Supplementary File for Transcriptionally active HPV-positive head and neck squamous cell carcinoma is associated with a coordinated meiotic gene programme and promoter hypomethylation

Suganth Gopinath Sangeetha[1#], Rejoy Besra[2#], Naisarg Patel[3,8], Prakash Lingasamy[4,5], Vino Sundararajan[1], Sajitha Lulu Sudhakaran[1], Andres Salumets[3,6,7*], Vijayachitra Modhukur[3,6*]

1. Integrative Multiomics Lab, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, India
2. Department of Bioscience and Biotechnology, IIT Kharagpur, Kharagpur, West Bengal, India
3. Celvia CC AS, 50411 Tartu, Estonia
4. Laboratory of Precision and Nanomedicine, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia
5. Nalam Biosciences OÜ, Tartu, Estonia
6. Department of Obstetrics and Gynecology, Institute of Clinical Medicine, University of Tartu, L. Puusepa 8, Tartu, Estonia
7. Division of Obstetrics and Gynecology, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, and Karolinska University Hospital, Stockholm, Sweden
8. Present Address: Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen, Auf der Morgenstelle 24, Tübingen, Germany

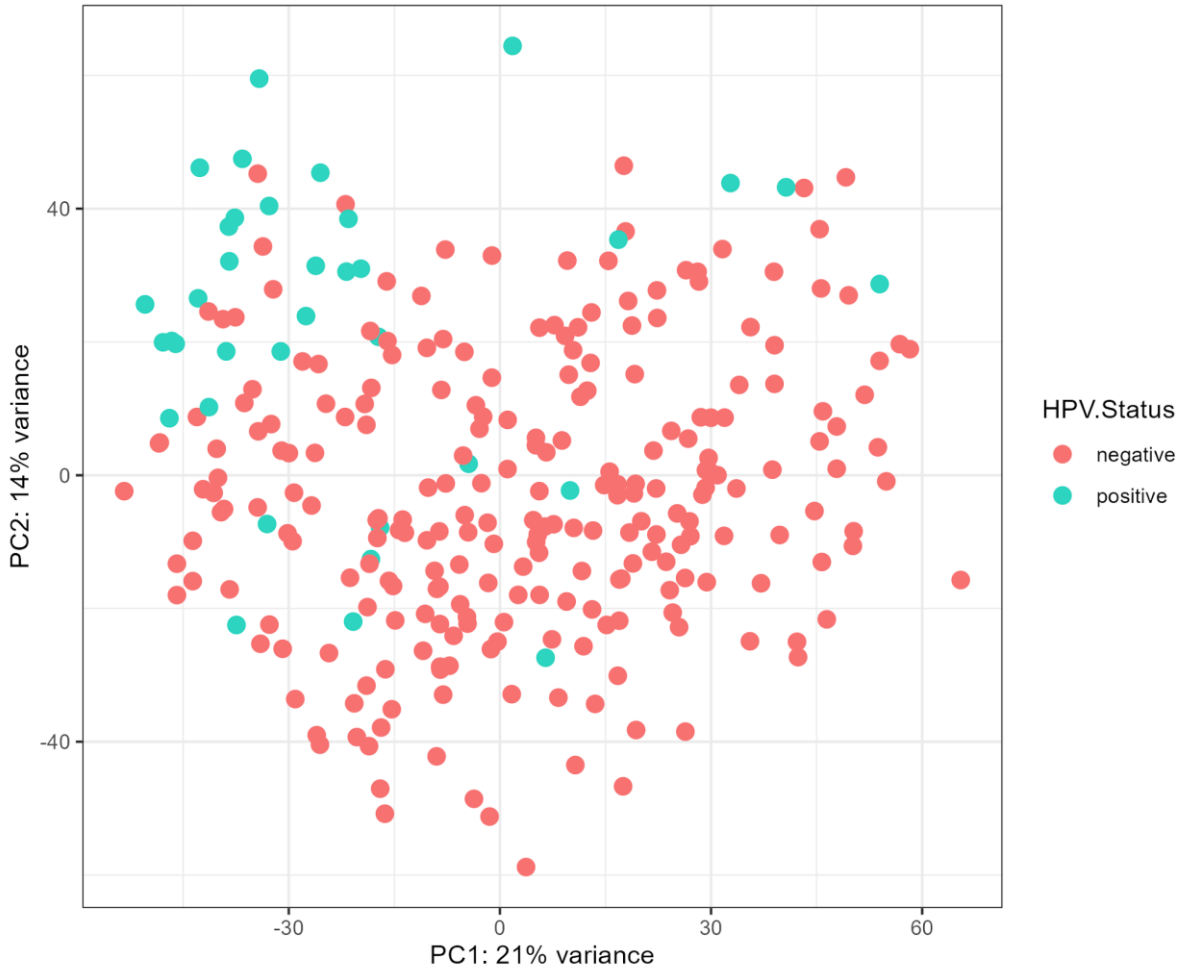
[#]These authors share joint first authorship

[*]To whom correspondence should be addressed:

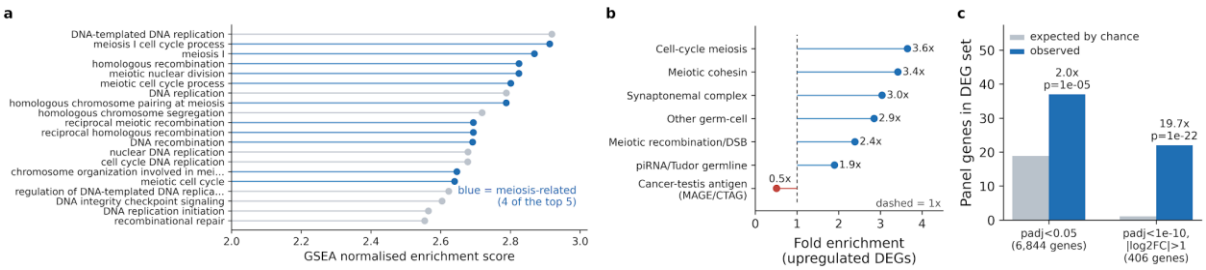
Dr. Vijayachitra Modhukur, Department of Obstetrics and Gynaecology, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia. E-mail: vijayachitra.modhukur@ut.ee

Prof. Andres Salumets, Division of Obstetrics and Gynaecology, Department of Clinical Science, Intervention and Technology (CLINTEC), Karolinska Institutet, and Karolinska University Hospital, Stockholm, Sweden. E-mail: andres.salumets@ki.se

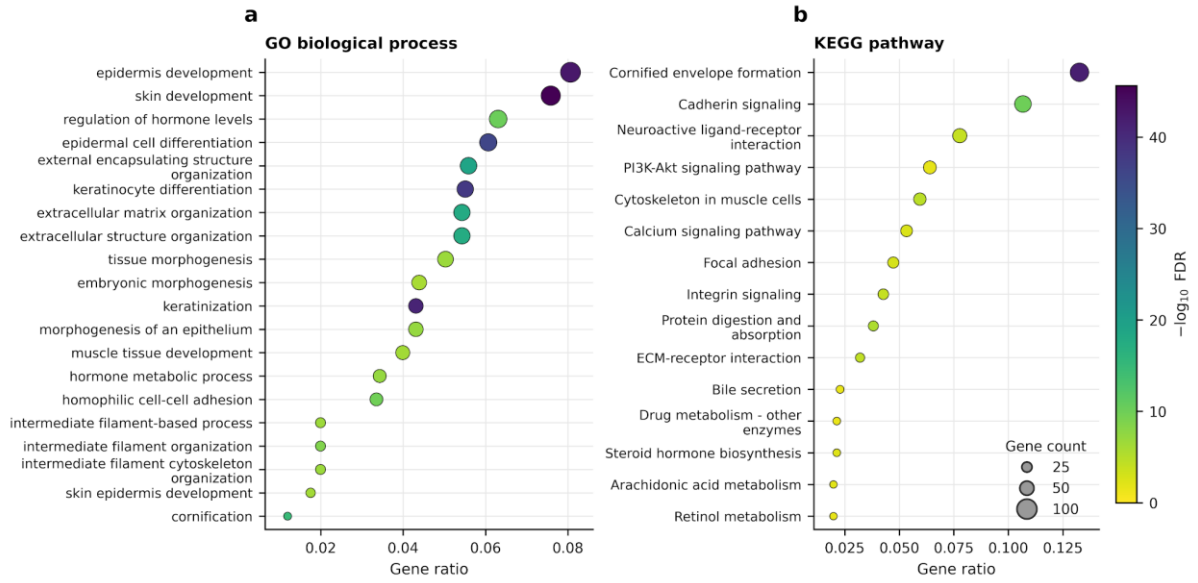
Supplementary figures



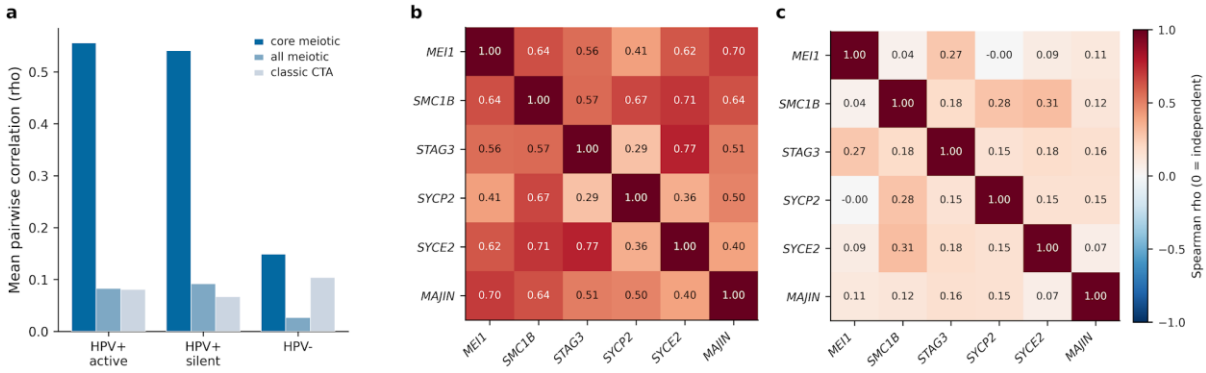
Supplementary Fig. S1. Principal component analysis of the 279 TCGA-HNSC tumours (variance-stabilised expression). HPV+ and HPV− tumours separate partially along PC1 (22% of variance) and PC2 (14%).



