Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Suganth Gopinath Sangeetha
Indian Institute of Technology Kharagpur
Rejoy Besra
Indian Institute of Technology Kharagpur
Naisarg Patel
Celvia CC AS
Prakash Lingasamy
University of Tartu
Vino Sundararajan
Vellore Institute of Technology University
Sajitha Lulu Sudhakaran
Vellore Institute of Technology University
Andres Salumets
Karolinska Institutet
Vijayachitra Modhukur
vijayachitra.modhukur@ut.ee

Celvia CC AS

Research Article

Keywords: Human papillomavirus, Head and neck squamous cell carcinoma, Meiotic gene reactivation, Promoter hypomethylation, Transcriptomics, Single-cell RNA sequencing, Cancer–testis antigens, Epigenetic regulation

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11688978/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Human papillomavirus-positive (HPV+) head and neck squamous cell carcinoma exhibits distinct molecular and clinical characteristics. Although aberrant expression of germ-cell genes has been reported, it remains unclear whether this expression reflects stochastic epigenetic de-repression or a coordinated transcriptional programme. This study investigated meiotic gene activation, promoter methylation, viral transcriptional activity and cellular localisation in HPV-associated tumours.

Methods

Transcriptomic profiles of 279 primary tumours from The Cancer Genome Atlas were analysed using differential expression analysis and functional enrichment. A literature-curated 95-gene panel and six-gene meiotic module were evaluated across independent expression (GSE65858), methylation (GSE38266) and two single-cell sequencing cohorts (GSE181919 and GSE182227). Promoter methylation, gene co-expression, immune-cell composition and malignant-cell localisation were assessed using Mann–Whitney tests, Spearman correlations, enrichment tests, receiver operating characteristic analysis and regression models, with multiple-testing correction where applicable.

Results

Transcriptomic comparison of HPV + and HPV− tumours revealed distinct expression programmes involving meiosis, immune responses and epithelial differentiation. A six-gene module comprising *MEI1, SMC1B, STAG3, SYCP2, SYCE2* and *MAJIN* showed coordinated expression and promoter hypomethylation. Module expression was inversely associated with methylation, and the methylation differences were replicated in an independent cohort. The module distinguished transcriptionally active HPV+ tumours from HPV− tumours (AUC = 0.959). In contrast, HPV DNA-positive tumours without detectable viral transcription showed no increase in module expression. Classical *MAGE/CTAG* cancer–testis antigens lacked comparable discrimination. Single-cell analysis and independent tumour-level validation confirmed that the meiotic signature was predominantly localised to malignant epithelial cells. HPV+ tumours also exhibited increased estimated plasma-cell and CD8-positive T-cell fractions. The association between HPV status and the meiotic signature remained significant after adjusting for anatomical site, proliferation, and tumour purity.

Conclusions

Transcriptionally active HPV+ head and neck carcinomas exhibit a coordinated, promoter-hypomethylated meiotic gene programme distinct from conventional cancer–testis antigen activation.

These findings define a tumour-cell transcriptional signature of active HPV infection and warrant mechanistic study of how the virus alters epigenetic regulation, as well as evaluation of the signature as a biomarker.

1. Background

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide, with roughly 900,000 new cases and 450,000 deaths annually [1]. Within HNSCC, tumours associated with human papillomavirus (HPV) infection have risen steeply in incidence over the past three decades, particularly in the oropharynx [2]. HPV+ HNSCC is now recognised as a disease entity distinct from its HPV− counterpart, with a different mutational landscape, a characteristic transcriptional profile and a markedly better prognosis [3–5]. Recent integrative transcriptomic analyses across cervical and head and neck cancers have further identified shared HPV-associated expression signatures involving cell-cycle regulation, DNA replication and epithelial-cell-associated processes [58]. Patients with HPV+ disease are typically younger and have less tobacco and alcohol exposure than patients with HPV− disease [5, 6]. The prognostic difference is large enough to have changed staging and to have motivated de-escalation trials [6], yet its biological basis is only partly understood, and viral status alone does not identify which patients can safely receive less intensive treatment.

Much of this distinction reflects viral oncogene activity. The HPV E6 and E7 oncoproteins inactivate p53 and RB, respectively, but their effects on the host cell extend beyond the loss of these checkpoints. E7 promotes E2F-dependent transcription and p16 accumulation, the latter serving as a clinical surrogate marker of HPV activity, while both oncoproteins interact with chromatin-modifying complexes [7]. These effects contribute to the distinct epigenetic landscape of HPV-positive tumours, characterised by reproducible methylation signatures [8], E7-mediated changes in promoter methylation [9], and widespread gains and losses of methylation at distinct loci [10–12]. The presence of viral DNA does not necessarily indicate active viral transcription, and integration status does not reliably reflect transcriptional activity [13, 14]. Associations between HPV and host gene expression should therefore be evaluated against viral transcriptional activity rather than HPV DNA status alone.

Expression of germ-cell genes in somatic tumours has traditionally been studied in the context of cancer–testis antigens (CTAs). Members of the *MAGE*, *GAGE* and *CTAG* families are typically silenced in normal somatic tissues through promoter methylation but become expressed in tumours following loss of epigenetic repression [15, 16]. Global hypomethylation is a well-established hallmark of malignancy [17], and promoter demethylation can activate cancer-germline genes such as *MAGEA1* [18]. Consequently, ectopic germ-cell gene expression is often regarded as a largely stochastic consequence of epigenetic dysregulation rather than a coordinated transcriptional programme. Because their expression in adult tissues is largely restricted to tumours and the testis, CTAs have attracted interest as immunotherapeutic targets [19, 20]. Immune checkpoint blockade has also shown clinical activity in recurrent or metastatic HNSCC [21].

Not all germ-cell genes expressed in tumours fit this model. Some encode functional components of the meiotic machinery. These include synaptonemal-complex proteins, which mediate homologous chromosome pairing, as well as meiosis-specific cohesins that hold sister chromatids together and factors involved in meiotic recombination and DNA double-strand break (DSB) repair. Unlike the *MAGE/CTAG* families, whose somatic functions remain poorly understood, these proteins have established structural and enzymatic roles. Ectopic expression of meiotic cohesins in mitotic cells can induce chromosomal instability [22], while aberrant expression of the meiotic protein *HORMAD1* in tumours has been associated with genomic

instability, homologous recombination deficiency [23, 24], and protection of stalled replication forks [25]. The concept of “meiomitosis” (partial meiotic activity in mitotic cells) further proposes that partial or mistimed activation of meiotic programmes in somatic cells may contribute to oncogenesis [26]. Thus, whereas isolated CTA de-repression may be a passenger event, coordinated activation of the meiotic machinery may represent a distinct programme with potential relevance to tumour biology.

Previous studies have implicated individual meiotic genes in HPV+ cancers. Pyeon et al. reported the overexpression of testis-specific genes normally restricted to meiotic cells in HPV+ head and neck and cervical cancers, demonstrating that *SYCP2* and *TCAM1* are synergistically induced by E6 and E7, while *STAG3* is upregulated in HPV+ head and neck cancer [27]. Subsequent studies identified *SYCP2* deregulation as a potential marker of early-stage HPV+ oropharyngeal carcinoma [28] and linked its expression to hypomethylation of a CpG-island-shore region in TCGA HNSCC. Both *SYCP2* and *SMC1B* were found to be hypomethylated and overexpressed [8], while an independent integrated analysis of TCGA expression and methylation data highlighted *SYCP2* and *TAF7L* [29]. Despite these findings, it remains unclear whether meiotic gene activation in HPV+ cancers represents a coordinated transcriptional programme or independent de-repression events. Specifically, whether this activation preferentially involves meiotic machinery rather than conventional cancer-testis antigens such as *MAGE/CTAG*, whether the genes are co-expressed, and whether their expression is associated with active viral transcription, promoter hypomethylation, and malignant-cell specificity remain unresolved. These questions have not been systematically investigated together within the same tumours.

Here, we compare the transcriptomes of HPV + and HPV− HNSCC to address these questions. We identify the genes and pathways that distinguish HPV+ from HPV− tumours, and test whether germ-cell gene expression is concentrated in the meiotic machinery and whether these genes form a coordinated transcriptional module. We further examine the relationship between their expression, promoter methylation, viral DNA status, and viral transcriptional activity. Additionally, we assess differences in the immune microenvironment between HPV + and HPV− tumours and use single-cell data to determine which cells express the meiotic programme. Finally, we explore associations between leading HPV-associated genes and overall survival. Throughout these analyses, the classical *MAGE/CTAG* cancer-testis antigens serve as an internal comparator to distinguish coordinated meiotic reactivation from stochastic de-repression. The study design and principal findings are summarised in Fig. 1.

2. Methods

2.1. Data sources and cohort selection

Raw gene-level RNA-seq counts for TCGA-HNSC (STAR–Counts workflow) were retrieved from the Genomic Data Commons using TCGAbiolinks. Analyses were restricted to primary tumours (sample type code 01), and cases lacking HPV status were excluded. Genomic and clinical records were matched using the first 15 characters of the TCGA barcode. Of 566 RNA-seq samples, 520 were primary tumours; 279 had HPV status available (36 HPV + and 243 HPV−). The 279 tumours formed the transcriptomic cohort; cBioPortal expression and methylation cohorts had different sample sizes (Table 2).

Table 1
Baseline clinical and demographic characteristics of the TCGA-HNSC transcriptomic cohort by HPV status.

Variable	**HPV+ (n = 36)**	**HPV− (n = 243)**
Age, years: mean ± SD	57.9 ± 10.8	61.8 ± 12.4
Age, years: median (range)	58.5 (35–82)	62.0 (19–90)
Sex: female / male	4 (11.1%) / 32 (88.9%)	72 (29.6%) / 171 (70.4%)
Stage I / II / III / IV	**2 (5.6%) / 5 (13.9%) / 4 (11.1%) / 12 (33.3%)**	**12 (4.9%) / 39 (16.0%) / 34 (14.0%) / 132 (54.3%)**
Stage unknown	13 (36.1%)	26 (10.7%)
Smoking: never / ever / unknown	10 (27.8%) / 26 (72.2%) / 0	42 (17.3%) / 194 (79.8%) / 7 (2.9%)
Site: oropharynx	22 (61.1%)	11 (4.5%)
Site: oral cavity	12 (33.3%)	160 (65.8%)
Site: larynx / hypopharynx	1 (2.8%) / 1 (2.8%)	71 (29.2%) / 1 (0.4%)
Vital status: alive / dead	26 (72.2%) / 10 (27.8%)	113 (46.5%) / 130 (53.5%)

Gene-panel and methylation analyses used separately processed TCGA-HNSC data obtained through the cBioPortal API [30, 31], comprising RSEM-normalised, Z-scored gene expression and HM450 DNA methylation β-values. Sample sizes were smaller owing to differences in panel-gene coverage and the availability of p16 and in situ hybridisation (ISH) annotations (Table 2). The two TCGA datasets were therefore treated as separate analytical cohorts, and the dataset used is stated for each analysis below. Four additional cohorts were analysed. GSE65858 [14] was used for independent expression validation, with probes mapped to gene symbols using GPL10558 and collapsed by maximum mean intensity. GSE38266 [10] provided promoter-proximal methylation data from oropharyngeal tumours, with probes

mapped using GPL13534. Two single-cell RNA-seq cohorts, GSE181919 [32] and GSE182227 [33], were used to investigate the cellular origin of the meiotic programme. GSE181919 was restricted to primary tumours. GSE182227 was deposited without cell-type or malignancy annotations, which were therefore derived de novo (Section 2.10). In both cohorts, counts were normalised to 10,000 per cell and log-transformed.

Table 2
Cohort characteristics, data platforms and analytical roles. Summary of the datasets used for transcriptomic and methylation analyses, including HPV + and HPV− sample sizes, profiling platforms and analytical purposes.

Cohort	Platform	HPV+	HPV−	Role
TCGA-HNSC, transcriptomic cohort	RNA-seq (STAR counts, GDC)	36	243	Differential expression, enrichment, immune deconvolution, survival
TCGA-HNSC, panel expression	RNA-seq (RSEM, cBioPortal)	41	80	Expression discovery
TCGA-HNSC, panel methylation	Illumina HM450	44	81	Methylation discovery
GSE65858	Illumina HT-12 v4 microarray	60 (35 active, 25 silent)	196	Independent expression validation
GSE38266	Illumina HM450	21	21	Independent methylation replication (oropharynx only)
GSE181919	Single-cell RNA-seq	7 patients	13 patients	Malignant-cell localisation
GSE182227	Single-cell RNA-seq	11 tumours	5 tumours	Independent single-cell replication (oropharynx)

2.2. HPV classification

In the transcriptomic cohort, HPV status was obtained from the TCGA HNSCC characterisation study [3], which primarily classified HPV status using viral transcripts detected by RNA-seq. For panel expression and methylation analyses, tumours were classified as HPV + if either p16 or in situ hybridisation (ISH) was positive; cases without results from any of the assays were excluded. Patient-level annotations were mapped to primary-tumour barcodes and cross-checked against the originating project's HPV classifications for overlapping samples. GSE65858 provided separate HPV DNA and transcriptional activity annotations, allowing classification into transcriptionally active HPV+, transcriptionally silent HPV DNA-positive, and HPV− groups. Cases without usable annotations were excluded. Given the

limited specificity of p16 outside the oropharynx, two stricter TCGA definitions were evaluated in sensitivity analyses. The first required ISH positivity alone, and the second applied the combined p16-or-ISH criterion only to oropharyngeal tumours (tonsil, base of tongue and oropharynx).

