Contents lists available at ScienceDirect

Neoplasia

The pan-cancer landscape of abnormal DNA methylation and intratumor microorganisms

Ping Zhou [a,b,c,d,1], Simon L. Lu [e,1], Liang Chang [a,b,c,d,1], Baoying Liao [a,b,c,d], Ming Cheng [a,b,c,d], Xiaolin Xu [a,b,c,d], Xin Sui [a,b,c,d], Fenting Liu [a,b,c,d], Mingshu Zhang [a,b,c,d], Yinxue Wang [a,b,c,d], Rui Yang [a,b,c,d], Rong Li [a,b,c,d,*], Heng Pan [a,b,c,d,*], Chao Zhang [f,*]

[a] Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China
[b] National Clinical Research Center for Obstetrics and Gynecology (Peking University Third Hospital), Beijing, China
[c] Key Laboratory of Assisted Reproduction (Peking University), Ministry of Education, Beijing, China
[d] Beijing Key Laboratory of Reproductive Endocrinology and Assisted Reproductive Technology, Beijing, China
[e] Boston University, Boston, MA, USA
[f] Department of Medicine, Boston University School of Medicine, Boston, MA, USA

ARTICLE INFO

ABSTRACT

Microorganisms play very important roles in carcinogenesis, tumor progression, and resistance upon treatment. Due to the challenge of accurately acquiring samples and quantifying low-biomass tissue microorganisms, most studies have focused on the effect of gut microorganisms on cancer treatments, especially the efficacy of immunotherapy. Although recent publications reveal the potential interactions between intratumor microorganisms and the immune microenvironment, whether and to what extent the intratumor microorganism could affect progression and treatment outcome remain controversial. This study is aiming to evaluate the associations among intratumor microorganisms, DNA methylation cancer driver genes, immune response, and clinical outcomes from a pan-cancer perspective, using 6,876 TCGA samples across 21 cancer types. We revealed that tumor microorganism dysbiosis is closely associated with the abnormal tumor methylome and/or tumor microenvironment, which might serve to enhance the proliferation ability and fitness for the therapy of tumors. These findings shed the light on a better understanding of the interactions between tumor cells and carcinogens during and after tumor formation, as well as microorganism-associated methylation alterations that could further serve as biomarkers for clinical outcome assessment.

Introduction

Cancer is the second most common cause of death and the main barrier to prolonged life expectancy globally, with a strikingly increased incidence in past decades [1,2]. As cancer was considered a disease primarily caused by abnormalities in the human genome, most studies to date have approached the topic of carcinogenesis from the human-centered perspectives, such as somatic mutations [3]. Furthermore, rapid development in high-throughput sequencing, computational biology, and tumor immunology have deepened our understanding of cancer genesis [4–6]. Particularly, recent studies have highlighted that microorganisms play a key role in cancer pathogenesis. For example, host-microbe interactions were identified as an important factor in the formation, diagnosis, prognosis, and treatment of cancer [7,8]. Additionally, multiple bacteria and viruses are ubiquitous in tumors and para-carcinoma tissues, which can directly affect the tumor microenvironment, exerting impacts on tumor recurrence and drug resistance [9–11]. The contributions of these microorganisms to the physiological stability of the human body are known to be significant. Their metagenomic genes are involved in a variety of metabolic and immune regulation pathways including anti-tumor immune surveillance [12–16]. The balance between host and microbiota is considered one of the preconditions for maintaining a healthy physiological state of the human body, and a perturbation might result in cancer development.

Several microorganisms have been proven to be the direct causal pathogen of cancer, while others indirectly initiate cancers by affecting the immune status and metabolites of the host. Recent studies have shown that such microorganism-induced immune responses can

---

also affect the host methylome. For instance, colorectal cancer (CRC)-related dysbiosis induces methylation changes of host genes directly, and the corresponding cumulative methylation index alongside associated bacteria might be potential biomarkers for CRC [17]. Together, epigenetic changes are an important way for microorganisms to regulate the transcriptional program in tumor cells and thus promote growth.

Cancer cells are characterized by a disrupted DNA methylation profile including site-specific hypermethylation and genome-wide hypomethylation [18,19]. The genome-wide analysis of DNA methylome and transcriptome has contributed to the identification of novel molecular subtypes within canonical subgroups [20,21]. Of note, DNA methylation has been reported to improve disease classification and is associated with microbiota composition [22]. Accurate identifications of tumor subtypes will not only improve the construction of preclinical models but also accelerate the development of personalized treatment [23]. In addition, epigenetic therapy has the chance to convert a tumor cell from an immune repressive status (immune cold) to an immune permissive status (immune hot) via regulating various factors of the tumor microenvironment that normally prevent the therapeutic effect of immune-checkpoint inhibition [24].

Public microbiome projects such as the Human Microbiome Project and the Metagenomics of the Human Intestinal Tract have provided tremendous insights into the diversity and function of human flora [25,26]. However, these databases are dominated by tissue swab and stool samples that do not necessarily reflect the microbial composition of local tissues [27]. The high content of human DNA in local tissue samples interferes the microbial identification and increases the challenge of accurately distinguishing microbiome fragments from tissue ones. Our previous study has presented a systematic framework of microbiome profiling directly from endoscopic biopsies by whole genome sequencing (WGS), allowing for the identification of the microbiome composition from primary tissues as well as the study of causative relationships between the microbiome and disease [28].

In our study, we utilized RNA-seq, WGS, and Infinium HumanMethylation450 BeadChip data in TCGA Pan-Cancer analysis project to characterize the microorganisms and methylomes of tumor samples. Two major clusters were identified based on tumor microorganism communities. Varied microorganism composition patterns were observed among different tumor types. Microorganism dysbiosis was speculated to be associated with abnormal DNA methylation and immune microenvironment in tumor cells. Patient samples of the cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) with different human papillomavirus (HPV) abundances showed distinct overall survival (OS) and the favorable outcome of patients with high HPV abundance might be associated with the promoter hypomethylation of specific genes such as *HKDC1*. This study aimed to provide an additional data source of intratumor microorganism structures and the landscape of DNA methylation cancer driver genes for 21 cancer types in TCGA data, laying a foundation for further analysis of the relationship between intratumor microorganisms and tumor epigenetics.

Methods

Study overview and data collection

To characterize the microorganisms and methylomes of tumor samples, we collected RNA-seq, WGS, and Infinium HumanMethylation450 BeadChip data in TCGA Pan-Cancer analysis project [29]. Microorganisms of 17,625 tumor samples were identified from RNA-seq or WGS data using several state-of-the-art methods [5,27,30]. To infer DNA methylation cancer driver genes, 9,664 methylation array data of 33 tumor types were obtained from TCGA Pan-Cancer analysis project [29]. MethSig [19] was used to detect DNA methylation cancer driver genes for each cancer type individually, which requires at least RNA-seq of

2 normal samples, methylation array data of 2 normal samples, and methylation array data of 40 tumor samples from the same cancer type. DNA methylation cancer drivers were successfully inferred from 7,052 tumor samples of 21 tumor types. In order to accurately evaluate the associations between DNA methylation and intratumor microorganisms, 6,876 samples of 21 tumor types with DNA methylation cancer driver inference and intratumor microorganism identification at the same time were selected for downstream analyses.

Identification of intratumor microorganisms and downstream analyses

We employed the microorganism identification pipelines from two recently published papers [5,27]. Kraken [31], Kraken2 [32], PathSeq [33], and Shogun [34] were used in the above papers for the microorganism profiling of TCGA samples. A total of 59,974 microbial genomes were downloaded via RepoPhlan (https://bitbucket.org/nsegata/repophlan) and filtered according to published criteria [5]. Potential lab contaminations were detected and removed based on published methods [5,27]. Principal coordinates analysis (PCoA) and hierarchical clustering analysis were applied to the microorganism profile for each cancer type based on Hellinger distance measurement. ConsensusClusterPlus (version 1.54.0) was used to determine the optimized cluster counts and membership of each sample [35]. In most cases, two major clusters were used for downstream association analysis since a larger number of subgroups is not conducive to the inference of DNA methylation cancer drivers in each subgroup, which requires a sufficient number of tumor samples. Of note, further stratification of each sub-cluster is feasible under the framework we provided such as in CESC when the sample size is sufficient and further stratification is biologically relevant.