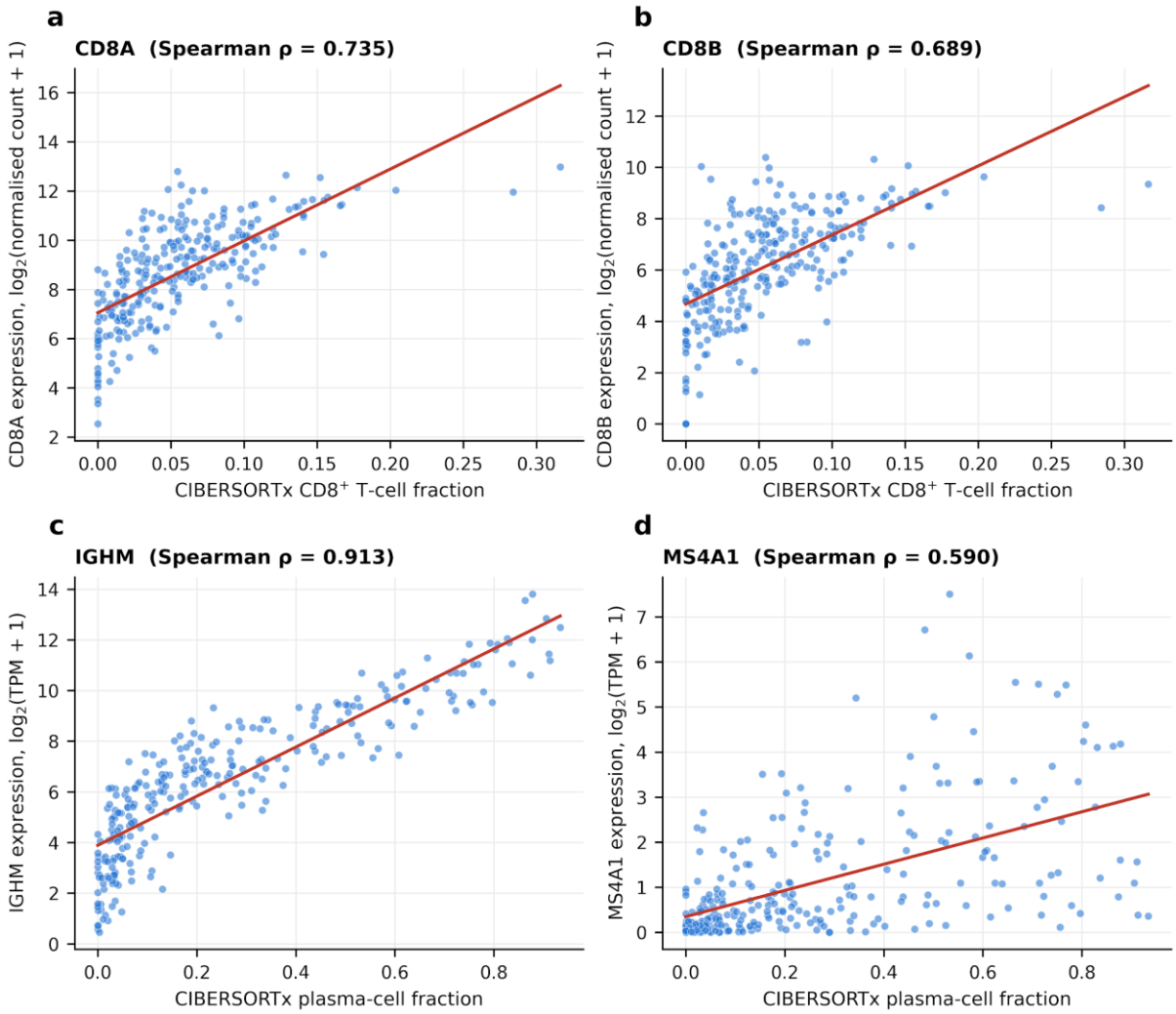
Supplementary Fig. S2. The meiotic hypothesis originates in an unbiased analysis. (a) The 20 most positively enriched GSEA biological-process terms in HPV+ versus HPV− TCGA-HNSC tumours; meiosis-related terms in blue. (b) Fold enrichment of each panel category among genes upregulated in HPV+ tumours (FDR < 0.05); the MAGE/CTAG comparator (red) is the only category not enriched. (c) Panel genes observed among upregulated differentially expressed genes versus the hypergeometric expectation, at two stringency thresholds.