2.3. Differential expression analysis

Differential expression between HPV + and HPV− tumours was assessed with DESeq2 (v1.52.0) [34] using a negative-binomial generalised linear model with HPV status as the explanatory variable. Size-factor normalisation and dispersion estimation followed the standard DESeq2 workflow, and P values were adjusted with the Benjamini–Hochberg procedure. Genes with FDR < 0.05 and $|\log_2 FC| > 1$ were considered differentially expressed. Because HPV+ tumours were predominantly oropharyngeal, a sensitivity analysis added primary site as a covariate, with sites grouped as oral cavity, oropharynx and other to avoid

sparse categories. Variance-stabilised expression was used for principal component analysis and survival modelling.

2.4. Functional enrichment and gene set enrichment analysis

Gene Ontology (GO) biological-process and KEGG pathway over-representation analyses were performed separately for up- and downregulated genes with clusterProfiler (v4.20.0) [35], using all genes passing expression filtering as the background; terms with adjusted P < 0.05 were considered significant. Gene set enrichment analysis (GSEA) [36] was performed with clusterProfiler (gseGO) against GO biological-process gene sets, using the full gene ranking by the DESeq2 Wald statistic (duplicate identifiers collapsed by maximum Wald statistic; gene-set size 10–500; Benjamini–Hochberg correction; random seed 123). Because gene symbols in the count matrix were incompletely annotated, identifiers were re-mapped to current HGNC symbols before panel genes were queried. Over-representation of panel genes among upregulated genes was assessed with the hypergeometric test at two significance thresholds; category-level enrichment was assessed with the hypergeometric test with Bonferroni correction across categories, and meiotic categories were compared with the *MAGE/CTAG* comparator by Fisher's exact test.

2.5. Gene panel and core module definition

The panel comprises 95 genes in seven categories. Its scope was motivated by the genome-wide enrichment analysis, which identified meiosis as one of the processes most strongly associated with HPV status. Gene membership was defined from the meiosis literature and CTdatabase family conventions, and category assignments were fixed before the analyses. Six categories cover the germline machinery. These are the synaptonemal complex (n = 10), meiosis-specific cohesin (n = 4), meiotic recombination and DSB factors (n = 23), piRNA/Tudor germline factors (n = 12), cell-cycle regulators with meiosis-specific roles (n = 5) and other germ-cell genes (n = 8). The seventh category, the classic *MAGE/CTAG* cancer–testis antigen family (n = 33), served as the comparator. The comparator

was used to test whether HPV-associated induction extended similarly to conventional cancer–testis antigens.

Because the discovery, validation, replication and single-cell datasets were generated on different platforms, a single core module was defined so that the same score could be computed in every cohort. A gene was retained only if it was measurable in the TCGA-HNSC expression HM450 methylation matrices, GSE65858 array after probe collapsing, GSE38266 HM450 matrix restricted to promoter-proximal probes and both single-cell count matrices. One core gene, *MAJIN*, is recorded under its previous HGNC symbol *C11orf85* in the deposited single-cell matrices, so both symbols were queried (Additional file 1: Table S1).

2.6. Promoter methylation and methylation–expression coupling

In TCGA-HNSC, methylation was summarised per gene with the cBioPortal gene-level HM450 β-value, which, for genes with multiple probes, is the probe most anti-correlated with expression. These expression-selected values were distinct from those used in the subsequent core-module analysis, in which promoter probes were defined using uniform criteria. Because the cBioPortal selection procedure inherently favours negative methylation–expression correlations, associations derived from these values were interpreted descriptively and assessed alongside the independent promoter-probe analyses. HPV+ versus HPV− differences were tested with the Mann–Whitney U test and Benjamini–Hochberg correction across all testable panel genes. Methylation–expression coupling was assessed per gene as the Spearman correlation between β and expression across samples with both measurements. At sample level, a composite methylation score (mean Z-scored β over the 18 hypomethylated genes) and a composite meiotic expression score (mean Z-scored expression over the core module) were correlated. Replication in GSE38266 was assessed in three ways. The first two used a gene-level summary in GSE38266 (mean β over promoter-proximal probes, 44 genes shared with the discovery cohort). The first was the Spearman correlation of per-gene HPV+ minus HPV− methylation differences between the two cohorts, and the second was a sign-concordance binomial test over the same genes, which is insensitive to platform differences in dynamic range. The third was a set-level comparison of mean methylation across the 227 promoter-proximal probes (GPL13534 manifest) mapping to 16 of the 18 genes hypomethylated in TCGA-HNSC.

To apply a single promoter definition in both methylation cohorts, the core-module analyses were repeated with probe-level β values, using GDC-harmonised (SeSAMe) HM450 β values for the 125 TCGA-HNSC tumours [37] and the deposited β values for GSE38266. Promoter CpGs were defined identically in both cohorts as those lying between 1,500 bp upstream and 500 bp downstream of an annotated transcription start site of the gene (GENCODE v36 HM450 manifest [38]). Because *MEI1* and *STAG3* are also expressed in B and plasma cells, HPV effects on core-module methylation were re-estimated in linear models adjusted for B-cell and plasma-cell content, measured by CIBERSORTx B-cell plus plasma-cell fractions where available and, for all tumours with expression data, by a marker-expression score

(mean Z-score of *CD19*, *CD79A*, *JCHAIN*, *MS4A1* and *MZB1*), which correlated with the CIBERSORTx B-cell plus plasma-cell fraction across the 279 transcriptomic tumours (Spearman ρ = 0.697).

2.7. Module scoring, specificity and coordination

In the bulk cohorts, a module score was defined as the mean Z-scored expression of its member genes. Discrimination between groups was quantified as the area under the receiver-operating-characteristic curve (AUC), with Mann–Whitney P values, for the module score, for each gene and for each panel category. Confidence intervals (95%) for AUCs were obtained with DeLong's method [39] and checked against a stratified bootstrap (2,000 resamples). For the category-level and single-cell comparator AUCs, for which per-sample scores were not re-derived, intervals were obtained with the Hanley–McNeil method from the AUC and group sizes [40]. In GSE65858, transcriptionally active HPV+ tumours and DNA-positive but transcriptionally silent tumours were each compared with HPV− tumours. Because stochastic de-repression predicts similar behaviour across germ-cell gene classes, specificity was assessed separately within each of the seven panel categories, and the meiotic categories were compared with the *MAGE/CTAG* comparator. Coordination was quantified as the mean off-diagonal Spearman correlation among module genes within each group. To account for the smaller HPV+ group, an empirical null distribution was generated from 1,000 random HPV− subsets matched to the HPV+ group size, giving an empirical 95% interval and permutation P value for the observed coherence of each HPV+ group (35 transcriptionally active and 25 transcriptionally silent tumours).

2.8. Immune-cell deconvolution

Immune-cell composition of the 279 TCGA tumours was estimated with CIBERSORTx [41] and the LM22 signature matrix of 22 immune-cell populations. Raw counts were converted to transcripts per million (TPM), Ensembl identifiers were mapped to HGNC symbols with org.Hs.eg.db (for duplicated symbols, the row with the highest mean expression was retained), and deconvolution was run in relative mode with 100 permutations, without quantile normalisation or batch correction. Fractions were compared between HPV groups with the Wilcoxon rank-sum test and Benjamini–Hochberg correction across all 22 populations (FDR

< 0.05). Estimates were validated by Spearman correlation of CD8 + T-cell fractions with variance-stabilised *CD8A* and *CD8B* expression and of plasma-cell fractions with *IGHM* and *MS4A1* expression in the TPM matrix supplied to CIBERSORTx. As an exploratory summary, a $\log_2$ CD8/M2 ratio, $\log_2[(\text{CD8 T cells} + 10^{-6})/(\text{M2 macrophages} + 10^{-6})]$, was compared between groups with the Wilcoxon test.

2.9. Confounder controls

Three potential confounders of the meiotic signal were addressed. In GSE65858 dataset (virus-active HPV + vs HPV− tumours), anatomical site was controlled by repeating the analysis in oropharyngeal tumours only, and proliferation, summarised as the mean Z-score of eight proliferation-associated genes (*MKI67*, *CCNB1*, *TOP2A*, *BUB1*, *CDK1, AURKA, PLK1* and *TYMS*), was included as a covariate in logistic regression models of HPV status on the core meiotic score, given the functional overlap between

meiotic and mitotic cell-cycle processes. In TCGA-HNSC, tumour purity was addressed by merging ABSOLUTE purity estimates [42], testing their association with HPV status and with the core meiotic score, and adding purity, and then purity and proliferation, as covariates to a linear model of the core meiotic score on HPV status. p16 expression was not used as a covariate, because p16 status is one of the two assays that define HPV status in the TCGA.

2.10. Single-cell analysis

In GSE181919, each deposited cell-type label was re-tested for enrichment of canonical lineage markers relative to all other labelled cell types, and per-cell-type sequencing depth was tabulated to detect compartment-specific dropout. Module activity was scored per cell with the control-gene-binned method of Tirosh et al. [43], which corrects the mean expression of a gene set against 50 expression-matched control genes drawn from 25 bins. Two meiotic scores were used. The first was the six-gene core score (*MEI1*, *SMC1B*, *STAG3*, *SYCP2*, *SYCE2*, *MAJIN*), and the second was a four-gene meiosis-specific score (*SMC1B*, *SYCP2*, *SYCE2*, *MAJIN*) that excludes *MEI1* and *STAG3*, the two core genes detected in lymphocytes. Scores were also computed for the wider meiotic panel, the *MAGE/CTAG* comparator and a six-gene proliferation score. Group differences were assessed with two-sided Mann–Whitney tests, with rank-biserial AUC as the effect size. Because cells from the same tumour are not independent, a linear mixed model was fitted with patient as a random effect, log-transformed per-cell gene count as a covariate and an HPV × malignant interaction term; in addition, malignant cells were aggregated into one pseudobulk profile per tumour. Per-gene detection rates were compared with Fisher's exact test and Benjamini–Hochberg correction, and the compartment specificity of each core gene was quantified as its detection rate in malignant cells divided by that in the non-malignant compartment with the highest detection rate.

Cell-type and malignant-cell assignments were not available for GSE182227 and therefore were derived de novo. Clusters were annotated with the same canonical marker panels; each lineage being required to show enrichment for its own markers relative to all other cells. The epithelial compartment was divided into malignant and normal epithelium using two criteria that were independent of HPV status and of any meiotic gene. These were the fraction of cells in each epithelial cluster that originated from adjacent-normal tissue and a per-cell copy-number score inferred from windowed expression along the genome. A cluster was classified as normal only when both criteria agreed. Malignant cells were aggregated into one profile per tumour, and the HPV contrast was tested at tumour level with the Mann–Whitney test and AUC. Proliferation was pre-specified as the principal confounder and included as a covariate in linear models of the per-tumour meiotic score, fitted within each cohort and in the pooled data; robustness was assessed with heteroskedasticity-consistent standard errors and Huber robust regression.

2.11. Exploratory survival analysis

Overall survival (months) and vital status for the 279-tumour transcriptomic cohort were derived from the cBioPortal TCGA-HNSC clinical annotation. The nine most significant HPV-upregulated genes (*LOC375196*, *STAG3*, *TAF7L*, *SMC1B*, *ZFR2*, *RAD9B*, *TDRD10*, *GBX1*

and *GRIN2C*) were analysed. For each gene, a univariable Cox proportional-hazards model (survival v3.8.6) was fitted to variance-stabilised expression scaled to unit variance, so that hazard ratios (HRs) are expressed per standard deviation. Kaplan–Meier curves were drawn for *ZFR2*, with tumours split at the median.

2.12. Statistical analysis and software

Unless stated otherwise, tests were two-sided and multiple testing was controlled with the Benjamini–Hochberg procedure. Bulk transcriptomic, enrichment and survival analyses were performed in R 4.6.0 with Bioconductor 3.23 (DESeq2, clusterProfiler, enrichplot, org.Hs.eg.db, TCGAbiolinks and survival). Code, processed data and output tables for all analyses are available in the project repository [44]; the panel, methylation, external-validation and single-cell analyses were performed in Python 3.11 (NumPy, pandas, SciPy and statsmodels).

3. Results

3.1. Cohorts and clinical characteristics

The transcriptomic cohort comprised 279 TCGA-HNSC primary tumours, 36 HPV + and 243 HPV−. HPV+ tumours arose predominantly in the oropharynx (61.1% vs 4.5%; P < 0.001), HPV+ patients were more often male (88.9% vs 70.4%; P = 0.033), and fewer HPV+ patients had died during follow-up (27.8% vs 53.5%; P = 0.007); age, stage and smoking status did not differ significantly (Table 1). Principal component analysis showed partial separation of HPV + and HPV− tumours on the first two components (22% and 14% of variance; Additional file 1: Fig. S1), indicating that HPV status contributes to, but does not dominate, global transcriptomic variation.

Five cohorts were included in the panel-based analyses (Table 2). In TCGA-HNSC, the p16/ISH rule (Section 2.2) identified 44 HPV + and 81 HPV− tumours with usable promoter methylation data and 41 HPV + and 80 HPV− tumours with usable expression data; these calls agreed with those of the originating project for all 53 overlapping samples. GSE65858 comprised 35 HPV+ tumours with active viral transcription, 25 HPV DNA-positive tumours without detectable viral transcription and 196 HPV− tumours. GSE38266 contributed 21 HPV + and 21 HPV− oropharyngeal tumours; 44 panel genes had gene-level methylation in both GSE38266 and the discovery cohort, and 16 of the 18 genes hypomethylated in TCGA-HNSC had promoter-proximal probes (227 probes). GSE181919 contributed 23,088 cells from 20 patients, and GSE182227 contributed 16 oropharyngeal tumours (11 HPV + and 5 HPV−).