To evaluate the biological and clinical relevance of the microorganism-derived major clusters, the cancer subtype information of the same TCGA study was downloaded from cBioPortal [36,37] and compared with the major clusters. Only 8 out of 21 cancer types have more than 2 cancer subtypes and a sufficient number of overlapped samples (> 100). The chi-square test was used to evaluate the dependency or association between microorganism clusters and known cancer subtypes.

Inference of DNA methylation cancer driver genes and differentially methylated promoters

Promoter (defined as ± 2 kb windows centered on RefSeq transcription start site) methylation was measured using the average methylation levels of all CpGs inside. Only promoters with at least 5 CpGs were included. MethSig, an in-house pipeline, was employed to infer DNA methylation cancer driver genes based on methylation array data [19]. DNA methylation cancer drivers are those DNA methylation changes that occur deterministically and drive the cancer phenotype. This concept was defined to discriminate driver promoter hypermethylation changes from the far larger number of stochastic DNA methylation changes without biological consequences (passenger DNA methylation changes). MethSig is a novel statistical inference framework that accounts for the varying stochastic hypermethylation rate across the genome and between samples, providing accurate and reproducible inference of methylation cancer drivers. For each tumor type, DNA methylation cancer drivers were inferred in either all the samples (drivers) or two major subgroups defined by microorganisms (sub-drivers). Promoters with a Benjamini-Hochberg false discovery rate (BH-FDR) Q value less than 0.05, inferred in all the samples, were defined as drivers. Sub-drivers were defined as promoters that were drivers in either major cluster defined by local microorganisms (BH-FDR Q < 0.05) while not drivers in all the samples. Differentially methylated promoters (DMPs) between two major microorganism-derived clusters were defined as a minimum of 5% methylation difference and a BH-FDR Q less than 0.05 (two-sided Mann-Whitney *U* test).

Enrichment analysis of immune cell types in tumor samples

To characterize the immunological landscape of tumor samples, xCell [38] was applied to RNA-seq data to quantify different types of immune cells for each sample. Differentially enriched immune cell types between two microorganism-derived clusters were defined as a BH-FDR $Q < 0.05$ (two-sided Mann-Whitney U test).

Pathway enrichment analysis and survival analysis

We performed functional analysis on the given gene list with DAVID [39] (version v2022q3). Biological process GO terms, KEGG pathways, and Reactome pathways with an enrichment p-value less than 0.1 were selected as overrepresented functions. Survival analysis was presented by the Kaplan-Meier plot and the p-value was calculated by a log-rank test.

Results

Different tumor types have varied intratumor microorganism landscapes

TCGA project provides a comprehensive genomic landscape for the most common human cancers. Due to the lack of a specifically designed assay to quantify intratumor microorganisms, associations among local microorganisms, microenvironments, and host genomics/epigenomics were not discussed in official TCGA publications from a pan-cancer perspective. Recent studies attempted to infer microorganism compositions from RNA-seq [5], whole exome sequencing (WES) [27,30], and WGS [5,27] data for a variety of TCGA cancer types (**Table S1**). A systematic inconsistency remains in the microorganism compositions identified by different studies due to varied data processing methods and different sequencing strategies. To avoid the bias introduced by methodologies and enable the evaluation of interactions between tumor microorganisms and methylomes, 6,876 TCGA tumor samples from 21 tumor types with RNA-seq and DNA methylation array data were used (Fig. 1). Raw counts of samples were downloaded, reported lab contaminants were

removed [5], and then relative microorganism abundances were calculated (Fig. 1).

Two major clusters determined by hierarchical clustering were colored in PCoA plots, which reflect the microorganism community structure of each cancer type (Fig. 2A-D, Fig. S1). To explore the biological and clinical relevance, we compared known cancer subtypes to our microorganism-derived clusters. The associations can be only evaluated in 8 out of 21 tumor types (Table S2). Of note, microorganism-derived clusters are only significantly associated with cancer subtypes in CESC and uterine corpus endometrial carcinoma (UCEC), indicating distinct contributions of the local microorganism community to tumor stratifications. In CESC, the C1 cluster is significantly enriched in adenocarcinoma and the C2 cluster is enriched in squamous cell carcinoma (Fig. 2E, P = 0.004). In UCEC, the C2 cluster is enriched in the subtypes with high copy numbers (UCEC_CN_High, Fig. 2F, $P = 1 \times 10^{-5}$). The different pathological roles of intratumor microorganisms in the above tumor types need further exploration.