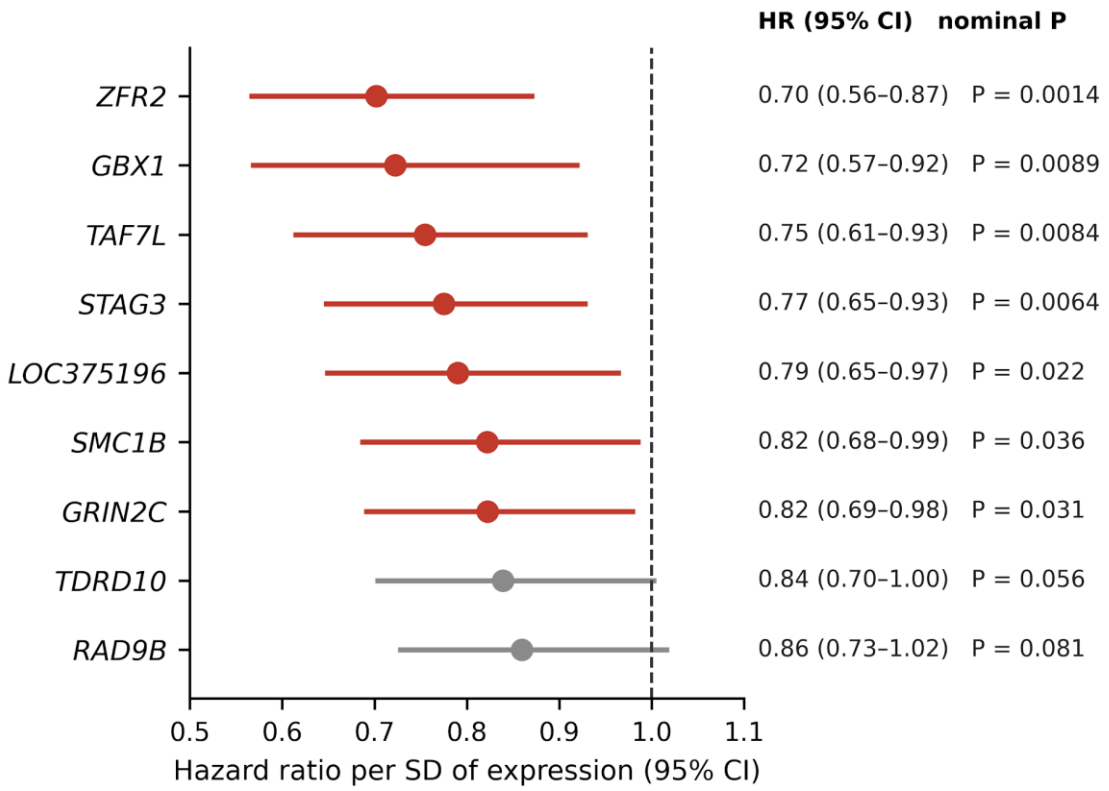
Supplementary Fig. S3. Pathways downregulated in HPV+ tumours. Left, GO biological-process over-representation among downregulated genes; right, KEGG pathway over-representation among downregulated genes, led by cornified envelope formation and cadherin signalling.



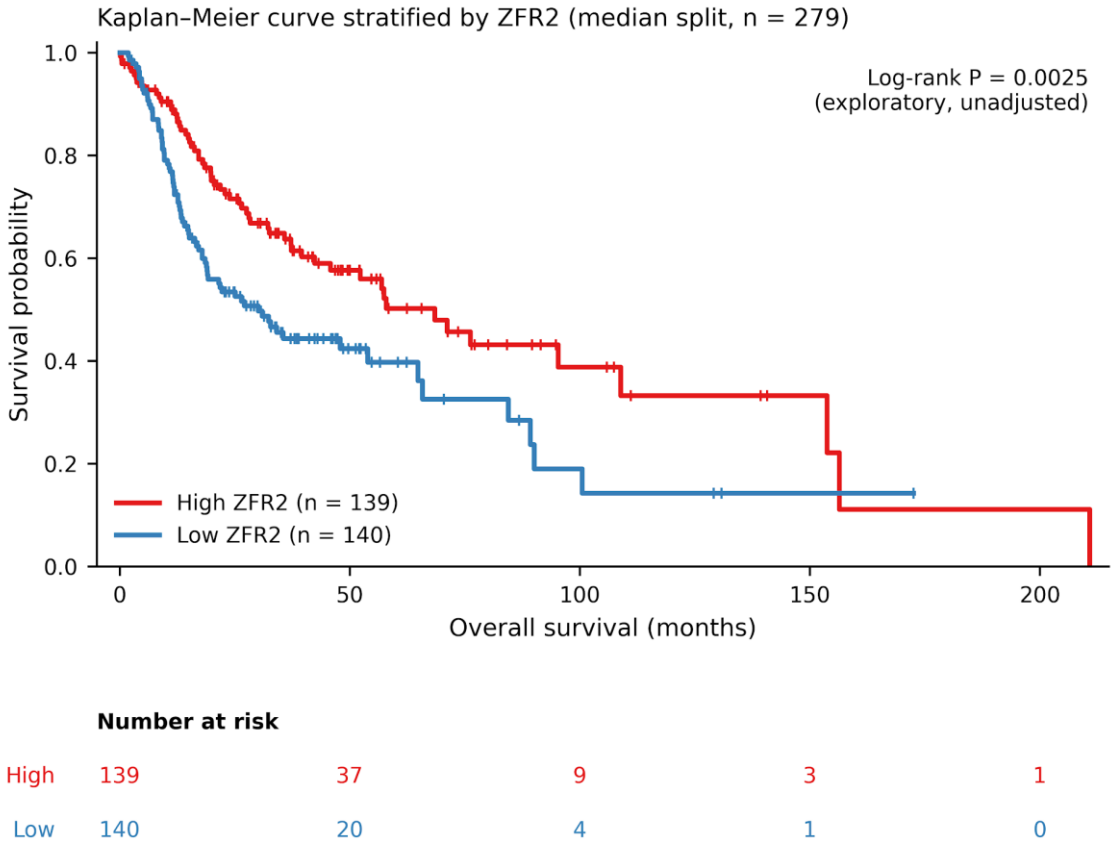
Supplementary Fig. S4. The core meiotic genes are co-expressed as a module (GSE65858). (a) Mean pairwise Spearman correlation within the six-gene core module, the wider meiotic panel and the classic cancer–testis antigens, by group. (b) Pairwise correlation matrix of the core genes in virus-active HPV+ tumours (n = 35). (c) The same matrix in HPV− tumours (n = 196), in which no module structure is present. The contrast is not attributable to group size (size-matched permutation P = 0.001; 1,000 random subsets of 35 HPV− tumours, null 95% interval 0.027–0.273).



