UNIVERSITY OF CALIFORNIA

SANTA CRUZ

# INVESTIGATING THE UNDERLYING CAUSES OF CANCER USING SEQUENCING

A dissertation submitted in partial satisfaction

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MOLECULAR, CELL AND DEVELOPMENTAL BIOLOGY

by

**Allison Cheney**

March 2024

The Dissertation of Allison Cheney is approved:

Professor Olena Morozova Vaske, Chair

Professor Joshua Stuart

Professor Sameer Agnihotri

Peter Biehl
Vice Provost and Dean of Graduate Studies

# Table of Contents

# List of Figures and Tables

# Abstract


Investigating the Underlying Causes of Cancer Using Sequencing

Allison Cheney

Cancer is a genetic disease. All cancers are caused by and dependent on mutations. The widespread adoption of NGS in cancer research has revolutionized the field by allowing us to catalog the variants within and across cancer types. This knowledge was used to create a new taxonomy of cancers, where many tumors are classified by genetic alterations in addition to tissue of origin. Despite the role of NGS in identifying the mutations behind these new classifications, pre-genomic methods such as ISH and IHC are still commonly used in the clinic to identify relevant genetic alterations in tumors.

About 20% of tumors worldwide are caused by viruses such as HPV, Epstein-Barr virus and Hepatitis B and C viruses. The virus may act as a mutagen by causing prolonged inflammation or by integration into the host genome. Viral infection can therefore be considered another type of genetic alteration used for the classification of tumors, and indeed some tumors are now classified by their causal viral infections.

Epimutations are a type of genetic alteration affects epigenetic marks such as DNA methylation rather than the DNA sequence. Epimutations are the genetic lesions behind several disorders including cancer predisposition syndromes.

Here I present my work developing an app that integrates with the UCSF 500 pipeline to detect known oncogenic viral sequences in DNA samples from tumors (Chapter 2). I also present my work analyzing long-read sequencing and methylation data of a patient with bilateral Wilms tumor to identify the epimutation causing their cancer predisposition syndrome previously missed by other

technologies (Chapter 3). My work reflects the utility of adapting state-of-the-art sequencing technologies for diagnostic use in the clinic.

H3K27M gliomas have a mutation in the genes encoding histone H3 that interferes with the deposition of the repressive epigenetic mark H3K27me3. This mark is important in regulating cell type development and EMT. In Chapter 4 I present my work using gene expression analysis to identify the role of EMT in the development and oncogenesis of H3K27M gliomas. Aberrant EMT may serve as a key dependency of these tumors.

# Dedication

This dissertation is dedicated to my childhood cat Nugget, who passed away in the late stages of writing it. She would have loved sitting on this manuscript while people were trying to read it.

May 15, 2004 – February 14, 2024

# Acknowledgements

I gratefully acknowledge the support of my advisor Dr. Olena Vaske, who gave me the freedom, time, and confidence to pursue new projects and ideas. I thank those who interrupted their extremely busy schedules to serve as my mentors, including Drs. Sameer Agnihotri, Josh Stuart, Holly Beale, Melissa Cline, and Patrick Devine.

The members of the Vaske lab and Treehouse Childhood Cancer Initiative, past and present, were extraordinary colleagues and friends. I thank Lauren Sanders, Geoff Lyle, Ellen Towle-Kephart, Holly Beale, Yvonne Vasquez, Krizia Chambers and Anouk van den Bout for their kind support, guidance, friendship, and creative suggestions. I thank Lauren Sanders for guiding me through my rotation almost entirely by herself. I would particularly like to thank Ellen Towle-Kephart for being incredibly patient with a nervous rotation student who had caused a catastrophic problem with the server on her first week in the lab.

I was fortunate to participate in many collaborations in my time as a student. I thank my collaborators for allowing me to participate in teams of incredible and creative scientists.

Last but not least, I am enormously grateful for the encouragement from my family and friends.

# Chapter One. Introduction

## Cancer and its Causes

Cancer is the second leading cause of death around the world[1]. In many cases, cancer is thought to be caused by the chance accumulation of somatic mutations in cells over time[2], modified by environmental and lifestyle factors[3]. The other 10-30% of cancers are thought to be attributable to a genetic cancer predisposition[4]. These mutations result in an advantage to the cancer cell by allowing uncontrolled cell growth.

### Overview of History of Research

Cancer may be our oldest medical mystery. Tumors have been found in 300 million year old fossils[5,6] of fish. A large-scale study of fossils from dinosaurs from the Jurassic and Cretaceous eras found tumors in 0.2% of specimens examined[6–8]. Neoplasms have been identified in the remains of Ancient Egyptian mummies[5]. In the year 160, Galen claimed cancer was caused by black bile that had become "trapped" in the body[7]. This idea was not overturned until the 1530s when Vesalius noted that none of his dissections of deceased cancer patients showed any evidence of black bile[9].

More credible theories emerged. The theory that cancers are caused by carcinogens took an early lead. In 1713, Bernardo Ramazzini noted several cases of occupation-associated cancers[7]. In 1761, John Hill was widely ridiculed for arguing that the use of snuff (oral tobacco) caused cancer[10]. In 1765, Percivall Pott noted that he had encountered many cases of skin cancers of the scrotum in boys who had worked as chimney sweeps[11]. Yet it was not until the twentieth century that scientists were able to successfully induce cancer in an experimental system. Jean Clunet was able to induce tumors in a single rat by exposing it to X-rays in 1910[12]. In

1915, the Japanese scientist Katsuaburo Yamagiwa reported that repeatedly painting the ears of rabbits with coal tar over many months results in tumors[13]. However, the mechanism behind this oncogenesis was not known.

In the early 1900s, Theodor Boveri, a scientist working with sea urchins, made a critical observation involving chromosomes[14]. He noted that sea urchin eggs fertilized by two sperm instead of one gave different numbers of chromosomes to their daughter cells. This led to differing effects on the cells. In many cases, the cells die, but in others, the cells grow and develop abnormally. Boveri was the first to propose chromosomes as the mechanism of heredity. Additionally, in a remarkable intuitive leap, Boveri suggested that cancer is caused by the inheritance of "a particular, incorrect chromosome combination"[14].

Others held that cancer was caused by infections. The citizens of Reims, France thought so: in 1779, the first cancer hospital was built a good distance from the city because they feared cancer was contagious[7]. In 1876, Mistislav Novinski, a researcher in Russia, showed that transplanting tumor cells from one dog to another could spread cancer[9,15]. This may be the first evidence of viruses causing cancer, but Novinski did not suggest this, as viruses would not be discovered until the 1890s.

In 1911, Boveri's contemporary Peyton Rous proved that a virus, later called Rous Sarcoma Virus (RSV), caused sarcomas in chickens[16,17]. Had the main causal agent of cancer been found? Initially, his observations weren't well received because cancers were thought to be of some unspecified "endogenous origin"[18]. When the evidence for RSV's role in avian cancer became irrefutable, it was dismissed by many as irrelevant to human cancers[16]. Many early attempts to identify cancer-causing viruses ("oncoviruses") in mammals failed[16]. Finally, in 1935, Richard Shope identified a type of papillomavirus in rabbits that could cause cancer[19]. Yet Shope's papillomavirus differed from RSV in a surprising way: it was a DNA virus, while RSV was an RNA virus.

In 1913, Johannes Fibiger published results showing that infection with nematodes in rats led to cancer[20]. Fibiger was awarded the Nobel prize for these results in 1927. At the time, cancer was "regarded essentially as a parasitic disease"[21]. Years later, a review of Fibiger's histology images and other evidence would prove that the lesions Fibiger found in the rats were not cancerous at all, but rather benign formations caused by an abnormal diet and nematode infection[21].

By the late 1950s, oncoviruses were back in favor. Viruses had recently been discovered as the culprits behind numerous diseases that had been mysterious for centuries: rabies, foot and mouth disease, and smallpox. It seemed plausible that viruses could be behind cancers as well. The first human oncovirus was identified in Epstein Barr Virus (EBV) in 1958 in Burkitt's lymphoma[16]. Largely on the strength of this discovery, the National Cancer Institute (NCI) created the "Special Virus Cancer Program" to discover new oncoviruses[22], with a significant portion of the NCI's budget behind it. In 1962, Life Magazine's headline announced: "Cancer may be infectious". And in 1966, Rous finally won his Nobel prize for the discovery of RSV 55 years earlier. In his Nobel lecture, he sharply dismissed what he termed "somatic mutation hypothesis": that somatic mutations can cause, or at least contribute to, cancer. Rous was well aware of Boveri's hypothesis, but he believed, mistakenly, that it was incompatible with viral oncogenesis[16]. The viral oncogenesis camp and the genetic oncogenesis camp were as far apart as ever.

While it was accepted that RSV and other viruses could cause cancer, what wasn't known was how this occurred. Little thought was given to a genetic basis for cancer and according to the virologist Harold Varmus, at the time "viruses seemed to be the only game in town"[23]. In the early 1970s, the labs of Sol Spiegelman, Robert Gallo, and others found retroviral RNA in many different types of cancers[24–27]: leukemias and lymphomas[28–35], breast cancer[36–41], lung cancer[42], sarcomas[43], nasopharyngeal carcinomas[44], brain tumors[45], and melanomas[46], and, oddly, human

milk[47]. But when these findings could not be replicated, the field of viral oncology suffered an embarrassment.

In 1970, Peter Vogt and Peter Duesberg showed that RSV strains with certain deletions could not cause cancer[48]. It was suspected that a certain gene in RSV caused cancer in its host, later called "*src*". *Src*, the first "oncogene", was sequenced in 1980[49].

Oncogenes had first been proposed by Huebner and Todaro in 1969[50]. They suggested that because viral DNA had accumulated over time in many vertebrate genomes including humans, cancer might be caused by the aberrant activation of this viral DNA[50]. While prescient, research did not bear out their theory: *src* did not have the hallmarks of a viral gene. In fact, *src* had both exons and introns; in other words, it had the structure of a typical gene in a normal, non-cancer cell[18,51,52]. Harold Varmus and Michael Bishop found *src* in normal, non-cancerous cells of chickens, mammals, and numerous other species [53]. It was clear that the *src* gene in RSV had originated from a normal cell infected by RSV. The *src* gene was then "kidnapped" by the virus and incorporated into its genome. RSV's *src* was simply a slightly mutated version of the proto-oncogenes we all carry. The mechanism of carcinogens was explained when Bruce Ames showed that many carcinogens, such as the compounds found in coal tar, are in fact mutagenic[54]. Finally, other studies showed that adding DNA from human cancers into normal cells induced cancer[55,56]. The most direct cause of cancer isn't an outside agent but our own mutated genes; in the words of Harold Varmus, "a distorted version of our normal selves"[23].

## Types of Mutations

Categories of mutations can include alterations at a single base (SNPs) or multiple bases (including small insertions/deletions (indels), copy number variants (CNVs),

translocations and inversions)[57]. A variant is considered a CNV if it is greater than 1000 bases in length.

In addition to the above changes occurring directly to the DNA sequence, gene expression is also controlled epigenetically, such as changes in DNA methylation or histone modifications[58].

Not all mutations seen in cancer cells contribute to the growth or aggressiveness of the cancer[57]. “Driver mutations” are clinically relevant mutations in cancers that are seen in two types of genes: oncogenes and tumor suppressor genes[57,59]. Proto-oncogenes are normal genes that become oncogenes following mutations that lead to increased activity. Proto-oncogenes can be activated by many mechanisms, including chromosomal translocation. In Burkitt’s lymphoma, the MYC proto-oncogene is translocated from its normal position on the q arm of chromosome 8 to a position on other chromosomes[60]. Activation of MYC leads to increased proliferation of the cell[60].

Tumor suppressor genes normally prevent the development of cancer, until they are mutated in such a way that their function is lost. For example, *RB1* is a tumor suppressor gene that normally regulates the cell’s entrance into the cell cycle. When it is mutated, cell proliferation may occur unchecked.

Tumors in adults typically carry two to eight driver gene mutations[57,61], but pediatric cancers may contain as few as two or none at all[61]. This may be explained by the fact that the oncogenic events behind pediatric cancers are hypothesized to occur in early progenitor cells that already have the capability of proliferation and self-renewal[61]. Therefore, pediatric cancers may be caused by a stall in differentiation rather than cells taking on the characteristics of stem cells.

## Germline Mutations

While the majority of cancers are caused by somatic mutations, approximately 10% of cancers are due to inherited genetic traits (germline mutations)[62]. In 2014, at least

114 cancer predisposition genes were known[63]. These inherited mutations may cause cancer in childhood or in adulthood.

### Retinoblastoma

An illustrative example of a cancer caused by frequent germline mutations is retinoblastoma, a tumor of the retina which usually affects young children. The first suggestion of a cancer predisposition syndrome occurred in 1886, when Hilário de Gouvêa, a Brazilian ophthalmologist, first reported retinoblastoma running in a family[64]. In 25%-40% of cases, patients carry a heterozygous (single) germline mutation in *RB1*[65]. These patients have a high risk of developing other cancers[62]. Alfred Knudson proposed his two-hit model to explain the mechanism of hereditary retinoblastoma[66]: one mutation in *RB1* is inherited, and the other occurs as a sporadic mutation in somatic cells of the retina. While the chance that a second mutation will occur in any one cell is low, the chance that they occur in at least one of the child's retinal cells is high because of their large population (about 100 million) and rapid proliferation. Occasionally, the form of the second mutation is epigenetic, such as methylation of the promoter leading to *RB1* silencing[67,68].

## Genomic Sequencing Methods

The first method developed to sequence DNA was the Sanger method (first-generation sequencing technology), first published in its final form in 1977[69]. Sanger sequencing was labor-intensive, costly, and inefficient. Still, Sanger sequencing was used to generate the first draft of the human genome through the Human Genome Project, which began in 1990 and finished in 2003[70].

In 2004, the first of many "next-generation sequencing" (NGS) methods debuted, called pyrosequencing[71]. Other methods followed, decreasing the cost of sequencing a human genome 50,000 fold when compared to the cost of the Human Genome Project[72].

NGS methods involve either PCR-based amplification or hybrid-capture targeting. The Illumina platform is currently the most commonly used amplification-based, short-read sequencing method[73]. DNA molecules are attached to a surface and amplified in situ[74]. Hybrid capture sequencing involves targeting specific loci or regions for sequencing rather than the entire genome. This method is commonly used in the clinic because it is more economical and allows for deeper sequencing coverage of the targeted regions[75]. In the research described in this thesis, a panel-based hybrid capture sequencing method was used on various tumor samples in Chapter 2.

### Long-read Sequencing

In addition to the short-read technologies mentioned above, long-read sequencing technologies such as Nanopore have been used to characterize genomic variants. Its advantages include longer reads (>10 kb on average), which allows for better characterization of repeat regions[76], and better characterization of longer variants such as CNVs[77,78]. Disadvantages include a higher error rate than other NGS methods[79]. In the research described in this thesis, Nanopore long-read sequencing was used in Chapter 3 to detect structural variants and methylation changes in a patient with Wilms tumor.

## Revolutions in the Diagnostics and Treatment of Cancer and Mendelian Disorders

Cancers have been classified by tissue characteristics and histopathology, or appearance under the microscope[80]. Reproducibility of histopathological classifications can be low in many tumor types[81–89]. For example, in one study, only 76% of patients with childhood brain tumors received the same diagnosis on two different readings by the same histopathologists[81].

Following the advent of NGS, the first whole cancer exome was sequenced in 2007, and the first whole cancer genome quickly followed in 2008[90]. Early whole exome sequencing (WES) cancer studies typically contained only 20 tumors at a cost of about $100,000 per case. The size of large genomic datasets has rapidly increased: a recent cancer study from the UK contained tumor samples from over 10,000 individuals in addition to their germline WGS data[91]. Detailed genomic characterization of tumors by NGS from large initiatives such as these has allowed us to identify and catalog the variants within[92–111] and across cancer types[112–118]. This knowledge has been used to create a new taxonomy of cancers, where many tumors are classified by genetic alterations[119] in addition to tissue of origin and histology.

The genetic vulnerabilities identified in these cancers have been used to develop targeted treatments. Targeted treatments are considered to be both more effective and less toxic than previous treatments. Following the discovery of mutations in the HER2 oncogene in a subset of breast cancer patients[120], the first genomic-based targeted cancer drug, trastuzumab (Herceptin), was approved to treat HER2 amplified breast cancer[121]. Patients with non-small-cell lung cancer are treated with inhibitors of EGFR, ALK and ROS1, depending on their mutations[122–124]. Melanoma patients may be treated with MAPK inhibitors if BRAF mutations are found[125,126].

However, WES and WGS are still not routinely used in all cancer patients. Barriers to widespread adoption include cost, turnaround time, and in some cases. clinical utility[122]. Pre-genomic methods such as in situ hybridization (ISH) and immunohistochemistry (IHC) are still commonly used in the clinic to identify relevant genetic alterations in tumors. [127]

In addition to cancer, NGS has also been applied to Mendelian genetic disorders. By 2009, primarily through the use of Sanger sequencing and linkage mapping, the gene or genes underlying only one third of Mendelian disorders had

been identified[128]. The arrival of “affordable exomes” was hailed in 2010, and by 2011, WES had been used to identify causal variants in at least “several dozen Mendelian disorders”[74], and many more variants in the following years[129–131]. WES was also successfully adapted for diagnostic purposes[132–137]. WGS is particularly helpful in diagnosing newborns who have been hospitalized, as the full clinical symptoms of a disorder may not manifest until later in life[138–140]. Long-read sequencing has been used to identify pathogenic mutations in numerous cases where short-read technologies could find no lesion[73,141–148]. The data generated by NGS changed the phenotypic spectrum considered to be diagnostic in several syndromes[149]. Identifying the pathognomonic mutation behind a disorder frequently leads to the discovery that the range of phenotypes associated with it are much broader than anticipated[150–153]. This is so common that there is a term for it: syndrome expansion.

The field of cancer predisposition syndromes has been profoundly affected by the advent of NGS. Where previously clinicians relied on individual tests for each gene, NGS allows for the testing of multiple genes or even the entire genome[63]. Currently, NGS is commonly used in the diagnosis of genetic disorders only when first line technologies return no result[154].

Interpreting NGS data may be a challenge, however. WGS, and to a lesser extent WES, produce many variants that require filtering and analysis. Exome sequencing will identify 20,000 – 24,000 SNVs on average, depending on ancestry[128]. The 1000 Genomes Project estimated that individuals carry about 2500 non-synonymous variants on average at conserved sites[155]. Other techniques include **1)** filtering out those variants which appear in databases of healthy individuals such as gnomAD and the Database of Genetic Variants (DGV), **2)** selecting for coding variants, especially those at conserved regions and **3)** prioritizing variants found to be pathogenic by other clinicians in databases such as ClinVar[156].

# Epigenetics

## DNA methylation

DNA methylation is the addition of a methyl group to DNA. When it occurs near a gene promoter, this methylation is associated with transcriptional repression[157]. Each cell in the body has the same number of genes in its DNA, more genes than any cell type will ever need. In order to regulate the expression of cell type-specific genes, many genes are methylated (and therefore silenced) early in embryogenesis[157,158]. The enzymes DNMT3A/3B and DNMT1 are responsible for depositing[159–162] and maintaining[163] DNA methylation, respectively.

DNA methylation occurs most commonly the CpG dinucleotide[164]. Regions with clusters of the CpG dinucleotide are called CpG islands[157,165]. Many CpG islands are associated with gene promoters. Most CpGs outside of CpG islands are methylated by default[159].

Methylation at CpG islands is often dysregulated in cancer[164]. Loss of DNA methylation, which is associated with gene activation, was the first epigenetic alteration found in cancer[166,167]. This hypomethylation has been seen in various cancers[168–171] including Wilms tumors[172], lymphomas[173], gastric cancers[174], and breast cancers[175].

Hypermethylation is seen in cancers as well. The promoters of tumor suppressor genes are frequently hypermethylated in tumors[176–178], including colorectal cancer[170,171,179,180] , prostate cancer[181], gastric cancer[182,183], and gliomas[184–190]. In sporadic retinoblastoma, *RB1* is frequently methylated[191]. Hypermethylation is also seen in the genomes of HPV and the Epstein-Barr virus in order to silence certain genes[166,192].

## Imprinting

In most cases, individuals inherit two copies of each gene, one from each parent. However, in the case of imprinted genes, genes are expressed only from one parent: either the paternal or maternal allele. First discovered in humans in the 1990s, currently over 200 imprinted genes are known in the human genome[193]. DNA methylation maintains allele-specific expression (the "imprint") by silencing one allele. Interestingly, imprinted regions share some similarities to endogenous retroviral DNA, including their repression by DNA methylation[194].

The first suggestion that imprinted genes might be found in humans followed the observation that Wilms tumors with loss of heterozygosity (LOH) always lost their maternal copy of the chromosome 11p15 alleles[195,196]. Additionally, Beckwith–Wiedemann syndrome (BWS), which predisposes children towards developing Wilms tumors, is only inherited when the affected parent is the mother[197]. The genes associated with BWS, H19 and IGF2, were the first imprinted genes identified in humans[198–202]. In the research described in this thesis, long-read sequencing was used in Chapter 3 to detect methylation changes in the imprinted genes H19 and IGF2 in a patient with Wilms tumor.

## Histone Modifications

Histone tails are subject to extensive post-translational marks, such as methylation, phosphorylation, and acetylation, which modify the expression of DNA. Histone methylation, like DNA methylation, is generally associated with transcriptional repression. One well known type of histone methylation is H3K27me3, or histone H3 lysine 27 trimethylation. H3K27me3 is essential for the regulation of cell type development and differentiation[203,204]. EZH2 is the catalytic subunit of Polycomb repressive complex 2 (PRC2)[205,206], which deposits H3K27me3[207,208].

It is not clear whether *EZH2* is a tumor suppressor or an oncogene. EZH2 is overexpressed in many cancers including breast[209–212], bladder[213], prostate[214],

endometrial[215], gastric[216], lung[217,218], pancreatic[219], and melanoma[220], and increased H3K27me3 has also been seen in prostate cancer[221], gastric cancer, lung cancer[222], and melanoma[223]. Confusingly, however, reduced levels were found in breast[210,224], bladder , and pancreatic cancers[224]. Both activating[226] and inactivating[227,228] mutations have been found in *EZH2* in B-cell lymphomas. About half of malignant peripheral nerve sheath tumors exhibit loss of H3K27me3[229]. Reduction of H3K27me3 is seen in many brain tumors, including meningiomas[230–232], ependymomas[233], oligodendrogliomas[234], and diffuse midline gliomas[235,236]. The reduction of H3K27me3 in diffuse midline gliomas is caused by the H3K27M mutation[235,237,238]. This mutation substitutes histone H3's residue K27 for methionine, which interferes with the deposition of H3K27me3. Notably, however, some loci in H3K27M gliomas gained the mark, particularly unmethylated CpG islands[239–241]. In the research described in this thesis, we examine the gene expression changes caused by the H3K27M mutation in diffuse midline gliomas.

The overall goal of my thesis is to uncover the causes of cancer using new genomic technologies. In Chapter 2, I will focus on detecting cancer-causing viruses in hybrid capture DNA samples from tumors associated with viral oncogenesis. I created an app that integrates into the UCSF 500 cancer genomic analysis pipeline.

In Chapter 3, I describe the detailed genomic analysis of a patient with bilateral Wilms tumor who was suspected to have a cancer predisposition syndrome. Previous analyses using short-read sequencing and methylation testing detected no causal variants in the patient's whole blood sample. This project was a collaboration with Dr. Vivian Chang at UCLA, and Miten Jain, Jean Monlong, Holly Beale and Olena Vaske at UCSC. Miten Jain and Jean Monlong led the sequencing and variant calling. Thousands of variants were identified, and I led the filtering, annotation, and interpretation of variants. I identified the epimutation putatively responsible for the patient's Beckwith-Wiedemann syndrome. Portions of the

introduction were previously published in an abstract that I wrote in *Cancer Research* in 2021[242].

In Chapter 4, we use gene expression data from a large cohort of diffuse midline gliomas to analyze the developmental origin of H3K27M tumors and the role of the Epithelial-Mesenchymal Transition (EMT). I led this study. I performed the analysis of the gene expression data that found the EMT pathway was upregulated, and I noted the expression of the EMT genes. I performed the experiments validating expression of these genes in glioma cell lines. This manuscript was published in *Giga Science* in 2020. Lauren Sanders was co-first author. Lauren Sanders performed the analysis of the single-cell data and developed the EMT score. Lucas Seninge contributed the analysis of the organoid data. Anouk van den Bout additionally performed RT-PCR on glioma cell lines to quantify expression of EMT genes.

# Chapter Two. Repurposing an Existing Gene Panel for Detecting Viral Sequences in Cancer

## Introduction

In 2009, the International Agency for Research on Cancer identified viruses classified as carcinogenic to humans[243,244]. Infections from these viruses account for 15-20% of cancer cases globally[245].

| Virus | Tumor types |
|---|---|
| Epstein-Barr Virus (EBV) | Nasopharyngeal carcinoma<br>Gastric carcinoma<br>Lymphomas |
| Human Papillomavirus (HPV) | Cervix uteri carcinoma<br>Oropharyngeal carcinoma<br>Penis carcinoma |
| Human Herpesvirus Type 8 (HHV8) | Kaposi's sarcoma |
| Hepatitis B Virus (HBV) | Hepatocellular carcinoma |
| Hepatitis C Virus (HCV) | Hepatocellular carcinoma<br>Non-Hodgkin's lymphoma |
| Human T-cell Lymphotropic Virus (HTLV) | T-cell leukemia and lymphoma |
| Merkel Cell Virus (MCV) | Merkel cell carcinoma |

Table 1: Human Oncogenic Viruses and the Tumor Types They Cause.
Adapted from "Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis"[245].

The first oncovirus was discovered in chicken leukemias by Ellerman and Bang in 1908[246]. This was largely ignored by researchers because at that time leukemia was not thought to be a type of cancer. Peyton Rous's discovery of RSV in chicken sarcomas was better appreciated, yet 20 years would pass before the next oncoviruses were found, in rabbits[247], mice[248], and frogs[249], not humans. The first human oncovirus discovered was Epstein-Barr virus[250]. In 1958, the surgeon Denis Burkitt noticed that a particular type of pediatric lymphoma was only seen in certain geographic regions[251]. This regional specificity suggested an infectious agent. While Burkitt initially theorized that mosquitoes were spreading the infection, the virologists Anthony Epstein and Yvonne Barr isolated the responsible oncovirus that now bears their names[252]. EBV DNA was soon found in Burkitt's lymphomas as well as nasopharyngeal carcinomas[253]. The role of EBV in cancer was proven following the induction of lymphomas in marmosets[254] and owl monkeys[255] in 1973. These discoveries received little attention from the field of oncology, however[250].

In the 1970s, the field of viral oncology was embarrassed by a large number of irreproducible studies from the labs of Sol Spiegelman, Robert Gallo, and others, reporting retroviral infections in virtually every cancer type they checked[24,25,35]. Many of these mistakes were caused by "cross contamination with an animal retrovirus"[256]. This was followed by a period of "intense antagonism towards research directed toward finding human tumor viruses"[256]. The role of viruses in causing cancer in humans was not widely accepted until the discoveries of Human papillomaviruses (HPV) and the hepatitis viruses in the 1980s.

The viral etiology of hepatitis and liver cancer was first theorized in the 1956[250,257]. Numerous epidemiological studies of these conditions linked Hepatitis B virus (HBV) to hepatocellular carcinomas[258–262]. The clearest evidence came from a study of over 22,000 people that showed that hepatitis B infection increased the risk of developing hepatocellular carcinoma over a hundred-fold[263].

A viral etiology for cervical cancer was long suspected. Unfortunately, for over a decade researchers focused on a role for Herpes simplex virus 2 (HSV-2). In the late 1960s, a report noted increased antibody levels against HSV-2 in cervical cancer patients[264], and a study of 245 patients with herpes infections found 12 cases of cervical cancer[265]. Antigens for HSV-2 were even proposed as cervical cancer markers[266,267]. HSV-2 RNA and DNA was found in one cervical cancer sample[268]. However, a large study found no HSV-2 antigens in the blood of cervical cancer patients[269], and other researchers could not reproduce the finding of HSV-2 DNA in cervical tumors.

In 1983, HPVs were finally linked to cancer when HPV16 was found in the majority of cervical cancer samples tested[270], followed by the discovery of HPV18[271,272]. HPV DNA is sufficient to transform cell lines[273,274] and induce tumors in mice[275–283]. The key role of the viral oncogenes E6 and E7 was quickly identified. Large case control studies have confirmed the oncogenicity of HPV[284–286]. The HSV-2 reports highlight the inherent difficulties in viral detection. HSV-2 may have been detected because of contamination or the partial homology of HSV-2 and HPV sequences. Additionally, some patients may truly have been co-infected with both HPV and HSV-2.

The first human oncogenic retrovirus, Human T-lymphotropic virus (HTLV), was found by the lab of Robert Gallo in a T-cell lymphoma in 1980[287,288], and this was quickly reproduced by researchers in Japan[289,290]. HTLV-1 causes adult T-cell leukemia, a rare malignancy which was seen almost exclusively in patients from Kyushu, an island of Japan[290,291], though it has since spread[292]. Despite the critical role of retroviruses such as RSV in discovering the mechanisms of oncogenesis, HTLV is the only retrovirus that has ever been shown to cause cancer in humans.

The most recently discovered oncovirus, Merkel cell virus (MCV), was found in 2008 in a rare form of neuroendocrine carcinoma, Merkel cell carcinoma. Like EBV and HPV, MCV infections are widespread[293,294]. It is interesting that the majority

of known oncoviruses are widespread or very common in some regions. This may reflect the difficulty in detecting rare oncoviruses, or perhaps oncogenicity confers an advantage in transmissibility.

## Epstein-Barr Virus

Epstein-Barr Virus (EBV) is a double stranded DNA virus that infects 90% of the world's population [295]. Like most members of the herpesvirus family, after infection, it is carried as an asymptomatic latent infection in the majority of people[296]. However, it can cause a wide range of conditions from multiple sclerosis[297] to mosquito bite hypersensitivity[298] in addition to cancers in adults and children. EBV is responsible for 300,000 cases of cancer per year globally, including Hodgkin's and non-Hodgkin's lymphomas, gastric carcinomas[299], and nasopharyngeal carcinomas[300]. It is estimated to have caused as many as 200,000 deaths in 2020[300].

EBV also has a causative role in pediatric Burkitt's lymphomas (BL)[252], though its detection rate in tumors varies by geographic region. It is detected in about 30% of Burkitt's lymphoma cases in the US[301] versus virtually all cases of BL in equatorial Africa[302,303] and New Guinea[304,305]. Widespread coinfection with malaria in these regions is believed to contribute to the oncogenesis of Burkitt's lymphoma[306].

The EBV genome is about 170kb long[307] and contains about 100 genes, but the 9 "latent genes" are most relevant in cancer cells: EBNA1, EBNA2, EBNA3A, EBNA3B, EBNA3C, EBNALP, LMP1, LMP2A, and LMP2B[308]. These genes influence cell proliferation, cell motility, metastasis, and immune response. The expression of these latent genes differs by malignancy: for example, Burkitt's lymphomas express only EBNA1, while Hodgkin's and diffuse large B cell lymphomas express EBNA1, LMP1, LMP2A, and LMP2B[309]. These expression patterns could be adapted for diagnostic use. In patients with EBV+ diffuse large B cell lymphoma, the expression

of EBNA2 and EBNA3 should prompt an investigation into underlying immunosuppression, according to the WHO[310–312]. The presence or absence of EBV in tumor cells is relevant clinically: 9 disease entities in the 2022 WHO classification of lymphoid neoplasms are defined primarily by its presence[312]. Additionally, higher amounts of EBV DNA detectable in the plasma of cancer patients correlates with an increased risk of recurrence or metastasis[313], but low interlaboratory concordance has limited the ability to establish clear cutoffs[314]. The presence of EBV DNA in plasma was successfully used to screen for nasopharyngeal cancer. Routine methods for EBV detection include IHC for the LMP1 protein and ISH for EBERs (EBV-encoded RNAs)[315,316]; the latter being described as the "gold standard"[317] despite its relative lack of sensitivity compared to PCR[315].

## Human Papillomavirus

Human Papillomavirus (HPV) causes 5% of all cancer cases globally[318], approximately 730,000 cases of cancer per year[245]. HPV was associated with 342,000 deaths in the year 2020[245,319,320]. HPV causes virtually all cases of cervical cancer and 70% of oropharyngeal cancers in the US[321,322].

There are at least 200 known strains or types* of HPV[323]. Only 12 are considered to be "high-risk" (HR-HPV) for cancer[243]. Of these, HPV16 and HPV18 alone cause about 70% of cervical cancer[324]. The risk of cervical cancer among HPV16+ and HPV18+ individuals is 17% and 14%, respectively[325]. The remaining types of HPV are considered "low-risk" for cancer. Approximately 10 HPV types are associated with genital warts called condylomas[326]. HPV6 and HPV11 alone cause about 90% of condylomas[197,327].

* HPV sequences are traditionally known as "types" within the HPV field but referred to as "strains" in publications of International Committee of Viruses (ICTV). "Type" is used throughout this document.

The HPV genome is 7-8 kbp long and has four to eight genes, depending on the type. The L1, L2, E1 and E2 genes are highly conserved, while E6 and E7 are oncogenes, found in only high-risk HPV types. While all of these genes can be found in some HPV-infected cancer cells, E6 and E7 are the only genes consistently expressed[328]. E6 targets p53, a critical tumor suppressor, for degradation[329–331]. Similarly, E7 is believed to initiate oncogenesis by inactivating the tumor suppressor RB1 and targeting it for degradation[328]. Additionally, E7 proteins epigenetically reprogram host cells by dramatically decreasing the amount of histone H3 lysine 27 trimethyl (H3K27me3) marks, which silence gene expression[332]. Both E6 and E7 contribute to EMT[333–335]. The L1 sequence is used to determine if a new HPV isolate sequence is a new type[336]. If the L1 sequence differs by 10% or more from the next closest type, it is considered a new type.

In 2020, at least 254 commercially available HPV detection methods were available, the majority of which were not clinically validated. Only 7 of these were FDA approved[337]. About 10-25% of HPV infections are missed[338,339] depending on the detection method. The most popular methods include PCR and hybridization. The choice of PCR primers emphasizes a trade-off between type-specificity and sensitivity. The most commonly used PCR primers include those targeting a conserved region in the L1 gene, or type-specific primers targeting E6 or E7[340]. Type-specific PCR detection of the E7 gene can detect HPV in many cases that L1 primers miss[341], possibly due to deletion of L1 upon integration into the host genome[342].

The presence of HPV is used as a prognostic factor in oropharyngeal squamous cell carcinoma (OPSCC)[343]. Remarkably, current ASCO guidelines for HPV detection in OPSCCs recommend the use of surrogate marker for HR-HPV, p16, via immunohistochemistry[344] rather than detecting the HPV sequence itself. However, p16 IHC has a 20-28% false positive rate for HPV in OPSCCs[345].

Some studies have found that HPV positive samples have a better prognosis than “HPV negative” samples, but it is not clear that there are any truly HPV negative cervical cancers[342]. Most, perhaps all, “HPV negative” samples are samples in which the HPV DNA has integrated into the genome in such a way that the L1 is deleted[339,341,342].

## Other Oncogenic Viruses

### Hepatitis Viruses

Hepatitis B and C viruses are associated with 80-90% of hepatocellular carcinomas (HCC)[245]. Despite their names, these viruses are not closely related[346]. Hepatitis C virus (HCV) is a single-stranded RNA virus[347], though not a retrovirus. About 3% of those infected with HCV will develop HCC[348]. Hepatitis B virus (HBV) is a DNA virus and 22% of chronic carriers of HBV will develop HCC in their lifetime[349]. Viral load is relevant clinically: the risk of developing HCC increases with increasing HBV viral load[350,351]. A higher viral load in HCC is also associated with a worse prognosis[352]. Antiviral therapy is crucial in the treatment of cancers caused by HBV[353], therefore detecting hepatitis viruses is critical in treating HCC.

### Polyomaviruses

Polyomaviruses can induce tumors in animal models, and therefore have long been suspected as human oncogenic agents. Four polyomaviruses have been associated with human cancers: BKV, JCV, SV40, and Merkel Cell Virus (MCV). Evidence for MCV’s causative role in cancer is the strongest of all the polyomaviruses, although the transforming capacity of MCV is not yet proven. MCV infections are widespread globally[293,294], and MCV is detected in 80% of Merkel cell carcinomas, a rare but highly aggressive type of skin cancer[354–357]. Despite extensive searches, MCV has not been linked to any other cancer type.

### HTLV-1

HTLV-1 is the only human retrovirus associated with cancer. Like the cancer it causes, HTLV-1 is rare, and was initially found almost exclusively in patients from the island of Kyushu, Japan[256]. It causes adult T-cell leukemias (ATL), a highly aggressive malignancy[358]. Of those infected with HTLV-1, about 5% will be diagnosed with ATL[359].

### HHV8 (KSHV)

Human herpesvirus 8 (HHV8), also known as Kaposi's sarcoma-associated herpesvirus (KSHV), is a double-stranded DNA virus[360]. While endemic in sub-Saharan Africa, it is found in only 1% of the population of the United States[360]. HHV8 was first identified as the cause of Kaposi's sarcoma, and has more recently been associated with various lymphomas, including diffuse large B-cell lymphoma[361] and primary effusion lymphoma[362]. HHV8 is detected in almost 100% of Kaposi's sarcomas[363]. The majority of Kaposi's sarcomas are found in immunocompromised individuals, especially those with AIDS.

Detection of HHV8 in leukemias and lymphomas is relevant diagnostically; confirmation of its presence is necessary to diagnose 8 lymphoproliferative disorders[360].

## Missteps in Oncogenic Viral Discovery

In the 1970s, numerous scientists searching for novel human retroviruses in cancer samples were misled by the endogenous retroviral DNA that makes up an estimated 8% of the human genome. These sequences, and similar retroviral DNA found in mice and other species, led to a number of false oncovirus discoveries in that era[365]. The advent of NGS has led to a similar number of false positives more recently. For example, the Cancer Genome Atlas (TCGA) is a large, public database of tumor samples[366], commonly used to detect novel oncogenic viruses. However, care must

be taken to eliminate technical artifacts and contaminants: numerous samples in TCGA are contaminated with bacteriophage DNA, viral vector DNA, and several samples are cross-contaminated[367]. In 2013, several studies detected HPV18 in cases of colon, rectal, or stomach adenocarcinoma in TCGA[368,369]. These cancers had not been previously associated with HPV. Analysis showed that these samples were contaminated by RNA or DNA from HeLa cells, which are known to be infected with HPV18[370]. The contamination was limited to samples from 2 specific TCGA sequencing centers and a limited number of specific dates in 2011 and 2012. Cross-contamination from samples at one of these sites also lead to false positive results involving HPV38 in another study[371].

More recently, a large-scale, widely publicized analysis of tumor samples from TCGA found that distinct microbial signatures were detected in 32 of 33 cancer types[372]. However, reanalysis showed that the vast majority of microbial reads detected were in fact human[373]. This error is possible in part because many sequences in GenBank and other popular sequence databases are contaminated or mislabeled [374,375]. For robust results, careful selection of reference genomes is necessary.

False positives in oncogenic viral detection are not limited to TCGA. In 2006, a novel virus, named xenotropic murine leukemia virus-related virus (XMRV), was detected in 10% of prostate cancer tumors in a cohort of 86 patients[376] at the Cleveland Clinic. It was considered plausible that XMRV might be oncogenic because the XMRV sequence was similar to murine leukemia viruses (MLVs) that are known to cause cancer in mice. In follow-up experiments, several laboratories detected XMRV in prostate cancer samples[377–381], but others were unable to replicate this[382–384]. In 2010, Oakes, *et al*. suggested that XMRV detection might be attributable to contamination from mouse DNA, as MLVs are widespread and commonly integrate into the mouse genome[385]. Supporting this, an RT-PCR kit was found to be contaminated with MLV DNA[386]. Subsequent studies noted that prostate

cancer samples which were positive for XMRV also tested positive for mouse DNA contamination[387–390]. Elegant analysis showed that the contamination of prostate cancer samples in the original study was due to RNA contamination from XMRV-infected cell lines (LNCaP and 22Rv1) at the Cleveland Clinic[391,392].

## Bioinformatic Tools for Viral Detection and Annotation Using Sequence Data

Next generation sequencing is used clinically to evaluate genomic mutations in neoplasms, but adoption for the purposes of microbial detection has been slow. A substantial number of programs exist to detect viral DNA in NGS samples. The majority are intended for the discovery of new oncogenic viruses or bacteria in research (metagenomics), not for clinical diagnostics. Many are computationally expensive, difficult for non-experts to use, unvalidated, or poorly maintained[393]. These methods vary in sensitivity and specificity, and false positive results are common[394–397].

### Assembly versus Alignment

Viral detection pipelines utilize either reference-based assembly, *de novo* assembly, or both. Reference-based assembly requires a pre-made reference genome or the use of viral sequence databases. *De novo* assembly can be used to discover novel viruses but may not be useful in diagnostic scenarios and increases the runtime.

### Viral Reference Genome Selection

Most viral detection pipelines use NCBI RefSeq or GenBank as their reference databases. As of February 2024, GenBank currently contains over 13 million viral sequences, over 11 million of which are from human hosts. Many of these sequences are incomplete, unannotated, or contaminated, and the majority are unvalidated. RefSeq contains only validated sequences but is missing a significant number of relevant viruses: it contains only 375 viral sequences taken from human

hosts. The massive size of GenBank will increase computation time for any pipeline that uses it as a reference genome.

I developed the ARCV (Analysis of Reads for Carcinogenic Viruses) app to detect known human oncogenic viruses in tumor samples sequenced by NGS. The app is designed to integrate with the results of the existing UCSF pipeline on the DNAnexus platform. ARCV was tested on known HPV+ and EBV+ tumor samples. Tumor samples with other viruses were not tested due to a lack of confirmed data. Here I describe a method for detecting viral sequences in UCSF 500 panel data and I evaluate the performance on real tumor and non-tumor specimens.

# Implementation

ARCV is written in bash and implemented for DNAnexus. ARCV is fully automated, with the option to input custom specifications, and high-throughput. The app is integrated with the existing UCSF analysis pipeline. ARCV removes human aligned sequences, aligns to a custom viral reference genome, and reports detailed viral coverage information, including viral gene coverage. ARCV was designed to be used for clinical/diagnostic purposes.

## Viral Reference Genome

A custom reference file was created consisting of 139 complete, annotated viral reference sequences from sequences from the Papillomavirus Episteme (PaVE)[398], RefSeq, and GenBank (Table 2, Supplementary File 1), including 109 types of HPV, 10 types of polyomavirus, 3 types of HTLV, Hepatitis B, Hepatitis C, EBV and HHV8. HPV genomes were derived from the Papillomavirus Episteme database. Other viruses were selected from RefSeq when a well annotated sequence was available. Annotated sequences were manually selected from GenBank. A custom viral gene bed file was used to generate gene coverage information. These gene annotations were derived from the original reference sequences.

Users may easily replace the reference genome and viral gene files with their own files. ARCV is available at https://github.com/allisoncheney/ARCV/.

