Data Supplement

Title: Association of HPV35 with cervical carcinogenesis among women of African ancestry: evidence of viral-host interaction with implications for disease intervention

Authors: Maisa Pinheiro, Ph.D.; Julia C. Gage, Ph.D.; Gary M. Clifford, Ph.D.; Maria Demarco, Ph.D.; Li C. Cheung, Ph.D.; Zigui Chen, Ph.D.; Meredith Yeager, Ph.D.; Michael Cullen, Ph.D.; Joseph F. Boland, Ph.D.; Xiaojian Chen, Ph.D. ; Tina Raine-Bennett, M.D.; Mia Steinberg Ph.D.; Sara Bass, Ph.D.; Brian Befano, B.S.; Yanzi Xiao, M.S.; Vanessa Tenet, MSc.; Joan Walker, M.D.; Rosemary Zuna, M.D.; Nancy E. Poitras, P.M.P.; Michael A. Gold, M.D.; Terence Dunn, Ph.D.; Kai Yu, Ph.D.; Bin Zhu, Ph.D.; Laurie Burdett, Ph.D.; Sevilay Turan, Ph.D.; Thomas Lorey, M.D.; Philip E. Castle, Ph.D.; Nicolas Wentzensen, M.D.; Robert D. Burk, Ph.D.; Mark Schiffman, M.D.; Lisa Mirabello, Ph.D.

Table of contents

Study population

PaP Cohort: The HPV Persistence and Progression (PaP) Cohort is a repository of residual cervical specimens from women who underwent cervical cancer screening from 2007-2011 at Kaiser Permanente Northern California (KPNC) using cytology (first specimen) and HPV (second specimen) cotesting. Specimens from 45,000 women testing HPV-positive and about 10,000 women testing HC2-negative by HC2 (Qiagen, Germantown, MD, USA) were collected. After the specimens were used for HR-HPV testing for clinical purposes, the residual specimens were neutralized and archived[1]. De-identified demographic, clinical and cervical cytology/histopathology information was obtained from electronic records. Histological classification was based on the Cervical Intraepithelial Neoplasia (CIN) system, and cases included precancer (defined liberally as CIN grade 2 (CIN2) as well as more definite cases of CIN grade 3 (CIN3) and adenocarcinoma *in situ* (AIS)), and ICC. Outcomes were ascertained through September 21, 2017. Specimens testing positive by HC2 at the KPNC Laboratory (~80% of women with CIN2+ and a third of non-cases) were typed using at least one of the following assays: Linear Array (LA; Roche Molecular Systems, Pleasanton, CA, USA), Cobas (Roche), and lab-specific polymerase chain reaction (PCR). After HPV-genotyping, a subset of 608 HPV35-positive specimens (all 274 CIN2+ cases and 334 of 363 randomly chosen benign infections), were selected for HPV35 whole-genome sequencing, with no restrictions on age or other HPV co-infections (Figure S1). Self-identified race/ethnicity was obtained from the KPNC cancer registry, mortality files, electronic medical records and previous large study databases. The KPNC institutional review board (IRB) approved use of the data, and the National Institutes of Health Office of Human Subjects Research deemed this study exempt from IRB review.

Additional Studies: ALTS, SUCCEED, Biopsy and TCGA: Data from CIN2+ cases in the PaP Cohort were pooled with data from similar cases from the ASCUS-LSIL Triage Study (ALTS), Study to Understand Cervical Cancer Early Endpoints and Determinants (SUCCEED), Biopsy Study and The Cancer Genome Atlas (TCGA) (Figure S1 and Table S5). ALTS was a two-year prospective, randomized clinical trial designed to evaluate three alternative methods (immediate colposcopy, repeat cytology, and HPV testing) for managing atypical squamous cells of undetermined significance (ASC-US) and low-grade squamous intraepithelial lesions (LSIL)[2,3]. SUCCEED was a cross-sectional study conducted to discover and validate biomarkers that can identify HPV infection leading to ICC[4]. The Biopsy Study was designed to understand cervical precancer on the lesion level with the objective of improving colposcopy biopsy practice[5]. HPV genotyping in ALTS, SUCCEED and Biopsy Study was conducted using LA and Line Blot Assay (a prototypic version of LA) from cervical samples collected into liquid medium (PreservCyt; Hologic, Boxborough MA). We considered as probably causal the HPV genotypes found in exfoliative cell specimens collected concurrently or preceding histopathologic diagnosis. TCGA was a convenience sample of resected or biopsied pathology specimens collected from 173 women diagnosed with ICC prior to any chemotherapy or radiation treatment[6]. In TCGA, HPV status was determined using real-time competitive PCR and mass spectroscopy. Local and National Cancer Institute IRBs approved all studies.