3.2. Meiotic genes are among the most strongly upregulated genes in HPV+ tumours

Of 27,662 tested genes, 4,744 were differentially expressed (FDR < 0.05, |log2FC| > 1), of which 2,139 were upregulated and 2,605 downregulated in HPV+ tumours (Fig. 2a). Meiotic genes were among the

most significantly upregulated. Nine of the twelve most significant upregulated genes belong to the panel, including *STAG3* (log2FC = 4.63, FDR = 1.0 × 10 − 100), *SMC1B* (5.91, 6.7 × 10 − 91), *ZFR2* (6.52, 4.5 × 10 − 86), *RAD9B* (2.79, 1.2 × 10 − 83), *TDRD10*

1. (4.48, 2.6 × 10 − 69) and *CNTD1* (3.18, 4.8 × 10 − 63). The synaptonemal-complex genes *SYCP2*
2. (4.64, 4.4 × 10 − 62), *MAJIN* (5.07, 4.7 × 10 − 55) and *SYCE2* (2.92, 9.1 × 10 − 52) and the

recombination initiator *MEI1* (3.32, 1.5 × 10 − 39) were upregulated to a similar extent (Fig. 2b; Additional file 1: Table S2), consistent with earlier reports of *SYCP2*, *STAG3* and *TAF7L* upregulation in HPV+ head and neck cancer [27–29]. Other leading upregulated genes included the germ-cell transcription factor gene *TAF7L* (5.81, 9.9 × 10 − 100) and the homeobox gene *GBX1* (6.95, 3.2 × 10 − 68). All reported log2 fold changes (log2FC) were statistically tested and met the specified false discovery rate (FDR) threshold of 0.05, unless otherwise indicated. Exact FDR-adjusted P values are provided in the Supplementary Materials.

Downregulated genes were dominated by epithelial differentiation and keratinisation, including *DSC1*, *MUCL1*, *SPRR4*, *KRTDAP*, *KRT1* and *S100A7* (all FDR < 0.05; full statistics in Additional file 1: Table S2), together with multiple small proline-rich, late cornified envelope and keratin genes (Additional file 1: Tables S2 and S3). This pattern is consistent with the tonsillar-crypt origin and non-keratinising histology of most HPV+ tumours. Because HPV+ tumours arise predominantly in the oropharynx, the analysis was repeated with primary site as a covariate. The site-adjusted model identified 4,800 differentially expressed genes at the same thresholds, slightly more than the unadjusted model, consistent with removal of between-site variance. Fold changes were highly concordant between the two models (Spearman ρ = 0.94 across all genes), 3,975 of the 4,744 unadjusted differentially expressed genes remained significant after adjustment, and the meiotic, immune (*IGKC*, *IGHA1*, *IGHM*, *MS4A1*, *CD79A*, *CD79B*, *CD8A*, *CD8B*) and keratinisation (*KRT1*, *KRT14*, *KRT16*, *KRT6A*,

KRTDAP) genes described here all remained significant (Additional file 1: Table S3).

Anatomical site did not account for the transcriptional differences.

differential expression between HPV + and HPV− tumours: 4,744 differentially expressed genes (2,139 upregulated, 2,605 downregulated). (b) Log_2 fold changes of the ten selected meiotic genes (range 2.8–6.5). (c) The eight most significantly enriched GO biological-process terms among upregulated genes, with meiotic terms in blue. n = 279 tumours (36 HPV+, 243 HPV−).

3.3. Pathway analysis identifies meiotic, adaptive-immune and epithelial-differentiation programmes

GO over-representation analysis of upregulated genes identified 69 enriched biological-process terms (FDR < 0.05), 13 of them related to meiosis (terms containing “meio” or “synaptonemal”). Two themes dominated the enrichment results: adaptive immune responses and meiosis. The first was adaptive immunity, including antigen receptor-mediated signalling

pathway (43 genes; FDR = 5.3×10^{-10}), lymphocyte-mediated immunity (56 genes; FDR = 1.8

$\times 10^{-9}$), B cell-mediated immunity (37 genes; FDR = 5.4×10^{-9}) and immunoglobulin-mediated immune response (36 genes; FDR = 8.5×10^{-9}). The second was meiosis, led by meiotic nuclear division (3.8-fold enrichment) and meiotic cell cycle process, which ranked third and fourth overall (Fig. 2c). KEGG analysis of upregulated genes was likewise led by immune pathways, including primary immunodeficiency (16 genes; FDR = 1.9×10^{-9}) and haematopoietic cell lineage (18 genes; FDR = 1.9×10^{-4}). GSEA of the full gene ranking placed meiosis higher still. Among the 245 positively enriched terms (FDR < 0.05), meiosis I cell cycle process (normalised enrichment score [NES] = 2.91), meiosis I (NES = 2.87) and meiotic nuclear division (NES = 2.82) ranked among the five most strongly enriched terms, with only DNA-templated DNA replication showing a higher enrichment score (NES = 2.92). Other highly ranked terms involved DNA replication and recombination processes that share genes with the meiotic programme (Additional file 1: Fig. S2a).

Downregulated genes were associated with squamous differentiation. GO analysis was dominated by skin development (95 genes; FDR = 2.4×10^{-46}), epidermis development (101 genes; FDR = 2.2×10^{-43}), keratinisation (54 genes; FDR = 8.9×10^{-42}) and keratinocyte differentiation (69 genes; FDR = 8.7×10^{-39}), and KEGG analysis by cornified envelope formation (87 genes; FDR = 1.4×10^{-42}), cadherin signalling (70 genes; FDR = 8.2×10^{-11}), ECM–receptor interaction (21 genes; FDR = 3.7×10^{-5}) and focal adhesion (31 genes; FDR = 1.2×10^{-3}) (Additional file 1: Fig. S3). Keratinisation was also the most negatively enriched GSEA term (NES = − 3.17).

Because the gene panel was constructed after this enrichment result was obtained, its overlap with the differentially expressed genes indicates the strength of the signal but is not independent evidence for it. Of the 86 panel genes testable after symbol re-annotation, 37 were upregulated at FDR < 0.05, a 2.0-fold enrichment over chance (P = 9.6×10^{-6}), rising to 19.7-fold among the 406 most strongly induced genes (22 panel genes observed vs 1.1 expected; P

= 1.4×10^{-22}; Additional file 1: Fig. S2c). The overlap itself does not denote an independent validation. Nevertheless, 49 of 86 testable panel genes were not upregulated, and the genome-wide class comparison favoured meiotic genes over classic CTAs. The meiotic categories were enriched among upregulated genes (34 of 59 genes), whereas the *MAGE/CTAG* comparator was not (3 of 27; odds ratio 10.9, Fisher P = 4.6×10^{-5}; Additional file 1: Fig. S2b; Additional

file 1: Table S4). The category comparison is informative but remains exploratory because the panel was curated after the initial pathway analysis.

3.4. A six-gene core module is measurable in every cohort

Applying the cross-platform intersection criterion to the meiotic-machinery categories identified six genes measurable in all five cohorts (*MEI1*, *SMC1B*, *STAG3*, *SYCP2*, *SYCE2* and *MAJIN*; Additional file 1:

Table S1). These genes represent three functional components of the meiotic apparatus. *SYCP2*, *SYCE2* and *MAJIN* belong to the synaptonemal complex and its attachment machinery, *SMC1B* and *STAG3* encode meiosis-specific cohesin subunits, and *MEI1* is required for recombination initiation. Re-mapping incomplete gene-symbol annotations increased the number of testable panel genes from 82 to 86, and the four recovered genes entered the candidate pool under the same criteria. *MAJIN* was recovered in both single-cell matrices only when its previous symbol, *C11orf85*, was also queried.

3.5. Meiotic promoters are hypomethylated in HPV+ tumours, and methylation is coupled to expression

Of 49 panel genes with testable gene-level methylation (cBioPortal) in TCGA-HNSC, 18 (37%) were significantly hypomethylated in HPV+ tumours (FDR < 0.05; Fig. 3a). The largest effect, at *MEI1*, was a 0.45 decrease in median β. Seventeen of the 18 hypomethylated genes showed a negative methylation–expression correlation (median ρ = −0.55; Fig. 3b), consistent with an inverse association, although the expression-selected probe summary can strengthen such correlations. At the module level, the composite methylation score of the 18 hypomethylated genes and the core meiotic expression score were inversely correlated across samples ($\rho = -0.755$, $P = 1.5 \times 10^{-23}$). A correlation between hypomethylation of a CpG-island-shore region at *SYCP2* and *SYCP2* expression has previously been reported in TCGA HNSCC

[8]; the present results extend this relationship to the whole module.

The methylation findings were replicated in GSE38266, an entirely oropharyngeal cohort with 21 HPV + and 21 HPV− tumours; because all tumours were oropharyngeal, anatomical site could not confound the comparison. Gene-level methylation differences agreed between the two cohorts ($\rho = 0.688$, $P = 2.5 \times 10^{-7}$ over 44 shared genes; Fig. 3c), with 29 of 44 genes sharing the same sign (binomial P = 0.05). At set level, mean methylation across the 227 promoter-proximal probes in 16 hypomethylated genes was lower in HPV+ than in HPV− tumours ($P = 6.8 \times 10^{-4}$). In this cohort of 42 samples, only one locus remained significant after per-gene Bonferroni correction; the replication is therefore supported by the concordance of effect sizes and by the set-level test, not by individual loci.

The core module showed the same pattern when a single promoter definition was applied to both cohorts (Section 2.6; Additional file 1: Table S5). In TCGA-HNSC (44 HPV + vs 81 HPV− tumours), promoter methylation was lower in HPV+ tumours at five of the six core genes (FDR < 0.05; Δβ from −0.049 at *STAG3* to − 0.207 at *SYCP2*), and the same five genes were hypomethylated in GSE38266 (Δβ from − 0.048 to − 0.135; FDR < 0.05 for *MEI1*, *STAG3*, *SYCP2* and *MAJIN*), whereas *SYCE2* differed in neither cohort. At set level, mean β across the 80 core-module promoter probes in GSE38266 was 0.257 in HPV + and 0.334 in HPV− tumours ($P = 1.1 \times 10^{-6}$). Using these promoter probes, core-module promoter methylation (β)

was inversely correlated with core-module expression across TCGA tumours ($\rho = -0.641$, n = 121, $P = 2.2 \times 10^{-15}$; four-gene module ρ = −0.632). Absolute differences in these GDC-harmonised data are smaller

than in the gene-level summary used for the panel-wide analysis (for example, Δβ = −0.126 at *MEI1*), reflecting differences in β-value processing and in probe selection (cBioPortal reports, per gene, the probe most anti-correlated with expression). Adjustment for B-cell and plasma-cell content left the HPV effect on the core-module promoter composite essentially unchanged (Δβ = −0.100 unadjusted and − 0.098 adjusted for the marker score, n = 121; −0.118 and − 0.129 adjusted for CIBERSORTx fractions, n = 53), as it did for the four-gene composite (− 0.102 and − 0.103). At *MEI1*, the marker score was independently associated with promoter methylation (P = 0.008) and adjustment attenuated the HPV effect by 16% (Δβ = −0.137 to − 0.115), but the effect remained significant ($P = 2.9 \times 10^{-7}$).

1. **The programme tracks viral transcriptional activity rather than viral DNA** GSE65858 records HPV DNA status and HPV transcriptional activity separately. The six-gene core meiotic score separated tumours with active viral transcription from HPV− tumours (AUC

= 0.959, 95% CI 0.90–1.00; $P = 5.2 \times 10^{-18}$), whereas tumours carrying HPV DNA without transcriptional activity did not differ detectably from HPV− tumours (AUC = 0.464, 95% CI 0.30–0.62; P = 0.56; Fig. 4a). With 25 silent tumours, this absence of a difference does not establish equivalence, and low-level viral transcription below the detection threshold in some of these tumours cannot be excluded. When tumours were grouped by viral DNA status alone, as in the conventional definition, discrimination fell to AUC = 0.753 (95% CI 0.66–0.85) because the silent tumours diluted the HPV+ group.

3.7. Specificity is confined to the meiotic machinery

Discrimination differed markedly between germ-cell gene classes. The synaptonemal complex (AUC = 0.930), meiotic cohesin (0.907) and recombination/DSB machinery (0.829) all discriminated strongly, whereas the classic *MAGE/CTAG* antigens, the largest category in the

panel (33 genes, 29 measurable on this platform), performed at chance (AUC = 0.411, 95% CI 0.31–0.51; P = 0.09; Fig. 4c; Table 3). The *MAGE/CTAG* point estimate was slightly below 0.5, indicating no trend towards higher expression in HPV+ tumours. The same ordering held for individual genes, and the four best discriminators were all meiotic structural or recombination components (*MAJIN*, 0.938; *SMC1B*, 0.937; *SYCP2*, 0.929; *STAG3*, 0.929; Fig. 4b). The class-specific expression pattern is not readily explained by uniform activation across germ-cell gene categories.

Table 3
Discrimination of virus-active HPV+ (n = 35) from HPV− (n = 196) tumours in GSE65858 by gene category. Category assignments were fixed before analysis. Genes, number of panel members measurable on the array; AUC, from the mean Z-scored expression of each category, with Hanley–McNeil 95% CI; P, Mann–Whitney U test.

Gene category	Genes	AUC (95% CI)	P
Synaptonemal complex	5	0.930 (0.87–0.99)	5.2×10^{-16}
Meiotic cohesin	4	0.907 (0.84–0.97)	1.7×10^{-14}
Meiotic recombination/DSB	16	0.829 (0.74–0.92)	5.7×10^{-10}
Other germ-cell	7	0.827 (0.74–0.91)	7.3×10^{-10}
piRNA/Tudor germline	11	0.740 (0.64–0.84)	6.1×10^{-6}
Cell-cycle meiosis	4	0.697 (0.59–0.80)	2.1×10^{-4}
Cancer–testis antigen (*MAGE/CTAG*)	29	0.411 (0.31–0.51)	0.09

(b) Per-gene AUC for virus-active HPV+ versus HPV− tumours; grey marks classic cancer–testis antigens. (c) Per-category AUC; the MAGE/CTAG family is at chance, whereas every meiotic category is elevated. Panels b and c are anchored at AUC = 0.5 (dashed line).

