JOHANNES GUTENBERG UNIVERSITÄT MAINZ

RNA-Seq Based Decomposition of Human Cell Lines and Primary Tumors for the Identification and Quantification of Viral Expression

Dissertation zur Erlangung des Grades
Doktor der Naturwissenschaften

am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Thomas Bukur
geb. 15.08.1977 in Frankenthal/Pfalz

Mainz, 2017

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung:

Zusammenfassung

Etwa zwanzig Prozent aller Krebserkrankungen werden durch Infektionserreger verursacht, dabei machen Viren den größten Teil aus. Viren zu detektieren kann sehr arbeits- oder zeitaufwändig sein, außerdem benötigt man genaues Vorwissen. Die Analyse humaner Proben mittels moderner Hochdurchsatzsequenzierung von Transkriptomen (RNA-Seq) ermöglicht unvoreingenommen die Sequenzierung aller beinhalteter mRNA-Sequenzen einer Probe, einschließlich viraler mRNA. Untersuchungen viraler Expressionsprofile und deren Einfluss auf den Wirtsorganismus sind ein wesentlicher Bestandteil um ein Verständnis für Virus-induzierte Entstehung von Krebs im Menschen zu erhalten. Die vorgelegte Arbeit repräsentiert einen Schritt in Richtung der Systemvirologie in der Krebsimmunologie.

Zunächst wurde die Software VirusID entwickelt um Viren in Säugetierzellen zu entdecken und zu identifizieren. VirusID wurde in Leistungsbewertungen überprüft und war bei der Detektion den anderen Methoden geringfügig überlegen. VirusID wurde weiter zur Untersuchung von internen und externen RNA-Seq-Daten im TRON eingesetzt. VirusID lieferte auch die identifizierten Viren in jeder der untersuchten 1.082 Zelllinien für das TRON Cell Line Portal.

Die Identifikation viraler Profile in RNA-Seq-Alignments führte zur Entwicklung von VIRGENE, einer Softwarepipeline um virale Genexpressionen zu berechnen. VIRGENE wurde auf 186 Zelllinienproben angewandt. Hier konnten bekannte Viren wie etwa Papillomviren oder Herpesviren bestätigt werden. In anderen Fällen wurden bislang unbekannte Viren entdeckt, wie etwa Mäuseretroviren oder das Bovine Polyomavirus. VIRGENE wurde dann auf Daten von Primärtumoren angewandt und offenbarte unterschiedliche Expressionsprofile des Epstein-Barr-Virus in Zelllinien verglichen mit Tumorproben. Die sogenannten „BamHI-A rightward transcripts“ (BARTs) sind jedoch in fast allen Proben gleichbleibend exprimiert. Ausserdem konnten im Falle der humanen Papillomviren unterschiedliche Expressionsmuster in Zervixkarzinomen identifiziert werden. Die Patientinnen unterschiedlicher Gruppen zeigen auch unterschiedliche Überlebenschancen. Das diagnostische und prognostische Potenzial dieser Biomarker wird noch ausgewertet werden müssen.

Diese Arbeit zeigt das Potenzial der Untersuchung bestehender RNA-Seq-Datensätze auf enthaltene Viren oder exprimierte virale Gene auf. Verschiedene Virusexpressionsmuster haben einen unterschiedlichen Einfluss auf die Genexpression des Wirtes und möglicherweise auch auf den Krankheitsverlauf. Die Ergebnisse von VIRGENE können umfangreiche Untersuchungen des Zusammenspiels von Wirts- und Erregertranskriptom ermöglichen. Die Verknüpfung mit klinischen Patientendaten führt zu einer Aufwertung der Ergebnisse in prognostischer, diagnostischer oder therapeutischer Hinsicht.

Summary

Around twenty percent of all cancer cases are contributed to infectious agents, mostly to viruses. However, detection of viruses can be laborious, time-consuming, or require foreknowledge. Next generation transcriptome sequencing (RNA-Seq) of human samples is unbiased in the way that all included mRNAs can be sequenced, including foreign mRNA like viral transcripts. The analysis of viral expression profiles and their influence on the host is fundamental to understanding virus-associated oncogenesis in human. The hereby-presented study represents a step towards systems virology in cancer immunology.

First, the software tool VirusID was developed for the qualified detection and identification of viruses in mammalian cells. VirusID was tested in benchmarks with known viral content and was slightly superior over other tools. It has since been in use for the identification of viral content in in-house and external sequencing RNA-Seq data at TRON and contributed to the TRON Cell Line Portal by delivering the identified viruses 1,082 cell lines.

Subsequently, identified virus alignment profiles in RNA-Seq data encouraged the development of VIRGENE, a pipeline to retrieve viral gene expression levels. We applied VIRGENE to 186 cell line samples and confirmed known viruses like papillomaviruses or herpesviruses. In other cell lines, yet unknown virus content was identified like traces of murine retroviruses or bovine polyomaviruses. VIRGENE was then applied to primary tumors and revealed distinct expression profiles for Epstein-Barr virus between cell lines and tumor samples. However, BamHI-A rightward transcripts (BARTs) are consistently expressed. Furthermore, different Human papillomavirus expression profiles were identified in cervical cancer, which were associated with distinct overall survival of the respective patients. The diagnostic and prognostic potential of these biomarkers will have to be further assessed.

This work shows the potential of screening existing RNA-Seq data for viruses and expressed viral genes. Diverse virus expression profiles have different effects on host gene expression and possibly also to disease outcome. The results produced by VIRGENE facilitate comprehensive studies of host and pathogen transcriptomic interplay. Prognostic, diagnostic, or therapeutic value is added by coupling the results with clinical annotation of cancer patients.

Acknowledgements

I conducted the hereby-presented research at the Faculty of Biology and the Medical Center of the Johannes Gutenberg-University Mainz, and at TRON – Translational Oncology at the University Medical Center of the Johannes Gutenberg-University Mainz gGmbH. It is a great pleasure to express my gratitude to all those who made this thesis possible.

…

I would like to thank the German ministry of education and research (Bundesministerium für Bildung und Forschung, BMBF) for funding parts of this study, in particular the VirusID project of Ci3 (FKZ 131A033).

Parts of this research were conducted using the supercomputer Mogon at the Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP and the Gauss Alliance e.V. I gratefully acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz.

I also acknowledge the Servier Medical Art database. Figures 1-3, 1-5, and 1-6 were created using provided components (http://www.servier.com/Powerpoint-image-bank). These are published under creative commons license 3.0 (https://creativecommons.org/licenses/by/3.0/). Some components were modified for the use in this thesis.

Contents

Zusammenfassung iii
Summary iv
Acknowledgements v
Contents vi
List of Figures viii
List of Tables ix
List of Abbreviations x
1 Introduction 1
1.1 Virus-induced Cancers 1
1.2 Viral Replication 3
1.3 Infection, Latency, and Virus-induced Tumorigenesis 5
1.4 Methods for the Detection of Viruses 8
1.5 Next Generation Sequencing 13
1.6 Aim of this Thesis 15
2 Materials and Methods 16
2.1 Data Sets and Cohorts 16
2.1.1 Samples for First Performance Review of VirusID 16
2.1.2 Samples for Final Benchmark of VirusID 19
2.1.3 Standard RNA-Seq 22
2.1.4 Cell Line Data 23
2.1.5 Cancer Cohort Data 24
2.1.6 Other Third Party Data 25
2.2 Databases 25
2.3 Third Party Software 28
2.4 Servers and Hardware 29
2.5 Data Analysis 30
2.5.1 Pre-processing of Data 30
2.5.2 Mapping 30
2.5.3 Gene Expression 31
2.5.4 VirusID 31
2.5.5 VIRGENE 34
2.5.6 Downstream Processing and Analyses 36
3 Results 38
3.1 VirusID: NGS-based Identification of Viruses 38
3.1.1 The Bioinformatics Platform VirusID 38
3.1.2 First Performance Review 43
3.1.3 Final Benchmark 49
3.2 Routinely Screening of RNA-Seq Data 54
3.3 Viral Abundance and Transcription Profile 57
3.4 VIRGENE: Virus Gene Expression in Transcriptomic Data of the Host 60
3.5 Viral Expression Patterns in Cell Lines 64
3.6 EBV Expression in Cell Lines and Primary Tumors 68
3.7 HPV Expression in Tumors 71
4 Discussion 75
4.1 *In Silico* Virus Detection 75
4.2 Performance of VirusID and VIRGENE 76
4.3 Detected Viruses 78
4.3.1 Viruses in Cell Lines 78
4.3.2 Re-analysis of EBL Samples 83

4.3.3 The Association of HCMV with GBM 87
4.4 HPV in Tumor Cohorts 87
4.5 Challenges and Potential 92
4.5.1 The Virus Reference 92
4.5.2 Incorporated Algorithms 93
4.5.3 Non-polyadenylated RNA 94
4.5.4 Viral Biomarkers 94
5 Conclusion 96
6 Literature 97
Appendix 111
Declaration of Authorship 135

List of Figures

Figure 1-1: Baltimore scheme of viral classification. 3
Figure 1-2: Progress of acute or persistent viral infections. 6
Figure 1-3: Common cellular targets for viral oncoproteins. 7
Figure 1-4: Scheme for enzyme-linked immunosorbent assay (ELISA). 10
Figure 1-5: Simplified workflow of a microarray experiment. 12
Figure 1-6: Simplified scheme of sequencing by synthesis. 14
Figure 2-1: Bioanalyzer results of virus spike-in samples. 18
Figure 2-2: Description of the utilized cell line cohort. 24
Figure 2-3: Host group distribution among the RefSeq virus database. 26
Figure 2-4: Workflow of data preparation for VIRGENE. 27
Figure 2-5: Workflow of the VIRGENE pipeline. 35
Figure 3-1: Alignment statistics of SARS mouse samples. 42
Figure 3-2: Bioanalyzer results for samples 12A and 14A. 44
Figure 3-3: Viral abundance and transcription profiles in cell lines. 59
Figure 3-4: Calculation of the entropy. 63
Figure 3-5: Viral abundance and transcription profiles in cell lines. 67
Figure 3-6: IGV view of alignments to Bovine polyomavirus. 68
Figure 3-7: LF3 and BART alignment profiles in eight cell lines. 69
Figure 3-8: Distinct expression profiles of EBV in cell lines and primary tumors. 70
Figure 3-9: HPV expression in HPV-positive HNSC and CESC patients. 72
Figure 3-10: Number of human-HPV chimeric sequence reads. 73
Figure 4-1: Expression profiles of EMCV and AbMLV in cell lines. 82
Figure 4-2: Human herpesvirus 1 expression in one EBL sample. 85
Figure 4-3: Alignment profiles of EBL viruses HTLV-1, HCMV, and KSHV. 86
Figure 4-4: HNSC cohort dissected by HPV status. 88
Figure 4-5: Overall survival in HNSC and CESC by HPV status. 90
Figure 4-6: IGV alignment plot on HPV16 of four exemplary CESC datasets. 90
Figure 4-7: Survival in HPV-positive CESC patients. 91

List of Tables

Table 1-1: Human oncogenic viral agents. 2
Table 2-1: Samples for the first performance review of VirusID 17
Table 2-2: Selected samples for the final benchmark of VirusID 20
Table 2-3: Quality check of cDNA and sequencing library. 22
Table 2-4: Used software and tools. 29
Table 3-1: Alignment statistics of human samples on virus genomes. 40
Table 3-2: Baxter samples and detected viruses using VirusID. 45
Table 3-3: Evaluation of four different broad virus detection methods. 48
Table 3-4: VirusID results for the nasal aspirate samples 50
Table 3-5: Results of the software benchmark for the detection of viruses. 53
Table 3-6: Sources and virus content of breast cancer cell line data. 56
Table 3-7: Human viruses in the TRON Cell Line Portal (TCLP). 57
Table 4-1: Comparison of detected viruses by Pandora and VIRGENE. 84
Table A-1: List of Cell Line Data Accession Numbers. 111
Table A-2: Accession numbers of EBL samples (project SRP062178). 115
Table A-3: HLA Alleles and expression determined by Seq2HLA. 116
Table A-4: Human viruses in 186 cell lines. 117
Table A-5: Non-human mammalian viruses in 186 cell lines 121
Table A-6: Bacteriophages in 186 cell lines. 126

List of Abbreviations

BMBF ..... *German Ministry of Education and Research*
bp ..... *base pair(s)*
CCLE ..... *Cancer Cell Line Encyclopedia*
cDNA ..... *complementary DNA*
CDS ..... *coding sequence*
CESC ..... *Cervical Squamous Cell Carcinoma*
CHO ..... *Chinese hamster ovarian cell line*
Ci3 ..... *Cluster for Individualized ImmuneIntervention*
CTL ..... *cytotoxic T lymphocyte*
DLBC ..... *Diffuse Large B Cell Lymphoma*
dsDNA ..... *double-stranded DNA*
EBV ..... *Epstein-Barr virus (HHV4)*
ELISA ..... *enzyme linked immunosorbent assay*
EM ..... *electron microscopy*
EMCV ..... *Encephalomyocarditis virus*
FVR ..... *fraction of virus reads [ppm]*
GBM ..... *Glioblastoma Multiforme*
GEO ..... *Gene Expression Omnibus*
HAdV-C ..... *Human Adenovirus C*
HBV ..... *Hepatitis B virus*
HCMV ..... *Human Cytomegalovirus (HHV5)*
HCoV-OC43 ..... *Human coronavirus OC43*
HCV ..... *Hepatitis C virus*
HERV ..... *Human endogenous retrovirus*
HHV1 ..... *Human herpesvirus 1 (HSV-1)*
HHV4 ..... *Human herpesvirus 4 (EBV)*
HHV5 ..... *Human herpesvirus 5 (HCMV)*
HHV8 ..... *Human herpesvirus 8 (KSHV)*
HIV-1 ..... *Human immunodeficiency virus 1*
HNSC ..... *Head and Neck Squamous Cell Carcinoma*
HPC ..... *high performance computing*
HPV ..... *Human papillomavirus*
HSV-1 ..... *Herpes simplex virus (HHV1)*
HTLV-1 ..... *Human T-lymphotropic virus 1*
i.e. ..... *id est (Latin for 'that is to say')*
IF ..... *immunofluorescence*
IGV ..... *Integrative Genomics Viewer*
ISH ..... *in situ hybridization*
ISVP ..... *intermediate sub-viral particle*
JGU ..... *Johannes Gutenberg-University Mainz*
KSHV ..... *Kaposi's sarcoma-associated herpesvirus (HHV8)*
MCRV ..... *Murine type C retrovirus*
MHC ..... *major histocompatibility complex*
mRNA ..... *messenger RNA*

mRNA-Seq ..... *high throughput sequencing of mRNA*
MuLV ..... *Murine leukemia virus*
MVM ..... *Minute virus of mice*
NA ..... *not applicable, not available*
NGS ..... *next-generation sequencing*
NMF ..... *Non-negative matrix factorization*
nt ..... *nucleotide(s)*
ORF ..... *open reading frame*
P53 ..... *Tumor Protein P53*
PCR ..... *polymerase chain reaction*
phiX174 ..... *Enterobacteria phage phiX174*
ppm ..... *parts per million*
qPCR ..... *quantitative (real-time) PCR*
RAM ..... *random access memory*
RB1 ..... *RB Transcriptional Corepressor 1*
RdRp ..... *RNA-dependent RNA polymerase*
RefSeq ..... *NCBI Reference Sequence Database*
ReoIII ..... *Reovirus type III*
RNA-Seq ..... *high throughput RNA sequencing*
RPKM ..... *reads per kilobase of exons per million mapped reads*
RSV ..... *Rous sarcoma virus*
RT ..... reverse transcriptase
SARS ..... *severe acute respiratory syndrome*
SARS-CoV ..... *severe acute respiratory syndrome coronavirus*
SCKV-M1 ..... *Saccharomyces cerevisiae killer virus M1*
ScV-M1 ..... *Saccharomyces cerevisiae killer virus M1*
SRA ..... *Sequence Read Archive*
SSDRV ..... *Sclerotinia sclerotiorum debilitation-associated RNA virus*
ssRNA ..... *single-stranded RNA*
STAD ..... *Stomach Adenocarcinoma*
STAR ..... *Spliced Transcripts Alignment to a Reference*
TB ..... *terabytes (trillion bytes)*
TCGA ..... *The Cancer Genome Atlas*
TCLP ..... *TRON Cell Line Portal*
TRON ..... *Translational Oncology at the Medical Center of the JGU Mainz gGmbH*
UCSC ..... *University of California Santa Cruz*
VERO ..... *African green monkey kidney cell line*
VGC ..... *viral genome coverage [%]*
X-MuLV ..... *Xenotropic murine leukemia virus*
ZDV ..... *Data Center of the JGU*

1 Introduction

"Tumors destroy man in a unique and appalling way, as flesh of his own flesh which has somehow been rendered proliferative, rampant, predatory, and ungovernable." This was the opening sentence of Peyton Rous's Nobel Lecture 1966. He received the honors 55 years after having been the first to scientifically report a transplantable avian tumor [1]. As it later turned out, the described neoplasms were associated with what we know today as the Rous sarcoma virus (RSV). This set the start point for further studies and discoveries regarding virus-associated tumorigenesis.

1.1 Virus-induced Cancers

Viruses are everywhere. The virosphere spans every environment populated by cellular life. For example, the abundance of viruses in the oceans exceeds that of prokaryotes by 15-fold [2]. A human being is exposed to millions of viruses in the environment, on the skin, and even inside the body, at any given time. Yet, human cells are only susceptible and permissive to a fraction of viruses, allowing infection and viral replication; and only a fraction of those viruses will cause recognizable infections or diseases. The human immune system can manage most viral infections. Very rarely will a virus lead to long-term infection, chronic disease, or even cancer. However, about 20% of all cancers are related to infectious agents with the majority being viruses [3].

Right now, we know of eight human viruses that have oncogenic potential [4, 5] (Table 1-1). The identified cancer associated viruses span a variety of viral classes including RNA viruses like the Hepatitis C virus, as well as DNA viruses, like Epstein-Barr virus. Almost all of these viruses have closely related virus species that are not known to be oncogenic. Hence, the class of oncogenic viruses cannot be defined by their molecular structure. Even the known suspects are necessary but not sufficient for driving emergence of cancer. Most humans infected with any of these viruses will never develop cancer; and if they do, it might be years after the initial infection. Further environmental co-factors are required [6].

Table 1-1: Human oncogenic viral agents. Table adapted from [5]. dsDNA: double-stranded DNA, ssRNA: single-stranded RNA.

Viral agent	Virus genome structure	Associated cancers
Epstein-Barr virus (EBV), also known as Human herpesvirus 4 (HHV4)	dsDNA herpesvirus	Nasopharyngeal carcinoma, Burkitt's lymphoma, some non-Hodgkin's lymphomas, Hodgkin's lymphoma, some gastrointestinal lymphomas, extranodal NK/T-cell lymphoma
Merkel cell polyomavirus (MCV)	dsDNA polyomavirus	Merkel cell carcinoma
Hepatitis B virus (HBV)	Partially double-stranded DNA hepadenovirus	Hepatocellular carcinoma
Hepatitis C virus (HCV)	Positive-strand ssRNA flavivirus	Hepatocellular carcinoma, some non-Hodgkin's lymphomas [7]
Kaposi's sarcoma herpes virus (KSHV), also known as Human herpesvirus 8 (HHV8)	dsDNA herpesvirus	Kaposi's sarcoma, primary effusion lymphoma, some multicentric Castleman's diseases
High-risk human papillomavirus (HPV) types	dsDNA papillomavirus	Carcinomas of the cervix, vulva, vagina, penis, anus, oral cavity, oropharynx, tonsil, head and neck cancers
Human T-lymphotropic virus type 1 (HTLV-1)	Positive-strand ssRNA retrovirus	Adult T cell leukemia and lymphoma
Human immunodeficiency virus type 1 (HIV-1)	Positive-strand ssRNA retrovirus	Immunosuppression promotes different types of cancers through other viruses

Viruses can be involved in oncogenesis through several mechanisms on genomic or proteomic levels. Firstly, some viruses express oncogenes that stimulate, for example, cellular proliferation. Secondly, viruses might express genes or proteins that interfere with tumor suppressor genes, thus enabling oncogenesis by suppressing the suppressor. Thirdly, the genome of some virus might be inserted into the host genome causing insertional mutagenesis and thus altering either oncogenes or tumor suppressors. These examples can be summarized as direct carcinogens and are usually present in every cancer cell. The Human immunodeficiency virus 1 (HIV-1) is not a direct carcinogen. It promotes neoplastic transformation through immunosuppression and consequent co-infection with Kaposi's sarcoma herpes virus (KSHV) [8]. Additionally, a virus induced chronic inflammation can lead to mutations that might

induce oncogenesis over the long term. This category can be described as indirect carcinogens and might include ‘hit-and-run’ virus infections; a process that is yet to be well described [4].

1.2 Viral Replication

Viruses infect organisms to replicate. The infection of a cellular organism is necessary for viruses to utilize the replicative mechanisms as viruses rely on the host ribosome to produce viral proteins from their DNA or RNA templates. Hence, the production of host readable mRNA is necessary for all viruses [9].

As viruses are obligate intracellular parasites, their infectious cycle start with the entry of a cell. Entry strategies rely on binding to cellular receptors or other attachment factors. Binding generally initiates uptake of the virion by the target cell through different mechanisms. After penetration of the cell, virus genomes have to be uncoated for replication and transcription. Most DNA viruses release their genome into the nucleus for transcription, while RNA virus genomes remain in the cytosol. Then, viral replication and transcription are initiated. The molecular structures of virus genomes are diverse (Figure 1-1) and thus replication and transcription follow different paths.

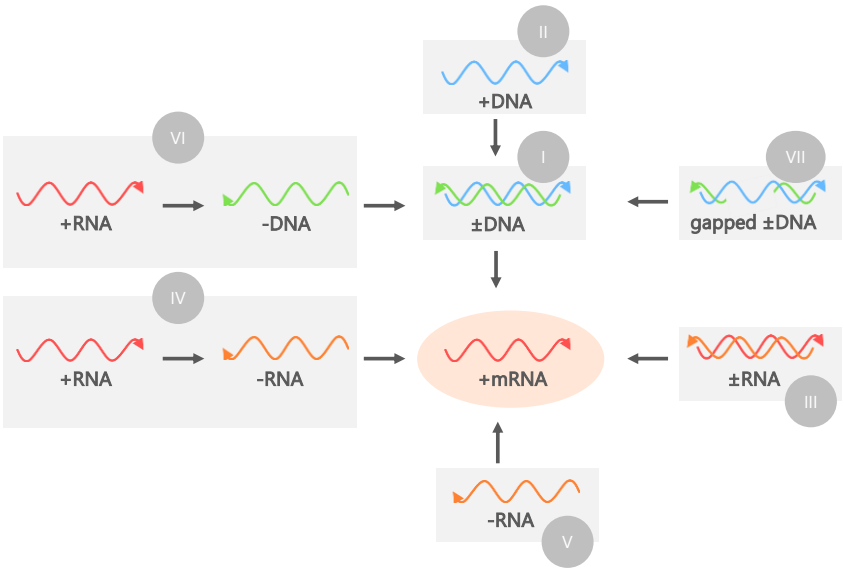


Figure 1-1: Baltimore scheme of viral classification. Adapted and extended from the original publication by David Baltimore (1971) [9]. Class I: dsDNA viruses; class II positive-strand single-stranded DNA viruses; class III: double-stranded RNA viruses; class IV: positive-strand ssRNA viruses, possibly with an RNA intermediate; class V: negative-strand ssRNA viruses; class VI: positive-strand ssRNA viruses with a DNA intermediate; class VII: gapped dsDNA virus.

In the end, all viruses have to produce mRNA that can be read by the host ribosome. The Baltimore scheme [9] is still in use today to classify viruses by their genomic features. dsDNA viruses like adenoviruses, papillomaviruses, or herpesviruses are grouped in class I. In the case of Human papillomavirus 16 (HPV16), the DNA will form episomes in the nucleus that are constantly replicated. The transcription is strongly regulated in concordance with cellular differentiation. However, only one strand of the DNA is transcriptionally active and transcription only occurs in one direction [3].

Class II contains positive-strand DNA viruses, for example parvoviruses. Their genome features palindromic termini, which are able to fold back and thus function as a primer for DNA polymerases to create the negative strand DNA intermediate. This intermediate template serves for transcription or replication with the synthesis of positive strand DNA genome [10].

Classes III, IV, and V cover RNA viruses, double-stranded, positive strand, and negative strand single-stranded RNA (ssRNA) viruses, respectively. The nature of such viral genomes necessitates an RNA-dependent RNA polymerase (RdRp) for replication that was first described in Poliovirus [11]. The viral genome contains the code for this enzyme, as animal genomes do not seem to contain it. Negative-stranded and double-stranded RNA viruses contain the RdPd readily in their virion. Here, uncoating activates the polymerase to produce the respective mRNAs. In general, the genomes of positive stranded RNA viruses function as mRNA, and translation is started immediately upon infection.

Class VI contains retroviruses. They feature a positive strand RNA genome and pass through a DNA intermediate for replication and transcription. This intermediate can persist over even long phases of viral latency by integration into the host genome (chapter 1.3). The RdRp is also packaged into the virion for replication. However, they also carry a reverse transcriptase (RT) [12, 13], an RNA-dependent DNA polymerase that also features ribonuclease and DNA-dependent DNA polymerase activities. The DNA intermediate is then transcribed into mRNA for synthesis of the proteins.

The youngest class of viruses, class VII, describes the Hepatitis B virus. This virus has a gapped double-stranded DNA genome that is covalently closed to a double-stranded DNA in the nucleus of the host cell [14]. The DNA is then transcribed into mRNAs,

including an RNA pre-genome. This mRNA codes for the viral core protein and the viral reverse transcriptase. The RT is required to reverse-transcribe the DNA of their own mRNA upon completion of their synthesis and packaging of the RT-mRNA complex into nucleocapsids.

The size restrictions of viral genomes often require regulated protein production from polycistronic mRNAs. However, the translational apparatus of eukaryotic cells synthesizes proteins from monocistronic mRNA. Viruses pursue different strategies of this translational control. Many of the synthesized viral proteins are controlled by frameshifting in the case of overlapping open reading frames (ORFs). In other cases, multiple ORFs are regulated via termination and reinitiation events. In influenza, for example, the termination codon of protein M1 overlaps the initiation codon of influenza protein BM2 [15]. Hence, the expression of the BM2 protein requires termination of M1 synthesis. Most large DNA virus mRNAs are monocistronic and do not require strategies of termination-reinitiation. Furthermore, many viruses induce a phenomenon termed 'host-shutoff' by interfering with host mRNA translation, either directly or indirectly. This leads to depleted antiviral responses in the host cell and favored viral protein synthesis [16].

Finally, assembly of virions, maturation, and virus release are not always distinct steps of the final phase of a viral infectious cycle. These mechanisms differ widely according to different viral features and strategies. When appropriate amounts of viral genomic nucleic acids and capsid proteins accumulate, viral self-assembly occurs at specific sites of a cell, depending on the virus. The assembly might include parts of the cell membrane in enveloped viruses. Maturation might include proteolytic cleavage by host or virus proteases from outside or inside the capsid. Naked (non-enveloped) viruses are usually released upon cell lysis. Enveloped viruses are shed via budding through the cellular membrane. Depending on the virus, budding might or might not damage the cell.

1.3 Infection, Latency, and Virus-induced Tumorigenesis

Throughout an infection, viruses express a variety of genes to hijack the host cell for their own needs. However, the host's immune system has developed a variety of strategies to eliminate viruses or viral disease.

Virus infections of a host can be classified into two categories: acute and persistent. Acute infections are often caused by rhinoviruses, rotaviruses, or influenza viruses. Acute infection is the most common form of viral infection and features a rapid rise of viral reproduction accompanied by a disease (Figure 1-2 upper panel). The cytolytic features of these infections destroy the host cells and are thus self-limiting: either the immune system encounters the virus and the host recovers, or the host dies during the progress of the disease. The first case leads to complete immune clearance of the virus from the body.

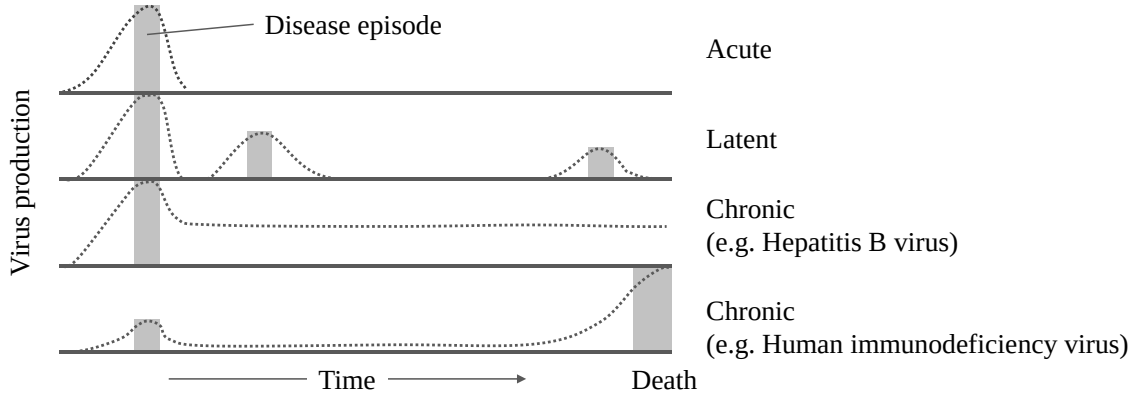


Figure 1-2: Progress of acute or persistent viral infections. Dotted lines show virus production over time; grey bars indicate disease episodes. Figure adapted from [17].

In persistent infections, the virus cannot be cleared completely from the infected individual. The virus usually persists in specific cells, often unrecognized by the immune system. Persistent infections can be roughly grouped into latent and chronic infections [17]. They commonly start with an acute phase (Figure 1-2). Herpesvirus infections, for example, might then result in a latent infection. In latent infections, viral protein production is silenced and the virus remains mostly unrecognized by the immune system followed by periodical reactivation and viral spreading. Chronic infections, on the other hand, are persistent asymptomatic infections. Virus particles are produced over the lifetime of the host. Often cytopathic effects are low and host immune reactions are reduced, despite potential expression of virus antigens on the surface of the infected cell. Other chronic virus infections with different disease and virus shedding profiles are associated with, for example, the Human immunodeficiency virus or the Human T-lymphotropic virus. After a mild acute infection, the infection often stays clinically latent for years with minimal viral

shedding. Reactivation of the virus might be associated with other factors, like drugs [18, 19], for instance .

An additional mechanism that allows persistent virus-host cell interactions is the integrated virus infection. This is frequently observed with retroviruses, but also occurs in other viral infections. At some point of the infectious cycle, the viral genome might get integrated into the host genome in part or completely. The remaining part of the viral genome might never produce complete virions; but often some viral proteins and antigens are still expressed and can lead to activation of immunity.

In order to modulate the cell for their needs, viruses express certain proteins. This includes regulation of the cell cycle [20] or metabolism [21] to exploit host replication machinery or to avoid metabolic exhaustion. Remarkably, tumor viruses show various strategies to target common tumor suppressor pathways (Figure 1-3) [4].

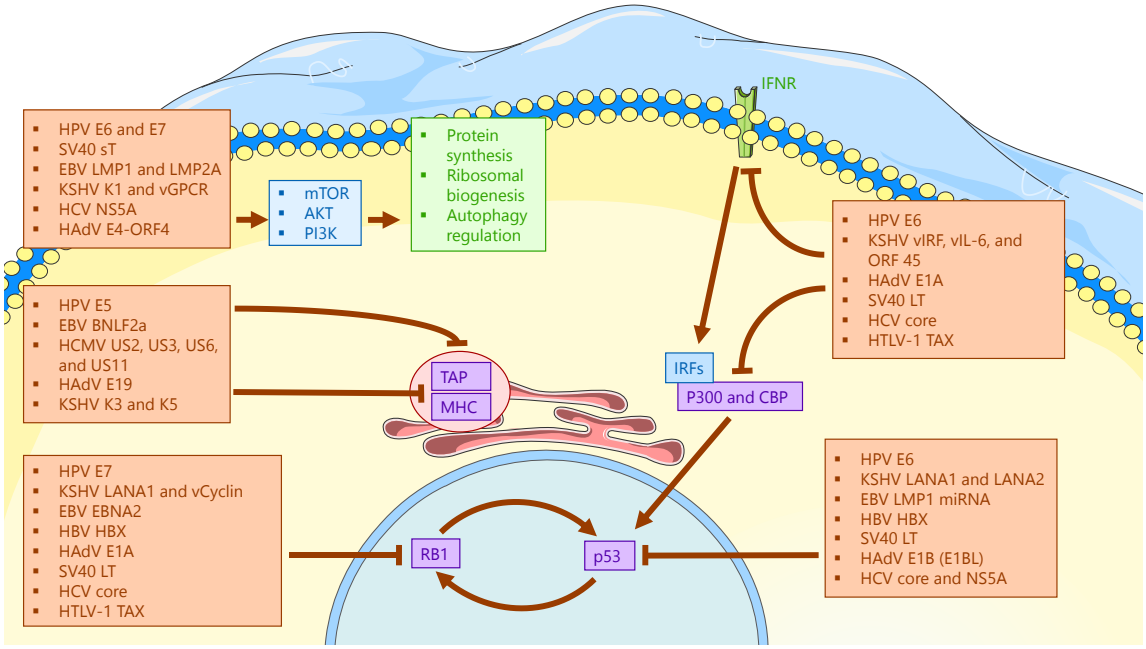


Figure 1-3: Common cellular targets for viral oncoproteins. Figure adapted from [4]. Tumor virus proteins can target RB1, p53, interferon and PI3K-mTOR signaling pathways. Furthermore, viruses have developed strategies to evade MHC class I presentation on the cell surface, including blocking of TAP, degradation of MHC, or retention of MHCI in the endoplasmic reticulum. Most of these viral proteins are evolutionarily unrelated and regulate pathways distinctly. HPV: Human papillomavirus; SV40: Simian virus 40; EBV: Epstein-Barr virus; HCV: Hepatitis C virus; KSHV: Kaposi's sarcoma virus; HAdV: human Adenovirus; HCMV; Human cytomegalovirus; HTLV-1: Human T-lymphotropic virus; HBV: Hepatitis B virus.

Although tumor viruses have developed different mechanisms, they target common tumor suppressors, like the tumor suppressor Tumor Protein P53 (P53) or the RB Transcriptional Corepressor 1 (RB1). The Human Papillomavirus (HPV) for example, expresses the genes E6 and E7, targeting both p53 and RB1. Tumor viruses also feature strategies to silence major histocompatibility complex (MHC) class I presentation. This could either occur through high intracellular mutation rate to modify presented peptides, also known as 'CTL escape mutants' (CTL: cytotoxic T lymphocyte) [22], or they do so by inhibition of the MHC-loading process or the transport to the cell surface [23, 24]. Other virus-host interactions include inhibition of pattern recognition receptors, interferon signaling, modulation of host protein expression or regulation of autophagy or apoptosis. Variations of host genes by insertional mutagenesis might also influence host regulatory pathways, but occur rarely. In summary, viral factors contribute to oncogenesis through the modulation of pathways that are also modulated by genomic alterations in mutation-driven oncogenesis [25].

1.4 Methods for the Detection of Viruses

In 1899, Martinus Beijerinck used ceramic filters to isolate infectious filterable agents [26]. He was not able to cultivate the filtrate and he excluded bacteria as a source. However, the fluidic agents seemed to replicate in plants and he called it a 'virus' (Latin: poison, slime, venom). Later, this agent became one of the most studied viruses, the Tobacco mosaic virus. Since then, numerous methods have been developed to detect and identify viruses. Some of these techniques measure viral infectivity, while others measure viral enzyme activity. Some detect anti-viral antibodies of a past infection, while others detect protein or nucleotide sequences of an existing infection. Some of the methods are suitable to identify viruses, others are used to quantify viral load. This chapter contains an introduction to some of these methods.

Hemagglutination

The HA (hemagglutinin) envelope protein of influenza viruses is able to bind to sialic acids on the surface of their target cells before entry. HA is also able to bind to erythrocytes in blood and form clusters. This process is called hemagglutination. Based on this property, George Hirst developed a rapid assay in the early 1940s to determine influenza virus load in a sample [27]. A variation of this assay, the HA

inhibition assay (HI), involves anti-influenza antibodies that prevents hemagglutination in titrations of serum [28]. This results in the HI titer of the serum at the highest dilution that prevents hemagglutination as a measure for anti-viral protection in epidemiological studies [29].

Plaque assay

Morphological changes of the host cell upon viral infection are called cytopathic effects. These include fusions with adjacent cells, rounding of the cells, or lysis of the infected cells. In 1953, the plaque assay for measuring titers of bacteriophage stocks was modified to determine titers of animal viruses [30]. In the plaque assay, virus titer is determined by inoculating susceptible cells with 10-fold dilutions of virus stock. The spread of the virus across the monolayered cells becomes visible as circular plaques. Then, the titer can be determined calculating plaque-forming units (PFU) per milliliter of the virus stock at given dilutions.

Electron microscopy

The development of electron microscopy (EM) enabled the detection and visualization of tobacco mosaic virus in 1939 [31]. Since then, EM has been used to discover new virus outbreaks [32]. However, restricted access and low throughput limit its use in general screening of viruses.

ELISA

Antibodies are in extensive use for the detection of proteins. They are also widely used for research and diagnosis in virology. One of these methods is the enzyme-linked immunosorbent assay (ELISA) [33, 34] to measure viral protein or anti-viral antibodies. In the first case, anti-viral antibodies are coated on a plastic surface. Then, a clinical specimen is added and the antibodies bind to their specific viral antigen. Antigens can then be detected by a second antibody labelled with a fluorescent indicator enzyme (Figure 1-4a).

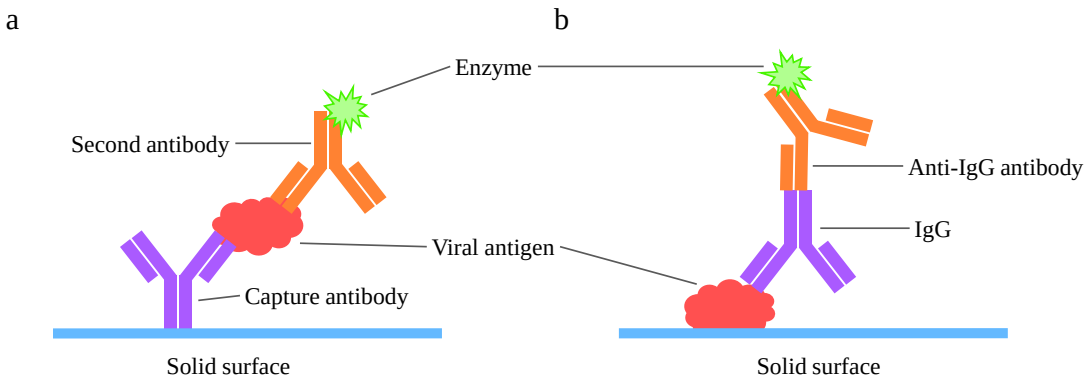


Figure 1-4: Scheme for enzyme-linked immunosorbent assay (ELISA). a) "Sandwich" ELISA, where the antigen is captured by an immobilized antibody and detected with a labelled second antibody. b) ELISA for the detection of antibodies that bind specific proteins immobilized on a surface.

ELISA can also be used to measure anti-viral antibodies in serum or nasal-wash after an infection. Here, antibodies bind to immobilized proteins and can be measured by labelled secondary antibodies (Figure 1-4b).

Western blot

The Western blot [35] is another antibody-based technique that can be used for the detection of viral proteins. Proteins are separated on a gel and transferred onto a nitro-cellulose sheet (blotted). This blot is then immuno-stained with enzyme-labelled antibodies. Bound antibodies become visible through addition of the enzyme's substrate.

Reverse transcriptase activity

The RT of retroviruses and herpesviruses is not present in uninfected host cells. The enzymatic activity of RT can be assayed to determine infection and viral propagation [36]. The RT is extracted from the virions. Then RNA, a primer, and radioactive-labelled nucleotides are added and incorporated nucleotides are quantified. Coupled with molecular detection and quantification (see PCR below), this assay enables sensitive identification of different viral species [37].

In situ hybidization

In situ hybridization (ISH) utilizes the hybridizing affinity of single-stranded DNA or RNA. Single-strand probes are added to a sample – often a tissue section or cells – and hybridize to their complementary DNA or RNA. The labeled probes can be localized

and intensities can be measured. Application to different time-points of an experiment reveals a spatio-temporal expression map of the investigated sample and is crucial for understanding gene regulations and functions, for example in embryonic studies. The method has its applications in virology [38, 39], as well, detecting viral DNA or RNA.

PCR

Molecular techniques are based on measurements of nucleic acids for the detection or identification of viruses in clinical samples. The polymerase chain reaction (PCR) is an extensively used method for the detection and quantification of viral nucleic acids in specimen [40]. PCR allows specific detection and measurement of small amounts of nucleic acids using a polymerase and specific primers to amplify the nucleic sequence of interest. Using an RT allows for the detection of RNA (RT-PCR), or the produced complementary DNA (cDNA). Sensitive quantification of produced DNA fragments is enabled by the use of real-time PCR, or quantitative PCR (qPCR) [41]. Labelled dyes binding unspecifically to dsDNA, or labelled DNA probes can be measured at each amplification step.

