universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Decoding the Molecular Landscape of HPV-Associated Cancers: Integrative Analysis of HPV Transcriptome Profiling using RNASeq"

verfasst von / submitted by

Ella Kathleen Cassidy BSc (Hons)

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2024 / Vienna 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Assoz. Prof. Dr. Oleg Simakov

Acknowledgements

During the process of the academic journey of this thesis I received a lot of support. I would like to sincerely thank Dr. rer. nat. Stephan Bernhart (Berni ☺), who was always there for any questions, regardless how silly. Without his guidance and patience, this thesis would not have been possible. I also want to express my gratitude to my supervisor Dr. Oleg Simakov who provided me with encouragement and hours' worth of zoom calls for guidance along the way.

I would like to thank my mum for the late night calls of encouragement from across the pond, as well as my uncles, Connie, John, Tadhg, and Pat. Each wonderful in their own ways, I would not have started this journey without them. Thank you for the endless love and support.

Lastly, I want to thank the group of the IZBI at Leipzig University, who made this process not only bearable, but fun. I cannot imagine having better people around me during this. Especially Bruno Schmidt, for always being patient and supportive.

Table of Contents

1 Abstract

Focusing on late-stage squamous cell carcinomas associated with Human Papillomavirus (HPV), this thesis utilizes RNAseq data for an in-depth exploration of the molecular landscape. The analysis uncovers intricate interactions involving PPAR signalling, HPV gene expression, and other key pathways that influence treatment responses. The findings suggest an upregulation of PPAR, coupled with a downregulation of HPV genes, may contribute to less favourable treatment outcomes. Despite HPV's well-established carcinogenicity, its presence generally correlates with improved treatment responses. The identified interplay between PPAR and HPV genes gains significance, especially within the context of incorporating the PPAR pathway into models predicting outcomes in cervical cancer. This discovery prompts further exploration of potential synergies between PPAR signalling and the expression profiles of HPV-associated genes. It illuminates a complex interplay in the molecular framework of cervical cancer prediction models, presenting a compelling avenue for deeper investigation into the intricate relationships shaping treatment responses in HPV-associated cancers. Additionally, this study highlights the overexpression of cytochrome P450 and discernible upregulation of the PI3K-Akt signalling pathway in specific clusters. This underscores the multifaceted nature of molecular alterations in HPV-induced cancers, and emphasizes the need for a comprehensive understanding to inform targeted therapeutic strategies.

Diese Arbeit konzentriert sich auf Plattenepithelkarzinome im Spätstadium, welche mit dem Humanen Papillomavirus (HPV) assoziiert sind und nutzt RNAseq-Daten für eine Untersuchung der daraus resultierenden molekularen Landschaft. Die Analyse deckt komplexe Wechselwirkungen zwischen PPAR-Signalen, HPV-Genexpression und anderen wichtigen Signalwegen auf, welche das Ansprechen auf eine medizinische Behandlung beeinflussen. Die in dieser Arbeit präsentierten Ergebnisse deuten darauf hin, dass eine Hochregulierung von PPAR in Verbindung mit einer Herunterregulierung von HPV-Genen zu schlechteren Behandlungsergebnissen beitragen kann. Obwohl HPV nachweislich karzinogen ist, korreliert sein Vorhandensein im Allgemeinen mit einem besseren Ansprechen auf eine Behandlung. Das festgestellte Zusammenspiel zwischen PPAR- und HPV-Genen gewinnt insbesondere im Zusammenhang mit der Einbeziehung des PPAR-Signalwegs in Modellen zur Vorhersage der Behandlungsergebnisse bei Gebärmutterhalskrebs an Bedeutung. Diese Entdeckung veranlasst weitere Erforschung potenzieller Synergien zwischen PPAR-Signalen und den Expressionsprofilen HPV-assoziierter Gene. Sie beleuchtet ein komplexes Zusammenspiel im molekularen Rahmen von Vorhersagemodellen für Gebärmutterhalskrebs und zeigt einen vielversprechenden Weg für eine intensivere Untersuchung der komplizierten Beziehungen auf, welche das Ansprechen auf die Behandlung von HPV-assoziierten Krebsarten bestimmen. Darüber hinaus hebt diese Studie die Überexpression von Cytochrom P450 und die erkennbare Hochregulierung des PI3K-Akt-Signalweges in bestimmten Patientengruppen hervor. Dies unterstreicht die Vielschichtigkeit der molekularen Veränderungen die mit HPV-induzierten Krebserkrankungen einhergehen und verdeutlicht die Notwendigkeit eines

umfassenden Verständnisses dieser Dynamiken um gezielte therapeutische Strategien zu entwickeln.

List of Abbreviations

HPV; **H**uman **P**apilloma **V**irus

HR-HPV; **H**igh **R**isk-**H**uman **P**apilloma **V**irus

SCC; **S**quamous **C**ell **C**arcinoma

ICI; **I**mmune **C**heckpoint **I**nhibitors

CIGAR; **C**ompact **I**diosyncratic **G**apped **A**lignment **R**eport

DGE; **D**ifferential **G**ene **E**xpression

SNP; **S**ingle **N**ucleotide **P**olymorphism

PPAR; **P**eroxisome **P**roliferator-**A**ctivated **R**eceptor

PCR; **P**olymerase **C**hain **R**eaction

NGS; **N**ext **G**eneration **S**equencing

2 Introduction

Human papillomaviruses (HPVs) constitute a diverse family of double-stranded DNA viruses characterized by an array of up to 222 distinct types, with continuous discoveries of novel variants [1]. Among these types, approximately 12 have been identified as oncogenic, earning classification as high-risk (HR) HPV genotypes. The persistence of infections with these oncogenic HR-HPV types is a well-established pivotal factor in the development of various cancers, and it has been estimated that HPV is a causative agent of approximately 4.5% of all cancers affecting humans [2]. HR-HPV types, including HPV16 and HPV18, have been prominently associated with the aetiology of multiple malignancies [3]. These include anal, penile, vaginal, and vulvar carcinomas, acting as primary causative agents in the majority of cases [4].

HR-HPVs possess compact circular, double-stranded DNA genomes (~8 kb) encoding eight genes, with oncogenes E6 and E7 exhibiting transformative properties. These proteins play diverse roles, including transmembrane signalling, cell cycle regulation, cell line transformation, primary cell line immortalization, and chromosomal stability control. Crucially, the viral E6 and E7 oncoproteins are essential for malignant conversion. Their interaction with tumour suppressors p53 and probe is proposed as a mechanism for inducing tumours [5]. These crucial oncogenes can be seen in Figure 1 below.

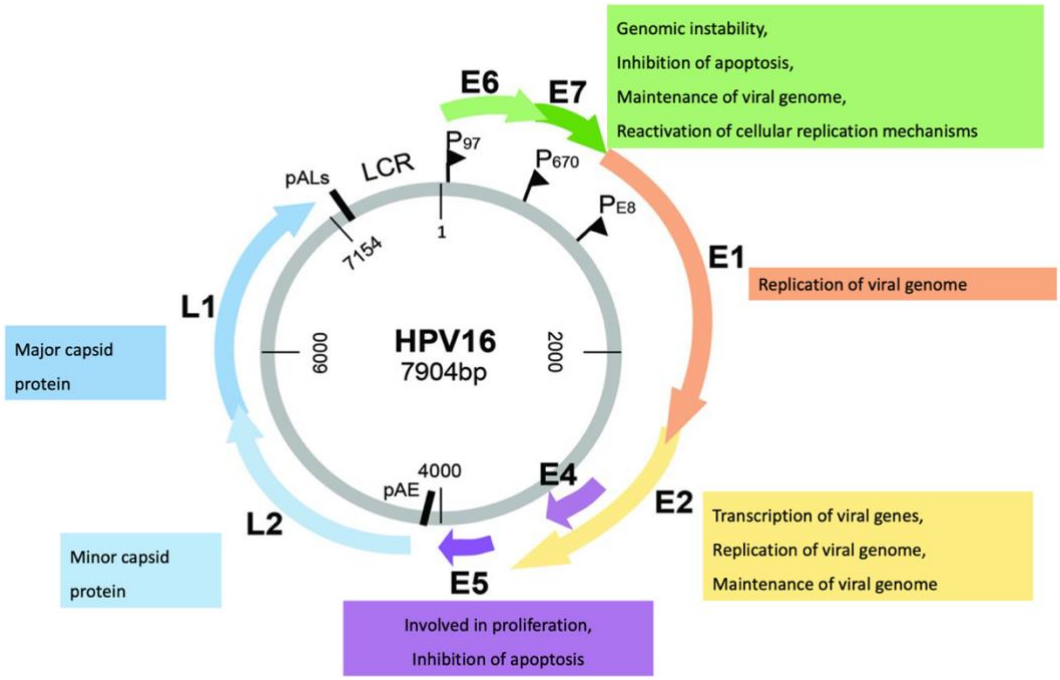


Figure 1. HPV16 life cycle. Viral life cycle contains 8 genes. E depicting the early genes (E6, E7, E1, E2, E4, and E5), L depicting the late genes (L1 and L2). Adapted from [6].

There are six genes located in the “early” region of the HPV genome and two in the “late”. Early proteins are mostly characterized by their regulatory functions, which are involved in processes such as apoptosis control, cell cycle, genome replication and transcription, and immune modulation. Through the infectious cycle most early proteins are expressed, although E4 is the first protein expressed during the late infection stages. The late proteins produce the capsid necessary for the survival, spread and transmission of the virus in the environment [7]. E5 serves as an accessory oncoprotein, providing support to the process of oncogenesis without being deemed essential for it [2]. E1, acting as a helicase, and E2, functioning as a DNA binding protein, constitute the fundamental viral proteins essential for replication and genome maintenance. In addition to tethering virus genomes to host chromosomes for distribution to daughter cells, E2 plays a role in downregulating E6 and E7 at specific stages during the viral life cycle. This is proposed to contribute to evading immune detection. E4 is believed to contribute to genome amplification and virion release, being among the most highly expressed open reading frames (ORFs). L1 constitutes the major structural protein, while L2 serves as the minor structural protein in the viral icosahedral capsid [2].

Traditional HPV detection methods have historically relied on broad-spectrum signal amplification assays, coupled with rapid high-throughput target-amplification techniques, such as polymerase chain reaction (PCR). However, the advent of next-generation sequencing (NGS) in the contemporary molecular research landscape has revolutionized our approach in terms of viral/bacterial detection and characterization [8]. NGS has the potential not only to discern the presence or absence of HPV infection, but also offers intricate insights into the HPV genome and infection dynamics. This provides a more comprehensive understanding of these viral agents [2]. RNAseq allows for the quantification of all expressed genes in a sample. Thereby capturing both known and novel transcripts, enabling a more detailed exploration of alternative splicing, gene fusion events, and the discovery of non-coding RNAs. This holistic approach provides a more comprehensive understanding of gene expression dynamics and regulatory networks, making it a powerful tool for unravelling complex biological processes compared to the targeted nature of traditional PCR. In the context of late-stage SCCs, leveraging RNAseq data becomes imperative for a holistic exploration of the molecular intricacies driving cancer progression. The multi-origin nature of the late-stage SCCs, encompassing diverse anatomical sites, demands an integrative approach to unravel the specific genetic, expression, and regulatory characteristics associated with HPV infection [9].

By elucidating the complex interplay between HR-HPVs and the host transcriptome, this thesis aims to contribute to our understanding of the molecular underpinnings of SCCs. The subsequent sections delve into the specific methodologies employed, the rationale behind the chosen analytical approaches, and the potential implications of the findings in the broader context of cancer research and precision medicine.

3 Materials and Methods

3.1 Data Acquisition

A dataset of 112 patients with late stage SCCs were provided from the PEVO-project, a breakdown of the number of patients per origin for each can be found in Figure 2 [10]. The data for this thesis was provided by a European funded basket trial – the PEVO project. PEVO stands as a European open-label, non-randomized, multi-center phase II basket trial focused on exploring the synergy between immunotherapy and an epidrug for the treatment of recurrent and/or metastatic SCCs. Enrolling patients with SCCs originating from diverse locations, including the lung, head and neck, cervix, anus, vulva, or penis, the trial specifically details the combination of pembrolizumab and vorinostat.

The trial's primary objective is to investigate the efficacy of this combination therapy through the systematic collection of sequential blood and tumour samples [10]. The aim of PEVO was to collect blood and tissue samples at key stages including baseline, under treatment, and at progression. Their sampling approach aims to uncover predictive biomarkers for both treatment response and resistance, allowing for the observation of treatment-related modifications. The overarching goal is to distinguish between short and long responders and unravel the intricate mechanisms underlying both resistance and sensitivity. PEVO adopts a basket trial framework, centering on a single histological subtype. This innovative approach serves as a bridge between traditional organ-specific clinical trials and the contemporary trend of treating cancer based on molecular alterations, irrespective of the organ of origin. By focusing on shared molecular similarities within the histological subtype across various anatomical locations, the trial tailors treatment strategies based on the histology itself rather than the primary location of the SCC. Despite the widespread use of immune checkpoint inhibitors (ICIs) in the first-line recurrent and/or metastatic setting for SCC, it is acknowledged that less than half of patients currently benefit from immunotherapy [10].

RNAseq and PCR genotyping was carried out on each patient. The fastq files from the RNAseq analysis containing reads that did not align to the human reference were utilized in the subsequent analysis. Due to the sensitive nature of the data in this thesis, authorization for access to the complete sequencing files was not granted. Therefore, it is important to note that only access to the unmapped reads was permitted in the following sections.

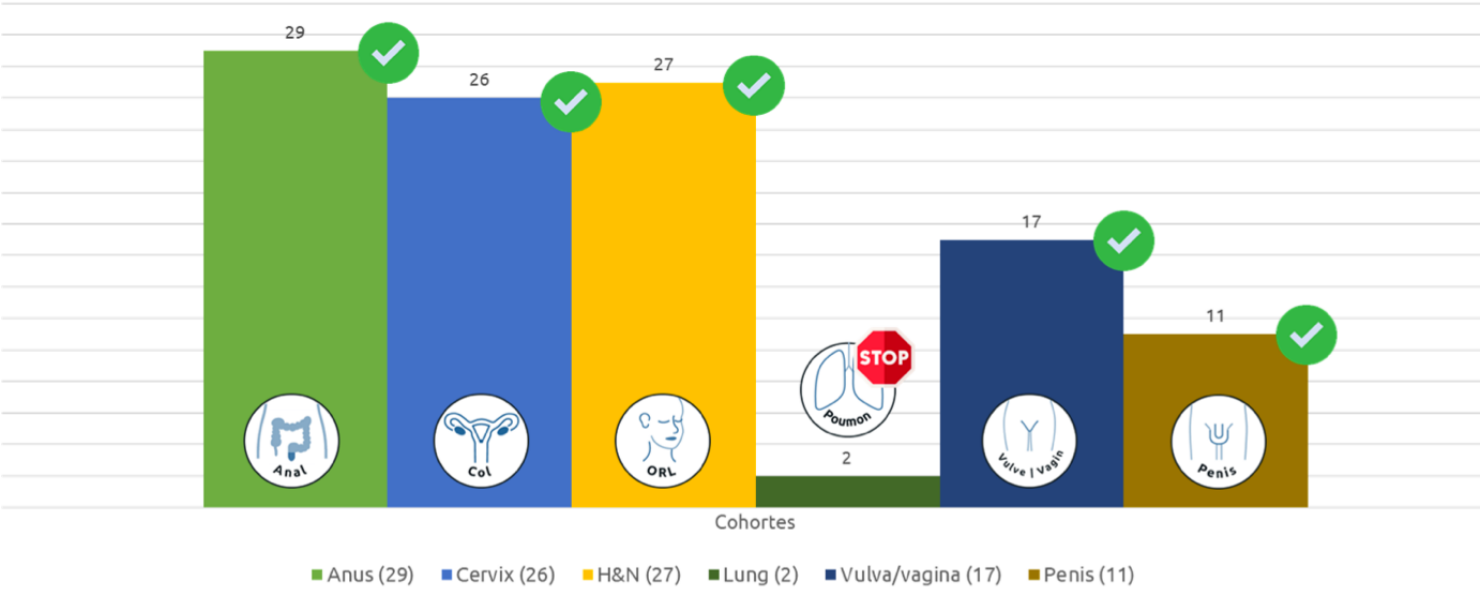


Figure 2. Bar plot of late stage squamous cell carcinoma samples by anatomical origin. Anus exhibits the highest representation (29), followed by Cervical (26), Head and Neck (27), Lung (2), Vulva/Vagina (17), and Penis (11) [10].

Virus-Host Database [11] was accessed on March 9th 2023 to acquire all viral genomes that are found to infect humans. A list of RefSeq accessions were downloaded from there and queried against the NCBI database [12] to obtain all genomes matched to the RefSeq list of accessions.

In total 3165 entries were downloaded from Virus-Host Database. Once a list of unique, non-duplicated entries was made this number was reduced to 1809. It was decided that no uniport entries should be included as they did not contain full genomes and had many duplicates. This reduced the number of viral genomes to 1408.

The PapillomaVirus Episteme (PaVE) [13][14] was accessed on March 21st 2023 to obtain multiple HR-HPVs that were not obtained using the Virus-Host Database. PaVE is designed with the goal to provide a system to organize and store curated genome information and tools focused on papillomaviruses for the scientific community. It encompasses a comprehensive database and web applications, facilitating the storage, annotation, analysis, and sharing of papillomavirus-related information [14].

Bacteria frequently play a role in human diseases, and for this study, the National Microbial Pathogen Database Resource (NMPDR) [15] was utilized to retrieve all pathogenic genomes (accessed on March 20th 2023). The initial selection comprised approximately 200,000 microbial genomes known to infect humans. To refine the dataset, filters were employed. Initially, only reference and representative genomes were considered, resulting in a total of 4 reference genomes and 377 representative genomes. Subsequently, only genomes

characterized by good quality assemblies, an option provided by the database, were included in the analysis.

All bacterial and viral species were integrated into a single fasta file and STAR (Spliced Transcripts Alignment to a Reference, v. 2.7.10b) was used for indexing [16]. STAR is a rapid RNAseq read mapper that incorporates support for detecting splice-junctions and fusion reads. The alignment process in STAR involves identifying Maximal Mappable Prefix (MMP) hits between reads (or read pairs) and the genome, utilizing a suffix array index. This method enables different segments of a read to be mapped to various genomic positions, indicative of splicing events or RNA fusions [16].

3.2 Alignment of Human Unmapped Reads

Entropy is a measure of the randomness or disorder within a sequence. In the context of sequence masking, low-entropy regions are often associated with repetitive or low-complexity sequences. Following the initial mapping steps using STAR (described above in section 3.1), high read counts for species such as tick-borne encephalitis appeared. Several bam files were investigated, and it was found that many of the reads were mapping to repetitive regions with polyA, polyT, polyC or GCGCGCGC repeats. BBMask was utilized to mask these regions.

BBMask, as part of the BBMap package (v. 39.6), was employed to mask low-complexity regions in the fasta file containing the integrated bacterial and viral species [17]. This tool analyses individual reads in a FASTA file by breaking them into overlapping k-mers and assessing their content. It then applies predefined criteria, such as identifying low-complexity regions or specific sequences, to independently mask or filter segments of each read. This tool is designed to mask sequences, particularly to counteract false-positive matches in highly conserved or low-complexity regions of genomes. BBMask offers three optional types of masking: low-entropy (complexity), tandem-repeated kmers, and Sam-file coverage. In this case, the low-entropy masking options were used and set to 0.7. This parameter controls the sensitivity of the masking process based on the entropy of the sequence regions. By setting entropy=0.7, it is instructing BBMask to mask regions of the input sequences where the entropy falls below the specified threshold of 0.7. Adjusting the entropy threshold allows you to control the trade-off between sensitivity and specificity in masking. Higher entropy thresholds result in more stringent masking, potentially reducing false positives but might also increase the chance of accidentally masking non-repetitive low complexity regions of the genome examined. Lower thresholds increase sensitivity but may lead to masking regions with slightly higher complexity [17].

The updated masked concatenated file containing all bacterial and viral species was indexed using STAR. Alignment of the BAM files to the concatenated reference was performed. To enhance alignment accuracy, adjustments were made to the default mismatch allowance. Some regions exhibited approximately seven mismatches, whereas HPV sequences typically showed only one or two at most. Consequently, the mismatch threshold was tightened to 80% (0.2). This combined approach, involving complexity masking and a refined mismatch threshold, effectively filtered out reads that originally mapped to species, such as tick-borne encephalitis, but constituted repetitive regions (such as polyA, polyT, polyC), ensuring the retention of HPV-specific reads.

3.3 Detection of HPV Strains using RNAseq

An additional bash script was written to analyse BAM files containing sequencing data and count the occurrences of HPV strains within these files. It generates a concise summary in the form of a .csv table listing the top three most frequently detected HPV strains in each BAM file. The script sifts through the input BAM files, extracts HPV-related lines, and then determines their HPV types and their respective counts. The resulting .csv file includes three columns: 'BAM_File' (the source BAM file), 'HPV_Type' (the identified HPV strain), and 'Count' (the read counts of the strain derived from the bam file). This script offers quick and effective means of summarizing the prevalence of HPV strains within data. Reads with less than 20 counts were excluded due to their potential for ambiguity and unreliable interpretation. This threshold helps mitigate the impact of random noise, sequencing errors, or other sources of uncertainty, ensuring that the retained reads are more likely to represent biologically relevant information. The resulting refined list was then cross-referenced with the provided PCR genotyping results for further analysis.

3.4 Alignment Using HPV16

Several HPV strains were found among the patients, although there was an abundance of matched HPV16 reads. Subsequently, the focus of further analysis was shifted towards HPV16. The annotation file corresponding to HPV16 was obtained from the PaVE database [13]. The alignment process was reinitiated, this time solely utilizing the HPV16 index file as the reference sequence. This approach aimed to facilitate a more precise alignment for identifying HPV16 reads, allowing for deeper insights into its genomic characteristics. GenBank accession number obtained from HPV16 on the PaVE database (K02718.1).

A custom Bash script was developed to automate the alignment and post-processing of the sequencing data, since there were 112 fastq files to process. It checks if HPV16 reads are present in the bam file, and if so, it copies the files to a directory where it sorts the generated BAM files are sorted and indexed for efficient data retrieval. The script again utilizes the STAR

aligner to map sequencing reads to the HPV16 reference genome. It is designed to handle a batch of input fastq files, facilitating a streamlined and consistent approach to the alignment process.

3.5 Sublineage Identification and SNP Analysis

A list of sublineages were obtained from NCBI for HPV16 (Table 1 below) [18]. Each sublineage fasta file was concatenated with the reference HPV16 genome. Clustalo (v. 1.2.2) was employed to perform multiple sequencing analysis using the concatenated file. All Clustal software is based on initiating a pairwise alignment using the progressive alignment method, generating a guide tree based on sequence similarity scores using the UPGMA/Neighbour-joining method, and leveraging this guide tree to perform a multiple sequence alignment. This process ensures a comprehensive alignment by systematically aligning sequences based on their similarity, facilitating robust comparisons and analyses [19].

A custom Python script was developed to extract single nucleotide polymorphisms (SNPs) from multiple sequence alignment files representing different subtypes. The core functionality of the script involves parsing each alignment file, identifying the subtype based on the file name, and generating a Variant Call Format (VCF) file. The script reads the alignment file, processes sequence information, and extracts SNP details.

Species	Type	Lineage	Sublineage	Variant genome ID	GenBank accession no.	Other names
Alpha-9	HPV16	A	A1	Ref	K02718	European (E)
			A2	W0122	AF536179	European (E)
			A3	AS411	HQ644236	E
			A4	W0724	AF534061	Asian, E(As)
		B	B1	W0236	AF536180	African-1, Afr1a
			B2	Z109	HQ644298	African-1, Afr1b
		C		R460	AF472509	African-2, Afr2a
		D	D1	QV00512	HQ644257	North American (NA)1
			D2	QV15321	AY686579	Asian–American (AA)2
			D3	QV00995	AF402678	Asian–American (AA)1

Table 1. **HPV16 lineages (A, B, C, D) and their respective sublineages.** The classification is adapted from a referenced source and was employed in the analysis of genomic data presented here. Each lineage is further delineated into distinct sublineages, providing a comprehensive view of the genetic diversity within the HPV16 genome. Adapted from [18].

The resulting VCF file contains essential information, such as chromosome, position, reference allele, and alternate allele for each identified SNP. Tools such as Bcftools that can create VCF files could not be applied here, as they do not allow FASTA file as an input for VCF generation.

For the SNP analysis, Samtools (v. 1.9) and Bcftools (v. 1.18) were instrumental in creating VCF files representing the intersection of SNPs found in BAM files with HPV16 reads exceeding 500 and the reference HPV16 genome. This intersection was computed using Bcftools isec. This tool generates unions, intersections, and complements of VCF files. Depending on the chosen options, the tool can produce output containing records from one or more files that either contain, or lack, corresponding records at the same position in the other files. This command compared the VCF files derived from the previously outlined steps for the sublineage with the VCF file generated from the BAM files containing HPV reads. This approach ensures an examination of shared SNPs within the sublineage and the specific subset of BAM files, contributing to a comprehensive understanding of the genetic landscape of HPV16 in the dataset. IGV was used to examine SNPs along the genomes for cervical, vulval, and anal samples. Common SNPs were then investigated further.

3.6 Splice Site Identification Analysis

For a comprehensive analysis of splice sites, a customized program was developed using Python (v. 3.8.18). The development of a custom solution for splice site analysis was necessitated by the unique nature of the dataset under investigation. In this study, the analysis extends beyond typical human genomic data, as it involves the examination of human unmapped reads to detect HPV16-related splicing events. This program was designed to quantify the number of reads that provided evidence of splicing at boundaries of intronic regions within each BAM file, employing a stringent threshold 500 HPV16 reads within the bam files to ensure data integrity.

The foundation of this analysis was the utilization of CIGAR (Concise Idiosyncratic Gapped Alignment Report) strings present within the BAM files. By scrutinizing these CIGAR strings, the program can distinguish between reads that indicate splicing and those that did not at a certain position. Specifically, reads displaying the pattern "xMxNxM," (where 'x' signified the number of base pairs) were indicative of spliced reads. ‘M’ signifying a match, and ‘N’ signifying a non-match are indicative for splicing if the non-match (N) region covers an intron. Conversely, CIGAR patterns deviating from this convention while matching (M) intron-exon/exon-intron boundaries pointed to non-spliced reads.

Each BAM file was evaluated, where reads covering an intron while lacking an "N" in the middle were categorized as non-spliced reads. Conversely, reads exhibiting the specified

splicing pattern were classified as spliced reads. Furthermore, boundary reads, a category encompassing ambiguous reads that defied clear assignment as either spliced or unspliced, were computed.

Unspliced reads were calculated separately for both the 5' and 3' ends, and a similar approach was employed for boundary reads. For the 5' end, a boundary read was defined as commencing within 5 base pairs of the intron's start position (excluding the first base pair of the intron) and extending into the intronic region by a minimum of 20 base pairs. These identified 5' boundary reads were subsequently deducted from the non-spliced read count to refine the analysis. This procedure was recurrently executed for the 3' end. In this instance, a read was classified as a boundary read if it initiated at least 20 base pairs prior to the end of the intron and extended beyond the initial base pair of the exon, without trespassing more than 5 base pairs into the exon. The culmination of this process yielded a comprehensive assessment of each BAM file, providing valuable insights into the intricacies of HPV16-related splicing patterns. BBMap's "BBMask" function analyses individual reads in a FASTA file by breaking them into overlapping k-mers and assessing their content. It then applies predefined criteria, such as identifying low-complexity regions or specific sequences, to independently mask or filter segments of each read.

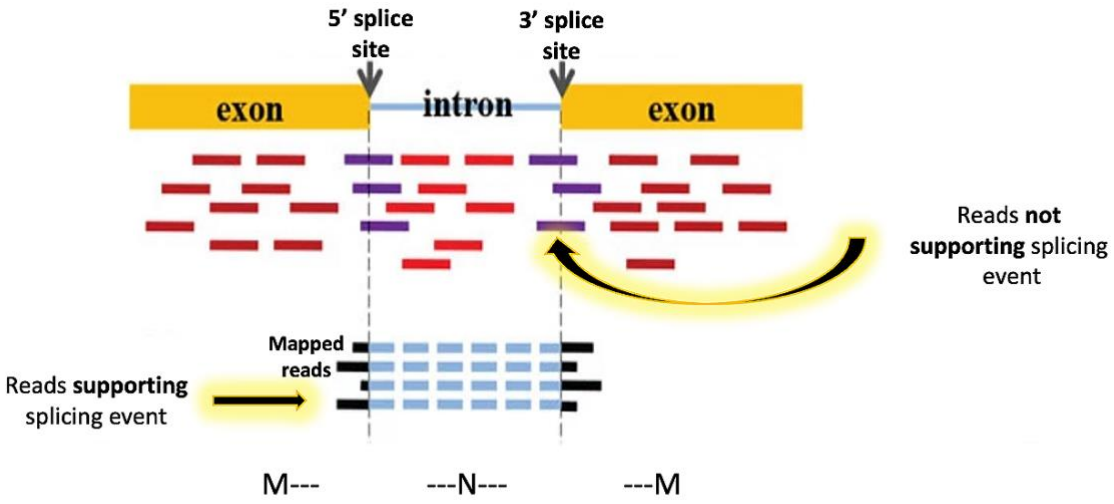


Figure 3. Calculating spliced and unspliced reads in the analysis of HPV16-related genomic data. The process centers around decoding CIGAR (Concise Idiosyncratic Gapped Alignment Report) strings within the BAM files. Adapted from [20].

3.7 Compositional Gene Usage and Splice Site Usage Analysis

3.7.1 K-means Clustering

In RNAseq data, library sizes can vary significantly between samples. Aitchison distances account for this library size variation by focusing on the proportions of gene expression rather than the absolute counts. This is crucial since, without such normalization, samples with large library sizes could dominate the analysis, and subtle compositional differences in samples with smaller library sizes might be overlooked [21]. Aitchison distances make it possible to compare samples with varying library sizes directly. This is because they measure the differences in gene expression proportions in a way that is scale-invariant. By using Aitchison distances, you can identify meaningful compositional differences between samples while mitigating the impact of library size variation [22]. The featureCounts output served as the foundation for calculating Aitchison distances. featureCounts stands as a remarkably efficient read summarization tool with broad utility. It tallies mapped reads for various genomic features, exons, promoters, and chromosomal locations [23]. Its versatility extends to counting reads from both RNAseq and genomic DNA-seq datasets.

For analysis within this thesis, gene-level read counts of HPV16 were extracted from the BAM files using featureCounts (v. 2.0.6)[23] and data pre-processing was performed. This involved handling zero values by introducing a small pseudo-count. The introduction of pseudo-counts was necessary for Aitchison distance calculations [24]. The pseudo-count utilized in this analysis was 1^{-6}. The Aitchison distances between samples were computed using the featureCounts table output between samples to quantify compositional differences. These compositional gene analysis steps were carried out using R Studio (v. 4.1.1).

To determine distinct gene expression patterns, k-means clustering was applied. The optimal number of clusters was determined using the elbow method (see Figure 5, section 4.3.1). As the distances were computed using individual featureCounts table, each sample is then assigned to a k-means cluster based on their compositional values. For the k-means clustering nstart option, which generates a number of random centroids and chooses the best for the algorithm, was set to 15.

Principal Component Analysis (PCA) was employed to visualize the overall gene expression variability across samples. The process encompassed the conversion of gene expression data, obtained as raw counts from featureCounts output, into compositional data. This transformation yielded proportions for each sample, ensuring a cumulative sum of 1. Utilizing the 'aDist' function from the compositions package in R Studio [25], a distance matrix was generated using Aitchison distance principles, reflecting the similarity or dissimilarity between samples. Subsequently, PCA was applied to this distance matrix to elucidate the relationships among samples. The decision to employ a distance matrix in our PCA analysis stems from the recognition that dissimilarity-based representations can uncover nuanced relationships

between samples that may not be apparent in the original feature space. Especially since the gene usage analysis is based on using Aitchison distances. By leveraging a distance matrix, our approach focuses on capturing the inherent structure encoded in the dissimilarities between samples. This is particularly advantageous when the relationships between samples are more meaningful in terms of dissimilarity, as opposed to the absolute values of the original features. The distance matrix is constructed using Aitchison distance computation. Additionally, explained variance ratios were calculated and visualized for each principal component, providing insight into the contribution of each component to the overall variance. The result can be seen in Figure 6, section 4.3.1. While well-separated clusters might indicate structure, the explained variance plot ensures that the chosen principal components not only separate clusters but also account for a substantial proportion of the overall variability in the dataset. It serves as a crucial metric for assessing the efficacy of the dimensionality reduction achieved by the PCA.

A fisher test was also utilized to test whether any particular cluster has a significant treatment response which could be linked to a gene usage pattern. Fishers exact test

The response to treatment was investigated within the five identified clusters. Fishers exact test was conducted iteratively, comparing each cluster against the combined response of the other clusters. For each iteration, a contingency table was conducted with two factors: 'Yes' or 'No' response to treatment. We repeated this process for all clusters, systematically replacing Cluster 1 with the subsequent clusters to evaluate if any cluster exhibited a significant overall response to treatment. Fisher's exact test, a statistical method for assessing the association between categorical variables, allowed us to determine whether specific clusters demonstrated a statistically significant response to the treatment compared to the rest of the clusters combined.

3.7.2 DIEGO Clustering

DIEGO (Detection of Differential Analysis using Aitchison’s Geometry, v. 0.1.2) is designed for the purpose of identifying splice junctions or exons with differential usage [26]. Central to its methodology is the employment of Aitchison's statistics in conjunction with a parameter-free statistical test, specifically the Wilcoxon test. The utilization of the Wilcoxon test necessitates a correction for multiple testing to achieve statistical significance. DIEGO operates on a tabular representation of splice junction usage or exon expression as its primary input [26]. This powerful tool employs hierarchical clustering to discern relationships among the samples based on their splice junction usage.

The data derived from the splice site analysis in section 3.2 served as input for DIEGO. The outcome is a dendrogram that provides an insightful depiction of the degree of relatedness among the samples concerning their splice junction patterns. The utility of this method was

extended to encompass gene usage compositions. Gene usage compositions refer to the proportional distribution of HPV16 genes within the genomic sequences. These compositions are analysed using Aitchison distances to quantitatively measure the relative abundance of specific genes, elucidating compositional variations across the dataset. Only the 8 HPV genes are included in this compositional analysis. Utilizing the featureCounts output table, a dendrogram was generated that elucidates the relationships among the samples based on their gene usage compositions. This comprehensive approach allows for a thorough exploration of the underlying structure within the data, shedding light on both splice junction usage and gene usage composition patterns. Figure 4 below depicts an over of the steps above.

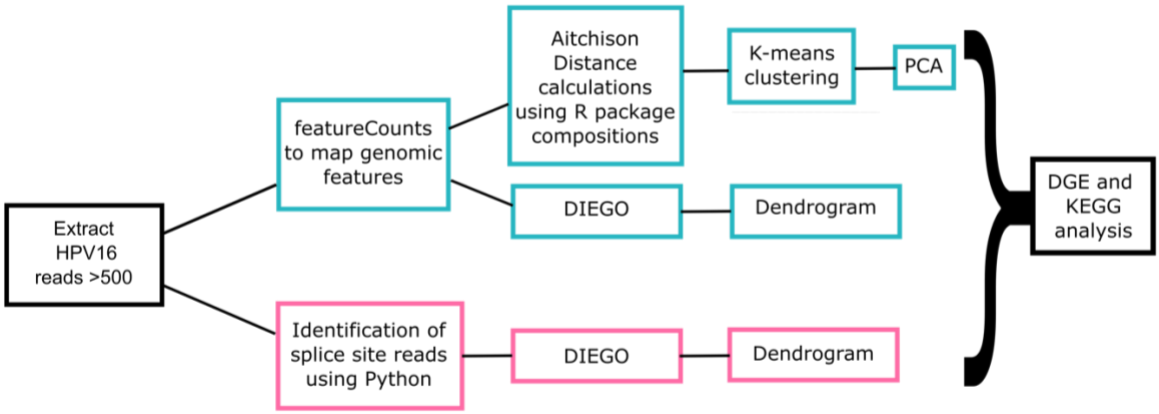


Figure 4. Compositional gene usage and splice site usage analysis workflow. In pink is the splice site usage steps, and in blue are the gene usage composition steps. Each outcome from the clustering method (DIEGO which is hierarchical, or K-means) is then passed on to the DGE and KEGG analysis.

3.8 Differential Gene Expression Analysis from Clustering

In the process of the clustering methods used (marked a, b, c below in Table 2), every sample was assigned to a unique cluster, facilitating the organization and interpretation of gene expression patterns. An overview of the clustering methods employed can be found in Table 2.

	a. Gene usage analysis	b. Gene usage analysis	c. Splice site usage
Clustering method	k-means	Hierarchical	Hierarchical
Program	R Studio (compositions)	DIEGO	DIEGO
Outcome	5 clusters	5 clusters	3 clusters

Table 2. Clustering methods applied to analyse gene expression patterns in the dataset. This table serves as a reference for the clustering methods employed, the corresponding programs utilized, and the outcomes achieved.

In order to gain insight into gene expression variations and patterns across different clusters of samples (within a single cluster method), DGE can be used. To achieve this, the cluster obtained were compared (only within the respective clustering methods a, b, or c) using DESeq2 (see Figure 5 for a schematic depiction).

For each cluster found within each clustering method (a, b and c), differential gene expression (DGE) analysis was performed by comparing each cluster with every other cluster, enabling us to explore nuanced gene expression variations and patterns, this is illustrated in Figure 3 below. DESeq2 was used for these steps. The DESeq2 package offers techniques for assessing differential expression through the application of negative binomial generalized linear models. The computations of dispersion and logarithmic fold changes integrate data-driven prior distributions for enhanced accuracy [27]. For this, access to the complete fastq files from the patients was necessary (not only unmapped reads), to compare and identify which genes are differentially expressed. For this purpose, as human-mapped reads could not be utilized for data protection reasons, Dr. Stephan Bernhart performed the DGE analysis using DESeq2. A list of the genes for further analysis was provided.

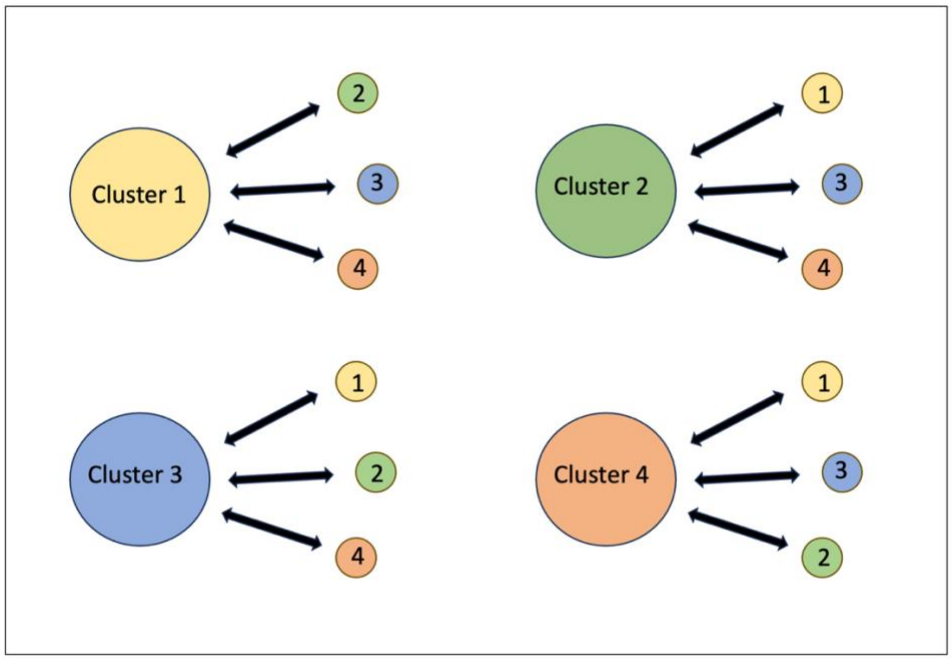


Figure 5. Pairwise comparison process within each clustering method. Each cluster, denoted by distinct identifiers, undergoes pairwise comparisons with every other cluster within the same clustering method.

In order to extract biologically relevant insights from the differentially expressed genes across clusters, the StringDB platform [28] was utilised, inputting each gene list generated from the comparative DGE analysis. STRING serves as a repository for both known and predicted protein-protein interactions, encompassing indirect (functional) and direct (physical) associations. These interactions arise from computational predictions, knowledge transfer across organisms, and the aggregation of interactions from various primary databases [28]. StringDB provided a comprehensive inventory of enriched pathways. An exploration of the KEGG (Kyoto Encyclopedia of Genes and Genomes)[29] pathways was undertaken to achieve a deeper understanding of the functional implications of the enriched pathways. KEGG stands as a database resource aimed at elucidating the overarching functions and utilities of biological systems, ranging from the cell and organism to ecosystems. It aims to achieve this by leveraging molecular-level information, particularly large-scale molecular datasets derived from genome sequencing [29].

It is important to note that the DESeq2 analysis was adjusted to account for potential confounding factors, particularly those arising from the diverse origins or tissue sources of the samples. This precaution ensures that the differential gene expression results remain robust and reliable in the face of such variables.

4 Results

4.1 HPV-RNAseq Detection vs. PCR Genotyping

The detection of HPV has been examined using two distinct approaches: RNAseq and PCR. The initial focus of this thesis was on determining the capability of RNAseq techniques to reliably identify the presence of HPV. Given the relative novelty of this application, the primary concern was to validate and assess the accuracy of HPV detection through RNAseq. Since the genotyping results were available along with the dataset, and it is a well-established method for detection of HPV, these can be cross referenced with the RNAseq results.

Analysis of HPV presence using both RNAseq and PCR techniques revealed remarkably consistent percentage matches across the majority of the samples. Notably, head and neck samples demonstrated a high matching percentage of 87%, although a significant proportion of results were negative for HPV, or HPV was not detected by PCR nor by RNAseq. These results can be seen in Table 3 below. The challenging nature of HPV detection in head and neck samples is attributed to certain complexities inherent in this anatomical region, such as tobacco and alcohol, or in general the heterogeneity of the tissue in the region [30]. The tumour microenvironment in the head and neck region can include various cell types and structures, making it challenging to isolate and identify HPV-specific transcripts amidst the complexity of the tissue. There was n=71 samples for genotyping data, and n=66 samples for the RNAseq available, as some patients did not have RNASeq data available. For the matching ratio calculation in Table 3, only patients which had samples for both genotyping and RNAseq were used (see column 3, Table 3).

Despite these challenges, the observed robust consistency in results underscores the reliable nature of RNAseq in HPV detection within the context of late-stage SCCs. The agreement in outcomes between these techniques instils confidence in their utility and accuracy for HPV detection in this specific cancer scenario.

Origin	Total number samples	Samples containing genotype and RNAseq data	Number of HPV+ samples	Number of HPV- samples	Number of matches	Matching ratio	Strains Detected
Anus	29	23	23	0	23	100.00%	HPV31,HPV16
Penis	11	6	3	3	5	83.33%	HPV-, HPV45, HPV16
Cervix	26	13	11	2	11	84.62%	HPV18, HPV35, HPV45, HPV16
Vulva	17	14	7	7	14	100.00%	HPV33, HPV-, HPV16
HeadAndNeck	27	15	4	11	13	86.67%	HPV-, HPV45, HPV16

Table 3. HPV detection results using RNAseq and PCR techniques across various sample origins. Total number of samples derived from [10], samples containing both data for the RNASeq and genotyping, number of HPV positive cases within that origin, number of matches (i.e. for each patient if there's the same result for genotyping and RNAseq, HPV- also counts as a match), the calculated matching ratio and the strains that were detected.

4.2 Sublineage Identification and SNP Analysis of HPV16

The identification of sublineages within the samples proved challenging, and it was not feasible to ascertain the specific sublineages present in each sample. Isolates belonging to the same HPV type are classified into variant lineages and sublineages based on distinctions in their complete genome nucleotide sequences, typically ranging from approximately 1% to 10% for variant lineages and 0.5% to 1% for sublineages [30,18]. Achieving accurate sublineage identification therefore requires high-quality coverage sequencing of their complete genome in order to discern the subtle nucleotide differences within this specified range. Common single nucleotide polymorphisms (SNPs) can then be investigated. Some sublineages exhibit distinctive signature SNPs [32], although the authors note that it may not be possible to definitively confirm the assignment to a specific sublineage based solely on these markers.

Upon inspection, it was observed that a subset of the samples exhibited a shared single nucleotide polymorphism (SNP) with the AF472509 sublineage. Notably, all instances of shared SNPs were confined to a singular genomic location, namely G7193T. Upon conducting a more detailed analysis, it became apparent that several common SNPs were shared among the samples. These shared SNPs are itemized in table 4 for reference.

Common SNP's found	Anus SNPs	Cervical/vulva SNPs
A4042G	C3684A	3409
T4228C	A3979C	T4114A
A5226C,G,T	T5041C	G4938A
A6434G		
G7521		
G7193		

Table 4. SNPs based on their occurrence in diverse sample origins. The table is segmented into three categories: firstly, SNPs common across all samples, reflecting shared genetic elements across all anatomical regions; secondly, SNPs exclusive to the anus origin, offering insights into region-specific genetic variations; and thirdly, SNPs specific to cervical and vulval origins, highlighting distinctive genetic features in these anatomical locations.

The C3409T SNP was particularly interesting as it has been linked to a persistent HPV16 infection [33]**.** Therefore, a fisher test [34] was conducted on patients which had data pertaining to whether they had a response to treatment. 20/49 of the samples (specifically containing HPV16 reads), have this SNP. There were 29 patients for penial, cervical and anal origins which had data about treatment response. Although this was limited to yes/no. These files were intersected. As not all these 29 samples which had data about treatment response were HPV16 positive. The result of the analysis was an odds Ratio = 0.28125, and a P-value = 0.1929. An odds ratio of 0.28125 suggests a negative association between the response to drug treatment and the presence of the SNP ('C3409T'). Although the p-value is not significant, likely due to lack of statistical power through the number of samples available. The non-significant p-value might be explained due to the lack of statistical power.

4.3 Compositional Gene and Splice Site Usage

4.3.1 Gene Usage Compositions using K-means

Among the 112 samples, 49 showed HPV16 read counts exceeding 500, prompting a focused analysis of their gene usage compositions. The majority of the remaining 63 samples had either no detectable HPV16, or 5 samples with HPV16 and <500 reads exhibited an average of 78 reads. In RNAseq data, discrepancies between low reads and HPV strain detection sometimes arise when cross-referenced with genotyping results. The genotype results provide

empirical support for the selection of this specific read count. This decision is rooted in the observation that samples with approximately 500 reads consistently yielded comparable outcomes during PCR validation. The recurrent concordance between computational analysis and genotype validated results underscores the biological significance associated with this threshold.

With these 49 samples, Aitchison distance [22] calculations were employed to generate a distance matrix of the data, and then construct an elbow plot (figure 6 below), helping determine the optimal number of clusters for k-means clustering required to effectively discern patterns within the dataset. The clusters remained stable with an nstart value of 15. From the elbow plot depicted in Figure 6 it was determined that 5 clusters were sufficient.

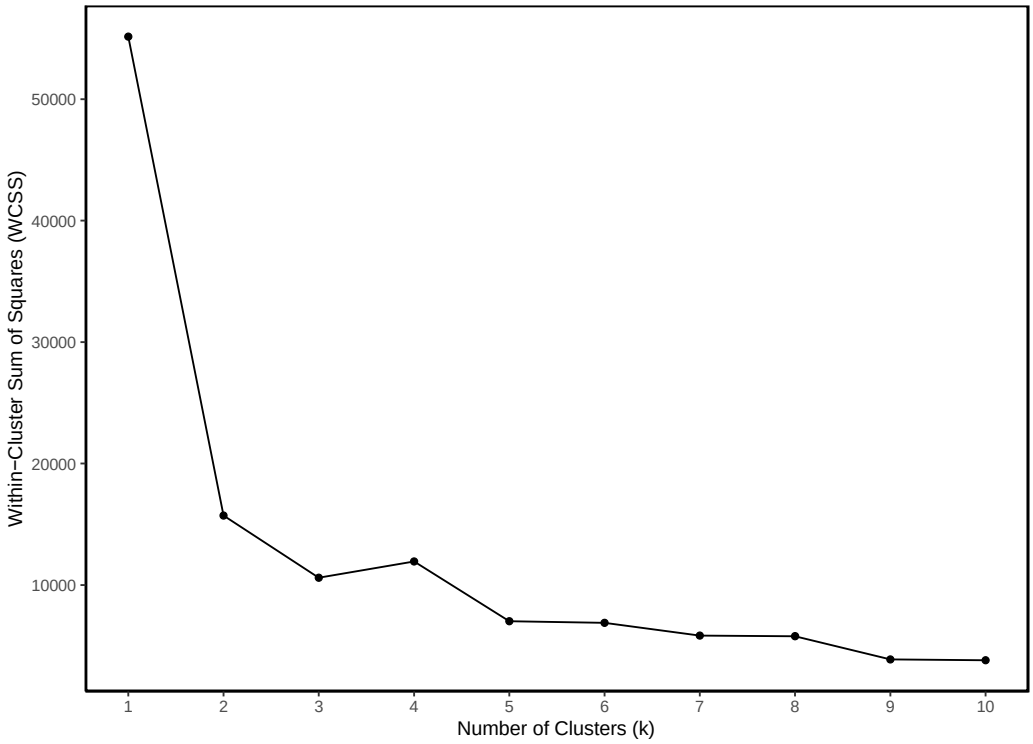


Figure 6. **Elbow plot to determine appropriate number of clusters for k-means analysis.** The number of clusters on the x-axis and the corresponding within-cluster sum of squares on the y-axis. Two distinct elbow regions can be seen on this plot.

The utilization of Principal Component Analysis (PCA) in our study serves to reinforce the robustness of the clustering results [35]. The well-separated clusters with no overlap strongly indicate that the data exhibits clear, distinct patterns that PCA can effectively uncover (see Figure 7. A). This finding underscores the validity of the clustering process and the meaningful distinctions between the data points within each cluster.

The high explained variance of around 90% for the first two principal components signifies that the chosen dissimilarity metric effectively summarizes the majority of the variability in the data illustrated in the Explained Variance by Principal Components plot (see Figure 7. B.). The Explained Variance plot not only provides insights into the cumulative explanatory power of principal components but also ensures that the retained components effectively capture meaningful information in the data.

This dissimilarity centric PCA provides a robust means of elucidating patterns based on sample relationships, contributing to a more comprehensive understanding of the underlying structure in our dataset. This high explanatory power in the early components suggests that the most critical features of the data are well-represented and that a reduced-dimensional representation of the data retains its essential characteristics. This emphasizes the quality of the clustering outcomes by showing that a concise set of principal components explains a substantial portion of the underlying data structure.

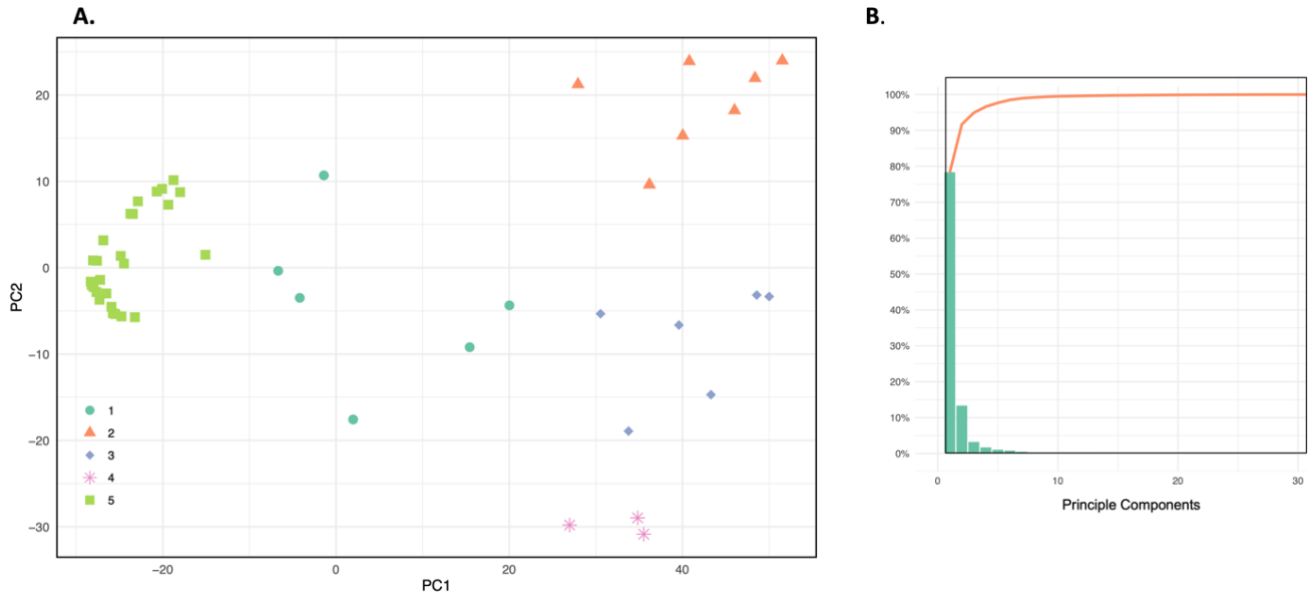
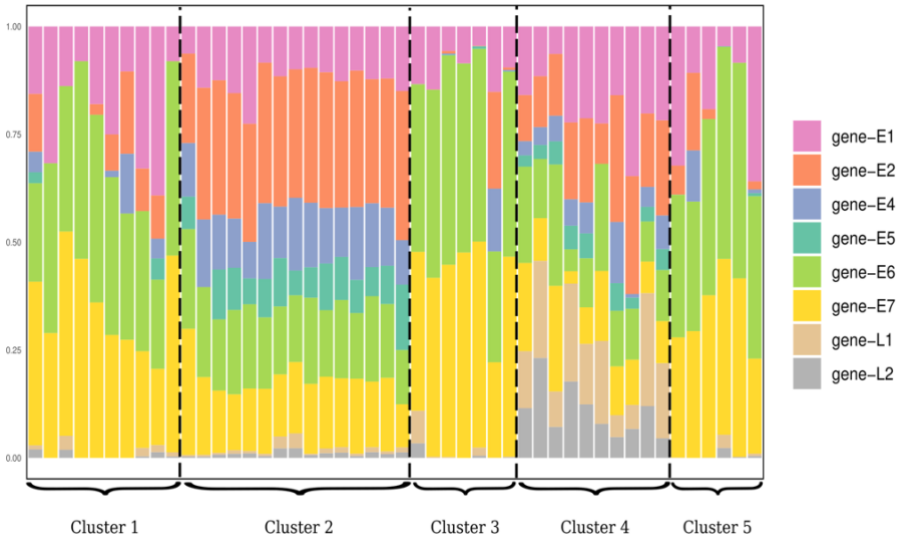


Figure 7. Principal component analysis (PCA) and explained variance plots using dissimilarity metrics. PCA with the left plot (A.), illustrating k-means clusters representing the datasets structure. On the right (B.), the Explained Variance by Principal Components graph demonstrates that PC1 and PC2 components collectively explain approximately 90% of the overall variance.

Given a clustering of samples by gene usage composition, one could expect that sample within each cluster follows a certain trend. This trend can be witnessed for the 5 k-means clusters that were computed in the process of this thesis, as can be seen in Figure 8, illustrating that each cluster exhibits a distinct gene usage. Especially for cluster 2, it is obvious that the samples represented within follow a specific pattern.

Across clusters, it becomes evident that there is a consistent and similar gene usage shared by a substantial portion of the clusters. This indicates a degree of commonality in gene usage patterns among various clusters, suggesting potential functional or regulatory relationships between these genes. The distinct gene usage within each cluster, coupled with the shared composition across clusters, offer a comprehensive perspective on the dataset, shedding light on both its diversity and commonalities regarding gene usage compositions. All clusters have some expression of the oncogenes E6 and E7, which are necessary for driving tumorigenesis [36]. Cluster 4 is the only cluster which has significant expression of late genes (L1, and L2). Late genes in viruses are typically associated with processes such as virion assembly, capsid formation, and release of viral particles [37]. Therefore, a cluster with higher expression of late genes may represent a subgroup of samples where the virus is in a more advanced stage of replication. Not all early genes (E1, E2, E4, E5), apart from the oncogenes (E6, E7), are expressed in each cluster [38]. Clusters 1, 2, 3, and 5 all have varying degrees of early gene expression. Cluster 2 shows a higher level of E4 expression, which could be indicative of progression. E4 is expressed before the late stage genes in the HPV lifecycle, serving functions related to the viral life cycle, such as genome amplification and the formation of viral particles. This early expression of E4 is consistent with its role in preparing the infected cells for later stages of the HPV lifecycle. Figure 8 depicting the gene usage bar plot and boxplots can be found below.

A.



B.

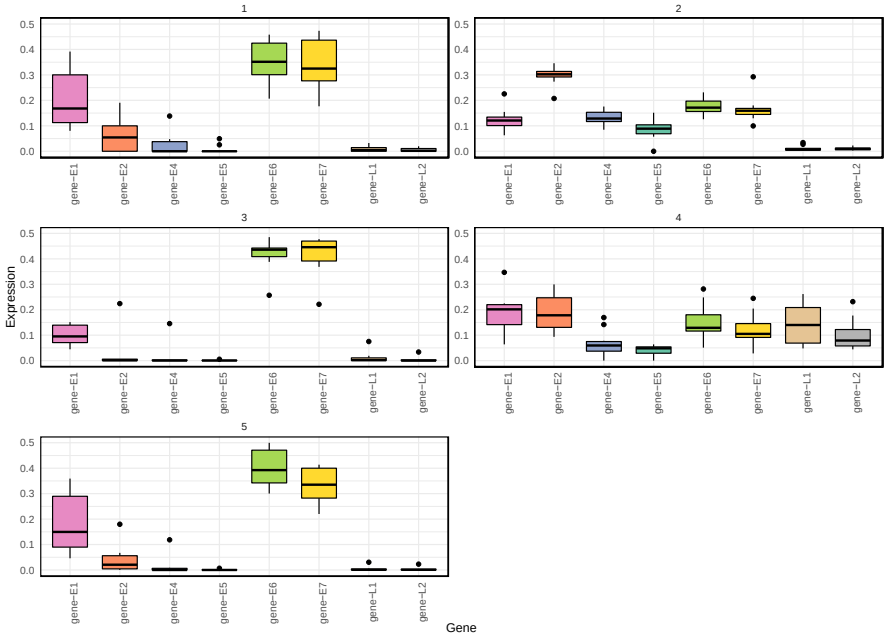


Figure 8. Barplot and boxplots depicting the gene usage compositions for HPV16 samples over 500 reads. The paired plots provide visualizations of gene usage compositions within distinct clusters. The bar plot on the top (A.) organizes genes by prevalence within each cluster, offering a Visual assessment of their unique compositions. The accompanying boxplots (B.) on the right further enhance the analysis by illustrating the distribution and characteristics of each cluster.

4.3.2 Treatment Response for Gene Usage using K-means Clustering Results

Since response to treatment data was available, the next step was to determine if any of the clusters have a significant response to treatment. For this, the Fisher exact test was used. Fisher's exact test assess the probability of obtaining a distribution of categorical data as extreme as the observed one, providing a statistical measure of the association between two categorical variables [34]. Since cluster 4 has higher expression of late genes, it could be hypothesized that samples from this cluster have a more progressed level of HPV16 infection, and thus a lowered treatment response. In total there are 49 samples in this k-means clustering analysis, but only 36 of those have response information available. Although there were no significant responses to treatment. The outcome of the Fisher test can be seen in Table 5 below. These gene usage pattern could potentially be linked to disease progression, but not to treatment response in this case.

	Odds ratio	P-value
Cluster 1	0.84	1
Cluster 2	2.57	0.43
Cluster 3	2.5	0.64
Cluster 4	0.49	0.65
Cluster 5	0.18	0.22

Table 5. Fisher exact test outcomes using k-means clustering for cluster assignment and response information. Odds ratio results in the first column, and p-value in the second. No significant p-value or odds ratios were observed.

4.3.3 Splice Site and Gene Usage Compositions Using DIEGO

The DIEGO output reveals a discernible clustering pattern, as illustrated in Figure 9. As mentioned in the methods above, DIEGO is designed for the purpose of identifying splice junctions or exons with differential usage [26]. DIEGO calculates the average Aitchison's distances for all genes within this highly variable set, generating a comprehensive distance matrix between samples. The key step involves subjecting this matrix to hierarchical agglomerative clustering, specifically utilizing the average linkage method. In clustering mode, DIEGO leverages this methodology to construct a dendrogram, providing users with a visual representation that facilitates the identification of both similar and dissimilar splicing patterns among samples. By focusing on the interplay of variance, Aitchison's distances, and the average linkage method, DIEGO's clustering approach offers a nuanced understanding of gene expression relationships in a hierarchical manner [26]. The output from this produces the splice site usage data. Notably, the cluster marked in blue for the splice site usage stands out as the only cluster lacking a clearly defined underlying pattern (see fig 7.A.). These samples

could represent outliers, or noisy data points. The same procedure was repeated for the gene usage, utilizing the featureCounts output for DIEGO.

Upon examining gene usage composition clustering of DIEGO, this tool identified 5 clusters that capture the underlying structure. Notably, the same number of clusters was determined to be descriptive using the elbow plot method in R Studio (see Figure, 7.B). Information was added pertaining to origin and response, as it becomes important in downstream analysis. DIEGO identified three clusters for the splice site usage analysis.

To ensure the robustness of the clustering results, a cluster validation step was performed by rerunning the analysis and confirming that the samples remained consistently assigned to their designated clusters. The validation process demonstrated stability across multiple iterations. While hierarchical clustering is generally deterministic regarding its rules for merging clusters, the outcomes may exhibit variability, even with identical parameters, due to the method's sensitivity to input data and the inherent randomness of the algorithm. This sensitivity becomes particularly apparent in cases of ties or equal distances between data points. Although the hierarchical agglomerative clustering algorithm employed by DIEGO is deterministic concerning its rules for merging clusters, the introduction of variability can occur when dealing with equal distances between clusters or samples. Factors contributing to result variations under the same parameters include ties in distance measurements, where the algorithm may deterministically break ties based on indices or labels, and data perturbation.

In order to gain a deeper understanding of the biological implications of the clusters DIEGO identified, a KEGG pathway analysis was performed. This analysis aims to unveil the enriched biological pathways associated with each cluster, providing insights into the potential functional distinctions among them.

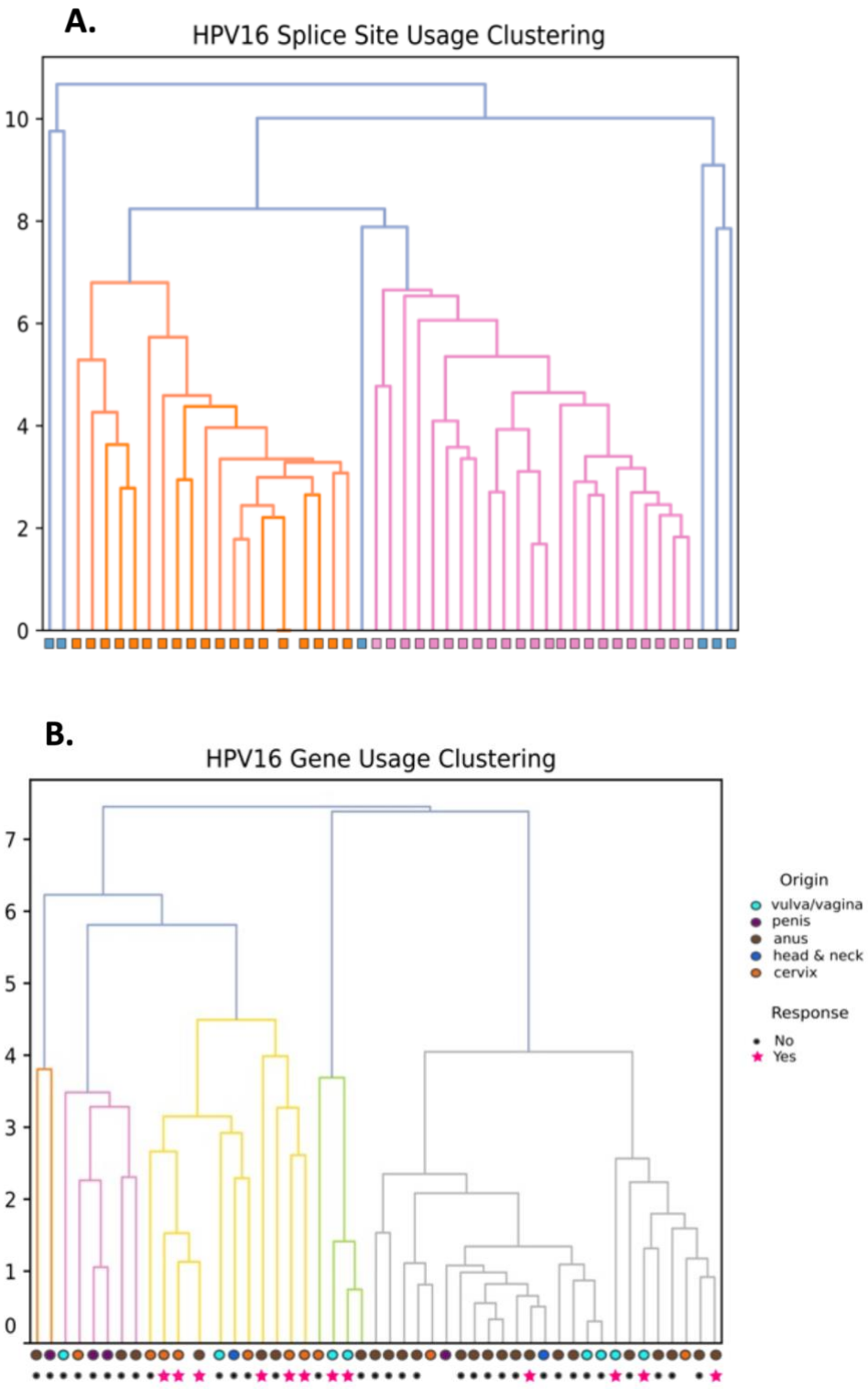
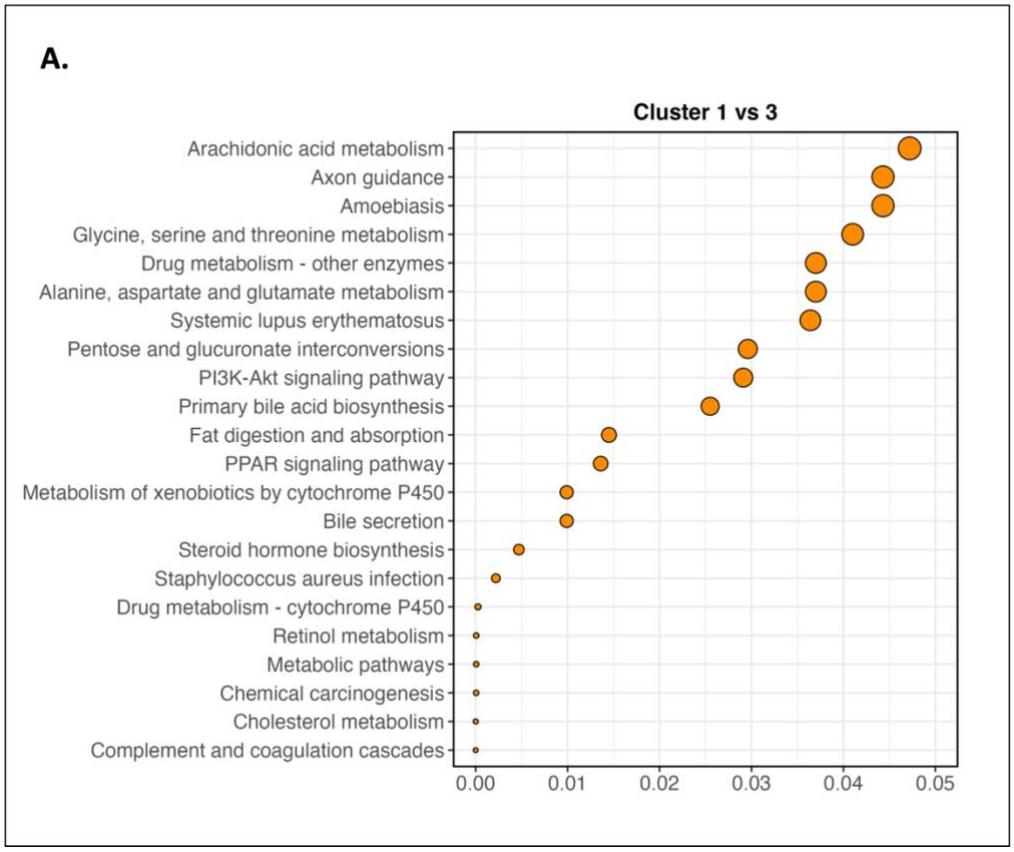


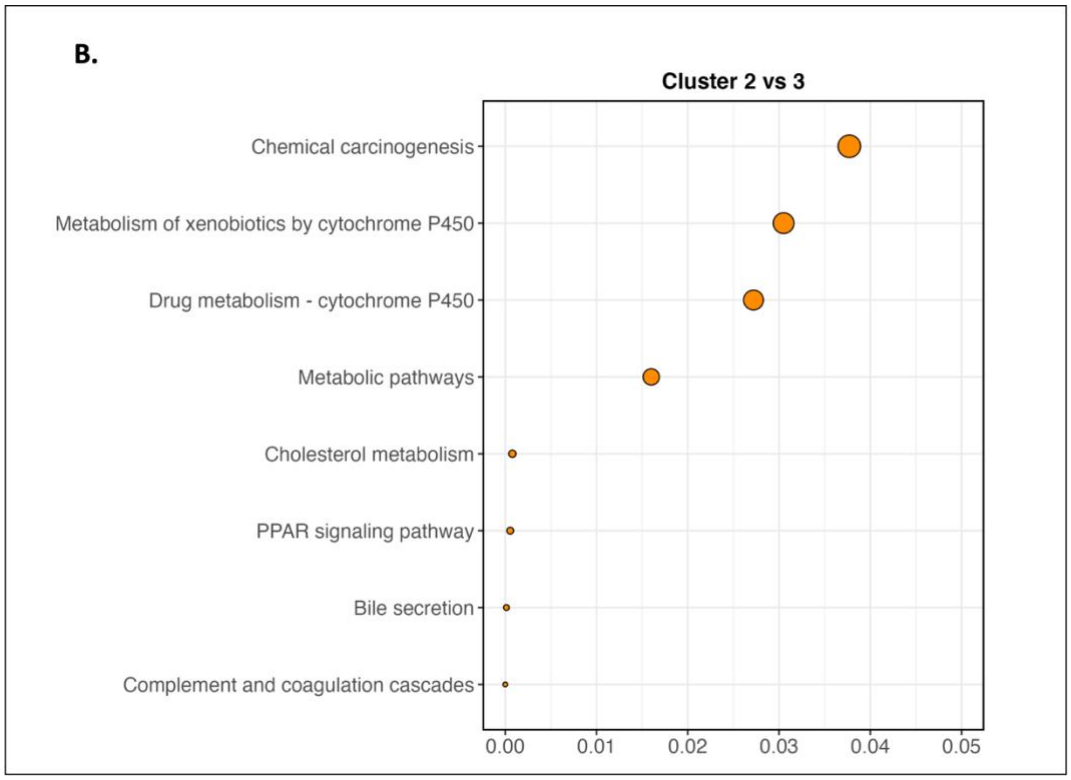
Figure 9. Hierarchical clustering results for splice site usage and gene usage composition. On the top (A.), Diego output reveals three distinct clusters for splice site data, with coloured boxes denoting sample assignments to specific clusters. On the bottom (B.), gene usage compositions exhibit five distinct groups, with non-blue branches/clusters indicating sample assignments based on splice site clustering. The color-coded origin circles in the legend provide insight into sample origins. The bottom of the legend denotes treatment response, with stars indicating a response and black dots indicating no response.

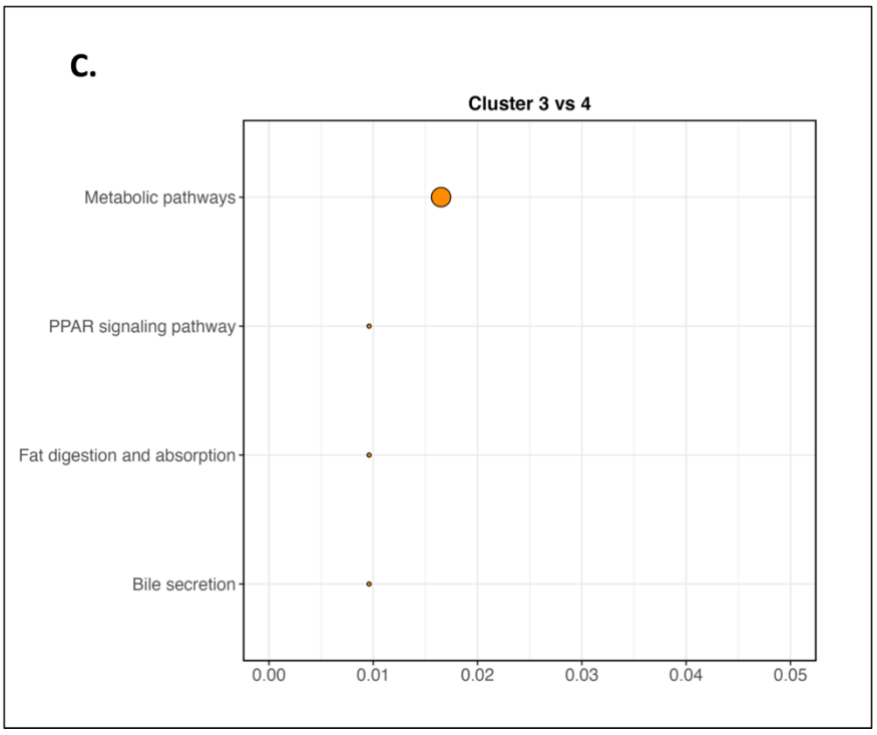
4.4 Differential Gene Expression and KEGG Pathway Analysis

The DGE dataset was stratified into five clusters for gene usage and three clusters for splice site usage, as was determined by DIEGO and R. Each sample was then assigned to one of these clusters for each dataset.

Subsequently, the focus of this thesis shifted to the gene usage analysis using DIEGO, prompted by the identification of an intriguing association with the PPAR pathway within the third cluster (see Figure 10 below). Upon examination, the third cluster yielded compelling outcomes, elucidating the modulation of various pathways. Notably, for each pairwise comparison involving the third cluster, the PPAR exhibited significant enrichment.




B.
Cluster 2 vs 3
Chemical carcinogenesis
Metabolism of xenobiotics by cytochrome P450
Drug metabolism - cytochrome P450
Metabolic pathways
Cholesterol metabolism
PPAR signaling pathway
Bile secretion
Complement and coagulation cascades
0.00
0.01
0.02
0.03
0.04
0.05


C.
Cluster 3 vs 4
Metabolic pathways
PPAR signaling pathway
Fat digestion and absorption
Bile secretion
0.00
0.01
0.02
0.03
0.04
0.05

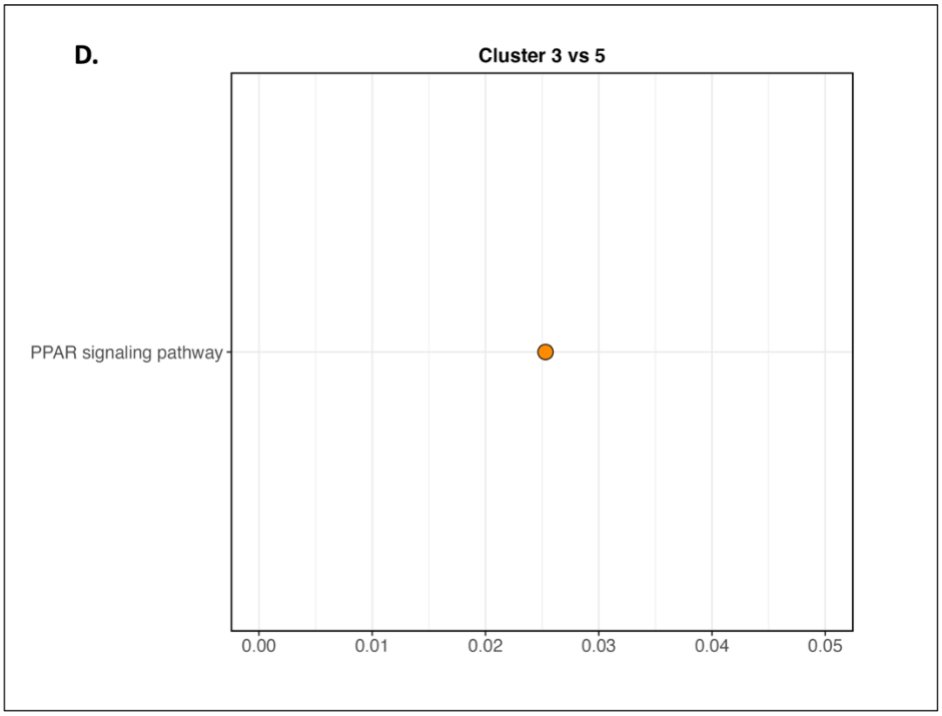


Figure 10. **KEGG pathway analysis for cluster comparisons.** These results for pairwise cluster comparisons emphasise the involvement of Cluster 3 (where each comparison is denoted by A, B, C and D for each pairwise comparison). Notably, the cluster-to-cluster comparison between Cluster 1 and Cluster 3 (A.) exhibits the highest number of entries. The x-axis denotes the p-value, and also the size of the bubbles is a visualize representation of the p-values found. PPAR is present in all comparisons of clusters against cluster 3.

Prior to delving into the modulation of the third cluster and examining how PPAR could influence treatment/response outcomes, the analysis was extended by introducing the HPV- subgroup as a sixth cluster. The HPV- cluster was comprised of the 58 negative samples. A similar process to the DGE workflow was followed in section 3.4 above, an adapted Figure 5 can be found below (in Figure 11), although the HPV- samples were dedicated their own cluster and differentially compared to the other samples in the clusters using DESeq2. There are 63 files which had no HPV16 reads detected.

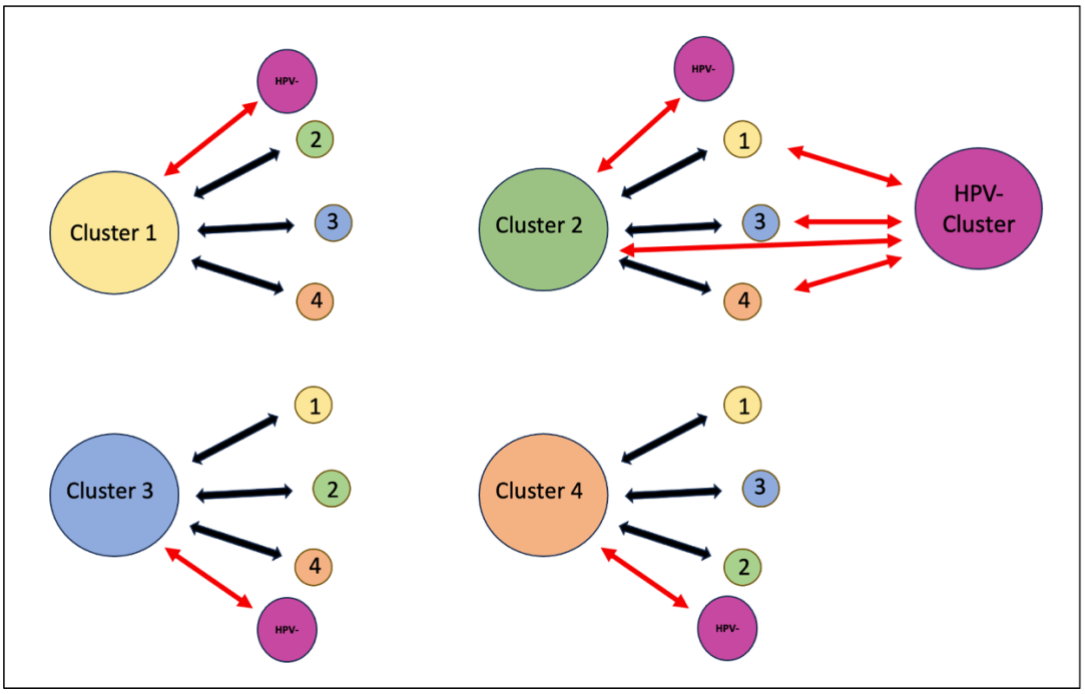


Figure 11. Pairwise comparison process within each clustering method with HPV- comparisons. Adaptation from Figure 5 above, with HPV negative workflow included. Red arrows denote the new comparisons using the HPV- subgroup.

However, no significant findings in terms of differential gene expression emerged within the HPV- cluster, when compared to the HPV+ clusters. This absence of noteworthy results stands in contrast to the distinctive patterns observed in the original five clusters determined by DIEGO. This result suggests that within the HPV- cluster, there were no significant changes in gene expression that stood out when compared to the other clusters. In other words, the genetic landscape within the HPV- subgroup did not show any distinctive patterns or alterations that could be linked to the specific pathways. This lack of distinctive patterns suggests that there are unique and potentially different molecular characteristics of the HPV- subgroup compared to the other clusters in this analysis, and the gene expression changes observed in the HPV+ clusters are not mirrored in the HPV- subgroup, indicating potential molecular distinctions between HPV-negative and HPV-positive cases in the studied clusters. The lack of differentially expressed genes in the HPV-negative subgroup may be attributed to its heterogeneity, characterized by varied gene expression patterns that diverge from the HPV16 positive clusters. The sample size imbalance among clusters could be a contributing factor, as well as the substantial size of the HPV-negative subgroup, in contrast to the smaller clusters of HPV16 positive patients, might impact statistical power. The HPV16 positive clusters range from as few as 2 samples to as many as 25 samples, introducing variability in the group sizes.

Fisher's exact test was applied to assess the response to treatment within the third cluster identified in the gene usage analysis conducted with DIEGO. A statistically significant value

was obtained, indicating that the response to treatment in this cluster was notably higher than in the other clusters. The number of samples within this cluster is 11, and the calculated p-value was 0.0445 for a noted treatment response, but lost significance after introducing corrections for multiple testing using Bonferroni. It is noteworthy that none of the remaining clusters exhibited a significant response to treatment or regulation of the PPAR pathway comparable to the distinctive pattern observed in this particular cluster.

The Peroxisome Proliferator-Activated Receptor (PPAR) plays a crucial role in regulating various biological functions, including lipid metabolism and inflammation [39]. The three isoforms—alpha, beta/delta, and gamma—have diverse functions and implications. In this step of the analysis, the aim was to elucidate the specific isoform involved. According to STRING-DB Wikipathways [28], three-fourths of the identified clusters are associated with the PPAR alpha isoform. However, for one cluster, determination of isoform association remains inconclusive which could be accounted to an insufficient enriched transcripts in the pathway. This is itemized in table 6 below.

Cluster 3 vs. 4	Cluster 2 vs. 3	Cluster 1 vs. 3	Cluster 3 vs. 5
Alpha	Alpha	Alpha	NA; pathway present in networks but no definite isoform on Wikipathways - could be due to lack of enriched transcripts (only 3 present)

Table 6. Isoform determination using STRING-DB for all pairwise comparisons using cluster 3. All belong to alpha, although the last clustering (3 vs. 5) is not available.

PPAR-α has been identified as a promoter of tumorigenesis in breast cancer, influencing proliferation and cell death through lipid metabolic modulation. It has been reported that the activation of PPAR-α promotes multiple signalling pathways, including the NF-κB/IL-6 axis, leading to the clonal expansion of breast cancer mammospheres [40]. In colon cancer, PPAR-α facilitates metastasis by inhibiting the expression of Cox-2, VEGF, and TGF-induced matrix metalloproteinase (MMP)-9, crucial factors in promoting metastasis [41]. Both PPAR-α and PPAR-γ have been shown to play widespread roles in the late stages of cancer, actively promoting metastasis [39].

The gene list associated with the PPAR pathway was extracted from the STRING-DB output and cross-referenced with the DESeq2 output files pertaining to the relevant clusters. As mentioned before, as human-mapped reads could not be utilized for data protection reasons,

Dr. Stephan Bernhart performed the DGE analysis using DESeq2. The aim was to discern the differential regulation of the pathway among these clusters. This investigation also encompassed the identification of upregulated/downregulated HPV genes from the DESeq2 results. The outcome as be seen in Figure 12 below.

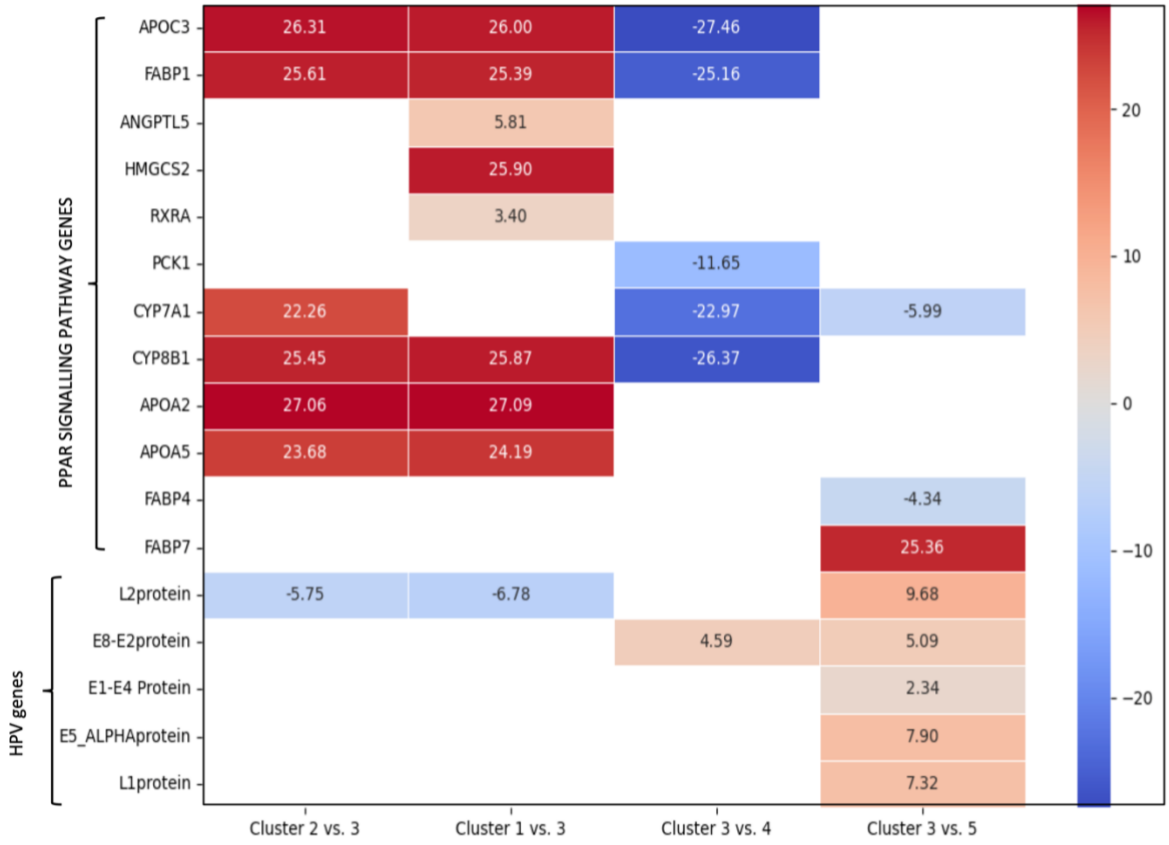


Figure 12. PPAR and HPV genes log-fold changes among clusters. This figure provides a detailed depiction of gene expression within the PPAR signalling pathway alongside differentially expressed HPV genes. The expression values are colour coded and offer a comprehensive overview of the regulatory landscape. The y-axis colour bar represents the log-fold change of expression. The inclusion of the PPAR signalling pathway sheds light on its potential role in the observed gene expression patterns. The differential expression of HPV genes adds a layer of specificity, highlighting key players in the viral-genomic interplay. This integrated view facilitates the identification of potential intersections between the PPAR signalling pathway and HPV-related transcriptional changes, contributing to a more nuanced understanding of the molecular dynamics within the dataset.

The results of this analysis suggest that an upregulation of PPAR signalling pathway associated genes, together with a downregulation of HPV genes, may contribute to a less favourable treatment response. Despite the well-established carcinogenicity of HPV, its presence

generally leads to improved treatment outcomes [42]. The prospect of an interaction between PPAR and HPV genes is plausible, especially in light of the successful incorporation of the PPAR pathway in constructing models for predicting outcomes in cervical cancer [43]. This finding prompts an exploration of potential interconnections or synergies between PPAR signalling and the expression profiles of HPV-associated genes. It underscores a complex interplay within the molecular framework of cervical cancer prediction models, offering a compelling avenue for further investigation.

It should also be highlighted that for cluster 1 compared to cluster 3, the cytochrome p450 and PI3K-Akt signalling pathway are overexpressed. An analysis of the data revealed a significant upregulation of cytochrome P450 genes in clusters 1 compared to 3, and 2 compared to 3. Notably, upregulation of two cytochrome P450 genes was identified in cluster 1 compared to cluster 3, which are implicated in the activation of carcinogens. One those genes is CYP2J2 epoxygenase, as highlighted in previous studies, which demonstrated its overexpression in human cancer cell lines and tissues. Moreover, overexpression of CYP2J2 is associated with increased tumour growth with elevated levels of epoxyeicosatrienoic acids (EETs), reduced apoptosis in cancer cells, and enhanced proliferation of carcinoma cells [44]. The resulting heatmap from this analysis can be seen in Figure 13 below. The upregulation of UGT2B7 and UGT2B4 in the cytochrome P450 pathway, as identified in the squamous cell carcinoma data, aligns with previous research on liver cancer, revealing miR-3664-3p-mediated regulation of UGT2B7 and miR-135a-5p and miR-410-3p-mediated regulation of UGT2B4 in liver cancer cells, highlighting potential molecular mechanisms contributing to the dysregulation of these genes in liver cancer and possibly implicating similar regulatory pathways in squamous cell carcinoma [45].

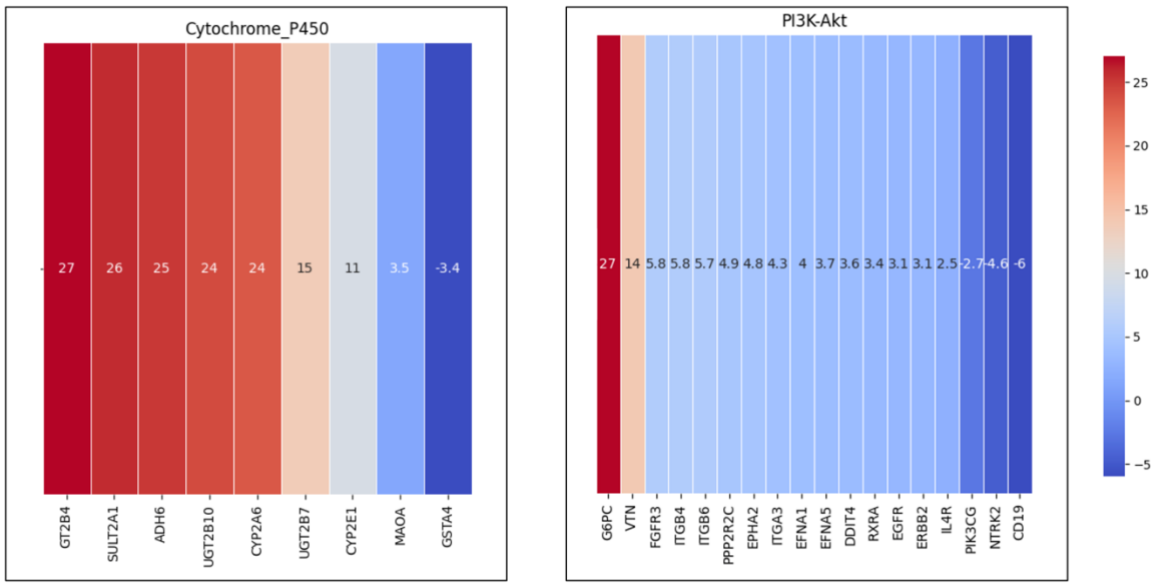


Figure 13. Cluster 1 vs 3 log fold changes in transcripts in the cytochrome P450 the PI3K-Akt pathway. The y-axis colour bar represents the log-fold change of expression.

As mentioned above, another pathway that has been found to be overexpressed in cluster 1 compared to cluster 3, was the PIK3-Akt signalling pathway. This pathway revealed a discernible pattern of upregulation across numerous genes, each demonstrating varying degrees of significance. This observation prompts a more in-depth exploration into the molecular intricacies governing the regulatory dynamics of this pathway. A growing body of evidence supporting these findings underscores a substantial and pervasive impact on cellular processes. Several studies have converged on the consensus that the hyperactivation of the PI3K/Akt pathway constitutes a critical factor in the multifaceted landscape of cancer biology. This pathway's influence extends across several key hallmarks of cancer, encompassing not only the initiation of carcinogenesis but also the facilitation of tumour cell proliferation, invasion, metastasis, and the development of resistance to therapeutic interventions [46].

However, it is crucial to note that cluster 1 contains only two samples. This raises a valid concern about the robustness of the findings, as such a small sample size may limit the generalizability and statistical power of the observed trends. Therefore, while the identified upregulation in cytochrome P450 genes and the PI3K-Akt signalling pathway is noteworthy, the interpretation should be approached with caution, considering the limited sample representation in cluster 1. Although cytochrome P450 was also found in cluster 2 vs. 3, which contains significantly more samples. Further investigation with a larger sample size is warranted to validate and strengthen the reliability of these findings. Moreover, there are other enriched pathways, as depicted in Figure 10 However, many of these involve typical physiological processes, such as fat digestion and absorption, or bile secretion, which may not be of particular interest in the context of this study.

5 Discussion

This study strategically concentrates on a singular histological subtype, specifically squamous SCCs, following the innovative framework of the PEVOsq trial—a European-funded phase II basket trial. This exclusive focus on SCCs found at diverse anatomical sites reflects a departure from traditional organ-specific clinical trials. By delving into the molecular landscape of this particular histological subtype, the aim of this thesis was to transcend conventional approaches and align with the contemporary trend of addressing cancer based on shared molecular characteristics. This focused exploration unveils similarities within the SCC histological subtype across various anatomical locations, and aimed to produce a nuanced understanding of its biological underpinnings. The significance of this approach lies in tailoring treatment strategies based on histology rather than the primary anatomical site, presenting potential avenues for better targeted and effective interventions.

The anticipated and fundamental outcome was the mostly successful identification of each HPV subtype through RNAseq, aligning with existing literature and expectations. This is a positive validation of the method's accuracy. However, the identification of sublineages for individual patients proved challenging due to limitations in genome coverage and completeness. Understanding the specific sublineage prevalent in an individual's HPV infection can provide insights into the associated phenotypic characteristics, such as varying carcinogenic risks and ethnic disparities. This information can contribute to a more personalized approach in assessing the potential outcomes and risks for the patient [31]. An intriguing avenue for future research involves a closer examination of common SNPs among patients. This approach could potentially reveal associations between specific SNPs and adverse treatment/response outcomes. Considering publications using machine learning method to predict HPV16 cervical cancer outcomes [32], employing similar models may uncover links between genetic variations, disease progression, and potential SNPs associated with treatment resistances.

A noteworthy finding is the identification of distinct gene usage composition clusters, with cluster 3 from the DIEGO results emerging as a significant contributor, displaying widespread differential expression across various gene sets and pathways. This cluster stands out due to its distinctive transcriptomic profile, suggesting a unique role in the molecular dynamics. This differentiation is evident in its potential association with the downregulation of HPV genes and the upregulation of PPAR pathway compared to other clusters. This aligns with existing literature, as one research group has utilized the PPAR signalling pathway as a primary indicator of cervical cancer outcomes [43]. Although, the role of PPAR in cancer research is a subject of debate, with conflicting opinions on whether it promotes tumorigenesis or exerts an opposing effect [47]. While supported by previous studies, the direct link between the differential expression and potential regulation of HPV genes and PPAR signalling is not well studied. Although one study postulates that HPV16 E7 (oncogene) upregulates miR-27b

(micro-RNA), leading to the inhibition of PPARγ expression and fostering proliferation and invasion in cervical carcinoma cells [48]. The findings indicated a significant role of PPARγ, identified as a target of miR-27b, in suppressing the progression of cervical cancer by downregulating sodium-hydrogen exchanger isoform 1 (NHE1). In summary, this study highlights that HPV16 E7 upregulates miR-27b, leading to the inhibition of PPARγ expression, ultimately promoting proliferation and invasion in cervical carcinoma cells [48]. Although these findings contrast with the findings of this thesis, it could potentially be explained due to a different PPAR isoform. Cluster 3 compared to cluster 5 from the gene usage DIEGO analysis (see table 6, section 4.4). faced challenges in assigning an isoform due to a deficiency of transcripts. In all other cluster comparisons PPAR-alpha was found to be the most likely isoform. While there is a possibility that it might be another isoform, the current evidence predominantly suggests that it is PPAR-alpha.

Several limitations regarding the study design and data of this PEVO project need consideration. Notably, the focus on late-stage patients poses challenges in deciphering survival outcomes due to the generally poor prognosis in this cohort. Additionally, the limited availability of samples at specific time points hinders the ability to draw robust conclusions about temporal changes in gene expression. Also being provided with only the unmapped reads restricted the scope of this thesis, since there was no access to the full human datasets.

In this study, comprehensive exploration of the molecular landscape has been undertaken within the confines of a single histological subtype. The approach follows the framework of the PEVO trial, focusing on SCCs originating from diverse anatomical sites. By narrowing the focus to this specific histological subtype, the aim was to bridge the gap between traditional organ-specific clinical trials and the contemporary trend of treating cancer based on molecular alterations. This focused approach allowed this thesis to uncover shared molecular similarities within the histological subtype across various anatomical locations. The study's unique contribution lies in identifying treatment strategies based on the histology itself rather than the primary location of the SCC, emphasizing the potential for more targeted and effective interventions.

Future analyses should additionally leverage increasingly popular technologies, such as nanopore sequencing, to complement RNASeq data. Exploring the episomal, or integrated nature of HPV in the genome could offer further insights into the dynamics of viral-host interactions. Understanding these aspects contributes to refining prognostic markers and therapeutic targets. The relevance of continued research on this topic is of importance to the broader population affected by HPV infections. As a prevalent and persistent issue, HPV-associated cancers impact a significant portion of the population. Unravelling the molecular landscape contributes to our understanding of the disease, potentially informing better diagnostic and therapeutic strategies.

In conclusion, this thesis sheds light on and explores the intricate molecular landscape of HPV-associated cancers through an integrative analysis of HPV transcriptomes. While acknowledging limitations and challenges, this thesis provides a foundation for future research avenues, emphasizing the importance of continued exploration to decipher the complexities of HPV-related transcriptomic changes.

6 References

[1] L. S. A. Mühr, C. Eklund, and J. Dillner, 'Towards quality and order in human papillomavirus research', *Virology*, vol. 519, pp. 74–76, Jun. 2018, doi: 10.1016/j.virol.2018.04.003.

[2] C. W. Nelson and L. Mirabello, 'Human papillomavirus genomics: Understanding carcinogenicity', *Tumour Virus Research*, vol. 15, p. 200258, Jun. 2023, doi: 10.1016/j.tvr.2023.200258.

[3] A. Bansal, M. P. Singh, and B. Rai, 'Human papillomavirus-associated cancers: A growing global problem.', *Int J Appl Basic Med Res*, vol. 6, no. 2, pp. 84–89, Jun. 2016, doi: 10.4103/2229-516X.179027.

[4] A. Ure, D. Mukhedkar, and L. S. Arroyo Mühr, 'Using HPV-meta for human papillomavirus RNA quality detection', *Scientific Reports*, vol. 12, no. 1, p. 13058, Jul. 2022, doi: 10.1038/s41598-022-17318-5.

[5] E.-K. Yim and J.-S. Park, 'The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis.', *Cancer Res Treat*, vol. 37, no. 6, pp. 319–324, Dec. 2005, doi: 10.4143/crt.2005.37.6.319.

[6] T. Nakahara and T. Kiyono, 'Interplay between NF-κB/interferon signaling and the genome replication of HPV', *Future Virology*, vol. 11, Feb. 2016, doi: 10.2217/fvl.16.2.

[7] S. V. Graham, 'Human papillomavirus: gene expression, regulation and prospects for novel diagnostic methods and antiviral therapies.', *Future Microbiol*, vol. 5, no. 10, pp. 1493–1506, Oct. 2010, doi: 10.2217/fmb.10.107.

[8] Y. Zhu *et al.*, 'Metagenomic Next-Generation Sequencing vs. Traditional Microbiological Tests for Diagnosing Varicella-Zoster Virus Central Nervous System Infection.', *Front Public Health*, vol. 9, p. 738412, 2021, doi: 10.3389/fpubh.2021.738412.

[9] A. Khan *et al.*, 'Detection of human papillomavirus in cases of head and neck squamous cell carcinoma by RNA-seq and VirTect.', *Mol Oncol*, vol. 13, no. 4, pp. 829–839, Apr. 2019, doi: 10.1002/1878-0261.12435.

[10] E. de Guillebon *et al.*, 'Combining immunotherapy with an epidrug in squamous cell carcinomas of different locations: rationale and design of the PEVO basket trial.', *ESMO Open*, vol. 6, no. 3, p. 100106, Jun. 2021, doi: 10.1016/j.esmoop.2021.100106.

[11] 'Virus-Host database'. [Online]. Available: https://www.genome.jp/virushostdb/

[12] 'National Center for Biotechnology Information (NCBI)[Internet]'. Accessed: Mar. 09, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[13] 'PapillomaVirus Episteme (PaVE)'. Accessed: Mar. 09, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[14] K. Van Doorslaer *et al.*, 'The Papillomavirus Episteme: a major update to the papillomavirus sequence database', *Nucleic Acids Research*, vol. 45, no. D1, pp. D499–D506, Jan. 2017, doi: 10.1093/nar/gkw879.

[15] ‘National Microbial Pathogen Database Resource (NMPDR)’. Accessed: Mar. 20, 2023. [Online]. Available: https://www.bv-brc.org
[16] A. Dobin *et al.*, ‘STAR: ultrafast universal RNA-seq aligner.’, *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
[17] B. Bushnell, ‘BBMap: A Fast, Accurate, Splice-Aware Aligner’, in *Report Number: LBNL-7065E*, Research Org.: Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Mar. 2014. [Online]. Available: https://www.osti.gov/biblio/1241166
[18] R. D. Burk, A. Harari, and Z. Chen, ‘Human papillomavirus genome variants.’, *Virology*, vol. 445, no. 1–2, pp. 232–243, Oct. 2013, doi: 10.1016/j.virol.2013.07.018.
[19] F. Sievers and D. G. Higgins, ‘Clustal Omega for making accurate alignments of many protein sequences’, *Protein Science*, vol. 27, no. 1, pp. 135–145, Jan. 2018, doi: 10.1002/pro.3290.
[20] Y. Bai, S. Ji, and Y. Wang, ‘IRcall and IRclassifier: two methods for flexible detection of intron retention events from RNA-Seq data.’, *BMC Genomics*, vol. 16 Suppl 2, no. Suppl 2, p. S9, 2015, doi: 10.1186/1471-2164-16-S2-S9.
[21] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, ‘Microbiome Datasets Are Compositional: And This Is Not Optional.’, *Front Microbiol*, vol. 8, p. 2224, 2017, doi: 10.3389/fmicb.2017.02224.
[22] T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley, ‘Understanding sequencing data as compositions: an outlook and review’, *Bioinformatics*, vol. 34, no. 16, pp. 2870–2878, Aug. 2018, doi: 10.1093/bioinformatics/bty175.
[23] Y. Liao, G. K. Smyth, and W. Shi, ‘featureCounts: an efficient general purpose program for assigning sequence reads to genomic features’, *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.
[24] F. Erhard, ‘Estimating pseudocounts and fold changes for digital expression measurements’, *Bioinformatics*, vol. 34, no. 23, pp. 4054–4063, Dec. 2018, doi: 10.1093/bioinformatics/bty471.
[25] K. G. van den Boogaart and R. Tolosana-Delgado, ‘“compositions”: A unified R package to analyze compositional data’, *Computers & Geosciences*, vol. 34, no. 4, pp. 320–338, 2008, doi: https://doi.org/10.1016/j.cageo.2006.11.017.
[26] G. Doose, S. H. Bernhart, R. Wagener, and S. Hoffmann, ‘DIEGO: detection of differential alternative splicing using Aitchison’s geometry’, *Bioinformatics*, vol. 34, no. 6, pp. 1066–1068, Mar. 2018, doi: 10.1093/bioinformatics/btx690.
[27] M. I. Love, W. Huber, and S. Anders, ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome Biology*, vol. 15, no. 12, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
[28] D. Szklarczyk *et al.*, ‘The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets’, *Nucleic Acids Research*, vol. 49, no. D1, pp. D605–D612, Jan. 2021, doi: 10.1093/nar/gkaa1074.

[29] M. Kanehisa and S. Goto, 'KEGG: kyoto encyclopedia of genes and genomes.', *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[30] C. Götz, C. Bischof, K.-D. Wolff, and A. Kolk, 'Detection of HPV infection in head and neck cancers: Promise and pitfalls in the last ten years - a meta-analysis', *Mol Clin Oncol*, vol. 10, no. 1, pp. 17–28, Jan. 2019, doi: 10.3892/mco.2018.1749.
[31] Z. Ou *et al.*, 'Genetic signatures for lineage/sublineage classification of HPV16, 18, 52 and 58 variants', *Virology*, vol. 553, pp. 62–69, Jan. 2021, doi: 10.1016/j.virol.2020.11.003.
[32] L. Asensio-Puig, L. Alemany, and M. A. Pavón, 'A Straightforward HPV16 Lineage Classification Based on Machine Learning.', *Front Artif Intell*, vol. 5, p. 851841, 2022, doi: 10.3389/frai.2022.851841.
[33] L. Zhang *et al.*, 'Variants of human papillomavirus type 16 predispose toward persistent infection.', *Int J Clin Exp Pathol*, vol. 8, no. 7, pp. 8453–8459, 2015.
[34] M.-A. C. Bind and D. B. Rubin, 'When possible, report a Fisher-exact P value and display its underlying null randomization distribution.', *Proc Natl Acad Sci U S A*, vol. 117, no. 32, pp. 19151–19158, Aug. 2020, doi: 10.1073/pnas.1915454117.
[35] M. Ringnér, 'What is principal component analysis?', *Nature Biotechnology*, vol. 26, no. 3, pp. 303–304, Mar. 2008, doi: 10.1038/nbt0308-303.
[36] A. Pal and R. Kundu, 'Human Papillomavirus E6 and E7: The Cervical Cancer Hallmarks and Targets for Therapy.', *Front Microbiol*, vol. 10, p. 3116, 2019, doi: 10.3389/fmicb.2019.03116.
[37] L. Yu, V. Majerciak, and Z.-M. Zheng, 'HPV16 and HPV18 Genome Structure, Expression, and Post-Transcriptional Regulation.', *Int J Mol Sci*, vol. 23, no. 9, Apr. 2022, doi: 10.3390/ijms23094943.
[38] T. P. Schrank *et al.*, 'Direct Comparison of HPV16 Viral Genomic Integration, Copy Loss, and Structural Variants in Oropharyngeal and Uterine Cervical Cancers Reveal Distinct Relationships to E2 Disruption and Somatic Alteration.', *Cancers (Basel)*, vol. 14, no. 18, Sep. 2022, doi: 10.3390/cancers14184488.
[39] Q. Gou, X. Gong, J. Jin, J. Shi, and Y. Hou, 'Peroxisome proliferator-activated receptors (PPARs) are potential drug targets for cancer therapy', *Oncotarget; Vol 8, No 36*, 2017, Accessed: Jan. 01, 2017. [Online]. Available: https://www.oncotarget.com/article/19610/text/
[40] Y. Wang, F. Lei, Y. Lin, Y. Han, L. Yang, and H. Tan, 'Peroxisome proliferator-activated receptors as therapeutic target for cancer', *Journal of Cellular and Molecular Medicine*, vol. n/a, no. n/a, Sep. 2023, doi: 10.1111/jcmm.17931.
[41] X. Yin *et al.*, 'PPARα Inhibition Overcomes Tumor-Derived Exosomal Lipid-Induced Dendritic Cell Dysfunction', *Cell Reports*, vol. 33, no. 3, p. 108278, Oct. 2020, doi: 10.1016/j.celrep.2020.108278.
[42] J. Zhang, Y. Zhang, and Z. Zhang, 'Prevalence of human papillomavirus and its prognostic value in vulvar cancer: A systematic review and meta-analysis.', *PLoS One*, vol. 13, no. 9, p. e0204162, 2018, doi: 10.1371/journal.pone.0204162.

[43] Y. Zhang *et al.*, 'Development and Validation of the Promising PPAR Signaling Pathway-Based Prognostic Prediction Model in Uterine Cervical Cancer', *PPAR Research*, vol. 2023, p. 4962460, May 2023, doi: 10.1155/2023/4962460.
[44] A. M. Alzahrani and P. Rajendran, 'The Multifarious Link between Cytochrome P450s and Cancer.', *Oxid Med Cell Longev*, vol. 2020, p. 3028387, 2020, doi: 10.1155/2020/3028387.
[45] D. D. Wijayakumara, P. I. Mackenzie, R. A. McKinnon, D. G. Hu, and R. Meech, 'Regulation of UDP-Glucuronosyltransferases UGT2B4 and UGT2B7 by MicroRNAs in Liver Cancer Cells.', *J Pharmacol Exp Ther*, vol. 361, no. 3, pp. 386–397, Jun. 2017, doi: 10.1124/jpet.116.239707.
[46] F. Rascio *et al.*, 'The Pathogenic Role of PI3K/AKT Pathway in Cancer Onset and Drug Resistance: An Updated Review.', *Cancers (Basel)*, vol. 13, no. 16, Aug. 2021, doi: 10.3390/cancers13163949.
[47] J. Youssef and M. Badr, 'Peroxisome proliferator-activated receptors and cancer: challenges and opportunities.', *Br J Pharmacol*, vol. 164, no. 1, pp. 68–82, Sep. 2011, doi: 10.1111/j.1476-5381.2011.01383.x.
[48] S. Zhang *et al.*, 'Elevation of miR-27b by HPV16 E7 inhibits PPARγ expression and promotes proliferation and invasion in cervical carcinoma cells.', *Int J Oncol*, vol. 47, no. 5, pp. 1759–1766, Nov. 2015, doi: 10.3892/ijo.2015.3162.