**Human Papilloma Virus Integration Strictly Correlates with Global Genome Instability in Head and Neck Cancer**

**Brandon LaBarge, Max Hennessey, Lijun Zhang, David Goldrich, Scott Chartrand, Carson Purnell, Sage Wright, David Goldenberg, James R. Broach**

**Supplementary Material**

**Tumor integration site descriptions**

**Table S1. Patient Demographics and Cancer Characteristics**

**Table S2. Whole Genome Sequencing and Whole Genome Imaging Metrics for Samples Analyzed in this Study**

**Table S3. HPV Copy Number.**

**Table S4. HPV Integration Sites**

**Table S5. Cancer Relevant Genes Affected by Structural Variants**

**Figure S1. Genome Copy Number Alterations**

**Figure S2. Mutational Burden**

**Genomic Rearrangements Surrounding Viral Integration Sites**

**7387LN:** Part of an HPV33 genome (4904-7909/0-3412) integrated at position 92.91 Mb of chromosome 15 with the concurrent deletion of 450 bp of chromosomal DNA at the site of integration and disruption of the CHD2 gene on the chromosome and the E2 gene on the virus. Subsequently, the 22 Kb chromosomal region spanning the viral insertion, along with the virus, underwent a tandem duplication, yielding two copies of the integrated viral fragment.

Of particular note, the tumor also contains a precise translocation of 173 Kb spanning a portion of the RB gene on chromosome 13 to position 15.7 Mb on chromosome 6, resulting in non-functional RB genes at both sites.

**20T:** The tumor contains both an intact HPV 16 genome and one deleted between 1595 and 2358, which eliminates portions of E1 and E2. Softclip reads and OGM indicate viral insertion at 68.719 Mb of chromosome 14 with deletion of 3260 bp of the chromosome at the site of insertion of the 7150 bp of the viral genome, resulting in a cumulative insertion of 3900 bp. Subsequently, 93 Kb segment surrounding the insertion duplicated twice, once in a direct tandem repeat and a second in a direct repeat separated from the first two by a 57 Kb fragment of the duplicated region.

**3922LN**: A single viral fragment spans a junction between chr11:94.52 and chr8:133.01. This is part of a concerted three way rearrangement among six fragments - Chr8 proximal (1-133.01Mb), Chr8 distal (133.01Mb-tel), Chr11 proximal A (1-94.52Mb), Chr11 distal A (94.52Mb-tel), Chr11 proximal B (1-99.75Mb) and Chr11 distal B (99.75Mb-tel) - that results in Chr8 proximal linked to Chr11 proximal B, Chr11 proximal A linked to Chr11 distal B and Chr8 distal linked through the virus to Chr11 distal B. Virus position 6196 connects to chr8:133,013,568 and virus position 296 connects to chr11:94,527,997.

We also observed softclip reads linking viral position 296 to chr2:32,916,251. However, while the WGI shows a 2130 bp deletion at that site, we do not observe a second set of softclip reads that would indicate linkage of the other arm of the virus to the chromosome. Accordingly, we cannot confirm the presence of the virus at this second site.

**5954LN**: Viral reads show approximately 15 copies of a complete genome and 15 copies of a partial genome, with the partial genome extending from ca. 4080 to 6700. However, softclip reads indicate staggered insertions on chr14 at 51,101,387 and 51,101,721 linked to virus position 4829, which is then connected to viral positions 5564 and 5794 linked to staggered position on chr14 at 51,103,716, 51,104,271 and 51,106,903. Optical mapping indicates a 7.8 kb insertion between 51.103Mb and 51.121Mb. The 36kb region surrounding the deletion is repeated in tandem at least five times and perhaps more, since no reads span all the repeats. In addition, the adjacent 162 kb, not including the viral insertion, is duplicated in a direct tandem repeat. We can only speculate as to the mechanism responsible for the staggered linkages between the virus and chromosome but one possibility is repeated NHEJ events at a multiply resected double strand break on chr 14.