IARC international collection: We included exfoliated cervical cell specimens and frozen or formalin fixed paraffin-embedded (FFPE) histopathologic tissues collected in International Agency for Research on Cancer (IARC)-coordinated projects, that studied worldwide HPV prevalence in cervical samples of women with and without ICC. Briefly, cervical cells were collected from women undergoing screening, diagnosis or treatment of cervical cancer. Detailed

sample selection and collection are described previously[7-10]. Cervical specimens were collected from seven regions (Europe, North Africa, Sub-Saharan Africa, East Asia, South Asia, Latin America and Oceania), genotyped using General Primer GP5+/6+ PCR amplification based on Enzyme-Linked Immunosorbent Assay (ELIZA), and then stored at the IARC-biobank. In total, 350 women without ICC (including <CIN1, low-grade lesions and precancer), and 95 women with ICC (histologically confirmed) samples were evaluated. "Benign infections" consist of <CIN1 and low-grade lesions combined. High-grade lesions include high-grade squamous intraepithelial lesions (HSIL), CIN2 and CIN3. Both local and IARC ethical committees approved this study.

HPV35 whole-genome sequencing and lineage assignment

We sequenced 1,053 HPV35-positive specimens from the PaP Cohort (N=608) and IARC biobank (N=445) (Figure S1). DNA was extracted from the stored specimens as previously reported[9,11]. HPV35 DNA was sequenced using the Ion Proton platform[12]. Libraries were prepared following the manufacturer's recommendation using AmpliSeq Library Preparation kit 2.0-96LV (Life Technologies, Part #4480441) and a custom Ion Torrent AmpliSeq panel with 48 pairs of primers targeting the HPV35 whole-genome. This panel was designed and optimized in-house. Raw sequencing data were trimmed and aligned using Ion Torrent Suite software (Life Technologies). The reference genome was the HPV35 complete genome GenBank accession no. X74477. Single Nucleotide Polymorphism (SNPs) were identified using the Torrent Variant Caller v.5.0.3 and annotated using snpEff v.3.6c[13]. The pipeline was executed using Snakemake, and settings and parameters can be found at https://github.com/cgrlab/cgrHPV35. The method and pipeline analysis were designed based on the HPV16 whole-genome assay, that has demonstrated of great efficiency[14]. On average, 40 thousand reads were generated for PaP and

114 thousand reads were generated for the IARC samples by gene ORF, with ~70% of the HPV35 genome covered at least 40x for both cohorts (Table S6).

All 1,053 HPV35 genomes were combined to perform a maximum likelihood (ML) phylogenetic tree, using RAxML MPI v7.2.8 software[15], and a neighbor joining (NJ) tree, using MEGA7 software[16]. For both trees, data were bootstrapped 1,000 times. MEGA7 was used to design the tree. HPV35 was classified into the two previously reported sublineages A1 and A2, and a new lineage named here as B. Divergence among lineages/sublineages was calculated using the p-distance method based on the HPV35 complete genome differences in MEGA7. Nucleotide divergence between A1 and A2 was 0.4% (~30 bp), and the new B lineage differed by 0.8% and 0.7% from A1 and A2, respectively.

Supplemental Tables

Table S1. Counts of IARC-collected specimens by geographic region, country, (sub)lineage and status.

Region Country Sublineage	Status				
	ICC	**HSIL/CIN/CIN3**	**Benign infections**	**Unknown**	**Total**
Africa, North					
Algeria					
A1	5		1		6
Morocco					
A1			1		1
A2	1				1
Africa, Sub-Saharan					
Guinea					
A1	1		10		11
A2	3		13		16
Kenya					
A1	5	1	1		7
A2	2	1			3
Mali					
A1	2				2
A2	2				2
Nigeria					
A1	2		4		6
A2	1	2	20		23
Rwanda					
A1	2	3	47	1	53
A2	3	3	40	1	47
South-Africa					
A1	6	4			10
A2	2				2
Asia, E					
Bhutan					
A1		5	17		22
B			2		2
China					
A1			4		4
A2			1		1
Korea					
A1	1	1			2
Mongolia					
A1	2	2	24	1	29
Philippines					
A1	1				1
Thailand					
A1	2		2		4
Vietnam					
A1			1		1

Asia, S					
India					
A1	4	1	4		9
B	3		12		15
Nepal					
A1			2		2
Pakistan					
A1			2		2
Europe					
Georgia					
A1	2		5		7
Germany					
A1	1				1
Poland					
A1			3		3
Spain					
A1		1			1
Latin America					
Argentina					
A1		1	7		8
A2		1	3		4
Brazil					
A1	10		2		12
A2	3		1		4
Chile					
A1	1				1
Colombia					
A1	3	3			6
Panama					
A1	2				2
Paraguay					
A1	2		3		5
Peru					
A1	6		1		7
A2	1				1
Oceania					
Fiji					
A1			11	1	12
Vanuatu					
A1	1	1	3		5
Total	**82**	**30**	**247**	**4**	**363**

ICC = Invasive cervical cancer; CIN2 = Cervical intraepithelial neoplasia (CIN) grade 2; CIN3 = CIN grade 3; HSIL = high-grade squamous intraepithelial lesions; Benign infections = HPV benign or transient infections with histology classification of CIN grade 1 or less.