Microarrays

qPCR and DNA microarrays are both methods to study gene expression. The DNA microarray technology was developed in the early 1990s. Arrays have been produced for a variety of different applications. The most relevant techniques for comparison in this setting are gene expression arrays [42] and the application to virus detection [43]. In any case, DNA probes of interest are printed on glass slides for testing. The probe sequences are known and the printing is performed in a grid such that the positions of all sequences are known as well. RNA of the investigated sample is reverse-transcribed into cDNA using oligo(d)T primers, or random primers. A fluorescent label is added to the library. The samples are applied to the microarray, where hybridization takes place (Figure 1-5).

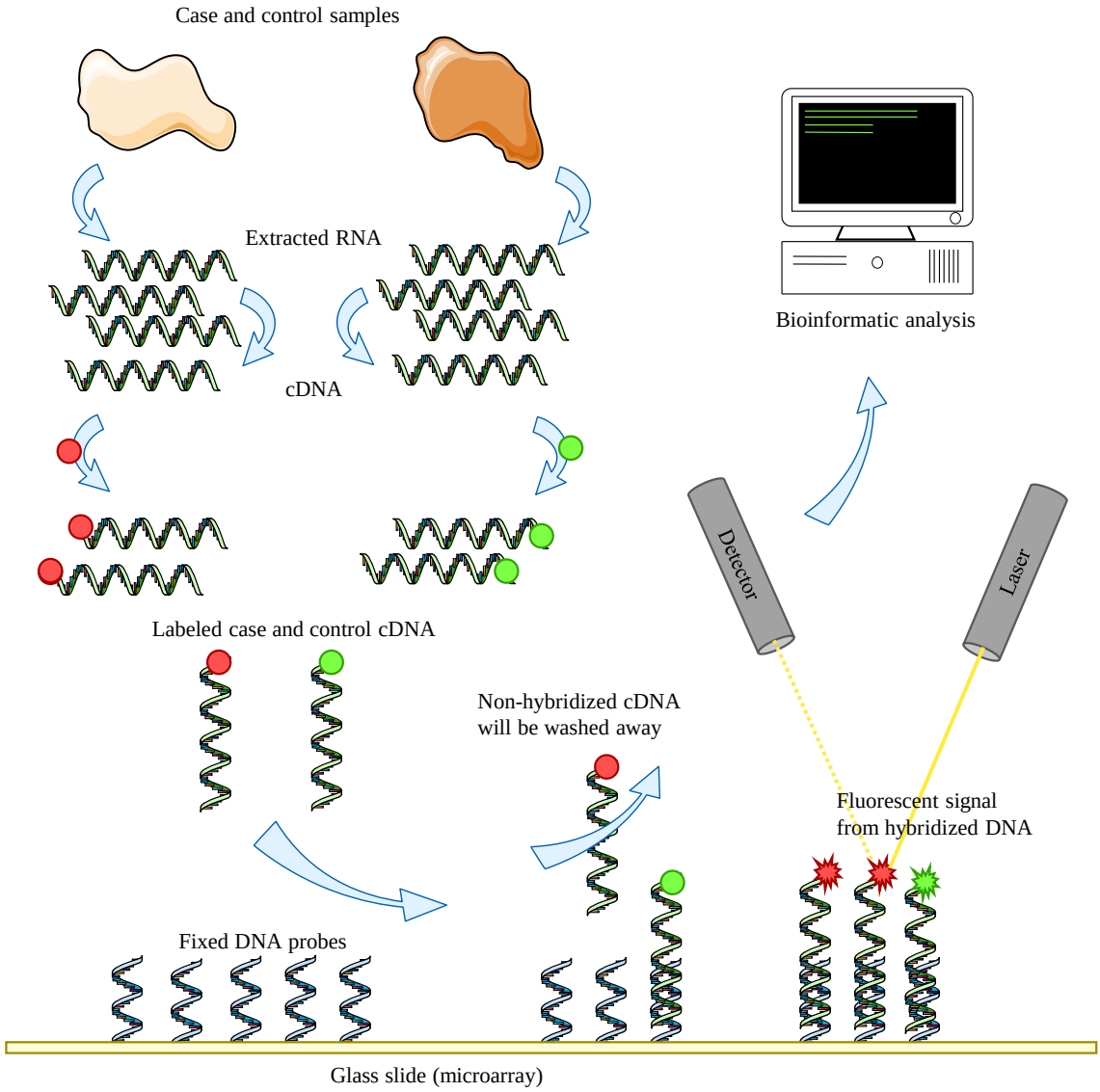


Figure 1-5: Simplified workflow of a microarray experiment. RNA is extracted from different samples, often case and control. RNA is reverse-transcribed into cDNA and labeled with a fluorescent dye. Labeled cDNA is added to fixed DNA probes on the microarray. After hybridization, nonbinding cDNA is washed off. The fluorescent dyes are excited by a laser and the emitted signal is measured and later analyzed using bioinformatics and statistics.

Any non-hybridized molecules will be washed off. The microarray is then scanned by a laser and the fluorescent emission levels are measured. Further, the raw data is normalized and the signal is quantified. Often the signal is calculated as a relative expression of one sample compared to another, or case compared to control.

1.5 Next Generation Sequencing

In 1977, Sanger and colleagues published the first ever complete genome sequence: the genome of the bacteriophage phiX174 [44]. Later that year, they published an improvement of the used method using DNA polymerase and chain termination [45]. This method is commonly referred to as Sanger sequencing. In the same year, Maxam and Gilbert published a comparable sequencing method based on chemical DNA modification and fragmentation techniques [46]. Because of its higher efficiency and fewer radioactive chemicals, Sanger sequencing was the prevalent method for DNA sequencing for the next 30 years.

Sanger sequencing was also used for the first complete human genome sequence [47]. The sequencing lasted more than a decade and required vast amounts of resources. This spurred the development of advanced technologies that were faster, cheaper, would allow a higher throughput, and would require less hands-on time. These next-generation sequencing (NGS) technologies were quickly adopted and led to a tremendous increase of the amount of available sequencing data [48].

Compared to Sanger sequencing, NGS technologies run massively in parallel with many million sequencing reactions simultaneously. The output of each reaction is detected directly without the need for electrophoresis. Using NGS, genomes can be sequenced at a higher throughput and at lower costs compared to Sanger sequencing. However, NGS methods produce shorter reads with a higher error rate [49], in general. Higher coverage enables compensation of these drawbacks, meaning that every genomic base has to be represented by several short sequences of the sequencing output. The short reads then have to be assembled by special algorithms to produce the full sequence of the analyzed genome.

Although other NGS systems might show higher accuracy and produce longer sequence reads [49], the high output and lower sequencing costs of the platforms produced by Illumina might explain the popularity and abundant usage of their sequencers. Illumina adopted a technology called sequencing by synthesis (SBS). Adapters are ligated to DNA fragments of the investigated genome (Figure 1-6a). These sequencing templates bind to immobilizing adapter oligonucleotides on the surface of a flow cell (Figure 1-6b).

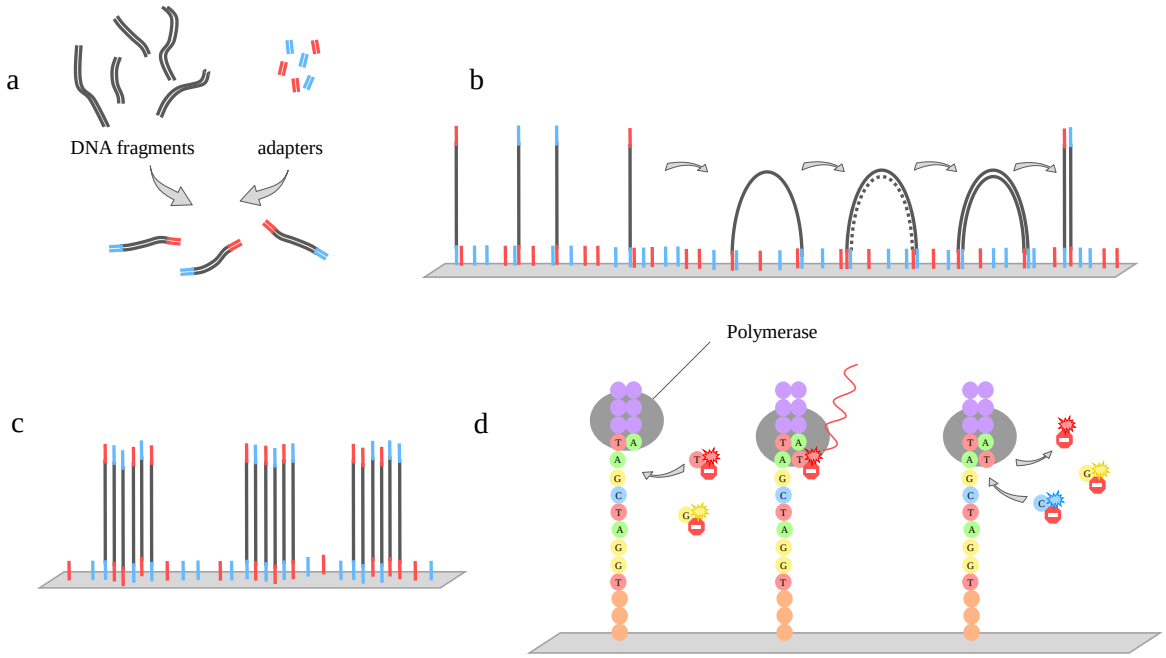


Figure 1-6: Simplified scheme of sequencing by synthesis on an Illumina flow cell. a) Adapters are ligated to double-stranded DNA fragments. b) Single stranded DNA templates are immobilized and multiplicated by bridge amplification. c) Each generated cluster contains up to 1,000 copies of the original template. d) Fluorescently labelled nucleotides are incorporated by a polymerase. Elongation is terminated until fluorescent signals are imaged. Enzymatic cleavage of labels enables further synthesis.

Solid-phase bridge amplification creates clusters with up to 1,000 identical copies of each template in close proximity (Figure 1-6c). Each single-stranded template is primed by the double-stranded adapter that serves as a binding site for a polymerase.

SBS generally uses four fluorescently labelled nucleotides (Figure 1-6d). Terminators ensure that a single labelled nucleotide is added to the nucleic acid strand during each sequencing cycle. The label terminates the polymerization process, once the nucleotide is incorporated. The fluorescent dyes of all clusters on the flow cell are imaged at each cycle to identify the added nucleotides. Then, the fluorescent labels and the blocking group are enzymatically cleaved to enable elongation of the sequence in the next sequencing cycle.

As opposed to the sequencing of genomic DNA, NGS can also be applied to the sequencing of RNA-derived cDNA [50, 51] (RNA-Seq). Several approaches have been developed to address different questions, including the sequencing of microRNAs, ribosomal RNAs, or targeted sequencing of specific RNA species. Sequencing of mRNA (mRNA-Seq) is the method of choice for analyzing the transcriptome. It has been widely adopted for the quantification of gene expression in

samples and is commonly simply referred to as RNA-Seq. In comparison to microarrays, sequencing of the transcriptome enables the detection of novel transcript isoforms, gene fusions, or sequence variations, in addition to gene expression. The total RNA of a sample is usually enriched for polyadenylated mRNA to remove a large share of ribosomal RNAs. However, this might also remove non-polyadenylated mRNAs or partially degraded mRNAs. For the analyses presented in this thesis, only data from mRNA-Seq for polyadenylated mRNAs were used.

1.6 Aim of this Thesis

The aim of the hereby-presented work is the development of a software pipeline for the detection of viruses in mammalian cells and tissues. The developed software should be less time consuming and less biased than existing methods for the identification of viruses. The deliverable of this project is a diagnostic platform that would thus exploit the yet unexplored unaligned sequence reads from high throughput sequencing experiments. The method should be applicable to standard RNA-Seq data and thus augment the generated knowledge from available sequencing data, as well as coming NGS experiments. Based on the rationale that all viruses express mRNA, which has to be processed by the host ribosome, viral transcripts are expected to be detectable and measurable in host RNA-Seq data.

Throughout the developmental phase, the platform should be challenged in benchmarks with samples containing known virus genome spike-in or known viral infections. The developmental phase includes a feedback-loop to improve the software iteratively, and thus to produce a qualitative platform for the identification of viruses. The developed pipeline should be applied to general screenings of RNA-Seq for quality check and diagnostic purposes. Further, the software application on well-characterized cell lines is planned with the anticipation of confirming known results and potentially identifying thus far unknown viral content. Finally, the application to clinically annotated cancer cohort data should enhance our knowledge of virus-associated malignant diseases. This can lead to the identification of potentially novel viral signatures as cancer restricted biomarkers.

2 Materials and Methods

The hereby-presented studies were conducted using data from high throughput sequencing experiments. All described laboratory techniques for handling of the samples (sections 2.1.1 and 2.1.2) were kindly performed by Jos de Graaf. I received the data for analysis after the sequencing run. At this point, I would like to give special thanks to him for his support throughout the project. TRON's core facility unit Medical Genomics or the unit for Next Generation Sequencing performed standard RNA-Seq for these samples. The following sections will describe the generation of the sequencing data from sample collection to FASTQ file creation. I will also describe the incorporated cell line and cancer cohort data, used databases, third party software, and the computer servers.

2.1 Data Sets and Cohorts

2.1.1 Samples for First Performance Review of VirusID

We obtained 20 samples from Baxter International containing 50 µL of cell line supernatant and spiked-in viral nucleic acids in a blinded fashion. Neither virus spike-in, nor the cell line background were known. Supernatants were taken from the African green monkey kidney cell line (VERO) or the Chinese hamster ovarian cell line (CHO). Four different virus species were used for this study: Minute virus of mice (MVM), Reovirus type III (ReoIII), Encephalomyocarditis virus (EMCV), and the Xenotropic murine leukemia virus (X-MuLV). The used virus genomes were added at two different concentrations with a 10-fold difference (Table 2-1) [52], i.e. roughly 10^4 genomes per 50µl and 10^5 genomes per 50 µL. The 20 samples, including four negative controls, were labeled 1A to 18A, 1C, and 10C.

Table 2-1: Samples for the first performance review of VirusID. Four different virus genomes were spiked-in into the supernatant of two different cell lines, VERO and CHO. The viruses were Minute virus of mice (MVM), Reovirus type III (ReoIII), Encephalomyocarditis virus (EMCV), and the Xenotropic murine leukemia virus (X-MuLV). Four samples contained only supernatants as negative controls.

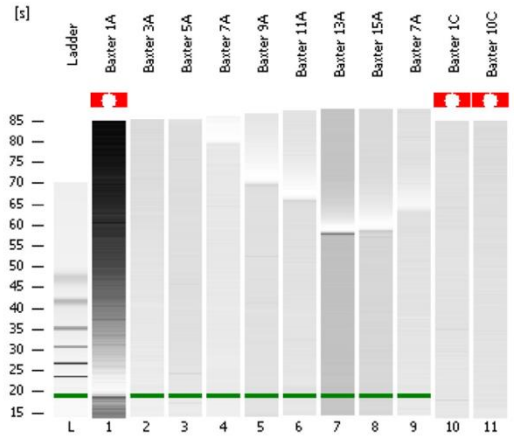
	negative controls	low concentration spike-in		high concentration spike-in	
VERO super-natant	1A: none	2A: MVM	3A: ReoIII	6A: MVM	7A: ReoIII
	1C: none	4A: EMCV	5A: X-MuLV	8A: EMCV	9A: X-MuLV
CHO super-natant	10A:none	11A: MVM	12A: ReoIII	15A: MVM	16A: ReoIII
	10C none	13A: EMCV	14A: X-MuLV	17A: EMCV	18A: X-MuLV

The samples were quality checked at arrival. Nucleic acids in samples 1A, 2A, 3A, 4A, and 5A were quantified using the Qubit dsDNA HS and RNA kit on a Bioanalyzer RNA pico Chip to measure both nucleic acid species. DNA and RNA quantities were however below the detection limits for all samples except 1A (DNA: 0.226 ng/µL) (Figure 2-1a). Further integrity checks of more samples confirmed the low DNA and RNA content (Figure 2-1b & c).

a

ID	DNA conc. [ng/µl]	RNA conc. [ng/µl]
1A	0,226	too low
2A	too low	too low
3A	too low	too low
4A	too low	too low
5A	too low	too low

b



c

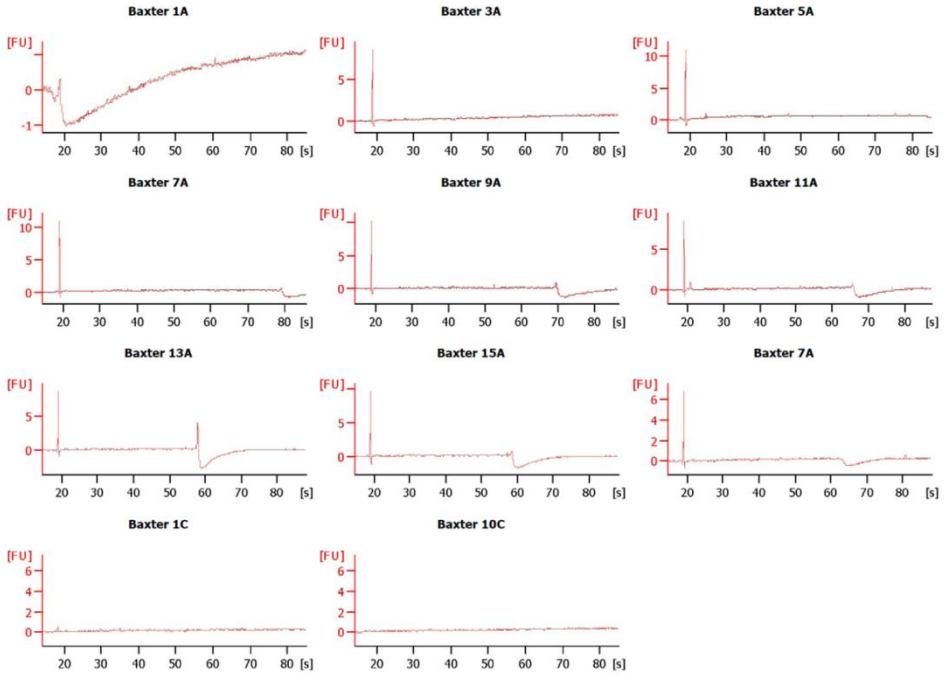


Figure 2-1: Bioanalyzer results of virus spike-in samples. In all samples except 1A, the nucleic acid concentration was below the detection limit (a). The gel-based image of further samples shows very weak, inconclusive bands (b). The results of the gel are mirrored in the plots of the electropherograms (c).

Sample processing

The low nucleic acid content of the samples necessitated a modified handling protocol compared to the standard in-house mRNA treatment. Before the cDNA synthesis, a vacuum concentration step was added. This procedure was shown to be superior to non-concentrated sample processing. The supplied liquid samples were vacuum-dried for 30 minutes and then resuspended in 5 µL of Qiagen nucleic acid elution buffer (Buffer EB). Further, RNA amplification was included after the second cDNA

synthesis step using the NuGEN Ovation RNA-Seq system, according to the manufacturer's instructions. After the amplification, cDNA was sheared in a Covaris S220. Barcoded libraries were prepared using the Illumina TruSeq DNA sample preparation kit using the standard protocol. The final libraries were amplified with 10 PCR cycles. The barcoded RNA-Seq libraries were clustered 10 pM on the cBot using TruSeq cluster kit v2.5 and 50 nucleotides (nt) were sequenced on the Illumina HiSeq 2000 with the TruSeq SBS kit-HS 50 cycles single and paired-end.

Data processing

All data was treated with the standard in-house processes for RNA-Seq data. Briefly, FASTQ files were generated and samples were demultiplexed using Illumina's software `bcl2fastq` (versions 1.8.4 – 2.17.1.14). The NGS runs resulted in an average of 23 million sequence reads per sample with a minimum of 13.2 million and a maximum of 31 million reads. Data was further processed as described in the VirusID section (3.1).

2.1.2 Samples for Final Benchmark of VirusID

We collaborated with Luis Terán Juárez from the National Institute for Respiratory Diseases, Mexico (Instituto Nacional de Enfermedades Respiratorias, INER), to obtain nasal aspirates from children with viral infections of the respiratory tract. The samples were virus typed at the INER by PCR or immunofluorescence (IF). Total RNA was extracted using the Qiagen RNeasy Micro kit before shipping. The 50 shipped samples were quality checked at reception. RNA concentrations were measured using the Qubit RNA kit and on a Bioanalyzer RNA pico Chip. Concentrations were below the detection limit in 26 of the 50 samples. Of the remaining 24 samples, the first 21 samples were selected for further processing (Table 2-2).

Table 2-2: Selected samples for the final benchmark of VirusID. Samples marked with '*' were not processed further.

Sample ID	Virus Typing Result	Detection Method	RNA conc. (Qubit ng/µL)
534	Human respiratory syncytial virus (RSV)	IF	5.9
536	Adenovirus	IF	8.9
537	Adenovirus	IF	9.8
544	Influenza A virus, Influenza B virus, Parainfluenza virus type 1, Parainfluenza virus type 3, Human respiratory syncytial virus (RSV)	IF	5,4
545	Influenza B virus, parainfluenza virus type 2, parainfluenza virus type 3	IF	6.8
546	Rhinovirus	PCR	6.9
548	Influenza A virus	IF	7.6
549	Rhinovirus	PCR	12.9
552	Human respiratory syncytial virus (RSV)	PCR	9.4
553	Rhinovirus	PCR	4.6
554	Rhinovirus	PCR	16.8
565	Rhinovirus	PCR	9.2
567	Rhinovirus	PCR	6,8
568	Rhinovirus	PCR	8.0
572	Rhinovirus	PCR	6.3
576	Rhinovirus	PCR	4.5
577	Rhinovirus	PCR	16.7
584	Corona Virus	PCR	6.7
585	Human respiratory syncytial virus (RSV), Rhinovirus	PCR	7.5
590	Rhinovirus	PCR	5,9
592	Corona Virus	PCR	9.4
593 *	Parainfluenza virus type 4	PCR	4,3
599 *	Coronavirus	PCR	3,11
605 *	Rhinovirus	PCR	8,6

RNA amplification using a modified SMART method

One microliter containing 1 ng of total RNA was resuspended in 4.8 µL H_2O containing 1 µL TSO-B (12 µM), 1 µL Oligo(dT) primer (12 µM), 0.125 µL RNase inhibitor (40 U/µL) and 2.675 µL H_2O.

The total RNA was heated to 70 °C for 5 minutes and immediately put on ice for an additional five minutes. 4.2 µL of reverse transcription buffer were added per sample (2 µL RT Buffer, 1 µL dNTP mix (10 mM), 0.25 µL RNase inhibitor (40 U/µL), 0.2 µL DTT (100 mM), 0.25 µL H_2O, 0.5 µL Mint-Reverse Transcriptase (200 U/µL).

Reverse transcription was carried out for 90 minutes at 42 °C then subsequently heated to 70 °C for 15 minutes. The cDNA was amplified without further purification. Five microliters of Advantage 2 PCR-Buffer, 2 µL dNTP Mix (10 mM), 2 µL TS-PCR-Oligo (12 µM), 2 µL Advantage 2 Polymerase, and 29 µL H_2O were added and mixed thoroughly by pipetting. The following program was used for amplification: 95 °C for 1 minute; 20 cycles: 95 °C for 1 minute, 65 °C for 30 seconds, 68 °C for 6 minutes; 72 °C for 10 minutes.

cDNA quality was checked after amplification. The results are reported in Table 2-3 in the column ‘cDNA conc.’.

Table 2-3: Quality check of cDNA and sequencing library. The three samples with comments were not processed further.

Sample ID	cDNA conc. (Qubit ng/µL)	Nextera XT lib. conc. (Qubit ng/µL)	Comment
534	2.7	10.1	
536	4.8	7.2	
537	3.3	5.6	
544	2.3	-	Only dimers
545	1.3	4.3	
546	2.6	2.9	
548	2.2	7.0	
549	0.5	5.2	
552	2.6	8.3	
553	1.7	16.6	
554	2.3	11.1	
565	2.6	5.1	
567	1.1	-	cDNA -
568	2.2	8.7	
572	1.7	10.8	
576	1.2	10.1	
577	1.8	9.2	
584	1.3	11.1	
585	0.8	9.3	
590	0.4	-	cDNA -
592	2.5	14.7	

Library preparation using Illumina Nextera XT

One nanogram of amplified cDNA was fragmented and tagged in a one-step reaction using a hyperactive transposase (tagmentation). The tagmented DNA was then amplified using the Nextera XT standard protocol. Dual index primers were used to prepare the samples for multiplex sequencing on a HiSeq 2500. The column 'Nextera XT lib. conc.' in Table 2-3 lists the concentrations of the sequencing libraries.

2.1.3 Standard RNA-Seq

For the quantification of gene expression values, the standard in-house sequencing protocol was applied. This entails preparing barcoded mRNA-Seq cDNA libraries from 1 µg of total RNA using a modified version of the Illumina mRNA-Seq protocol: the mRNA was isolated using Seramag Oligo(dT) magnetic beads (Thermo Scientific). Isolated mRNA was fragmented using divalent cations and heat resulting in fragments

ranging from 160 to 220 bp. Fragmented mRNA was converted into cDNA using random primers and SuperScriptII (Invitrogen) followed by second strand synthesis using DNA polymerase I and RNaseH. cDNA was end-repaired using T4 DNA polymerase and Klenow DNA polymerase, and 5' phosphorylated using T4 polynucleotide kinase. Blunt-ended cDNA fragments were 3' adenylated using Klenow fragment (3' to 5' exo minus). cDNA insert and 3' single T-overhang Illumina multiplex-specific adapters were ligated using a 10:1 molar ratio of adapter to insert using the T4 DNA ligase.

Enrichment and addition of Illumina six base index and flow cell specific sequences was done by PCR using Phusion DNA polymerase (Finnzymes). All cleanups were done using 1.8x volume of Agencourt AMPure XP magnetic beads. All quality controls were performed using Invitrogen's Qubit HS assay and fragment size was determined using Agilent's 2100 Bioanalyzer HS DNA assay.

Barcoded RNA-Seq libraries were clustered 7 pM on the cBot using TruSeq SR cluster kit v2.5 and 50 bases were sequenced on the Illumina HiSeq 2000 using TruSeq SBS kit-HS 50 cycles single or paired-end.

2.1.4 Cell Line Data

Raw sequencing data were collected from either the Sequence Read Archive (SRA) [53] or the Gene Expression Omnibus (GEO) including RNA-Seq data of 186 human normal and cancer cell line data sets (Appendix Table A-1) from 22 different projects. In cases where cell lines were treated or modified in the course of the experiments, only data from cell lines at base line or the controls were downloaded. Some of the cell lines were sequenced as replicates, either in one project or in different projects (Figure 2-2). In total, the 186 data sets contained sequencing data of 143 different cell lines. The SRA formatted and compressed files were processed into raw files in FASTQ format using the SRA toolkit (versions 2.3.4- 2.5.4), if needed. The complete dataset contained paired-end reads of varying lengths, ranging from 30 to 107 bases per read. To ensure comparability, data were restricted to those derived from Illumina platforms.

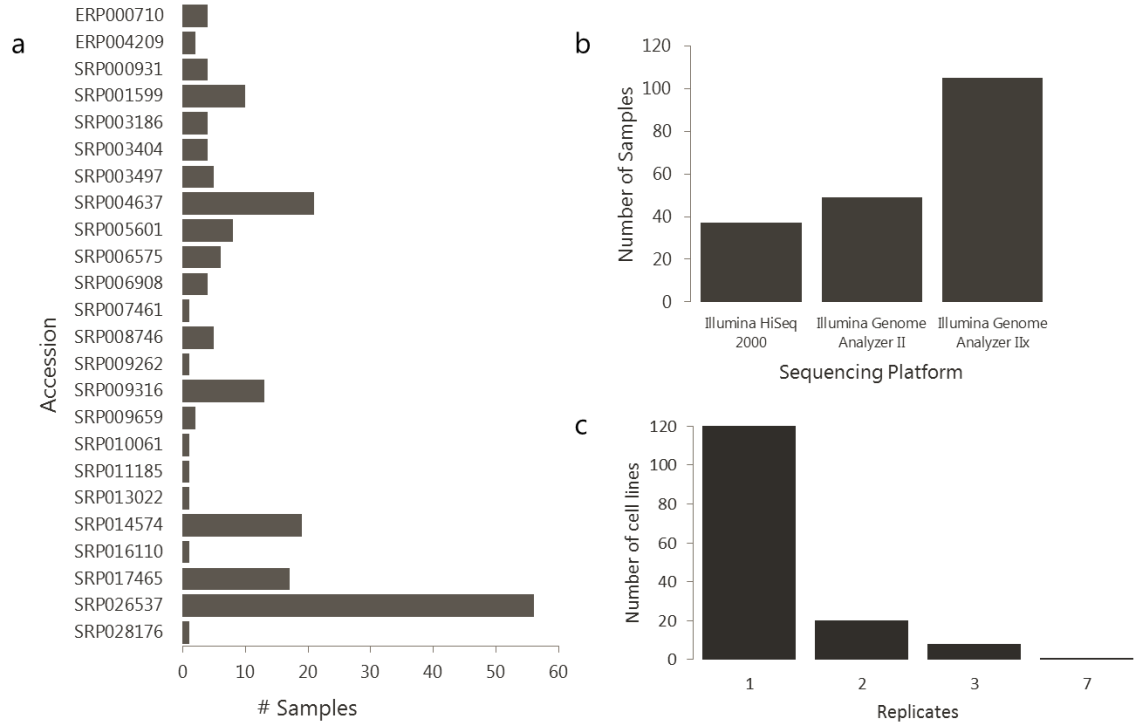


Figure 2-2: Description of the utilized cell line cohort. The incorporated samples were downloaded from 22 different projects with varying sample numbers (a). Sample selection was restricted to data from Illumina platforms only (b). The majority of the cell lines were not available as replicates (c).

2.1.5 Cancer Cohort Data

For this analysis, we downloaded and processed raw RNA-Seq data of the following cohorts from The Cancer Genome Atlas (TCGA): cervical squamous cell carcinoma (CESC; 304 samples), diffuse large B cell lymphoma (DLBC; 48 samples), glioblastoma (GBM; 160 samples), head and neck squamous cell carcinoma (HNSC; 398 samples), and stomach adenocarcinoma (STAD; 190 samples). We retrieved the raw data in BAM format, a binary format containing the aligned sequences. The files were processed to FASTQ format using `bam2fastq` to utilize them for further processing. No additional pre-processing was applied.

Additionally, MRI-localized brain biopsies [54] and endemic Burkitt's lymphoma samples [55] were collected from the SRA. Data were pre-processed as described in section 2.1.4.

2.1.6 Other Third Party Data

For a first test of feasibility of virus detection in mammalian samples, we selected a dataset of seven murine lung samples, four of which were infected with severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV, NC_004718.3) [56]. The raw sequencing data were kindly provided by John C. Castle. The samples had been sequenced single-end with 36 bases per read on an Illumina Genome Analyzer II. Each sample produced 24.3 to 41.7 million sequencing reads.

2.2 Databases

Host Genomes and Transcriptomes

In the presented studies, I used the human genome version hg19/GRCh37 or the murine genome mm9/NCBI37 as reference genomes for the mapping of short NGS reads using alignment algorithms. These genomes had been downloaded as standard references from the University of California Santa Cruz (UCSC, http://hgdownload.cse.ucsc.edu/ downloads.html).

As reference genes and transcripts, I used the UCSC Known Genes tables, available on the UCSC servers (http://genome.ucsc.edu/cgi-bin/hgTables). Genes and transcripts were available for both, human and murine genomes. The tables were in BED format, containing the gene identifier, the locus, exon information, and genome strand specification.

Virus Genomes

The virus detection tools VirusID and VIRGENE rely on the incorporation of viral genomes and annotation for the analysis of samples. Therefore I downloaded the curated NCBI Reference Sequence Database (RefSeq) [57] genomes of the NCBI Viral Genomes Resource [58]. They are available for download in FASTA format from NCBI's FTP server site ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/. The viral database contains a variety of viruses including human viruses, vertebrate and invertebrate viruses, plant viruses, bacteriophages, and fungal viruses (Figure 2-3). Segmented viral genomes are represented by one sequence per segment. The FASTA file for the first version of VirusID contained 4,715 viruses (retrieved 2013-06-18).

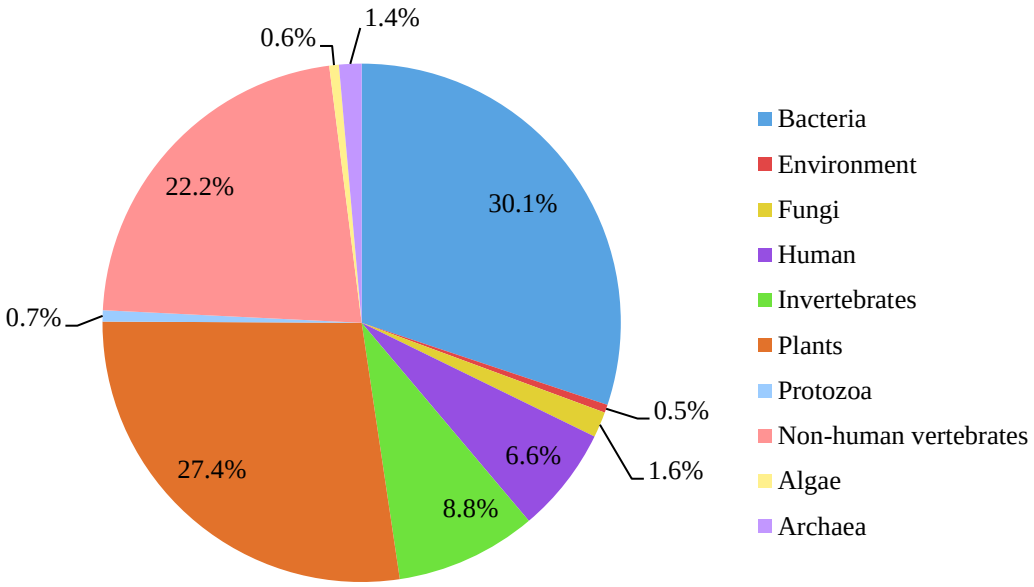


Figure 2-3: Relative host group distribution among viral genomes in the RefSeq virus database. Figure modified from Brister *et al.* (2014) [58].

In November 2016, the database contained 7,807 sequences. To mirror the actuality of the curated and updated database, the virus database was downloaded biannually.

Viral Genes and Transcripts

To calculate the expression values of viral genes from RNA-Seq data, the corresponding genome locations have to be mapped to the genome sequences and the sequencing profiles. However, the genomic features for coding sequences (CDS) or mRNA of the RefSeq virus genomes database cannot be readily downloaded. The NCBI Nucleotide database is a collection of sequences, including genome, gene, and transcript sequence data. The information can be retrieved in GenBank flat file format via the HTTPS protocol. VIRGENE contains the script `gbFeatures.py`. This script loops over the accession numbers of the RefSeq virus database and retrieves the corresponding CDS and mRNA features from the NCBI Nucleotide database (Figure 2-4).

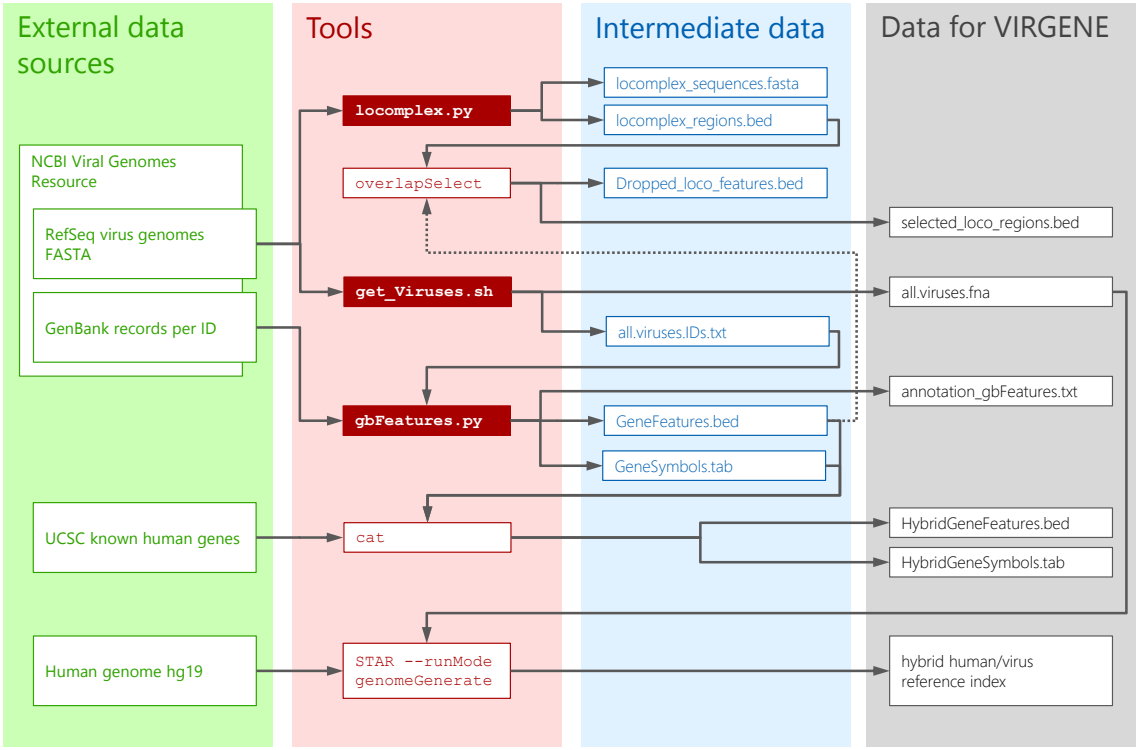


Figure 2-4: Workflow of data preparation for VIRGENE. The scheme depicts the used external data sources that are downloaded or transformed by the listed tools. Intermediate data might be used during the further process. Finally, datasets are prepared and stored for use by the VIRGENE pipeline. Tools highlighted with a red background are self-developed. Sources and versions of the used third-party software are listed in Table **2-4**.

For the convenient use with software that quantifies gene expressions, the data was stored in BED format for later use. This download is not part of the analysis process and need only be performed once during the generation of the virus database.

The virus genomes database downloaded 2016-11-25 contained 7,807 FASTA sequences. Of these sequences, 2,057 originate from phages. The term ‘segment’ in 1,592 of the sequence descriptions indicated that the virus genome was segmented and hence the complete sequence was composed of several entries in the FASTA file. The script `gbFeatures.py` retrieved the CDS or mRNA loci of 7,640 virus sequences. The resulting `GeneFeatures.bed` file contains 298,353 entries. Of these features, 99.16 % (295,854 entries) consisted of only one exon, 2,499 entries contained at least 2 exons.

Virus Annotation

The RefSeq virus genomes database contains sequence information from a variety of viruses from a wide range of hosts. Some of the viruses are evolutionary closely

related, resulting in high similarity of their genomic sequences. The alignment of short sequence reads to related or similar reference sequences can lead to ambiguous alignments. This will affect the analysis as several sequences might be reported as potentially interesting. However, the evolutionary distance is reflected neither in the accession numbers of the genomic sequences, nor in the names of the viruses. Therefore, the VIRGENE script `gbFeatures.py` also retrieves the taxonomy as part of the annotation of each contained virus (Figure 2-4). This information is stored for use during the filtering of potentially closely related viruses. The resulting semicolon-separated text file is comprised of one line for each virus containing the accession number, the organism name of the virus, as well as its taxonomy information. The hierarchical steps of the taxonomy are separated by a hyphen. The following example shows the kingdom, genome structure, order, family, subfamily, and the genus of the Human herpesvirus 1 (accession number NC_001806.2):

```
Viruses - dsDNA viruses, no RNA stage - Herpesvirales - Herpesviridae - Alphaherpesvirinae - Simplexvirus
```

2.3 Third Party Software

VirusID and VIRGENE were both designed as partial pipelines incorporating a set of different software and tools (Table 2-4). These tools were used for pre-processing of the raw data, alignment of the sequence reads, post-processing, and visualization of the results as well as for the creation of the used databases. In some cases, several versions of a tool were used throughout this project. In those cases, the ranges of version numbers are indicated, or only the latest available version is listed (Biopython).

Table 2-4: Used software and tools.

Name	Version	Task	Reference/Source
bam2fastq	1.1.0	Transforming alignment files (BAM) to sequence files (FASTQ)	https://gsl.hudsonalpha.org/information/software/bam2fastq
bcl2fastq	1.8.4 – 2.17.1.14	Demultiplexing of sequenced samples and conversion of Illumina BCL files to FASTQ text files	https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html/
Biopython	1.68	Retrieve and dissect virus genes and annotation from NCBI	Cock *et al*. [59]
Bowtie	0.12.8	Exact mapping of sequence reads	Langmead *et al.* [60]
ContextMap	V2.3.3	Virus detection software	Bonfert *et al.* [61]
IGV – Integrative Genomics Viewer	2.3.55	Visualization mapping profiles	Robinson *et al.* [62]
NCBI SRA toolkit	2.3.4- 2.5.4	Conversion of SRA formatted files to FASTQ format	Leinonen *et al.* [53]
OverlapSelect	NA	Intersect gene locations with annotations	http://hgdownload.cse.ucsc.edu/admin/exe/
PRINSEQ	0.20.3	Filter, reformat, or trim genomic data	Schmieder and Edwards [63]
R	3.1.2_SL	Downstream analyses and visualizations	R core team [64]
Samtools	0.1.16-1.3.1	Operations on alignment files in SAM or BAM format	Li *et al.* [65]
seq2HLA	V2.4	HLA typing software	Boegel *et al.* [66]
STAR	2.1.4a-2.5.1b	Gapped/spliced alignment of sequence reads	Dobin *et al.* [67]
VirusSeq	2015-01-05	Virus detection software	Chen *et al.* [68]

2.4 Servers and Hardware

The hereby-presented analyses were all performed on the high performance computing (HPC) servers of TRON and at the Data Center (ZDV) of the Johannes Gutenberg-University Mainz (JGU).

TRON owns a collection of computing servers hosted by the ZDV. The server landscape was extended over the time course of this project. At the end of 2016, five servers were available with a total of 320 computing cores and 1.25 terabytes (TB) of random access memory (RAM) in total. The servers are connected to storage servers

with a total capacity of 378 TB of data storage. All servers run on Linux operating systems, four of which use SUSE Linux Enterprise Server 11 Service Pack 4, one uses Scientific Linux release 7.2.

Additionally, the ZDV at the JGU is hosting MOGON, a computing cluster for high performance scientific computing. MOGON consists of 555 nodes with 64 cores each, totaling to 35,520 cores with 89 TB of RAM. The nodes have access to 1.5 TB of local storage, each. All servers are running on the operating system Scientific Linux 6.4.

2.5 Data Analysis

The following sections will describe general handling of the used data before, during, and after analysis. This chapter will also address the developed pipelines and algorithms.

2.5.1 Pre-processing of Data

The sequence alignment software described in the next section used FASTQ formatted files as an input. The raw data files retrieved from TCGA were compressed FASTQ files and could thus be used for alignment without manipulation. The SRA provided all sequence data in their own SRA format. This format allows the addition of metadata to raw files. To use the sequences for alignment, the files were converted first using the SRA toolkit [53]. The command `fastq-dump` extracted all sequence information and stored it in FASTQ format. Files were then compressed using the Linux command line tool `gzip`.

2.5.2 Mapping

One of the key processing steps of RNA-Seq data analysis is the correct allocation of a short sequence read to its genomic origin i.e. the alignment of reads. Spliced Transcripts Alignment to a Reference (STAR) [67] is a splice aware alignment tool enabling the correct assignment of sequence reads, even across exon-exon junctions and was the preferred tool in this analysis.

For the alignment of sequence reads of all data sets, I created a hybrid genome reference. The hybrid reference consisted of the human reference genome (version hg19, GRCh37) and the respective viral genomes sequence database (see chapter 2.2). The reference index was created using STAR in run mode ‘genomeGenerate’. No gene

annotations or splice junction annotations were used for the creation of this index. Alignments were valid with a mismatch ratio of up to 0.2, allowing up to 100 alignments, if the mapping scores were equal to the best alignment score.

2.5.3 Gene Expression

Gene expression values were calculated and further processed as part of the VIRGENE pipeline using the TRON RNA-Seq pipeline (TRSP). After allocation of the short sequence reads to genomic coordinates, the mapped reads per gene locus were quantified as described in our publication of the TRON Cell Line Portal (TCLP) [69]. Briefly, to enable viral gene expression quantification the process was adjusted as follows: A 'hybrid' gene locus database in BED format was created containing UCSC known human genes [70] and the viral genes retrieved from NCBI by `gbFeatures.py` (chapter 2.2). All gene models were built by a union of all transcript isoforms per gene symbol. Additionally, an isoform-to-gene dictionary was created for both, human and viral genes. Reads overlapping transcripts or genes were counted. Total counts were then normalized by sequence library size and gene length resulting in gene expression levels as reads per kilobase of exons per million mapped reads (RPKM) [71]. Results were printed in a semicolon-separated table with expression values for each gene and sample.

2.5.4 VirusID

The deliverable of the VirusID project was a diagnostic platform that would detect viruses in mammalian cells in an unbiased fashion. The software was developed over the course of the project and beyond, and improved iteratively. The VirusID software pipeline `virusID.py` is composed of Python code executing third party software, processing the intermediate results and printing the results of the analyses. The following sections will address general handling of data by VirusID.

Alignment and counting

First, short sequence reads of an RNA-Seq experiment are mapped to a created reference to identify their origin. The reference is a hybrid of the host genome and the virus genomes (see chapter 2.2). For the analyses described in this thesis, I used the human genome (hg19/GRCh37) or the mouse genome (mm9/NCBI37) as host genomes. The virus genomes and their retrieval are described in chapter 2.2. The

alignment of the sequence reads is described in section 2.5.2 and can be executed from inside the pipeline or independently of VirusID as well. Thus, VirusID can also be applied to a prior alignment.

In the next step, the pipeline reads in the virus genomes FASTA database. Then, the aligned reads are then read from the SAM/BAM alignment file, as well as the alignment statistics stored in `Log.final.out` generated by the aligner STAR. VirusID utilizes the Pysam package (https://github.com/pysam-developers/pysam), a python wrapper for SAMtools [65]. This allows for the direct accession of the binary format (BAM). Thus, for each genomic coordinate of the virus genomes the mapped reads can be accessed without looping over the whole alignment file. Information is stored for a variety of parameters:

- alignment coverage with number of reads per base
- consensus sequence of all aligned reads
- mean contig size per virus created by overlapping reads
- length of the longest contig
- number or reads per virus genome
- number of bases of all mapped reads per virus
- normalized read count per genome in parts per million (ppm) of all mapped reads, including host sequences, as fraction of viral reads (FVR)
- percentage of virus genome covered with sequence reads (VGC)

All parameters will be stored for uniquely aligned reads as well as for all mapped reads including ambiguously aligned reads. Virus annotation is also stored, including name, length of the genome, genome sequence, description, and taxonomy.

Selection process and filtering

Whenever sequence reads map to a virus genome, the virus is further considered for the selection and filtering process. First, viruses with a VGC of less than 5 % are filtered out. Of all remaining viruses, the taxonomic annotation is compared in a pairwise manner. If the taxonomy of two viruses is identical to the level of viral sub-family, the virus with lower genome coverage is discarded.

Viruses are considered as ‘detected’ if their genome coverage exceeds 20 % VGC and if FVR is greater than or equal to two. The VGC cutoff was intentionally set well below

the determined median of all virus genome coverages above 0 % in all cell line samples plus two times the standard deviation, which roughly equaled 30 %. All other viruses with coverages greater than 5 % and below 20 % will be considered potentially suspicious and reported for further consideration. The FVR cutoff appropriately determined the positive samples (for example see Figure 3-3), except the extremely low expression signals, and has been used by other research groups as well [72].

Output

By default, output files will be prefixed by the job name defined by the option handle `-j/--job` (default: 'VirusID'). The output file `<job>_scores.csv` is a semicolon-separated text file. It contains collected information for each of the viruses: the virus identifier, the description, the length of the genome, the genome hit count in number of reads and in number of bases, the number of created contigs, the mean contig size and the longest contig size, the VGC in percent calculated from all mapped reads, the VGC with only the uniquely mapped reads, and the FVR in ppm of all aligned reads.

Reported contigs resemble the consensus sequence of assembled overlapping mapped reads. The contigs are stored in the file `<job>_contigs.fasta` in a FASTA sequence format. The script reports contigs if they are equal or longer than the specified minimum length. This value can be varied by the command line argument `-m/--min_contig` and is set to 101 nucleotides (nt) by default. Coverages of one read per base are already considered for the assembly of contigs. For each nucleotide position of the contig, the consensus nucleotide is selected if it has a minimum prevalence of 70 %. Otherwise, the base is printed as 'N' (any).

The genome coverage output file `<job>_coverages.csv` stores the coverages of each virus genome with a hit count of greater than one. The file is a semicolon-separated file containing one line per virus. Each line starts with the virus identifier. This is followed by the number of reads covering the first base of the virus, then the number of reads for the second base of the virus, continuing until the coverage of all virus genome nucleotide positions are reported.

VirusID logs its actions in a log file named `log_<job>.txt`. Apart from general events, like the start of the software itself as well as commands to start third-party

software, the log file also contains information about detected viruses and potentially suspicious sequences.

The virus scores file reports metrics for all of the viruses for consideration and further inspection. The reported contigs can be incorporated into genome or transcriptome assembly attempts. The coverages can be used for further inspections of the expression profiles.

2.5.5 VIRGENE

Data processing

As a first step toward viral gene expression profiles, sequence reads are aligned to a human-virus hybrid reference genome index (see chapter 2.2). VIRGENE integrates STAR [67] for the alignments, with the same parameters as VirusID, as described in section 2.5.2. VIRGENE then starts the processing of the aligned reads with TRSP as described in section 2.5.3. Read counts and normalized counts are printed as results. VIRGENE reads the alignments and the normalized gene expression table for further processing.

The filtering of potential false positive signals is visualized in Figure 2-5. First, the FVR and VGC are calculated for each of the virus genomes. Next VIRGENE filters low abundance viruses and discards viruses with an FVR $\leq$ 2 ppm, a VGC $\leq$ 5 %, and that do not express at least one gene at > 1 RPKM. In the following step, each virus is discarded, if it only expresses genes that contain low complexity sequences. All remaining viruses are then further selected according to their relatedness using the taxonomy filter as described above (section 2.5.4).

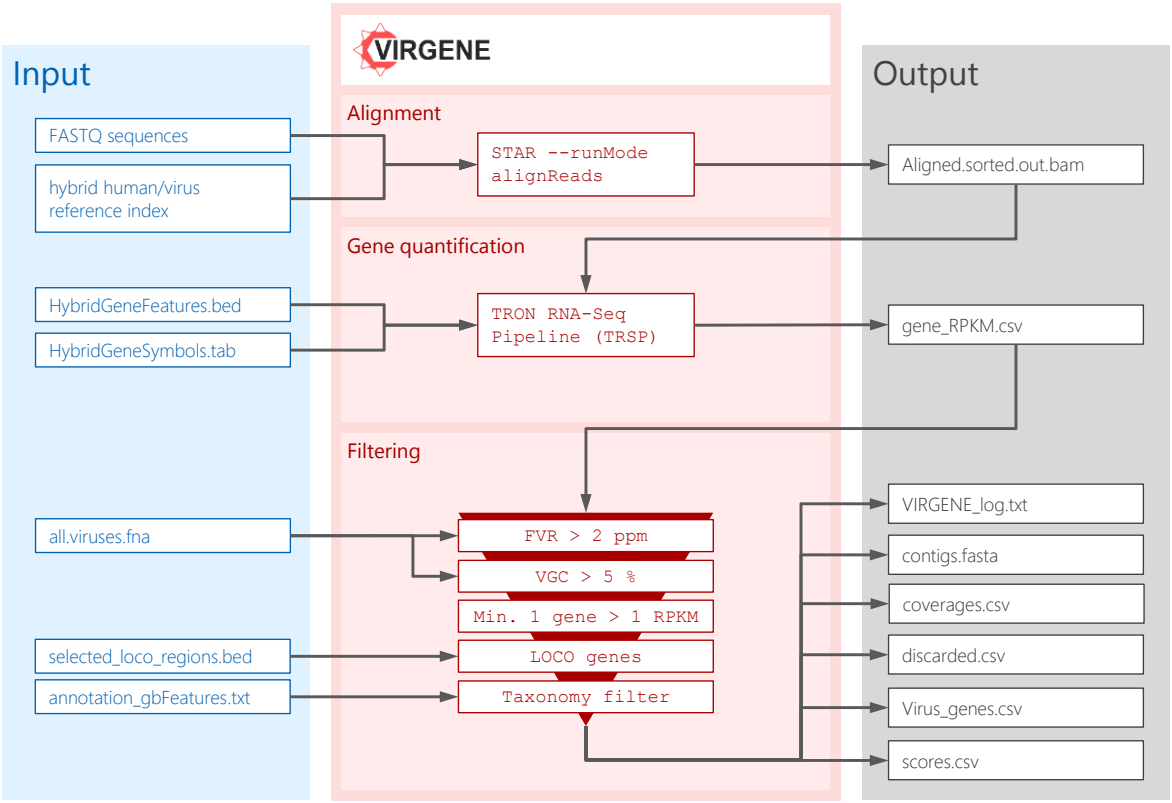


Figure 2-5: Workflow of the VIRGENE pipeline. Sequence reads are aligned to a human-virus hybrid reference. Aligned reads are processed for gene quantification. Finally, gene expression values are filtered before output of detected viruses and their expression profiles.

Low complexity genomic regions

To determine and locate virus genome sections with a low sequence complexity, a custom script was created. The script `locomplex.py` runs a sliding window approach with an adjustable window size (default: 40 nucleotides) and computes the sequence complexity for each of the iterations. The genome sequence complexity is computed using an adaptation of the sequence entropy [73]. For this, the distribution of character counts has been replaced by the distribution of distance counts between the single characters of the nucleotide alphabet. Hence, we calculate the entropy $H(X)$ with

$$H(X) = \sum_i p_i * \log_2\left(\frac{1}{p_i}\right),$$

where p_i is the probability of a distance i from one letter to its next repetition inside the sliding window. Thus, homopolymers or repetitions of a small set of characters would lead to small values of $H(X)$, while random distributions of distances between characters would result in high values. We experienced a cut off of $H(X) = 1.8$ being

sufficient for a window of 40 bases. Adjacent windows with low sequence complexity are joined before being output in FASTA and BED formats. The output is then used for downstream processing to filter potentially false alignments.

Output

VIRGENE outputs all of the files that are also created by VirusID (2.5.4): the scores table, the contigs FASTA file, and the coverages file. In addition, VIRGENE reports the expressed viral genes in a separate semicolon-separated table with the corresponding read counts normalized by number of total reads and gene length. VIRGENE prints the filtered genes and viruses as well for the user's information and for possible further consideration.

2.5.6 Downstream Processing and Analyses

The following methods were used for further analysis of the results, visualization, and interpretation of the data. These analyses were performed downstream of VirusID or VIRGENE.

Heatmaps

Data tables can be visualized as a heatmap, where the data values are translated to a color key. Gene expression heatmaps were plotted using the computing environment R [64]. Final plots were generated using the function `heatmap.2` from the `gplots` package [74] (version 2.17.0).

Non-negative matrix factorization

The ascertained co-expression of HPV genes was investigated by applying non-negative matrix factorization (NMF). NMF was provided in the package `NMF` [75] (version 0.20.6) inside the statistical computing environment R [64]. The basis map was plotted using the `NMF` function `basismap`.

Survival analysis

Five year overall survival for TCGA cohorts HNSC and CESC were calculated where annotation was available. The samples were stratified by their Human papillomavirus (HPV) status with a cutoff of 2 ppm on the fraction of viral reads (FVR). For the HPV subgroup analyses, groups were stratified by their mean expression of HPV genes E2, E4, and E5 with a cutoff of 20 RPKM. All survival analyses were performed in the

statistical computing environment R [64] using the package `survival` [76] (version 2.38-3) and fitting a cox proportional hazards regression model [77]. The Kaplan-Meyer survival plots were printed within this environment.

BLAST

VirusID and VIRGENE both output partial virus genomes that are covered with aligned sequence reads in a FASTA format. Some of the reported sequences were further analyzed using the Basic Local Alignment Search Tool (BLAST) [78] provided online by the National Center for Biotechnology Information (NCBI). To account for the expected variety of sequence sources, I used nucleotide BLAST for 'somewhat similar sequences' (blastn) [79, 80] against the database 'Nucleotide collection (nr/nt)'.

3 Results

High throughput sequencing methods like NGS provide an unbiased view into a sample's nucleic acids, including pathogenic transcriptomes in the case of RNA-Seq. We set out to implement a method for the detection of viral sequences from mammalian transcriptomic data. This chapter describes the development of a software for virus detection, its further development incorporating viral transcriptomes, and their application to cell lines and tumor samples.

3.1 VirusID: NGS-based Identification of Viruses

TRON participated in the Spitzencluster Award of the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) for a regional Cluster for Individualized Immune Intervention (Ci3). *VirusID: the detection of viruses with NGS* was one of Ci3's projects and benefited from the cluster's support, both structurally and financially. The aim of the VirusID project was the development of an NGS-based diagnostic pipeline to detect viral nucleic acids in mammalian samples. The used samples are described in sections 2.1.1 and 2.1.2. The initial lab and software-sided methods were performance reviewed in the beginning using cell-free supernatants of mammalian cell lines containing spiked-in virus genomes. For the final benchmark, we used nasal swabs of virus-infected children to challenge VirusID and compared the results to other virus detection software. The results are reported in the following sections of this chapter.

3.1.1 The Bioinformatics Platform VirusID

VirusID is one of the software pipelines developed during my studies to detect and identify viral nucleic acids in mammalian cells using NGS. The software is written in the programming language Python. It incorporates third party software and performs counting and filtering of the intermediate results (see section 2.5.4).

The initial step of most RNA-Seq analyses is the alignment to a reference genome. Several tools have been implemented to cope with this task and a collection of them has recently been benchmarked [81]. One of the best-benchmarked tools regarding accuracy and reliability is STAR. This tool has been used for standard RNA-Seq

analysis by our group and was the aligner for VirusID from the very beginning. STAR is designed to perform gapped alignments to retrieve transcript isoforms or genome break points. The latter might become important in downstream analyses of virus genome insertions into the host genome.

Analogies in human and virus sequencing data

When studying the microbiome or the virome of a host organism using NGS data, it is common to remove all the host-derived sequence reads from a dataset. This process called host subtraction is usually performed by mapping the reads to the host genome and then proceeding with the remaining sequences. To assess the abundance and structure of potential human reads mapping to viruses, I aligned RNA-Seq data from the Illumina Human Body Map 2.0 project (GEO accession GSE30611) to the virus reference genomes (downloaded 2013-06-18). Reads were aligned using STAR (version 2.3.0) with a mismatch to match ratio of 0.02 allowing multiple mappings of up to 100 loci. The Illumina Human Body Map contains data from 16 different healthy tissues or cell types: adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testis, thyroid, and white blood cells. The dataset contained single end 75 nt reads. Each sequence file contained an average of 78.7 million reads with a minimum of 64 million and a maximum of 83 million sequence reads. A mean of 0.67 % of the reads was uniquely aligned to the viruses (minimum: 0.5 %, maximum: 0.8 %) (Table 3-1). On average, another 0.0042 % of the reads did align to multiple viral loci.

Table 3-1: Alignment statistics of human samples on virus genomes. Listed are samples of the Illumina Human Body Map 2.0 project (GEO accession GSE30611), the number of unaligned reads, the absolute number of reads that aligned uniquely to viruses and their percentage, the absolute number of reads mapping to multiple loci on viruses and their percentage of total reads.

Sample	Number of reads	Uniquely mapped to viruses	Uniquely mapped %	Multi mappers	Multi mappers %
adipose	76,060,765	601,728	0.79	3,889	0.01
adrenal	75,917,971	604,070	0.80	4,004	0.01
brain	64,130,391	323,755	0.50	1,990	< 0.01
breast	76,929,180	589,400	0.77	3,925	0.01
colon	79,990,865	550,526	0.69	3,589	< 0.01
kidney	79,111,516	555,342	0.70	3,526	< 0.01
heart	76,609,389	407,519	0.53	2,403	< 0.01
liver	77,297,801	458,388	0.59	2,758	< 0.01
lung	80,893,859	567,816	0.70	3,647	< 0.01
lymph node	81,606,337	539,604	0.66	3,321	< 0.01
prostate	82,933,631	600,859	0.72	3,856	< 0.01
skeletal muscle	82,648,918	576,918	0.70	3,657	< 0.01
white blood cell	82,443,339	519,143	0.63	3,064	< 0.01
ovary	80,777,009	522,513	0.65	3,367	< 0.01
testis	81,766,985	507,646	0.62	3,234	< 0.01
thyroid	80,030,048	539,836	0.67	3,368	< 0.01

Virus genome sequences of the covered regions were located, extracted, and printed in FASTA format. Mapped loci were also reported in BED format. Between 31 (for heart) and 135 (for colon) viral loci (mean 82) were mapped by data from the Human Body Map 2.0 to the virus database. The alignment revealed mappings to the complete genome of Enterobacteria phage phiX174 (NC_001422.1) in all samples with a genome coverage of 100 %. According to the sample annotation deposited at GEO, all lanes included phiX174 DNA spike-ins at levels of roughly 0.5 % as a sequencing quality control. Following in terms of mapped genome length are species of phages like the Enterobacteria phage P1 (NC_005856.1) and Escherichia phage TL-2011b (NC_019445.1). Both have partially similar genome sequences to phiX174 and a maximum viral genome coverage (VGC) of 1.27 % and 0.6 %, respectively.

The breast tissue sample of the Human Body Map showed expression of a section of 414 bases of Human adenovirus C (HAdV-C, NC_001405.1). In total 1,242 reads mapped to HAdV-C, compared to between 1 and 21 reads in the other tissue samples.

The mapped region was the section from nucleotide 558 to 972, corresponding to the coding sequence of the first exon of the E1A gene of HAdV-C. A BLAST analysis did not reveal significant similarities to human genome or transcriptome sequences.

Other viruses showed some sequence similarities to human expressed genes, like Abelson murine leukemia virus, Woolly monkey sarcoma virus, or Shamonda virus segment S. However, these viruses show covered regions of only up to 100 nt in a stretch with counts from 5 to 500 reads. Other virus genomes with alignments had stretches of low sequence complexity with repetitions of one or few nucleotides. Some representative examples are given below:

```
>NC_006641.1_part
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTATG
>NC_012783.2_part
ACACACACACACACACACACACACACACACACACACACACACACACCGT
>NC_020101.1_part
GTCTGGCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
>NC_019491.1_part
CACACACACACACACACACACACACACACACACACACACACACACACCTT
>NC_006966.1_part
CATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCAAT
>NC_002794.1_part
CACACACACACACACACACACACACACACACACACACACACACACACGA
>NC_021312.1_part
CCTCTTCATCCTCCTCTTCCTCTTCCTCTTCCTCTTCCTCTTCCTCTTCTT
```

Depending on the quality of the sequencing data, these loci of low sequence complexity might have accumulated a large proportion of the short reads. Hence, I sought a method to identify and mark low complexity regions in viral genomes (see chapter 3.4).

Host alignments and filtering of reads

Pre-alignment filtering of low quality or low complexity reads is time consuming and might affect virus detection or viral expression profiles. Host-derived reads can be filtered by host subtraction, if the data is analyzed for host gene expression beforehand. Then the remaining reads can be processed for viral identification. However, if the data is only analyzed for viral content, host subtraction will consume great amounts of time and computing power.

To assess the impact of both factors, I compared the alignments using a dataset with known viral content. The selected dataset consisted of seven murine lung samples, four of which were infected with severe acute respiratory syndrome coronavirus (SARS-

CoV, NC_004718.3) [56]. Data of all seven samples were aligned to virus genomes alone, and a mouse-virus hybrid reference genome, each once with pre-alignment filtering using PRINSEQ, and once without filtering (Figure 3-1).

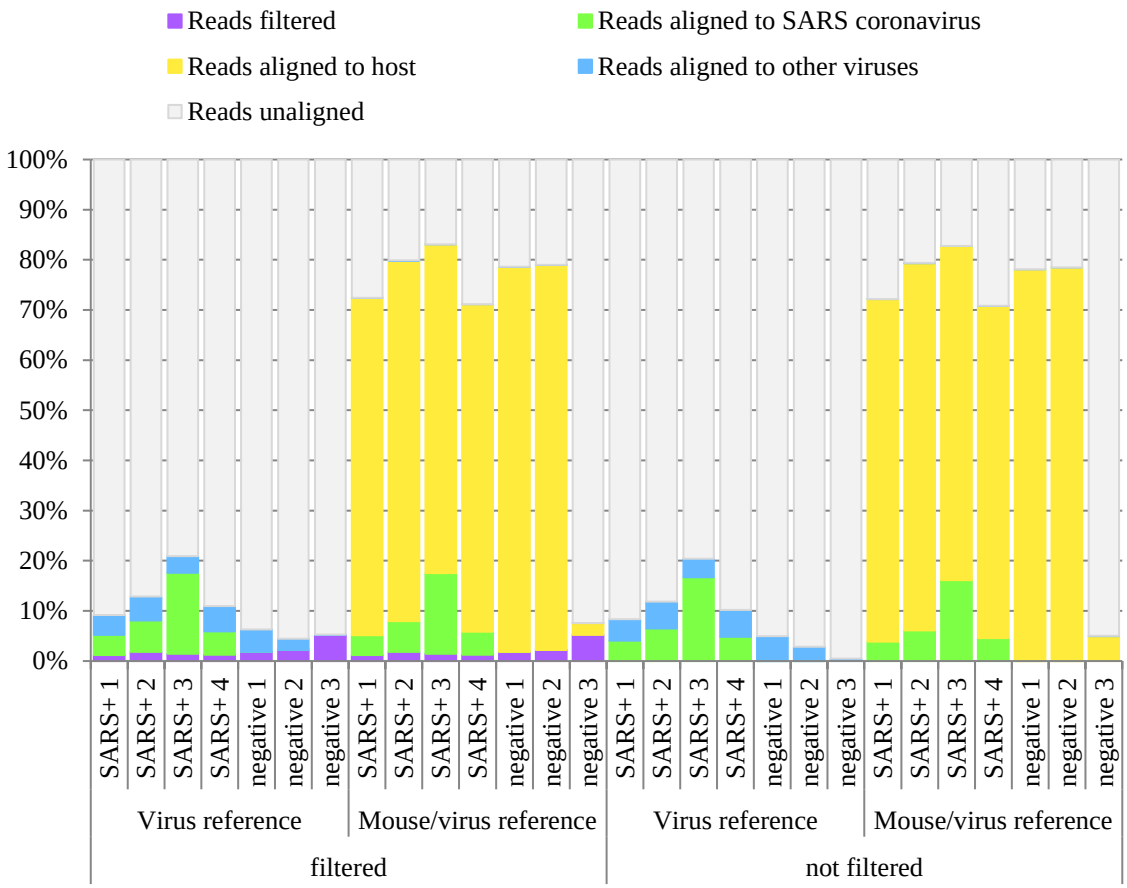


Figure 3-1: Alignment statistics of SARS mouse samples. Figures are given as a percentage of the total read counts for each sample. Seven mouse samples were analyzed, four of which were positive for SARS coronavirus (SARS+), and three were virus free (negative). Reads were aligned to virus reference genomes or a mouse/virus hybrid reference genome with or without prior filtering of low quality reads.

Each sample consisted of 31 million reads on average, with a maximum of 41.7 million (sample 'SARS+ 3') and a minimum of 13.8 million for sample 'negative 3'. Alignments of sample 'negative 3' suggest corrupt sequence data of some kind with only 5 % of reads mapping to the mouse genome. This was not investigated further. The alignment statistics revealed remarkable differences in the alignments to viruses other than SARS-CoV between the virus-only alignment and the mouse/virus reference alignment (Figure 3-1, blue category). An average of 1.7 million reads mapped to other virus genomes in the first case, compared to 5,410 reads on average in the latter case. Pre-alignment filtering using PRINSEQ depleted 580,754 reads

(1.87 %), on average (minimum: 341,376; maximum: 800,463). However, the usage of PRINSEQ for filtering added around 200 % of computing time, compared to alignment and virus detection without filtering.

Filtering the sequencing data for low quality reads showed only insignificant improvements for the virus detection. Additionally, computing time increased massively with the filtering process. However, mapping the sequence reads to the host/virus hybrid reference genomes almost completely diminished alignments to viruses other than the expected SARS-CoV. Thus, further analyses of viruses in mammalian samples were performed by alignment to host/virus reference genomes without filtering, if host sequences were not subtracted beforehand. Using a spliced alignment tool like STAR, mapping to a hybrid genome would also enable the downstream localization of expressed virus insertion sites, if applicable.

3.1.2 First Performance Review

For the first performance review of VirusID we used cell line supernatant samples from our collaborator Baxter International. We received twenty liquid samples from Baxter International in a blinded fashion. Thus, we knew neither the cellular background, nor the virus spike-in until after the unblinding. As described in section 2.1.1, the samples were labelled 1A to 18A, 1C, and 10C. Measurements revealed very low concentrations of nucleic acids, mostly below the detection limit. This necessitated a modified handling protocol for the laboratory platform of VirusID (evaluation and sample handling by Jos de Graaf). Two samples were picked randomly (samples 12A and 14A) to test vacuum drying against the standard of preparation. Of each of the two samples, 5 µL were prepared without the vacuum drying procedure (samples 12A_5 and 14A_5). The remaining 45 µL were vacuum concentrated and then taken up in 5 µL of buffer (samples 12A conc. and 14A conc.). Nucleic acid yield and quality was higher in the concentrated samples (Figure 3-2). However, virus detection worked equally well in both and neither method exceeded the other in data quality or alignment statistics (Table 3-2).

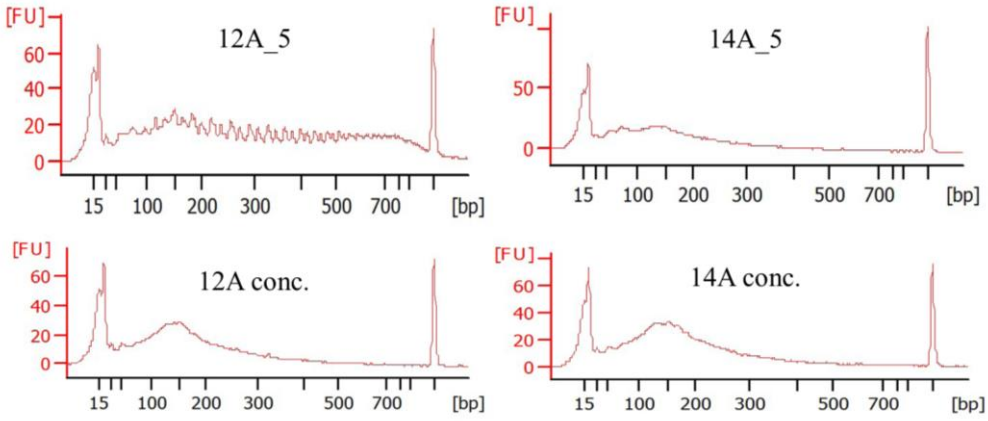


Figure 3-2: Bioanalyzer results for samples 12A and 14A. We randomly picked two samples out of all samples (12A and 14A). Five microliters from each were aliquoted and processed without increasing the concentration (12A_5 and 14A_5). The remaining 45 µl were vacuum dried and taken up in 5 µl of buffer for further processing (12A conc. and 14A conc.). The latter samples led to a higher nucleic acid yield and higher quality.

After these results, we decided to process all other samples using the vacuum concentration protocol. However, we were not able to detect any nucleic acids in five of the twenty samples during sample preparation (1A, 1C, 10A, 10C, 6A). These samples were excluded from further processing and not sequenced. The remaining liquid samples were vacuum dried, re-suspended and the nucleic acids were amplified in addition to the standard RNA-Seq library preparation protocol (see section 2.1.1). Libraries were sequenced to a length of 50 base pairs (bp), single-end or paired-end reads. The NGS run resulted in 23.1 million sequences per sample on average, with a minimum of 13.2 million and a maximum of 31 million reads. NGS sequences were aligned to virus genomes without host subtraction, because host or background genomes were unknown. The cutoff for a detected virus was raised to a VGC of 50 % to account for higher expected noise. This has been experienced in prior alignments without host subtraction (Figure 3-1). The raised cutoff roughly represented the median of all virus genome coverages that were greater than 0 % plus two standard deviations. A mean of 1.8 % of the reads were aligned to the virus genomes, with a minimum of 0.06 % and a maximum of 5.64 % (Table 3-2).

VirusID detected four different virus species in the fifteen analyzed samples (Table 3-2): Minute virus of mice (MVM) in three samples, Mammalian orthoreovirus III

(ReoIII) in three samples, Encephalomyocarditis virus (EMCV) in five samples, and Pre-Xenotropic murine leukemia virus-related virus 1 provirus (PreXMRV-1) in two samples. The results showed more than one virus in six of the samples. No viruses was reported in samples 9A and 18A, both spiked-in with X-MuLV (10^4/50 µL).

Table 3-2: Baxter samples and detected viruses using VirusID. Number of available reads after sequencing are given as well as the percentage of mapped reads to virus genomes database. I calculated the percentage of all virus-mapped reads to the target virus. The right-hand section of the table (black heading) shows virus spike-in and cellular background for comparison (information received after unblinding). PE: paired-end sequenced; NA: not applicable; CHO: Chinese hamster ovary cell line; VERO: African green maquaque kidney epithelial cell line; MVM: Minute virus of mice; EMCV: Encephalomycarditis virus; ReoIII: Mammalian orthoreovirus III; X-MuLV: Xenotropic murine leukemia virus; PreXMRV-1: Pre-Xenotropic murine leukemia virus-related virus 1 provirus.

							Revealed after unblinding		
Sample	**Detected virus**	**Further findings**	**# of reads**	**% mapped**	**% on target**	**Comment**	**Cellular background**	**Virus load**	**Virus spike-in**
1A	none		NA	NA	NA	not sequenced	VERO	NA	none
1C	none		NA	NA	NA	not sequenced	VERO	NA	none
10A	none		NA	NA	NA	not sequenced	CHO	NA	none
10C	none		NA	NA	NA	not sequenced	CHO	NA	none
2A	MVM	EMCV	22,463,879	0.84	59.17		VERO	10^5/50µl	MVM
6A	none		NA	NA	NA	not sequenced	VERO	10^4/50µl	MVM
11A	MVM	EMCV	18,342,445	0.73	54.04		CHO	10^5/50µl	MVM
15A	MVM	EMCV	25,029,198	1.09	95.52	PE	CHO	10^4/50µl	MVM
3A	ReoIII	EMCV	29,422,619	0.08	35.87	PE	VERO	10^5/50µl	ReoIII
7A	ReoIII	EMCV, MVM	20,595,706	0.29	6.72		VERO	10^4/50µl	ReoIII
12A	ReoIII		17,857,979	1.47	2.84		CHO	10^5/50µl	ReoIII
12A (12_5)	ReoIII		25,088,838	4.46	1.81		CHO	10^5/50µl	ReoIII
16A	EMCV		28,355,975	0.06	0.00	PE	CHO	10^4/50µl	ReoIII
4A	EMCV		27,910,969	4.11	99.26	PE	VERO	10^5/50µl	EMCV
8A	EMCV		20,640,967	4.43	95.38		VERO	10^4/50µl	EMCV
13A	EMCV		31,031,050	0.98	93.20	PE	CHO	10^5/50µl	EMCV
17A	EMCV		30,861,114	0.16	75.83	PE	CHO	10^4/50µl	EMCV
5A	PreXMRV-1	EMCV	17,139,367	0.88	68.95		VERO	10^5/50µl	X-MuLV
9A	none		17,804,392	0.63	8.32		VERO	10^4/50µl	X-MuLV
14A	PreXMRV-1		24,219,178	4.64	2.32		CHO	10^5/50µl	X-MuLV
14A (14_5)	PreXMRV-1		22,597,772	5.64	5.09		CHO	10^5/50µl	X-MuLV
18A	none		13,226,228	0.30	20.07		CHO	10^4/50µl	X-MuLV

After analysis, the samples were unblinded using the sample information provided by our collaborator. The supernatants of two cell lines were used for the creation of the samples: an African green monkey kidney cell line (VERO) and a Chinese hamster ovary cell line (CHO). Both represent common biotechnological production cell lines and are thus suitable as background for a virus detection test. Four different virus genomes were spiked in at two different concentrations, each. They were, namely, the Minute virus of mice (MVM), Reovirus type III (ReoIII), Encephalomyocarditis virus (EMCV), and the Xenotropic murine leukemia virus (X-MuLV) (Table 2-1).

VirusID had reported viruses in thirteen samples (not counting the 5 µL aliquots of samples 12A and 14A). Of these, twelve were assigned correctly. However, in six of the samples VirusID had detected more than one virus. Remarkably, all additional viruses were also contained in the panel of spike-in virus species. A wrong virus was identified in sample 16A. Here, not a single sequence of the expected virus spike-in ReoIII was detected. EMCV showed the highest VGC in this sample, indicating some sort of mix-up at some stage of the process. The two negatively identified X-MuLV samples (9A and 18A) showed low signal of PreXMRV-1 below the set cutoff.

Of the aligned reads, between 0.00 % and 99.26 % were on-target hits (mean 42.61 %). Sample 4A contained the most correct hits with only 0.74 % off-target hits of all mapped reads. In total, the EMCV-positive samples showed the highest on-target rate.

The detection of viruses in the provided cell line supernatants was part of a comprehensive evaluation of broad virus detection methods in different laboratories [52]. The study was conducted by scientists from the Global Pathogen Safety Department of Baxter International. Four investigators received the samples in a blinded fashion (Table 3-3). The samples were analyzed with different detection methods such as PCR-electrospray ionization mass spectrometry (PCR-ESI/MS) by IBIS Biosciences (Carlsbad, CA, USA), a microarray-based approach by the Lawrence Livermore National Laboratory (LLNL, Livermore, CA, USA), and two NGS approaches, by TRON (Mainz, Germany) and Eurofins Medigenomics (Ebersberg, Germany). Both NGS approaches used the Illumina HiSeq 2500 sequencing platform.

Table 3-3: Evaluation of four different broad virus detection methods. Table adapted from [52]. Spike-in: (✓) sample correctly identified as virus positive or negative; (–) incorrect results; Further findings: (+) unexpected virus(es) identified. MVM: Minute virus of mice; EMCV: Encephalomycarditis virus; ReoIII: Mammalian orthoreovirus III; X-MuLV: Xenotropic murine leukemia virus.

Virus	Concentration [log copies / mL]	Medium	IBIS Biosciences		LLNL		TRON		Eurofins	
			Spike-in	Further findings	Spike-in	Further findings	Spike-in	Further findings	Spike-in	Further findings
MVM	4.7 / 5.0	1	✓		-		✓	+	-	+
		2	✓		✓		-		✓	
	5.5 / 5.6	1	✓		✓		✓	+	-	+
		2	✓		✓		✓	+	-	+
ReoIII	4.1 / 4.6	1	✓	+	-	+	✓	+	✓	
		2	✓	+	-	+	✓	+	✓	
	5.4 / 6.0	1	✓		✓	+	-	+	✓	
		2	✓		-		✓		✓	
EMCV	5.8 / 6.2	1	✓		✓		✓		✓	
		2	✓	+	✓		✓		✓	
	7.3 / 7.5	1	✓		✓		✓		✓	+
		2	✓	+	✓		✓		✓	+
X-MuLV	4.8 / 5.3	1	-		-		-		✓	+
		2	-		✓		-		✓	+
	6.5 / 6.8	1	-		✓	+	✓		✓	
		2	-		✓	+	✓		-	+
Negative controls	NA	1	✓		✓		✓		✓	
		2	-	+	✓		✓		-	+

The comparison revealed that none of the methods outperformed the others in sensitivity or specificity (Table 3-3). Each method identified the majority of the samples correctly (marked '✓'). Similarly, each method had its challenges detecting some of the viruses or negative controls correctly ('-'). All methods reported a variety of false positive results ('+', between 20 and 45 %).

The evaluated broad virus detection methods proved to be suitable to detect even relatively low concentrations of virus genomes. It seems noteworthy that the PCR-based and the microarray-based method reported several samples correctly only after unblinding and by changing the used amplification method [52]. VirusID only failed to label four samples correctly (compared to five in all other methods), one of which was not sequenced (MVM sample 2).

3.1.3 Final Benchmark

The Ci3 project plan for VirusID implied an iterative optimization process after the first performance review. Several improvements distinguished the final version from the version in section 3.1.2. These include the filtering for taxonomically closely related genomic sequences, the addition of the fraction of viral reads (FVR) to the detection limit of viral genome coverage, the inclusion of Pysam for faster processing of the alignment files, and the restriction of ambiguous alignments during mapping of the reads.

The final milestone of the project plan was to trial VirusID for another benchmark. A collaboration with Luis Terán Juárez from the National Institute for Respiratory Diseases, Mexico (Instituto Nacional de Enfermedades Respiratorias, INER) enabled us to obtain nasal aspirates from children with viral infections of the respiratory tract. Fifty samples were virus typed at the INER by polymerase chain reaction (PCR) or immunofluorescence (IF). Eighteen of the shipped samples were selected for further sequencing and further analyses (see section 2.1.2).

The samples were sequenced on three lanes of an Illumina 2500 flow cell. Sequencing yielded between 23.3 and 37.7 million paired-end reads (2x50 nt) per sample (mean: 31.4 M). I mapped the reads to a human-virus hybrid reference genome. The reference included human genome hg19/GRCh37 and the viral genomes database accessed 2014-12-05 (see chapter 2.2). Between 78 and 90 % of the reads aligned to the hybrid reference. Between 23 thousand (0.04 %) and 1.2 million (1.77 %) reads mapped to the virus genomes. Table 3-4 shows the results of the virus typing using VirusID.

Table 3-4: VirusID results for the nasal aspirate samples. HRSV: Human respiratory syncytial virus, HERV-K113: Human endogenous retrovirus K113, ScV-M1: Saccharomyces cerevisiae killer virus M1. Color code: green: > 20 % and concordant with virus typing; yellow: < 20 % and concordant with virus typing; blue: > 20 % and contrary to virus typing.

Sample ID	Virus Typing Result	Detected Viruses (VirusID)	VGC [%]
534	HRSV	HERV-K113	14.49
536	Adenovirus	HERV-K113	6.19
537	Adenovirus	HERV-K113 Human rhinovirus 89	12.13 6.64
545	Influenza B virus, Parainfluenza virus type 2, Parainfluenza virus type 3	HERV-K113 ScV-M1	7.72 6.27
546	Rhinovirus	HERV-K113	12.01
548	Influenza A virus	HERV-K113	13.87
549	Rhinovirus	HERV-K113 ScV-M1	11.59 6.00
552	HRSV	Human coronavirus OC43 HERV-K113 Streptococcus phage K13	19.56 11.74 8.38
553	Rhinovirus	HERV-K113 Human rhinovirus 89	17.91 7.62
554	Rhinovirus	Human coronavirus OC43 HERV-K113 ScV-M1	87.72 11.31 5.89
565	Rhinovirus	Torque teno virus 7 HERV-K113 ScV-M1 Human rhinovirus C	99.95 6.86 6.16 6.16
568	Rhinovirus	HERV-K113 Human enterovirus D	8.02 7.69
572	Rhinovirus	HERV-K113 Human rhinovirus 14 Human coronavirus OC43	14.00 13.09 5.55
576	Rhinovirus	Human rhinovirus 89 HERV-K113 HRSV ScV-M1	15.65 7.52 6.62 5.55
577	Rhinovirus	Human rhinovirus 89 HERV-K113 HRSV	28.89 10.61 5.73
584	Corona Virus	Human coronavirus OC43 HERV-K113	31.15 10.38
585	HRSV, Rhinovirus	HRSV HERV-K113	75.36 12.69
592	Corona Virus	Human coronavirus OC43 HERV-K113	99.15 8.86

The analysis revealed sequences of Human endogenous retrovirus (HERV) K113 (HERV-K113, NC_022518.1) in all samples (Table 3-4) with genome coverages between 6 and 18 % (568 – 1705 nt). This was an expected result and had been experienced in the majority of human samples so far. HERV-K113 is the only Human

endogenous retrovirus in the virus genome database. Sequences originating from any type of HERV in the human genome are likely to map to similar loci on K113.

Only six of eighteen samples (33.3 %) could be typed successfully *in silico*. Four out of the six results match the original virus typing results (Samples 577, 584, 585, and 592). These results are highlighted in green in Table 3-4. Human coronavirus OC43 (HCoV-OC43, NC_005147.1) sequence was measured in sample 592 with high viral load of 14,235 ppm and 99.2 % viral genome coverage. The same virus was detected in sample 584 with only 31.2 % genome coverage (FVR: 951 ppm). The Human respiratory syncytial virus that had been successfully detected in sample 585 also showed high expression and viral load (VGC: 75.36 %, FVR: 7,564 ppm). Sample 577 contained a low load of Rhinovirus (FVR: 52.14 ppm) and only less than a third of the genome was expressed (VGC: 28.9 %).

HCoV-OC43 was also measured in sample 554 with a viral read fraction of 23,787 ppm and a genome coverage of 87.7 %. Torque teno virus 7 in sample 565 has the highest genome coverage (99.95 %), but a rather low viral load (198 ppm). Both results were significant but different from the original virus typing (marked blue in Table 3-4).

Further, in four samples the expected virus was detected with low signal only (highlighted in yellow). Some of the other low signal results are potentially false positives, like the Saccharomyces cerevisiae killer virus M1 (ScV-M1, NC_001782.1). This fungal virus genome contains a 110 nt poly-Adenine stretch. All reads mapping to ScV-M1were aligned to this region.

Software benchmark of VirusID

Since these results were difficult to interpret, further virus detection software was considered for comparison and understanding of the results. One of the first available software for the detection of pathogenic sequences from NGS data was PathSeq [82]. The software was implemented as a gradual sequence subtraction algorithm with a final *de novo* assembly of leftover sequence reads. In a first step, low quality, low complexity, and duplicate sequences were removed. This is followed by a three-step host-subtraction of potential human reads screening genomic and transcriptomic references. Potential pathogen-derived reads then enter the analytical phase including *de novo* assembly and metagenomic analysis. In contrast, the tool RINS, the Rapid

Identification of Non-human Sequences [83], was designed to run without host subtraction. The algorithm first directly maps 25-mers of single-end sequence reads against the query dataset. These are then further filtered for human homology. Potential pathogen-derived sequences are assembled into contigs in the end. Other algorithms like VirusSeq [68], ViralFusionSeq [84], or VirusFinder [85] focus on virus detection and detection of viral integration sites. Further, I considered VIRANA [86] and ContextMap [61], the first using the STAR aligner against a human-viral reference followed by assembly and phylogenetic annotation. The latter is a context-based method trying to predict the most probable alignment per read out of several positive matches. ContextMap also reports the scores of virus genome coverage and fraction of viral reads.

In preparation of the software benchmark, the packages were downloaded in January 2015 in their respective versions. However, the usage of the majority of the algorithms was compounded by the fact that they, or software they depended on, could not be sufficiently installed or run on our servers. I was able to install RINS properly, but failed to run the code due to its inability to incorporate the Trinity package, which is required for the proper execution of the software. Hence, VirusID was only compared to ContextMap (V2.3.3) and VirusSeq (retrieved 2015-01-05). I executed both methods using the same viral database as for VirusID. Separate reference indices were created for both, human and viruses, for both algorithms. Both methods were applied to the same sequence data as analyzed with VirusID. Table 3-5 contains the results for all three tools.

Table 3-5: Results of the software benchmark for the detection of viruses. HRSV: Human respiratory syncytial virus, HERV-K113: Human endogenous retrovirus K113, SsDRV: Sclerotinia sclerotiorum debilitation-associated RNA virus, HCV6: Hepatitis C virus genotype 6, HCV2: Hepatitis C virus genotype 2. Color code: green: concordant with virus typing; yellow: low VGC, but concordant with virus typing; blue: contrary to virus typing; purple: concordant with VirusID result.

Sample ID	Virus Typing	VirusID	ContextMap	VirusSeq
534	HRSV	HERV-K113	HERV-K113	HCV2
536	Adenovirus	HERV-K113	SsDRV	HCV2
537	Adenovirus	HERV-K113 Human rhinovirus 89	HCV6	HCV2
545	Influenza B virus, Parainfluenza virus type 2, Parainfluenza virus type 3	HERV-K113 ScV-M1	HERV-K113	HCV2
546	Rhinovirus	HERV-K113	SsDRV	HCV2
548	Influenza A virus	HERV-K113	HERV-K113	HCV2
549	Rhinovirus	HERV-K113 ScV-M1	HERV-K113	HCV2
552	HRSV	Human coronavirus OC43 HERV-K113 Streptococcus phage K13	HERV-K113	HCV2
553	Rhinovirus	HERV-K113 Human rhinovirus 89	HERV-K113	HCV2
554	Rhinovirus	Human coronavirus OC43 HERV-K113 ScV-M1	Human coronavirus OC43	HCV2
565	Rhinovirus	Torque teno virus 7 HERV-K113 ScV-M1 Human rhinovirus C	SsDRV	HCV2
568	Rhinovirus	HERV-K113 Human enterovirus D	SsDRV	HCV2
572	Rhinovirus	HERV-K113 Human rhinovirus 14 Human coronavirus OC43	HERV-K113	HCV2
576	Rhinovirus	Human rhinovirus 89 HERV-K113 HRSV ScV-M1	SsDRV	HCV2
577	Rhinovirus	Human rhinovirus 89 HERV-K113 HRSV	HERV-K113	HCV2
584	Corona Virus	Human coronavirus OC43 HERV-K113	Human coronavirus OC43	HCV2
585	HRSV, Rhinovirus	HRSV HERV-K113	HRSV	HCV2
592	Corona Virus	Human coronavirus OC43 HERV-K113	Human coronavirus OC43	HCV2

The output of ContextMap reports all viruses with mapped sequence reads. I only considered the best-reported virus for this comparison. ContextMap confirmed the

VirusID results of the correctly identified samples 584, 585, and 592 (Table 3-5). Additionally, ContextMap confirmed the result of VirusID of HCoV-OC43 in sample 554, which was not detected by classical virus typing. However, it failed to identify Torque teno virus 7 in sample 565. Although a high confidence level had been computed, the coverage of only 46 % and the low read count compared to other viral genomes let it come in at fifth place. ContextMap also reports HERV-K113 as most probable in eight out of eighteen samples. The reported genomes of Sclerotinia sclerotiorum debilitation-associated RNA virus (NC_007415.1) and Hepatitis C virus (HCV) genotype 6 (NC_009827.1) both contain loci with repetitions of adenines or thymines, as described for the Saccharomyces cerevisiae killer virus in the VirusID section previously.

VirusSeq reported the detection of HCV genotype 2 (NC_009823.1) for all samples. This virus genome features a stretch of repetitions of thymine over 84 bases followed by a thymine-rich region with only few interspersed cysteines or adenines. VirusSeq reports the virus genome with the highest number of alignments as a potential hit. Here, for each sample several hundred thousand sequence reads mapped to the poly-T region resulting in reports of only HCV in each sample. The reads likely originate from the poly-A tail-derived cDNA.

Overall, only 3 of 18 (16.7 %) classical virus typing results could successfully be confirmed in congruence by two *in silico* virus detection methods: VirusID and ContextMap. One additional sample was reported to contain Human coronavirus consistently by both methods. In addition, VirusID reported several viruses in concordance with the classical virus typing, most of which were only detected with low VGC. VirusSeq was not able to identify any virus correctly.

3.2 Routinely Screening of RNA-Seq Data

Since the implementation of VirusID, the code has been in regular use for all RNA-Seq data processed in the bioinformatics core facility of TRON. All in-house sequenced samples were screened as well as external data retrieved from different public or restricted access databases. In three years (2013-2015), the bioinformatics core facility of TRON had analyzed RNA-Seq data from 11,900 human samples and 254 murine samples, including human and murine cell lines. A majority of the human samples (67.16 %) contained measurable amounts of HERV-K113. Almost 6,000

samples contained sequences of phiX-174. VirusID detected different human viruses in several samples: EBV (130 samples), HPV16 (114 samples), HPV18 (40 samples), and HHV1 (33 samples). One hundred twenty five of the samples contained viruses associated with cell line contaminations, such as murine or simian viruses. Results were similar for the murine samples. Three of the murine samples contained signs of HERV-K113. Further analysis revealed contamination with human sequences or human material of the sample. Here, I will show exemplary results from the standard screening of RNA-Seq data.

Cell line replicates

Cell lines are an important tool in biological research, for instance as disease models. Many groups have performed genomic characterization of their used cell lines and published their results including the sequencing data. Here, I highlight the viral content of the breast cancer cell lines MDA-MB-231 and MCF7. Sequencing data was available from four different laboratories.

RNA-Seq data for both cell lines were available from different sources (Table 3-6). The transcriptomes of both cell lines had been sequenced in-house at TRON paired-end with 100 nt. Further, the SRA projects SRP005601 and SRP026537 contained both cell lines and were sequenced 2x50 nt and 2x76 nt, respectively. Finally, SRA project SRP013022 contained sequencing data of MDA-MB-231 and SRP003186 contained data of MCF7. These samples were sequenced paired-end 2x100 nt and 2x50 nt, respectively.

To reduce the risk of potential misidentification, the samples were HLA-typed using seq2HLA [66]. Results were definite in the case of MDA-MB-231 with good confidence scores and there were only minor discrepancies due to locus ambiguity (Appendix Table A-3). However, due to very low expression of HLA-B in MCF7, the results were less clear for this cell line. HLA-A and HLA-C did match for all the MCF7 samples. Hence, cell line identity can be assumed for all four datasets of both cell lines.

Table 3-6: Sources and virus content of breast cancer cell line data. Listed are the respective projects, sample accession numbers, cell line names, and the detected viruses. NA: Not applicable.

Project	Accession number	Cell line	Detected virus
TRON	NA	MDA-MB-231	-
SRP013022	SRR496398	MDA-MB-231	-
SRP005601	SRR097790	MDA-MB-231	-
SRP026537	SRR925726	MDA-MB-231	Murine type C retrovirus
TRON	NA	MCF7	-
SRP003186	SRR064286	MCF7	-
SRP005601	SRR097789	MCF7	-
SRP026537	SRR925723	MCF7	Murine type C retrovirus

VirusID detected MCRV in two samples, one of the MDA-MB-231 samples and one of the MCF7 samples (Table 3-6). Both samples were sequenced within the same project. All other replicates were determined to be virus-free. The six public samples were also part of further studies in this thesis (chapter 3.5).

Virus detection in 1,082 cell lines

At TRON, we have developed different bioinformatic workflows to further annotate samples regarding their mutations, HLA types [87], expression levels, and viral content. We downloaded the publicly available raw data of two cell line cohorts, Cancer Cell Line Encyclopedia (CCLE) [88] and Klijn *et al.* [89], with the aim to annotate these cell lines and make the results publicly available.

The CCLE dataset contained RNA-Seq data from 781 cell lines; Klijn *et al.* had sequenced 675 transcriptomes, out of which 374 samples overlapped with CCLE. Hence, we only used the 301 unique cell lines from the latter dataset. In total, we analyzed 1,082 cancer cell lines. All produced results and further available annotation were integrated into a cell line annotation database called the TRON Cell Line Portal (TCLP) [69]. The portal is available and consultable at http://celllines.tron-mainz.de.

We searched the 1,082 cell lines for viruses using VirusID. A large majority of all samples expressed HERV-K113 (921 samples, 85 %) (Table 3-7). Other detected human viruses included HTLV-1, EBV (HHV4), Hepatitis B virus (HBV), HPV18 and

HPV16, and Kaposi's sarcoma-associated herpesvirus (KSHV), also known as Human Herpesvirus 8 (HHV8).

Table 3-7: Human viruses in the TRON Cell Line Portal (TCLP). VirusID results for 1,082 cell lines from the CCLE [88] and from Klijn *et al.* [89].

Virus	Occurrence
NC_022518.1; Human endogenous retrovirus K113 (HERV-K113)	921
NC_001436.1; Human T-lymphotropic virus 1 (HTLV-1)	19
NC_007605.1; Human herpesvirus 4 (HHV4)	16
NC_003977.1; Hepatitis B virus (HBV)	10
NC_001357.1; Human papillomavirus – 18 (HPV18)	4
NC_001526.2; Human papillomavirus type 16 (HPV16)	2
NC_009333.1; Human herpesvirus 8 (HHV8)	1

As well as human viruses, different murine retroviruses were also detected. VirusID also reported detected Parainfluenzavirus 5 (PIV5) and the Enterobacteria phage phiX174.

The data presented in this section demonstrated the large-scale applicability of viral screening to RNA-Seq datasets. We applied the method to in-house sequenced samples as well as to downloaded public datasets. The correct interpretation of the results required further investigation in some cases, but the output of VirusID allowed for detailed analyses, for instance based on expressed virus contigs and alignment profiles.

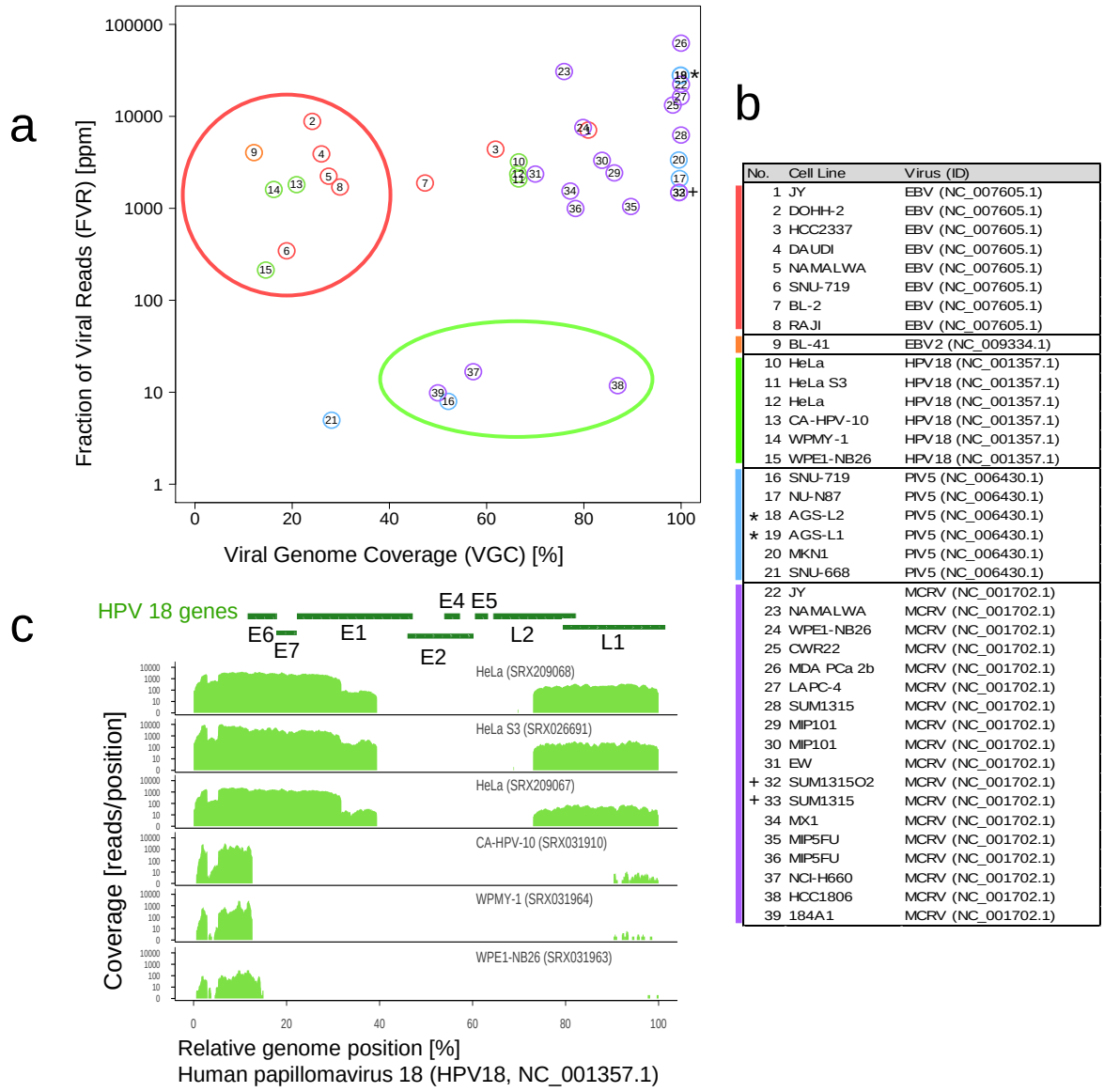
3.3 Viral Abundance and Transcription Profile

RNA-Seq has demonstrated reproducible results in various applications using mammalian transcriptomes [71, 90]. In the preceding chapters, I have shown its applicability to the detection of expressed viruses. Alignment profiles suggest a variety of expression profiles. To further investigate the possibility of utilization in this regard, I first explored a subset of the cell lines reported more thoroughly in chapter 3.5.

VirusID uses different measures to determine the presence of viral content in host RNA-Seq data. The fraction of viral reads (FVR) in parts per million (ppm) of all sequence reads is a measure of viral load in a sample. The viral genome coverage (VGC) measures the relative amount of virus genome expressed. Both measures are uncorrelated (Figure 3-3a), as some viruses only express a fraction of their genome

with a high viral load (red cluster), and vice versa (green cluster). However, both measures are important parameters for interpreting viral expression.

Some viruses, for example MCRV (Figure 3-3a and b, purple dots), showed an overall high genome expression (VGC > 50) with high variability in viral load. The expression of HPV18 in HeLa is very consistent and reproducible in replicates (VGC around 66 %, samples 10, 11, and 12). However, CA-HPV-10 (sample 13), WPMY-1 (sample 14), and WPE1-NB26 (sample 15) showed expression of HPV18 genes E6, E7, and E1 only. VirusID would have missed the expression of HPV18 in these samples due to a VGC cutoff. The first two of these samples were transfected using only HPV18 genes E6 and E7, which are necessary and sufficient for the immortalization of cells [91]. Hence, the VGC of these cell lines is lower with ranges from 14.5 % to 21 % but with comparably high FVRs. By generating alignment profiles (Figure 3-3c), I was able to confirm these findings. The incorporated virus database contained 48 different Human papillomavirus genomes. We only detected signals from Human papillomavirus 18 (HPV18, NC_001357.1) in these six samples.



No.	Cell Line	Virus (ID)
1	JY	EBV (NC_007605.1)
2	DOHH-2	EBV (NC_007605.1)
3	HCC2337	EBV (NC_007605.1)
4	DAUDI	EBV (NC_007605.1)
5	NAMALWA	EBV (NC_007605.1)
6	SNU-719	EBV (NC_007605.1)
7	BL-2	EBV (NC_007605.1)
8	RAJI	EBV (NC_007605.1)
9	BL-41	EBV2 (NC_009334.1)
10	HeLa	HPV18 (NC_001357.1)
11	HeLa S3	HPV18 (NC_001357.1)
12	HeLa	HPV18 (NC_001357.1)
13	CA-HPV-10	HPV18 (NC_001357.1)
14	WPMY-1	HPV18 (NC_001357.1)
15	WPE1-NB26	HPV18 (NC_001357.1)
16	SNU-719	PIV5 (NC_006430.1)
17	NU-N87	PIV5 (NC_006430.1)
* 18	AGS-L2	PIV5 (NC_006430.1)
* 19	AGS-L1	PIV5 (NC_006430.1)
20	MKN1	PIV5 (NC_006430.1)
21	SNU-668	PIV5 (NC_006430.1)
22	JY	MCRV (NC_001702.1)
23	NAMALWA	MCRV (NC_001702.1)
24	WPE1-NB26	MCRV (NC_001702.1)
25	CWR22	MCRV (NC_001702.1)
26	MDA PCa 2b	MCRV (NC_001702.1)
27	LAPC-4	MCRV (NC_001702.1)
28	SUM1315	MCRV (NC_001702.1)
29	MIP101	MCRV (NC_001702.1)
30	MIP101	MCRV (NC_001702.1)
31	EW	MCRV (NC_001702.1)
+ 32	SUM1315O2	MCRV (NC_001702.1)
+ 33	SUM1315	MCRV (NC_001702.1)
34	MX1	MCRV (NC_001702.1)
35	MIP5FU	MCRV (NC_001702.1)
36	MIP5FU	MCRV (NC_001702.1)
37	NCI-H660	MCRV (NC_001702.1)
38	HCC1806	MCRV (NC_001702.1)
39	184A1	MCRV (NC_001702.1)

Figure 3-3: Viral abundance and transcription profiles in cell lines. a) Fraction of viral reads (FVR) in parts per million (ppm) (log-scale) of all mapped sequence reads plotted against the viral genome coverage (VGC). Red circle: high FVR, low VGC, green circle: high VGC, low FVR. b) List of cell lines and viruses in (a) color coded by virus genome. Replicate cell lines AGS (*) and SUM1315 (+) completely overlap in (a). c) Human papillomavirus 18 (HPV18) expression profiles are similar in replicate cell lines (HeLa/HeLa S3) but differ from transfected samples (CA-HPV-10, WPMY-1, and WPE1-NB26). EBV: Epstein-Barr virus, EBV2: Epstein-Barr virus type 2, HPV18: Human papillomavirus 18, PIV5: Parainfluenza virus 5, MCRV: Murine type C retrovirus.

The data in this section demonstrated the importance of different metrics for the determination of the presence of a virus in sequencing data. The used VGC cutoff in VirusID was not suitable for the detection of expression signal from short stretches of

viral genomes. The gained insights fostered the development of a pipeline for the detection and quantification of expressed viral genes in RNA-Seq data (VIRGENE).

3.4 VIRGENE: Virus Gene Expression in Transcriptomic Data of the Host

In the scope of Ci3's VirusID project, I developed a pipeline for the detection and identification of viral signatures in NGS transcriptomic data. Studying a variety of samples and investigating the alignment profiles revealed diverse expression patterns of viruses in cell lines and tumor samples (see chapter 3.3). These patterns indicated diversely expressed viral genes. To exploit these RNA-Seq based data, I developed VIRGENE, a pipeline for the identification and quantification of viral gene expression.

The development of VIRGENE was based on VirusID. I integrated the same sequence aligner, STAR [67], and re-purposed several other parts of VirusID's code. Section 2.5.5 and its Figure 2-5 contain a detailed description of the workflow.

Viral genes and transcripts

Quantification of viral gene expression requires a database of virus genes, transcripts, and/or coding sequences. However, such a database is not readily available, unlike, for instance, for the human genome. Since I already used the NCBI virus genomes, I set out to use a database of genes for the used virus genomes. This led to the development of the script `gbFeatures.py`. Biopython [59] provides a set of tools for the handling of genomic data in common formats and their processing. It also contains routines to extract genomic features for given genomes from the NCBI database using the NCBI genomic identifiers. I used these methods to parse features for each viral genome. Features were extracted if they are labelled either 'CDS' (coding sequence) or 'mRNA'. Exact loci and names for genes, transcripts, and exons were then parsed from GenBank XML format into standard 12-column BED format [92].

Some viruses feature circular genomes. The genomic sequences of such viruses need to be displayed linearly in a FASTA file format. However, some genes in those viruses span the breakpoint of the virus genome sequence either in a positive or negative-strand direction. These transcript loci cannot be properly handled by gene expression quantification software. The used quantification pipeline, the TRON RNA-Seq

Pipeline (TRSP), used a union of all transcripts model for each gene. Hence, the described script stores transcripts overlapping the genomic breakpoints as two transcript parts, one of each coming from either end of the genome. The gene model then incorporates both transcripts to calculate gene expression values.

VIRGENE workflow

VIRGENE was designed as a pipeline incorporating different other software. I detailed the workflow in section 2.5.5 and in Figure 2-5. After preparation of the input data (Figure 2-4), the workflow starts with an alignment using the tool STAR [67]. VIRGENE reads the alignment statistics file and directs the alignment output into TRSP for gene expression quantification. Then, VIRGENE reads the resulting gene expression table and the virus annotation, before starting the filtering process.

As the initial step of filtering, VIRGENE calculates the virus genome coverage for each virus and the viral abundance as a fraction of viral reads. Simultaneously, potential contigs are computed by preparing the consensus sequence of all overlapping aligned sequence reads. Then, low virus signals are filtered out using thresholds for FVR, VGC, and gene expression. Further, if only genes containing low complexity loci are expressed, these viruses are filtered as well. Finally, the software compares closely related virus species and keeps only the species with the highest expression. VIRGENE then writes all results to separate files: scores for each virus genome, all contigs in FASTA format, the genome coverage, expressed virus genes, and all discarded viruses. The final step is a cleanup of all temporary files.

Low complexity in virus genomes

During the final performance review of VirusID (section 3.1.3), three virus genomes appeared prominently as potential false positives (Table 3-4 and Table 3-5): Saccharomyces cerevisiae killer virus M1 (SCKV-M1), Sclerotinia sclerotiorum debilitation-associated RNA virus (SSDRV), and Hepatitis C virus genotype 2 (HCV2). These loci were stretches of repetitions of a single nucleotide or repetitions of two or three nucleotides, like the 38 repetitions of a ‘GTT’ motif in the Pandoravirus dulcis genome. Sequence reads with low sequence complexity in an RNA-Seq dataset would map to these loci, if not filtered *a priori*. However, the low complexity locus in the Pandoravirus dulcis genome spans the majority of the coding sequence of gene N376 (gp0668). In case of expression of such a gene, low complexity sequence reads

would be expected and a pre-filtering of sequence reads would be undesired. Depending on the quality of the sequencing data, these loci of low sequence complexity might accumulate a large proportion of the short reads. Hence, I sought a method to identify and mark low complexity regions in viral genomes. Several algorithms have been developed to overcome these challenges, either by filtering the sequence reads or by masking the respective genomic loci.

The well-known tool RepeatMasker [93] screens genomic sequences for annotated interspersed repeats and low complexity DNA sequences. However, the tool primarily screens for poly-purine/poly-pyrimidine stretches as low complexity regions and would thus miss several of the experienced issues. Other repeat motifs would have to be included in the incorporated Repeatbase database. Hence, prior knowledge of the motifs is required. The software tool PRINSEQ [63] facilitates a different approach. It enables the filtering of low quality or low complexity reads from the user's sequencing data file. However, this would eliminate potentially correct expression values of low complexity genomic regions in viruses.

PRINSEQ also enables the search for reads with low sequence entropy. In information theory, the Shannon entropy of a message can be used to infer the expected value of the information from a sequence of letters [73]. However, the Shannon entropy is based on the numerical distribution of the letters of an alphabet across the message. Given three example sequences of DNA code:

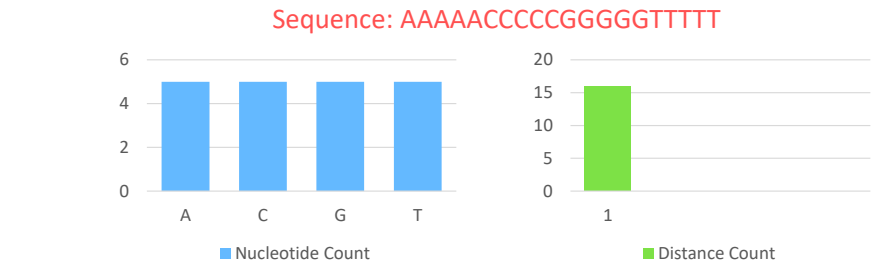
1) ACGTACGTACGTACGTACGT
2) AAAAACCCCCGGGGGTTTTT
3) AGATCCGTTGACGATCCTAG

In all cases, the quantities of the letters are equal across the stretch of the message, five of each, adenine, thymine, guanine, and cytosine. All three examples will result in perfect entropy of $H(X) = 2$ with probabilities for each letter of 0.25 for each position.

I applied the calculation of the entropy to the distribution of the distances between the occurrences of each letter. Then, the entropy was calculated based on the histogram of all distances (see 2.5.5 and Figure 3-4). Equal distributions from the first two examples lead to very low values of $H(X)$, while random distribution leads to a high value of

entropy (Figure 3-4b, right panel). The resulting virus genome loci of low sequence complexity were then used to further investigate potentially false expression profiles.

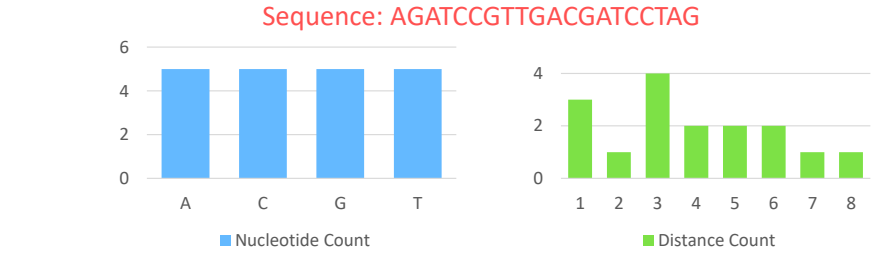
a



$H(X)$ = 5/20 * log (20/5) + 5/20 * log(20/5) + … = 2

$H(X)$ = 16/16 * log (16/16) = 0

b



$H(X)$ = 5/20 * log (20/5) + 5/20 * log(20/5) + … = 2

$H(X)$ = 3/16 * log (16/3) + 1/16 * log(16/1) + … = 2.82782

Figure 3-4: Calculation of the entropy based on nucleotide count and distance count. Entropy is calculated for sequences two (a) and three (b) of the example sequences displayed above.

I applied the developed scripts `locomplex.py` to the viral genomes FASTA file with the expectation of finding all earlier mentioned virus genomes from the VirusID benchmark: SCKV-M1 (NC_001782.1), HCV2 (NC_009823.1), SSDRV (NC_007415.1), and Pandoravirus dulcis (NC_021858.1). Using a sliding window size of 40 bases and a cutoff for the entropy $H(X)$ of 1.8, the software reported 32,960 low complexity loci. The four desired loci were included:

```
>NC_001782.1:1023-1228
AATATAGTAGGCACAAAATAAAATAAAAATTAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAGAAAAGAGAGAAGAAGAAGAAAAAGAAAAAACAAAAGAAA
CAGAAAAAGAGAGAACAGG
>NC_009823.1:9458-9601
TACACTCCATAGCTAACTGTCCCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCTTTTTTTCTCTTTTCC
TTCTTTCTTACCTTATTTT
>NC_007415.1:5391-5470
TTGAAGGTTTACTTTCTTTTATAATGGGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAA
```

```
>NC_021858.1:1026444-1026572
CCTTTTCTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTT
GTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGA
CAGGC
```

The script prints the detected loci in BED format as well. VIRGENE incorporates an overlap of this file with viral coding sequences. Expressed viral genes are compared to the low complexity file. If only a low complexity gene is expressed, it will be considered as a false expression. If more than the low complexity regions are expressed, it is assumed that the expressions are true.

The implementation of a pipeline for virus identification and viral gene expression, VIRGENE, involved the extension of the before used tool VirusID in various points. Compared to the virus genomes database, a virus gene database is not readily available. VIRGENE incorporates several scripts to produce input data for analysis. Additionally, several filtering steps had to be developed. The final pipeline reports expressed viral genes from RNA-Seq data of the host.

3.5 Viral Expression Patterns in Cell Lines

Many cell lines carry known and unknown viruses with distinct viral expression profiles. Apart from natural infection of the source tissue or through transduction, cell lines might carry viruses as a consequence of contamination [94]. Viral proteins influence the host in different ways. Thus, analysis of viral expression profiles is a key step in understanding viral-host interactions and virus impact on host gene expression pathways. To study virus expression in cell lines, I collected a set of 186 samples from 22 different projects and publications (Appendix Table A-1). The RNA-Seq data were retrieved from the SRA and the GEO.

The application of VIRGENE to published cell line data revealed variations in expression profiles between different cell lines, in addition to virus detection data. Of the examined 186 samples, 77 samples were negative for viruses. The remaining 109 samples showed signatures from human viruses, other mammalian viruses, and phages.

To avoid comparison of data from different sequencing platforms, I only selected samples sequenced on Illumina platforms (Figure 2-2). The well-studied genome of Enterobacteria phage phiX174 (phiX174, NC_001422.1) [44] has been used as spike-

in quality control by Illumina for several years now, and especially on the Genome Analyzer platforms. In total, 49 samples were positive for phiX174 with an FVR greater than 2 ppm. The maximum detected load was 73,935 ppm in cell line CA-46.

Using TRSP, VIRGENE calculated the expression values for all viral genes. The final expression table contained all expressed genes after filtering. Exemplarily, I selected a subset of nine viruses and plotted all of their 115 expressed genes in the respective 55 samples as a heatmap (Figure 3-5). The tables in Appendix chapter C show all examined cell lines and their virus content after filtering.

Seven lymphoma cell lines and one gastric cancer cell line displayed heterogeneous expression profiles of Epstein-Barr virus (EBV) (Figure 3-5). The expression profiles varied between the samples, which did not uniformly express the commonly noted latency marker genes EBNA-1, EBNA-2, LMP-1, and LMP-2 (indicated by arrows). The virus gene database only included mRNA and coding regions, excluding potential latency associated micro-RNAs and non-coding RNAs. I provide more detail on these findings and a comparison to EBV in primary tumor samples in chapter 3.6.

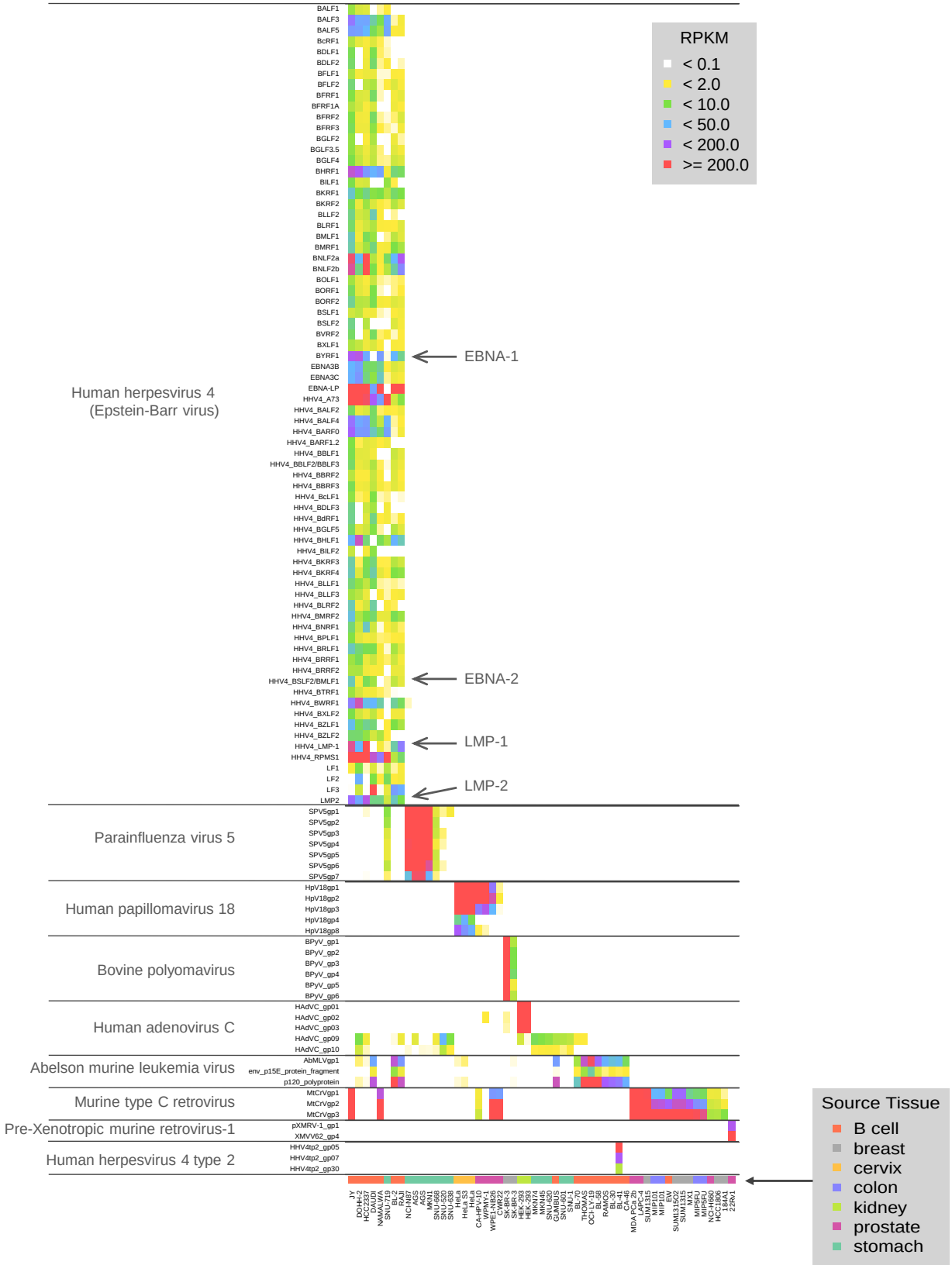

RPKM
< 0.1
< 2.0
< 10.0
< 50.0
< 200.0
>= 200.0
Human herpesvirus 4 (Epstein-Barr virus)
EBNA-1
EBNA-2
LMP-1
LMP-2
Parainfluenza virus 5
Human papillomavirus 18
Bovine polyomavirus
Human adenovirus C
Abelson murine leukemia virus
Murine type C retrovirus
Pre-Xenotropic murine retrovirus-1
Human herpesvirus 4 type 2
Source Tissue
B cell
breast
cervix
colon
kidney
prostate
stomach

Figure 3-5: Viruses show distinct expression patterns in cell lines. Heatmap of sequence read counts for 115 virus genes (vertical axis) from nine selected viruses in 55 cell line samples (horizontal axis) normalized for sequencing library size and gene length. Listed are only the genes that show expression > 1 RPKM in at least one sample. If two viruses from the same taxonomic subfamily were detected in one sample, only the viral genes of the virus with higher virus genome coverage (VGC) are shown. Epstein-Barr virus latency genes are indicated with arrows.

I detected high levels of Parainfluenza virus 5 (PIV5, NC_006430.1) in two samples of the gastric cancer cell line AGS. Expression of PIV5 is comparable in both samples with FVRs of 28,070 and 27,865 ppm. In addition, VIRGENE identified PIV5 in six other gastric cancer cell lines processed within the same project with lower fractions between 1.2 and 3351 ppm.

As stated in chapter 3.3, VIRGENE confirmed HPV18 in three HeLa samples. Additionally, HPV18 expression signals were also measured in cell lines CA-HPV-10, WPMY-1, and WPE1-NB26.

Three different projects have analyzed and sequenced the breast cancer cell line SK-BR-3. VIRGENE detected Bovine polyomavirus (BPyV, NC_001442.1) in two of three samples of SK-BR-3 (13,516 ppm and 16 ppm, Figure 3-5), but no trace of BPyV in the third sample. Inspection in the Integrative Genomics Viewer (IGV) suggested a variety of splice variants in the junction plot and several single nucleotide variations compared to the virus reference (Figure 3-6). Here, even splice variants and single nucleotide variations are visible. The third SK-BR-3 sample (SRX329206) contained only one paired-end read mapping to BPyV.

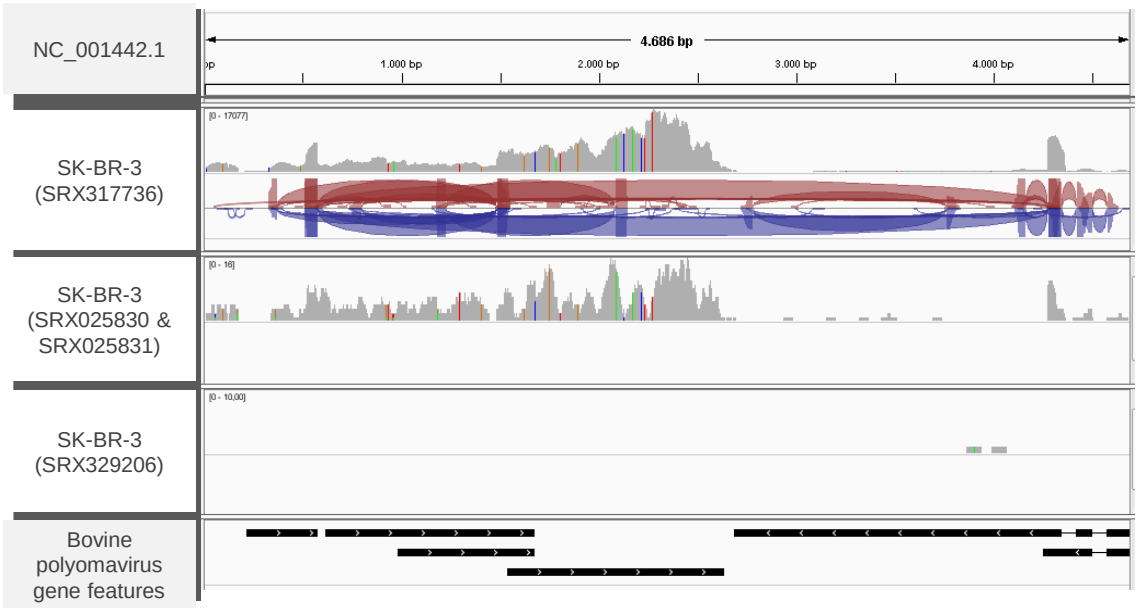


Figure 3-6: IGV visualization of alignments to Bovine polyomavirus. Displayed are the alignment profiles to Bovine polyomavirus (BPyV) in three SK-BR-3 samples. Sample SRX317736 showed the highest virus load (FVR: 13.516 ppm, top lane). The second sample (SRX025830-1) contained significantly lower amount of viral nucleic acids (FVR: 16 ppm). Sample SRX329206 contained one paired-end read of BPyV (bottom lane).

Further detected viruses included different murine retrovirus species. Their presence and other discovered patterns are discussed in section 4.3.1. In summary, mapping sequence reads to a database of virus genomes revealed a variety of alignment profiles. Using a gradual filtering procedure, I was able to identify expressed viral genes and to measure their expression values. In some replicate samples, for example Hela or AGS, these values were comparable. However, expression profiles between different cell lines remained variable and featured distinct profiles.

3.6 EBV Expression in Cell Lines and Primary Tumors

Studies of the Epstein-Barr virus (EBV, Human herpesvirus 4 (HHV4)) and its transformational effects started half a century ago [95] and have revealed the extreme effectiveness in the immortalization of B lymphocytes. It has also become clear, that latent EBV infection in humans is ubiquitous and is associated with some cancer types such as Burkitt's lymphoma or gastric carcinomas [96].

As shown in the previous chapter, I tested seven B lymphocyte cell lines (JY, DOHH-2, HCC2337, DAUDI, NAMALWA, BL-2, and RAJI) and one stomach carcinoma cell line (SNU-719) positive for EBV (NC_007605.1) with varying viral gene

expression profiles. EBV latency type III associated genes BKRF1 (EBNA-1), LMP-1, LMP-2, and BYRF1 (EBNA-2) were expressed in all but one of the eight cell lines (Figure 3-5), as DAUDI lacks the expression of LMP-1 and EBNA-2, for which it has a reported genomic deletion [97]. Conversely, DAUDI has one of the highest expressions of LF3 (11,660 RPKM), while for example JY lacks LF3 expression completely. This is visualized in the IGV plot in Figure 3-7.

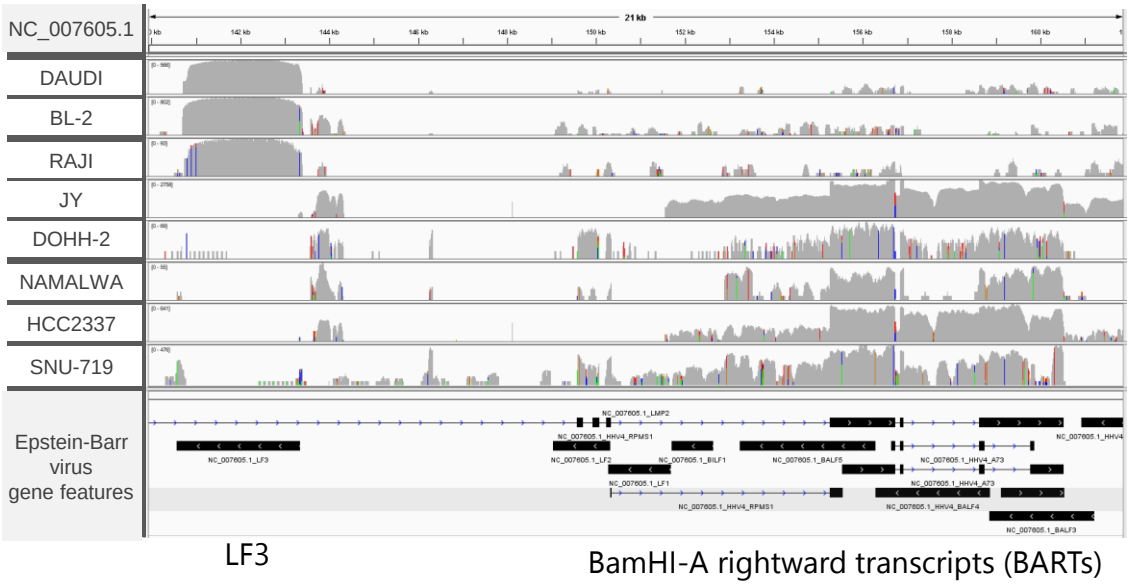


Figure 3-7: LF3 and BART alignment profiles in eight cell lines. IGV plot of seven lymphoma cell lines (DAUDI, BL-2, RAJI, JY, DOHH-2, NAMALWA, and HCC2337) and a gastric carcinoma cell line (SNU-719) expressing Epstein-Barr virus (EBV, NC_007605.1).

I compared viral expression in the EBV-positive cell lines to primary tumor samples from different cohorts with known associations to EBV infection (Figure 3-8): endemic Burkitt's lymphoma (EBL), stomach adenocarcinoma (STAD), diffuse large B cell lymphoma (DLBC), and head and neck squamous cell carcinoma (HNSC).

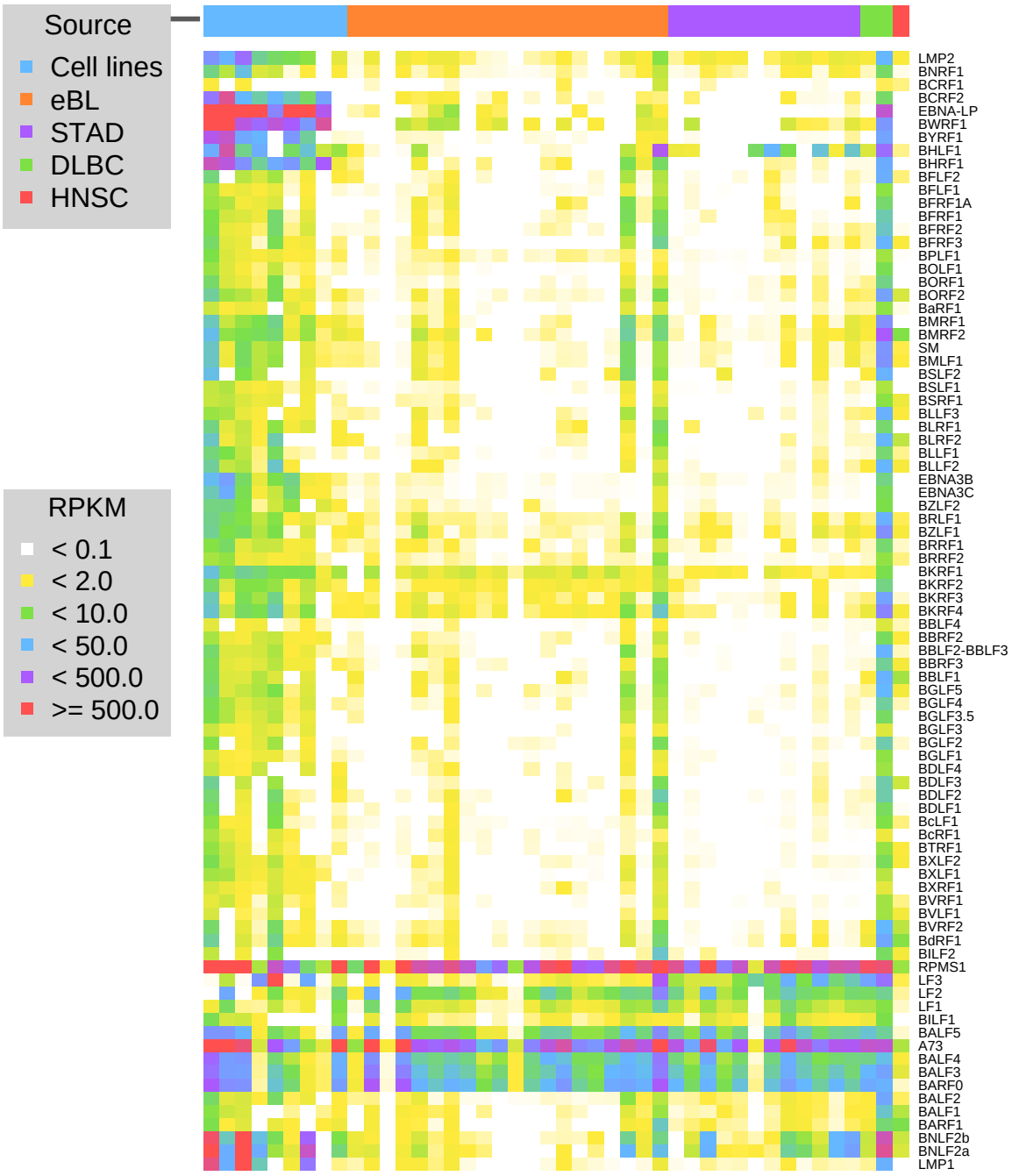


Figure 3-8: Distinct expression profiles of EBV in cell lines and primary tumors. Sequence read counts normalized for gene length and library size for all Epstein-Barr virus (EBV) positive cell lines and primary tumors. Gene names are sorted according to transcription start sites on the EBV genome. Cell lines left to right: JY, DOHH-2, HCC2337, BL-2, DAUDI, NAMALWA, RAJI, BL-41, SNU-719. (EBL: Endemic Burkitt Lymphoma; STAD: Stomach Adenocarcinoma; DLBC: Diffuse Large B cell-Lymphoma; HNSC: Head and Neck Squamous Cell Carcinoma).

EBV was detected in fifteen samples from three cohorts from The Cancer Genome Atlas (TCGA): two out of forty-eight DLBC samples, one oral cavity sample out of 398 HNSC patients, and twelve out of 190 STAD samples. In addition, I included 20

EBL samples [55]. Compared to cell lines, I could not define distinct latent gene expression in primary tumor samples. However, expression of BamHI-A rightward transcripts (BARTs), including potential open reading frames BARF0, A73, and RPMS1, could be measured across most of the primary tumors and cell lines (Figure 3-8) confirming earlier findings on the expression of BARTs in all EBV-infected cells with diverse patterns [98].

Two types of EBV were included in the used virus genome database, EBV1 (NC_007605.1) and EBV2 (NC_009334.1). The two differ only in transcript structure and genomic sequences of latent genes. These genes were hardly expressed in the tumor samples, measured for both virus types (data not shown). Hence, I could not draw a conclusion regarding the distinction of the underlying EBV type. For simplicity, only EBV type 1 has been used for the presented investigation.

The data showed that expression of EBV genes differed between cell lines and primary tumors. The expression of BARTs was comparable in all EBV-positive tumor samples, but varied in the cell lines. However, I could only measure significant expression of further EBV genes in the cell lines, while most of the tumor samples only displayed expression of BARTs.

3.7 HPV Expression in Tumors

Human papillomaviruses (HPVs) are recognized as major infectious agents in human carcinogenesis and are strongly associated with cervical cancers and a subgroup of head and neck cancers [99, 100]. To evaluate the clinical utility of viral sample annotation using *in silico* virus typing, we sought to investigate the HPV status and expression in two cancer cohorts: a head and neck squamous cell carcinoma (HNSC) and a cervical squamous cell carcinoma (CESC) cohort. RNA-Seq data sets for both cohorts were retrieved from TCGA. In total, I analyzed 701 primary tumor samples from the HNSC and the CESC cohort.

RNA-Seq data for 402 HNSC samples were downloaded in October 2014. Alignments failed for five of the samples, showing extremely low mapping rates. The remaining 397 samples were analyzed with a pre-final version of VIRGENE. Briefly, data was downloaded in BAM format. Sequence reads were extracted using `bam2fastq`. Reads were aligned to a hybrid human-virus reference genome using STAR as described in

section 2.5.2. Virus alignments were analyzed and filtered as with VIRGENE (section 2.5.5) but using the statistical computing environment R. After evaluation of the hereby-produced results, this led to the development of VIRGENE as a pipeline.

All samples contained measurable loads of phiX174 sequences (43 to 25324 ppm). HERV-K113 was detected in several samples with a viral load of up to 50 ppm. One sample expressed EBV BARTs at low levels (1.0 to 2.3 RPKM). Another sample was positive for HHV1 with gene expression levels between 1.0 and 29.6 RPKM. HHV1 was also detected in one further sample, but with low gene expression (< 1.5 RPKM). Seven samples were positive for HCMV. Six of these samples expressed HCMV genes with levels up to 5 RPKM. The seventh sample exceeded this six-fold. All seven samples showed the strongest expression of glycoprotein RL5A. Lower expression levels were detected for the genes RL8A, RL9A, UL22A, and UL41A. Other samples showed comparable read distributions, but were considered negative due to very low viral load. Very low signals of plant viruses, fish viruses, or phages were not evaluated. The EBV, HHV1, and HCMV-positive samples were negative for HPV. Thirty-six of the HNSC samples were positive for HPV16 (9 %) with a majority of the HPV genes expressed (Figure 3-9a). Of all HPV types, VIRGENE only reported HPV16 infections in the HNSC cohort.

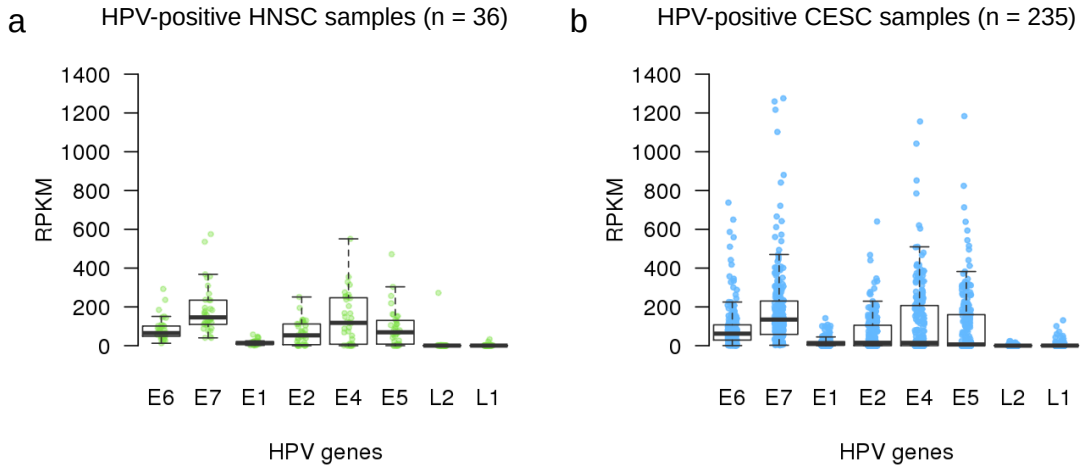


Figure 3-9: HPV expression in HPV-positive HNSC and CESC patients. a) Of 397 head and neck squamous cell carcinoma (HNSC) samples, 36 were determined positive for HPV16. A majority of HPV genes were expressed in most samples. b) In the cervical squamous cell carcinoma (CESC) cohort, 235 samples (77 %) tested positive for HPV. All samples show expression of at least E6 and E7 genes. However, other HPV genes are not expressed by all samples.

As cervical carcinomas have a strong association with HPV infection [99]. I incorporated the TCGA cervical squamous cell carcinoma (CESC) cohort for further analysis of HPV expression in cancers. The CESC cohort contained 304 samples. I was able to type 235 of the 304 CESC samples (77 %) as positive for HPV. HPV types 16 and 18 were present most frequently, with 171 samples and 59 samples, respectively. Two patients were positive for HPV26, two were positive for HPV34, and one sample contained HPV53. The majority of HPV genes were expressed in several samples (Figure 3-9b). However, compared to the HNSC cohort, more samples expressed only genes E6 and E7 and less of all other genes.

One outcome of infection with HPV might be chromosomal instability and subsequent integration of the HPV genome into the human genome [101, 102]. The STAR alignment reported chimeric reads, i.e. sequence reads, where one part mapped to one chromosome and the other part mapped to another chromosome. These sequence reads might indicate an integration event, but are restricted to expressed integration sites in RNA-Seq data. The virus genomes have been added to the used reference genome as individual sequences and have been interpreted by STAR as individual chromosomes. Hence, STAR reported sequence reads mapping to human and virus chromosomes as human-viral chimeric reads. Figure 3-10 shows the number of human-HPV reads in three groups: the HPV-negative HNSC and CESC samples, the HPV-positive HNSC samples, and the HPV-positive CESC samples.

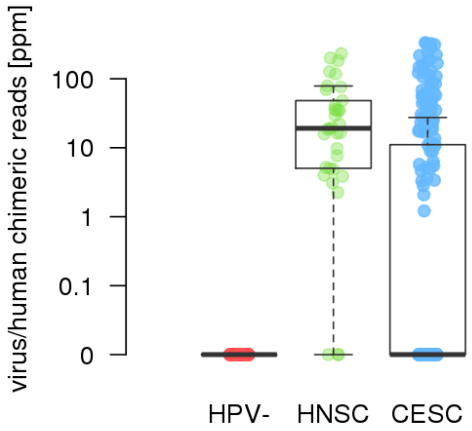


Figure 3-10: Number of human-HPV chimeric sequence reads. Values are given for HPV-negative HNSC and CESC samples, combined, and for HPV-positive HNSC and CESC samples, separately. Numbers of chimeric reads, i.e. sequence reads indicating an integration event, are given as relative number of reads mapped to the human genome in parts per million (ppm).

Of the 36 HNSC samples, two did not show chimeric sequence reads. In the CESC cohort, 160 samples were free of human-HPV chimeric reads (68 % of HPV-positive CESC samples). The relative chimeric read counts varied between 2.2 and 231.3 ppm.

HPV could frequently be detected in the analyzed tumor cohorts. Only a fraction of the analyzed HNSC patients was positive for HPV16, while different HPV strains were detected in the majority of CESC samples. The investigated cohorts revealed a variety of HPV expression profiles and different numbers of integration events. Both observations warrant further investigation.

4 Discussion

This study showed that host transcriptomic data could be exploited to detect and identify contained viral sequences (Table 3-2). The results slightly exceeded PCR and microarray-based platforms, as well as other NGS-based methods (Table 3-3), even when nucleic acid content was extremely low (Table 3-4 and Table 3-5). The alignment patterns suggested diversely expressed viral genes (Figure 3-3). I developed VIRGENE, a pipeline for the identification and quantification of viral gene expression, to assess these expression profiles. Measuring expression in 186 cell line samples revealed presence of different virus species and diverse expression across cell lines (Figure 3-5). A comparison of EBV expression showed differences in expression of a majority of the genes between cell lines and different primary tumor cohorts, but consistent expression of BARTs across most of the samples (Figure 3-8). Application of VIRGENE to RNA-Seq data from HNSC and CESC patients revealed different expression profiles (Figure 3-9) and variable evidence for genomic integration (Figure 3-10).

4.1 *In Silico* Virus Detection

Detection and identification of viral infections in cell cultures remain challenging. Classical detection methods, like ELISA, PCR, or ISH can be laborious or require foreknowledge and thus might be biased [94, 103, 104]. However, the screening cell lines that are used in biotechnological production is essential and recommended to prevent severe consequences to the recipient of the medical product or economic losses to the producer [103, 104]. The application of RNA-Seq to human samples has produced a plethora of data in the past few years. Viruses produce mRNA that can be read by the host ribosome. Next-generation mRNA sequencing is unbiased regarding the source of the nucleic acids. Hence, we sought to develop an RNA-Seq based method that would allow for the detection and identification of viral mRNA in sequenced mammalian samples. This would enable analysis of already sequenced tumor cohorts for their viral content, or the screening for viruses in potentially contaminated cell lines. It would also enable the detection of yet unknown sequences. In the light of this thesis, this empowers the detection of known and novel viral infections from RNA-Seq data [105, 106].

The detection of viral sequences in high-throughput sequencing data can be initiated by host subtraction[68, 83, 107], i.e. the alignment to the host genome or transcriptome in one or a few steps and the retention of remaining unaligned reads for further processing [71, 85, 109]. The remaining sequence reads can then be directly mapped to viruses or assembled to produce contigs that are then aligned to determine their origin. Another approach could be to simultaneously mine host and viral genomes by adding one or more viral genomes to the reference genome [84, 86], or by mapping to host and viruses in parallel [61].

In a subsequent step, the virus-derived sequence reads can be quantified. To determine the included viral species, VirusSeq, for instance, uses a simple cutoff of 1,000 aligned reads to consider a virus as detected [68]. In other approaches, the hit virus genomes are ranked and the top hit is selected [85]. ContextMap [61] calculates a confidence score for each screened species based on genome coverage, read numbers, and mismatch distribution of aligned reads. VIRANA [86] includes a post-alignment taxonomy identification of contigs followed by several steps of BLAST alignments to determine the correct pathogenic species. To determine viral integration and to recognize the insertion site, some methods analyze chimeric host-virus reads and their alignment loci [84, 85].

VirusID and VIRGENE were both designed to run a single step alignment against a combined host and virus reference genome. A comparison has shown a huge decrease of signal from additional viruses by simultaneously mapping to the host genome (Figure 3-1). Low complexity regions in virus genomes often attract many thousand sequence reads. Hence, I implemented the low complexity filter and declined the use of a cutoff for the number of aligned reads per virus as a single selection criterion. Finally, VirusID and VIRGENE enable downstream analysis of viral integration or virus assembly, if desired, by reporting chimeric reads and viral contigs for further processing.

4.2 Performance of VirusID and VIRGENE

Immortalized and biotechnologically modified cell lines are used to produce clinically relevant products, such as hormones, enzymes, and monoclonal antibodies [108]. Together with the collaborator, Baxter International, we and three other investigators set out to assess whether molecular testing was applicable for the detection of viruses

in the supernatant of production cell lines (section 3.1.2). We received twenty samples in a blinded fashion. Five of the twenty samples were not sequenced, because no nucleic acids were detected during library preparation. Four of these samples were negative controls, the fifth was supposed to be positive for MVM (sample 6A, Table 3-2). VirusID was able to identify the correct virus in twelve of the fifteen sequenced samples. In one case, no reads of the virus spike-in ReoIII were detected. Instead, a high load of EMCV was measured and reported. All other methods failed to identify five samples correctly; in two cases, they even reported viruses in the negative controls.

In two of the samples that were prepared with X-MuLV (samples 5A and 14A), VirusID detected the strongest signal from Pre-Xenotropic murine leukemia virus-related virus 1 provirus (PreXMRV-1, NC_007815.2). The spiked-in murine leukemia virus is closely related to other murine retroviruses like Murine type C retrovirus (MCRV) or other murine leukemia viruses. The original sequence of the MuLV genome was not contained in the used virus database. The reads aligned to other related viruses instead. At the time of the analysis, taxonomic filtering was not yet implemented in VirusID. Both factors lead to an accumulation of detected murine leukemia viruses and retroviruses for the respective samples. This shows the need for precisely selected virus genomes for the database, as is further discussed in section 4.5.1. The additional murine retroviruses are not listed in Table 3-2.

All four methods reported further detected viruses, i.e. viruses in addition to the expected results (Table 3-3). However, VirusID only detected viruses from within the study panel (EMCV and MVM), indicating some sort of cross-contamination. The other three methods additionally detected viruses not included in the test set [52].

For the final benchmark, nasal washes of virus-infected children were collected. After extraction of RNA, samples were shipped to our laboratories in Mainz. Due to very low nucleic acid content, only eighteen of 50 samples were sequenced and analyzed, and four of these could be typed correctly and with high confidence by VirusID. One virus was detected with high VGC that did not match the previous typing result. Only one other software was able to confirm four out of five detected viruses in samples with sufficient viral load (Table 3-5).

Samples with a low nucleic acid content are difficult to analyze and the results remain difficult to interpret. One reason for the low recognition rate of the virus genomes could be the low input of RNA for sequencing. The PCR amplification might expand different sequences, while others might get lost. This can lead to low library complexity resulting in sequence duplicates of some of the sequences. Additionally, the dilution of the nasopharyngeal secretion and the subsequent concentration might impede the detection of rare sequence species by stochastic chance. A higher viral load would lead to more viral sequences in the samples. Consequently, these would be easier to detect. Nasal aspirates might not contain the necessary amounts of viral mRNA for their detection via RNA-Seq based methods.

The final benchmark revealed the challenges of the VirusID platform with low concentration samples (Table 3-4). All tested pipelines detected viral signal from low complexity genome regions. Further development of the software enabled us to circumvent this issue (chapter 3.4).

4.3 Detected Viruses

For this study, various samples have been analyzed. Some of the samples contained known virus spike-ins or were tested for their viral content beforehand (chapter 4.2). For some of the downloaded cell line data the virus content was also known. In other cell lines, VirusID or VIRGENE detected yet unknown viral infections or contaminations.

4.3.1 Viruses in Cell Lines

Viruses in cell lines have been described in detail [109, 110] and viruses have been known for decades to be a source of cell culture contamination [94, 111]. Hence, an analysis of cell line NGS data and *in silico* virus detection would enable validation of the method by finding known viruses. It could possibly lead to findings of yet unknown viral content of cell lines.

In this work, 186 cell line samples were analyzed (chapter 3.5). These samples contained 143 distinct cell lines. Some cell lines were sequenced either as replicates in the same project, or as duplicate samples in different projects (Figure 2-2). HeLa cells were the first immortalized human-derived cells to be cultured successfully [112]. They were derived from a cervical adenocarcinoma caused by viral infection and

integration. The contained HPV18 is well described in the literature [113, 114]. Using VirusID and VIRGENE, I was able to confirm the presence and expression profiles of HPV18 in all three HeLa samples (Figure 3-3, Figure 3-5).

Fourteen of the analyzed samples contained sequences of Human endogenous retrovirus K113 (HERV-K113, NC_022518.1) (FVR > 2 ppm). Thousands of human endogenous retroviruses are known. However, the used virus database only contained HERV-K113, a member of the HERV-K family. These retroviruses do express viral particles [115]. Due to the lack of further HERV genomes for the described analyses, we could not consider the results of HERV-K113 as reliably identified.

The detection of high levels of Parainfluenza virus 5 (PIV5, NC_006430.1) in two samples of the cell line AGS confirms earlier findings [116]. Unexpectedly, VIRGENE identified PIV5 in six other cell lines processed within the same project. Infection with PIV5 has not been reported before in these cell lines, suggesting cellular or viral cross-contamination.

VIRGENE reported BPyV (NC_001442.1) in two out of three SK-BR-3 cell line samples; one sample was virus free (Figure 3-6). BPyV is frequently detected as a contaminant of calf serum [110], suggesting the used serum as a source for contamination.

Cell line HEK293 is known to contain Human adenovirus 5 (HAdV5) [117]. The utilized virus database contained eight species of human adenoviruses, lacking the genome of HAdV5. Thus, VIRGENE reported high expressions of the closely related Human Adenovirus C serotype 2 (HAdVC, NC_001405.1) for HEK293.

Murine retroviruses are also known to have the capability of infecting and contaminating human cell cultures [94, 118, 119]. The Xenotropic murine leukaemia virus-related virus (XMRV, NC_007815.2) is present in cell line 22Rv1, confirming previous findings [119]. Expression of Murine type C retrovirus (MCRV, NC_001702.1) could be detected in nineteen cell lines of different tissue origin and across several projects with viral fractions ranging from 8 to 62,476 ppm (median 2,357) (Figure 3-5). These findings elucidate the widespread contamination of cell lines with murine retroviruses across several laboratories.

Some of the examined samples showed signatures of more than one virus. JY and NAMALWA are both positive for EBV [120, 121]. Both are positive for MCRV as well. Additionally, cell line JY is reportedly positive for Murine leukemia virus (MuLV) [120], a virus that is not present in the used virus genome database. A significant number of reads mapped to XMRV (VGC: 43 %, FVR: 667 ppm). Due to the close taxonomic relation of XMRV and MCRV, VIRGENE reported only one of the viruses here. The HPV18 E6/E7-tranduced cell lines CA-HPV-10 and WPE1-NB26 both show contamination with MCRV. The gastric carcinoma cell line SNU-719 is positive for EBV and contains sequences from PIV5 as well. This cell line was processed alongside the PIV5-positive cell line AGS in project SRP014574. VIRGENE has a filtering step to prevent misidentification of taxonomically closely related viruses. As observed with the cell line JY, this could hinder the detection of co-infection with two closely related viruses.

OCI-LY-19 has been typed positive for EBV by PCR before, but the signal was too low for fluorescence in situ hybridization (FISH) [109]. Using the RNA-Seq data, we do not see EBV expression in this cell line. The original publication on cell line DOHH-2 from 1991 reported it as EBV-negative [122]. A later publication reported EBV-signal by PCR and FISH [109]. VIRGENE revealed expression of a majority of EBV genes in the examined sample. However, "*the original culture was a mixture of EBV- and EBV+ cells, of which the EBV- cells were isolated and clonally expanded"* (https://www.dsmz.de/catalogues/details/culture/ACC-47.html; last accessed March 22, 2017), according to information on the website of the vendor DSMZ. BL-2 is another cell line that is reportedly negative for EBV [123]. The original publication evaluated absense of EBV by immunoflourescence using antibodies againt EBV nuclear antigen (EBNA) protein, but detected normal EBV serological titres. Again, according to VIRGENE, the majority of EBV genes are expressed in the analyzed RNA-Seq sample (Figure 3-5).

The cell line 22Rv1 was derived from cell line CWR22 that was serially passaged in mice [124]. Both samples were sequenced in project SRP004637. I can confirm XMRV-positivity in 22Rv1, as stated by the vendor ATCC (https://www.lgcstandards-atcc.org/products/all/CRL-2505.aspx; last accessed March 22, 2017). However, CWR22 is additionally positive for MCRV, which has not been reported before.

WPMY-1 had been transformed using the large T antigen of Simian virus 40 (SV40) [125], of which we couldn't detect more than one paired-end sequence read. Further investigations would be necessary to determine contamination or mislabeling of the cell line.

Several other samples of this study contained HPV18 derived sequence reads. The cell lines are Hepg2 in duplicates, RT4 in duplicates, and one of the HUVEC samples. The viral genome coverages in these samples ranged from 7-20 %. However, the viral abundance was below 2 ppm in these samples. Only between 8 and 23 reads mapped to HPV18. These cell lines are not known to be positive for HPV18. However, the low expression might be a hint to partial contamination with HeLa cells or HeLa-derived RNA. This has been observed before for cell lines as well as for TCGA RNA-Seq data [126, 127].

Detected AbMLV and EMCV

The analysis revealed expressed loci of Abelson murine leukemia virus (AbMLV, NC_001499.1) in twelve cell lines, eleven of which had been sequenced in the same project (SRP009316). The viral load ranged from 132 to 8,930 ppm with relatively constant genomic regions expressed (VGC: 29 – 39 %). Additionally, the remaining two samples from project SRP009316 contained AbMLV expression, as well. The report for this virus was suppressed due to a co-infection with MCRV and its taxonomic proximity to AbMLV. The AbMLV expression profile was not as expected as only incomplete transcripts are expressed (Figure 4-1a). Interestingly, almost all of the AbMLV-positive samples contained sequences of the Encephalomyocarditis virus (EMCV, NC_001479.1) (Figure 4-1b), although not all signals were above the chosen threshold. The expression profiles of EMCV suggest expression of the promoter region only. Of the depicted cell lines, LAPC-4 is the only cell line not expressing the EMCV promoter locus.

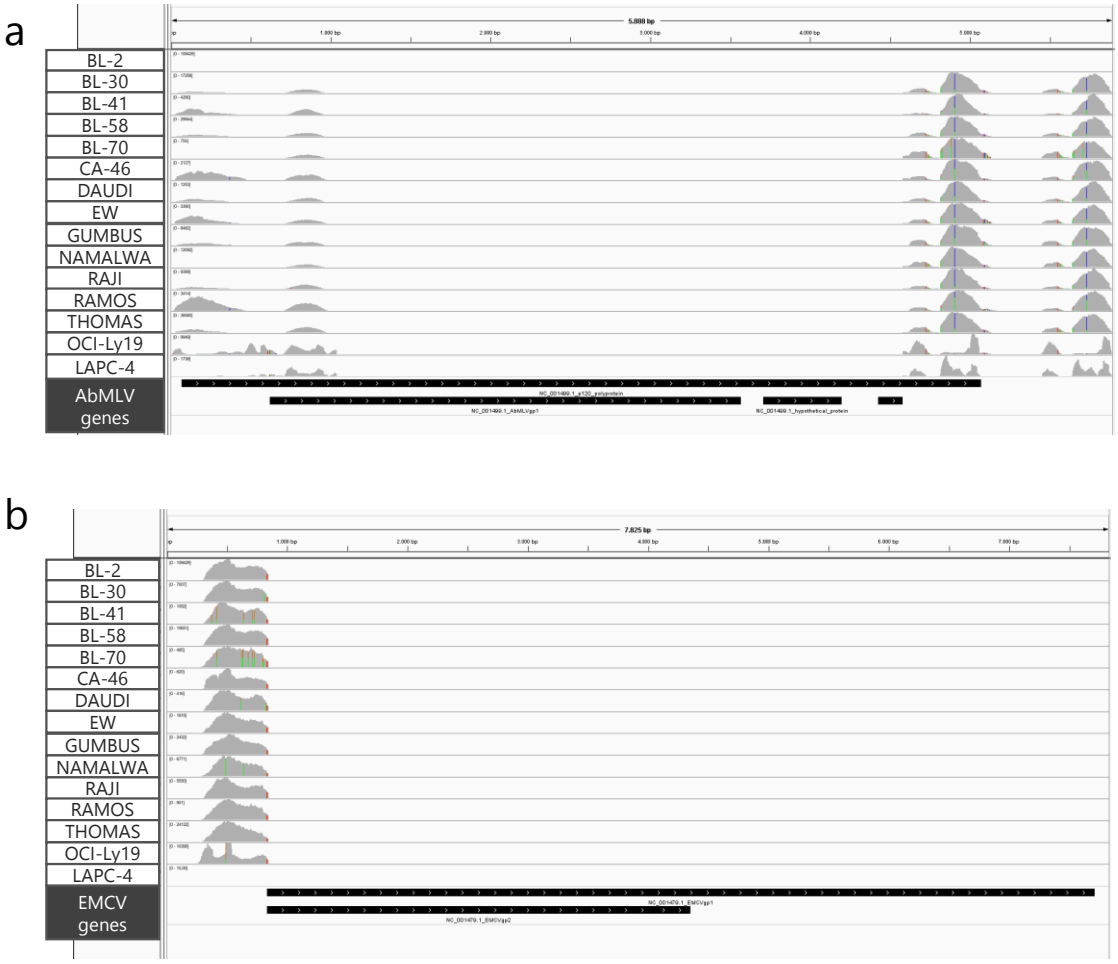


Figure 4-1: Expression profiles of EMCV and AbMLV in cell lines. Visualization of the alignments using the Integrative Genomics Viewer (IGV). Profiles shown of Abelson murine leukemia virus (AbMLV) (a) and Encephalomyocarditis virus (EMCV) (b) in all samples of SRP009316, in cell line OCI-Ly19, and in cell line LAPC-4.

VirusID and VIRGENE both report expressed loci of virus genomes as contigs in FASTA format. To gain deeper insights into the sources of the detected sequences I performed a BLAST analysis of the expressed genome fragments using the shortest contig representing the common sequence expressed in the samples (Appendix chapter D). The EMCV contig of cell line BL-41 had a length of 562 nt. The BLAST search resulted in different transgenic lacZ-tagged mouse mutant alleles with 100 % query cover and 99 % identity in all of the top 100 hits. Four loci of AbMLV were analyzed with BLAST using the sequences from the cell line NAMALWA. The BLAST search of one of the contigs revealed similarities to human ABL proto-oncogene, ABL in other species, and to AbMLV, as expected. The result shows a sequence similarity of 100 % to the human ABL across the 165 nt sequence contig.

The remaining three sequences show high similarities to different types of vectors, including gene trapping vectors, retroviral cloning vectors, and expression vectors hinting at vector contamination of the cell lines. I was able to confirm this using NCBI's online tool VecScreen that identifies vector contamination in samples (https://www.ncbi.nlm.nih.gov/tools/vecscreen/). All three sequences matched to retroviral vector pLXSN (gnl|uv|M28248.1). As the expression profile of OCI-LY-19 differs from the others, I repeated the search with the sequences from OCI-LY-19. The results remained the same with little sequence variation in some cases.

These results suggest contamination of the cell lines with a viral cloning or expression vector. Since the original publications of these datasets do not mention any type of vectoral transformation of the cell lines, it is likely that their presence in the cell lines has not been previously discovered.

4.3.2 Re-analysis of EBL Samples

The incorporated endemic Burkitt lymphoma (EBL) cohort with 20 samples from Uganda had been analyzed before and viruses were detected and quantified using the RNA-Seq data and the pipeline Pandora [55]. The sample accession numbers are listed in Appendix Table A-2. Abate et al. reported EBV-positivity for all twenty EBL samples. Indeed, I can confirm EBV for nineteen samples, whereas measured viral load was below 2 ppm in one sample. Figure 3-8 contains the expression profiles of all twenty samples, including the sample with a very low signal.

The original research also reported co-infections with other viruses in eight samples: Human cytomegalovirus (HCMV, Human herpesvirus 5 (HHV5)) in six samples, Kaposi's sarcoma-associated herpesvirus (KSHV, Human herpesvirus 8 (HHV8)) in five samples, and Human T-lymphotropic virus 1 (HTLV-1) in one sample. I compared these results to the results produced by VIRGENE (Table 4-1).

Table 4-1: Comparison of detected viruses by Pandora and VIRGENE. RNA-Seq data of twenty endemic Burkitt lymphoma samples were analyzed using Pandora in the publication of Abate et al. [55]. VIRGENE was used to re-analyze the same data. ⭘: reported identified viruses; ★: virus signal below the detection cutoff. (EBV: Epstein-Barr virus, HCMV: Human cytomegalovirus, HTLV-1: Human T-lymphotropic virus 1, KSHV: Kaposi's sarcoma-associated herpesvirus, HHV-1: Human herpesvirus 1)

	Pandora				VIRGENE				
Sample	EBV	HCMV	HTLV-1	KSHV	EBV	HCMV	HTLV-1	KSHV	HHV-1
BL15	⭘	⭘	⭘	⭘	⭘	⭘	⭘		
BL19	⭘			★	⭘				
BL20	⭘				★				
BL22	⭘	⭘			⭘	★			⭘
BL23	⭘				⭘				
BL27	⭘			⭘	⭘				
BL30	⭘				⭘				
BL35	⭘	★			⭘				
BL40	⭘				⭘				
BL43	⭘			⭘	⭘	★			
BL45	⭘				⭘				
BL48	⭘	⭘			⭘	★			
BL49	⭘			⭘	⭘				
BL50	⭘				⭘				
BL60	⭘	⭘			⭘	⭘			
BL62	⭘				⭘				
BL69	⭘	⭘			⭘	★			
BL80	⭘				⭘				
BL81	⭘				⭘				
BL84	⭘				⭘				

According to our results, two HCMV genes are expressed in samples BL15 and BL60 with expression levels of 1.3 and 1.6 RPKM for RL9A and 4.4 and 5.3 RPKM for RL5A, respectively. RL5A is also expressed in sample BL48 (1.3 RPKM), and below 1 RPKM in the other samples marked with the star (★) (Table 4-1). Furthermore, I can confirm HTLV-1 with an FVR of 71 ppm in sample BL15. All samples originally reported to be positive for KSHV contained viral signals below 1 ppm FVR and below 1 % VGC before filtering. Filtering reduced the signal to zero. Additionally, I detected Human herpesvirus 1 (HHV1, NC_001806.1) in sample BL22 (20 ppm), which was not reported in the original publication. The detected signal from HHV-1, however, is higher than all HCMV signals. Sequence reads are scattered across the genome

suggesting expression of the majority of transcripts (Figure 4-2). All other samples were negative for HHV-1, only BL60 contained one sequence read that mapped to HHV-1.