3.8. The core meiotic genes are co-expressed as a module

The mean pairwise correlation among the six core genes was higher in virus-active HPV+ tumours than in HPV− tumours (0.556 vs 0.149; Table 4; Additional file 1: Fig. S4). In 1,000 random HPV− subsets of the same size as the HPV+ group, the 95% interval was 0.027–0.273, well below the observed value (P = 0.001). Coherence was specific to the core module. In the same tumours, the wider meiotic panel reached only 0.083 and the classic CTAs 0.081. Coherence was almost as high in the transcriptionally silent HPV+ group (0.541), although its mean expression was not elevated. Against its own size-matched null (1,000 random subsets of 25 HPV− tumours; 95% interval − 0.003 to 0.299), this coherence was also higher than expected by chance (P = 0.001), suggesting that the core genes remain co-regulated in these tumours even when the module is not activated. The four-gene meiosis-specific score, which excludes the lymphocyte-expressed *MEI1* and *STAG3* (Section 3.11), behaved in the same way. It separated virus-active HPV+ from HPV− tumours (AUC = 0.957, 95% CI 0.90–1.00), and its coherence was 0.546 in virus-active and 0.521 in silent HPV+ tumours but 0.179 in HPV− tumours, both HPV+ values exceeding their size-matched nulls (P = 0.001 and 0.002).

Table 4
Module coherence (mean off-diagonal Spearman ρ among member genes) by group and gene set in GSE65858.

Group	Gene set	Samples	Genes	Mean ρ
HPV+ active	Core meiotic	35	6	0.556
HPV+ active	All meiotic	35	25	0.083
HPV+ active	Classic CTA	35	29	0.081
HPV+ silent	Core meiotic	25	6	0.541
HPV+ silent	All meiotic	25	25	0.092
HPV+ silent	Classic CTA	25	29	0.067
HPV−	Core meiotic	196	6	0.149
HPV−	All meiotic	196	25	0.027
HPV−	Classic CTA	196	29	0.104

1. **Site, proliferation and tumour purity do not account for the meiotic signal** Anatomical site, proliferation and tumour purity were then tested as potential confounders (Additional file 1: Table S5). In GSE65858, restricting the comparison of virus-active HPV + and HPV− tumours to the oropharynx (28 vs 47 tumours) left discrimination by the core meiotic score essentially unchanged (AUC = 0.960, 95% CI 0.90–1.00; P = 3.3×10^{-11}), so the effect is not attributable to the anatomical distribution of HPV+ disease. In logistic regression with virus-active HPV status as the outcome and the core meiotic score as the predictor (GSE65858; 35 virus-active HPV + and 196 HPV− tumours; n = 231), adding the proliferation score did not attenuate the meiotic-score coefficient (β = 4.45 alone vs 4.52 adjusted; model

AUC 0.959 vs 0.961). Tumour purity was examined in TCGA-HNSC using ABSOLUTE purity estimates [42], available for 277 of the 279 tumours. HPV+ tumours were purer than HPV− tumours (mean purity 0.63 vs 0.48; Mann–Whitney P = 8.0×10^{-6}), making purity a potential confounder; however, purity was not correlated with the core meiotic score (Spearman ρ = 0.027, P = 0.66). In a linear model with the core meiotic score as the outcome and HPV status as the predictor, adding purity left the HPV coefficient unchanged (β = 2.29 unadjusted and 2.29 adjusted; purity coefficient − 0.022, P = 0.89), and adding purity and proliferation together gave an HPV coefficient of 2.25 (95% CI 2.08–2.41; P = 1.2×10^{-77}). The higher purity of HPV+ tumours may seem inconsistent with their lymphocyte-rich composition (Section 3.10). However, ABSOLUTE purity reflects all non-malignant content, including stroma, and lower stromal content in oropharyngeal HPV+ tumours could outweigh their greater immune infiltration. Purity was therefore used as a covariate and not as a measure of immune infiltration. A p16-adjusted model was not fitted because p16 immunohistochemistry is one of the two assays defining the HPV grouping; instead, the TCGA analyses were repeated with stricter HPV definitions. Under the main p16-or-ISH rule,

the core meiotic expression score separated HPV+ from HPV− tumours with AUC = 0.933 (95% CI 0.87–1.00; 41 vs 80). Requiring ISH positivity (21 vs 65) gave complete separation (AUC = 1.000) and a larger core-module promoter hypomethylation effect (22 vs 67; Δβ = −0.116 vs

−0.092 under the main rule; AUC = 0.964, 95% CI 0.93–1.00), and restricting the p16-or-ISH rule to oropharyngeal tumours (34 vs 8) gave AUC = 0.974 (95% CI 0.92–1.00). The GSE65858 analysis (Section 3.6) is not affected by this issue, because HPV DNA and transcriptional activity were determined there independently of p16.

3.10. HPV+ tumours have a plasma-cell- and CD8 + T-cell-rich immune

microenvironment

Consistent with the adaptive-immune terms among upregulated genes, CIBERSORTx deconvolution of the 279 TCGA tumours identified five of the 22 LM22 populations whose estimated fractions differed by HPV status after correction for multiple testing (Fig. 5; Table 5; all 22 populations in Additional file 1: Table S6). Plasma cells showed the largest increase in HPV+ tumours (median fraction 0.488 vs 0.145; FDR = 8.1×10^{-4}), and CD8 + T cells were also higher (0.071 vs 0.043; FDR = 1.7×10^{-3}). HPV+ tumours had lower fractions of M0 macrophages (0.033 vs 0.131; FDR = 5.6×10^{-5}), resting NK cells (0.005 vs 0.029; FDR = 1.7

$\times 10^{-3}$) and resting CD4 + memory T cells (0.041 vs 0.068; FDR = 0.042). The exploratory log2 CD8/M2 ratio was correspondingly higher in HPV+ tumours (median 3.29, IQR 1.82 to 15.35, vs 0.78, IQR − 0.63 to 3.16; P = 1.0×10^{-4}; Fig. 5f; Additional file 1: Table S7).

Marker-gene expression supported the deconvolution estimates. CD8 + T-cell fractions correlated with *CD8A* (Spearman ρ = 0.735) and *CD8B* (ρ = 0.689) expression (Additional file 1: Fig. S5), and plasma-cell fractions correlated with *IGHM* (ρ = 0.913) and *MS4A1* (ρ = 0.590) expression (Additional file 1: Table S8). At the transcript level, immunoglobulin and B-cell genes were upregulated in HPV+ tumours, including *IGHM* ($\log_2$FC = 2.44, FDR = 2.0×10^{-8}), *MS4A1* (2.07, 1.8×10^{-5}), *IGHA1* (1.51, 6.4×10^{-4}), *CD79A* (1.36, 6.7×10^{-4}) and *IGKC* (1.14,

8.3×10^{-3}), as were *CD8A* (1.26, 5.3×10^{-5}) and *CD8B* (1.18, 2.5×10^{-4}); all remained

significant after adjustment for primary site (Additional file 1: Table S3).

HPV+ tumours therefore also differ from HPV− tumours in cellular composition, with a larger lymphoid and plasma-cell component. Two core meiotic genes, *MEI1* and *STAG3*, have documented functions in lymphocytes, so part of the bulk meiotic signal could come from these cells. This possibility was tested with single-cell data (Sections 3.11 and 3.12).

Table 5
Immune-cell populations whose CIBERSORTx-estimated fractions differed between HPV+ (n = 36) and HPV− (n = 243) TCGA-HNSC tumours. Wilcoxon rank-sum test; Benjamini–Hochberg FDR across all 22 LM22 populations.

Immune-cell population	Median HPV+	Median HPV−	P	FDR	Direction in HPV+
Plasma cells	0.488	0.145	7.4×10^{-5}	8.1×10^{-4}	Higher
CD8 + T cells	0.071	0.043	2.8×10^{-4}	1.7×10^{-3}	Higher
M0 macrophages	0.033	0.131	2.5×10^{-6}	5.6×10^{-5}	Lower
Resting NK cells	0.005	0.029	3.1×10^{-4}	1.7×10^{-3}	Lower
Resting CD4 + memory T cells	0.041	0.068	9.4×10^{-3}	0.042	Lower

M0 macrophages, (d) resting NK cells and (e) resting CD4 + memory T cells, the five populations that differed after Benjamini–Hochberg correction across 22 cell types (FDR shown above each panel). (f) Exploratory log $_2$ CD8/M2 macrophage ratio. Boxes show the median and interquartile range; points are individual tumours (36 HPV+, 243 HPV−).

3.11. The HPV-associated meiotic contrast is concentrated in malignant cells

Because the cohorts analysed so far consisted of bulk tissue and HPV+ tumours contain more lymphocytes and plasma cells (Section 3.10), the meiotic signal could reflect a larger population of non-malignant cells expressing these genes rather than reactivation within tumour cells. This was tested in GSE181919, a single-cell HNSCC atlas annotated for HPV status and malignant versus stromal identity. After restriction to primary tumours, 2,463 malignant cells from 7 HPV+ patients, 2,366 malignant cells from 13 HPV− patients and 18,259 non-malignant cells from the same tumours were available. The deposited cell-type labels were checked against canonical markers, which confirmed all nine reported cell types. Malignant cells had the greatest sequencing depth (median 21,880 UMIs and 3,885 detected genes per cell), making differential dropout an unlikely explanation for a compartment-specific signal.

The HPV contrast was confined to the malignant compartment. With the six-gene core score computed per cell, malignant cells differed by 0.100 between HPV groups (AUC = 0.868), whereas every non-malignant cell type, including the T, B and plasma-cell compartments, differed by at most 0.011 in either direction (Fig. 6a, b). In a mixed model of the six-gene core score over all cells, with patient as a random

effect and per-cell detection depth as a covariate, the HPV × malignant interaction was 0.081 ($P = 5.5 \times 10^{-128}$); within malignant cells alone, the HPV+ effect was 0.108 (95% CI 0.071–0.144). Omitting *MAJIN*, which is deposited under its previous symbol *C11orf85*, did not alter the result. With the five core genes matched by current symbol alone, the per-cell contrast was 0.115 (AUC 0.866).

The effect was not driven by a single tumour. With malignant cells aggregated per patient, all 7 HPV+ tumours scored above all 13 HPV− tumours on the six-gene core score (two-sided exact Mann–Whitney $P = 2.6 \times 10^{-5}$; Fig. 6e), and separation persisted after removal of any single HPV+ patient (largest leave-one-out $P = 7.4 \times 10^{-5}$). Pseudobulk analysis, which does not depend on per-cell scoring, gave the same result. Among 6 HPV + and 13 HPV− tumours with at least 20 malignant cells, core-module pseudobulk expression (six genes) was 6.6-fold higher in HPV+ malignant cells ($P = 7.4 \times 10^{-5}$), whereas the cancer–testis comparator was unchanged (0.96-fold, P = 0.52) (Fig. 6c). For individual genes, *SYCP2* was detected in 35.8% of HPV+ versus 1.5% of HPV− malignant cells and *SMC1B* in 29.5% versus 0.6% (Fig. 6d); across the panel, 15 of 30 meiotic genes but only 4 of 13 cancer–testis antigens were detected significantly more often in HPV+ malignant cells (FDR < 0.05), mirroring the class difference seen in bulk tissue.

Not all core genes were specific to malignant cells, and the exceptions followed a clear pattern. *SYCP2* and *SMC1B* were detected 7.7- and 11.0-fold more often in malignant cells than in the most-detecting non-malignant compartment, whereas *STAG3* showed little enrichment (1.18-fold) and *MEI1* was detected more often in B/plasma cells than in malignant cells (0.22-fold; Additional file 1: Table S9). Both genes have documented functions in lymphocytes in addition

to their meiotic roles, which is consistent with the B-cell and plasma-cell enrichment of HPV+ tumours. Excluding them gave the four-gene meiosis-specific score (*SMC1B*, *SYCP2*, *SYCE2*, *MAJIN*), which is free of this lymphoid contribution; all conclusions were the same for both scores (Table 6).

1. **The malignant-cell localisation replicates in an independent single-cell cohort** GSE181919 contained only seven HPV+ tumours, so the analysis was repeated in GSE182227, an oropharyngeal cohort of 16 tumours (11 HPV+, 5 HPV−) deposited without cell-type or malignant-cell annotations. De novo annotation with the same canonical marker panels resolved nine lineages (Fig. 7a), and the epithelial compartment was divided into malignant and normal epithelium using adjacent-normal tissue fraction and inferred copy number, neither of which involves a meiotic gene. Malignant cells clustered by patient, whereas normal

epithelium from different patients clustered together (Fig. 7b). One of 16 epithelial clusters contained 75% of all adjacent-normal epithelial cells in the cohort and had an inferred copy-number score of 0.094, compared with a median of 0.126 across malignant clusters; both criteria classified it as normal. Without this classification, cells in this cluster that were sampled from tumour tissue would have been counted as malignant. With this assignment, 3.9% of normal epithelial cells (n = 2,127) detected any of the four meiosis-specific genes, compared with 37.2% of 23,709 malignant cells (Fig. 7c).