Table S2. Prevalence of HPV35 among CIN2, CIN3 and cancer cases testing positive for carcinogenic HPV from 5 NCI studies.

Histology Race/ethnicity	# Tested	**Single HPV35 infection** N	Row % (95% CI)		**HPV35 co-infection, most prevalent** N	Cumulative row % (95% CI)		**HPV35 co-infection, less prevalent** N	Cumulative row % (95% CI)	
Cancer										
White	351	3	0.9	(0-1.8)	0	0.0	-	2	0.6	(0.0-1.4)
Hispanic	70	0	0.0	-	0	0.0	-	1	1.4	(0.0-4.2)
Asian	40	0	0.0	-	1	2.5	(0.0-7.3)	0	0.0	-
African-American	32	1	3.1	(0-9.2)	0	0.0	-	0	0.0	-
CIN3										
White	1858	35	1.9	(1.3-2.5)	4	0.2	(0.0-0.4)	56	3.0	(2.2-3.8)
Hispanic	532	10	1.9	(0.7-3.0)	0	0.0	-	10	1.9	(0.7-3.0)
Asian	392	9	2.3	(0.8-3.8)	1	0.3	(0.0-0.8)	3	0.8	(0-1.6)
African-American	271	15	5.5	(2.8-8.3)	5	1.8	(0.2-3.4)	10	3.7	(1.4-5.9)
CIN2										
White	1821	77	4.2	(3.3-5.2)	15	0.8	(0.4-1.2)	64	3.5	(2.7-4.4)
Hispanic	565	13	2.3	(1.1-3.5)	1	0.2	(0-0.5)	16	2.8	(1.5-4.2)
Asian	434	8	1.8	(0.6-3.1)	4	0.9	(0.0-1.8)	8	1.8	(0.6-3.1)
African-American	369	48	13.0	(9.6-16.4)	21	5.7	(3.3-8.1)	15	4.1	(2.1-6.1)

CIN2 = Cervical intraepithelial neoplasia (CIN) grade 2; CIN3 = CIN grade 3; ICC = Invasive cervical cancer. For each woman with HPV35, her HPV35 infection was classified as either: 1) "Single HPV35 infection" due to single type infection with no other concurrent high-risk HPV type, 2) "HPV35 co-infection, most prevalent" due to positive with multiple high-risk HPV infections but HPV35 was most prevalent, 3) "HPV35 co-infection, less prevalent" due to positive with multiple high-risk HPV infections and one or more were ranked more prevalent or 4) HPV35 negative (not shown).

Table S3. Associations with CIN2+ and CIN3+ by sublineage within HPV35 among women in the Persistence and Progression (PaP) Cohort at Kaiser Permanente Northern California.

Sublineage Race/ethnicity	**Controls (%)**		**CIN2+ (%)**		**OR (95%CI)**	**CIN3+ (%)**		**OR (95%CI)**
A1								
Non-White	126	(55.5)	76	(40.6)	ref	23	(41.1)	ref
White	101	(44.5)	111	(59.4)	**1.8 (1.2-2.7)**	33	(58.9)	1.8 (1.0- 3.2)
Non-Hispanic	177	(78.0)	163	(87.2)	ref	46	(82.1)	ref
Hispanic	50	(22.0)	24	(12.8)	0.5 (0.3-0.9)	10	(17.9)	0.8 (0.4-1.6)
Non-Asian	196	(86.3)	169	(90.4)	ref	47	(83.9)	ref
Asian	31	(13.7)	18	(9.6)	0.7 (0.4-1.2)	9	(16.1)	1.2 (0.5-2.7)
Non-African-American	182	(80.2)	153	(81.8)	ref	52	(92.9)	ref
African-American	45	(19.8)	34	(18.2)	0.9 (0.6-1.5)	4	(7.1)	**0.3 (0.1-0.9)**
p-value (Wald-test)					**<0.001**			**0.03**
A2								
Non-White	23	(53.5)	18	(85.7)	ref	9	(90.0)	ref
White	20	(46.5)	3	(14.3)	**0.2 (0.0-0.7)**	1	(10.0)	0.1 (0.0-1.1)
Non-Hispanic	33	(76.7)	16	(76.2)	ref	7	(70.0)	ref
Hispanic	10	(23.3)	5	(23.8)	1.0 (0.3-3.5)	3	(30.0)	1.4 (0.3-6.5)
Non-Asian	40	(93.0)	19	(90.5)	ref	9	(90.0)	ref
Asian	3	(7.0)	2	(9.5)	1.4 (0.2-9.1)	1	(10.0)	1.5 (0.1-15.9)
Non-African-American	33	(76.7)	10	(47.6)	ref	5	(50.0)	ref
African-American	10	(23.3)	11	(52.4)	**3.6 (1.2-11.0)**	5	(50.0)	3.3 (0.8-13.8)
p-value (Wald-test)					**0.01**			0.11

CIN2+ = Cervical intraepithelial neoplasia (CIN) grade 2, CIN grade 3 and cancer; CIN3+ = CIN grade 3 and cancer; OR = Odds ratio and 95% confidence intervals (CI) from logistic regression. Bold indicates a p-value <0.05 for logistic regression or Wald-test for heterogeneity.