## Table 2: Viral Sequences Included in the Reference File of ARCV and their Source Databases.

| virus | general type | source | accession |
|---|---|---|---|
| EBV | EBV | RefSeq | NC_007605.1 |
| HTLV type1 | HTLV | GenBank | MH399769.1 |
| HTLV type1 | HTLV | GenBank | J02029.1 |
| HTLV type2 | HTLV | GenBank | Y14365.1 |
| HTLV type3 | HTLV | GenBank | EU649782.1 |
| Hepatitis B | Hepatitis | GenBank | MH818373.1 |
| Hepatitis B | Hepatitis | GenBank | MT114172.1 |
| Hepatitis B | Hepatitis | GenBank | MT437386.1 |
| Hepatitis B | Hepatitis | GenBank | MN683729.1 |
| Hepatitis C | Hepatitis | GenBank | AF165053.1 |
| Hepatitis C | Hepatitis | GenBank | AF207754.1 |
| Hepatitis C | Hepatitis | GenBank | MT212178.1 |
| Hepatitis C | Hepatitis | GenBank | LC368448.1 |
| Hepatitis C | Hepatitis | GenBank | NC_009824.1 |
| Hepatitis C | Hepatitis | GenBank | D84263.2 |
| Hepatitis C | Hepatitis | GenBank | MH155319.1 |
| Hepatitis C | Hepatitis | GenBank | D63822.1 |
| Hepatitis C | Hepatitis | GenBank | MG428679.1 |
| Hepatitis C | Hepatitis | GenBank | MG406988.1 |
| HHV8 | human herpes virus | RefSeq | NC_009333.1 |
| WU | human polyomavirus | RefSeq | NC_009539.1 |
| MW | human polyomavirus | RefSeq | NC_018102.1 |
| BK | human polyomavirus | RefSeq | NC_001538.1 |
| MCV | human polyomavirus | RefSeq | NC_10277.2 |
| HPyV 9 | human polyomavirus | RefSeq | NC_15150.1 |
| HPyV 7 | human polyomavirus | RefSeq | NC_14407.1 |
| HPyV 6 | human polyomavirus | RefSeq | NC_14406.1 |
| HPyV 8 | human polyomavirus | RefSeq | NC_14361.1 |
| JC | human polyomavirus | RefSeq | NC_001699.1 |
| KI | human polyomavirus | RefSeq | NC_009238.1 |
| 444 HPV types | HPV | PaVE consortium | see appendix |

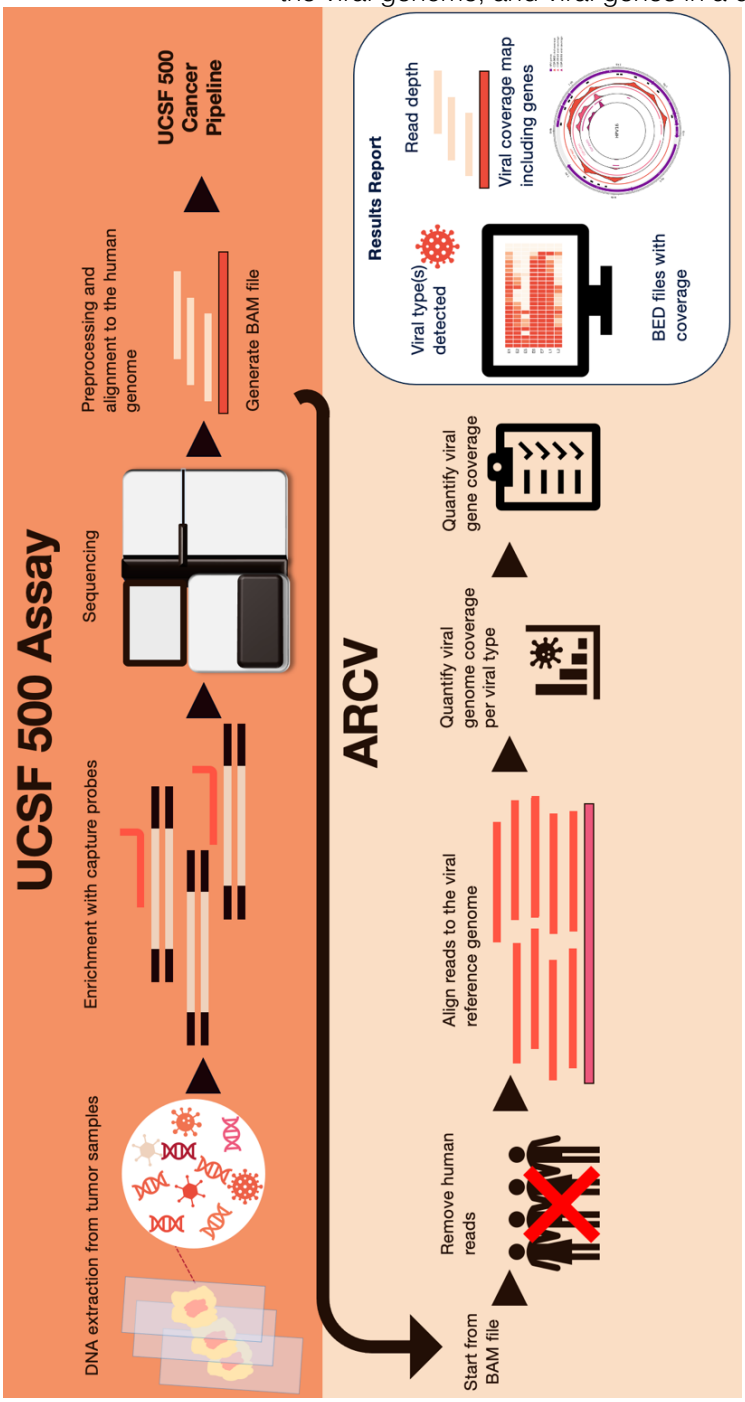
## Viral Detection and Quantitation Workflow

The ARCV workflow is shown in **Figure 1**.

The app takes as input BAM files produced by the UCSF 500 pipeline containing deduplicated reads mapped to the human reference genome (GRh37/hg19). In the first step, reads not mapped to the human genome are extracted using samtools[399] view (samtools version: 1.10-3) and then sorted using samtools sort. The resultant BAM file is converted into fastq files using samtools fastq. The fastq files are aligned to the viral reference file using BWA-MEM (BWA version: 0.7.17-4)[400]. The aligned reads are converted to BAM files, sorted by samtools sort. The user is given the option to filter out reads with a low mapq score (default of 10). If the total number of viral reads are below the customizable cutoff, the sample is reported negative for viral reads. To generate viral genotype information, samtools idxstats is used. Next, bedtools[401] genomecov (bedtools version: 2.27.1) in combination with custom scripts are used to generate detailed viral read sequence depth by region as well as depth by position. Coverage statistics generated include the minimum coverage level, maximum coverage level and average coverage depth per viral type detected, as well as the percent of the viral sequence at each level of coverage. Finally, using bedtools coverage and the custom viral gene bed file, viral gene read statistics are generated including which genes have coverage, the fraction of the gene with coverage, and the read count per gene. Custom scripts package this viral read information into an easily interpretable format. A text report is generated containing four files: a summary file giving basic viral read information (**Figure 2**), a file with coverage statistics, a gene coverage results file, and a bed file with detailed coverage information per base.

# Figure 1: ARCV Viral Detection Workflow.

This is a workflow of the ARCV viral detection app, which is part of the UCSF 500 assay. Top panel: The UCSF 500 assay includes the following steps: DNA extraction, enrichment with cancer probes, sequencing, preprocessing and alignment to the human genome. Bottom panel: The ARCV app receives BAM files as input and includes the following steps: human read subtraction, alignment to the viral reference genomes, and quantification of viral coverage and viral gene coverage. The report includes viral types detected, the coverage of the viral genome, and viral genes in a dataset.



# Figure 2: Example Output from the ARCV App for an HPV+ Sample.

## Results summary

sample ID: CGP24666
total reads: 47370745
reads that mapped to hg19: 46963417
reads that did not map to hg19: 407328
viral-mapped reads: 57736

### Sequence mapping and coverage

| sample name | virus name | mapped reads | lowest coverage depth | highest coverage depth | mean coverage depth | percent with no coverage | percent < 50X coverage | percent <100X coverage | percent >100X coverage |
|---|---|---|---|---|---|---|---|---|---|
| CGP24666 | HPV16 | 57736 | 0 | 9981 | 715.27 | 4.20% | 59.80% | 64.80% | 31.00% |

### Viral gene coverage

| virus name | gene name | read count | bases covered | percent of bases covered | gene length | start position | end position |
|---|---|---|---|---|---|---|---|
| HPV16 | L1 | 2287 | 1596 | 100.00% | 1596 | 5558 | 7154 |
| HPV16 | L2 | 143 | 1307 | 91.91% | 1422 | 4234 | 5656 |
| HPV16 | E6 | 33198 | 477 | 100.00% | 477 | 82 | 559 |
| HPV16 | E7 | 12840 | 297 | 100.00% | 297 | 561 | 858 |
| HPV16 | E1 | 14287 | 1938 | 99.44% | 1949 | 864 | 2813 |
| HPV16 | E2 | 98 | 958 | 87.25% | 1098 | 2754 | 3852 |
| HPV16 | E4 | 20 | 234 | 81.25% | 288 | 3331 | 3619 |
| HPV16 | E5 | 26 | 210 | 88.61% | 237 | 3862 | 4099 |

## Patient Cohort

The tumor samples tested included 25 UCSF cervical cancer samples (23 known to be HR-HPV positive by previous ISH testing and 2 negative controls) and 18 lymphomas (13 known to be EBV positive and 5 negative controls), for a total of 43 tumor samples. I was blinded to EBV or HPV status of these samples when performing my analysis.

An additional 25 condyloma (non-tumor) samples were run through ARCV. They had not been tested for HPV by any previous methods including ISH, but were presumed to be positive for HPV6 or HPV11.

## Targeted Gene Sequencing

Hybrid capture-based NGS was performed on formalin-fixed paraffin embedded (FFPE) tissue samples at the University of California, San Francisco Clinical Cancer Genomics Laboratory, using a clinically validated targeted sequencing panel, the UCSF 500 Cancer Gene Panel, as previously described[402–404]. UCSF 500 targets over 500 human variants relevant to cancer, in addition to containing 1069 probes derived from viruses known to cause cancers in humans (Table 3). The viral probes were designed using the MSSPE algorithm[405] in order to capture as many divergent viral sequences as possible.

### Table 3: Viral Probes in the UCSF 500 Cancer Gene Panel.

Number of probes per virus, according to labels.

| Virus | Number of probes |
|---|---|
| EBV | 22 |
| HPV16 | 20 |
| HPV18 | 15 |
| Other high risk HPV types | 855 |
| Human polyomaviruses | 100 |
| HTLV-1 | 17 |
| HHV4 | 18 |
| Hepatitis B virus | 8 |
| Hepatitis C virus | 14 |

# Results

## UCSF 500 HPV Viral Probes are Specific to High-Risk HPV Types.

At the time I was assigned this project, little was known about the development of the viral probes. I had access to a fasta file containing the sequence of the viral probes and a corresponding label, i.e. “HPV16_1” or “HPV_20”. 20 probes were labeled “HPV16”, 15 were labeled “HPV18”, and 855 were labeled as HPV but had no strain specified.

To determine which specific viral types the probes were targeting, I ran BLAST[406] on all probe sequences against both my reference genome file and all human viruses in GenBank. Of the 15 probes labeled “HPV18”, only 8 mapped to HPV18. 7 mapped to other HPV types (including HPV16, HPV45, HPV84, HPV39, and HPV6). Similar results were found for the HPV16 probes: only ten aligned to HPV16. All 855 of the remaining HPV probes aligned to high-risk HPV types. No probes mapped to low-risk strains.

After contacting the original designer of the probes, he explained that the probes "were not designed to be specific for individual viral strains but were designed as highly conserved probes to broadly capture as many divergent viral sequences as possible.” The probes were not designed with diagnostic purposes in mind, but rather novel oncogenic viral discovery. Thus, it may not be possible to accurately detect all HPV viral types with the current probes.

## HPV and EBV are Detected by a Targeted Sequencing Panel

ARCV successfully detected HPV in 23 known HPV-positive samples (Table 4) and detected EBV in 13 known EBV-positive samples (Table 5). No viral reads were detected in 7 negative control samples. The number of viral reads detected in a sample varied, ranging from 179 to 2 million, with a mean of 177,426.

Of the HPV positive samples, the majority (21) had reads aligned to HPV16, one sample aligned to HPV18, and one sample aligned to HPV56. Notably, the read count for the HPV18 sample was very low (178).

# Table 4: ARCV Results of 25 Cervical Cancer Samples: Basic Viral Read Counts.

| Sample ID | Cancer type | HPV status by ISH | Virus detected by ARCV | HPV read count | Viral read percent of total reads |
|---|---|---|---|---|---|
| CGP19306 | Cervical | HR HPV positive | HPV16 | 6,291 | 0.0152% |
| CGP20349 | Cervical | HR HPV positive | HPV16 | 2,433 | 0.0030% |
| CGP20350 | Cervical | HR HPV positive | HPV16 | 37,249 | 0.1114% |
| CGP20351 | Cervical | HR HPV positive | HPV16 | 2,505 | 0.0071% |
| CGP20541 | Cervical | HR HPV positive | HPV16 | 459,052 | 0.5276% |
| CGP20542 | Cervical | HR HPV positive | HPV16 | 16,864 | 0.0157% |
| CGP20543 | Cervical | HR HPV positive | HPV16 | 33,907 | 0.0677% |
| CGP20544 | Cervical | HR HPV positive | HPV16 | 171,606 | 0.3242% |
| CGP23573 | Cervical | HR HPV positive | HPV16 | 2,656 | 0.0081% |
| CGP23591 | Cervical | HR HPV positive | HPV16 | 15,270 | 0.0393% |
| CGP24465 | Cervical | HR HPV positive | HPV16 | 603 | 0.0015% |
| CGP24666 | Cervical | HR HPV positive | HPV16 | 57,730 | 0.1219% |
| CGP25458 | Cervical | HR HPV positive | HPV16 | 162,822 | 0.3466% |
| CGP26453 | Cervical | HR HPV positive | HPV16 | 48,603 | 0.1494% |
| CGP27684 | Cervical | HR HPV positive | HPV16 | 21,138 | 0.0582% |
| CGP28096 | Cervical | HR HPV positive | HPV16 | 7,611 | 0.0155% |
| CGP28876 | Cervical | HR HPV positive | HPV16 | 432,687 | 1.0386% |
| CGP29919 | Cervical | HR HPV positive | HPV16 | 80,407 | 0.1473% |
| CGP30431 | Cervical | HR HPV positive | HPV16 | 86,859 | 0.1899% |
| CGP31034 | Cervical | HR HPV positive | HPV16 | 750,743 | 1.9266% |
| CGP9680 | Cervical | HR HPV positive | HPV16 | 2,094,635 | 5.4875% |
| CGP26780 | Cervical | HR HPV positive | HPV18 | 178 | 0.0005% |
| CGP18006 | Cervical | HR HPV positive | HPV56 | 13,060 | 0.0258% |
| CGP19698 | Cervical | HR HPV negative control | no virus detected | 0 | 0.0000% |
| CGP22291 | Cervical | HR HPV negative control | no virus detected | 0 | 0.0000% |

## Table 5: ARCV Results of 15 Lymphoma Samples: Basic Viral Read Counts.

| Sample ID | Cancer type | Viral ISH results | Virus detected by ARCV | Viral read percent of total reads | EBV read count |
|---|---|---|---|---|---|
| CGP14605 | Lymphoma | EBV positive | EBV | 0.005846 | 2,252 |
| CGP15944 | Lymphoma | EBV positive | EBV | 0.906719 | 451,438 |
| CGP17765 | Lymphoma | EBV positive | EBV | 0.591918 | 288,189 |
| CGP19553 | Lymphoma | EBV positive | EBV | 0.063132 | 27,642 |
| CGP21070 | Lymphoma | EBV positive | EBV | 0.055786 | 18,151 |
| CGP23360 | Lymphoma | EBV positive | EBV | 0.013264 | 6,469 |
| CGP23819 | Lymphoma | EBV positive | EBV | 0.406916 | 226,106 |
| CGP26223 | Lymphoma | EBV positive | EBV | 0.079863 | 37,249 |
| CGP29936 | Lymphoma | EBV positive | EBV | 0.233148 | 141,960 |
| CGP30202 | Lymphoma | EBV positive | EBV | 0.446120 | 150,485 |
| CGP30527 | Lymphoma | EBV negative control | no virus detected | | 0 |
| CGP30899 | Lymphoma | EBV negative control | no virus detected | | 0 |
| CGP26254 | Lymphoma | EBV negative control | no virus detected | | 0 |
| CGP26777 | Lymphoma | EBV negative control | no virus detected | | 0 |
| CGP28135 | Lymphoma | EBV negative control | no virus detected | | 0 |

## Viral Genome Coverage

Coverage of the HPV16 viral genome varied by sample (Table 6, Figure 3), ranging from 100% coverage to 40% coverage. Read depth was highest in regions covered by UCSF 500 capture probes (Figure 3), and lower or non-existent in regions far from probe coverage. However, several probes labeled as “HPV16” had low sequence identity with the HPV16 reference genome. These probes did not appear to impact coverage at their locations. Notably, a probe labeled “HPV18” aligned to HPV16 with 100% sequence identity and there is an identifiable peak in coverage corresponding to that region.

Only one sample was HPV56+, and 48% of the HPV56 genome was covered (Figure 4). Coverage for the single HPV18+ sample was lowest, at 22% (Supplementary Figure 1).

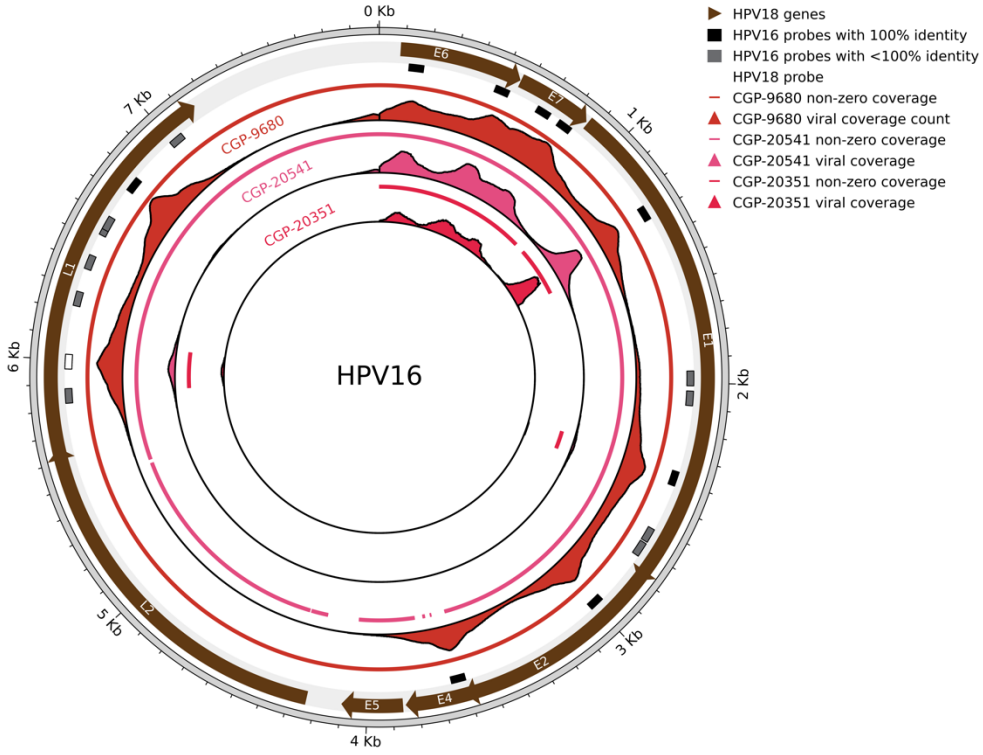
# Table 6: HPV Coverage Results of 25 Cervical Cancer Samples Run on ARCV.

| Sample ID | Cancer type | HPV status by ISH | Virus detected by ARCV | HPV read count | Viral read percent of total reads | Highest coverage depth | Mean coverage depth | Lowest coverage depth | Percent of HPV genome with no coverage |
|---|---|---|---|---|---|---|---|---|---|
| CGP19306 | Cervical | HR HPV positive | HPV16 | 6,291 | 0.0152% | 663 | 78.51 | 0 | 36% |
| CGP20349 | Cervical | HR HPV positive | HPV16 | 2,433 | 0.0030% | 308 | 30.34 | 0 | 53% |
| CGP20350 | Cervical | HR HPV positive | HPV16 | 37,249 | 0.1114% | 6,995 | 465.68 | 0 | 35% |
| CGP20351 | Cervical | HR HPV positive | HPV16 | 2,505 | 0.0071% | 635 | 30.57 | 0 | 60% |
| CGP20541 | Cervical | HR HPV positive | HPV16 | 459,052 | 0.5276% | 61,474 | 5769.96 | 0 | 1% |
| CGP20542 | Cervical | HR HPV positive | HPV16 | 16,864 | 0.0157% | 2,149 | 213.23 | 0 | 30% |
| CGP20543 | Cervical | HR HPV positive | HPV16 | 33,907 | 0.0677% | 5,227 | 427.36 | 0 | 48% |
| CGP20544 | Cervical | HR HPV positive | HPV16 | 171,606 | 0.3242% | 38,698 | 2134.24 | 0 | 1% |
| CGP23573 | Cervical | HR HPV positive | HPV16 | 2,656 | 0.0081% | 444 | 32.84 | 0 | 43% |
| CGP23591 | Cervical | HR HPV positive | HPV16 | 15,270 | 0.0393% | 2,845 | 194.57 | 0 | 12% |
| CGP24465 | Cervical | HR HPV positive | HPV16 | 603 | 0.0015% | 77 | 7.36 | 0 | 58% |
| CGP24666 | Cervical | HR HPV positive | HPV16 | 57,730 | 0.1219% | 9,981 | 720.7 | 0 | 4% |
| CGP25458 | Cervical | HR HPV positive | HPV16 | 162,822 | 0.3466% | 25,796 | 2061.54 | 0 | 20% |
| CGP26453 | Cervical | HR HPV positive | HPV16 | 48,603 | 0.1494% | 12,317 | 597.69 | 0 | 22% |
| CGP27684 | Cervical | HR HPV positive | HPV16 | 21,138 | 0.0582% | 2,401 | 262.43 | 0 | 42% |
| CGP28096 | Cervical | HR HPV positive | HPV16 | 7,611 | 0.0155% | 801 | 95.92 | 0 | 34% |
| CGP28876 | Cervical | HR HPV positive | HPV16 | 432,687 | 1.0386% | 51,190 | 5286.43 | 0 | 8% |
| CGP29919 | Cervical | HR HPV positive | HPV16 | 80,407 | 0.1473% | 11,521 | 1017.47 | 0 | 18% |
| CGP30431 | Cervical | HR HPV positive | HPV16 | 86,859 | 0.1899% | 8,516 | 1099.14 | 0 | 18% |
| CGP31034 | Cervical | HR HPV positive | HPV16 | 750,743 | 1.9266% | 80,853 | 9483.02 | 0 | 2% |
| CGP9680 | Cervical | HR HPV positive | HPV16 | 2,094,635 | 5.4875% | 93,507 | 26475.95 | 31 | 0% |
| CGP26780 | Cervical | HR HPV positive | HPV18 | 178 | 0.0005% | 42 | 2.23 | 0 | 78% |
| CGP18006 | Cervical | HR HPV positive | HPV56 | 13,060 | 0.0258% | 2,861 | 167.43 | 0 | 52% |
| CGP19698 | Cervical | HR HPV negative control | no virus detected | 0 | 0.0000% | 0 | 0 | 0 | 100% |
| CGP22291 | Cervical | HR HPV negative control | no virus detected | 0 | 0.0000% | 0 | 0 | 0 | 100% |

# Figure 3: HPV16 Coverage Varied by Sample.

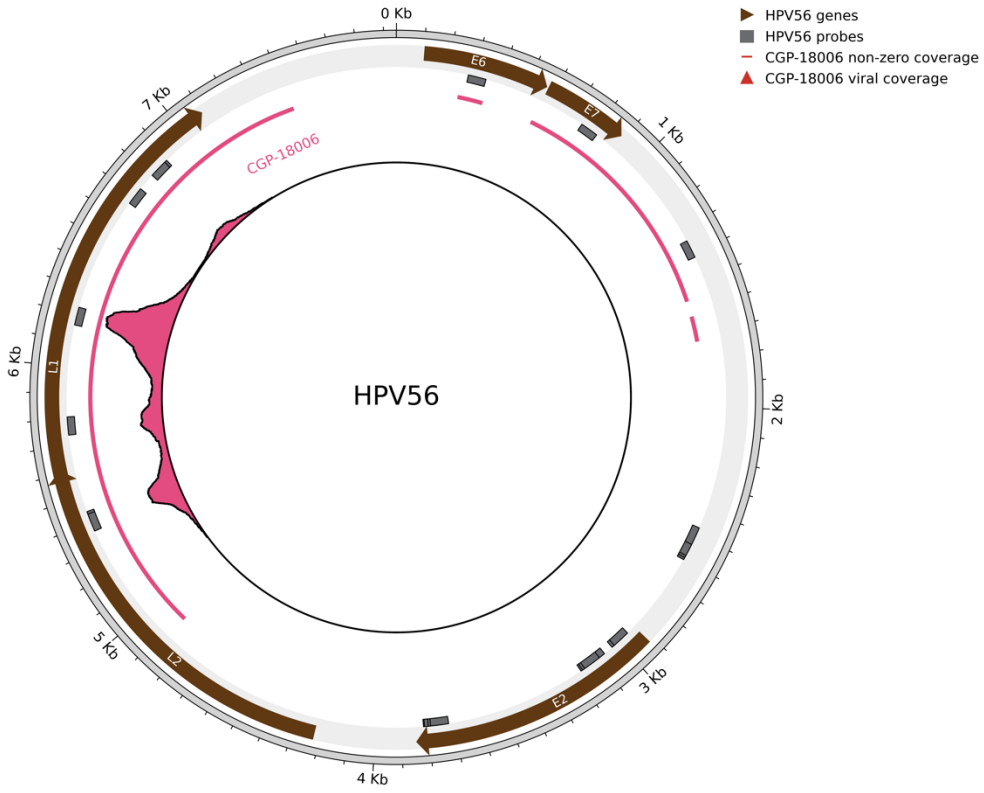
Circular visualization of viral coverage in three representative HPV16+ samples. The outermost ring represents the HPV16 reference viral genome (accession: K02718) with seven genes shown with brown arrows. The E1 and E2 genes partially overlap and the E4 gene sequence entirely overlaps with the longer E2 gene. Black bars represent locations of the probes labeled HPV16 that aligned to the HPV16 reference genome with 100% sequence identity. Gray bars represent locations of putative HPV16 probes that aligned to other HPV types with 100% sequence identity. The white bar represents a putative HPV18 probe that aligned to the HPV16 reference genome with 100% sequence identity. The E6 and E7 genes had two probes each. E1 had six probes, but only two of these probes (black bars) aligned to the HPV16 reference genome with a high sequence identity. The other four E6 probes (gray bars) ranged from 85 to 90 percent sequence identities. E2 had one probe and there was one probe aligning to the region shared by both E2 and E4. L1 had eight probes. Only two of these probes had 100% sequence identity with the HPV16 reference genome (black bars). The white bar represents a putatively “HPV18” probe, which has 100% sequence identity with the HPV16 genome and 85% sequence identity with the HPV18 genome. E5 and L2 had no probes.

The next three rings represent viral coverage maps in three representative samples. Read counts were normalized per sample, and y-axes are not equivalent. CGP-9680 (second ring) had non-zero coverage over the entire HPV16 genome (denoted by unbroken red line). CGP-20541 (third ring) had coverage over 99% of the HPV16 genome, denoted by the broken pink line. CGP-20351 (innermost ring) had coverage, over 40% of the genome, denoted by the broken fuchsia lines. Coverage was highest in regions surrounding probes except for the probes with less than 100% sequence identity (gray bars). These probes did not appear to increase coverage. Samples were chosen to accurately display the range of coverage: CGP-9680 had the highest HPV16 coverage, CGP-20351 had the lowest, and CGP20541 represented the median level of coverage of all samples. The plot was created using the Pycirclize package in python[407].


0 Kb
1 Kb
2 Kb
3 Kb
4 Kb
5 Kb
6 Kb
7 Kb
HPV16
CGP-9680
CGP-20541
CGP-20351
E6
E7
E1
E2
E4
E5
L2
L1
HPV18 genes
HPV16 probes with 100% identity
HPV16 probes with <100% identity
HPV18 probe
CGP-9680 non-zero coverage
CGP-9680 viral coverage count
CGP-20541 non-zero coverage
CGP-20541 viral coverage
CGP-20351 non-zero coverage
CGP-20351 viral coverage

# Figure 4: HPV56 Coverage was Highest in the L1 and L2 Region.

Circular visualization of viral coverage in the only HPV56+ sample. The outermost ring represents the HPV56 reference viral genome (accession: X74483) with five genes shown with brown arrows. The E1 gene is fragmented in the HPV56 genome and not shown. Gray bars represent locations of 25 “HPV” probes with no type specified that aligned to the HPV56 reference genome. Only the portion of probes that aligned to the HPV56 genome is shown. Shorter bars represent shorter aligned sequences in probes. The percent sequence identities of these probes ranged from 98.5 to 81.4%.
The E6 and E7 genes had one probe each. The fragmented E1 region had six probes, some of which overlapped. E2 had ten probes. L1 had five probes. L2 had two probes.
The next ring represents the viral coverage map in CGP-18006. CGP-18006 had coverage over 48% of the HPV16 genome, denoted by the broken pink line. Coverage was highest in regions surrounding probes. The plot was created using the Pycirclize package in python[407].


HPV56 genes
HPV56 probes
CGP-18006 non-zero coverage
CGP-18006 viral coverage
0 Kb
1 Kb
2 Kb
3 Kb
4 Kb
5 Kb
6 Kb
7 Kb
E6
E7
E2
L2
L1
CGP-18006
HPV56

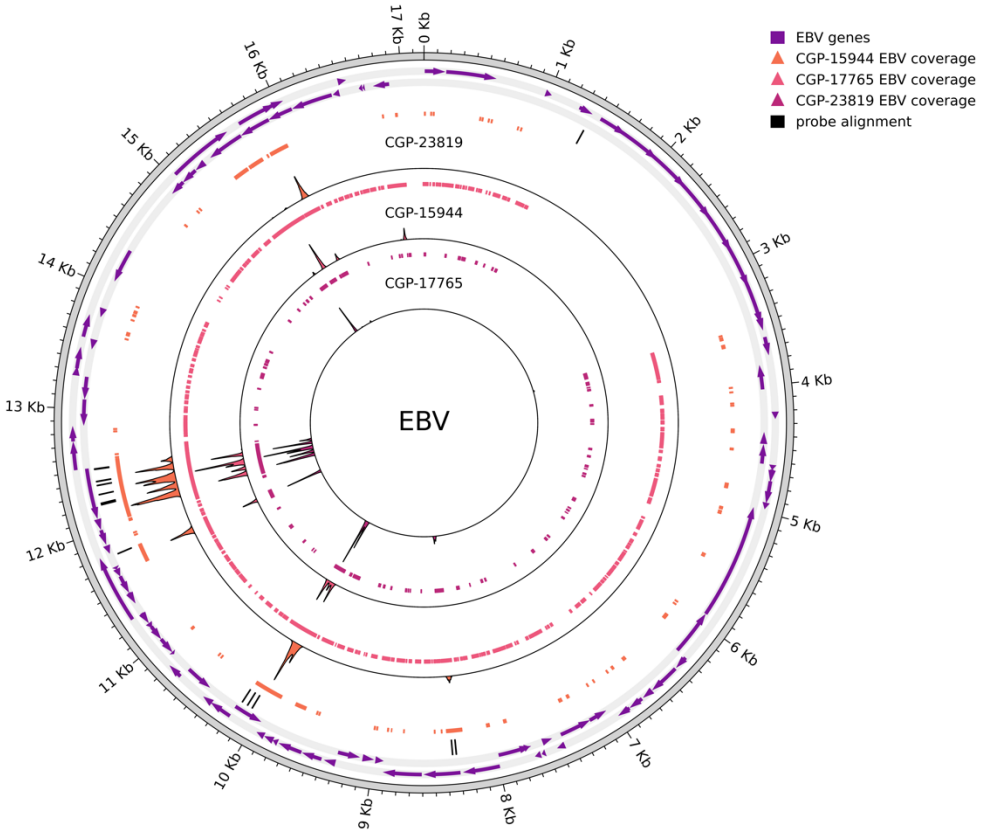
Due to the larger relative size of the EBV genome (17 kb), a greater fraction of the EBV genome had zero coverage. Coverage ranged from 49% to 7% (Table 7, Figure 5). Given the relatively small size of the probes (60 bp), it is interesting that the 22 EBV probes were able to obtain coverage in such a high fraction of the genome.

## Table 7: EBV Coverage Results of 15 Lymphoma Samples Run on ARCV.

| Sample ID | EBV status by ISH | Virus detected by ARCV | EBV read count | Highest coverage depth | Mean coverage depth | Lowest coverage depth | Percent of EBV genome with no coverage |
|---|---|---|---|---|---|---|---|
| CGP14605 | EBV positive | EBV | 2,252 | 160 | 1.3 | 0 | 93% |
| CGP15944 | EBV positive | EBV | 451,438 | 38,347 | 264.21 | 0 | 51% |
| CGP17765 | EBV positive | EBV | 288,189 | 19,682 | 168.25 | 0 | 83% |
| CGP19553 | EBV positive | EBV | 27,642 | 2,029 | 16.18 | 0 | 90% |
| CGP21070 | EBV positive | EBV | 18,151 | 1,029 | 10.6 | 0 | 91% |
| CGP23360 | EBV positive | EBV | 6,469 | 397 | 3.79 | 0 | 92% |
| CGP23819 | EBV positive | EBV | 226,106 | 12,355 | 132.65 | 0 | 85% |
| CGP26223 | EBV positive | EBV | 37,249 | 2,732 | 21.76 | 0 | 91% |
| CGP29936 | EBV positive | EBV | 141,960 | 10,292 | 82.65 | 0 | 87% |
| CGP30202 | EBV positive | EBV | 150,485 | 25,347 | 86.82 | 0 | 89% |
| CGP30527 | EBV negative control | no virus detected | 0 | 0 | 0 | 0 | 0% |
| CGP30899 | EBV negative control | no virus detected | 0 | 0 | 0 | 0 | 0% |
| CGP26254 | EBV negative control | no virus detected | 0 | 0 | 0 | 0 | 0% |
| CGP26777 | EBV negative control | no virus detected | 0 | 0 | 0 | 0 | 0% |
| CGP28135 | EBV negative control | no virus detected | 0 | 0 | 0 | 0 | 0% |

# Figure 5: EBV Coverage Varied by Sample.

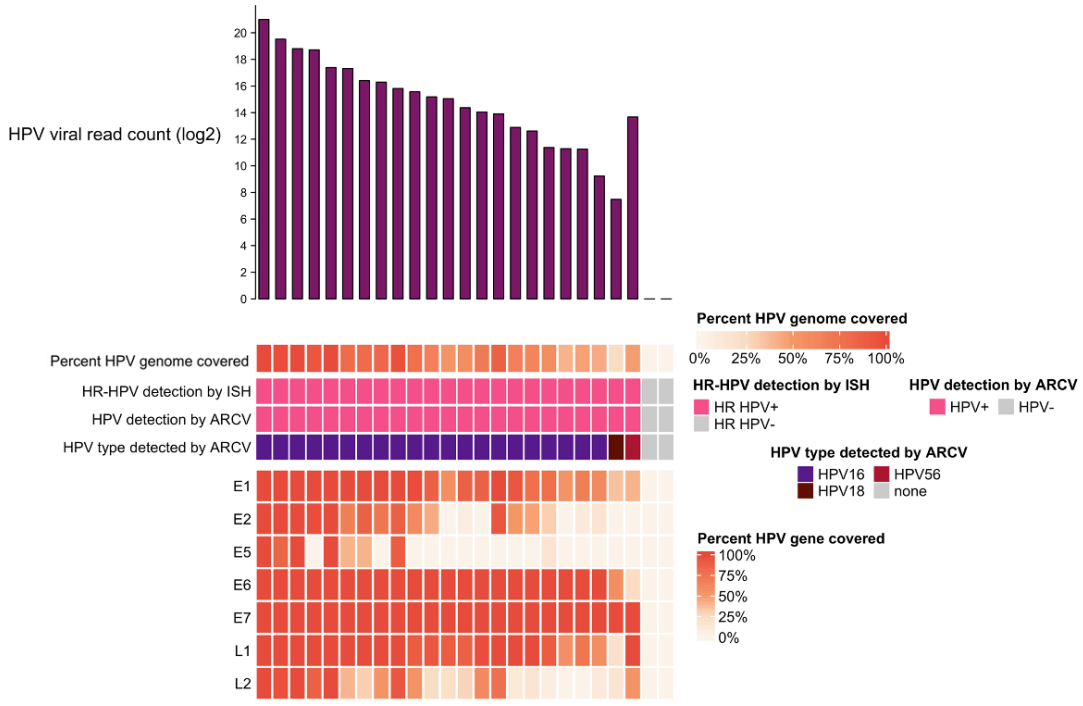
Circular visualization of viral coverage in three representative EBV+ samples. The outermost rings represents the EBV reference viral genome with genes shown with purple arrows. Black bars represent locations of EBV-specific probes. The next three rings represent viral coverage maps in three representative samples. Read counts were normalized per sample, and y-axes are not equivalent. CGP-23819 (second ring) had coverage over the entire HPV16 genome (denoted by broken orange line). CGP-15944 (third ring) had coverage over XX genome, denoted by the broken pink line. CGP-17765 (innermost ring) had coverage, over 40% of the genome, denoted by the broken purple lines. Coverage was highest in regions surrounding probes The plot was created using the Pycirclize package in python[407].



HPV viral gene coverage varied by sample and by gene (Figure 6). All HPV+ samples had 100% coverage of the E7 oncogene sequence. E6 also had high coverage, and to a lesser extent L1 and E1. 2 samples had no coverage of L2, 5 samples had no coverage of E2, and only 8 samples had any coverage at all of E5.

# Figure 6: ARCV Results of 25 Cervical Cancer Samples Visualized Using the OncoPlot package in R.

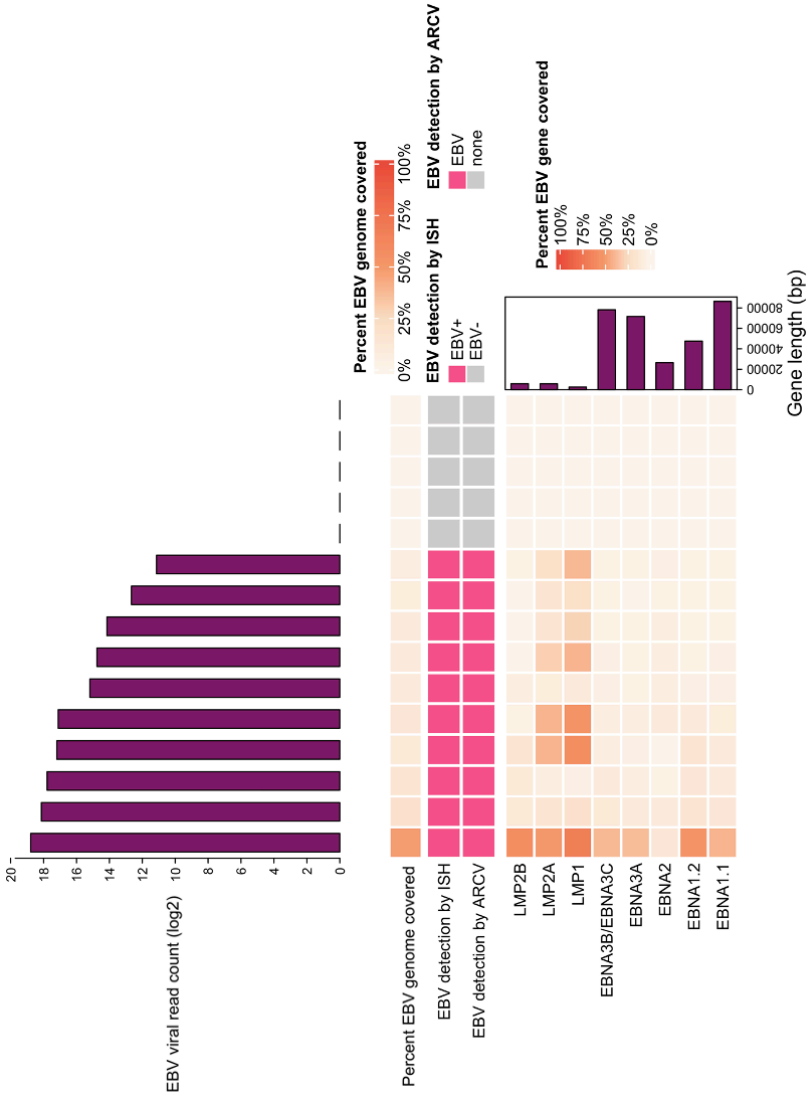
Top panel: Bar plot of total viral reads in each sample. Second panel: Heatmap containing percentage HPV genome covered as well as detection status by ISH and ARCV. Bottom panel: Heatmap visualizing coverage across HPV genes. Light orange indicates the lowest relative levels of expression and dark orange the highest, as defined by the color key. Overall, we had perfect concordance in terms of whether HPV was detected or not, but the coverage was not uniform. The only HPV18 positive sample had the worst coverage of any sample.




ARCV provides viral gene coverage information for EBV latent genes. As expected, gene coverage decreased with increasing gene length. In general, greater viral gene coverage corresponded to a higher viral read count per sample.

# Figure 7: ARCV Results of 15 Lymphoma Samples Visualized using the OncoPlot Package in R

Top row: Bar plot with EBV read count per sample. Middle panel: Heatmap containing percentage of EBV genome covered as well as detection status by ISH and ARCV. Bottom panel: Heatmap of percent sequence coverage per EBV latent gene: LMP-1, LMP-2A, EBNA-1.2, LMP-2B, EBNA1.1, EBNA-3B/3C, EBNA-3A, and EBNA-2. Light orange indicates the lowest relative levels of expression and dark orange the highest, as defined by the color key. Right box: Bar chart with gene length in base pairs.



## High-risk HPV detected in condyloma samples.

Condylomas are genital warts caused by HPV. An estimated 90% are caused by the low-risk types HPV6 and HPV11, although coinfection with HPV16 or other high-risk types is frequently seen[197,327]. ARCV was run on 25 condyloma samples with unknown HPV status. HPV6 or 11 was detected in only 7 samples (Table 8). In 3 of those samples, both high-risk and low-risk HPV types were detected. No HPV type was detected in 3 samples. Given that there are no probes in the UCSF 500 panel explicitly labeled for HPV6 or HPV11, it is not surprising that these types were not consistently detected. There was a probe labeled “HPV18” that had 100% sequence identity with HPV6. In cases where both HPV16 and HPV6 are actually present, HPV6 reads may be outcompeted by HPV16 reads, especially following the use of the probes.