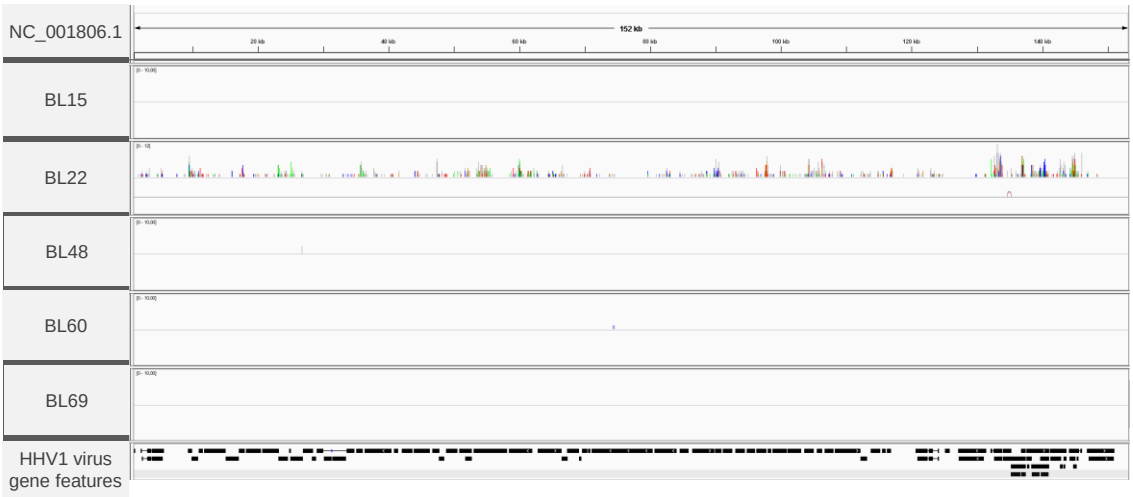


Figure 4-2: Human herpesvirus 1 expression in five EBL samples. Depicted are the alignment profiles on Human herpesvirus 1 (HHV-1) genome of five HCMV-positive samples from [55]. Only sample BL22 was reported positive for HHV-1 by VIRGENE.

It is noteworthy that the signal of some of the detected viruses was very low in these EBL samples. Figure 4-3a depicts the alignment profiles of HCMV in five samples: BL15, BL22, BL48, BL60, and BL69. VIRGENE reported only two expressed genes. Sequence reads are scarce and widely distributed over the genome. Even fewer reads aligned to KSHV (Figure 4-3b). HTLV-1, which was detected by both methods, displayed an unexpected expression of an intron with a previously undescribed splice site (Figure 4-3c).

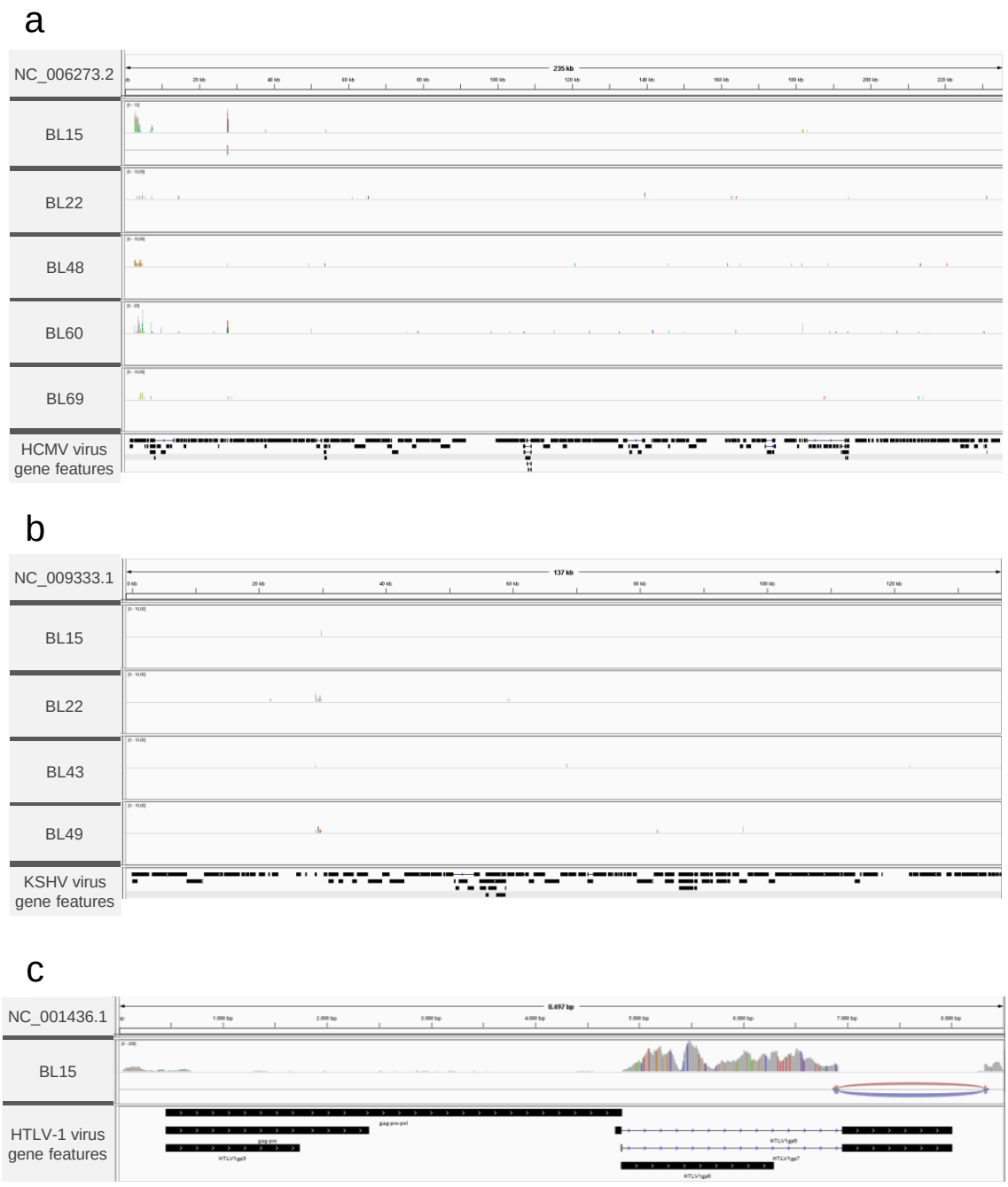


Figure 4-3: Alignment profiles of EBL viruses HCMV, KSHV, and HTLV-1. a) HCMV alignments and b) KSHV alignments in all samples that were typed positive for viruses other than Epstein-Barr virus in the original analyses by Abate et al. [55] c) HTLV-1 Expression in sample BL15.

In summary, VIRGENE was able to confirm expression of EBV in almost all samples. Expression values of KSHV and HCMV were very low and thus discarded by VIRGENE in most cases. The histochemical images in the original publication show detection of HCMV and KSHV in only a few cells, while EBV is detected abundantly. The main reason for the discrepancies might be the different cutoffs in both methods for virus detection. Expression of HTLV-1 is restricted to one intron. The

interpretation of this result remains challenging. VIRGENE detected HHV-1 in one sample with higher abundance than HCMV or KSHV. This had not been detected by Pandora.

4.3.3 The Association of HCMV with GBM

An association of Human cytomegalovirus (HCMV, NC_006273.2) with Glioblastoma multiforme (GBM) was first proposed more than a decade ago [128]. Since then this association has been a controversial point of discussion. HCMV in GBM samples has been shown to be detectable on a protein level and with PCR-based methods [129], but has failed to be detected by high-throughput sequencing based screening methods [130–132]. I had analyzed 160 untreated primary GBM samples from TCGA and 92 additional brain biopsy samples [54], including samples from 27 glioma patients. VIRGENE could not report HCMV in any of these RNA-Seq samples. This is in concordance with the described lack of evidence by any previous NGS analysis.

4.4 HPV in Tumor Cohorts

HPV can induce squamous cell carcinomas. These tumors occur most commonly at anogenital (cervix, vulva, penis, anus, or vagina) or oropharyngeal (oral cavity or tonsils) sites [3]. Although several polyvalent preventative HPV-vaccines are expected to show high impact [133], therapeutic approaches against HPV-induced carcinomas are not yet available. However, immunotherapeutic or immunomodulatory studies have already shown promising results [134–136]. The use of high-throughput sequencing technologies and bioinformatics analyses can help to identify novel diagnostic, prognostic, or therapeutic virus-derived markers [137]. Their restriction to tumor cells make them valuable biomarkers for the development of future tools and therapies.

To assess the clinical applicability of VIRGENE, I sought to compare the results of VIRGENE to the well-studied HNSC TCGA cohort [72, 138]. As described in chapter 3.7, 36 of the 402 samples tested positive for HPV16. To further evaluate the results, I linked the gene expression results to the available clinical annotation of the samples.

For 344 samples, the anatomic origin of the sample was known. Many of the HPV-positive samples were tonsil-derived (17 out of 36), followed by 'Base of the tongue' (7 samples) (Figure 4-4a), whereas HPV-negative samples were mostly oral tongue-derived.

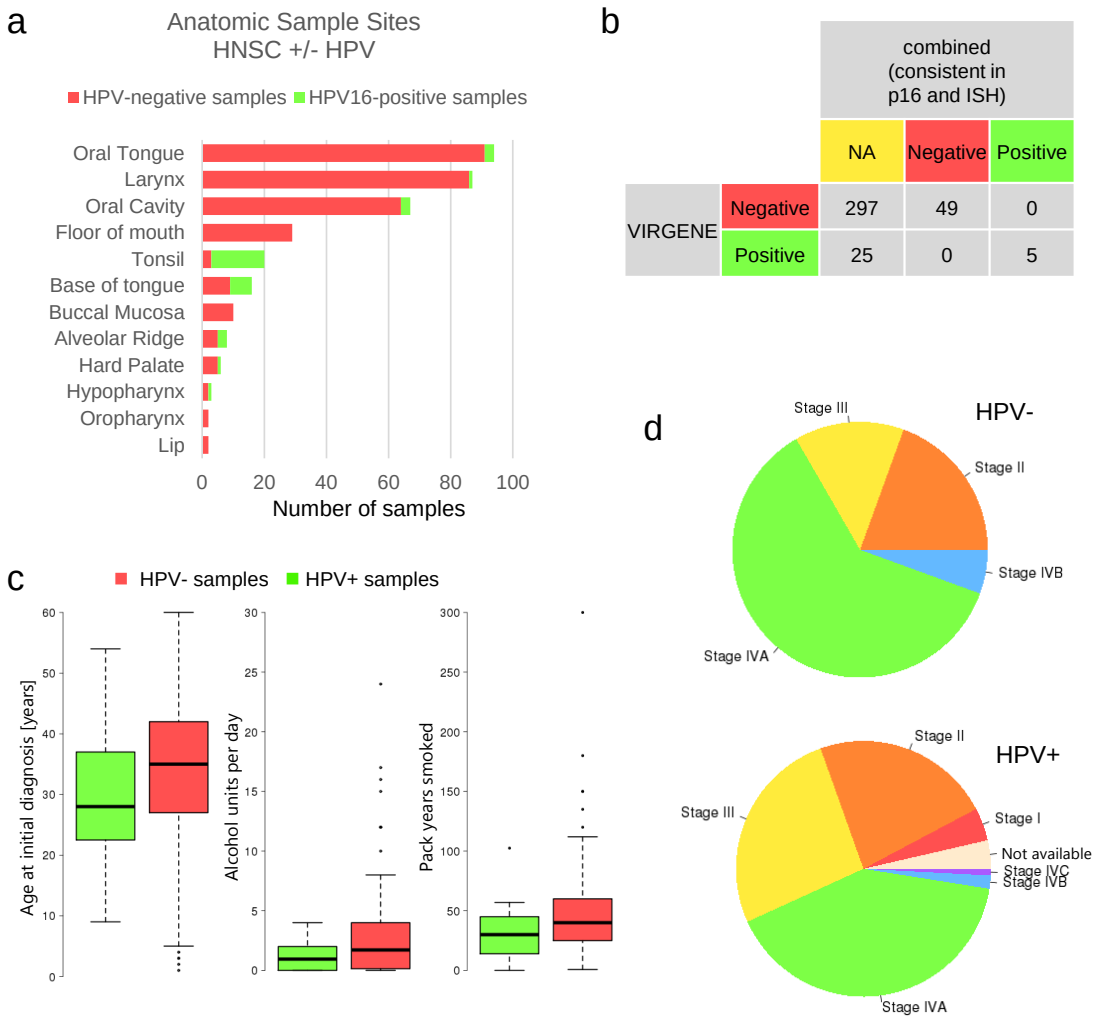


Figure 4-4: HNSC cohort dissected by HPV status. a) Samples in the HNSC cohort were dissected from a range of anatomical sites. b) VIRGENE confirmed clinical HPV typing by p16-typing or by *in situ*-hybridization (ISH), where results were consistent. Previously untyped or inconsistently marked samples were typed successfully. c) A comparison of groups stratified by HPV-status did not reveal significant differences for other risk factors (Boxes: median and quartiles; whiskers: minimum or maximum up to 1.5 inter-quartile ranges). d) Tumor stages differed, with less stage IVa tumors in the HPV+ groups. NA: clinical annotation not available or inconsistent.

The provided clinical sample annotation contained HPV typing based on p16-typing and *in situ*-hybridization (ISH). p16, or more precisely $p16^{INK4A}$, is the gene product of the gene CDKN2A and serves as a tumor suppressor. Its overexpression has been shown in cervical neoplasia [139], and in tonsillar carcinomas with an involvement of

high-risk papillomaviruses [140]. However, clinical information on HPV status was scarce for the underlying dataset. Typing results were available and consistent in both typing methods for 54 samples (49 negative and 5 positive). In all cases, *in silico*-typing was able to confirm the results (Figure 4-4b). Additionally, the HPV typing could be extended for the 301 samples where no clinical typing was available ('Not available' or 'Not evaluated'), and the 21 samples where either only one typing result was available, or the results were inconclusive. VIRGENE identified 297 samples as HPV-negative and 25 samples as HPV-positive.

Other risk factors for head and neck carcinomas are smoking or alcohol consumption [141]. A comparison of these groups, as well as the age of the patients and the tumor stage, showed no significant differences between the HPV-positive and the HPV-negative group (Figure 4-4c & d).

It is well known that HPV-positive head and neck cancer patients have an overall increased survival compared to virus free patients [142, 143]. In an attempt to reproduce the survival analysis of the TCGA HNSC cohort [138] (supplementary figure S1.4c), I performed a Kaplan-Meier survival analysis of the HNSC patients stratified by their HPV status. The analysis confirmed beneficial overall survival in HPV-positive HNSC patients (hazard ratio (HR) = 0.28, 95 % confidence interval (CI): 0.12 - 0.63, p-value (p) < 0.001) compared to the virus-free group (Figure 4-5a). As opposed to the HNSC cohort, HPV status in CESC did not correlate significantly with improved survival (HR = 1.21, 95 % CI: 0.66 - 2.25, p = 0.53) (Figure 4-5b).

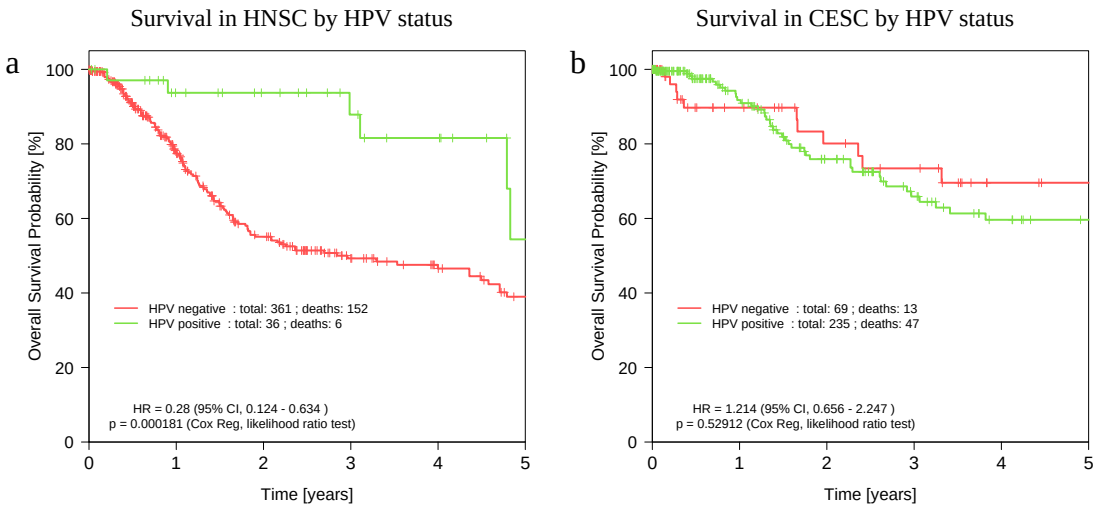


Figure 4-5: Overall survival in HNSC and CESC by HPV status. a) Overall survival in HNSC shows favorable outcome for the HPV-positive group compared to the virus-free patients (Cox regression likelihood ratio test, p = 0.000181, HR = 0.28). b) In CESC, overall survival does not vary significantly when stratified for HPV status.

Expression of HPV genes E6 and E7 is sufficient to induce carcinogenesis [91]. All analyzed samples express these two genes. However, parts of the HPV genome might be lost during viral integration. Hence, some of the viral coding sequences are only partially expressed or are not expressed at all (Figure 4-6).

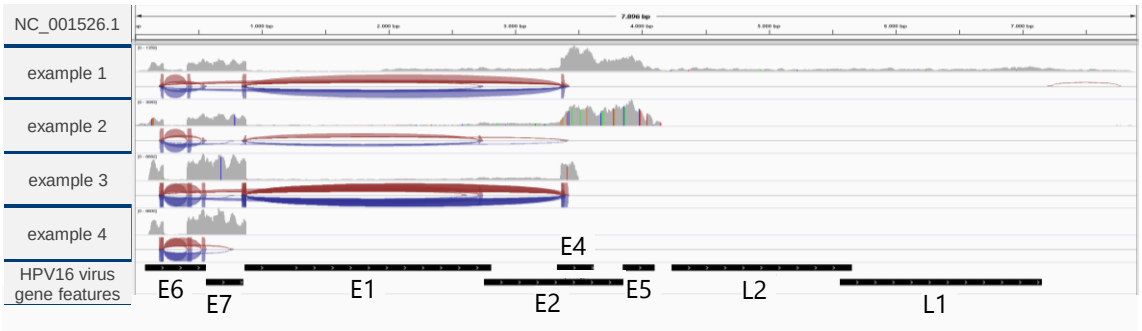


Figure 4-6: IGV alignment plot on HPV16 of four exemplary CESC datasets. Example 1 shows a completely expressed HPV genome with low expression of E1, E2, L2, and L1. The other displayed genomes are only partially expressed.

A non-negative matrix factorization of all HPV gene expression patterns revealed three basis components of co-expressed HPV genes. As L2 and L1 are expressed at very low levels, only the two sub-groups E6-E7 and E2-E4-E5 were further investigated (basis components 2 and 3) (Figure 4-7a). The latter three loci overlap either partially or completely. Thus, a complete distinction of the expression is not possible with the used

methods. However, alignment profiles indicate a stronger expression of E4 and E5 (Figure 4-6).

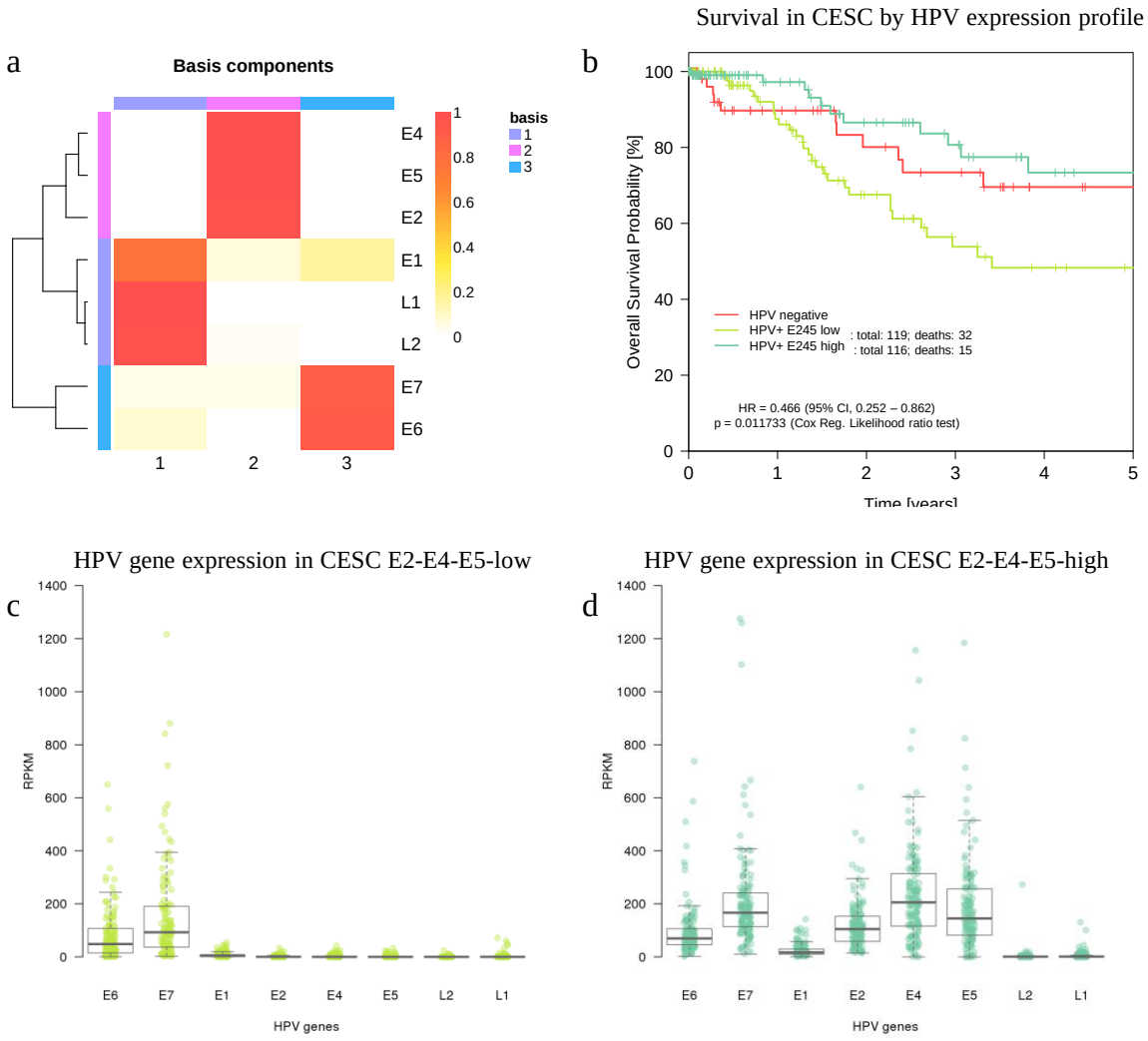


Figure 4-7: Survival in HPV-positive CESC patients stratified by HPV expression profile. a) HPV expression was dissected into basis components using non-negative matrix factorization. b) Survival analysis reveals significant differences when comparing (c) the E2-E4-E5 low expression subgroup to (d) the high expression group (HR = 0.466, p = 0.011733). HPV-negative survival plotted for comparison.

The subgroups among the HPV-positive CESC samples were classified by their mean expression of E2, E4, and E5 (E245-low and E245-high; section 2.5.6). The survival analysis of the defined subgroups showed significantly increased survival for the E245-high group compared to the E245-low group (HR: 0.466, 95 % CI: 0.25-0.86, p = 0.01) (Figure 4-7b). Expression patterns in the two subgroups show a clear distinction (Figure 4-7c & d) by expression of genes of the basis component 2 (E2, E4,

and E5). Differences in survival between the subgroups could not be reproduced in HNSC due to the low HPV-positive sample number.

The displayed results showed the capability of VIRGENE to reproduce clinical virus typing results and provide missing sample annotation. The gain in information for each sample increases the power to predict outcome in head and neck cancer patients. In addition to virus typing alone, viral expression profiles stratify cervical cancer patients and enable the prediction of outcome in subgroups of patient cohorts. The application of VIRGENE to clinical specimen can increase the depth of gained actionable knowledge and serve as a diagnostic and prognostic tool in virus-associated malignancies.

4.5 Challenges and Potential

4.5.1 The Virus Reference

For the identification of viruses using molecular tests, it is essential to use an appropriate reference. The reference could be a well-defined panel for microarray screening, a sufficient set of PCR primers to identify viruses, or a well-curated virus genome database as a reference for NGS alignments. For this work, I used the NCBI RefSeq viral genomes database in different versions from 2013 to 2016 (chapter 2.2).

During the time course of this thesis, the database changed significantly. In the beginning, the database contained 4,715 genomic sequences. By the end of 2016, this number had grown to 7,807 sequences. However, the maintenance of the database might introduce changes other than just additions of further virus types. For instance, the human coronavirus (NC_004718.3) detected in the SARS+ mouse samples (section 3.1.1) was removed from the latest version. Other virus genomes like HPV16 have changed over time from genome version NC_001526.1 to version NC_001526.4 including changes in genome structure and gene order. Regular updates of the database will impede comparability of experiments at different time points.

Influenza viruses, for example, have a segmented genome [144]. Each segment is represented by a single FASTA record in the virus database. This does not impair the alignment process, but will have an effect due to VIRGENE's taxonomy filter. The

segments would appear closely related in terms of the taxonomy and only the segment with highest VGC would be reported.

Some virus genomes are very similar, like for instance EBV (NC_007605.1) and EBV2 (NC_009334.1) [145]. Especially in the case of latent infection with the expression of only a subset of genes, a differentiation of both remains challenging. For other viruses, like papillomaviruses for example, representative sequences of different virus genotypes are stored in the database. HPV16 is the model species of Alphapapillomavirus 9. Its genome represents HPV genotypes 16, 31, 33, 35, 52, 58, and 67. HPV18 is representative for genotypes 18, 39, 45, 59, 68, 70, 85 and 97 of Alphapapillomavirus 7. Each genotype differs from the others by at least 10 % genome sequence difference of the L1 ORF [144]. In HPV, if only the E6 and E7 genes were present and expressed, a distinction could be challenging and might require significant downstream analyses.

These examples illustrate the need for a well-curated database of reference virus genomes. For the screening of cancer cohorts, the contained plant viruses or bacterial phages might be less critical. Only around 31 % of the contained genomes were derived from vertebrate host viruses, in total (Figure 2-3). The dataset might be reduced to improve alignment speed and accuracy. However, the detected phage sequences in some of the cell line samples (section 4.3.1) justify the retention of representative genomes of foreign hosts. On the other hand, some of the investigated viruses were missing from the database. These might be added manually to curate a cancer virome database or a human disease database.

4.5.2 Incorporated Algorithms

VIRGENE has been developed to determine the expression levels of viral genes. Methods for the quantification of gene expression have been proposed in the past and are under constant development. Recent bioinformatic improvements include the development of algorithms for alignment-free quantification of gene expression, for instance Sailfish [146], Salmon [147], or kallisto [148], compared to RNA-Seq aligners like STAR [67], TopHat2 [149], or ContextMap2 [150]. While most methods perform similarly [151], alignment-free quantification algorithms usually run faster and require less memory. For the further development of VIRGENE, different aligners or alignment-free methods might be evaluated. However, the downstream analysis of

alignment-derived sequence contigs has been critical in some cases of my study, for instance in detecting the sequence similarity of the AbMLV and EMCV-derived contigs to vectors in section 4.3.1. This would not have been possible using an alignment-free approach. Hence, the use of alignment-free software will have to be judged critically.

4.5.3 Non-polyadenylated RNA

For this study, only poly-A-enriched RNA-Seq data were used. This is sufficient for human mRNA studies and most viruses. *Flaviviridae* are a family of viruses with a single-stranded positive-direction RNA genome [152, 153]. The family includes members like the Yellow fever virus, Hepatitis C virus, Dengue virus, or Zika virus. The genome is not polyadenylated and serves as mRNA. Translation starts immediately from the genome. The genome contains only one transcription start-site and viral proteins are produced from a single premature poly-protein. Hence, the genomes of *Flaviviridae* are infectious by themselves. Replication is then initiated by the virus's RdRp. Since the RNA lacks a poly-A tail, the expression of these viruses could not be measured by poly-A enriched RNA-Seq. Rather, a total RNA-approach would be suitable with subsequent depletion of ribosomal RNA. Especially in the light of virus-induced carcinogenesis, expression analysis of the liver carcinoma-associated Hepatitis C virus might be interesting.

4.5.4 Viral Biomarkers

Prognostic, diagnostic, and therapeutic biomarkers are being developed to identify the risk of cancer, detect disease, to monitor disease or progression, or to target cancer cells in the patient. Some cancer-associated viruses are considered direct carcinogens. They express oncogenes that directly promote cancer cell transformation, for instance HPV, EBV, or KSHV. These viruses are present in every cancer cell and have to maintain the malignant status of the cell. They represent ideal biomarkers in the sense that they are exclusively expressed in a patient's cancer cells. The HPV-status in HNSC patients, for instance, is a prognostic marker for prolonged survival (chapter 4.4) [142, 143]. The cancer-restricted property has led to targeted therapies as well, like HPV E6-targeted T cell transfer [154], EBV LMP2-pulsed dendritic cell transfer [155], or modified vaccinia virus therapies targeting HPV [135] or EBV genes [156].

In this work, a comparison of EBV in cell line and primary tumor data revealed a diverse expression pattern across all samples. I saw an expression across large parts of the EBV genome in cell lines but less in most tumors. One reason for the discrepancy could be immune surveillance in tumors. The selective pressure of the human immune system leads to downregulation of EBV genes in tumors, but not in cell lines. However, I could confirm consistent expression of EBV BamHI-A rightward transcripts (BARTs) in EBV-positive tumor samples and in most analyzed cell lines (chapter 3.6) [98]. Although BARTs have been known for almost 30 years now [157], their protein functions are not completely resolved yet. However, their consistent expression across all samples suggest a role as markers for viral persistence. Different groups have recently shown that EBV associated tumors express PD-L1 and can thus be potentially treated with checkpoint blockade anti-PD-1/PD-L1 therapy [158, 159]. Very recently, a group has shown that the expression of PD-L1 is sufficient to reduce cytotoxicity in immunogenic tumors [160]. Hence, measurements of BARTs using low-cost and fast screening methods like PCR-based tests could lead to an improvement in patient therapies for EBV-associated malignancies in future, in contrast to the commonly performed measurements of EBV latency genes. One advantage of using qPCR instead of immunohistochemical evaluation would be that the viruses could be detected even if proteins were downregulated. The expression of BARTs might not lead to detectable proteins; however, the transcripts are detectable with RNA-Seq and qPCR. The application of VIRGENE to further cancer cohorts or other disease entities might enable identification of further biomarkers and the development of routine screening methods directed against these novel biomarkers.

5 Conclusion

Within the scope of this thesis, murine and human RNA-Seq data were analyzed for their content of viral nucleic acids. Two pipelines were developed throughout the course of this thesis: VirusID for the detection and identification of viral signals in host RNA-Seq data, and VIRGENE, enabling the identification and quantification of viral gene expression.

The identification of viruses in cell lines illustrates the importance of such tools. If cell lines are not routinely screened for viruses by conventional methods, *in silico* screening of the sequencing data can reveal viral or vectoral contaminations. The application of VIRGENE to cancer cohorts revealed the potential of virus detection to stratify patients by viral status or virus expression profiles. Further, identified viral genes could serve as diagnostic biomarkers or therapeutic targets.

VirusID has been challenged in two benchmarks. The low nucleic acid content of the test samples aggravated the analyses; however, VirusID was slightly superior to other investigator's methods or algorithms. Regular evaluation and subsequent iterative improvements have revealed and eradicated some of the pitfalls in detecting viruses from RNA-Seq data. Further development could include evaluation of the software performance in terms of speed and memory usage, or consideration and evaluation of other alignment or gene quantification software.

The presented study provides evidence for the as yet neglected differential virus expression in host malignant diseases. The design of the pipelines enables the use of standard high-throughput sequencing data. Hence, the platform can be applied to existing data, as well as to future transcriptomic studies to enable identification of viral expression profiles in disease studies.

6 Literature

1. Rous P (1911) A Sarcoma of the Fowl Transmissible by an Agent Seperable from the Tumor Cells. J Exp Med 13(4): 397–411
2. Suttle CA (2007) Marine viruses — major players in the global ecosystem. Nat Rev Micro 5(10): 801–812. doi: 10.1038/nrmicro1750
3. Zur Hausen H, Fox JG (2006) Infections causing human cancer. Wiley-VCH, Weinheim
4. Moore PS, Chang Y (2010) Why do viruses cause cancer? Highlights of the first century of human tumour virology. Nat Rev Cancer 10(12): 878–889. doi: 10.1038/nrc2961
5. Castle JC, Boegel S, Bukur T et al. Genomics Meets Cancer Immunotherapy. In: Britten CM, Huber C (eds) Cancer immunotherapy meets oncology: In honor of Christoph Huber, pp 229–236
6. Butel JS (2000) Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. Carcinogenesis 21(3): 405–426. doi: 10.1093/carcin/21.3.405
7. Peveling-Oberhag J, Arcaini L, Hansmann M-L et al. (2013) Hepatitis C-associated B-cell non-Hodgkin lymphomas. Epidemiology, molecular signature and clinical management. Journal of Hepatology 59(1): 169–177. doi: 10.1016/j.jhep.2013.03.018
8. Gallo RC (1998) BIOMEDICINE: The Enigmas of Kaposi's Sarcoma. Science 282(5395): 1837–1839. doi: 10.1126/science.282.5395.1837
9. Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35(3): 235–241
10. Berns KI (1990) Parvovirus replication. Microbiol. Rev. 54(3): 316–329
11. Baltimore D, Eggers HJ, Franklin RM et al. (1963) Poliovirus-induced RNA Polymerase and the Effects of Virus-specific Inhibitors on its Production. Proc Natl Acad Sci U S A 49(6): 843–849
12. Temin HM, Mizutani S (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226(5252): 1211–1213
13. Baltimore D (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature 226(5252): 1209–1211

14. Seeger C, Mason WS (2000) Hepatitis B Virus Biology. Microbiol. Mol. Biol. Rev. 64(1): 51–68. doi: 10.1128/MMBR.64.1.51-68.2000
15. Horvath CM, Williams MA, Lamb RA (1990) Eukaryotic coupled translation of tandem cistrons: identification of the influenza B virus BM2 polypeptide. EMBO J 9(8): 2639–2647
16. Walsh D, Mathews MB, Mohr I (2013) Tinkering with Translation: Protein Synthesis in Virus-Infected Cells. Cold Spring Harb Perspect Biol 5(1): a012351. doi: 10.1101/cshperspect.a012351
17. Boldogh I, Albrecht T, Porter DD (1996) Persistent Viral Infections
18. Spina CA, Anderson J, Archin NM et al. (2013) An in-depth comparison of latent HIV-1 reactivation in multiple cell model systems and resting CD4+ T cells from aviremic patients. PLoS Pathog 9(12): e1003834. doi: 10.1371/journal.ppat.1003834
19. Ho Y-C, Shan L, Hosmane NN et al. (2013) Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell 155(3): 540–551. doi: 10.1016/j.cell.2013.09.020
20. Bagga S, Bouchard MJ (2014) Cell cycle regulation during viral infection. In: Noguchi E, Gadaleta MC (eds) Cell Cycle Control: Mechanisms and Protocols, 2nd ed. 2014, vol 1170. Springer New York, New York, NY, pp 165–227
21. Lévy P, Bartosch B (2016) Metabolic reprogramming: a hallmark of viral oncogenesis. Oncogene 35(32): 4155–4164. doi: 10.1038/onc.2015.479
22. Leslie AJ, Pfafferott KJ, Chetty P et al. (2004) HIV evolution: CTL escape mutation and reversion after transmission. Nat Med 10(3): 282–289. doi: 10.1038/nm992
23. Ashrafi GH, Brown DR, Fife KH et al. (2006) Down-regulation of MHC class I is a property common to papillomavirus E5 proteins. Virus Research 120(1–2): 208–211. doi: 10.1016/j.virusres.2006.02.005
24. Hansen TH, Bouvier M (2009) MHC class I antigen presentation: learning from viral evasion strategies. Nat Rev Immunol 9(7): 503–513. doi: 10.1038/nri2575
25. Mesri EA, Feitelson MA, Munger K (2014) Human viral oncogenesis: a cancer hallmarks analysis. Cell Host Microbe 15(3): 266–282. doi: 10.1016/j.chom.2014.02.011
26. Beijerinck MW (1899) Uber Ein Contagium Vivum Fluidum Als Ursache Der Fleckenkrankheit Des Tabahsblattes. Rivista di Patologia Vegetale 7: 387–389

27. Hirst GK (1942) The Quantitative Determination of Influenza Virus and Antibodies by Means of Red Cell Agglutination. Journal of Experimental Medicine 75(1): 49–64. doi: 10.1084/jem.75.1.49
28. POTTER CW, OXFORD JS (1979) Determinants of Immunity to Influenza Infection in Man. Br Med Bull 35(1): 69–75. doi: 10.1093/oxfordjournals.bmb.a071545
29. Cowling BJ, Chan KH, Fang VJ et al. (2010) Comparative Epidemiology of Pandemic and Seasonal Influenza A in Households. N Engl J Med 362(23): 2175–2184. doi: 10.1056/NEJMoa0911530
30. Dulbecco R, Vogt M (1953) Some Problems of Animal Virology as Studied by the Plaque Technique. Cold Spring Harb Symp Quant Biol 18: 273–279. doi: 10.1101/SQB.1953.018.01.039
31. Kausche GA, Pfankuch E, Ruska H Die Sichtbarmachung von pflanzlichem Virus im Übermikroskop. Naturwissenschaften 27(18): 292–299. doi: 10.1007/BF01493353
32. Goldsmith CS, Miller SE (2009) Modern Uses of Electron Microscopy for Detection of Viruses. Clin. Microbiol. Rev. 22(4): 552–563. doi: 10.1128/CMR.00027-09
33. Engvall E, Perlmann P (1971) Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. Immunochemistry 8(9): 871–874. doi: 10.1016/0019-2791(71)90454-X
34. van Weemen BK, Schuurs A (1971) Immunoassay using antigen-enzyme conjugates. FEBS Letters 15(3): 232–236. doi: 10.1016/0014-5793(71)80319-8
35. Burnette W (1981) “Western Blotting”: Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Analytical Biochemistry 112(2): 195–203. doi: 10.1016/0003-2697(81)90281-5
36. Roy-Burman P, Dougherty M, Pal BK et al. (1976) Assay for type C virus in mouse sera based on particulate reverse transcriptase activity. J. Virol. 19(3): 1107–1110
37. Pyra H, Böni J, Schüpbach J (1994) Ultrasensitive retrovirus detection by a reverse transcriptase assay based on product enhancement. PNAS 91(4): 1544–1548. doi: 10.1073/pnas.91.4.1544

38. Brahic M, Haase AT (1978) Detection of viral sequences of low reiteration frequency by in situ hybridization. PNAS 75(12): 6125–6129
39. Loni MC, Green M (1973) Detection of Viral DNA Sequences in Adenovirus-Transformed Cells by In Situ Hybridization. J. Virol. 12(6): 1288–1292
40. Clementi M, Menzo S, Bagnarelli P et al. (1993) Quantitative PCR and RT-PCR in virology. PCR Methods Appl 2(3): 191–196
41. Heid CA, Stevens J, Livak KJ et al. (1996) Real time quantitative PCR. Genome Research 6(10): 986–994. doi: 10.1101/gr.6.10.986
42. Schena M, Shalon D, Davis RW et al. (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270(5235): 467–470. doi: 10.1126/science.270.5235.467
43. Wang D, Coscoy L, Zylberberg M et al. (2002) Microarray-based detection and genotyping of viral pathogens. PNAS 99(24): 15687–15692. doi: 10.1073/pnas.242579699
44. Sanger F, Air GM, Barrell BG et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265(5596): 687–695
45. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12): 5463–5467
46. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74(2): 560–564
47. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011): 931–945. doi: 10.1038/nature03001
48. Muir P, Li S, Lou S et al. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol 17: 53. doi: 10.1186/s13059-016-0917-0
49. Liu L, Li Y, Li S et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364. doi: 10.1155/2012/251364
50. Nagalakshmi U, Wang Z, Waern K et al. (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 320(5881): 1344–1349. doi: 10.1126/science.1158441
51. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1): 57–63. doi: 10.1038/nrg2484

52. Modrof J, Berting A, Kreil TR (2014) Parallel Evaluation of Broad Virus Detection Methods. PDA Journal of Pharmaceutical Science and Technology 68(6): 572–578. doi: 10.5731/pdajpst.2014.01014
53. Leinonen R, Sugawara H, Shumway M (2010) The Sequence Read Archive. Nucleic Acids Research 39(Database): D19–D21. doi: 10.1093/nar/gkq1019
54. Gill BJ, Pisapia DJ, Malone HR et al. (2014) MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. Proceedings of the National Academy of Sciences 111(34): 12550–12555. doi: 10.1073/pnas.1405839111
55. Abate F, Ambrosio MR, Mundo L et al. (2015) Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma. PLoS Pathog 11(10): e1005158. doi: 10.1371/journal.ppat.1005158
56. Peng X, Gralinski L, Armour CD et al. (2010) Unique Signatures of Long Noncoding RNA Expression in Response to Virus Infection and Altered Innate Immune Signaling. mBio 1(5): e00206-10–e00206-18. doi: 10.1128/mBio.00206-10
57. Pruitt KD, Brown GR, Hiatt SM et al. (2013) RefSeq: an update on mammalian reference sequences. Nucl. Acids Res. 42(D1): D756–D763. doi: 10.1093/nar/gkt1114
58. Brister JR, Ako-adjei D, Bao Y et al. (2015) NCBI Viral Genomes Resource. Nucleic Acids Research 43(D1): D571–D577. doi: 10.1093/nar/gku1207
59. Cock, P. J. A., Antao T, Chang JT et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11): 1422–1423. doi: 10.1093/bioinformatics/btp163
60. Langmead B, Trapnell C, Pop M et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3): R25. doi: 10.1186/gb-2009-10-3-r25
61. Bonfert T, Csaba G, Zimmer R et al. (2013) Mining RNA–Seq Data for Infections and Contaminations. PLOS ONE 8(9): e73071. doi: 10.1371/journal.pone.0073071
62. Robinson JT, Thorvaldsdóttir H, Winckler W et al. (2011) Integrative genomics viewer. Nat Biotechnol 29(1): 24–26. doi: 10.1038/nbt.1754

63. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6): 863–864. doi: 10.1093/bioinformatics/btr026
64. R Core Team (2014) R: A Language and Environment for Statistical. R Foundation for Statistical Computing, Vienna, Austria
65. Li H, Handsaker B, Wysoker A et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079. doi: 10.1093/bioinformatics/btp352
66. Boegel S, Lower M, Schafer M et al. (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12): 102. doi: 10.1186/gm403
67. Dobin A, Davis CA, Schlesinger F et al. (2012) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1): 15–21. doi: 10.1093/bioinformatics/bts635
68. Chen Y, Yao H, Thompson EJ et al. (2013) VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29(2): 266–267. doi: 10.1093/bioinformatics/bts665
69. Scholtalbers J, Boegel S, Bukur T et al. (2015) TCLP: an online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. Genome Med 7(1): 1081. doi: 10.1186/s13073-015-0240-5
70. Hsu F, Kent WJ, Clawson H et al. (2006) The UCSC Known Genes. Bioinformatics 22(9): 1036–1046. doi: 10.1093/bioinformatics/btl048
71. Mortazavi A, Williams BA, McCue K et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5(7): 621–628. doi: 10.1038/nmeth.1226
72. Tang K-W, Alaei-Mahabadi B, Samuelsson T et al. (2013) The landscape of viral expression and host gene fusion and adaptation in human cancer. Nature Communications 4: 2513. doi: 10.1038/ncomms3513
73. Shannon CE (1948) A Mathematical Theory of Communication. The Bell System Technical Journal 27: 379-423, 623-656
74. Warnes GR, Bolker B, Bonebakker L et al. gplots: Various R Programming Tools for Plotting Data
75. Gaujoux R, Seoighe C (2010) A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 11(1): 367. doi: 10.1186/1471-2105-11-367
76. Terry M. Therneau (2015) A Package for Survival Analysis in S

77. Cox DR (1992) Regression Models and Life-Tables. http://link.springer.com/content/pdf/10.1007%2F978-1-4612-4380-9_37.pdf
78. Altschul S (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 25(17): 3389–3402. doi: 10.1093/nar/25.17.3389
79. Zhang Z, Schwartz S, Wagner L et al. (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1-2): 203–214. doi: 10.1089/10665270050081478
80. Morgulis A, Coulouris G, Raytselis Y et al. (2008) Database indexing for production MegaBLAST searches. Bioinformatics 24(16): 1757–1764. doi: 10.1093/bioinformatics/btn322
81. Baruzzo G, Hayer KE, Kim EJ et al. (2016) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Meth. doi: 10.1038/nmeth.4106
82. Kostic AD, Ojesina AI, Pedamallu CS et al. (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature Biotechnology 29(5): 393–396. doi: 10.1038/nbt.1868
83. Bhaduri A, Qu K, Lee CS et al. (2012) Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics 28(8): 1174–1175. doi: 10.1093/bioinformatics/bts100
84. Li JW, Wan R, Yu CS et al. (2013) ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics 29(5): 649–651. doi: 10.1093/bioinformatics/btt011
85. Wang Q, Jia P, Zhao Z (2013) VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One 8(5): e64465. doi: 10.1371/journal.pone.0064465
86. Schelhorn S-E, Fischer M, Tolosi L et al. (2013) Sensitive Detection of Viral Transcripts in Human Tumor Transcriptomes. PLOS Computational Biology 9(10): e1003228. doi: 10.1371/journal.pcbi.1003228
87. Boegel S, Lower M, Bukur T et al. (2014) A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 3(8): e954893. doi: 10.4161/21624011.2014.954893

88. Barretina J, Caponigro G, Stransky N et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391): 603–607. doi: 10.1038/nature11003
89. Klijn C, Durinck S, Stawiski EW et al. (2015) A comprehensive transcriptional portrait of human cancer cell lines. Nature Biotechnology 33(3): 306–312. doi: 10.1038/nbt.3080
90. Marioni JC, Mason CE, Mane SM et al. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18(9): 1509–1517. doi: 10.1101/gr.079558.108
91. Münger K, Phelps WC, Bubb V et al. (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(10): 4417–4421
92. UCSC Genome Browser Data File Formats: BED format. http://www.genome.ucsc.edu/FAQ/FAQformat.html#format1. Accessed 13 Mar 2017
93. Smit A, Hubley R, Green P (2013-2015) RepeatMasker Open-4.0
94. Merten O-W (2002) Virus contaminations of cell cultures – A biotechnological view. Cytotechnology 39(2): 91–116. doi: 10.1023/A:1022969101804
95. Henle W, Diehl V, Kohn G et al. (1967) Herpes-type virus and chromosome marker in normal leukocytes after growth with irradiated Burkitt cells. Science 157(3792): 1064–1065
96. Thompson MP, Kurzrock R (2004) Epstein-Barr virus and cancer. Clin Cancer Res 10(3): 803–821
97. Jones MD, Foster L, Sheedy T et al. (1984) The EB virus genome in Daudi Burkitt's lymphoma cells has a deletion similar to that observed in a non-transforming strain (P3HR-1) of the virus. EMBO J 3(4): 813–821
98. Yamamoto T, Iwatsuki K (2012) Diversity of Epstein-Barr virus BamHI-A rightward transcripts and their expression patterns in lytic and latent infections. J Med Microbiol 61(Pt 10): 1445–1453. doi: 10.1099/jmm.0.044727-0
99. Zur Hausen H (1996) Papillomavirus infections — a major cause of human cancers. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1288(2): F55-F78. doi: 10.1016/0304-419X(96)00020-0

100. Powell NG, Evans M (2015) Human papillomavirus-associated head and neck cancer: Oncogenic mechanisms, epidemiology and clinical behaviour. Diagnostic Histopathology 21(2): 49–64. doi: 10.1016/j.mpdhp.2015.02.003
101. Durst M, Kleinheinz A, Hotz M et al. (1985) The physical state of human papillomavirus type 16 DNA in benign and malignant genital tumours. J Gen Virol 66 (Pt 7): 1515–1522. doi: 10.1099/0022-1317-66-7-1515
102. Corden SA, Sant-Cassia LJ, Easton AJ et al. (1999) The integration of HPV-18 DNA in cervical carcinoma. Mol Pathol 52(5): 275–282
103. ICH Harmonised Tripartite Guideline (1997) viral safety evaluation of biotechnology products derived from cell lines of human or animal origin Q5A (R1). Current Step 4
104. Vatsan RS, Bross PF, Liu K et al. (2013) Regulation of immunotherapeutic products for cancer and FDA's role in product development and clinical evaluation. J Immunother Cancer 1: 5. doi: 10.1186/2051-1426-1-5
105. Chiu CY (2013) Viral pathogen discovery. Host–microbe interactions: fungi/parasites/viruses 16(4): 468–478. doi: 10.1016/j.mib.2013.05.001
106. Lipkin WI (2010) Microbe hunting. Microbiol Mol Biol Rev 74(3): 363–377. doi: 10.1128/MMBR.00007-10
107. Moore RA, Warren RL, Freeman JD et al. (2011) The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS One 6(5): e19838. doi: 10.1371/journal.pone.0019838
108. Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotechnol 22(11): 1393–1398. doi: 10.1038/nbt1026
109. Uphoff CC, Denkmann SA, Steube KG et al. (2010) Detection of EBV, HBV, HCV, HIV-1, HTLV-I and -II, and SMRV in human and other primate cell lines. J Biomed Biotechnol 2010: 904767. doi: 10.1155/2010/904767
110. Schuurman R, van Steenis B, Sol C (1991) Bovine polyomavirus, a frequent contaminant of calf serum. Biologicals 19(4): 265–270. doi: 10.1016/S1045-1056(05)80014-4
111. Fogh J, Holmgren NB, Ludovici PP (1971) A review of cell culture contaminations. In Vitro 7(1): 26–41
112. Gey, G., Coffman, W. D., & Kubicek, M. T. (1952) Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Research 12(4): 264–265

113. Meissner JD (1999) Nucleotide sequences and further characterization of human papillomavirus DNA present in the CaSki, SiHa and HeLa cervical carcinoma cell lines. J Gen Virol 80 (Pt 7): 1725–1733. doi: 10.1099/0022-1317-80-7-1725
114. Wentzensen N, Vinokurova S, Doeberitz MvK (2004) Systematic Review of Genomic Integration Sites of Human Papillomavirus Genomes in Epithelial Dysplasia and Invasive Cancer of the Female Lower Genital Tract. Cancer Res 64(11): 3878–3884. doi: 10.1158/0008-5472.CAN-04-0009
115. Löwer R, Löwer J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. PNAS 93(11): 5177–5184
116. Young DF, Carlos TS, Hagmaier K et al. (2007) AGS and other tissue culture cells can unknowingly be persistently infected with PIV5; a virus that blocks interferon signalling by degrading STAT1. Virology 365(1): 238–240. doi: 10.1016/j.virol.2007.03.061
117. Graham FL, Smiley J, Russell WC et al. (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J Gen Virol 36(1): 59–74. doi: 10.1099/0022-1317-36-1-59
118. Deichmann M, Huder JB, Kleist C et al. (2005) Detection of reverse transcriptase activity in human melanoma cell lines and identification of a murine leukemia virus contaminant. Arch Dermatol Res 296(8): 345–352. doi: 10.1007/s00403-004-0501-4
119. Hue S, Gray ER, Gall A et al. (2010) Disease-associated XMRV sequences are consistent with laboratory contamination. Retrovirology 7(1): 111. doi: 10.1186/1742-4690-7-111
120. Lin Z, Puetter A, Coco J et al. (2012) Detection of Murine Leukemia Virus in the Epstein-Barr Virus-Positive Human B-Cell Line JY, Using a Computational RNA-Seq-Based Exogenous Agent Detection Pipeline, PARSES. J. Virol. 86(6): 2970–2977. doi: 10.1128/JVI.06717-11
121. Moar MH, Klein G (1978) Detection of Epstein-Barr virus (EBV) DNA sequences using in situ hybridization. Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis 519(1): 49–64. doi: 10.1016/0005-2787(78)90061-8

122. Kluin-Nelemans HC, Limpens J, Meerabux J et al. (1991) A new non-Hodgkin's B-cell line (DoHH2) with a chromosomal translocation t(14;18)(q32;q21). Leukemia 5(3): 221–224
123. Bertrand S, Berger R, Philip T et al. (1981) Variant translocation in a non endemic case of Burkitt's lymphoma: T (8;22) in an Epstein-Barr virus negative tumour and in a derived cell line. European Journal of Cancer (1965) 17(5): 577–581. doi: 10.1016/0014-2964(81)90060-8
124. Sramkoski RM, Pretlow TG, Giaconia JM et al. (1999) A new human prostate carcinoma cell line, 22Rv1. In Vitro: Journal of the Tissue Culture Association 35(7): 403–409
125. Webber MM (1999) A human prostatic stromal myofibroblast cell line WPMY-1: A model for stromalepithelial interactions in prostatic neoplasia. Carcinogenesis 20(7): 1185–1192. doi: 10.1093/carcin/20.7.1185
126. BUEHRING GC, EBY EA, EBY MJ (2004) Cell Line Cross-Contamination: How Aware are Mammalian Cell Culturists of the Problem and how to Monitor it? In Vitro Cell Dev Biol Anim 40(7): 211. doi: 10.1290/1543-706X(2004)40<211:CLCHAA>2.0.CO;2
127. Cantalupo PG, Katz JP, Pipas JM (2015) HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18. J Virol 89(8): 4051–4057. doi: 10.1128/JVI.03365-14
128. Cobbs CS, Harkins L, Samanta M et al. (2002) Human Cytomegalovirus Infection and Expression in Human Malignant Glioma. Cancer Res 62(12): 3347–3350
129. Dziurzynski K, Chang SM, Heimberger AB et al. (2012) Consensus on the role of human cytomegalovirus in glioblastoma. Neuro Oncol 14(3): 246–255. doi: 10.1093/neuonc/nor227
130. Lau SK, Chen Y-Y, Chen W-G et al. (2005) Lack of association of cytomegalovirus with human brain tumors. Mod Pathol 18(6): 838–843. doi: 10.1038/modpathol.3800352
131. Yamashita Y, Ito Y, Isomura H et al. (2014) Lack of presence of the human cytomegalovirus in human glioblastoma. Modern Pathology 27(7): 922–929. doi: 10.1038/modpathol.2013.219

132. Tang K-W, Hellstrand K, Larsson E (2015) Absence of cytomegalovirus in high-coverage DNA sequencing of human glioblastoma multiforme. Int J Cancer 136(4): 977–981. doi: 10.1002/ijc.29042
133. van de Velde N, Boily M-C, Drolet M et al. (2012) Population-level impact of the bivalent, quadrivalent, and nonavalent human papillomavirus vaccines: a model-based analysis. J Natl Cancer Inst 104(22): 1712–1723. doi: 10.1093/jnci/djs395
134. Rice AE, Latchman YE, Balint JP et al. (2015) An HPV-E6/E7 immunotherapy plus PD-1 checkpoint inhibition results in tumor regression and reduction in PD-L1 expression. Cancer Gene Ther 22(9): 454–462. doi: 10.1038/cgt.2015.40
135. Borysiewicz L, Fiander A, Nimako M et al. (1996) A recombinant vaccinia virus encoding human papillomavirus types 16 and 18, E6 and E7 proteins as immunotherapy for cervical cancer. The Lancet 347(9014): 1523–1527. doi: 10.1016/S0140-6736(96)90674-1
136. Zur Hausen H (2002) Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer 2(5): 342–350. doi: 10.1038/nrc798
137. Flippot R, Malouf GG, Su X et al. (2016) Oncogenic viruses: Lessons learned using next-generation sequencing technologies. European Journal of Cancer 61: 61–68. doi: 10.1016/j.ejca.2016.03.086
138. The Cancer Genome Atlas Network (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517(7536): 576–582. doi: 10.1038/nature14129
139. Lesnikova I, Lidang M, Hamilton-Dutoit S et al. (2009) p16 as a diagnostic marker of cervical neoplasia: a tissue microarray study of 796 archival specimens. Diagnostic Pathology 4(1): 22. doi: 10.1186/1746-1596-4-22
140. Klussmann JP, Gültekin E, Weissenborn SJ et al. (2003) Expression of p16 Protein Identifies a Distinct Entity of Tonsillar Carcinomas Associated with Human Papillomavirus. The American Journal of Pathology 162(3): 747–753. doi: 10.1016/S0002-9440(10)63871-0
141. Ram H, Sarkar J, Kumar H et al. (2011) Oral Cancer: Risk Factors and Molecular Pathogenesis. J. Maxillofac. Oral Surg. 10(2): 132–137. doi: 10.1007/s12663-011-0195-z

142. Ang KK, Harris J, Wheeler R et al. (2010) Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. N Engl J Med 363(1): 24–35. doi: 10.1056/NEJMoa0912217
143. Muller M, Demeret C (2012) The HPV E2-Host Protein-Protein Interactions: A Complex Hijacking of the Cellular Network. The Open Virology Journal 6(1)
144. Bouvier NM, Palese P (2008) The biology of influenza viruses. Influenza Vaccines: Research, Development and Public Health Challenges 26, Supplement 4: D49-D53. doi: 10.1016/j.vaccine.2008.07.039
145. Dolan A, Addison C, Gatherer D et al. (2006) The genome of Epstein–Barr virus type 2 strain AG876. Virology 350(1): 164–170. doi: 10.1016/j.virol.2006.01.015
146. Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32(5): 462–464. doi: 10.1038/nbt.2862
147. Patro R, Duggal G, Love MI et al. (2015) Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference
148. Bray NL, Pimentel H, Melsted P et al. (2016) Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34(5): 525–527. doi: 10.1038/nbt.3519
149. Kim D, Pertea G, Trapnell C et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4): R36. doi: 10.1186/gb-2013-14-4-r36
150. Bonfert T, Kirner E, Csaba G et al. (2015) ContextMap 2: fast and accurate context-based RNA-seq mapping. BMC Bioinformatics 16: 122. doi: 10.1186/s12859-015-0557-5
151. Teng M, Love MI, Davis CA et al. (2016) A benchmark for RNA-seq quantification pipelines. Genome Biol 17: 74. doi: 10.1186/s13059-016-0940-1
152. Chambers TJ, Hahn CS, Galler R et al. (1990) Flavivirus genome organization, expression, and replication. Annu Rev Microbiol 44: 649–688. doi: 10.1146/annurev.mi.44.100190.003245
153. Khromykh AA, Meka H, Guyatt KJ et al. (2001) Essential role of cyclization sequences in flavivirus RNA replication. J Virol 75(14): 6719–6728. doi: 10.1128/JVI.75.14.6719-6728.2001

154. Draper LM, Kwong MLM, Gros A et al. (2015) Targeting of HPV-16+ Epithelial Cancer Cells by TCR Gene Engineered T Cells Directed against E6. Clin Cancer Res 21(19): 4431–4439. doi: 10.1158/1078-0432.CCR-14-3341
155. Lin C-L, Lo W-F, Lee T-H et al. (2002) Immunization with Epstein-Barr Virus (EBV) Peptide-pulsed Dendritic Cells Induces Functional CD8+ T-Cell Immunity and May Lead to Tumor Regression in Patients with EBV-positive Nasopharyngeal Carcinoma. Cancer Res 62(23): 6952–6958
156. Taylor GS, Haigh TA, Gudgeon NH et al. (2004) Dual Stimulation of Epstein-Barr Virus (EBV)-Specific CD4+- and CD8+-T-Cell Responses by a Chimeric Antigen Construct: Potential Therapeutic Vaccine for EBV-Positive Nasopharyngeal Carcinoma. J Virol 78(2): 768–778. doi: 10.1128/JVI.78.2.768-778.2004
157. Hitt MM, Allday MJ, Hara T et al. (1989) EBV gene expression in an NPC-related tumour. EMBO J 8(9): 2639–2651
158. Green MR, Rodig S, Juszczynski P et al. (2012) Constitutive AP-1 activity and EBV infection induce PD-L1 in Hodgkin lymphomas and posttransplant lymphoproliferative disorders: implications for targeted therapy. Clin Cancer Res 18(6): 1611–1618. doi: 10.1158/1078-0432.CCR-11-1942
159. Chen BJ, Chapuy B, Ouyang J et al. (2013) PD-L1 expression is characteristic of a subset of aggressive B-cell lymphomas and virus-associated malignancies. Clin Cancer Res 19(13): 3462–3473. doi: 10.1158/1078-0432.CCR-13-0855
160. Juneja VR, McGuire KA, Manguso RT et al. (2017) PD-L1 on tumor cells is sufficient for immune evasion in immunogenic tumors and inhibits CD8 T cell cytotoxicity. J Exp Med. doi: 10.1084/jem.20160801

Appendix

A. Data Accession Numbers

Table A-1: List of Cell Line Data Accession Numbers.

Cell line	Experiment accession	Project accession
PEO1	ERX013520, ERX013534, ERX013531	ERP000710
PEO14	ERX013521, ERX013535, ERX013527	ERP000710
PEO23	ERX013519, ERX013530, ERX013528	ERP000710
PEO4	ERX013524, ERX013525, ERX013533	ERP000710
C4-2B	ERX333833	ERP004209
LNCaP	ERX333831	ERP004209
Mel501	SRX006132	SRP000931
MeWo	SRX006122	SRP000931
MeWo	SRX006129	SRP000931
MeWo	SRX006130	SRP000931
DB	SRX079566	SRP001599
DOHH-2	SRX079565	SRP001599
Karpas 422	SRX079567	SRP001599
NU-DHL-1	SRX079568	SRP001599
NU-DUL-1	SRX079574	SRP001599
OCI-LY1	SRX079571	SRP001599
OCI-LY-19	SRX079573	SRP001599
OCI-LY7	SRX079572	SRP001599
SU-DHL-6	SRX079569	SRP001599
WSU-DLCL2	SRX079570	SRP001599
BT-474	SRX025828, SRX025829	SRP003186
KPL4	SRX025832	SRP003186
MCF7	SRX025827	SRP003186
SK-BR-3	SRX025830, SRX025831	SRP003186
MIP101	SRX026157	SRP003404
MIP101	SRX026158	SRP003404
MIP5FU	SRX026159	SRP003404
MIP5FU	SRX026160	SRP003404
HeLa S3	SRX026691	SRP003497
Hep G2	SRX026684	SRP003497
HUVEC	SRX026678	SRP003497
HUVEC	SRX026687	SRP003497
NHEK	SRX026673	SRP003497
22Rv1	SRX031908	SRP004637
C4-2B	SRX031909	SRP004637
CA-HPV-10	SRX031910	SRP004637
CWR22	SRX031911	SRP004637
DU 145	SRX031915	SRP004637
DU 145	SRX031916	SRP004637

DU 145	SRX031917	SRP004637
DU 145	SRX031918	SRP004637
DU 145	SRX031919	SRP004637
DU 145	SRX031920	SRP004637
DU 145	SRX031921	SRP004637
LAPC-4	SRX031922	SRP004637
MDA PCa 2b	SRX031934	SRP004637
NCI-H660	SRX031935	SRP004637
PC-3	SRX031936	SRP004637
PrEC	SRX031941	SRP004637
PrEC	SRX031942	SRP004637
PrSMC	SRX031943	SRP004637
PWR-1E	SRX031938	SRP004637
WPE1-NB26	SRX031963	SRP004637
WPMY-1	SRX031964	SRP004637
BT-20	SRX040501	SRP005601
BT-474	SRX040502	SRP005601
MCF 10A	SRX040503	SRP005601
MCF7	SRX040504	SRP005601
MDA-MB-231	SRX040505	SRP005601
MDA-MB-468	SRX040506	SRP005601
T-47D	SRX040507	SRP005601
ZR751	SRX040508	SRP005601
CAMA-1	SRX176115	SRP006575
HCC1419	SRX176116	SRP006575
HCC1500	SRX176117	SRP006575
UACC-812	SRX176119	SRP006575
ZR-75-30	SRX176120	SRP006575
HCC3153	SRX066556	SRP006908
MBC647	SRX066578	SRP006908
SUM1315	SRX066560	SRP006908
SUM149	SRX066554	SRP006908
SK-N-SH	SRX084673	SRP007461
HCC2337	SRX101336	SRP008746
HCC3153	SRX101334	SRP008746
MCF 10A	SRX099963	SRP008746
SUM1315O2	SRX101335	SRP008746
JY	SRX105330	SRP009262
BL-2	SRX105541	SRP009316
BL-30	SRX105538	SRP009316
BL-41	SRX105531	SRP009316
BL-58	SRX105542	SRP009316
BL-70	SRX105537	SRP009316
CA-46	SRX105539	SRP009316
DAUDI	SRX105540	SRP009316
EW	SRX105536	SRP009316
GUMBUS	SRX105530	SRP009316
NAMALWA	SRX105535	SRP009316
RAJI	SRX105532	SRP009316

RAMOS	SRX105534	SRP009316
THOMAS	SRX105533	SRP009316
U-87 GM	SRX110671	SRP009659
U-87 GM	SRX110672	SRP009659
K-562	SRX113647	SRP010061
MDA-MB-231	SRX147674	SRP013022
AGS	SRX181260	SRP014574
AGS	SRX181261	SRP014574
KATOIII	SRX181269	SRP014574
MKN1	SRX181267	SRP014574
MKN28	SRX181268	SRP014574
MKN45	SRX181251	SRP014574
MKN74	SRX181252	SRP014574
NCI-N87	SRX181255	SRP014574
SNU-1	SRX181250	SRP014574
SNU-16	SRX181248	SRP014574
SNU-216	SRX181246	SRP014574
SNU-484	SRX181245	SRP014574
SNU-5	SRX181249	SRP014574
SNU-520	SRX181243	SRP014574
SNU-601	SRX181244	SRP014574
SNU-620	SRX181247	SRP014574
SNU-638	SRX181256	SRP014574
SNU-668	SRX181257	SRP014574
SNU-719	SRX181259	SRP014574
A-431	SRX209056	SRP017465
A-431	SRX209057	SRP017465
Caco-2	SRX209063	SRP017465
Caco-2	SRX209064	SRP017465
HEK-293	SRX209065	SRP017465
HEK-293	SRX209066	SRP017465
HeLa	SRX209067	SRP017465
HeLa	SRX209068	SRP017465
Hep G2	SRX209069	SRP017465
Hep G2	SRX209070	SRP017465
PC-3	SRX209073	SRP017465
PC-3	SRX209074	SRP017465
RT-4	SRX209075	SRP017465
RT-4	SRX209076	SRP017465
U-2 OS	SRX209060	SRP017465
U-251 MG	SRX209058	SRP017465
U-251 MG	SRX209059	SRP017465
184A1	SRX317694	SRP026537
184B5	SRX317695	SRP026537
21MT1	SRX317696	SRP026537
21MT2	SRX317697	SRP026537
21NT	SRX317698	SRP026537
21PT	SRX317699	SRP026537
600MPE	SRX317700	SRP026537

BT-474	SRX317702	SRP026537
BT-483	SRX317703	SRP026537
BT-549	SRX317704	SRP026537
CAMA-1	SRX317705	SRP026537
EFM192A	SRX317706	SRP026537
EFM192B	SRX317707	SRP026537
EFM192C	SRX317708	SRP026537
HCC1143	SRX317709	SRP026537
HCC1395	SRX317710	SRP026537
HCC1419	SRX317711	SRP026537
HCC1428	SRX317712	SRP026537
HCC1569	SRX317713	SRP026537
HCC1599	SRX317714	SRP026537
HCC1806	SRX317715	SRP026537
HCC1937	SRX317716	SRP026537
HCC1954	SRX317717	SRP026537
HCC202	SRX317718	SRP026537
HCC2218	SRX317719	SRP026537
HCC3153	SRX317720	SRP026537
HCC38	SRX317721	SRP026537
HCC70	SRX317722	SRP026537
Hs 578T	SRX317723	SRP026537
JIMT1	SRX317724	SRP026537
LY2	SRX317725	SRP026537
MB157	SRX317726	SRP026537
MCF 10A	SRX317727	SRP026537
MCF 10F	SRX317728	SRP026537
MCF-12A	SRX317729	SRP026537
MCF7	SRX317730	SRP026537
MDA-MB-134-VI	SRX317731	SRP026537
MDA-MB-175	SRX317732	SRP026537
MDA-MB-231	SRX317733	SRP026537
MDA-MB-361	SRX317734	SRP026537
MX1	SRX317735	SRP026537
SK-BR-3	SRX317736	SRP026537
SUM1315	SRX317737	SRP026537
SUM149PT	SRX317738	SRP026537
SUM159PT	SRX317739	SRP026537
SUM225CWN	SRX317740	SRP026537
SUM229PE	SRX317741	SRP026537
SUM52PE	SRX317742	SRP026537
T-47D	SRX317743	SRP026537
T47D-KBluc	SRX317744	SRP026537
UACC-812	SRX347745	SRP026537
UACC-893	SRX317746	SRP026537
ZR-75-1	SRX317747	SRP026537
ZR-75-30	SRX317748	SRP026537
ZR-75-B	SRX317749	SRP026537
SK-BR-3	SRX329206	SRP028176

Table A-2: Accession numbers of EBL samples (project SRP062178).

Patient	Run	Experiment
BL15	SRR2149844	SRX1137085
BL19	SRR2149896	SRX1137154
BL20	SRR2149897	SRX1137156
BL22	SRR2149935	SRX1137158
BL23	SRR2149936	SRX1137161
BL27	SRR2149937	SRX1137163
BL30	SRR2149954	SRX1137255
BL35	SRR2149938	SRX1137191
BL40	SRR2149940	SRX1137194
BL43	SRR2149942	SRX1137195
BL45	SRR2149943	SRX1137198
BL48	SRR2149944	SRX1137201
BL49	SRR2149945	SRX1137204
BL50	SRR2149946	SRX1137205
BL60	SRR2149947	SRX1137206
BL62	SRR2149948	SRX1137207
BL69	SRR2149949	SRX1137209
BL80	SRR2149950	SRX1137210
BL81	SRR2149951	SRX1137212
BL84	SRR2149952	SRX1137213

B. Results of Seq2HLA

Table A-3: HLA Alleles and expression determined by Seq2HLA.

Sample	Locus	Allele 1	Confidence	Allele 2	Confidence	Locus Expression (RPKM)
TRON	A	A*02:17'	0,00	A*02:17	0,0574801	384,45
	B	B*41:01	0,000780081	B*40:02'	0,05311753	278,99
	C	C*02:02	0,000131754	C*17:01	0,007869289	118,88
SRR496398	A	A*02:17'	6,66E-14	A*02:17	0,0574801	110,26
	B	B*41:01	0,000513658	B*40:06'	0,007323567	35,40
	C	C*17:01	0,01595271	C*02:02	0,01280146	24,82
SRR097790	A	A*02:17'	2,10E-11	A*24:13'	0,07008888	187,47
	B	B*41:01	2,48E-07	B*40:02'	0,006225645	109,17
	C	C*17:01	6,75E-07	C*02:02	0,01056092	35,39
SRR925726	A	A*02:01'	0,00	A*02:01	0,0001648	128,98
	B	B*41:01	5,48E-08	B*40:02'	0,005038569	44,20
	C	C*17:01	5,77E-15	C*02:02	0,01047586	35,58
TRON	A	A*02:01'	3,03E-06	A*02:01	NA	9,19
	B	no	NA	B*41:01	NA	0,25
	C	C*05:01	0,001924597	C*05:01	0,0077544	20,65
SRR064286	A	A*02:01'	1,07E-09	A*02:01	1,00	8,05
	B	B*15:33	0,002615615	B*35:09'	1,00	1,63
	C	C*05:01	0,000916171	C*05:01	8,41E-05	51,64
SRR097789	A	A*02:01	1,22E-09	A*02:01	0,2116656	38,86
	B	B*15:02'	1,00	B*18:01'	0,001471015	3,05
	C	C*05:01	0,00485058	C*05:01	1,38E-05	78,68
SRR925723	A	A*02:01	0,0	A*02:01	0,0223903	18,44
	B	B*44:02'	4,52E-13	B*44:02	1,00	0,69
	C	C*05:01	0,009417334	C*05:01	4,10E-06	30,25

C. Viruses in Cell Lines

Table A-4: Human viruses in 186 cell lines.

			Human herpesvirus 4 (Epstein-Barr virus)	Human herpesvirus 4 type 2 (Epstein-Barr virus type 2)	Parainfluenza virus 5	Human adenovirus C	Human papillomavirus - 18	Encephalomyocarditis virus	Human endogenous retrovirus K113
cell_line	exp_acc	proj_acc	NC_007605.1	NC_009334.1	NC_006430.1	NC_001405.1	NC_001357.1	NC_001479.1	NC_022518.1
PEO1	ERX013520, ERX013534, ERX013531	ERP000710	0	0	0	0	0	0	0
PEO14	ERX013521, ERX013535, ERX013527	ERP000710	0	0	0	0	0	0	0
PEO23	ERX013519, ERX013530, ERX013528	ERP000710	0	0	0	0	0	0	0
PEO4	ERX013524, ERX013525, ERX013533	ERP000710	0	0	0	0	0	0	29,717
C4-2B	ERX333833	ERP004209	0	0	0	0	0	0	0
LNCaP	ERX333831	ERP004209	0	0	0	0	0	0	0
Mel501	SRX006132	SRP000931	0	0	0	0	0	0	0
MeWo	SRX006122	SRP000931	0	0	0	0	0	0	216,036
MeWo	SRX006129	SRP000931	0	0	0	0	0	0	219,142
MeWo	SRX006130	SRP000931	0	0	0	0	0	0	235,531
DB	SRX079566	SRP001599	0	0	0	0	0	0	0
DOHH-2	SRX079565	SRP001599	8811,275	0	0	5,427	0	0	0
Karpas 422	SRX079567	SRP001599	0	0	0	0	0	0	0
NU-DHL-1	SRX079568	SRP001599	0	0	0	0	0	0	0
NU-DUL-1	SRX079574	SRP001599	0	0	0	0	0	0	0
OCI-LY1	SRX079571	SRP001599	0	0	0	0	0	0	0
OCI-LY-19	SRX079573	SRP001599	0	0	0	0	0	1964,937	0
OCI-LY7	SRX079572	SRP001599	0	0	0	0	0	0	0
SU-DHL-6	SRX079569	SRP001599	0	0	0	0	0	0	0
WSU-DLCL2	SRX079570	SRP001599	0	0	0	0	0	0	0
BT-474	SRX025828, SRX025829	SRP003186	0	0	0	0	0	0	0
KPL4	SRX025832	SRP003186	0	0	0	0	0	0	0
MCF7	SRX025827	SRP003186	0	0	0	0	0	0	0
SK-BR-3	SRX025830, SRX025831	SRP003186	0	0	0	0	0	0	0
MIP101	SRX026157	SRP003404	0	0	0	0	0	0	0
MIP101	SRX026158	SRP003404	0	0	0	0	0	0	0

MIP5FU	SRX026159	SRP003404	0	0	0	0	0	0	0
MIP5FU	SRX026160	SRP003404	0	0	0	0	0	0	0
HeLa S3	SRX026691	SRP003497	0	0	0	0	2085,27	0	0
Hep G2	SRX026684	SRP003497	0	0	0	0	0	0	0
HUVEC	SRX026678	SRP003497	0	0	0	0	0	0	0
HUVEC	SRX026687	SRP003497	0	0	0	0	0	0	0
NHEK	SRX026673	SRP003497	0	0	0	0	0	0	0
22Rv1	SRX031908	SRP004637	0	0	0	0	0	0	0
C4-2B	SRX031909	SRP004637	0	0	0	0	0	0	0
CA-HPV-10	SRX031910	SRP004637	0	0	0	0	1809,026	0	0
CWR22	SRX031911	SRP004637	0	0	0	0	0	0	0
DU 145	SRX031915	SRP004637	0	0	0	0	0	0	34,363
DU 145	SRX031916	SRP004637	0	0	0	0	0	0	47,825
DU 145	SRX031917	SRP004637	0	0	0	0	0	0	0
DU 145	SRX031918	SRP004637	0	0	0	0	0	0	0
DU 145	SRX031919	SRP004637	0	0	0	0	0	0	0
DU 145	SRX031920	SRP004637	0	0	0	0	0	0	0
DU 145	SRX031921	SRP004637	0	0	0	0	0	0	0
LAPC-4	SRX031922	SRP004637	0	0	0	0	0	0	0
MDA PCa 2b	SRX031934	SRP004637	0	0	0	0	0	0	0
NCI-H660	SRX031935	SRP004637	0	0	0	0	0	0	0
PC-3	SRX031936	SRP004637	0	0	0	0	0	0	0
PrEC	SRX031941	SRP004637	0	0	0	0	0	0	25,811
PrEC	SRX031942	SRP004637	0	0	0	0	0	0	36,022
PrSMC	SRX031943	SRP004637	0	0	0	0	0	0	0
PWR-1E	SRX031938	SRP004637	0	0	0	0	0	0	0
WPE1-NB26	SRX031963	SRP004637	0	0	0	0	213,141	0	0
WPMY-1	SRX031964	SRP004637	0	0	0	0	1603,39	0	0
BT-20	SRX040501	SRP005601	0	0	0	0	0	0	0
BT-474	SRX040502	SRP005601	0	0	0	0	0	0	0
MCF 10A	SRX040503	SRP005601	0	0	0	0	0	0	37,339
MCF7	SRX040504	SRP005601	0	0	0	0	0	0	0
MDA-MB-231	SRX040505	SRP005601	0	0	0	0	0	0	0
MDA-MB-468	SRX040506	SRP005601	0	0	0	0	0	0	26,112
T-47D	SRX040507	SRP005601	0	0	0	0	0	0	0
ZR751	SRX040508	SRP005601	0	0	0	0	0	0	0
CAMA-1	SRX176115	SRP006575	0	0	0	0	0	0	0
HCC1419	SRX176116	SRP006575	0	0	0	0	0	0	0
HCC1500	SRX176117	SRP006575	0	0	0	0	0	0	0
UACC-812	SRX176119	SRP006575	0	0	0	0	0	0	0
ZR-75-30	SRX176120	SRP006575	0	0	0	0	0	0	0
HCC3153	SRX066556	SRP006908	0	0	0	0	0	0	0
MBC647	SRX066578	SRP006908	0	0	0	0	0	0	0
SUM1315	SRX066560	SRP006908	0	0	0	0	0	0	0
SUM149	SRX066554	SRP006908	0	0	0	0	0	0	0

SK-N-SH	SRX084673	SRP007461	0	0	0	0	0	0	0
HCC2337	SRX101336	SRP008746	4399,993	0	0	0	0	0	0
HCC3153	SRX101334	SRP008746	0	0	0	0	0	0	0
MCF 10A	SRX099963	SRP008746	0	0	0	0	0	0	24,364
SUM1315O2	SRX101335	SRP008746	0	0	0	0	0	0	0
JY	SRX105330	SRP009262	7095,322	0	0	0	0	0	0
BL-2	SRX105541	SRP009316	1884,221	0	0	0	0	2939,026	0
BL-30	SRX105538	SRP009316	0	0	0	0	0	0	0
BL-41	SRX105531	SRP009316	0	4034,113	0	0	0	0	0
BL-58	SRX105542	SRP009316	0	0	0	0	0	1121,223	0
BL-70	SRX105537	SRP009316	0	0	0	0	0	0	0
CA-46	SRX105539	SRP009316	0	0	0	0	0	0	0
DAUDI	SRX105540	SRP009316	3903,062	0	0	0	0	0	0
EW	SRX105536	SRP009316	0	0	0	0	0	0	0
GUMBUS	SRX105530	SRP009316	0	0	0	1,467	0	0	0
NAMALWA	SRX105535	SRP009316	2234,626	0	0	0	0	0	0
RAJI	SRX105532	SRP009316	1704,523	0	0	0	0	819,598	0
RAMOS	SRX105534	SRP009316	0	0	0	0	0	0	0
THOMAS	SRX105533	SRP009316	0	0	0	0	0	2604,053	0
U-87 GM	SRX110671	SRP009659	0	0	0	0	0	0	0
U-87 GM	SRX110672	SRP009659	0	0	0	0	0	0	0
K-562	SRX113647	SRP010061	0	0	0	0	0	0	0
MDA-MB-231	SRX147674	SRP013022	0	0	0	0	0	0	0
AGS	SRX181260	SRP014574	0	0	27865,249	0	0	0	0
AGS	SRX181261	SRP014574	0	0	28069,729	0	0	0	0
KATOIII	SRX181269	SRP014574	0	0	1,575	0	0	0	0
MKN1	SRX181267	SRP014574	0	0	3351,033	0	0	0	0
MKN28	SRX181268	SRP014574	0	0	1,202	0	0	0	0
MKN45	SRX181251	SRP014574	0	0	0	2,175	0	0	0
MKN74	SRX181252	SRP014574	0	0	0	2,25	0	0	0
NCI-N87	SRX181255	SRP014574	0	0	2107,089	0	0	0	0
SNU-1	SRX181250	SRP014574	0	0	0	1,485	0	0	0
SNU-16	SRX181248	SRP014574	0	0	0	0	0	0	0
SNU-216	SRX181246	SRP014574	0	0	0	0	0	0	0
SNU-484	SRX181245	SRP014574	0	0	0	0	0	0	0
SNU-5	SRX181249	SRP014574	0	0	0	0	0	0	0
SNU-520	SRX181243	SRP014574	0	0	0	14,701	0	0	45,496
SNU-601	SRX181244	SRP014574	0	0	0	1,2	0	0	0
SNU-620	SRX181247	SRP014574	0	0	0	2,104	0	0	0
SNU-638	SRX181256	SRP014574	0	0	0	1,79	0	0	0
SNU-668	SRX181257	SRP014574	0	0	4,967	0	0	0	0
SNU-719	SRX181259	SRP014574	345,075	0	7,95	0	0	0	0
A-431	SRX209056	SRP017465	0	0	0	0	0	0	0
A-431	SRX209057	SRP017465	0	0	0	0	0	0	0
Caco-2	SRX209063	SRP017465	0	0	0	0	0	0	0

Caco-2	SRX209064	SRP017465	0	0	0	0	0	0	0
HEK-293	SRX209065	SRP017465	0	0	0	8369,452	0	0	0
HEK-293	SRX209066	SRP017465	0	0	0	7834,867	0	0	0
HeLa	SRX209067	SRP017465	0	0	0	0	2357,969	0	0
HeLa	SRX209068	SRP017465	0	0	0	0	3200,028	0	0
Hep G2	SRX209069	SRP017465	0	0	0	0	0	0	0
Hep G2	SRX209070	SRP017465	0	0	0	0	0	0	0
PC-3	SRX209073	SRP017465	0	0	0	0	0	0	0
PC-3	SRX209074	SRP017465	0	0	0	0	0	0	0
RT-4	SRX209075	SRP017465	0	0	0	0	0	0	0
RT-4	SRX209076	SRP017465	0	0	0	0	0	0	0
U-2 OS	SRX209060	SRP017465	0	0	0	0	0	0	0
U-251 MG	SRX209058	SRP017465	0	0	0	0	0	0	0
U-251 MG	SRX209059	SRP017465	0	0	0	0	0	0	0
184A1	SRX317694	SRP026537	0	0	0	0	0	0	0
184B5	SRX317695	SRP026537	0	0	0	0	0	0	31,596
21MT1	SRX317696	SRP026537	0	0	0	0	0	0	0
21MT2	SRX317697	SRP026537	0	0	0	0	0	0	0
21NT	SRX317698	SRP026537	0	0	0	0	0	0	0
21PT	SRX317699	SRP026537	0	0	0	0	0	0	0
600MPE	SRX317700	SRP026537	0	0	0	0	0	0	0
BT-474	SRX317702	SRP026537	0	0	0	0	0	0	0
BT-483	SRX317703	SRP026537	0	0	0	0	0	0	0
BT-549	SRX317704	SRP026537	0	0	0	0	0	0	0
CAMA-1	SRX317705	SRP026537	0	0	0	0	0	0	0
EFM192A	SRX317706	SRP026537	0	0	0	0	0	0	0
EFM192B	SRX317707	SRP026537	0	0	0	0	0	0	0
EFM192C	SRX317708	SRP026537	0	0	0	0	0	0	0
HCC1143	SRX317709	SRP026537	0	0	0	0	0	0	0
HCC1395	SRX317710	SRP026537	0	0	0	0	0	0	0
HCC1419	SRX317711	SRP026537	0	0	0	0	0	0	0
HCC1428	SRX317712	SRP026537	0	0	0	0	0	0	0
HCC1569	SRX317713	SRP026537	0	0	0	0	0	0	0
HCC1599	SRX317714	SRP026537	0	0	0	0	0	0	0
HCC1806	SRX317715	SRP026537	0	0	0	0	0	0	0
HCC1937	SRX317716	SRP026537	0	0	0	0	0	0	0
HCC1954	SRX317717	SRP026537	0	0	0	0	0	0	0
HCC202	SRX317718	SRP026537	0	0	0	0	0	0	0
HCC2218	SRX317719	SRP026537	0	0	0	0	0	0	0
HCC3153	SRX317720	SRP026537	0	0	0	0	0	0	0
HCC38	SRX317721	SRP026537	0	0	0	0	0	0	0
HCC70	SRX317722	SRP026537	0	0	0	0	0	0	20,001
Hs 578T	SRX317723	SRP026537	0	0	0	0	0	0	0
JIMT1	SRX317724	SRP026537	0	0	0	0	0	0	0
LY2	SRX317725	SRP026537	0	0	0	0	0	0	0

MB157	SRX317726	SRP026537	0	0	0	0	0	0	0
MCF 10A	SRX317727	SRP026537	0	0	0	0	0	0	0
MCF 10F	SRX317728	SRP026537	0	0	0	0	0	0	0
MCF-12A	SRX317729	SRP026537	0	0	0	0	0	0	0
MCF7	SRX317730	SRP026537	0	0	0	0	0	0	0
MDA-MB-134-VI	SRX317731	SRP026537	0	0	0	0	0	0	0
MDA-MB-175	SRX317732	SRP026537	0	0	0	0	0	0	0
MDA-MB-231	SRX317733	SRP026537	0	0	0	0	0	0	0
MDA-MB-361	SRX317734	SRP026537	0	0	0	0	0	0	0
MX1	SRX317735	SRP026537	0	0	0	0	0	0	0
SK-BR-3	SRX317736	SRP026537	0	0	0	0	0	0	0
SUM1315	SRX317737	SRP026537	0	0	0	0	0	0	0
SUM149PT	SRX317738	SRP026537	0	0	0	0	0	0	0
SUM159PT	SRX317739	SRP026537	0	0	0	0	0	0	0
SUM225CWN	SRX317740	SRP026537	0	0	0	0	0	0	0
SUM229PE	SRX317741	SRP026537	0	0	0	0	0	0	0
SUM52PE	SRX317742	SRP026537	0	0	0	0	0	0	0
T-47D	SRX317743	SRP026537	0	0	0	0	0	0	0
T47D-KBluc	SRX317744	SRP026537	0	0	0	0	0	0	0
UACC-812	SRX347745	SRP026537	0	0	0	0	0	0	0
UACC-893	SRX317746	SRP026537	0	0	0	0	0	0	0
ZR-75-1	SRX317747	SRP026537	0	0	0	0	0	0	0
ZR-75-30	SRX317748	SRP026537	0	0	0	0	0	0	0
ZR-75-B	SRX317749	SRP026537	0	0	0	0	0	0	0
SK-BR-3	SRX329206	SRP028176	0	0	0	0	0	0	0

Table A-5: Non-human mammalian viruses in 186 cell lines.