Table S4. Individual SNPs associated with CIN2+ or CIN3+ by a women's race/ethnicity.

All races			CIN2+										CIN3+									
Position	**Gene**	**Nucleotide**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**
127	E6	C	0.45	0	280	227	0	**0.002**	1.75	1.23	2.51	**0.017**	0.41	0	214	148	0	0.124	1.5	0.9	2.6	-
136	E6	C	0.45	0	280	227	0	**0.002**	1.75	1.23	2.51	**0.017**	0.41	0	214	148	0	0.124	1.5	0.9	2.6	-
163	E6	A*	0.02	7	0	0	436	**0.045**	0.00	0.00	Inf	0.186	0.02	7	0	0	313	0.987	0.0	0.0	Inf	-
341	E6	C	0.48	0	227	250	0	**0.000**	1.94	1.34	2.81	**0.016**	0.49	0	173	165	0	**0.011**	2.1	1.2	3.6	0.250
1386	E1	T	0.33	339	164	0	0	**0.045**	0.68	0.46	0.99	0.186	0.33	119	1	0	240	**0.045**	0.5	0.3	1.0	0.328
2728	E1-E2	G	0.39	0	202	0	312	**0.019**	0.65	0.45	0.93	0.120	0.42	0	154	0	215	0.133	0.7	0.4	1.1	-
2980	E2	C	0.29	337	0	140	0	**0.040**	0.65	0.43	0.98	0.186	0.32	230	0	106	0	0.212	0.7	0.4	1.3	-
4304	L2	A	0.48	261	0	0	245	**0.001**	1.84	1.29	2.64	**0.016**	0.48	177	0	0	188	**0.012**	2.0	1.2	3.4	0.250
4384	L2	G	0.46	238	0	0	274	**0.001**	1.86	1.30	2.65	**0.016**	0.50	183	0	0	185	**0.015**	1.9	1.1	3.3	0.250
5030	L2	G	0.00	0	512	0	2	0.176	-	-	-	-	0.01	0	367	0	2	**0.037**	-	-	-	0.305
5101	L2	T	0.31	0	160	350	0	**0.042**	0.67	0.45	0.99	0.186	0.32	0	119	248	0	**0.038**	0.5	0.3	1.0	0.305
5894	L1	A	0.48	268	0	0	248	**0.001**	1.80	1.26	2.57	**0.017**	0.49	181	0	0	189	**0.015**	1.9	1.1	3.3	0.250
6500	L1	C	0.29	328	0	135	1	**0.037**	0.64	0.42	0.97	0.186	0.31	228	0	103	1	0.155	0.6	0.3	1.2	-
6642	L1	A	0.47	270	235	0	0	**0.001**	1.79	1.25	2.57	**0.017**	0.50	182	180	0	0	**0.027**	1.8	1.1	3.1	0.297
7719	URR	G	0.16	0	424	0	83	**0.049**	1.61	1.00	2.58	0.190	0.15	0	310	0	53	0.270	1.5			-
7758	URR	A	0.38	192	0	0	317	**0.014**	0.63	0.43	0.91	0.099	0.40	145	0	0	219	0.053	0.6			-

White			CIN2+										CIN3+									
Position	**Gene**	**Nucleotide**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**
127	E6	C	0.47	0	108	124	0	**0.012**	1.96	1.16	3.31	**0.035**	0.46	0	81	70	0	0.642	1.2	0.6	2.6	-
136	E6	C	0.47	0	108	124	0	**0.012**	1.96	1.16	3.31	**0.035**	0.46	0	81	70	0	0.642	1.2	0.6	2.6	-
341	E6	C	0.36	0	77	139	0	**0.001**	2.59	1.45	4.60	**0.011**	0.41	0	57	82	0	**0.021**	3.0	1.2	7.4	1
3441	E2-E4	C	0.09	0	209	21	0	**0.009**	0.22	0.07	0.69	**0.030**	0.12	0	132	18	0	0.125	0.2	0.0	1.6	-
4197	NC	T*	0.05	0	11	192	0	**0.020**	0.08	0.01	0.67	**0.047**	0.08	0	10	117	0	0.990	0.0	0.0	Inf	-
4264	L2	G	0.06	219	0	0	14	**0.019**	0.16	0.04	0.74	**0.047**	0.08	143	0	0	12	0.989	0.0	0.0	Inf	-
4304	L2	A	0.37	146	0	0	87	**0.013**	1.98	1.15	3.41	**0.036**	0.41	91	0	0	64	0.086	2.0	0.9	4.6	-