# Table 8: ARCV Results for 25 Condyloma Samples of Unknown HPV type.

| Sample ID | Sample type | Viral read percent of total reads | Virus detected by ARCV | HPV type risk | HPV read count | Highest coverage depth | Mean coverage depth | Lowest coverage depth | Percent of HPV genome with no coverage |
|---|---|---|---|---|---|---|---|---|---|
| CGP23440 | Condyloma | 0.0509% | HPV6 | low-risk | 21,594 | 16,103 | 232.39 | 0 | 5% |
| CGP23016 | Condyloma | 0.0313% | HPV6 | low-risk | 18,592 | 2,932 | 220.23 | 0 | 0% |
| CGP23008 | Condyloma | 0.0119% | HPV6 | low-risk | 4,072 | 569 | 45.7 | 0 | 1% |
| CGP23015 | Condyloma | 0.1076% | HPV11 | low-risk | 43,846 | 1,539 | 498.41 | 32 | 0% |
| | | | HPV45 | high-risk | 1,012 | 384 | 12.43 | 0 | 58% |
| CGP23025 | Condyloma | 0.9001% | HPV16 | high-risk | 423,680 | 82,761 | 5281.9 | 0 | 5% |
| CGP23011 | Condyloma | 0.5127% | HPV16 | high-risk | 324,642 | 85,429 | 4010.23 | 1 | 0% |
| CGP23031 | Condyloma | 0.2387% | HPV16 | high-risk | 210,600 | 29,611 | 2656.16 | 2 | 0% |
| CGP23028 | Condyloma | 0.1064% | HPV16 | high-risk | 50,068 | 23,502 | 597.14 | 0 | 30% |
| | | | HPV11 | low-risk | 14 | 14 | 0.04 | 0 | 99% |
| CGP23010 | Condyloma | 0.0246% | HPV16 | high-risk | 11,696 | 2,183 | 143.41 | 0 | 3% |
| CGP23017 | Condyloma | 0.0194% | HPV16 | high-risk | 7,728 | 1,750 | 96.01 | 0 | 2% |
| CGP23441 | Condyloma | 0.0174% | HPV16 | high-risk | 5,471 | 1,085 | 67.66 | 0 | 63% |
| | | | HPV6 | low-risk | 130 | 65 | 3.94 | 0 | 85% |
| CGP23439 | Condyloma | 0.0066% | HPV16 | high-risk | 2,496 | 549 | 30.74 | 0 | 62% |
| CGP23013 | Condyloma | 0.0019% | HPV16 | high-risk | 564 | 119 | 7.11 | 0 | 70% |
| CGP23012 | Condyloma | 0.0009% | HPV16 | high-risk | 542 | 209 | 6.81 | 0 | 84% |
| CGP23014 | Condyloma | 0.0002% | HPV16 | high-risk | 92 | 78 | 1.09 | 0 | 94% |
| CGP23026 | Condyloma | 0.0102% | HPV18 | high-risk | 3,819 | 1,222 | 48.31 | 0 | 53% |
| CGP23442 | Condyloma | 0.0011% | HPV18 | high-risk | 308 | 131 | 3.92 | 0 | 89% |
| CGP23029 | Condyloma | 0.0007% | HPV18 | high-risk | 296 | 122 | 3.61 | 0 | 85% |
| CGP23019 | Condyloma | 0.1410% | HPV33 | high-risk | 52,043 | 14,547 | 644.21 | 20 | 0% |
| CGP24552, 23027 | Condyloma | 0.0001% | HPV33 | high-risk | 32 | 16 | 0.41 | 0 | 92% |
| | | | HPV6 | low-risk | 24 | 24 | 0.19 | 0 | 99% |
| CGP23023 | Condyloma | 0.0061% | HPV45 | high-risk | 2,625 | 651 | 32.06 | 0 | 60% |
| | | | HPV59 | high-risk | 51 | 12 | 0.63 | 0 | 87% |
| CGP23024 | Condyloma | 0.0038% | HPV68 | high-risk | 1,994 | 460 | 23.83 | 0 | 11% |
| CGP23009 | Condyloma | 0.0000% | none | - | 0 | 0 | 0 | 0 | 0% |
| CGP23018 | Condyloma | 0.0000% | none | - | 0 | 0 | 0 | 0 | 0% |
| CGP23030 | Condyloma | 0.0000% | none | - | 0 | 0 | 0 | 0 | 0% |

## Discussion

In this Chapter I have described a computational pipeline, ARCV, for detecting viral sequences in cancer gene panel data. This pipeline was able to accurately detect HPV and EBV viruses in clinical samples, outputs coverage statistics including coverage of key viral genes, and runs quickly. No viral reads were detected in the negative control samples.

Additionally, I have generated a viral reference genome that contains validated reference sequence for all known human oncogenic viruses, including 109 types of HPV, 10 types of polyomaviruses, 3 types of HTLV, Hepatitis B, Hepatitis C, EBV and HHV8.

Notably, few reads or no reads at all were found in regions where the probes had low percent identity to the reference sequence. This suggests that the strategy of adjusting the probe sequence to capture a broader spectrum of viral types may not be reliable if used for diagnostic purposes. However, these non-specific probes were able to capture HPV6 reads in a portion of the samples.

Given the relative sizes of the EBV genome and the small size (60 bp) and number (22) of EBV probes, covering about 7% of the genome, it is surprising that the coverage of the EBV genome ranged to as high as 49% in one sample.

While the accuracy of sequencing-based assays for viral detection purposes is well established, those in the clinic commonly use less accurate, obsolete methods such as IHC for surrogate markers rather than HPV itself. Adoption of sequencing based-assays like ARCV will decrease the number of samples where viral presence is missed.

# Chapter Three. Detection of a Putative Aberration Indicative of BWS Spectrum Disorder by Nanopore Sequencing

## Introduction

Wilms tumor (WT), also known as nephroblastoma, is the most common childhood kidney cancer[408]. Bilateral WT, in which WT is found in both kidneys, has both a lower survival rate and a higher rate of kidney failure.

While 15% of patients with Wilms tumor have germline pathogenic variants in genes or regions such as WT1 or the 11p15 region, up to 75% of patients with bilateral Wilms tumor had a germline predisposition syndrome[409]. Among these is Beckwith-Wiedemann syndrome (BWS). BWS is an overgrowth syndrome: symptoms of BWS can include hemihypertrophy, abdominal wall defects, and hyperinsulinism. Recently, BWS was recognized as a spectrum (BWSp) in that in some patients only one of those symptoms is seen[149]. BWSp is caused by genetic or epigenetic alterations in either of two imprinted regions in the chromosome 11p15 region (ICR1 or ICR2)[410].

The clinical understanding of BWS has evolved over time. Original reports emphasized macroglossia, exomphalos, gigantism and neonatal hypoglycemia[411], without finding any cases of WT, though one author did note that a cousin of a patient with BWS had WT. Ten years later, clinicians noted that phenotypes were

frustratingly variable between patients and no clear diagnostic criteria could be proposed[412]. Additionally, though BWS was clearly heritable, no clear mechanism of inheritance could be discerned. The first linkage of BWS to Wilms tumor occurred when Sotelo-Avila proposed that hemihyperplasia was an "incomplete form of BWS"[412]. They also noted the strong association of hemihyperplasia with WT[413–416]. Wiedemann himself noted this in 1983[417].

In the search for a candidate gene (or genes) associated with BWS, it was noted that mutations in the chromosome 11p15 region were found in 2% of BWs patients[418]. Researchers quickly focused on IGF2, an imprinted gene in the 11p15 region. In 1993, Weksberg, *et al.* noted biallelic expression of IGF2 in BWS patients and monoallelic expression in normal controls[419]. In a patient with BWS and WT, biallelic expression of IGF2 was also found in samples taken from her peripheral blood, the normal part of the kidney, and from her tumor[202]. A follow up study found 10 BWS patients with increased levels of methylation in this region detectable in peripheral blood samples, in addition to biallelic expression of IGF2[420]. Notably, all 10 of these patients were female. The authors hypothesized that BWS in these cases could be caused by a mutation occurring in the "early embryo", leading to some cells with increased methylation and other cells with normal methylation.

## ICR1

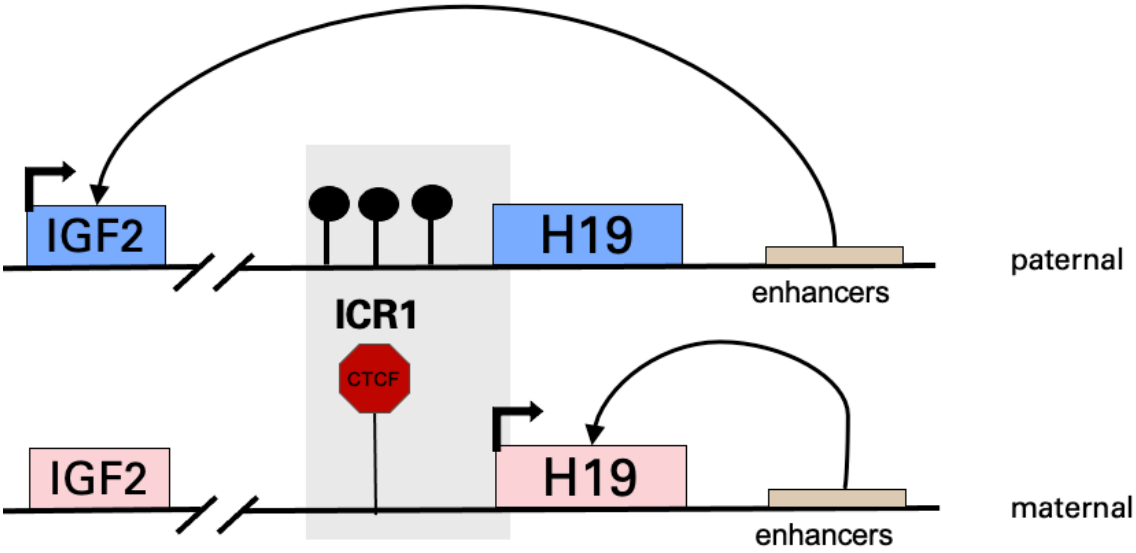
In humans, the majority of autosomal genes are expressed equally from both alleles. However, imprinted genes have monoallelic expression, depending on the parental origin of the allele (maternal or paternal). Imprinting control region 1 (ICR1, also known as H19/IGF2 intergenic differentially methylated region) consists of two genes, H19 and IGF2, which are imprinted in a reciprocal manner, and a methylated region upstream of H19[421].

On the maternal unmethylated allele, CTCF proteins bind the ICR1, forming an insulator[421,422]. This insulator blocks activation of IGF2 by enhancers downstream

of H19. Therefore, IGF2 is silenced and H19 is activated on the maternal (unmethylated) allele. On the paternal allele, the ICR1 is methylated and CTCF binding is blocked, so no insulator is formed. The enhancers activate expression of IGF2. As a result, on the paternal (methylated) allele, IGF2 is expressed and H19 is silenced (**Figure 8**).

## Figure 8: ICR1 is Methylated on the Paternal Allele.

Schematic representation of the ICR1 in humans and the genes it regulates, including the IGF2 gene and H19 lncRNA.




While loss of methylation at the maternal ICR2 allele is the most common cause of BWSp, gain of methylation at the maternal ICR1 has been seen in 5% of BWSp patients[149]. These patients with BWSp caused by gain of methylation at ICR1 have a 25% chance or greater of developing a tumor[410].

The diagnostic criteria for BWS were reevaluated following the identification of the epigenetic/genetic defects that cause BWS. Many patients with isolated hemihyperplasia were found to have the same genetic defects as BWS[423]. Therefore, in 2018, the consensus recommendation was that BWS should be

considered a spectrum with variable phenotypes, including patients with isolated hemihyperplasia[149]. Additionally, all individuals with 11p15 anomalies are considered to be on the BWS spectrum, even if no symptoms associated with BWS are found[149].

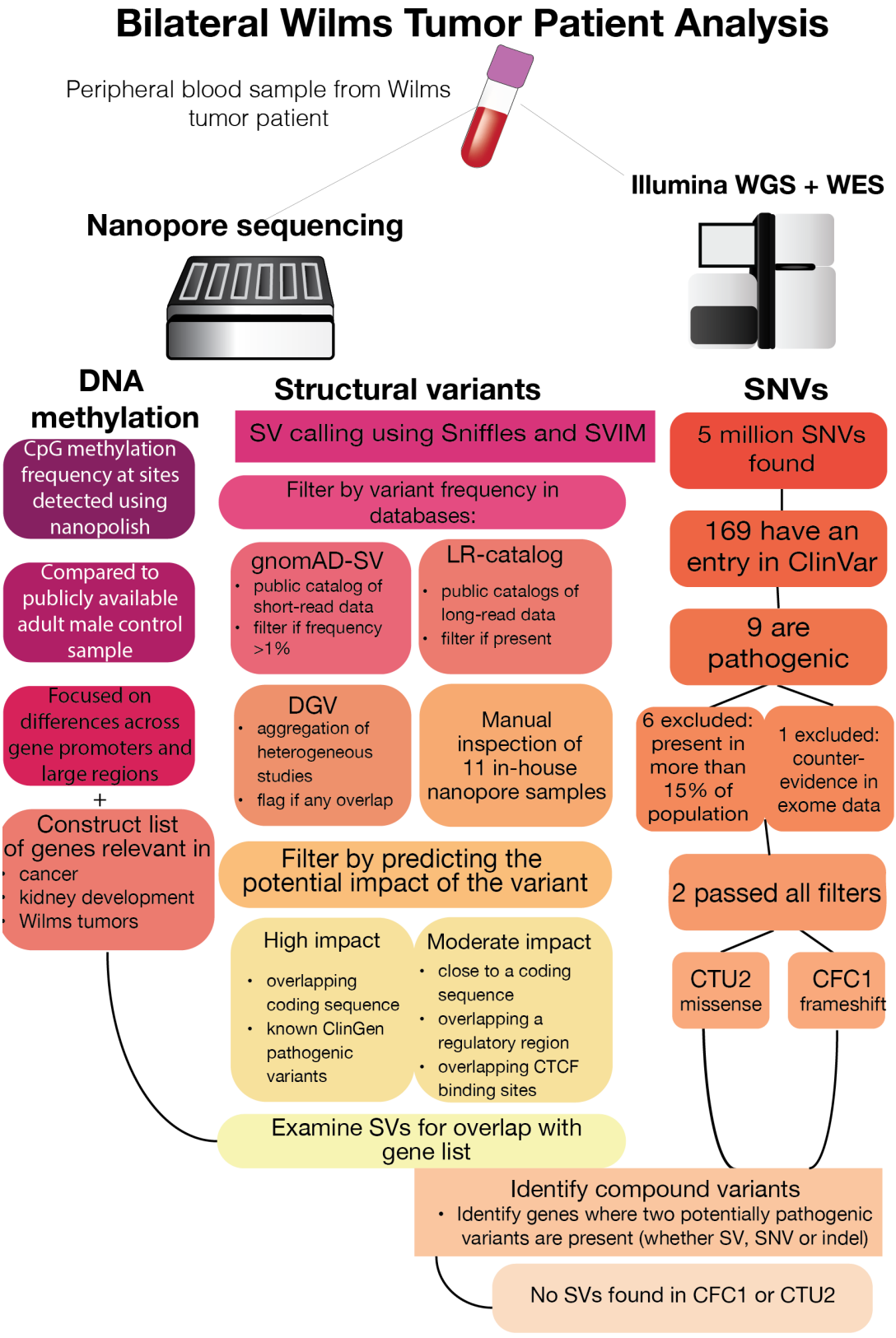
The most commonly used assay for BWS detection is methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA), which can detect changes in methylation status and some copy number variants[149]. However, other more sensitive methods are needed to detect low-level increases in hypermethylation in the blood[424], as MS-MLPA has a detection threshold of 10%[425]. If a methylation abnormality is detected, the current consensus recommends follow-up testing to identify any CNVs using chromosome microarray analysis[149]. It is estimated that 20% of patients with ICR1 hypermethylation might also carry SNVs or small CNVs in ICR1, and these patients have a higher risk of recurrence[149]. Recent studies[424,426] of the peripheral blood of patients with WT have demonstrated that low-level gain of methylation at ICR1 is detectable in the blood of patients with no other genetic abnormalities. The majority of patients positive for ICR1 gain of methylation had bilateral WT and notably, nearly all were female.

In many cases of Wilms tumor, no known pathogenic variant could be found using the state-of-the-art technologies, including comprehensive approaches such as Illumina whole exome sequencing. One explanation for this is that such technologies have difficulties in detecting structural variants (SVs) in areas associated with repeat or low complexity sequences. ICR1 contains 2 repeat sequences. Long-read sequencing detects SVs with higher accuracy than short read methods[427]. In addition, Illumina technology does not support direct methylation detection[428]. In 2023, a study of 20 individuals with Prader-Willi syndrome, another syndrome caused by changes in methylation at an imprinted region, detected increased methylation in all samples using Nanopore long-read sequencing[429]. Therefore, we hypothesized that analysis of bilateral Wilms tumor of

unknown etiology using long-read sequencing could reveal molecular events of potential clinical interest.

We performed in-depth genomic analysis on a whole blood DNA sample from a patient with a bilateral Wilms tumor (Figure 9). This patient had no significant family history of cancer, and previously tested negative for Beckwith-Wiedemann syndrome by methylation testing of the 11p15 region; clinical exome sequencing of the patient's germline detected no variants associated with Wilms tumors. We detected low-level gain of methylation at ICR1, consistent with BWSp.

## Figure 9: Workflow of Bilateral Wilms Tumor Patient Genomic Analysis.


Bilateral Wilms Tumor Patient Analysis
Peripheral blood sample from Wilms tumor patient
Nanopore sequencing
Illumina WGS + WES
DNA methylation
CpG methylation frequency at sites detected using nanopolish
Compared to publicly available adult male control sample
Focused on differences across gene promoters and large regions
+
Construct list of genes relevant in
cancer
kidney development
Wilms tumors
Structural variants
SV calling using Sniffles and SVIM
Filter by variant frequency in databases:
gnomAD-SV
public catalog of short-read data
filter if frequency >1%
LR-catalog
public catalogs of long-read data
filter if present
DGV
aggregation of heterogeneous studies
flag if any overlap
Manual inspection of 11 in-house nanopore samples
Filter by predicting the potential impact of the variant
High impact
overlapping coding sequence
known ClinGen pathogenic variants
Moderate impact
close to a coding sequence
overlapping a regulatory region
overlapping CTCF binding sites
Examine SVs for overlap with gene list
SNVs
5 million SNVs found
169 have an entry in ClinVar
9 are pathogenic
6 excluded: present in more than 15% of population
1 excluded: counter-evidence in exome data
2 passed all filters
CTU2 missense
CFC1 frameshift
Identify compound variants
Identify genes where two potentially pathogenic variants are present (whether SV, SNV or indel)
No SVs found in CFC1 or CTU2

# Methods

## Case Presentation

A 3-year-old female patient presented to the University of California, Los Angeles with stage IV bilateral Wilms tumor with metastases in her lungs and hemihypertrophy. Due to the presence of hemihypertrophy and bilateral Wilms tumor, the patient underwent a genetic workup. Methylation analysis of 11p15 was normal and germline clinical exome sequencing identified no clinically significant variants.

We sequenced a whole blood DNA sample from this patient using long-read whole genome sequencing (WGS), Illumina WGS and exome sequencing. Long-read sequencing was performed at 40x depth using PromethION Nanopore sequencing.

## SVs

We performed SV calling using Sniffles and SVIM. Only high-confidence SVs that were detected by both methods were considered for downstream analysis. SVs were annotated and filtered based on predicted functional impact and frequency in the population. SVs found in gnomAD-SV (v2.1.1) at a frequency greater than 1% were filtered out. All SVs found in LR catalog were filtered. Additionally, SVs found in our 11 in-house Nanopore samples were filtered out based on manual inspection. Compound heterozygotes were detected using phased variant calls.
SVs from regulatory regions and CTCF peaks in kidney cells from ENCODE were flagged.

## Epigenetic Marks

For the methylation analysis, we used a publicly available normal blood sample from an adult male as a control. The frequency of methylation in CpG sites genome-wide

was detected using nanopolish. We searched for extreme differences across large regions and gene promoter regions.

### Clinical Annotation

I led the annotation of variants following filtering. I created a list of genes of interest that are implicated in WTs, kidney cancers in general, kidney development, and cancer in general and focused on variants in genes found on that list.

## Results

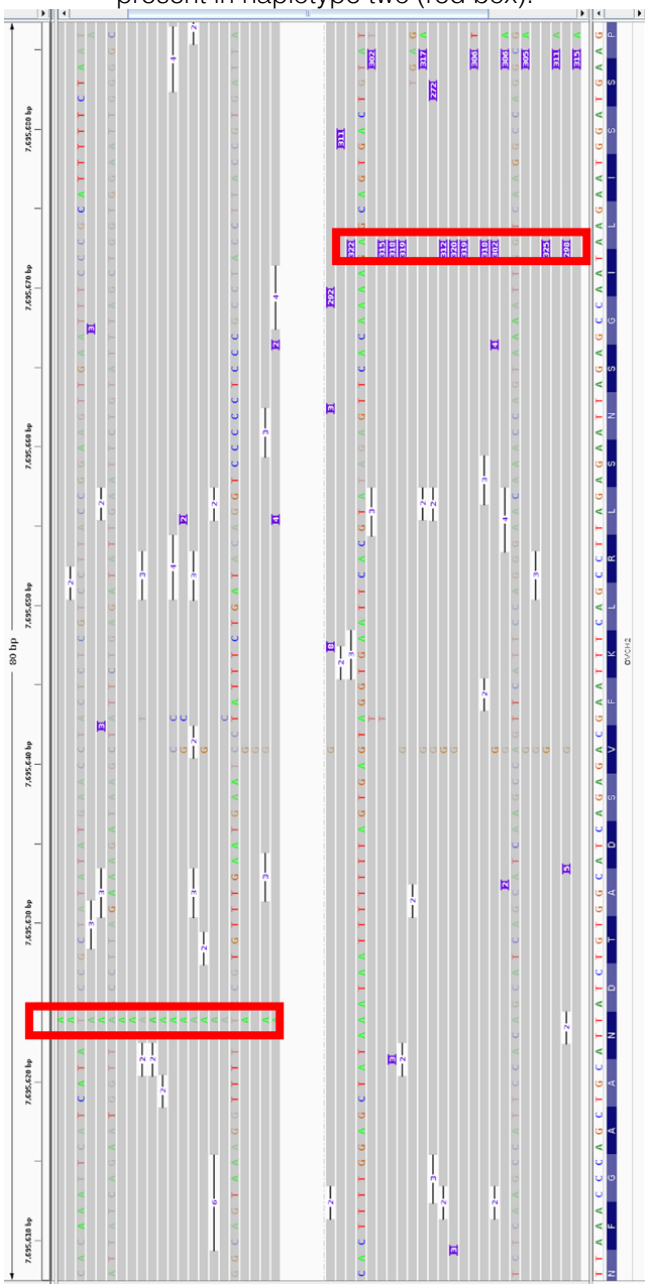
While gain of chromosome 1q and loss of chromosome 22q are commonly found in bilateral WT samples as somatic mutations, no chromosomal abnormalities were detected in the patient's blood sample. Additionally, we detected no loss of heterozygosity or copy number variants in the 11p15.5 region.

I filtered the SVs found using the list of genes I created, which included genes relevant to Wilms tumors and kidney development. No SVs were found in the coding regions, promoters, or UTRs of genes from that list. A 42 bp deletion was found in the intronic region of CTNNB1. CTNNB1 is commonly mutated in WTs, but this deletion was found in healthy controls.

Because we looked for both SVs and small mutations, we were able to identify compound heterozygous variants. Compound heterozygosity refers to the presence of two different mutations on two different alleles of the same gene. Two compound heterozygous variants were found overlapping an exon of the OVCH2 gene (**Figure 10**). A heterozygous missense mutation was identified in haplotype one, while a 300 base pair insertion in an ALU element was present in haplotype two. Notably, the ALU element was not detected by prior Illumina analysis. Deletions in OVCH2 have been associated with multiple endocrine neoplasia type 2, a cancer predisposition syndrome[430].

# Figure 10: Compound Variant and ALU Insertion in OVCH2.

Screenshot of IGV viewer of Nanopore sequencing reads mapped to the OVCH2 locus. Gray bars show individual reads with indel mutations in black lines or single nucleotide mutations in colored letters. Top: haplotype one, a heterozygous missense mutation is detected in OVCH2 (red box). Bottom: a 300 base pair insertion in an ALU element was present in haplotype two (red box).



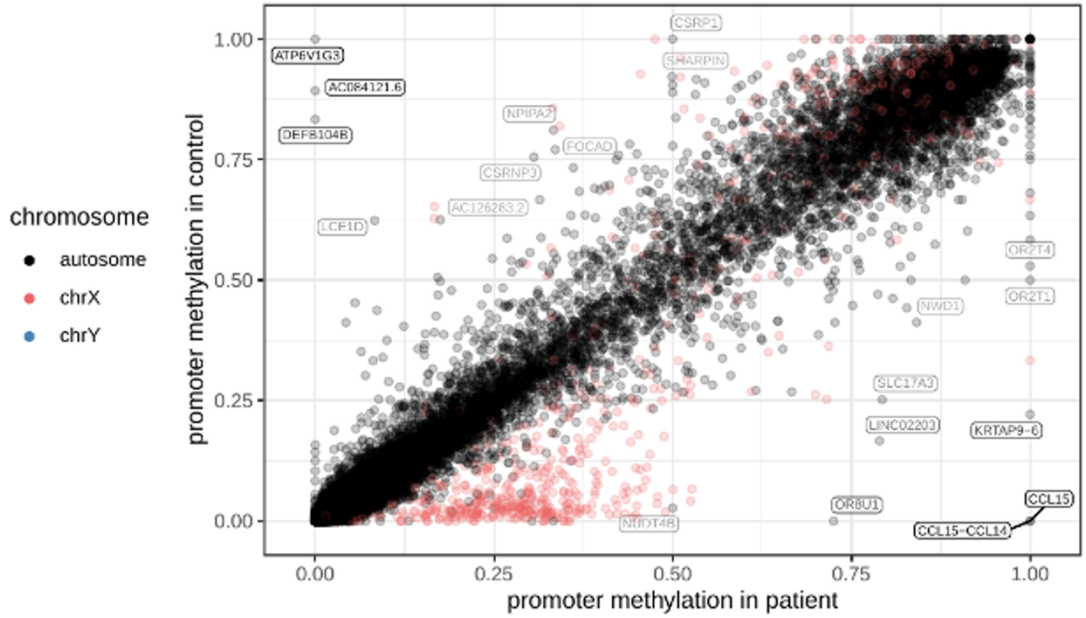
## Comparison of Methylation Levels Between Patient and Control Samples

Comparing the patient with the control sample, we found that overall methylation patterns were consistent. However, some outliers were visible. As expected, methylation levels differed on X chromosome genes due to X inactivation. Immune gene methylation was also higher in the patient (**Figure 11**). This is consistent with previous studies which have found the composition of certain immune cell types, including monocytes, varies with age[431].

DNA methylation levels at specific loci vary between individuals based on factors such as age[432–436], diet[437–439], and exposure to stress[440,441]. DNA methylation has been used to predict the age of a DNA source. However, this could be explained by the changes in proportion of white blood cell types as we age[442].

### Figure 11. Gene Promoter Methylation Differences Between Patient and Control Samples.

Black dots represent methylation intensity of autosomes. Red dots represent methylation intensity of X chromosome. Blue dots represent methylation intensities from the Y chromosome. I noted the differential methylation of the X chromosome, which we expected because the patient and control were different sexes.

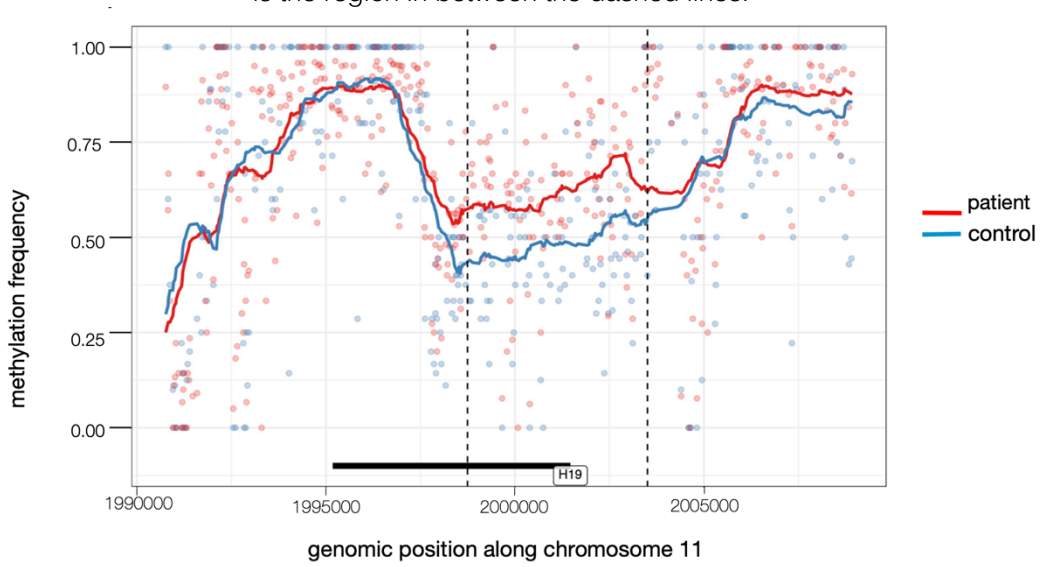


Methylation was higher in olfactory receptor genes in the patient. Changes in methylation of olfactory receptor genes have previously been associated with dietary intake including carbohydrate intake[443]. Therefore, these changes may be explained in differences in diet between patient and control.

Notably, hypermethylation over the ICR1 locus was observed in the patient as compared to the control (**Figure 12**). Gain of methylation over ICR1 is associated with BWS, and significantly increases the risk of developing Wilms tumors. Even relatively low-level hypermethylation meets the criteria for BWSp[424]. No significant difference was detected in methylation over ICR2 (**Figure 13**).

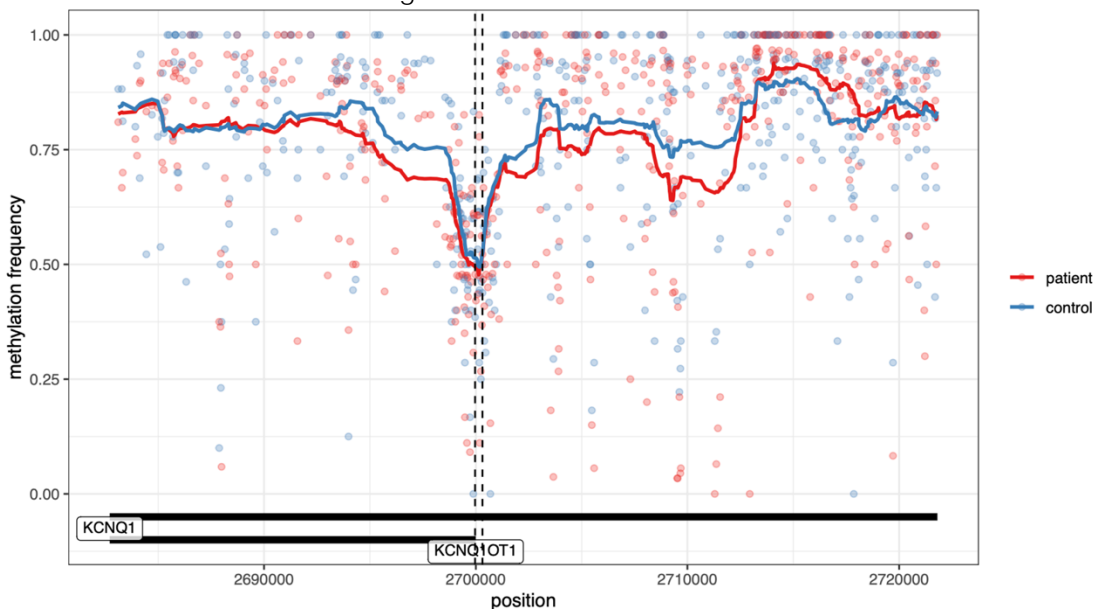
## Figure 12: Low-Level Hypermethylation Over ICR1 was Detected in the Patient.

Methylation frequency over ICR1 and surrounding region in patient vs control samples. ICR1 is the region in between the dashed lines.



## Figure 13: Methylation Frequency Over ICR2 in the Patient is Similar to the Control.

Methylation frequency over ICR2 and surrounding region in patient vs control samples. ICR2 is the region in between the dashed lines.




# Discussion

Germline cancer predisposition syndromes are associated with about 10% of pediatric cancers[444,445], but this number excludes predispositions caused by alterations in methylation or epimutations[424]. Increased usage of sequencing technologies which can support methylation detection may allow us to identify new epimutations relating to cancer susceptibility which may otherwise be missed, such as in this case. In addition to detecting mosaic hypermethylation other technologies missed, nanopore sequencing detected a gene with compound heterozygosity that was not detectable by Illumina sequencing.

We detected hypermethylation over the ICR1 in our patient's blood sample, indicating BWS. Our results are remarkably consistent with recent studies of patients with WT that found low-level gain of methylation at ICR1 detectable in the blood[424,426]. Of the combined 20 patients without a detectable germline mutation, 19

were female and 18 had bilateral Wilms tumor. While bilateral Wilms tumor is more common in females, this dramatic gendered pattern warrants further research. Evidence suggests that this mosaic gain of methylation at ICR1 is the first hit in a two-hit model[424,426]. Murphy, *et al*. theorized that the somatic mosaic epimutation occurs at an early embryonic stage in a mesodermal progenitor cell before the embryonic kidney is lateralized[426]. Mesodermal cells are progenitors of both kidney and lymphocyte cells. This explains why 11p15.5 hypermethylation is detectable in the peripheral blood of some patients. Oncogenesis does not occur until a second mutation occurs in the kidney. No tumor samples from our patient were available to us to test for somatic mutations.

A similarly gendered pattern of mosaic constitutional epimutation has been found in children with SDHC gastrointestinal stromal tumors. 12 out of 12 patients with SDHC hypermethylation and without coding mutations were female[446]. This hypermethylation was detectable, although to a lesser degree, in the blood[446]. MS-MLPA, the most commonly used technique to detect changes in methylation associated with BWSp, may not be sensitive enough to detect the low-level gain of methylation present in the blood of some patients with BWSp. MS-MLPA has difficulty in detecting hypermethylation at levels <20% relative to the expected normal 50% methylation value[447–450]. A recent study of the peripheral blood of 97 patients with WT found detectable gain of methylation in only one patient using MS-MLPA[425]. This female patient had hemihypertrophy. The authors attributed the low detection rate to the relatively low sensitivity of MS-MLPA compared to other methods[425,451–453].

In conclusion, Nanopore technology is able to detect variants missed by Illumina sequencing and has the potential to yield new findings of interest in a case of a child with suspected cancer predisposition syndrome. To my knowledge, this was the first use of Nanopore sequencing to detect an epimutation.

# Chapter Four. Identification of a Differentiation Stall in Epithelial Mesenchymal Transition in Histone H3-Mutant Diffuse Midline Glioma

In Chapter 4, we use gene expression data from a large cohort of diffuse midline gliomas to analyze the developmental origin of H3K27M tumors and the role of the Epithelial-Mesenchymal Transition (EMT). I led this study. I performed the analysis of the gene expression data that found the EMT pathway was upregulated, and I noted the expression of the EMT genes. I performed the experiments validating expression of these genes in glioma cell lines. This manuscript was published in *Giga Science* in 2020. Lauren Sanders was co-first author. Lauren Sanders performed the analysis of the single-cell data and developed the EMT score. Lucas Seninge contributed the analysis of the organoid data. Anouk van den Bout additionally performed RT-PCR on glioma cell lines to quantify expression of EMT genes.

OXFORD (GIGA)n SCIENCE

*GigaScience*, 9, 2020, 1–14

doi: 10.1093/gigascience/giaa136
RESEARCH

RESEARCH

# Identification of a differentiation stall in epithelial mesenchymal transition in histone H3–mutant diffuse midline glioma

Lauren M. Sanders[1,2,*,†], Allison Cheney[3,†], Lucas Seninge[1,2], Anouk van den Bout[2,3], Marissa Chen[2,3], Holly C. Beale[2,3], Ellen Towle Kephart[2], Jacob Pfeil[1,2], Katrina Learned[2], A. Geoffrey Lyle[2,3], Isabel Bjork[2], David Haussler[1,2,4], Sofie R. Salama[1,2,4,‡] and Olena M. Vaske[2,3,‡]

[1]Department of Biomolecular Engineering, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA; [2]University of California Santa Cruz Genomics Institute, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA; [3]Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA and [4]Howard Hughes Medical Institute, 1156 High Street, Santa Cruz, CA 95064, USA

*Correspondence address: Lauren M. Sanders, 1156 High Street, 220 Sinsheimer Labs, University of California Santa Cruz, Santa Cruz, CA 95064, USA. Tel: +530 409 2174; E-mail: lmsh@ucsc.edu
†Co–first author.
‡Co-senior author.

## Abstract

**Background:** Diffuse midline gliomas with histone H3 K27M (H3K27M) mutations occur in early childhood and are marked by an invasive phenotype and global decrease in H3K27me3, an epigenetic mark that regulates differentiation and development. H3K27M mutation timing and effect on early embryonic brain development are not fully characterized. **Results:** We analyzed multiple publicly available RNA sequencing datasets to identify differentially expressed genes between H3K27M and non-K27M pediatric gliomas. We found that genes involved in the epithelial-mesenchymal transition (EMT) were significantly overrepresented among differentially expressed genes. Overall, the expression of pre-EMT genes was increased in the H3K27M tumors as compared to non-K27M tumors, while the expression of post-EMT genes was decreased. We hypothesized that H3K27M may contribute to gliomagenesis by stalling an EMT required for early brain development, and evaluated this hypothesis by using another publicly available dataset of single-cell and bulk RNA sequencing data from developing cerebral organoids. This analysis revealed similarities between H3K27M tumors and pre-EMT normal brain cells. Finally, a previously published single-cell RNA sequencing dataset of H3K27M and non-K27M gliomas revealed subgroups of cells at different stages of EMT. In particular, H3.1K27M tumors resemble a later EMT stage compared to H3.3K27M tumors. **Conclusions:** Our data analyses indicate that this mutation may be associated with a differentiation stall evident from the failure to proceed through the EMT-like developmental processes, and that H3K27M cells preferentially exist in a pre-EMT cell phenotype. This study demonstrates how novel biological insights could be

**Received:** 23 April 2020; **Revised:** 17 August 2020; **Accepted:** 5 November 2020

derived from combined analysis of several previously published datasets, highlighting the importance of making genomic data available to the community in a timely manner.

*Keywords:* glioma; H3K27M mutation; epithelial mesenchymal transition

## Background

Pediatric high-grade gliomas (pHGGs) are aggressive brain tumors occurring at a median age of 6 years [1]. Sixty percent of pHGGs harbor a histone H3 K27M (H3K27M) mutation, which is associated with an aggressive phenotype and dismal survival rates [2]. H3K27M-mutant pHGG tumors are located along the midline, including in the pons, cerebellum, and brainstem. A diffuse phenotype and delicate location leave them unsuitable for surgery, and their pronounced chemoresistance renders the standard treatments for gliomas ineffective, resulting in a median survival time of only 12 months [3, 4]. The prognostic significance of the H3K27M mutation in these gliomas resulted in a new World Health Organization tumor classification, diffuse midline glioma with H3K27M mutation [5].

The H3K27M mutation results in a global decrease in H3K27me3, an epigenetic repressive mark and posttranslational histone modification [6]. Seventy-five percent of gene loci lose or have reduced H3K27me3, although a few loci gain the mark as a result of the H3K27M mutation [2, 7]. H3K27me3 is deposited predominantly by EZH2, the catalytic subunit of the PRC2 methyltransferase complex. By regulating H3K27me3, EZH2 maintains cell identity and regulates cellular differentiation [8–11]. Silencing EZH2 in neuroepithelial cells before their differentiation alters the distribution of the progeny cell types [12]. EZH2 also maintains neuroepithelial cell integrity and midbrain identity [13, 14].

Because H3K27me3 is globally lost in H3K27M-mutant glioma, the subsequent deregulation of gene expression is thought to lead to tumorigenesis, although the developmental timing of the mutational event is important [15]. H3K27M expression in neural stem cells has led to tumorigenesis in mice when accompanied by *TP53* knockout and/or *PDGFRA* amplification, but this combination of molecular aberrations failed to result in tumorigenesis when introduced in mature astrocytes [16, 17]. However, the precise cell type of origin for H3K27M gliomas is not yet known. Candidate cell types include neuroepithelial cells (also known as neural stem cells), radial glia (also known as neural progenitor cells), and oligodendrocyte precursor cells (OPCs) [16–18].

Many important brain developmental processes are regulated by H3K27me3 deposition and could contribute to gliomagenesis if not well controlled. One of these is the epithelial-mesenchymal transition (EMT) pathway, which is essential for gastrulation, migration of neural crest cells, and neural tube formation [19–22]. The EMT is regulated by SNAI1, a transcription factor master regulator [23–25]. By regulating EMT, SNAI1 plays a critical role in many developmental processes, including gastrulation and differentiation of embryonic stem cells [26–28]. SNAI1 induces EMT through direct recruitment of PRC2, resulting in H3K27 trimethylation of key epithelial genes, as well as concurrently upregulating mesenchymal genes [29, 30].

In the brain, cellular transitions driven by EMT-like transcriptional programs are involved in key developmental steps such as the differentiation of neuroepithelial cells to both neuronal and glial cells [31, 32]. These transitional transcriptional programs, which control cell fate and identity in early neural cell development, are regulated by EZH2 [33].

Given the regulation of EMT-associated gene transcription by H3K27me3 deposition in the brain, and the disruption of this deposition by the H3K27M mutation, we sought to investigate EMT-related gene expression in pHGGs with and without the H3K27M mutation. We analyzed RNA sequencing (RNA-seq) data from 78 pHGGs obtained from several different studies (Supplementary Table S1). First, we performed differential expression analysis using RNA-seq–derived gene expression from bulk tumor samples and found that H3K27M gliomas differentially express pre-EMT genes [34]. Second, we examined previously published cerebral organoid data and observed transcriptional similarities between pretransition neural stem cells and H3K27M gliomas [35]. Finally, we leveraged a recent single-cell RNA-seq dataset to uncover multiple EMT-related transcriptional states in H3K27M tumor cells [18]. Overall, our results suggest that the H3K27M mutation may cause an arrest in development of a neural stem cell type due to lack of H3K27me3 transcriptional control of EMT-related cellular transitions, indicating a developmental window of opportunity for H3K27M mutations to induce gliomagenesis.

Our study highlights the importance of genomic data sharing for rare diseases, such as pHGGs. By combining RNA-seq data from multiple previously published studies, we were able to assemble a cohort of 78 pHGGs, large enough for the differential expression analysis of pHGGs with and without the H3K27M mutation. We used this new cohort of previously published data to derive a novel biological model to describe the molecular pathogenesis of the disease.

## Data Description

The RNA-seq data from bulk clinical pediatric glioma samples used in these analyses were downloaded from the Treehouse cancer compendium v8, which is publicly available at the Treehouse website [36]. All samples passed the RNA-seq quality control analysis used in the curation of the Treehouse cancer compendium [34]. The single-cell glioma RNA-seq data were downloaded from the Gene Expression Omnibus (accession: GSE102130), where they are publicly available. The dataset was log-normalized and filtered for low-expression and low-variability genes. The RNA-seq data from glioma cell lines were accessed with permission from dbGap phs000900.v1.p1, where they are available to other researchers with permission, and all samples passed the RNA-seq quality control analysis used in the curation of the Treehouse cancer compendium [34]. The bulk and single-cell organoid RNA-seq data were downloaded from the Gene Expression Omnibus (accession: GSE106245), which is publicly available. The datasets were log-normalized and filtered for low-expression and low-variability genes.

## Analyses

### Differential expression analysis of pediatric gliomas with and without H3K27M mutation reveals deregulation of genes involved in epithelial-mesenchymal transition

We obtained RNA-seq data from 33 H3K27M pediatric/young adult (ages 0–29 years) high-grade gliomas (pHGGs) and 45 non-

K27M pHGGs from the Treehouse Childhood Cancer Initiative public cancer compendium v8 [36] (Supplementary Table S1). These data came from several cohorts including the Pacific Pediatric Neuro-Oncology Consortium (PNOC), Dr. Michelle Monje's studies, and The Cancer Genome Atlas [37–42].

Using the limma package in R [43], we conducted differential expression analysis between the H3K27M and non-K27M pHGG cohorts. A total of 1,905 genes are differentially expressed between the 2 tumor types (Supplementary Table S2). Using Gene Set Enrichment Analysis (GSEA) and the Molecular Signatures Database (MSigDB) [44], we found 23 biological signaling pathways with significant enrichment in protein-coding genes overexpressed in the H3K27M cohort (Supplementary Table S2). The top 5 most significantly enriched gene pathways included "Hallmark KRAS Signaling Down" (genes repressed by KRAS activation) and the "Hallmark Epithelial Mesenchymal Transition" (Fig. 1A). KRAS pathway enrichment is consistent with a recent study that found RAS signaling to be activated in H3K27M gliomas [45].

Because genes involved in the EMT are regulated by deposition of H3K27me3, an epigenetic transcriptional repressive mark that is lost in H3K27M cells, we were particularly interested in the differential expression of genes involved in the EMT pathway. The hallmark EMT pathway gene list is limited to 200 genes [46], so to comprehensively characterize expression of EMT-associated genes in H3K27M-mutant versus non-K27M tumors, we generated a master list of non-redundant EMT-related genes (n = 437) by merging several MSigDB developmental and cellular EMT-related gene sets (Supplementary Table S2). We included only genes from gene sets focused on EMT as a developmental process and eliminated gene sets that were derived from published studies of EMT in adult carcinomas as per MSigDB [46] because the epithelial nature of those cancers makes those gene sets inapplicable to pediatric gliomas. This list includes genes implicated in both pre- and post-EMT cell states, as well as intermediate EMT cell states and EMT-like processes.

