TRANSCRIPTOMICS IN THE STUDY OF PATHOGENS AND HUMAN MALIGNANCIES

A DISSERTATION

SUBMITTED ON THE SIXTH DAY OF MARCH 2015

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE GRADUATE SCHOOL

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

MICHAEL JAMES STRONG

APPROVED:

Erik K. Flemington, Ph.D.

Barbara S. Beckman

Barbara S. Beckman, Ph.D.

Cindy A Morris

Cindy A. Morris, Ph.D.

Alton O. Sartor, M.D.

Marcus L. Ware, M.D., Ph.D.

ABSTRACT

Next generation sequencing (NGS) is a relatively new technology that has revolutionized the way scientists discover and investigate pathogens. It has been estimated that a staggering one in every five cancers worldwide is linked to an infectious agent. An understanding of the pathogen biology as well as the interactions with the host will lead to better therapies and outcomes for patients suffering from pathogen-associated malignancies. Despite the promise for this phenomenon through NGS-based approaches, we are still in the infancy of sequence analysis and are unable to fully appreciate the potential of NGS.

To facilitate data mining, an automated computational pipeline for the simultaneous analysis of pathogen and host transcripts called RNA CoMPASS was developed. Using RNA CoMPASS to investigate a variety of sequencing datasets over the years, substantial bacterial contamination have been routinely identified in human-derived RNA-seq datasets that likely arose from environmental sources. Based on this analysis, a need for more stringent sequencing and analysis protocols to investigate sequence-based microbial signatures in clinical samples is crucial.

NGS-based approaches were utilized to investigate the role of Epstein-Barr virus (EBV) in the pathogenesis of gastric carcinoma. A comprehensive assessment of the virome of various brain tissue samples was also performed, with the notion that an NGS-based detection method would be unbiased, sensitive, specific, and accurate. Taken together, these studies provide a framework for using NGS technology to study oncogenic pathogens and bring awareness to contamination issues within sequencing datasets.

TRANSCRIPTOMICS IN THE STUDY OF PATHOGENS AND HUMAN MALIGNANCIES

A DISSERTATION

SUBMITTED ON THE SIXTH DAY OF MARCH 2015

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE GRADUATE SCHOOL

OF TULANE UNIVERSITY

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

BY

____

MICHAEL JAMES STRONG

APPROVED:

____
Erik K. Flemington, Ph.D.

______________________________ Barbara S. Beckman, Ph.D.	______________________________ Cindy A. Morris, Ph.D.
______________________________ Alton O. Sartor, M.D.	______________________________ Marcus L. Ware, M.D., Ph.D.

ACKNOWLEDGEMENTS

Over the past four years I have been very fortunate to receive support and encouragement from many individuals. First, I would like to express my gratitude to my advisor, Dr. Erik Flemington for the unwavering support, guidance, and encouragement required for me to complete my dissertation. The wisdom he has bestowed upon me will serve as a solid foundation to which I can build a successful career as a physician-scientist. His patience, mentoring, intense desire to learn, ability to listen, and genius level knowledge truly makes him a great mentor and it was a privilege to have worked in his laboratory.

I am thankful to my committee members, Drs. Barbara Beckman, Cindy Morris, Alton Sartor, and Marcus Ware for their guidance and support. Each member had substantial influence on my development as a scientist.

I would like to thank the many collaborators that have helped me with various projects over the years: Drs. Yao-Zhong Liu, Michelle Lacey, Carl Baribault, Lisa Morici, Dass Vinay, MaryBeth Ferris, Kenneth Swan, Deborah Sullivan, Matthew Burow, Bruce Bunnell, Alan Tucker, Dina Gaupp, Prescott Deininger, and Paula Gregory.

I would also like to give a special thanks to Dr. Christopher Taylor and Gene Blanchard at LSU for helping me with various aspects of bioinformatics analysis. I would also like to thank Dr. Guorong Xu at UCSD for helping with my data analysis and for writing the code for RNA CoMPASS.

In addition, I would like to thank the many members, former and current, of the Flemington laboratory for their help and support throughout the years: Claire Roberts, Xia Wang, Adriane Puetter, Jennifer Cameron, Qinyan Yin, Melody Baddoo, Monica Concha, Christina O'Grady, Hani Nakhoul, Yi Yu, Thomas Laskow, and Zhen Lin. I would like to acknowledge Zhen for his tremendous help and guidance on all of my projects.

Finally, I would like to thank my parents and brother for their unconditional love, support, and encouragement throughout my journey. I would also like to thank my in-laws for their love and support. Last but not least, I would like to thank my wife, Amy Strong, for being patient and her unwavering supporting for me, especially on those long nights in the laboratory. She truly embodies the meaning of partner and I am blessed to have had the privilege to share this journey with her.

I would like to thank the National Institutes of Health and National Cancer Institute for their financial support through the Ruth L. Kirschstein National Research Service Award predoctoral fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AD: Alzheimer’s disease

BART: BamHI A rightward transcripts

BL: Burkitt’s lymphoma

CCLE: Cancer Cell Line Encyclopedia

CD: Cluster of differentiation

cDNA: Complementary DNA

CGCI: Cancer Genome Characterization Initiative

CGHub: Cancer Genomics Hub

CIMP: CpG island methylator phenotype

CISH: Chromogenic in situ hybridization

CMV: Cytomegalovirus

CRC: Colorectal carcinoma

CTL: Cytotoxic T-cell

dATP: Deoxyadenosine triphosphate

ddNTP: Dideoxynucleotides

DLBCL: Diffuse large B-cell lymphoma

DNA: Deoxyribonucleic Acid

DNMT1: DNA methyltransferase 1

EBNA: Epstein–Barr nuclear antigen

EBV: Epstein-Barr Virus

EBVaGC: EBV associated gastric carcinoma

FFPE: Formalin-fixed, paraffin-embedded

GBM: Glioblastoma multiforme

GEO: Gene Expression Omnibus

GUI: Graphical user interface

HAdV-C: Human Adenovirus C

HBV: Hepatitis B

HBRR: Human brain reference RNA

HCMV: Human Cytomegalovirus

HCV: Hepatitis C

HHV: Human Herpesvirus

HiEBVaGC: High EBV associated gastric carcinoma

HIV: Human Immunodeficiency Virus

HPV: Human Papillomavirus

HRP: Horseradish peroxidase

HTLV-1: Human T-cell Lymphoma Virus 1

IDO: Indoleamine 2,3-dioxygenase

IE: Immediate Early

IFNG: Interferon gamma

IGV: Integrative Genomics Viewer

IHC: Immunohistochemistry

IPA: Ingenuity Pathway Analysis

ISH: In situ hybridization

JNK: C-Jun N-terminal kinase

JPPF: Java Parallel Processing Framework

NK: Natural killer cell

KSVH: Kaposi’s Sarcoma Virus

LCL: Lymphoblastoid cell line

LCRC: Louisiana Cancer Research Consortium

LMP1: Latent membrane protein 1

LMP2: Latent membrane protein 2

LncRNA: Long non-coding RNA

LTR: Long terminal repeat

MAQC: MicroArray quality control

MCPyV: Merkel Cell Polyomavirus

MuLV: Murine Leukemia Virus

NF-kB: Nuclear factor kappa-light-chain-enhancer of activated B cells

NGS: Next generation sequencing

NHL: Non-Hodgkin’s lymphoma

NIH-CQV: National Institutes of Health-Chongqing virus

PBS: Phosphate buffered saline

PCR: Polymerase chain reaction

PHV: Parvo-like hybrid virus

PI(3): Phosphatidylinositol-3-OH

RNA: Ribonucleic acid

RPMH: Reads per million human mapped reads

SOP: Standard operating procedure

STAR: Spliced Transcripts Alignment to a Reference

STAT: Signal transducer and activator of transcription

SV40: Simian Virus 40

TCGA: The Cancer Genome Atlas

Trp: Tryptophan

UHRR: Universal human reference RNA

VIGAS: Valganciclovir treatment of glioblastoma patients in Sweden

WGS: Whole genome sequencing

ZMW: Zero-mode waveguide

Chapter 1: Introduction

1.1 Oncogenic Pathogens

It has been estimated that a staggering one in five cancers worldwide (20%) is linked to an infectious agent [1]. There are currently eleven pathogens that are known to be oncogenic and associated with human malignancies, including seven viruses, hepatitis B (HBV) and C (HCV) (linked to hepatocellular carcinoma), human papillomavirus (HPV) (linked to cervical carcinoma), Epstein-Barr virus (EBV) or human herpesvirus 4 (associated with B- and T-cell lymphomas, post-transplant lymphoproliferative disease, leiomyosarcoma, nasopharyngeal carcinoma, and gastric carcinoma), human T-cell lymphoma virus 1 (HTLV-1) (linked to T-cell leukemia), Merkel cell polyomavirus (MCPyV) (linked to Merkel cell carcinoma), Kaposi's sarcoma virus (KSVH) or human herpesvirus 8 (associated with Kaposi's sarcoma). There are also three parasites including *Schistosoma haematobium* (linked to squamous cell carcinoma of the bladder), *Opisthorchis viverrini* (linked to cholangiocarcinoma), and *Opisthorchis viverrini* (linked to cholangiocarcinoma) and one bacterium, *Helicobacter pylori* (associated with gastric carcinoma) [2] (Table 1).

Among the eleven known oncogenic pathogens, five viruses (HPV, HTLV-1, EBV, MCPyV, and KSVH) have direct links with oncogenesis modulating the host gene expression through the expression of viral genes. For instance, HPV encodes oncoproteins E6 and E7, which bind to and inhibit p53 and pRB, respectively [3-5]. Disrupting these vital tumor suppressor proteins leads to the inhibition of apoptotic signaling pathways and the increase in cellular proliferation. EBV encodes an oncoprotein, latent membrane protein 1 (LMP1), which mimics a constitutively active cellular receptor, CD40 [6, 7]. As a result of this ligand-independent activation, multiple signaling pathways are activated including NF-kB (nuclear factor kappa-light-chain-enhancer of activated B cells), JNK (c-Jun N-terminal kinase), and STAT (Signal Transducer and Activator of Transcription) [8-10]. In contrast, the other six pathogens (HBV, HCV, *H. pylori*, *Schistosoma haematobium, Opisthorchis viverrini, Opisthorchis viverrini)* have all been indirectly linked to oncogenesis, whereby a persistent infection results in a chronic inflammatory state that promotes tumorigenesis.

1.2 History of DNA Sequencing

1.2.1 Sanger Sequencing

When Frederick Sanger first described a technique in 1977, in which the exchange of the 3’ hydroxyl group of nucleotides with a hydrogen atom (termed dideoxynucleotides; ddNTP) led to the termination of nucleotide chain elongation, this was the birth of what would be referred to as Sanger sequencing and essentially pioneered the ear of modern genomics. Sanger sequencing is based on a chain-termination approach that incorporates ddNTPs during DNA synthesis and effectively prevents nucleotide elongation by DNA polymerase because the

Table 1. Oncogenic pathogens and their associated malignancies

Oncogenic Pathogen	Associated Malignancy
Viruses	
Hepatitis B	Hepatocellular carcinoma
Hepatitis C	Hepatocellular carcinoma
Human papillomavirus	Cervical carcinoma
Epstein-Barr virus (Human herpesvirus 4)	B- and T-cell lymphomas, post-transplant lymphoproliferative disease, leiomyosarcoma, nasopharyngeal carcinoma, and gastric carcinoma
Human T-cell lymphoma virus 1	T-cell leukemia
Merkel cell polyomavirus	Merkel cell carcinoma
Kaposi's sarcoma virus (Human herpesvirus 8)	Kaposi's sarcoma
Parasites	
Schistosoma haematobium	Squamous cell carcinoma of the bladder
Opisthorchis viverrini	Cholangiocarcinoma
Opisthorchis viverrini	Cholangiocarcinoma
Bacteria	
Helicobacter pylori	Gastric carcinoma

hydrogen atom is fundamentally unable to participate in the elongation reaction with the incoming nucleotide [11]. During this process, a radiolabeled dATP, which is supplied in the sequence reaction, is incorporated and serves as a visual to determine fragment position on a polyacrylamide gel when exposed to X-ray. Despite transforming the field of genomics, Sanger sequencing at this current state was laborious, slow, and not scalable with X-ray visualized gel-separated fragments, requiring manual entry of the corresponding nucleotide sequences (Figure 1) [12].

1.2.2 First Generation Sequencing

In the pursuing years, several advancements to the standard Sanger sequencing were introduced. First, the transition from using radiolabeled dATP to fluorescent-labeled dATP (different fluorochromes for each nucleotide) replaced the need for X-ray, manual visualization, and entry of the sequence reaction [13]. Improvements to sequencing enzymology were also introduced including the use of thermostable DNA polymerases, which were first applied to polymerase chain reaction (PCR) by Mullis and colleagues [14]. Cycled sequencing reactions or amplification, catalyzed by these thermostable DNA polymerases, enabled the use of lower input template DNA. Finally, using capillaries for separation through a process known as electrokinetic injection, instead of the traditional polyacrylamide gel, eliminated several previously lengthy steps including gel imaging, visualization, and manual entry of corresponding nucleotides. These advances were realized with the introduction of the first automated DNA

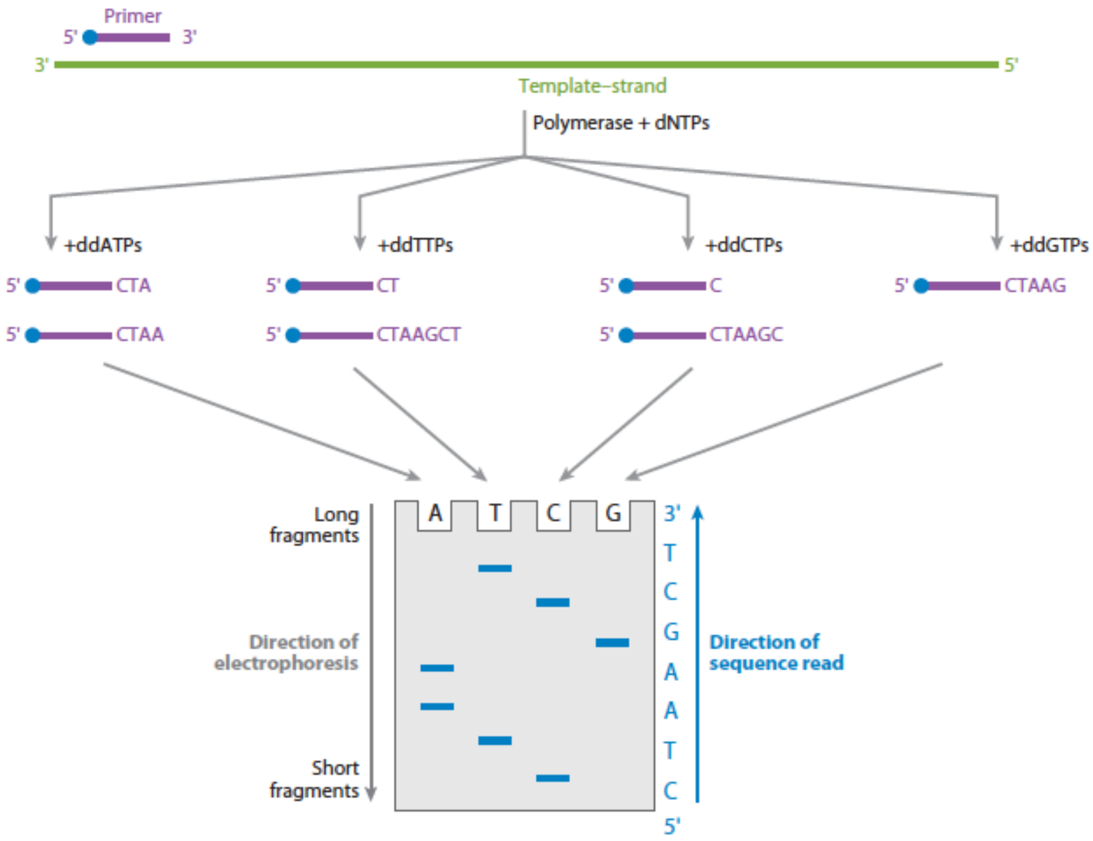


Figure 1. Overview of Sanger sequencing. (Adopted from Mardis, 2013).

sequencer from Applied Biosystems (AB370), which incorporated Sanger sequencing with capillary electrophoresis. For the first time, DNA sequencing was faster and more accurate than previous methods and marked the beginning of the “first generation” of DNA sequencing. Automated Sanger sequencing using capillary sequencing machines were the mainstay for the completion of the human genome project in 2001 [15].

1.3 Second Generation Sequencing

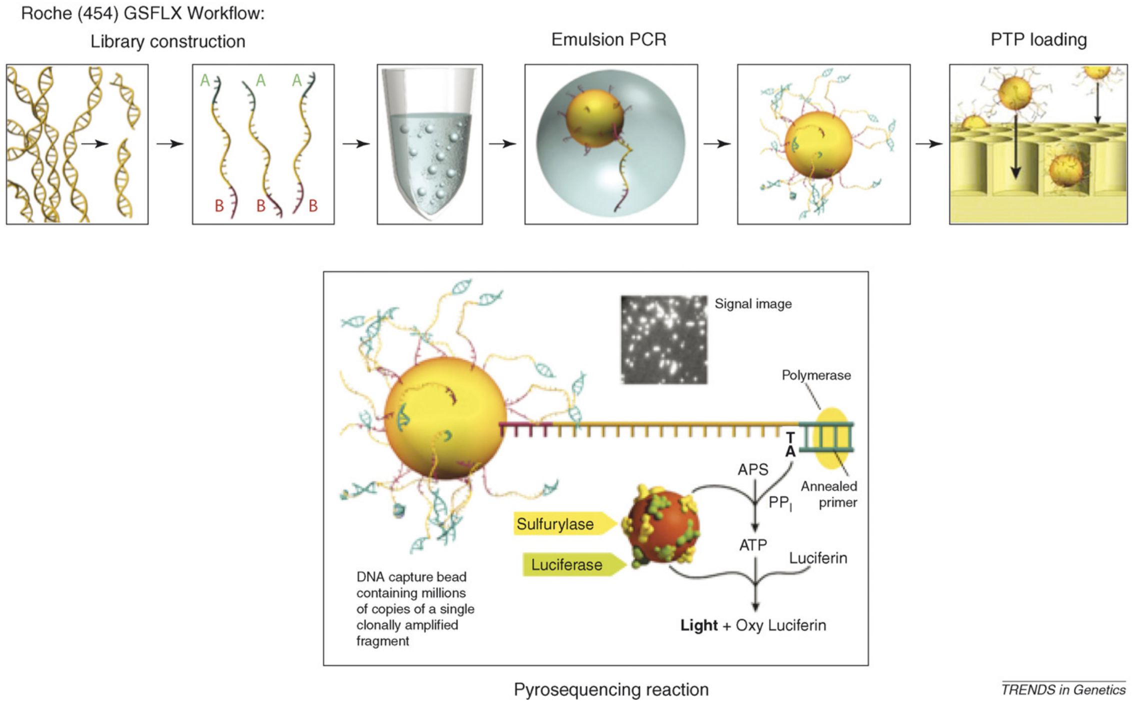
The advent of “second generation” or next generation sequencing (NGS) technology launched with the introduction of three new sequencing systems, all providing cost effective and accurate sequencing at high throughput capacity compared to Sanger sequencing [16]. In stark contrast to Sanger sequencing, NGS technology do not require a cloning step but rather DNA is sequenced through the use of fragment libraries. These libraries are constructed by adding universal adapters to the end of fragmented DNA with subsequent anchorage to a solid surface (e.g. glass slide or bead). A clonal amplification step (e.g. emulsion PCR or solid-phase) is required for all second generation sequencing systems due to the limitation of detecting single fluorescent events. The advantage of anchoring hundreds of millions of fragments on a solid surface is the ability to sequence each of these fragments simultaneously (often referred to as massively parallel sequencing). The great improvement of these NGS systems is their ability to sequence and detect incorporated nucleotides simultaneously without the need for additional steps.

1.3.1 Roche 454

Launched first in 2005, 454, later purchased by Roche, was the first to successfully commercialize a NGS-based sequencing system. The Roche 454 sequencer uses pyrosequencing technology, which detects the pyrophosphate released during nucleotide incorporation [17]. Fragments are anchored to agarose beads and amplified using emulsion PCR relying on an oil-aqueous emulsion to capture bead-DNA complexes into single aqueous droplets where several thousand copies of the same template DNA is produced [18]. (Figure 2)

1.3.2 Illumina

One year later Solexa, later purchased by Illumina, introduced their sequencer. The Illumina sequencer relies on a solid-phase amplification method, which involves clonal amplification of the fragmented DNA on a glass slide, often referred to as bridging amplification [19]. Universal adapters are ligated to randomly fragmented DNA and bound to the surface of a glass slide called a flow cell. (Figure 3)



TRENDS in Genetics

Figure 2. Workflow for Roche 454 sequencing. The top panels illustrate sequence library preparation including emulsion PCR to amplify fragments prior to sequencing. The DNA fragments coupled to beads are loaded into the picotiter plate (PTP). The bottom panel shows the pyrosequencing reaction. (Adopted from Mardis, 2008) [20].

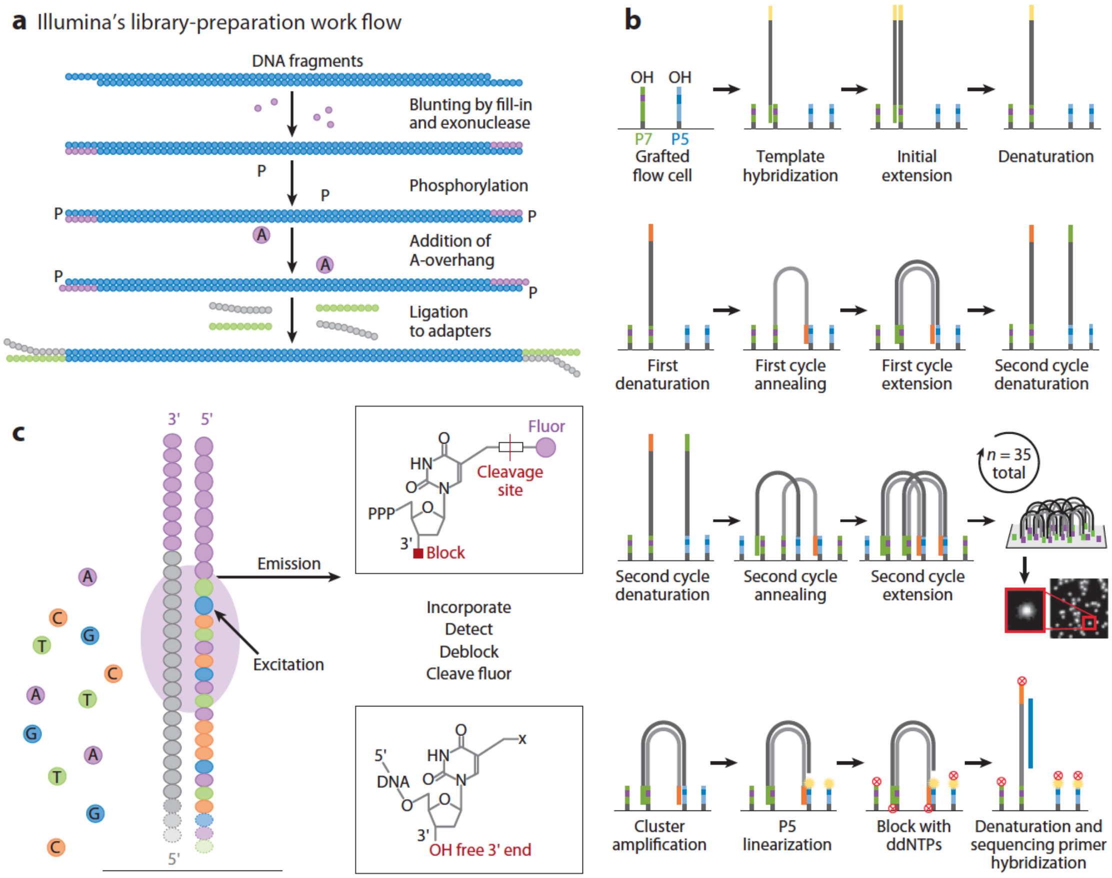


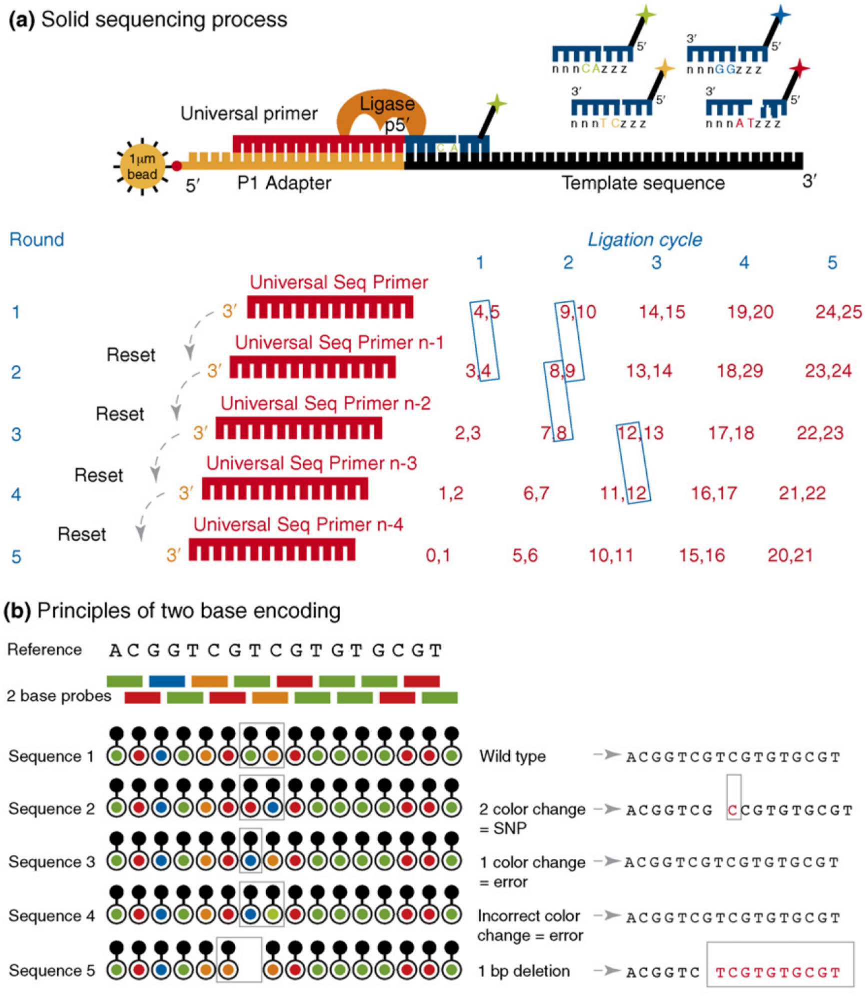
Figure 3. Workflow for Illumina sequencing. (A) Sequence library preparation steps including ligating adapters. (B) Clonal amplification of DNA fragments by bridge amplification. (C) Sequencing of DNA fragments. (Adopted from Mardis, 2013).

1.3.3 SOLiD

The SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing platform introduced by Applied Biosystems uses an emulsion PCR to amplify adapter-ligated fragments bound to magnetic beads. Using a ligase-mediated sequencing approach, whereby each nucleotide is called twice, the SOLiD platform is unique from the other NGS platforms and theoretically reduces base-calling errors with this two base encoding sequence strategy. (Figure 4)

1.3.4 Ion Torrent

In 2010, Ion Torrent, later purchased by Life Technologies, developed and introduced their DNA sequencer, a NGS system completely different from the other sequencers on the market. Using a semiconductor sensor, the Ion Torrent, detects changes in pH from the release of hydrogen ions as a byproduct of nucleotide incorporation to determine the fragment sequence (Figure 5) [21]. Similar to other platforms, the Ion Torrent relies on emulsion PCR for fragment amplification.



TRENDS in Genetics

Figure 4. Overview of SOLiD sequencing. (A) Universal sequence primer is ligated to end of DNA fragment. The fragment then goes through subsequent ligation cycles of the appropriate labeled 8 mer. (B) Schematic of the two base encoding approach. (Adopted from Mardis, 2008).

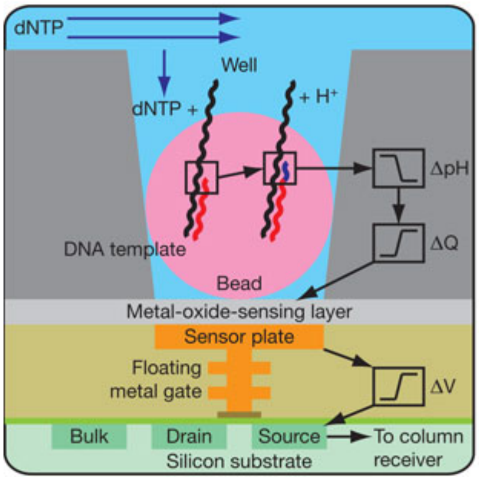


Figure 5. Overview of Ion Torrent sequencing. Sequencing takes place within silicon wells and changes in pH are detected using a semiconductor pH sensor device. (Modified from Rothberg et al., 2011) [21].

1.4 Third Generation Sequencing

The concept of single-molecule sequencing has paved the way for the third generation of DNA sequencing. Pacific Biosystems (SMRT Cell) and Oxford Nanopore Technologies are among the companies utilizing third generation sequencing. The commonality between these third generation sequencing platforms is the elimination of a PCR amplification step, capturing the detection signal in real time and utilizing nanotechnology. Although Pacific Biosystems and Oxford Nanopore Technologies both provide single-molecule sequencing, their approaches are very different. Pacific Biosystems utilizes fluorophores to detect incoming nucleotides during DNA synthesis, while Oxford Nanopore does not rely on DNA polymerase but rather detects alterations in electrical current from nucleotides pass through the nanopore.

1.4.1 Pacific Biosystems

The nanotechnology utilized by Pacific Biosystems is the zero-mode waveguide (ZMW) [22-24]. Within the ZMW, a single polymerase molecule is attached to which a primed template molecule is bound and synthesis and detection takes place in real-time. There are thousands of ZMW spatially distributed on the surface of a silicon wafer called the SMRT Cell. (Figure 6)

1.4.2 Oxford Nanopore Technologies

In contrast, Oxford Nanopore Technologies utilizes an α-hemolysin nanopore [25, 26] inserted into a lipid bilayer that allows only single stranded

DNA or RNA to traverse through. At the same time, a current is passed through the lipid bilayer and alterations in the electrical current are monitored as the nucleotides pass through the nanopore [27, 28]. (Figure 7)

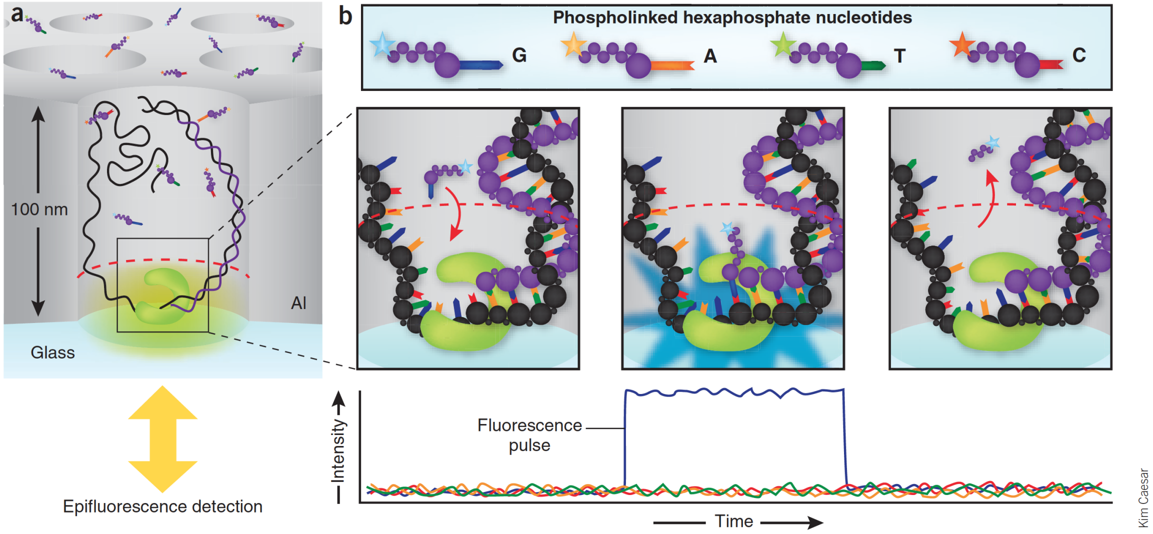

a
100 nm
Al
Glass
Epifluorescence detection
b
Phospholinked hexaphosphate nucleotides
G
A
T
C
Fluorescence pulse
Intensity
Time
Kim Caesar

Figure 6. Overview of Pacific Biosystems sequencing. Real-time single-molecule sequencing is accomplished using zero-mode wavelength (ZMW) nanostructures. (A) One DNA polymerase is anchored at the bottom of the ZMW and adds fluorescently tagged nucleotides to a primed DNA template (depicted in black). (B) As the phospholinked nucleotides enter the region below the red broken line, multiple lasers excite the fluorophores and excitation and emission wavelengths are detected. (Adopted from Metzker, 2009, with permission) [29].

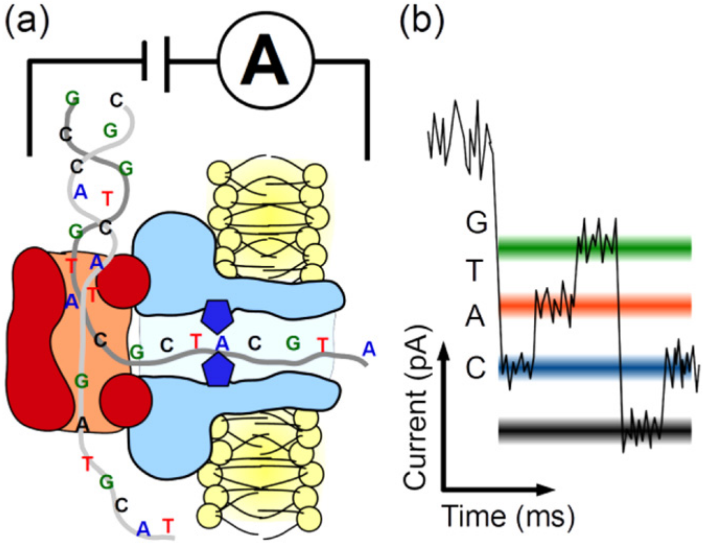


Figure 7. Overview of Oxford Nanopore sequencing. (A) Single-stranded DNA (ssDNA) is fed through a nanopore that possesses a constriction within the channel (dark blue diamonds), which facilitates the reading of the ssDNA. (B) A current is applied across the membrane and as a nucleotide passes through, changes in current are detected corresponding to the various nucleotides. (Adopted from Steinbock and Radenovic, 2015) [30].

1.5 RNA Sequencing

Prior to RNA sequencing (RNA-seq), transcriptomic studies were accomplished using microarray technology [31, 32]. Although transformative at the time, microarray technology has been met with several limitations that have been overcome with RNA-seq technology. Instead of relying on a set of DNA probes hybridizing to cDNA libraries and measuring changes in fluorescence to infer relative abundance of transcripts, RNA-seq massively parallelizes sequencing of fragmented RNA resulting in a readout of sequence stretches (referred to as a read). The move from a probe-dependent to a probe-independent approach essentially transformed gene expression analysis from an analogue methodology of relying on fluorescent changes to a digital methodology that directly counts the sequenced reads. By directly counting the sequenced reads, the dynamic range of detection and quantification is theoretically infinite and greatly enhances the ability to detect very low abundant gene expression. The other advantage of using a probe-independent approach is the potential to discover novel transcripts, alternative splicing events, and gene fusion events. Finally, the ability to retain the direction of transcripts through directional sequencing protocols greatly enhances the ability to interrogate transcriptomes. This capacity is especially important given the recent discoveries of noncoding and antisense transcripts throughout human and pathogen genomes [33-38]. For example, directional sequencing approaches have been applied to study the EBV genome during viral reactivation. The extensive bidirectional transcription extending across nearly the entire EBV genome with the discovery of hundreds of more viral transcripts than was previously known. Most newly identified transcribed regions do not encode proteins but instead likely function as noncoding RNA molecules which could participate in regulating gene expression, gene splicing or even activities such as viral genome processing [38].

1.6 Using RNA Sequencing to Study Pathogen Transcriptome and Host-Pathogen Interaction

The detection/discovery of etiological agents associated with cancers or other diseases is not always an easy task due in part to the broad spectrum of candidate infectious agents that exist. However, through its capacity to delve deeply into the genetic composition of a biological specimen, next generation sequencing (NGS) technology presents an unprecedented approach to pathogen discovery in the context of human disease [39]. This unbiased, sensitive and accurate approach to identify potential causal pathogens has shown promise, resulting in the discovery of a novel Merkel cell polyomavirus in Merkel cell carcinoma [40]. More recently, the discovery of an association between *Fusobacterium* and colorectal carcinoma was made using two different NGS approaches [41, 42]. These discoveries were facilitated by the use of computational subtraction approaches where reads aligning to reference genomes were subtracted from the sequence file, leaving behind sequences from undiscovered organisms.

The genome of an organism contains all the programming information necessary to manufacture the organism and facilitate its life cycle. This genomic information thereby specifies the organism's identity. Sequencing the genomic content of a clinical sample should be a means to accurately identify exogenous agents with potential etiology in the disease. Nevertheless, because the human genome is larger and has a sparser coding density than most microbial and viral genomes, transcriptome analysis of mixed human/microbial/viral communities has become a more sensitive and cost efficient means to detect ectopic organisms than whole genome analysis. This issue is well illustrated in Figure 8. Analysis of genome sequence data from a follicular lymphoma sample in the Cancer Genome Characterization Initiative (CGCI) revealed 1.2×10^{-5} bacterial and 6.7×10^{-6} *Acinetobacter* reads per human mapped reads using genome sequencing data (DNA-seq), whereas 0.018 bacterial and 0.0045 *Acinetobacter* reads per human mapped reads were observed in the corresponding RNA-seq data. This corresponds to a 1,515 fold higher sensitivity of RNA-seq compared to DNA-seq for detecting bacteria in mixed community specimens. RNA level analysis therefore has the potential to be more sensitive than genome sequencing in the identification of ectopic organisms.

An added benefit of RNA-seq analysis of tissue samples is the ability to simultaneously assess both pathogen and host transcriptomes [43]. This dual RNA-seq approach was applied to the study of the fungus *C. albicans* and the interaction with mouse dendritic cells. Although the authors did not full utilize the potential of this approach (e.g. perform an in-depth characterization of the global response to infection), this study represents a successful application of the concept of analyzing the host and pathogen transcriptomes in parallel [44].

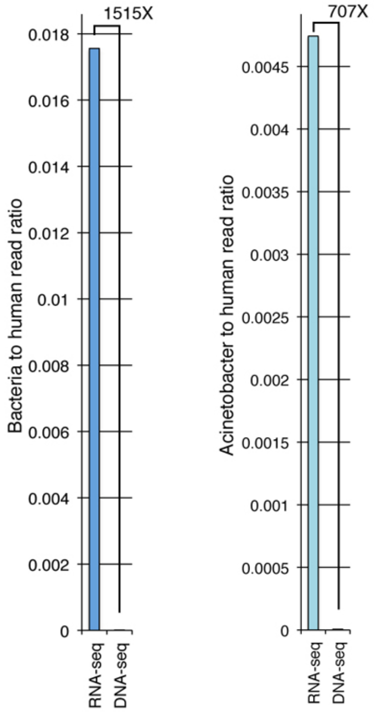


Figure 8. RNA-seq is more sensitive than DNA-seq for the identification of exogenous organisms within human specimens. Bacterial to human and *Acinetobacter* to human read ratios were calculated. The available RNA-seq and DNA-seq data from the single CGCI FL sample were analyzed. RNA-seq analysis shows significantly greater concentrations of both bacterial and *Acinetobacter* reads than DNA-seq analysis.

With rapid advancements in technological approaches in the field of genomics, we now have the capacity to understand and begin to unravel the complexity of pathogens and the intimate interplay with their host organisms. Although the promise of undiscovered possibilities is infinite, our current ability to analyze and fully appreciate the depth to which we now are able to analyze genomic data is still in its infancy. Our ability to efficiently sequence organisms has greatly surpassed our capability of analyzing all of this sequence data. Automated computational pipelines designed to sift through millions of lines of sequence data are being developed in order to assist scientist in analyzing what has been referred to as Big Data. With a seemingly endless supply of data and connects to be discovered, the next formidable task for scientists studying human pathogens is to piece everything together and garner information on the pathogen-host interaction.

PART 1

THE UTILITY OF RNA-SEQ TECHNOLOGY FOR THE DISCOVERY AND INVESTIGATION OF ONCOGENIC PATHOGENS IN THE CONTEXT OF HUMAN MALIGNANCIES

CHAPTER 2: Developing an Automated Computational Pipeline for the Comprehensive Interrogation of RNA-seq Datasets

RNA CoMPASS: A Dual Approach for Pathogen and Host Transcriptome Analysis of RNA-Seq Datasets.

#Guorong Xu, #Michael J. Strong, Michelle R. Lacey, Carl Baribault, Erik K. Flemington and Christopher M. Taylor

#Contributed equally

Published in PLoS One 2014 Feb 25;9(2):e89445. doi: 10.1371/journal.pone.0089445. eCollection 2014

2.1 Abstract

High-throughput RNA sequencing (RNA-seq) has become an instrumental assay for the analysis of multiple aspects of an organism's transcriptome. Further, the analysis of a biological specimen's associated microbiome can also

be performed using RNA-seq data and this application is gaining interest in the scientific community. There are many existing bioinformatics tools designed for analysis and visualization of transcriptome data. Despite the availability of an array of next generation sequencing (NGS) analysis tools, the analysis of RNA-seq data sets poses a challenge for many biomedical researchers who are not familiar with command-line tools. Here we present RNA CoMPASS, a comprehensive RNA-seq analysis pipeline for the simultaneous analysis of transcriptomes and metatranscriptomes from diverse biological specimens. RNA CoMPASS leverages existing tools and parallel computing technology to facilitate the analysis of even very large datasets. RNA CoMPASS has a web-based graphical user interface with intrinsic queuing to control a distributed computational pipeline. RNA CoMPASS was evaluated by analyzing RNA-seq data sets from 45 B-cell samples. Twenty-two of these samples were derived from lymphoblastoid cell lines (LCLs) generated by the infection of naïve B-cells with the Epstein Barr virus (EBV), while another 23 samples were derived from Burkitt’s lymphomas (BL), some of which arose in part through infection with EBV. Appropriately, RNA CoMPASS identified EBV in all LCLs and in a fraction of the BLs. Cluster analysis of the human transcriptome component of the RNA CoMPASS output clearly separated the BLs (which have a germinal center-like phenotype) from the LCLs (which have a blast-like phenotype) with evidence of activated MYC signaling and lower interferon and NF-kB signaling in the BLs. Together, this analysis illustrates the utility of RNA CoMPASS in the simultaneous analysis of transcriptome and metatranscriptome data. RNA CoMPASS is freely available at http://rnacompass.sourceforge.net/.

2.2 Introduction

Through its capacity to delve deeply into the genetic composition of a biological specimen, next generation sequencing (NGS) technology presents an unprecedented approach to pathogen discovery in the context of human disease. This unbiased approach to identify undiscovered human disease causing pathogens has already shown promise, resulting in the discovery of a novel Merkel cell polyomavirus in Merkel cell carcinoma [40], for example. More recently, the discovery of an association between *Fusobacterium* and colorectal carcinoma was made using two different NGS approaches [41, 42]. These discoveries were facilitated by the use of computational subtraction approaches where reads aligning to reference genomes were subtracted from the sequence file leaving behind sequences from undiscovered organisms. Using this general approach, several groups, including ours, have previously reported computational pipelines for the analysis of exogenous sequences and for pathogen discovery [39, 41, 45-48].

While current sequence-based computational subtraction pipelines are used solely for pathogen discovery, RNA CoMPASS, takes advantage of the richness of RNA-seq data to provide host transcript expression data in addition to pathogen analysis. This concept, recently coined “dual RNA-seq’ by Westermann and colleagues [43] allows the user to simultaneously investigate cellular signaling pathways. It also allows the user to investigate associations between differences in cellular signaling pathways and the presence or absence of discovered pathogens. RNA CoMPASS leverages some of the most useful freely available tools and automates distribution of the computational burden over the available computing resources. It is designed to be deployable on either a local cluster or a grid environment managed by Portable Batch System (PBS) submission. RNA CoMPASS provides a web-based graphical user interface, making the program accessible to most biological researchers. Here we present RNA CoMPASS and demonstrate its utility in dual analysis of RNA-seq data sets from different B-cell types with different EBV infection status.

2.3 Materials and Methods

2.3.1 Sequence data acquisition

RNA-seq data sets from 22 Human B-Cell samples (lymphoblastoid cell lines [LCLs]) immortalized with Epstein-Barr Virus (EBV) were downloaded from the NCBI Sequence Read Archive (SRA010302). Samples were sequenced using an Illumina Genome Analyzer II machine running single end 50 base sequencing reactions. Similarly, 22 Human Burkitt's Lymphoma (BL) samples were obtained from the NCBI Sequence Read Archive (SRA048058). Samples were sequenced using an Illumina Genome Analyzer II machine running paired end 107 and 102 base sequencing reactions. The Akata RNA-seq data set was generated previously in our lab (SRA047981) [49]. The Akata sample was sequenced using an Illumina HiSeq instrument running paired end 100 base sequencing reactions.

2.3.2 RNA CoMPASS

RNA CoMPASS (RNA comprehensive multi-processor analysis system for sequencing) is a graphical user interface (GUI) based parallel computation pipeline for the analysis of both exogenous and human sequences from RNA-seq data. Several open source programs and a single commercial program are utilized in this automated pipeline. For the deduplication steps, an in-house de-duplication algorithm is used. Alignments to the reference genome are carried out using Novoalign V2.07.18 (www.novocraft.com) [-o SAM, default options] with a reference genome (e.g. human (hg19; UCSC)), splice junctions (which is generated using the make transcriptome application from Useq [50]; splice junction radius is set to the read length minus 4), and abundant sequences (which include sequence adapters, mitochondrial, ribosomal, enterobacteria phage phiX174, poly-A, and poly-C sequences). Human mapped reads are analyzed using SAMMate [51] to quantify gene expression and to generate genome coverage information. Nonmapped reads are separated following this alignment and subjected to consecutive BLAST V2.2.27 searches against the Human RefSeq RNA database (a final filtering step) and then to the NCBI NT database to identify reads corresponding to known exogenous organisms [52]. Results from the NT BLAST searches are filtered to eliminate matches with an E-value of greater than $10e^{-6}$. The results are fed into MEGAN 4 V4.62.3 [53] for convenient visualization and taxonomic classification of BLAST search results.

2.3.3 Statistics and Cluster analysis

Human transcript counts were imported into the R software environment and analyzed using the edgeR package [54]. Genes with low transcript counts (less than 1 CPM (count per million)) in the majority of samples were filtered out. The Manhattan (L-1) distance matrix for the samples was computed using the remaining transcript counts, and this was taken as input for hierarchical clustering using the Ward algorithm. After assigning each sample to one of two groups identified by hierarchical clustering (Human B-Cell or Burkitt's Lymphoma), the glmFit function was used to fit the mean log(CPM) for each group and likelihood ratio tests were used to identify those genes that were differentially expressed, with adjusted $P<0.05$ following the Benjamini-Hochberg correction for multiple testing. The fitted log(CPM) values for the subset of genes that were differentially expressed in the LCL samples relative to the BL samples were then clustered using the Euclidean distance and complete linkage algorithm to detect groups of co-expressed genes.

2.4 Results

2.4.1 RNA CoMPASS Architecture

RNA CoMPASS facilitates the analysis of small and large RNA sequencing studies through an automated dataflow management and acceleration of processing via distributed computing over a cluster (Figure 1). It has the capability to analyze fastq sequence files generated from single-end, paired-end, and/or directional sequencing strategies. After an initial deduplication step, the first phase of RNA CoMPASS is to perform the alignment of millions of short reads against the host genome using an accurate aligner, Novoalign (http://www.novocraft.com/) [-o SAM, default options]. Any host genome can be uploaded to RNA CoMPASS. In our case, we used the human reference genome, hg19 (UCSC), plus splice junctions (which is generated using the make transcriptome application from Useq [50]; splice junction radius is set to the read length minus 4), and abundant sequences (which include sequence adapters, mitochondrial, ribosomal, enterobacteria phage phiX174, poly-A, and poly-C sequences). After alignment, Novoalign categorizes reads into four classes: uniquely mapped reads, repeat mapped reads, unmapped reads and quality controlled reads. Further processing is bifurcated into the analysis of endogenous sequences (uniquely mapped reads and repeat mapped reads) and the investigation of exogenous reads (unmapped reads) (Figure 1).

Endogenous sequence analysis is performed via the SAMMate transcript analysis software (Figure 1) [51]. A gene annotation file of interest is uploaded to facilitate the calculation of expression abundance scores for annotated genes and transcripts using the uniquely mapped reads (this includes spliced reads) and the best hits of repeat mapped reads from Novoalign. Gene expression is calculated using Reads/Fragments Per Kilobase of exon model per million Mapped reads (RPKM/FPKM) [55]. Isoform quantification is also computed via the RAEM algorithm [56] or the iQuant procedure [57] to estimate the relative isoform proportions and abundance scores. RNA CoMPASS also generates useful files for visualization of RNA-seq data. Read coverage files are produced in Wiggle format for coverage viewing in a genome browser and signal map files are produced with single base pair resolution which can be used with peak detection algorithms [51].

Exogenous sequence analysis proceeds concurrently with the endogenous analysis (Figure 1). Utilizing BLAST [58], unmapped reads are searched against the NCBI NT database for identification using an E-value of better than $10e^{-6}$. This process is extremely computationally intensive and is distributed across the computing cluster to minimize processing time and memory requirements. BLAST run time and memory requirements depend not only on the size of the database being searched but also on the number of input reads. The filtering of reads originating from the human genome prior to searching against the NCBI NT database is the first major step in managing this burden. Despite this step, we have discovered that many host reads remain unmapped and are subsequently identified by BLAST (since BLAST is substantially more permissive). To further reduce the computational burden incurred by BLASTing these unmapped host reads, RNA CoMPASS offers an optional stage prior to NT database BLASTing where the user can BLAST against a host transcript database. Because host transcript databases are much smaller than the NT database, host reads not aligned by Novoalign can be

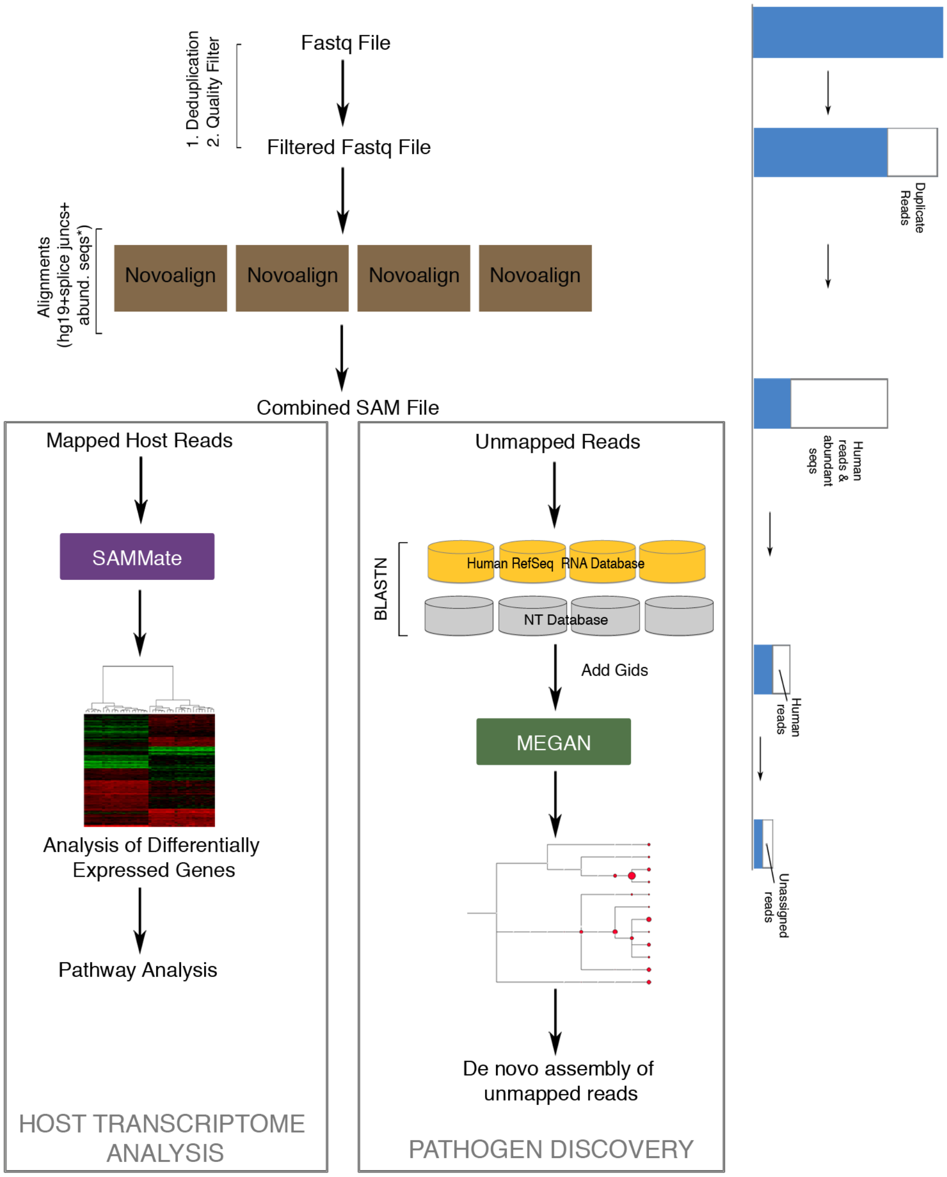


Figure 1. Schematic of RNA CoMPASS (RNA comprehensive multi-processor analysis system for sequencing) architecture. RNA CoMPASS is a graphical user interface (GUI) based parallel computation pipeline for the analysis of both exogenous and human sequences from RNA-seq data. It employs a commercial and several open-source programs to analyze RNA-seq data sets including Novoalign, SAMMate, BLAST, and MEGAN. Each step results in the subtraction of reads in order to further analyze the unmapped reads for pathogen discovery. The mapped reads are analyzed separately. The end result from this pipeline is pathogen discovery and host transcriptome analysis.

filtered out at lower computational cost than would otherwise be incurred by BLASTing these reads against the NT database.

After BLASTing against the NT database, taxonomic analysis is performed by importing the BLAST results into MEGAN [53]. To allow MEGAN to determine the taxon associated with each match, the NCBI taxon id number is appended to each BLAST hit. This is accomplished by looking up the GI accession number in the GI to TaxID file using a custom script. MEGAN then determines the taxon associated with matches based on the hit table using a lowest common ancestor algorithm. MEGAN categorizes the exogenous sequences and outputs an NCBI taxonomy tree. Each node of the output tree is labeled by a taxon and the size of a given node represents the number of reads assigned to that taxon. This provides the researcher with an overview of reads of possible exogenous origin. The researcher can then evaluate the exogenous sequence content in the context of their own biological knowledge of the experiment at hand. The researcher can also formulate hypotheses to test given the taxonomic classification displayed by MEGAN and then export all reads that were assigned to a specific taxon for further analysis. For example, the reads can be assembled into longer transcripts [59] using a de novo parallel sequence assembler. This provides the researcher with a broader view of the particular transcripts that were found within a given taxon. De novo assembly can be repeated for each taxon of interest and the researcher can search the longer assembled transcripts against the databases again to get more precise hits.

In RNA CoMPASS, we have implemented both the Java Parallel Processing Framework (JPPF) API and Portable Batch System (PBS) API in order to deploy it on either a small local cluster or a grid system managed by PBS submission. Our testing of RNA CoMPASS in both environments (data not shown for grid system) showed that our pipeline could efficiently analyze RNA-seq data sets achieving a significant speedup over analysis on a single machine. This will allow other investigators to use RNA CoMPASS on whichever type of computational environment they have access to. In our case, we employed RNA CoMPASS on a local 4-node cluster environment (Intel Xeon Mac Pros with 64-96GB RAM).

2.4.2 RNA CoMPASS performance

To evaluate the performance of RNA CoMPASS on a cluster environment versus a single node environment, we benchmarked 6 RNA-seq data sets with incrementally varied file sizes on both a single machine and on a local cluster with 4 nodes. The 6 files used for this analysis were extracted from a previously generated RNA-seq data set from a BL cell line (1 sequence pair from the Akata RNA-seq data set) [49] and the file sizes varied from 1.4 to 51 million reads. All 6 samples, run on the single node or the 4-node cluster, were processed using identical parameters. As expected, run time increased with file size with the 51 million read file taking approximately 1,400 minutes on a single machine but only 400 minutes on the cluster (Figure 2). Speedup increased with file size (up to 3.4, Figure 2) supporting a benefit of a cluster environment for large-scale projects. Overall speedup was attributed primarily to the parallelization of the two more computationally intensive steps, Novoalign and BLAST (Figure 3).

2.4.3 Pathogen Discovery and Analysis

To test the utility of RNA CoMPASS to identify pathogens within biological specimens we used RNA-seq data sets from two distinct B-cell types, lymphoblastoid cell lines (LCLs) and Burkitt's lymphoma (BL) samples. LCLs are not tumor cells but have instead been immortalized by infection with EBV. In contrast, the Burkitt's lymphoma cells lines are tumor cell lines, some of which underwent tumorigenesis in part through natural infection with EBV. Notably, however, although some Burkitt's lymphoma cell lines are infected with EBV, the EBV gene expression pattern and the cell phenotype of Burkitt's lymphomas and LCLs are distinct.

Single-end RNA-seq data sets from 45 B-cell lines (22-Lymphoblastoid cell lines, 23-Burkitt's lymphomas) were analyzed using RNA CoMPASS. Most samples contained relatively low numbers of non-human viral reads (e.g. enterobacteria phage) that most likely represent environmental contamination (Figure 4A). EBV was the primary mammalian virus detected in these samples (displayed as Human herpesvirus 4 in examples shown in Figure 4A). Nevertheless, related viruses were sometimes displayed in the MEGAN output such as that for sample SRR032270 where 88 reads were classified as Macacine herpesvirus 4 reads and 32 were classified as Papiline herpesvirus 1 (Figure 4A). Further analysis of these reads using manual BLAST revealed that EBV ranked

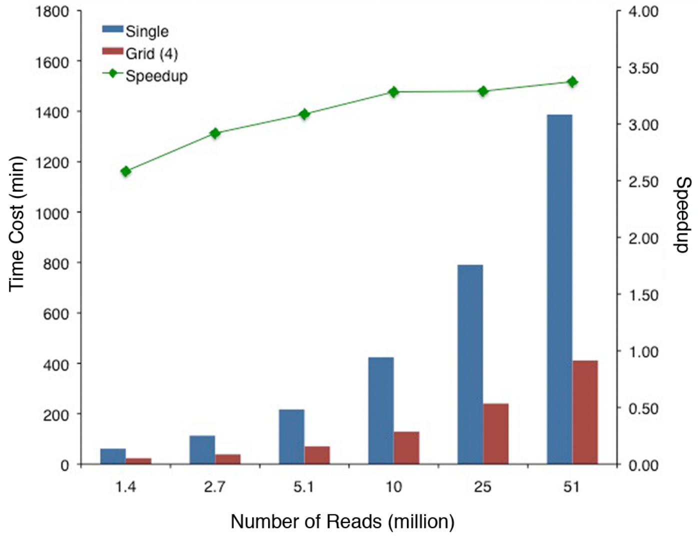


Figure 2. Performance Analysis of RNA CoMPASS. RNA CoMPASS was deployed on a local cluster and benchmarking was performed. An Akata RNA-seq data set was split into six files of varying sizes: 1 – 393.4 MB, 1,397,139 reads, 2 – 757 MB, 2,685,149 reads, 3 – 1.44 GB, 5,120,805 reads, 4 – 2.72 GB, 9,651,466 reads, 5 – 5.01 GB, 25,465,406 reads, sample 6 – 8.99 GB, 50,930,812 reads. Overall time was calculated for each file on a single machine (blue column) and on the local 4-node cluster (red column). Speedup time is represented as a green line.

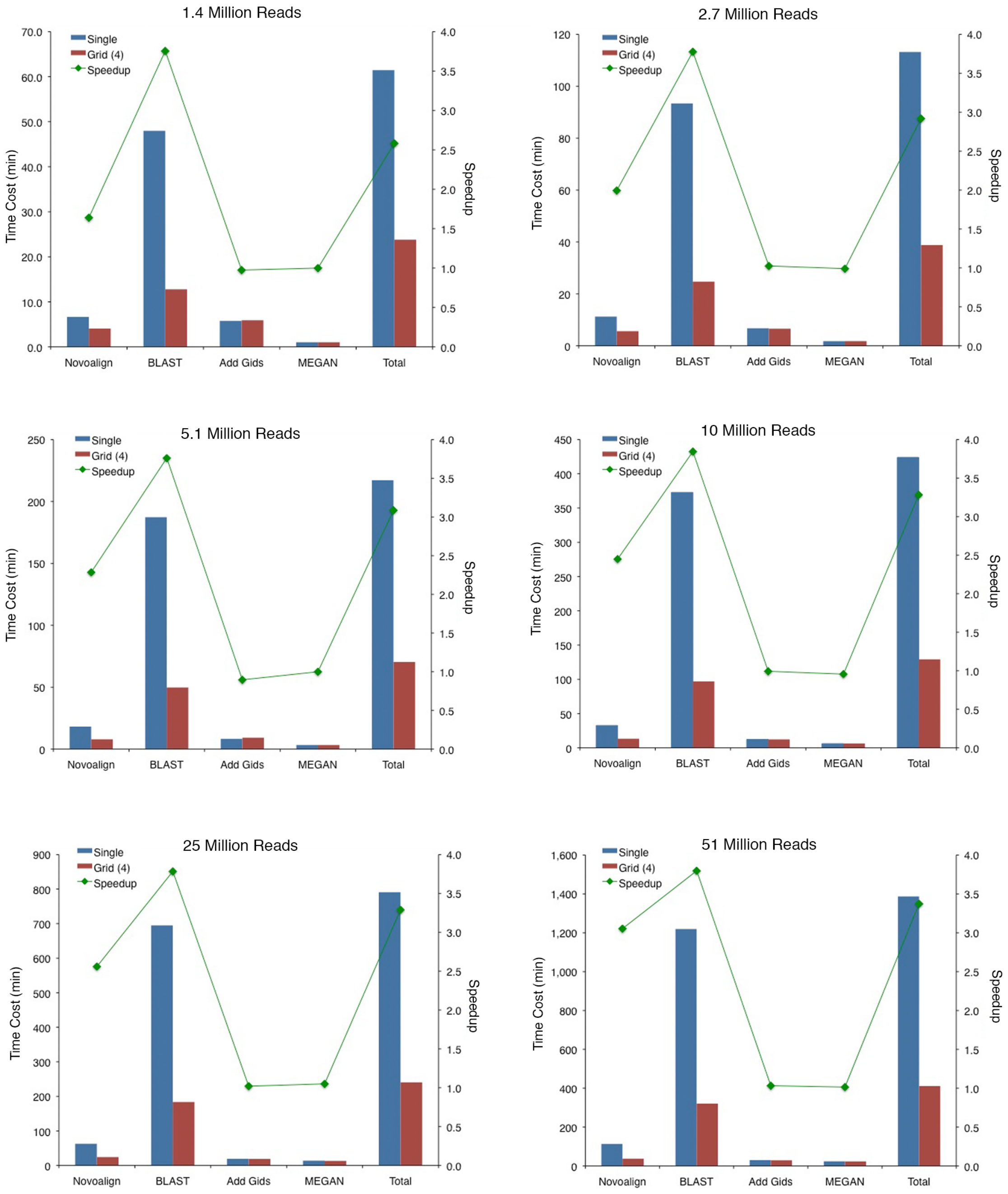


Figure 3. Performance of RNA CoMPASS based on individual tasks. The six Akata RNA-seq data set files used previously were benchmarked on completion of individual tasks and represented in the graphs. Runs on a single node are represented using blue columns while runs on a 4-node cluster are represented using red columns. The green line represents speedup time between the single node and 4-node environment. Note in particular that speedup of the BLAST portion of RNA CoMPASS and overall speedup approaches the theoretical limit of 4 as the data size is increased.

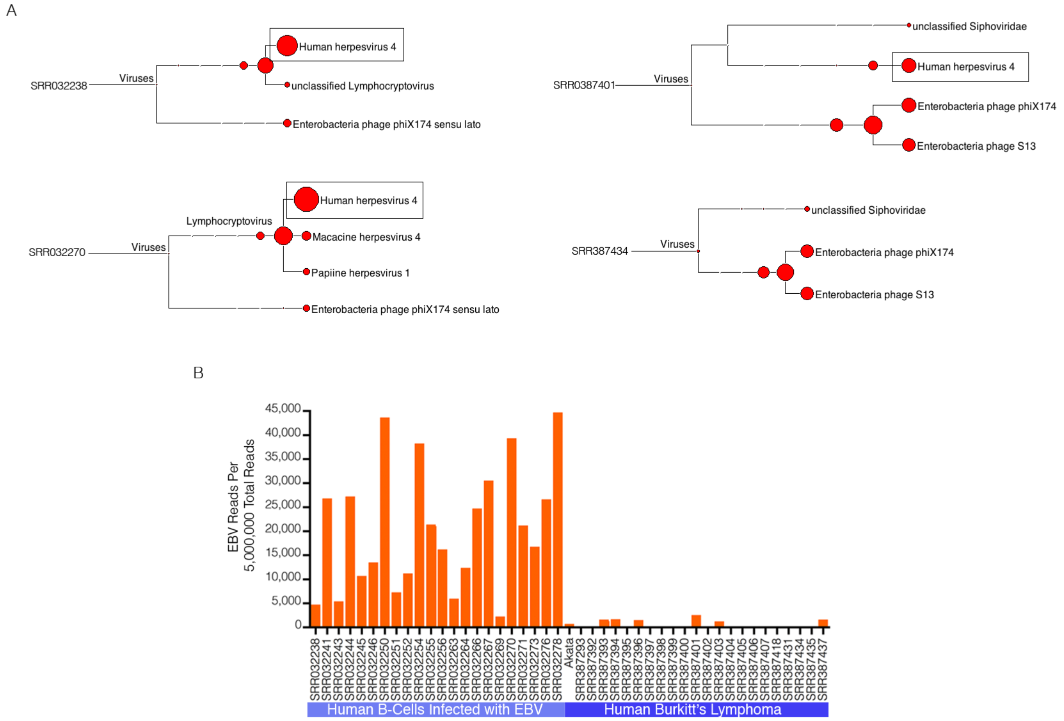


Figure 4. Detection of EBV in Human B-Cells using RNA CoMPASS. Analysis of all 45 single-end RNA-seq data sets (22-Lymphoblastoid cell lines, 23-Burkitt's lymphomas) was performed using RNA CoMPASS. (A) The virome branch of the taxonomy trees for two representative LCLs and Burkitt's lymphomas were generated using the metagenome analysis tool, MEGAN 4. (B) EBV reads were quantified in all 45 RNA-seq data sets and are represented as per 5,000,000 total sequence reads.

among the top 2 hits suggesting that these reads are likely EBV but were misclassified. Most importantly, RNA CoMPASS identified all 22 LCLs and 7 of the 23 Burkitt's lymphoma samples as being positive for EBV (Figure 4B).

As expected, EBV gene expression in LCLs is generally more robust and shows the expression profile expected in this cell type with all the latent proteins including EBNA 1,2, and 3 and LMP 1 and 2 being expressed (Figure 5). In contrast, the BL samples showed the expected more restricted gene expression pattern (referred to as type 1 latency) with regions in the BamHI A and the EBNA 1 loci showing coverage (Figure 5).

2.4.4 Host Transcriptome Analysis

The host transcriptome analysis component of RNA CoMPASS generates gene expression output files that can be used for cluster and pathway analysis. Gene expression output from RNA CoMPASS analysis of the 22 LCL and 23 BL samples was subjected to hierarchical clustering and differential gene expression analysis. Using the Ward criterion, the samples separated in two well defined clusters with one cluster representing the LCL phenotype and the other representing the BL phenotype (Figure 6). Furthermore, within the BL cluster, biopsies separated from the cell lines, possibly caused by the contribution of stromal signals in the biopsies ads/or by genetic drift in the cell lines.

To investigate differences in LCLs compared to BLs, Ingenuity Pathway Analysis software (IPA: Ingenuity Systems) was used to assist in the analysis of signaling pathways and molecular functions associated with the differentially expressed cellular genes. Upstream regulator analysis within IPA predicted activation of MYC (z-score: 3.375), MYCN (z-score: 2.813), MAPK9 (z-score: 2.414), and MAPK1 (z-score: 2.138) pathways with an inhibition of Interferon alpha (z-score: -2.916), interferon gamma (z-score: -2.788), NF-kB (z-score: -2.746), interferon alpha-2 (z-score: -2.723), and interferon lambda (z-score: -2.000) pathways in BL relative to LCL samples (Figures 7-8). TCF3 (5.4-fold) and TOP2A (9.0-fold) were both increased in BLs relative to LCLs.

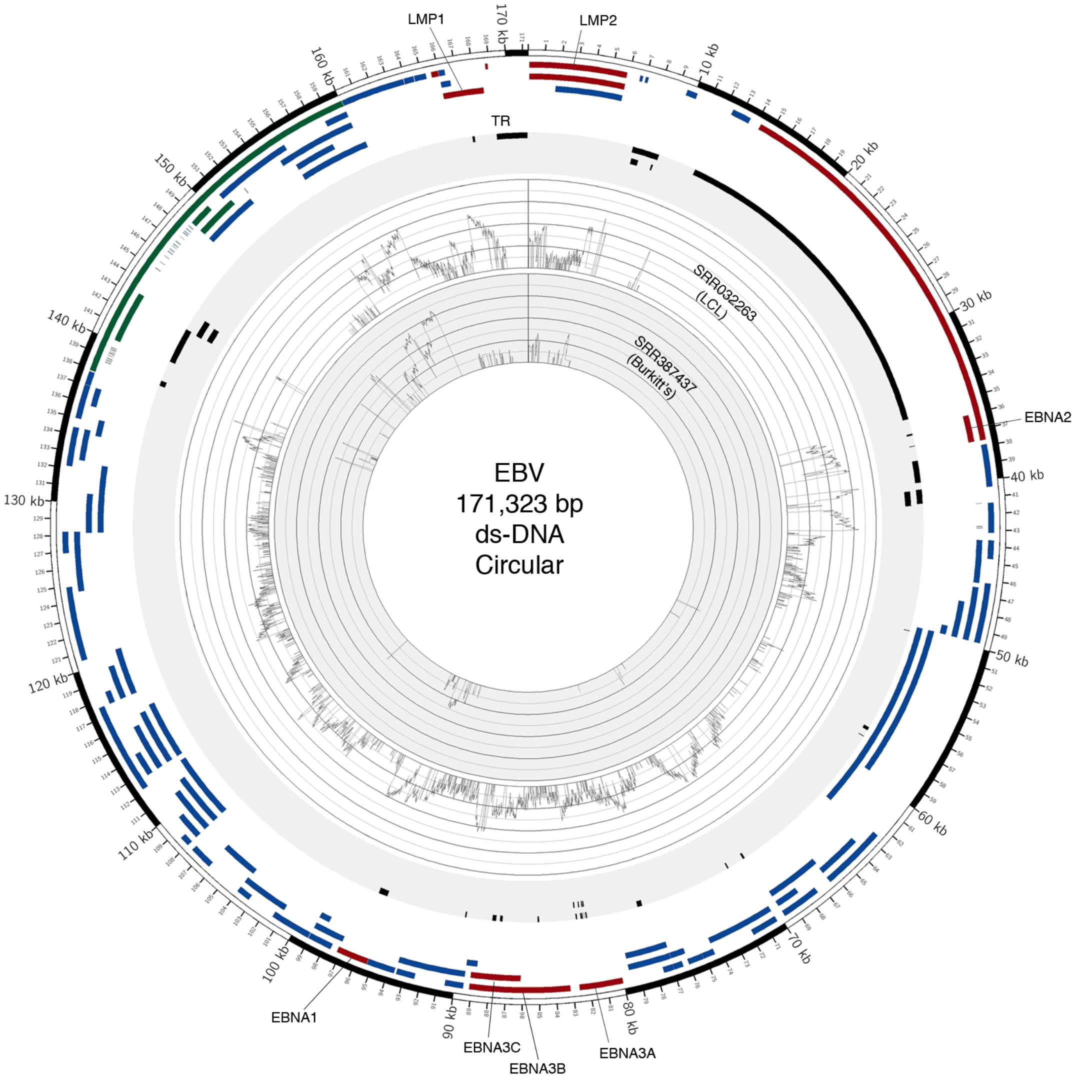


Figure 5. Circos plot of two EBV samples shows distinct gene expression. An annotated Circos plot depicts the EBV read coverage across the EBV genome of two samples. The graph displays the number of reads mapped to each nucleotide position of the genome and are depicted in log scale. Blue features represent lytic genes, red features represent latency genes, green features represent potential non-coding genes, and black features represent non-gene features (e.g. repeat regions and origins of replication).

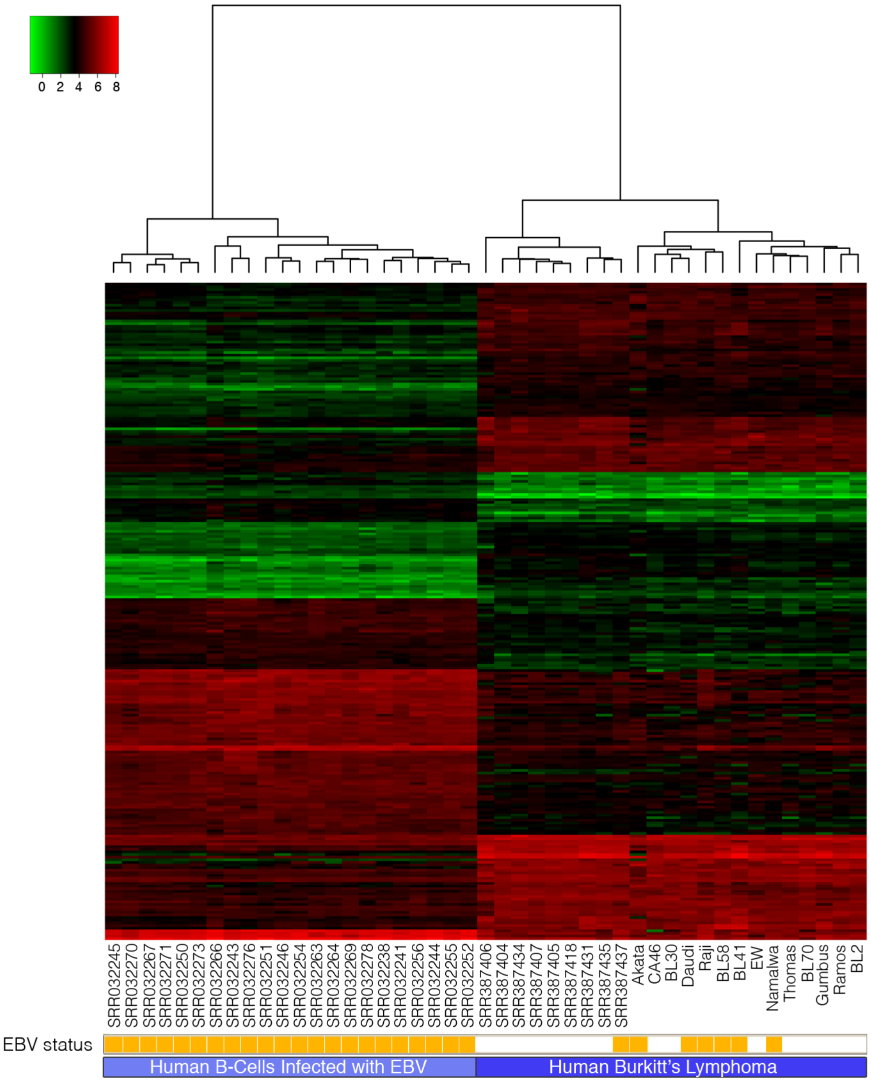

0
2
4
6
8
SRR032245
SRR032270
SRR032267
SRR032271
SRR032250
SRR032273
SRR032266
SRR032243
SRR032276
SRR032251
SRR032246
SRR032254
SRR032263
SRR032264
SRR032269
SRR032278
SRR032238
SRR032241
SRR032256
SRR032244
SRR032255
SRR032252
SRR387406
SRR387404
SRR387434
SRR387407
SRR387405
SRR387418
SRR387431
SRR387435
SRR387437
Akata
CA46
BL30
Daudi
Raji
BL58
BL41
EW
Namalwa
Thomas
BL70
Gumbus
Ramos
BL2
EBV status
Human B-Cells Infected with EBV
Human Burkitt’s Lymphoma

Figure 6. Heat Map representing Human B-Cells analyzed using RNA CoMPASS. Human transcript counts from the 45 B-cell samples were imported into the R software environment and analyzed using the edgeR package [54]. Genes with low transcript counts (less than 1 CPM (count per million)) in the majority of samples were filtered out. The Manhattan (L-1) distance matrix for the samples was computed using the remaining transcript counts, and this was taken as input for hierarchical clustering using the Ward algorithm. After assigning each sample to one of two groups identified by hierarchical clustering (Human B-Cell or Burkitt's Lymphoma), the glmFit function was used to fit the mean log(CPM) for each group and likelihood ratio tests were used to identify those genes that were differentially expressed, with adjusted *P*<0.05 following the Benjamini-Hochberg correction for multiple testing. The fitted log(CPM) values for the subset of genes that were differentially expressed in the LCL samples relative to the Burkitt's lymphoma samples were then clustered using the Euclidean distance and complete linkage algorithm to detect groups of co-expressed genes. The expression heat map displays the top 250 differentially expressed genes.

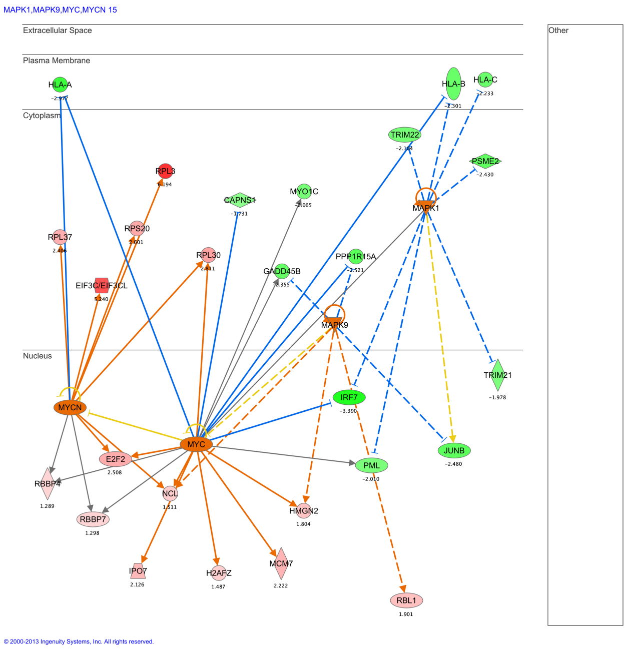


Figure 7. Predicted top activated upstream pathway of top 250 differentially expressed genes

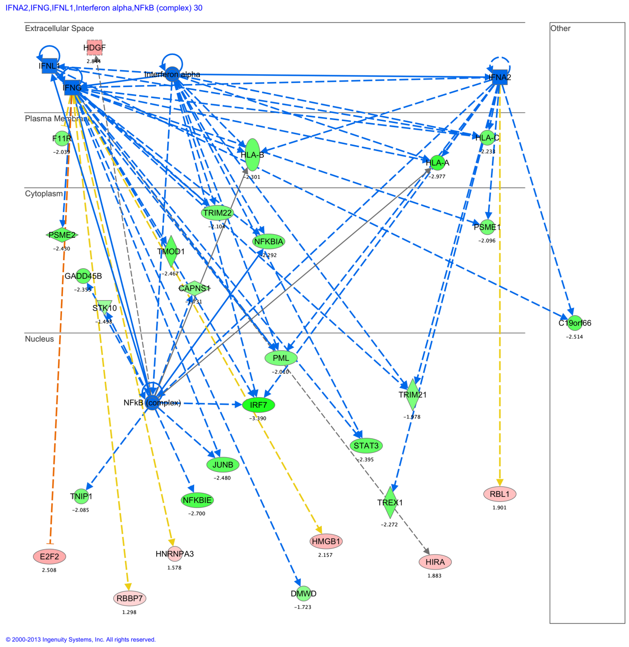


Figure 8. Predicted top inhibited upstream pathway of top 250 differentially expressed genes

2.5 Discussion

RNA CoMPASS is designed to take advantage of several open source programs in order to streamline and accelerate RNA-seq data analysis. RNA CoMPASS helps the researcher to manage the computational burden of processing large sets of RNA-Seq data by parallelizing the most compute intensive steps of the process and automatically managing files through each step of the pipeline. The simultaneous analysis of the host transcriptome along with the discovery of pathogens allows investigators to not only detect pathogens but also study the relationship between the pathogen and host transcription.

In RNA CoMPASS, we have implemented both the Java Parallel Processing Framework (JPPF) API and Portable Batch System (PBS) API in order to deploy it on either a small local cluster or a grid system managed by PBS submission. Our testing of RNA CoMPASS in both environments showed that our pipeline could efficiently analyze RNA-seq data sets achieving a significant speedup over analysis on a single machine. This will allow other investigators to use RNA CoMPASS on whichever type of computational environment they have access to. In our case, we employed RNA CoMPASS on a local 4-node cluster environment (Intel Xeon Mac Pros with 64-96GB RAM) which achieves a speedup approaching the theoretical limit of 4 by splitting the computational tasks over 4 machines as the file size increases. This speedup of a 4 node cluster serves as a proof of principle that an even greater speedup could be obtained using a significant computational cluster involving hundreds of nodes. A recently published study [60] also outlines a different approach using the Bowtie aligner (which is significantly faster than novoalign) to align against the human plus virus genomes. As an alternative approach, they use BLASTing of de novo assembled reads instead of all unaligned reads. These are computationally more efficient approaches but the first method is constrained by index size limitations, which preclude the inclusion of a broad array of organisms such as bacteria and fungi, for example. In contrast, the BLAST approach of RNA CoMPASS surveys the entire NT database. The blasting of only de novo assembled reads would also significantly speed up our approach, however, BLASTing raw reads allows us to quantify relative levels of each exogenous agent found, which is an important read-out for these studies.

Though the BLASTing step of RNA CoMPASS incurs moderate limits on the size of input file that can be processed (depending on access to a large computational cluster), it allows for a more comprehensive analysis. In previous work, we would that sampling 10 million reads from an RNA sequencing experiment is likely to be well beyond the number needed to detect meaningful levels of exogenous agents [61]. A future enhancement of RNA CoMPASS will be to leverage this result and first align all reads from a sample for analysis of human reads, but then to carry forward only an adequately sized sample of unmapped reads through the more computationally burdensome analysis of exogenous agents (BLASTing). Future implementations of RNA CoMPASS are also under development which will leverage large computing clouds (like Amazon EC2) and will also provide the option of using alternative aligners such as Bowtie or STAR to significantly speed up the alignment process.

We used LCLs and BL samples to evaluate the pathogen and host transcriptome analysis arms of RNA CoMPASS because of their differences in phenotypes. The LCLs were generated by infecting human B-cells with EBV, which typically display an activated B-cell like phenotype (type III latency – expressing all 9 EBV latency genes (LMP1, LMP2A, LMP2B, EBNA1, EBNA-LP, EBNA2, EBNA3A, EBNA3B, and EBNA3C) and BART transcripts). The BL samples typically display a germinal center-like phenotype (type I latency – expressing EBNA1 and BART transcripts). The entire sequence file for all samples was used as input for RNA CoMPASS. Although the BL samples were sequenced using a paired-end approach, only one of the reads from each pair was analyzed in order to remain consistent among all samples and because a single-end read should provide sufficient evidence for pathogen discovery.

RNA CoMPASS discovered a significantly larger proportion of EBV (Human Herpesvirus 4) reads within the LCLs as compared to the BLs. This is an anticipated result, which validates the usage of RNA CoMPASS to interrogate genetic material of exogenous origin.

One of the most highly active pathways within BLs is the MYC pathway. In our study, the MYC and MYCN pathways were predicted to be the top two activated pathways in BLs relative to LCLs according to IPA’s upstream regulator analysis. Several MYC targets have been reported in the literature [62-66] and we see many of these targets regulated in our study including the MYC-induced genes, BUB1, CENPF, CCNB1, PLK1, PCNA, AURKB; and the MYC-repressed genes: STAT1, IL10RA, and HLA-DRA. In addition, the MYC pathway has emerged as one of the central regulators of cell growth and ribosome biogenesis by inducing several genes encoding ribosomal proteins [67]. In our study, we observe that the top differentially expressed gene targets of MYC and MYCN are related to ribosomal protein synthesis (NCL, RPL30, RPL37, RPS20, and RPL3).

Transcriptome analysis has shown that the MYC signature is the hallmark signaling difference between Burkitt's lymphomas and diffuse large B-cell lymphomas with: upregulation of MYC-target genes and downregulation of genes involved in the NF-kB and interferon responses [62, 63, 66]. This hallmark signaling is recapitulated in our study using transcriptome analysis of BL and LCL samples. Taken together, the results presented here as well as from others indicate that a single master transcriptional pathway, MYC, mainly governs the growth potential of BLs with the help of other oncogenes as cofactors [66].

On the other hand, the most highly inhibited pathways within BLs compared to LCLs were the interferon response pathway and the NF-kB pathway. The NF-kB pathway has been shown to play a vital role in EBV's ability to transform naïve B-cells as the EBV transforming latency protein, LMP1 continuously activates NF-kB [68, 69]. Among other genes, we observe an increased expression of antigen presentation molecules in LCLs relative to BLs, possibly through the LMP1/NF-kB pathway [68]. The inhibition of the interferon response pathways seen in BLs lends to the overexpression of MYC contributing to immune escape through repression of the interferon response [70].

A few other noteworthy genes that were observed as being differentially expressed include TCF3 and TOP2A, both of which have increased expression in BLs relative to LCLs. In a recent study using RNA-seq with RNA interference screening of BLs, Schmitz et al was able to identify mutations affecting the transcription factor TCF3 [71]. TCF3 has been shown to activate the pro-survival phosphatidylinositol-3-OH (PI(3)) kinase in part by augmenting B-cell receptor signaling [71]. The authors suggest that the MYC and PI(3) kinase pathways may act synergistically in BL oncogenesis, and that the PI(3) kinase pathway may be a new target for drug development [71].

The other molecule, TOP2A has been shown to determine anthracycline-based drug (e.g. doxorubicin) response *in vitro* and *in vivo* [72]. Dose intensity of doxorubicin was evaluated by Kwak and colleagues in a retrospective analysis of 115 patients with diffuse large B-cell lymphoma [73]. The outcome of this study determined that doxorubicin should be used for the treatment of aggressive non-Hodgkin's lymphomas and dose intensity of doxorubicin was a key factor in predicting patient survival. Further, a meta-analysis of published randomized controlled trials comparing chemotherapy regimens incorporating doxorubicin at a high dose with standard CHOP therapy was conducted and their conclusions were consistent with the Kwak and colleagues study [74]. Altogether, levels of TOP2A in BLs are elevated relative to immortalized B cells and that high dose doxorubicin in addition to standard CHOP therapy may improve Burkitt's lymphoma patient outcome through TOP2A mediated doxorubicin response.

2.6 Conclusion

In summary, our results demonstrate the utility of RNA CoMPASS in analyzing large sequence datasets for the discovery of pathogens and host transcriptome analysis. The use of this pipeline is expected to enable more researchers to enter the filed of RNA Sequencing and to yield novel associations between pathogens and human diseases with important medical implications. This study shows the disparate expression profiles between Lymphoblastoid cell lines and the Burkitt's lymphoma samples thereby exhibiting the ability of RNA CoMPASS to analyze endogenous sequencing. RNA CoMPASS is publically available under the GPL: http://rnacompass.sourceforge.net.

We are planning to implement the Circos plot capability for discovered pathogens as well as clustering analysis for host gene expression in later versions of RNA CoMPASS. These improvements will further streamline and complement the analysis of RNA-seq data in the discovery and analysis of pathogens associated with malignancies. In addition, we are investigating ways to further improve the speedup of the pipeline.

Chapter 3: The Feasibility and Effectiveness of RNA CoMPASS in Evaluating Human Pathogens Associated With Disease

Epstein-Barr Virus and Human Herpesvirus 6 Detection in a non-Hodgkin's Diffuse Large B-Cell Lymphoma Cohort using RNA-Seq

Michael J. Strong, Tina O'Grady, Zhen Lin, Guorong Xu, Melody Baddoo, Chris Parsons, Kun Zhang, Christopher M. Taylor, and Erik K. Flemington

Published in J Virol 2013 87(23):13059-62. doi: 10.1128/JVI.02380-13. Epub 2013 Sep 18

3.1 Abstract

Comprehensive virome analysis of RNA-seq data sets from 118 non-Hodgkin's B-cell lymphomas revealed a small subset that are positive for Epstein-Barr virus (EBV) or human herpesvirus-6B (HHV-6B), with one co-infection. EBV transcriptome analysis revealed expression of the latency genes RPMS1, LMP1 and LMP2, with one sample additionally showing high early lytic expression and another sample showing high EBNA2 expression. HHV-6B transcriptome analysis revealed that the majority of genes were transcribed.

3.2 Introduction

Herpesviridae is a large family of DNA viruses that can infect and cause disease in humans. Epstein-Barr Virus (EBV) and Human Herpesvirus-6 (HHV-6) are two members of this family that are highly ubiquitous and have been associated with mononucleosis and exanthema subitum (roseola), respectively. In addition, EBV is a well-known oncovirus that is associated with several malignancies including nasopharyngeal carcinoma, gastric carcinoma, and lymphomas. HHV-6 is an emerging pathogen that has not been defined as an oncogenic pathogen but has been variably associated with lymphomas using traditional detection methods (e.g. polymerase chain reaction (PCR), southern blot, and immunohistochemistry (IHC)) [75].

For many years, associations between cancers and infectious agents have been made through epidemiological approaches and methods such as IHC and PCR. Although IHC and PCR approaches have been important for the detection of infectious agents in cancers, they have also led to false discovery and/or controversy. Several groups, including ours, have utilized RNA-seq for the discovery and investigation of infectious agents; for example, Merkel cell virus linked to Merkel cell carcinoma [40], Fusobacterium associated with colorectal carcinoma [41, 42], EBV associated with gastric carcinoma [61], MuLV in human B-cell lines [76], and the screening of large sequencing databases for oncoviruses [77]. Next generation sequencing (NGS) approaches have several advantages over previous detection methods for this type of study. In addition to high sensitivity, NGS is highly specific since the sequence for each read represents a fingerprint for a particular organism. Another key advantage is that a broad relatively unbiased assessment of all known organisms can be performed in a single assay. Not only does this technology help better identify etiological agents, but it can also better define cancers/specimens that are truly not associated with any known viruses.

Previous associations between EBV and non-Hodgkin's lymphomas [78-81], prompted us to explore the links between diffuse large B-cell lymphomas (DLBCLs) and human viruses using next generation sequencing. Using this approach, we comprehensively assessed the virome of a large non-AIDS non-Hodgkin's lymphoma (NHL) RNA-seq cohort from the Cancer Genome Characterization Initiative (CGCI).

3.3 Results and Discussion

3.3.1 EBV and HHV-6B are detected in a small percentage of diffuse large B-cell lymphomas

RNA-seq data sets from 118 NHLs (105 diffuse large B-cell lymphomas (DLBCLs) and 13 follicular lymphomas (FL)) [82] were downloaded from the NIH database of Genotypes and Phenotypes (dbGap; http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) using accession phs000235.v2.p1 (additional details pertaining to the samples can be obtained through controlled access). Virome analysis of these polyA-selected RNA-seq data sets was performed by running roughly 27 million reads from each sample through our automated RNA-seq exogenous organism analysis software, RNA

CoMPASS [83]. Within RNA CoMPASS, reads were aligned to the human reference genome, hg19 (UCSC), plus a splice junction database (which was generated using the make transcriptome application from Useq [50]; splice junction radius set to the read length minus 4) using Novoalign V3.00.05 (www.novocraft.com) [-o SAM, default options]. Nonmapped reads were isolated and subjected to consecutive BLAST V2.2.27 searches against the Human RefSeq RNA database (an additional "pre-clearing" step) and then to the NCBI NT database to identify reads corresponding to known exogenous organisms [52]. Results from the NT BLAST searches are filtered to eliminate matches with an E-value of greater than $10e^{-6}$. The results are then fed into MEGAN 4 V4.70.4 [53] for visualization of taxonomic classifications.

Most of the samples analyzed contained low levels of bacteriophage sequences, which likely represent either environmental contamination or quality control spike-ins (Figure 1A). Of the 118 samples analyzed, 113 of them showed no evidence of eukaryotic viral polyadenylated RNA expression suggesting a different mechanism for tumor progression in these cases. Nevertheless, five DLBCL samples were positive for EBV (4 samples (3.4%)) or HHV-6B (2 samples (1.7%)) with one of these samples, SRS405443, being co-infected (Figure 1A).

The findings in virus positive samples were further analyzed by combining all sequencing runs for each EBV and/or HHV-6B positive tumor and aligning them directly to the human reference genome (hg19; UCSC) plus the Akata strain of the EBV genome (Genbank accession number KC207813) [84] and the

HHV-6B genome (Genbank accession number NC000898). Alignments were performed using Spliced Transcripts Alignment to a Reference (STAR) aligner V 2.3.0 [default options] [85]. From this analysis, samples SRS405439 and SRS405443 were found to have the highest EBV read numbers (432 and 37 reads per million human mapped reads, respectively), while samples SRS405392 and SRS405456 had relatively low EBV read numbers (3 and 0.5 reads per million human mapped reads, respectively) (Figure 1B). Samples SRS405408 and SRS405443 showed 19 and 99 HHV-6B reads per million human mapped reads, respectively (Figure 1B).

3.3.2 Viral transcriptome analysis

In a recent study we showed that gastric carcinomas with high EBV read numbers exhibited unique signaling effects on cellular and microenvironmental pathways compared to samples with either low or no EBV gene expression [61]. Cluster analysis of these EBV positive samples based on EBV gene expression alone showed unique clustering of the samples with high versus low EBV read counts. The distinct EBV gene expression patterns in these two groups suggested distinct infection types, which may partly explain differences in signaling effects. To similarly assess global differences in EBV gene expression patterns in the EBV positive DLBCL samples, we performed cluster analysis. Transcript quantification of EBV genes was performed using Sammate [51]. Transcript counts and RPKMs (reads per kilobase per million mapped reads) were imported into MeV [86] for hierarchical clustering analysis. The Manhattan

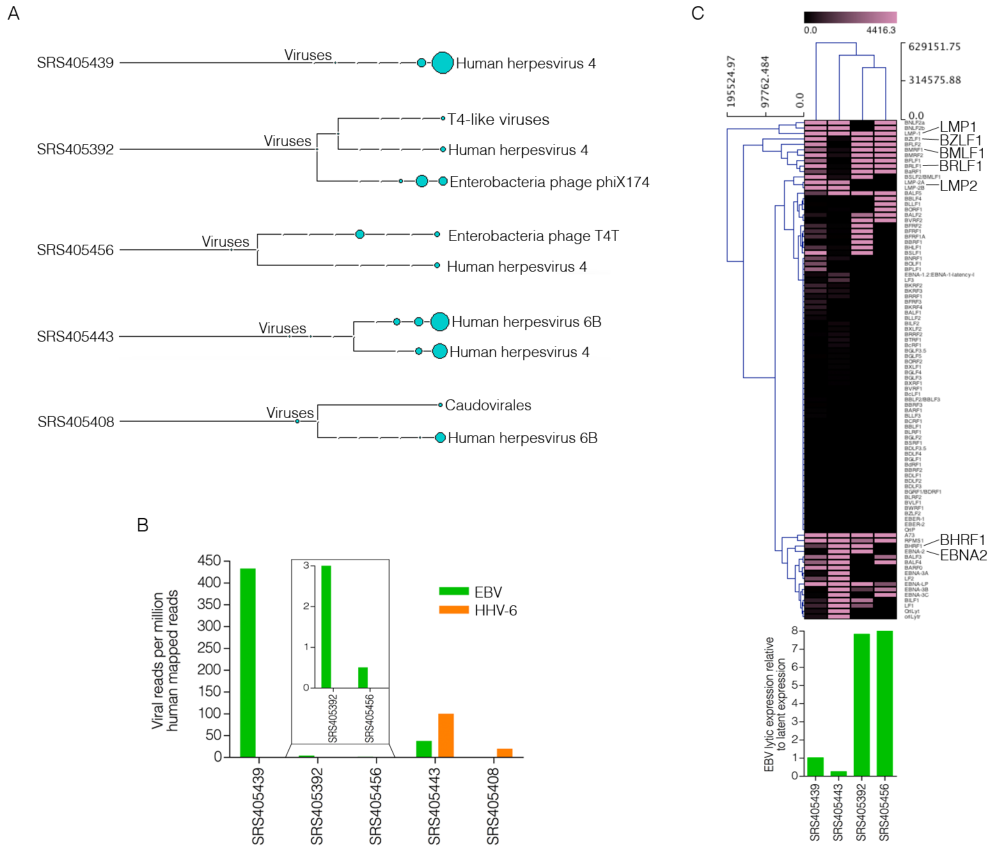


Figure 1. EBV and HHV-6B detection in diffuse large B-cell lymphomas. (A) Virus branch of the taxonomy trees for the five virus positive DLBCL samples. (B) Number of viral reads per million human mapped reads. (C) Cluster analysis of EBV transcripts along with a bar graph representing the ratio of EBV lytic to latent gene expression for each positive sample. (Adapted from Strong et al., 2013).

distance matrix was computed for the samples and used as input for hierarchical clustering using the complete linkage-clustering algorithm. The samples with low EBV read counts, SRS405392 and SRS405456, were found to cluster together (Figure 1C). Visualization of reads across the EBV genome using the Integrative Genomics Viewer (IGV) [87] showed latency gene peaks in the two samples with high EBV read counts (Figure 2A). In contrast, only scattered reads were observed across the entire genome in the two samples with low EBV read counts (data not shown). This observation is illustrated by the finding of high lytic to latent read ratios in the samples with low EBV versus high EBV read counts (bottom of Figure 1C). The lack of distinct latency gene expression along with the observed overall low EBV transcript levels for the samples with low EBV read numbers raises the possibility that the finding of EBV in these samples is less consequential than it is in samples SRS405439 and SRS405443, possibly reflecting low level reactivation in infiltrating latent B-cells.

Detailed analysis of gene expression in the two EBV positive samples with higher read counts showed expression of the EBV latency genes, RPMS1, LMP1 and LMP2 genes in both cases (Figure 2A). In contrast to these similarities, EBNA2 was found to be expressed in sample SRS405443 but not in sample SRS405439 (Figure 2A). On the other hand, sample SRS405439 was unique in the detection of lytic transcripts with a disproportionately high level of the immediate early/early genes BZLF1 and BMLF1 relative to the bulk of other lytic genes (Figure 2A). This predominant expression of early genes without other lytic genes is suggestive of an abortive lytic cycle, which has previously been linked to tumor progression [61, 88, 89]. In contrast to the hallmark distinctive expression of EBV latency genes in samples SRS405439 and SRS405443, HHV-6B gene expression showed a more broad expression profile across the entire genome, consistent with lytic transcription (Figure 2B).

In conclusion, based on the samples tested here, most non-AIDS NHLs are free of known eukaryotic viruses expressing polyadenylated RNAs. In the two EBV positive DLBCL samples, SRS405439 and SRS405443, the high read numbers in conjunction with the finding of clear expression of oncogenic latency genes [90-93] is consistent with an etiological role for EBV in these cases. In contrast, it is much less clear whether EBV contributes to the tumor phenotype in the two samples with lower read numbers where there is a lack of pronounced oncogenic latent gene expression. Similarly, the general observation of broad lytic HHV-6B gene expression in the two HHV-6B positive samples rather than expression of any particular potentially oncogenic latency gene suggests that at a minimum, any contribution of HHV-6B to tumor progression likely occurs through a different mechanism (e.g. through a mechanism involving persistent smoldering stimulation of an inflammatory response to HHV-6B lytic antigens).

It is possible that moderate disease related immunosuppression could lead to HHV-6B reactivation in HHV-6B positive tumors, which may or may not contribute to the tumor phenotype. The finding of EBV and HHV-6 co-infection in one case raises the possibility that this patient may in fact have some level of immunosuppression. The expression of the highly immunogenic EBNA2 gene in this case further supports the suspicion of immunosuppression.

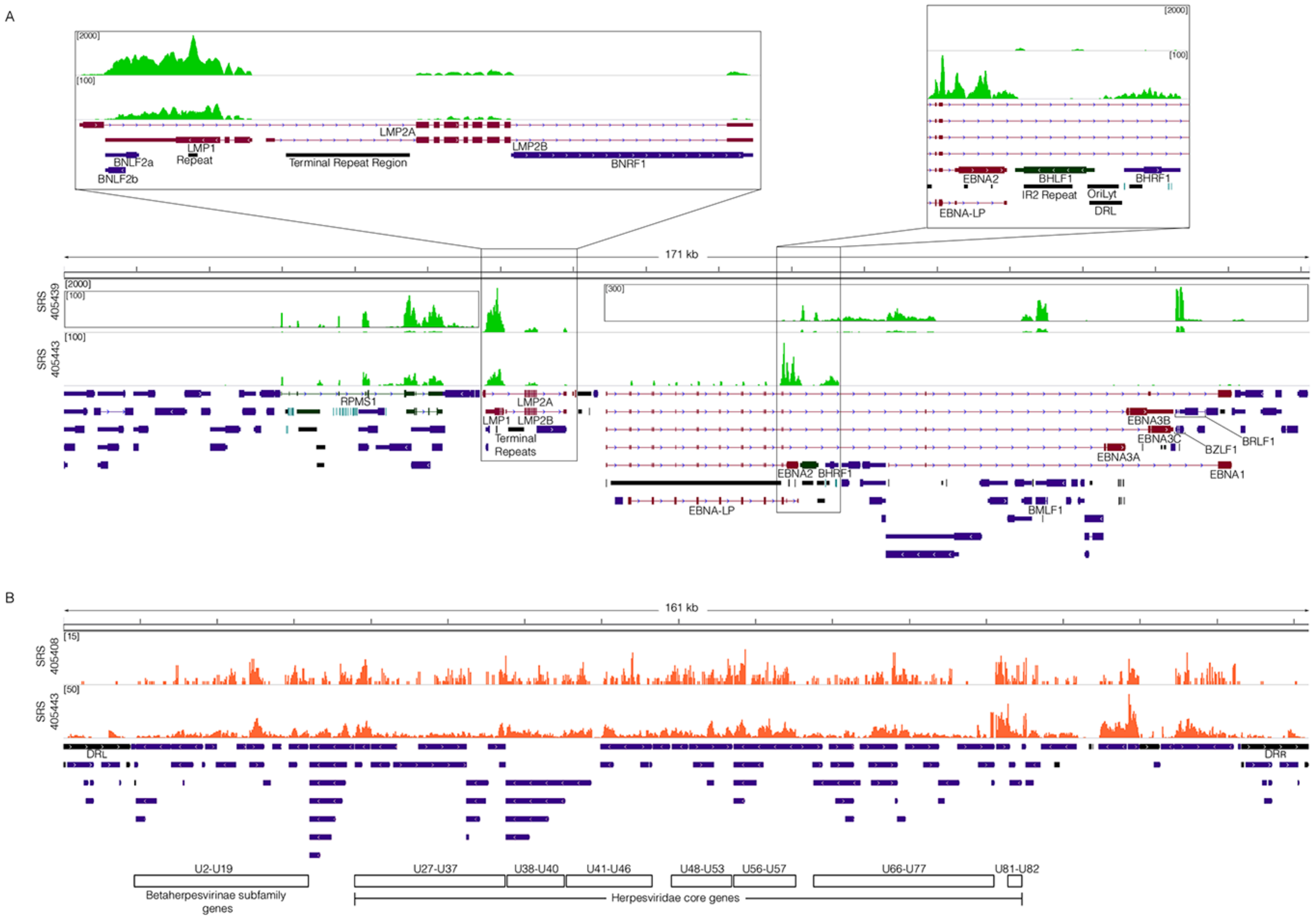

A
[2000]
[100]
LMP2A
LMP1
Repeat
BNLF2a
BNLF2b
Terminal Repeat Region
LMP2B
BNRF1
[2000]
[100]
EBNA2
BHLF1
BHRF1
IR2 Repeat
OriLyt
DRL
EBNA-LP
171 kb
SRS 405439
[2000]
[100]
[300]
SRS 405443
[100]
RPMS1
LMP2A
LMP1
LMP2B
Terminal Repeats
EBNA2
BHRF1
EBNA-LP
BMLF1
EBNA3B
EBNA3C
EBNA3A
BZLF1
BRLF1
EBNA1
B
161 kb
SRS 405408
[15]
SRS 405443
[50]
DRL
DRR
U2-U19
Betaherpesvirinae subfamily genes
U27-U37
U38-U40
U41-U46
U48-U53
U56-U57
U66-U77
U81-U82
Herpesviridae core genes

Figure 2. **Viral transcriptome.** (A) EBV genome coverage data for EBV positive DLBCLs. The y-axis represents the number of reads at each nucleotide position. The modified EBV Akata genome was split between the BBLF2/3 and the BGLF3.5 lytic genes rather than at the terminal repeats to accommodate coverage of splice junctions for the latency membrane protein, LMP2. The scale for sample SRS405439 is set to a max read level of 2000 reads with the inset displays set to a max read level of 100 and 300 reads. The scale for sample SRS405443 is set to a max read level of 100 reads. Blue features represent lytic genes, red features represent latent genes, green features represent potential non-coding genes, aquamarine features represent microRNAs, and black features represent non-gene features (e.g. repeat regions). (B) HHV-6B genome coverage data for HHV-6B positive DLBCLs. The scale for sample SRS405408 is set to a max read level of 15 reads, while SRS405443 is set to 50 reads. (Adapted from Strong et al., 2013).

Despite this possibility, it seems likely that EBV latency genes contribute to tumor progression [90-93] in this patient. Whether HHV-6B plays a role in tumor progression or whether expression is just a bystander effect of possible immunosuppression is unclear and will require further investigation. Regardless, HHV-6B is a component of the tumor microenvironment and it is appropriate to consider its presence in potential tailored future therapeutic design.

Chapter 4: Assessment of Contaminating/Superfluous Sequences in Next Generation Sequencing Datasets

Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples

Michael J. Strong, Guorong Xu, Lisa Morici, Sandra Splinter Bon-Durant, Melody Baddoo, Zhen Lin, Claire Fewell, Christopher M. Taylor, and Erik K. Flemington

Published in Plos Pathog 2014 Nov 20;10(11):e1004437. doi: 10.1371/journal.ppat.1004437. eCollection 2014.

4.1 Abstract

The high level of accuracy and sensitivity of next generation sequencing for quantifying genetic material across organismal boundaries gives it tremendous potential for pathogen discovery and diagnosis in human disease. Despite this promise, substantial bacterial contamination is routinely found in existing human-derived RNA-seq datasets that likely arises from environmental sources. This raises the need for stringent sequencing and analysis protocols for studies investigating sequence-based microbial signatures in clinical samples.

4.2 Introduction

The advent of next generation sequencing (NGS) technology has revolutionized the way pathogens can be detected, studied, and discovered. NGS lends itself to highly sensitive, relatively unbiased, global assessments of all known exogenous agents within biological specimens, including human biopsies. Several laboratories, including ours, have successfully utilized NGS for the discovery and investigation of exogenous agents associated with several human diseases such as the recent association of fusiform bacteria with colorectal carcinoma [40, 42, 61, 76, 94-96]. NGS-based approaches also have great potential in the clinic for the diagnosis of symptomatic infections. Early studies examined microbial sequence-based signatures in feces from patients with diarrheal disease and in urine from patients suspected of having a urinary tract infection to identify the infectious cause [97, 98]. In a recent case report, NGS was used to diagnose a patient with a rare but treatable bacterial meningoencephalitis caused by leptospirosis, a condition which was undetectable using current clinical assays [99].

With the great potential of NGS for pathogen analysis of clinical samples, opportunities are being discussed and bioinformatics challenges are being addressed [100, 101]. While the discussion of opportunities and bioinformatics challenges is highly appropriate, data reliability and contamination, issues that are especially relevant to the inquisitive nature of this application, are scarcely discussed. For some of the current mainstream applications of NGS, such as host transcriptome quantification, reproducibility studies across sequencing centers are being performed to assess data veracity [102]. At a minimum, data reliability in pathogen sleuthing also needs to be thoroughly tested and analyzed, and potential obstacles need to be addressed.

4.3 Results and Discussion

4.3.1 Bacterial reads in multiple human derived RNA-seq datasets

During the course of DNA and RNA sequencing experiments performed in our laboratory over the past several years we invariably noted surprising levels of bacterial reads whether the genetic material was derived from human clinical specimens, tissue culture cells, or animal tissues. The extent and pervasiveness of this observation led us to investigate this issue using data from a variety of other publically available data sources. As a first line of investigation, we downloaded RNA-seq datasets from 93 invasive breast carcinomas [103], 15 kidney renal papillary cell carcinomas, 18 lung adenocarcinomas [104], 38 lung squamous cell carcinomas, and 50 rectum adenocarcinomas [105] from the TCGA cohort which (originally made available from the database of Genotypes and Phenotypes (dbGaP) (phs000178)). Colorectal carcinoma (CRC) RNA-seq datasets from Castellarin et al. were downloaded from the NCBI Sequence Read Archive (accession number SRP007584) [42]. We also downloaded RNA-seq datasets from normal human tissue samples from the Illumina Human Body Map 2.0 project (from the NCBI Gene Expression Omnibus database (GEO accession number: GSE30611)). In total, we analyzed RNA-seq datasets from 244 different specimens from different sources and from different specimen types (Table 1).

Ten specimens were identified as outliers based on poor alignment percentages to the human genome (using the ROUT outlier test in GraphPad Prism (version 6 Mac, www.graphpad.com)) and excluded from the analysis.

Metatranscriptome analysis was performed using our computational pathogen detection pipeline, RNA CoMPASS [83]. Briefly, reads ranging from 42-101 nucleotides long were aligned to the human reference genome, hg19 (UCSC), plus a splice junction database (which was generated using the make transcriptome application from Useq [50]; splice junction radius set to the read length minus 4), and abundant sequences (which include sequence adapters, mitochondrial, ribosomal, enterobacteria phage phiX174, poly-A, and poly-C sequences) using Novoalign V3 (www.novocraft.com) [-o SAM, default options]. Nonmapped reads were isolated and subjected to consecutive BLAST V2.2.28 searches against the Human RefSeq RNA database and then to the NCBI NT database to identify reads corresponding to known exogenous organisms [52, 58]. Results from the NT BLAST searches were filtered to eliminate matches with an E-value greater than 10e-6. The results were then fed into MEGAN 4 V4 [53] for visualization of taxonomic classifications.

RNA CoMPASS, analysis revealed fairly extensive levels of bacterial reads across all RNA-seq studies analyzed, with average numbers ranging from 1,406 reads per million human mapped reads (RPMHs) in the TCGA datasets to 11,106 RPMHs in the normal tissue from the CRC dataset (Table 2 and Figure 1). Despite the widespread presence of bacteria across groups, different taxa displayed substantial heterogeneity across studies with high levels of *Paracoccus* *denitrificans* SD1 in the TCGA and BodyMap datasets but not in the CRC dataset, and *Pseudomonas* showing generally high levels in the CRC but not the TCGA or BodyMap studies (Table 2 and Figure 2). The substantial bacterial read numbers for each of these diverse datasets suggest a fairly ubiquitous nature to these findings, and taxa specific differences across centers raises the possibility of multiple center specific issues.

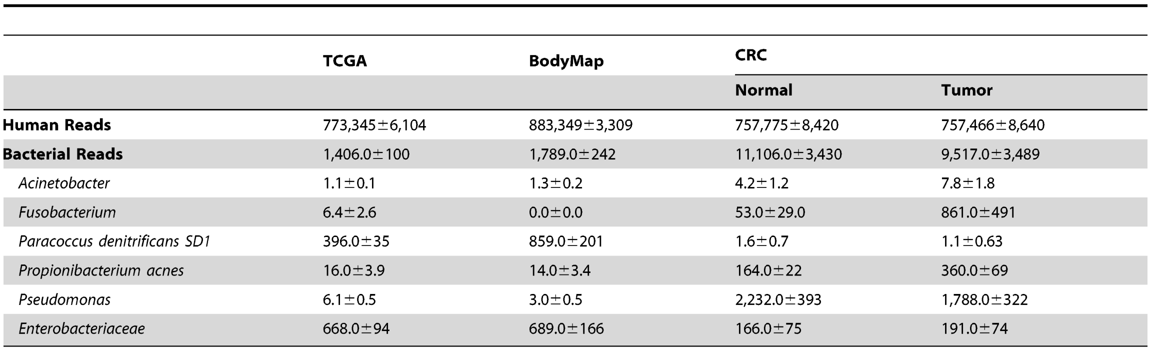
4.3.2 Identical cell lines analyzed in separate studies show differences in bacterial read profiles

To shed light on possible contamination sources we analyzed bacterial reads in cell lines, which we presumed to be free from microbial contamination. RNA-seq data from 7 different diffuse large B-cell lymphoma (DLBCL) cell lines that were analyzed independently in the Cancer Genome Characterization Initiative (CGCI) and the Cancer Cell Line Encyclopedia (CCLE) studies were analyzed. CGCI and CCLE RNA-seq datasets were downloaded from dpGaP (phs000235) and the Cancer Genomics Hub (managed by the University of California, Santa Cruz), respectively.

Table 1. List of Databases

Databases	Number of samples analyzed
TCGA	
BRCA	88
KIRP	15
LUAD	18
LUSC	38
READ	48
BodyMap	13
Adipose	
Adrenal	
Breast	
Colon	
Heart	
Kidney	
Lymph Node	
Ovary	
Prostate	
Skeletal Muscle	
Thyroid	
Testes	
White Blood Cells	
COAD	
Normal	12
Tumor	12

Table 2. Bacterial profile among various human RNA-seq datasets



	TCGA	BodyMap	CRC	
			Normal	Tumor
Human Reads	773,345±6,104	883,349±3,309	757,775±8,420	757,466±8,640
Bacterial Reads	1,406.0±100	1,789.0±242	11,106.0±3,430	9,517.0±3,489
Acinetobacter	1.1±0.1	1.3±0.2	4.2±1.2	7.8±1.8
Fusobacterium	6.4±2.6	0.0±0.0	53.0±29.0	861.0±491
Paracoccus denitrificans SD1	396.0±35	859.0±201	1.6±0.7	1.1±0.63
Propionibacterium acnes	16.0±3.9	14.0±3.4	164.0±22	360.0±69
Pseudomonas	6.1±0.5	3.0±0.5	2,232.0±393	1,788.0±322
Enterobacteriaceae	668.0±94	689.0±166	166.0±75	191.0±74

The average of five RNA-seq datasets represent values for TCGA. Similarly, the average of thirteen RNA-seq datasets represent values for BodyMap. Colorectal (CRC) RNA-seq datasets were obtained from Castellarin et al. accession number SRP007584. All values shown as mean±s.e.m.

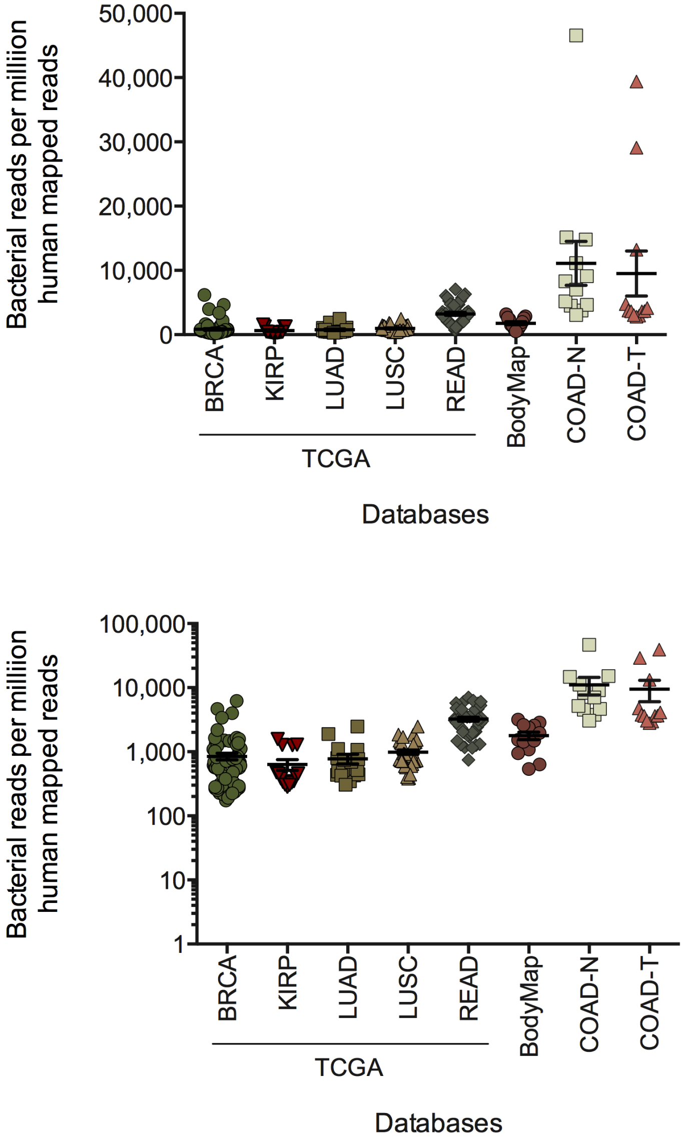


Figure 1. Bacterial reads across RNA-seq datasets. Data displayed in linear and log scales.

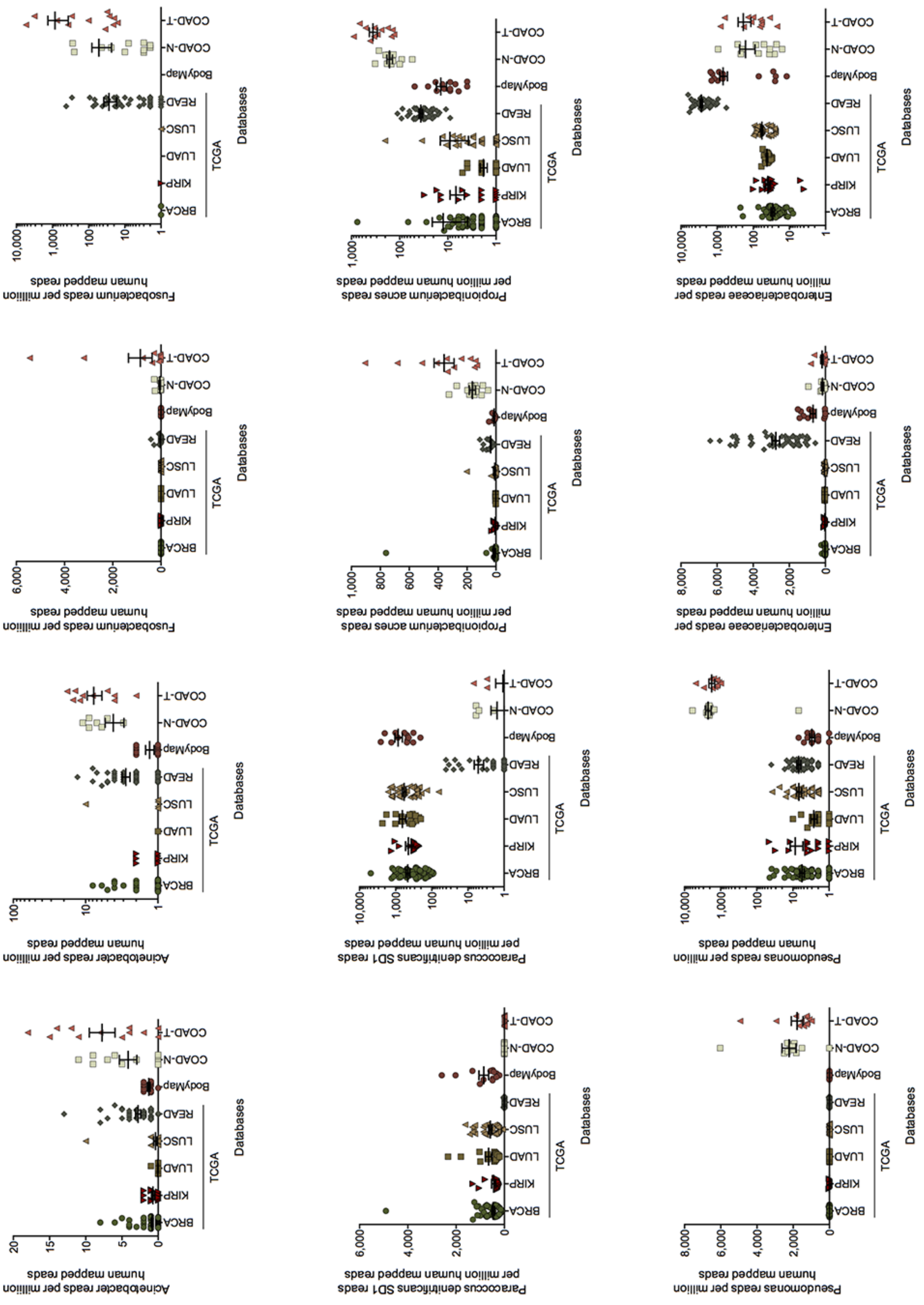


Figure 2. Various bacterial species reads across RNA-seq datasets. Data displayed in linear and log scales.

Based on averaging RPMHs across all cell lines for each study, bacterial reads were found in all datasets, with a considerably greater number in the CGCI study (Figure 3A). *Acinetobacter* was found to contribute to the bulk of bacterial reads in the CGCI data and *Paracoccus denitrificans* SD1 made up the majority of bacterial reads in the CCLE study (Figure 3A). Higher bacterial reads were consistently found in all of the CGCI cell lines except for NU-DUL-1 (Figure 3B). In CCLE data, all cell lines were found to be enriched for *Paracoccus denitrificans* SD1 reads relative to the CGCI data, whereas the converse was true for *Acinetobacter* (Figure 3C).

The discovery of bacterial reads in cell line data and the finding of different bacterial taxa in data from different sequencing initiatives supports the idea that a good portion of bacterial reads are not derived from the specimens themselves. It is noteworthy that most of these datasets were derived from RNA samples that were polyA selected, a process that selects against most bacterial transcripts (which are typically poorly polyadenylated) [106-109]. Contamination that occurs upstream from the polyA selection step, then, is expected to be removed during this purification step. Nevertheless, inefficiencies in polyA selection can result in carry-through of non-polyadenylated bacterial RNAs. If inefficient polyA selection accounted for the majority of bacterial read findings then we would expect that differences in levels of bacterial reads would relate to differences in polyA selection efficiencies between samples. We assessed polyA selection efficiencies by determining the number of ribosomal RNA reads for each sample and we found little correlation between the numbers of bacterial reads and the levels of human ribosomal reads (Figure 3A and B) supporting the contention that downstream contamination is likely a key source of bacterial reads in these datasets.

4.3.3 Different bacterial read profiles across sequencing centers using identical RNA samples and library preparation kits

To more directly address whether downstream contamination can occur, we took advantage of a well-controlled study performed by the GEUVADIS consortium [102, 110]. In their pilot study, ERP000177, RNA from five Epstein Barr virus (EBV) positive lymphoblastoid cell line samples was delivered to seven different sequencing laboratories across Europe to evaluate the reproducibility of sequencing data across various centers. We restricted our analysis to the six laboratories that used Illumina sequencing. For these datasets, library construction at all institutes was performed utilizing identical library preparation kits. Across these labs the level of bacterial RPMHs differed by as much as 30-fold, with Lab 5 showing an average of 18 bacterial RPMHs while Labs 1 and 6 showed an average of 542 and 570 bacterial RPMHs, respectively (Figure 4A). Also noteworthy is the tight clustering of bacterial read numbers in different samples within each lab, suggesting the attribution of bacterial contamination to laboratory practices and/or the environment. Similar to our findings in the DLBCL data, the levels of bacterial reads across centers did not correlate with the levels of human ribosomal RNA contamination, indicating that these differences were not due to polyA-selection disparities. Finally, differences in read levels for

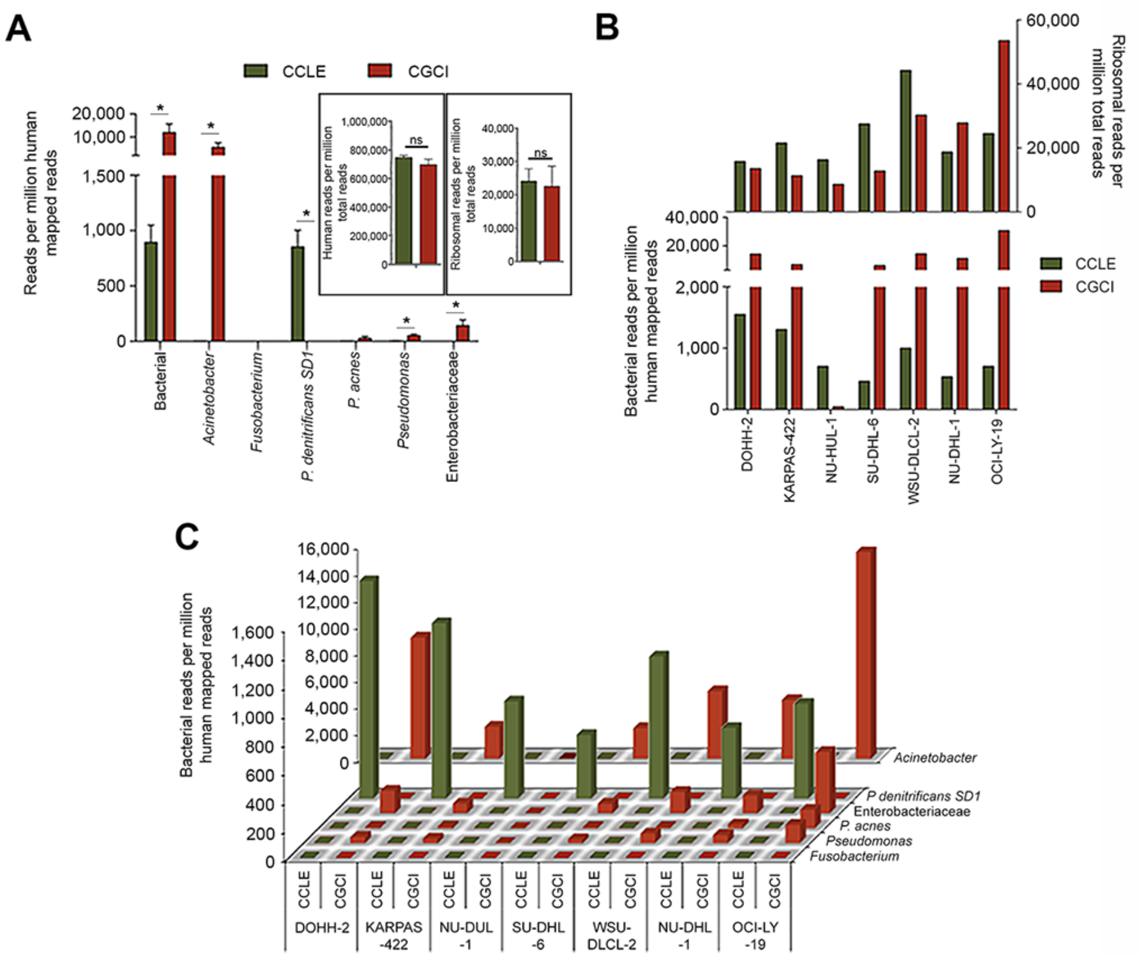


Figure 3. Seven RNA-seq DLBCL cell line datasets sequenced in two different studies (CCLE and CGCI) were analyzed using RNA CoMPASS. (A) Bacterial reads per human mapped reads. For insets, human and ribosomal reads are normalized to total reads. Green columns represent the average RNA-seq reads from the CCLE dataset, while red columns represent the average RNA-seq reads from the CGCI dataset. (B) Mean bacterial RPMHs for each cell line analyzed in the CCLE (green) and CGCI (red) studies with the corresponding mean ribosomal reads (upper graph). (C) Mean RPMHs of various taxa for each cell line analyzed in the CCLE (green) and CGCI (red) studies. *, $P < 0.05$.

different bacterial taxa were found across labs (Figure 4B-E and Figure 5) including the presence of high Xanthomonadaceae read numbers in all five cell line datasets from Lab 1 (Figure 4E (inset)). In contrast, the read levels for endogenously expressed Epstein Barr virus transcripts were similar across labs for each cell line (Figure 4F).

4.3.4 Contamination levels

Based on our own observations as well as the observations of others [111, 112] we think that bacterial contamination is a relevant issue that needs to be extensively addressed for NGS-based pathogen detection and diagnostic approaches. The amplitudes of contaminating bacterial reads in RNA-seq datasets are likely high enough to be a confounding factor. For example, our analysis of the data from the CRC study which previously reported the association between *Fusobacterium* and colorectal cancer [42] showed an average of 861 *Fusobacterium* RPMHs in the tumor samples (Table 2). This is comparable to the levels of *Paracoccus denitrificans* SD1 and Enterobacteriaceae found in the Human BodyMap study (859 and 689 RPMHs, respectively) (Table 2). This observation is more notable considering the fact that the data from the BodyMap study was derived from polyA-selected RNAs, whereas the data from the CRC data was generated using ribodepleted RNA (which does not select against bacterial reads).

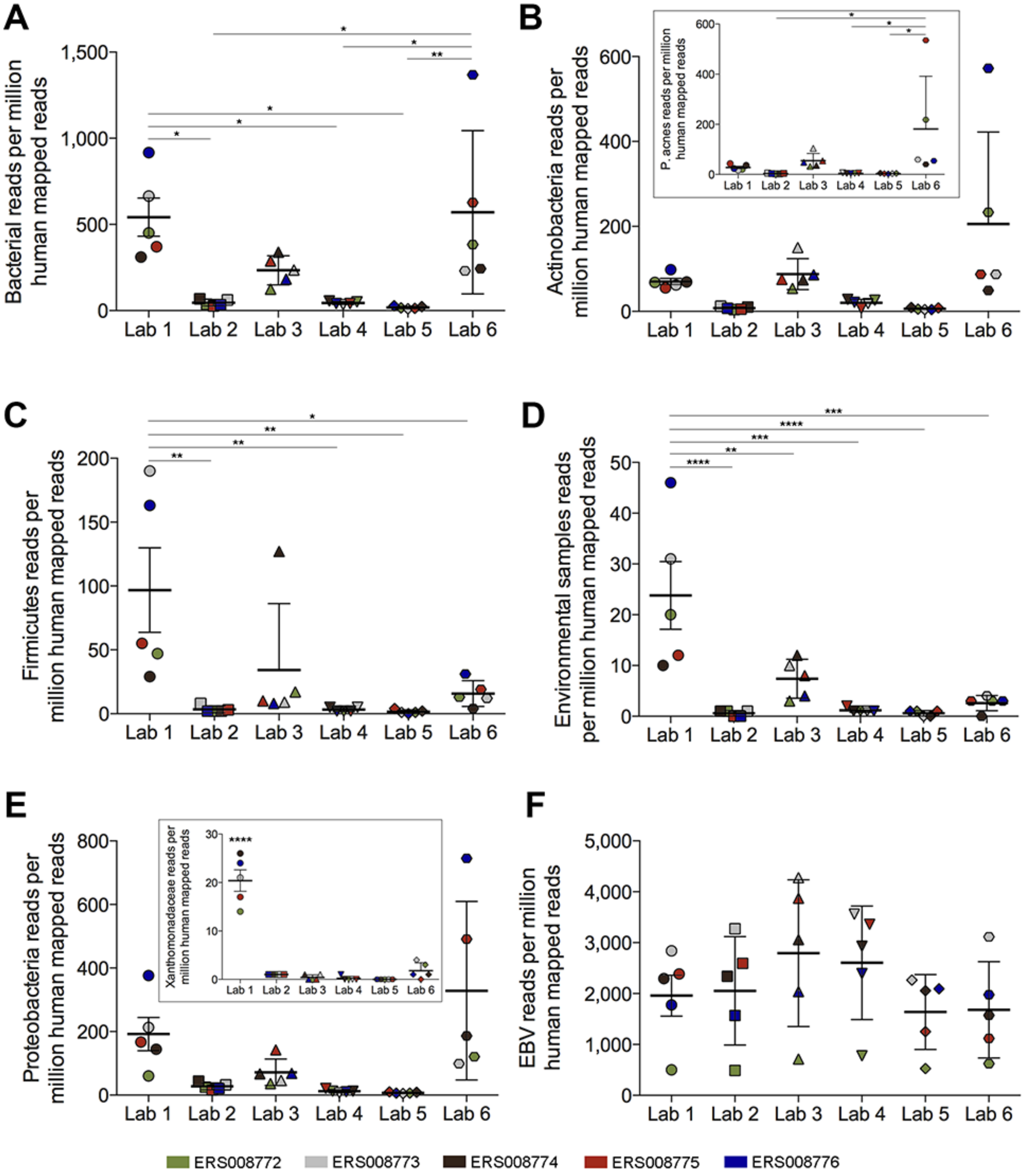


Figure 4. Metatranscriptomic profiles of five RNA sequencing datasets vary across laboratories. Five LCL RNA-seq datasets, sequenced at six sequencing centers across Europe, were analyzed using RNA CoMPASS. Various classification groups within the bacteria domain for each sample were compared across sequencing centers (A) bacteria (B) Actinobacteria (C) Firmicutes (D) environmental samples and (E) Proteobacteria. (F) As a control, EBV read numbers were also analyzed. All reads are normalized to million mapped human reads. The five LCL RNA samples are represented by unique respective colors. *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001.

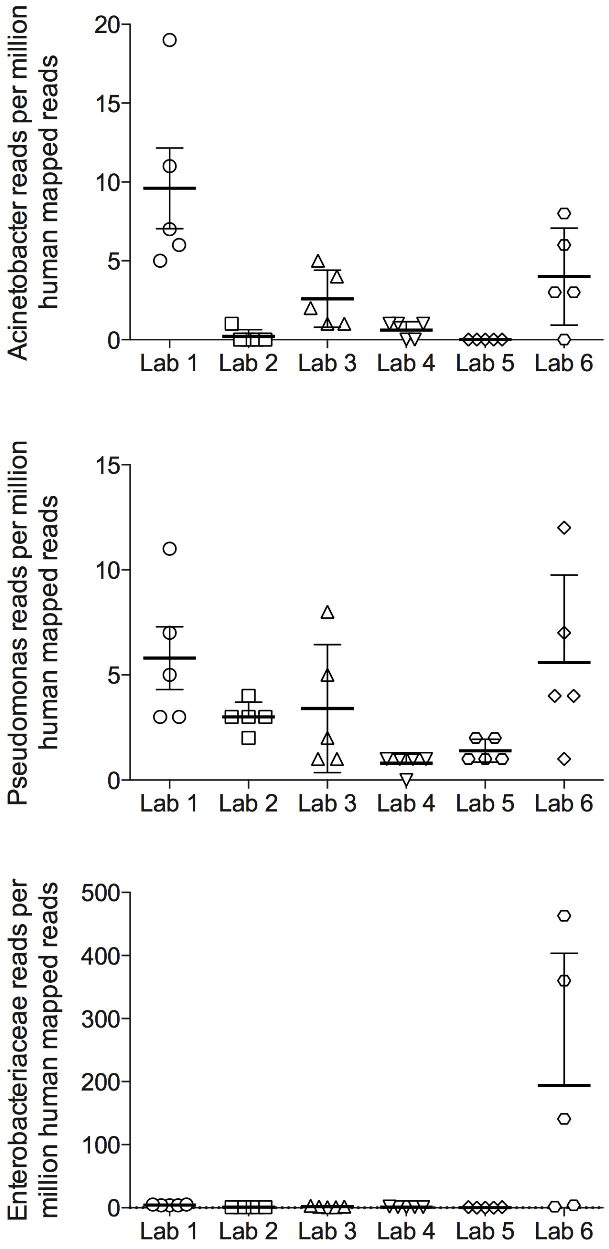


Figure 5. Major bacterial contributors to Proteobacteria taxa.

4.3.5 Is contamination a threat to all microbial sequencing studies?

There are several different approaches to sequencing-based microbial examination which vary based on the starting material; for example, RNA versus DNA, or the investigation of relatively pure microbial samples versus the assessment of heterogeneous samples where the microbial genetic material is a minor component (such as much of the clinical human tissue-based work). The impact of contamination on data interpretation varies depending on the approach because different methodologies inherently traject different signal to noise ratios. Contamination is less relevant for studies utilizing relatively homogeneous microbial communities but it can be a confounding factor in the assessment of samples where the predominant genetic material is human (for example, tumor biopsies) or where the offending microbe is in the minority.

A somewhat less obvious effect on signal to noise ratio is the difference between sequencing RNA versus DNA. Assuming contamination that occurs downstream from the nucleic acid preparation step, there is a larger impact of contaminating microbial DNA on RNA sequencing relative to DNA sequencing approaches. This difference arises due to the inefficiencies in converting RNA to cDNA. Since contaminating DNA does not require this step, the signal to noise ratio for RNA-seq is lower than for DNA-seq.

So why not just sequence DNA? There are certainly advantages to sequencing DNA including its greater stability and the ability to retrieve genetic material from archived samples. Nevertheless, there are also advantages to sequencing RNA for some applications. There is an abundance of publicly available RNA-seq datasets that are potentially useful for future pathogen studies. Another advantage is relevant to the study of human biopsies where the microbial material is a minor component of the sample. The bacterial to human transcriptome size ratio is typically greater than the bacterial to human genome size ratio because of the abundance of extra human DNA that is poorly or not expressed. In these cases, it is more cost effective to assess the microbial component through RNA sequencing. An added benefit of RNA-seq for clinical diagnosis is the ability to simultaneously obtain information on expressed pathogenic and resistance markers that can inform treatment options.

In the end, when it makes sense for a particular study, one way to obviate the impact of potential contamination is to use a viable approach that maximizes the signal to noise ratio. On the other hand, when methods are required that have inherently lower expected signal to noise ratios, alternative approaches are necessary to combat this issue.

4.3.6 Dealing with contamination issues

For some cases, contamination can potentially be dealt with bioinformatically. One approach would be to utilize a repository of common contaminating organisms (although this could potentially result in oversight of a relevant organism that happens to be a common contaminant). Alternatively, for investigations where negative controls are available (and/or suitable), statistics can be used to prove an association (although contamination could result in the requirement for larger sample sets than would otherwise be necessary to attain statistical significance). Despite the utility of informatics approaches to alleviate contamination issues in some cases, minimizing contamination sources is more cost effective and will minimize the chances of data misinterpretation.

Interestingly, contamination has already had an impact on the very databases that are used for bioinformatics work. Laurence et al. identified Bradyrhizobium sequences in assembled genomes in the NCBI Genome database [111]. Bradyrhizobium species along with other microbes, have been reported in ultrapure water systems and may help explain the presence of this microbe in several deposited genome assemblies. Another group found *Leucobacter sp.* sequences in assembled genomes of *Caenorhabditis sp.* [112]. These two cases exemplify the need to sequence contaminant genomes in order to exclude them from the host genome assembly.

Furthermore, in a recent study, Xu et al. discovered National Institutes of Health-Chongqing virus (NIH-CQV) in patients with seronegative hepatitis using NGS [113]. However, two later studies demonstrated that the presence of parvo-like hybrid virus (PHV) and NIH-CQV was actually contamination from silica column-based nucleic acid extraction kits and not bona fide viral infection, indicating that contamination is not restricted to bacterial sequences [114-116]. Subsequently, in a follow up study, the authors of the initial report confirmed that the finding of NIH-CQV in human plasma was due to contamination from the columns [117]. This example underscores the importance of rigorously validating novel pathogen discoveries, and when possible, identifying any potential contaminating sources.

The route between clinical specimen collection to the sequencing reaction is complex with many candidate points of contamination ranging from specimen contamination in the operating room to storage, sample processing, RNA preparation, library preparation, etc. Another key consideration is the purity of library preparation reagents, many of which (e.g. ligases, polymerases, nucleotides) are purified from bacteria during their manufacture. Depending on the level of purity for these reagents, there is the potential for different levels of bacterial genetic material to be present. Nevertheless, the analysis of the data from the highly controlled GEUVADIS study suggests that laboratory SOPs specific to different sequencing centers is also a critical consideration.

The relative contribution of this panorama of potential contamination sources needs to be parsed in future expressly designed studies. Until these sources are better understood, we propose the following recommendations:

1) Detection studies, especially with a diagnostic focus, should incorporate stringent SOPs across the entire experimental pipeline from sample collection to sequencing.

2) Highly purified metabolic enzymes and other reagents used in sequence library preparation should be used whenever possible.

3) Establishment of standards for the curation of microbial sequences submitted to Genbank and other large-scale databases in order to assess completeness and quality of the assembled genomes.

4) Contamination controls such as mock sequence library preparations should be used to help guide the development of appropriate and effective SOPs for metagenomic and metatranscriptomic studies.

PART II

THE ROLE OF EPSTEIN-BARR VIRUS (EBV) IN GASTRIC CARCINOMA

Chapter 5: EBV Transcriptomics Reveal Distinct Molecular Pathways in the Pathogenesis of EBV Associated Gastric Carcinoma

Differences in gastric carcinoma microenvironment stratify according to EBV infection intensity; implications for possible immune adjuvant therapy.

Michael J. Strong, Guorong Xu, Joseph Coco, Carl Baribault, Dass S. Vinay, Michelle R. Lacey, Amy L. Strong, Teresa A. Lehman, Michael B. Seddon, Zhen Lin, Monica Concha, Melody Baddoo, MaryBeth Ferris, Kenneth F. Swan, Deborah E. Sullivan, Matthew E. Burow, Christopher M. Taylor, and Erik K. Flemington

Published in PLoS Pathog 9(5):e1003341. doi: 10.1371/journal.ppat.1003341. Epub 2013 May 9.

5.1 Abstract

Epstein-Barr virus (EBV) is associated with roughly 10% of gastric carcinomas worldwide (EBVaGC). Although previous investigations provide a strong link between EBV and gastric carcinomas, these studies were performed

using selected EBV gene probes. Using a cohort of gastric carcinoma RNA-seq data sets from The Cancer Genome Atlas (TCGA), we performed a quantitative and global assessment of EBV gene expression in gastric carcinomas and assessed EBV associated cellular pathway alterations. EBV transcripts were detected in 17% of samples but these samples varied significantly in EBV coverage depth. In four samples with the highest EBV coverage (hiEBVaGC – high EBV associated gastric carcinoma), transcripts from the BamHI A region comprised the majority of EBV reads. Expression of LMP2, and to a lesser extent, LMP1 were also observed as was evidence of abortive lytic replication. Analysis of cellular gene expression indicated significant immune cell infiltration and a predominant IFNG response in samples expressing high levels of EBV transcripts relative to samples expressing low or no EBV transcripts. Despite the apparent immune cell infiltration, high levels of the cytotoxic T-cell (CTL) and natural killer (NK) cell inhibitor, IDO1, was observed in the hiEBVaGCs samples suggesting an active tolerance inducing pathway in this subgroup. These results were confirmed in a separate cohort of 21 Vietnamese gastric carcinoma samples using qRT-PCR and on tissue samples using in situ hybridization and immunohistochemistry. Lastly, a panel of tumor suppressors and candidate oncogenes were expressed at lower levels in hiEBVaGC versus EBV-low and EBV-negative gastric cancers suggesting the direct regulation of tumor pathways by EBV.

5.2 Introduction

Epstein-Barr virus (EBV) is a herpes virus that infects most humans by adulthood. EBV is associated with several human malignancies, including malignancies of epithelial origin. The first report showing EBV's association with lymphoepithelioma-like carcinomas of the stomach was in 1990 by Burke and colleagues using polymerase chain reaction (PCR) [118]. Since that time, several studies have investigated the association between EBV and gastric carcinomas using a variety of methods (PCR, Southern blotting, and in situ hybridization (ISH)). In 1992, Shibata and Weiss reported EBV infection in 16% of gastric adenocarcinomas using PCR primers to the EBNA 1 gene and by ISH using probes against the EBV encoded small RNAs, EBERs [119]. Another report from Japan detected EBV in 6.9% of gastric carcinoma cases using EBER ISH [120]. Attributed to regional/country differences, the highest incidence of EBV-associated gastric carcinoma (EBVaGC) (16%) has been reported from the United States [119] while the lowest incidence (1.3%) is from Papua New Guinea [121]. Despite these landmark studies showing the association between gastric carcinomas and EBV, the mechanisms of EBV pathogenesis in gastric carcinoma are unclear.

Previous studies have shown the sensitivity of high throughout sequencing for detecting infectious agents [39, 41, 76] and for the new discovery of exogenous agents associating with human cancer [40, 41]. Merkel cell virus has been linked to Merkel carcinoma [40] and Fusobacterium has recently been associated with colorectal carcinoma [41]. In line with other reported methods for investigating pathogen associations in human cancers, we have previously developed a computational pipeline for the identification of exogenous sequences in RNA-seq data called PARSES [45]. Using PARSES, we examined two B-cell lines, Akata and JY, which are commonly used as model systems for EBV studies. Analysis of these cell lines revealed the presence of EBV in both cell lines as expected, but it also revealed the presence of the murine leukemia virus, MuLV in the JY but not Akata cell lines [76].

We have improved PARSES to include the utilization of parallel computing either on a local cluster or large-scale clusters, and we have included features that allow the user to simultaneously analyze the human cellular genes in addition to pathogen discovery (recently coined as ‘dual RNA-seq’ by Westermann and colleagues [43]). Here we utilized this pipeline, RNA CoMPASS (RNA comprehensive multi-processor analysis system for sequencing) [83], for the detection of viral pathogens in clinical tumor samples by analyzing a cohort of gastric carcinomas generated by the Cancer Genome Atlas initiative (SRA035410). EBV was detected in 12 out of 71 gastric carcinoma samples and the depth of coverage was sufficient to assess overall transcriptome structure in four cases. To our knowledge, this is the first study to globally assess both the EBV and host transcriptomes in gastric carcinomas using RNA-seq (although a recent paper has shed light onto this EBV specific host cell changes using a real time RT-PCR approach [122]). Our analysis led to insights into viral-host interactions and mechanisms through which EBV alters its local tumor environment. Further, this analysis revealed significant differences in the degree of host responses depending on the level of EBV gene expression. This raises the idea that the magnitude may be a more important clinical indicator than the simple detection of EBV in the selection of therapeutic regimens and the prediction of therapeutic responses in gastric carcinomas.

5.3 Materials and methods

5.3.1 Clinical tumor sample and sequence data acquisition

All human specimens were de-identified prior to acquisition. Total RNA from 21 Vietnamese gastric carcinoma samples and 5 normal adjacent samples were obtained from Biospecimen Repository at Bioserve (Beltsville, MD). RNA-seq data from 71 gastric carcinoma samples generated through the National Institutes of Health, The Cancer Genome Atlas (TCGA) project were obtained from the NCBI Sequence Read Archive (SRA035410, now available through the Cancer Genomics Hub managed by the University of California, Santa Cruz (UCSC)). Demographic and clinical data for each sample is available through the TCGA data portal (http://cancergenome.nih.gov/). Briefly, samples were obtained from non-Hispanic White Russians with no previous treatment. The mean age was 69 years with a range of 43 to 90 years. Total RNA was isolated from each sample using mirVana RNA kit according to TCGA. High quality RNA was polyA selected and sequenced using an Illumina Genome Analyzer II machine running paired end 51 base sequencing reactions with two lanes of sequence per sample.

5.3.2 Cell Culture

SNU-719 gastric adenocarcinoma cells were obtained from the Korean Cell Line Bank. They were grown in RPMI 1640 (Thermo Scientific; Waltham, MA) plus 10% fetal bovine serum (Invitrogen-Gibco; Grand Island, NY) with 1% penicillin-streptomycin (Invitrogen-Gibco; Grand Island, NY). Cells were grown at 37ºC in a humidified, 5% CO_2 incubator.

5.3.3 Sample preparation and next generation DNA sequencing

Total RNA was extracted from SNU-719 cells using the miRNeasy Mini Kit (Qiagen, Hilden, Germany) according to manufacturer’s instructions. Two separate cDNA libraries were prepared from polyA selected and from Ribo-Zero selected RNAs using the Illumina Truseq Stranded Total RNA Sample Prep Kit (RS-122-2101). 101-base paired-end sequencing was performed using an Illumina HiSeq 2000 instrument. The SNU-719 RNA-seq data used in this publication have been deposited in NCBI's Gene Expression Omnibus [123] and are accessible through GEO Series accession number GSE45453 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45453).

5.3.4 RNA CoMPASS

RNA CoMPASS (RNA comprehensive multi-processor analysis system for sequencing) is a graphical user interface (GUI) based parallel computation pipeline for the analysis of both exogenous and human sequences from RNA-seq data [83] (Figure 1 in Chapter 2). Briefly, for the analysis of both exogenous and human sequences, raw sequence data is first processed through an in house de-duplication algorithm. Following de-duplication, reads are aligned to a reference genome containing human (hg19; UCSC) and abundant sequences (which include sequence adapters, mitochondrial, ribosomal, enterobacteria phage phiX174, poly-A, and poly-C sequences). Novoalign V2.07.18 (www.novocraft.com) [-o SAM, default options] is used to map reads to the reference genome and to eliminate low-quality reads (QC < 20). TopHat V1.4.0 [default options] [55] is used to identify and isolate all sequences that map to human splice junctions. The results from these programs are compiled and separated into mapped reads (used for human transcriptome analysis) and unmapped reads (used for exogenous sequence analysis). Human mapped reads are analyzed using SAMMate [51] to quantify gene expression and to generate genome coverage information. Unmapped reads are subjected to consecutive BLAST V2.2.24 searches against the Human RefSeq RNA database (an additional “pre-clearing” step) and then to the NCBI NT database to identify reads corresponding to known exogenous organisms [52]. Results from the NT BLAST searches are filtered to eliminate matches with an E-value of greater than $10e^{-6}$. The results are fed into MEGAN 4 [53] for convenient visualization and taxonomic classification of BLAST search results.

RNA CoMPASS is designed to take advantage of parallel processing at several key steps to speed processing times. In our case, we used a four node, 12 core, Intel Xeon Mac Pro (64GB of memory per node) cluster.

5.3.5 Human and EBV Transcriptome quantification

Samples containing evidence of EBV were identified using RNA CoMPASS. Since each sample contained sequence data from two runs, data from both runs were combined in order to generate a greater sequencing depth for transcript quantification. In addition, 20 EBV negative samples were randomly chosen for analysis. Samples were aligned to a reference genome containing human (hg19) and a modified EBV B95-8 genome that contains Raji genome sequences inserted into a deleted region of the B95-8 genome (Genbank accession number AJ507799) using Novoalign V2.07.18 (www.novocraft.com) [-o SAM, default options]. Transcript data from Novoalign was analyzed using SAMMate for transcript quantification of human and EBV genes and to generate coverage (wiggle) files for visualization of read distributions. Splice junction data was generated using the junction aligner, TopHat V1.4.0 [default options]. Coverage data was visualized using the Integrative Genomics Viewer (IGV) [87].

5.3.6 Quantitative RT-PCR

Total RNA was reverse-transcribed using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA). Random hexamers were used along with 250 ng RNA in a 20µl reaction volume according to manufacturer’s instructions. For the incubation steps (25˚C for 10 min followed by 50˚C for 50 min) a Mastercycler ep (Eppendorf, Hamburg, Germany) was used. For real-time PCR, 1µl of the resulting cDNA was used in a 10µl reaction mixture that included 5µl of 10x SsoFast EvaGreen supermix (Bio-Rad, Hercules, CA), 1µl of 10µM forward and reverse primer mix (Integrated DNA Technologies, Coralville, IA), and 3µl of PCR grade water. The IDO1 primers amplified a 112 base pair product. Forward primer 5'-CAAATCCACGATCATGTGAACC-3' and reverse primer 5'-AGAACCCTTCATACACCAGAC-3' were used previously by Prachason et al [124]. The RPMS1 primers amplified a 181 base pair product consisting of exon 6 and exon 7. Forward primer 5'-CCAGGTCAAAGACGTTGGAG-3' and reverse primer 5'-CACCACGGTGCAGCCTAC-3' were used. The GAPDH primers amplified a 297 base pair product. Forward primer 5'- CAATGACCCCTTCATTGACC-3' and reverse primer 5'- GACAAGCTTCCCGTTCTCAG-3' were used. Each sample was performed in triplicates. No-template controls and no-reverse transcription controls were also included in each PCR run. Thermal cycling was performed on a CFX96 Real Time System (Bio-Rad, Hercules, CA) and data analysis was performed using the CFX Manager 3.0 software. Cycling conditions included an initial incubation at 95°C for 30 seconds followed by 40 cycles consisting of 95°C for 5 seconds, and 60°C for 5 seconds. Melting curve analysis was performed at the end of every qRT-PCR run.

5.3.7 In-situ hybridization

Chromogenic In Situ Hybridization (CISH) was performed by the Tulane Molecular Pathology Lab using the HistoSonda EBER probe kit (American Master Tech, Lodi, CA) according to manufacture's instructions. The tissue array was deparaffinized and rehydrated in a graded solution of Xylene and alcohol.

Tissue array was deproteinized using Proteinase K and incubated with Digoxigenin EBER probe. Tissue array was subsequently washed with water and PBS. The tissue array was incubated with Anti-digoxin and anti-mouse horseradish peroxidase to form a duplex with the Digoxigenin EBER probe. For colorimetric staining, slides were then incubated in 3,3'-Diaminobenzidine (DAB; Vector Laboratories), washed with dH_2O, counterstained with hematoxylin, and rinsed with PBS (pH 7.4). Slides were dehydrated in a graded solution of alcohol and Xylene and sealed with Permount Mounting Medium (Sigma). To visualize the tissue array, slides were scanned into ScanScope CS2 (Aperio, Vista, CA) and images were acquired with ImageScope (Aperio).

5.3.8 Immunohistochemistry

Formalin-fixed, paraffin-embedded (FFPE) gastric tumor tissue array (ST2901) was purchased from U.S. BioMax (Rockville, MD). Demographic and clinical data can be found on the U.S. BioMax website (http://www.biomax.us/tissue-arrays/Stomach/ST2091). The tissue array was deparaffinized, and rehydrated in a graded solution of Sub-X clearing medium (Leica Biosystems, Buffalo Grove, IL). Antigen retrieval was performed with Tris-EDTA Buffer, consisting of 10mM Tris Base, 1mM EDTA Solution, and 0.5% Tween 20 (pH 9.0), for 30 minutes. The tissue array was then quenched with 3% H_2O_2 (Sigma), rinsed with TNT washing buffer made of 0.1M Tris-HCl, 0.15M NaCl, and 0.5% Tween-20 (pH 7.5), blocked with blocking reagent purchased from Perkin Elmer (Waltham, MA) and stained with goat-anti-human IDO

(Abcam, Cambridge, MA) overnight at 4^{o}C. Tumor sections were subsequently washed in TNT, incubated with donkey-anti-goat HRP conjugated secondary antibody (Santa Cruz, Dallas, TX) for 1 hour at room temperature, and washed with TNT. For colorimetric staining, slides were then incubated in 3,3'-Diaminobenzidine (DAB; Vector Laboratories), washed with dH_2O, counterstained with hematoxylin, and rinsed with PBS (pH 7.4). Slides were dehydrated in a graded solution of Sub-X clearing medium and sealed with Permount Mounting Medium (Sigma). To visualize the tissue array, slides were scanned into ScanScope CS2 (Aperio, Vista, CA) and images were acquired with ImageScope (Aperio).

5.3.9 Statistics and Cluster analysis

Transcript counts were imported into the R software environment and analyzed using the edgeR package [54]. Genes with low transcript counts (less than 1 CPM (count per million)) in the majority of samples were filtered. The Manhattan (L-1) distance matrix for the samples was computed using the remaining transcript counts, and this was taken as input for hierarchical clustering using the Ward algorithm. The well separated cluster of four EBV positive samples was found to be those with the highest numbers of EBV reads and was classified as “high EBV”. The remaining samples were then classified as “EBV-negative” or “low EBV”. The glmFit function was then used to fit the mean log(CPM) for each group and likelihood ratio tests were used to identify those genes that were differentially expressed in any of the three possible comparisons, with adjusted $p<0.05$ following the Benjamini-Hochberg correction for multiple testing. The fitted log(CPM) values for the subset of genes that were differentially expressed in the high EBV samples relative to both the low EBV and EBV-negative samples were then clustered using the Euclidean distance and complete linkage algorithm to detect groups of co-expressed genes. The EBV transcript counts from all positive samples were imported in MeV [86] for hierarchical clustering using the Manhattan distance matrix and average linkage clustering algorithm.

5.4 Results

5.4.1 Detection of EBV in gastric adenocarcinoma samples using RNA CoMPASS

RNA-seq data from The Cancer Genome Atlas (TCGA) gastric adenocarcinoma cohort (SRA035410) was first analyzed using RNA CoMPASS [83] to assess the virome for each of the 71 data sets. This initial screening was performed using a single lane of sequencing data from each patient. Most samples contained relatively low numbers of reads matching non-human viral sources (e.g. enterobacteria phage T4T) that possibly represent environmental contamination (Figure 1A-B). Of the known human viruses detected, one sample (BR-4298, Figure 1A) contained 6 reads attributed to Hepatitis C virus. Further inspection of these reads showed high homology to the human immunoglobulin light chain variable region. These reads likely represent human sequences rather than reads derived from Hepatitis C virus. Twelve samples showed evidence of human cytomegalovirus (HCMV) with read numbers ranging from 5 to 132. Individual BLASTing of selected HCMV reads showed high homology to HCMV genomes but not to human sequences indicating that these are bona fide HCMV derived reads. The relatively low numbers of HCMV reads in these samples (relative to the numbers of EBV in some samples, see below) suggests that these reads are derived from a low number of HCMV infected cells or that the virus is not expressing substantial numbers of polyadenylated RNAs in these tumor samples.

EBV was detected in 12 out of the 71 (17%) gastric carcinoma cases with varying levels of reads. To further analyze the EBV-associated gastric carcinoma (EBVaGC) samples, the two lanes of sequence per sample were combined to attain greater sequencing depth. These sequence files were aligned against a modified EBV B95-8 genome that contains Raji genome sequences inserted into a deleted region of the B95-8 genome (Genbank accession number AJ507799) plus the hg19 assembly of the human genome. Alignments were carried out using Novoalign V2.07.18 [-o SAM, paired-end, default options]. Based on the assembly to the human genome, sample quality and throughput was found to be consistent across all samples with the numbers of human mapped reads ranging from 128 to 159 million. Eight of the 12 EBV positive samples were found to have less than 200 reads per sample (inset of Figure 1C), three were found to contain more than 30,000 reads and one sample was found to contain 1,194 reads (Figure 1C). We tentatively considered the 8 cases with less than 200 reads to represent nominal infections similar to that observed with CMV (above). The 4 samples with higher read numbers, BR-4253, BR-4271, BR-4376, and BR-4298, were taken for more in depth transcriptome analysis.

Notably, while three of the four EBV positive samples with high numbers of EBV reads were classified as the more common Type I strain of EBV, one of these samples, BR-4253, was classified as the type II strain (Figure 1A). Since the strain defining regions of EBV, EBNA2 and EBNA3A/3B/3C [125] are largely not expressed in EBVaGC, we were concerned that the reads from sample BR-4253 could be misclassified as type II. We analyzed a few of the reads defined by MEGAN as type II from sample BR-4253 using manual BLAST and the majority of reads aligned to both type I (B95-8/Raji) and type II strains (AG876 (Genbank accession number DQ279927)) with some of these showing better homology to AG876 (data not shown). Despite this, the small number of reads derived from the EBNA2 and EBNA3A/3B/3C loci were more homologous to the type I than the type II strain. Therefore, this sample was likely misclassified as the type II strain because of greater similarity to the AG876 genome at highly expressed regions outside of the EBNA2 and EBNA3A/3B/3C loci.

5.4.2 EBV gene expression in gastric carcinomas

EBV transcript quantification and genome coverage information was generated for samples, BR-4253, BR-4271, BR-4376, and BR-4298 using the transcriptome analysis software, SAMMate (note that the sequencing libraries were generated from polyA selected RNA which precludes the sequencing of EBER genes) [51]. Genome coverage information was first visualized by

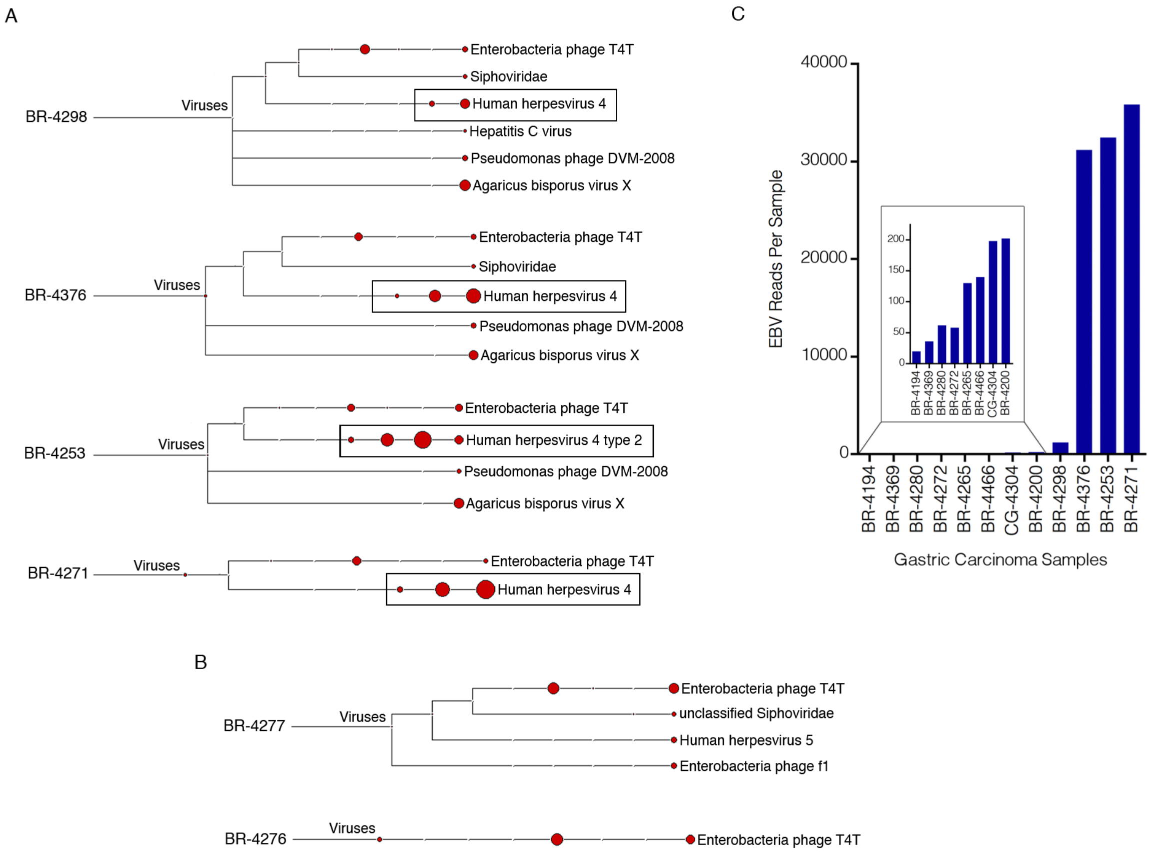


Figure 1. Detection of EBV in gastric carcinoma samples. Four gigabytes of deduplicated RNA-seq read data from each of the seventy-one gastric carcinoma samples were analyzed using RNA CoMPASS. The virome branch of the taxonomy trees for the four samples with the highest number of EBV reads (A) and two EBV-negative samples (B) were generated using the metagenome analysis tool, MEGAN 4. (C) For a more in depth analysis of EBV reads, the combined sequence read files for each sample were aligned to the EBV genome and the hg19 human genome assembly using the genome aligner, Novoalign. Of the EBV-positive samples, four samples were identified as having high numbers of EBV reads while eight were found to have low but detectable numbers of EBV reads.

displaying the number of reads across each genomic position in the Circos plot shown in Figure 2A. Because of disparate coverage intensities, the Circos graph in Figure 2A is plotted in log scale to allow simultaneous visualization of the less abundantly expressed regions of the genome. Notably, coverage across the BamHI A region was high relative to other parts of the genome with greater than 96% of total reads corresponding to the BamHI A region in each case (Figure 2B).

Evidence for transcription of the essential episomal replication factor, EBNA1 is observed in samples BR-4253, BR-4271 and BR-4376 (Figure 2A (upper left region of the figure) and Figure 2B). No EBNA1 reads were detected in BR-4298 most likely owing to the significantly lower read numbers in this sample (Figures 2A-B).

Evidence for transcription of the immediate early genes, BZLF1 and BRLF1, is similarly seen in BR-4253, BR-4271 and BR-4376 but not in BR-4298 (again, possibly due to the low overall read numbers). Despite the detection of BZLF1 and BRLF1 reads in samples BR-4253, BR-4271 and BR-4376, there is a remarkable absence of reads for most other downstream lytic genes in these samples. In Figure 3, we plotted the ratio of lytic gene transcripts (*sans* lytic genes in the BamHI A region) relative to the level of BZLF1 RNAs in BR-4253, BR-4271 and BR-4376 and compared this to the corresponding relative levels of these gene transcripts in reactivating Akata cells [49]. This comparison indicates that while the BZLF1 and BRLF1 immediate early genes are expressed in these

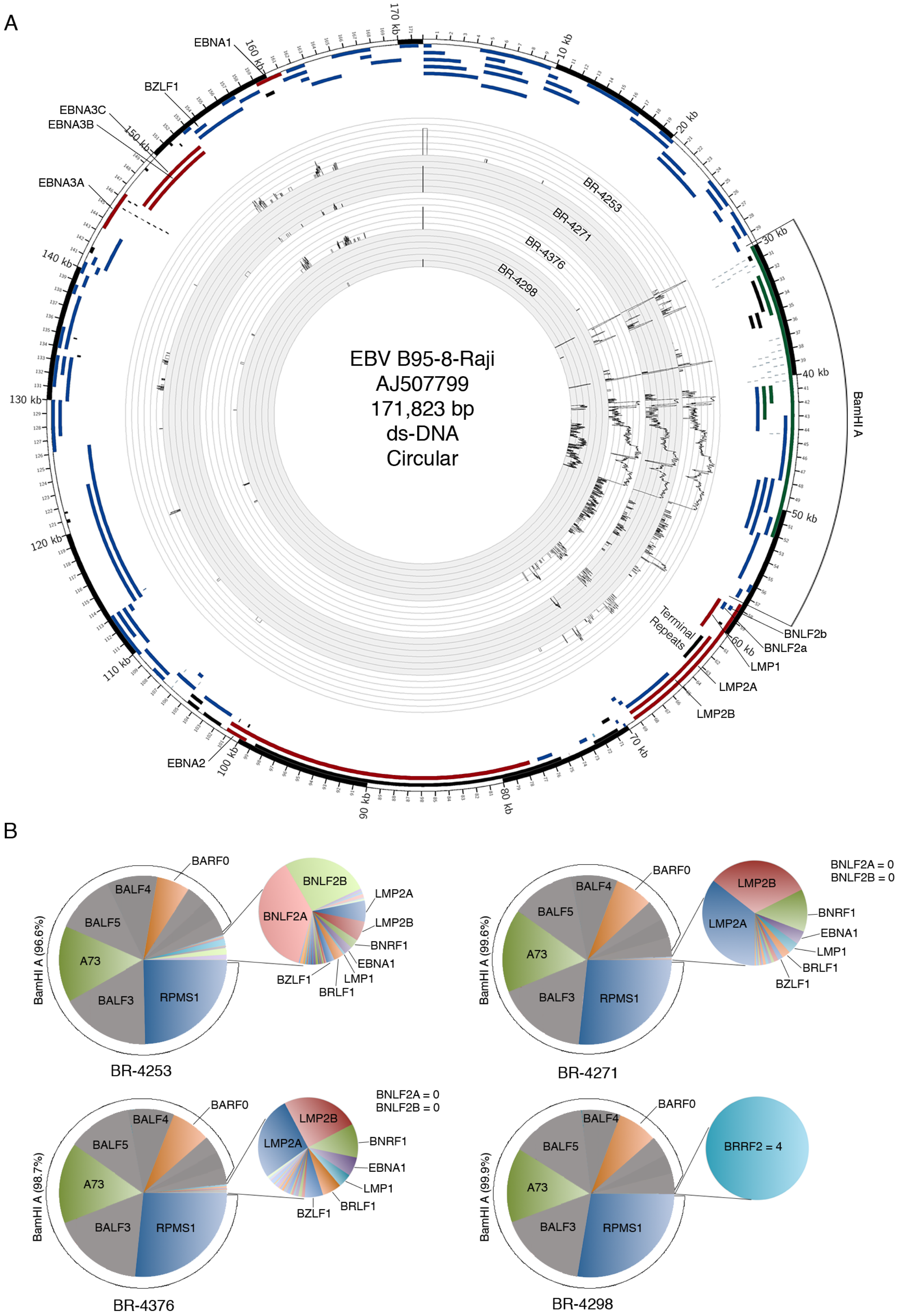

A
EBV B95-8-Raji
AJ507799
171,823 bp
ds-DNA
Circular
BR-4253
BR-4271
BR-4376
BR-4298
EBNA1
BZLF1
EBNA3C
EBNA3B
EBNA3A
EBNA2
Terminal Repeats
BNLF2b
BNLF2a
LMP1
LMP2A
LMP2B
BamHI A
10 kb
20 kb
30 kb
40 kb
50 kb
60 kb
70 kb
80 kb
90 kb
100 kb
110 kb
120 kb
130 kb
140 kb
150 kb
160 kb
170 kb
B
BamHI A (96.6%)
BARF0
BALF4
BALF5
A73
BALF3
RPMS1
BNLF2B
BNLF2A
LMP2A
LMP2B
BNRF1
EBNA1
LMP1
BRLF1
BZLF1
BR-4253
BamHI A (99.6%)
BARF0
BALF4
BALF5
A73
BALF3
RPMS1
BNLF2A = 0
BNLF2B = 0
LMP2B
LMP2A
BNRF1
EBNA1
LMP1
BRLF1
BZLF1
BR-4271
BamHI A (98.7%)
BARF0
BALF4
BALF5
A73
BALF3
RPMS1
BNLF2A = 0
BNLF2B = 0
LMP2B
LMP2A
BNRF1
EBNA1
LMP1
BRLF1
BZLF1
BR-4376
BamHI A (99.9%)
BARF0
BALF4
BALF5
A73
BALF3
RPMS1
BRRF2 = 4
BR-4298

Figure 2. Genome wide analysis of EBV gene expression. (A) An annotated Circos plot depicting EBV read coverage across the EBV genome. Coverage graphs display the number of reads mapping to each nucleotide position of the genome and are depicted in log scale. Note that alignments were performed using a genome that was split between the BBLF2/3 and the BGLF3.5 lytic genes rather than at the terminal repeats to accommodate coverage of splice junctions for the latency membrane protein, LMP2. The natural termini of the linear genome, the terminal repeats, are shown in the lower right quadrant of the graph. Coverage data is plotted relative to the modified B95-8 genome containing Raji genome sequences (Genbank accession number AJ507799). Blue features represent lytic genes, red features represent latency genes, green features represent potential non-coding genes, aquamarine features represent microRNAs, and black features represent non-gene features (repeat regions and origins of replication, for example). (B) Pie charts displaying read counts across EBV gene features. Because the BNLF2a/b region is contained within the LMP1 gene, total LMP1 read counts were inferred by determining the counts within the unique LMP1 sequences, multiplying by the total length of LMP1, and dividing by the length of the unique region. BNLF2a/b counts were calculated by determining the number of reads within the BNLF2a/b locus and subtracting the inferred number of LMP1 reads derived from within the BNLF2a/b coordinates (i.e. number of LMP1 reads within the unique region times the length of the overlap region divided by the length of the unique region). Leftward oriented genes within the BamHI A region are shown in grey. This representation indicates uncertainty due to the finding of primarily rightward transcription across these genes in the gastric carcinoma cell line SNU-719 using directional sequencing methods.

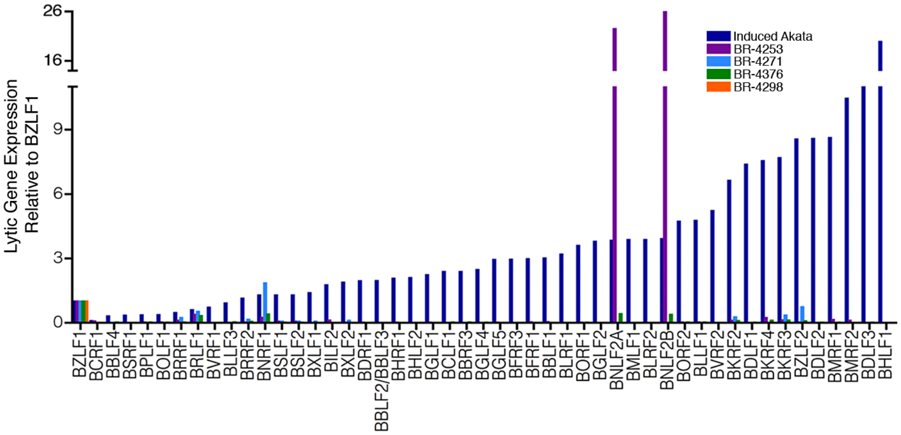


Figure 3. Abortive lytic gene expression. EBV lytic gene expression in EBVaGC samples. Lytic gene expression relative to BZLF1 represents RPKMs (reads per kilobase of exon model per million mapped reads) for each indicated gene divided by the RPKMs of BZLF1 for the respective biological sample. For reference to a productive replication setting, samples were compared to the lytic gene expression profile in reactivated Akata cells.

tumors, there is a clear lack of lytic cycle progression; reflecting abortive lytic replication in this *in vivo* setting.

Consistent with previous reports of LMP2 expression in gastric carcinomas [126, 127], we similarly see evidence of LMP2 transcription in samples BR-4253, BR-4271 and BR-4376 (Figures 2A-B and Figure 4A). LMP1 has been previously reported to be expressed at low levels or to be not expressed in gastric carcinomas [128-130]. We similarly find low albeit detectable levels of LMP1 in BR-4253, BR-4271 and BR-4376 (Figures 2B and 4A). Strikingly, however, sample BR-4253 has a very high number of reads corresponding to the early BNLF2A/B locus, which overlaps the LMP1 3' untranslated region (Figures 2B, 3 and 4A). No BNLF2A/B reads are detected in BR-4271, BR-4376, and BR-4298 (Figure 2B) suggesting that this is unique to BR-4253. The high expression level of the early BNLF2A/B genes in BR-4253 is surprising because it occurs in the absence of most other early genes. This suggests the possibility that BNLF2A/B is expressed in this patient through an alternative mechanism possibly mediated through a viral genetic alteration.

5.4.3 Analysis of the highly expressed BamHI A region

The most actively polyA transcribed region of the EBV genome, the BamHI A region (Figures 2A-B), shows excellent coverage across most of the RPMS1/A73 exons with apparent additional coverage observed for the regions spanning the leftward transcribed genes, BALF5, BALF3, and BALF4 (Figure 4B). Coverage across these leftward genes is unexpected because they are

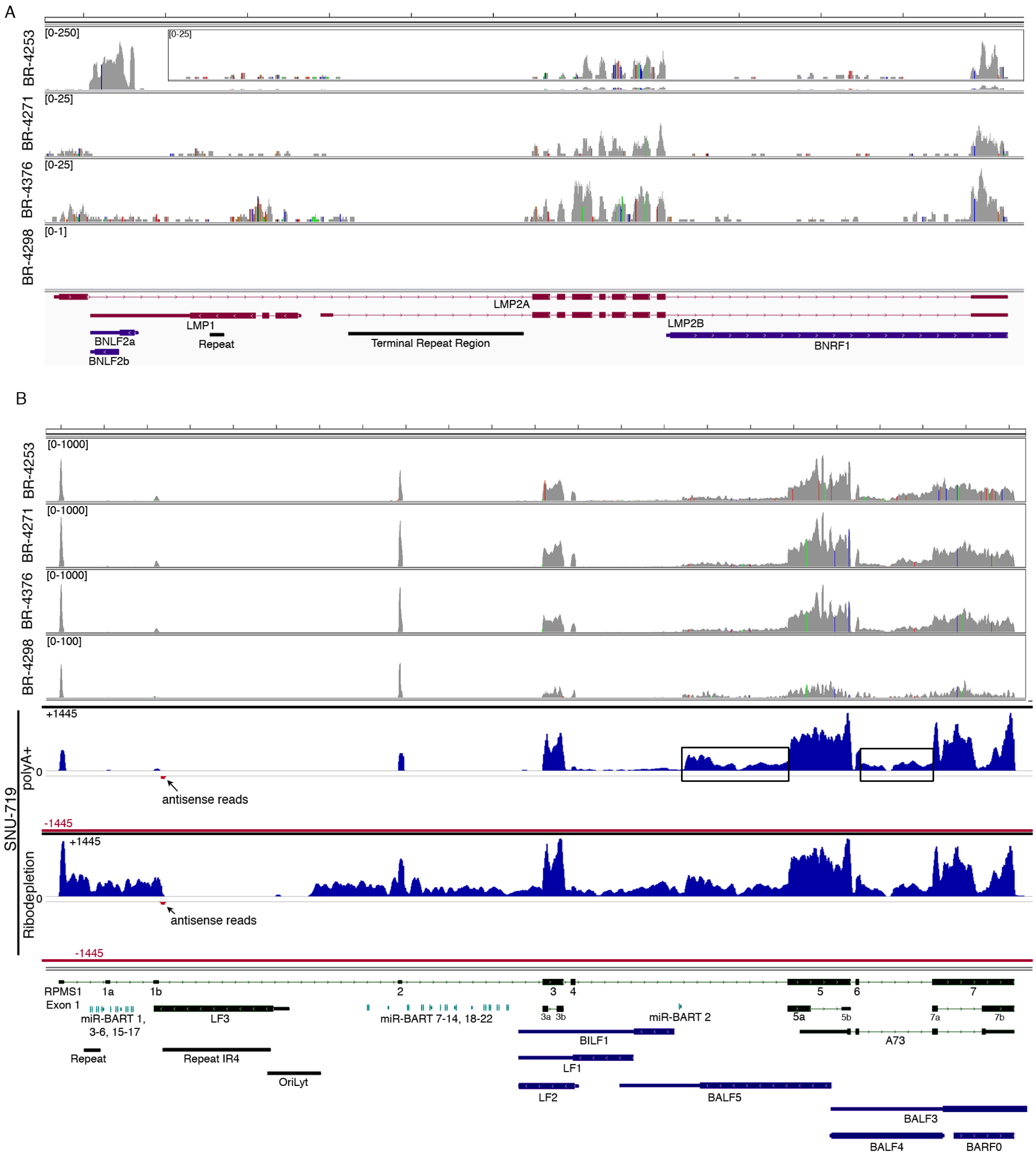

A
BR-4253
[0-250]
[0-25]
BR-4271
[0-25]
BR-4376
[0-25]
BR-4298
[0-1]
LMP2A
LMP1
LMP2B
BNLF2a
BNLF2b
Repeat
Terminal Repeat Region
BNRF1
B
BR-4253
[0-1000]
BR-4271
[0-1000]
BR-4376
[0-1000]
BR-4298
[0-100]
SNU-719
polyA+
+1445
0
antisense reads
-1445
Ribodepletion
+1445
0
antisense reads
-1445
RPMS1
Exon 1
1a
1b
2
3
4
5
6
7
miR-BART 1,
3-6, 15-17
LF3
miR-BART 7-14, 18-22
3a
3b
miR-BART 2
5a
5b
7a
7b
BILF1
A73
Repeat
Repeat IR4
LF1
OriLyt
LF2
BALF5
BALF3
BALF4
BARF0

Figure 4. EBV gene expression analysis. Detailed read coverage data for the LMP2a, LMP1, and BNLF2a/b genes (A) and the RPMS1/BamHI A regions (B) of the EBV genome. Data was displayed using the Integrative Genomics Viewer (IGV) using the modified B95-8 genome containing Raji genome sequences (Genbank accession number AJ507799). The y-axis represents the number of reads at each nucleotide position of the genome. Blue features represent lytic genes, red features represent latency genes, green features represent potential non-coding genes, aquamarine features represent microRNAs, and black features represent non-gene features (repeat regions and origins of replication, for example). In panel (A), coverage graphs for BR-4253 is scaled to a maximum read level of 250 reads (the BR-4253 inset displays the data with a max read level of 25), the BR-4271 and BR-4376 graphs are scaled to a max read level of 25, while the max read level for BR-4298 is 1. For coverage across the RPMS1/BamHI A region (B), BR-4253, BR-4271, and BR-4376 are scaled to 1,000 reads, while BR-4298 is scaled to 100. Strand specific sequencing from SNU-719 cells of the RPMS1/BamHI A region is also displayed. The top 2 tracks are from poly(A) selected RNA and the bottom 2 tracks are from Ribo-Zero depleted RNA. The read coverage for the sense strand is displayed in blue with positive values while the antisense strand is displayed in red with negative values. The scale is + or - 1,445 reads for the sense and antisense strands.

thought to be lytic genes and not expressed during latency. We therefore performed directional sequencing of a naturally occurring EBV positive gastric carcinoma cell line, SNU-719, to allow us to determine the orientation of transcripts across this region. EBV read coverage for SNU-719 was remarkably similar to that observed for the tumor specimens (Figure 4B). Outside of a small blip of leftward transcription noted near the RPMS1 exon 1b, there is little leftward transcription across this region. This indicates that the transcription observed across this region in the tumor specimens are likely rightward oriented and to a large extent related to RPMS1 and/or A73 but not BALF5, BALF3, BALF4, BILF1, LF1, or LF2.

Also notable in Figure 4B is rightward coverage across the introns between exons 4 and 5 and exons 6 and 7 of the RPMS1 gene (boxed regions in SNU-719 tracks). This coverage likely does not represent intron fragments generated after transcript splicing because this coverage is observed in sequencing libraries generated from polyA selected RNA (upper SNU-719 tracks). In contrast, there is no coverage of the first 4 RPMS1 introns on the polyA track whereas there is substantial coverage across these regions when ribo-depleted RNA was used for sequencing (Figure 4B). Therefore, the rightward coverage between exons 4 and 5 and between exons 6 and 7 likely represent bona fide previously unannotated rightward exons/transcripts. The read coverage between exons 6 and 7 may arise from mature RPMS1 isoforms that retain this intron (forming a unique RPMS1 isoform). The coverage between exons 4 and 5 starts near the middle of this intron suggesting that this is a site of transcription initiation or a that it is a splice acceptor site. Since splice mapping (see below) did not identify candidate splicing events near the beginning of this intron coverage, it is possible that this coverage arises from transcription initiation from an unknown upstream promoter.

As mentioned above, more than 96% of all EBV reads align to the BamHI A region. Further, RPMS1 exon coverage ranks within the top seven percent of expressed cellular genes in samples BR-4253, BR-4271, and BR-4376 with expression that is more than five times the median cellular gene expression level (Figure 5). We conclude that not only is expression of this region high relative to other EBV encoded genes, but the expression is also high relative to cellular genes. In contrast, it is notable that the LF3 gene which is within the BamHI A locus and which has been found to be expressed at very high levels in other systems [131], shows no evidence of expression in these *in vivo* gastric carcinoma tumor datasets.

To assess splicing events in this region, alignments were performed using the junction mapper, TopHat [55]. Consideration of the most abundant splice junction reads indicates the predominance of sequential splicing from exons 1-2-3-4-5-6-7 (Figure 6). Nevertheless, there is significant evidence of intra-exonal splicing at exons 3 (3a to 3b), 5 (5a to 5b), and 7 (7a to 7b) (Figure 6). Although splicing from exons 1 to 2 is the most predominant 5' region splicing order, there is also good evidence of alternative splicing to exon 1a (i.e. splicing of exon 1 to exon 1a to exon 2) (Figure 6). In samples BR-4253 (Figure 6) and SNU-719 (data not shown), we also noted evidence of splicing initiating from the middle of

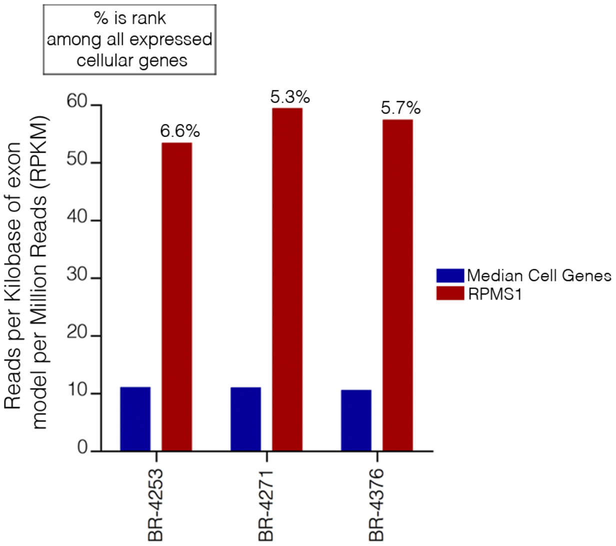


Figure 5. EBV transcripts from RPMS1 are among the highest expressed genes in EBVaGCs. RPKM values calculated using reads across all RPMS1 exons are shown with respect to the median expression of all expressed cellular genes (expressed genes defined as cellular genes with greater than 1 RPKM). The percentage values above each RPMS1 bar represents the rank of RPMS1 expression in the respective sample among all expressed cellular genes in that sample.

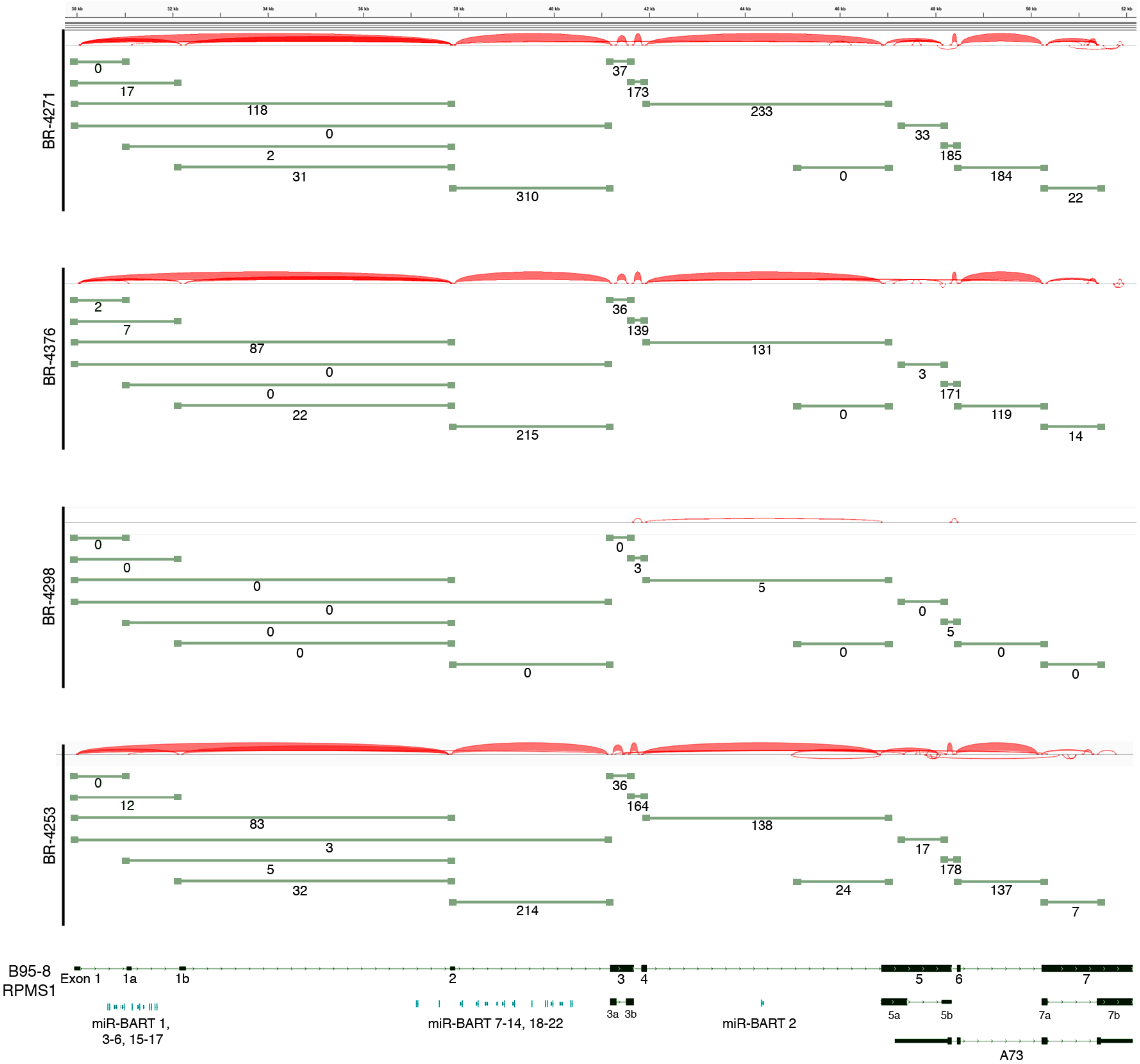


Figure 6. Alternative splicing in the EBV BamH1 A region in EBVaGCs. RNA-seq data from BR-4271, BR-4376, and BR-4298 and BR-4253 was analyzed using the TopHat aligner to obtain splice junction information. Samples with the type I strain of EBV, BR-4271, BR-4376, BR-4298, and BR-4253 were aligned to the type I genome, B95-8/Raji (Genbank accession number AJ507799). Junctions were visualized using Integrative Genomic Viewer (IGV) [87]. Thickness of red junction features correlates with the number of reads for the respective junction. The number of junction spanning reads for each junction is indicated below each olive green junction feature.

the newly identified coverage in the intron between exons 4 to 5 to the start of exon 5. This indicates additional complexity in this new region whereby some of these transcripts splice to exon 5 while some read through to exon 5.

5.4.4 Differential cellular gene expression patterns in EBV associated and EBV low/negative gastric adenocarcinomas

EBV likely contributes to gastric carcinoma through the subversion of at least some of the oncogenic pathways required for the development of gastric carcinoma. However, the way that EBV subverts these pathways is likely distinct from the mechanism of pathway disruption in the absence of EBV (e.g. through genetic alterations). Since cellular gene expression is typically responsive to altered signaling mechanisms, differences in gene expression profiles can be used to not only classify cell populations but also infer upstream signaling events within certain cell populations.

To investigate influences of EBV dependent alterations in tumor signaling pathways, we analyzed global cellular gene expression in all 12 EBV positive specimens plus an additional 20 randomly selected EBV negative samples. EBV gene expression data was not included in this analysis to ensure that clustering occurred based only on differences in cellular gene expression (i.e. that it occurred independently of biases incurred by the presence of EBV gene expression signatures). Strikingly, when the set of samples were analyzed using hierarchical clustering, the four gastric carcinoma samples with higher numbers of EBV reads (BR-4253, BR-4271, BR-4376, and BR-4298) formed its own well-separated group (Figure 7A). One of the EBV negative samples, BR-4294, clustered independently of the others and subsequent analysis revealed that this sample was likely an outlier. Nevertheless, this sample was retained in the subsequent differential expression analysis as a conservative measure.

Human transcript counts from the EBVaGCs with high EBV read levels were compared to the EBVaGCs with low EBV read numbers and with the EBVnGCs. Using this approach, 490 genes were found to have statistically significant differential expression in the “high” EBVaGC (hiEBVaGC) samples relative to both EBVnGC and “low” EBVaGCs (loEBVaGC) samples (Figure 7A-B). These genes separated into five distinct clusters with clusters 1, 3, and 5 showing genes that were predominately expressed at higher levels in hiEBVaGCs and clusters 2 and 4 containing genes that were predominantly expressed at lower levels in hiEBVaGCs (Figure 7A). We also performed an additional clustering analysis using only the EBV genes across the 12 EBVaGC. This analysis revealed that the 4 hiEBVaGC samples cluster distinctly from the other EBVaGC samples (Figure 8). This apparently distinct gene expression pattern observed in the 4 hiEBVaGC samples raises the possibility that these samples represent infection of a unique cell type relative to the other samples (possibly tumor cells versus stroma or B-cells).

Ingenuity Pathway Analysis software (IPA: Ingenuity Systems) was used to assist the analysis of pathways and known molecular functions associated with differentially expressed genes. Twenty four percent (116) of the 490 genes with statistically significant differential expression were found to be immunologically

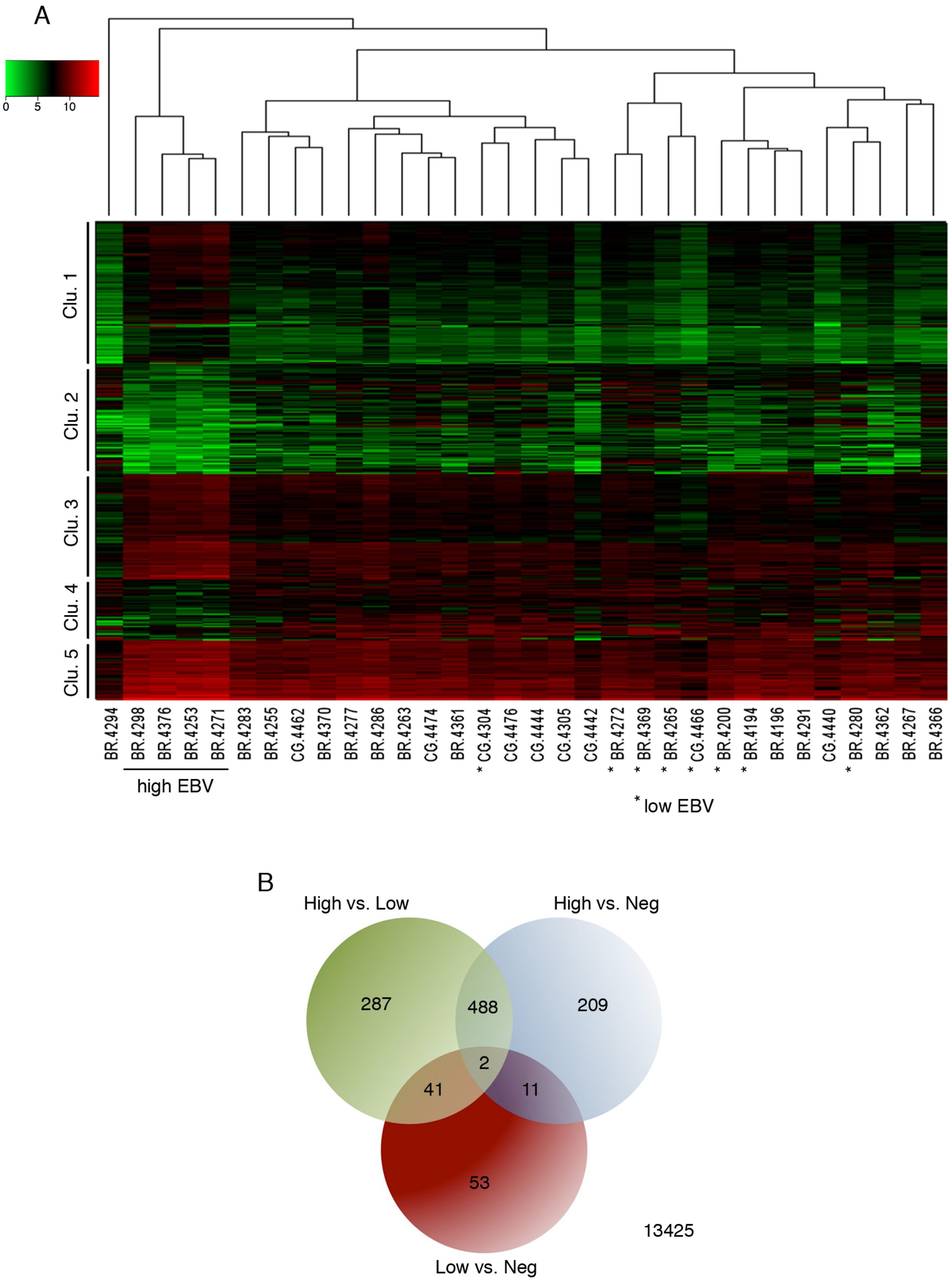

A
0
5
10
Clu. 1
Clu. 2
Clu. 3
Clu. 4
Clu. 5
BR.4294
BR.4298
BR.4376
BR.4253
BR.4271
BR.4283
BR.4255
CG.4462
BR.4370
BR.4277
BR.4286
BR.4263
CG.4474
BR.4361
* CG.4304
CG.4476
CG.4444
CG.4305
CG.4442
* BR.4272
* BR.4369
* BR.4265
* CG.4466
* BR.4200
* BR.4194
BR.4196
BR.4291
CG.4440
* BR.4280
BR.4362
BR.4267
BR.4366
high EBV
* low EBV
B
High vs. Low
High vs. Neg
Low vs. Neg
287
488
209
2
41
11
53
13425

Figure 7. Cluster analysis of EBV-associated gastric carcinoma samples. (A) A representative cohort of 32 gastric carcinoma samples (12 EBV-positive and 20 EBV-negative) were grouped using hierarchical clustering and are displayed with an expression heat map of the 490 genes that were found to be significantly differentially expressed in high EBV. (B) The cohort of 32 gastric carcinoma samples was divided into three categories (high EBV, low EBV, and negative). These categories were subjected to differential gene expression analysis using edgeR. The Venn diagram displays the numbers of all statistically significant differentially expressed genes. Statistical significance was determined by an adjusted *P* value < 0.05.

related genes (Figure 9A). The vast majority of these genes were expressed at higher levels in hiEBVaGCs with IDO1 and IFNG ranking among the top (38-fold and 16-fold, high v. negative). The differentiation and other cell surface marker profiles are consistent with the presence of cytotoxic T-cells (CTLs) and/or natural killer (NK) cells in hiEBVaGC. Further, CTLs and NK cells are key producers of granzymes and perforin, which are found to be elevated in the hiEBVaGC (Figure 9A).

The interferon gamma (IFNG) pathway was analyzed further using IPA to determine the extent of IFNG pathway involvement in hiEBVaGC. We observed marked involvement of the IFNG pathway with 156 of the 490 differentially expressed genes associated with the IFNG pathway, the majority of which were elevated (Figure 9B).

The analysis of IDO1 levels for each of the 32 gastric carcinomas showed that the samples with the highest number of EBV reads had the highest levels of IDO1 expression (Figure 10A). To further explore the link between EBV and IDO1, we analyzed a separate cohort of Vietnamese gastric carcinoma samples by real time RT-PCR. RPMS1 was detected in two of these samples (CZRDPREA and WZQ1TALM) (Figure 10B) and these samples ranked among the highest for expression of IDO1 (27 and 17 fold relative to the average of the 5 normal adjacent tissue samples). Further, in these samples, normal adjacent tissue showed lower RPMS1 expression and lower IDO1 expression compared to their tumor counterparts. Notably, one of the EBV negative samples, W31AB410, showed the highest level of IDO1 (43 fold). Nevertheless, this

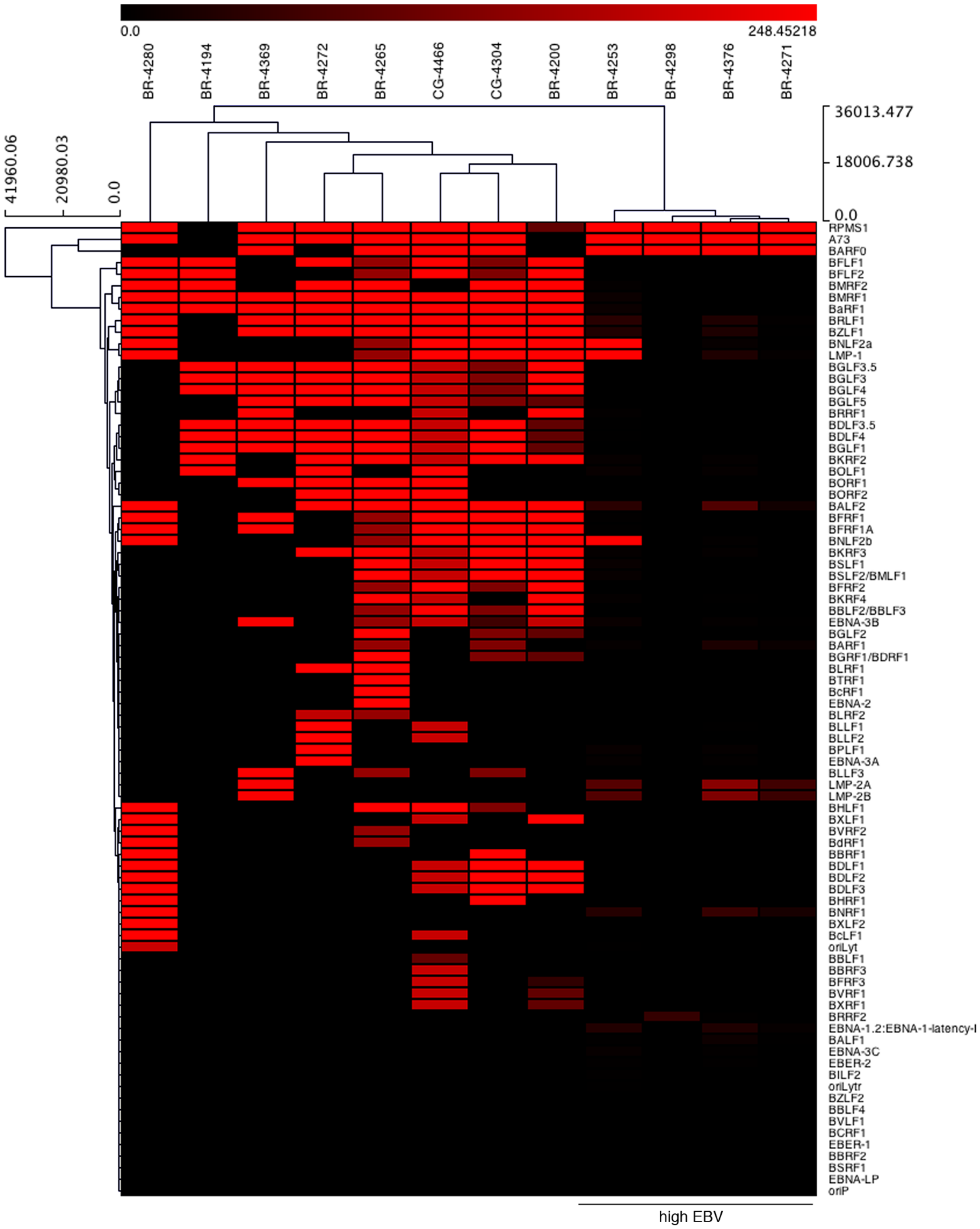


Figure 8. Cluster analysis of EBV genes from EBV-associated gastric carcinoma samples. The EBV genes from the 12 EBVaGC samples were subjected to hierarchical clustering and displayed with an expression heat map of all EBV genes.

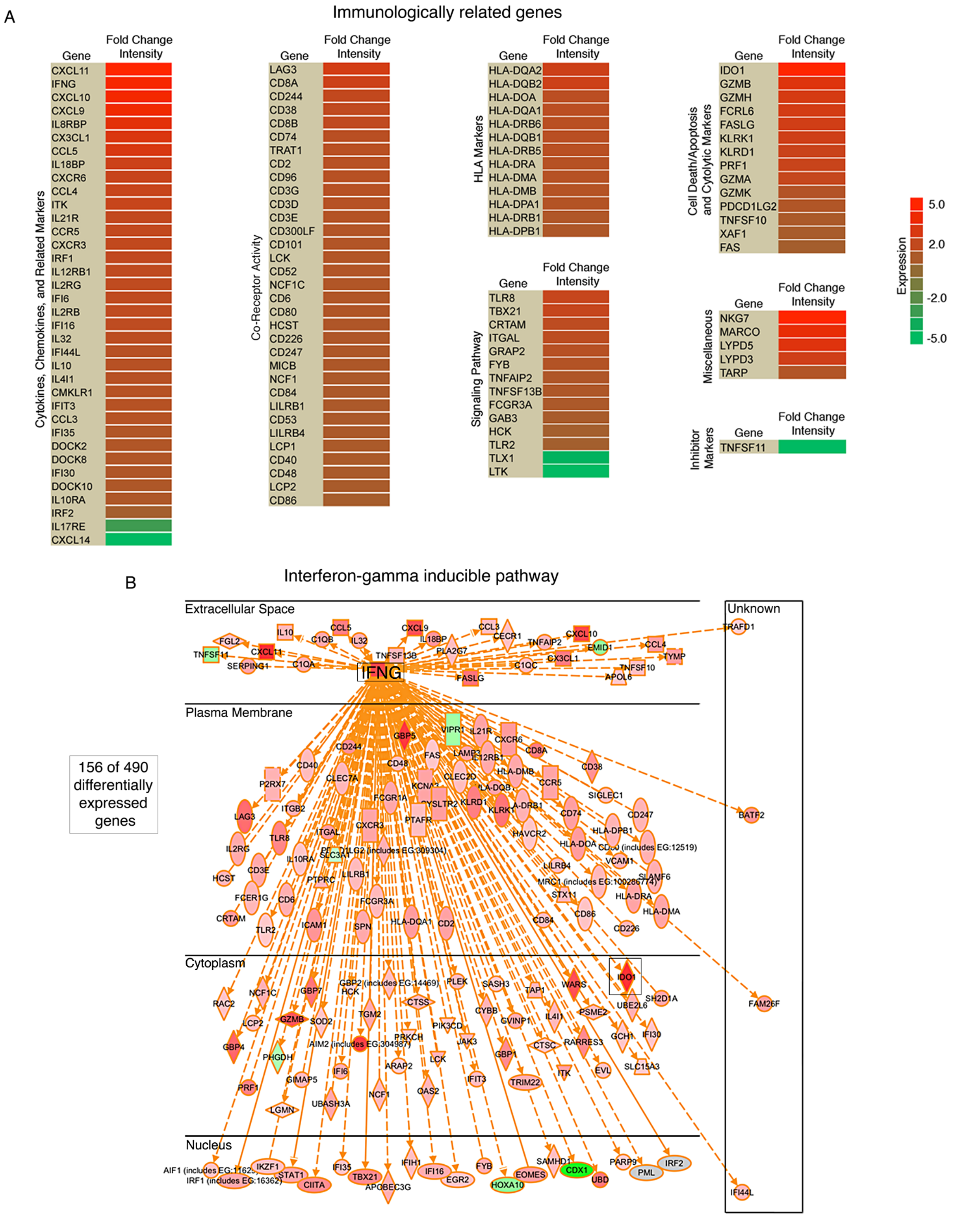

A
Immunologically related genes
Cytokines, Chemokines, and Related Markers
Gene
Fold Change Intensity
CXCL11
IFNG
CXCL10
CXCL9
IL8RBP
CX3CL1
CCL5
IL18BP
CXCR6
CCL4
ITK
IL21R
CCR5
CXCR3
IRF1
IL12RB1
IL2RG
IFI6
IL2RB
IFI16
IL32
IFI44L
IL10
IL4I1
CMKLR1
IFIT3
CCL3
IFI35
DOCK2
DOCK8
IFI30
DOCK10
IL10RA
IRF2
IL17RE
CXCL14
Co-Receptor Activity
Gene
Fold Change Intensity
LAG3
CD8A
CD244
CD38
CD8B
CD74
TRAT1
CD2
CD96
CD3G
CD3D
CD3E
CD300LF
CD101
LCK
CD52
NCF1C
CD6
CD80
HCST
CD226
CD247
MICB
NCF1
CD84
LILRB1
CD53
LILRB4
LCP1
CD40
CD48
LCP2
CD86
HLA Markers
Gene
Fold Change Intensity
HLA-DQA2
HLA-DQB2
HLA-DOA
HLA-DQA1
HLA-DRB6
HLA-DQB1
HLA-DRB5
HLA-DRA
HLA-DMA
HLA-DMB
HLA-DPA1
HLA-DRB1
HLA-DPB1
Signaling Pathway
Gene
Fold Change Intensity
TLR8
TBX21
CRTAM
ITGAL
GRAP2
FYB
TNFAIP2
TNFSF13B
FCGR3A
GAB3
HCK
TLR2
TLX1
LTK
Cell Death/Apoptosis and Cytolytic Markers
Gene
Fold Change Intensity
IDO1
GZMB
GZMH
FCRL6
FASLG
KLRK1
KLRD1
PRF1
GZMA
GZMK
PDCD1LG2
TNFSF10
XAF1
FAS
Miscellaneous
Gene
Fold Change Intensity
NKG7
MARCO
LYPD5
LYPD3
TARP
Inhibitor Markers
Gene
Fold Change Intensity
TNFSF11
Expression
5.0
2.0
-2.0
-5.0
B
Interferon-gamma inducible pathway
Extracellular Space
Plasma Membrane
Cytoplasm
Nucleus
Unknown
IFNG
156 of 490 differentially expressed genes
TRAFD1
BATF2
FAM26F
IFI44L

Figure 9. High numbers of infiltrating immune cellular genes are detected in EBVaGC. (A) Significant immunologically related genes differentially expressed in EBVaGC are represented in a heat map. The $\log^2$ fold change intensities are represented by the color gradient with red corresponding to the highest intensity and green corresponding to the lowest. (B) Interferon-gamma (IFNG) associated genes differentially expressed in EBVaGC are displayed in a diagram.

sample was notable in that like the two EBV positive samples, the pathology report for this sample similarly noted high levels of immune cell infiltration which may result from the presence of another infectious agent.

In Situ Hybridization for EBER was performed on a gastric carcinoma tissue array (ST2091; US Biomax) in order to assess the presence of EBV. In the strongly EBV positive cases, EBV was detected in the epithelial cells (e.g. F8 in Figure 10C). A high level of immune cell infiltration is observed in EBV positive (e.g. F8, Figure 10C) but not the tumor grade matched EBV negative sample, A15 (Figure 10C) with a high proportion of the immune cells in F8 showing intense IDO1 staining.

Analysis of the 178 down regulated genes showed that 19 tumor suppressor genes and 13 candidate oncogenes were found to be expressed at lower levels in hiEBVaGC (Table 1). Furthermore, we observed several inhibitors of the hedgehog and Wnt pathways to be expressed at lower levels in hiEBVaGCs suggesting additional components to the complex interactions involved in EBVaGC pathogenesis.

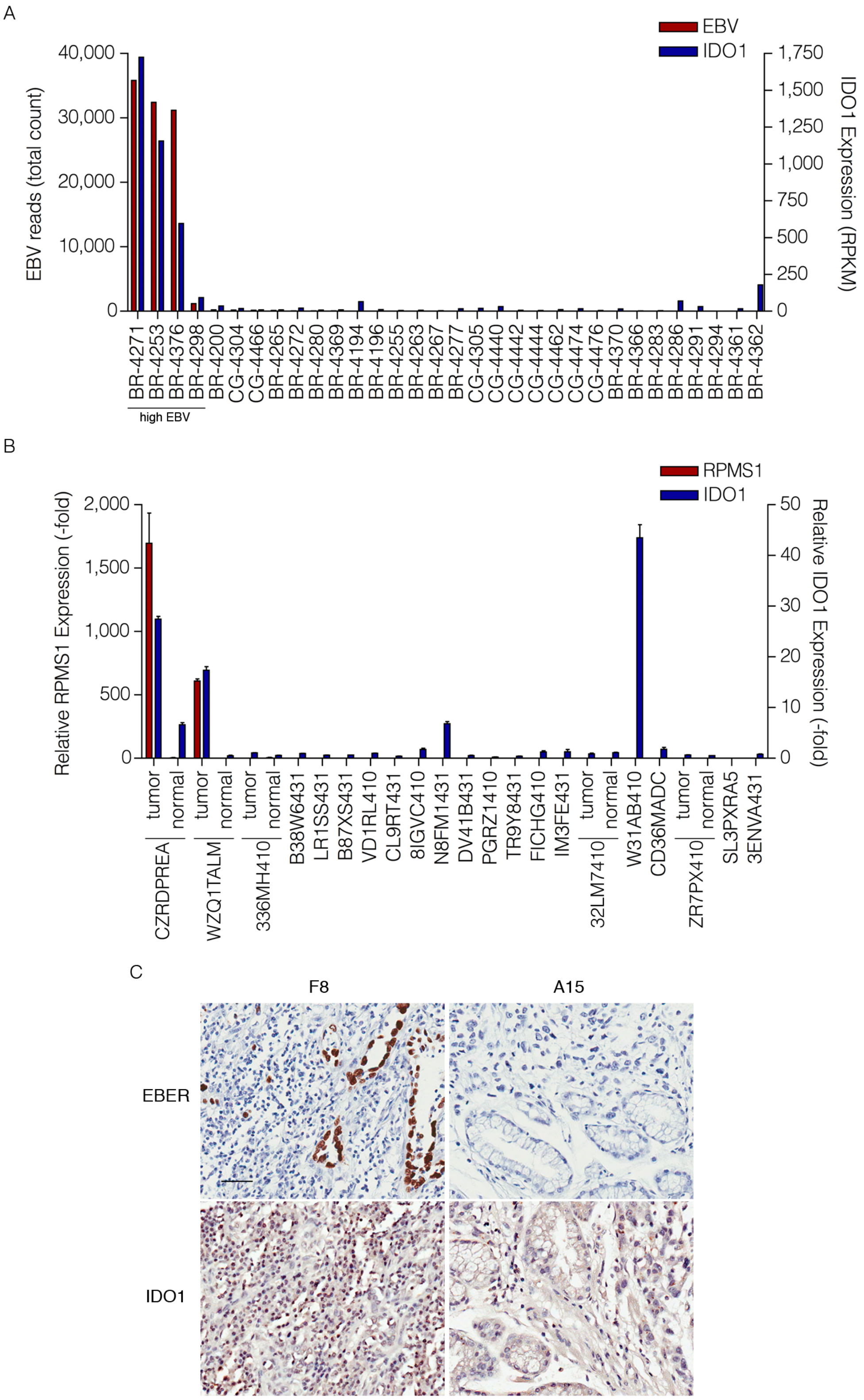

A
EBV
IDO1
EBV reads (total count)
40,000
30,000
20,000
10,000
0
1,750
1,500
1,250
1,000
750
500
250
0
IDO1 Expression (RPKM)
BR-4271
BR-4253
BR-4376
BR-4298
BR-4200
CG-4304
CG-4466
BR-4265
BR-4272
BR-4280
BR-4369
BR-4194
BR-4196
BR-4255
BR-4263
BR-4267
BR-4277
CG-4305
CG-4440
CG-4442
CG-4444
CG-4462
CG-4474
CG-4476
BR-4370
BR-4366
BR-4283
BR-4286
BR-4291
BR-4294
BR-4361
BR-4362
high EBV
B
RPMS1
IDO1
Relative RPMS1 Expression (-fold)
2,000
1,500
1,000
500
0
50
40
30
20
10
0
Relative IDO1 Expression (-fold)
tumor
normal
CZRDPREA
tumor
normal
WZQ1TALM
tumor
normal
336MH410
B38W6431
LR1SS431
B87XS431
VD1RL410
CL9RT431
8IGVC410
N8FM1431
DV41B431
PGRZ1410
TR9Y8431
FICHG410
IM3FE431
tumor
normal
32LM7410
W31AB410
CD36MADC
tumor
normal
ZR7PX410
SL3PXRA5
3ENVA431
C
F8
A15
EBER
IDO1

Figure 10. High levels of IDO1 in high EBV positive gastric carcinomas (A) Gene expression profile of the cohort of 32 gastric carcinoma samples (12 EBV-positive and 20 EBV-negative). Both total EBV reads and IDO1 expression (RPKM-reads per kilobase of exon model per million mapped reads) are represented as red and blue columns, respectively. (B) Gene expression profile of the cohort of 21 Vietnamese gastric carcinomas and 5 normal adjacent samples. Both relative RPMS1 expression (-fold) and relative IDO1 expression (-fold) are represented as red and blue columns and are the fold difference compared to the average of normal adjacent control values. (C) Images of paraffin-embedded human gastric carcinoma probed for EBER using in situ hybridization or IDO1 staining with immunohistochemistry. F8 and A15 each represent a specific gastric carcinoma on the tissue array selected to be closely matched with respect to age, tumor grade and stage. Scale bar represents 50μm.

Table 1 Representative genes with decreased expression in EBVaGC relative to EBVnGC

Gene	Log Fold	*P*-value	Function	Refs.
Tumor Suppressors				
GKN2	-8.1	4.70E-02	Down regulated in gastric carcinoma	[132]
TFF2	-6.9	2.91E-02	Hyper-methylated in gastric carcinoma	[133]
EFNA2	-4.8	1.31E-04	Tumor suppressor in gastrointestinal cancers	[134]
CLDN3	-3.9	1.40E-02	Down regulated in gastric carcinoma leads to proliferative potential	[135]
HOXA10	-3.4	1.31E-02	Up regulation in gastric carcinoma results in favorable prognosis	[136]
PTCH1	-2.6	1.94E-02	Tumor suppressor in medulloblastoma	[137]
CNTNAP2	-4.1	6.90E-04	Acts as tumor suppressor in glioma	[138]
SCARA3	-2.9	3.14E-02	Tumor suppressor in prostate cancer	[139]
WNK2	-3.5	4.51E-03	Tumor suppressor	[140]
VIPR1	-2.9	5.19E-03	Candidate tumor suppressor	[141]
REEP6	-2.5	2.38E-02	Tumor suppressor	[142]
B3GALT5	-6.5	9.86E-04	Down regulated in colon cancer	[143]
RBP4	-5.1	3.83E-03	Hyper-methylated in esophageal carcinoma	[144]
SORBS2	-2.7	2.44E-02	Down regulated in pancreatic cancer	[145]
HOXA9	-4.9	1.51E-02	Hyper-methylated in lung cancer	[146]
LRRN1	-4.5	3.09E-03	Hyper-methylated in non-small cell lung cancer	[147]
FOXA2	-2.8	1.45E-03	Tumor suppressor in lung cancer	[148]
HNF4A	-1.9	1.67E-03	Candidate tumor suppressor	[149]
RAP1GAP	-2.7	1.88E-02	Hyper-methylated in thyroid cancer	[150]

(Table continues)

Gene	Log Fold	*P*-value	Function	Refs.
Oncogenes				
CDH17	-3.7	2.23E-02	Up regulated in gastric carcinoma	[151]
CDX1	-7.3	1.21E-03	Up regulated in gastric carcinoma	[152]
ETV4	-1.9	1.32E-02	Up regulated in gastric carcinoma	[153]
PPP1R1B	-4.6	5.07E-03	Up regulated in gastrointestinal cancers	[154]
TM4SF5	-3.9	1.29E-02	Candidate oncogene	[155]
GPC3	-3.8	2.52E-04	Up regulated in hepatocellular carcinoma	[156]
TLX1	-3.7	1.79E-02	Oncogene	[157]
PEG10	-3.7	2.32E-02	Up regulated in hepatocellular carcinoma	[158]
WNT4	-3.4	2.66E-02	Candidate oncogene	[159]
CA8	-4.6	2.04E-03	Promotes colon cancer cell growth	[160]
BCAS1	-3.7	4.33E-02	Oncogene	[161]
FAM84A	-3.4	2.29E-03	Promotes colon cancer	[162]
USP2	-2.6	1.19E-02	Candidate oncogene – negative regulator of p53	[163]
Miscellaneous				
HHIP	-5.3	6.06E-03	Inhibits Hedgehog signaling	[164]
SHISA3	-3.9	1.70E-02	Inhibits Wnt and FGF signaling	[165]
NKD2	-3.1	1.02E-02	Antagonist of Wnt signaling	[166]
LRP4	-3.5	4.25E-04	Negative regulator of Wnt signaling	[167]
DUSP8	-1.6	2.51E-02	Inhibits JNK pathway	[168]
SLC26A3	-4.9	3.24E-02	Expression inhibited by IFNG	[169]
TNFSF11	-3.0	2.29E-02	Regulator of T cells and dendritic cells	[170]

5.5 Discussion

Consistent with the Shibata and Weiss study for the incidence of EBVaGC in the United States using ISH against EBERs [119], we detected EBV in 12 of the 71 (17%) gastric carcinoma samples from The Cancer Genome Atlas (TCGA) cohort using RNA CoMPASS. The detection of EBV using EBER ISH is widely used and the similar detection levels between the Shibata and Weiss study [119] and our work suggest that both methods are accurate for determining the presence of EBV in biological specimens. Importantly, however, the use of RNA-seq data allowed us to also infer the magnitude of local environmental signaling influences for different levels of EBV infection/viral gene expression. While the four samples with higher levels of EBV transcripts formed a clearly distinct cellular gene expression cluster, the eight samples with low numbers of EBV reads clustered in a mixed fashion among the EBV negative specimens. We propose that these two classes of EBV infection should be considered functionally distinct with possible implications in therapeutic intervention decisions and/or therapeutic response predictions.

RNA CoMPASS has the potential to simultaneously allow for the investigation of all pathogens present in tumor samples. In addition to EBV, we detected low levels of enterobacteria phage T4T, HCMV, Hepatitis C virus, and *Helicobacter pylori* (data not shown). The detection of enterobacteria phage T4T and Hepatitis C virus transcripts should be met with caution due to the likely possibility of environmental contamination and misclassification of these reads, respectively. While the HCMV reads likely represent true HCMV infection of cells within the tumor sample, the low read levels suggest either low numbers of HCMV infected cells or limited expression of polyadenylated viral RNAs. Finally, we detected *H. pylori* in three of the gastric carcinoma samples but the number of reads was very low in each case. Since bacterial RNAs are typically not polyadenylated or have limited numbers of polyadenylated RNAs [106-109], this low detection level probably results from the sequencing libraries being prepared from polyA selected RNA rather than an absence of *H. pylori* in these samples.

Of the 12 EBVaGCs, there was sufficient EBV read coverage in four of the samples to carry out more detailed transcriptomic analysis. LMP2, EBNA1, and LMP1 expression was detected in three of the EBVaGCs and these results are generally consistent with the findings of other groups [126-129]. The magnitude of expression from the BamHI A region relative to the transcription levels of other EBV genes is striking, however. This result is consistent with a previous report using a naturally infected EBV positive gastric carcinoma cell line [171]. Nevertheless, our analysis makes this observation in the natural *in vivo* setting of the tumor, and the use of RNA-seq facilitated the evaluation of transcript structures and the magnitudes of BamHI A region gene expression relative to other viral and cellular genes.

Although others have been unable to detect protein from naturally expressed BamHI A rightward transcripts [172, 173], the high expression level of these transcripts in hiEBVaGC samples suggests a functional role in gastric adenocarcinomas; possibly as long non-coding RNAs (lncRNA). These rightward BamHI A transcripts also encode as many as 44 intronic microRNAs (miRNAs)

[174, 175]. The function of the BART miRNAs in the EBV life cycle and in EBV associated malignancies is currently unclear but a recent study by Raab-Traub's group provided evidence that the BART miRNAs contribute to the tumor phenotype in EBVaGC [176]. In Raab-Traub's study, several lines of evidence supported this contention. First, very little EBV latent protein expression was detected and inhibition of the small amount of LMP1 expressed did not affect the cell's phenotype. Second, they observed that the majority of the significant cellular gene expression changes following infection of AGS (a gastric carcinoma cell line) cells with EBV were down regulated, many of which were significantly enriched in both experimentally and bioinformatically predicted BART miRNA targets [176, 177]. Based on this evidence and the fact that the BamHI A rightward transcripts are expressed at high levels in gastric carcinomas, it seems likely that the BART miRNAs play an important role in modulating the cellular phenotype in this tumor type. Nevertheless, many lncRNAs are involved in repressive complexes raising the possibility that the high levels of spliced rightward BamHI A transcripts that we detect *in vivo* may function as lncRNAs which similarly contribute to repression of cellular gene expression in hiEBVaGCs. Our strand specific RNA-seq analysis of SNU-719 cells further support our contention of high level expression of the rightward RPMS1 and A73 related transcripts in gastric carcinomas. This analysis also demonstrated the presence of additional rightward exons/genes within this region that may similarly play a role in lncRNA mediated regulation of viral and/or cellular signaling.

Although EBV primarily exhibits latent gene expression patterns in EBV associated tumors, recent studies using EBV associated lymphoma models suggest that a small portion of tumor cells express lytic transcripts that promote tumor growth [88, 89, 178, 179]. The Kenney lab showed that B cells harboring an EBV BZLF1 knock out mutant grew slower than wild type infected cells in a SCID mouse xenograft model [88]. In a separate study, they showed that a mutant EBV over expressing BZLF1 induces lymphomas with abortive lytic EBV infection in a humanized mouse model [89]. By assessing global EBV gene expression, we provide evidence for an abortive lytic phase *in vivo*; in the context of the natural setting of a human tumor. This supports the lymphoma animal studies from the Kenney group and raises the possibility that an abortive lytic phase may also play a role in EBV associated epithelial tumors.

One EBVaGC sample (BR-4253) was found to express high levels of BNLF2A/B. In the absence of significant expression of other lytic genes, the detection of BNLF2A/B expression in this sample was unexpected. One of the simplest models to explain this observation is a possible viral genetic alteration that juxtaposes this gene with an active viral promoter; in a manner reminiscent of the previously identified hetDNA (BZLF1 gene recombined to an active latency promoter) [180-182]. Alternatively, this could result from a rare viral integration event positioning the BNLF2A/B gene downstream from an active cellular promoter. Just as advantageous genetic alterations evolve in the cellular genome during cancer progression, a genetic event that resulted in the activation of BNLF2A/B may be an example of an advantageous viral genetic alteration that was selected during tumor evolution. BNLF2A was shown previously to function as an immune evasion protein through HLA class I down regulation (via blocking of TAP activity) [183]. This anti-immune function may have been selected for during tumor evolution and may support viral/tumor survival in this patient.

Cellular RNA expression profiling provided strong evidence for immune cell infiltration in hiEBVaGCs. This can be seen in tissue sections from EBV positive specimens (e.g. see Figure 10C) and is further supported by the pathology reports from the two EBV positive gastric carcinoma samples from the Vietnamese cohort which indicate high levels of immune cells. This observation is in line with previous studies using standard hematoxylin and eosin staining of tumor sections [120, 184] where lymphocyte infiltration was found to be predominately CD8+ T cells [185, 186]. Notably, however, despite this apparent robust immune response in hiEBVaGC, EBV and the infected tumor cells are able to persist in these patients. This suggests that these tumors may have compensatory immune evasion strategies that allow virus/tumor survival in this setting [187]. First, the limited expression of viral protein coding genes in EBVaGC likely contributes to the avoidance of viral antigen targeting [188]. Second, although the EBV encoded protein, EBNA1 is required for viral episomal replication/maintenance and therefore must be expressed in proliferating cells, it encodes a glycine-alanine repeat domain that blocks its proteasomal processing for CTL presentation [189, 190]. Third, here we found that expression of the interferon-gamma (IFNG) inducible CTL and NK inhibitor, indoleamine 2,3-dioxygenase (IDO1) is high in hiEBVaGC. IDO1 is a rate-limiting enzyme involved in the catabolism of tryptophan (Trp) [191]. CTLs and NK cells are uniquely sensitive to Trp depletion leading to the induction of stress responses and the inhibition of proliferation and activation [192, 193]. IDO1 functions to cause local tryptophan depletion under physiological and pathogenic immune tolerance settings such as during placentation and cancer [194, 195] where it is considered to be critical for establishing local immune tolerance. Among other candidate effectors, increased IFNG has been shown to induce IDO1 expression [196, 197]. Therefore, despite the apparent increase in CTL and NK cells in hiEBVaGCs, the activated IFNG signaling may counteract this response through IDO1 mediated Trp depletion (Figure 11); allowing tumor survival.

The findings of high IDO1 levels in several cancers and studies showing that IDO1 is critical for tumor survival has led to intense interest in the potential of anti-IDO1 based immunomodulatory therapeutics [198-201]. IDO1 inhibitors, such as the small molecule inhibitor, 1MT, have shown anti-tumor potential in combination with conventional chemotherapeutic drugs [200, 201]. This raises the important possibility that the therapeutic response for at least the subset of hiEBVaGCs may similarly be enhanced by the addition of IDO1 targeting therapeutics.

In our study, 156 of the genes found to be differentially expressed in EBVaGCs are linked to the IFNG pathway. The EBV encoded small RNAs, EBERs, have been shown to induce the expression of IFNG [202], and they likely play a significant role in the active IFNG response observed here. Despite this, the extensive level of secondary structure guiding the processing of the BamHI A

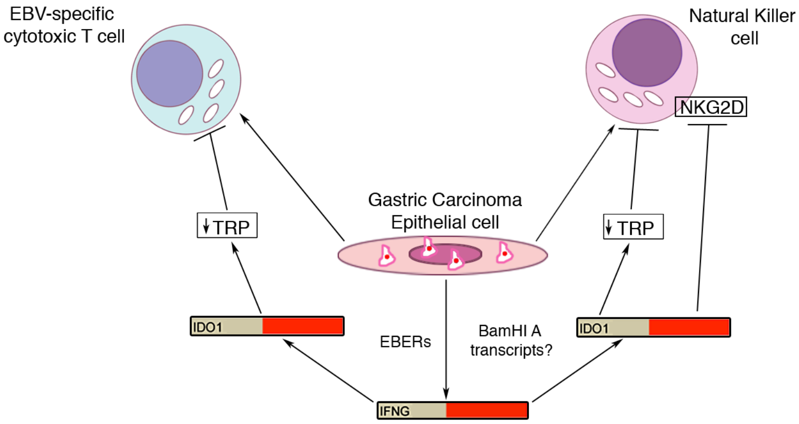


Figure 11. Model for EBV modulation of cytotoxic T-cell and natural killer cell function in tumor microenvironment. EBV infected gastric carcinoma cells recruit cytolytic immune cells such as T-cell and natural killer cells via unclear mechanisms. In addition, these cells induce an increase in interferon-gamma (IFNG) via EBERs and possibly BamHI A transcripts. Increased IFNG results in increased IDO1 resulting in depleted tryptophan. Depleted tryptophan results in T-cell and natural killer cell inhibition.

rightward introns during the miRNA processing steps may similarly contribute to the IFNG response in EBVaGCs observed here (Figure 9).

EBVaGCs exhibit extensive nonrandom DNA methylation at the promoter regions of various cancer-related genes [203, 204] and has been classified as having the CpG island methylator phenotype (CIMP) [205]. Several studies have investigated possible mechanisms of promoter hypermethylation of host genes in association with EBV infection. LMP1 mediated activation of DNA methyltransferase 1 (DNMT1), through either activation of c-Jun NH_2-terminal kinase (JNK)-activator protein-1 (AP-1) signaling [206] or through RB-E2F pathway activation [207], have been proposed as mechanisms in some systems. However, EBVaGCs do not typically express significant levels of LMP1. A study by Hino et al. demonstrated DNMT1 activation via LMP2A [208] raising the possibility that a LMP2A/DNMT1 mechanism could be involved. Nevertheless, a study by Chong et al. showed that DNMT expression was suppressed in EBVaGC and that the methylation of specific genes occurs through a mechanism independent of DNMT1 activation [209]. Based on this observation and on our findings of relatively low levels of LMP1 and LMP2A expression in EBVaGCs, we propose that methylation/imprinting may be downstream of more direct EBV inhibitory mechanisms. The robust expression levels of the BamHI A transcripts in EBVaGCs put them high on the radar as candidates for this type of regulation, possibly through lncRNA mediated chromatin imprinting based mechanisms.

Multiple tumor suppressors were expressed at lower levels in EBVaGCs including five (TFF2, RBP4, HOXA9, LRRN1, and RAP1GAP) that are known to be hypermethylated in cancers [133, 144, 146, 147, 150]. Another gene expressed at lower levels in EBVaGCs was HNF4A, a cell-specific transcription factor known to regulate a large number of genes in liver, intestine, pancreas, and stomach [149]. Decreased expression of HNF4A has been shown in renal cell carcinoma [210] and has recently been shown to regulate key genes involved in cellular proliferation [149]. A recent study by Lucas and colleagues suggest that HNF4A acts as a tumor suppressor [149].

In addition to tumor suppressors, we also observed several candidate oncogenes to be expressed at lower levels in EBVaGCs including 4 (CDH17, CDX1, ETV4, and PPP1R1B) known to be over expressed in gastric carcinoma and gastrointestinal cancers [151-154]. Although EBV clearly contributes to cancers, its oncogenic properties are a byproduct of its life cycle rather than an evolved tumor promoting function. In line with this concept, the lower levels of these oncogenes in EBVaGCs may be a byproduct of EBV's life cycle. Conversely, it is possible that the non-EBV mediated gastric carcinoma oncogenic pathway occurs through the up-regulation of these genes whereas the EBV assisted oncogenic path does not. Regardless of which of these principles may explain this observation, the lower levels of oncogenes in EBVaGC may partly explain the more favorable prognosis that is often observed in EBVaGC. Similarly, the lower levels of USP2, a negative regulator of p53, may help explain the normal to elevated levels of p53 found in EBVaGC [211, 212] and possibly the better responses to chemotherapeutics.

An increase in sonic hedgehog (SHH) expression and its activation in gastric carcinoma, especially *H. pylori* associated gastric carcinomas has been well established [213]. In our study, several inhibitors of both the SHH and Wnt pathways were found to be lower in hiEBVaGC including HHIP (SHH) and SHISA3, NKD2 and LRP4 (Wnt). The decrease in SSH inhibitor, HHIP, [164] suggests that Hedgehog activity may be higher in hiEBVaGC. Down regulation of HHIP in pancreatic cancer has been shown to be mediated through epigenetic CpG hypermethylation within the promoter region [214]. This raises the possibility of a specific methylation process by EBV, since we observe a significantly lower level of HHIP reads in the hiEBVaGC compared to loEBVaGC and EBVnGC. Hypermethylation of the promoter region of NKD2 has been established in malignant astrocytic gliomas [215], and a CpG island within the SHISA3 and LRP4 promoter regions have been identified [216]. This suggests that epigenetic silencing of these Wnt pathway inhibitors may also occur through an EBV mediated mechanism.

PART III

RESOLUTION OF LINGERING CONTROVERSY OF A VIRAL ASSOCIATION WITH GLIOBLASTOMAS

Chapter 6: No Virus and Tumor Association Established For Gliomas Using Next Generation Sequencing

A comprehensive next generation sequencing-based virome assessment in brain tissue suggests no major virus - tumor association

Michael J. Strong, Eugene Blanchard, Zhen Lin, Cindy A. Morris, Melody Baddoo, Christopher M. Taylor, Marcus L. Ware, and Erik K. Flemington

Published in Acta Neuropathol Commun. 2016 Jul 11;4(1):71. doi: 10.1186/s40478-016-0338-z.

6.1 Abstract

Next generation sequencing (NGS) can globally interrogate the genetic composition of biological samples in an unbiased yet sensitive manner. The objective of this study was to utilize the capabilities of NGS to investigate the reported association between glioblastoma multiforme (GBM) and human cytomegalovirus (HCMV). A large-scale comprehensive virome assessment was

performed on publicly available sequencing datasets from the Cancer Genome Atlas (TCGA), including RNA-seq datasets from primary GBM (n=157), recurrent GBM (n=13), low-grade gliomas (n=514), recurrent low-grade gliomas (n=17), and normal brain (n=5), and whole genome sequencing (WGS) datasets from primary GBM (n=51), recurrent GBM (n=10), and normal matched blood samples (n=20). In addition, RNA-seq datasets from MRI-guided biopsies (n=92) and glioma stem-like cell cultures (n=9)) were analyzed. Sixty-four DNA-seq datasets from 11 meningiomas and their corresponding blood control samples were also analyzed. Finally, three primary GBM tissue samples were obtained, sequenced using RNA-seq, and analyzed. After in-depth analysis, the most robust virus findings were the detection of papillomavirus (HPV) and hepatitis B reads in the occasional LGG sample (4 samples and 1 sample, respectively). In addition, low numbers of virus reads were detected in several datasets but detailed investigation of these reads suggest that these findings likely represent artifacts or non-pathological infections. For example, all of the sporadic low level HCMV reads were found to map to the immediate early promoter intimating that they likely originated from laboratory expression vector contamination. Despite the detection of low numbers of Epstein-Barr virus reads in some samples, these likely originated from infiltrating B-cells. Finally, human herpesvirus 6 and 7 aligned viral reads were identified in all DNA-seq and a few RNA-seq datasets but detailed analysis demonstrated that these were likely derived from the homologous human telomeric-like repeats. Other low abundance viral reads were detected in some samples but for most viruses, the reads likely represent artifacts or incidental infections. This analysis argues against associations between most known viruses and GBM or meningiomas. Nevertheless, there may be a low percentage association between HPV and/or hepatitis B and LGGs.

6.2 Introduction

Glioblastoma multiforme (GBM) is the most common malignant primary brain tumor in adults. An estimated 77,670 new cases of primary CNS tumors are expected to be diagnosed in the United States in 2016 [217]. Of these, 24,790 will be diagnosed as malignant [217]. Although the incidence of primary brain tumors is low compared to other cancer types, primary brain tumors give rise to a disproportionate amount of morbidity and mortality, often robbing patients of basic and critical functions such as movement and speech [218]. The median survival of newly diagnosed patients is only 12-15 months, making it one of the most devastating types of cancers [219]. In fact, the five-year survival rate for primary malignant brain and central nervous system tumors is the sixth lowest among all types of cancers after pancreatic, liver & intrahepatic bile duct, lung, esophageal, and stomach [218]. Unfortunately, despite substantial investigations into disease mechanisms and at least some advances in currently available treatment options, the outcomes for GBM patients remain dismal [219].

Although an association between human cytomegalovirus (HCMV) and GBM was first observed in 2002 [220], there is still a high degree of discordance in the literature regarding the detection of viral agents in CNS tumors [60, 77, 220-242]. These discrepancies have been attributed to a number of issues including the use of different cohorts, differences in sensitivities of different PCR assays for low levels of viral gene expression, and the exquisite sensitivities of assays such as IHC to slight differences in experimental conditions.

In an attempt to remedy the high degree of discordance regarding the detection of HCMV in CNS tumors, an HCMV and glioma symposium was convened in Washington, DC on April 17, 2011. At the conclusion of this workshop, a summary paper was published reporting the consensus position in 4 major areas: 1) the existence of HCMV in gliomas, 2) the role of HCMV in gliomas, 3) HCMV as a therapeutic target, and 4) key future investigative directions [243]. Based on the evidence presented at the workshop, it was concluded that HCMV sequences and viral gene expression exist in many malignant gliomas and that in vitro studies support the idea that HCMV can modulate key signaling pathways in glioblastomas [243].

Next generation sequencing (NGS) has the ability to globally interrogate the genetic composition of biological samples in an unbiased manner and with relatively high sensitivity. Applying this technology to pathogen discovery has already shown promise, resulting in the discovery of a novel Merkel cell polyomarvirus in Merkel cell carcinoma [40], for example. In our laboratory, we have utilized NGS technology in the interrogation of Epstein-Barr virus (EBV) in diffuse large B-cell lymphomas [95] and gastric carcinoma [61]. The goal for the study presented here was to help resolve the lingering controversy pertaining to the presence of HCMV in GBM while at the same time providing a comprehensive and unbiased assessment of the viral genetic composition of brain tumor biopsies. This analysis failed to find convincing evidence for an association between HCMV or other known viruses and GBM or mengiomas. Nevertheless, we detected human papillomavirus (HPV) and hepatitis B in some low-grade gliomas (LGG). In addition, we expand on our previous reporting of potential contamination and/or interpretational artifacts that need to be considered in next generation sequencing based metagenomic and metatranscriptomic studies [244, 245].

6.3 Materials and Methods

6.3.1 Clinical tumor sample and sequence data acquisition

All human specimens were de-identified prior to acquisition. Fresh frozen tissue from 2 GBM samples was obtained from the Louisiana Cancer Research Consortium (LCRC) Biospecimen Core. An RNA-seq dataset from a lymphoblastoid cell line immortalized with EBV (JY) was used as a control for downstream analysis [76].

Publically available sequence datasets were obtained from various sources. Next generation sequencing datasets from The Cancer Genome Atlas (TCGA) initiative were downloaded from the Cancer Genomics Hub (CGHub; https://cghub.ucsc.edu) and included RNA-seq datasets (unaligned fastq files) from primary GBM tumors [246-248] (n=157), recurrent GBM tumors (n=13), low grade gliomas [249] (LGG; n=514), recurrent low grade gliomas (n=17), and normal brain (n=5), and TCGA whole genome sequencing datasets (aligned bam files) from primary GBM tumors (n=51), recurrent GBM tumors (n=10), and normal matched blood samples (n=20) (Table 1). The aligned bam files were converted to fastq files using bam2fastq (http://www.hudsonalpha.org/gsl/information/software/bam2fastq, default parameters).

Additional brain tissue sequencing datasets were obtained from the NCBI Sequence Read Archive (Table 1). RNA-seq datasets from tumor and peripheral brain tissue of a GBM patient were obtained using accession number SRP009144 [250]. Normal brain tissue RNA-seq dataset from the Illumina Human Body Map 2.0 project was obtained using (NCBI GEO accession GE30611). RNA-seq datasets from a cohort of short-term cultures of glioma stem-like cells freshly isolated from nine patients diagnosed with primary GBM were downloaded using accession number SRP016798 [251, 252]. A cohort of RNA-seq datasets from MRI-localized biopsies of the tumor core and margins from multiple glioma patients and non-neoplastic brain tissue specimens were downloaded using accession number SRP044668 analyzed [253]. Non-neoplastic brain tissue samples were obtained from multiple patients undergoing procedures to alleviate epilepsy symptoms or to place shunts to treat normal pressure hydrocephalus. In total, RNA-seq datasets from 39 contrast-enhancing glioma core samples, 36 non-enhancing FLAIR glioma margin samples, and 17 non-neoplastic brain tissue samples were analyzed. A cohort of 64 whole genome sequencing datasets from 11 Grade I meningiomas and 11 matched blood samples were obtained using accession number SRP016129 [254]. Finally,

Table 1. Sequencing Datasets

Dataset	Disease	Analyte Type	Library Type	Number of Samples	Accession Number
TCGA-GBM	GBM	Total RNA (polyA)	RNA-seq	157	
TCGA-GBM	GBM (recurrence)	Total RNA (polyA)	RNA-seq	13	
TCGA-GBM	Normal Brain	Total RNA (polyA)	RNA-seq	5	
TCGA-GBM	GBM	DNA	WGS	51	
TCGA-GBM	GBM (recurrence)	DNA	WGS	10	
TCGA-GBM	Normal Blood	DNA	WGS	20	
TCGA-LGG	LGG	Total RNA (polyA)	RNA-seq	514	
TCGA-LGG	LGG (recurrence)	Total RNA (polyA)	RNA-seq	17	
00RTS3	GBM	Total RNA (polyA)	RNA-seq	1	
CAURPRVE	GBM	Total RNA (ribodepleted)	RNA-seq	1	
H8CPFRSJ	GBM	Total RNA (ribodepleted)	RNA-seq	1	
GBM	GBM	Total RNA (polyA)	RNA-seq	1 GBM 1 Normal	SRP009144
BodyMap	Normal Brain	Total RNA (polyA)	RNA-seq	1	
Glioma stem cells	GBM		RNA-seq	9	SRP016798
MRI-localized biopsy	GBM	Total RNA (polyA)	RNA-seq	39 CE 36 NE 17 NB	SRP044668
Meningioma	Meningioma	DNA	WGS	11 tumor 11 blood	SRP016129
HCMV	Fibroblast cell line	Total RNA (polyA)	RNA-seq	3	SRP016143

CE: contrast-enhancing glioma core samples, NE: non-enhancing FLAIR glioma margin samples, NB: non-neoplastic brain tissue, HCMV: human cytomegalovirus

RNA-seq datasets from HCMV infected fibroblasts were downloaded using accession number SRP016143 [255].

6.3.2 Sample preparation and next generation RNA sequencing

Total RNA was extracted from 3 primary GBM biopsies (00RTS3 – from the LCRC biospecimen core; and CAURPRVE and H8CPFRSJ – from Bioserve) using Trizol (Invitrogen, Carlsbad, CA) according to manufacturer's instructions. Total RNA from sample 00RTS3 was subjected to polyA selection, and the library was prepared using the ScriptSeq Protocol (Epicentre, Madison, WI) and subjected to 2x101 base paired-end sequencing on an Illumina Hi-seq 2000 machine. Total RNA from samples CAURPRVE, H8CPFRSJ, and JY were subjected to ribosomal RNA depletion using the Ribo-Zero kit (Epicentre, Madison, WI) and cDNA libraries were prepared using the Illumina Truseq Stranded Total RNA Sample Prep Kit and subjected to 1x101 base single-end multiplexed sequencing on an Illumina Hi-seq 2000 machine. The RNA-seq data used in this publication is available through GEO Series accession number (in process).

6.3.3 Metatranscriptomic Analysis using RNA CoMPASS

Metatranscriptome analysis was performed by running single-end or one pair from paired-end sequencing data through our automated RNA-seq exogenous organism analysis software, RNA CoMPASS [83]. Within RNA CoMPASS, reads were aligned to the human reference genome, hg19 (UCSC), plus a splice junction database (which was generated using the make transcriptome application from Useq [50]; splice junction radius set to the read length minus 4) using Novoalign V3.00.05 (www.novocraft.com) [-o SAM, default options]. Non-mapped reads were subjected to a BLAST V2.2.30 search against the Human RefSeq RNA database to identify and remove human reads that fail to be identified through the Novoalignment. Remaining non-human reads were then subjected to a BLAST V2.2.30 search against the NCBI NT database to identify reads corresponding to known exogenous organisms [52]. Results from the NT BLAST searches were filtered to eliminate matches with an E-value of greater than 10e-6. The results were then fed into MEGAN 4 V4.70.4 for visualization of taxonomic classifications [53]. RNA CoMPASS was run in parallel on three 2x2.66 GHz 12 core Intel Xeon Mac Pro computers with 64-96GBs of memory each.

6.3.4 Viral Transcriptome Analysis

Raw sequence data from RNA-seq and DNA-seq were aligned to a reference containing a human genome (hg19; Genome Reference Consortium GRCH37) plus a library of 740 virus genomes (including sequences from all known human viruses documented by NCBI) [256]. The alignments were performed using Spliced Transcripts Alignment to a Reference (STAR) aligner version 2.3.0 [--clip5pNbases 6 (only used if reads were longer than 36 base pairs), default options] [85]. Uniquely mapped viral and human reads were quantified using in-house computational pipelines. Signal maps (i.e. the total number of reads covering each nucleotide position) from viruses of interest were generated using IGV tools and were subjected to manual visual inspection using the IGV genome browser [87].

6.3.5 Quantitative RT-PCR

Total RNA was reverse-transcribed using the SuperScript III First-Strand Synthesis System for RT-PCR (Invitrogen, Carlsbad, CA). Random hexamers were used with 1 ug total RNA in a 20μl reaction volume according to manufacturer's instructions. For the incubation steps (25˚C for 10 min followed by 50˚C for 50 min) a Mastercycler ep (Eppendorf, Hamburg, Germany) was used. For real-time PCR, 1μl of the resulting cDNA (diluted to 10ng/ul) was used in a 10μl reaction mixture that included 5μl of 2x iQ SYBR Green Supermix (Bio-Rad, Hercules, CA), 1μl of 10μM forward and reverse primer mix (Integrated DNA Technologies, Coralville, IA), and 3μl of PCR grade water.

A list of PCR primer oligos can be found in Table 2. Each sample was PCR'ed in triplicate. No-template controls and no-reverse transcription controls were also included in each PCR run. Thermal cycling was performed on a CFX96 Real Time System (Bio-Rad, Hercules, CA) and data analysis was performed using the CFX Manager 3.0 software. Cycling conditions included an initial incubation at 95°C for 3 minutes followed by 40 cycles consisting of 95°C for 15 seconds, and 60°C for 60 seconds. Melting curve analysis was performed at the end of every qRT-PCR run.

Table 2. List of PCR primers

Primer	Forward Primer	Reverse Primer
EBV (EBER1)[257]	5'- GGACCTACGCTGCCCTAGA -3'	5'- CAGCTGGTACTTGACCGAAGA -3'
HCMV[257]	5'- TAATACAAGCCATCCACA -3'	5'- TAGATAAGGTTCATGAGCCT -3'
MuLV[76]	5'- CTGAGAGAAGTCAACAAGCG -3'	5'- CTGGATCTCTCCACTCAAAGGC -3'
GAPDH	5'- GCCAAAAGGGTCATCATCTC -3'	5'- GGGGCCATCCACAGTCTTCT -3'

6.4 Results

6.4.1 Lack of virus association in The Cancer Genome Atlas GBM RNA-seq datasets

Previous studies have used human read subtraction-based approaches to assess the metatranscriptomic profile of primary and recurrent glioma RNA-seq datasets from The Cancer Genome Atlas (TCGA) where a lack of association with HCMV was reported [60, 77, 236, 238]. To investigate this issue further, we first performed a global virome analysis on these samples (as well as 5 normal brain tissue samples), using a more directed, non-subtraction based approach that we have reported previously [244, 256, 258]. For this analysis, we directly aligned all reads from these datasets to an alignment index containing the human genome plus a library of 740 virus genomes (including sequences from all known human viruses documented by NCBI) using the Spliced Transcripts Alignment to a Reference (STAR) aligner. Running this virome pipeline on a known EBV-associated gastric cancer tissue biopsy cohort from TCGA [259] showed EBV read levels which ranged from 7-400 viral Reads Per Mapped Human reads (RPMH) [258]. Like PCR, next generation sequencing is susceptible to low level contamination issues [245]. Nevertheless, in an attempt to capture potential low abundance infections, we set a viral read threshold that was 10 times lower than the lowest viral RPMH value for EBV in the gastric cancer cohort [258, 259]. Based on this requirement of at least 0.7 viral RPMH, no virus associations were called in 157 primary and 13 recurrent gliomas or in 5 normal brain RNA-seq datasets analyzed except for the finding of 2.1 human herpesvirus (HHV) -6 or

HHV-7 RPMH (note: HHV-6 and HHV-7 reads are considered together here due to all reads mapping to a homologous region of these two viruses) in one recurrent glioma (Figure 1). Notably, all HHV6/7 reads in this sample were found to contain the simple sequence, TAACCC, a repeat found in both of these virus genomes and in human telomeric repeats. Since no HHV6/7 assigned reads mapped outside of the TAACCC repeat region of the viral genomes, we conclude that these reads likely originated from contaminating genomic sequences from human telomeric repeats.

Analysis of RNA-seq data from a cohort of time course HCMV infected fibroblasts [255] demonstrated robust numbers of HCMV reads (ranging from 68,875 – 578,635 RPMH) that escalated roughly proportionally to the length of infection. Since our detection cutoff of 0.7 RPMH for TCGA cohort is about 5 to 6 orders of magnitude less than these read numbers, this data supports a lack of association between HCMV and GBMs.

6.4.2 Assessment of sequence library preparation strategies

Although the bulk of viral RNAs are polyadenylated, some viral transcripts are not. To assess whether we were missing viral infections because of a lack of detecting non-polyadenylated viral transcripts, we wanted to test whether sequencing of ribodepleted RNAs (versus polyA selected RNAs as per the TCGA cohort) might yield the detection of viral reads. HCMV has been reported to have a high penetrance in GBM with some studies reporting as high as 90-100% positivity [220, 222, 225, 233, 243]. We sequenced two of our own GBM samples

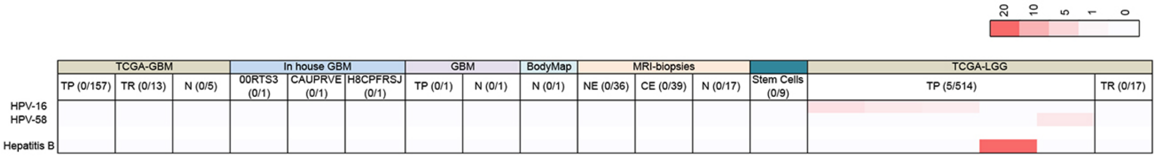


Figure 1. Heat map showing the number of viral reads per million human mapped (RPHM) reads for brain tissue RNA-seq datasets. Color intensity represents relative viral RPHM across all datasets.

using ribosomal depleted RNA as well as an additional sample using poly-A selected RNA. Because virome analysis was not previously performed on these samples, we implemented two virus detection approaches. First we analyzed each of these samples through our custom subtraction based pathogen detection pipeline, RNA CoMPASS, which analyzes the entire metatranscriptome. For a more focused virome analysis, we then implemented the STAR aligner approach.

No associations were found with any known viruses in the polyA selected RNA sample (00RTS3) (Figure 1) using either RNA CoMPASS or the STAR/virome analysis approach. Additional publically available poly-A selected RNA datasets from a GBM study, in which they sequenced one primary GBM and the matched normal brain, and a normal brain sample from the BodyMap project were obtained and analyzed. Assessing these public poly-A selected RNA datasets similarly showed no association with any known virus (Figure 1).

Analysis of the ribosomal depleted RNA samples (CAUPRVE and H8CPFRSJ) revealed reads from the murine leukemia virus (MuLV) family (2.4 and 4.3 RPMH, respectively) and a low abundance of EBV reads (0.8 and 1.9 RPMH, respectively). Further analysis of the viral read coverage for MuLV and EBV demonstrated near identical coverage patterns to another sample (JY – an EBV-immortalized B cell lymphoblastoid cell line, which we have previously shown to be infected with MuLV [76]) that was sequenced in the same sequencing lane (Figure 2). Our suspicion of contamination across barcodes was confirmed by real-time PCR for samples CAUPRVE and H8CPFRSJ, in which neither MuLV nor EBV transcripts were detected (Figure 3). Real-time RT-PCR for HCMV transcripts were similarly negative in these samples, thus confirming the lack of RNA-seq based HCMV findings in these datasets.

6.4.3 Analysis of tumor tissue sampling

Since GBM is a very heterogeneous solid tumor, differential sampling of the tumor mass may result in different transcriptomic and metatranscriptomic profiles. To take this issue into account for the detection of HCMV, we took advantage of a well-designed study in which the authors sequenced MRI-localized biopsies of the tumor core and margins from multiple GBM patients. The authors also sequenced several non-neoplastic brain tissue samples [253]. Virome assessment of this cohort using RNA CoMPASS detected no viruses. Assessment of this cohort using the STAR/virome approach detected Human Immunodeficiency virus type 1 (HIV-1) at levels greater than 0.7 viral RPMH in 2 glioma samples (5.3 and 5.5 RPMH) taken from the non-enhancing FLAIR portion (margins) of the tumor and 2 glioma samples (3.1 and 4.2 RPMH) taken from the contrast enhancing portion (core) of the tumor. To investigate these findings further, we analyzed the genome coverage of HIV reads in these samples plus three samples that showed HIV read levels that are below our 0.7 RPMH threshold (1 non-neoplastic sample (0.37 RPMH) and 2 additional contrast enhancing core samples (0.04 and 0.16 RPMH)). Inspection of the HIV-1 read coverage from all seven glioma samples revealed that the majority of the HIV-1 reads aligned to the long terminal repeat regions and additional

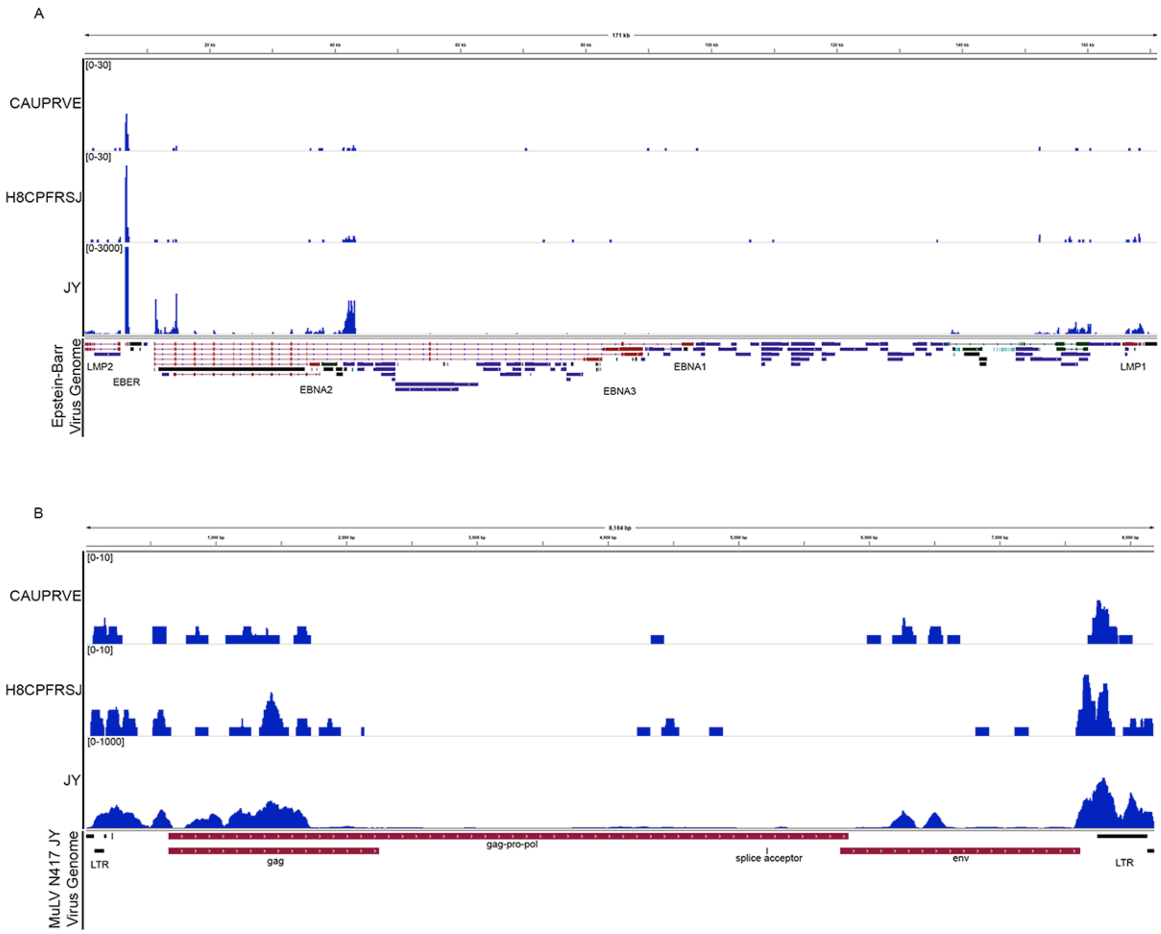


Figure 2. EBV and MuLV gene coverage analysis on RNA-seq datasets from GBM samples and JY cell line. Data was displayed using the Integrative Genomics Viewer (IGV) using the (A) EBV Akata genome (GenBank accession number KC207813) and (B) MuLV Abelson genome (GenBank accession number NC_001499). The y-axis represents the number of reads at each nucleotide position in the genome.

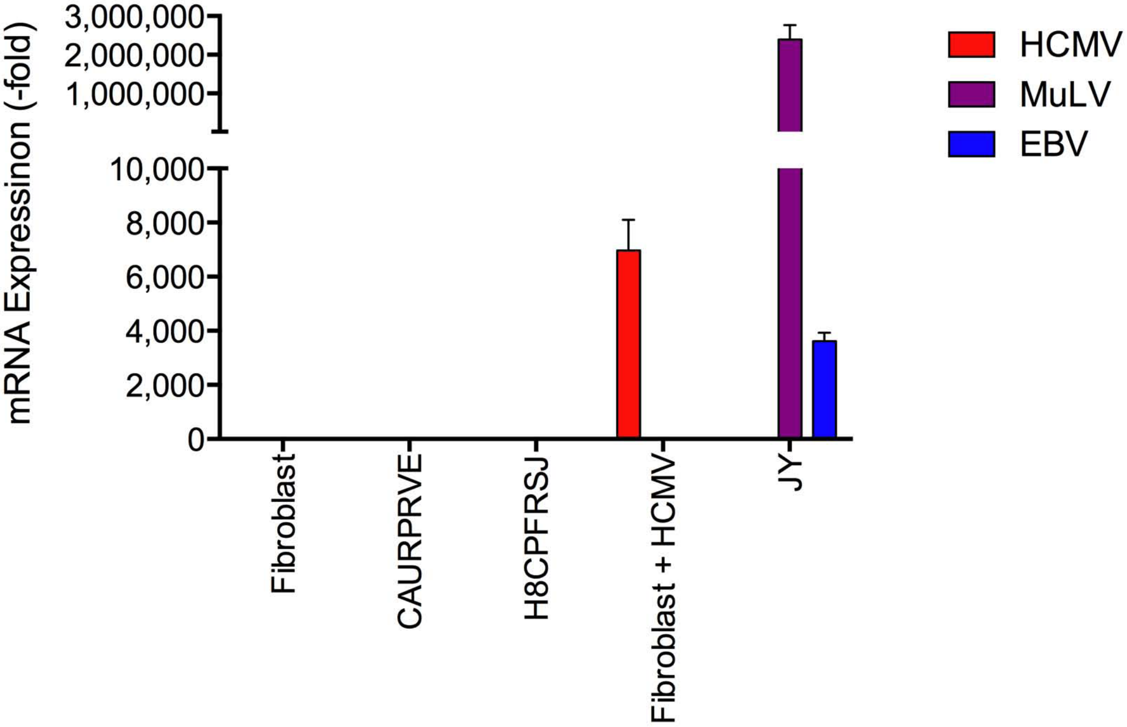


Figure 3. Gene expression profile of several tissue samples. These included a fibroblast-derived cell line, fibroblast-derived cell line infected with HCMV, JY cell line, and 2 primary GBMs. Relative viral mRNA expression (-fold) is represented as red (HCMV), purple (MuLV), and blue (EBV) columns and are the fold difference compared to the fibroblast-derived cell line values.

homologous regions of the expression vector pH1TO, as shown in Figure 4. MuLV reads were also detected in 33 out of the 36 glioma samples (0.03 – 5.74 RPMH) taken from the non-enhancing FLAIR portion of the tumor, 35 out of the 39 glioma samples (0.03 – 12.72 RPMH) taken from the contrast enhancing portion of the tumor, and 15 out of the 17 non-neoplastic samples (0.03 – 3.33 RPMH). Inspection of the MuLV read coverage showed a sporadic genomic coverage pattern that was similar across the tumor samples (Figures 5-7). The observed similar coverage profiles, the finding of low read numbers, and our previous observations of sample cross contamination of human samples with MuLV, lead us to suspect that the MuLV read findings here are most likely due to cross contamination. Finally, although the level of HCMV reads did not exceed 0.7 viral RPMH in any given sample, for completeness, all detected HCMV reads including those from 14 glioma samples from non-enhancing FLAIR portions (0.03 – 0.21 RPMH) and 12 glioma samples from contrast enhancing portions (0.04 – 0.25 RPMH) were analyzed further. All HCMV reads were found to align to the HCMV immediate early promoter (Figures 8-9). As mentioned above for the HCMV findings in the TCGA cohort, this is suggestive of sample contamination with laboratory expression plasmids bearing the HCMV promoter. After accounting for the artifactual viral reads, we conclude that this cohort shows no likely association with any known viruses (Figure 1).

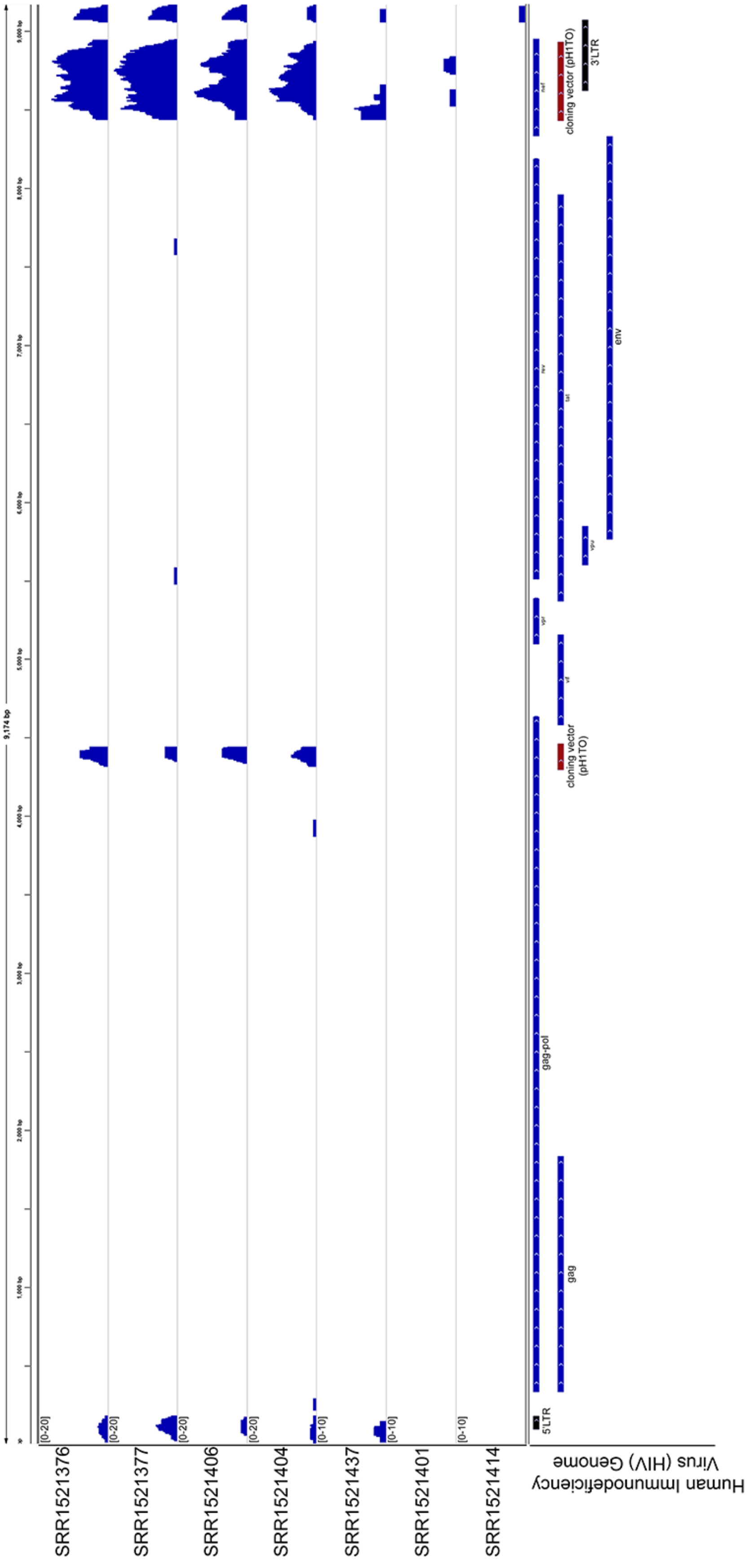

9,174 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
6,000 bp
7,000 bp
8,000 bp
9,000 bp
SRR1521376
SRR1521377
SRR1521406
SRR1521404
SRR1521437
SRR1521401
SRR1521414
[0-20]
[0-10]
Human Immunodeficiency Virus (HIV) Genome
5'LTR
gag
gag-pol
cloning vector (pH1TO)
vif
vpr
vpu
tat
rev
env
nef
3'LTR

Figure 4. HIV gene coverage analysis for MRI-localized biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the HIV-1 genome (GenBank accession number NC_001802). The y-axis represents the number of reads at each nucleotide position in the genome. Samples SRR1521376, SRR1521377, SRR1521406, and SRR1521404 are displayed with a max read level of 20, while the max read level for the other samples is 10.

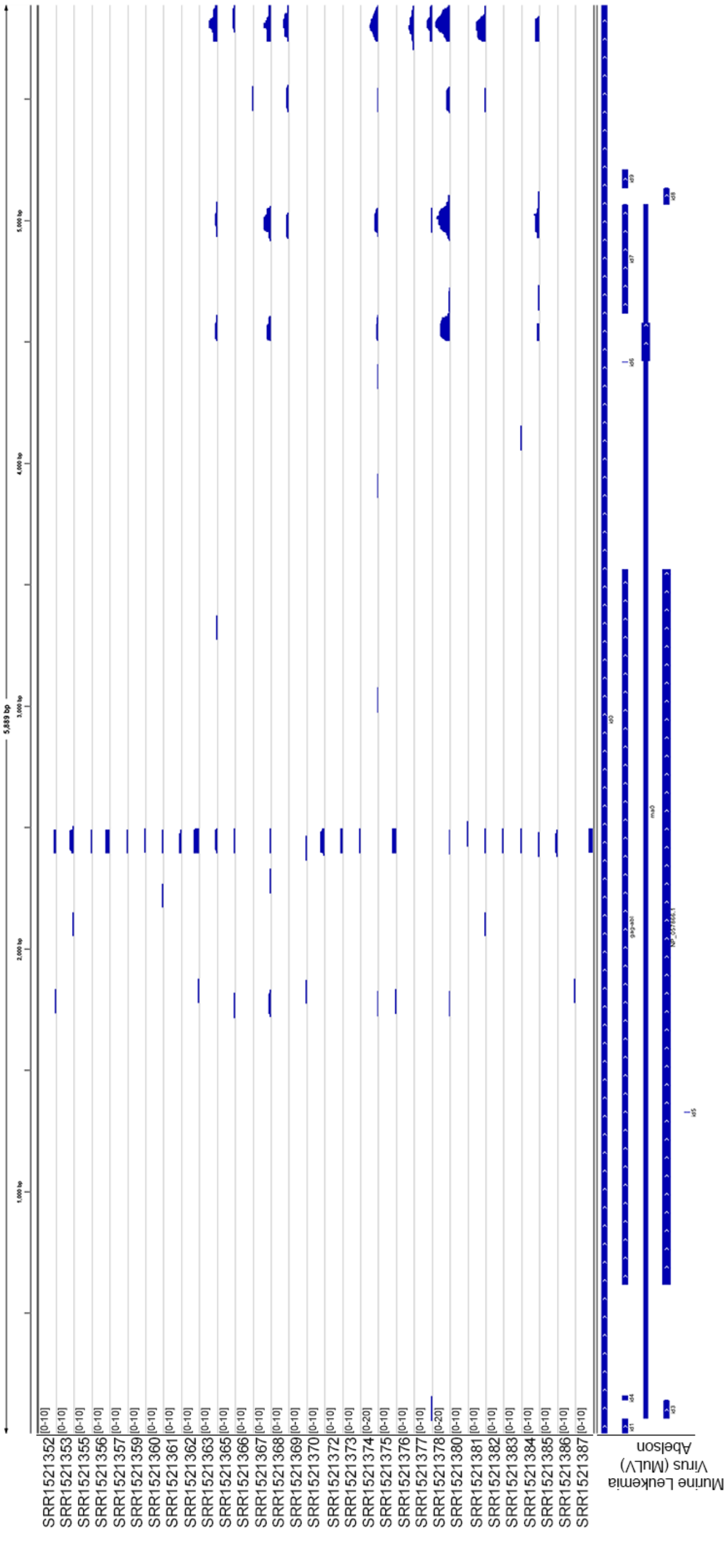

5,889 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
SRR1521352 [0-10]
SRR1521353 [0-10]
SRR1521355 [0-10]
SRR1521356 [0-10]
SRR1521357 [0-10]
SRR1521359 [0-10]
SRR1521360 [0-10]
SRR1521361 [0-10]
SRR1521362 [0-10]
SRR1521363 [0-10]
SRR1521365 [0-10]
SRR1521366 [0-10]
SRR1521367 [0-10]
SRR1521368 [0-10]
SRR1521369 [0-10]
SRR1521370 [0-10]
SRR1521372 [0-10]
SRR1521373 [0-10]
SRR1521374 [0-20]
SRR1521375 [0-10]
SRR1521376 [0-10]
SRR1521377 [0-10]
SRR1521378 [0-20]
SRR1521380 [0-10]
SRR1521381 [0-10]
SRR1521382 [0-10]
SRR1521383 [0-10]
SRR1521384 [0-10]
SRR1521385 [0-10]
SRR1521386 [0-10]
SRR1521387 [0-10]
Murine Leukemia Virus (MuLV) Abelson
gag-abl
NP_057866.1

Figure 5. MuLV gene coverage analysis for MRI-localized non-enhancing biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the MuLV Abelson genome (GenBank accession number NC_001499). The y-axis represents the number of reads at each nucleotide position in the genome. Samples SRR1521374 and SRR1521378 are displayed with a max read level of 20, while the max read level for the other samples is 10.

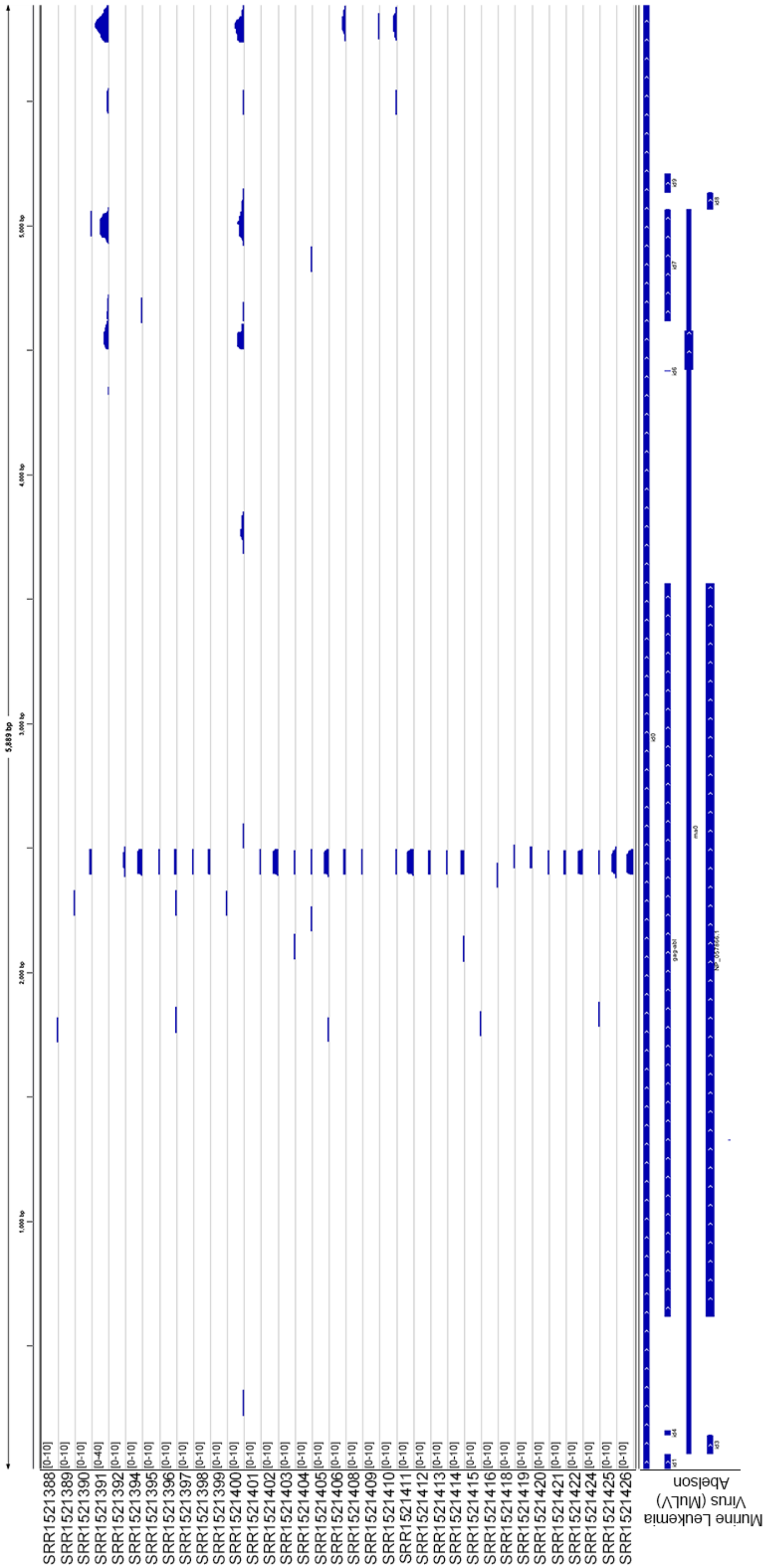

5,889 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
SRR1521388 [0-10]
SRR1521389 [0-10]
SRR1521390 [0-10]
SRR1521391 [0-40]
SRR1521392 [0-10]
SRR1521394 [0-10]
SRR1521395 [0-10]
SRR1521396 [0-10]
SRR1521397 [0-10]
SRR1521398 [0-10]
SRR1521399 [0-10]
SRR1521400 [0-10]
SRR1521401 [0-10]
SRR1521402 [0-10]
SRR1521403 [0-10]
SRR1521404 [0-10]
SRR1521405 [0-10]
SRR1521406 [0-10]
SRR1521408 [0-10]
SRR1521409 [0-10]
SRR1521410 [0-10]
SRR1521411 [0-10]
SRR1521412 [0-10]
SRR1521413 [0-10]
SRR1521414 [0-10]
SRR1521415 [0-10]
SRR1521416 [0-10]
SRR1521418 [0-10]
SRR1521419 [0-10]
SRR1521420 [0-10]
SRR1521421 [0-10]
SRR1521422 [0-10]
SRR1521424 [0-10]
SRR1521425 [0-10]
SRR1521426 [0-10]
Murine Leukemia Virus (MuLV) Abelson
gag-abl

Figure 6. MuLV gene coverage analysis for MRI-localized contrast biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the MuLV Abelson genome (GenBank accession number NC_001499). The y-axis represents the number of reads at each nucleotide position in the genome. Sample SRR1521391 is displayed with a max read level of 40, while the max read level for the other samples is 10.

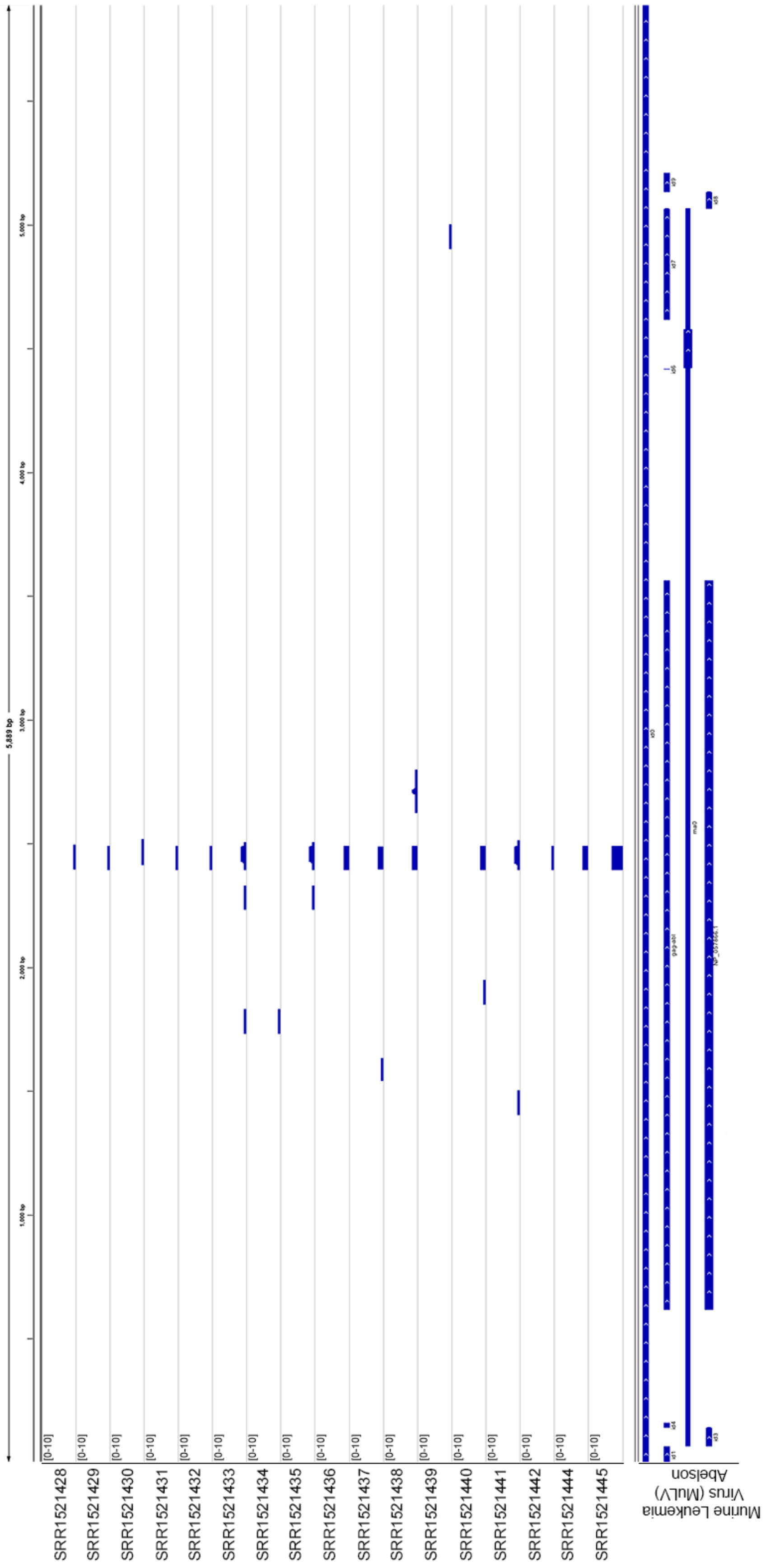

5,859 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
SRR1521428 [0-10]
SRR1521429 [0-10]
SRR1521430 [0-10]
SRR1521431 [0-10]
SRR1521432 [0-10]
SRR1521433 [0-10]
SRR1521434 [0-10]
SRR1521435 [0-10]
SRR1521436 [0-10]
SRR1521437 [0-10]
SRR1521438 [0-10]
SRR1521439 [0-10]
SRR1521440 [0-10]
SRR1521441 [0-10]
SRR1521442 [0-10]
SRR1521444 [0-10]
SRR1521445 [0-10]
Murine Leukemia Virus (MuLV) Abelson
gag-abl
NP_057866.1

Figure 7. MuLV gene coverage analysis for MRI-localized non-neoplasm biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the MuLV Abelson genome (GenBank accession number NC_001499). The y-axis represents the number of reads at each nucleotide position in the genome. Samples SRR1521374 and SRR1521378 are displayed with a max read level of 20, while the max read level for the other samples is 10.

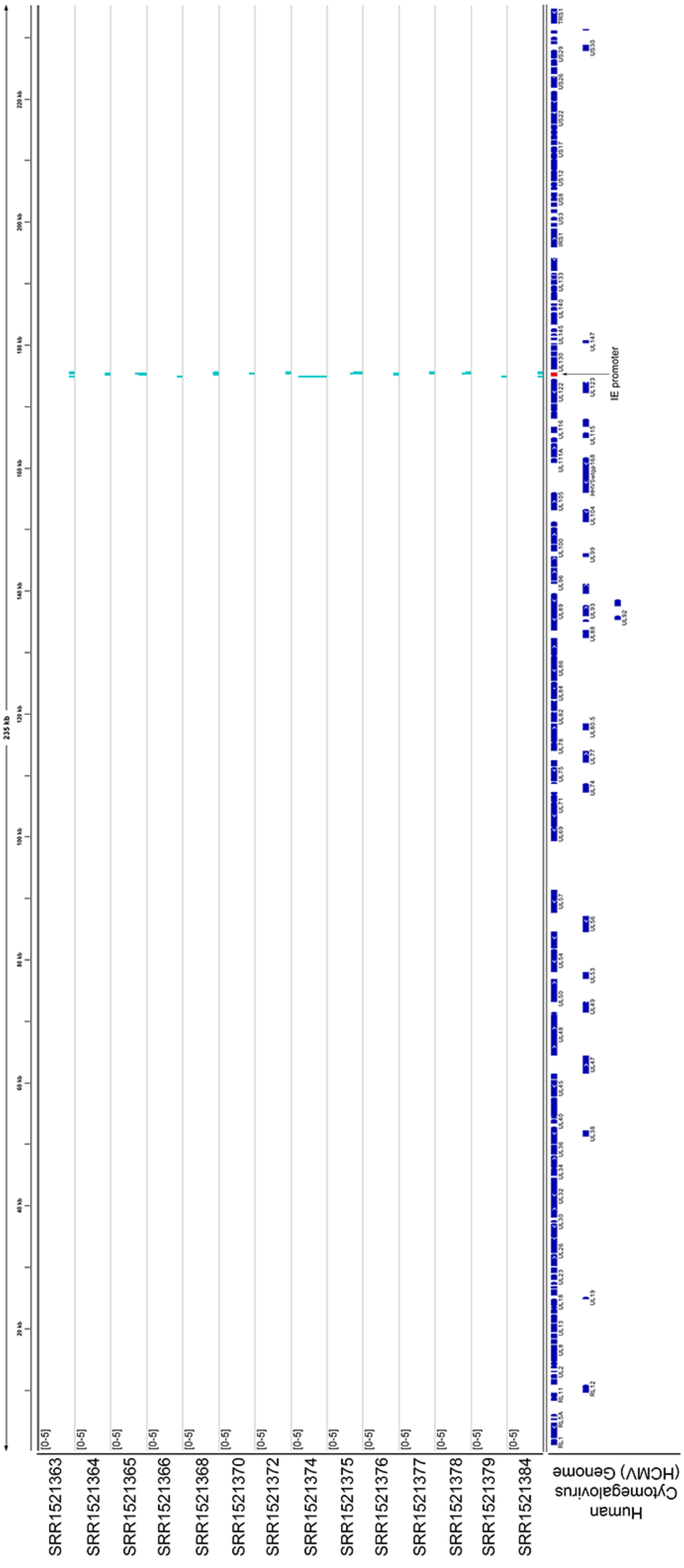

235 kb
20 kb
40 kb
60 kb
80 kb
100 kb
120 kb
140 kb
160 kb
180 kb
200 kb
220 kb
SRR1521363
SRR1521364
SRR1521365
SRR1521366
SRR1521368
SRR1521370
SRR1521372
SRR1521374
SRR1521375
SRR1521376
SRR1521377
SRR1521378
SRR1521379
SRR1521384
[0-5]
Human Cytomegalovirus (HCMV) Genome
RL1
RL5A
RL11
UL2
UL8
UL13
UL18
UL23
UL26
UL30
UL32
UL34
UL36
UL40
UL43
UL48
UL50
UL54
UL57
UL69
UL71
UL75
UL79
UL82
UL84
UL86
UL89
UL96
UL100
UL105
UL111A
UL116
UL122
UL130
UL145
UL140
UL133
IRS1
US3
US8
US12
US17
US22
US26
US29
TRS1
RL12
UL19
UL38
UL47
UL49
UL53
UL56
UL74
UL77
UL80.5
UL88
UL93
UL92
UL99
UL104
HHV5wtgp168
UL115
UL123
UL147
US30
IE promoter

Figure 8. HCMV gene coverage analysis for MRI-localized non-enhancing biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the HCMV genome (GenBank accession number NC_006273). The y-axis represents the number of reads at each nucleotide position in the genome. Samples are displayed with a max read level of 5.

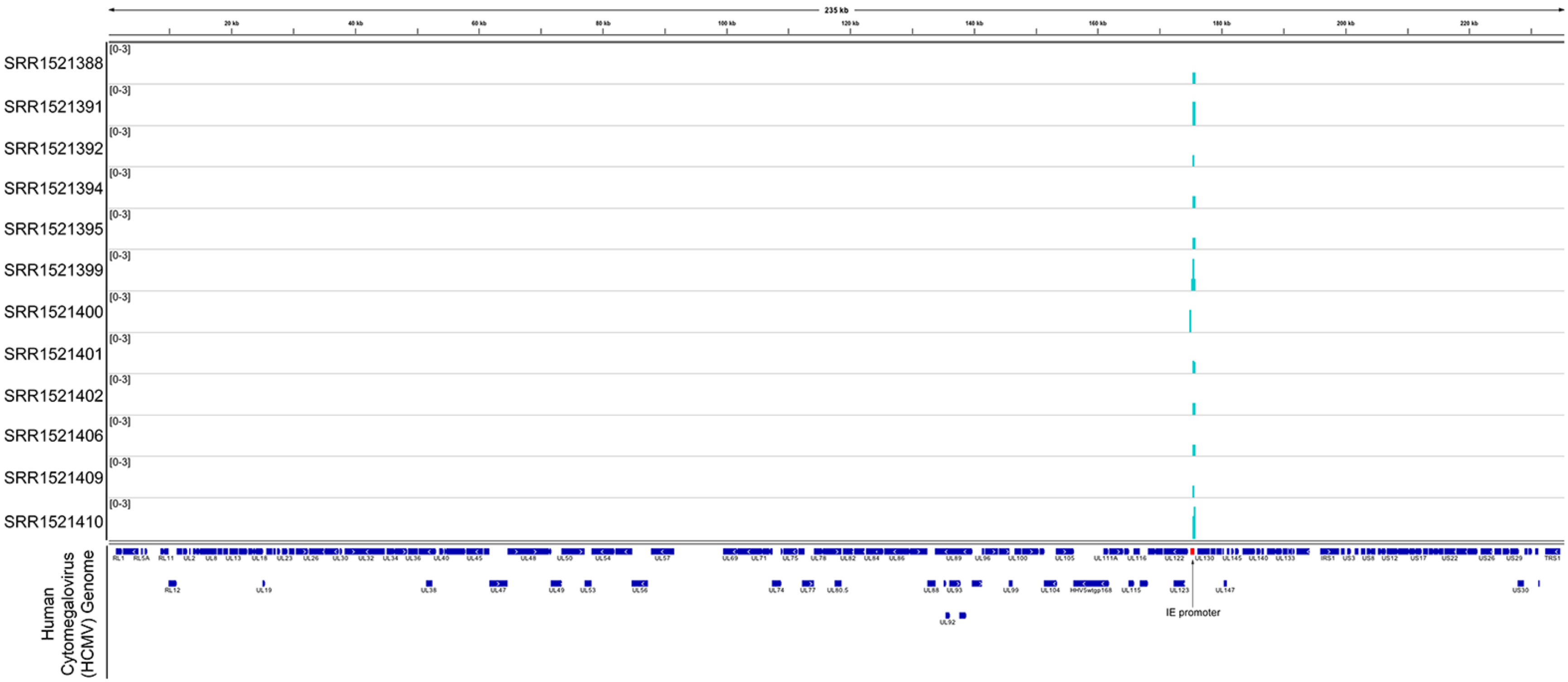

235 kb
20 kb
40 kb
60 kb
80 kb
100 kb
120 kb
140 kb
160 kb
180 kb
200 kb
220 kb
SRR1521388
SRR1521391
SRR1521392
SRR1521394
SRR1521395
SRR1521399
SRR1521400
SRR1521401
SRR1521402
SRR1521406
SRR1521409
SRR1521410
[0-3]
Human Cytomegalovirus (HCMV) Genome
RL1 RL5A RL11 UL2 UL8 UL13 UL18 UL23 UL26 UL30 UL32 UL34 UL36 UL40 UL45 UL48 UL50 UL54 UL57 UL69 UL71 UL75 UL78 UL82 UL84 UL86 UL89 UL96 UL100 UL105 UL111A UL116 UL122 UL130 UL145 UL140 UL133 IRS1 US3 US8 US12 US17 US22 US26 US29 TRS1
RL12 UL19 UL38 UL47 UL49 UL53 UL56 UL74 UL77 UL80.5 UL88 UL93 UL99 UL104 HHV5wtgp168 UL115 UL123 UL147 US30
UL92
IE promoter

Figure 9. HCMV gene coverage analysis for MRI-localized contrast biopsies. Data was displayed using the Integrative Genomics Viewer (IGV) using the HCMV genome (GenBank accession number NC_006273). The y-axis represents the number of reads at each nucleotide position in the genome. Samples are displayed with a max read level of 3.

6.4.4 Analysis of GBM stem cell subpopulations

Due to the low abundance and limited detection of HCMV in GBMs reported in the literature, some groups have proposed that HCMV is harbored in only a small number of tumor cells, specifically the CD133+ tumor stem cells [260-262]. To address this possibility, we analyzed RNA-seq datasets generated from a cohort of short-term glioma stem-like cell cultures freshly isolated from nine patients diagnosed with primary GBM. Analysis of these datasets using our STAR/virome pipeline showed the detection of human Adenovirus C (HAdV-C) in 9 out of the 24 samples with the remaining 15 samples showing low abundance HAdV-C read levels. HAdV-C was also detected in the same 9 samples using RNA CoMPASS. Despite detecting HAdV-C in these samples, viral read coverage was primarily limited to the same three small regions of the genome in the majority of samples (Figure 10). Further manual blast analysis of the HAdV-C reads demonstrated homology with laboratory adenovirus vectors (data not shown). Although we detected only low level HCMV reads in 3 glioma stem-like cell culture samples (0.03 – 0.1 RPMH), we further analyzed these reads due to the historical significance of this virus with GBM. Similar to our previous HCMV findings discussed above, viral read coverage analysis again showed that all reads aligned with the HCMV IE1 promoter (Figure 11). RNA CoMPASS analysis did not identify HCMV reads in these samples.

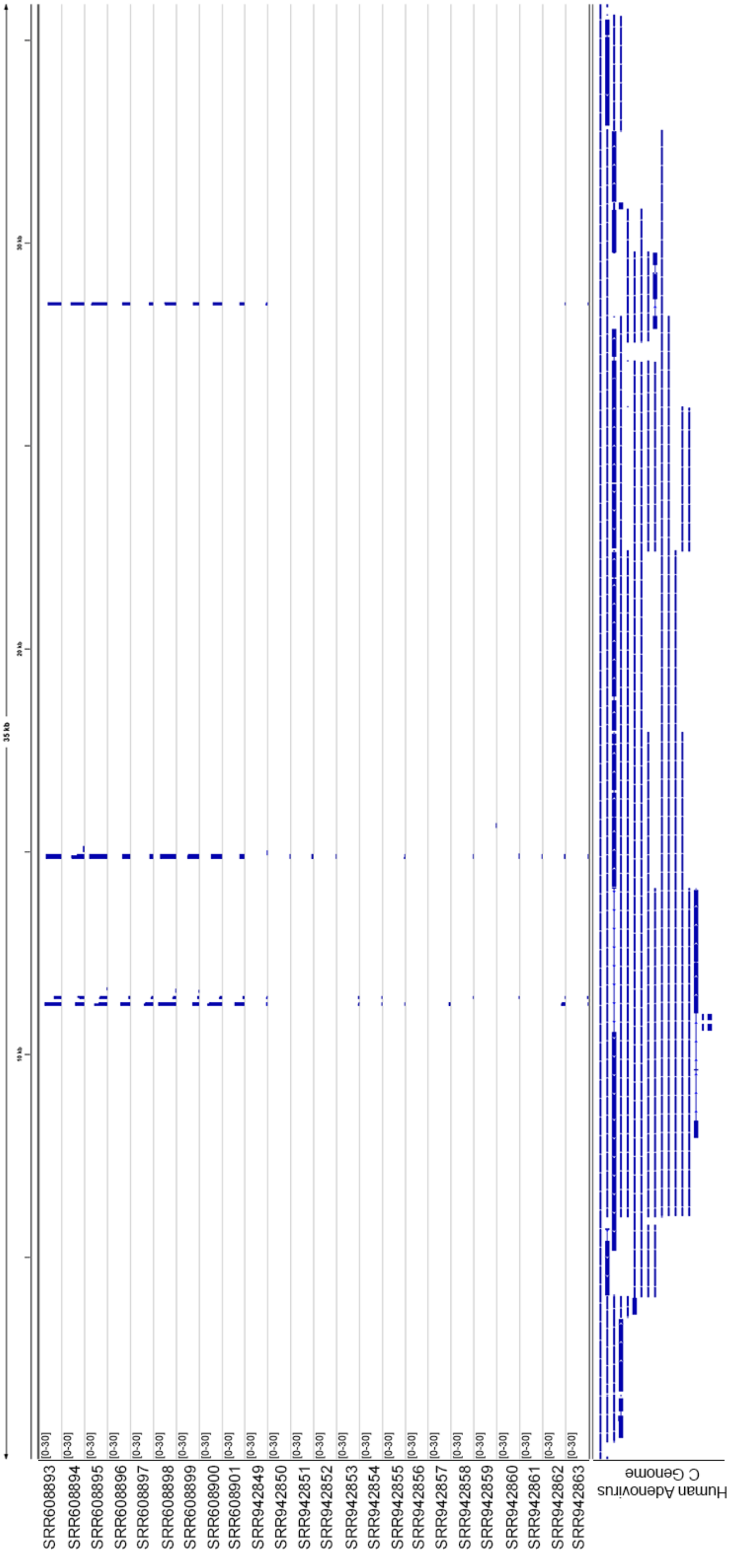

35 kb
10 kb
20 kb
30 kb
SRR608893 [0-30]
SRR608894 [0-30]
SRR608895 [0-30]
SRR608896 [0-30]
SRR608897 [0-30]
SRR608898 [0-30]
SRR608899 [0-30]
SRR608900 [0-30]
SRR608901 [0-30]
SRR942849 [0-30]
SRR942850 [0-30]
SRR942851 [0-30]
SRR942852 [0-30]
SRR942853 [0-30]
SRR942854 [0-30]
SRR942855 [0-30]
SRR942856 [0-30]
SRR942857 [0-30]
SRR942858 [0-30]
SRR942859 [0-30]
SRR942860 [0-30]
SRR942861 [0-30]
SRR942862 [0-30]
SRR942863 [0-30]
Human Adenovirus C Genome

Figure 10. Human Adenovirus C gene coverage analysis for gliomas stem cell samples. Data was displayed using the Integrative Genomics Viewer (IGV) using the Human Adenovirus C genome (GenBank accession number NC_001405). The y-axis represents the number of reads at each nucleotide position in the genome. Samples are displayed with a max read level of 30.

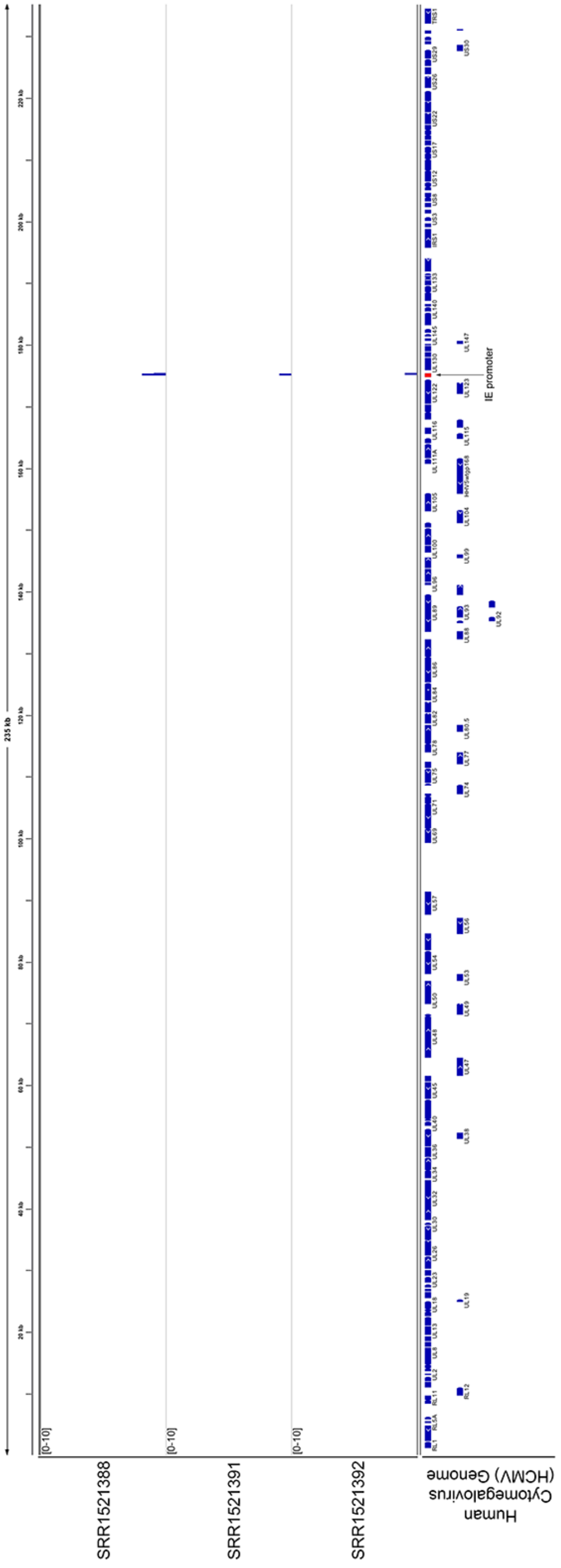

235 kb
20 kb
40 kb
60 kb
80 kb
100 kb
120 kb
140 kb
160 kb
180 kb
200 kb
220 kb
[0-10]
[0-10]
[0-10]
SRR1521388
SRR1521391
SRR1521392
Human Cytomegalovirus (HCMV) Genome
IE promoter
RL1
RL5A
RL11
UL2
UL8
UL13
UL18
UL23
UL26
UL30
UL32
UL34
UL36
UL40
UL45
UL48
UL50
UL54
UL57
UL69
UL71
UL75
UL78
UL82
UL84
UL86
UL89
UL96
UL100
UL105
UL111A
UL116
UL122
UL130
UL145
UL140
UL133
IRS1
US3
US8
US12
US17
US22
US26
US29
TRS1
RL12
UL19
UL38
UL47
UL49
UL53
UL56
UL74
UL77
UL80.5
UL88
UL93
UL92
UL99
UL104
HHV5wtgp168
UL115
UL123
UL147
US30

Figure 11. HCMV gene coverage analysis for gliomas stem cell samples. Data was displayed using the Integrative Genomics Viewer (IGV) using the HCMV genome (GenBank accession number NC_006273). The y-axis represents the number of reads at each nucleotide position in the genome. Samples are displayed with a max read level of 10.

6.4.5 Virome analysis of TCGA low-grade gliomas

To study the virome of low-grade gliomas (LGG), RNA-seq datasets from 514 primary and 17 recurrent LGGs from the TCGA were analyzed [60]. Due to the magnitude of the sample number, we exclusively used the sensitive yet rapid approach of our STAR/virome method. Based on this analysis, Human Papillomavirus (HPV) 16 was detected in 3 out of the 514 samples (1.5 – 2.4 RPMH), although lower HPV-16 read numbers were observed in 22 additional samples (Figure 1). Inspection of HPV-16 read coverage for the 3 positive samples showed expression of the HPV-16 E6 and E7 oncogenes with lower coverage of the E1, E2 and E4 regions (Figure 12). The lack of any coverage of the right half of the HPV-16 genome is consistent with deletion of this region which is frequently observed in oncogenic HPV genome integrations (where viral integration and the concomitant deletion of these negative regulatory genes results in increased expression of the oncogenic E6 and E7 genes while at the same time preventing productive viral infection and host cell destruction). We also detected HPV-58 in 1 sample (1.9 RPMH) with lower HPV-58 reads in 2 additional samples (Figure 1). Inspection of viral read coverage for the HPV-58 positive samples showed that the single sample with read levels above the 0.7 RPMH threshold had good read coverage of the E6, E7, and the E2/E4/E5 region (Figure 13). Finally, Hepatitis B reads were detected in 1 sample (13 RPMH), although below threshold Hepatitis B reads were detected in 2 additional samples (Figure 1). Inspection of Hepatitis B read coverage for the above threshold sample showed good read coverage within the

HBVgp1/HBVgp2/HBVgp3 region where coverage abruptly stops. Furthermore, the majority of HBV reads align to the HBVgp3 gene, which encodes the regulatory HBx protein (Figure 14). Like the above findings of read mapping being primarily limited to viral HPV oncogenes, the HBV mapping data are suggestive of oncogenic viral integration. No virus associations were observed with the 17 recurrent LGGs (Figure 1).

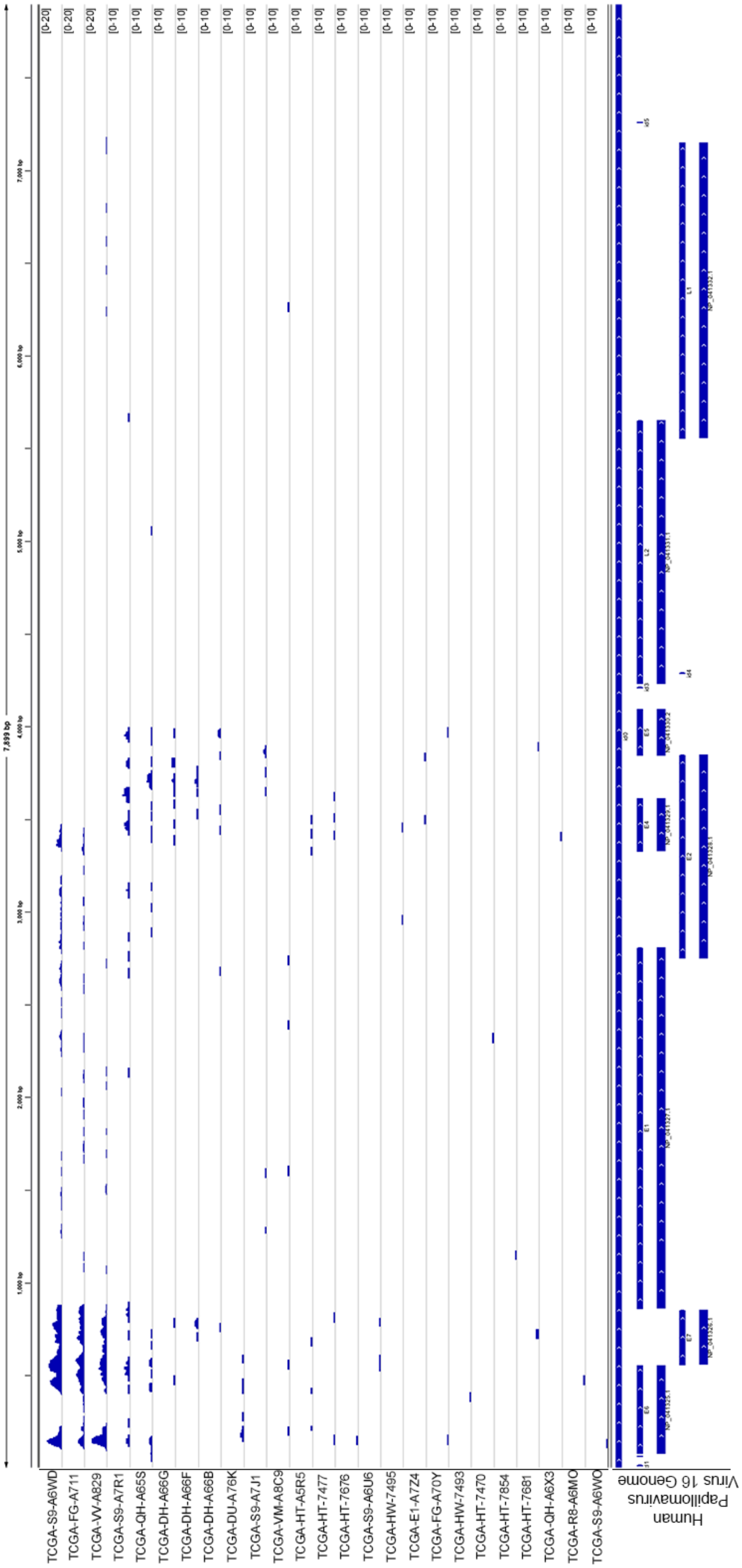

7,899 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
6,000 bp
7,000 bp
TCGA-S9-A6WD [0-20]
TCGA-FG-A711 [0-20]
TCGA-VV-A829 [0-20]
TCGA-S9-A7R1 [0-10]
TCGA-QH-A65S [0-10]
TCGA-DH-A66G [0-10]
TCGA-DH-A66F [0-10]
TCGA-DH-A66B [0-10]
TCGA-DU-A76K [0-10]
TCGA-S9-A7J1 [0-10]
TCGA-VM-A8C9 [0-10]
TCGA-HT-A5R5 [0-10]
TCGA-HT-7477 [0-10]
TCGA-HT-7676 [0-10]
TCGA-S9-A6U6 [0-10]
TCGA-HW-7495 [0-10]
TCGA-E1-A7Z4 [0-10]
TCGA-FG-A70Y [0-10]
TCGA-HW-7493 [0-10]
TCGA-HT-7470 [0-10]
TCGA-HT-7854 [0-10]
TCGA-HT-7681 [0-10]
TCGA-QH-A6X3 [0-10]
TCGA-R8-A6MO [0-10]
TCGA-S9-A6WO [0-10]
Human Papillomavirus 16 Genome
id0
id1
E6
NP_041325.1
E7
NP_041326.1
E1
NP_041327.1
E2
NP_041328.1
E4
NP_041329.1
E5
NP_041330.2
id3
id4
L2
NP_041331.1
L1
NP_041332.1
id5

Figure 12. HPV-16 gene coverage analysis for low grade gliomas. Data was displayed using the Integrative Genomics Viewer (IGV) using the HPV-16 genome (GenBank accession number NC_001526). The y-axis represents the number of reads at each nucleotide position in the genome. Samples TCGA-S9-A6WD, TCGA-FG-A711, and TCGA-VV-A829 are displayed with a max read level of 20, while the max read level for the other samples is 10.

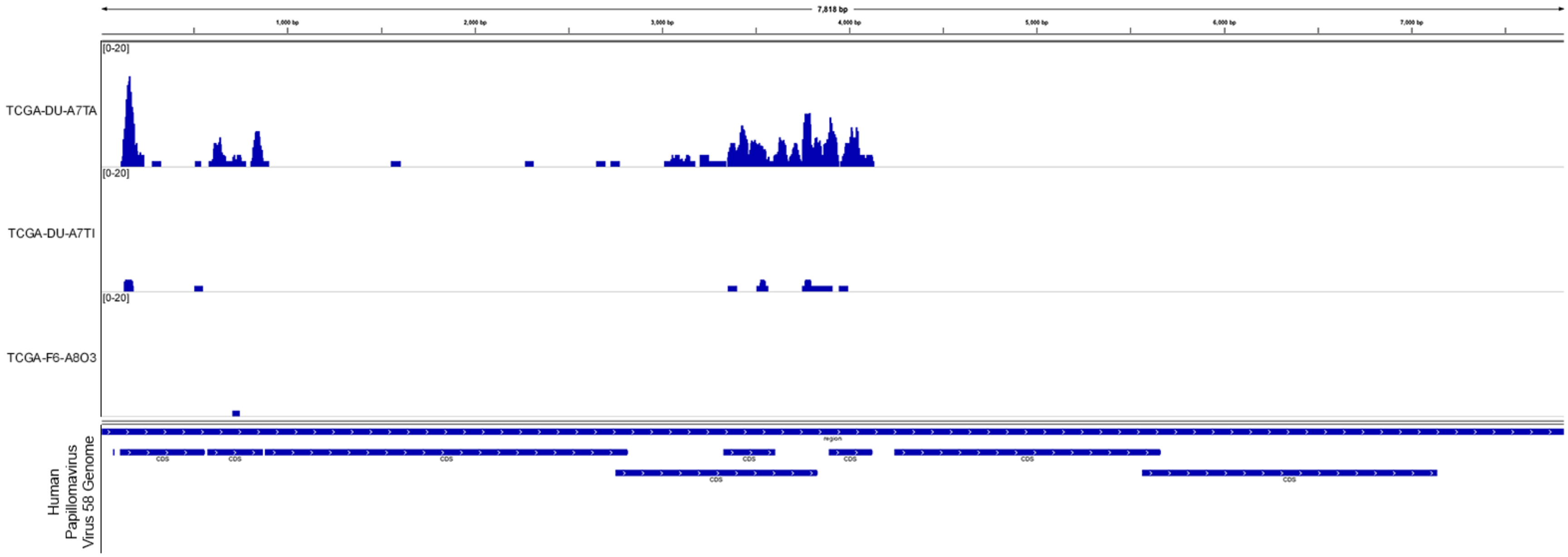

7,818 bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
6,000 bp
7,000 bp
[0-20]
TCGA-DU-A7TA
[0-20]
TCGA-DU-A7TI
[0-20]
TCGA-F6-A8O3
Human Papillomavirus Virus 58 Genome
region
CDS
CDS
CDS
CDS
CDS
CDS
CDS
CDS

Figure 13. HPV-58 gene coverage analysis for low grade gliomas. Data was displayed using the Integrative Genomics Viewer (IGV) using the HPV-58 genome (GenBank accession number NC_00). The y-axis represents the number of reads at each nucleotide position in the genome. Samples are displayed with a max read level of 20.

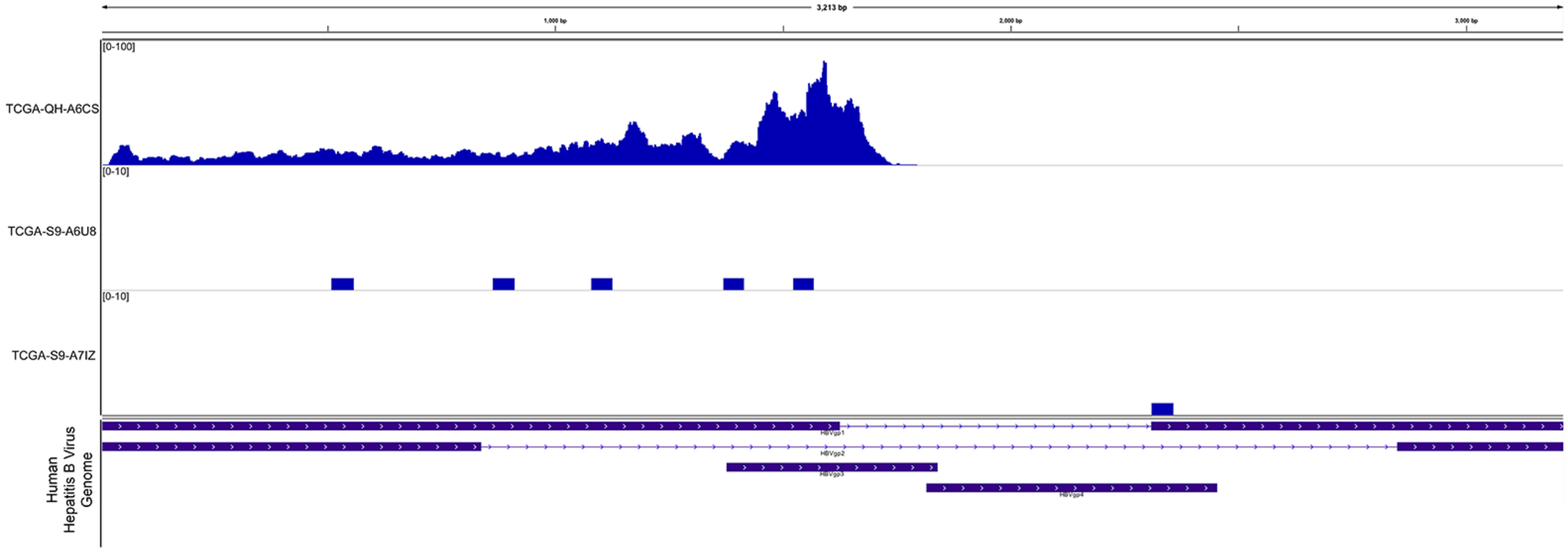

3,213 bp
1,000 bp
2,000 bp
3,000 bp
[0-100]
TCGA-QH-A6CS
[0-10]
TCGA-S9-A6U8
[0-10]
TCGA-S9-A7IZ
Human
Hepatitis B Virus
Genome
HBVgp1
HBVgp2
HBVgp3
HBVgp4

Figure 14. Hepatitis B gene coverage analysis for low grade gliomas. Data was displayed using the Integrative Genomics Viewer (IGV) using the Hepatitis B genome (GenBank accession number NC_003977). The y-axis represents the number of reads at each nucleotide position in the genome. Sample TCGA-QH-A6CS is displayed with a max read level of 100, while the max read level for the other samples is 10.

6.4.6 Lack of virus association in The Cancer Genome Atlas GBM whole genome datasets.

To explore the possibility that viruses infecting brain tissue become transcriptionally dormant, resulting in the lack of virus detection in RNA-seq datasets, we subjected the TCGA GBM whole genome sequencing (WGS) datasets to our virome analysis pipeline. Analysis of viruses at the DNA level was relatively unremarkable with the highest read numbers in the primary (TP) GBM WGS datasets being derived from EBV and HAdV-C (Figure 15). The highest viral read numbers identified in a recurrent (TR) GBM WGS datasets was derived from EBV with 1,454 total viral reads (Figure 15). Two HCMV viral reads were detected in 1 TP GBM WGS dataset (TCGA-14-1823) but not the corresponding blood matched normal (N) control. Manual blast of these HCMV reads from the tumor demonstrated homology to HCMV laboratory expression vector sequences, which was also demonstrated by Tang et al [236].

EBV reads were detected in 9 TP GBM WGS datasets for which 6 out of 9 samples analyzed had corresponding blood matched samples. We detected EBV in 4 additional normal blood samples and 3 TR GBM WGS datasets. Upon further analysis of the raw EBV reads from the TP GBM datasets, the viral reads were found to be homologous to the EBV genome based on blasting analysis. EBV read coverage analysis for the 3 TR GBM samples displayed viral genome coverage across the entire genome for 2 of the samples (Figure 16). Torque Teno Virus reads were incidentally identified in three normal blood samples. Finally, although a high level of HHV- 6 and 7 reads were detected in all WGS

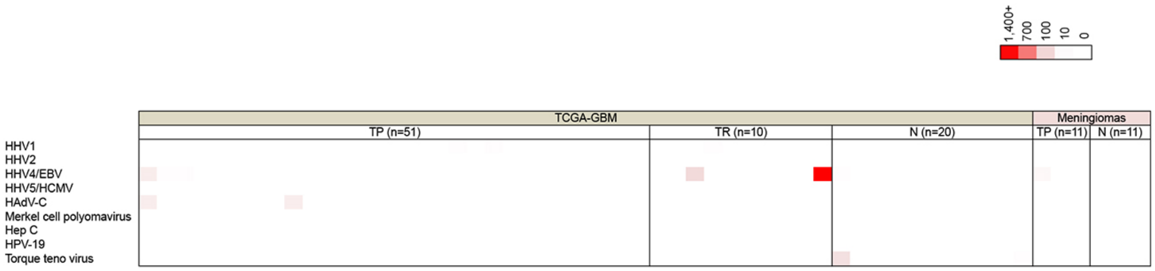


Figure 15. Heat map showing the total reads for brain tissue whole genome sequencing datasets. Color intensity represents relative viral reads across all datasets.

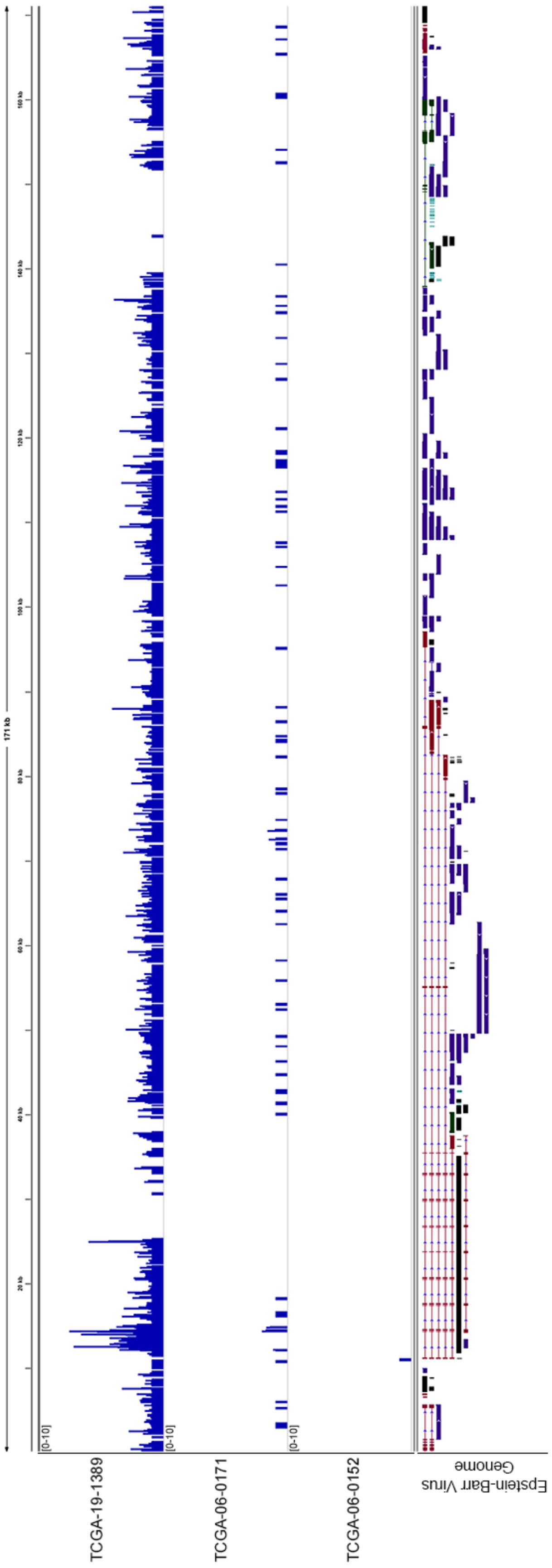

171 kb
20 kb
40 kb
60 kb
80 kb
100 kb
120 kb
140 kb
160 kb
[0-10]
[0-10]
[0-10]
TCGA-19-1389
TCGA-06-0171
TCGA-06-0152
Epstein-Barr Virus Genome

Figure 16. EBV gene coverage for recurrent GBM samples. Data was displayed using the Integrative Genomics Viewer (IGV) using the (A) EBV Akata genome (GenBank accession number KC207813). The y-axis represents the number of reads at each nucleotide position in the genome.

samples, manual inspection of the raw sequence reads showed that they are likely derived from human chromosomal telomeric-like repeats, TAACCC (data not shown).

6.4.7 Lack of virus association in meningioma whole genome datasets

To determine if other brain tumors are associated with viruses, we analyzed 64 WGS datasets from 11 patients with grade I meningiomas and their matched blood control samples. Analysis of the WGS datasets from the meningioma samples demonstrated no confirmed association with viruses (Figure 15). Similar to our analysis of the TCGA GBM WGS datasets, a low abundance of EBV reads were detected in 3 tumor samples and 4 normal blood samples, raising the possibility that these reads came from infiltrating B-cells. Furthermore, HHV-6 and HHV-7 reads were detected in all 22 samples but consisted of the simple sequence repeat, TAACCC suggesting that these reads likely originated from human telomeric repeats (data not shown).

6.5 Discussion

Although there was an agreement reached from the HCMV/GBM symposium in 2011, emerging studies using NGS to assess the viral association with GBM has been unable to recapitulate this association [60, 77, 236-238]. In line with these previous studies, our data further supports no direct viral association with GBM. There may be a possible association of HPV-16, HPV-58, and Hepatitis B with LGGs, however additional validation studies are required before any conclusions can be drawn from our initial assessment. Furthermore, based on the low abundance of viral reads that were identified in these cases, whether these viruses are truly associated with LGGs and not derived from sequencing contamination is unclear. Finally, although the viral detection threshold that we set for the RNA-seq datasets is relatively low (0.07 RPMH), all HCMV read findings were analyzed further irrespective of how low the HCMV read level and were found to be likely derived from laboratory plasmid contamination.

The hallmark of herpesviruses, and their key to persistence within their host, is their ability to switch to highly restricted gene expression patterns that allow avoidance of the immune system. To overcome the potential problem of missing viral infections due to this type of viral adaptation, WGS datasets were analyzed. Nevertheless, this approach also failed to identify any meaningful virus associations in the analyzed samples. This is in contrast to a report by Amirian et al. in which they identified HHV-6A and HHV-6B in the WGS datasets from TCGA [242]. Another study conducted by Cimino et al. also identified HHV-6 and

EBV DNA when they analyzed unmapped reads from a NGS-based comprehensive oncology panel [237]. Although our initial investigation detected the presence of HHV-6 and HHV-7 viruses, further analysis of these viral reads revealed all reads consisted of human chromosomal telomeric-like repeats, TAACCC. Although HHV 6 and 7 have sequences homologous to this region, no other regions of the viral genome were represented in the sequence datasets. This is highly suggestive that these reads originated from the telomeric region of human chromosomes rather than representing bona fide HHV6 or HHV7 infection.

EBV DNA reads were identified in a number of the TCGA DNA-seq datasets including 9 TP GBM WGS samples, 6 normal matched blood WGS samples plus 4 additional normal blood WGS samples, and 3 TR GBM WGS samples. In addition, we identified EBV DNA reads in 3 grade I meningioma samples and 4 normal blood samples. All EBV DNA reads identified were low in abundance with 1 – 39 reads detected in primary GBM samples, 1-5 reads detected in normal blood samples, and 1 – 15 reads detected in grade 1 meningiomas, a result similar to the findings of Cimino et al in which they identified 1 – 18 EBV reads in 5 GBM samples [237]. We identified 3 TR GBM samples using WGS datasets that were EBV positive, with 1 of these datasets showing moderate EBV levels (1454 viral reads), another showing minimal EBV levels (80 viral reads), and the last dataset had 1 EBV viral read. Although these three TR GBM WGS datasets were positive for EBV, the corresponding RNA-seq datasets for these samples failed to validate these findings. Without tissue to confirm these findings, it is impossible to determine the origin of these viral reads and we do not feel confident in associating EBV with these TR GBM samples. In addition, based on our past experience in the field of EBV, if EBV was truly associated, we would likely see greater than 10 viral RPMH for RNA-seq and thousands of viral reads for DNA-seq [61, 258]. Finally, given the ubiquitous nature of EBV, the low viral read counts, and the presence of EBV in both tumor and blood samples in relatively equal proportions, we postulate that the EBV reads that were detected likely originated from EBV infected B-cells localized in the tumor stroma and/or from library or sequencing sample cross-contamination. Due to the nature of GBM, there is a possibility for a preponderance of necrotic tissue within the tumor bulk, resulting in the effective dilution of tumor cells and tumor associated viruses; which could be argued as an explanation for the lack of strong viral detection. However, given the large number of samples analyzed and the careful procurement protocols utilized by TCGA, it is unlikely that the majority of samples fall within this scenario. Further supporting this contention, our analysis of the MRI-localized GBM biopsies from Gill et al [253] did not detect any known viruses and there were no differences between samples obtained from the core (presumably more necrotic) and those samples obtained from the tumor margin (presumably less necrotic, with active tumor growth and neoangiogenesis).

The identification of HPV-16, HPV-58 and HBV in a small portion of LGG RNA-seq datasets is a potentially interesting finding. Analysis of the clinical data from these patients using cBioPortal [263, 264] demonstrated that the majority of virus positive samples were oligodendrogliomas (3 out of the 5 samples) from White males with an average age of 42 (Supplemental File 8). The demographics are relatively consistent with the whole LGG cohort (55% males, 92% White, and average age 43). Tumor type varied slightly from the whole cohort, which consisted of 193 astrocytomas (38%), 130 oligoastrocytomas (25%), and 191 oligodendrogliomas (37%). In addition, although the genetic profile of these patients demonstrates a variety of alterations, some of the more common alterations observed in the entire cohort (e.g., IDH1, IDH2, ATRX, and TP53) were not observed in these patients with HPV or HBV reads (see reference [249] for additional details regarding LGG samples). The lack of mutation of one or more of these in tumors with detected virus could be due to viral subversion of the corresponding pathways, obviating the need for somatic mutations (for example, through HPV E6 mediated inhibition of the p53 pathway). Nevertheless, further investigation into the association between viruses, HPV and HBV and LGGs is warranted.

Both HPV-16 and HPV-58 are considered high-risk HPV types, which are causative agents in the development of cervical carcinoma. The likely mechanism of action for both HPV-16 and HPV-58 is viral integration into the host genome [265, 266]. Coverage analysis of the HPV positive LGG datasets indicate that some of the samples display evidence of integration with disruption of the viral E1 gene (Figures 12-13) with all samples with HPV reads showing the majority of read coverage mapping to viral E6 and E7 oncogenes. Due to the low viral read numbers detected in our study, additional validation experiments are warranted to determine if there is truly an association between LGGs and HPV or whether these findings represent sample cross-contamination with true HPV associated samples.

Like HPV, the mechanism of action for HBV is also integration into the host genome. Visual analysis of the HBV positive LGG datasets demonstrated robust gene coverage within the HBVgp1/HBVgp2/HBVgp3 region with an abrupt termination of gene coverage after HBVgp3 (Figure 14). Further, the majority of reads align within the HBVgp3 gene, which encodes the regulatory HBx protein. Previous studies have shown that HBx plays a critical role in the pathogenesis of hepatocellular carcinoma [267, 268]. While this observation is also of potential interest, given the fact that adequate HBV reads were detected in only 1 sample out of 514 LGG datasets (0.19%), further analysis is necessary to validate this observation.

The RNA CoMPASS analysis of the auxiliary brain tissue sequencing datasets provided a full metatranscriptomic profile including bacterial, fungal, and viral reads. Although we only presented data on the virome in this study, a complete metatranscriptomic analysis was performed. Although reads for several bacterial species were identified in the datasets, it has been our experience that the source of many of these reads are from environmental contamination [244, 245, 269] and do not represent true associations.