Within the malignant compartment defined in this way, the HPV contrast was reproduced. Per tumour, the four-gene meiosis-specific score was 0.176 higher in HPV+ than in HPV− tumours (AUC = 0.945, 95% CI 0.84–1.00; two-sided P = 0.003; Fig. 7d), compared with 0.112 (AUC

= 1.000, P = 7.4×10^{-5}) for the same score in GSE181919; the effect therefore had the same direction and a similar size in two cohorts annotated by different procedures (Fig. 7e; Table 6). The six-gene core score gave equivalent results in both cohorts (Table 6). The cancer–testis comparator discriminated in neither cohort (AUC = 0.282, 95% CI 0.04–0.52, P = 0.15; AUC

= 0.564, 95% CI 0.26–0.87, P = 0.74), nor did the proliferation score alone (AUC = 0.731 and 0.782). Tumour-level separation in GSE182227 was near-complete, with two HPV+ tumours falling within the HPV− range (Fig. 7d). OP13, which contributed only 155 malignant cells, scored below two HPV− tumours (OP8 and OP12), and OP35 (270 malignant cells) scored below OP8, giving 3 discordant pairs of 55 (AUC = 0.945). When the analysis was repeated with independently written code, in which epithelial cells were annotated from marker genes without a copy-number split, OP13 versus OP8 was the only discordant pair (1 of 55; AUC = 0.982, P = 9.2×10^{-4}), and per-tumour scores from the two implementations were highly correlated (Pearson r = 0.94). The proportion of malignant cells detecting any of the four meiosis-specific genes separated the groups completely (AUC = 1.000, P = 4.6×10^{-4}), suggesting that the prevalence of programme-positive cells is a more stable readout than the mean score when few cells are available.

Proliferation is an important potential confounder because meiotic and mitotic programmes share components and HPV+ tumours are more proliferative. In the pooled analysis of 35 tumours from both cohorts (the 19 GSE181919 tumours with at least 20 malignant cells and all 16 GSE182227 tumours), adding the proliferation score reduced the HPV coefficient for the four-gene meiosis-specific score from 0.141 (P = 4.5×10^{-6}) to 0.109 (P = 2.7×10^{-5}), a 22% attenuation (Fig. 7f; Additional file 1: Table S10). The attenuation was larger in GSE182227 (34%) than in GSE181919 (6%), and in the pooled model proliferation had an independent coefficient of 0.359 (P = 2.8×10^{-4}), indicating that proliferation and HPV status were independently associated with the meiotic score. The pooled adjusted estimate was robust to the regression method. Heteroskedasticity-consistent standard errors gave P = 1.3×10^{-4}, and Huber robust regression gave a coefficient of 0.097 (P = 9.1×10^{-8}).

Table 6
HPV contrast in per-tumour module scores in both single-cell cohorts, malignant cells only. Δ, difference in mean score between HPV + and HPV− tumours; P, two-sided Mann–Whitney test on tumour-level means. These tumour-level contrasts are not directly comparable with the per-cell contrasts in Fig. 6a.

Cohort	Score	Δ	AUC	P	Complete separation
GSE181919	Six-gene core	0.113	1.000	7.4×10^{-5}	Yes
GSE181919	Four-gene meiosis-specific	0.112	1.000	7.4×10^{-5}	Yes
GSE181919	CTA control (5 of 10 genes detected)	−0.089	0.282	0.15	No
GSE181919	Proliferation control (6 genes)	0.038	0.731	0.13	No
GSE182227	Six-gene core	0.162	0.945	0.003	No
GSE182227	Four-gene meiosis-specific	0.176	0.945	0.003	No

(c) Percentage of cells detecting any of the four meiosis-specific genes, by compartment, after the epithelial split. (d) Per-tumour four-gene meiosis-specific score by HPV status; two HPV+ tumours fall within the HPV− range (see text). (e) The same four-gene contrast in both single-

cell cohorts. (f) HPV effect on the four-gene meiosis-specific score with and without adjustment for proliferation, by cohort and pooled; bars are 95% confidence intervals.

3.13. Exploratory univariable survival analysis

All 279 tumours in the transcriptomic cohort had overall-survival annotation (140 deaths; Table 1). In exploratory univariable Cox models, higher expression of seven of the nine leading HPV-upregulated genes was nominally associated with longer overall survival, with HRs per standard deviation between 0.70 (*ZFR2*; 95% CI 0.56–0.87) and 0.82 (*SMC1B* and *GRIN2C*); associations for *TDRD10* and *RAD9B* did not reach nominal significance (Additional file 1: Table S11; Additional file 1: Fig. S6). Kaplan–Meier curves for *ZFR2*, split at median expression, are shown in Additional file 1: Fig. S7. These models were not adjusted for HPV status, stage, smoking or treatment and were not corrected for multiple testing. Because the genes are strongly HPV-associated and HPV+ patients had markedly better outcomes in this cohort (Table 1), the associations are likely to reflect HPV status to a large extent and should not be interpreted as independent prognostic effects.

4. Discussion

This study aimed to characterise the molecular differences between HPV-positive and HPV-negative head and neck squamous cell carcinomas, with particular emphasis on germ-cell gene expression, epigenetic regulation and the cellular origin of the associated transcriptional changes. Through

integrated transcriptomic, DNA methylation and single-cell analyses, we identified a coordinated meiotic gene programme associated with transcriptionally active HPV-positive tumours. The programme was enriched for genes involved in meiotic chromosome organisation and recombination, showed coordinated expression across tumours and was accompanied by promoter hypomethylation. In contrast, the classical *MAGE/CTAG* cancer–testis antigens did not show comparable HPV-associated expression patterns. We also observed differences in the tumour immune microenvironment and epithelial differentiation, consistent with established features of HPV-associated HNSCC. Recent work by Kiruba et al. [58] identified shared HPV-associated transcriptional alterations across cervical and head and neck cancers, highlighting genes involved in cell-cycle regulation, DNA replication and epithelial biology. Our findings extend this understanding by identifying a distinct meiotic gene programme associated with promoter hypomethylation and active viral transcription in HPV-positive HNSCC. Together, these findings suggest that the expression of meiotic genes in HPV-positive tumours may reflect a more coordinated molecular process than the broader, potentially stochastic de-repression of cancer–testis antigens described previously [15, 18].

The association with viral transcriptional activity is a central observation. HPV DNA-positive tumours without viral transcription resembled HPV− tumours on the core meiotic score (P = 0.56), whereas virus-active tumours were clearly separated. This pattern is consistent with an HPV transcription-associated host programme rather than an effect of viral DNA detection alone, but does not establish that E6/E7 directly activates the meiotic genes. E7-mediated promoter methylation changes have been reported at other loci [9]. In practice, studies that rely on HPV DNA status alone may underestimate such associations. In the same samples, DNA-based classification reduced the AUC from 0.959 to 0.753.

Two explanations could account for the selective and coordinated activation of the meiotic machinery. The first is shared regulation. If the core meiotic genes are silenced in somatic cells by a common, methylation-dependent mechanism, relief of that repression would activate them together and would explain both the specificity and the coherence of the programme. A plausible but untested candidate is *E2F6*-dependent repression. In somatic cells, meiosis-specific genes carry a conserved *E2F6*-binding element in their promoters [45], *E2F6* silences *SMC1B* and *STAG3* [46], and *E2F6* recruits *DNMT3B* to establish methylation at germline gene promoters [47]; no comparable *E2F6* dependence has been described for the *MAGE/CTAG f*amily, whose silencing is attributed mainly to promoter DNA methylation [18]. Because E7 disrupts RB–E2F control [7], interference with *E2F6*-dependent repression by E7 could explain, through a single mechanism, the class specificity, the coordination and the dependence on viral transcription observed here. This hypothesis predicts reduced *E2F6* or *DNMT3B* occupancy at core meiotic promoters in E7-expressing cells, and de-repression of the core genes, but not of *MAGE/CTAG* genes, after *E2F6* depletion. Shared regulation would also explain why coherence was almost as high in transcriptionally silent HPV+ tumours (0.541) as in active ones (0.556), with both values exceeding a size-matched null, even though only the active tumours showed increased expression.

The second explanation is functional selection. Ectopic meiotic cohesin generates chromosomal instability [22], ectopically expressed *HORMAD1* has been linked to genomic instability and homologous-

recombination deficiency [23, 24] and to protection of stalled replication forks [25], and partial activation of the meiotic programme has been proposed to be oncogenic [26]. Under this model, expression of the meiotic module would be maintained because it benefits tumour cells, not merely because of passive demethylation. Distinguishing these explanations will require perturbation experiments.

HPV+ tumours showed higher estimated plasma-cell and CD8 + T-cell fractions and lower M0 macrophage, resting NK-cell and resting CD4 + memory T-cell fractions, supported by correlations of the estimates with lineage marker genes and by upregulation of immunoglobulin and B-cell transcripts that persisted after adjustment for primary site. B-cell- and plasma-cell-rich infiltrates and tertiary lymphoid structures have been associated with favourable outcome and immunotherapy response in other tumour types [48, 49], and a lymphocyte-rich microenvironment is a recognised feature of HPV+ HNSCC [5]. The lymphoid-rich oropharynx, where most HPV+ tumours arise, probably contributes to the plasma-cell signal. The estimates are also relative, computationally inferred fractions, not direct cell counts. The immune differences further raised the possibility that the bulk meiotic signal came partly from infiltrating lymphocytes, which the single-cell data allowed us to test. The HPV contrast was confined to malignant cells and was essentially absent from the lymphocyte populations that differ most in abundance between HPV + and HPV− tumours. Within HPV+ tumours, core meiotic genes were also detected more often in malignant than in non-malignant cells of the same specimen, arguing against a tumour-wide field effect. The two exceptions to malignant-cell specificity, *MEI1* and *STAG3*, were detected in B and plasma cells, the populations that are expanded in HPV+ tumours, and excluding them did not change the results. The same localisation was observed in a second cohort that was annotated de novo, with the malignant compartment defined by copy number and tissue of origin rather than by deposited labels, which makes an annotation artefact unlikely.

The reduced expression of keratinisation, cornified-envelope and epithelial-differentiation genes in HPV+ tumours is consistent with the non-keratinising histology and tonsillar-crypt origin of most HPV+ oropharyngeal carcinomas [5]. This difference persisted after adjustment for site, and its agreement with known HPV-associated histology supports the validity of the transcriptomic comparison. Whether loss of squamous differentiation is related to meiotic reactivation cannot be determined from these data.

These findings are relevant to the use of germ-cell antigens as therapeutic targets. Cancer–testis antigens are pursued as immunotherapeutic targets largely because of their tumour-restricted expression [19, 20]. In HPV+ HNSCC, the *MAGE/CTAG* family did not distinguish tumours by HPV status, whereas the meiotic structural and recombination components did, were more tightly co-expressed, and were expressed in malignant cells within a lymphocyte-rich microenvironment. This conclusion rests on expression specificity alone; immunogenicity was not assessed.

The single-cell data do not indicate uniform activation across malignant cells. Even in HPV+ tumours, individual meiotic transcripts were detected in a minority of malignant cells; *SYCP2* had the highest detection rate (35.8%). Limited detection partly reflects the sensitivity of 3' droplet-based sequencing, but activation restricted to a malignant subpopulation cannot be excluded. Module scores and patient-

level aggregates are therefore more robust measures of programme activation than single-cell detection, and higher-sensitivity approaches such as full-length or targeted single-cell sequencing will be needed to distinguish between these possibilities.

The panel was constructed after the genome-wide enrichment analysis had identified meiosis as one of the processes most strongly associated with HPV status, so the overlap statistics in Additional file 1: Fig. S2c quantify the magnitude of an effect in the same samples that motivated panel construction and are not independent evidence. The panel categories were defined from literature and CTdatabase sources, but the within-TCGA enrichment statistic should not be considered independent confirmation. The main findings were subsequently assessed in independent expression and methylation cohorts and in two single-cell datasets, which did not contribute to the original panel construction.

This study has several limitations. The TCGA transcriptomic cohort included only 36 HPV+ tumours, and the single-cell analyses were limited to 36 tumours across two cohorts. Droplet-based single-cell sequencing also has limited sensitivity for low-abundance transcripts. Malignant-cell annotations in GSE182227 lacked external validation, although the two annotation criteria were concordant. The methylation replication cohort was also small (42 samples), and evidence linking promoter demethylation to gene expression remains correlative, requiring experimental validation. Differences between RNA-seq and microarray platforms limit direct comparison of effect sizes across cohorts. HPV status was defined differently for transcriptomic (RNA-seq-based) and panel (p16/ISH-based) analyses, resulting in different sample sizes. The p16-or-ISH definition may misclassify non-oropharyngeal tumours, although associations strengthened when restricted to ISH-positive cases. The oropharynx-restricted analysis included only eight HPV− tumours. Immune-cell fractions were computationally inferred and represent relative rather than absolute abundances, while the

CD8/M2 ratio remains exploratory. Residual confounding by tumour site cannot be excluded despite site-adjusted sensitivity analyses. Finally, the retrospective, observational design and predominance of North American and European populations limit causal interpretation and broader generalisability.

5. Conclusions

This study identifies a coordinated meiotic gene programme associated with transcriptionally active HPV-positive head and neck squamous cell carcinoma. The programme is characterised by selective activation of meiotic genes, promoter hypomethylation and predominant expression in malignant epithelial cells. Unlike conventional cancer–testis antigens, the six-gene core module strongly distinguished transcriptionally active HPV-positive tumours from HPV-negative tumours, with findings supported across independent cohorts. These results suggest that meiotic gene reactivation represents a distinct molecular feature of HPV-associated carcinogenesis and is not mirrored by the classic *MAGE/CTAG* comparator. Functional studies are needed to test its regulation and biological role; prospective studies would be required to assess clinical biomarker utility.