To investigate differential EMT gene expression, we calculated the overlap between the EMT master list and the differentially expressed genes (Supplementary Table S2). We found 49 differentially expressed genes from the EMT master list, indicating potential differential activity of the EMT pathway in H3K27M-mutant gliomas ($P < 7.89^{-14}$, hypergeometric test). Of these genes, 26 were more highly expressed in H3K27M tumors, and the remaining 23 were more highly expressed in non-K27M tumors (Fig. 1B). Further investigation via manual inspection revealed that, in general, the EMT-related genes overexpressed in the H3K27M cohort are associated with the transcriptional profile of cells prior to an EMT-like transition. In contrast, many of the EMT genes underexpressed in H3K27M tumors were associated with a post-EMT cell state.

To statistically quantify the association of the 26 EMT-related genes overexpressed in the H3K27M cohort with pre-EMT cell states in the brain, we manually identified 9 gene sets relating to epithelial cells and early brain development (Supplementary Table S2). The H3K27M-high EMT genes had significant enrichment in 8 of 9 gene sets ($P < 0.1$). In contrast, we calculated the enrichment of 26 randomly selected genes in these 9 gene sets and it was not significant (Supplementary Table S2). The enriched epithelial gene sets include "GO Epithelium Development" ($P < 3.4^{-12}$, hypergeometric test), "GO Epithelial Cell Differentiation" ($P < 3.587^{-4}$), and "GO Neural Tube Formation" ($P < 0.001$). In the developing brain, some of the cells of the neural tube, a pseudostratified epithelium, undergo an EMT in order to migrate [21]. *RHOB*, which plays a role in epithelial cell maintenance in the neural tube [47], is more highly expressed in H3K27M tumors and belongs to 3 of 9 epithelial gene sets. Additionally, *SFRP1* and *SFRP2*, which are crucial in neural tube formation [48], are more highly expressed in H3K27M tumors and belong to 6 of 9 epithelial gene sets.

Importantly, we noted that *SNAI1*, a transcription factor and key regulator of the EMT transcriptional program, is significantly overexpressed in H3K27M tumors ($\log_2$ fold-change [LFC] = 0.6; Fig. 1C). High expression of *SNAI1* is a marker of the beginning of the induction of EMT or EMT-like cellular transitions. If the transition is successful, this is followed by high expression of post-EMT markers *TWIST1* [49], fibronectin (*FN1*) [50], N-cadherin (*CDH2*) [51], and cadherin-11 (*CDH11*) [52]. Using a Mann-Whitney nonparametric significance test, we found significantly reduced expression of all of these genes in H3K27M tumors (*TWIST1* LFC = −1.2, *FN1* LFC = −0.2, *CDH2* LFC = −0.2, *CDH11* LFC = −0.3; Fig. 1C). *TWIST1*, *CDH2*, and *CDH11* are also underexpressed in the H3K27M cohort by the limma analysis.

Because SNAI1 induces EMT-like processes in the developing brain by directly recruiting PRC2 methyltransferase activity for H3K27-trimethylation, a process blocked by the H3K27M mutation, we hypothesized that the occurrence of the H3K27M mutation may promote tumorigenesis by stalling EMT during early neuroepithelial differentiation. To further investigate this hypothesis, we performed comparative RNA-seq expression outlier analysis developed by the Treehouse Childhood Cancer Initiative, which identifies genes with outlier expression in individual samples as compared to a background cohort of highly correlated and disease-matched samples (pan-disease analysis, see Methods) [34]. We identified genes with outlier expression only in non-K27M pHGG samples (but not H3K27M pHGG samples) as compared to a background glioma cohort and noted that 4 of the top 10 enriched pathways were related to EMT, including "TGF-Beta regulation of the extracellular matrix" (adjusted $P = 4.01^{-9}$) and "Extracellular matrix organization" (adjusted $P = 2.98^{-5}$) (Supplementary Fig. S1, Supplementary Table S2).

Finally, because EMT is associated with invasiveness in gliomas, and diffuse midline gliomas are by nature more invasive than hemispheric glioma, we performed an additional analysis restricted to diffuse intrinsic pontine glioma (DIPG) to elucidate the role of the H3K27M mutation in the observed EMT-related transcriptional profiles. The goal of this analysis was to remove any potential histological or location signal that may be influencing the EMT-related gene expression. We used 10 H3 wild-type DIPG samples and 47 H3K27M DIPG samples from Treehouse Cancer Compendium v11.

Limma differential expression analysis revealed 48 genes with higher expression in H3K27M DIPG compared with non-K27M DIPG (Supplementary Table S2). We again computed statistical overlap of these genes with 9 gene sets relating to epithelial cells and early brain development and found significant overlap with 4 of the 9 gene sets ($P < 0.1$, Supplementary Table S2). The enrichment of 48 randomly selected genes in these 9 gene sets was not significant (Supplementary Table S2).

Overall, our multiple analyses of the pHGG RNA-seq cohort suggest that H3K27M pHGG tumors are characterized by a transcriptional profile typically expressed by cells before undergoing an EMT-like transitional process, while non-K27M pHGG tumors are characterized by post-EMT gene expression.

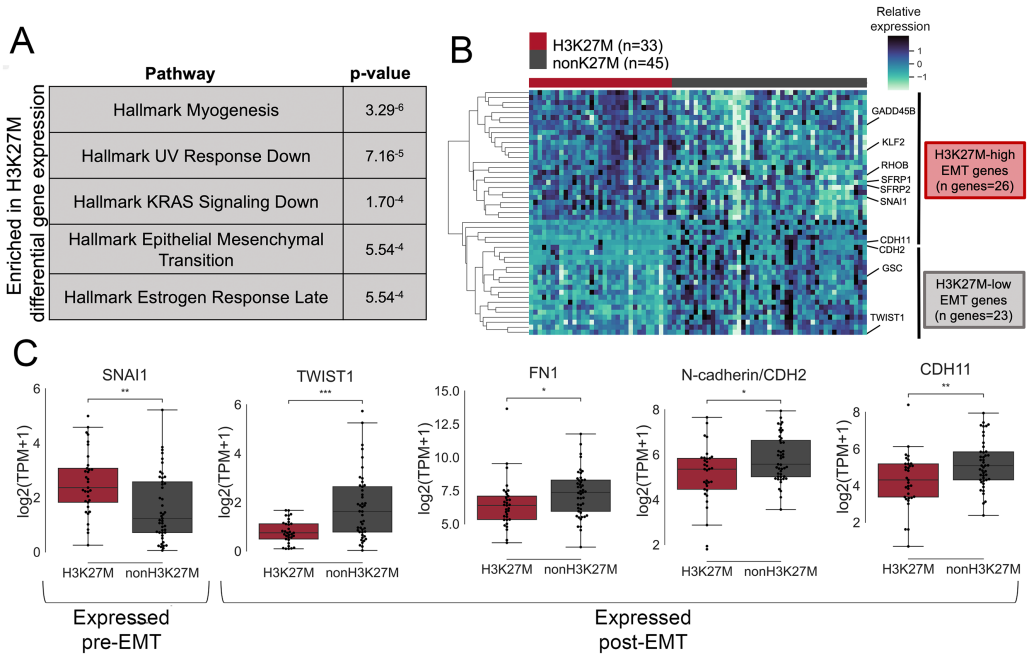



**Figure 1.** The EMT pathway is differentially expressed in H3K27M gliomas as compared with non-K27M gliomas. A, Differential expression analysis of a cohort of H3K27M and non-K27M pHGGs revealed significant enrichment of Hallmark Epithelial Mesenchymal Transition in genes overexpressed in H3K27M gliomas. B, Heat map of differentially expressed EMT genes between H3K27M and non-K27M pHGGs. C, *SNAI1* is overexpressed in H3K27M glioma, while *TWIST1*, *FN1*, *CDH2*, and *CDH11* are underexpressed in H3K27M glioma as compared with non-K27M gliomas (Mann-Whitney significance test; $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$). Boxplot shows median and quartiles, whiskers are at 1.5 interquartile range.

## H3K27M-mediated gliomagenesis is associated with pre-EMT cell types

Consistent with our differential expression analysis, a review of the literature revealed that H3K27M-associated gliomagenesis has been experimentally recapitulated only in cell types that are poised to undergo an EMT differentiation event (Fig. 2A). For example, a combination of H3K27M, *p53* loss, and *PDGFRA* constitutive activation in human neural progenitor cells (NPCs) induced low-grade gliomas when injected into the pons of neonatal mice [16]. These gliomas expressed markers of pre-EMT neuroepithelial cells. Another study found that H3K27M and *Trp53* loss was sufficient for gliomagenesis in the NPCs of embryonic mice in the forebrain and hindbrain [17]. Strikingly, when introduced postnatally, H3K27M and *p53* loss in pre-EMT NPCs was not sufficient for gliomagenesis, although postnatal induction of H3K27M, *Trp53* loss, and *PDGFRA* amplification in pre-EMT NPCs resulted in glioma formation [53, 54]. Additionally, no tumorigenesis was observed upon introduction of H3K27M, *p53* loss, and *PDGFRA* constitutive activation in mature astrocytes, a post-EMT cell type [16]. These observations indicate that experimental H3K27M-mediated gliomagenesis occurs in a pre-EMT cell type.

On the basis of our gene expression analysis and review of the literature, we hypothesized that H3K27M gliomas arise in pre-EMT cell types and retain the EMT-related transcriptional profile of the cell type in which the mutation arises. To compare the expression of the EMT-related genes of interest between H3K27M tumors and normal developing brain cells, we examined total and single-cell RNA-seq data from a human embryonic stem cell–derived cerebral cortex organoid time course experiment (Fig. 2B) [35]. These organoid cultures mimic the early weeks of human prenatal cortical development and generate relevant cell types, uniquely allowing us to investigate early time points in development that are not available in existing human fetal brain datasets. After induction of neural epithelium by week 1, at week 2 radial glia cells and Cajal-Retzius neurons are present, in addition to some remaining neuroepithelial cells. By week 5, the organoids contain populations of radial glia, intermediate progenitors, and deep-layer neurons.

When we investigated EMT-related gene expression in cerebral organoids during gestational weeks 1–6, we noted the presence of 2 EMT-related transcriptional transitions (Fig. 2A, bottom). The first transition starts as *SNAI1* expression peaks in neural stem cells (week 1), coincident with low expression of post-EMT markers *TWIST1*, *CDH2*, *CDH11*, and *FN1*. As differentiation from neural epithelial cells to early radial glia occurs, *SNAI1* expression decreases while post-EMT marker expression increases. In the second transition, as radial glia cells differentiate intermediate progenitor cells, *SNAI1* expression increases once again.

To further characterize the EMT-like transcriptional profiles represented in cerebral organoids, we utilized single-cell RNA-seq data from the cerebral organoids at gestational weeks 3 and 6 [35]. These sample collection times effectively covered all relevant cell type diversity because gestation week 3 organoids con-

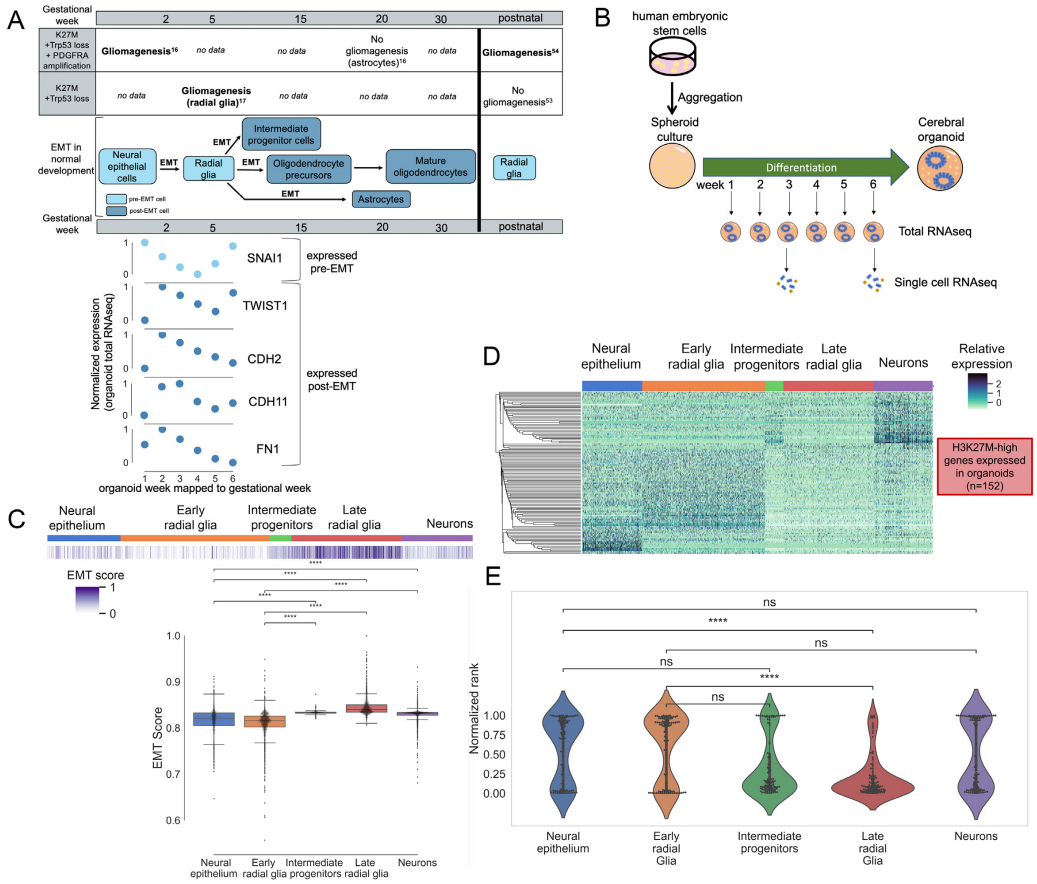



**Figure 2.** H3K27M-specific EMT transcriptional signature is similar to pre-EMT neural stem cell expression in cerebral organoids. A), *In vitro* and *in vivo* experimental H3K27M-associated gliomagenesis occurs exclusively in pre-EMT cell types (top). These cell types are represented in our cerebral organoid assay, and a time course of these organoid cultures represents 2 EMT events in early brain development (bottom). B, Experimental workflow for total RNA-seq and single-cell RNA-seq from a human embryonic stem cell–derived cerebral cortex organoid time course experiment. C, Single cells from cerebral organoids were scored for EMT completion. Pre-EMT neural epithelium and early radial glia were least enriched for the EMT score, while post-EMT intermediate progenitors, late radial glia, and neurons were the most enriched. Boxplot shows median and quartiles, whiskers are at 1.5 interquartile range. D, A signature of genes differentially expressed in H3K27M gliomas and expressed in cerebral organoids shows highest expression in pre-EMT neural epithelium and early radial glia. E, EMT-related genes highly expressed in H3K27M-mutant gliomas are also highly expressed in neural epithelium and early radial glia (Mann-Whitney significance test; $^{*}P$ value $< 0.05$, $^{**}P$ value $< 0.01$, $^{****}P$ value $< 0.0001$). The violin plots show a kernel density estimation of the data distribution.

tain substantial populations of neural epithelial cells, early radial glia cells, and Cajal-Retzius neurons, while week 6 organoids are composed of late radial glia cells, intermediate progenitors, and immature neurons. We scored the EMT status of each cell using a gene signature representing EMT completion and a previously published scoring method based on aggregate expression of the gene set as compared with a control gene set (Fig. 2C, Supplementary Table S3, see Methods) [18, 55–58]. Neural epithelial and early radial glia cells show significantly lower EMT scores than post-EMT intermediate progenitors, late radial glia, and neurons (Mann-Whitney test, $P < 0.0001$). This shows that our assay contains distinct populations of pre- and post-EMT cerebral cells and is consistent with the levels of *SNAI1*, *CDH2*, *CDH11*, *FN1*, and *TWIST1* in the bulk weeks 1–6 organoid data. This dataset enables us to investigate transcriptional similarities between H3K27M-mutant gliomas and normal pre-EMT cell types during neural development.

We then examined the expression of genes overexpressed in H3K27M gliomas in the single-cell organoid RNA-seq dataset to see which normal cell type is most similar to H3K27M glioma cells. Of the 1,180 H3K27M-overexpressed genes, 152 genes passed the single-cell RNA-seq expression filter (Supplementary Table S3, see Methods). Hierarchical clustering of the expression profiles of these genes in normal cell types during neural development revealed highest expression in pre-EMT neural epithelium and early radial glia (Fig. 2D). We then ranked this gene signature on the basis of each gene's expression in each cell type (see Methods). We found that this signature is ranked most

highly in pre-EMT neural epithelium and in early radial glia ($P < 0.05$, Fig. 2E).

Overall, these results suggest that the differential EMT-related gene expression observed in our tumor cohort is consistent with identifiable stages in cyclic EMT-like transcriptional programs in the normal developing brain and that H3K27M tumor cells resemble normal developing brain cells at a point where they are expressing a pre-EMT transcriptional profile.

## Single-cell profiling of H3K27M gliomas reveals groups of cells with different EMT-related transcriptional profiles

We used recently published single-cell RNA-seq data from 6 H3K27M and 2 H3 wild-type (H3WT) gliomas to directly investigate the EMT-related transcriptional profiles of single-cell populations within each tumor type [18]. One of the H3K27M tumors harbors the mutation in the *HIST1H3B* gene (referenced as H3.1K27M), while the remaining 5 H3K27M tumors harbor the mutation in the *H3F3A* gene (referenced as H3.3K27M).

We performed hierarchical clustering of 3,057 tumor cells using 207 genes from the EMT master list that passed expression filters (see Methods, Supplementary Table S4) [59]. Nine EMT-related clusters were discovered and named A–I (Fig. 3A, Supplementary Table S4). Cluster gene signatures were identified by assigning each cluster the genes with maximum mean expression in that cell cluster across the dataset (Supplementary Table S4).

We assigned cluster function on the basis of manual review of genes in each signature, and observed several populations of cells whose presence in this dataset has already been noted [18]. Cluster I is composed predominantly of non-malignant immune cells, indicated by comparatively high expression of immune markers such as *CD68* [60]. Cluster H resembles oligodendrocytic cells, with highest expression of *PADI2*, *PMP22*, and *RHOA*, and Cluster G resembles oligodendrocyte precursor cells with the highest expression of *PDGFRA* [61–64]. The presence of each of these cell types has already been noted in H3K27M gliomas [18].

However, the remaining clusters are defined by genes associated with EMT. We again scored the EMT status of each cell with a gene signature representing EMT completion (Fig. 3A, see Methods) [18, 55–58]. Clusters D, E, and F scored the lowest overall, while Clusters A, B, and C scored the highest overall. Cluster relationships are shown with Uniform Manifold Approximation and Projection (UMAP) in Fig. 3B, and expression patterns of selected genes relating to transcriptional stages of EMT-like transitions are shown in the lower panel of Fig. 3B. Of the genes identified in the bulk RNA-seq analysis (Fig. 1C), only *FN1*, *CDH2*, and *CDH11* were expressed in the glioma single-cell RNA-seq data, so we also visualized *VIM* as a post-EMT marker and *SFRP1* as a pre-EMT marker.

In keeping with our previous analysis, we noted that Cluster A, which is composed mainly of H3WT glioma cells, strongly resembles post-EMT cells and most highly expresses canonical post-EMT markers including *CDH2*, *CDH6*, and *VIM* [65, 66]. This is consistent with our observation that non-K27M gliomas transcriptionally resemble a post-EMT state as compared to H3K27M in the bulk RNA-seq pHGG cohort.

Interestingly, within the clusters composed predominantly of H3K27M cells, we observed multiple EMT-related transcriptional profiles. Cluster B, composed of H3K27M cells, had highest expression of post-EMT markers including *CDH11* and *FN1*, potentially indicating a subclonal population of cells that differentiated through alternative means. Thus, we defined Clusters A and B "post-EMT."

In contrast, H3K27M-expressing Clusters E and F cells exhibit comparatively the highest expression of several genes known for their expression in pre-EMT cell types, including *CADM1*, *PTEN*, *CTNNB1*, and *SFRP1* [48, 67–70]. Therefore, we defined Clusters E and F "pre-EMT." In contrast, Cluster C has comparatively the highest expression of only 10 genes and has no clear expression profile of any stage of EMT, so we defined Cluster C "EMT-ambiguous."

Cluster D was defined "EMT-intermediate" because it displays high expression of genes normally expressed while the EMT process is taking place, without a clear bias towards epithelial or mesenchymal gene expression, including *SMAD2* and *VCAN*, which are activated during the EMT process rather than before or after [71, 72].

## Histone H3.1K27M glioma cells express a different EMT-related transcriptional profile than H3.3K27M glioma cells

Further examination revealed that cluster D mainly consists of cells from the H3.1K27M-mutant tumor. H3.1 and H3.3K27M characterize 2 functionally different subtypes of H3K27M gliomas; H3.1K27M gliomas are comparatively rare but have a slightly better prognosis [40, 73]. Normally, histone H3.3 is preferentially located at active chromatin [74–76]. This leads to distinct patterns of epigenetic reprogramming in each histone variant, where loss of the H3.3K27me3 mark is directly correlated with areas of H3.3 genomic enrichment, while H3.1K27me3 loss is higher at intergenic regions [76, 77]. Because the H3K27M mutation is known to induce dose-dependent inhibition of PRC2 methyltransferase, this suggests that the localized distribution of histone H3.3 may result in higher local inhibition of PRC2 and loss of H3K27me3 at H3.3K27M sites [53, 76]. Because precise control of gene transcription via active chromatin is necessary for EMT-like developmental cell state transitions, a H3.3K27M mutation would be particularly damaging to proper regulation of these processes. Indeed, functional analysis of enhancer regions in H3.3K27M-expressing NPCs revealed enrichment of regions positively regulating EMT-related genes, indicating that H3.3 active chromatin regions are directly involved in transcriptional control of EMT-related genes [76]. This suggests that EMT-poised H3.3K27M cells will be unable to properly complete the transition owing to lack of transcriptional control.

Accordingly, we observed EMT-intermediate or E/M hybrid expression genes in glioma single-cell Cluster D, which has a substantial number of H3.1K27M glioma cells. We hypothesized that H3.1K27M cells may be more differentiated than H3.3K27M cells.

To investigate this hypothesis further, we subset the single-cell glioma RNA-seq data to 2,458 cells with H3.1K27M or H3.3K27M mutation and performed the Wilcoxon rank-sum test to identify genes overexpressed in each variant group (Supplementary Table S4, Supplementary Fig. S2). Consistent with our previous observations, GSEA of Gene Ontology (GO) gene sets (Fig. 4B, Supplementary Table S4) revealed enrichment of epithelial gene sets in H3.3K27M compared with H3.1K27M (GO Adhesion pathways, GO Neurogenesis, GO Embryo Development) and mesenchymal gene sets in H3.1K27M compared with H3.3K27M (GO EMT pathway, GO Mesenchymal Cell Differentiation, and GO Mesenchyme Development). Additionally, scoring of all cells for EMT completeness shows that H3.1K27M cells score significantly higher overall than H3.3K27M cells,

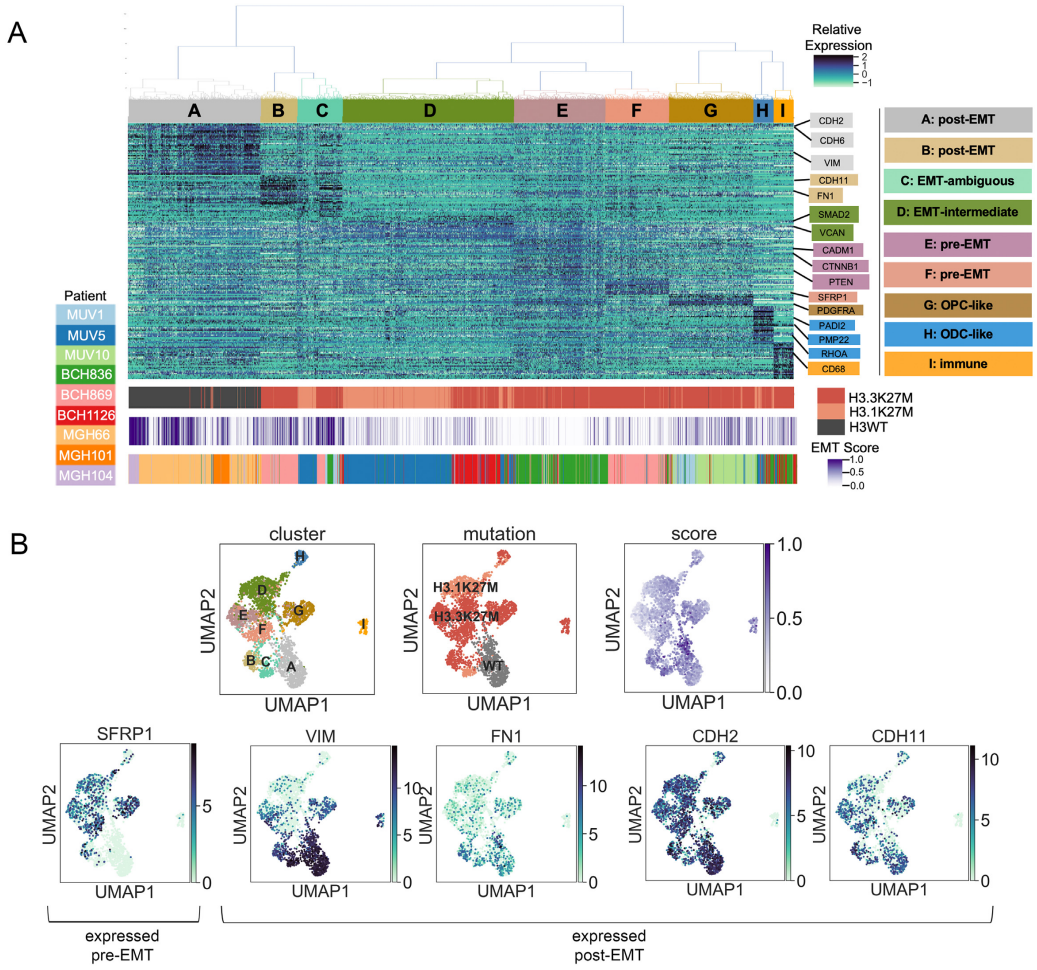



**Figure 3.** Single-cell RNA sequencing of H3K27M and non-K27M gliomas reveals multiple EMT stages within tumors. A, Expression heat map showing hierarchical clustering of 3,057 cells from 6 H3K27M and 2 non-K27M high-grade gliomas, with a master list of EMT genes. Nine clusters (A–I) were assigned gene signatures on the basis of maximum mean gene expression in each cluster, and clusters were classified on the basis of manual review of each gene signature. Histone H3 mutation status and EMT score are shown at the bottom of the heat map (ODC: oligodendrocyte; OPC: oligodendrocyte precursor). B, UMAP dimensionality reduction projection of the same expression data as the heat map and labeled by cluster, histone H3 mutation status, and EMT score. Expression of selected pre-EMT and post-EMT genes is shown in the bottom panel.

while non-K27M cells score significantly higher than either mutant cell type (Supplementary Fig. S3). However, because the H3.1K27M cells come from a single tumor, we performed additional analysis to investigate this observation.

We cultured DIPG primary cell lines isolated in a previous study to investigate the expression of EMT markers in H3.3K27M, H3.1K27M, and non-K27M glioma cells [78]. Morphologically, we observed that when cultured in serum-free conditions, the H3.1K27M cell lines preferentially grow attached to the flask (4 of 5 cell lines), while the H3.3K27M cells preferentially grow as neurospheres (8 of 9 cell lines) (Fig. 4C). Because differentiation out of the neurosphere state is accompanied by attachment and increased expression of N-cadherin, this morphological trend is consistent with our hypothesis that H3.1K27M cells exist in a more differentiated state than H3.3K27M cells [79].

We analyzed RNA-seq data from 3 DIPG cell lines to compare the expression of EMT genes (SU-DIPG-IV is H3.1K27M mutant; SU-DIPG-VI and JHH-DIPG1 are H3.3K27M mutant). We used 4 replicate samples from each of SU-DIPG-IV and SU-DIPG-VI and 3 replicate samples from JHH-DIPG1. Each sample was scored using a gene signature of EMT completion (see Methods), and the H3.1K27M samples scored significantly higher than the H3.3K27M samples (Fig. 4D, $P < 0.05$).

We then performed RT-PCR to quantify the expression of *FN1* and *CDH2*, canonical post-EMT genes that were previously identified as differentially expressed by Mann-Whitney test in the

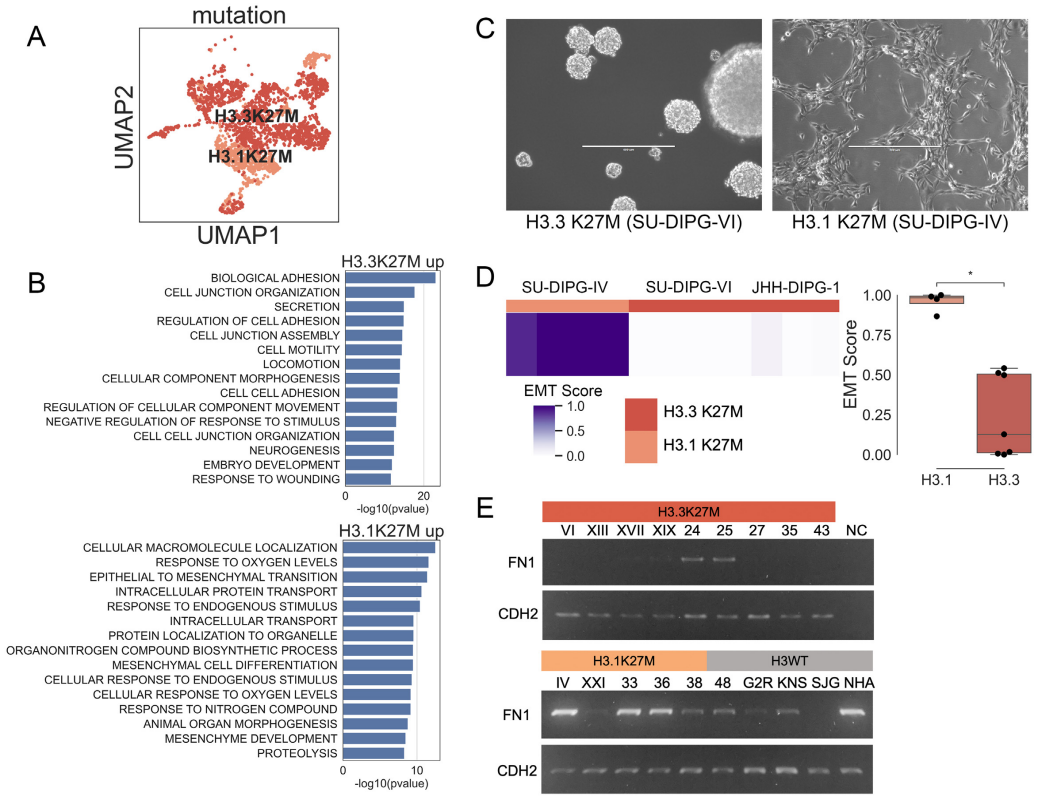



**Figure 4.** H3.1K27M glioma cells express a different EMT-related transcriptional profile than H3.3K27M glioma cells. A, UMAP dimensionality reduction of 2,458 histone-mutant glioma single cells. B, Gene set enrichment analysis of genes overexpressed in H3.3K27M vs H3.1K27M (top) or H3.1K27M vs H3.3K27M (bottom) by Wilcoxon rank-sum test using glioma single-cell RNA-seq data. C, Representative images of H3.1K27M and H3.3K27M glioma–derived cell cultures. Scale bar 400 $\mu$m. D, Total RNA sequencing datasets from glioma cell lines were scored for EMT completeness (4 samples from SU-DIPG-IV, 4 samples from SU-DIPG-VI, and 3 samples from JHH-DIPG1). Scoring is shown in a heat map and a box plot (Mann-Whitney significance test; $^*P < 0.05$). E, RT-PCR of *FN1* and *CDH2* expression in glioma primary cell cultures (all numbered lines are SU-DIPG).

bulk glioma RNA-seq analysis (Fig. 4E, full-length gel in Supplementary Fig. S4). We attempted to quantify E-cadherin/*CDH1* because it is a canonical pre-EMT marker, but the levels were so low as to be undetectable by RT-PCR in these cell lines [RNA-seq $< 1.0$ $\log_2(\text{TPM} + 1)$]. We compared 9 H3.3K27M cell lines (SU-DIPG-VI, XIII, XVII, XIX, 24, 25, 27, 35, and 43) with 5 H3.1K27M cell lines (SU-DIPG-IV, XXI, 33, 36, and 38) and included 5 H3 wild-type lines (SU-DIPG-48, pcGBM2R, KNS42, SJ-GBM2, and normal human astrocytes hTERT) and a negative RT-PCR control (NC). Overall, the H3 wild-type and H3.1K27M cell lines appear to more highly express both post-EMT markers, in keeping with the bulk and single-cell RNA-seq analyses. Our computational and *in vitro* observations are consistent with a recent study indicating that H3.1K27M tumor cells are overall more differentiated than H3.3K27M tumor cells [76]. Our results are also consistent with previous studies on EMT in pediatric gliomas, which first found a mesenchymal subtype of DIPG and subsequently discovered that H3.1K27M-mutant gliomas express genes associated with a more mesenchymal subtype of glioblastoma [73, 80].

Overall, these data suggest that the histone H3K27M mutation is associated with a preferentially early or pre-EMT cell state as compared with non-K27M cells but that H3.1K27M cells may represent a somewhat later or intermediate-EMT cell state as compared with H3.3K27M cells.

## Discussion

H3K27M diffuse midline gliomas are aggressive tumors generally occurring in early childhood in the hindbrain or midline. These tumors have poor prognosis and do not respond to standard chemotherapies for adult gliomas [81]. In contrast to most adult cancers, pediatric cancers, including pediatric gliomas, are thought to occur due to a developmental stall relating to epigenetic dysregulation of normal cellular differentiation pathways [15, 40, 82]. H3K27M diffuse midline glioma cells lose EZH2-deposited H3K27me3 epigenetic transcriptional control markers, which are known to play crucial roles in cell differentiation and development in the brain [6]. In particular, normal H3K27me3 deposition controls neural cell differentiation through multiple EMT processes [22]. Research has implicated the EMT in pediatric gliomas [80, 83, 84], particularly those with a more invasive phenotype. Histone deacetylase inhibitor

treatment of *in vitro* pHGG cells reversed mesenchymal phenotypes, in keeping with a model in which the interplay between H3K27 acetylation and methylation controls EMT-related transcriptional states [38]. We hypothesized that loss of H3K27me3 in H3K27M-mutant gliomas may lead to a stall in EMT processes in normal brain development.

In this study, we observed that various canonical EMT-inducing genes are significantly overexpressed in H3K27M-mutant pHGGs, compared with non-K27M pHGGs, while many canonical mesenchymal markers are underexpressed in H3K27M pHGGs as compared with the non-K27M tumors. In particular, we noted higher expression of the pre-EMT transcription factor *SNAI1* in H3K27M-mutant gliomas. Because *SNAI1* relies on PRC2 and H3K27me3 to facilitate EMT through gene expression regulation, this may indicate an arrest in the EMT process. The existence of a hybrid epithelial/mesenchymal phenotype is well established: the result of a partial EMT is the expression of both epithelial and mesenchymal genes [85]. Studies have shown that a hybrid E/M phenotype may indicate a worse prognosis than mesenchymal-only states in solid tumors [85–87].

We hypothesized that if H3K27M mutation prevents full EMT, neural stem cells harboring H3K27M may be forced to retain a proliferative, stem cell phenotype, eventually leading to tumorigenic development. Accordingly, we observed from extensive literature review that experimental induction of H3K27M-associated gliomas has occurred exclusively in pre-EMT cell types, and that 2 consecutive EMT-like transcriptional transitions occur early in normal brain development.

Single-cell RNA-seq from H3K27M and non-K27M tumors confirmed a post-EMT expression signature in the non-K27M cells and also revealed subsets of H3K27M cells with different EMT-related transcriptional profiles. Specifically, we observed an intermediate EMT signature in the H3.1K27M cells as compared with the H3.3K27M cells. This was also observed in bulk RNA-seq and *in vitro* RT-PCR analysis. We hypothesize that because the H3.1K27M mutation is not concentrated at active chromatin, it has less repressive power as specific developmental processes such as EMT are activated over time. If a subset of H3.1K27M cells are able to differentiate, this may explain why H3.1K27M gliomas have a slightly better prognosis.

To conclude, we mined 3 publicly available RNA-seq datasets from pediatric gliomas and cerebral organoids to generate a hypothesis for the gliomagenesis of H3K27M gliomas. We propose that the H3K27M mutation is tumorigenic when the mutational hit occurs in a cell poised to undergo an EMT-like cell state transition, due to the dependence of EMT-associated transcriptional activity on the correct timing of the H3K27me3 mark (Fig. 5). More work is needed to characterize the observed difference in the EMT-associated transcriptional profiles between the H3.1 and H3.3K27M variants. Additionally, a limitation of our study is that it is difficult to isolate the role of the H3K27M mutation from other factors such as histology and tumor location. Future studies will focus on EMT-related transcriptional programs in cellular models with inducible H3K27M expression to further characterize the molecular interplay between the H3K27M mutation and EMT in developing brain cells.

Taken together, our results hold important implications for better understanding the developmental origin and timing of these aggressive and untreatable cancers. Furthermore, the presence of an epigenetically driven differentiation stall may imply that a pharmacological methylation agent or a pro-differentiation therapy may aid in future treatment of H3K27M-mutant tumors [88].

## Potential Implications

Our study holds implications for other diseases because H3K27M mutation is not exclusive to diffuse midline gliomas. It can also be found in a fraction of pediatric ependymomas and medulloblastomas [89]. Interestingly, ependymomas located in the posterior fossa typically do not harbor the H3K27M mutation but exhibit the K27M-associated H3K27 hypomethylation phenotype. Thus, the proposed differentiation stall and an associated EMT transcriptional signature as a result of H3K27me3 loss may also apply to these cancers. Beyond the SNAI1-H3K27me3 axis, EMT is also regulated by other epigenetic marks [90]. Given the epigenetically dysfunctional nature of many pediatric cancers [15], EMT arrest could conceivably play a role in the oncogenesis of these tumors as well.

## Methods

### Glioma bulk RNA sequencing data

Gene expression data from 78 pHGG samples were downloaded from the Treehouse Childhood Cancer Initiative public compendium v8 (Tumor Compendium v8 Public) [36]. All samples in the compendium have been uniformly processed using the UC Santa Cruz TOIL RNA-seq pipeline (v3.3.4) [91]. This dataset ($n = 58{,}581$ genes) is in transcripts per million (TPM) and normalized by $\log_2(\text{TPM} + 1)$. We divided the dataset into 33 H3K27M-mutant samples and 45 non-K27M samples and performed differential expression analysis of all genes between the 2 groups using R library limma v3.34.9 in R v3.3.4. We performed GSEA of the resulting 1,905 differentially expressed genes ($P < 0.1$) with MSigDB v7.0 on the GSEA/MSigDB website v6.4 (Supplementary Table 2). Because the EMT pathway was in the top 5 most significantly enriched pathways in H3K27M overexpressed genes, we created a non-redundant master list of EMT genes ($n = 437$) by merging 7 EMT-related MSigDB pathways and by identifying EMT-related genes through manual literature curation (Supplementary Table 2).

We performed pan-disease outlier analysis on all the pHGG samples using Treehouse CARE (see Availability of Source Code and Requirements section) against the Treehouse Cancer Compendium v10. Pan-disease outlier analysis identifies genes with outlier expression in each sample of interest as compared to a background cohort of tumors identified as most similar (in this analysis, the background cohort was 37 pediatric gliomas, 19 young adult gliomas, 18 pediatric glioblastomas, and 4 young adult glioblastomas) [34]. We identified a list of genes with outlier expression in the non-K27M pHGG samples that did not also have outlier expression in the H3K27M pHGG samples, and performed GSEA using Enrichr in the GSEApy package (gseapy-v0.9.17) [92] against the BioPlanet_2019 library with *P*-value cutoff 0.05 (outlier genes and enriched pathways in Supplementary Table 2). We used the EnrichmentMap app in Cytoscape to visualize functionally similar clusters of enriched pathways [93].

### Cerebral organoid RNA sequencing data (bulk and single cell)

Gene expression data (TPM) from 6 weekly time points of human cerebral organoid growth were downloaded from accession GSE106245 [35]. Organoid weeks 0–5 were converted to gestational weeks 1–6 and duplicate gene measurements were averaged. For Fig. 2A, expression of each gene was normalized in the range 0–1. Single-cell RNA-seq data from weeks 2 and 5 (gestational weeks 3 and 6) cerebral organoids were downloaded from

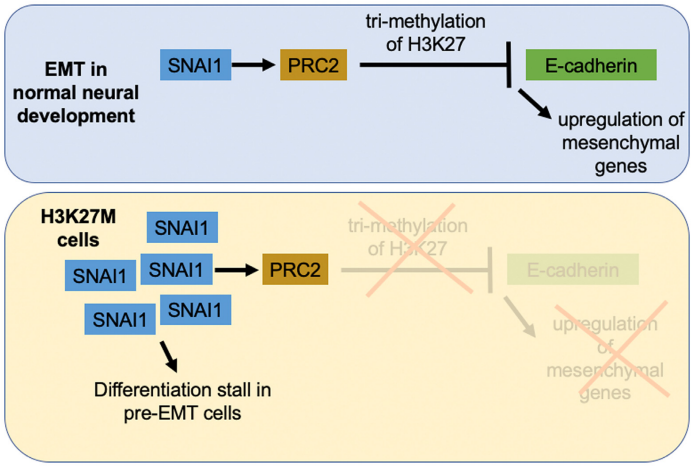



**Figure 5.** Proposed model for EMT stall in H3K27M cells. We propose that H3K27M cells retain high levels of SNAI1 expression but remain stalled in a pre-EMT state owing to inability of PCR2 to tri-methylate H3K27.

accession GSE106245 [35]. Expression data were filtered to remove genes with expression in <10% of cells. Cell types were assigned using a list of marker genes (Supplementary Table S3).

## Glioma single-cell RNA sequencing data

Smart-seq2 RSEM TPM single-cell RNA-seq data from 3,057 glioma cells were downloaded from accession GSE102130 [18]. Data were $\log_2$-normalized and filtered to remove genes with expression in <20% of cells. The cells per tumor remaining after filtering are as follows: MUV1: 146, MUV5: 708, MUV10: 286, BCH836: 527, BCH869: 492, BCH1126: 299, MGH66: 442, MGH101: 92, MGH104: 65. Hierarchical clustering of all cells was performed using the Python scipy.cluster.hierarchy function (scipy v1.4.1) after subsetting to a non-redundant master list of EMT genes (n = 437, Supplementary Table S2). Of these genes, 207 passed the expression filter and were included in the hierarchical clustering. The clustering results were plotted using the scipy.cluster.hierarchy.dendrogram function with threshold set to 3.5. Gene signatures for each cluster were assigned by identifying the cluster in which each gene has maximum mean expression, and assigning that gene to that cluster. For UMAP visualizations, Leiden clustering was performed on the single-cell data using the scanpy.tl.leiden function (scanpy v1.4.5.post1) with resolution set to 0.5 and top 10 principal components used as input.

## DIPG cell lines

The patient-derived DIPG cell lines (SU-DIPG-IV, SU-DIPG-VI, SU-DIPG-XIII, SU-DIPG-XVII, SU-DIPG-XIX, SU-DIPG-XXI, SU-DIPG-24, SU-DIPG-25, SU-DIPG-27, SU-DIPG-33, SU-DIPG-35, SU-DIPG-36, SU-DIPG-38, SU-DIPG-48) were kindly provided by Dr. Michelle Monje (Stanford University School of Medicine, Stanford, CA) [38]. SU-DIPG-IV, SU-DIPG-XXI, SU-DIPG-33, SU-DIPG-36, and SU-DIPG-38 cells harbor a H3.1K27M mutation while SU-DIPG-VI, SU-DIPG-XIII, SU-DIPG-XVII, SU-DIPG-XIX, SU-DIPG-24, SU-DIPG-25, SU-DIPG-27, SU-DIPG-35, and SU-DIPG-43 cells harbor a H3.3K27M mutation. SU-DIPG-48 and glioblastoma cell line SU-pcGBM-2 are H3WT. Glioblastoma H3WT cell lines, as well as KNS-42 (BCRJ Cat# 0295, RRID:CVCL_0378), SJ-GBM2 (RRID:CVCL_M141), and 1 normal astrocyte cell line NHA hTERT, were kindly provided by Prof. Sameer Agnihotri (University of Pittsburgh Medical Center Children's Hospital of Pittsburgh, Pittsburgh, PA). The Universal Mycoplasma Detection Kit (ATTC, Manassas, VA) was used for testing SU-DIPG-XIII, XVII, XIX, and VI latest on 10 January 2020. All cells were cultured in tumor stem medium containing 50X B-27 Supplement Minus Vitamin A (Invitrogen, Waltham, MA), H-EGF at 20 ng/mL (Shenandoah Biotechnology, Warwick, PA), H-FGF-basic-154 at 20 ng/mL (Shenandoah Biotechnology, Warwick, PA), H-PDGF-AA at 10 ng/mL (Shenandoah Biotechnology, Warwick, PA), H-PDGF-BB at 10 ng/mL (Shenandoah Biotechnology, Warwick, PA), and 0.2% heparin solution at 2 $\mu$g/mL (STEMCELL Technologies, BC, Canada). All experiments used cells collected within 5 passages after thawing. The cells were passaged by the treatment of TrypLE (Gibco, Waltham, MA) and DNAse I (Worthington, Lakewood, NJ) rocking at 37°C for 5–15 minutes, then HBSS (Corning, Corning, NY) was added to deactivate TrypLE. The cells were transferred to new Nunc EasYFlask Cell Culture Flasks (ThermoFisher Scientific, Waltham, MA) and grown in tumor stem medium as previously described. The bulk RNA-seq data from lines SU-DIPG-VI, SU-DIPG-IV, and JHH-DIPG1 were obtained with permission from Dr. Michelle Monje from dbGap accession phs000900.v1.p1.

## RNA extraction and RT-PCR

Total RNA was extracted from cell pellets using the Quick-RNA Miniprep Kit (Zymo Research, Irvine, CA). Complementary DNA was synthesized from 1 $\mu$g of total RNA using Oligo(dT)20 primers and the SuperScript III First Strand Synthesis System (Invitrogen, Waltham, MA). PCR was performed using KAPA HiFi HotStart ReadyMixPCR Kit (KAPA Biosystems, Wilmington, MA), 50 ng of template DNA, and the appropriate primers and 27 PCR cycles and an annealing temperature of 64°C. *CDH2* primer sequences were as follows: forward: ggcttaatggtgattttgctcag; reverse: tccataccacaaacatcagcac. *FN1* primer sequences were as follows: forward: cttgaaccaacctacggatgac; reverse: tcccatcat-

cataacacgttgc. Primer oligos were purchased from Integrated DNA Technologies.

### Data Analysis

All statistical comparisons were performed with a 2-sided Mann-Whitney test, with measurements taken from distinct samples without assumption of normality, and Benjamini-Hochberg multiple testing correction was applied. Single-cell and bulk tumor samples were scored for EMT activity using a manually curated set of mesenchymal genes and a previously published scoring method based on aggregate expression of the gene set as compared to a control gene set (Supplementary Table S3) [18, 56, 94].

## Availability of Source Code and Requirements

Code for figures and data analysis: https://github.com/lauren-sanders/EMT-paper
Code for outlier analysis: https://github.com/UCSC-Treehouse/CARE/
License: Apache-2.0
Operating system: Platform independent
Programming languages: Python, R
Other requirements: Python 3.8 or higher, R 3.3 or higher

## Data Availability

All data used for this article are available at the following websites or accession numbers: publicly available: (i) bulk glioma RNA-seq: [36], (ii) cerebral organoid RNA-seq: GSE106245, (iii) glioma single-cell RNA-seq: GSE102130. Data available with permission for the glioma cell line RNA-seq data dbGap phs000900.v1.p1. All supporting data and materials are available in the *GigaScience* GigaDB database [95].

## Additional Files

Supplementary Figure S1. Comparative gene expression analysis shows enrichment of post-EMT and mesenchymal pathways in genes with outlier expression in H3WT pHGG tumors.
Supplementary Figure S2. Top 30 differentially expressed genes by Wilcoxon rank-sum test comparing H3.3K27M mutant glioma cells with H3.1K27M mutant glioma cells.
Supplementary Figure S3. Continuum of EMT completeness scores in glioma single cell RNA-seq data.
Supplementary Figure S4. Full-length RT-PCR gel images for FN1 and CDH2 quantification.
Supplementary Table S1. Patient characteristics.
Supplementary Table S2. Differentially expressed genes between H3K27M and H3WT gliomas.
Supplementary Table S3. Genes used for EMT scoring.
Supplementary Table S4. Glioma single-cell clusters and genes.

## Abbreviations

DIPG: diffuse intrinsic pontine glioma; E/M: epithelial/mesenchymal; EMT: epithelial-mesenchymal transition; GO: gene ontology; GSEA: gene set enrichment analysis; H3WT: histone 3 wild-type; LFC: $\log_2$ fold-change; MSigDB: Molecular Signatures Database; NPC: neural progenitor cell; ODC: oligodendrocyte cell; OPC: oligodendrocyte precursor cell; pHGG: pediatric high-grade glioma; PNOC: Pacific Pediatric Neuro-Oncology Consortium; TPM: transcripts per million; UMAP: Uniform Manifold Approximation and Projection; UCSC: University of California Santa Cruz.

## Ethics Statement

The protocols for the PNOC-003 trial, Dr. Michelle Monje's studies, Dr. Mariella Filbin's studies, The Cancer Genome Atlas, the Children's Brain Tumor Tissue Consortium, the International Cancer Genome Consortium, and the University of Michigan Clinical Sequencing Exploratory Research have been previously described [18, 37–42]. The UCSC Treehouse Childhood Cancer Initiative protocol was approved by the UCSC Institutional Review Board (No. HS2648) [34].

## Competing Interests

The authors declare that they have no competing interests.

## Funding

This study was funded by American Association for Cancer Research NextGen Grant for Transformative Cancer Research Award (O.M.V.), St Baldrick's Foundation Consortium Award and Emily Beazley Kures for Kids Fund Hero Award (D.H., O.M.V., S.R.S.), Alex's Lemonade Stand Foundation for Childhood Cancer Research, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, and Live for Others Foundation, The Schmidt Futures Foundation (D.H.), CIRM Shared Stem Cell Facilities (CL1-00506) award to UCSC. A.C. is supported by the T32GM133391 Training Program in Molecular, Cell, and Developmental Biology. D.H. is a Howard Hughes Medical Institute Investigator. O.M.V. holds the Colligan Presidential Chair in Pediatric Genomics.

## Authors' Contributions

Analysis and manuscript authorship: L.M.S. and A.C.
Single-cell organoid cell type gene ranking: L.S.
Experimental work: A.C., A.v.d.B., and M.C.
Treehouse cancer compendium and manuscript review: H.C.B., E.T.K., J.P., K.L., A.G.L., and I.B.
Funding, scientific oversight, and manuscript review: D.H., S.R.S., and O.M.V.

## Acknowledgments

We gratefully acknowledge Dr. Michelle Monje and Prof. Sameer Agnihotri, who provided cell lines used in this study.

## References

1. Juratli TA, Qin N, Cahill DP, et al. Molecular pathogenesis and therapeutic implications in pediatric high-grade gliomas. Pharmacol Ther 2018;**182**:70–79.
2. Chan K-M, Fang D, Gan H, et al. The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes Dev 2013;**27**:985–90.
3. Johung TB, Monje M. Diffuse intrinsic pontine glioma: New pathophysiological insights and emerging therapeutic targets. Curr Neuropharmacol 2017;**15**:88–97.
4. Jones C, Baker SJ. Unique genetic and epigenetic mechanisms driving paediatric diffuse high-grade glioma. Nat Rev Cancer 2014;**14**, doi:10.1038/nrc3811.

5. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: A summary. Acta Neuropathol 2016;**131**:803–20.
6. de Vries NA, Hulsman D, Akhtar W, et al. Prolonged Ezh2 depletion in glioblastoma causes a robust switch in cell fate resulting in tumor progression. Cell Rep 2015;**10**:383–97.
7. Mohammad F, Weissmann S, Leblanc B, et al. EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. Nat Med 2017;**23**:483–92.
8. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature 2011;**469**:343–9.
9. Mohn F, Weber M, Rebhan M, et al. Lineage-specific Polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 2008;**30**:755–66.
10. Roidl D, Hacker C. Histone methylation during neural development. Cell Tissue Res 2014;**356**:539–52.
11. Sher F, Boddeke E, Olah M, et al. Dynamic changes in Ezh2 gene occupancy underlie its involvement in neural stem cell self-renewal and differentiation towards oligodendrocytes. PLoS One 2012;**7**:e40399.
12. Sher F, Rössler R, Brouwer N, et al. Differentiation of neural stem cells into oligodendrocytes: Involvement of the Polycomb group protein Ezh2. Stem Cells 2008;**26**:2875–83.
13. Akizu N, Martínez-Balbás MA. EZH2 orchestrates apicobasal polarity and neuroepithelial cell renewal. Neurogenesis (Austin) 2016;**3**:e1250034.
14. Zemke M, Draganova K, Klug A, et al. Loss of Ezh2 promotes a midbrain-to-forebrain identity switch by direct gene derepression and Wnt-dependent regulation. BMC Biol 2015;**13**:103.
15. Filbin M, Monje M. Developmental origins and emerging therapeutic opportunities for childhood cancer. Nat Med 2019;**25**:367–76.
16. Funato K, Major T, Lewis PW, et al. Use of human embryonic stem cells to model pediatric gliomas with H3.3K27M histone mutation. Science 2014;**346**:1529–33.
17. Pathania M, De Jay N, Maestro N, et al. H3.3K27M cooperates with Trp53 loss and PDGFRA gain in mouse embryonic neural progenitor cells to induce invasive high-grade gliomas. Cancer Cell 2017;**32**:684–700.e9.
18. Filbin MG, Tirosh I, Hovestadt V, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science 2018;**360**:331–5.
19. Viebahn C. Epithelio-mesenchymal transformation during formation of the mesoderm in the mammalian embryo. Acta Anal 1995;**154**:79–97.
20. Duband J-L. Diversity in the molecular and cellular strategies of epithelium-to-mesenchyme transitions: Insights from the neural crest. Cell Adh Migr 2010;**4**:458–82.
21. Kalcheim C. Epithelial-mesenchymal transitions during neural crest and somite development. J Clin Med Res 2015;**5**, doi:10.3390/jcm5010001.
22. Zou S, Zhang D, Xu Z, et al. JMJD3 promotes the epithelial-mesenchymal transition and migration of glioma cells via the CXCL12/CXCR4 axis. Oncol Lett 2019;**18**:5930–40.
23. Bolós V, Peinado H, Pérez-Moreno MA, et al. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: A comparison with Snail and E47 repressors. J Cell Sci 2003;**116**:499–511.
24. Cano A, Pérez-Moreno MA, Rodrigo I, et al. The transcription factor Snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nat Cell Biol 2000;**2**: 76–83.
25. Lin Y, Dong C, Zhou BP. Epigenetic regulation of EMT: The Snail story. Curr Pharm Des 2014;**20**:1698–705.
26. Galvagni F, Lentucci C, Neri F, et al. Snai1 promotes ESC exit from the pluripotency by direct repression of self-renewal genes. Stem Cells 2015;**33**:742–50.
27. Murray SA, Gridley T. Snail family genes are required for left-right asymmetry determination, but not neural crest formation, in mice. Proc Natl Acad Sci U S A 2006;**103**:10300–4.
28. Carver EA, Jiang R, Lan Y, et al. The mouse Snail gene encodes a key regulator of the epithelial-mesenchymal transition. Mol Cell Biol 2001;**21**:8184–8.
29. Motta FJN, Valera ET, Lucio-Eterovic AKB, et al. Differential expression of E-cadherin gene in human neuroepithelial tumors. Genet Mol Res 2008;**7**:295–304.
30. Howng S-L, Wu C-H, Cheng T-S, et al. Differential expression of Wnt genes, beta-catenin and E-cadherin in human brain tumors. Cancer Lett 2002;**183**:95–101.
31. Itoh Y, Moriyama Y, Hasegawa T, et al. Scratch regulates neuronal migration onset via an epithelial-mesenchymal transition-like mechanism. Nat Neurosci 2013;**16**:416–25.
32. Ohayon D, Garcès A, Joly W, et al. Onset of spinal cord astrocyte precursor emigration from the ventricular zone involves the Zeb1 transcription factor. Cell Rep 2016;**17**:1473–81.
33. Hirabayashi Y, Suzki N, Tsuboi M, et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. Neuron 2009;**63**:600–13.
34. Vaske OM, Bjork I, Salama SR, et al. Comparative tumor RNA sequencing analysis for difficult-to-treat pediatric and young adult patients with cancer. JAMA Netw Open 2019;**2**:e1913968.
35. Field AR, Jacobs FMJ, Fiddes IT, et al. Structurally conserved primate LncRNAs are transiently expressed during human cortical differentiation and influence cell-type-specific genes. Stem Cell Rep 2019;**12**:245–57.
36. Treehouse Public Data. 2020. https://treehousegenomics.soe.ucsc.edu/public-data/. Accessed 21 April 2020.
37. Mueller S, Jain P, Liang WS, et al. A pilot precision medicine trial for children with diffuse intrinsic pontine glioma - PNOC003: A report from the Pacific Pediatric Neuro-Oncology Consortium. Int J Cancer 2019;**145**(7):1889–901.
38. Grasso CS, Tang Y, Truffaux N, et al. Functionally defined therapeutic targets in diffuse intrinsic pontine glioma. Nat Med 2015;**21**:555–9.
39. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 2016;**164**:550–63.
40. Mackay A, Burford A, Carvalho D, et al. Integrated molecular meta-analysis of 1,000 pediatric high-grade and diffuse intrinsic pontine glioma. Cancer Cell 2017;**32**:520–537.e5.
41. Robinson DR, Wu Y-M, Lonigro RJ, et al. Integrative clinical genomics of metastatic cancer. Nature 2017;**548**:297–303.
42. Sturm D, Orr BA, Toprak UH, et al. New brain tumor entities emerge from molecular classification of CNS-PNETs. Cell 2016;**164**:1060–72.
43. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;**43**:e47.
44. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;**102**:15545–50.
45. Koncar RF, Dey BR, Stanton A-CJ, et al. Identification of novel RAS signaling therapeutic vulnerabilities in diffuse intrinsic pontine gliomas. Cancer Res 2019;**79**:4026–41.

46. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;**1**:417–25.
47. Kerosuo L, Bronner-Fraser M. What is bad in cancer is good in the embryo: Importance of EMT in neural crest development. Semin Cell Dev Biol 2012;**23**:320–32.
48. Chung M-T, Lai H-C, Sytwu H-K, et al. SFRP1 and SFRP2 suppress the transformation and invasion abilities of cervical cancer cells through Wnt signal pathway. Gynecol Oncol 2009;**112**:646–53.
49. Tran DD, Corsa CAS, Biswas H, et al. Temporal and spatial cooperation of Snail1 and Twist1 during epithelial-mesenchymal transition predicts for human breast cancer recurrence. Mol Cancer Res 2011;**9**:1644–57.
50. Stanisavljevic J, Porta-de-la-Riva M, Batlle R, et al. The p65 subunit of NF-$\kappa$B and PARP1 assist Snail1 in activating fibronectin transcription. J Cell Sci 2011;**124**:4161–71.
51. Javaid S, Zhang J, Anderssen E, et al. Dynamic chromatin modification sustains epithelial-mesenchymal transition following inducible expression of Snail-1. Cell Rep 2013;**5**:1679–89.
52. Tanaka S, Kobayashi W, Haraguchi M, et al. Snail1 expression in human colon cancer DLD-1 cells confers invasive properties without N-cadherin expression. Biochem Biophys Rep 2016;**8**:120–6.
53. Lewis PW, Müller MM, Koletsky MS, et al. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. Science 2013;**340**:857–61.
54. Larson JD, Kasper LH, Paugh BS, et al. Histone H3.3 K27M accelerates spontaneous brainstem glioma and drives restricted changes in bivalent gene expression. Cancer Cell 2019;**35**:140–155.e7.
55. Tirosh I, Venteicher AS, Hebert C, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;**539**:309–13.
56. Neftel C, Laffy J, Filbin MG, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell 2019;**178**, doi:10.1016/j.cell.2019.06.024.
57. Tan TZ, Miow QH, Miki Y, et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med 2014;**6**:1279–93.
58. Mak MP, Tong P, Diao L, et al. A patient-derived, pan-cancer EMT signature identifies global molecular alterations and immune target enrichment following epithelial-to-mesenchymal transition. Clin Cancer Res 2016;**22**:609–20.
59. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods 2020;**17**:261–72.
60. Holness CL, Simmons DL. Molecular cloning of CD68, a human macrophage marker related to lysosomal glycoproteins. Blood 1993;**81**:1607–13.
61. Richardson WD, Pringle N, Mosley MJ, et al. A role for platelet-derived growth factor in normal gliogenesis in the central nervous system. Cell 1988;**53**:309–19.
62. Li J, Parker B, Martyn C, et al. The PMP22 gene and its related diseases. Mol Neurobiol 2013;**47**:673–98.
63. Ackerman SD, Garcia C, Piao X, et al. The adhesion GPCR Gpr56 regulates oligodendrocyte development via interactions with G$\alpha$12/13 and RhoA. Nat Commun 2015;**6**: 6122.
64. Falcão AM, Meijer M, Scaglione A, et al. PAD2-Mediated citrullination contributes to efficient oligodendrocyte differentiation and myelination. Cell Rep 2019;**27**:1090–102.e10.
65. Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. J Clin Invest 2009;**119**:1429–37.
66. Sancisi V, Gandolfi G, Ragazzi M, et al. Cadherin 6 is a new RUNX2 target in TGF-$\beta$ signalling pathway. PLoS One 2013;**8**:e75489.
67. Vallath S, Sage EK, Kolluri KK, et al. CADM1 inhibits squamous cell carcinoma progression by reducing STAT3 activity. Sci Rep 2016;**6**:24006.
68. Sakurai-Yageta M, Masuda M, Tsuboi Y, et al. Tumor suppressor CADM1 is involved in epithelial cell structure. Biochem Biophys Res Commun 2009;**390**:977–82.
69. Kim J, Kang HS, Lee Y-J, et al. EGR1-dependent PTEN upregulation by 2-benzoyloxycinnamaldehyde attenuates cell invasion and EMT in colon cancer. Cancer Lett 2014;**349**:35–44.
70. Sun Y, Shen S, Liu X, et al. MiR-429 inhibits cells growth and invasion and regulates EMT-related marker genes by targeting Onecut2 in colorectal carcinoma. Mol Cell Biochem 2014;**390**:19–30.
71. Xu J, Lamouille S, Derynck R. TGF-beta-induced epithelial to mesenchymal transition. Cell Res 2009;**19**:156–72.
72. Lv Q-L, Huang Y-T, Wang G-H, et al. Overexpression of RACK1 promotes metastasis by enhancing epithelial-mesenchymal transition and predicts poor prognosis in human glioma. Int J Environ Res Public Health 2016;**13**, doi:10.3390/ijerph13101021.
73. Castel D, Philippe C, Calmon R, et al. Histone H3F3A and HIST1H3B K27M mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes. Acta Neuropathol 2015;**130**:815–27.
74. Szenker E, Ray-Gallet D, Almouzni G. The double face of the histone variant H3.3. Cell Res 2011;**21**:421–34.
75. Goldberg AD, Banaszynski LA, Noh K-M, et al. Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell 2010;**140**:678–91.
76. Nagaraja S, Quezada MA, Gillespie SM, et al. Histone variant and cell context determine H3K27M reprogramming of the enhancer landscape and oncogenic state. Mol Cell 2019;**76**, doi:10.1016/j.molcel.2019.08.030.
77. Castel D, Philippe C, Kergrohen T, et al. Transcriptomic and epigenetic profiling of "diffuse midline gliomas, H3 K27M-mutant" discriminate two subgroups based on the type of histone H3 mutated and not supratentorial or infratentorial location. Acta Neuropathol Commun 2018;**6**:117.
78. Lin GL, Monje M. A protocol for rapid post-mortem cell culture of diffuse intrinsic pontine glioma (DIPG). J Vis Exp 2017, doi:10.3791/55360.
79. Kim MY, Kaduwal S, Yang DH, et al. Bone morphogenetic protein 4 stimulates attachment of neurospheres and astrogenesis of neural stem cells in neurospheres via phosphatidylinositol 3 kinase-mediated upregulation of N-cadherin. Neuroscience 2010;**170**:8–15.
80. Puget S, Philippe C, Bax DA, et al. Mesenchymal transition and PDGFRA amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. PLoS One 2012;**7**:e30313.
81. Jones C, Karajannis MA, Jones DTW, et al. Pediatric high-grade glioma: Biologically and clinically in need of new thinking. Neuro Oncol 2017;**19**:153–61.
82. Hargrave D, Bartels U, Bouffet E. Diffuse brainstem glioma in children: Critical review of clinical trials. Lancet Oncol 2006;**7**:241–8.
83. Meel MH, Schaper SA, Kaspers GJL, et al. Signaling pathways and mesenchymal transition in pediatric high-grade glioma. Cell Mol Life Sci 2018;**75**:871–87.

84. Tam WL, Weinberg RA. The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med 2013;**19**:1438–49.
85. Christiansen JJ, Rajasekaran AK. Reassessing epithelial to mesenchymal transition as a prerequisite for carcinoma invasion and metastasis. Cancer Res 2006;**66**:8319–26.
86. Grosse-Wilde A, Fouquier d'Hérouël A, McIntosh E, et al. Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival. PLoS One 2015;**10**:e0126522.
87. Jolly MK, Mani SA, Levine H. Hybrid epithelial/mesenchymal phenotype(s): The "fittest" for metastasis? Biochim Biophys Acta Rev Cancer 2018;**1870**:151–7.
88. Pan M-R, Hsu M-C, Chen L-T, et al. Orchestration of H3K27 methylation: Mechanisms and therapeutic implication. Cell Mol Life Sci 2018;**75**:209–23.
89. Gröbner SN, Worst BC, Weischenfeldt J, et al. The landscape of genomic alterations across childhood cancers. Nature 2018;**555**:321–7.
90. Sun L, Fang J. Epigenetic regulation of epithelial-mesenchymal transition. Cell Mol Life Sci 2016;**73**:4493–515.
91. Vivian J, Rao AA, Nothaft FA, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol 2017;**35**:314–6.
92. Fang Z. GSEApy.
93. Merico D, Isserlin R, Stueker O, et al. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. PLoS One 2010;**5**:e13984.
94. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;**352**:189–96.
95. Sanders LM, Cheney A, Seninge L, et al. Supporting data for "Identification of a differentiation stall in epithelial mesenchymal transition in histone H3–mutant diffuse midline glioma." GigaScience Database 2020. http://dx.doi.org/10.5524/100818.

# Appendix

Supplementary table 1

| virus | general type | source | accession |
|---|---|---|---|
| EBV | EBV | RefSeq | NC_007605.1 |
| HTLV type1 | HTLV | GenBank | MH399769.1 |
| HTLV type1 | HTLV | GenBank | J02029.1 |
| HTLV type2 | HTLV | GenBank | Y14365.1 |
| HTLV type3 | HTLV | GenBank | EU649782.1 |
| Hepatitis B | Hepatitis | GenBank | MH818373.1 |
| Hepatitis B | Hepatitis | GenBank | MT114172.1 |
| Hepatitis B | Hepatitis | GenBank | MT437386.1 |
| Hepatitis B | Hepatitis | GenBank | MN683729.1 |
| Hepatitis C | Hepatitis | GenBank | AF165053.1 |
| Hepatitis C | Hepatitis | GenBank | AF207754.1 |
| Hepatitis C | Hepatitis | GenBank | MT212178.1 |
| Hepatitis C | Hepatitis | GenBank | LC368448.1 |
| Hepatitis C | Hepatitis | GenBank | NC_009824.1 |
| Hepatitis C | Hepatitis | GenBank | D84263.2 |
| Hepatitis C | Hepatitis | GenBank | MH155319.1 |
| Hepatitis C | Hepatitis | GenBank | D63822.1 |
| Hepatitis C | Hepatitis | GenBank | MG428679.1 |
| Hepatitis C | Hepatitis | GenBank | MG406988.1 |
| HHV8 | human herpes virus | RefSeq | NC_009333.1 |
| WU | human polyomavirus | RefSeq | NC_009539.1 |
| MW | human polyomavirus | RefSeq | NC_018102.1 |
| BK | human polyomavirus | RefSeq | NC_001538.1 |
| MCV | human polyomavirus | RefSeq | NC_10277.2 |
| HPyV 9 | human polyomavirus | RefSeq | NC_15150.1 |
| HPyV 7 | human polyomavirus | RefSeq | NC_14407.1 |
| HPyV 6 | human polyomavirus | RefSeq | NC_14406.1 |
| HPyV 8 | human polyomavirus | RefSeq | NC_14361.1 |
| JC | human polyomavirus | RefSeq | NC_001699.1 |
| KI | human polyomavirus | RefSeq | NC_009238.1 |

| | | | |
|---|---|---|---|
| HPV1 | HPV | PaVE consortium | V01116 |
| HPV2 | HPV | PaVE consortium | X55964 |
| HPV3 | HPV | PaVE consortium | X74462 |
| HPV4 | HPV | PaVE consortium | X70827 |
| HPV5 | HPV | PaVE consortium | M17463 |
| HPV6 | HPV | PaVE consortium | X00203 |
| HPV7 | HPV | PaVE consortium | X74463 |
| HPV8 | HPV | PaVE consortium | M12737 |
| HPV9 | HPV | PaVE consortium | X74464 |
| HPV10 | HPV | PaVE consortium | X74465 |
| HPV11 | HPV | PaVE consortium | M14119 |
| HPV12 | HPV | PaVE consortium | X74466 |
| HPV13 | HPV | PaVE consortium | X62843 |
| HPV14 | HPV | PaVE consortium | X74467 |
| HPV15 | HPV | PaVE consortium | X74468 |
| HPV16 | HPV | PaVE consortium | K02718 |
| HPV17 | HPV | PaVE consortium | X74469 |
| HPV18 | HPV | PaVE consortium | X05015 |
| HPV19 | HPV | PaVE consortium | X74470 |
| HPV20 | HPV | PaVE consortium | U31778 |
| HPV21 | HPV | PaVE consortium | U31779 |
| HPV22 | HPV | PaVE consortium | U31780 |
| HPV23 | HPV | PaVE consortium | U31781 |
| HPV24 | HPV | PaVE consortium | U31782 |
| HPV25 | HPV | PaVE consortium | X74471 |
| HPV26 | HPV | PaVE consortium | X74472 |
| HPV27 | HPV | PaVE consortium | X74473 |
| HPV28 | HPV | PaVE consortium | U31783 |
| HPV29 | HPV | PaVE consortium | U31784 |
| HPV30 | HPV | PaVE consortium | X74474 |
| HPV31 | HPV | PaVE consortium | J04353 |
| HPV32 | HPV | PaVE consortium | X74475 |
| HPV33 | HPV | PaVE consortium | M12732 |
| HPV34 | HPV | PaVE consortium | X74476 |
| HPV35 | HPV | PaVE consortium | X74477 |
| HPV36 | HPV | PaVE consortium | U31785 |
| HPV37 | HPV | PaVE consortium | U31786 |
| HPV38 | HPV | PaVE consortium | U31787 |
| HPV39 | HPV | PaVE consortium | M62849 |
| HPV40 | HPV | PaVE consortium | X74478 |
| HPV41 | HPV | PaVE consortium | X56147 |
| HPV42 | HPV | PaVE consortium | M73236 |
| HPV43 | HPV | PaVE consortium | AJ620205 |

| | | | |
|---|---|---|---|
| HPV44 | HPV | PaVE consortium | U31788 |
| HPV45 | HPV | PaVE consortium | X74479 |
| HPV47 | HPV | PaVE consortium | M32305 |
| HPV48 | HPV | PaVE consortium | U31789 |
| HPV49 | HPV | PaVE consortium | X74480 |
| HPV50 | HPV | PaVE consortium | U31790 |
| HPV51 | HPV | PaVE consortium | M62877 |
| HPV52 | HPV | PaVE consortium | X74481 |
| HPV53 | HPV | PaVE consortium | X74482 |
| HPV54 | HPV | PaVE consortium | U37488 |
| HPV56 | HPV | PaVE consortium | X74483 |
| HPV57 | HPV | PaVE consortium | X55965 |
| HPV58 | HPV | PaVE consortium | D90400 |
| HPV59 | HPV | PaVE consortium | X77858 |
| HPV60 | HPV | PaVE consortium | U31792 |
| HPV61 | HPV | PaVE consortium | U31793 |
| HPV62 | HPV | PaVE consortium | AY395706 |
| HPV63 | HPV | PaVE consortium | X70828 |
| HPV65 | HPV | PaVE consortium | X70829 |
| HPV66 | HPV | PaVE consortium | U31794 |
| HPV67 | HPV | PaVE consortium | D21208 |
| HPV68 | HPV | PaVE consortium | DQ080079 |
| HPV69 | HPV | PaVE consortium | AB027020 |
| HPV70 | HPV | PaVE consortium | U21941 |
| HPV71 | HPV | PaVE consortium | AB040456 |
| HPV72 | HPV | PaVE consortium | X94164 |
| HPV73 | HPV | PaVE consortium | X94165 |
| HPV74 | HPV | PaVE consortium | AF436130 |
| HPV75 | HPV | PaVE consortium | Y15173 |
| HPV76 | HPV | PaVE consortium | Y15174 |
| HPV77 | HPV | PaVE consortium | Y15175 |
| HPV78 | HPV | PaVE consortium | KC138720 |
| HPV80 | HPV | PaVE consortium | Y15176 |
| HPV81 | HPV | PaVE consortium | AJ620209 |
| HPV82 | HPV | PaVE consortium | AB027021 |
| HPV83 | HPV | PaVE consortium | AF151983 |
| HPV84 | HPV | PaVE consortium | AF293960 |
| HPV85 | HPV | PaVE consortium | AF131950 |
| HPV86 | HPV | PaVE consortium | AF349909 |
| HPV87 | HPV | PaVE consortium | AJ400628 |

| | | | |
|---|---|---|---|
| HPV88 | HPV | PaVE consortium | EF467176 |
| HPV89 | HPV | PaVE consortium | AF436128 |
| HPV90 | HPV | PaVE consortium | AY057438 |
| HPV91 | HPV | PaVE consortium | AF419318 |
| HPV92 | HPV | PaVE consortium | AF531420 |
| HPV93 | HPV | PaVE consortium | AY382778 |
| HPV94 | HPV | PaVE consortium | AJ620211 |
| HPV95 | HPV | PaVE consortium | AJ620210 |
| HPV96 | HPV | PaVE consortium | AY382779 |
| HPV97 | HPV | PaVE consortium | DQ080080 |
| HPV98 | HPV | PaVE consortium | FM955837 |
| HPV99 | HPV | PaVE consortium | FM955838 |
| HPV100 | HPV | PaVE consortium | FM955839 |
| HPV101 | HPV | PaVE consortium | DQ080081 |
| HPV102 | HPV | PaVE consortium | DQ080083 |
| HPV103 | HPV | PaVE consortium | DQ080078 |
| HPV104 | HPV | PaVE consortium | FM955840 |
| HPV105 | HPV | PaVE consortium | FM955841 |
| HPV106 | HPV | PaVE consortium | DQ080082 |
| HPV107 | HPV | PaVE consortium | EF422221 |
| HPV108 | HPV | PaVE consortium | FM212639 |
| HPV109 | HPV | PaVE consortium | EU541441 |
| HPV110 | HPV | PaVE consortium | EU410348 |
| HPV111 | HPV | PaVE consortium | EU410349 |
| HPV112 | HPV | PaVE consortium | EU541442 |
| HPV113 | HPV | PaVE consortium | FM955842 |
| HPV114 | HPV | PaVE consortium | GQ244463 |
| HPV115 | HPV | PaVE consortium | FJ947080 |
| HPV116 | HPV | PaVE consortium | FJ804072 |
| HPV117 | HPV | PaVE consortium | GQ246950 |
| HPV118 | HPV | PaVE consortium | GQ246951 |
| HPV119 | HPV | PaVE consortium | GQ845441 |
| HPV120 | HPV | PaVE consortium | GQ845442 |
| HPV121 | HPV | PaVE consortium | GQ845443 |
| HPV122 | HPV | PaVE consortium | GQ845444 |
| HPV123 | HPV | PaVE consortium | GQ845445 |
| HPV124 | HPV | PaVE consortium | GQ845446 |
| HPV125 | HPV | PaVE consortium | FN547152 |
| HPV126 | HPV | PaVE consortium | AB646346 |
| HPV127 | HPV | PaVE consortium | HM011570 |

| | | | |
|---|---|---|---|
| HPV128 | HPV | PaVE consortium | GU225708 |
| HPV129 | HPV | PaVE consortium | GU233853 |
| HPV130 | HPV | PaVE consortium | GU117630 |
| HPV131 | HPV | PaVE consortium | GU117631 |
| HPV132 | HPV | PaVE consortium | GU117632 |
| HPV133 | HPV | PaVE consortium | GU117633 |
| HPV134 | HPV | PaVE consortium | GU117634 |
| HPV135 | HPV | PaVE consortium | HM999987 |
| HPV136 | HPV | PaVE consortium | HM999988 |
| HPV137 | HPV | PaVE consortium | HM999989 |
| HPV138 | HPV | PaVE consortium | HM999990 |
| HPV139 | HPV | PaVE consortium | HM999991 |
| HPV140 | HPV | PaVE consortium | HM999992 |
| HPV141 | HPV | PaVE consortium | HM999993 |
| HPV142 | HPV | PaVE consortium | HM999994 |
| HPV143 | HPV | PaVE consortium | HM999995 |
| HPV144 | HPV | PaVE consortium | HM999996 |
| HPV145 | HPV | PaVE consortium | HM999997 |
| HPV146 | HPV | PaVE consortium | HM999998 |
| HPV147 | HPV | PaVE consortium | HM999999 |
| HPV148 | HPV | PaVE consortium | GU129016 |
| HPV149 | HPV | PaVE consortium | GU117629 |
| HPV150 | HPV | PaVE consortium | FN677755 |
| HPV151 | HPV | PaVE consortium | FN677756 |
| HPV152 | HPV | PaVE consortium | JF304768 |
| HPV153 | HPV | PaVE consortium | JN171845 |
| HPV154 | HPV | PaVE consortium | JN211193 |
| HPV155 | HPV | PaVE consortium | JF906559 |
| HPV156 | HPV | PaVE consortium | JX429973 |
| HPV157 | HPV | PaVE consortium | KT698166 |
| HPV158 | HPV | PaVE consortium | KT698168 |
| HPV159 | HPV | PaVE consortium | HE963025 |
| HPV160 | HPV | PaVE consortium | AB745694 |
| HPV161 | HPV | PaVE consortium | JX413109 |
| HPV162 | HPV | PaVE consortium | JX413108 |
| HPV163 | HPV | PaVE consortium | JX413107 |
| HPV164 | HPV | PaVE consortium | JX413106 |
| HPV165 | HPV | PaVE consortium | JX444072 |
| HPV166 | HPV | PaVE consortium | JX413104 |
| HPV167 | HPV | PaVE consortium | KC862318 |

| | | | |
|---|---|---|---|
| HPV168 | HPV | PaVE consortium | KC862317 |
| HPV169 | HPV | PaVE consortium | JX413105 |
| HPV170 | HPV | PaVE consortium | JX413110 |
| HPV171 | HPV | PaVE consortium | KF006398 |
| HPV172 | HPV | PaVE consortium | KF006399 |
| HPV173 | HPV | PaVE consortium | KF006400 |
| HPV174 | HPV | PaVE consortium | HF930491 |
| HPV175 | HPV | PaVE consortium | KC108721 |
| HPV176 | HPV | PaVE consortium | KR816167 |
| HPV177 | HPV | PaVE consortium | KR816168 |
| HPV178 | HPV | PaVE consortium | KJ130020 |
| HPV179 | HPV | PaVE consortium | HG421739 |
| HPV180 | HPV | PaVE consortium | KC108722 |
| HPV181 | HPV | PaVE consortium | KR816169 |
| HPV182 | HPV | PaVE consortium | KR816170 |
| HPV183 | HPV | PaVE consortium | KR816171 |
| HPV184 | HPV | PaVE consortium | HG530535 |
| HPV185 | HPV | PaVE consortium | KR816172 |
| HPV186 | HPV | PaVE consortium | KR816173 |
| HPV187 | HPV | PaVE consortium | KR816174 |
| HPV188 | HPV | PaVE consortium | KR816175 |
| HPV189 | HPV | PaVE consortium | KR816176 |
| HPV190 | HPV | PaVE consortium | KR816177 |
| HPV191 | HPV | PaVE consortium | KR816178 |
| HPV192 | HPV | PaVE consortium | KR816179 |
| HPV193 | HPV | PaVE consortium | KR816180 |
| HPV194 | HPV | PaVE consortium | KR816181 |
| HPV195 | HPV | PaVE consortium | KR816182 |
| HPV196 | HPV | PaVE consortium | KR816183 |
| HPV197 | HPV | PaVE consortium | KM085343 |
| HPV199 | HPV | PaVE consortium | KJ913662 |
| HPV200 | HPV | PaVE consortium | KP692114 |
| HPV201 | HPV | PaVE consortium | KP692115 |
| HPV202 | HPV | PaVE consortium | KP692116 |
| HPV203 | HPV | PaVE consortium | MG921180 |
| HPV204 | HPV | PaVE consortium | KP769769 |
| HPV205 | HPV | PaVE consortium | KT698167 |
| HPV206 | HPV | PaVE consortium | U85660 |
| HPV207 | HPV | PaVE consortium | MK645900 |
| HPV208 | HPV | PaVE consortium | MK645901 |

| | | | |
|---|---|---|---|
| HPV209 | HPV | PaVE consortium | KY242583 |
| HPV210 | HPV | PaVE consortium | MH460956 |
| HPV211 | HPV | PaVE consortium | MF509816 |
| HPV212 | HPV | PaVE consortium | MF509817 |
| HPV213 | HPV | PaVE consortium | MF509818 |
| HPV214 | HPV | PaVE consortium | MF509819 |
| HPV215 | HPV | PaVE consortium | MF509820 |
| HPV216 | HPV | PaVE consortium | MF509821 |
| HPV219 | HPV | PaVE consortium | MH172376 |
| HPV220 | HPV | PaVE consortium | MH172377 |
| HPV221 | HPV | PaVE consortium | MH172378 |
| HPV222 | HPV | PaVE consortium | MH172379 |
| HPV223 | HPV | PaVE consortium | MG063749 |
| HPV224 | HPV | PaVE consortium | MF356498 |
| HPV225 | HPV | PaVE consortium | MG520499 |
| HPV226 | HPV | PaVE consortium | MG813996 |
| HPV227 | HPV | PaVE consortium | MK080568 |
| HPV228 | HPV | PaVE consortium | ON482334 |
| HPV229 | HPV | PaVE consortium | MW535770 |
| HPV-mCG2 | HPV | PaVE consortium | JF966378 |
| HPV-mCG3 | HPV | PaVE consortium | JF966379 |
| HPV-mCGG5 | HPV | PaVE consortium | MG869605 |
| HPV-mCH2 | HPV | PaVE consortium | KF791917 |
| HPV-mDysk1 | HPV | PaVE consortium | KX781280 |
| HPV-mDysk2 | HPV | PaVE consortium | KX781281 |
| HPV-mDysk3 | HPV | PaVE consortium | KX781282 |
| HPV-mDysk5 | HPV | PaVE consortium | KX781284 |
| HPV-mDysk6 | HPV | PaVE consortium | KX781285 |
| HPV-mEV03c05 | HPV | PaVE consortium | MF588698 |
| HPV-mEV03c09 | HPV | PaVE consortium | MF588687 |
| HPV-mEV03c104 | HPV | PaVE consortium | MF588724 |
| HPV-mEV03c188 | HPV | PaVE consortium | MF588729 |
| HPV-mEV03c212 | HPV | PaVE consortium | MF588735 |
| HPV-mEV03c40 | HPV | PaVE consortium | MF588746 |
| HPV-mEV03c434 | HPV | PaVE consortium | MF588755 |
| HPV-mEV03c45 | HPV | PaVE consortium | MF588721 |
| HPV-mEV06c107 | HPV | PaVE consortium | MF588747 |
| HPV-mEV06c118 | HPV | PaVE consortium | MF588736 |
| HPV-mEV06c12b | HPV | PaVE consortium | MF588676 |
| HPV-mEV07c367 | HPV | PaVE consortium | MF588716 |

| | | | |
|---|---|---|---|
| HPV-mEV07c382 | HPV | PaVE consortium | MF588723 |
| HPV-mEV07c390 | HPV | PaVE consortium | MF588705 |
| HPV-mFD1 | HPV | PaVE consortium | JF966375 |
| HPV-mFD2 | HPV | PaVE consortium | JF966376 |
| HPV-mFS1 | HPV | PaVE consortium | JF966373 |
| HPV-mFi864 | HPV | PaVE consortium | KC311731 |
| HPV-mHIVGc36 | HPV | PaVE consortium | MF588677 |
| HPV-mJDFY01 | HPV | PaVE consortium | MW311489 |
| HPV-mKC5 | HPV | PaVE consortium | JX444073 |
| HPV-mKN1 | HPV | PaVE consortium | JF966371 |
| HPV-mKN2 | HPV | PaVE consortium | JF966372 |
| HPV-mKN3 | HPV | PaVE consortium | JF966374 |
| HPV-mL55 | HPV | PaVE consortium | KF482069 |
| HPV-mLCOSOc196 | HPV | PaVE consortium | MF588688 |
| HPV-mSD2 | HPV | PaVE consortium | KC113191 |
| HPV-mSE379 | HPV | PaVE consortium | KP692117 |
| HPV-mSE383 | HPV | PaVE consortium | KP692118 |
| HPV-mSK001 | HPV | PaVE consortium | MH675888 |
| HPV-mSK008 | HPV | PaVE consortium | MH777156 |
| HPV-mSK010 | HPV | PaVE consortium | MH777158 |
| HPV-mSK011 | HPV | PaVE consortium | MH777159 |
| HPV-mSK012 | HPV | PaVE consortium | MH777160 |
| HPV-mSK013 | HPV | PaVE consortium | MH777161 |
| HPV-mSK014 | HPV | PaVE consortium | MH777162 |
| HPV-mSK015 | HPV | PaVE consortium | MH777163 |
| HPV-mSK019 | HPV | PaVE consortium | MH777167 |
| HPV-mSK020 | HPV | PaVE consortium | MH777168 |
| HPV-mSK022 | HPV | PaVE consortium | MH777170 |
| HPV-mSK023 | HPV | PaVE consortium | MH777171 |
| HPV-mSK025 | HPV | PaVE consortium | MH777173 |
| HPV-mSK027 | HPV | PaVE consortium | MH777175 |
| HPV-mSK028 | HPV | PaVE consortium | MH777176 |
| HPV-mSK029 | HPV | PaVE consortium | MH777177 |
| HPV-mSK030 | HPV | PaVE consortium | MH777178 |
| HPV-mSK034 | HPV | PaVE consortium | MH777182 |
| HPV-mSK037 | HPV | PaVE consortium | MH777185 |
| HPV-mSK038 | HPV | PaVE consortium | MH777186 |
| HPV-mSK039 | HPV | PaVE consortium | MH777187 |
| HPV-mSK040 | HPV | PaVE consortium | MH972557 |
| HPV-mSK041 | HPV | PaVE consortium | MH777188 |

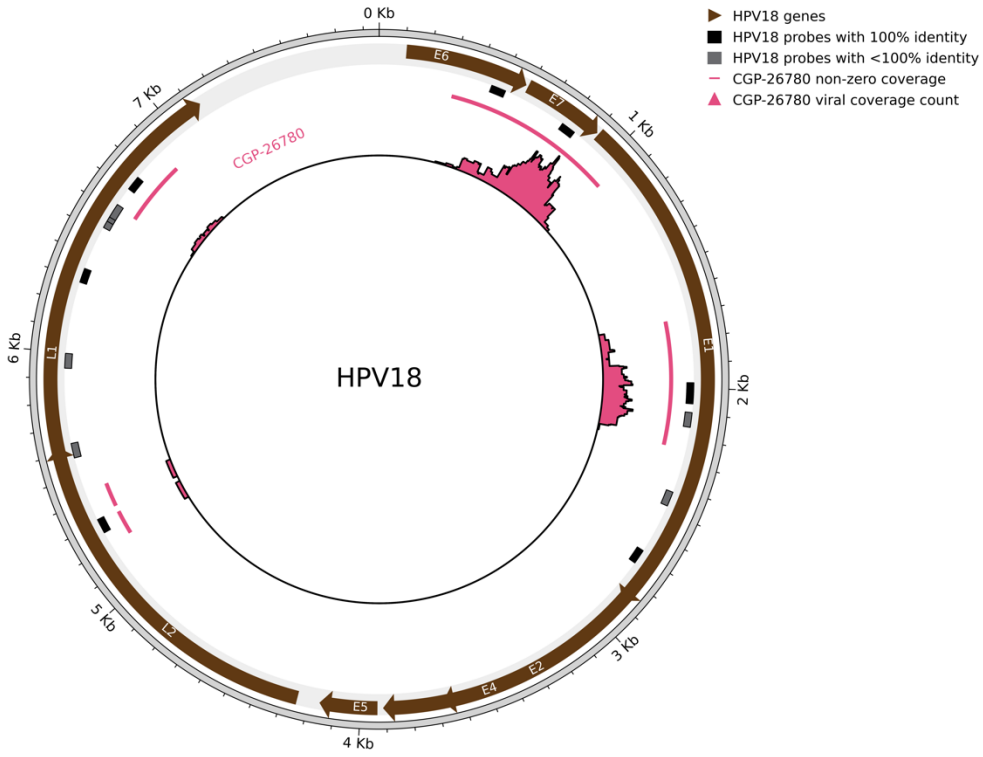
| | | | |
|---|---|---|---|
| HPV-mSK043 | HPV | PaVE consortium | MH777190 |
| HPV-mSK045 | HPV | PaVE consortium | MH972558 |
| HPV-mSK046 | HPV | PaVE consortium | MH777192 |
| HPV-mSK049 | HPV | PaVE consortium | MH777194 |
| HPV-mSK050 | HPV | PaVE consortium | MH777195 |
| HPV-mSK051 | HPV | PaVE consortium | MH777196 |
| HPV-mSK054 | HPV | PaVE consortium | MH777199 |
| HPV-mSK055 | HPV | PaVE consortium | MH777200 |
| HPV-mSK056 | HPV | PaVE consortium | MH777201 |
| HPV-mSK057 | HPV | PaVE consortium | MH777202 |
| HPV-mSK058 | HPV | PaVE consortium | MH777203 |
| HPV-mSK061 | HPV | PaVE consortium | MH777206 |
| HPV-mSK062 | HPV | PaVE consortium | MH777207 |
| HPV-mSK063 | HPV | PaVE consortium | MH777208 |
| HPV-mSK064 | HPV | PaVE consortium | MH777209 |
| HPV-mSK065 | HPV | PaVE consortium | MH777210 |
| HPV-mSK067 | HPV | PaVE consortium | MH777212 |
| HPV-mSK075 | HPV | PaVE consortium | MH777219 |
| HPV-mSK076 | HPV | PaVE consortium | MH777220 |
| HPV-mSK078 | HPV | PaVE consortium | MH777222 |
| HPV-mSK081 | HPV | PaVE consortium | MH972561 |
| HPV-mSK083 | HPV | PaVE consortium | MH777226 |
| HPV-mSK085 | HPV | PaVE consortium | MH777228 |
| HPV-mSK086 | HPV | PaVE consortium | MH777229 |
| HPV-mSK087 | HPV | PaVE consortium | MH777230 |
| HPV-mSK089 | HPV | PaVE consortium | MH777232 |
| HPV-mSK090 | HPV | PaVE consortium | MH777233 |
| HPV-mSK091 | HPV | PaVE consortium | MH777234 |
| HPV-mSK093 | HPV | PaVE consortium | MH777236 |
| HPV-mSK094 | HPV | PaVE consortium | MH777237 |
| HPV-mSK095 | HPV | PaVE consortium | MH777238 |
| HPV-mSK096 | HPV | PaVE consortium | MH777239 |
| HPV-mSK098 | HPV | PaVE consortium | MH777241 |
| HPV-mSK100 | HPV | PaVE consortium | MH777243 |
| HPV-mSK101 | HPV | PaVE consortium | MH972562 |
| HPV-mSK102 | HPV | PaVE consortium | MH777244 |
| HPV-mSK103 | HPV | PaVE consortium | MH777245 |
| HPV-mSK107 | HPV | PaVE consortium | MH777249 |
| HPV-mSK109 | HPV | PaVE consortium | MH777251 |
| HPV-mSK111 | HPV | PaVE consortium | MH777253 |

| | | | |
|---|---|---|---|
| HPV-mSK115 | HPV | PaVE consortium | MH777257 |
| HPV-mSK116 | HPV | PaVE consortium | MH777258 |
| HPV-mSK117 | HPV | PaVE consortium | MH777259 |
| HPV-mSK118 | HPV | PaVE consortium | MH777260 |
| HPV-mSK119 | HPV | PaVE consortium | MH777261 |
| HPV-mSK120 | HPV | PaVE consortium | MH777262 |
| HPV-mSK121 | HPV | PaVE consortium | MH777263 |
| HPV-mSK126 | HPV | PaVE consortium | MH777268 |
| HPV-mSK128 | HPV | PaVE consortium | MH777270 |
| HPV-mSK131 | HPV | PaVE consortium | MH777273 |
| HPV-mSK132 | HPV | PaVE consortium | MH777274 |
| HPV-mSK138 | HPV | PaVE consortium | MH777280 |
| HPV-mSK139 | HPV | PaVE consortium | MH777281 |
| HPV-mSK146 | HPV | PaVE consortium | MH777288 |
| HPV-mSK147 | HPV | PaVE consortium | MH777289 |
| HPV-mSK150 | HPV | PaVE consortium | MH777292 |
| HPV-mSK152 | HPV | PaVE consortium | MH777294 |
| HPV-mSK155 | HPV | PaVE consortium | MH777297 |
| HPV-mSK156 | HPV | PaVE consortium | MH777298 |
| HPV-mSK160 | HPV | PaVE consortium | MH777302 |
| HPV-mSK164 | HPV | PaVE consortium | MH777306 |
| HPV-mSK165 | HPV | PaVE consortium | MH777307 |
| HPV-mSK166 | HPV | PaVE consortium | MH777308 |
| HPV-mSK167 | HPV | PaVE consortium | MH777309 |
| HPV-mSK168 | HPV | PaVE consortium | MH777310 |
| HPV-mSK169 | HPV | PaVE consortium | MH777311 |
| HPV-mSK170 | HPV | PaVE consortium | MH777312 |
| HPV-mSK171 | HPV | PaVE consortium | MH777313 |
| HPV-mSK173 | HPV | PaVE consortium | MH777315 |
| HPV-mSK174 | HPV | PaVE consortium | MH777316 |
| HPV-mSK176 | HPV | PaVE consortium | MH777318 |
| HPV-mSK183 | HPV | PaVE consortium | MH777325 |
| HPV-mSK185 | HPV | PaVE consortium | MH777327 |
| HPV-mSK186 | HPV | PaVE consortium | MH777328 |
| HPV-mSK189 | HPV | PaVE consortium | MH777331 |
| HPV-mSK190 | HPV | PaVE consortium | MH777332 |
| HPV-mSK191 | HPV | PaVE consortium | MH777333 |
| HPV-mSK192 | HPV | PaVE consortium | MH777334 |
| HPV-mSK197 | HPV | PaVE consortium | MH777339 |
| HPV-mSK199 | HPV | PaVE consortium | MH972563 |

| | | | |
|---|---|---|---|
| HPV-mSK202 | HPV | PaVE consortium | MH777343 |
| HPV-mSK204 | HPV | PaVE consortium | MH777345 |
| HPV-mSK205 | HPV | PaVE consortium | MH777346 |
| HPV-mSK206 | HPV | PaVE consortium | MH777347 |
| HPV-mSK209 | HPV | PaVE consortium | MH777350 |
| HPV-mSK210 | HPV | PaVE consortium | MH777351 |
| HPV-mSK211 | HPV | PaVE consortium | MH777352 |
| HPV-mSK216 | HPV | PaVE consortium | MH777355 |
| HPV-mSK217 | HPV | PaVE consortium | MH777356 |
| HPV-mSK218 | HPV | PaVE consortium | MH777357 |
| HPV-mSK219 | HPV | PaVE consortium | MH777358 |
| HPV-mSK227 | HPV | PaVE consortium | MH777366 |
| HPV-mSK229 | HPV | PaVE consortium | MH777368 |
| HPV-mSK231 | HPV | PaVE consortium | MH777370 |
| HPV-mSK238 | HPV | PaVE consortium | MH777377 |
| HPV-mSK239 | HPV | PaVE consortium | MH777378 |
| HPV-mSK240 | HPV | PaVE consortium | MH777379 |
| HPV-mSK241 | HPV | PaVE consortium | MH777380 |
| HPV-mSK242 | HPV | PaVE consortium | MH777381 |
| HPV-mSK244 | HPV | PaVE consortium | MH777383 |
| HPV-mSK245 | HPV | PaVE consortium | MH777384 |
| HPV-mSK246 | HPV | PaVE consortium | MH777385 |
| HPV-mSK249 | HPV | PaVE consortium | MH777388 |
| HPV-mSK250 | HPV | PaVE consortium | MH777389 |
| HPV-mTG550 | HPV | PaVE consortium | MF176072 |
| HPV-mTVMBSFc09 | HPV | PaVE consortium | MF588684 |
| HPV-mTVMBSGc529 | HPV | PaVE consortium | MF588706 |
| HPV-mTVMBSGc2024 | HPV | PaVE consortium | MF588686 |
| HPV-mTVMBSGc2450 | HPV | PaVE consortium | MF588732 |
| HPV-mTVMBSHc13 | HPV | PaVE consortium | MF588738 |
| HPV-mTVMBSHc33 | HPV | PaVE consortium | MF588707 |
| HPV-mTVMBSWc141 | HPV | PaVE consortium | MF588739 |
| HPV-mZJ01 | HPV | PaVE consortium | KX082661 |
| HPV-md01c06 | HPV | PaVE consortium | MF588757 |
| HPV-mdo1c02 | HPV | PaVE consortium | MF588734 |
| HPV-mdo1c232 | HPV | PaVE consortium | MF588704 |
| HPV-mga2c01 | HPV | PaVE consortium | MF588744 |
| HPV-mga2c70 | HPV | PaVE consortium | MF588754 |
| HPV-mm090c09 | HPV | PaVE consortium | MF588681 |
| HPV-mm090c10 | HPV | PaVE consortium | MF588745 |

| | | | |
|---|---|---|---|
| HPV-mm090c145 | HPV | PaVE consortium | MF588737 |
| HPV-mm090c66 | HPV | PaVE consortium | MF588748 |
| HPV-mm292c100 | HPV | PaVE consortium | MF588685 |
| HPV-mm292c14 | HPV | PaVE consortium | MF588682 |
| HPV-mm292c88 | HPV | PaVE consortium | MF588683 |
| HPV-mw02c24a | HPV | PaVE consortium | MF588696 |
| HPV-mw07c34d | HPV | PaVE consortium | MF588753 |
| HPV-mw07c68b | HPV | PaVE consortium | MF588717 |
| HPV-mw11C24 | HPV | PaVE consortium | MF588709 |
| HPV-mw11C39 | HPV | PaVE consortium | MF588710 |
| HPV-mw11C51 | HPV | PaVE consortium | MF588689 |
| HPV-mw11c13 | HPV | PaVE consortium | MF588718 |
| HPV-mw18c07 | HPV | PaVE consortium | MF588740 |
| HPV-mw18c11d | HPV | PaVE consortium | MF588733 |
| HPV-mw18c134 | HPV | PaVE consortium | MF588725 |
| HPV-mw18c25 | HPV | PaVE consortium | MF588714 |
| HPV-mw18c39 | HPV | PaVE consortium | MF588741 |
| HPV-mw20c01a | HPV | PaVE consortium | MF588701 |
| HPV-mw20c01b | HPV | PaVE consortium | MF588708 |
| HPV-mw20c02c | HPV | PaVE consortium | MF588702 |
| HPV-mw20c03a | HPV | PaVE consortium | MF588715 |
| HPV-mw20c04 | HPV | PaVE consortium | MF588695 |
| HPV-mw20c08a | HPV | PaVE consortium | MF588742 |
| HPV-mw20c09 | HPV | PaVE consortium | MF588703 |
| HPV-mw20c10a | HPV | PaVE consortium | MF588726 |
| HPV-mw21c693 | HPV | PaVE consortium | MF588743 |
| HPV-mw22c09 | HPV | PaVE consortium | MF588750 |
| HPV-mw23c08c | HPV | PaVE consortium | MF588690 |
| HPV-mw23c101c | HPV | PaVE consortium | MF588711 |
| HPV-mw23c77 | HPV | PaVE consortium | MF588719 |
| HPV-mw27c03a | HPV | PaVE consortium | MF588758 |
| HPV-mw27c04c | HPV | PaVE consortium | MF588728 |
| HPV-mw27c157c | HPV | PaVE consortium | MF588692 |
| HPV-mw27c39c | HPV | PaVE consortium | MF588712 |
| HPV-mw34c04a | HPV | PaVE consortium | MF588693 |
| HPV-mw34c11a | HPV | PaVE consortium | MF588720 |
| HPV-mw34c14a | HPV | PaVE consortium | MF588694 |
| HPV-mw34c28a | HPV | PaVE consortium | MF588751 |
| HPV-mw34c34a | HPV | PaVE consortium | MF588731 |
| HPV-mwg1c05 | HPV | PaVE consortium | MF588756 |
| HPV-mwg1c09 | HPV | PaVE consortium | MF588727 |

Supplementary figure 1: HPV18 viral coverage map in one sample.



# Bibliography


1. Hulvat, M.C. (2020). Cancer Incidence and Trends. Surg. Clin. *100*, 469–481. 10.1016/j.suc.2020.01.002.

2. Tomasetti, C., and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science *347*, 78–81. 10.1126/science.1260825.

3. Danaei, G., Hoorn, S.V., Lopez, A.D., Murray, C.J., and Ezzati, M. (2005). Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet *366*, 1784–1793. 10.1016/S0140-6736(05)67725-2.

4. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483–1489. 10.1126/science.aab4082.

5. Moodie, R.L. (1927). Tumors in the Lower Carboniferous. Science *66*, 540–540.

6. Capasso, L.L. (2005). Antiquity of cancer. Int. J. Cancer *113*, 2–13. 10.1002/ijc.20610.

7. Faguet, G.B. (2015). A brief history of cancer: Age-old milestones underlying our current knowledge database. Int. J. Cancer *136*, 2022–2036. 10.1002/ijc.29134.

8. Rothschild, B.M., Tanke, D.H., Helbling, M., and Martin, L.D. (2003). Epidemiologic study of tumors in dinosaurs. Naturwissenschaften *90*, 495–500. 10.1007/s00114-003-0473-9.

9. Weinstein, I.B., and Case, K. (2008). The History of Cancer Research: Introducing an AACR Centennial Series. Cancer Res. *68*, 6861–6862. 10.1158/0008-5472.CAN-08-2827.

10. Redmond, D.E. (1970). Tobacco and Cancer: The First Clinical Report, 1761. N. Engl. J. Med. *282*, 18–23. 10.1056/NEJM197001012820105.

11. Brown, J.R., and Thornton, J.L. (1957). Percivall Pott (1714-1788) and Chimney Sweepers' Cancer of the Scrotum. Br. J. Ind. Med. *14*, 68–70.

12. Mustacchi, P., and Shimkin, M.B. (1956). Radiation cancer and Jean Clunet. Cancer *9*, 1073–1074. 10.1002/1097-0142(195611/12)9:6<1073::AID-CNCR2820090602>3.0.CO;2-9.

13. Yamagiwa, K., and Ichikawa, K. (1918). Experimental Study of the Pathogenesis of Carcinoma1. J. Cancer Res. *3*, 1–29. 10.1158/jcr.1918.1.

14. Balmain, A. (2001). Cancer genetics: from Boveri and Mendel to microarrays. Nat. Rev. Cancer *1*, 77–82. 10.1038/35094086.

15. Weiss, R.A., and Vogt, P.K. (2011). 100 years of Rous sarcoma virus. J. Exp. Med. *208*, 2351–2355. 10.1084/jem.20112160.

16. Vogt, P.K. (1996). MILESTONES IN BIOLOGICAL RESEARCH Peyton Rous: Homage and Appraisal. FASEB J. *10*, 1559–1562. 10.1096/fasebj.10.13.8940303.

17. Rous, P. (1911). A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. J. Exp. Med. *13*, 397–411.

18. Martin, G.S. (2001). The hunting of the Src. Nat. Rev. Mol. Cell Biol. *2*, 467–475. 10.1038/35073094.

19. Shope, R.E. (1936). A CHANGE IN RABBIT FIBROMA VIRUS SUGGESTING MUTATION: II. BEHAVIOR OF THE VARIANT VIRUS IN COTTONTAIL RABBITS. J. Exp. Med. *63*, 173–178. 10.1084/jem.63.2.173.

20. Bashford, E.F. (1913). Fresh Light on the Cause of Cancer. Nature *90*, 701–702. 10.1038/090701a0.

21. Stolt, C.-M., Klein, G., and Jansson, A.T.R. (2004). An Analysis of a Wrong Nobel Prize—Johannes Fibiger, 1926: A Study in the Nobel Archives. In Advances in Cancer Research (Academic Press), pp. 1–12. 10.1016/S0065-230X(04)92001-5.

22. Wade, N. (1971). Special virus cancer program: travails of a biological moonshot. Science *174*, 1306–1311. 10.1126/science.174.4016.1306.

23. Varmus, H.E. (1990). Retroviruses and Oncogenes I (Nobel Lecture). Angew. Chem. Int. Ed. Engl. *29*, 707–715. 10.1002/anie.199007073.

24. Spiegelman, S. (1976). Molecular Evidence for the Association of RNA Tumor Viruses with Human Mesenchymal Malignancies. In Modern Trends in Human Leukemia II Hämatologie und Bluttransfusion., R. Neth, R. C. Gallo, K. Mannweiler, and W. C. Moloney, eds. (Springer), pp. 391–429. 10.1007/978-3-642-87524-3_38.

25. Spiegelman, S. (1976). The Search for Viruses in Human Cancer. Proc. Am. Philos. Soc. *120*, 69–86.

26. Spiegelman, S., Baxt, W., Kufe, D., Peters, W.P., and Schlom, J. (1975). Sequences related to the RNA tumor viruses in the RNA and DNA of human leukemias and lymphomas. Bibl. Haematol., 3–25. 10.1159/000397514.

27. Spiegelman, S., Axel, R., and Schlom, J. (1972). Virus-Related RNA in Human and Mouse Mammary Tumors12. JNCI J. Natl. Cancer Inst. *48*, 1205–1211. 10.1093/jnci/48.4.1205.

28. Hehlmann, R., Kufe, D., and Spiegelman, S. (1972). Viral-Related RNA in Hodgkins' Disease and Other Human Lymphomas. Proc. Natl. Acad. Sci. *69*, 1727–1731. 10.1073/pnas.69.7.1727.

29. Baxt, W.G., and Spiegelman, S. (1972). Nuclear DNA Sequences Present in Human Leukemic Cells and Absent in Normal Leukocytes. Proc. Natl. Acad. Sci. *69*, 3737–3741. 10.1073/pnas.69.12.3737.

30. Baxt, W., Hehlmann, R., and Spiegelman, S. (1972). Human Leukaemic Cells contain Reverse Transcriptase associated with a High Molecular Weight Virus-related RNA. Nature. New Biol. *240*, 72–75. 10.1038/newbio240072a0.

31. Baxt, W., Yates, J.W., Wallace, H.J., Holland, J.F., and Spiegelman, S. (1973). Leukemia-Specific DNA Sequences in Leukocytes of the Leukemic Member of Identical Twins. Proc. Natl. Acad. Sci. *70*, 2629–2632. 10.1073/pnas.70.9.2629.

32. Gallagher, R.E., Salahuddin, S.Z., Hall, W.T., McCredie, K.B., and Gallo, R.C. (1975). Growth and differentiation in culture of leukemic leukocytes from a patient with acute myelogenous leukemia and re-identification of type-C virus. Proc. Natl. Acad. Sci. *72*, 4137–4141. 10.1073/pnas.72.10.4137.

33. Reitz, M.S., Miller, N.R., Wong-Staal, F., Gallagher, R.E., Gallo, R.C., and Gillespie, D.H. (1976). Primate type-C virus nucleic acid sequences (woolly monkey and baboon types) in tissues from a patient with acute myelogenous leukemia and in viruses isolated from cultured cells of the same patient. Proc. Natl. Acad. Sci. *73*, 2113–2117. 10.1073/pnas.73.6.2113.

34. Wong-Staal, F., Gillespie, D., and Gallo, R.C. (1976). Proviral sequences of baboon endogenous type C RNA virus in DNA of human leukaemic tissues. Nature *262*, 190–195. 10.1038/262190a0.

35. Aulakh, G.S., and Gallo, R.C. (1977). Rauscher-leukemia-virus-related sequences in human DNA: presence in some tissues of some patients with hemotopoietic neoplasias and absence in DNA from other tissues. Proc. Natl. Acad. Sci. *74*, 353–357. 10.1073/pnas.74.1.353.

36. Axel, R., Schlom, J., and Spiegelman, S. (1972). Presence in Human Breast Cancer of RNA homologous to Mouse Mammary Tumour Virus RNA. Nature *235*, 32–36. 10.1038/235032a0.

37. Axel, R., Gulati, S.C., and Spiegelman, S. (1972). Particles Containing RNA-Instructed DNA Polymerase and Virus-Related RNA in Human Breast Cancers. Proc. Natl. Acad. Sci. *69*, 3133–3137. 10.1073/pnas.69.11.3133.

38. Axel, R., Schlom, J., and Spiegelman, S. (1972). Evidence for Translation of Viral-Specific RNA in Cells of a Mouse Mammary Carcinoma. Proc. Natl. Acad. Sci. *69*, 535–538. 10.1073/pnas.69.3.535.

39. Colcher, D., Spiegelman, S., and Schlom, J. (1974). Sequence Homology Between the RNA of Mason-Pfizer Monkey Virus and the RNA of Human Malignant Breast Tumors. Proc. Natl. Acad. Sci. *71*, 4975–4979. 10.1073/pnas.71.12.4975.

40. Mesa-Tejada, R., Keydar, I., Ramanarayanan, M., Ohno, T., Fenoglio, C., and Spiegelman, S. (1978). Detection in human breast carcinomas of an antigen immunologically related to a group-specific antigen of mouse mammary tumor virus. Proc. Natl. Acad. Sci. *75*, 1529–1533. 10.1073/pnas.75.3.1529.

41. Michalides, R., Spiegelman, S., and Schlom, J. (1975). Biochemical Characterization of Putative Subviral Particulates from Human Malignant Breast Tumors1. Cancer Res. *35*, 1003–1008.

42. Perk, K., Michalides, R., Spiegelman, S., and Schlom, J. (1974). Biochemical and Morphologic Evidence for the Presence of an RNA Tumor Virus in Pulmonary Carcinoma of Sheep (Jaagsiekte)2. JNCI J. Natl. Cancer Inst. *53*, 131–135. 10.1093/jnci/53.1.131.

43. Kufe, D., Hehlmann, R., and Spiegelman, S. (1972). Human Sarcomas Contain RNA Related to the RNA of a Mouse Leukemia Virus. Science *175*, 182–185. 10.1126/science.175.4018.182.

44. Kufe, D., Hehlmann, R., and Spiegelman, S. (1973). RNA Related to That of a Murine Leukemia Virus in Burkitt's Tumors and Nasopharyngeal Carcinomas. Proc. Natl. Acad. Sci. *70*, 5–9. 10.1073/pnas.70.1.5.

45. Cuatico, W., Cho, J.-R., and Spiegelman, S. (1976). Molecular evidence for a viral etiology of human CNS tumors. Acta Neurochir. (Wien) *35*, 149–160. 10.1007/BF01405943.

46. Balda, B.R., Hehlmann, R., Cho, J.R., and Spiegelman, S. (1975). Oncornavirus-like particles in human skin cancers. Proc. Natl. Acad. Sci. *72*, 3697–3700. 10.1073/pnas.72.9.3697.

47. Feldman, S.P., Schlom, J., and Spiegelman, S. (1973). Further Evidence for Oncornaviruses in Human Milk: The Production of Cores. Proc. Natl. Acad. Sci. *70*, 1976–1980. 10.1073/pnas.70.7.1976.

48. Duesberg, P.H., and Vogt, P.K. (1970). Differences between the Ribonucleic Acids of Transforming and Nontransforming Avian Tumor Viruses*. Proc. Natl. Acad. Sci. *67*, 1673–1680. 10.1073/pnas.67.4.1673.

49. Czernilofsky, A.P., Levinson, A.D., Varmus, H.E., Bishop, J.M., Tischer, E., and Goodman, H.M. (1980). Nucleotide sequence of an avian sarcoma virus oncogene (src) and proposed amino acid sequence for gene product. Nature *287*, 198–203. 10.1038/287198a0.

50. Huebner, R.J., and Todaro, G.J. (1969). Oncogenes of rna tumor viruses as determinants of cancer. Proc. Natl. Acad. Sci. *64*, 1087–1094. 10.1073/pnas.64.3.1087.

51. Takeya, T., and Hanafusa, H. (1983). Structure and sequence of the cellular gene homologous to the RSV *src* gene and the mechanism for generating the transforming virus. Cell *32*, 881–890. 10.1016/0092-8674(83)90073-9.

52. Shalloway, D., Zelenetz, A.D., and Cooper, G.M. (1981). Molecular cloning and characterization of the chicken gene homologous to the transforming gene of rous sarcoma virus. Cell *24*, 531–541. 10.1016/0092-8674(81)90344-5.

53. Stehelin, D., Varmus, H.E., Bishop, J.M., and Vogt, P.K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature *260*, 170–173. 10.1038/260170a0.

54. McCann, J., Choi, E., Yamasaki, E., and Ames, B.N. (1975). Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. Proc. Natl. Acad. Sci. *72*, 5135–5139. 10.1073/pnas.72.12.5135.

55. Shih, C., Padhy, L.C., Murray, M., and Weinberg, R.A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. Nature *290*, 261–264. 10.1038/290261a0.

56. Krontiris, T.G., and Cooper, G.M. (1981). Transforming activity of human tumor DNAs. Proc. Natl. Acad. Sci. *78*, 1181–1184. 10.1073/pnas.78.2.1181.

57. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724. 10.1038/nature07943.

58. Banerjee, H.N., and Verma, M. (2009). Epigenetic mechanisms in cancer. Biomark. Med. *3*, 397–410. 10.2217/bmm.09.26.

59. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158. 10.1038/nature05610.

60. Smith, S.M., Anastasi, J., Cohen, K.S., and Godley, L.A. (2010). The impact of MYC expression in lymphoma biology: Beyond Burkitt lymphoma. Blood Cells. Mol. Dis. *45*, 317–323. 10.1016/j.bcmd.2010.08.002.

61. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science *339*, 1546–1558. 10.1126/science.1235122.

62. Saletta, F., Dalla Pozza, L., and Byrne, J.A. (2015). Genetic causes of cancer predisposition in children and adolescents. Transl. Pediatr. *4*, 67–75. 10.3978/j.issn.2224-4336.2015.04.08.

63. Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308. 10.1038/nature12981.

64. Monteiro, A.N.A., and Waizbort, R. (2007). The accidental cancer geneticist: Hilário de Gouvêa and hereditary retinoblastoma. Cancer Biol. Ther. *6*, 811–813. 10.4161/cbt.6.5.4420.

65. Rodriguez-Galindo, C., Orbach, D.B., and VanderVeen, D. (2015). Retinoblastoma. Pediatr. Clin. *62*, 201–223. 10.1016/j.pcl.2014.09.014.

66. Knudson, A.G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. Proc. Natl. Acad. Sci. *68*, 820–823. 10.1073/pnas.68.4.820.

67. Richter, S., Vandezande, K., Chen, N., Zhang, K., Sutherland, J., Anderson, J., Han, L., Panton, R., Branco, P., and Gallie, B. (2003). Sensitive and Efficient Detection of *RB1* Gene Mutations Enhances Care for Families with Retinoblastoma. Am. J. Hum. Genet. *72*, 253–269. 10.1086/345651.

68. Houdayer, C., Gauthier-Villars, M., Laugé, A., Pagès-Berhouet, S., Dehainault, C., Caux-Moncoutier, V., Karczynski, P., Tosi, M., Doz, F., Desjardins, L., et al.

(2004). Comprehensive screening for constitutional RB1 mutations by DHPLC and QMPSF. Hum. Mutat. *23*, 193–202. 10.1002/humu.10303.

69. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. *74*, 5463–5467. 10.1073/pnas.74.12.5463.

70. Collins, F.S., and McKusick, V.A. (2001). Implications of the Human Genome Project for Medical Science. JAMA *285*, 540–544. 10.1001/jama.285.5.540.

71. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380. 10.1038/nature03959.

72. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. 10.1038/nrg.2016.49.

73. Logsdon, G.A., Vollger, M.R., and Eichler, E.E. (2020). Long-read human genome sequencing and its applications. Nat. Rev. Genet. *21*, 597–614. 10.1038/s41576-020-0236-x.

74. Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46. 10.1038/nrg2626.

75. Samorodnitsky, E., Datta, J., Jewell, B.M., Hagopian, R., Miya, J., Wing, M.R., Damodaran, S., Lippus, J.M., Reeser, J.W., Bhatt, D., et al. (2015). Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. J. Mol. Diagn. *17*, 64–75. 10.1016/j.jmoldx.2014.09.009.

76. Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nat. Rev. Genet. *24*, 464–483. 10.1038/s41576-023-00590-0.

77. Mantere, T., Kersten, S., and Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Front. Genet. *10*.

78. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784. 10.1038/s41467-018-08148-z.

79. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345. 10.1038/nbt.4060.

80. Kloth, M., and Buettner, R. (2014). Changing Histopathological Diagnostics by Genome-Based Tumor Classification. Genes *5*, 444–459. 10.3390/genes5020444.

81. The Childhood Brain Tumor Consortium (1989). Intraobserver reproducibility in assigning brain tumors to classes in the world health organization diagnostic scheme. J. Neurooncol. *7*, 211–224. 10.1007/BF00172914.

82. Rorke, L.B. (1997). Pathologic diagnosis as the gold standard. Cancer *79*, 665–667. 10.1002/(SICI)1097-0142(19970215)79:4<665::AID-CNCR1>3.0.CO;2-D.

83. Tihan, T., Zhou, T., Holmes, E., Burger, P.C., Ozuysal, S., and Rushing, E.J. (2008). The prognostic value of histological grading of posterior fossa ependymomas in children: a Children's Oncology Group study and a review of prognostic factors. Mod. Pathol. *21*, 165–177. 10.1038/modpathol.3800999.

84. Furness, P.N., Taub, N., Assmann, K.J.M., Banfi, G., Cosyns, J.-P., Dorman, A.M., Hill, C.M., Kapper, S.K., Waldherr, R., Laurinavicius, A., et al. (2003). International Variation in Histologic Grading Is Large, and Persistent Feedback Does Not Improve Reproducibility. Am. J. Surg. Pathol. *27*, 805.

85. Lessey, B.A. (2013). The pathologists are free to go, or are they? Fertil. Steril. *99*, 350–351. 10.1016/j.fertnstert.2012.11.043.

86. Stenkvist, B., Bengtsson, E., Eriksson, O., Jarkrans, T., Nordin, B., and Westman-Naeser, S. (1983). Histopathological systems of breast cancer classification: reproducibility and clinical significance. J. Clin. Pathol. *36*, 392–398.

87. Han, G., Sidhu, D., Duggan, M.A., Arseneau, J., Cesari, M., Clement, P.B., Ewanowich, C.A., Kalloger, S.E., and Köbel, M. (2013). Reproducibility of histological cell type in high-grade endometrial carcinoma. Mod. Pathol. *26*, 1594–1604. 10.1038/modpathol.2013.102.

88. Gilks, C.B., Oliva, E., and Soslow, R.A. (2013). Poor Interobserver Reproducibility in the Diagnosis of High-grade Endometrial Carcinoma. Am. J. Surg. Pathol. *37*, 874. 10.1097/PAS.0b013e31827f576a.

89. Talhouk, A., and McAlpine, J.N. (2016). New classification of endometrial cancers: the development and potential applications of genomic-based classification in research and clinical care. Gynecol. Oncol. Res. Pract. *3*, 14. 10.1186/s40661-016-0035-4.

90. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66–72. 10.1038/nature07485.

91. Sosinsky, A., Ambrose, J., Cross, W., Turnbull, C., Henderson, S., Jones, L., Hamblin, A., Arumugam, P., Chan, G., Chubb, D., et al. (2024). Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Nat. Med. *30*, 279–289. 10.1038/s41591-023-02682-0.

92. McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., M. Mastrogianakis, G., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K., et al.

(2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068. 10.1038/nature07385.

93. Weinstein, J.N., Akbani, R., Broom, B.M., Wang, W., Verhaak, R.G.W., McConkey, D., Lerner, S., Morgan, M., Creighton, C.J., Smith, C., et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322. 10.1038/nature12965.

94. Bass, A.J., Thorsson, V., Shmulevich, I., Reynolds, S.M., Miller, M., Bernard, B., Hinoue, T., Laird, P.W., Curtis, C., Shen, H., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209. 10.1038/nature13480.

95. Agrawal, N., Akbani, R., Aksoy, B.A., Ally, A., Arachchi, H., Asa, S.L., Auman, J.T., Balasundaram, M., Balu, S., Baylin, S.B., et al. (2014). Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell *159*, 676–690. 10.1016/j.cell.2014.09.050.

96. Burk, R.D., Chen, Z., Saller, C., Tarvin, K., Carvalho, A.L., Scapulatempo-Neto, C., Silveira, H.C., Fregnani, J.H., Creighton, C.J., Anderson, M.L., et al. (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378–384. 10.1038/nature21386.

97. Robertson, A.G., Shih, J., Yau, C., Gibb, E.A., Oba, J., Mungall, K.L., Hess, J.M., Uzunangelov, V., Walter, V., Danilova, L., et al. (2017). Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell *32*, 204-220.e15. 10.1016/j.ccell.2017.07.003.

98. Abeshouse, A., Adebamowo, C., Adebamowo, S.N., Akbani, R., Akeredolu, T., Ally, A., Anderson, M.L., Anur, P., Appelbaum, E.L., Armenia, J., et al. (2017). Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell *171*, 950-965.e28. 10.1016/j.cell.2017.10.014.

99. Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. 10.1038/nature11412.

100. Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353–360. 10.1038/nature11143.

101. Stephens, P.J., McBride, D.J., Lin, M.-L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005–1010. 10.1038/nature08645.

102. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400–404. 10.1038/nature11017.

103. Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399. 10.1038/nature10933.

104. Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D.W., Dao, F., Dhir, R., DiSaia, P., Gabra, H., Glenn, P., et al. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615. 10.1038/nature10166.

105. Muzny, D.M., Bainbridge, M.N., Chang, K., Dinh, H.H., Drummond, J.A., Fowler, G., Kovar, C.L., Lewis, L.R., Morgan, M.B., Newsham, I.F., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337. 10.1038/nature11252.

106. Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., et al. (2012). Recurrent R-spondin fusions in colon cancer. Nature *488*, 660–664. 10.1038/nature11282.

107. Hammerman, P.S., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., Gabriel, S., et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525. 10.1038/nature11404.

108. Berger, M.F., Hodis, E., Heffernan, T.P., Deribe, Y.L., Lawrence, M.S., Protopopov, A., Ivanova, E., Watson, I.R., Nickerson, E., Ghosh, P., et al. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature *485*, 502–506. 10.1038/nature11071.

109. Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The Origin and Evolution of Mutations in Acute Myeloid Leukemia. Cell *150*, 264–278. 10.1016/j.cell.2012.06.023.

110. Agrawal, N., Frederick, M.J., Pickering, C.R., Bettegowda, C., Chang, K., Li, R.J., Fakhry, C., Xie, T.-X., Zhang, J., Wang, J., et al. (2011). Exome Sequencing of Head and Neck Squamous Cell Carcinoma Reveals Inactivating Mutations in NOTCH1. Science *333*, 1154–1157. 10.1126/science.1206923.

111. The Cancer Genome Atlas Research Network (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N. Engl. J. Med. *372*, 2481–2498. 10.1056/NEJMoa1402121.

112. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339. 10.1038/nature12634.

113. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385.e18. 10.1016/j.cell.2018.02.060.

114. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell *158*, 929–944. 10.1016/j.cell.2014.06.049.

115. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304.e6. 10.1016/j.cell.2018.03.022.

116. Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., et al. (2018). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Rep. *23*, 282-296.e4. 10.1016/j.celrep.2018.01.088.

117. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep. *23*, 239-254.e6. 10.1016/j.celrep.2018.03.076.

118. Gao, Q., Liang, W.-W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.-W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L., et al. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. *23*, 227-238.e3. 10.1016/j.celrep.2018.03.050.

119. Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B.A., and Curigliano, G. (2019). Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer. In Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics Advances in Experimental Medicine and Biology., E. Ruiz-Garcia and H. Astudillo-de la Vega, eds. (Springer International Publishing), pp. 9–30. 10.1007/978-3-030-24100-1_2.

120. Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., and McGuire, W.L. (1987). Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/neu Oncogene. Science *235*, 177–182. 10.1126/science.3798106.

121. Fischer, O.M., Streit, S., Hart, S., and Ullrich, A. (2003). Beyond Herceptin and Gleevec. Curr. Opin. Chem. Biol. *7*, 490–495. 10.1016/S1367-5931(03)00082-6.

122. Berger, M.F., and Mardis, E.R. (2018). The emerging clinical relevance of genomics in cancer medicine. Nat. Rev. Clin. Oncol. *15*, 353–365. 10.1038/s41571-018-0002-6.

123. Kwak, E.L., Bang, Y.-J., Camidge, D.R., Shaw, A.T., Solomon, B., Maki, R.G., Ou, S.-H.I., Dezube, B.J., Jänne, P.A., Costa, D.B., et al. (2010). Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. N. Engl. J. Med. *363*, 1693–1703. 10.1056/NEJMoa1006448.

124. Shaw, A.T., Ou, S.-H.I., Bang, Y.-J., Camidge, D.R., Solomon, B.J., Salgia, R., Riely, G.J., Varella-Garcia, M., Shapiro, G.I., Costa, D.B., et al. (2014). Crizotinib in ROS1-Rearranged Non–Small-Cell Lung Cancer. N. Engl. J. Med. *371*, 1963–1971. 10.1056/NEJMoa1406766.

125. Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al. (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. N. Engl. J. Med. *364*, 2507–2516. 10.1056/NEJMoa1103782.

126. Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., Lichinitser, M., Dummer, R., Grange, F., Mortier, L., et al. (2015). Improved Overall Survival in Melanoma with Combined Dabrafenib and Trametinib. N. Engl. J. Med. *372*, 30–39. 10.1056/NEJMoa1412690.

127. Szymiczek, A., Lone, A., and Akbari, M.R. (2021). Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. Clin. Genet. *99*, 613–637. 10.1111/cge.13900.

128. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755. 10.1038/nrg3031.

129. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al. (2010). Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. N. Engl. J. Med. *362*, 1181–1191. 10.1056/NEJMoa0908094.

130. Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L., et al. (2011). Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med. *13*, 255–262. 10.1097/GIM.0b013e3182088158.

131. Meder, B., Haas, J., Keller, A., Heid, C., Just, S., Borries, A., Boisguerin, V., Scharfenberger-Schmeer, M., Stähler;, P., Beier, M., et al. (2011). Targeted Next-Generation Sequencing for the Molecular Genetic Diagnostics of Cardiomyopathies. Circ. Cardiovasc. Genet. *4*, 110–122. 10.1161/CIRCGENETICS.110.958322.

132. Meng, L., Pammi, M., Saronwala, A., Magoulas, P., Ghazi, A.R., Vetrini, F., Zhang, J., He, W., Dharmadhikari, A.V., Qu, C., et al. (2017). Use of Exome Sequencing for Infants in Intensive Care Units: Ascertainment of Severe Single-Gene Disorders and Effect on Medical Management. JAMA Pediatr. *171*, e173438. 10.1001/jamapediatrics.2017.3438.

133. Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. *106*, 19096–19101. 10.1073/pnas.0910672106.

134. Hoischen, A., van Bon, B.W.M., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., de Vries, P., de Reuver, R., Wieskamp, N., Mortier, G., et al. (2010). De

novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. *42*, 483–485. 10.1038/ng.581.

135. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276. 10.1038/nature08250.

136. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N. Engl. J. Med. *369*, 1502–1511. 10.1056/NEJMoa1306555.

137. Eldomery, M.K., Coban-Akdemir, Z., Harel, T., Rosenfeld, J.A., Gambin, T., Stray-Pedersen, A., Küry, S., Mercier, S., Lessel, D., Denecke, J., et al. (2017). Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. *9*, 26. 10.1186/s13073-017-0412-6.

138. Saunders, C.J., Miller, N.A., Soden, S.E., Dinwiddie, D.L., Noll, A., Alnadi, N.A., Andraws, N., Patterson, M.L., Krivohlavek, L.A., Fellis, J., et al. (2012). Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units. Sci. Transl. Med. *4*, 154ra135-154ra135. 10.1126/scitranslmed.3004041.

139. Farnaes, L., Hildreth, A., Sweeney, N.M., Clark, M.M., Chowdhury, S., Nahas, S., Cakici, J.A., Benson, W., Kaplan, R.H., Kronick, R., et al. (2018). Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. Npj Genomic Med. *3*, 1–8. 10.1038/s41525-018-0049-4.

140. Willig, L.K., Petrikin, J.E., Smith, L.D., Saunders, C.J., Thiffault, I., Miller, N.A., Soden, S.E., Cakici, J.A., Herd, S.M., Twist, G., et al. (2015). Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. Lancet Respir. Med. *3*, 377–387. 10.1016/S2213-2600(15)00139-3.

141. Merker, J.D., Wenger, A.M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K.S., et al. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet. Med. *20*, 159–163. 10.1038/gim.2017.86.

142. Mizuguchi, T., Suzuki, T., Abe, C., Umemura, A., Tokunaga, K., Kawai, Y., Nakamura, M., Nagasaki, M., Kinoshita, K., Okamura, Y., et al. (2019). A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J. Hum. Genet. *64*, 359–368. 10.1038/s10038-019-0569-5.

143. Miao, H., Zhou, J., Yang, Q., Liang, F., Wang, D., Ma, N., Gao, B., Du, J., Lin, G., Wang, K., et al. (2018). Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. Hereditas *155*, 32. 10.1186/s41065-018-0069-1.

144. Reiner, J., Pisani, L., Qiao, W., Singh, R., Yang, Y., Shi, L., Khan, W.A., Sebra, R., Cohen, N., Babu, A., et al. (2018). Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a Bardet–Biedl

Syndrome 9 (BBS9) deletion. Npj Genomic Med. *3*, 1–5. 10.1038/s41525-017-0042-3.

145. Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T., Takahashi, M., Matsuura, T., Flanigan, K.M., Iwasaki, S., et al. (2009). Spinocerebellar Ataxia Type 31 Is Associated with "Inserted" Penta-Nucleotide Repeats Containing (TGGAA)n. Am. J. Hum. Genet. *85*, 544–557. 10.1016/j.ajhg.2009.09.019.

146. de Jong, L.C., Cree, S., Lattimore, V., Wiggins, G.A.R., Spurdle, A.B., Miller, A., Kennedy, M.A., Walker, L.C., and kConFab Investigators (2017). Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. Breast Cancer Res. *19*, 127. 10.1186/s13058-017-0919-1.

147. Wenzel, A., Altmueller, J., Ekici, A.B., Popp, B., Stueber, K., Thiele, H., Pannes, A., Staubach, S., Salido, E., Nuernberg, P., et al. (2018). Single molecule real time sequencing in ADTKD-MUC1 allows complete assembly of the VNTR and exact positioning of causative mutations. Sci. Rep. *8*, 4170. 10.1038/s41598-018-22428-0.

148. Sone, J., Mitsuhashi, S., Fujita, A., Mizuguchi, T., Hamanaka, K., Mori, K., Koike, H., Hashiguchi, A., Takashima, H., Sugiyama, H., et al. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat. Genet. *51*, 1215–1221. 10.1038/s41588-019-0459-y.

149. Brioude, F., Kalish, J.M., Mussa, A., Foster, A.C., Bliek, J., Ferrero, G.B., Boonen, S.E., Cole, T., Baker, R., Bertoletti, M., et al. (2018). Clinical and molecular diagnosis, screening and management of Beckwith–Wiedemann syndrome: an international consensus statement. Nat. Rev. Endocrinol. *14*, 229–249. 10.1038/nrendo.2017.166.

150. Doherty, E.S., Lacbawan, F., Hadley, D.W., Brewer, C., Zalewski, C., Kim, H.J., Solomon, B., Rosenbaum, K., Domingo, D.L., Hart, T.C., et al. (2007). Muenke syndrome (FGFR3-related craniosynostosis): Expansion of the phenotype and review of the literature. Am. J. Med. Genet. A. *143A*, 3204–3215. 10.1002/ajmg.a.32078.

151. Wojcik, M.H., Srivastava, S., Agrawal, P.B., Balci, T.B., Callewaert, B., Calvo, P.L., Carli, D., Caudle, M., Colaiacovo, S., Cross, L., et al. (2023). Jansen-de Vries syndrome: Expansion of the PPM1D clinical and phenotypic spectrum in 34 families. Am. J. Med. Genet. A. *191*, 1900–1910. 10.1002/ajmg.a.63226.

152. Lieberman, S., Walsh, T., Schechter, M., Adar, T., Goldin, E., Beeri, R., Sharon, N., Baris, H., Ben Avi, L., Half, E., et al. (2017). Features of Patients With Hereditary Mixed Polyposis Syndrome Caused by Duplication of GREM1 and Implications for Screening and Surveillance. Gastroenterology *152*, 1876-1880.e1. 10.1053/j.gastro.2017.02.014.

153. Wright, C.F., FitzPatrick, D.R., and Firth, H.V. (2018). Paediatric genomics: diagnosing rare disease in children. Nat. Rev. Genet. *19*, 253–268. 10.1038/nrg.2017.116.

154. Adams, D.R., and Eng, C.M. (2018). Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. N. Engl. J. Med. *379*, 1353–1362. 10.1056/NEJMra1711801.

155. McVean, G.A., Altshuler (Co-Chair), D.M., Durbin (Co-Chair), R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65. 10.1038/nature11632.

156. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. 10.1093/nar/gkx1153.

157. Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022. 10.1101/gad.2037511.

158. Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905. 10.1126/science.1237905.

159. Smallwood, S.A., and Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. Trends Genet. *28*, 33–42. 10.1016/j.tig.2011.09.004.

160. Hsieh, C.-L. (1999). In Vivo Activity of Murine De Novo Methyltransferases, Dnmt3a and Dnmt3b. Mol. Cell. Biol. *19*, 8211–8218. 10.1128/MCB.19.12.8211.

161. Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. *19*, 219–220. 10.1038/890.

162. Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257. 10.1016/S0092-8674(00)81656-6.

163. Pradhan, S., Bacolla, A., Wells, R.D., and Roberts, R.J. (1999). Recombinant Human DNA (Cytosine-5) Methyltransferase: I. EXPRESSION, PURIFICATION, AND COMPARISON OF DE NOVO AND MAINTENANCE METHYLATION*. J. Biol. Chem. *274*, 33002–33010. 10.1074/jbc.274.46.33002.

164. Laird, P.W. (2003). The power and the promise of DNA methylation markers. Nat. Rev. Cancer *3*, 253–266. 10.1038/nrc1045.

165. Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282. 10.1016/0022-2836(87)90689-9.

166. Feinberg, A.P., and Tycko, B. (2004). The history of cancer epigenetics. Nat. Rev. Cancer *4*, 143–153. 10.1038/nrc1279.

167. Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89–92. 10.1038/301089a0.

168. Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. *43*, 768–775. 10.1038/ng.865.

169. Feinberg, A.P., Gehrke, C.W., Kuo, K.C., and Ehrlich, M. (1988). Reduced Genomic 5-Methylcytosine Content in Human Colonic Neoplasia1. Cancer Res. *48*, 1159–1161.

170. Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P.E., van Dijk, C.M., Tollenaar, R.A.E.M., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. Nat. Genet. *44*, 40–46. 10.1038/ng.969.

171. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186. 10.1038/ng.298.

172. Qu, G., Grundy, P.E., Narayan, A., and Ehrlich, M. (1999). Frequent Hypomethylation in Wilms Tumors of Pericentromeric DNA in Chromosomes 1 and 16. Cancer Genet. Cytogenet. *109*, 34–39. 10.1016/S0165-4608(98)00143-5.

173. Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M.A., Richter, J., Soler, E., et al. (2010). Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat. Med. *16*, 571–579. 10.1038/nm.2129.

174. Nishigaki, M., Aoyagi, K., Danjoh, I., Fukaya, M., Yanagihara, K., Sakamoto, H., Yoshida, T., and Sasaki, H. (2005). Discovery of Aberrant Expression of R-RAS by Cancer-Linked DNA Hypomethylation in Gastric Cancer Using Microarrays. Cancer Res. *65*, 2115–2124. 10.1158/0008-5472.CAN-04-3340.

175. Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. *22*, 246–258. 10.1101/gr.125872.111.

176. Herman, J.G., and Baylin, S.B. (2003). Gene Silencing in Cancer in Association with Promoter Hypermethylation. N. Engl. J. Med. *349*, 2042–2054. 10.1056/NEJMra023075.

177. Sproul, D., Kitchen, R.R., Nestor, C.E., Dixon, J.M., Sims, A.H., Harrison, D.J., Ramsahoye, B.H., and Meehan, R.R. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. Genome Biol. *13*, R84. 10.1186/gb-2012-13-10-r84.

178. Esteller, M., Hamilton, S.R., Burger, P.C., Baylin, S.B., and Herman, J.G. (1999). Inactivation of the DNA Repair Gene O6-Methylguanine-DNA Methyltransferase by Promoter Hypermethylation is a Common Event in Primary Human Neoplasia1. Cancer Res. *59*, 793–797.

179. Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., et al. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. *38*, 787–793. 10.1038/ng1834.

180. Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B., and Issa, J.-P.J. (1999). CpG island methylator phenotype in colorectal cancer. Proc. Natl. Acad. Sci. *96*, 8681–8686. 10.1073/pnas.96.15.8681.

181. Li, J., Xu, C., Lee, H.J., Ren, S., Zi, X., Zhang, Z., Wang, H., Yu, Y., Yang, C., Gao, X., et al. (2020). A genomic and epigenomic atlas of prostate cancer in Asian populations. Nature *580*, 93–99. 10.1038/s41586-020-2135-x.

182. Grady, W.M., Willis, J., Guilford, P.J., Dunbier, A.K., Toro, T.T., Lynch, H., Wiesner, G., Ferguson, K., Eng, C., Park, J.-G., et al. (2000). Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. Nat. Genet. *26*, 16–17. 10.1038/79120.

183. Fleisher, A.S., Esteller, M., Tamura, G., Rashid, A., Stine, O.C., Yin, J., Zou, T.-T., Abraham, J.M., Kong, D., Nishizuka, S., et al. (2001). Hypermethylation of the hMLH1 gene promoter is associated with microsatellite instability in early human gastric neoplasia. Oncogene *20*, 329–335. 10.1038/sj.onc.1204104.

184. Tepel, M., Roerig, P., Wolter, M., Gutmann, D.H., Perry, A., Reifenberger, G., and Riemenschneider, M.J. (2008). Frequent promoter hypermethylation and transcriptional downregulation of the NDRG2 gene at 14q11.2 in primary glioblastoma. Int. J. Cancer *123*, 2080–2086. 10.1002/ijc.23705.

185. Bachman, K.E., Herman, J.G., Corn, P.G., Merlo, A., Costello, J.F., Cavenee, W.K., Baylin, S.B., and Graff, J.R. (1999). Methylation-associated silencing of the tissue inhibitor of metalloproteinase-3 gene suggest a suppressor role in kidney, brain, and other human cancers. Cancer Res. *59*, 798–802.

186. Kim, T.-Y., Zhong, S., Fields, C.R., Kim, J.H., and Robertson, K.D. (2006). Epigenomic Profiling Reveals Novel and Frequent Targets of Aberrant DNA Methylation-Mediated Silencing in Malignant Glioma. Cancer Res. *66*, 7490–7501. 10.1158/0008-5472.CAN-05-4552.

187. Nakamura, M., Yonekawa, Y., Kleihues, P., and Ohgaki, H. (2001). Promoter Hypermethylation of the *RB1* Gene in Glioblastomas. Lab. Invest. *81*, 77–82. 10.1038/labinvest.3780213.

188. Hegi, M.E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., de Tribolet, N., Weller, M., Kros, J.M., Hainfellner, J.A., Mason, W., Mariani, L., et al. (2005). MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. N. Engl. J. Med. *352*, 997–1003. 10.1056/NEJMoa043331.

189. Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S.N., Hidalgo, O.F., Vanaclocha, V., Baylin, S.B., and Herman, J.G. (2000). Inactivation of the DNA-Repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents. N. Engl. J. Med. *343*, 1350–1354. 10.1056/NEJM200011093431901.

190. Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P., et al. (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Cancer Cell *17*, 510–522. 10.1016/j.ccr.2010.03.017.

191. Greger, V., Passarge, E., Höpping, W., Messmer, E., and Horsthemke, B. (1989). Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. Hum. Genet. *83*, 155–158. 10.1007/BF00286709.

192. Li, H., and Minarovits, J. (2003). Host Cell-Dependent Expression of Latent Epstein–Barr Virus Genomes: Regulation by DNA Methylation. In Advances in Cancer Research (Academic Press), pp. 133–156. 10.1016/S0065-230X(03)01004-2.

193. Tucci, V., Isles, A.R., Kelsey, G., Ferguson-Smith, A.C., Tucci, V., Bartolomei, M.S., Benvenisty, N., Bourc'his, D., Charalambous, M., Dulac, C., et al. (2019). Genomic Imprinting and Physiological Processes in Mammals. Cell *176*, 952–965. 10.1016/j.cell.2019.01.043.

194. Ferguson-Smith, A.C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. *12*, 565–575. 10.1038/nrg3032.

195. Pal, N., Wadey, R.B., Buckle, B., Yeomans, E., Pritchard, J., and Cowell, J.K. (1990). Preferential loss of maternal alleles in sporadic Wilms' tumour. Oncogene *5*, 1665–1668.

196. Schroeder, W.T., Chao, L.Y., Dao, D.D., Strong, L.C., Pathak, S., Riccardi, V., Lewis, W.H., and Saunders, G.F. (1987). Nonrandom loss of maternal chromosome 11 alleles in Wilms tumors. Am. J. Hum. Genet. *40*, 413–420.

197. Brown, D.R., Schroeder, J.M., Bryan, J.T., Stoler, M.H., and Fife, K.H. (1999). Detection of Multiple Human Papillomavirus Types in Condylomata Acuminata Lesions from Otherwise Healthy and Immunosuppressed Patients. J. Clin. Microbiol. *37*, 3316–3322. 10.1128/jcm.37.10.3316-3322.1999.

198. Zhang, Y., and Tycko, B. (1992). Monoallelic expression of the human H19 gene. Nat. Genet. *1*, 40–44. 10.1038/ng0492-40.

199. Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C.G., and Polychronakos, C. (1993). Parental genomic imprinting of the human IGF2 gene. Nat. Genet. *4*, 98–101. 10.1038/ng0593-98.

200. Ohlsson, R., Nyström, A., Pfeifer-Ohlsson, S., Töhönen, V., Hedborg, F., Schofield, P., Flam, F., and Ekström, T.J. (1993). IGF2 is parentally imprinted during human embryogenesis and in the Beckwith–Wiedemann syndrome. Nat. Genet. *4*, 94–97. 10.1038/ng0593-94.

201. Rainier, S., Johnson, L.A., Dobry, C.J., Ping, A.J., Grundy, P.E., and Feinberg, A.P. (1993). Relaxation of imprinted genes in human cancer. Nature *362*, 747–749. 10.1038/362747a0.

202. Ogawa, O., Becroft, D.M., Morison, I.M., Eccles, M.R., Skeen, J.E., Mauger, D.C., and Reeve, A.E. (1993). Constitutional relaxation of insulin–like growth

factor II gene imprinting associated with Wilms' tumour and gigantism. Nat. Genet. *5*, 408–412. 10.1038/ng1293-408.

203. Orlando, V. (2003). Polycomb, Epigenomes, and Control of Cell Identity. Cell *112*, 599–606. 10.1016/S0092-8674(03)00157-0.

204. Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353. 10.1038/nature04733.

205. Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. *16*, 2893–2905. 10.1101/gad.1035902.

206. Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. Cell *111*, 197–208. 10.1016/S0092-8674(02)00976-5.

207. Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. Science *298*, 1039–1043. 10.1126/science.1076997.

208. Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites. Cell *111*, 185–196. 10.1016/S0092-8674(02)00975-3.

209. Kleer, C.G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S.A., Ghosh, D., Sewalt, R.G.A.B., Otte, A.P., Hayes, D.F., et al. (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc. Natl. Acad. Sci. *100*, 11606–11611. 10.1073/pnas.1933744100.

210. Holm, K., Grabau, D., Lövgren, K., Aradottir, S., Gruvberger-Saal, S., Howlin, J., Saal, L.H., Ethier, S.P., Bendahl, P.-O., Stål, O., et al. (2012). Global H3K27 trimethylation and EZH2 abundance in breast tumor subtypes. Mol. Oncol. *6*, 494–506. 10.1016/j.molonc.2012.06.002.

211. Pietersen, A.M., Horlings, H.M., Hauptmann, M., Langerød, A., Ajouaou, A., Cornelissen-Steijger, P., Wessels, L.F., Jonkers, J., Vijver, M.J. van de, and van Lohuizen, M. (2008). EZH2 and BMI1 inversely correlate with prognosis and TP53 mutation in breast cancer. Breast Cancer Res. *10*, R109. 10.1186/bcr2214.

212. Collett, K., Eide, G.E., Arnes, J., Stefansson, I.M., Eide, J., Braaten, A., Aas, T., Otte, A.P., and Akslen, L.A. (2006). Expression of enhancer of zeste homologue 2 is significantly associated with increased tumor cell proliferation and is a marker of aggressive breast cancer. Clin. Cancer Res. *12*, 1168–1174. 10.1158/1078-0432.CCR-05-1533.

213. Raman, J.D., Mongan, N.P., Tickoo, S.K., Boorjian, S.A., Scherr, D.S., and Gudas, L.J. (2005). Increased Expression of the Polycomb Group Gene, EZH2, in Transitional Cell Carcinoma of the Bladder. Clin. Cancer Res. *11*, 8570–8576. 10.1158/1078-0432.CCR-05-1047.

214. Varambally, S., Dhanasekaran, S.M., Zhou, M., Barrette, T.R., Kumar-Sinha, C., Sanda, M.G., Ghosh, D., Pienta, K.J., Sewalt, R.G.A.B., Otte, A.P., et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature *419*, 624–629. 10.1038/nature01075.

215. Zhou, J., Roh, J.-W., Bandyopadhyay, S., Chen, Z., Munkarah, A.R., Hussein, Y., Alosh, B., Jazaerly, T., Hayek, K., Semaan, A., et al. (2013). Overexpression of enhancer of zeste homolog 2 (EZH2) and focal adhesion kinase (FAK) in high grade endometrial carcinoma. Gynecol. Oncol. *128*, 344–348. 10.1016/j.ygyno.2012.07.128.

216. Matsukawa, Y., Semba, S., Kato, H., Ito, A., Yanagihara, K., and Yokozaki, H. (2006). Expression of the enhancer of zeste homolog 2 is correlated with poor prognosis in human gastric cancer. Cancer Sci. *97*, 484–491. 10.1111/j.1349-7006.2006.00203.x.

217. Breuer, R.H.J., Snijders, P.J.F., Smit, E.F., Sutedja, T.G., Sewalt, R.G.A.B., Otte, A.P., van Kemenade, F.J., Postmus, P.E., Meijer, C.J.L.M., and Raaphorst, F.M. (2004). Increased Expression of the *EZH2* Polycomb Group Gene in *BMI-1*-Positive Neoplastic Cells during Bronchial Carcinogenesis. Neoplasia *6*, 736–743. 10.1593/neo.04160.

218. Kikuchi, J., Kinoshita, I., Shimizu, Y., Kikuchi, E., Konishi, J., Oizumi, S., Kaga, K., Matsuno, Y., Nishimura, M., and Dosaka-Akita, H. (2010). Distinctive expression of the polycomb group proteins Bmi1 polycomb ring finger oncogene and enhancer of zeste homolog 2 in nonsmall cell lung cancers and their clinical and clinicopathologic significance. Cancer *116*, 3015–3024. 10.1002/cncr.25128.

219. Ougolkov, A.V., Bilim, V.N., and Billadeau, D.D. (2008). Regulation of Pancreatic Tumor Cell Proliferation and Chemoresistance by the Histone Methyltransferase Enhancer of Zeste Homologue 2. Clin. Cancer Res. *14*, 6790–6796. 10.1158/1078-0432.CCR-08-1013.

220. Bachmann, I.M., Halvorsen, O.J., Collett, K., Stefansson, I.M., Straume, O., Haukaas, S.A., Salvesen, H.B., Otte, A.P., and Akslen, L.A. (2006). EZH2 Expression Is Associated With High Proliferation Rate and Aggressive Tumor Subgroups in Cutaneous Melanoma and Cancers of the Endometrium, Prostate, and Breast. J. Clin. Oncol. *24*, 268–273. 10.1200/JCO.2005.01.5180.

221. Ngollo, M., Lebert, A., Daures, M., Judes, G., Rifai, K., Dubois, L., Kemeny, J.-L., Penault-Llorca, F., Bignon, Y.-J., Guy, L., et al. (2017). Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression. BMC Cancer *17*, 261. 10.1186/s12885-017-3256-y.

222. Sato, T., Kaneda, A., Tsuji, S., Isagawa, T., Yamamoto, S., Fujita, T., Yamanaka, R., Tanaka, Y., Nukiwa, T., Marquez, V.E., et al. (2013). PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. Sci. Rep. *3*, 1911. 10.1038/srep01911.

223. Sengupta, D., Byrum, S.D., Avaritt, N.L., Davis, L., Shields, B., Mahmoud, F., Reynolds, M., Orr, L.M., Mackintosh, S.G., Shalin, S.C., et al. (2016). Quantitative Histone Mass Spectrometry Identifies Elevated Histone H3 Lysine 27 (Lys27) Trimethylation in Melanoma *. Mol. Cell. Proteomics *15*, 765–775. 10.1074/mcp.M115.053363.

224. Wei, Y., Xia, W., Zhang, Z., Liu, J., Wang, H., Adsay, N.V., Albarracin, C., Yu, D., Abbruzzese, J.L., Mills, G.B., et al. (2008). Loss of trimethylation at lysine 27 of histone H3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers. Mol. Carcinog. *47*, 701–706. 10.1002/mc.20413.

225. Ellinger, J., Bachmann, A., Göke, F., Behbahani, T.E., Baumann, C., Heukamp, L.C., Rogenhofer, S., and Müller, S.C. (2014). Alterations of Global Histone H3K9 and H3K27 Methylation Levels in Bladder Cancer. Urol. Int. *93*, 113–118. 10.1159/000355467.

226. McCabe, M.T., Graves, A.P., Ganji, G., Diaz, E., Halsey, W.S., Jiang, Y., Smitheman, K.N., Ott, H.M., Pappalardi, M.B., Allen, K.E., et al. (2012). Mutation of A677 in histone methyltransferase EZH2 in human B-cell lymphoma promotes hypertrimethylation of histone H3 on lysine 27 (H3K27). Proc. Natl. Acad. Sci. *109*, 2989–2994. 10.1073/pnas.1116418109.

227. Morin, R.D., Johnson, N.A., Severson, T.M., Mungall, A.J., An, J., Goya, R., Paul, J.E., Boyle, M., Woolcock, B.W., Kuchenbauer, F., et al. (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat. Genet. *42*, 181–185. 10.1038/ng.518.

228. Bödör, C., O'Riain, C., Wrench, D., Matthews, J., Iyengar, S., Tayyib, H., Calaminici, M., Clear, A., Iqbal, S., Quentmeier, H., et al. (2011). EZH2 Y641 mutations in follicular lymphoma. Leukemia *25*, 726–729. 10.1038/leu.2010.311.

229. Schaefer, I.-M., Fletcher, C.D., and Hornick, J.L. (2016). Loss of H3K27 trimethylation distinguishes malignant peripheral nerve sheath tumors from histologic mimics. Mod. Pathol. *29*, 4–13. 10.1038/modpathol.2015.134.

230. Katz, L.M., Hielscher, T., Liechty, B., Silverman, J., Zagzag, D., Sen, R., Wu, P., Golfinos, J.G., Reuss, D., Neidert, M.C., et al. (2018). Loss of histone H3K27me3 identifies a subset of meningiomas with increased risk of recurrence. Acta Neuropathol. (Berl.) *135*, 955–963. 10.1007/s00401-018-1844-9.

231. Nassiri, F., Wang, J.Z., Singh, O., Karimi, S., Dalcourt, T., Ijad, N., Pirouzmand, N., Ng, H.-K., Saladino, A., Pollo, B., et al. (2021). Loss of H3K27me3 in meningiomas. Neuro-Oncol. 10.1093/neuonc/noab036.

232. Ammendola, S., and Barresi, V. (2022). Timing of H3K27me3 loss in secondary anaplastic meningiomas. Brain Tumor Pathol. *39*, 179–181. 10.1007/s10014-021-00422-1.

233. Bayliss, J., Mukherjee, P., Lu, C., Jain, S.U., Chung, C., Martinez, D., Sabari, B., Margol, A.S., Panwalkar, P., Parolia, A., et al. (2016). Lowered H3K27me3

and DNA hypomethylation define poorly prognostic pediatric posterior fossa ependymomas. Sci. Transl. Med. *8*, 366ra161. 10.1126/scitranslmed.aah6904.

234. Ammendola, S., Caldonazzi, N., Simbolo, M., Piredda, M.L., Brunelli, M., Poliani, P.L., Pinna, G., Sala, F., Ghimenton, C., Scarpa, A., et al. (2021). H3K27me3 immunostaining is diagnostic and prognostic in diffuse gliomas with oligodendroglial or mixed oligoastrocytic morphology. Virchows Arch. *479*, 987–996. 10.1007/s00428-021-03134-1.

235. Wu, G., Broniscer, A., McEachron, T.A., Lu, C., Paugh, B.S., Becksfort, J., Qu, C., Ding, L., Huether, R., Parker, M., et al. (2012). Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nat. Genet. *44*, 251–253. 10.1038/ng.1102.

236. Venneti, S., Garimella, M.T., Sullivan, L.M., Martinez, D., Huse, J.T., Heguy, A., Santi, M., Thompson, C.B., and Judkins, A.R. (2013). Evaluation of Histone 3 Lysine 27 Trimethylation (H3K27me3) and Enhancer of Zest 2 (EZH2) in Pediatric Glial and Glioneuronal Tumors Shows Decreased H3K27me3 in H3F3A K27M Mutant Glioblastomas. Brain Pathol. *23*, 558–564. 10.1111/bpa.12042.

237. Bender, S., Tang, Y., Lindroth, A.M., Hovestadt, V., Jones, D.T.W., Kool, M., Zapatka, M., Northcott, P.A., Sturm, D., Wang, W., et al. (2013). Reduced H3K27me3 and DNA Hypomethylation Are Major Drivers of Gene Expression in K27M Mutant Pediatric High-Grade Gliomas. Cancer Cell *24*, 660–672. 10.1016/j.ccr.2013.10.006.

238. Schwartzentruber, J., Korshunov, A., Liu, X.-Y., Jones, D.T.W., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A.M., Quang, D.-A.K., Tönjes, M., et al. (2012). Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature *482*, 226–231. 10.1038/nature10833.

239. Deshmukh, S., Ptack, A., Krug, B., and Jabado, N. Oncohistones: a roadmap to stalled development. FEBS J. *n/a*. 10.1111/febs.15963.

240. Mohammad, F., Weissmann, S., Leblanc, B., Pandey, D.P., Højfeldt, J.W., Comet, I., Zheng, C., Johansen, J.V., Rapin, N., Porse, B.T., et al. (2017). EZH2 is a potential therapeutic target for H3K27M-mutant pediatric gliomas. Nat. Med. *23*, 483–492. 10.1038/nm.4293.

241. Chan, K.-M., Fang, D., Gan, H., Hashizume, R., Yu, C., Schroeder, M., Gupta, N., Mueller, S., James, C.D., Jenkins, R., et al. (2013). The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes Dev. *27*, 985–990. 10.1101/gad.217778.113.

242. Cheney, A.R., Monlong, J., Beale, H.C., Olsen, H., Kephart, E.T., Learned, K., White, S., Martinez-Agosto, J.A., Federman, N., Akeson, M., et al. (2021). Abstract 261: Long-read sequencing characterization of a patient with bilateral Wilms tumor of unknown etiology. Cancer Res. *81*, 261. 10.1158/1538-7445.AM2021-261.

243. Bouvard, V., Baan, R., Straif, K., Grosse, Y., Secretan, B., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., et al. (2009). A

review of human carcinogens—Part B: biological agents. Lancet Oncol. *10*, 321–322. 10.1016/S1470-2045(09)70096-8.

244. Bouvard, V., Baan, R.A., Grosse, Y., Lauby-Secretan, B., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., and Straif, K. (2012). Carcinogenicity of malaria and of some polyomaviruses. Lancet Oncol. *13*, 339–340. 10.1016/S1470-2045(12)70125-0.

245. Martel, C. de, Georges, D., Bray, F., Ferlay, J., and Clifford, G.M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob. Health *8*, e180–e190. 10.1016/S2214-109X(19)30488-7.

246. Ellermann, V., and Bang, O. (1909). Experimentelle Leukämie bei Hühnern. II. Z. Für Hyg. Infekt. *63*, 231–272. 10.1007/BF02227892.

247. Shope, R.E. (1935). Serial Transmission of Virus of Infectious Papillomatosis in Domestic Rabbits. Proc. Soc. Exp. Biol. Med. *32*, 830–832. 10.3181/00379727-32-7875.

248. Gross, L. (1953). A Filterable Agent, Recovered from Ak Leukemic Extracts, Causing Salivary Gland Carcinomas in C3H Mice. Proc. Soc. Exp. Biol. Med. *83*, 414–421. 10.3181/00379727-83-20376.

249. Lucké, B. (1938). CARCINOMA IN THE LEOPARD FROG: ITS PROBABLE CAUSATION BY A VIRUS. J. Exp. Med. *68*, 457–468.

250. Zur Hausen, H. (2006). Infections causing human cancer (Wiley-VCH) 10.1002/3527609318.

251. Burkitt, D. (1958). A sarcoma involving the jaws in african children. BJS Br. J. Surg. *46*, 218–223. 10.1002/bjs.18004619704.

252. Epstein, M.A., Achong, B.G., and Barr, Y.M. (1964). VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. Lancet Lond. Engl. *1*, 702–703. 10.1016/s0140-6736(64)91524-7.

253. Hausen, H.Z., and Schulte-Holthausen, H. (1970). Presence of EB Virus Nucleic Acid Homology in a "Virus-free" Line of Burkitt Tumour Cells. Nature *227*, 245–248. 10.1038/227245a0.

254. Shope, T., Dechairo, D., and Miller, G. (1973). Malignant Lymphoma in Cottontop Marmosets after Inoculation with Epstein-Barr Virus. Proc. Natl. Acad. Sci. *70*, 2487–2491. 10.1073/pnas.70.9.2487.

255. Epstein, M.A., Hunt, R.D., and Rabin, H. (1973). Pilot experiments with EB virus in owl monkeys (aotus trivirgatus). I. Reticuloproliferative disease in an inoculated animal. Int. J. Cancer *12*, 309–318. 10.1002/ijc.2910120202.

256. Gallo, R.C. (2005). History of the discoveries of the first human retroviruses: HTLV-1 and HTLV-2. Oncogene *24*, 5926–5930. 10.1038/sj.onc.1208980.

257. Payet, M., Camain, R., and Pene, P. (1956). [Primary cancer of the liver; critical study of 240 cases]. Rev. Int. Hepatol. *6*, 1–86.

258. Vogel, C.L., Mody, N., Anthony, P.P., and Barker, L.F. (1970). HEPATITIS-ASSOCIATED ANTIGEN IN UGANDAN PATIENTS WITH HEPATOCELLULAR CARCINOMA. The Lancet *296*, 621–624. 10.1016/S0140-6736(70)91396-6.

259. Denison, E.K., Peters, R.L., and Reynolds, T.B. (1971). Familial Hepatoma with Hepatitis-Associated Antigen. Ann. Intern. Med. *74*, 391–394. 10.7326/0003-4819-74-3-391.

260. Trichopoulos, D., Violaki, M., Sparros, L., and Xirouchaki, E. (1975). EPIDEMIOLOGY OF HEPATITIS B AND PRIMARY HEPATIC CARCINOMA. The Lancet *306*, 1038–1039. 10.1016/S0140-6736(75)90322-0.

261. Larouze, B., Saimot, G., Lustbader, E.D., London, W.T., Werner, B.G., Payet, M., and Blumberg, B.S. (1976). HOST RESPONSES TO HEPATITIS-B INFECTION IN PATIENTS WITH PRIMARY HEPATIC CARCINOMA AND THEIR FAMILIES: A Case/Control Study in Senegal, West Africa. The Lancet *308*, 534–538. 10.1016/S0140-6736(76)91792-X.

262. Terés, J., Guardia, J., Bruguera, M., and Rodes, J. (1971). HEPATITIS-ASSOCIATED ANTIGEN AND HEPATOCELLULAR CARCINOMA. The Lancet *298*, 215. 10.1016/S0140-6736(71)90926-3.

263. Beasley, R.P., Lin, C.-C., Hwang, L.-Y., and Chien, C.-S. (1981). HEPATOCELLULAR CARCINOMA AND HEPATITIS B VIRUS: A Prospective Study of 22 707 Men in Taiwan. The Lancet *318*, 1129–1133. 10.1016/S0140-6736(81)90585-7.

264. Rawls, W.E., Tompkins, W.A.F., Figueroa, M.E., and Melnick, J.L. (1968). Herpesvirus Type 2: Association with Carcinoma of the Cervix. Science *161*, 1255–1256. 10.1126/science.161.3847.1255.

265. Naib, Z.M., Nahmias, A.J., Josey, W.E., and Kramer, J.H. (1969). Genital herpetic infection association with cervical dysplasia and carcinoma. Cancer *23*, 940–945. 10.1002/1097-0142(196904)23:4<940::AID-CNCR2820230432>3.0.CO;2-E.

266. Aurelian, L., Strnad, B.C., and Smith, M.F. (1977). Immunodiagnostic potential of a virus-coded, tumor-associated antigen (AG-4) in cervical cancer. Cancer *39*, 1834–1849. 10.1002/1097-0142(197704)39:4+<1834::AID-CNCR2820390816>3.0.CO;2-L.

267. Gupta, P.K., Aurelian, L., Frost, J.K., Carpenter, J.M., Klacsmann, K.T., Rosenshein, N.B., and Tyrer, H.W. (1981). Herpesvirus antigens as markers for cervical cancer. Gynecol. Oncol. *12*, S232–S258. 10.1016/0090-8258(81)90077-9.

268. Frenkel, N., Roizman, B., Cassai, E., and Nahmias, A. (1972). A DNA Fragment of Herpes Simplex 2 and Its Transcription in Human Cervical Cancer Tissue. Proc. Natl. Acad. Sci. U. S. A. *69*, 3784–3789.

269. Vonka, V., Kanaka, J., Hirsch, I., Závadová, H., Krčmář, M., Suchánková, A., Řezácová, D., Brouček, J., Press, M., Domorázková, E., et al. (1984). Prospective study on the relationship between cervical neoplasia and herpes

simplex type-2 virus. II. Herpes simplex type-2 antibody presence in sera taken at enrolment. Int. J. Cancer *33*, 61–66. 10.1002/ijc.2910330111.

270. Dürst, M., Gissmann, L., Ikenberg, H., and zur Hausen, H. (1983). A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc. Natl. Acad. Sci. *80*, 3812–3815. 10.1073/pnas.80.12.3812.

271. Boshart, M., Gissmann, L., Ikenberg, H., Kleinheinz, A., Scheurlen, W., and zur Hausen, H. (1984). A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. EMBO J. *3*, 1151–1157. 10.1002/j.1460-2075.1984.tb01944.x.

272. Schwarz, E., Freese, U.K., Gissmann, L., Mayer, W., Roggenbuck, B., Stremlau, A., and Hausen, H. zur (1985). Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. Nature *314*, 111–114. 10.1038/314111a0.

273. Barbosa, M.S., and Schlegel, R. (1989). The E6 and E7 genes of HPV-18 are sufficient for inducing two-stage in vitro transformation of human keratinocytes. Oncogene *4*, 1529–1532.

274. Münger, K., Phelps, W.C., Bubb, V., Howley, P.M., and Schlegel, R. (1989). The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. J. Virol. *63*, 4417–4421.

275. Kondoh, G., Murata, Y., Aozasa, K., Yutsudo, M., and Hakura, A. (1991). Very high incidence of germ cell tumorigenesis (seminomagenesis) in human papillomavirus type 16 transgenic mice. J. Virol. *65*, 3335–3339. 10.1128/jvi.65.6.3335-3339.1991.

276. Arbeit, J.M., Münger, K., Howley, P.M., and Hanahan, D. (1993). Neuroepithelial carcinomas in mice transgenic with human papillomavirus type 16 E6/E7 ORFs. Am. J. Pathol. *142*, 1187–1197.

277. Lambert, P.F., Pan, H., Pitot, H.C., Liem, A., Jackson, M., and Griep, A.E. (1993). Epidermal cancer associated with expression of human papillomavirus type 16 E6 and E7 oncogenes in the skin of transgenic mice. Proc. Natl. Acad. Sci. *90*, 5583–5587. 10.1073/pnas.90.12.5583.

278. Arbeit, J.M., Münger, K., Howley, P.M., and Hanahan, D. (1994). Progressive squamous epithelial neoplasia in K14-human papillomavirus type 16 transgenic mice. J. Virol. *68*, 4358–4368. 10.1128/jvi.68.7.4358-4368.1994.

279. Auewarakul, P., Gissmann, L., and Cid-Arregui, A. (1994). Targeted Expression of the E6 and E7 Oncogenes of Human Papillomavirus Type 16 in the Epidermis of Transgenic Mice Elicits Generalized Epidermal Hyperplasia Involving Autocrine Factors. Mol. Cell. Biol. *14*, 8250–8258. 10.1128/mcb.14.12.8250-8258.1994.

280. Greenhalgh, D.A., Wang, X.J., Rothnagel, J.A., Eckhardt, J.N., Quintanilla, M.I., Barber, J.L., Bundman, D.S., Longley, M.A., Schlegel, R., and Roop, D.R. (1994). Transgenic mice expressing targeted HPV-18 E6 and E7 oncogenes in

the epidermis develop verrucous lesions and spontaneous, rasHa-activated papillomas. Cell Growth Differ. Mol. Biol. J. Am. Assoc. Cancer Res. *5*, 667–675.

281. Sasagawa, T., Kondoh, G., Inoue, M., Yutsudo, M., and Hakura, A. (1994). Cervical/vaginal dysplasias of Transgenic Mice Harbouring Human Papillomavirus Type 16 E6-E7 Genes. J. Gen. Virol. *75*, 3057–3065. 10.1099/0022-1317-75-11-3057.

282. Arbeit, J.M., Howley, P.M., and Hanahan, D. (1996). Chronic estrogen-induced cervical and vaginal squamous carcinogenesis in human papillomavirus type 16 transgenic mice. Proc. Natl. Acad. Sci. *93*, 2930–2935. 10.1073/pnas.93.7.2930.

283. Herber, R., Liem, A., Pitot, H., and Lambert, P.F. (1996). Squamous epithelial hyperplasia and carcinoma in mice transgenic for the human papillomavirus type 16 E7 oncogene. J. Virol. *70*, 1873–1881.

284. Munoz, N., Bosch, F.X., de Sanjose, S., Tafur, L., Izarzugaza, I., Gili, M., Viladiu, P., Navarro, C., Martos, C., Ascunce, N., et al. (1992). The causal link between human papillomavirus and invasive cervical cancer: A population-based case-control study in colombia and spain. Int. J. Cancer *52*, 743–749. 10.1002/ijc.2910520513.

285. Bosch, F.X., Muñoz, N., de Sanjosé, S., Izarzugaza, I., Gili, M., Viladiu, P., Tormo, M.J., Moreo, P., Ascunce, N., Gonzalez, L.C., et al. (1992). Risk factors for cervical cancer in Colombia and Spain. Int. J. Cancer *52*, 750–758. 10.1002/ijc.2910520514.

286. Bosch, F.X., Manos, M.M., Muñoz, N., Sherman, M., Jansen, A.M., Peto, J., Schiffman, M.H., Moreno, V., Kurman, R., Shan, K.V., et al. (1995). Prevalence of Human Papillomavirus in Cervical Cancer: a Worldwide Perspective. JNCI J. Natl. Cancer Inst. *87*, 796–802. 10.1093/jnci/87.11.796.

287. Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., and Gallo, R.C. (1980). Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc. Natl. Acad. Sci. *77*, 7415–7419. 10.1073/pnas.77.12.7415.

288. Poiesz, B.J., Ruscetti, F.W., Reitz, M.S., Kalyanaraman, V.S., and Gallo, R.C. (1981). Isolation of a new type C retrovirus (HTLV) in primary uncultured cells of a patient with Sézary T-cell leukaemia. Nature *294*, 268–271. 10.1038/294268a0.

289. Miyoshi, I., Kubonishi, I., Yoshimoto, S., Akagi, T., Ohtsuki, Y., Shiraishi, Y., Nagata, K., and Hinuma, Y. (1981). Type C virus particles in a cord T-cell line derived by co-cultivating normal human cord leukocytes and human leukaemic T cells. Nature *294*, 770–771. 10.1038/294770a0.

290. Hinuma, Y., Nagata, K., Hanaoka, M., Nakai, M., Matsumoto, T., Kinoshita, K.I., Shirakawa, S., and Miyoshi, I. (1981). Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. Proc. Natl. Acad. Sci. *78*, 6476–6480. 10.1073/pnas.78.10.6476.

291. Uchiyama, T., Yodoi, J., Sagawa, K., Takatsuki, K., and Uchino, H. (1977). Adult T-Cell Leukemia: Clinical and Hematologic Features of 16 Cases. Blood *50*, 481–492. 10.1182/blood.V50.3.481.481.

292. Chihara, D., Ito, H., Katanoda, K., Shibata, A., Matsuda, T., Tajima, K., Sobue, T., and Matsuo, K. (2012). Increase in incidence of adult T-cell leukemia/lymphoma in non-endemic areas of Japan and the United States. Cancer Sci. *103*, 1857–1860. 10.1111/j.1349-7006.2012.02373.x.

293. Tolstov, Y.L., Pastrana, D.V., Feng, H., Becker, J.C., Jenkins, F.J., Moschos, S., Chang, Y., Buck, C.B., and Moore, P.S. (2009). Human Merkel cell polyomavirus infection II. MCV is a common human infection that can be detected by conformational capsid epitope immunoassays. Int. J. Cancer *125*, 1250–1256. 10.1002/ijc.24509.

294. Chen, T., Hedman, L., Mattila, P.S., Jartti, T., Ruuskanen, O., Söderlund-Venermo, M., and Hedman, K. (2011). Serological evidence of Merkel cell polyomavirus primary infections in childhood. J. Clin. Virol. *50*, 125–129. 10.1016/j.jcv.2010.10.015.

295. Thompson, M.P., and Kurzrock, R. (2004). Epstein-Barr Virus and Cancer. Clin. Cancer Res. *10*, 803–821. 10.1158/1078-0432.CCR-0670-3.

296. Arai, A. (2019). Advances in the Study of Chronic Active Epstein-Barr Virus Infection: Clinical Features Under the 2016 WHO Classification and Mechanisms of Development. Front. Pediatr. *7*.

297. Soldan, S.S., and Lieberman, P.M. (2023). Epstein–Barr virus and multiple sclerosis. Nat. Rev. Microbiol. *21*, 51–64. 10.1038/s41579-022-00770-5.

298. Ishihara, S., Ohshima, K., Tokura, Y., Yabuta, R., Imaishi, H., Wakiguchi, H., Kurashige, T., Kishimoto, H., Katayama, I., Okada, S., et al. (1997). Hypersensitivity to Mosquito Bites Conceals Clonal Lymphoproliferation of Epstein-Barr Viral DNA-positive Natural Killer Cells. Jpn. J. Cancer Res. *88*, 82–87. 10.1111/j.1349-7006.1997.tb00305.x.

299. Borozan, I., Zapatka, M., Frappier, L., and Ferretti, V. (2018). Analysis of Epstein-Barr Virus Genomes and Expression Profiles in Gastric Adenocarcinoma. J. Virol. *92*, 10.1128/jvi.01239-17. 10.1128/jvi.01239-17.

300. Wong, Y., Meehan, M.T., Burrows, S.R., Doolan, D.L., and Miles, J.J. (2022). Estimating the global burden of Epstein–Barr virus-related cancers. J. Cancer Res. Clin. Oncol. *148*, 31–46. 10.1007/s00432-021-03824-y.

301. Mbulaiteye, S.M., Pullarkat, S.T., Nathwani, B.N., Weiss, L.M., Rao, N., Emmanuel, B., Lynch, C.F., Hernandez, B., Neppalli, V., Hawes, D., et al. (2014). Epstein–Barr virus patterns in US Burkitt lymphoma tumors from the SEER residual tissue repository during 1979–2009. APMIS *122*, 5–15. 10.1111/apm.12078.

302. Hausen, H.Z., and Schulte-Holthausen, H. (1970). Presence of EB Virus Nucleic Acid Homology in a “Virus-free” Line of Burkitt Tumour Cells. Nature *227*, 245–248. 10.1038/227245a0.

303. Niedobitek, G., Agathanggelou, A., Rowe, M., Jones, E., Jones, D., Turyaguma, P., Oryema, J., Wright, D., and Young, L. (1995). Heterogeneous expression of Epstein-Barr virus latent proteins in endemic Burkitt's lymphoma. Blood *86*, 659–665. 10.1182/blood.V86.2.659.bloodjournal862659.

304. Epstein, M.A., Achong, B.G., and Pope, J.H. (1967). Virus in cultured lymphoblasts from a New Guinea Burkitt lymphoma. Br. Med. J. *2*, 290–291.

305. Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nat. Rev. Neurosci. *8*, 427–437. 10.1038/nrn2151.

306. Molyneux, E.M., Rochford, R., Griffin, B., Newton, R., Jackson, G., Menon, G., Harrison, C.J., Israels, T., and Bailey, S. (2012). Burkitt's lymphoma. The Lancet *379*, 1234–1244. 10.1016/S0140-6736(11)61177-X.

307. Weidner-Glunde, M., Kruminis-Kaszkiel, E., and Savanagouder, M. (2020). Herpesviral Latency—Common Themes. Pathogens *9*, 125. 10.3390/pathogens9020125.

308. Kaye, K.M., Izumi, K.M., and Kieff, E. (1993). Epstein-Barr virus latent membrane protein 1 is essential for B-lymphocyte growth transformation. Proc. Natl. Acad. Sci. *90*, 9150–9154. 10.1073/pnas.90.19.9150.

309. Saha, A., and Robertson, E.S. (2011). Epstein-Barr Virus–Associated B-cell Lymphomas: Pathogenesis and Clinical Outcomes. Clin. Cancer Res. *17*, 3056–3063. 10.1158/1078-0432.CCR-10-2578.

310. Quintanilla-Martinez, L., Swerdlow, S.H., Tousseyn, T., Barrionuevo, C., Nakamura, S., and Jaffe, E.S. (2023). New concepts in EBV-associated B, T, and NK cell lymphoproliferative disorders. Virchows Arch. *482*, 227–244. 10.1007/s00428-022-03414-4.

311. Campo, E., Jaffe, E.S., Cook, J.R., Quintanilla-Martinez, L., Swerdlow, S.H., Anderson, K.C., Brousset, P., Cerroni, L., de Leval, L., Dirnhofer, S., et al. (2022). The International Consensus Classification of Mature Lymphoid Neoplasms: a report from the Clinical Advisory Committee. Blood *140*, 1229–1253. 10.1182/blood.2022015851.

312. Alaggio, R., Amador, C., Anagnostopoulos, I., Attygalle, A.D., Araujo, I.B. de O., Berti, E., Bhagat, G., Borges, A.M., Boyer, D., Calaminici, M., et al. (2022). The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia *36*, 1720–1748. 10.1038/s41375-022-01620-2.

313. Lo, Y.M.D., Chan, A.T.C., Chan, L.Y.S., Leung, S.-F., Lam, C.-W., Huang, D.P., and Johnson, P.J. (2000). Molecular Prognostication of Nasopharyngeal Carcinoma by Quantitative Analysis of Circulating Epstein-Barr Virus DNA1. Cancer Res. *60*, 6878–6881.

314. Le, Q.-T., Zhang, Q., Cao, H., Cheng, A.-J., Pinsky, B.A., Hong, R.-L., Chang, J.T., Wang, C.-W., Tsao, K.-C., Lo, Y.D., et al. (2013). An International Collaboration to Harmonize the Quantitative Plasma Epstein-Barr Virus DNA

Assay for Future Biomarker-Guided Trials in Nasopharyngeal Carcinoma. Clin. Cancer Res. *19*, 2208–2215. 10.1158/1078-0432.CCR-12-3702.

315. Mundo, L., Del Porro, L., Granai, M., Siciliano, M.C., Mancini, V., Santi, R., Marcar, L., Vrzalikova, K., Vergoni, F., Di Stefano, G., et al. (2020). Frequent traces of EBV infection in Hodgkin and non-Hodgkin lymphomas classified as EBV-negative by routine methods: expanding the landscape of EBV-related lymphomas. Mod. Pathol. *33*, 2407–2421. 10.1038/s41379-020-0575-3.

316. Lauritzen, A.F., Hørding, U., and Nielsen, H.W. (1994). Epstein-Barr virus and Hodgkin's disease: A comparative immunological, in situ hybridization, and polymerase chain reaction study. APMIS *102*, 495–500. 10.1111/j.1699-0463.1994.tb05196.x.

317. Abusalah, M.A.H., Gan, S.H., Al-Hatamleh, M.A.I., Irekeola, A.A., Shueb, R.H., and Yean Yean, C. (2020). Recent Advances in Diagnostic Approaches for Epstein–Barr Virus. Pathogens *9*, 226. 10.3390/pathogens9030226.

318. Forman, D., de Martel, C., Lacey, C.J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., Vignat, J., Ferlay, J., Bray, F., Plummer, M., et al. (2012). Global Burden of Human Papillomavirus and Related Diseases. Vaccine *30*, F12–F23. 10.1016/j.vaccine.2012.07.055.

319. Tanzi, E., Bianchi, S., Fappani, C., Gori, M., Colzani, D., Passera, I., Tincati, C., Canuti, M., Raviglione, M., and Amendola, A. (2023). Detection of human papillomavirus in fresh and dried urine through an automated system for cervical cancer screening in low- and middle-income countries. J. Med. Virol. *95*, e28802. 10.1002/jmv.28802.

320. Singh, D., Vignat, J., Lorenzoni, V., Eslahi, M., Ginsburg, O., Lauby-Secretan, B., Arbyn, M., Basu, P., Bray, F., and Vaccarella, S. (2023). Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. Lancet Glob. Health *11*, e197–e206. 10.1016/S2214-109X(22)00501-0.

321. Chaturvedi, A.K., Engels, E.A., Pfeiffer, R.M., Hernandez, B.Y., Xiao, W., Kim, E., Jiang, B., Goodman, M.T., Sibug-Saber, M., Cozen, W., et al. (2011). Human Papillomavirus and Rising Oropharyngeal Cancer Incidence in the United States. J. Clin. Oncol. *29*, 4294–4301. 10.1200/JCO.2011.36.4596.

322. Mork, J., Lie, A.K., Glattre, E., Clark, S., Hallmans, G., Jellum, E., Koskela, P., Møller, B., Pukkala, E., Schiller, J.T., et al. (2001). Human Papillomavirus Infection as a Risk Factor for Squamous-Cell Carcinoma of the Head and Neck. N. Engl. J. Med. *344*, 1125–1131. 10.1056/NEJM200104123441503.

323. Bernard, H.-U. (2005). The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. J. Clin. Virol. *32*, 1–6. 10.1016/j.jcv.2004.10.021.

324. Lowy, D.R., and Schiller, J.T. (2012). Reducing HPV-Associated Cancer Globally. Cancer Prev. Res. (Phila. Pa.) *5*, 18–23. 10.1158/1940-6207.CAPR-11-0542.

325. Khan, M.J., Castle, P.E., Lorincz, A.T., Wacholder, S., Sherman, M., Scott, D.R., Rush, B.B., Glass, A.G., and Schiffman, M. (2005). The Elevated 10-Year Risk of Cervical Precancer and Cancer in Women With Human Papillomavirus (HPV) Type 16 or 18 and the Possible Utility of Type-Specific HPV Testing in Clinical Practice. JNCI J. Natl. Cancer Inst. *97*, 1072–1079. 10.1093/jnci/dji187.

326. Ljubojevic, S., and Skerlev, M. (2014). HPV-associated diseases. Clin. Dermatol. *32*, 227–234. 10.1016/j.clindermatol.2013.08.007.

327. Ball, S.L.R., Winder, D.M., Vaughan, K., Hanna, N., Levy, J., Sterling, J.C., Stanley, M.A., and Goon, P.K.C. (2011). Analyses of human papillomavirus genotypes and viral loads in anogenital warts. J. Med. Virol. *83*, 1345–1350. 10.1002/jmv.22111.

328. Munger, K., and Jones, D.L. (2015). Human Papillomavirus Carcinogenesis: an Identity Crisis in the Retinoblastoma Tumor Suppressor Pathway. J. Virol. *89*, 4708–4711. 10.1128/jvi.03486-14.

329. Moody, C.A., and Laimins, L.A. (2010). Human papillomavirus oncoproteins: pathways to transformation. Nat. Rev. Cancer *10*, 550–560. 10.1038/nrc2886.

330. Werness, B.A., Levine, A.J., and Howley, P.M. (1990). Association of Human Papillomavirus Types 16 and 18 E6 Proteins with p53. Science *248*, 76–79. 10.1126/science.2157286.

331. Scheffner, M., Werness, B.A., Huibregtse, J.M., Levine, A.J., and Howley, P.M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. Cell *63*, 1129–1136. 10.1016/0092-8674(90)90409-8.

332. McLaughlin-Drubin, M.E., Crum, C.P., and Münger, K. (2011). Human papillomavirus E7 oncoprotein induces KDM6A and KDM6B histone demethylase expression and causes epigenetic reprogramming. Proc. Natl. Acad. Sci. *108*, 2130–2135. 10.1073/pnas.1009933108.

333. Duffy, C.L., Phillips, S.L., and Klingelhutz, A.J. (2003). Microarray analysis identifies differentiation-associated genes regulated by human papillomavirus type 16 E6. Virology *314*, 196–205. 10.1016/S0042-6822(03)00390-8.

334. Hellner, K., Mar, J., Fang, F., Quackenbush, J., and Münger, K. (2009). HPV16 E7 oncogene expression in normal human epithelial cells causes molecular changes indicative of an epithelial to mesenchymal transition. Virology *391*, 57–63. 10.1016/j.virol.2009.05.036.

335. Mesri, E.A., Feitelson, M.A., and Munger, K. (2014). Human Viral Oncogenesis: A Cancer Hallmarks Analysis. Cell Host Microbe *15*, 266–282. 10.1016/j.chom.2014.02.011.

336. de Villiers, E.-M., Fauquet, C., Broker, T.R., Bernard, H.-U., and zur Hausen, H. (2004). Classification of papillomaviruses. Virology *324*, 17–27. 10.1016/j.virol.2004.03.033.

337. Poljak, M., Oštrbenk Valenčak, A., Gimpelj Domjanič, G., Xu, L., and Arbyn, M. (2020). Commercially available molecular tests for human papillomaviruses: a

global overview. Clin. Microbiol. Infect. *26*, 1144–1150. 10.1016/j.cmi.2020.03.033.

338. Volesky-Avellaneda, K.D., Laurie, C., Tsyruk-Romano, O., El-Zein, M., and Franco, E.L. (2023). Human Papillomavirus Detectability and Cervical Cancer Prognosis: A Systematic Review and Meta-analysis. Obstet. Gynecol. *142*, 1055. 10.1097/AOG.0000000000005370.

339. Arroyo Mühr, L.S., Lagheden, C., Lei, J., Eklund, C., Nordqvist Kleppe, S., Sparén, P., Sundström, K., and Dillner, J. (2020). Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR. Br. J. Cancer *123*, 1790–1795. 10.1038/s41416-020-01111-0.

340. Santos, F.L.S.G., Invenção, M.C.V., Araújo, E.D., Barros, G.S., and Batista, M.V.A. (2021). Comparative analysis of different PCR-based strategies for HPV detection and genotyping from cervical samples. J. Med. Virol. *93*, 6347–6354. 10.1002/jmv.27118.

341. Walboomers, J.M.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J.F., Peto, J., Meijer, C.J.L.M., and Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J. Pathol. *189*, 12–19. 10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F.

342. Walboomers, J.M.M., and Meijer, C.J.L.M. (1997). Do HPV-negative cervical carcinomas exist? J. Pathol. *181*, 253–254. 10.1002/(SICI)1096-9896(199703)181:3<253::AID-PATH755>3.0.CO;2-0.

343. Gillison, M.L., Chaturvedi, A.K., Anderson, W.F., and Fakhry, C. (2015). Epidemiology of Human Papillomavirus–Positive Head and Neck Squamous Cell Carcinoma. J. Clin. Oncol. *33*, 3235–3242. 10.1200/JCO.2015.61.6995.

344. Fakhry, C., Lacchetti, C., Rooper, L.M., Jordan, R.C., Rischin, D., Sturgis, E.M., Bell, D., Lingen, M.W., Harichand-Herdt, S., Thibo, J., et al. (2018). Human Papillomavirus Testing in Head and Neck Carcinomas: ASCO Clinical Practice Guideline Endorsement of the College of American Pathologists Guideline. J. Clin. Oncol. *36*, 3152–3161. 10.1200/JCO.18.00684.

345. Jordan, R.C., Lingen, M.W., Perez-Ordonez, B., He, X., Pickard, R., Koluder, M., Jiang, B., Wakely, P., Xiao, W., and Gillison, M.L. (2012). Validation of methods for oropharyngeal cancer HPV status determination in United States cooperative group trials. Am. J. Surg. Pathol. *36*, 945–954. 10.1097/PAS.0b013e318253a2d1.

346. Tsai, W.-L., and Chung, R.T. (2010). Viral hepatocarcinogenesis. Oncogene *29*, 2309–2324. 10.1038/onc.2010.36.

347. Lindenbach, B.D., and Rice, C.M. (2005). Unravelling hepatitis C virus replication from genome to function. Nature *436*, 933–938. 10.1038/nature04077.

348. El-Serag, H.B. (2002). Hepatocellular carcinoma and hepatitis C in the United States. Hepatology *36*, s74–s83. 10.1053/jhep.2002.36807.

349. Chen, C.-J., and Yang, H.-I. (2011). Natural history of chronic hepatitis B REVEALed. J. Gastroenterol. Hepatol. *26*, 628–638. 10.1111/j.1440-1746.2011.06695.x.

350. Iloeje, U.H., Yang, H., Su, J., Jen, C., You, S., and Chen, C. (2006). Predicting Cirrhosis Risk Based on the Level of Circulating Hepatitis B Viral Load. Gastroenterology *130*, 678–686. 10.1053/j.gastro.2005.11.016.

351. Chen, C.-J., Yang, H.-I., Su, J., Jen, C.-L., You, S.-L., Lu, S.-N., Huang, G.-T., Iloeje, U.H., and REVEAL-HBV Study Group, for the (2006). Risk of Hepatocellular Carcinoma Across a Biological Gradient of Serum Hepatitis B Virus DNA Level. JAMA *295*, 65–73. 10.1001/jama.295.1.65.

352. Yang, T., Lu, J.-H., Zhai, J., Lin, C., Yang, G.-S., Zhao, R.-H., Shen, F., and Wu, M.-C. (2012). High viral load is associated with poor overall and recurrence-free survival of hepatitis B virus-related hepatocellular carcinoma after curative resection: A prospective cohort study. Eur. J. Surg. Oncol. EJSO *38*, 683–691. 10.1016/j.ejso.2012.04.010.

353. Zamor, P.J., deLemos, A.S., and Russo, M.W. (2017). Viral hepatitis and hepatocellular carcinoma: etiology and management. J. Gastrointest. Oncol. *8*. 10.21037/jgo.2017.03.14.

354. Feng, H., Shuda, M., Chang, Y., and Moore, P.S. (2008). Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma. Science *319*, 1096–1100. 10.1126/science.1152586.

355. Moshiri, A.S., Doumani, R., Yelistratova, L., Blom, A., Lachance, K., Shinohara, M.M., Delaney, M., Chang, O., McArdle, S., Thomas, H., et al. (2017). Polyomavirus-Negative Merkel Cell Carcinoma: A More Aggressive Subtype Based on Analysis of 282 Cases Using Multimodal Tumor Virus Detection. J. Invest. Dermatol. *137*, 819–827. 10.1016/j.jid.2016.10.028.

356. Pastrana, D.V., Tolstov, Y.L., Becker, J.C., Moore, P.S., Chang, Y., and Buck, C.B. (2009). Quantitation of Human Seroresponsiveness to Merkel Cell Polyomavirus. PLOS Pathog. *5*, e1000578. 10.1371/journal.ppat.1000578.

357. Schrama, D., Peitsch, W.K., Zapatka, M., Kneitz, H., Houben, R., Eib, S., Haferkamp, S., Moore, P.S., Shuda, M., Thompson, J.F., et al. (2011). Merkel Cell Polyomavirus Status Is Not Associated with Clinical Course of Merkel Cell Carcinoma. J. Invest. Dermatol. *131*, 1631–1638. 10.1038/jid.2011.115.

358. Tajima, K., The T- and B-Cell Malignancy Study Group, and Co-Authors (1990). The 4th nation-wide study of adult T-cell leukemia/lymphoma (ATL) in Japan: Estimates of risk of ATL and its geographical and clinical features. Int. J. Cancer *45*, 237–243. 10.1002/ijc.2910450206.

359. Coffin, J.M. (2015). The discovery of HTLV-1, the first pathogenic human retrovirus. Proc. Natl. Acad. Sci. *112*, 15525–15529. 10.1073/pnas.1521629112.

360. Vega, F., Miranda, R.N., and Medeiros, L.J. (2020). KSHV/HHV8-positive large B-cell lymphomas and associated diseases: a heterogeneous group of

lymphoproliferative processes with significant clinicopathological overlap. Mod. Pathol. *33*, 18–28. 10.1038/s41379-019-0365-y.

361. Dupin, N., Diss, T.L., Kellam, P., Tulliez, M., Du, M.-Q., Sicard, D., Weiss, R.A., Isaacson, P.G., and Boshoff, C. (2000). HHV-8 is associated with a plasmablastic variant of Castleman disease that is linked to HHV-8–positive plasmablastic lymphoma. Blood *95*, 1406–1412. 10.1182/blood.V95.4.1406.004k26_1406_1412.

362. Nador, R., Cesarman, E., Chadburn, A., Dawson, D., Ansari, M., Sald, J., and Knowles, D. (1996). Primary effusion lymphoma: a distinct clinicopathologic entity associated with the Kaposi's sarcoma-associated herpes virus. Blood *88*, 645–656. 10.1182/blood.V88.2.645.bloodjournal882645.

363. Hengge, U.R., Ruzicka, T., Tyring, S.K., Stuschke, M., Roggendorf, M., Schwartz, R.A., and Seeber, S. (2002). Update on Kaposi's sarcoma and other HHV8 associated diseases. Part 2: pathogenesis, Castleman's disease, and pleural effusion lymphoma. Lancet Infect. Dis. *2*, 344–352. 10.1016/S1473-3099(02)00288-8.

364. Schwartz, R.A. (2004). Kaposi's sarcoma: An update. J. Surg. Oncol. *87*, 146–151. 10.1002/jso.20090.

365. Voisset, C., Weiss, R.A., and Griffiths, D.J. (2008). Human RNA "Rumor" Viruses: the Search for Novel Human Retroviruses in Chronic Disease. Microbiol. Mol. Biol. Rev. *72*, 157–196. 10.1128/mmbr.00033-07.

366. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120. 10.1038/ng.2764.

367. Cantalupo, P.G., Katz, J.P., and Pipas, J.M. (2018). Viral sequences in human cancer. Virology *513*, 208–216. 10.1016/j.virol.2017.10.017.

368. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M., and Larsson, E. (2013). The landscape of viral expression and host gene fusion and adaptation in human cancer. Nat. Commun. *4*, 2513. 10.1038/ncomms3513.

369. Salyakina, D., and Tsinoremas, N.F. (2013). Viral expression associated with gastrointestinal adenocarcinomas in TCGA high-throughput sequencing data. Hum. Genomics *7*, 23. 10.1186/1479-7364-7-23.

370. Cantalupo, P.G., Katz, J.P., and Pipas, J.M. (2015). HeLa Nucleic Acid Contamination in The Cancer Genome Atlas Leads to the Misidentification of Human Papillomavirus 18. J. Virol. *89*, 4051–4057. 10.1128/jvi.03365-14.

371. Kazemian, M., Ren, M., Lin, J.-X., Liao, W., Spolski, R., and Leonard, W.J. (2015). Possible Human Papillomavirus 38 Contamination of Endometrial Cancer RNA Sequencing Samples in The Cancer Genome Atlas Database. J. Virol. *89*, 8967–8973. 10.1128/jvi.00822-15.

372. Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., et al. (2020). Microbiome

analyses of blood and tissues suggest cancer diagnostic approach. Nature *579*, 567–574. 10.1038/s41586-020-2095-1.

373. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., Brewer, D.S., Pertea, M., and Salzberg, S.L. (2023). Major data analysis errors invalidate cancer microbiome findings. Preprint at bioRxiv, 10.1101/2023.07.28.550993 10.1101/2023.07.28.550993.

374. Breitwieser, F.P., Pertea, M., Zimin, A.V., and Salzberg, S.L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. *29*, 954–960. 10.1101/gr.245373.118.

375. Steinegger, M., and Salzberg, S.L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. *21*, 115. 10.1186/s13059-020-02023-1.

376. Urisman, A., Molinaro, R.J., Fischer, N., Plummer, S.J., Casey, G., Klein, E.A., Malathi, K., Magi-Galluzzi, C., Tubbs, R.R., Ganem, D., et al. (2006). Identification of a Novel Gammaretrovirus in Prostate Tumors of Patients Homozygous for R462Q RNASEL Variant. PLOS Pathog. *2*, e25. 10.1371/journal.ppat.0020025.

377. Fischer, N., Hellwinkel, O., Schulz, C., Chun, F.K.H., Huland, H., Aepfelbacher, M., and Schlomm, T. (2008). Prevalence of human gammaretrovirus XMRV in sporadic prostate cancer. J. Clin. Virol. *43*, 277–283. 10.1016/j.jcv.2008.04.016.

378. Danielson, B.P., Ayala, G.E., and Kimata, J.T. (2010). Detection of Xenotropic Murine Leukemia Virus-Related Virus in Normal and Tumor Tissue of Patients from the Southern United States with Prostate Cancer Is Dependent on Specific Polymerase Chain Reaction Conditions. J. Infect. Dis. *202*, 1470–1477. 10.1086/656146.

379. Switzer, W.M., Jia, H., Zheng, H., Tang, S., and Heneine, W. (2011). No Association of Xenotropic Murine Leukemia Virus-Related Viruses with Prostate Cancer. PLOS ONE *6*, e19065. 10.1371/journal.pone.0019065.

380. Verhaegh, G.W., de Jong, A.S., Smit, F.P., Jannink, S.A., Melchers, W.J.G., and Schalken, J.A. (2011). Prevalence of human xenotropic murine leukemia virus-related gammaretrovirus (XMRV) in dutch prostate cancer patients. The Prostate *71*, 415–420. 10.1002/pros.21255.

381. Schlaberg, R., Choe, D.J., Brown, K.R., Thaker, H.M., and Singh, I.R. (2009). XMRV is present in malignant prostatic epithelium and is associated with prostate cancer, especially high-grade tumors. Proc. Natl. Acad. Sci. *106*, 16351–16356. 10.1073/pnas.0906922106.

382. Hohn, O., Krause, H., Barbarotto, P., Niederstadt, L., Beimforde, N., Denner, J., Miller, K., Kurth, R., and Bannert, N. (2009). Lack of evidence for xenotropic murine leukemia virus-related virus(XMRV) in German prostate cancer patients. Retrovirology *6*, 92. 10.1186/1742-4690-6-92.

383. Furuta, R.A., Miyazawa, T., Sugiyama, T., Kuratsune, H., Ikeda, Y., Sato, E., Misawa, N., Nakatomi, Y., Sakuma, R., Yasui, K., et al. (2011). No association

of xenotropic murine leukemia virus-related virus with prostate cancer or chronic fatigue syndrome in Japan. Retrovirology *8*, 20. 10.1186/1742-4690-8-20.

384. Stieler, K., Schindler, S., Schlomm, T., Hohn, O., Bannert, N., Simon, R., Minner, S., Schindler, M., and Fischer, N. (2011). No Detection of XMRV in Blood Samples and Tissue Sections from Prostate Cancer Patients in Northern Europe. PLOS ONE *6*, e25592. 10.1371/journal.pone.0025592.

385. Oakes, B., Tai, A.K., Cingöz, O., Henefield, M.H., Levine, S., Coffin, J.M., and Huber, B.T. (2010). Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. Retrovirology *7*, 109. 10.1186/1742-4690-7-109.

386. Sato, E., Furuta, R.A., and Miyazawa, T. (2010). An Endogenous Murine Leukemia Viral Genome Contaminant in a Commercial RT-PCR Kit is Amplified Using Standard Primers for XMRV. Retrovirology *7*, 110. 10.1186/1742-4690-7-110.

387. Groom, H.C.T., Warren, A.Y., Neal, D.E., and Bishop, K.N. (2012). No Evidence for Infection of UK Prostate Cancer Patients with XMRV, BK Virus, Trichomonas vaginalis or Human Papilloma Viruses. PLOS ONE *7*, e34221. 10.1371/journal.pone.0034221.

388. Robinson, M.J., Tuke, P.W., Erlwein, O., Tettmar, K.I., Kaye, S., Naresh, K.N., Patel, A., Walker, M.M., Kimura, T., Gopalakrishnan, G., et al. (2011). No Evidence of XMRV or MuLV Sequences in Prostate Cancer, Diffuse Large B-Cell Lymphoma, or the UK Blood Donor Population. Adv. Virol. *2011*, e782353. 10.1155/2011/782353.

389. Sakuma, T., Hué, S., Squillace, K.A., Tonne, J.M., Blackburn, P.R., Ohmine, S., Thatava, T., Towers, G.J., and Ikeda, Y. (2011). No evidence of XMRV in prostate cancer cohorts in the Midwestern United States. Retrovirology *8*, 23. 10.1186/1742-4690-8-23.

390. Aloia, A.L., Sfanos, K.S., Isaacs, W.B., Zheng, Q., Maldarelli, F., De Marzo, A.M., and Rein, A. (2010). XMRV: A New Virus in Prostate Cancer? Cancer Res. *70*, 10028–10033. 10.1158/0008-5472.CAN-10-2837.

391. Lee, D., Gupta, J.D., Gaughan, C., Steffen, I., Tang, N., Luk, K.-C., Qiu, X., Urisman, A., Fischer, N., Molinaro, R., et al. (2012). In-Depth Investigation of Archival and Prospectively Collected Samples Reveals No Evidence for XMRV Infection in Prostate Cancer. PLOS ONE *7*, e44954. 10.1371/journal.pone.0044954.

392. Knouf, E.C., Metzger, M.J., Mitchell, P.S., Arroyo, J.D., Chevillet, J.R., Tewari, M., and Miller, A.D. (2009). Multiple Integrated Copies and High-Level Production of the Human Retrovirus XMRV (Xenotropic Murine Leukemia Virus-Related Virus) from 22Rv1 Prostate Carcinoma Cells. J. Virol. *83*, 7353–7356. 10.1128/jvi.00546-09.

393. Rose, R., Constantinides, B., Tapinos, A., Robertson, D.L., and Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. Virus Evol. *2*, vew022. 10.1093/ve/vew022.

394. Tan, C.C.S., Ko, K.K.K., Chen, H., Liu, J., Loh, M., Chia, M., and Nagarajan, N. (2023). No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. Nat. Microbiol. *8*, 973–985. 10.1038/s41564-023-01350-w.

395. Kennedy, K.M., de Goffau, M.C., Perez-Muñoz, M.E., Arrieta, M.-C., Bäckhed, F., Bork, P., Braun, T., Bushman, F.D., Dore, J., de Vos, W.M., et al. (2023). Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. Nature *613*, 639–649. 10.1038/s41586-022-05546-8.

396. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., and Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. *12*, 87. 10.1186/s12915-014-0087-z.

397. de Goffau, M.C., Lager, S., Sovio, U., Gaccioli, F., Cook, E., Peacock, S.J., Parkhill, J., Charnock-Jones, D.S., and Smith, G.C.S. (2019). Human placenta has no microbiome but can contain potential pathogens. Nature *572*, 329–334. 10.1038/s41586-019-1451-5.

398. Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., Sun, Q., Kaur, R., Huyen, Y., and McBride, A.A. (2017). The Papillomavirus Episteme: a major update to the papillomavirus sequence database. Nucleic Acids Res. *45*, D499–D506. 10.1093/nar/gkw879.

399. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

400. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv, 10.48550/arXiv.1303.3997 10.48550/arXiv.1303.3997.

401. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

402. Kline, C.N., Joseph, N.M., Grenert, J.P., van Ziffle, J., Talevich, E., Onodera, C., Aboian, M., Cha, S., Raleigh, D.R., Braunstein, S., et al. (2017). Targeted next-generation sequencing of pediatric neuro-oncology patients improves diagnosis, identifies pathogenic germline mutations, and directs targeted therapy. Neuro-Oncol. *19*, 699–709. 10.1093/neuonc/now254.

403. Sirohi, D., Vaske, C., Sanborn, Z., Smith, S.C., Don, M.D., Lindsey, K.G., Federman, S., Vankalakunti, M., Koo, J., Bose, S., et al. (2018). Polyoma virus-associated carcinomas of the urologic tract: a clinicopathologic and molecular study. Mod. Pathol. *31*, 1429–1441. 10.1038/s41379-018-0065-z.

404. The Regents of the University of California UCSF 500 Cancer Gene Panel Test (UCSF500 / UC500). UCSF Health Cent. Clin. Genet. Genomics. https://genomics.ucsf.edu/UCSF500.

405. Deng, X., Achari, A., Federman, S., Yu, G., Somasekar, S., Bártolo, I., Yagi, S., Mbala-Kingebeni, P., Kapetshi, J., Ahuka-Mundeke, S., et al. (2020). Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat. Microbiol. *5*, 443–454. 10.1038/s41564-019-0637-9.

406. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. 10.1016/S0022-2836(05)80360-2.

407. Shimoyama, Y. (2022). pyCirclize: Circular visualization in Python.

408. Nakata, K., Colombet, M., Stiller, C.A., Pritchard-Jones, K., Steliarova-Foucher, E., and Contributors, I.-3 (2020). Incidence of childhood renal tumours: An international population-based study. Int. J. Cancer *147*, 3313–3327. 10.1002/ijc.33147.

409. van Peer, S.E., Hol, J.A., van der Steeg, A.F.W., Grotel, M. van, Tytgat, G.A.M., Mavinkurve-Groothuis, A.M.C., Janssens, G.O.R., Littooij, A.S., de Krijger, R.R., Jongmans, M.C.J., et al. (2021). Bilateral Renal Tumors in Children: The First 5 Years' Experience of National Centralization in The Netherlands and a Narrative Review of the Literature. J. Clin. Med. *10*, 5558. 10.3390/jcm10235558.

410. Choufani, S., Shuman, C., and Weksberg, R. (2010). Beckwith–Wiedemann syndrome. Am. J. Med. Genet. C Semin. Med. Genet. *154C*, 343–354. 10.1002/ajmg.c.30267.

411. Irving, I.M. (1967). Exomphalos with macroglossia: A study of eleven cases. J. Pediatr. Surg. *2*, 499–507. 10.1016/S0022-3468(67)80003-4.

412. Sotelo-Avila, C., Gonzalez-Crussi, F., and Fowler, J.W. (1980). Complete and incomplete forms of Beckwith-Wiedemann syndrome: Their oncogenic potential. J. Pediatr. *96*, 47–50. 10.1016/S0022-3476(80)80322-2.

413. Lenstrup, E., and Monrad, P.S. (1926). Eight Cases of Hemi-Hypertrophy. Acta Paediatr. *6*, 205–213. 10.1111/j.1651-2227.1926.tb09347.x.

414. RIEDEL, H.A. (1952). ADRENOGENITAL SYNDROME IN A MALE CHILD DUE TO ADRENOCORTICAL TUMOR : Report of Case with Hemihypertrophy and Subsequent Development of Embryoma (Wilms' Tumor). Pediatrics *10*, 19–27. 10.1542/peds.10.1.19.

415. Benson, P.F., Vulliamy, D.G., and Taubman, J.O. (1963). CONGENITAL HEMIHYPERTROPHY AND MALIGNANCY. The Lancet *281*, 468–469. 10.1016/S0140-6736(63)92360-2.

416. Björklund, S.-I. (1955). Hemihypertrophy and Wilms's Tumour. Acta Paediatr. *44*, 287–292. 10.1111/j.1651-2227.1955.tb04141.x.

417. Wiedemann, H.-R. (1983). Tumours and hemihypertrophy associated with Wiedemann-Beckwith syndrome. Eur. J. Pediatr. *141*, 129–129. 10.1007/BF00496807.

418. Maher, E.R., and Reik, W. (2000). Beckwith-Wiedemann syndrome: imprinting in clusters revisited. J. Clin. Invest. *105*, 247–252. 10.1172/JCI9340.

419. Weksberg, R., Ren Shen, D., Ling Fei, Y., Li Song, Q., and Squire, J. (1993). Disruption of insulin–like growth factor 2 imprinting in Beckwith–Wiedemann syndrome. Nat. Genet. *5*, 143–150. 10.1038/ng1093-143.

420. Reik, W., Brown, K.W., Schneid, H., Le Bouc, Y., Bickmore, W., and Maher, E.R. (1995). Imprinting mutations in the Beckwith—Wiedemann syndrome suggested by an altered imprinting pattern in the IGF2–H19 domain. Hum. Mol. Genet. *4*, 2379–2385. 10.1093/hmg/4.12.2379.

421. Nordin, M., Bergman, D., Halje, M., Engström, W., and Ward, A. (2014). Epigenetic regulation of the Igf2/H19 gene cluster. Cell Prolif. *47*, 189–199. 10.1111/cpr.12106.

422. Thorvaldsen, J.L., Fedoriw, A.M., Nguyen, S., and Bartolomei, M.S. (2006). Developmental Profile of H19 Differentially Methylated Domain (DMD) Deletion Alleles Reveals Multiple Roles of the DMD in Regulating Allelic Expression and DNA Methylation at the Imprinted H19/Igf2 Locus. Mol. Cell. Biol. *26*, 1245–1258. 10.1128/MCB.26.4.1245-1258.2006.

423. Martin, R.A., Grange, D.K., Zehnbauer, B., and DeBaun, M.R. (2005). LIT1 and H19 methylation defects in isolated hemihyperplasia. Am. J. Med. Genet. A. *134A*, 129–131. 10.1002/ajmg.a.30578.

424. Fiala, E.M., Ortiz, M.V., Kennedy, J.A., Glodzik, D., Fleischut, M.H., Duffy, K.A., Hathaway, E.R., Heaton, T., Gerstle, J.T., Steinherz, P., et al. (2020). 11p15.5 epimutations in children with Wilms tumor and hepatoblastoma detected in peripheral blood. Cancer *126*, 3114–3121. 10.1002/cncr.32907.

425. Hol, J.A., Kuiper, R.P., van Dijk, F., Waanders, E., van Peer, S.E., Koudijs, M.J., Bladergroen, R., van Reijmersdal, S.V., Morgado, L.M., Bliek, J., et al. (2022). Prevalence of (Epi)genetic Predisposing Factors in a 5-Year Unselected National Wilms Tumor Cohort: A Comprehensive Clinical and Genomic Characterization. J. Clin. Oncol. *40*, 1892–1902. 10.1200/JCO.21.02510.

426. Murphy, A.J., Cheng, C., Williams, J., Shaw, T.I., Pinto, E.M., Dieseldorff-Jones, K., Brzezinski, J., Renfro, L.A., Tornwall, B., Huff, V., et al. (2023). Genetic and epigenetic features of bilateral Wilms tumor predisposition in patients from the Children's Oncology Group AREN18B5-Q. Nat. Commun. *14*, 8006. 10.1038/s41467-023-43730-0.

427. Coster, W.D., Rijk, P.D., Roeck, A.D., Pooter, T.D., D'Hert, S., Strazisar, M., Sleegers, K., and Broeckhoven, C.V. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. *29*, 1178–1187. 10.1101/gr.244939.118.

428. Minervini, C.F., Cumbo, C., Orsini, P., Anelli, L., Zagaria, A., Specchia, G., and Albano, F. (2020). Nanopore Sequencing in Blood Diseases: A Wide Range of Opportunities. Front. Genet. *11*.

429. Deest, M., Brändl, B., Rohrandt, C., Eberlein, C., Bleich, S., Müller, F.-J., and Frieling, H. (2022). Long-read nanopore sequencing reveals novel common

genetic structural variants in Prader-Willi syndrome and associated psychosis. Preprint at medRxiv, 10.1101/2022.07.18.22277235 10.1101/2022.07.18.22277235.

430. Du, Z.-F., Li, P.-F., Zhao, J.-Q., Cao, Z.-L., Li, F., Ma, J.-M., and Qi, X.-P. (2017). Genetic diagnosis of a Chinese multiple endocrine neoplasia type 2A family through whole genome sequencing. J. Biosci. *42*, 209–218. 10.1007/s12038-017-9686-5.

431. Alisch, R.S., Barwick, B.G., Chopra, P., Myrick, L.K., Satten, G.A., Conneely, K.N., and Warren, S.T. (2012). Age-associated DNA methylation in pediatric populations. Genome Res. *22*, 623–632. 10.1101/gr.125187.111.

432. Yousefi, P.D., Suderman, M., Langdon, R., Whitehurst, O., Davey Smith, G., and Relton, C.L. (2022). DNA methylation-based predictors of health: applications and statistical considerations. Nat. Rev. Genet. *23*, 369–383. 10.1038/s41576-022-00465-w.

433. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol. Cell *49*, 359–367. 10.1016/j.molcel.2012.10.016.

434. Do, W.L., Whitsel, E.A., Costeira, R., Masachs, O.M., Le Roy, C.I., Bell, J.T., Staimez, L.R., Stein, A.D., Smith, A.K., Horvath, S., et al. (2021). Epigenome-wide association study of diet quality in the Women's Health Initiative and TwinsUK cohort. Int. J. Epidemiol. *50*, 675–684. 10.1093/ije/dyaa215.

435. Crimmins, E.M., Thyagarajan, B., Levine, M.E., Weir, D.R., and Faul, J. (2021). Associations of Age, Sex, Race/Ethnicity, and Education With 13 Epigenetic Clocks in a Nationally Representative U.S. Sample: The Health and Retirement Study. J. Gerontol. Ser. A *76*, 1117–1123. 10.1093/gerona/glab016.

436. Bocklandt, S., Lin, W., Sehl, M.E., Sánchez, F.J., Sinsheimer, J.S., Horvath, S., and Vilain, E. (2011). Epigenetic Predictor of Age. PLOS ONE *6*, e14821. 10.1371/journal.pone.0014821.

437. Mandaviya, P.R., Joehanes, R., Brody, J., Castillo-Fernandez, J.E., Dekkers, K.F., Do, A.N., Graff, M., Hänninen, I.K., Tanaka, T., de Jonge, E.A., et al. (2019). Association of dietary folate and vitamin B-12 intake with genome-wide DNA methylation in blood: a large-scale epigenome-wide association analysis in 5841 individuals. Am. J. Clin. Nutr. *110*, 437–450. 10.1093/ajcn/nqz031.

438. Gensous, N., Garagnani, P., Santoro, A., Giuliani, C., Ostan, R., Fabbri, C., Milazzo, M., Gentilini, D., di Blasio, A.M., Pietruszka, B., et al. (2020). One-year Mediterranean diet promotes epigenetic rejuvenation with country- and sex-specific effects: a pilot study from the NU-AGE project. GeroScience *42*, 687–701. 10.1007/s11357-019-00149-0.

439. Ma, J., Rebholz, C.M., Braun, K.V.E., Reynolds, L.M., Aslibekyan, S., Xia, R., Biligowda, N.G., Huan, T., Liu, C., Mendelson, M.M., et al. (2020). Whole Blood DNA Methylation Signatures of Diet Are Associated With Cardiovascular Disease Risk Factors and All-Cause Mortality. Circ. Genomic Precis. Med. *13*, e002766. 10.1161/CIRCGEN.119.002766.

440. Boks, M.P., Mierlo, H.C. van, Rutten, B.P.F., Radstake, T.R.D.J., De Witte, L., Geuze, E., Horvath, S., Schalkwyk, L.C., Vinkers, C.H., Broen, J.C.A., et al. (2015). Longitudinal changes of telomere length and epigenetic age related to traumatic stress and post-traumatic stress disorder. Psychoneuroendocrinology *51*, 506–512. 10.1016/j.psyneuen.2014.07.011.

441. Zannas, A.S., Arloth, J., Carrillo-Roa, T., Iurato, S., Röh, S., Ressler, K.J., Nemeroff, C.B., Smith, A.K., Bradley, B., Heim, C., et al. (2015). Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. Genome Biol. *16*, 266. 10.1186/s13059-015-0828-5.

442. Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. *15*, R31. 10.1186/gb-2014-15-2-r31.

443. Ramos-Lopez, O., Riezu-Boj, J.I., Milagro, F.I., Zulet, M.A., Santos, J.L., Martinez, J.A., Alonso, A., Arancibia, C., Arós, F., Astrup, A., et al. (2019). Associations between olfactory pathway gene methylation marks, obesity features and dietary intakes. Genes Nutr. *14*, 11. 10.1186/s12263-019-0635-9.

444. Zhang, J., Walsh, M.F., Wu, G., Edmonson, M.N., Gruber, T.A., Easton, J., Hedges, D., Ma, X., Zhou, X., Yergeau, D.A., et al. (2015). Germline Mutations in Predisposition Genes in Pediatric Cancer. N. Engl. J. Med. *373*, 2336–2346. 10.1056/NEJMoa1508054.

445. Parsons, D.W., Roy, A., Yang, Y., Wang, T., Scollon, S., Bergstrom, K., Kerstein, R.A., Gutierrez, S., Petersen, A.K., Bavle, A., et al. (2016). Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. JAMA Oncol. *2*, 616–624. 10.1001/jamaoncol.2015.5699.

446. Killian, J.K., Miettinen, M., Walker, R.L., Wang, Y., Zhu, Y.J., Waterfall, J.J., Noyes, N., Retnakumar, P., Yang, Z., Smith, W.I., et al. (2014). Recurrent epimutation of SDHC in gastrointestinal stromal tumors. Sci. Transl. Med. *6*, 268ra177-268ra177. 10.1126/scitranslmed.3009961.

447. Baker, S.W., Duffy, K.A., Richards-Yutz, J., Deardorff, M.A., Kalish, J.M., and Ganguly, A. (2021). Improved molecular detection of mosaicism in Beckwith-Wiedemann Syndrome. J. Med. Genet. *58*, 178–184. 10.1136/jmedgenet-2019-106498.

448. Lee, B.H., Kim, G.-H., Oh, T.J., Kim, J.H., Lee, J.-J., Choi, S.H., Lee, J.Y., Kim, J.-M., Choi, I.H., Kim, Y.-M., et al. (2013). Quantitative analysis of methylation status at 11p15 and 7q21 for the genetic diagnosis of Beckwith–Wiedemann syndrome and Silver–Russell syndrome. J. Hum. Genet. *58*, 604–610. 10.1038/jhg.2013.67.

449. van Veghel-Plandsoen, M.M., Wouters, C.H., Kromosoeto, J.N.R., den Ridder-Klünnen, M.C., Halley, D.J.J., and van den Ouweland, A.M.W. (2011). Multiplex ligation-depending probe amplification is not suitable for detection of low-grade mosaicism. Eur. J. Hum. Genet. *19*, 1009–1012. 10.1038/ejhg.2011.60.

450. Kalish, J.M., Conlin, L.K., Mostoufi-Moab, S., Wilkens, A.B., Mulchandani, S., Zelley, K., Kowalski, M., Bhatti, T.R., Russo, P., Mattei, P., et al. (2013).

Bilateral Pheochromocytomas, Hemihyperplasia, and Subtle Somatic Mosaicism: The Importance of Detecting Low-Level Uniparental Disomy. Am. J. Med. Genet. A. *161*, 993–1001. 10.1002/ajmg.a.35831.

451. Turner, J.T., Hill, D.A., and Dome, J.S. (2022). Revisiting the Threshold for Cancer Genetics Referral in Patients With Wilms Tumor. J. Clin. Oncol. *40*, 1853–1860. 10.1200/JCO.22.00411.

452. Coffee, B., Muralidharan, K., Highsmith, W.E., Lapunzina, P., and Warren, S.T. (2006). Molecular diagnosis of Beckwith-Wiedemann Syndrome using quantitative methylation-sensitive polymerase chain reaction. Genet. Med. *8*, 628–634. 10.1097/01.gim.0000237770.42442.cc.

453. Russo, S., Calzari, L., Mussa, A., Mainini, E., Cassina, M., Di Candia, S., Clementi, M., Guzzetti, S., Tabano, S., Miozzo, M., et al. (2016). A multi-method approach to the molecular diagnosis of overt and borderline 11p15.5 defects underlying Silver–Russell and Beckwith–Wiedemann syndromes. Clin. Epigenetics *8*, 23. 10.1186/s13148-016-0183-8.