Average relative abundance differences at the family level between the two major clusters were presented (Fig. 2G). Overall, there are two distinguished scenarios among different tumor types in TCGA data (Fig. S2). In the first one, two major clusters carried distinct profiles, and one cluster is predominated by 1 - 2 specific microorganisms which have a more than 20% relative abundance difference between the two clusters. For example, the *Papillomaviridae* and *Hepadnaviridae* are overrepresented in one group of patients in CESC or liver hepatocellular carcinoma (LIHC), respectively (Fig. S2). According to a CDC report, about 80% - 90% of cervical cancers are related to HPV infection [40,41], and hepatitis B virus (HBV) might be the leading cause of approximately 65% of liver cancer [42]. Identification of known carcinogens in cancer samples proves the reliability of microorganism detection from sequencing data of primary tissues. In another scenario, no single family had more than 20% relative abundance differences between two clusters in most cancer types, while small consistent relative abundance shifts of multiple microorganisms between two clusters remained (Fig. S2). These results confirmed the reliability of intratumor microorganism inference through tissue RNA-seq data and revealed
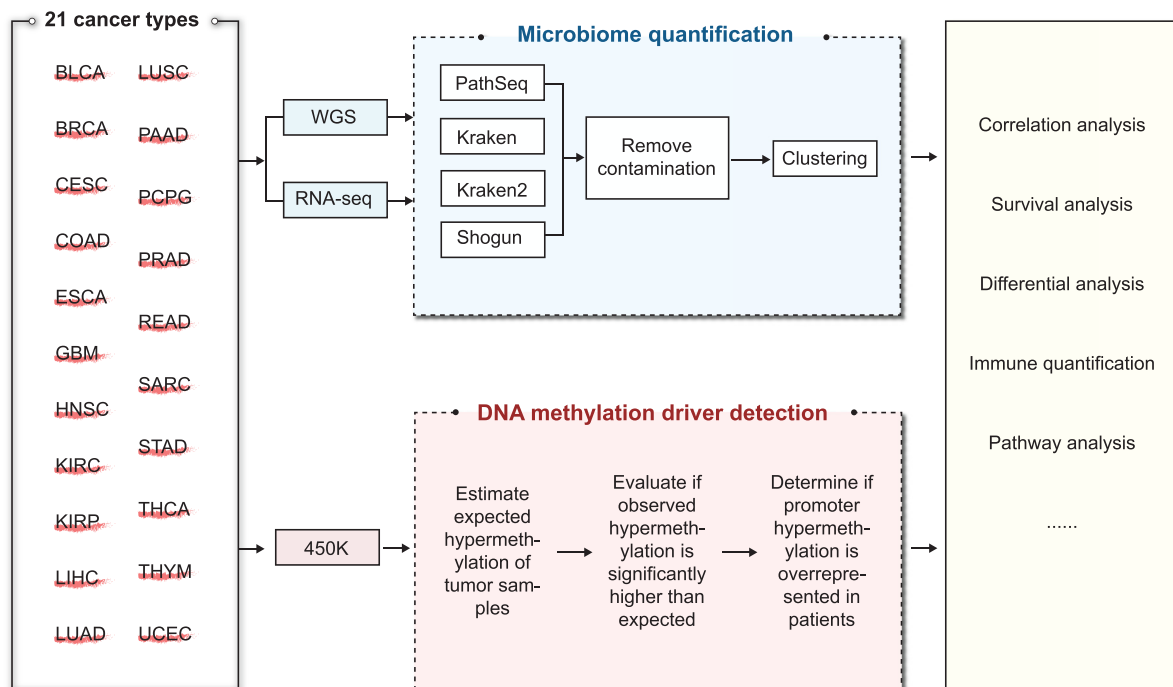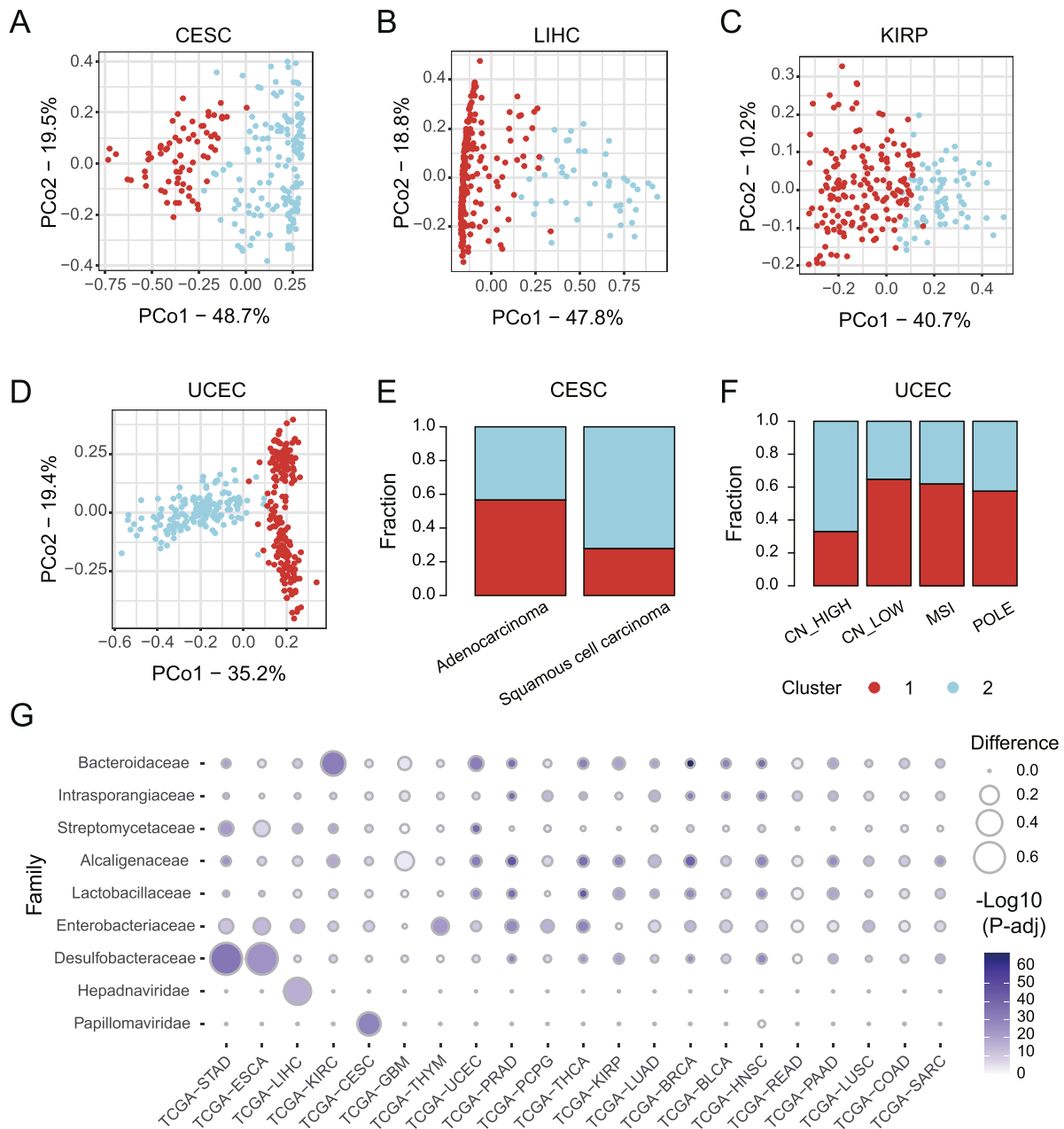

Fig. 1. Workflow of the data and analysis used in this project.

Fig. 2. Different tumor types have varied intratumor microorganism landscapes. (**A-D**) Scatter plots showing principal coordinate analysis based on Hellinger distance of microorganism profiles among samples from (**A**) CESC, (**B**) LIHC, (**C**) KIRP, and (**D**) UCEC. (**E-F**) Segmentation plots showing the association between microorganism-derived major clusters and known cancer subtypes in (**E**) CESC and (**F**) UCEC. (**G**) A bubble plot showing microorganism composition differences between the top two major clusters at a family level. Bubble size is the relative difference between the top two major clusters, and color presents the -log10 p-adjust values of the t-test.

the varied microorganism composition patterns among different tumor types.

MethSig identifies distinct numbers of DNA methylation driver genes among different cancer types

Epigenetic landscape changes are a hallmark of cancer [43] and play a crucial role in response to extracellular stimuli such as infection of local microorganisms [44]. Linking local microorganism composition patterns to tumor epigenetic landscape might result in an enhanced understanding of the role of local microorganisms in cancer initiation and progression. However, our ability to differentiate driver DNA methylation changes from passenger events is limited. Thus, it is challenging to explore the functional association between microorganisms and key

epigenomic abnormalities. MethSig, an in-house novel statistical framework, was designed to infer DNA methylation cancer drivers. Compared with benchmarked methods, MethSig delivers well-calibrated quantile-quantile plots and more reproducible identification in independent cancer cohorts (Fig. 3A). Importantly, in comparison with extant methods, MethSig achieves higher sensitivity and specificity in the inference of likely DNA methylation drivers, defined as close association with gene repression and clinical outcome (Fig. 3A).

MethSig was applied to the same 6,876 tumor samples for microorganism discovery. Different cancer types had a varied prevalence of DNA methylation cancer driver genes, from 1.4% to 13.5% (Fig. 3B, BH-FDR Q < 0.05). Moreover, the significance of top DNA methylation drivers of specific tumor types was higher than others (Fig. 3C). Using thymoma (THYM) as an example, all the top 500 DNA methylation drivers had a