Supplementary Fig. S5. Validation of CIBERSORTx CD8+ T-cell and plasma-cell estimates. CD8+ T-cell fraction versus CD8A (a; Spearman ρ = 0.735) and CD8B (b; ρ = 0.689) expression, plotted as log2(DESeq2-normalised count + 1); plasma-cell fraction versus IGHM (c; ρ = 0.913) and MS4A1 (d; ρ = 0.590) expression, plotted as log2(TPM + 1) of the values supplied to CIBERSORTx; n = 279 tumours. Because Spearman ρ is rank-based, the coefficients are identical to those obtained with the DESeq2 variance-stabilised expression reported in Supplementary Table S8 and Methods 2.8.



Supplementary Fig. S6. Exploratory univariable Cox models of overall survival for the nine leading HPV-upregulated genes (n = 279). HRs per standard deviation of variance-stabilised expression with 95% confidence intervals; red marks nominal $P < 0.05$; estimates are listed at right. Models are not adjusted for HPV status or other covariates.



Supplementary Fig. S7. Kaplan–Meier overall survival of the 279 TCGA-HNSC tumours split at the median of *ZFR2* expression (139 high, 140 low; exploratory; unadjusted). The log-rank P value is 0.0025 and the numbers at risk at 0, 50, 100, 150 and 200 months are shown below the axis.

Supplementary tables

Supplementary Table S1. The six-gene core meiotic module, its functional components, gene symbols in the single-cell matrices and membership of the four-gene meiosis-specific score.

Gene	Functional component	Symbol in GSE181919 / GSE182227	Four-gene score
SYCP2	Synaptonemal complex	*SYCP2*	Yes
SYCE2	Synaptonemal complex	*SYCE2*	Yes
MAJIN	Synaptonemal-complex attachment machinery	*C11orf85* (previous symbol)	Yes
SMC1B	Meiosis-specific cohesin	*SMC1B*	Yes
STAG3	Meiosis-specific cohesin	*STAG3*	No (detected in B/plasma cells)
MEI1	Recombination initiation	*MEI1*	No (detected in B/plasma cells)

Supplementary Table S2. The twelve most significant up- and ten most significant downregulated genes in HPV+ versus HPV− TCGA-HNSC tumours (DESeq2; ranked by FDR, ties broken by P value). Panel, member of the 95-gene panel (cta_geneset.csv).

Upregulated gene	$\log_2$FC	FDR	Panel	Downregulated gene	$\log_2$FC	FDR
LOC375196	5.48	5.1×10^{-107}		*LINC02700*	−5.78	7.4×10^{-33}
STAG3	4.63	1.0×10^{-100}	Yes	*DSC1*	−5.20	1.1×10^{-32}
TAF7L	5.81	9.9×10^{-100}	Yes	*FOLR3*	−4.67	6.3×10^{-32}
SMC1B	5.91	6.7×10^{-91}	Yes	*AL033397.1*	−7.59	1.7×10^{-31}
ZFR2	6.52	4.5×10^{-86}	Yes	*MUCL1*	−5.44	2.3×10^{-31}
RAD9B	2.79	1.2×10^{-83}	Yes	*RDH12*	−4.41	2.4×10^{-31}
TDRD10	4.48	2.6×10^{-69}	Yes	*SLC3A2*	−1.46	8.2×10^{-31}
GBX1	6.95	3.2×10^{-68}	Yes	*IGFL3*	−5.44	1.0×10^{-29}
GRIN2C	3.04	2.3×10^{-64}	Yes	*SLC35F3*	−3.25	1.2×10^{-28}
ABCA17P	4.98	1.2×10^{-63}		*SPRR4*	−4.49	4.0×10^{-27}
CNTD1	3.18	4.8×10^{-63}	Yes			
C8G	4.06	1.9×10^{-62}				

Supplementary Table S3. Differential expression of immune and epithelial marker genes in the unadjusted and primary-site-adjusted DESeq2 models (TCGA-HNSC; 36 HPV+ vs 243 HPV−).