4384	L2	G	0.36	84	0	0	152	**0.007**	2.11	1.22	3.65	**0.026**	0.40	63	0	0	93	0.109	1.9	0.9	4.4	-
5735	L1	A	0.09	20	0	0	204	**0.004**	0.16	0.05	0.56	**0.017**	0.12	18	0	0	130	0.094	0.2	0.0	1.3	-
5894	L1	A	0.37	149	0	0	88	**0.007**	2.10	1.22	3.59	**0.026**	0.42	91	0	0	65	0.078	2.1	0.9	4.7	-
6470	L1	C	0.06	0	222	14	0	**0.018**	0.16	0.04	0.73	**0.047**	0.08	0	144	12	0	0.989	0.0	0.0	Inf	-
6642	L1	A	0.34	153	80	0	0	**0.012**	2.03	1.17	3.53	**0.035**	0.39	94	59	0	0	0.104	2.0	0.9	4.7	-
6743	L1	G	0.05	201	0	0	11	**0.048**	0.21	0.04	0.98	0.100	0.07	129	0	0	9	0.990	0.0	0.0	Inf	-
7034	L1	C	0.08	0	209	19	0	**0.032**	3.17	1.10	9.12	0.074	0.05	0	140	8	0	0.274	2.3	0.5	10.2	-
7719	URR	G	0.18	0	191	0	43	**0.043**	2.02	1.02	3.98	0.095	0.15	0	131	0	23	0.299	1.7	0.6	4.5	-

Hispanic			**CIN2+**										**CIN3+**									
Position	**Gene**	**Nucleotide**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**
7758	URR	A	0.35	31	0	0	58	**0.019**	0.27	0.09	0.81	0.609	0.39	29	0	0	45	0.142	0.4	0.1	1.4	-
326	E6	G	0.20	66	0	0	16	**0.028**	0.10	0.01	0.77	0.609	0.24	51	0	0	16	0.158	0.2	0.0	1.8	-
7418	URR	G	0.18	72	0	0	16	**0.033**	0.10	0.01	0.84	0.609	0.22	57	0	0	16	0.168	0.2	0.0	1.9	-
5245	L2	C	0.19	0	63	15	0	**0.037**	0.11	0.01	0.88	0.609	0.23	0	49	15	0	0.199	0.2	0.0	2.1	-
6500	L1	C	0.33	51	0	25	0	**0.034**	0.27	0.08	0.91	0.609	0.37	39	0	23	0	**0.034**	0.3	0.1	0.9	-

African-American			**CIN2+**										**CIN3+**									
Position	**Gene**	**Nucleotide**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**	**MAF**	**No. A**	**No. T**	**No. C**	**No. G**	**P-value**	**OR**	**CI lower**	**CI upper**	**FDR p-value**
91	URR	G	0.04	96	0	0	4	0.250	3.86	0.39	38.43	-	0.05	61	0	0	3	**0.034**	15.4	1.2	193.0	0.114
748	E7	A*	0.10	10	0	0	90	0.322	1.96	0.52	7.43	-	0.11	7	0	0	57	**0.035**	6.4	1.1	35.6	0.114
1386	E1	T	0.49	48	50	0	0	0.321	0.67	0.30	1.48	-	0.49	32	31	0	0	**0.036**	0.1	0.0	0.9	0.114
1464	E1	C	0.06	79	0	5	0	0.144	5.29	0.57	49.54	-	0.07	50	0	4	0	**0.008**	27.0	2.3	311.2	0.106
2185	E1	C	0.04	0	0	4	85	0.231	4.08	0.41	40.88	-	0.05	0	0	3	55	**0.032**	16.3	1.3	208.3	0.114
3142	E2	G	0.05	0	0	90	5	0.135	5.47	0.59	50.95	-	0.07	0	0	57	4	**5.76E-03**	31.2	2.7	358.7	0.106
3431	E2-E4	T	0.04	0	4	95	0	0.242	3.95	0.40	39.38	-	0.05	0	3	61	0	**0.034**	15.4	1.2	193.0	0.114
3441	E2-E4	C	0.18	0	81	18	0	0.298	1.73	0.62	4.84	-	0.20	0	51	13	0	**0.010**	7.3	1.6	33.4	0.108
3465	E2-E4	A	0.47	52	0	0	47	0.393	0.71	0.32	1.57	-	0.50	32	0	0	32	**0.033**	0.1	0.0	0.8	0.114
4197	NC	T	0.08	0	7	77	0	0.165	3.33	0.61	18.26	-	0.08	0	4	47	0	**0.021**	14.7	1.5	143.7	0.114
5101	L2	T	0.49	0	51	49	0	0.237	0.62	0.28	1.37	-	0.50	0	32	32	0	**0.033**	0.1	0.0	0.8	0.114
5888	L1	A	0.05	5	0	0	95	0.144	5.27	0.57	48.93	-	0.06	4	0	0	60	**0.007**	27.0	2.4	302.2	0.106