4
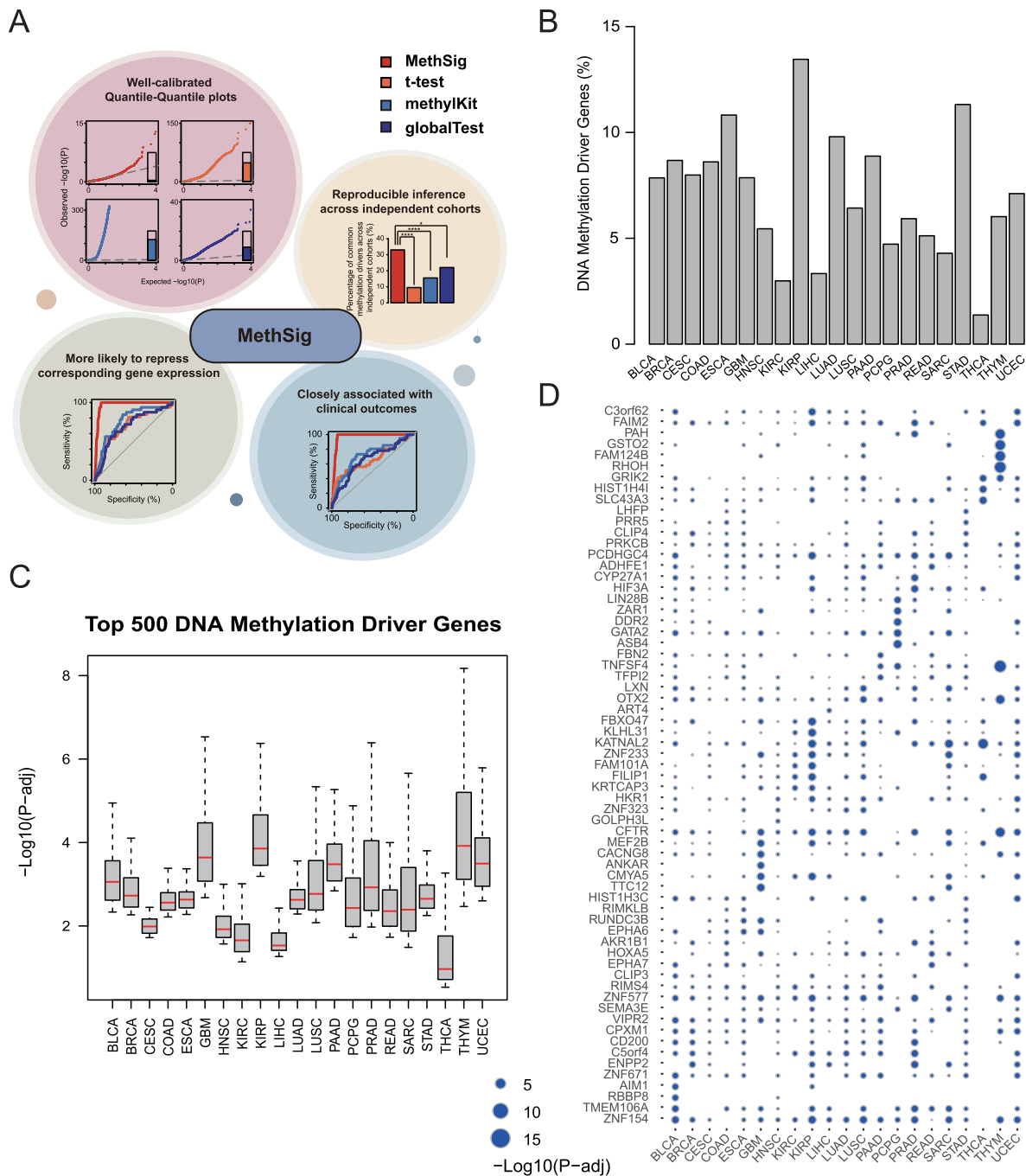
Fig. 3. Different cancer types have distinct numbers of DNA methylation driver genes. (**A**) Simplified illustration of the benefits of the inference of DNA methylation drivers using MethSig. (**B**) A barplot showing the percentage of DNA methylation cancer drivers out of all the tested promoters across 21 cancer types. (**C**) A boxplot showing the significant levels of the top 500 DNA methylation driver genes across 21 cancer types. (**D**) A bubble plot showing the significance of the top 5 drivers derived from each cancer type across all tumors.

BH-FDR Q value less than 0.005 while only the top 76 drivers in thyroid carcinoma (THCA) had the same degree of significance (Fig. 3C). By comparing the top 5 DNA methylation drivers across all the tumor types, we observed a plethora of methylation drivers that were highly tumor-specific, indicating heterogeneous landscapes of DNA methylation drivers among different tumor types (Fig. 3D). For example, a tumor suppressor gene (TSG), *RBBP8*, was only identified as a driver gene in the bladder urothelial carcinoma (BLCA) and the head and neck squamous cell carcinoma (HNSC), whose hypermethylation will disrupt DNA repair function [45]. In summary, most tumor types showed a distinct landscape of DNA methylation drivers in comparison with each other,

revealing a heterogenous methylome evolution among different cancer types.

Microorganism dysbiosis is associated with abnormal DNA methylation and immune microenvironment in tumor cells

To explore the association between local microorganisms and tumor epigenetics, we identified DMPs between two major microorganism-derived clusters for all cancer types. The numbers of statistically significant DMPs are highly variable across different cancer types (**Table S3**). Four cancer types with the most DMPs (> 100) were further selected
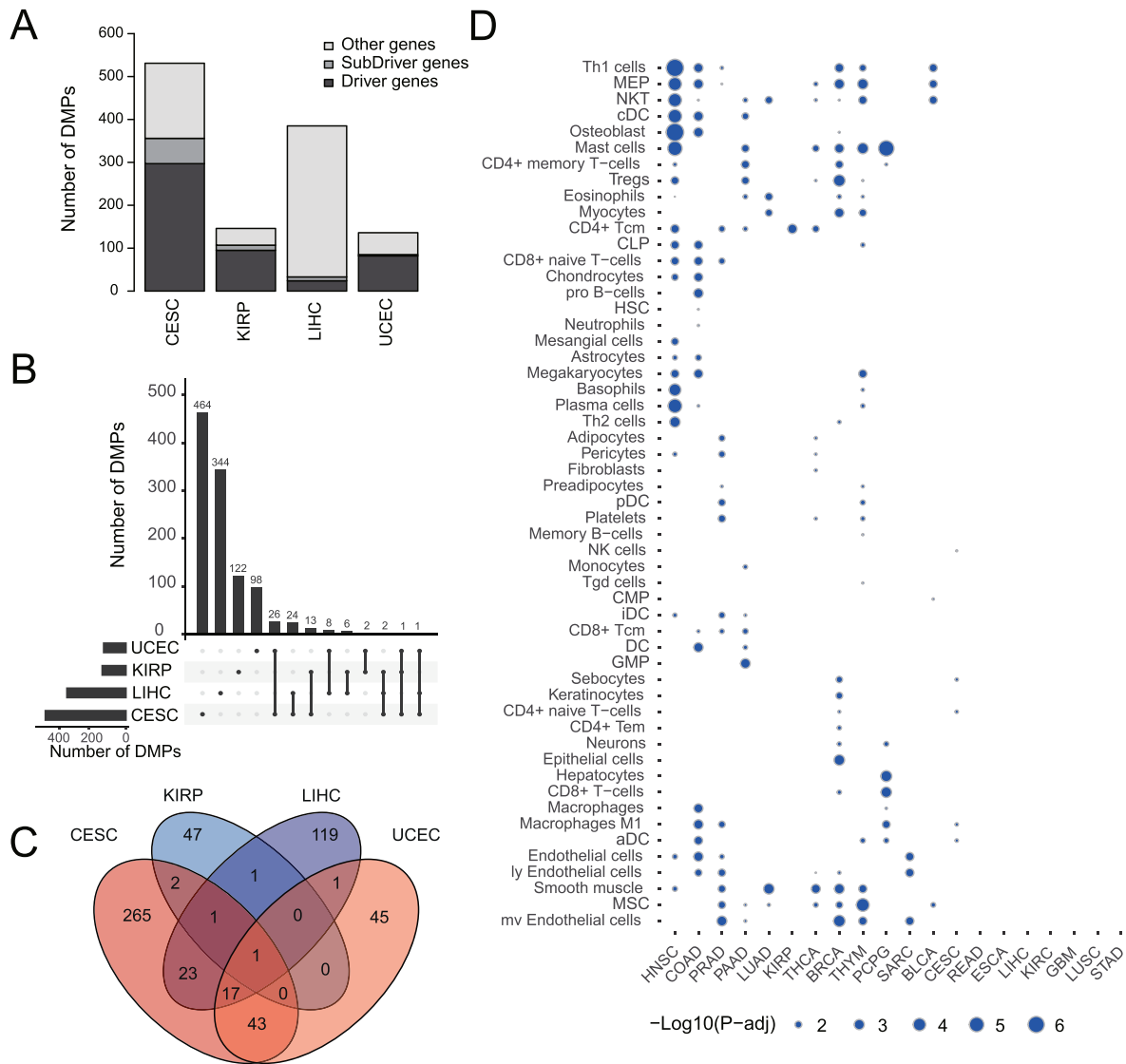
Fig. 4. Microorganism dysbiosis is associated with methylome and the immune microenvironment of tumor cells. (**A**) A barplot showing the number of DMPs between the two major microorganism-derived clusters in CESC, KIRP, LIHC, and UCEC. (**B**) An Upset plot showing the overlap of DMPs across multiple cancer types. (**C**) A Venn diagram showing the overlap of DMP-enriched functions and pathways across multiple cancer types. (**D**) A bubble plot showing differences in the immune cell abundances between the top two major clusters defined by microorganisms.

