Divergence of HPV16 variants reflects loci undergoing inter-host positive selection, potentially immunologic selection

Chase W. Nelson*[1], *Apurva Narechania*[1], *Robert D. Burk*[2], *Mark Schiffman*[3], *Michael Cullen*[3], *Joseph F. Boland*[3], *Zigui Chen*[4], *Nicolas Wentzensen*[3], *Qi Yang*[3], *Jason Mitchell*[3], *David Roberson*[3], *Sara Bass*[3], *Laurie Burdett*[3], *Tina Raine-Bennett*[5], *Thomas Lorey*[5], *Meredith Yeager*[3], *Lisa Mirabello*[3]

[1]**Sackler Institute for Comparative Genomics, American Museum of Natural History (New York);** [2]Albert Einstein College of Medicine (Bronx, NY); [3]Divison of Cancer Epidemiology and Genetics, National Cancer Institute (Rockville, MD); [4]The Chinese University of Hong Kong (Hong Kong); [5]Kaiser Permanente (Oakland, CA); ***Presenting author**: cnelson@amnh.org

Introduction

Human papillomaviruses (HPVs) are highly prevalent (~40%)[1] in adults. Most of the **>200 HPV types** are benign. **Cervical cancer** is caused by **~12 high-risk HPV types**,[2] with >**50% caused by HPV type 16 (HPV16) alone**. The genetic basis of HPV16's unique carcinogenicity is unsolved. HPV16 consists of **~13 sublineages** across 4 lineages (A: A1-4; B: B1-2; C: C1-3; and D: D1-4) which exhibit **>100x cancer risk differences**.[3,4] HPV16 sublineages contain **lineage defining variants** (*i.e.*, diagnostic characters)[4] in their genomes of ~8kb circular dsDNA comprised of 8 genes (**Fig. 1**).

Between-host mean viral **nonsynonymous** (**N**; amino acid-altering) and **synonymous** (**S**; amino acid-conserving) nucleotide distances can be measured as $\overline{d_N}$ and $\overline{d_S}$. Regions under **diversifying positive selection** may exhibit $\overline{d_N} > \overline{d_S}$.[5] One target of selection is **immune epitopes**, viral peptide fragments identified by antibodies or T-cells.[6] Individuals differ in their ability to detect epitopes based on immune alleles, allowing **inter-host selection**.

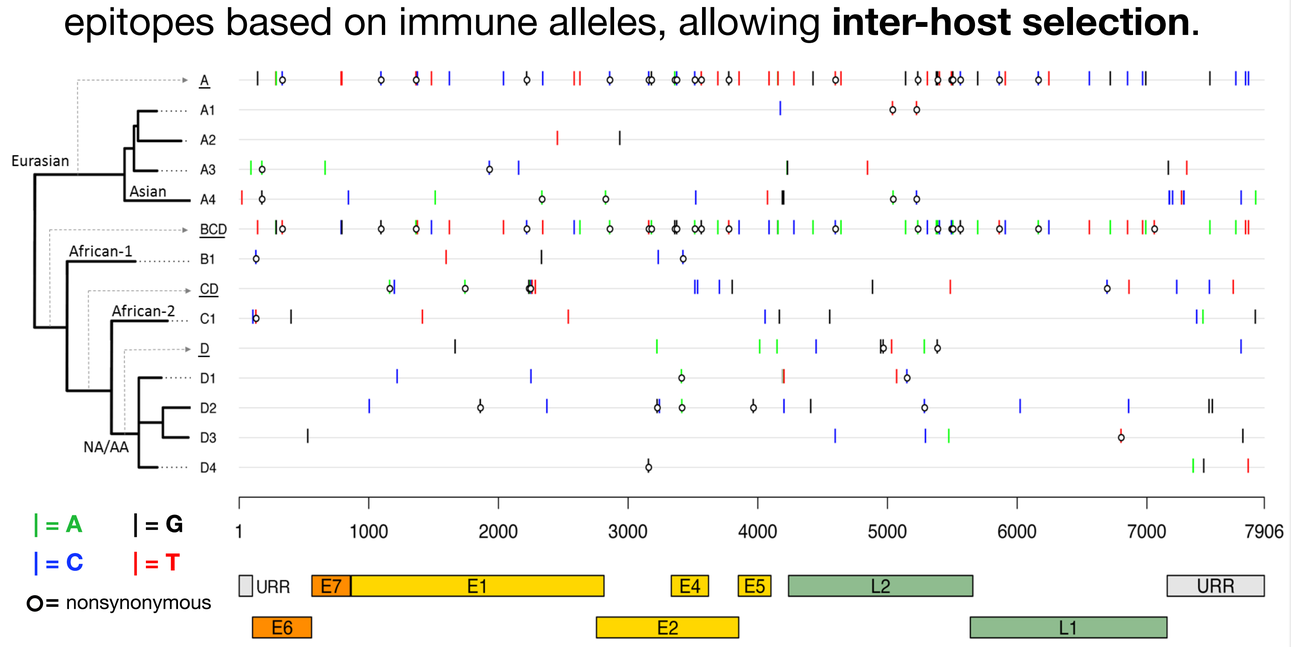

Figure 1 Phylogenetic tree of HPV16 sublineages, adapted from Mirabello *et al*.[4] Lineage-defining variants (diagnostic characters) are shown (|) along the genome, cumulative toward terminal nodes (*e.g.*, all A variants are present in A1-4). The HPV16 genome contains oncogenes **E6** and **E7**; early genes **E1**, **E2**, **E4**, and **E5**; late genes **L1** and **L2**; and a **URR**.

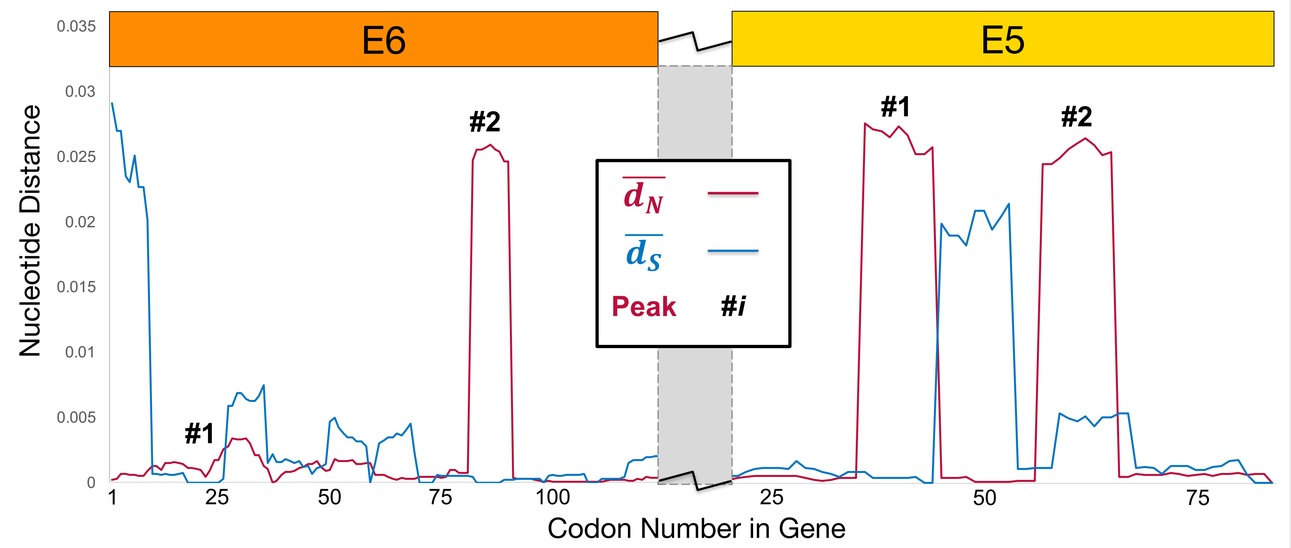

Figure 2 Nucleotide distance 9-codon sliding window results for portions of **E6** and **E5**. Nonsynonymous **peak** numbers (**#*i***) and details are as described in **Methods** and **Table 1**. **E6** begins at genome position 83, and **E5** begins at genome position 3,850.

Methods

We used a custom Ion Torrent sequencing method[7] to obtain whole HPV16 viral genomes from 3,215 HPV-positive specimens from women undergoing cervical cancer screening at Kaiser Permanente Northern California in the Persistence and Progression (PaP) cohort.[4] From these, we identified 795 sublineage A1 **cases** and 839 A1 **controls**. Aligned sequences were analyzed with SNPGenie[8] to estimate $\overline{d_N}$ and $\overline{d_S}$ for all viral codons using a Nei-Gojobori approach.[9,10] Sliding windows of 9 codons (step size of 1) were performed to reflect T-cell epitope size.[6] *Z*-test bootstraps and permutation tests were performed with 1000 replicates. Candidate epitopes were queried as all contiguous octamers at Immune Epitope Database (IEDB)[11] and HPV T-Cell Antigen Database.[12] Results were analyzed and visualized in Excel, R version 3.4.0, Perl, and PowerPoint.

Results & Discussion

- **Whole-Gene $\overline{d_N}/\overline{d_S}$.** **E6** did not differ significantly from neutrality ($\overline{d_N} \approx \overline{d_S}$). All other genes were under purifying selection ($\overline{d_N} < \overline{d_S}$), but this was significantly **relaxed in E7 for controls** (**Table 1**).[13]
- **Sliding Windows for Nonsynonymous Peak Discovery.** Fourteen (14) nine-codon peaks were identified in which $\overline{d_N}$ exceeded $\overline{d_S}$ of both the peak and the gene. These were concatenated, end-trimmed, and considered **candidate epitopes** (**Table 1; Fig. 2**).
- **Benchmarking Peaks as Candidate Epitopes.** Prior *in vivo* work shows peak **E6#1** overlaps an HLA-B7 epitope and a p53 degrading region.[14,15] **E6#2** overlaps L90[83]V (T350G), implicated in differential cancer risk; HLA-B44, -B51, and -B57 epitopes; and antibody epitopes (**Fig. 2**).[16,17] Moreover, **9/14 candidates matched *in vitro* or *in vivo* verified epitopes** in online databases (**Table 1**).
- **Candidate Epitopes Overlap Lineage-Defining Sites.** When controlling for lineage signal by limiting analyses to A1, **peaks nevertheless overlap 37/360 lineage defining sites** in coding regions, representing significant enrichment compared to random peak placement (11.34 ± 0.12 S.E. expected; permutation test; Z=210; P≈0). Excluding **E2#1** (longest) did not change this conclusion.
- **Conclusions and Implications.** Ongoing immune selection may have played a role in the **historical divergence** of the HPV16 sublineages. Population genetic detection of natural selection may allow **rapid epitope characterization** from sequence data alone.
- **Future Direction.** Include other sublineages; bootstrap peak $\overline{d_N}/\overline{d_S}$; manufacture candidate epitopes to assay MHC binding.

Table 1 HPV16 sublineage A1 $\overline{d_N}/\overline{d_S}$ in nonsynonymous peaks