5939	L1	G	0.46	53	0	0	46	0.164	0.56	0.25	1.26	-	0.47	34	0	0	30	**0.046**	0.1	0.0	1.0	0.123
6431	L1	C#	0.05	0	95	5	0	0.144	5.27	0.57	48.93	-	0.05	0	61	3	0	**0.034**	15.4	1.2	193.0	0.114
6500	L1	C	0.45	51	0	41	1	0.155	0.55	0.24	1.26	-	0.46	32	0	26	1	**0.049**	0.1	0.0	1.0	0.126
7315	URR	G	0.05	95	0	0	5	0.144	5.27	0.57	48.93	-	0.06	60	0	0	4	**0.007**	27.0	2.4	302.2	0.106
7333	URR	A	0.05	5	0	0	95	0.144	5.27	0.57	48.93	-	0.06	4	0	0	60	**0.007**	27.0	2.4	302.2	0.106
7535	URR	T	0.47	0	47	53	0	0.206	0.60	0.27	1.33	-	0.47	0	30	34	0	**0.046**	0.1	0.0	1.0	0.123

CIN2+ = Cervical intraepithelial neoplasia (CIN) grade 2, CIN grade 3 and cancer; CIN3+ = CIN grade 3 and cancer; MAF = minor allele frequency; SNP = single nucleotide polymorphism. *potential APOBEC3-induced variant. #Independent SNP. Correction for multiple comparisons were performed using false discovery rate (FDR) method for all SNPs with a minor allele frequency (MAF) >0.03 for each race/ethnic group. No SNPs were significantly associated with CIN2+/CIN3+ among Asian women.

Table S5. Number of CIN2, CIN3 and cancer cases from 5 NCI studies: HPV Persistence and Progression (PaP) Cohort, ASCUS-LSIL Triage Study (ALTS), Study to Understand Cervical Cancer Early Endpoints and Determinants (SUCCEED), Biopsy Study and The Cancer Genome Atlas (TCGA).

Histology Race/ethnicity	**Total**	**PaP Cohort**	**ALTS**	**SUCCEED/ Biopsy**	**TCGA**
Cancer	**493**	**170**	**-**	**166**	**157**
White	351	95	-	142	114
Hispanic	70	47	-	13	10
Asian	40	20	-	2	18
African-American	32	8	-	9	15
CIN3	**3053**	**2174**	**499**	**380**	**-**
White	1858	1210	357	291	-
Hispanic	532	459	12	61	-
Asian	392	372	17	3	-
African-American	271	133	113	25	-
CIN2	**3189**	**2359**	**332**	**498**	**-**
White	1821	1253	210	358	-
Hispanic	565	473	17	75	-
Asian	434	419	10	5	-
African-American	369	214	95	60	-

CIN2 = Cervical intraepithelial neoplasia (CIN) grade 2; CIN3 = CIN grade 3; ICC = Invasive cervical cancer.

Table S6. Number of reads generated by gene ORF across samples from each study.

Number of reads				
	PaP		IARC	
Gene	Total reads	Average reads by sample	Total reads	Average reads by sample
URR	23128388	42205.1	24611094	57772.5
E6	13819137	25217.4	17601155	41317.3
E7	21654015	39514.6	31827826	74713.2
E1	33641091	61388.9	179172117	420591.8
E2	10348300	18883.8	56319314	132205.0
E4	7379624	13466.5	5923536	13905.0
E5	1012413	1847.5	6526132	15319.6
L2	42606623	77749.3	45594152	107028.5
L1	47643575	86940.8	70833456	166275.7
Average reads by gene		40801.5		114347.6