to evaluate the potential DNA methylation alterations along with intra-tumor microorganism dysbiosis (Fig. 4A). Identified DMPs were further annotated to DNA methylation drivers and sub-drivers (Fig. 4A, **Table S3**). More than 60% of DMPs in CESC, kidney renal papillary cell carcinoma (KIRP), and UCEC were drivers or sub-drivers, implying a close association between microorganism dysbiosis and tumor initiation. In LIHC, less than 10% of DMPs were drivers or sub-drivers, reflecting the independent role of microorganisms in driving tumors through regulating DNA methylation, which requires further exploration. An Upset plot was generated for visualizing common DMPs across multiple cancer types (Fig. 4B). It is not surprising to observe that most DMPs are unique to a single cancer type. This observation indicated a heterogeneous degree of association between local microorganisms and tumor methylome, which is expected considering the varied microorganism composition and dominant family across cancer types. Nevertheless, we still revealed that CESC shared 16, 27, and 28 common DMPs with KIRP, LIHC, and UCEC, respectively (Fig. 4B). To find the functional commonality, we performed the functional enrichment analysis of DMPs for each cancer type. CESC, LIHC, and UCEC shared many common functions and pathways (Fig. 4C, **Table S4**), while DMPs in KIRP were enriched in a

distinct set of pathways compared with other cancer types. We hypothesized that the commonality of DMP-enriched pathways might be due to the similarity of pathogens in CESC, LIHC, as well as UCEC since HPV and HBV infection are very prevalent in those three types of cancers. In the contrast, the dominant microorganisms in KIRP are different and thus targeted a different set of promoters.

Besides abnormal tumor DNA methylation, the local microorganism dysbiosis may also interact with the immune microenvironment of tumor cells directly. We employed xCell to infer the immunological landscapes of 21 cancer types (Fig. 4D). Similarly, we observed heterogeneous changes in the abundance of immune cell types across different cancer types. In HNSC, the abundance of a variety of immune cell types was changed between clusters such as Th1 cells, Th2 cells, mast cells, plasma cells, and Tregs, which were identified to be associated with HNSC prognosis. For instance, the abundance of mast cells can serve as a prognostic predictor in HNSC [46]. In a plethora of cancer types including CESC, KIRP, and LIHC, whose methylomes were changed along with tumor microorganism shifts, we did not observe a significant abundance difference in most of inferred immune cell types.

Thus, we speculated that tumor microorganism dysbiosis is associated with abnormal tumor methylome and/or tumor microenvironment, enhancing the proliferation ability and fitness for the therapy of tumor cells together.

HPV-associated DNA methylation and transcriptional changes in CESC

Although persistent HPV infection is the leading cause of cervical cancer, the viral load could be highly variable in different CESC samples. According to the HPV abundance, we named three clusters HPV-H (high), HPV-M (medium), and HPV-L (low, Fig. 5A-B). The significance of different clinical outcomes across clusters was tested. Higher HPV abundance was significantly associated with longer OS (Fig. 5C), indicating HPV infection as a favorable predictor of OS in CESC. Even though such a favorable association has been reported in several carcinomas including HNSC [65], the impact of HPV infection on the survival rate of CESC was not carefully determined previously.

Promoter methylation and gene expression levels of tumor samples were compared with normal tissues, aiming to explore the potential molecular differences. A group of perturbed genes in CESC tumor samples including *SYCP2* and *PCDH10* were identified (Fig. 5D). *SYCP2* is recently found to be involved in the initiation of HPV-related cancers [47], especially significantly upregulated in cervical cancer and oropharyngeal squamous cell carcinoma [48–50]. This is consistent with our observations in TCGA CESC samples (Fig. 5D). In addition, the positive effect on *SYCP2* gene expression [47,51] might be a result of promoter hypomethylation associated with HPV infection (Fig. 5D). Our observation provides new insights into the potential regulatory mechanism of HPV infection on *SYCP2* gene expression through methylation in tumorigenesis of CESC. Another notable candidate *PCDH10* could inhibit proliferation, migration, and epithelial-to-mesenchymal transition of tumor cells via the Wnt/β-catenin signaling pathway [52]. As a TSG, *PCDH10* was reported as hypermethylated or down-regulated in cervical, colorectal, and esophageal cancers [52–54], which was also confirmed in the TCGA CESC cohort (Fig. 5D). However, there were no clear associations between promoter methylation levels of the above genes and clinical outcomes, implying that the dysregulated methylation status was the residue of HPV infection during tumorigenesis with limited impacts on tumor progression upon treatment.

Besides the above cancer drivers, a plethora DMPs, such as *HKDC1* and *SEMA3E*, are associated with survival differences across clusters. Up-regulation of *HKDC1* can be triggered by the overexpression of HPV8 E7 protein [66], implying a potential regulatory linkage between HPV infection and *HKDC1* gene expression. Moreover, *HKDC1* has been reported to be expressed in multiple cancers and proposed as a favorable prognosis biomarker in intrahepatic cholangiocarcinoma [55,56]. As a key regulator for cell-cell communication, cancer cell invasion and metastasis, angiogenesis, as well as inflammation [57–60], Sema3E was reported to be highly expressed in metastatic cancer cells and is considered a clinical marker for breast and ovarian cancer [57,58]. We observed greater *SEMA3E* promoter methylation levels in HPV-H CESC patients against remaining samples and corresponding expression repression in HPV-H versus HPV-M (Fig. 5E). Moreover, patients with higher *SEMA3E* promoter methylation levels had a favorable OS compared with the remaining samples (P = 0.14), implying *SEMA3E* is an important candidate target of HPV infection and through which HPV can affect the progression of CESC. Further experimental validations are needed to confirm our speculations.

In conclusion, our results revealed new insights into the explanation of how HPV could affect transcriptional disruption via the modification of promoter methylation levels in tumor cells, thus contributing to the promotion of tumorigenesis in cervical cancer.

Discussion

During the past decade, several studies revealed that gut microorganisms can affect cancer treatment, especially the efficacy of im-

munotherapy [61,62]. Nonetheless, the associations among the intratumor microorganisms, immune microenvironment, and treatment outcomes have been neglected due to the challenge of obtaining samples and quantifying low-biomass tissue microorganisms. In this study, we evaluated the survival differences between two major clusters of patients defined by the tumor microorganism community in 21 TCGA cancer types. Around 30% - 40% of cancer types showed different OS or progression-free survival with varied significance levels (Table S5). Glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), stomach adenocarcinoma (STAD), and UCEC have consistent trends in both survival measurements, implying that different microorganism structures are not only associated with OS but also the treatment outcome. Associations between intratumor microorganism presence and clinical features in lung cancer have been reported recently, which are consistent with our findings [63,64]. Further work will be needed to reveal the mechanism of how intratumor microorganisms can affect the clinical outcomes of tumor patients.

To better understand the interactions between intratumor microorganisms and cancer cells, we evaluated two principal axes along which microorganism dysbiosis can interact with tumors: tumor immune microenvironment and epigenome. Out of the 21 cancer types, 4 have numerous DNA methylation cancer drivers with significant promoter methylation differences between two major clusters defined by microorganisms, while they did not show abundance changes in most major immune cell types (Fig. 6). In contrast, 11 cancer types with significant differences in at least three immune cell types do not carry any DMPs. Thus, we hypothesized that intratumor microorganisms might be only associated with either abnormal immunological or epigenetic pathways. However, whether those mutual-exclusive alterations in the immune environment or methylome of tumor cells are direct effects of microorganism dysbiosis remains controversial and further experimental validations are needed to confirm the causal relationship. It is also unclear how tumor cells determine to undergo specific mechanisms to regulate the proliferation with or without treatment in different cancer types and whether those distinct patterns are due to the response to infection or required to recruit specific bacteria or viruses to accelerate the dysbiosis. Understanding the underlying mechanism will help provide novel insights into the treatment of cancers via monitoring or regulating local microorganisms.