Abbreviations

AUC
Area under the receiver operating characteristic curve
CI
Confidence interval
CTA
Cancer–testis antigen
DSB
Double-strand break
FC
Fold change
FDR
False discovery rate
GDC
Genomic Data Commons
GEO
Gene Expression Omnibus
GO
Gene Ontology
GSEA
Gene set enrichment analysis
HM450
Illumina Infinium HumanMethylation450 BeadChip
HNSCC
Head and neck squamous cell carcinoma
HPV
Human papillomavirus
IQR
Interquartile range
ISH
In situ hybridisation
KEGG
Kyoto Encyclopedia of Genes and Genomes
NES
Normalised enrichment score
NK
Natural killer
RNA-seq
RNA sequencing

RSEM
RNA-Seq by Expectation-Maximization
TCGA
The Cancer Genome Atlas
TPM
Transcripts per million

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified data. No new patient samples were collected and no additional ethical approval was required.

Consent for publication

Not applicable.

Availability of data and materials

All analysed data are publicly available. TCGA-HNSC RNA-seq counts and HM450 methylation data were obtained from the NCI Genomic Data Commons [50], while RSEM
expression, clinical annotations, HPV status and gene-level methylation were retrieved through cBioPortal [51, 52]. Independent cohorts were obtained from GEO Database with accessions GSE65858 [53], GSE38266 [54], GSE181919 [55] and GSE182227 [56]. The tumour-purity

estimates were extracted from the PanCancer Atlas ABSOLUTE calls [42,57]. All processed data, analysis scripts, results, supplementary tables and figure-generation code are available in the project repository at https://github.com/rejoy2004-rgb/TCGA-HNSC-HPV-Transcriptomics and archived at https://doi.org/10.5281/zenodo.23815232 [44].

Competing interests

The authors declare no competing interests.

Funding

No external funding was received for this work.

Use of AI-assisted tools

Large language model (LLM)-based tools were used to assist with manuscript drafting and editing. The authors are responsible for the study design, analyses, interpretation and final manuscript, and confirm that all AI-assisted text have been reviewed and checked before submission.

Authors' contributions

Suganth Gopinath Sangeetha (SGS): Methodology, Formal analysis, Investigation, Data curation, Visualization, Software, Writing – original draft.

Rejoy Besra (RB): Methodology, Formal analysis, Investigation, Data curation, Visualization, Software, Writing – original draft.

Naisarg Patel (NP): Conceptualization, Supervision, Validation, Writing – review & editing.

Prakash Lingasamy (PL): Investigation, Resources, Writing – review & editing. **Vino Sundararajan (VS):** Investigation, Resources, Writing – review & editing. **Sajitha Lulu S (SLS):** Investigation, Resources, Writing – review & editing.

Andres Salumets (AS): Conceptualization, Supervision, Project administration, Writing – review & editing.

Vijayachitra Modhukur (VM): Conceptualization, Supervision, Project administration, Writing – review & editing.

All authors read and approved the final manuscript.

Acknowledgements

Not Applicable

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49. https://doi.org/10.3322/caac.21660.
2. Chaturvedi AK, Engels EA, Pfeiffer RM, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol. 2011;29:4294–301. https://doi.org/10.1200/JCO.2011.36.4596.

3. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82. https://doi.org/10.1038/nature14129.
4. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363:24–35. https://doi.org/10.1056/NEJMoa0912217.
5. Leemans CR, Snijders PJF, Brakenhoff RH. The molecular landscape of head and neck cancer. Nat Rev Cancer. 2018;18:269–82. https://doi.org/10.1038/nrc.2018.11.
6. Lechner M, Liu J, Masterson L, Fenton TR. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. Nat Rev Clin Oncol. 2022;19:306–27. https://doi.org/10.1038/s41571-022-00603-7.
7. Moody CA, Laimins LA. Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer. 2010;10:550–60. https://doi.org/10.1038/nrc2886.
8. Degli Esposti D, Sklias A, Lima SC, et al. Unique DNA methylation signature in HPV-positive head and neck squamous cell carcinomas. Genome Med. 2017;9:33. https://doi.org/10.1186/s13073-017-0419-z.
9. Chalertpet K, Pakdeechaidan W, Patel V, Mutirangura A, Yanatatsaneejit P. Human papillomavirus type 16 E7 oncoprotein mediates CCNA1 promoter methylation. Cancer Sci. 2015;106:1333–40. https://doi.org/10.1111/cas.12761.
10. Lechner M, Fenton T, West J, et al. Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. Genome Med. 2013;5:15. https://doi.org/10.1186/gm419.
11. Anayannis NV, Schlecht NF, Belbin TJ. Epigenetic mechanisms of human papillomavirus-associated head and neck cancer. Arch Pathol Lab Med. 2015;139:1373–8. https://doi.org/10.5858/arpa.2014-0554-RA.
12. Xu Y, Wang Z, Wei P, et al. Hypermethylation of nc886 in HPV-positive oropharyngeal cancer and its clinical implications: an epigenome-wide association study. Mol Ther Nucleic Acids. 2022;30:596–605. https://doi.org/10.1016/j.omtn.2022.11.012.
13. Parfenov M, Pedamallu CS, Gehlenborg N, et al. Characterization of HPV and host genome interactions in primary head and neck cancers. Proc Natl Acad Sci U S A. 2014;111:15544–9. https://doi.org/10.1073/pnas.1416074111.
14. Wichmann G, Rosolowski M, Krohn K, et al. The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. Int J Cancer. 2015;137:2846–57. https://doi.org/10.1002/ijc.29649.
15. Zhang C, Kawakami T, Okada Y, Okamoto K. Distinctive epigenetic phenotype of cancer testis antigen genes among seminomatous and nonseminomatous testicular germ-cell tumors. Genes Chromosomes Cancer. 2005;43:104–12. https://doi.org/10.1002/gcc.20160.
16. Almeida LG, Sakabe NJ, deOliveira AR, et al. CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. Nucleic Acids Res. 2009;37:D816–9. https://doi.org/10.1093/nar/gkn673.

17. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301:89–92. https://doi.org/10.1038/301089a0.
18. Cannuyer J, Loriot A, Parvizi GK, De Smet C. Epigenetic hierarchy within the *MAGEA1* cancer-germline gene: promoter DNA methylation dictates local histone modifications. PLoS ONE. 2013;8:e58743. https://doi.org/10.1371/journal.pone.0058743.
19. Wei X, Chen F, Xin K, et al. Cancer-testis antigen peptide vaccine for cancer immunotherapy: progress and prospects. Transl Oncol. 2019;12:733–8. https://doi.org/10.1016/j.tranon.2019.02.008.
20. Chiappinelli KB, Zahnow CA, Ahuja N, Baylin SB. Combining epigenetic and immunotherapy to combat cancer. Cancer Res. 2016;76:1683–9. https://doi.org/10.1158/0008-5472.CAN-15-2125.
21. Ferris RL, Blumenschein G, Fayette J, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. N Engl J Med. 2016;375:1856–67. https://doi.org/10.1056/NEJMoa1602252.
22. Boukaba A, Liu J, Ward C, et al. Ectopic expression of meiotic cohesin generates chromosome instability in cancer cell line. Proc Natl Acad Sci U S A. 2022;119:e2204071119. https://doi.org/10.1073/pnas.2204071119.
23. Watkins J, Weekes D, Shah V, et al. Genomic complexity profiling reveals that *HORMAD1* overexpression contributes to homologous recombination deficiency in triple-negative breast cancers. Cancer Discov. 2015;5:488–505. https://doi.org/10.1158/2159-8290.CD-14-1092.
24. Gantchev J, Messina-Pacheco J, Martínez Villarreal A, et al. Ectopically expressed meiosis-specific cancer testis antigen HORMAD1 promotes genomic instability in squamous cell carcinomas. Cells. 2023;12:1627. https://doi.org/10.3390/cells12121627.
25. Herrera LR, Johnson RA, McGlynn K, et al. The cancer testes antigen, HORMAD1, limits genomic instability in cancer cells by protecting stalled replication forks. J Biol Chem. 2023;299:105348. https://doi.org/10.1016/j.jbc.2023.105348.
26. Gantchev J, Martínez Villarreal A, Gunn S, et al. The ectopic expression of meiCT genes promotes meiomitosis and may facilitate carcinogenesis. Cell Cycle. 2020;19:837–54. https://doi.org/10.1080/15384101.2020.1743902.
27. Pyeon D, Newton MA, Lambert PF, et al. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. Cancer Res. 2007;67:4605–19. https://doi.org/10.1158/0008-5472.CAN-06-3619.
28. Masterson L, Sorgeloos F, Winder D, et al. Deregulation of SYCP2 predicts early stage human papillomavirus-positive oropharyngeal carcinoma: a prospective whole transcriptome analysis. Cancer Sci. 2015;106:1568–75. https://doi.org/10.1111/cas.12809.
29. Hinić S, Rich A, Anayannis NV, Cabarcas-Petroski S, Schramm L, Meneses PI. Gene expression and DNA methylation in human papillomavirus positive and negative head and neck squamous cell carcinomas. Int J Mol Sci. 2022;23:10967. https://doi.org/10.3390/ijms231810967.
30. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.

https://doi.org/10.1158/2159-8290.CD-12-0095.

31. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:pl1. https://doi.org/10.1126/scisignal.2004088.
32. Choi JH, Lee BS, Jang JY, et al. Single-cell transcriptome profiling of the stepwise progression of head and neck cancer. Nat Commun. 2023;14:1055. https://doi.org/10.1038/s41467-023-36691-x.
33. Puram SV, Mints M, Pal A, et al. Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. Nat Genet. 2023;55:640–50. https://doi.org/10.1038/s41588-023-01357-3.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.
35. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7. https://doi.org/10.1089/omi.2011.0118.
36. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
37. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. 2018;46:e123. https://doi.org/10.1093/nar/gky691.
38. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45:e22. https://doi.org/10.1093/nar/gkw967.
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45. https://doi.org/10.2307/2531595.
40. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36. https://doi.org/10.1148/radiology.143.1.7063747.
41. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37:773–82. https://doi.org/10.1038/s41587-019-0114-2.
42. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21. https://doi.org/10.1038/nbt.2203.
43. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96. https://doi.org/10.1126/science.aad0501.
44. TCGA-HNSC-HPV-Transcriptomics. analysis code and processed data [code repository]. GitHub. https://github.com/rejoy2004-rgb/TCGA-HNSC-HPV-Transcriptomics. Accessed 30 Oct 2026.
45. Kehoe SM, Oka M, Hankowski KE, et al. A conserved E2F6-binding element in murine meiosis-specific gene promoters. Biol Reprod. 2008;79:921–30.

https://doi.org/10.1095/biolreprod.108.067645.

46. Storre J, Schäfer A, Reichert N, et al. Silencing of the meiotic genes SMC1β and *STAG3* in somatic cells by E2F6. J Biol Chem. 2005;280:41380–6. https://doi.org/10.1074/jbc.M506797200.
47. Velasco G, Hubé F, Rollin J, et al. Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proc Natl Acad Sci U S A. 2010;107:9281–6. https://doi.org/10.1073/pnas.1000473107.
48. Cabrita R, Lauss M, Sanna A, et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature. 2020;577:561–5. https://doi.org/10.1038/s41586-019-1914-8.
49. Petitprez F, de Reyniès A, Keung EZ, et al. B cells are associated with survival and immunotherapy response in sarcoma. Nature. 2020;577:556–60. https://doi.org/10.1038/s41586-019-1906-8.
50. National Cancer Institute Genomic Data Commons. TCGA-HNSC: The Cancer Genome Atlas Head and Neck Squamous Cell Carcinoma project. GDC Data Portal. https://portal.gdc.cancer.gov/projects/TCGA-HNSC. Accessed 30 Oct 2026.
51. The Cancer Genome Atlas Network. Head and Neck Squamous Cell Carcinoma. (TCGA, Nature 2015) [study hnsc_tcga_pub]. cBioPortal for Cancer Genomics. https://www.cbioportal.org/study/summary?id=hnsc_tcga_pub. Accessed 30 Oct 2026.
52. The Cancer Genome Atlas Research Network. Head and Neck Squamous Cell Carcinoma (TCGA, Firehose Legacy) [study hnsc_tcga]. cBioPortal for Cancer Genomics. https://www.cbioportal.org/study/summary?id=hnsc_tcga. Accessed 30 Oct 2026.
53. Wichmann G, Rosolowski M. Gene expression patterns and TP53 mutations are associated with HPV RNA status, lymph node metastasis, and survival in head and neck cancer. Gene Expression Omnibus. 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65858. Accessed 30 Oct 2026.
54. Lechner M, Fenton T, West J et al. 450k Analysis of 42 FFPE HPV + and HPV-HNSCC samples. Gene Expression Omnibus. 2012. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38266. Accessed 30 Oct 2026.
55. Woo HG. Single-cell transcriptome profiling of the stepwise progression of head and neck cancer. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181919. Accessed 30 Oct 2026.
56. Mints M, Tirosh I. Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. Gene Expression Omnibus. 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182227. Accessed 30 Oct 2026.
57. TCGA PanCancer Atlas. ABSOLUTE purity/ploidy calls (TCGA_mastercalls.abs_tables_JSedit.fixed.txt). NCI Genomic Data Commons, PanCanAtlas Publications. 2018. https://gdc.cancer.gov/about-data/publications/pancanatlas. Accessed 30 Oct 2026.
58. Kiruba B, Niranjan SB, Bright D, et al. HPV-Associated transcriptional landscape in cervical and head and neck cancers. Infect Agents Cancer. 2026. https://doi.org/10.1186/s13027-026-00811-1.

