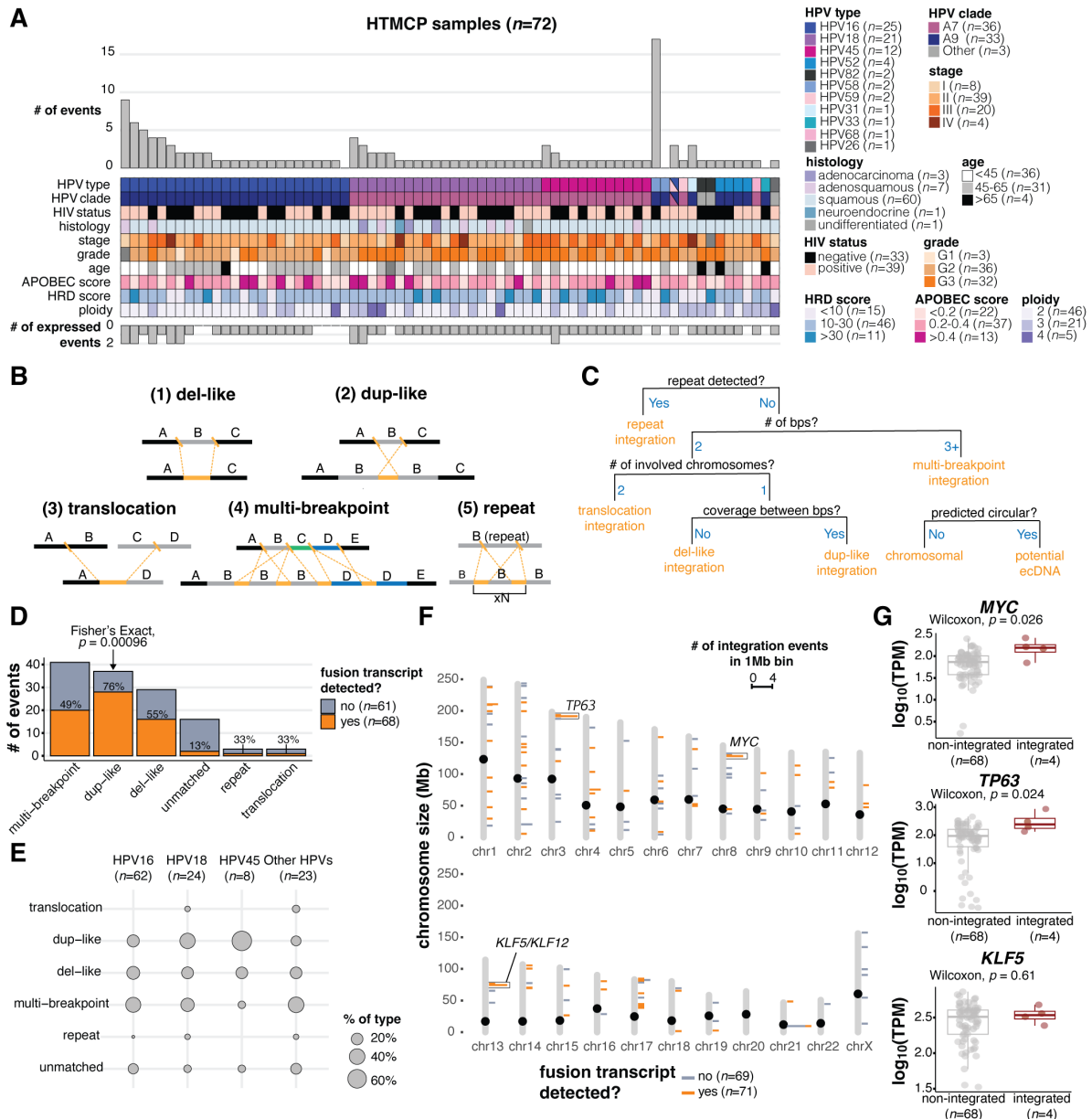
- 1003 Bruni L, Diaz M, Barrionuevo-Rosas L, Herrero R, Bray F, Xavier Bosch F, de Sanjosé S,  
1004 Castellsagué X. 2016. Global estimates of human papillomavirus vaccination  
1005 coverage by region and income level: a pooled analysis. *The Lancet Global Health* **4**:  
1006 e453–e463. [http://dx.doi.org/10.1016/s2214-109x\(16\)30099-7](http://dx.doi.org/10.1016/s2214-109x(16)30099-7).
- 1007 Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical  
1008 Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman  
1009 Research Institute of City of Hope, Buck Institute for Research on Aging, Canada's  
1010 Michael Smith Genome Sciences Centre, Harvard Medical School, Helen F. Graham  
1011 Cancer Center & Research Institute at Christiana Care Health Services, et al. 2017.  
1012 Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**:  
1013 378–384.
- 1014 Chang G, Porter VL, O'Neill K, Culibrk L, Akbari V, Marra MA, Jones SJM. 2023. IMPALA: A  
1015 Comprehensive Pipeline for Detecting and Elucidating Mechanisms of Allele Specific  
1016 Expression in Cancer. *bioRxiv* 2023.09.11.555771.  
1017 <https://www.biorxiv.org/content/biorxiv/early/2023/09/12/2023.09.11.555771>  
1018 (Accessed September 18, 2023).
- 1019 Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences.  
1020 *Curr Protoc Bioinformatics* **Chapter 4**: Unit 4.10.
- 1021 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak  
1022 S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for  
1023 germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–1222.  
1024 <http://dx.doi.org/10.1093/bioinformatics/btv710>.
- 1025 Cheung JLK, Cheung TH, Tang JWT, Chan PKS. 2008. Increase of integration events and  
1026 infection loads of human papillomavirus type 52 with lesion severity from low-grade  
1027 cervical lesion to invasive cancer. *J Clin Microbiol* **46**: 1356–1362.
- 1028 Clarke MA, Gradissimo A, Schiffman M, Lam J, Sollecito CC, Fetterman B, Lorey T, Poitras  
1029 N, Raine-Bennett TR, Castle PE, et al. 2018. Human Papillomavirus DNA  
1030 Methylation as a Biomarker for Cervical Precancer: Consistency across 12  
1031 Genotypes and Potential Impact on Management of HPV-Positive Women. *Clin  
1032 Cancer Res* **24**: 2194–2202.
- 1033 Clarke MA, Wentzensen N, Mirabello L, Ghosh A, Wacholder S, Harari A, Lorincz A,  
1034 Schiffman M, Burk RD. 2012. Human papillomavirus DNA methylation as a potential  
1035 biomarker for cervical cancer. *Cancer Epidemiol Biomarkers Prev* **21**: 2125–2137.
- 1036 Culibrk L, Grewal JK, Pleasance ED, Williamson L, Mungall K, Laskin J, Marra MA, Jones  
1037 SJM. 2021. Ploidetect enables pan-cancer analysis of the causes and impacts of  
1038 chromosomal instability. *bioRxiv* 2021.08.06.455329.  
1039 <https://www.biorxiv.org/content/biorxiv/early/2021/08/08/2021.08.06.455329>  
1040 (Accessed September 18, 2023).
- 1041 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,  
1042 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools.  
1043 *Gigascience* **10**. <http://dx.doi.org/10.1093/gigascience/giab008>.
- 1044 de Sanjosé S, Serrano B, Tous S, Alejo M, Lloveras B, Quirós B, Clavero O, Vidal A,  
1045 Ferrándiz-Pulido C, Pavón MA, et al. 2018. Burden of human papillomavirus (HPV)-  
1046 related cancers attributable to HPVs 6/11/16/18/31/33/45/52 and 58. *JNCI cancer  
1047 spectrum* **2**: ky045.