**7331LN**: The tumor contains two integrations both followed by focal amplification. Viral reads show an intact genome present at approximately 5 copies, one partial genome lacking region 2050 to 4300 also at 5 copies and a third partial genome in two copies covering 400 to 1325. The first integration site on chromosome 8 involved two intrachromosomal translocations with the virus inserted at the junction between the one of the translocations, linking position 11,855,901 to viral position 6833 through position 1643 and then to chr8:11,828,568. This 35 Kb region along with the viral fragment was subsequently triplicated in tandem and then that triplicated region and the adjacent 86 Kb segment duplicated in tandem.

The second integration occurred at chr19:28,377,176 linking it to viral position 3972 and chr19:28,377,585 linked to viral position 4061. This result in a small deletion of chromosomal sequences and insertion of a nearly intact copy of the virus. Subsequently, the 78 kb region surrounding the insertion site was duplicated in tandem.

**3718LN**: The tumor contains two separate viral insertions. The first insertion results in an intrachromosomal translocation linking chr20 11,453,956 to chr20 11,105,927 via a 5577 bp fragment of virus extending from position 4627 to 2298. The 19 Kb region spanning the junction subsequently underwent local amplification yielding five copies of the sequence and the virus in tandem with one of the copies containing an additional stretch of the proximal translocation domain. The immediate adjacent region also underwent a single tandem duplication.

The second insertion occurred in conjunction with a deletion on chr11 from 67,598,149 to 67,607,927, with the deleted region replace by a fragment of the virus extending from 5767 to 2594. To account for the size of the insertion identified at this site by optical mapping, we surmise that the insertion involves six tandem repeats of the viral fragment insertion.

**5785T**: The tumor contains two separate insertions. Viral reads indicate the presence of an essentially complete HPV genome, except for a small deletion at the end of E2, at a copy number of approximately 5.6. Softclip reads link the virus to chr5 from position 135,027,888 to 135,034,828 at which site optical mapping indicates a 5.5 Kb insertion, without any additional rearrangements, suggesting a simple integration event. We also obtained a large number of softclip reads at chr1:143,264,527. However, this lies in the centromeric region of chromosome 1 for which, due to its repetitive sequence, optical mapping data is not available.

**7122LN**: The tumor contains two separate insertions. Viral reads show two partial genomes in essentially equal amounts, with left breakpoints at 410 and 840 and right breakpoints at 3060 and 6800. Softclip read indicate integration on chromosome 18 with the partial viral genomes abridging an unequal crossover linking position 3.80 Mb to 3.47 Mb to generate a duplication of the intervening 326 Kb with a junction structure of chr18:3.80-HPV(410-0/7909-3060)-chr18:3.47. The 50 kb region surrounding the junction fragment and including the viral genome is duplicated in direct and inverted orientation at least 26 times in an array interspersed with additional sequences from the overlap region to generate a complex focal amplification of 2.5 Mb with the following structure, using this nomenclature, Chr18: 3.474 A 3.498 B 3.603 C 3.733 D 3.775 E 3.800, with the viral insert designated as “v”.

EvAADEvAE’vAEvAADEvAƎAEvAƎⱯʌƎⱯⱯ●ⱯʌƎⱯʌƎⱯⱯ●ⱯʌƎⱭꞚⱯⱯʌƎⱯʌ,ƎⱯʌƎⱭEvAEvAAAAⱭEvAAEvAE’vAAEvABCDEvAE’vAAE’vAAⱭDEvAE’vAEvAADEvA

Softclip reads indicate a second integration at chr22:0.14 linked to viral positions 6800 and 840, perhaps accounting for the second partial HPV genome. However, this position corresponds to the telomeric region of chromosome 22, which due to its repetitive nature, we were unable to recover optical mapping reads over this region.

**Table S1. Patient Demographics and Cancer Characteristics**

|  |  |
| --- | --- |
|  | HPV Genomic Status |
|  | Integrated (n=8) | Episomal (n=4) |
| Mean Age at Diagnosis  | 59.88 | 56.75 |
| Percent Male  | 7/8 (87.5%) | 3/4 (75%) |
| Mean Tumor Size (cm) Greatest Dimension | 3.50 | 2.43 |
| T1 | 2/8 (25%) | 2/4 (50%) |
| T2 | 3/8 (37.5%) | 2/4 (50%) |
| T4 | 3/8 (37.5%) | 0/4 (0%) |
| N0 | 1/8 (14%) | 1/4 (25%) |
| N1 | 5/8 (62.5%) | 1/4 (25%) |
| N2 | 2/8 (25%) | 2/4 (50%) |
| Stage I | 4/8 (50%) | 2/4 (50%) |
| Stage II | 3/8 (37.5%) | 2/4 (50%) |
| Stage III | 1/8 (14%) | 0/4 (0%)  |
| Positive Margins | 2/8 (25%) | 1/4 (25%) |
| Cancer Recurrence  | 3/8 (37.5%) | 1/4 (25%) |
| Extranodal Extension (ENE) | 4/8 (50%) | 2/4 (50%) |
| Perinteural Invasion (PNI) | 4/8 (50%) | 0/4 (0%) |
| Lympho-vascular Invasion (LVI) | 5/8 (62.5%) | 0/4 (0%) |
| Mean Size (cm) Greatest Dimension Largest Metastatic Focus in Lymph Node | 3.62 | 2.9 |
| Mean Number of Positive Lymph Nodes | 3.25 | 3.5 |
| Mean Total Number of Nodes  | 22.625 | 18.75 |
| Presence of Positive Lymph Node(s) | 7/8 (87.5%) | 3/4 (75%) |

AJCC 8th T, N, M Staging

**Table S2. Whole Genome Sequencing and Optical Genome Mapping Metrics for Samples Analyzed in this Study.**

A. Optical Genome Mapping

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **N50 Mbp (>20Kbp)** | **N50 Mbp (>150Kbp)** | **Labels/ 100kbp** | **Map Rate (%)** | **Total DNA Scanned (Gbp)** | **Effective Coverage** |
| 20T | 0.20 | 0.28 | 15.79 | 54.60 | 264.00 | 107.16 |
| 20 Blood | 0.24 | 0.29 | 14.69 | 89.90 | 1307.00 | 368.33 |
| 3718LN | 0.20 | 0.27 | 16.12 | 58.70 | 410.00 | 75.72 |
| 3718 Blood | 0.28 | 0.32 | 15.08 | 89.10 | 1335.00 | 371.95 |
| 3726T | 0.24 | 0.28 | 15.61 | 88.60 | 2001.19 | 415.92 |
| 3726 Blood | 0.17 | 0.24 | 16.26 | 65.00 | 643.55 | 78.18 |
| 3922LN | 0.14 | 0.27 | 15.45 | 80.30 | 1303.00 | 329.98 |
| 3922 Blood | 0.29 | 0.29 | 14.75 | 89.50 | 1304.00 | 363..58 |
| 3943LN | 0.27 | 0.28 | 15.51 | 83.10 | 1318.55 | 344.19 |
| 3943 Blood | 0.25 | 0.30 | 14.72 | 88.50 | 1305.00 | 361.50 |
| 5785T | 0.18 | 0.28 | 15.26 | 74.30 | 498.00 | 115.99 |
| 5785 Blood | 0.25 | 0.32 | 14.31 | 84.90 | 1313.00 | 347.00 |
| 5954LN | 0.12 | 0.27 | 16.78 | 68.60 | 1230.00 | 267.17 |
| 5954 Blood | 0.23 | 0.29 | 13.56 | 77.00 | 1341.00 | 321.37 |
| 7122LN | 0.16 | 0.26 | 15.45 | 83.20 | 1304.00 | 341.70 |
| 7122 Blood | 0.27 | 0.33 | 14.52 | 90.30 | 1106.00 | 311.48 |
| 7309LN | 0.17 | 0.29 | 17.46 | 71.50 | 947.00 | 214.84 |
| 7309 Blood | 0.24 | 0.29 | 14.09 | 82.00 | 1312.00 | 336.03 |
| 7313LN | 0.22 | 0.33 | 15.73 | 84.90 | 960.00 | 257.95 |
| 7313 Blood | 0.26 | 0.30 | 14.75 | 87.60 | 1307.00 | 356.95 |
| 7331LN | 0.17 | 0.29 | 16.62 | 67.30 | 1235.00 | 146.87 |
| 7331 Blood | 0.22 | 0.28 | 14.18 | 78.00 | 1119.00 | 271.98 |
| 7387T | 0.16 | 0.25 | 16.27 | 74.10 | 1922.71 | 235.86 |
| 7387 Blood | 0.24 | 0.30 | 15.37 | 81.70 | 806.93 | 153.23 |
| **Average (all)** | 0.21 | 0.29 | 15.41 | 78.46 | 1089.35 | 251.46 |
| **Average (tissue)** | 0.18 | 0.27 | 16.12 | 73.30 | 995.41 | 208.80 |
| **Average (blood)** | 0.25 | 0.30 | 14.69 | 83.63 | 1183.29 | 298.00 |