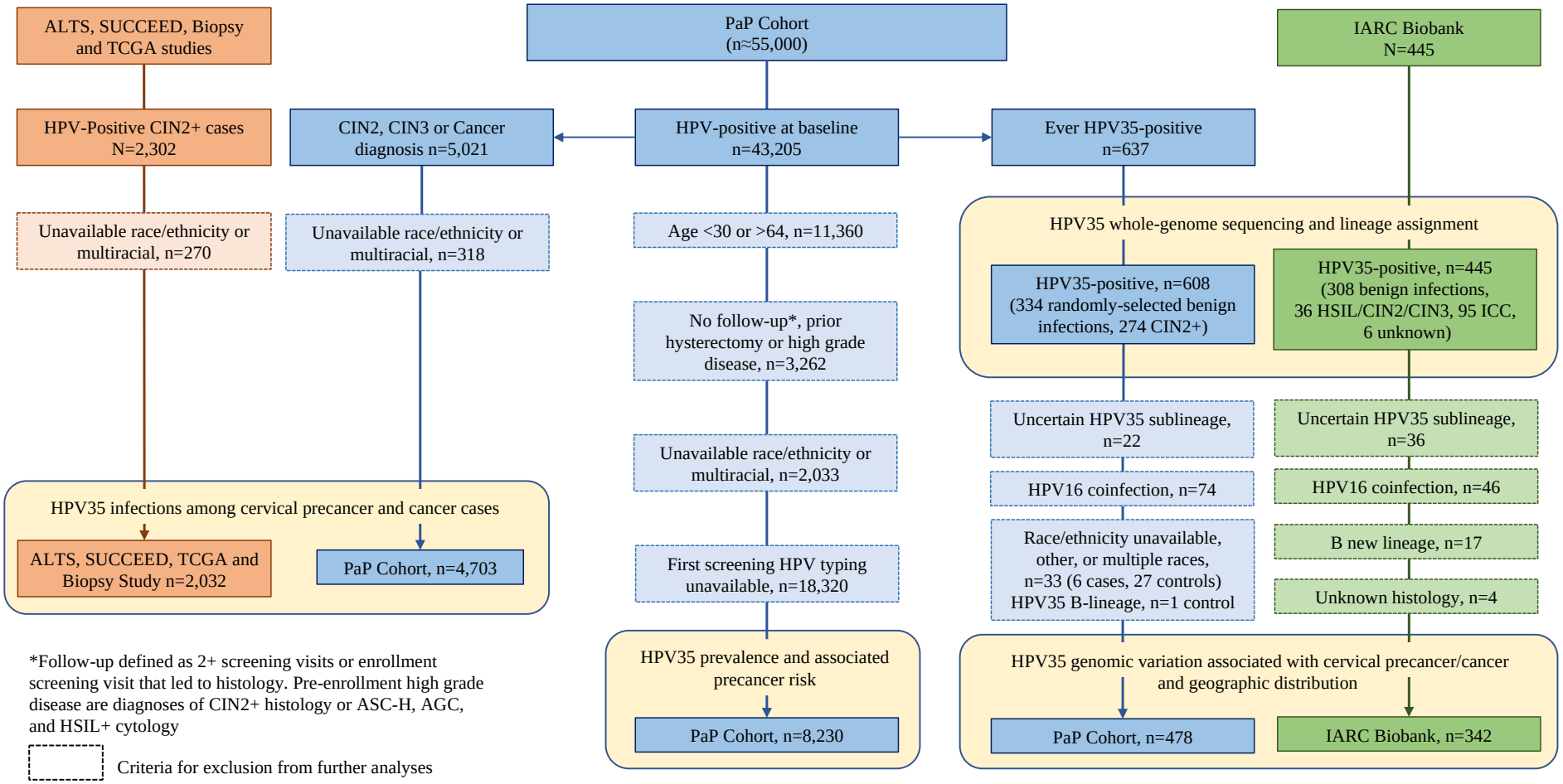

Figure S1. Analytical datasets of four analyses related to HPV35 infection from the IARC Biobank, PaP Cohort, ALTS, SUCCEED, Biopsy and TCGA studies.

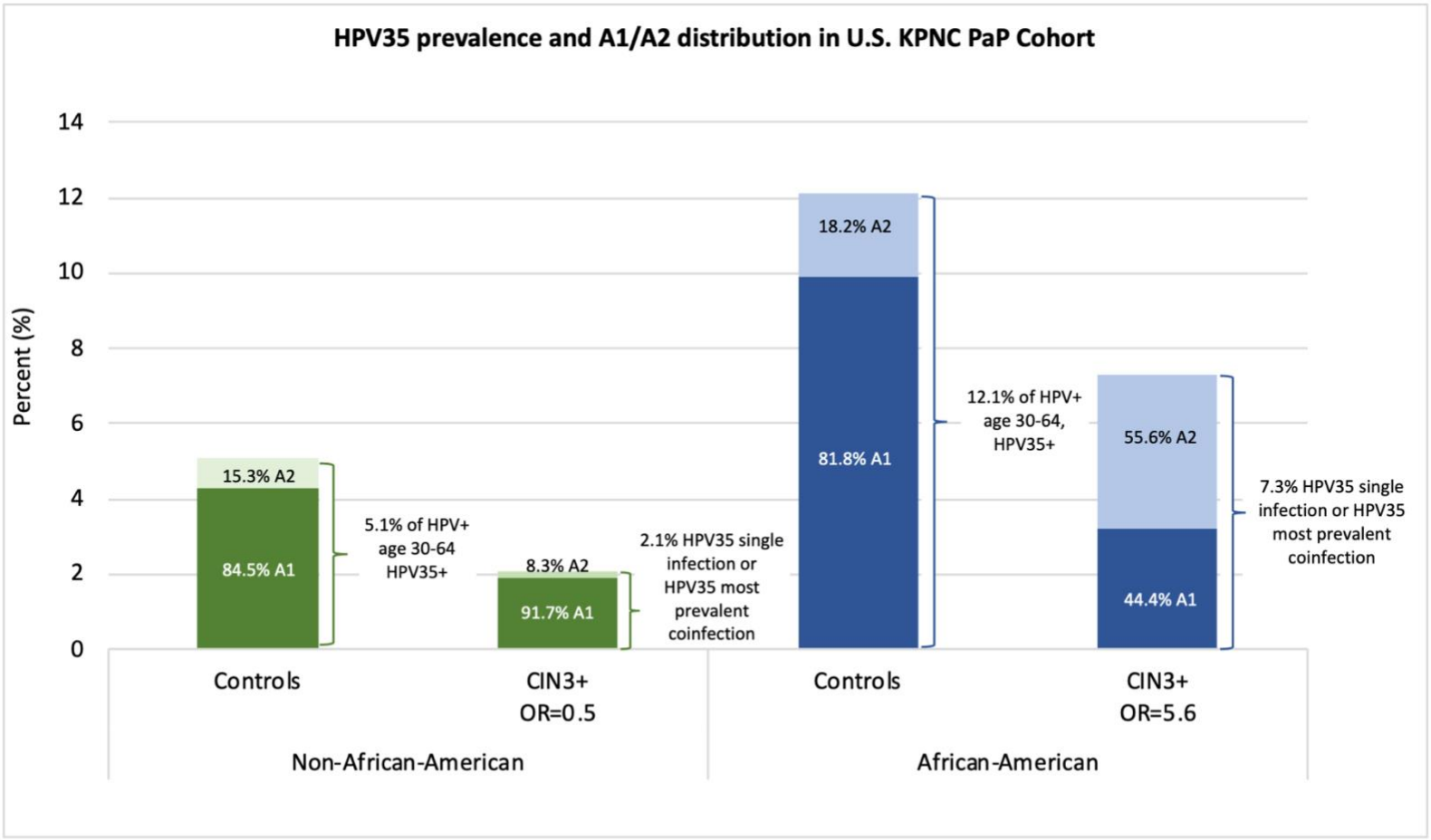

Figure S2. HPV35 prevalence among HPV-positive women and CIN3+ cases and distribution of A1/A2 sublineages within cases and controls for African-American and non-African-American Women in the U.S. KPNC PaP Cohort.

Supplementary references

1. LaMere BJ, Howell R, Fetterman B, Shieh J, Castle PE. Impact of 6-month frozen storage of cervical specimens in alkaline buffer conditions on human papillomavirus genotyping. *J Virol Methods.* 2008;151(2):298-300.
2. Results of a randomized trial on the management of cytology interpretations of atypical squamous cells of undetermined significance. *American journal of obstetrics and gynecology.* 2003;188(6):1383-1392.
3. A randomized trial on the management of low-grade squamous intraepithelial lesion cytology interpretations. *American journal of obstetrics and gynecology.* 2003;188(6):1393-1400.
4. Wang SS, Zuna RE, Wentzensen N, et al. Human papillomavirus cofactors by disease progression and human papillomavirus types in the study to understand cervical cancer early endpoints and determinants. *Cancer Epidemiol Biomarkers Prev.* 2009;18(1):113-120.
5. Wentzensen N, Walker JL, Gold MA, et al. Multiple biopsies and detection of cervical cancer precursors at colposcopy. *J Clin Oncol.* 2015;33(1):83-89.
6. Cancer Genome Atlas Research N, Albert Einstein College of M, Analytical Biological S, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature.* 2017;543(7645):378-384.
7. Clifford GM, Gallus S, Herrero R, et al. Worldwide distribution of human papillomavirus types in cytologically normal women in the International Agency for Research on Cancer HPV prevalence surveys: a pooled analysis. *The Lancet.* 2005;366(9490):991-998.
8. Muñoz N, Bosch FX, de Sanjosé S, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med.* 2003;348(6):518-527.
9. Cornet I, Gheit T, Franceschi S, et al. Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR. *J Virol.* 2012;86(12):6855-6861.
10. Bosch FX, Manos MM, Muñoz N, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. *J Natl Cancer Inst.* 1995;87(11):796-802.
11. Burk RD, Ho GY, Beardsley L, Lempa M, Peters M, Bierman R. Sexual behavior and partner characteristics are the predominant risk factors for genital human papillomavirus infection in young women. *J Infect Dis.* 1996;174(4):679-689.
12. Hyun N, Cheung LC, Pan Q, Schiffman M, Katki HA. Flexible risk prediction models for left or interval-censored data from electronic health records. *Ann Appl Stat.* 2017;11(2):1063-1084.
13. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly.* 2012;6(2):80-92.
14. Cullen M, Boland JF, Schiffman M, et al. Deep sequencing of HPV16 genomes: A new high-throughput tool for exploring the carcinogenicity and natural history of HPV16 infection. *Papillomavirus Res.* 2015;1:3-11.
15. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006;22(21):2688-2690.

16. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution.* 2016;33(7):1870-1874.