- 1048 Dixon K, Shen Y, O'Neill K, Mungall KL, Chan S, Bilobram S, Zhang W, Bezeau M, Sharma  
1049 A, Fok A, et al. 2023. Defining the heterogeneity of unbalanced structural variation  
1050 underlying breast cancer susceptibility by nanopore genome sequencing. *Eur J Hum*  
1051 *Genet* **31**: 602–606.
- 1052 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,  
1053 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:  
1054 15–21.
- 1055 Falcaro M, Castañón A, Ndlela B, Checchi M, Soldan K, Lopez-Bernal J, Elliss-Brookes L,  
1056 Sasieni P. 2021. The effects of the national HPV vaccination programme in England,  
1057 UK, on cervical cancer and grade 3 cervical intraepithelial neoplasia incidence: a  
1058 register-based observational study. *Lancet* **398**: 2084–2092.
- 1059 Fan J, Fu Y, Peng W, Li X, Shen Y, Guo E, Lu F, Zhou S, Liu S, Yang B, et al. 2023. Multi-  
1060 omics characterization of silent and productive HPV integration in cervical cancer.  
1061 *Cell Genomics* **3**: 100211.
- 1062 Feng H, Wu H. 2019. Differential methylation analysis for bisulfite sequencing using DSS.  
1063 *Quant Biol* **7**: 327–334.
- 1064 Gagliardi A, Porter VL, Zong Z, Bowlby R, Titmuss E, Namirembe C, Griner NB, Petrello H,  
1065 Bowen J, Chan SK, et al. 2020. Analysis of Ugandan cervical carcinomas identifies  
1066 human papillomavirus clade-specific epigenome and transcriptome landscapes. *Nat*  
1067 *Genet* **52**: 800–810.
- 1068 Groves IJ, Drane ELA, Michalski M, Monahan JM, Scarpini CG, Smith SP, Bussotti G,  
1069 Várnai C, Schoenfelder S, Fraser P, et al. 2021. Short- and long-range cis  
1070 interactions between integrated HPV genomes and cellular chromatin dysregulate  
1071 host gene expression in early cervical carcinogenesis. *PLoS Pathog* **17**: e1009875.
- 1072 Groves IJ, Knight ELA, Ang QY, Scarpini CG, Coleman N. 2016. HPV16 oncogene  
1073 expression levels during early cervical carcinogenesis are determined by the balance  
1074 of epigenetic chromatin modifications at the integrated virus genome. *Oncogene* **35**:  
1075 4773–4786. <http://dx.doi.org/10.1038/onc.2016.8>.
- 1076 Haller F, Bieg M, Will R, Körner C, Weichenhan D, Bott A, Ishaque N, Lutsik P, Moskalev  
1077 EA, Mueller SK, et al. 2019a. Enhancer hijacking activates oncogenic transcription  
1078 factor NR4A3 in acinic cell carcinomas of the salivary glands. *Nat Commun* **10**: 368.
- 1079 Haller F, Skálová A, Ihrler S, Märkl B, Bieg M, Moskalev EA, Erber R, Blank S, Winkelmann  
1080 C, Hebele S, et al. 2019b. Nuclear NR4A3 Immunostaining Is a Specific and  
1081 Sensitive Novel Marker for Acinic Cell Carcinoma of the Salivary Glands. *Am J Surg*  
1082 *Pathol* **43**: 1264–1272.
- 1083 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell  
1084 D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome  
1085 annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.
- 1086 Hu Z, Zhu D, Wang W, Li W, Jia W, Zeng X, Ding W, Yu L, Wang X, Wang L, et al. 2015.  
1087 Genome-wide profiling of HPV integration in cervical cancer identifies clustered  
1088 genomic hot spots and a potential microhomology-mediated integration mechanism.  
1089 *Nat Genet* **47**: 158–163.
- 1090 Huang L-W, Chao S-L, Lee B-H. 2008. Integration of human papillomavirus type-16 and  
1091 type-18 is a very early event in cervical carcinogenesis. *J Clin Pathol* **61**: 627–631.