Gene	Role	$\log_2$FC	FDR	$\log_2$FC, site-adjusted	FDR, site-adjusted
IGHM	Immunoglobulin heavy chain	2.44	2.0×10^{-8}	2.63	1.8×10^{-8}
MS4A1	Mature B cell (CD20)	2.07	1.8×10^{-5}	2.03	4.2×10^{-5}
IGHA1	Immunoglobulin heavy chain	1.51	6.4×10^{-4}	1.57	8.6×10^{-4}
CD79A	B-cell receptor	1.36	6.7×10^{-4}	1.40	8.3×10^{-4}
CD79B	B-cell receptor	1.16	7.6×10^{-5}	1.10	3.1×10^{-4}
IGKC	Immunoglobulin light chain	1.14	8.3×10^{-3}	1.22	7.5×10^{-3}
IGHG1	Immunoglobulin heavy chain	0.93	0.035	0.99	0.038
IGHG3	Immunoglobulin heavy chain	0.98	0.022	1.29	3.6×10^{-3}

Gene	Role	log_2FC	FDR	log_2FC, site-adjusted	FDR, site-adjusted
CD8A	Cytotoxic T cell	1.26	5.3×10^{-5}	1.12	9.2×10^{-4}
CD8B	Cytotoxic T cell	1.18	2.5×10^{-4}	1.07	2.2×10^{-3}
KRT1	Keratinisation	−4.90	4.8×10^{-21}	−4.09	1.1×10^{-13}
KRTDAP	Keratinocyte differentiation	−4.25	1.6×10^{-23}	−3.72	8.3×10^{-17}
KRT14	Basal keratin	−1.87	1.5×10^{-9}	−1.96	1.1×10^{-9}
KRT16	Keratinisation	−1.61	1.2×10^{-6}	−1.60	2.7×10^{-6}
KRT6A	Keratinisation	−1.31	9.5×10^{-9}	−1.25	2.0×10^{-7}
S100A7	Keratinisation	−1.92	1.1×10^{-5}	−1.89	3.3×10^{-5}

Supplementary Table S4. Panel-category enrichment among genes upregulated in HPV+ tumours. Counts are genes testable in the re-annotated table; P values are hypergeometric; P_{adj} is Bonferroni-corrected across the seven categories.

Category	Tested	Up	Expected	Fold	P	P_{adj}
Cell-cycle meiosis	5	4	1.10	3.65	0.010	0.07
Meiotic cohesin	4	3	0.88	3.42	0.04	0.25
Synaptonemal complex	9	6	1.97	3.04	0.005	0.03
Other germ-cell	8	5	1.75	2.85	0.02	0.11
Meiotic recombination/DSB	21	11	4.60	2.39	0.002	0.01
piRNA/Tudor germline	12	5	2.63	1.90	0.10	0.71
Cancer–testis antigen (MAGE/CTAG)	27	3	5.92	0.51	0.95	1.00

Supplementary Table S5. Confounder and sensitivity analyses for the core meiotic signal. Site and proliferation analyses are in GSE65858 (virus-active HPV+ vs HPV−); logistic models predict HPV status from the core meiotic score with the stated covariates (β, meiotic-score coefficient). Purity analyses are in TCGA-HNSC with ABSOLUTE purity estimates (277 of 279 tumours); linear models regress the core meiotic score on HPV status with the stated covariates (β, HPV coefficient). A p16-adjusted model is omitted because p16 contributes to the TCGA HPV-status definition; stricter HPV definitions are used instead. Promoter analyses use GDC-harmonised (TCGA) or deposited (GSE38266) β values and CpGs 1,500 bp upstream to 500 bp downstream of an annotated TSS; the B/plasma marker score is the mean Z-score of *CD19*, *CD79A*, *JCHAIN*, *MS4A1* and *MZB1*. AUC 95% CIs by DeLong's method; Δβ, HPV+ minus HPV− mean β.