Out of four cancer types with strong associations between microorganism structures and epigenetic alterations, CESC and LIHC are highly related to the infection of HPV and HBV, respectively. As known pathogens, those viruses account for more than 65% of all cancers. However, only a small proportion of samples in each cancer type carry a predominant level (> 30%) of HPV or HBV (28.7% in CESC and 12.5% in LIHC). To evaluate the post-tumorigenesis function of these pathogens, we further classified samples into three clusters according to relative abundances. The pathogen-dominating cluster has a better survival rate in both cancer types. Our findings of the DNA methylation alterations along with dysbiosis have the potential to shed light on a deeper understanding of the interactions between tumor cells and carcinogens after tumor formation. Furthermore, carcinogen-derived methylation changes could serve as biomarkers for clinical outcome assessment. Our observations also confirm that the major clusters derived from microorganism structures can be further stratified to match biologically or clinically well-defined sub-clusters, deepening our understanding of the associations between microorganism communities and tumor heterogeneities.

In addition to the above findings, we also discovered some potential false identifications from published studies, due to the limitation of sequencing data, microorganism detection methods, and databases used for identification. For example, partial samples in STAD and esophageal carcinoma (ESCA) are dominated by *Desulfobacteraceae sp*, which uses sulfur compounds as the main energy source and has a particularly low level in the human body. In addition, the *Helicobacter pylori* levels were found to be consistently low across all STAD samples, which is
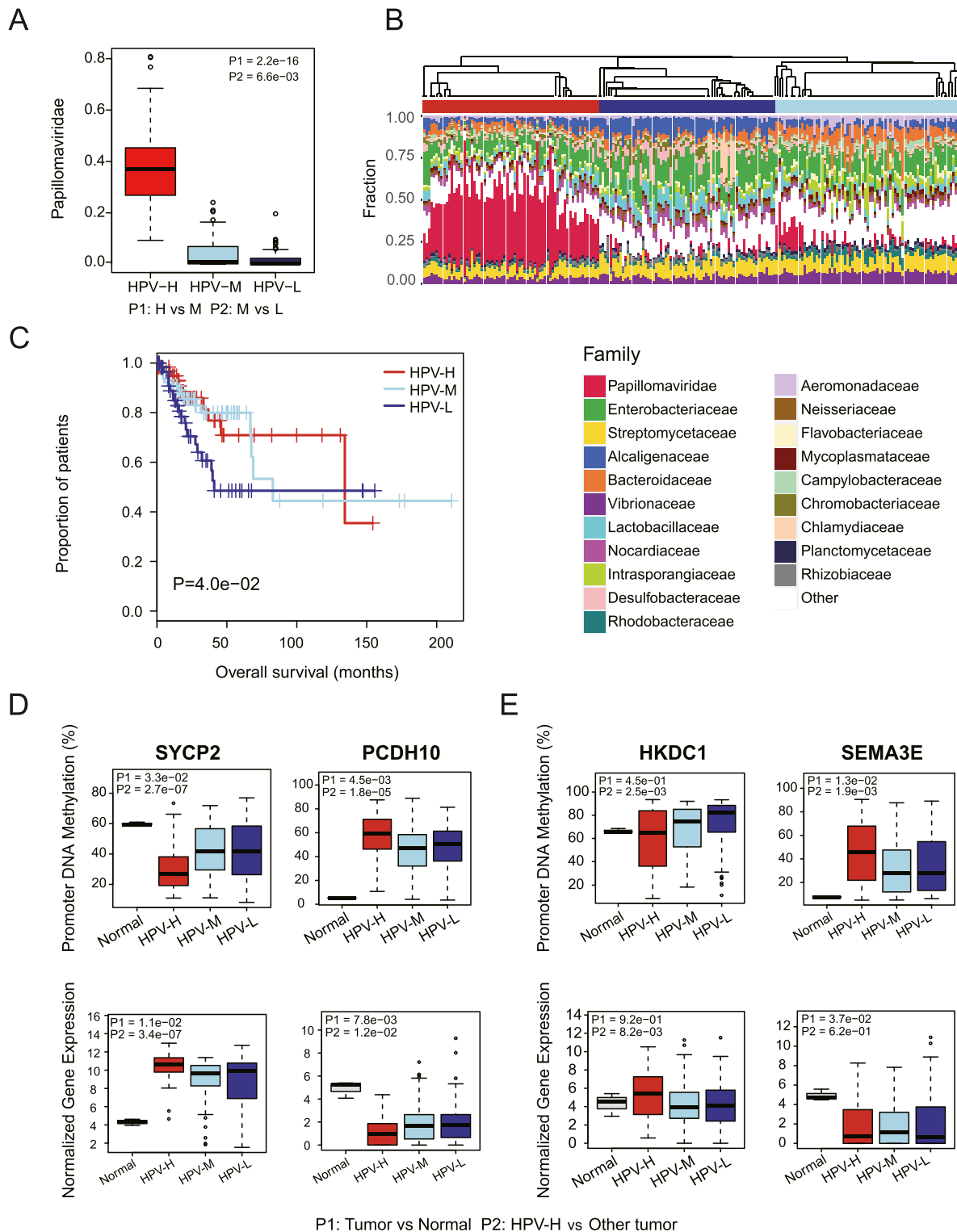
Fig. 5. HPV-associated DNA methylation and transcription changes in CESC. (**A**) A boxplot showing *Papillomaviridae* abundance of three major clusters in CESC: HPV-H (red), HPV-M (light blue), and HPV-L (dark blue). (**B**) Consensus clustering results of microorganism profile indicating three major clusters with different HPV fractions. (**C**) A Kaplan-Meier plot showing overall survival in CESC patients with a different abundance of HPV infection. (**D**) Boxplots showing promoter methylation levels and gene expression levels of *SYCP2* and *PCDH10* in normal samples as well as tumor samples with a different abundance of HPV infection. (**E**) Boxplots showing promoter methylation levels and gene expression levels of *HKDC1* and *SEMA3E* in normal samples as well as tumor samples with a different abundance of HPV infection.
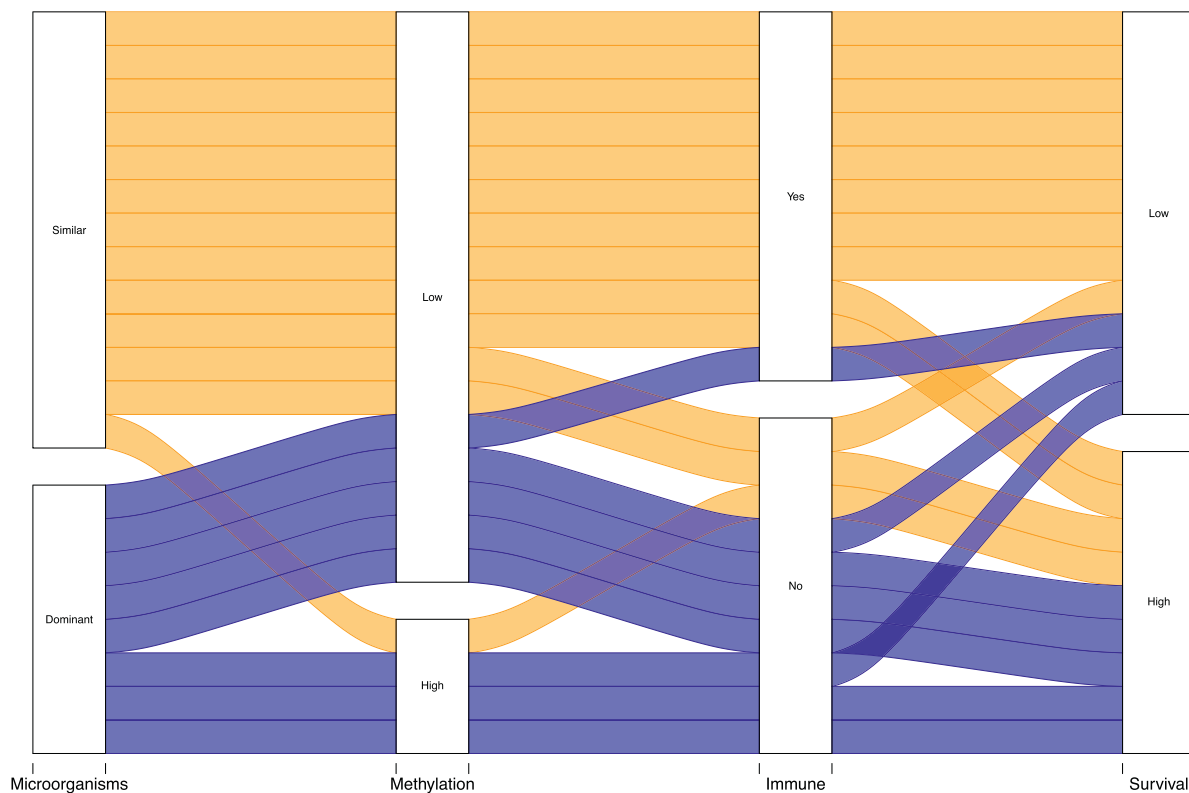
Fig. 6. An alluvial diagram reveals the associations among microorganism composition shifts, methylation alterations, immune response changes, and survival differences between two major microorganism clusters across 21 TCGA cancer types.