- 1092 Iden M, Tsaih S-W, Huang Y-W, Liu P, Xiao M, Flister MJ, Rader JS. 2021. Multi-omics  
1093 mapping of human papillomavirus integration sites illuminates novel cervical cancer  
1094 target genes. *Br J Cancer* **125**: 1408–1419.
- 1095 Jähner D, Jaenisch R. 1985. Retrovirus-induced de novo methylation of flanking host  
1096 sequences correlates with gene inactivity. *Nature* **315**: 594–597.
- 1097 Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based  
1098 human genomic structural variation detection with cuteSV. *Genome Biol* **21**: 189.
- 1099 Kadaja M, Isok-Paas H, Laos T, Ustav E, Ustav M. 2009. Mechanism of genomic instability  
1100 in cells infected with the high-risk human papillomaviruses. *PLoS Pathog* **5**.  
1101 <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2666264/>.
- 1102 Kamal M, Lameiras S, Deloger M, Morel A, Vacher S, Lecerf C, Dupain C, Jeannot E, Girard  
1103 E, Baulande S, et al. 2021. Human papilloma virus (HPV) integration signature in  
1104 Cervical Cancer: identification of MACROD2 gene as HPV hot spot integration site.  
1105 *Br J Cancer* **124**: 777–785.
- 1106 Karimzadeh M, Arlidge C, Rostami A, Lupien M, Bratman SV, Hoffman MM. 2023. Human  
1107 papillomavirus integration transforms chromatin to drive oncogenesis. *Genome Biol*  
1108 **24**: 142.
- 1109 Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004.  
1110 The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-6.
- 1111 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using  
1112 repeat graphs. *Nat Biotechnol* **37**: 540–546.
- 1113 Kovaka S, Ou S, Jenike KM, Schatz MC. 2023. Approaching complete genomes,  
1114 transcriptomes and epi-omes with accurate long-read sequencing. *Nat Methods* **20**:  
1115 12–16.
- 1116 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:  
1117 3094–3100.
- 1118 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin  
1119 R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence  
1120 Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- 1121 Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S,  
1122 Korbel JO, Haber JE, et al. 2020. Patterns of somatic structural variation in human  
1123 cancer genomes. *Nature* **578**: 112–121.
- 1124 Liu M, Han Z, Zhi Y, Ruan Y, Cao G, Wang G, Xu X, Mu J, Kang J, Dai F, et al. 2023. Long-  
1125 read sequencing reveals oncogenic mechanism of HPV-human fusion transcripts in  
1126 cervical cancer. *Transl Res* **253**: 80–94.
- 1127 Liu P, Iden M, Fye S, Huang Y-W, Hopp E, Chu C, Lu Y, Rader JS. 2017. Targeted, Deep  
1128 Sequencing Reveals Full Methylation Profiles of Multiple HPV Types and Potential  
1129 Biomarkers for Cervical Cancer Progression. *Cancer Epidemiol Biomarkers Prev* **26**:  
1130 642–650.
- 1131 Liu Y, Lu Z, Xu R, Ke Y. 2016. Comprehensive mapping of the human papillomavirus (HPV)  
1132 DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget*  
1133 **7**: 5852–5864.

- 1134 Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W. Clair: Exploring the  
1135 limit of using a deep neural network on pileup data for germline variant calling.  
1136 <http://dx.doi.org/10.1101/865782>.
- 1137 Maki H, Fujikawa-Adachi K, Yoshie O. 1996. Evidence for a promoter-like activity in the short  
1138 non-coding region of human papillomaviruses. *J Gen Virol* **77** ( Pt 3): 453–458.
- 1139 Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall  
1140 T. 2016. WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050.  
1141 <https://www.biorxiv.org/content/10.1101/085050v2> (Accessed May 1, 2023).
- 1142 Mima M, Okabe A, Hoshii T, Nakagawa T, Kurokawa T, Kondo S, Mizokami H, Fukuyo M,  
1143 Fujiki R, Rahmutulla B, et al. 2023. Tumorigenic activation around HPV integrated  
1144 sites in head and neck squamous cell carcinoma. *Int J Cancer* **152**: 1847–1862.
- 1145 Mirabello L, Frimer M, Harari A, McAndrew T, Smith B, Chen Z, Wentzensen N, Wacholder  
1146 S, Castle PE, Raine-Bennett T, et al. 2015. HPV16 methyl-haplotypes determined by  
1147 a novel next-generation sequencing method are associated with cervical precancer.  
1148 *Int J Cancer* **136**: E146-53.
- 1149 Mirabello L, Schiffman M, Ghosh A, Rodriguez AC, Vasiljevic N, Wentzensen N, Herrero R,  
1150 Hildesheim A, Wacholder S, Scibior-Bentkowska D, et al. 2013. Elevated methylation  
1151 of HPV16 DNA is associated with the development of high grade cervical  
1152 intraepithelial neoplasia. *Int J Cancer* **132**: 1412–1422.
- 1153 Noyez L, van de Wal HJ. 1989. Perioperative morbidity and mortality of coronary artery  
1154 surgery after the age of 70 years. *J Cardiovasc Surg* **30**: 981–984.
- 1155 Quinlan AR. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr*  
1156 *Protoc Bioinformatics* **47**: 11.12.1-34.
- 1157 Rossi NM, Dai J, Xie Y, Wangsa D, Heselmeyer-Haddad K, Lou H, Boland JF, Yeager M,  
1158 Orozco R, Freites EA, et al. 2023. Extrachromosomal Amplification of Human  
1159 Papillomavirus Episomes Is a Mechanism of Cervical Carcinogenesis. *Cancer Res*  
1160 **83**: 1768–1781.
- 1161 Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz  
1162 MC. 2018. Accurate detection of complex structural variations using single-molecule  
1163 sequencing. *Nat Methods* **15**: 461–468.
- 1164 Simpson J. 2018. Nanopolish: Signal-level algorithms for MinION data. *Github Available at:*  
1165 <https://github.com/jts/nanopolish> [Accessed January 10, 2019].
- 1166 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA  
1167 cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.
- 1168 Singh AK, Walavalkar K, Tavernari D, Ciriello G, Notani D, Sabarinathan R. 2024. Cis-  
1169 regulatory effect of HPV integration is constrained by host chromatin architecture in  
1170 cervical cancers. *Mol Oncol* **18**: 1189–1208.
- 1171 Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E,  
1172 Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level  
1173 structural variants with Sniffles2. *Nat Biotechnol*. <http://dx.doi.org/10.1038/s41587-023-02024-y>.  
1174
- 1175 Symer DE, Akagi K, Geiger HM, Song Y, Li G, Emde A-K, Xiao W, Jiang B, Corvelo A,