Figures

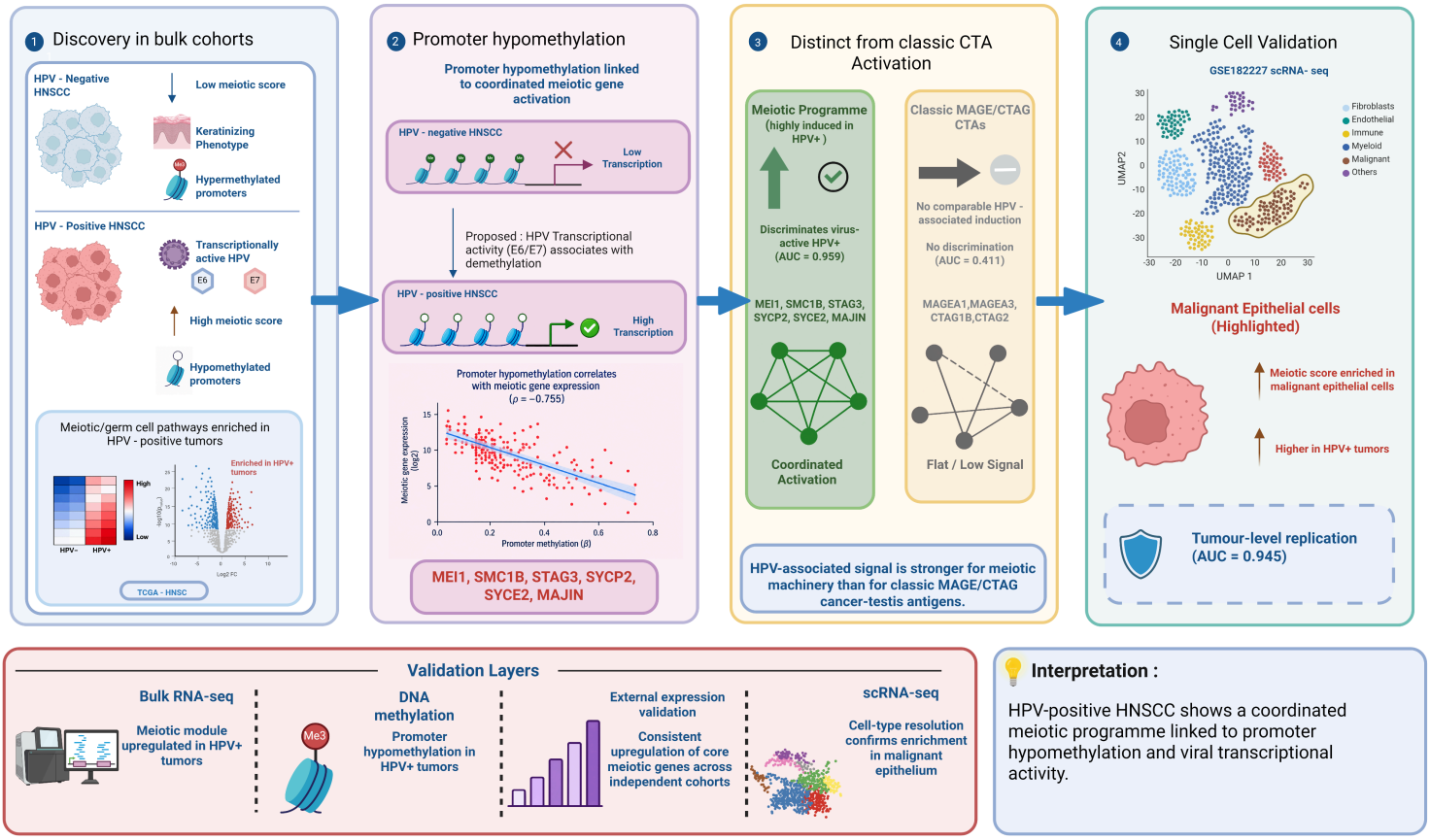


Figure 1

Overview of the study design and principal findings. *(1) Bulk transcriptomic analysis identifies meiotic pathway enrichment in HPV+ HNSCC. (2) Promoter hypomethylation is associated with coordinated meiotic gene expression. (3) Meiotic genes show stronger HPV-associated activation than classical MAGE/CTAG cancer-testis antigens. (4) Single-cell RNA-seq localises the meiotic programme to malignant epithelial cells. The lower panel summarises the validation layers. Embedded plots are schematic.*

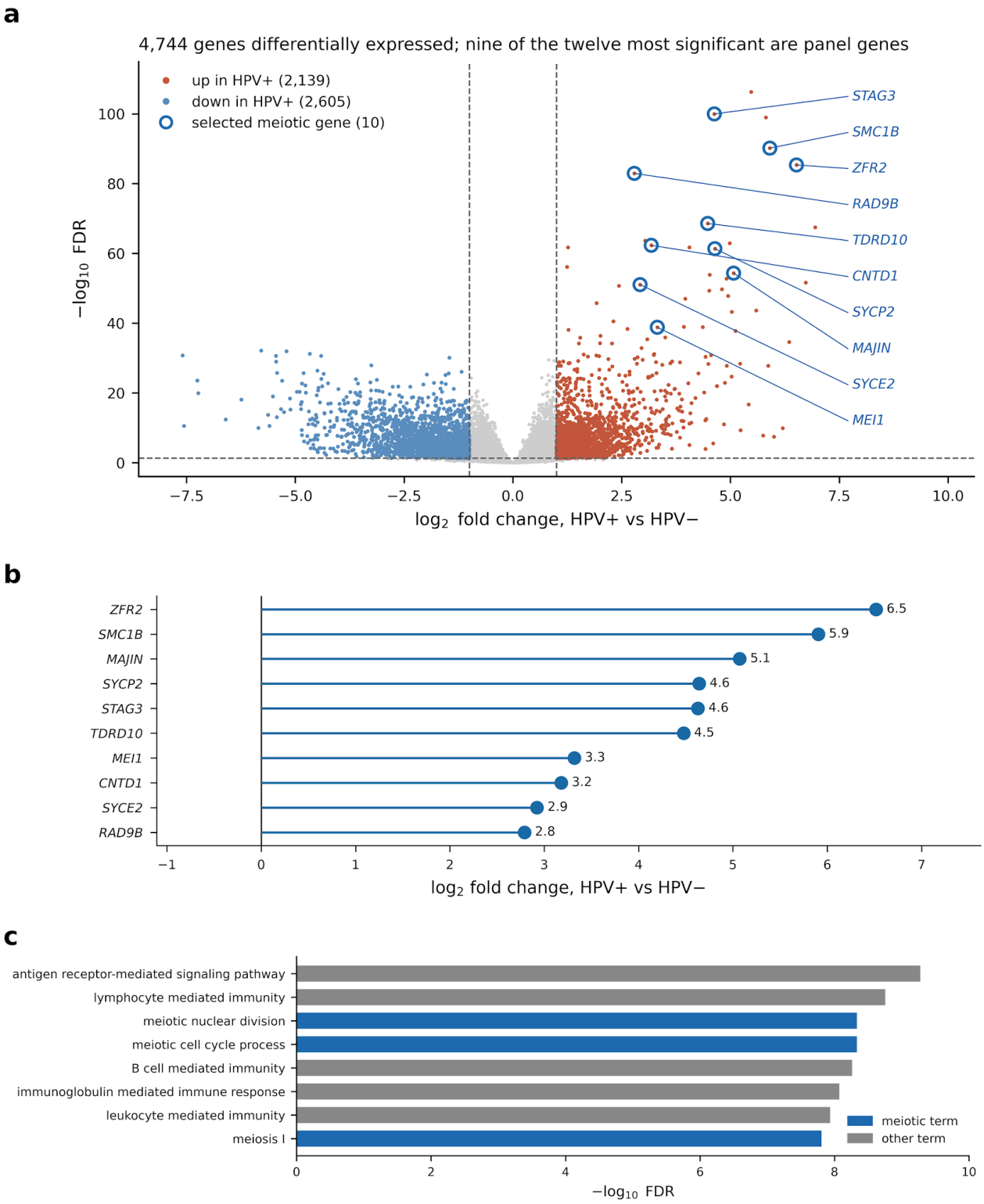


Figure 2

Differential expression in HPV+ versus HPV− TCGA-HNSC tumours places the meiotic machinery among the most strongly induced programmes. (a) Volcano plot of

differential expression between HPV+ and HPV− tumours: 4,744 differentially expressed genes (2,139 upregulated, 2,605 downregulated). (b) Log_2 fold changes of the ten selected meiotic genes (range 2.8–

6.5). (c) The eight most significantly enriched GO biological-process terms among upregulated genes, with meiotic terms in blue. n = 279 tumours (36 HPV+, 243 HPV−).

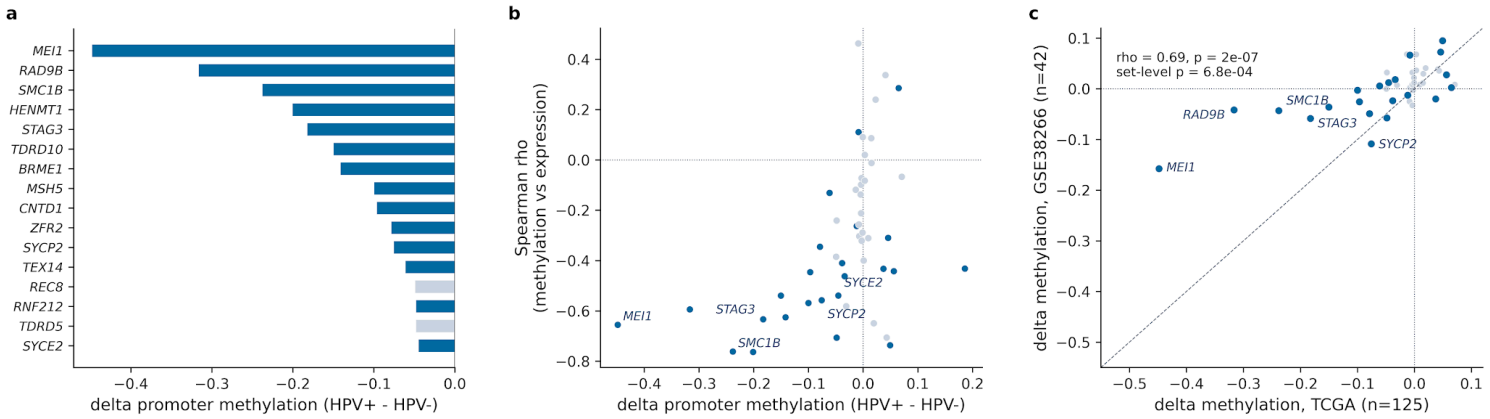


Figure 3

Promoter hypomethylation at meiotic loci in HPV+ HNSCC and its coupling to expression. (a) Panel genes with differential promoter methylation in TCGA-HNSC (44 HPV+ vs 81 HPV−), ranked by effect size; blue indicates FDR < 0.05. (b) Promoter methylation difference versus methylation–expression Spearman correlation; 17 of 18 FDR-significant genes lie in the lower-left quadrant (hypomethylated and inversely coupled to expression). (c) Replication of per-gene methylation differences in GSE38266 (21 HPV+ vs 21 HPV−, all oropharyngeal); the dashed line marks identity.

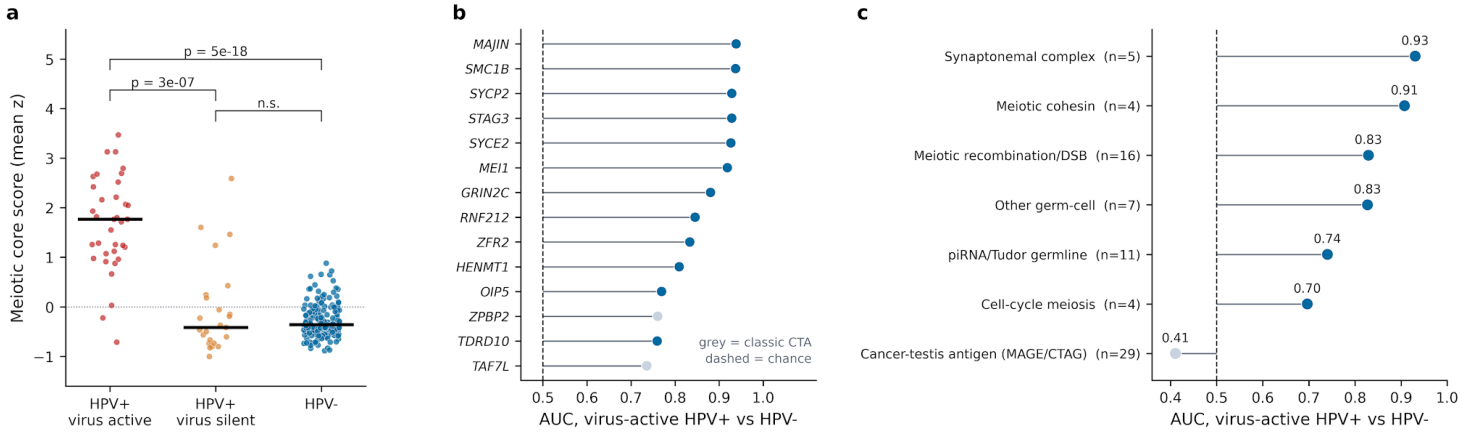


Figure 4

The meiotic programme is associated with transcriptionally active HPV and is specific to the meiotic machinery (GSE65858). *(a) Six-gene core meiotic score across the three groups: virus-active HPV+ tumours (n = 35) are elevated, whereas virus-silent HPV DNA+ tumours (n = 25) showed no detectable score difference from HPV− tumours (n = 196).*

(b) Per-gene AUC for virus-active HPV+ versus HPV− tumours; grey marks classic cancer–testis antigens. (c) Per-category AUC; the MAGE/CTAG family is at chance, whereas every meiotic category is elevated. Panels b and c are anchored at AUC = 0.5 (dashed line).