inconsistent with previous studies [28]. We speculated that these inconsistencies might be a result of the limitation in the filtering steps of the current protocol where the human reads in the sequencing data may not be removed adequately. Improvement of computational methods in microorganism identification is necessary to yield more accurate results. Specifically designed experimental protocols for capturing, amplifying, and sequencing the microorganisms from tumor samples will serve to improve the identification accuracy and minimize the biases introduced by batch effects, lab contaminations, and computational errors. Also, bacteria and viruses may affect distinct pathways during infections and the pathogenesis role of microorganisms beyond HPV and HBV requires further and extensive evidence to support.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ping Zhou: Investigation, Methodology, Software, Visualization, Formal analysis, Data curation, Writing – original draft. **Simon L. Lu:** Software, Data curation, Writing – review & editing. **Liang Chang:** Investigation, Data curation, Writing – original draft. **Baoying Liao:** Investigation, Writing – original draft. **Ming Cheng:** Visualization, Data curation. **Xiaolin Xu:** Data curation, Writing – original draft. **Xin Sui:** Visualization, Data curation. **Fenting Liu:** Data curation, Writing – review & editing. **Mingshu Zhang:** Data curation, Writing – original draft. **Yinxue Wang:** Data curation. **Rui Yang:** Data curation. **Rong Li:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Heng Pan:** Conceptualization, Investigation, Methodology, Formal analysis, Data curation, Visualization, Supervision, Writing – original draft. **Chao Zhang:** Conceptualization, Investigation, Methodology,

Formal analysis, Data curation, Visualization, Supervision, Writing – original draft.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neo.2023.100882.

References

[1] The global challenge of cancer, Nature Cancer 1 (2020) 1–2, doi:10.1038/s43018-019-0023-9.

[2] R.A. Hiatt, N. Beyeler, Cancer and climate change, Lancet. Oncol. 21 (2020) e519–e527, doi:10.1016/s1470-2045(20)30448-4.

[3] J. Zhang, et al., Pan-cancer analyses reveal genomics and clinical characteristics of the melatonergic regulators in cancer, J. Pineal Res. 71 (2021) e12758, doi:10.1111/jpi.12758.

[4] D. Nejman, et al., The human tumor microbiome is composed of tumor type-specific intracellular bacteria, Science 368 (2020) 973–980, doi:10.1126/science.aay9189.

[5] G.D. Poore, et al., Microbiome analyses of blood and tissues suggest cancer diagnostic approach, Nature 579 (2020) 567–574, doi:10.1038/s41586-020-2095-1.

[6] L.T. Geller, et al., Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine, Science 357 (2017) 1156–1160, doi:10.1126/science.aah5043.

[7] N. Cullin, C. Azevedo Antunes, R. Straussman, C.K. Stein-Thoeringer, E. Elinav, Microbiome and cancer, Cancer Cell 39 (2021) 1317–1341, doi:10.1016/j.ccell.2021.08.006.

[8] N. Pernigoni, et al., Commensal bacteria promote endocrine resistance in prostate cancer through androgen biosynthesis, Science (New York, N.Y.) 374 (2021) 216–224, doi:10.1126/science.abf8403.

[9] S. Yuan, et al., Translatomic profiling reveals novel self-restricting virus-host interactions during HBV infection, J. Hepatol. 75 (2021) 74–85, doi:10.1016/j.jhep.2021.02.009.

[10] A.A. McBride, Human papillomaviruses: diversity, infection and host interactions, Nat. Rev. Microbiol. 20 (2022) 95–108, doi:10.1038/s41579-021-00617-5.

[11] M.C. Kordahi, et al., Genomic and functional characterization of a mucosal symbiont involved in early-stage colorectal cancer, Cell Host Microbe 29 (2021) 1589-1598.e1586, doi:10.1016/j.chom.2021.08.013.

[12] C. Jin, et al., Commensal microbiota promote lung cancer development via gammadelta T Cells, Cell 176 (2019) 998-1013 e1016, doi:10.1016/j.cell.2018.12.040.

[13] B. Aykut, et al., The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL, Nature 574 (2019) 264–267, doi:10.1038/s41586-019-1608-2.

[14] J.J. Tsay, et al., Lower airway dysbiosis affects lung cancer progression, Cancer Discov. 11 (2021) 293–307, doi:10.1158/2159-8290.CD-20-0263.

[15] C.M. Dejea, et al., Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria, Science 359 (2018) 592–597, doi:10.1126/science.aah3648.

[16] W.S. Garrett, Cancer and the microbiota, Science 348 (2015) 80–86, doi:10.1126/science.aaa4972.

[17] I. Sobhani, et al., Colorectal cancer-associated microbiota contributes to oncogenic epigenetic signatures, Proc. Nat. Acad. Sci. U.S.A. 116 (2019) 24285–24295, doi:10.1073/pnas.1912129116.

[18] A. Nishiyama, M. Nakanishi, Navigating the DNA methylation landscape of cancer, Trends Genet. 37 (2021) 1012–1027, doi:10.1016/j.tig.2021.05.002.

[19] H. Pan, et al., Discovery of candidate DNA methylation cancer driver genes, Cancer Discov. 11 (2021) 2266–2281, doi:10.1158/2159-8290.Cd-20-1334.

[20] C.M.S. Tesileanu, et al., Prognostic significance of genome-wide DNA methylation profiles within the randomized, phase 3, EORTC CATNON trial on non-1p/19q deleted anaplastic glioma, Neuro-oncol. 23 (2021) 1547–1559, doi:10.1093/neuonc/noab088.

[21] F.M.G. Cavalli, et al., Intertumoral Heterogeneity within Medulloblastoma Subgroups, Cancer Cell 31 (2017) 737-754.e736, doi:10.1016/j.ccell.2017.05.005.

[22] F.J. Ryan, et al., Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease, Nat. Commun. 11 (2020) 1512, doi:10.1038/s41467-020-15342-5.

[23] Y.C. Liu, et al., Demethylation and Up-Regulation of an Oncogene after Hypomethylating Therapy, N. Engl. J. Med. 386 (2022) 1998–2010, doi:10.1056/NEJMoa2119771.

[24] M.J. Topper, M. Vaz, K.A. Marrone, J.R. Brahmer, S.B. Baylin, The emerging role of epigenetic therapeutics in immuno-oncology, Nature Rev. Clin. Oncol. 17 (2020) 75–90, doi:10.1038/s41571-019-0266-5.

[25] J. Qin, et al., A human gut microbial gene catalogue established by metagenomic sequencing, Nature 464 (2010) 59–65, doi:10.1038/nature08821.

[26] K.L. Hu, et al., Predictive value of serum kisspeptin concentration at 14 and 21 days after frozen-thawed embryo transfer, Reprod. Biomed. Online 39 (2019) 161–167, doi:10.1016/j.rbmo.2019.03.202.

[27] A.B. Dohlman, et al., The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants, Cell Host Microbe 29 (2021) 281-298.e285, doi:10.1016/j.chom.2020.12.001.

[28] C. Zhang, et al., Identification of low abundance microbiome in clinical samples using whole genome sequencing, Genome Biol. 16 (2015) 265, doi:10.1186/s13059-015-0821-z.

[29] J.N. Weinstein, et al., The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (2013) 1113–1120, doi:10.1038/ng.2764.

[30] R.M. Rodriguez, B.Y. Hernandez, M. Menor, Y. Deng, V.S. Khadka, The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing, Comput. Struct. Biotechnol. J. 18 (2020) 631–641, doi:10.1016/j.csbj.2020.03.003.

[31] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (2014) R46, doi:10.1186/gb-2014-15-3-r46.

[32] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257, doi:10.1186/s13059-019-1891-0.

[33] A.D. Kostic, et al., PathSeq: software to identify or discover microbes by deep sequencing of human tissue, Nat. Biotechnol. 29 (2011) 393–396, doi:10.1038/nbt.1868.

[34] B. Hillmann, et al., SHOGUN: a modular, accurate and scalable framework for microbiome quantification, Bioinformatics 36 (2020) 4088–4090, doi:10.1093/bioinformatics/btaa277.

[35] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (2010) 1572–1573, doi:10.1093/bioinformatics/btq170.

[36] E. Cerami, et al., The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (2012) 401–404, doi:10.1158/2159-8290.Cd-12-0095.

[37] J. Gao, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal 6 (2013) pl1, doi:10.1126/scisignal.2004088.

[38] D. Aran, Z. Hu, A.J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape, Genome Biol. 18 (2017) 220, doi:10.1186/s13059-017-1349-1.

[39] B.T. Sherman, et al., DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic. Acids. Res. 50 (2022) W216–W221, doi:10.1093/nar/gkac194.

[40] A.J. Kombe Kombe, et al., Epidemiology and burden of human papillomavirus and related diseases, molecular pathogenesis, and vaccine evaluation, Front. Public Health 8 (2020) 552028, doi:10.3389/fpubh.2020.552028.

[41] N. Muñoz, et al., Epidemiologic classification of human papillomavirus types associated with cervical cancer, N. Engl. J. Med. 348 (2003) 518–527, doi:10.1056/NEJMoa021641.

[42] A.B. Ryerson, et al., Annual Report to the Nation on the Status of Cancer, 1975-2012, featuring the increasing incidence of liver cancer, Cancer 122 (2016) 1312–1337, doi:10.1002/cncr.29936.

[43] W.A. Flavahan, E. Gaskell, B.E. Bernstein, Epigenetic plasticity and the hallmarks of cancer, Science (New York, N.Y.) 357 (2017), doi:10.1126/science.aal2380.

[44] X. Zhang, J. Liu, X. Cao, Metabolic control of T-cell immunity via epigenetic mechanisms, Cell. Mol. Immunol. 15 (2018) 203–205, doi:10.1038/cmi.2017.115.

[45] Y. Yu, et al., RBBP8/CtIP suppresses P21 expression by interacting with CtBP and BRCA1 in gastric cancer, Oncogene 39 (2020) 1273–1289, doi:10.1038/s41388-019-1060-7.

[46] Z. Cai, B. Tang, L. Chen, W. Lei, Mast cell marker gene signature in head and neck squamous cell carcinoma, BMC Cancer 22 (2022) 577, doi:10.1186/s12885-022-09673-3.

[47] N. Hosoya, K. Miyagawa, Synaptonemal complex proteins modulate the level of genome integrity in cancers, Cancer Sci. 112 (2021) 989–996, doi:10.1111/cas.14791.

[48] Z. Li, et al., Discovery and validation of novel biomarkers for detection of cervical cancer, Cancer Med. 10 (2021) 2063–2074, doi:10.1002/cam4.3799.

[49] A.M. Espinosa, et al., Mitosis is a source of potential markers for screening and survival and therapeutic targets in cervical cancer, PLoS One 8 (2013) e55975, doi:10.1371/journal.pone.0055975.

[50] L. Masterson, et al., Deregulation of SYCP2 predicts early stage human papillomavirus-positive oropharyngeal carcinoma: a prospective whole transcriptome analysis, Cancer Sci. 106 (2015) 1568–1575, doi:10.1111/cas.12809.

[51] A. Berglund, et al., Characterization of epigenomic alterations in HPV16+ head and neck squamous cell carcinomas, Cancer Epidemiol. Biomarkers Prev. 31 (2022) 858–869, doi:10.1158/1055-9965.EPI-21-0922.

[52] J. Zhou, et al., Hsa_circ_0001666 suppresses the progression of colorectal cancer through the miR-576-5p/PCDH10 axis, Clin. Transl. Med. 11 (2021) e565, doi:10.1002/ctm2.565.

[53] G. Narayan, et al., Protocadherin PCDH10, involved in tumor progression, is a frequent and early target of promoter hypermethylation in cervical cancer, Genes Chromosomes Cancer 48 (2009) 983–992, doi:10.1002/gcc.20703.

[54] K.H. Wang, H.W. Liu, S.R. Lin, D.C. Ding, T.Y. Chu, Field methylation silencing of the protocadherin 10 gene in cervical carcinogenesis as a potential specific diagnostic test from cervical scrapings, Cancer Sci. 100 (2009) 2175–2180, doi:10.1111/j.1349-7006.2009.01285.x.

[55] A.G. Evstafieva, I.E. Kovaleva, M.S. Shoshinova, A.V. Budanov, P.M. Chumakov, Implication of KRT16, FAM129A and HKDC1 genes as ATF4 regulated components of the integrated stress response, PLoS One 13 (2018) e0191107, doi:10.1371/journal.pone.0191107.

[56] L. Dong, et al., Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma, Cancer Cell 40 (2022) 70-87 e15, doi:10.1016/j.ccell.2021.12.006.

[57] R. Mastrantonio, H. You, L. Tamagnone, Semaphorins as emerging clinical biomarkers and therapeutic targets in cancer, Theranostics 11 (2021) 3262–3277, doi:10.7150/thno.54023.

[58] T. Worzfeld, S. Offermanns, Semaphorins and plexins as therapeutic targets, Nat. Rev. Drug Discov. 13 (2014) 603–621, doi:10.1038/nrd4337.

[59] A. Casazza, et al., Sema3E-Plexin D1 signaling drives human cancer cell invasiveness and metastatic spreading in mice, J. Clin. Invest. 120 (2010) 2684–2698, doi:10.1172/JCI42118.

[60] J. Luchino, et al., Semaphorin 3E suppresses tumor cell death triggered by the plexin D1 dependence receptor in metastatic breast cancers, Cancer Cell 24 (2013) 673–685, doi:10.1016/j.ccr.2013.09.010.

[61] C.G. Buffie, et al., Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile, Nature 517 (2015) 205–208, doi:10.1038/nature13828.

[62] A. Sivan, et al., Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy, Science 350 (2015) 1084–1089, doi:10.1126/science.aac4255.

[63] G. Yu, et al., Characterizing human lung tissue microbiota and its relationship to epidemiological and clinical features, Genome Biol. 17 (2016) 163, doi:10.1186/s13059-016-1021-1.

[64] K.L. Greathouse, et al., Interaction between the microbiome and TP53 in human lung cancer, Genome Biol. 19 (2018) 123, doi:10.1186/s13059-018-1501-6.

[65] D.N. Frank, et al., A dysbiotic microbiome promotes head and neck squamous cell carcinoma, Oncogene 41 (2022) 1269–1280, doi:10.1038/s41388-021-02137-1.

[66] X. Chen, et al., Gene expression profile analysis of human epidermal keratinocytes expressing human papillomavirus type 8 E7, Pathol. Oncol. Res. 28 (2022), doi:10.3389/pore.2022.1610176.