- 1176 Toussaint NC, et al. 2022. Diverse tumorigenic consequences of human  
1177 papillomavirus integration in primary oropharyngeal cancers. *Genome Res* **32**: 55–  
1178 70.
- 1179 Telli ML, Timms KM, Reid J, Hennessy B, Mills GB, Jensen KC, Szallasi Z, Barry WT, Winer  
1180 EP, Tung NM, et al. 2016. Homologous Recombination Deficiency (HRD) Score  
1181 Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients  
1182 with Triple-Negative Breast Cancer. *Clin Cancer Res* **22**: 3764–3773.
- 1183 Tian R, Huang Z, Li L, Yuan J, Zhang Q, Meng L, Lang B, Hong Y, Zhong C, Tian X, et al.  
1184 2023. HPV integration generates a cellular super-enhancer which functions as  
1185 ecDNA to regulate genome-wide transcription. *Nucleic Acids Res.*  
1186 <http://dx.doi.org/10.1093/nar/gkad105>.
- 1187 Van Doorslaer K, Li Z, Xirasagar S, Maes P, Kaminsky D, Liou D, Sun Q, Kaur R, Huyen Y,  
1188 McBride AA. 2017. The Papillomavirus Episteme: a major update to the  
1189 papillomavirus sequence database. *Nucleic Acids Res* **45**: D499–D506.
- 1190 Vasiljević N, Scibior-Bentkowska D, Brentnall A, Cuzick J, Lorincz A. 2014. A comparison of  
1191 methylation levels in HPV18, HPV31 and HPV33 genomes reveals similar  
1192 associations with cervical precancers. *J Clin Virol* **59**: 161–166.
- 1193 Vinokurova S, Wentzensen N, Kraus I, Klaes R, Driesch C, Melsheimer P, Kisseljov F, Dürst  
1194 M, Schneider A, von Knebel Doeberitz M. 2008. Type-dependent integration  
1195 frequency of human papillomavirus genomes in cervical lesions. *Cancer Res* **68**:  
1196 307–313.
- 1197 Warburton A, Markowitz TE, Katz JP, Pipas JM, McBride AA. 2021. Recurrent integration of  
1198 human papillomavirus genomes at transcriptional regulatory hubs. *NPJ Genom Med*  
1199 **6**: 101.
- 1200 Warburton A, Redmond CJ, Dooley KE, Fu H, Gillison ML, Akagi K, Symer DE, Aladjem MI,  
1201 McBride AA. 2018. HPV integration hijacks and multimerizes a cellular enhancer to  
1202 generate a viral-cellular super-enhancer that drives high viral oncogene expression.  
1203 *PLoS Genet* **14**: e1007179.
- 1204 Watanabe Y, Yamamoto H, Oikawa R, Toyota M, Yamamoto M, Kokudo N, Tanaka S, Aii S,  
1205 Yotsuyanagi H, Koike K, et al. 2015. DNA methylation at hepatitis B viral integrants is  
1206 associated with methylation at flanking human genomic sequences. *Genome Res* **25**:  
1207 328–337.
- 1208 Wentzensen N, Sun C, Ghosh A, Kinney W, Mirabello L, Wacholder S, Shaber R, LaMere B,  
1209 Clarke M, Lorincz AT, et al. 2012. Methylation of HPV18, HPV31, and HPV45  
1210 genomes and cervical intraepithelial neoplasia grade 3. *J Natl Cancer Inst* **104**:  
1211 1738–1749.
- 1212 Wenzl K, Troppan K, Neumeister P, Deutsch AJA. 2015. The nuclear orphan receptor  
1213 NR4A1 and NR4A3 as tumor suppressors in hematologic neoplasms. *Curr Drug*  
1214 *Targets* **16**: 38–46.
- 1215 Zhu Y, Gujar AD, Wong C-H, Tjong H, Ngan CY, Gong L, Chen Y-A, Kim H, Liu J, Li M, et  
1216 al. 2021. Oncogenic extrachromosomal DNA functions as mobile enhancers to  
1217 globally amplify chromosomal transcription. *Cancer Cell* **39**: 694-707.e7.
- 1218

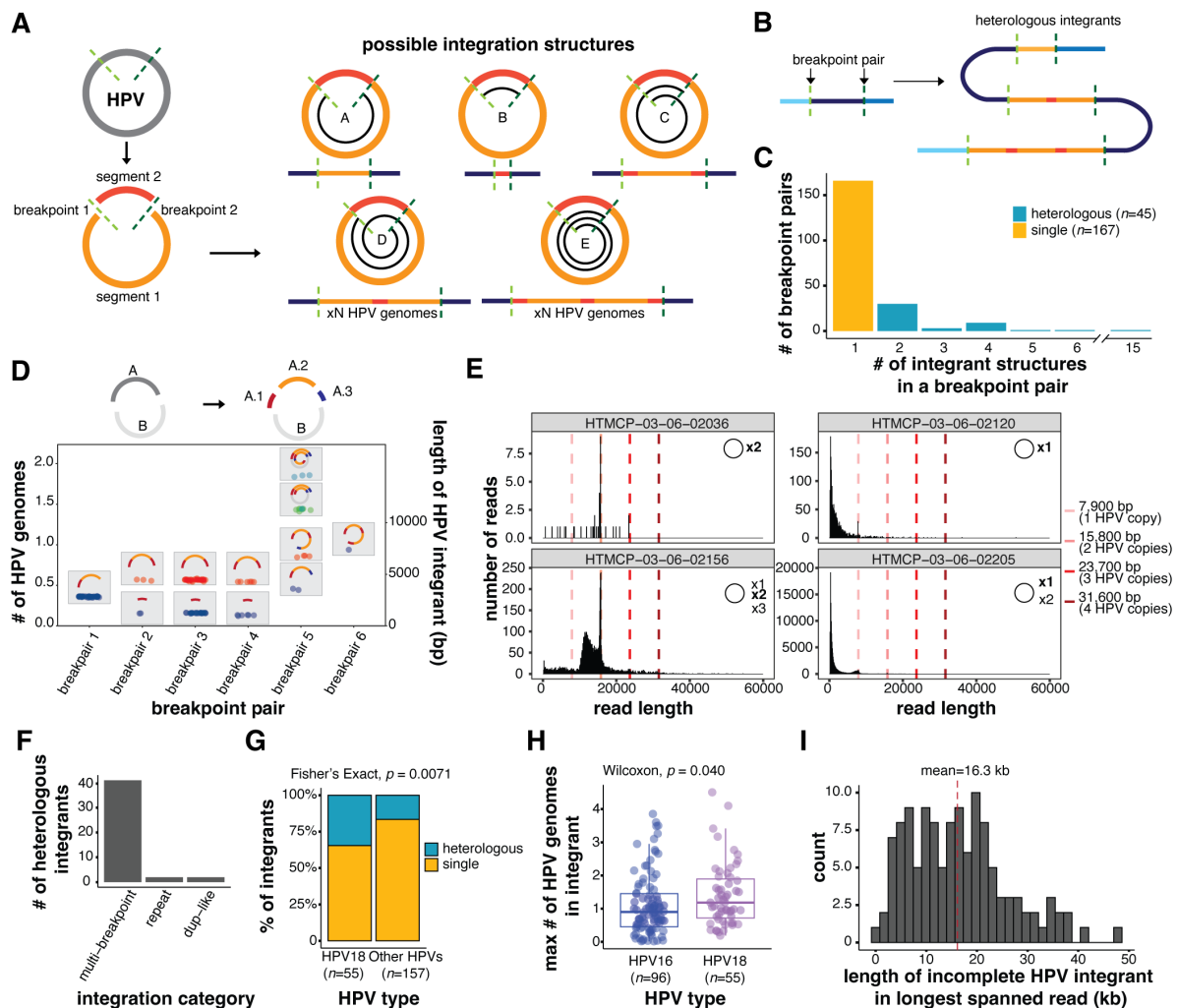


1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233

**Figure 1.** Detection and categorization of HPV integration events in cervical cancers. (A) The number of HPV integration events detected in the DNA (# of events) and detected in DNA and RNA (# of expressed events) across the HTMCP samples, as well as clinical and molecular characteristics. (B) Schematic illustrations of integration event categories. Orange segments indicate the insertion of HPV integrants. (C) Decision chart for categorizing HPV integration events (“bps” = breakpoints). (D) The frequency of the HPV integration categories across the samples. The percentage of events that produce an HPV-human fusion transcript is indicated for each integration type. (E) The percentage of events belonging to each integration category for HPV16, HPV18, HPV45, and all other HPV types. (F) The genomic locations of integration events across the cohort, colored by the transcriptional status of the event. Bins with 2+ integration events are highlighted with boxes. Notable cancer genes within bins recurrently affected by integration are indicated. (G) Gene expression differences between samples with HPV integration within 1 Mb of *MYC*, *TP63*, and *KLF12* compared to the remaining samples in the dataset. Box plots represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 IQR below Q1 or 1.5 IQR above Q3). *P* values were calculated using the Wilcoxon rank-sum test.

1234

1235  
1236

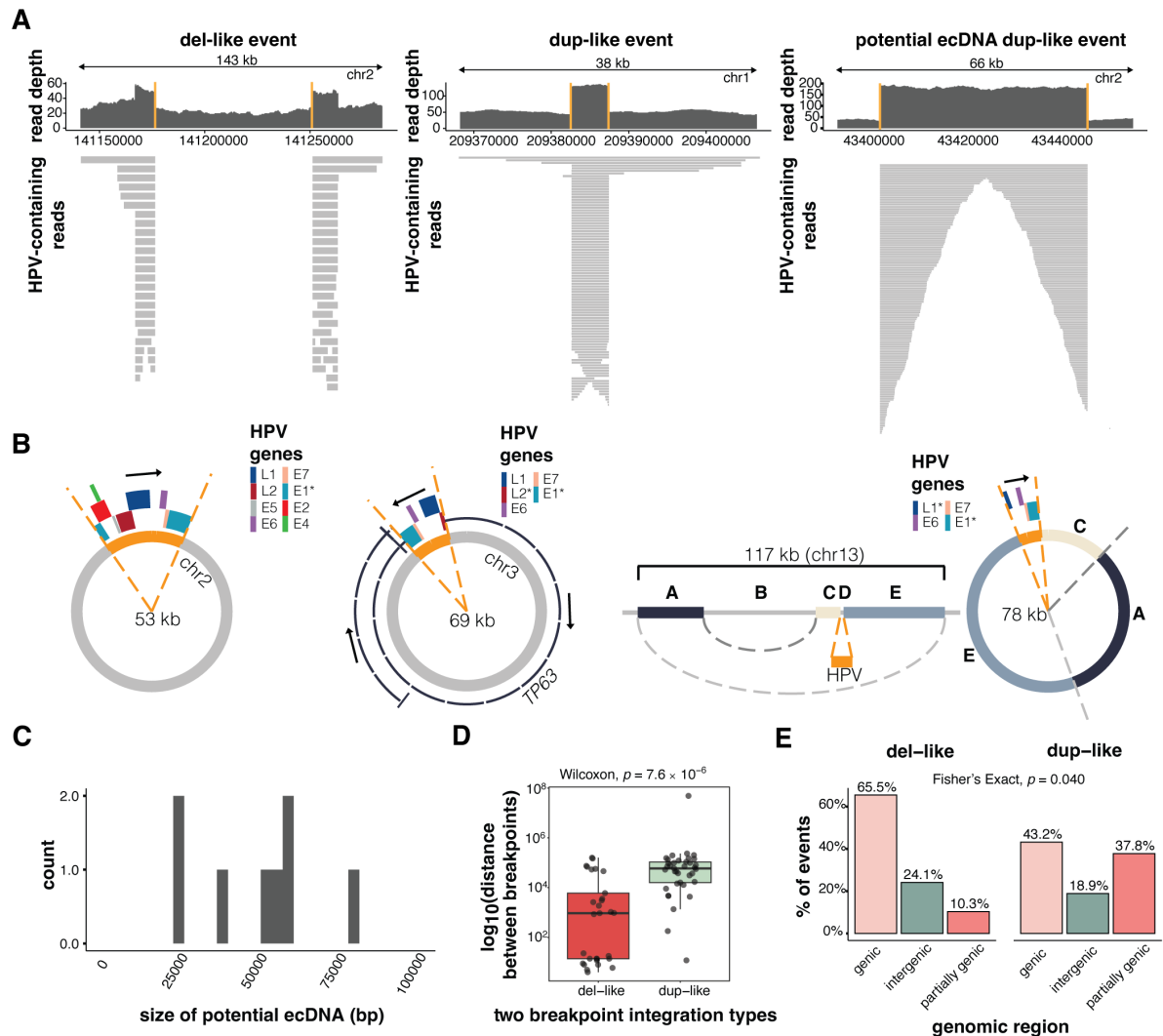


**Figure 2.** Heterologous structures of the HPV genome before and after integration. (A) Schematic of possible HPV integrant structures. The spirals in structures A-E show the portion(s) of the HPV genome that could be contained within an integrant. (B) Schematic showing how several heterologous integrants can exist between a single breakpoint pair, with the size of the integrant varying by  $n$  HPV copies. The colors correspond to the regions of the HPV genome as depicted in A. (C) The number of integrant structures between all identified breakpoint pairs within the cohort. Integrants with 2+ identified structures were classified as heterologous. (D) The sizes of the HPV integrants in a multi-breakpoint event with schematics depicting the various integrant structures detected. The HPV genome in this case was broken into two segments A and B, and the A segment was further broken into three segments (A.1, A.2, A.3). These segments were variably rearranged into new structures across the breakpoint pairs. Each point represents the size of an HPV integrant contained on an individual read, which are then grouped together by color (e.g. blue, red, green, light blue) if they do not differ in size by more than 300 bp. Each color in a breakpoint pair thus represents one unique integrant structure, as indicated in the accompanying schematics. (E) The lengths of HPV-aligned reads in four predominantly episomal samples. The existence of HPV episomes and episome concatemers are supported by the accumulation of read counts in bin sizes corresponding to one or more HPV genome copies, as indicated by the dotted lines. (F) Frequency of heterologous integrants in the different integration categories. Only categories harboring heterologous integrants are shown. (G) The percentage of integrants from different HPV types that form single or heterologous structures. (H) The maximum size of the integrant structure in each breakpoint pair in HPV16 and HPV18 integrants, represented as the number of HPV genome copies. (I) Distribution showing the maximum number of HPV copies found in the longest spanning read for each incomplete integrant. The x-axis shows the length of the longest spanning read. Box plots represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 IQR below Q1 or 1.5 IQR above Q3).  $P$  values were calculated using the Wilcoxon rank-sum test and the Fisher's exact test, as indicated on the figure.

1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261

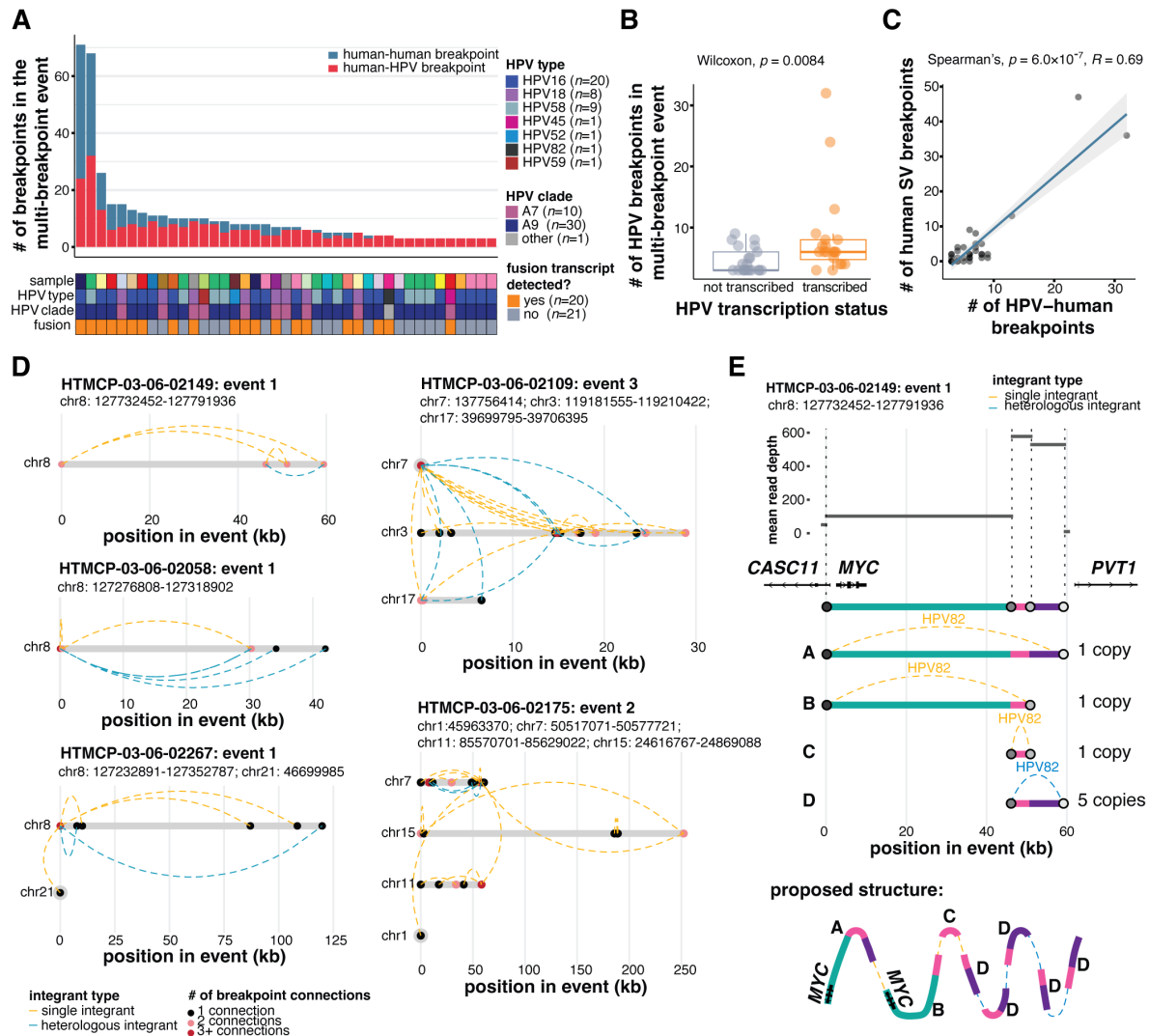
1262

1263



**Figure 3.** The characteristics of two-breakpoint events and potential ecDNAs. (A) Examples of read coverage patterns from a del-like event, a dup-like event, and a potential ecDNA dup-like event. Orange lines indicate the integration breakpoints. (B) Circular assemblies from three potential ecDNA integration events. The orange portions show the integrated HPV segment including the viral genes. The direction of HPV gene transcription is shown by a black arrow. The right-most example depicts a complex event in which three non-adjacent human segments have been combined in the potential ecDNA. (C) The size distribution of potential HPV-human hybrid ecDNAs ( $n=8$ ). (D) The genomic distance between breakpoints in del-like versus dup-like events. Box plots represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 IQR below Q1 or 1.5 IQR above Q3). (E) The percentage of events occurring in genic (>90% within a gene), intergenic (<10% within a gene), and partially genic (10-90% within a gene) regions, plotted by integration category. The  $P$  value in D was calculated using a Wilcoxon ranked-sum test. The  $P$  value in E was calculated using a Fisher's Exact test.

1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288



**Figure 4.** Complex structural variation is associated with multi-breakpoint integrations. (A) The number of human-HPV and human-human SV breakpoints across multi-breakpoint integration events, and the HPV type and clade in each. (B) The number of breakpoints per event in transcribed and non-transcribed multi-breakpoint events. Box plots represent the median and upper and lower quartiles of the distribution; whiskers represent the limits of the distribution (1.5 IQR below Q1 or 1.5 IQR above Q3). (C) Spearman's correlation between the number of HPV-human breakpoints and the number of human-human SV breakpoints in multi-breakpoint events. (D) Examples illustrating the connectivity between HPV breakpoint pairs in five multi-breakpoint events: three from the *MYC*-associated locus on Chromosome 8 and the two most rearranged multi-chromosomal events. Dots denote HPV breakpoints along the event, and dotted lines represent the HPV integrants that connect the breakpoints. The dots are colored according to the number of connections that converge at that position in the event. The integrants are colored according to whether single (orange) or heterologous (blue) integrant structures connect the breakpoints. (E) An example breaking down the copy number changes and the proposed structure of an event overlapping *MYC*. Each color represents a genomic segment on Chromosome 8 in between two human-HPV breakpoints. *P* values in B and C were calculated using the Wilcoxon rank-sum test and a Spearman's correlation test.

1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305

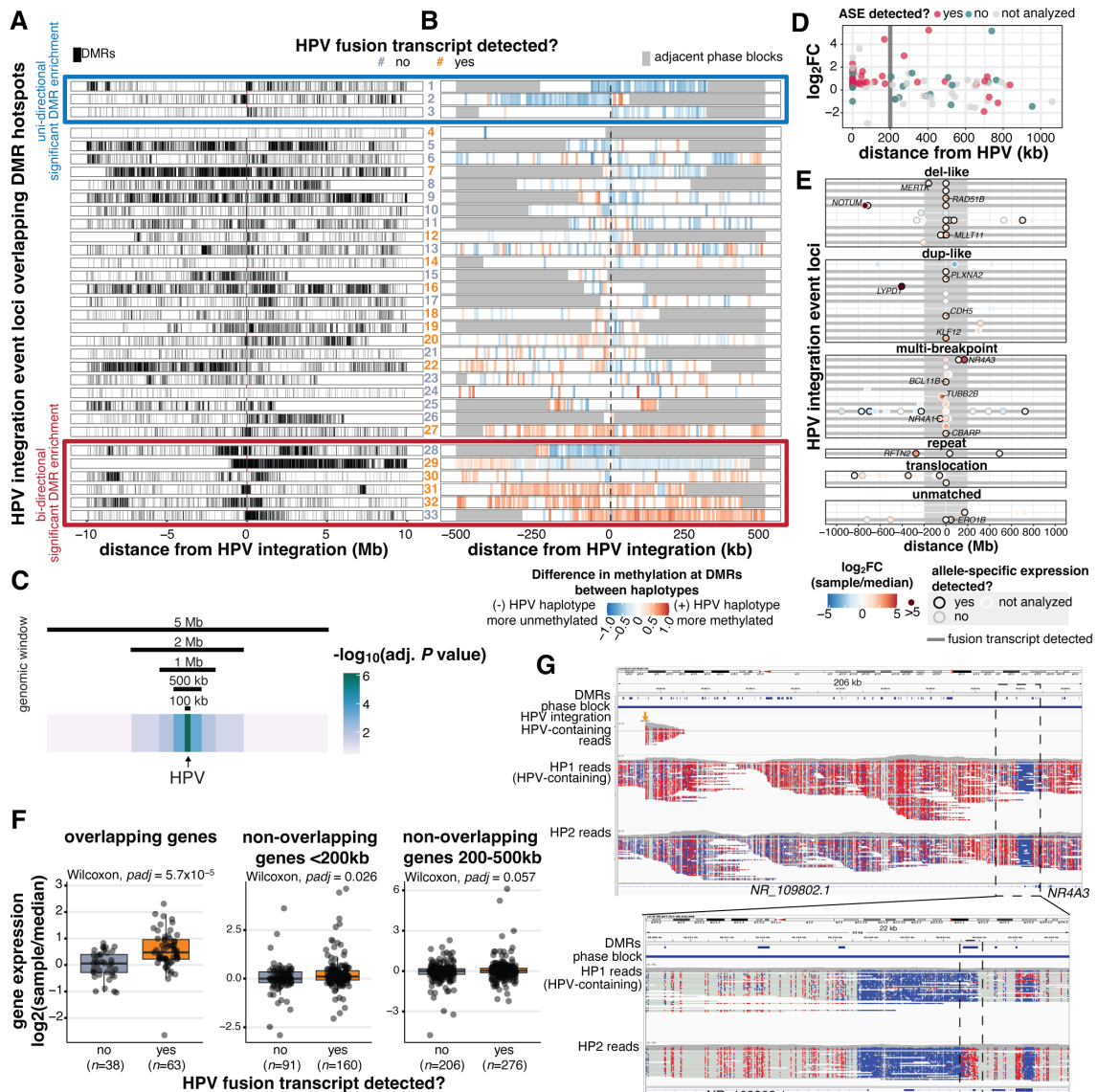
1306

1307

1308

1309





1333  
 1334  
 1335 **Figure 6.** HPV integration is associated with dysregulation of the methylome and nearby genes on the integrated  
 1336 haplotype. (A) The distribution of DMRs across 10 Mb on either side of HPV integrations in the 33 events that  
 1337 overlapped a DMR hotspot. Each tick represents a DMR. (B) The direction of methylation changes in the HPV-  
 1338 containing haplotype with respect to the unintegrated haplotype within the phase block containing HPV integration.  
 1339 Adjacent phase blocks are shown as flanking gray bars. Events are ordered identically in A and B. The color of  
 1340 event numbers (centre column) indicate the transcriptional status of the event. Events within the red (bottom) and  
 1341 blue (top) boxes show significant DMR enrichment ( $padj < 0.05$ ) either uni-directionally (blue box) or bi-directionally  
 1342 (red box) relative to the HPV integration event, as determined by a permutation test of the 500 kb bins flanking the  
 1343 event. (C) The significance of the association between HPV integration and high DMR density at all 147 HPV-  
 1344 integrated regions, using window sizes of 100,000 bp, 500,000 bp, 1,000,000 bp, 2,000,000 bp, and 5,000,000 bp  
 1345 around HPV. (D) The fold change and allele specific expression (ASE) status of outlier genes (1.5 IQR below Q1  
 1346 or 1.5 IQR above Q3) within 1 Mb of integration events. The  $\log_2$  fold change ( $\log_2\text{FC}$ ) of the integrated sample is  
 1347 relative to the median of the cohort. The ASE status, integration event type, and transcriptional status of the event are also  
 1348 indicated. (E) The position of genes with outlier expression (1.5 IQR below Q1 or 1.5 IQR above Q3) relative to sites of HPV  
 1349 integration. Expression fold change in the integrated sample relative to the median of the cohort. The ASE status,  
 1350 integration event type, and transcriptional status of the event are also indicated. (F) The difference in gene  
 1351 expression fold change (integrated sample/median) at transcribed ("yes") and non-transcribed ("no")  
 1352 integration events. All box plots represent the median and upper and lower quartiles of the distribution;  
 1353 whiskers represent the limits of the distribution (1.5 IQR below Q1 or 1.5 IQR above Q3). (G) Integrative  
 1354 Genomics Viewer snapshots showing wide (top) and zoomed (bottom) views of the haplotype-specific methylation  
 1355 changes around *NR4A3* and HPV integration in HTMCP-03-06-02428 (#31), with reads separated into the two  
 1356 haplotypes (HP1 and HP2). The sample's DMRs, phase blocks, and HPV integration breakpoints are also indicated  
 1357 in the top three tracks. Reads are colored by CpG methylation status, with red indicating methylated and blue  
 indicating unmethylated. Adjusted  $P$  values in F were calculated using Benjamini-Hochberg-corrected Wilcoxon  
 rank-sum tests.