			Mason-Pfizer monkey virus	Simian virus 40	Squirrel monkey retrovirus	Bovine polyomavirus	Abelson murine leukemia virus	PreXMRV-1 provirus	Murine type C retrovirus	Rous sarcoma virus
cell_line	exp_acc	proj_acc	NC_001550.1	NC_001669.1	NC_001514.1	NC_001442.1	NC_001499.1	NC_007815.2	NC_001702.1	NC_001407.1
PEO1	ERX013520, ERX013534, ERX013531	ERP000710	0	0	0	0	0	0	0	0
PEO14	ERX013521, ERX013535, ERX013527	ERP000710	0	0	0	0	0	0	0	0
PEO23	ERX013519, ERX013530, ERX013528	ERP000710	0	0	0	0	0	0	0	0

PEO4	ERX013524, ERX013525, ERX013533	ERP000710	0	0	0	0	0	0	0	0
C4-2B	ERX333833	ERP004209	0	0	0	0	0	0	0	0
LNCaP	ERX333831	ERP004209	0	0	0	0	0	0	0	0
Mel501	SRX006132	SRP000931	0	0	0	0	0	0	0	0
MeWo	SRX006122	SRP000931	0	0	0	0	0	0	0	0
MeWo	SRX006129	SRP000931	0	0	0	0	0	0	0	0
MeWo	SRX006130	SRP000931	0	0	0	0	0	0	0	0
DB	SRX079566	SRP001599	0	0	0	0	0	0	0	0
DOHH-2	SRX079565	SRP001599	0	0	0	0	0	0	0	0
Karpas 422	SRX079567	SRP001599	0	0	0	0	0	0	0	0
NU-DHL-1	SRX079568	SRP001599	0	0	0	0	0	0	0	0
NU-DUL-1	SRX079574	SRP001599	0	0	0	0	0	0	0	0
OCI-LY1	SRX079571	SRP001599	0	0	0	0	0	0	0	0
OCI-LY-19	SRX079573	SRP001599	0	0	0	0	3118,993	0	0	0
OCI-LY7	SRX079572	SRP001599	0	0	0	0	0	0	0	0
SU-DHL-6	SRX079569	SRP001599	0	0	0	0	0	0	0	0
WSU-DLCL2	SRX079570	SRP001599	0	0	0	0	0	0	0	0
BT-474	SRX025828, SRX025829	SRP003186	0	0	0	0	0	0	0	0
KPL4	SRX025832	SRP003186	0	0	0	0	0	0	0	0
MCF7	SRX025827	SRP003186	0	0	0	0	0	0	0	0
SK-BR-3	SRX025830, SRX025831	SRP003186	0	0	0	16,313	0	0	0	0
MIP101	SRX026157	SRP003404	0	0	0	0	0	0	3325,345	0
MIP101	SRX026158	SRP003404	0	0	0	0	0	0	2426,722	0
MIP5FU	SRX026159	SRP003404	0	0	0	0	0	0	1046,658	0
MIP5FU	SRX026160	SRP003404	0	0	0	0	0	0	995,748	0
HeLa S3	SRX026691	SRP003497	0	0	0	0	0	0	0	0
Hep G2	SRX026684	SRP003497	0	0	0	0	0	0	0	0
HUVEC	SRX026678	SRP003497	0	0	0	0	0	0	0	0
HUVEC	SRX026687	SRP003497	0	0	0	0	0	0	0	0
NHEK	SRX026673	SRP003497	0	0	0	0	0	0	0	0
22Rv1	SRX031908	SRP004637	0	0	0	0	0	52887,322	0	0
C4-2B	SRX031909	SRP004637	0	0	0	0	0	0	0	0
CA-HPV-10	SRX031910	SRP004637	0	0	0	0	0	0	8,202	0
CWR22	SRX031911	SRP004637	0	0	0	0	0	0	13256,66	0
DU 145	SRX031915	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031916	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031917	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031918	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031919	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031920	SRP004637	0	0	0	0	0	0	0	0
DU 145	SRX031921	SRP004637	0	0	0	0	0	0	0	0
LAPC-4	SRX031922	SRP004637	0	231,303	0	0	0	0	16306,969	0
MDA PCa 2b	SRX031934	SRP004637	0	0	0	0	0	0	62475,513	0

NCI-H660	SRX031935	SRP004637	0	0	0	0	0	0	16,717	0
PC-3	SRX031936	SRP004637	0	0	0	0	0	0	0	0
PrEC	SRX031941	SRP004637	0	0	0	0	0	0	0	0
PrEC	SRX031942	SRP004637	0	0	0	0	0	0	0	0
PrSMC	SRX031943	SRP004637	0	0	0	0	0	0	0	0
PWR-1E	SRX031938	SRP004637	0	257,774	0	0	0	0	0	0
WPE1-NB26	SRX031963	SRP004637	0	0	0	0	0	0	7570,002	0
WPMY-1	SRX031964	SRP004637	0	0	0	0	0	0	0	0
BT-20	SRX040501	SRP005601	0	0	0	0	0	0	0	0
BT-474	SRX040502	SRP005601	0	0	0	0	0	0	0	0
MCF 10A	SRX040503	SRP005601	0	0	0	0	0	0	0	0
MCF7	SRX040504	SRP005601	0	0	0	0	0	0	0	0
MDA-MB-231	SRX040505	SRP005601	0	0	0	0	0	0	0	0
MDA-MB-468	SRX040506	SRP005601	0	0	0	0	0	0	0	0
T-47D	SRX040507	SRP005601	0	0	0	0	0	0	0	0
ZR751	SRX040508	SRP005601	0	0	0	0	0	0	0	0
CAMA-1	SRX176115	SRP006575	0	0	0	0	0	0	0	0
HCC1419	SRX176116	SRP006575	0	0	0	0	0	0	0	0
HCC1500	SRX176117	SRP006575	0	0	0	0	0	0	0	0
UACC-812	SRX176119	SRP006575	0	0	0	0	0	0	0	0
ZR-75-30	SRX176120	SRP006575	0	0	0	0	0	0	0	0
HCC3153	SRX066556	SRP006908	0	0	0	0	0	0	0	0
MBC647	SRX066578	SRP006908	0	0	0	0	0	0	0	0
SUM1315	SRX066560	SRP006908	0	0	0	0	0	0	1483,746	0
SUM149	SRX066554	SRP006908	5,835	0	0	0	0	0	0	0
SK-N-SH	SRX084673	SRP007461	0	0	0	0	0	0	0	0
HCC2337	SRX101336	SRP008746	0	0	0	0	0	0	0	0
HCC3153	SRX101334	SRP008746	0	0	0	0	0	0	0	0
MCF 10A	SRX099963	SRP008746	0	0	0	0	0	0	0	0
SUM1315O2	SRX101335	SRP008746	0	0	0	0	0	0	1483,746	0
JY	SRX105330	SRP009262	0	0	0	0	0	0	22304,951	0
BL-2	SRX105541	SRP009316	0	0	0	0	8929,808	0	0	0
BL-30	SRX105538	SRP009316	0	0	0	0	720,735	0	0	0
BL-41	SRX105531	SRP009316	0	0	0	0	634,33	0	0	0
BL-58	SRX105542	SRP009316	0	0	0	0	3599,416	0	0	0
BL-70	SRX105537	SRP009316	0	0	0	0	132,43	0	0	0
CA-46	SRX105539	SRP009316	0	0	23712,275	0	205,656	0	0	0
DAUDI	SRX105540	SRP009316	0	0	0	0	1350,746	0	0	0
EW	SRX105536	SRP009316	0	0	0	0	0	0	2357,27	0
GUMBUS	SRX105530	SRP009316	0	0	0	0	1494,437	0	0	0
NAMALWA	SRX105535	SRP009316	0	0	0	0	0	0	30703,756	0
RAJI	SRX105532	SRP009316	0	0	0	0	1755,865	0	0	0
RAMOS	SRX105534	SRP009316	0	0	0	0	720,408	0	0	0
THOMAS	SRX105533	SRP009316	0	0	0	0	5726,326	0	0	0

U-87 GM	SRX110671	SRP009659	0	0	0	0	0	0	0	0
U-87 GM	SRX110672	SRP009659	0	0	0	0	0	0	0	0
K-562	SRX113647	SRP010061	0	0	0	0	0	0	0	0
MDA-MB-231	SRX147674	SRP013022	0	54,192	0	0	0	0	0	8,496
AGS	SRX181260	SRP014574	0	0	0	0	0	0	0	0
AGS	SRX181261	SRP014574	0	0	0	0	0	0	0	0
KATOIII	SRX181269	SRP014574	0	0	0	0	0	0	0	0
MKN1	SRX181267	SRP014574	0	0	0	0	0	0	0	0
MKN28	SRX181268	SRP014574	0	0	0	0	0	0	0	0
MKN45	SRX181251	SRP014574	0	0	0	0	0	0	0	0
MKN74	SRX181252	SRP014574	0	0	0	0	0	0	0	0
NCI-N87	SRX181255	SRP014574	0	0	0	0	0	0	0	0
SNU-1	SRX181250	SRP014574	0	0	0	0	0	0	0	0
SNU-16	SRX181248	SRP014574	0	0	0	0	0	0	0	0
SNU-216	SRX181246	SRP014574	0	0	0	0	0	0	0	0
SNU-484	SRX181245	SRP014574	0	0	0	0	0	0	0	0
SNU-5	SRX181249	SRP014574	0	0	0	0	0	0	0	0
SNU-520	SRX181243	SRP014574	0	0	0	0	0	0	0	0
SNU-601	SRX181244	SRP014574	0	0	0	0	0	0	0	0
SNU-620	SRX181247	SRP014574	0	0	0	0	0	0	0	0
SNU-638	SRX181256	SRP014574	0	0	0	0	0	0	0	0
SNU-668	SRX181257	SRP014574	0	0	0	0	0	0	0	0
SNU-719	SRX181259	SRP014574	0	0	0	0	0	0	0	0
A-431	SRX209056	SRP017465	0	0	0	0	0	0	0	0
A-431	SRX209057	SRP017465	0	0	0	0	0	0	0	0
Caco-2	SRX209063	SRP017465	0	0	0	0	0	0	0	0
Caco-2	SRX209064	SRP017465	0	0	0	0	0	0	0	0
HEK-293	SRX209065	SRP017465	0	0	0	0	0	0	0	0
HEK-293	SRX209066	SRP017465	0	0	0	0	0	0	0	0
HeLa	SRX209067	SRP017465	0	0	0	0	0	0	0	0
HeLa	SRX209068	SRP017465	0	0	0	0	0	0	0	0
Hep G2	SRX209069	SRP017465	0	0	0	0	0	0	0	0
Hep G2	SRX209070	SRP017465	0	0	0	0	0	0	0	0
PC-3	SRX209073	SRP017465	0	0	0	0	0	0	0	0
PC-3	SRX209074	SRP017465	0	0	0	0	0	0	0	0
RT-4	SRX209075	SRP017465	0	0	0	0	0	0	0	0
RT-4	SRX209076	SRP017465	0	0	0	0	0	0	0	0
U-2 OS	SRX209060	SRP017465	0	0	0	0	0	0	0	0
U-251 MG	SRX209058	SRP017465	0	0	0	0	0	0	0	0
U-251 MG	SRX209059	SRP017465	0	0	0	0	0	0	0	0
184A1	SRX317694	SRP026537	0	0	0	0	0	0	9,848	0
184B5	SRX317695	SRP026537	0	0	0	0	0	0	0	0
21MT1	SRX317696	SRP026537	0	0	0	0	0	0	0	0
21MT2	SRX317697	SRP026537	0	0	0	0	0	0	0	0

21NT	SRX317698	SRP026537	0	0	0	0	0	0	0	0
21PT	SRX317699	SRP026537	0	0	0	0	0	0	0	0
600MPE	SRX317700	SRP026537	7,752	0	0	0	0	0	0	0
BT-474	SRX317702	SRP026537	0	0	0	0	0	0	0	0
BT-483	SRX317703	SRP026537	0	0	0	0	0	0	0	0
BT-549	SRX317704	SRP026537	0	0	0	0	0	0	0	0
CAMA-1	SRX317705	SRP026537	0	0	0	0	0	0	0	0
EFM192A	SRX317706	SRP026537	0	0	0	0	0	0	0	0
EFM192B	SRX317707	SRP026537	0	0	0	0	0	0	0	0
EFM192C	SRX317708	SRP026537	0	0	0	0	0	0	0	0
HCC1143	SRX317709	SRP026537	0	0	0	0	0	0	0	0
HCC1395	SRX317710	SRP026537	0	0	0	0	0	0	0	0
HCC1419	SRX317711	SRP026537	0	0	0	0	0	0	0	0
HCC1428	SRX317712	SRP026537	0	0	0	0	0	0	0	0
HCC1569	SRX317713	SRP026537	0	0	0	0	0	0	0	0
HCC1599	SRX317714	SRP026537	0	0	0	0	0	0	0	0
HCC1806	SRX317715	SRP026537	0	0	0	0	0	0	11,785	0
HCC1937	SRX317716	SRP026537	0	0	0	0	0	0	0	0
HCC1954	SRX317717	SRP026537	0	0	0	0	0	0	0	0
HCC202	SRX317718	SRP026537	0	0	0	0	0	0	0	0
HCC2218	SRX317719	SRP026537	0	0	0	0	0	0	0	0
HCC3153	SRX317720	SRP026537	0	0	0	0	0	0	0	0
HCC38	SRX317721	SRP026537	2,618	0	0	0	0	0	0	0
HCC70	SRX317722	SRP026537	0	0	0	0	0	0	0	0
Hs 578T	SRX317723	SRP026537	0	0	0	0	0	0	0	0
JIMT1	SRX317724	SRP026537	0	0	0	0	0	0	0	0
LY2	SRX317725	SRP026537	0	0	0	0	0	0	0	0
MB157	SRX317726	SRP026537	0	0	0	0	0	0	0	0
MCF 10A	SRX317727	SRP026537	0	0	0	0	0	0	0	0
MCF 10F	SRX317728	SRP026537	0	0	0	0	0	0	0	0
MCF-12A	SRX317729	SRP026537	0	0	0	0	0	0	0	0
MCF7	SRX317730	SRP026537	0	0	0	0	0	0	0	0
MDA-MB-134-VI	SRX317731	SRP026537	0	0	0	0	0	0	0	0
MDA-MB-175	SRX317732	SRP026537	0	0	0	0	0	0	0	0
MDA-MB-231	SRX317733	SRP026537	0	0	0	0	0	0	0	0
MDA-MB-361	SRX317734	SRP026537	0	0	0	0	0	0	0	0
MX1	SRX317735	SRP026537	0	0	0	0	0	0	1540,929	0
SK-BR-3	SRX317736	SRP026537	0	0	0	13516,47	0	0	0	0
SUM1315	SRX317737	SRP026537	0	0	0	0	0	0	6259,412	0
SUM149PT	SRX317738	SRP026537	0	0	0	0	0	0	0	0
SUM159PT	SRX317739	SRP026537	0	0	0	0	0	0	0	0
SUM225CWN	SRX317740	SRP026537	0	0	0	0	0	0	0	0
SUM229PE	SRX317741	SRP026537	0	0	0	0	0	0	0	0
SUM52PE	SRX317742	SRP026537	0	0	0	0	0	0	0	0

T-47D	SRX317743	SRP026537	0	0	0	0	0	0	0	0
T47D-KBluc	SRX317744	SRP026537	0	0	0	0	0	0	0	0
UACC-812	SRX347745	SRP026537	0	0	0	0	0	0	0	0
UACC-893	SRX317746	SRP026537	0	0	0	0	0	0	0	0
ZR-75-1	SRX317747	SRP026537	0	0	0	0	0	0	0	0
ZR-75-30	SRX317748	SRP026537	0	0	0	0	0	0	0	0
ZR-75-B	SRX317749	SRP026537	0	0	0	0	0	0	0	0
SK-BR-3	SRX329206	SRP028176	0	0	0	0	0	0	0	0

Table A-6: Bacteriophages in 186 cell lines.

			Shigella phage SfIV	Enterobacteria phage HK630	Enterobacteria phage phiX174	Enterobacteria phage P1	Enterobacteria phage lambda
cell_line	exp_acc	proj_acc	NC_022749.1	NC_019723.1	NC_001422.1	NC_005856.1	NC_001416.1
PEO1	ERX013520, ERX013534, ERX013531	ERP000710	0	0	0	0	0
PEO14	ERX013521, ERX013535, ERX013527	ERP000710	0	0	1,468	0	0
PEO23	ERX013519, ERX013530, ERX013528	ERP000710	0	0	0	0	0
PEO4	ERX013524, ERX013525, ERX013533	ERP000710	0	0	0	0	0
C4-2B	ERX333833	ERP004209	0	0	0	0	0
LNCaP	ERX333831	ERP004209	0	0	0	0	0
Mel501	SRX006132	SRP000931	0	0	0	0	13,058
MeWo	SRX006122	SRP000931	0	0	0	0	0
MeWo	SRX006129	SRP000931	0	0	0	0	2,948
MeWo	SRX006130	SRP000931	0	0	0	0	3,678
DB	SRX079566	SRP001599	0	0	0	0	0
DOHH-2	SRX079565	SRP001599	0	0	24,963	0	0
Karpas 422	SRX079567	SRP001599	0	0	0	0	0
NU-DHL-1	SRX079568	SRP001599	0	0	0	0	0
NU-DUL-1	SRX079574	SRP001599	0	0	0	0	0
OCI-LY1	SRX079571	SRP001599	0	0	0	0	0
OCI-LY-19	SRX079573	SRP001599	0	0	0	0	0
OCI-LY7	SRX079572	SRP001599	0	0	0	0	0

SU-DHL-6	SRX079569	SRP001599	0	0	0	0	0
WSU-DLCL2	SRX079570	SRP001599	0	0	0	0	0
BT-474	SRX025828, SRX025829	SRP003186	0	0	0	0	0
KPL4	SRX025832	SRP003186	0	0	0	0	0
MCF7	SRX025827	SRP003186	0	0	0	0	0
SK-BR-3	SRX025830, SRX025831	SRP003186	0	0	0	0	0
MIP101	SRX026157	SRP003404	0	0	0	0	5,096
MIP101	SRX026158	SRP003404	0	0	0	0	2,262
MIP5FU	SRX026159	SRP003404	0	0	0	0	4,03
MIP5FU	SRX026160	SRP003404	0	0	0	0	4,139
HeLa S3	SRX026691	SRP003497	0	391,346	0	0	0
Hep G2	SRX026684	SRP003497	0	595,724	0	0	0
HUVEC	SRX026678	SRP003497	0	428,957	0	0	0
HUVEC	SRX026687	SRP003497	0	500,938	0	0	0
NHEK	SRX026673	SRP003497	0	569,601	0	0	0
22Rv1	SRX031908	SRP004637	0	0	0	0	0
C4-2B	SRX031909	SRP004637	0	0	0	0	0
CA-HPV-10	SRX031910	SRP004637	0	0	0	0	0
CWR22	SRX031911	SRP004637	0	0	0	0	0
DU 145	SRX031915	SRP004637	0	0	0	0	0
DU 145	SRX031916	SRP004637	0	0	0	0	0
DU 145	SRX031917	SRP004637	0	0	0	0	0
DU 145	SRX031918	SRP004637	0	0	0	0	0
DU 145	SRX031919	SRP004637	0	0	0	0	0
DU 145	SRX031920	SRP004637	0	0	0	0	0
DU 145	SRX031921	SRP004637	0	0	0	0	0
LAPC-4	SRX031922	SRP004637	0	0	0	0	0
MDA PCa 2b	SRX031934	SRP004637	0	0	0	0	0
NCI-H660	SRX031935	SRP004637	0	0	0	0	0
PC-3	SRX031936	SRP004637	0	0	0	0	0
PrEC	SRX031941	SRP004637	0	0	0	0	0
PrEC	SRX031942	SRP004637	0	0	0	0	0
PrSMC	SRX031943	SRP004637	0	0	0	0	0
PWR-1E	SRX031938	SRP004637	0	0	25390,029	0	0
WPE1-NB26	SRX031963	SRP004637	0	0	8085,77	0	0
WPMY-1	SRX031964	SRP004637	0	0	15342,292	0	0
BT-20	SRX040501	SRP005601	0	0	17631,681	0	0
BT-474	SRX040502	SRP005601	0	0	15628,155	0	0
MCF 10A	SRX040503	SRP005601	0	0	20843,841	0	0
MCF7	SRX040504	SRP005601	0	0	21040,831	0	0
MDA-MB-231	SRX040505	SRP005601	0	0	9232,161	0	0
MDA-MB-468	SRX040506	SRP005601	0	0	16840,959	0	0
T-47D	SRX040507	SRP005601	0	0	18967,077	0	0
ZR751	SRX040508	SRP005601	0	0	19170,621	0	0

CAMA-1	SRX176115	SRP006575	0	0	0	0	0
HCC1419	SRX176116	SRP006575	0	0	0	0	0
HCC1500	SRX176117	SRP006575	0	0	0	0	0
UACC-812	SRX176119	SRP006575	0	0	0	0	0
ZR-75-30	SRX176120	SRP006575	0	0	0	0	0
HCC3153	SRX066556	SRP006908	0	0	19063,673	0	0
MBC647	SRX066578	SRP006908	0	0	1,794	0	0
SUM1315	SRX066560	SRP006908	0	0	23109,656	0	0
SUM149	SRX066554	SRP006908	0	0	0	0	0
SK-N-SH	SRX084673	SRP007461	0	0	15946,629	0	0
HCC2337	SRX101336	SRP008746	0	0	16631,322	2,483	0
HCC3153	SRX101334	SRP008746	0	0	19063,673	0	0
MCF 10A	SRX099963	SRP008746	0	0	14422,641	0	0
SUM1315O2	SRX101335	SRP008746	0	0	23109,656	0	0
JY	SRX105330	SRP009262	0	0	19909,304	0	0
BL-2	SRX105541	SRP009316	0	0	20781,613	8,718	0
BL-30	SRX105538	SRP009316	3,252	0	31894,014	6,426	0
BL-41	SRX105531	SRP009316	0	0	0	5,649	0
BL-58	SRX105542	SRP009316	4,544	0	32149,913	9,892	0
BL-70	SRX105537	SRP009316	0	0	0	4,944	0
CA-46	SRX105539	SRP009316	3,386	0	73935,34	6,035	0
DAUDI	SRX105540	SRP009316	0	0	27340,847	0	0
EW	SRX105536	SRP009316	0	0	0	0	0
GUMBUS	SRX105530	SRP009316	6,235	0	0	11,957	0
NAMALWA	SRX105535	SRP009316	0	0	0	0	0
RAJI	SRX105532	SRP009316	0	0	0	0	0
RAMOS	SRX105534	SRP009316	0	0	0	0	0
THOMAS	SRX105533	SRP009316	0	0	0	0	0
U-87 GM	SRX110671	SRP009659	0	0	0	0	0
U-87 GM	SRX110672	SRP009659	0	0	0	0	0
K-562	SRX113647	SRP010061	0	0	0	0	0
MDA-MB-231	SRX147674	SRP013022	0	0	0	0	0
AGS	SRX181260	SRP014574	0	0	0	0	0
AGS	SRX181261	SRP014574	0	0	0	0	0
KATOIII	SRX181269	SRP014574	0	0	0	0	0
MKN1	SRX181267	SRP014574	0	0	0	0	0
MKN28	SRX181268	SRP014574	0	0	0	0	0
MKN45	SRX181251	SRP014574	0	0	0	0	0
MKN74	SRX181252	SRP014574	0	0	0	0	0
NCI-N87	SRX181255	SRP014574	0	0	0	0	0
SNU-1	SRX181250	SRP014574	0	0	0	0	0
SNU-16	SRX181248	SRP014574	0	0	0	0	0
SNU-216	SRX181246	SRP014574	0	0	0	0	0
SNU-484	SRX181245	SRP014574	0	0	0	0	0
SNU-5	SRX181249	SRP014574	0	0	0	0	0

SNU-520	SRX181243	SRP014574	0	0	0	0	0
SNU-601	SRX181244	SRP014574	0	0	0	0	0
SNU-620	SRX181247	SRP014574	0	0	0	0	0
SNU-638	SRX181256	SRP014574	0	0	0	0	0
SNU-668	SRX181257	SRP014574	0	0	0	0	0
SNU-719	SRX181259	SRP014574	0	0	0	0	0
A-431	SRX209056	SRP017465	0	0	141,343	0	0
A-431	SRX209057	SRP017465	0	0	154,44	0	0
Caco-2	SRX209063	SRP017465	3,562	0	46,94	5,658	0
Caco-2	SRX209064	SRP017465	0	0	42,479	0	0
HEK-293	SRX209065	SRP017465	0	0	52,619	0	0
HEK-293	SRX209066	SRP017465	0	0	60,404	0	0
HeLa	SRX209067	SRP017465	0	0	42,486	0	0
HeLa	SRX209068	SRP017465	0	0	55,211	0	0
Hep G2	SRX209069	SRP017465	0	0	58,181	0	0
Hep G2	SRX209070	SRP017465	0	0	48,254	0	0
PC-3	SRX209073	SRP017465	0	0	75,949	0	0
PC-3	SRX209074	SRP017465	0	0	48,868	0	0
RT-4	SRX209075	SRP017465	0	0	42,563	0	0
RT-4	SRX209076	SRP017465	0	0	42,198	0	0
U-2 OS	SRX209060	SRP017465	0	0	82,362	0	0
U-251 MG	SRX209058	SRP017465	0	0	162,954	0	0
U-251 MG	SRX209059	SRP017465	0	0	111,896	0	0
184A1	SRX317694	SRP026537	0	0	0	0	0
184B5	SRX317695	SRP026537	0	0	0	0	0
21MT1	SRX317696	SRP026537	0	0	0	0	0
21MT2	SRX317697	SRP026537	0	0	0	0	0
21NT	SRX317698	SRP026537	0	0	0	0	0
21PT	SRX317699	SRP026537	0	0	0	0	0
600MPE	SRX317700	SRP026537	0	0	0	0	0
BT-474	SRX317702	SRP026537	0	0	0	0	0
BT-483	SRX317703	SRP026537	0	0	0	0	0
BT-549	SRX317704	SRP026537	0	0	0	0	0
CAMA-1	SRX317705	SRP026537	0	0	0	0	0
EFM192A	SRX317706	SRP026537	0	0	0	0	0
EFM192B	SRX317707	SRP026537	0	0	0	0	0
EFM192C	SRX317708	SRP026537	0	0	0	0	0
HCC1143	SRX317709	SRP026537	0	0	0	0	0
HCC1395	SRX317710	SRP026537	0	0	0	0	0
HCC1419	SRX317711	SRP026537	0	0	0	0	0
HCC1428	SRX317712	SRP026537	0	0	0	0	0
HCC1569	SRX317713	SRP026537	0	0	0	0	0
HCC1599	SRX317714	SRP026537	0	0	0	0	0
HCC1806	SRX317715	SRP026537	0	0	0	0	0
HCC1937	SRX317716	SRP026537	0	0	0	0	0

HCC1954	SRX317717	SRP026537	0	0	0	0	0
HCC202	SRX317718	SRP026537	0	0	0	0	0
HCC2218	SRX317719	SRP026537	0	0	0	0	0
HCC3153	SRX317720	SRP026537	0	0	0	0	0
HCC38	SRX317721	SRP026537	0	0	0	0	0
HCC70	SRX317722	SRP026537	0	0	0	0	0
Hs 578T	SRX317723	SRP026537	0	0	0	0	0
JIMT1	SRX317724	SRP026537	0	0	2194,092	0	0
LY2	SRX317725	SRP026537	0	0	0	0	0
MB157	SRX317726	SRP026537	0	0	0	0	0
MCF 10A	SRX317727	SRP026537	0	0	0	0	0
MCF 10F	SRX317728	SRP026537	0	0	0	0	0
MCF-12A	SRX317729	SRP026537	0	0	0	0	0
MCF7	SRX317730	SRP026537	0	0	0	0	0
MDA-MB-134-VI	SRX317731	SRP026537	0	0	0	0	0
MDA-MB-175	SRX317732	SRP026537	0	0	2158,245	0	0
MDA-MB-231	SRX317733	SRP026537	0	0	0	0	0
MDA-MB-361	SRX317734	SRP026537	0	0	2745,113	0	0
MX1	SRX317735	SRP026537	0	0	0	0	0
SK-BR-3	SRX317736	SRP026537	0	0	0	0	0
SUM1315	SRX317737	SRP026537	0	0	0	0	0
SUM149PT	SRX317738	SRP026537	0	0	0	0	0
SUM159PT	SRX317739	SRP026537	0	0	2376,213	0	0
SUM225CWN	SRX317740	SRP026537	0	0	0	0	0
SUM229PE	SRX317741	SRP026537	0	0	0	0	0
SUM52PE	SRX317742	SRP026537	0	0	0	0	0
T-47D	SRX317743	SRP026537	0	0	0	0	0
T47D-KBluc	SRX317744	SRP026537	0	0	0	0	0
UACC-812	SRX347745	SRP026537	0	0	0	0	0
UACC-893	SRX317746	SRP026537	0	0	0	0	0
ZR-75-1	SRX317747	SRP026537	0	0	0	0	0
ZR-75-30	SRX317748	SRP026537	0	0	0	0	0
ZR-75-B	SRX317749	SRP026537	0	0	0	0	0
SK-BR-3	SRX329206	SRP028176	0	0	3836,708	0	0

D. Sequences

Clustal O multiple sequence alignment of AbMLV contigs

```
CLUSTAL O(1.2.4) multiple sequence alignment
Daudi         ------------------------------------------------------------ 0
BL70          ------------------------------------------------------------ 0
Namalwa       ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCNCCCCNCCCCNNCCCCCCCCCCCCCC 60
BL41          ------------------------------------------------------------ 0
BL30          ------------------------------------------------------------ 0
Thomas        ------------------------------------------------------------ 0
EW            ------------------------------------------------------------ 0
CA46          ------------------------------------------------------------ 0
BL58          ------------------------------------------------------------ 0
BL2           ------------------------------------------------------------ 0
OCI-Ly19      ------------------------------------------------------------ 0
Gumbus        ------------------------------------------------------------ 0
Raji          ------------------------------------------------------------ 0
Ramos         ------------------------------------------------------------ 0


Daudi         ---------------------------------------------------CCCCTCTC 8
BL70          ------------------------------------------------------------ 0
Namalwa       CNCNCCNCNCNCCCNCCCCNCNCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCNC 120
BL41          ------------------------------------------------------------ 0
BL30          ---------------------------------------------------CCCCTCTC 9
Thomas        ----CCCCCCCCCCTCCCCCCCCCNCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCNC 56
EW            -------------------------CCCCCCCCCCCCCCCCCCCCCCCCCCCCCTCCC 34
CA46          ------------------------------------------------------CCTCTN 6
BL58          ---------------------CCCCCCCCCCCCCCCCCCCCCCCCCCCCNCCCCCTCTC 39
BL2           ---------------------------------------------------CCCCCCTCTC 11
OCI-Ly19      ----------------------CCCCCCCTCTCCCCCNCCCCNCNCCNCCCCCCCCTCTC 38
Gumbus        ---------------------------------------------------CCCCCCTCTC 10
Raji          ------------------------------------------------------------ 0
Ramos         ------------------------------------------------------------ 0


Daudi         CCTCCCCCCCCCCTAACNNTNNNNGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 68
BL70          CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 60
Namalwa       CCNCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 180
BL41          ----------------CGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 44
BL30          CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 69
Thomas        CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 116
EW            CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 94
CA46          CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 66
BL58          CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 99
BL2           CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 71
OCI-Ly19      CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 98
Gumbus        CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 70
Raji          CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 60
Ramos         CCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTG 60
                           *    ************************************


Daudi         TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 128
BL70          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 120
Namalwa       TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 240
BL41          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 104
BL30          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 129
Thomas        TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 176
EW            TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 154
CA46          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 126
BL58          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 159
BL2           TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 131
OCI-Ly19      TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 158
Gumbus        TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 130
Raji          TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 120
Ramos         TCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG 120
              ************************************************************


Daudi         GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 188
BL70          GCCCTGTCTTCTTGACGAGCATTCCTAGNGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 180
Namalwa       GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 300
BL41          GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 164
```

```
BL30          GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 189
Thomas        GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 236
EW            GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 214
CA46          GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 186
BL58          GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 219
BL2           GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 191
OCI-Ly19      GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 218
Gumbus        GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 190
Raji          GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 180
Ramos         GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAG 180
              ************************** ******************************

Daudi         GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 248
BL70          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 240
Namalwa       GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTAAAGACAAACAACGT 360
BL41          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 224
BL30          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 249
Thomas        GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 296
EW            GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 274
CA46          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 246
BL58          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 279
BL2           GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 251
OCI-Ly19      GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 278
Gumbus        GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 250
Raji          GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 240
Ramos         GTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT 240
              ****************************************** **************

Daudi         CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 308
BL70          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 300
Namalwa       CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 420
BL41          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 284
BL30          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 309
Thomas        CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 356
EW            CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 334
CA46          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 306
BL58          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 339
BL2           CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 311
OCI-Ly19      CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 338
Gumbus        CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 310
Raji          CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 300
Ramos         CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCC 300
              ************************************************************

Daudi         AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAATGCCACGTTGTGA 368
BL70          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTNA 360
Namalwa       AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 480
BL41          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 344
BL30          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 369
Thomas        AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 416
EW            AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 394
CA46          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 366
BL58          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 399
BL2           AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 371
OCI-Ly19      AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 398
Gumbus        AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 370
Raji          AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 360
Ramos         AAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTTGTGA 360
              ********************************************** *********** *

Daudi         GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 428
BL70          GTTNGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAANGGGCTGA 420
Namalwa       GTTGGATAGTTGTAGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 540
BL41          GTTGGATAGTTGTNGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 404
BL30          GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 429
Thomas        GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 476
EW            GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 454
CA46          GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 426
BL58          GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 459
BL2           GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 431
OCI-Ly19      GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 458
Gumbus        GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 430
Raji          GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 420
Ramos         GTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGA 420
              *** ********* ************************************ *******
```

```
Daudi         AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 488
BL70          AGGATGCCCAGAAGGTACCCCATTGTATNGGATCTGATCTNGGGCCTCGGTGCACATGCT 480
Namalwa       AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 600
BL41          AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 464
BL30          AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 489
Thomas        AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 536
EW            AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 514
CA46          AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 486
BL58          AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 519
BL2           AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 491
OCI-Ly19      AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 518
Gumbus        AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 490
Raji          AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 480
Ramos         AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCT 480
              ************************** *********** ******************

Daudi         TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 548
BL70          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGNGGACGTNGT 540
Namalwa       TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 660
BL41          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 524
BL30          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 549
Thomas        TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 596
EW            TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 574
CA46          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 546
BL58          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 579
BL2           TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 551
OCI-Ly19      TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 578
Gumbus        TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 550
Raji          TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 540
Ramos         TTACATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGT 540
              ************************************************ ****** **

Daudi         TTTCCTTTGAAAAACACAATGATAAGCTTGCCACAACC---------------------- 586
BL70          TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 578
Namalwa       TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 698
BL41          TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 562
BL30          TTTCCTTTAAAAAACACGATGATAAGCTTGCCACAACCA--------------------- 588
Thomas        TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACCACGG------------------- 638
EW            TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 612
CA46          TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 584
BL58          TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACCA--------------------- 618
BL2           TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACCANGNAANNNNNNTGGGGCGCCC 611
OCI-Ly19      TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACCA--------------------- 617
Gumbus        TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 588
Raji          TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 578
Ramos         TTTCCTTTGAAAAACACGATGATAAGCTTGCCACAACC---------------------- 578
              ******** ******** ********************
```

Daudi	---	586
BL70	---	578
Namalwa	---	698
BL41	---	562
BL30	---	588
Thomas	---	638
EW	---	612
CA46	---	584
BL58	---	618
BL2	TCT	614
OCI-Ly19	---	617
Gumbus	---	588
Raji	---	578
Ramos	---	578

AbMLV contig sequence in cell line BL-41

>NC_001479.1|contig1

CGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTGTCTATATGTTATTTTCCACCATAT
TGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCTGTCTTCTTGACGAGCATTCCTAGGGGT
CTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGC
TTCTTGAAGACAAACAACGTCTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGT
GCCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTGCCACGTT
GTGAGTTGGATAGTTGTNGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGA
TGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTT
AGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGAT
GATAAGCTTGCCACAACC

Declaration of Authorship

I hereby affirm that the submitted doctoral thesis

RNA-Seq Based Decomposition of Human Cell Lines and Primary Tumors for the Identification and Quantification of Viral Expression

is, to the best of my knowledge and belief, in all parts my original work and that I have not received assistance from outside other than acknowledged. I have clearly indicated and referenced all used material and sources. This work has not been submitted, either substantially or in whole, for examination purposes at this or any other University before.

Mainz, ..…….. …….

Thomas Bukur