Analysis	Cohort (n)	Estimate	P
Within oropharynx only: core score AUC	GSE65858 (28 vs 47)	AUC = 0.960 (95% CI 0.90–1.00)	3.3×10^{-11}
Logistic (outcome: HPV status): meiotic score alone	GSE65858 (231)	β = 4.45; model AUC = 0.959	7.6×10^{-9}
Logistic (outcome: HPV status): meiotic score + proliferation	GSE65858 (231)	β = 4.52; model AUC = 0.961	9.1×10^{-9}
Purity by HPV status (mean ± SD, ABSOLUTE)	TCGA-HNSC (36 vs 241)	0.63 ± 0.16 vs 0.48 ± 0.18	8.0×10^{-6} (Mann–Whitney)
Purity vs core meiotic score	TCGA-HNSC (277)	Spearman ρ = 0.027	0.66
Linear model (outcome: core meiotic score) + purity: HPV coefficient	TCGA-HNSC (277)	β = 2.29 (95% CI 2.13–2.45) unadjusted vs 2.29 (2.13–2.46) adjusted; purity coefficient = −0.022 (−0.33 to 0.28)	0.89 (purity)

Analysis	Cohort (n)	Estimate	P
Linear model (outcome: core meiotic score) + purity + proliferation: HPV coefficient	TCGA-HNSC (277)	β = 2.25 (95% CI 2.08–2.41); purity −0.075 (−0.38 to 0.23), P = 0.63; proliferation 0.085 (0.017–0.154), P = 0.015	1.2×10^{-77} (HPV)
Four-gene score (excludes *MEI1*, *STAG3*): virus-active HPV+ vs HPV− AUC	GSE65858 (35 vs 196)	AUC = 0.957 (95% CI 0.90–1.00)	8.1×10^{-18}
Core-module coherence, transcriptionally silent HPV+ vs size-matched HPV− null	GSE65858 (25)	ρ = 0.541; null 95% interval −0.003 to 0.299	0.001 (permutation)
Four-gene coherence, virus-active / silent HPV+ vs size-matched nulls	GSE65858 (35; 25)	ρ = 0.546 / 0.521 (HPV−: 0.179)	0.001 / 0.002 (permutation)
Core expression score AUC, HPV by p16 or ISH (main rule)	TCGA-HNSC (41 vs 80)	AUC = 0.933 (95% CI 0.87–1.00)	8.0×10^{-15}
Core expression score AUC, HPV by ISH only	TCGA-HNSC (21 vs 65)	AUC = 1.000 (complete separation)	7.1×10^{-12}
Core expression score AUC, p16 or ISH in oropharyngeal tumours only	TCGA-HNSC (34 vs 8)	AUC = 0.974 (95% CI 0.92–1.00)	7.6×10^{-7}
Core promoter composite, HPV by p16 or ISH (main rule)	TCGA-HNSC (44 vs 81)	Δβ = −0.092; AUC = 0.844 (95% CI 0.76–0.93)	2.3×10^{-10}
Core promoter composite, HPV by ISH only	TCGA-HNSC (22 vs 67)	Δβ = −0.116; AUC = 0.964 (95% CI 0.93–1.00)	8.0×10^{-11}
Core promoter composite, p16 or ISH in oropharyngeal tumours only	TCGA-HNSC (35 vs 8)	Δβ = −0.084; AUC = 0.918 (95% CI 0.83–1.00)	5.6×10^{-5}
Promoter Δβ, *MEI1* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.126; −0.116	1.4×10^{-6}; 2.9×10^{-4}
Promoter Δβ, *SMC1B* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.073; −0.048	8.7×10^{-8}; 0.056
Promoter Δβ, *STAG3* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.049; −0.055	7.4×10^{-4}; 0.006
Promoter Δβ, *SYCP2* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.207; −0.110	3.4×10^{-13}; 0.001
Promoter Δβ, *SYCE2* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.007; 0.005	0.40; 0.34
Promoter Δβ, *MAJIN* (TCGA; GSE38266)	44 vs 81; 21 vs 21	−0.092; −0.135	9.1×10^{-6}; 9.7×10^{-5}
Set-level promoter β, 80 core-module probes (HPV+ vs HPV−)	GSE38266 (21 vs 21)	0.257 vs 0.334	1.1×10^{-6}
Promoter methylation vs expression, core-module composites (four-gene)	TCGA-HNSC (121)	Spearman ρ = −0.641 (−0.632)	2.2×10^{-15} (8.1×10^{-15})
Core promoter composite: HPV Δβ unadjusted vs adjusted for B/plasma marker score	TCGA-HNSC (121)	−0.100 vs −0.098	9.4×10^{-15} vs 8.5×10^{-13}
Core promoter composite: HPV Δβ unadjusted vs adjusted for CIBERSORTx B + plasma fraction	TCGA-HNSC (53)	−0.118 vs −0.129	1.5×10^{-7} vs 1.2×10^{-7}

Analysis	Cohort (n)	Estimate	P
Four-gene promoter composite: HPV Δβ unadjusted vs adjusted for B/plasma marker score	TCGA-HNSC (121)	−0.102 vs −0.103	3.8×10^{-15} vs 1.2×10^{-13}
MEI1 promoter: HPV Δβ unadjusted vs adjusted for B/plasma marker score	TCGA-HNSC (121)	−0.137 vs −0.115	4.1×10^{-10} vs 2.9×10^{-7}

Supplementary Table S6. CIBERSORTx (LM22) estimated fractions of all 22 immune-cell populations in HPV+ (n = 36) and HPV− (n = 243) TCGA-HNSC tumours. Wilcoxon rank-sum test; Benjamini–Hochberg FDR across the 22 populations.

Population	Median HPV+	Median HPV−	P	FDR
Macrophages M0	0.033	0.131	2.5×10^{-6}	5.6×10^{-5}
Plasma cells	0.488	0.145	7.4×10^{-5}	8.1×10^{-4}
T cells CD8	0.071	0.043	2.8×10^{-4}	0.0017
NK cells resting	0.005	0.029	3.1×10^{-4}	0.0017
T cells CD4 memory resting	0.041	0.068	0.0094	0.042
Mast cells activated	0.000	0.000	0.016	0.059
Macrophages M2	0.009	0.022	0.025	0.077
Monocytes	0.000	0.009	0.036	0.100
NK cells activated	0.002	0.000	0.105	0.256
T cells follicular helper	0.057	0.050	0.118	0.260
T cells regulatory (Tregs)	0.012	0.004	0.141	0.283
B cells naive	0.000	0.000	0.179	0.322
B cells memory	0.011	0.014	0.200	0.322
Dendritic cells resting	0.022	0.026	0.205	0.322
Mast cells resting	0.064	0.067	0.287	0.421
T cells CD4 naive	0.000	0.000	0.374	0.514
Macrophages M1	0.041	0.049	0.585	0.756
T cells gamma delta	0.000	0.000	0.678	0.820
Eosinophils	0.000	0.000	0.708	0.820
Dendritic cells activated	0.000	0.000	0.778	0.830
T cells CD4 memory activated	0.002	0.003	0.792	0.830
Neutrophils	0.009	0.009	0.898	0.898

Supplementary Table S7. Log_2 CD8/M2 macrophage ratio (exploratory) by HPV status; Wilcoxon rank-sum $P = 1.0 \times 10^{-4}$.

Group	n	Median	Q1	Q3
HPV−	243	0.78	−0.63	3.16
HPV+	36	3.29	1.82	15.35

Supplementary Table S8. Spearman correlation between CIBERSORTx-estimated fractions and lineage marker-gene expression (n = 279). *CD8A and CD8B, log2(DESeq2-normalised count + 1), as plotted in Supplementary Fig. S5 (Spearman ρ is rank-based and identical for variance-stabilised values)*; *IGHM* and *MS4A1*, TPM values supplied to CIBERSORTx.

Cell population	Marker	Expression	Spearman ρ	P
CD8+ T cells	*CD8A*	log2(DESeq2-normalised count + 1)	0.735	1.2×10^{-48}
CD8+ T cells	*CD8B*	log2(DESeq2-normalised count + 1)	0.689	1.3×10^{-40}
Plasma cells	*IGHM*	TPM	0.913	6.5×10^{-110}
Plasma cells	*MS4A1*	TPM	0.590	1.4×10^{-27}

Supplementary Table S9. Malignant-cell specificity of the six core genes in GSE181919. Detection rates are percentages of cells; the specificity ratio is detection in malignant cells divided by detection in the highest-detecting non-malignant cell type.

Gene	Malignant (%)	Highest other (%)	That cell type	Specificity ratio	Four-gene score
SMC1B	15.3	1.4	T	10.96	Yes
SYCP2	19.0	2.5	T	7.68	Yes
MAJIN	3.5	0.5	Endothelial	6.86	Yes
SYCE2	5.1	2.0	T	2.51	Yes
STAG3	15.4	13.1	B/plasma	1.18	No
MEI1	8.2	37.4	B/plasma	0.22	No

Supplementary Table S10. HPV effect on the per-tumour four-gene meiosis-specific score before and after adjustment for proliferation, by cohort and pooled (linear models on per-tumour means; n, tumours).

Cohort	n	Model	HPV β (SE)	P (HPV)	Proliferation β	P (proliferation)
GSE181919	19	Unadjusted	0.112 (0.025)	2.9×10^{-4}	n/a	n/a
GSE181919	19	+ Proliferation	0.105 (0.025)	6.5×10^{-4}	0.180	0.23
GSE182227	16	Unadjusted	0.176 (0.047)	0.002	n/a	n/a
GSE182227	16	+ Proliferation	0.115 (0.040)	0.01	0.414	0.005
Pooled	35	Unadjusted	0.141 (0.026)	4.5×10^{-6}	n/a	n/a
Pooled	35	+ Proliferation	0.109 (0.022)	2.7×10^{-5}	0.359	2.8×10^{-4}

Supplementary Table S11. Exploratory univariable Cox proportional-hazards models of overall survival (TCGA-HNSC, n = 279; 140 deaths). HR per standard deviation of variance-stabilised expression. Models are not adjusted for HPV status, stage, smoking or treatment, and P values are not corrected for multiple testing.

Gene	HR	95% CI	P (nominal)
ZFR2	0.70	0.56–0.87	0.0014
STAG3	0.77	0.65–0.93	0.0064
TAF7L	0.75	0.61–0.93	0.0084
GBX1	0.72	0.57–0.92	0.0089
LOC375196	0.79	0.65–0.97	0.022
GRIN2C	0.82	0.69–0.98	0.031
SMC1B	0.82	0.68–0.99	0.036
TDRD10	0.84	0.70–1.00	0.056
RAD9B	0.86	0.73–1.02	0.081