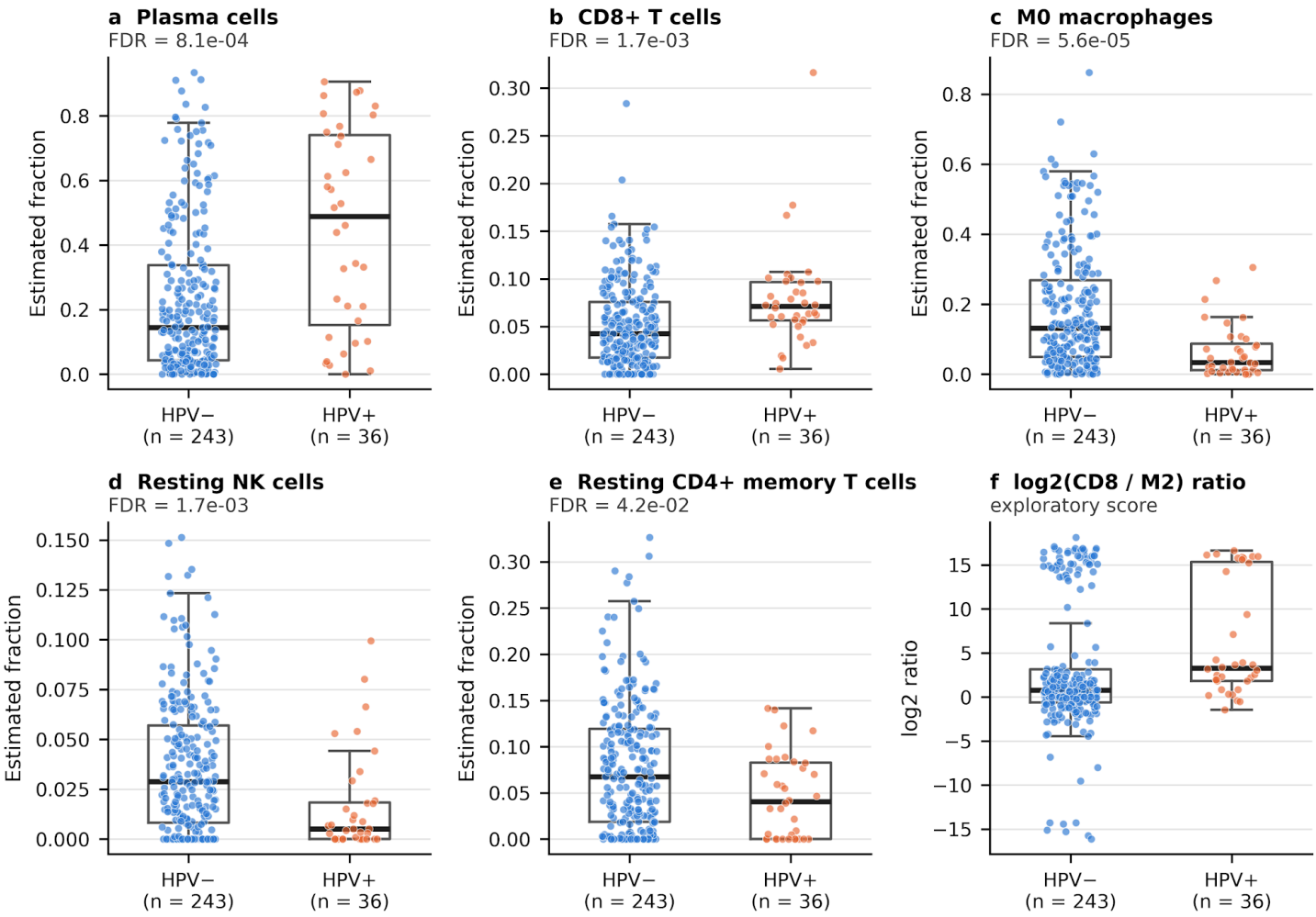


Figure 5

Tumour immune microenvironment of HPV+ versus HPV− TCGA-HNSC tumours. *CIBERSORTx (LM22) estimated fractions of (a) plasma cells, (b) CD8+ T cells, (c)*

M0 macrophages, (d) resting NK cells and (e) resting CD4+ memory T cells, the five populations that differed after Benjamini–Hochberg correction across 22 cell types (FDR shown above each panel). (f) Exploratory $\log_2$ CD8/M2 macrophage ratio. Boxes show the median and interquartile range; points are individual tumours (36 HPV+, 243 HPV−).

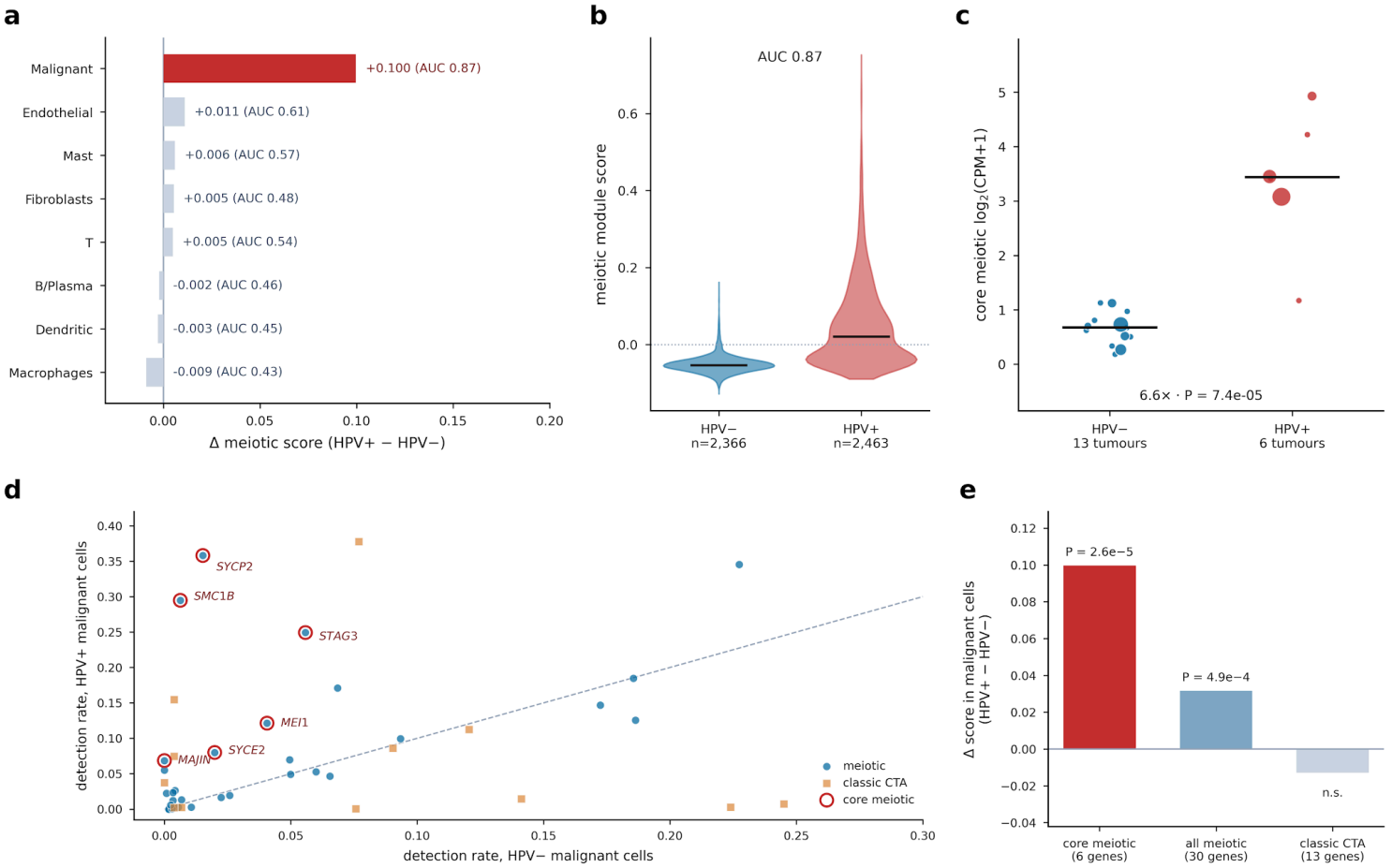


Figure 6

The HPV-associated meiotic programme is predominantly localised to malignant cells (GSE181919; 23,088 primary-tumour cells). *The core meiotic module comprises six genes (MEI1, SMC1B, STAG3, SYCP2, SYCE2 and MAJIN. (a) Difference in the six-gene core module score between HPV-positive and HPV-negative cells across annotated cell types. (b) Distribution of per-cell core module scores in HPV-positive and HPV-negative malignant cells. (c) Patient-level pseudobulk expression of the core module in malignant cells; point size is proportional to the number of contributing cells, and patients with fewer than 20 malignant cells were excluded. (d) Proportion of HPV-positive and HPV-negative malignant cells expressing individual panel genes. (e) Patient-level comparisons of the core meiotic module, wider meiotic panel and classical cancer–testis antigen comparator. Cell-level summaries and detection rates are descriptive, whereas patient-level comparisons were assessed using two-sided Mann–Whitney tests.*

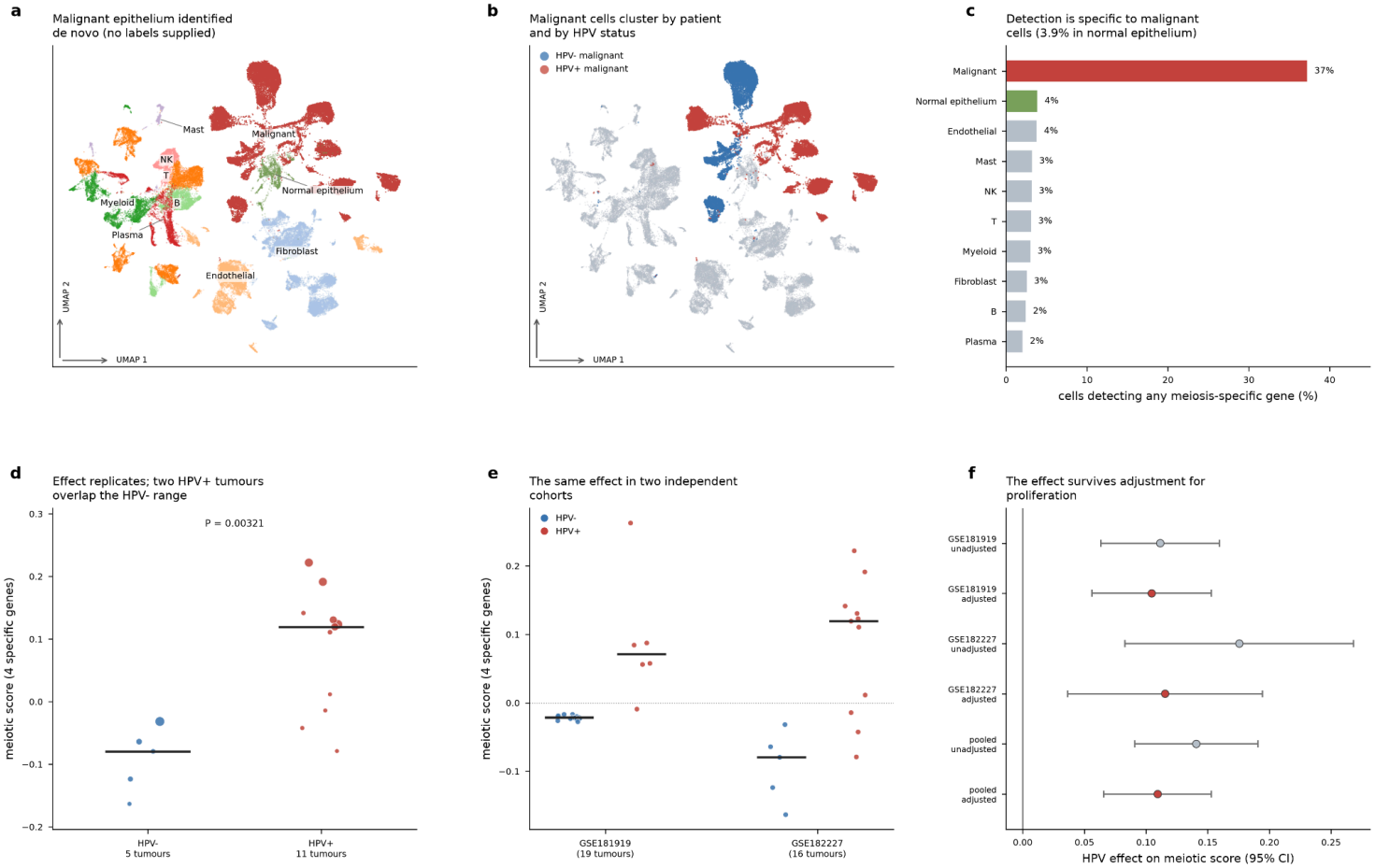


Figure 7

Independent replication in GSE182227 (16 oropharyngeal tumours, annotated de novo). *(a) Lineage annotation from canonical markers. (b) Malignant cells, coloured by HPV status, cluster by patient, whereas normal epithelium from different patients clusters together.*

(c) Percentage of cells detecting any of the four meiosis-specific genes, by compartment, after the epithelial split. (d) Per-tumour four-gene meiosis-specific score by HPV status; two HPV+ tumours fall within the HPV− range (see text). (e) The same four-gene contrast in both single-

cell cohorts. (f) HPV effect on the four-gene meiosis-specific score with and without adjustment for proliferation, by cohort and pooled; bars are 95% confidence intervals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf