Genome characterization of human papillomavirus (HPV) 16

Andrew K. Teng*, Han Zhang*, Yanzi Xiao†, Lisa Mirabello†, Kai Yu*

*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville MD
†Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville MD

Introduction

HPV is a common sexually transmitted viral infection that targets basal epithelial cells and has a double-stranded DNA structure with around 8000 base pairs that codes for eight genes. There are 13 different carcinogenic HPV types and HPV16 is the most prevalent carcinogenic HPV, causing over half of cervical cancer cases and is the third most frequent cancer among women around the globe[2]. Not all cases and variants of HPV16 will lead to cervical cancer and it is unknown why only small fraction of the common benign infections will progress. HPV16 can be classified into four main variant lineages based on common phylogenetic patterns of SNPs (A, B, C, D), where B, C, D are non-European lineages and the A variants represent European-Asian lineages.

Objective

To further the understanding and determination of cervical cancer risk, full analysis of all the genomic characteristics of HPV16 was performed. The primary focus was to identify the viral genetic basis of HPV carcinogenicity and to globally analyze the differences between cases and controls.

Methods

Study Population[1]:

- 3569 HPV16 samples from NCI/DCEG Kaiser Permanente Northern California (KPNC) HPV Persistence and Progression (PaP) cohort, testing from 12/06-01/11
- Controls are HPV16 sequences from subjects with <CIN2, cases are the ones from subjects with CIN3
- Sublineages A1, A2, A3 (A) and D2, D3 (D) were grouped together- remaining lineages were not further explored due to small sample size

	A	B	C	D
Control (<CIN2)	988	37	14	38
Case (CIN3)	847	17	25	81
Total	1835	54	39	119

Genomic Variation Exploration:

- Entropy measures the genetic variation at a particular locus, defined as:

$$H(x) = -\sum_{i=1}^{4} p(x_i) log(p(x_i))$$

- Fisher exact p-values to show the correlation of genetic variants at two loci

Results

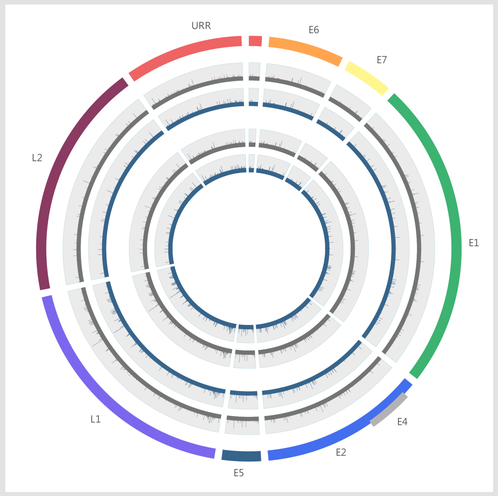

Figure 1: Group A Entropy and MAF, Case and Control Calculation

Middle Ring Pair: Entropy of all Group A SNPs for the control and case groups. **Inner Ring Pair:** MAF Calculation for the same control and case groups. Wedges are the viral gene regions E1, E2-7, L1-2.

The entropy and minor allele frequency (MAF) for both the control and case groups are similar, with only a few points of differences indicating that both groups have similar signals and are hard to differentiate. Entropy data was scaled by $-log(0.25)$, the largest possible value for entropy.

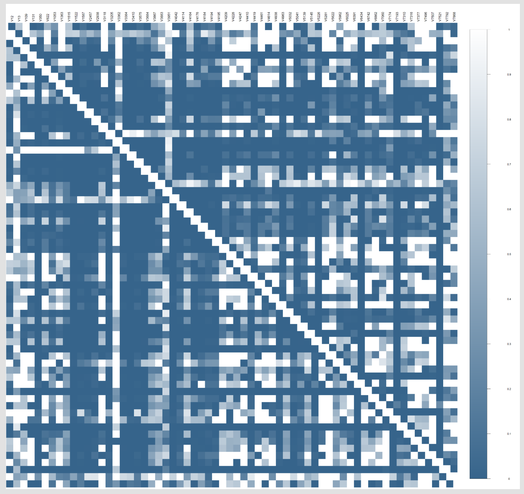

Figure 2: Group A Heatmap based on Fisher Exact P-Values, MAF > 2%

Upper Triangle: Genetic correlation between two loci within the control group. **Lower Triangle:** Genetic correlation between two loci within case group.

The dark blue indicates higher correlation (lower Fisher's p-value) between two loci. There are very few differences between the case and control groups, as expected, since the entropy and MAF signals, as seen in Figure 1, are not very strong.

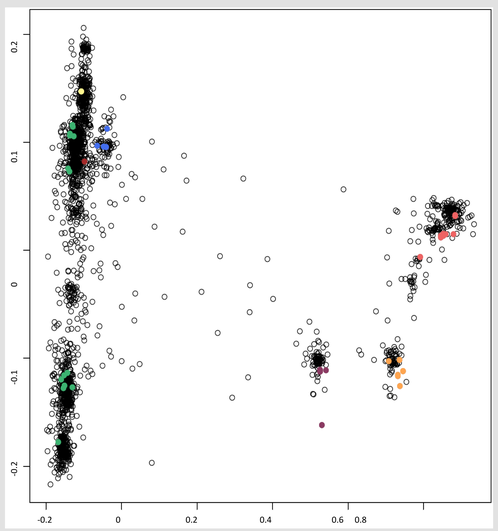

Figure 3: Lineage Distribution

Green: Lineage A1 **Yellow:** A2 **Brown:** A3 **Blue:** A4 **Purple:** B1 **Orange:** C1 **Red:** D1, D2, D3

Principle component analysis (PCA) separating the various lineages. There are four distinct clusters indicating the four main evolutionarily divergent lineages. Some samples are not well clustered indicating that characteristics of two lineages may be present, likely due to coinfection.

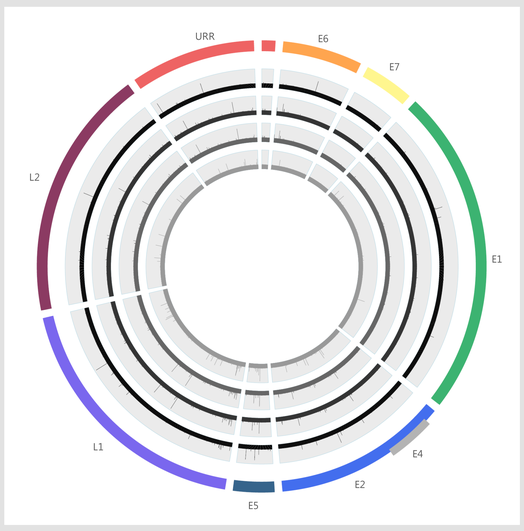

Figure 4: MAF Calculation Amongst Lineages

Inner to Outer Ring: MAF Calculation for lineages Group A, B1, C1, and Group D

Contrary to the comparison of MAF between case and control groups in Group A, there are distinct and subtle differences in MAF between the four lineages.

Prediction Model

- Missing data imputed by mode
- Random forest used for classification, independent predictors are alleles at loci with MAF > 2%
- Training (67%) and independent testing (33%) were created

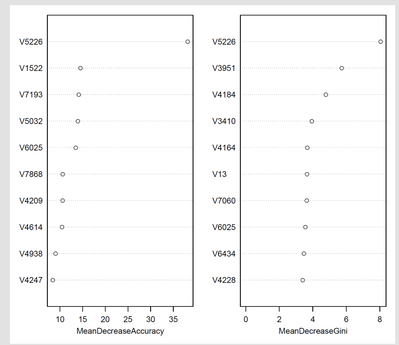

Figure 5: Variable Importance Plot, Top 10 Loci

Left: How much the model fit decreases when you drop a variable- greater the drop the more significant the variable. **Right:** Overall explanatory power (relationship) of the variables. Gini impurity measures how often an element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. (node impurity)

Area under the curve: 56.11%

Out-of-Bag Error Confusion Matrix

	0	1	Class Error
0	297	239	0.445
1	213	254	0.456

Conclusion

Discussion:

- When focusing on the A lineage, there was no obvious difference between cases and controls for genetic variants observed at individual loci.
- Looking jointly at genetic variants at multiple loci might be more helpful in distinguishing between cases and controls
- The random forest model using all common generic variants has an AUC of 56.11% based on a separate testing set
- A larger sample size might be critical for generating prediction model with more accurate prediction

References

1. Mirabello L, Yeager M, Cullen M, et al. HPV16 Sublineage Associations With Histology-Specific Cancer Risk Using HPV Whole-Genome Sequences in 3200 Women. JNCI J Natl Cancer Inst (2016) 108 (9): djw100 doi:10.1093/jnci/djw100
2. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. GLOBOCAN 2012 v1.1, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2014