	WHOLE GENES		NONSYNONYMOUS PEAKS (CANDIDATE EPITOPES)					
Gene[a]	CASE $\overline{d_N}/\overline{d_S}$	CTRL $\overline{d_N}/\overline{d_S}$	Peak #	Num Lineage Defining Sites[b]	CASE $\overline{d_N}/\overline{d_S}$	CTRL $\overline{d_N}/\overline{d_S}$	Consensus Amino Acid Sequence[c]	IEDB[11] / HPV T-cell Ag DB[12] Hits
E6	0.828	0.923	1	2	∞	2.79	**Q**TTI**H**D**I**I	18/5
			2	2	59.44	28.05	…**H**YCYS**L**YG…	12/5
E7**	0.050	0.374	—	—	—	—	—	—/—
E1	0.234	0.281	1	0	5.65	1.93	**…L**TQA**E**T**E**T…	0/0
			2	2	2.75	12.74	SG**G**S**G**G**GC**	0/0
			3	1	∞	1.09	…VSF**S**ELVR**P**	0/0
			4	1	2.22	—	**S**LF**G**MSL**M**	0/0
E2	0.676	0.698	1	13	20.44	13.66	**…SSPEI**IRQ**H…**	3/0
			2	3	23.96	14.26	**P**IL**T**A**F**NS**S**	1/4
			3	1	—	1.38	YRFKK**H**C**T**L	6/8
			4	2	—	5.36	**W**QR**DQ**FL**SQ**	2/5
E4	0.035	0.052	—	—	—	—	—	—/—
E5	0.647	0.532	1	4	48.94	19.14	…**TS**LI**I**LVL**L**L…	4/0
			2	4	7.74	3.15	**IV**YI**I**FVYIPL**FL**	1/1
L2	0.189	0.222	1	2	7.20	1.79	**K**VV**DPAFV**T…	0/0
L1	0.264	0.276	1	0	33.33	12.51	…**T**VGEN**V**P…	1/0

$\overline{d_N}/\overline{d_S}$ tends to be higher in **controls** (**CTRL**) in whole genes but higher in **cases** (**CASE**) in peaks. [a]Asterisks indicate a difference in $\overline{d_N}/\overline{d_S}$ between **cases** and **controls**: **α<0.05 (Z-tests, Bonferroni correction). [b]As identified by Smith *et al.*[18] [c]Variable sites in **bold**.

Acknowledgments

C.W.N. was supported by a Gerstner Scholars Fellowship from the Gerstner Family Foundation at the American Museum of Natural History. He thanks James Hussey, Jeff Witmer, Dan Graur, an anonymous genius, and the Sackler Institute for Comparative Genomics Computational Genomics and Bioinformatics workgroup for valuable feedback.

References

[1] CDC, https://www.cdc.gov/nchs/products/databriefs/db280.htm **[2]** Bouvard V, *et al.* 2009. A review of human carcinogens… *Lancet Oncol.* **10**:321–2. **[3]** Burk RD, *et al.* 2013. Human papillomavirus genome variants. *Virology* **445**:232–43. **[4]** Mirabello L, *et al.* 2016. HPV16 Sublineage Associations… *J. Natl. Cancer Inst.* **108**:1–9. **[5]** Hughes AL. 1999. *Adaptive Evolution of Genes and Genomes*. New York, NY: Oxford University Press. **[6]** Rammensee H-G, *et al.* 1995. MHC ligands and peptide motifs… *Immunogenetics* **41**:178–228. **[7]** Cullen M, *et al.* 2015. Deep sequencing of HPV16 genomes… *Papillomavirus Res.* **1**:3–11. **[8]** Nelson CW, *et al.* 2015. SNPGenie: estimating evolutionary parameters… *Bioinformatics* **31**:3709–11. **[9]** Nelson CW, Hughes AL. 2015. Within-host nucleotide diversity… *Infect. Genet. Evol.* **30**:1–7. **[10]** Nei M, Gojobori T. 1986. Simple methods for estimating… *Mol. Biol. Evol.* **3**:418–26. **[11] Immune Epitope Database**, http://www.iedb.org/ **[12] HPV T-Cell Antigen Database**, http://cvc.dfci.harvard.edu/hpv/index.php **[13]** Mirabello L, *et al.* 2017. HPV16 E7 genetic conservation… *in revision*. **[14]** Ellis JRM, *et al.* 1995. The association of an HPV16 oncogene… *Nat. Med.* **1**:464–70. **[15]** Crook T, *et al.* 1991. Degradation of p53… *Cell* **67**:547–56. **[16]** Chen Z, *et al.* Diversifying selection… *J. Virol.* **79**:7014–23. **[17]** Dillner J. 1990. Mapping of linear epitopes… *Int. J. Cancer* **46**:703–11. **[18]** Smith B, *et al.* 2011. Sequence imputation of HPV16… *PLoS One* **6**:e21375.