B. Whole genome Sequencing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **No. of Primary Reads** | **Unique Map Rate (%)** | **Q30 (%)** | **Effective Coverage** |
| 20T | 738,396,370 | 82.67 | 94.88 | 52.89x |
| 20 Blood | 387,421,874 | 83.26 | 95.04 | 27.79x |
| 3718LN | 408,309,083 | 90.89 | 94.47 | 30.38x |
| 3718 Blood | 610,549,967 | 89.23 | 94.75 | 46.21x |
| 3726T | 647,108,324 | 70.00 | 94.63 | 39.08x |
| 3726LN | 595,215,698 | 74.22 | 95.21 | 38.17x |
| 3726 Blood | 251,663,984 | 73.88 | 94.57 | 16.23 |
| 3922 LN | 543,081,256 | 84.24 | 94.00 | 39.55x |
| 3922 Blood | 415,898,602 | 84.63 | 95.18 | 30.53x |
| 3943T | 370,055,404 | 91.59 | 94.29 | 27.98x |
| 3943 Blood | 483,394,001 | 90.09 | 94.46 | 37.07x |
| 5785T | 554,128,188 | 83.40 | 94.47 | 40.11x |
| 5785 Blood | 486,977,313 | 81.44 | 95.34 | 34.03x |
| 5954LN | 399,616,394 | 90.17 | 93.94 | 29.89x |
| 5954 Blood | 410,622,572 | 90.97 | 94.46 | 31.63x |
| 7122LN | 473,467,888 | 84.15 | 94.64 | 33.95x |
| 7122 Blood | 605,397,144 | 81.26 | 95.33 | 42.04x |
| 7309LN | 479,227,100 | 84.09 | 94.97 | 33.76x |
| 7309 Blood | 501,664,150 | 83.59 | 93.95 | 35.99x |
| 7313LN | 568,503,215 | 83.82 | 94.55 | 40.58x |
| 7313 Blood | 321,917,279 | 85.89 | 95.08 | 24.00x |
| 7331LN | 490,075,862 | 83.32 | 95.39 | 34.80x  |
| 7331 Blood | 449,699,675 | 84.89 | 95.23 | 32.72x |
| 7387T | 252,570,746 | 96.48 | 94.67 | 20.61x |
| 7387 Blood | 328,752,272 | 96.17 | 95.31 | 26.68x |

**Table S3. Viral Copy Number**

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Intact Copy | Fragment 1 | Fragment 2 |
|  | Level\* | Level† | Length | Level† | Length |
|  |  |  |  |  |  |
| 7387LN |  | 1.0 | 6550 |  |  |
| 20T | 0.5 | 1.0 | 7250 |  |  |
| 3718LN | 0.1 | 6.0 | 6300 | 6.0 | 2200 |
| 3726T | 14.6 | 5.4 | 4650 |  |  |
| 3922T |  | 0.6 | 1050 |  |  |
| 5785T | 4.3 | 1.6 | 450 |  |  |
| 5954LN | 16.3 | 14 | 2225 |  |  |
| 7122LN |  | 18 | 5660 | 18 | 1650 |
| 7331LN | 4.3 | 4.3 | 5700 | 3.8 | 1000 |
|  |  |  |  |  |  |
| 3943T | 1.0 | 1.0 | 3480 |  |  |
| 7309LN | 3.5 |  |  |  |  |
| 7313LN | 0.6 | 1.3 | 1400 |  |  |

\*Copy level was determined from data presented in Figure 1 by calculating the average coverage of the viral genome over its entire length, if the coverage was uniform across the entire length, or of only that region with uniformly lower coverage. Values were normalized to the average coverage of single copy human sequences in the same sequenced sample.

†Fragments were defined as the region between discontinuities in sequence coverage across the viral genome. Copy levels were determined by calculating the average coverage between those discontinuities, relative to single copy human sequences in the same sample, after subtracting the level of the intact virus, if present, and any larger subgenomic fragment subsuming the fragment in question.

**Table S4. Viral Integration Sites**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Hchr** | **HIntegPos** | **VirusID** | **VIntePos** | **Discordant Reads** | **Softclip Location Reads** | **Softclip Total Reads** | **Nearest Gene** | **Dist to Gene** |
| **3922LN** | chr11 | 94527997 | HPV16 | 296 | 16 | 6 | 12 | C11orf97 | 0 |
| chr11 | 94528460 | HPV16 | 296 | 6 | 6 | 11 | C11orf97 | 0 |
| chr8 | 133013568 | HPV16 | 6196 | 10 | 6 | 213 | TG | 0 |
| chr2 | 32916251 | HPV16 | 296 | 4 | 17 | 20 | LINC00486 | 0 |
| **20T** | chr14 | 68722498 | HPV16 | 1595 | 28 | 14 | 19 | RAD51B | 39392 |
| chr14 | 68719238 | HPV16 | 2358 | 25 | 21 | 40 | RAD51B | 36132 |
| **7331LN** | chr8 | 11855901 | HPV16 | 6835 | 159 | 1 | 2 | CTSB | 0 |
| chr8 | 11858153 | HPV16 | 4142 | 2 | 14 | 14 | CTSB | 0 |
| chr8 | 11828568 | HPV16 | 5762 | 12 | 4 | 7 | FDFT1 | 0 |
| chr19 | 28377176 | HPV16 | 3972 | 22 | 17 | 18 | LOC100420587 | 58212 |
| chr19 | 28377585 | HPV16 | 4061 | 15 | 13 | 17 | LOC100420587 | 57803 |
| **7122LN** | chr18 | 3474193 | HPV16 | 3059 | 283 | 183 | 299 | GAPLINC | 0 |
| chr18 | 3802328 | HPV16 | 492 | 241 | 308 | 401 | DLGAP1 | 0 |
| chr18 | 3802129 | HPV16 | 366 | 92 | 3 | 369 | DLGAP1 | 0 |
| **5785T** | chr1 | 143264527 | HPV16 | 1449 | 4 | 58 | 947 | LSP1P5 | 135443 |
| chr5 | 135034828 | HPV16 | 1157 | 3 | 6 | 7 | C5orf66 | 0 |
| chr5 | 135027888 | HPV16 | 6154 | 2 | 5 | 6 | PITX1 | 0 |
| chr5 | 135035424 | HPV16 | 1392 | 1 | 0 | 0 | C5orf66 | 0 |
| **5954LN** | chr14 | 51103716 | HPV16 | 5794 | 9 | 2 | 133 | TRIM9 | 8567 |
| chr14 | 51104271 | HPV16 | 5564 | 14 | 95 | 115 | TRIM9 | 8567 |
| chr14 | 51106903 | HPV16 | 5564 | 82 | 43 | 84 | TRIM9 | 11199 |
| chr14 | 51101387 | HPV16 | 4829 | 15 | 4 | 21 | TRIM9 | 5683 |
| chr14 | 51101721 | HPV16 | 4829 | 4 | 2 | 5 | TRIM9 | 6017 |
| **3718LN** | chr20 | 11105927 | HPV16 | 4626 | 45 | 29 | 34 | C20orf187 | 76561 |
| chr20 | 11453956 | HPV16 | 2298 | 87 | 41 | 42 | LOC339593 | 180573 |
| chr11 | 67598149 | HPV16 | 2594 | 5 | 4 | 6 | C11orf72 | 4730 |
| chr11 | 67607927 | HPV16 | 5767 | 5 | 9 | 9 | NDUFV1 | 0 |
| **7387LN** | chr15 | 92910685 | HPV33 | 5139 | 1 | 2 | 3 | CHD2 | 0 |
| chr15 | 92911786 | HPV33 | 4904 | 11 | 9 | 10 | CHD2 | 0 |
| chr15 | 92912334 | HPV33 | 3455 | 6 | 9 | 20 | CHD2 | 0 |

**Table S5. Cancer Relevant Genes Affected by Structural Variants**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample** | **Oncogene** | **Tumor Suppressor** | **Gene Fusion** |
| 20T |  | PTEN () | PTEN-RNLS |
|  |  | ARID1A (T) | ARID1A-RAB3GAP2 |
|  |  |  |  |
| 3922LN |  | SDHAF2 () |  |
|  |  | FEN1 () |  |
|  |  |  |  |
| 5785T |  | LEPROTL1 () |  |
|  |  | WRN () |  |
|  |  | NRG1 () |  |
|  |  | ARHGEF12 (T) |  |
|  |  |  |  |
| 5954LN |  | RAD51B (T) |  |
|  |  |  |  |
| 7122LN | NFE2L2 (D)\* |  | STOX1-BCL7A |
|  |  | CYLD (T) | CYLD-ZNF778 |
|  |  |  |  |
| 7331LN |  | LRP1B () |  |
|  |  | GPC5 () |  |
|  |  | PTPRT ()\* |  |
|  |  |  | PALM2-AKAP2-DCC |
|  |  |  |  |
| 7387T |  | TNFAIP3 () |  |
|  |  | PTEN () |  |
|  |  | RB1 () |  |
|  | SYK (D) | CHD2 (I) |  |

(), deletion; T, translocation; D, duplication; \*previously associated with HNSCC

**Figure S1. Genome Copy Number Alterations in HPV Tumors**



Segmental chromosomal copy number alterations are indicated by chromosomal position for each of the OPSCC samples. The fractional increased (red) or decreased (blue) copy number relative to the patient’s normal genome are indicated by color intensity.

**Figure S2. Somatic Mutational Burden of HPV Tumors.**



Points on the left hand graph show the mutational frequency in exomes in 512 individual Head and Neck Squamous Cell Carcinoma tumor-normal pairs (Kyle Ellrott, Matthew H. Bailey, Gordon Saksena, et. al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018 Mar 28; 6(3): 271–281.e7. <https://doi.org/10.1016/j.cels.2018.03.002>). The same data is shown on the right for the samples reported in this study. Green dots represent samples in which HPV is present only as an episome. Red bars indicate median values.