Due to reports of an association between HCMV and GBM, immunotherapy treatments against HCMV were considered a logical next step as an exhilarating new avenue for cancer therapy. There are several clinical trials in the United States in various stages of completion focused on targeted HCMV therapy in GBM patients. While we await the results of these clinical trials, the results from the valganciclovir treatment of glioblastoma patients in Sweden (VIGAS) study, a randomized, double-blinded, placebo-controlled trial was recently published showing trends but no significant differences in tumor volumes between the valganciclovir (an anti-CMV drug) and placebo groups at 3 and 6 months [270]. However, in a retrospective analysis of the same cohort with additional patients taking valganciclovir for compassionate reasons, the rate of survival of treated patients at 2 years was 62% as compared with 18% of contemporary controls with a similar disease stage, surgical-resection grade, and baseline treatment [271]. Although these are remarkable results, questions have been raised as to the interpretation of the data and whether this survival rate is misleading [272].

Several recent publications have highlighted the lack of association between viruses, specifically HCMV, and GBM [60, 77, 221, 223, 236, 238, 239, 241, 273]. Based on our comprehensive analysis, we substantiate these claims. Given the austerity of recent evidence against a HCMV etiology for GBM, moving forward, we caution against the use of anti-CMV therapy for GBM patients until this issue is completely resolved.

Chapter 7: Discussion

7.1 NGS technology in deciphering oncogenic pathogens in the context of human malignancies

NGS is revolutionizing the way scientists discover and investigate oncogenic pathogens. Through the analysis of Big Data, several recent discoveries have validated the potential that NGS technology has on investigating oncogenic pathogens: the discovery of a novel Merkel cell polyomavirus in Merkel cell carcinoma [40] and the discovery of an association between *Fusobacterium* and colorectal carcinoma was made possible with two different NGS approaches [41, 42]. Both of these approaches were facilitated by the use of computational subtraction approaches, whereby reads aligning to reference genomes were subtracted from the sequence file, resulting in sequences from undiscovered organisms. By applying an automated computational pipeline to pathogen sleuthing, scientists are able to assess hundreds to thousands of biological samples relatively quickly. Several automated computational pipelines have been designed, including the work conducted in Chapter 2 on RNA CoMPASS, for the analysis of exogenous sequences and for pathogen discovery [39, 41, 46-48].

Although several sequence-based computational subtraction pipelines are used mainly for pathogen discovery, RNA CoMPASS takes advantage of the richness of RNA-seq data to provide host transcript expression data in addition to pathogen analysis. This simultaneous dual assessment of host and pathogen transcripts leverages the unique characteristics of RNA-seq technology. In designing RNA CoMPASS, its capability was evaluated by analyzing a cohort of Burkitt's lymphoma samples and human B-cells infected with EBV. In this study, we were able to demonstrate the gene coverage of EBV and determine the differentially expressed human genes between the two cohorts clearly representing the lymphoblastoid and Burkitt's phenotypes.

Another example of the utility of RNA CoMPASS is presented in Chapter 3 in which we analyzed 118 non-AIDS non-Hodgkin's lymphoma samples (NHLs) and 13 follicular lymphomas samples from the Cancer Genome Characterization Initiative (CGCI) using RNA CoMPASS. As expected, we detected EBV in 4/118 NHLs (3.4%), which is relatively consistent with previously published reports [274-276]. In addition, we identified 2 samples with HHV-6B infections with one of these samples being co-infected with EBV. Serendipitously, cluster analysis of these EBV positive samples based on EBV gene expression alone showed unique clustering of the samples with high versus low EBV read counts. Further analysis demonstrated a high lytic to latent read ratio in the samples with low EBV versus high EBV read counts (Figure 1C in Chapter 3), suggesting these reads possibly reflect low level reactivation in infiltrating latent B-cells. The ability to use NGS-based technology to determine level of viral infection will be discussed further in section 7.3.

7.2 Contamination Issues in Sequence Datasets

During the course of metatranscriptomic studies performed over the past several years, we invariably noted surprising levels of bacterial reads whether the genetic material was derived from human clinical specimens, tissue culture cells, or animal tissues. The extent and pervasiveness of this observation led to the work conducted in Chapter 4. We identified fairly extensive levels of bacterial reads across a variety of RNA-seq datasets analyzed with *Paracoccus denitrificans SD1* and *Acinetobacter* among the most prevalent bacteria. Interestingly, when the same RNA was prepared and sequenced at six different laboratories, the metatranscriptomic profile varies with bacterial reads differing as much as 30-fold [245]. Based on this analysis, we concluded that the bacterial reads were not derived from the specimens themselves but likely associated with environmental contamination in the operating room, during sample storage, sample processing, RNA preparation, or sequence library preparation.

Contamination issues have already had an impact on the very databases that are used for bioinformatics work. For example, Laurence et al. identified Bradyrhizobium sequences in assembled genomes in the NCBI Genome database [111]. Interestingly, Bradyrhizobium species along with other microbes, have been reported in ultrapure water systems and may help explain the presence of this microbe in several deposited genome assemblies. Another group found *Leucobacter sp.* sequences in assembled genomes of *Caenorhabditis sp.* [112]. Still another source of contaminating reads was discovered from silica column-based nucleic acid extraction kits, which harbored the NIH-CQV virus [113-117].

Due to the sensitive nature of NGS, microbial reads derived from sample/sequencing procedures have the potential to lead to data misinterpretations and false positive findings. Therefore, microbial contamination issues are relevant in sequencing experiments and warrant steps to minimize the source of this contamination. As outlined in Chapter 4, we proposed the following recommendations to combat potential sources of contamination:

1) Detection studies, especially with a diagnostic focus, should incorporate stringent SOPs across the entire experimental pipeline from sample collection to sequencing.

2) Highly purified metabolic enzymes and other reagents used in sequence library preparation should be used whenever possible.

3) Establishment of standards for the curation of microbial sequences submitted to Genbank and other large-scale databases in order to assess completeness and quality of the assembled genomes.

4) Contamination controls such as mock sequence library preparations should be used to help guide the development of appropriate and effective SOPs for metagenomic and metatranscriptomic studies.

7.3 Dual assessment of pathogen and host transcripts

As mentioned previously, the power of RNA-seq is the ability to simultaneously assess host and pathogen transcripts. This principle is exemplified well in the work conducted in Chapter 5. Using RNA CoMPASS, we investigated the role of EBV in the pathogenesis of gastric carcinoma using the TCGA gastric carcinoma cohort. Several important observations were garnered from this work that included both viral and host changes. EBV transcripts were detected in 17% of gastric carcinoma (EBVaGC) samples, but these samples varied significantly in EBV coverage depth. EBV transcript analysis demonstrated that transcripts from the BamHI A region of the EBV genome comprised the majority of EBV reads. Expression of LMP2 and LMP1, to a lesser extent, were also observed as was EBNA1 and evidence of abortive lytic replication. Although transcripts from the BamHI region of EBV were reported previously in a naturally infected EBV positive gastric carcinoma cell line [171], our analysis identified this observation in the setting of solid tumors. Based on our analysis, we were able to measure the magnitude of BamHI A region gene expression relative to other viral and cellular genes, which demonstrated strikingly high expression levels. Although others have been unable to detect protein from naturally expressed BamHI A rightward transcripts [172, 173], the high expression level of these transcripts in EBVaGC samples suggests a functional role in gastric adenocarcinomas, possibly as long non-coding RNAs (lncRNA).

In addition to EBV transcriptome analysis, host cellular RNA expression analysis was performed to determine EBV/host pathway interactions. The cellular analysis indicated high levels of immune cells, which has been reported previously using standard hematoxylin and eosin staining of tumor sections [120, 184]. These histological analyses confirmed the lymphocyte infiltration was predominately CD8+ T cells [185, 186]. Despite the high level of immune cell infiltrate observed in EBVaGCs, EBV and tumor cells are able to survive. Based on previous data and the work from Chapter 5, we proposed four separate compensatory immune evasion strategies employed by EBV infected tumor cells that allow the tumor cells to survive in the setting of high immune infiltrate. First, the limited expression of viral protein coding genes in EBVaGC may contribute to the avoidance of viral antigen processing and targeting [188]. Second, although the EBV encoded protein, EBNA1 is required for viral episomal replication/maintenance and therefore must be expressed in proliferating cells, it encodes a glycine-alanine repeat domain that blocks its proteasomal processing for CTL presentation [189, 190]. Third, in Chapter 5, we found that the levels of expression of the interferon-gamma (IFNG) inducible cytotoxic T-cell (CTL) and natural killer (NK) cell inhibitor, indoleamine 2,3-dioxygenase (IDO1) are high in EBVaGC. This is an important finding because IDO1 is a rate-limiting enzyme involved in the catabolism of tryptophan (Trp) [191]. CTLs and NK cells are uniquely sensitive to Trp depletion leading to the induction of stress responses and the inhibition of proliferation and activation [192, 193]. IDO1 functions to cause local tryptophan depletion under physiological and pathogenic immune tolerance settings such as during placentation and cancer [194, 195], where it is considered to be critical for establishing local immune tolerance. Among other candidate effectors, increased IFNG has been shown to induce IDO1 expression [196, 197].

As mentioned in Section 7.1, using RNA-seq, we are able to not only detect viral infections but also measure the magnitude of the infection. This is particularly important for EBV due to it unique latency gene profiles and ubiquitous nature. As previously mentioned, we were able to distinguish between tumors infected with EBV (demonstrating an EBV latency profile) and those samples that were EBV positive due to reactivated infiltrating B-cells harboring EBV (demonstrating a more lytic gene profile) (Figure 1C in Chapter 3). Using the same approach, we performed clustering analysis using only the EBV genes across the EBVaGC samples. This analysis revealed 4 samples clustering separately from the other EBVaGC samples (Figure 7 in Chapter 5). This apparently distinct gene expression profile observed in the 4 EBVaGC samples raises the possibility that these samples represent infection of a unique cell type relative to the other samples (possibly tumor cells versus stroma or reactivated B-cells). This new insight into determining the magnitude and type of EBV infection may provide clinical clues into treating and monitoring EBV association malignancies. EBV associated malignancies are typically diagnosed using traditional detection methods (e.g., IHC and PCR) relying mostly on the detection of the EBV gene EBER. Although EBER is highly expressed by EBV, resulting in easy detection, one limitation of this approach is the lack of viral tropism information regarding which cell types are infected with EBV. Since EBER is expressed regardless of cell type or latency gene expression, it is impossible to determine if the tumor and/or stroma harbor EBV. Using NGS-based approaches we are now able to glean this information that may lead to improved clinical outcomes for these patients.

7.4 Unbiased nature of NGS has potential to resolve controversies

RNA-seq also has the potential to resolve controversies surrounding previous findings in an unbiased manner. For example, the association between human cytomegalovirus (HCMV) and glioblastoma multiforme (GBM) was first reported in 2002 [220]. Since that time, there have been a lot of discrepancies in the literature surrounding this issue [60, 77, 220-242, 277]. The work conducted in Chapter 6 revolved around the concept that a NGS-based analysis of brain tumors, including GBM, would facilitate resolution of this lingering controversy. Based on our comprehensive analysis of several different brain tumor sequencing datasets, we conclude that there is no viral association with GBM. This is in line with several recent publications that also report the lack of association between viruses, specifically HCMV, and GBM [60, 77, 221, 223, 236, 238, 239, 241].

LIST OF REFERENCES

1. De Flora, S. and P. Bonanni, *The prevention of infection-associated cancers.* Carcinogenesis, 2011.

2. Liao, J., *Viruses and human cancer.* Yale Journal of Biology and Medicine, 2006. **79**(3-4): p. 115-122.

3. Scheffner, M., et al., *The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53.* Cell, 1990. **63**(6): p. 1129-1136.

4. McLaughlin-Drubin, M.E. and K. Münger, *The human papillomavirus E7 oncoprotein.* Virology, 2009. **384**(2): p. 335-344.

5. Howie, H.L., R.A. Katzenellenbogen, and D.A. Galloway, *Papillomavirus E6 proteins.* Virology, 2009. **384**(2): p. 324-334.

6. Gires, O., et al., *Latent membrane protein 1 of Epstein-Barr virus mimics a constitutively active receptor molecule.* The EMBO Journal, 1997. **16**(20): p. 6131-6140.

7. Uchida, J. and T. Yasui, *Mimicry of CD40 Signals by Epstein-Barr Virus LMP1 in B Lymphocyte Responses.* Science, 1999. **286**(5438): p. 300.

8. Eliopoulos, A.G., et al., *Epstein-Barr Virus-Encoded Latent Membrane Protein 1 Activates the JNK Pathway through Its Extreme C Terminus via a Mechanism Involving TRADD and TRAF2.* Journal of Virology, 1999. **73**(2): p. 1023-1035.

9. Gires, O., et al., *Latent membrane protein 1 of Epstein-Barr virus interacts with JAK3 and activates STAT proteins.* The EMBO Journal, 1999. **18**(11): p. 3064-3073.

10. Mitchell, T. and B. Sugden, *Stimulation of NF-kappa B-mediated transcription by mutant derivatives of the latent membrane protein of Epstein-Barr virus.* Journal of Virology, 1995. **69**(5): p. 2968-76.

11. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences of the United States of America, 1977. **74**(12): p. 5463-5467.

12. Mardis, E.R., *Next-Generation Sequencing Platforms.* Annual Review of Analytical Chemistry, 2013. **6**(1): p. 287-303.

13. Smith, L.M., et al., *Fluorescence detection in automated DNA sequence analysis.* Nature, 1986. **321**(6071): p. 674-679.

14. Mullis, K.B. and F.A. Faloona, *Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction*, in *Methods in Enzymology*, W. Ray, Editor. 1987, Academic Press. p. 335-350.

15. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: Lessons from Large-Scale Biology.* Science, 2003. **300**(5617): p. 286-290.

16. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

17. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-380.

18. Dressman, D., et al., *Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations.* Proceedings of the National Academy of Sciences, 2003. **100**(15): p. 8817-8822.

19. Fedurco, M., et al., *BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.* Nucleic Acids Research, 2006. **34**(3): p. e22.

20. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends in Genetics, 2008. **24**(3): p. 133-141.

21. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-352.

22. Levene, M.J., et al., *Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations.* Science, 2003. **299**(5607): p. 682-686.

23. Korlach, J., et al., *Chapter 20 - Real-Time DNA Sequencing from Single Polymerase Molecules*, in *Methods in Enzymology*, G.W. Nils, Editor. 2010, Academic Press. p. 431-455.

24. Eid, J., et al., *Real-Time DNA Sequencing from Single Polymerase Molecules.* Science, 2009. **323**(5910): p. 133-138.

25. Gouaux, J.E., et al., *Subunit stoichiometry of staphylococcal alpha-hemolysin in crystals and on membranes: a heptameric transmembrane pore.* Proceedings of the National Academy of Sciences, 1994. **91**(26): p. 12828-12831.

26. Song, L., et al., *Structure of Staphylococcal α-Hemolysin, a Heptameric Transmembrane Pore.* Science, 1996. **274**(5294): p. 1859-1866.

27. Kasianowicz, J.J., et al., *Characterization of individual polynucleotide molecules using a membrane channel.* Proceedings of the National Academy of Sciences, 1996. **93**(24): p. 13770-13773.

28. Wang, Y., Q. Yang, and Z. Wang, *The evolution of nanopore sequencing.* Frontiers in Genetics, 2014. **5**: p. 449.

29. Metzker, M.L., *Sequencing in real time.* Nat Biotech, 2009. **27**(2): p. 150-151.

30. Steinbock, L.J. and A. Radenovic, *The emergence of nanopores in next-generation sequencing.* Nanotechnology, 2015. **26**(7): p. 074003.

31. Fodor, S.P.A., et al., *Multiplexed biochemical assays with biological chips.* Nature, 1993. **364**(6437): p. 555-556.

32. Schena, M., et al., *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.* Science, 1995. **270**(5235): p. 467-470.

33. Faghihi, M.A. and C. Wahlestedt, *Regulatory roles of natural antisense transcripts.* Nat Rev Mol Cell Biol, 2009. **10**(9): p. 637-643.

34. Group, R.G.E.R., et al., *Antisense Transcription in the Mammalian Transcriptome.* Science, 2005. **309**(5740): p. 1564-1566.

35. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes & Development, 2011. **25**(18): p. 1915-1927.

36. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-227.

37. Li, S., et al., *Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data.* Genome Research, 2013. **23**(10): p. 1730-1739.

38. O'Grady, T., et al., *Global Bidirectional Transcription of the Epstein-Barr Virus Genome During Reactivation.* Journal of Virology, 2013.

39. Moore, R.A., et al., *The Sensitivity of Massively Parallel Sequencing for Detecting Candidate Infectious Agents Associated with Human Tissue.* PLoS ONE, 2011. **6**(5): p. e19838.

40. Feng, H., et al., *Clonal Integration of a Polyomavirus in Human Merkel Cell Carcinoma.* Science, 2008. **319**(5866): p. 1096-1100.

41. Kostic, A.D., et al., *PathSeq: software to identify or discover microbes by deep sequencing of human tissue.* Nat Biotech, 2011. **29**(5): p. 393-396.

42. Castellarin, M., et al., *Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.* Genome Research, 2012. **22**(2): p. 299-306.

43. Westermann, A.J., S.A. Gorski, and J. Vogel, *Dual RNA-seq of pathogen and host.* Nat Rev Micro, 2012. **10**(9): p. 618-630.

44. Tierney, L., et al., *An interspecies regulatory network inferred from simultaneous RNA-seq of Candida albicans invading innate immune cells.* Frontiers in Microbiology, 2012. **3**.

45. Coco, J.R., E. K. Flemington, and C. M. Taylor, *PARSES: A Pipeline for Analysis of RNA-Seq Exogenous Sequences*, in *Proceedings of the ISCA 3rd International Conference on Bioinformatics and Computational Biology*. 2011, BICoB-2011: Holiday Inn Downtown-Superdome, New Orleans, Louisiana, USA 2011. p. 196-200.

46. Weber, G., et al., *Identification of foreign gene sequences by transcript filtering against the human genome.* Nat Genet, 2002. **30**(2): p. 141-142.

47. Xu, Y., et al., *Pathogen discovery from human tissue by sequence-based computational subtraction.* Genomics, 2003. **81**(3): p. 329-335.

48. Feng, H., et al., *Human Transcriptome Subtraction by Using Short Sequence Tags To Search for Tumor Viruses in Conjunctival Carcinoma.* Journal of Virology, 2007. **81**(20): p. 11332-11340.

49. Concha, M., et al., *Identification of New Viral Genes and Transcript Isoforms during Epstein-Barr Virus Reactivation using RNA-Seq.* Journal of Virology, 2012. **86**(3): p. 1458-1467.

50. Nix, D., S. Courdy, and K. Boucher, *Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks.* BMC Bioinformatics, 2008. **9**(1): p. 523.

51. Xu, G., et al., *SAMMate: a GUI tool for processing short read alignments in SAM/BAM format.* Source Code for Biology and Medicine, 2011. **6**(1): p. 2.

52. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Research, 2012. **40**(D1): p. D130-D135.

53. Huson, D.H., et al., *Integrative analysis of environmental sequences using MEGAN4.* Genome Research, 2011. **21**(9): p. 1552-1560.

54. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

55. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.

56. Deng, N., et al., *Isoform-level microRNA-155 target prediction using RNA-seq.* Nucleic Acids Research, 2011. **39**(9): p. e61-e61.

57. Nguyen, T., et al. *iQuant: A fast yet accurate GUI tool for transcript quantification*. in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*. 2011.

58. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.

59. Birol, I., et al., *De novo transcriptome assembly with ABySS.* Bioinformatics, 2009. **25**(21): p. 2872-2877.

60. Tang, K.-W., et al., *The landscape of viral expression and host gene fusion and adaptation in human cancer.* Nat Commun, 2013. **4**.

61. Strong, M.J., et al., *Differences in Gastric Carcinoma Microenvironment Stratify According to EBV Infection Intensity: Implications for Possible Immune Adjuvant Therapy.* PLoS Pathog, 2013. **9**(5): p. e1003341.

62. Dave, S.S., et al., *Molecular Diagnosis of Burkitt's Lymphoma.* New England Journal of Medicine, 2006. **354**(23): p. 2431-2442.

63. Hummel, M., et al., *A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling.* New England Journal of Medicine, 2006. **354**(23): p. 2419-2430.

64. Li, Z., et al., *A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells.* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8164-8169.

65. Schuhmacher, M., et al., *The transcriptional program of a human B cell line in response to Myc.* Nucleic Acids Research, 2001. **29**(2): p. 397-406.

66. Faumont, N., et al., *c-Myc and Rel/NF-kB Are the Two Master Transcriptional Systems Activated in the Latency III Program of Epstein-Barr Virus-Immortalized B Cells.* Journal of Virology, 2009. **83**(10): p. 5014-5027.

67. Adhikary, S. and M. Eilers, *Transcriptional regulation and transformation by Myc proteins.* Nature Reviews Molecular Cell Biology, 2005. **6**(8): p. 635-645.

68. Huen, D., et al., *The Epstein-Barr virus latent membrane protein-1 (LMP1) mediates activation of NF-kappa B and cell surface phenotype via two effector regions in its carboxy-terminal cytoplasmic domain.* Oncogene, 1995. **10**(3): p. 549-560.

69. Cahir-McFarland, E.D., et al., *Role of NF-kB in Cell Survival and Transcription of Latent Membrane Protein 1-Expressing or Epstein-Barr Virus Latency III-Infected Cells.* Journal of Virology, 2004. **78**(8): p. 4108-4119.

70. Schlee, M., et al., *c-MYC Impairs Immunogenicity of Human B Cells*, in *Advances in Cancer Research*, F.V.W. George and K. George, Editors. 2007, Academic Press. p. 167-188.

71. Schmitz, R., et al., *Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics.* Nature, 2012. **490**(7418): p. 116-120.

72. Burgess, D.J., et al., *Topoisomerase levels determine chemotherapy response in vitro and in vivo.* Proceedings of the National Academy of Sciences, 2008. **105**(26): p. 9053-9058.

73. Kwak, L.W., et al., *Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis.* Journal of Clinical Oncology, 1990. **8**(6): p. 963-77.

74. Azim, H.A., et al., *High dose intensity doxorubicin in aggressive non-Hodgkin's lymphoma: a literature-based meta-analysis.* Annals of Oncology, 2010. **21**(5): p. 1064-1071.

75. Ogata, M., *Human Herpesvirus 6 in Hematological Malignancies.* Journal of Clinical and Experimental Hematopathology, 2009. **49**(2): p. 57-67.

76. Lin, Z., et al., *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.* Journal of Virology, 2012. **86**(6): p. 2970-2977.

77. Khoury, J.D., et al., *Landscape of DNA Virus Associations across Human Malignant Cancers: Analysis of 3,775 Cases Using RNA-Seq.* Journal of Virology, 2013. **87**(16): p. 8916-8926.

78. Mueller, N.E., A. Mohar, and A. Evans, *Viruses Other than HIV and Non-Hodgkin's Lymphoma.* Cancer Research, 1992. **52**(19 Supplement): p. 5479s-5481s.

79. de Sanjosé, S., et al., *Epstein-Barr virus infection and risk of lymphoma: Immunoblot analysis of antibody responses against EBV-related proteins in a large series of lymphoma subjects and matched controls.* International Journal of Cancer, 2007. **121**(8): p. 1806-1812.

80. Hardell, K., et al., *Concentrations of organohalogen compounds and titres of antibodies to Epstein-Barr virus antigens and the risk for non-Hodgkin lymphoma.* Oncology Reports, 2009. **21**(6): p. 1567-1576.

81. Pozdnyakova, O., et al., *Epstein-Barr Virus-associated Diffuse Large B-Cell Lymphoma in an Immunocompetent Woman.* Journal of Clinical Oncology, 2010. **28**(5): p. e75-e78.

82. Morin, R.D., et al., *Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma.* Nature, 2011. **476**(7360): p. 298-303.

83. Xu, G., et al., *RNA CoMPASS: A Dual Approach for Pathogen and Host Transcriptome Analysis of RNA-Seq Datasets.* PLoS ONE, 2014. **9**(2): p. e89445.

84. Lin, Z., et al., *Whole-Genome Sequencing of the Akata and Mutu Epstein-Barr Virus Strains.* Journal of Virology, 2013. **87**(2): p. 1172-1182.

85. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2012.

86. Saeed, A., et al., *TM4: a free, open-source system for microarray data management and analysis.* Biotechniques, 2003. **34**(2): p. 374-8.

87. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotech, 2011. **29**(1): p. 24-26.

88. Hong, G.K., et al., *Epstein-Barr Virus Lytic Infection Contributes to Lymphoproliferative Disease in a SCID Mouse Model.* Journal of Virology, 2005. **79**(22): p. 13993-14003.

89. Ma, S., et al., *An Epstein-Barr Virus (EBV) Mutant with Enhanced BZLF1 Expression Causes Lymphomas with Abortive Lytic EBV Infection in a Humanized Mouse Model.* Journal of Virology, 2012. **86**(15): p. 7976-7987.

90. Portis, T., et al., *The LMP2A signalosome--a therapeutic target for Epstein-Barr virus latency and associated disease.* Front Biosci., 2002. **1**(7): p. d414-26.

91. Dawson, C.W., R.J. Port, and L.S. Young, *The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC).* Seminars in Cancer Biology, 2012. **22**(2): p. 144-153.

92. Frappier, L., *Role of EBNA1 in NPC tumourigenesis.* Seminars in Cancer Biology, 2012. **22**(2): p. 154-161.

93. Saha, A. and E.S. Robertson, *Impact of EBV essential nuclear protein EBNA-3C on B-cell proliferation and apoptosis.* Future Microbiology, 2013. **8**(3): p. 323-352.

94. Kostic, A.D., et al., *Genomic analysis identifies association of Fusobacterium with colorectal carcinoma.* Genome Research, 2012. **22**(2): p. 292-298.

95. Strong, M.J., et al., *Epstein-Barr Virus and Human Herpesvirus 6 Detection in a non-Hodgkin's Diffuse Large B-Cell Lymphoma Cohort using RNA-Seq.* Journal of Virology, 2013.

96. Bhatt, A.S., et al., *Sequence-Based Discovery of Bradyrhizobium enterica in Cord Colitis Syndrome.* New England Journal of Medicine, 2013. **369**(6): p. 517-528.

97. Loman, N.J., et al., *A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic escherichia coli o104:h4.* JAMA, 2013. **309**(14): p. 1502-1510.

98. Hasman, H., et al., *Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples.* Journal of Clinical Microbiology, 2013.

99. Wilson, M.R., et al., *Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing.* New England Journal of Medicine, 2014. **370**(25): p. 2408-2417.

100. Fricke, W.F. and D.A. Rasko, *Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions.* Nat Rev Genet, 2014. **15**(1): p. 49-55.

101. Köser, C.U., et al., *Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology.* PLoS Pathogens, 2012. **8**(8): p. e1002824.

102. t Hoen, P.A.C., et al., *Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories.* Nat Biotech, 2013. **31**(11): p. 1015-1022.

103. Network, T.C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

104. Network, T.C.G.A.R., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-525.

105. Network, T.C.G.A., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-337.

106. Nakazato, H., S. Venkatesan, and M. Edmonds, *Polyadenylic acid sequences in E. coli messenger RNA.* Nature, 1975. **256**(5513): p. 144-6.

107. Srinivasan, P., M. Ramanarayanan, and E. Rabbani, *Presence of polyriboadenylate sequences in pulse-labeled RNA of Escherichia coli.* Proceedings of the National Academy of Sciences, 1975. **72**(8): p. 2910-4.

108. Ohta, N., M. Sanders, and A. Newton, *Poly(adenylic acid) sequences in the RNA of Caulobacter crescenus.* Proc Natl Acad Sci U S A, 1975. **72**(6): p. 2343-6.

109. Sarkar, N., *Polyadenylation of mRNA in bacteria.* Microbiology, 1996. **142**: p. 3125-3133.

110. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-511.

111. Laurence, M., C. Hatzis, and D.E. Brash, *Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes.* PLoS ONE, 2014. **9**(5): p. e97876.

112. Percudani, R., *A Microbial Metagenome (Leucobacter sp.) in Caenorhabditis Whole Genome Sequences.* Bioinformatics and Biology Insights, 2013. **7**(3557-BBI-A-Microbial-Metagenome-Leucobacter-sp.-in-Caenorhabditis-Whole-Genom2.pdf): p. 55-72.

113. Xu, B., et al., *Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing.* Proceedings of the National Academy of Sciences, 2013. **110**(25): p. 10264-10269.

114. Smuts, H., et al., *Novel Hybrid Parvovirus-Like Virus, NIH-CQV/PHV, Contaminants in Silica Column-Based Nucleic Acid Extraction Kits.* Journal of Virology, 2014. **88**(2): p. 1398.

115. Naccache, S.N., et al., *The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns.* Journal of Virology, 2013. **87**(22): p. 11966-11977.

116. Naccache, S.N., et al., *Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis.* Proceedings of the National Academy of Sciences, 2014. **111**(11): p. E976.

117. Zhi, N., et al., *Reply to Naccache et al: Viral sequences of NIH-CQV virus, a contamination of DNA extraction method.* Proceedings of the National Academy of Sciences, 2014. **111**(11): p. E977.

118. Burke, A., et al., *Lymphoepithelial carcinoma of the stomach with Epstein-Barr virus demonstrated by polymerase chain reaction.* Mod Pathol, 1990. **3**(3): p. 377-80.

119. Shibata, D. and L. Weiss, *Epstein-Barr virus-associated gastric adenocarcinoma.* Am J pathol, 1992. **140**(4): p. 769-774.

120. Tokunaga, M., et al., *Epstein-Barr virus in gastric carcinoma.* Am J Pathol, 1993. **143**(5): p. 1250-4.

121. Morewaya, J., et al., *Epstein-Barr virus-associated gastric carcinoma in Papua New Guinea.* Oncol Rep, 2004. **12**(5): p. 1093-8.

122. Tang, W., et al., *Epstein-barr virus infected gastric adenocarcinoma expresses latent and lytic viral transcripts and has a distinct human gene expression profile.* Infectious Agents and Cancer, 2012. **7**(1): p. 21.

123. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207-210.

124. Prachason, T., et al., *Activation of Indoleamine 2,3-Dioxygenase in Patients with Scrub Typhus and Its Role in Growth Restriction of Orientia tsutsugamushi.* PLoS Negl Trop Dis, 2012. **6**(7): p. e1731.

125. Dolan, A., et al., *The genome of Epstein-Barr virus type 2 strain AG876.* Virology, 2006. **350**(1): p. 164-170.

126. Sugiura, M., et al., *Transcriptional analysis of Epstein-Barr virus gene expression in EBV-positive gastric carcinoma: unique viral latency in the tumour cells.* Br J Cancer, 1996. **74**(4): p. 625-631.

127. Luo, B., et al., *Expression of Epstein-Barr virus genes in EBV-associated gastric carcinomas* World J Gastroenterol, 2005. **11**(5): p. 629-633.

128. Shin, W., et al., *Epstein-Barr virus-associated gastric adenocarcinomas among Koreans.* Am J pathol, 1996. **105**(2): p. 174-81.

129. Harn, H., et al., *Epstein-Barr virus-associated gastric adenocarcinoma in Taiwan.* Hum Pathol, 1995. **26**(3): p. 267-71.

130. Lee, M., et al., *Detection of Epstein-Barr virus by PCR and expression of LMP1, p53, CD44 in gastric cancer.* Korean J Intern Med, 2004. **19**(1): p. 43-7.

131. Lin, Z., et al., *Quantitative and Qualitative RNA-Seq-Based Evaluation of Epstein-Barr Virus Transcription in Type I Latency Burkitt's Lymphoma Cells.* Journal of Virology, 2010. **84**(24): p. 13053-13058.

132. Chu, G., et al., *Gastrointestinal tract specific gene GDDR inhibits the progression of gastric cancer in a TFF1 dependent manner.* Molecular and Cellular Biochemistry, 2012. **359**(1): p. 369-374.

133. Hong, S.-J., et al., *DNA Methylation Patterns of Ulcer-Healing Genes Associated with the Normal Gastric Mucosa of Gastric Cancers.* J Korean Med Sci, 2010. **25**(3): p. 405-417.

134. Katuri, V., et al., *Inactivation of ELF//TGF-[beta] signaling in human gastrointestinal cancer.* Oncogene, 2005. **24**(54): p. 8012-8024.

135. Okugawa, T., et al., *Down-Regulation of Claudin-3 Is Associated with Proliferative Potential in Early Gastric Cancers.* Digestive Diseases and Sciences, 2012. **57**(6): p. 1562-1567.

136. Sentani, K., et al., *Upregulation of HOXA10 in gastric cancer with the intestinal mucin phenotype: reduction during tumor progression and favorable prognosis.* Carcinogenesis, 2012. **33**(5): p. 1081-1088.

137. Schofield, D., et al., *Correlation of loss of heterozygosity at chromosome 9q with histological subtype in medulloblastomas.* Am J Pathol, 1995. **146**(2): p. 472-480.

138. Bralten, L.B.C., et al., *The CASPR2 cell adhesion molecule functions as a tumor suppressor gene in glioma.* Oncogene, 2010. **29**(46): p. 6138-6148.

139. Yu, G., et al., *CSR1 Suppresses Tumor Growth and Metastasis of Prostate Cancer.* Am J Pathol, 2006. **168**(2): p. 597-607.

140. Moniz, S., et al., *Protein kinase WNK2 inhibits cell proliferation by negatively modulating the activation of MEK1//ERK1//2.* Oncogene, 2007. **26**(41): p. 6071-6081.

141. Mlakar, V., et al., *Oligonucleotide DNA Microarray Profiling of Lung Adenocarcinoma Revealed Significant Downregulation and Deletions of Vasoactive Intestinal Peptide Receptor 1.* Cancer Investigation, 2009. **28**(5): p. 487-494.

142. Koenig-Hoffmann, K., et al., *High throughput functional genomics: Identification of novel genes with tumor suppressor phenotypes.* International Journal of Cancer, 2005. **113**(3): p. 434-439.

143. Caretti, A., et al., *DNA methylation and histone modifications modulate the B1,3 galactosyltransferase B3Gal-T5 native promoter in cancer cells.* The International Journal of Biochemistry &amp; Cell Biology, 2012. **44**(1): p. 84-90.

144. Tsunoda, S., et al., *Methylation of CLDN6, FBN2, RBP1, RBP4, TFPI2, and TMEFF2 in esophageal squamous cell carcinoma.* Oncology Reports, 2009. **21**(4): p. 1067-1073.

145. Taieb, D., et al., *ArgBP2-Dependent Signaling Regulates Pancreatic Cell Migration, Adhesion, and Tumorigenicity.* Cancer Research, 2008. **68**(12): p. 4588-4596.

146. Hwang, S., et al., *Detection of HOXA9 gene methylation in tumor tissues and induced sputum samples from primary lung cancer patients.* Clin Chem Lab Med, 2011. **49**(4): p. 699-704.

147. Dmitriev, A., et al., *Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays.* Epigenetics, 2012. **7**(5): p. 502-13.

148. Tang, Y., et al., *FOXA2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers.* Cell Res, 2011. **21**(2): p. 316-326.

149. Lucas, B., et al., *HNF4[alpha] reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma.* Oncogene, 2005. **24**(42): p. 6418-6431.

150. Zuo, H., et al., *Downregulation of Rap1GAP through Epigenetic Silencing and Loss of Heterozygosity Promotes Invasion and Progression of Thyroid Tumors.* Cancer Research, 2010. **70**(4): p. 1389-1397.

151. Liu, Q.-S., et al., *Lentiviral-mediated miRNA against liver-intestine cadherin suppresses tumor growth and invasiveness of human gastric cancer.* Cancer Science, 2010. **101**(8): p. 1807-1812.

152. Kang, J.M., et al., *CDX1 and CDX2 Expression in Intestinal Metaplasia, Dysplasia and Gastric Cancer.* J Korean Med Sci, 2011. **26**(5): p. 647-653.

153. Keld, R., et al., *PEA3/ETV4-related transcription factors coupled with active ERK signalling are associated with poor prognosis in gastric adenocarcinoma.* Br J Cancer, 2011. **105**(1): p. 124-130.

154. Vangamudi, B., et al., *Regulation of B-catenin by t-DARPP in upper gastrointestinal cancer cells.* Mol Cancer, 2011. **10**: p. 32.

155. Lee, S.-A., et al., *Transmembrane 4 L six family member 5 (TM4SF5) enhances migration and invasion of hepatocytes for effective metastasis.* Journal of Cellular Biochemistry, 2010. **111**(1): p. 59-66.

156. Ruan, J., et al., *Inhibition of glypican-3 expression via RNA interference influences the growth and invasive ability of the MHCC97-H human hepatocellular carcinoma cell line.* Int J Mol Med, 2011. **28**(4): p. 497-503.

157. Hatano, M., et al., *Deregulation of a Homeobox Gene, HOX11, by the t(10;14) in T Cell Leukemia.* Science, 1991. **253**(5015): p. 79-82.

158. Tsuji, K., et al., *PEG10 is a probable target for the amplification at 7q21 detected in hepatocellular carcinoma.* Cancer Genetics and Cytogenetics, 2010. **198**(2): p. 118-125.

159. Louis, I., et al., *The Signaling Protein Wnt4 Enhances Thymopoiesis and Expands Multipotent Hematopoietic Progenitors through beta-Catenin-Independent Signaling.* Immunity, 2008. **29**(1): p. 57-67.

160. Nishikata, M., et al., *Carbonic anhydrase-related protein VIII promotes colon cancer cell growth.* Molecular Carcinogenesis, 2007. **46**(3): p. 208-214.

161. Collins, C., et al., *Positional cloning of ZNF217 and NABC1: Genes amplified at 20q13.2 and overexpressed in breast carcinoma.* Proceedings of the National Academy of Sciences, 1998. **95**(15): p. 8703-8708.

162. Kobayashi, T., et al., *A gene encoding a family with sequence similarity 84, member A (FAM84A) enhanced migration of human colon cancer cells.* Int J Oncol, 2006. **29**(2): p. 341-7.

163. Stevenson, L.F., et al., *The deubiquitinating enzyme USP2a regulates the p53 pathway by targeting Mdm2.* EMBO J, 2007. **26**(4): p. 976-986.

164. Olsen, C., et al., *Hedgehog-interacting protein is highly expressed in endothelial cells but down-regulated during angiogenesis and in several human tumors.* BMC Cancer, 2004. **4**(1): p. 43.

165. Furushima, K., et al., *Mouse homologues of Shisa antagonistic to Wnt and Fgf signalings.* Developmental Biology, 2007. **306**(2): p. 480-492.

166. Hu, T., et al., *Myristoylated Naked2 Antagonizes Wnt-B-Catenin Activity by Degrading Dishevelled-1 at the Plasma Membrane.* Journal of Biological Chemistry, 2010. **285**(18): p. 13561-13568.

167. Li, Y., et al., *LRP4 Mutations Alter Wnt-Catenin Signaling and Cause Limb and Kidney Malformations in Cenani-Lenz Syndrome.* American journal of human genetics, 2010. **86**(5): p. 696-706.

168. Muda, M., et al., *The Dual Specificity Phosphatases M3/6 and MKP-3 Are Highly Selective for Inactivation of Distinct Mitogen-activated Protein Kinases.* Journal of Biological Chemistry, 1996. **271**(44): p. 27205-27208.

169. Saksena, S., et al., *Mechanisms of transcriptional modulation of the human anion exchanger SLC26A3 gene expression by IFN-Œ≥.* American Journal of Physiology - Gastrointestinal and Liver Physiology, 2010. **298**(2): p. G159-G166.

170. Anderson, D.M., et al., *A homologue of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function.* Nature, 1997. **390**(6656): p. 175-179.

171. Jang, B.-G., E.J. Jung, and W.H. Kim, *Expression of BamHI-A Rightward Transcripts in Epstein-Barr Virus-Associated Gastric Cancers.* Cancer Res Treat, 2011. **43**(4): p. 250-254.

172. Al-Mozaini, M., et al., *Epstein-Barr virus BART gene expression.* Journal of General Virology, 2009. **90**(2): p. 307-316.

173. Smith, P.R., et al., *Structure and Coding Content of CST (BART) Family RNAs of Epstein-Barr Virus.* Journal of Virology, 2000. **74**(7): p. 3082-3092.

174. Cai, X., et al., *Epstein-Barr Virus MicroRNAs Are Evolutionarily Conserved and Differentially Expressed.* PLoS Pathog, 2006. **2**(3): p. e23.

175. Pfeffer, S., et al., *Identification of Virus-Encoded MicroRNAs.* Science, 2004. **304**(5671): p. 734-736.

176. Marquitz, A.R., et al., *Infection of Epstein-Barr virus in a gastric carcinoma cell line induces anchorage independence and global changes in gene expression.* Proceedings of the National Academy of Sciences, 2012. **109**(24): p. 9593-9598.

177. Gottwein, E., et al., *Viral MicroRNA Targetome of KSHV-Infected Primary Effusion Lymphoma Cell Lines.* Cell Host &amp; Microbe, 2011. **10**(5): p. 515-526.

178. Jones, R.J., et al., *Roles of lytic viral infection and IL-6 in early versus late passage lymphoblastoid cell lines and EBV-associated lymphoproliferative disease.* International Journal of Cancer, 2007. **121**(6): p. 1274-1281.

179. Ma, S.-D., et al., *A New Model of Epstein-Barr Virus Infection Reveals an Important Role for Early Lytic Viral Protein Expression in the Development of Lymphomas.* Journal of Virology, 2011. **85**(1): p. 165-177.

180. Taylor, N., et al., *Expression of the BZLF1 latency-disrupting gene differs in standard and defective Epstein-Barr viruses.* Journal of Virology, 1989. **63**(4): p. 1721-1728.

181. Jenson, H.B., P.J. Farrell, and G. Miller, *Sequences of the Epstein-Barr Virus (EBV) large internal repeat form the center of a 16-kilobase-pair palindrome of EBV (P3HR-1) heterogeneous DNA.* Journal of Virology, 1987. **61**(5): p. 1495-1506.

182. Jenson, H.B., M.S. Rabson, and G. Miller, *Palindromic structure and polypeptide expression of 36 kilobase pairs of heterogeneous Epstein-Barr virus (P3HR-1) DNA.* Journal of Virology, 1986. **58**(2): p. 475-486.

183. Horst, D.I., et al., *Specific Targeting of the EBV Lytic Phase Protein BNLF2a to the Transporter Associated with Antigen Processing Results in Impairment of HLA Class I-Restricted Antigen Presentation.* The Journal of Immunology, 2009. **182**(4): p. 2313-2324.

184. Oda, K., et al., *Association of Epstein-Barr virus with gastric carcinoma with lymphoid stroma.* Am J Pathol, 1993. **143**(4): p. 1063-1071.

185. van Beek, J., et al., *Morphological Evidence of an Activated Cytotoxic T-Cell Infiltrate in EBV-Positive Gastric Carcinoma Preventing Lymph Node Metastases.* The American Journal of Surgical Pathology, 2006. **30**(1): p. 59-65.

186. Saiki, Y., et al., *Immunophenotypic characterization of Epstein-Barr virus-associated gastric carcinoma:massive infiltration by proliferating CD8+ T-lymphocytes.* Lab Invest, 1996. **75**(1): p. 67-76.

187. Drake, C.G., E. Jaffee, and D.M. Pardoll, *Mechanisms of Immune Evasion by Tumors*, in *Advances in Immunology*, G.D. James P. Allison and W.A. Frederick, Editors. 2006, Academic Press. p. 51-81.

188. Thorley-Lawson, D.A. and A. Gross, *Persistence of the Epstein-Barr Virus and the Origins of Associated Lymphomas.* New England Journal of Medicine, 2004. **350**(13): p. 1328-1337.

189. Levitskaya, J., et al., *Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1.* Nature, 1995. **375**(6533): p. 685-688.

190. Levitskaya, J., et al., *Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala repeat domain of the Epstein-Barr virus nuclear antigen 1.* Proceedings of the National Academy of Sciences, 1997. **94**(23): p. 12616-12621.

191. Mellor, A.L. and D.H. Munn, *Ido expression by dendritic cells: tolerance and tryptophan catabolism.* Nature Reviews Immunology, 2004. **4**(10): p. 762-774.

192. Hwu, P., et al., *Indoleamine 2,3-Dioxygenase Production by Human Dendritic Cells Results in the Inhibition of T Cell Proliferation.* The Journal of Immunology, 2000. **164**(7): p. 3596-3599.

193. Munn, D.H., et al., *Inhibition of T Cell Proliferation by Macrophage Tryptophan Catabolism.* The Journal of Experimental Medicine, 1999. **189**(9): p. 1363-1372.

194. Munn, D.H., et al., *Prevention of Allogeneic Fetal Rejection by Tryptophan Catabolism.* Science, 1998. **281**(5380): p. 1191-1193.

195. Uyttenhove, C., et al., *Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase.* Nat Med, 2003. **9**(10): p. 1269-1274.

196. King, N.J.C. and S.R. Thomas, *Molecules in focus: Indoleamine 2,3-dioxygenase.* The International Journal of Biochemistry &amp; Cell Biology, 2007. **39**(12): p. 2167-2172.

197. Puccetti, P., *On watching the watchers: IDO and type I/II IFN.* European Journal of Immunology, 2007. **37**(4): p. 876-879.

198. Yen, M.-C., et al., *A Novel Cancer Therapy by Skin Delivery of Indoleamine 2,3-Dioxygenase siRNA.* Clinical Cancer Research, 2009. **15**(2): p. 641-649.

199. Muller, A.J., et al., *Inhibition of indoleamine 2,3-dioxygenase, an immunoregulatory target of the cancer suppression gene Bin1, potentiates cancer chemotherapy.* Nat Med, 2005. **11**(3): p. 312-319.

200. Hou, D.-Y., et al., *Inhibition of Indoleamine 2,3-Dioxygenase in Dendritic Cells by Stereoisomers of 1-Methyl-Tryptophan Correlates with Antitumor Responses.* Cancer Research, 2007. **67**(2): p. 792-801.

201. Yang, H.-J., et al., *A combination of the metabolic enzyme inhibitor APO866 and the immune adjuvant L-1-methyl tryptophan induces additive antitumor activity.* Experimental Biology and Medicine, 2010. **235**(7): p. 869-876.

202. Iwakiri, D., et al., *Epstein-Barr virus (EBV)-encoded small RNA is released from EBV-infected cells and activates signaling from toll-like receptor 3.* The Journal of Experimental Medicine, 2009. **206**(10): p. 2091-2099.

203. Fukayama, M., R. Hino, and H. Uozaki, *Epstein–Barr virus and gastric carcinoma: virus–host interactions leading to carcinoma.* Cancer Science, 2008. **99**(9): p. 1726-1733.

204. Chang, M.-S., et al., *CpG Island Methylation Status in Gastric Carcinoma with and without Infection of Epstein-Barr Virus.* Clinical Cancer Research, 2006. **12**(10): p. 2995-3002.

205. Uozaki, H. and M. Fukayama, *Epstein-Barr Virus and Gastric Carcinoma – Viral Carcinogenesis through Epigenetic Mechanisms.* Int J Clin Exp Pathol, 2008. **1**(3): p. 198-216.

206. Tsai, C.-L., et al., *Activation of DNA Methyltransferase 1 by EBV LMP1 Involves c-Jun NH2-Terminal Kinase Signaling.* Cancer Research, 2006. **66**(24): p. 11668-11676.

207. Seo, S.Y., E.-O. Kim, and K.L. Jang, *Epstein-Barr virus latent membrane protein 1 suppresses the growth-inhibitory effect of retinoic acid by inhibiting retinoic acid receptor-B2 expression via DNA methylation.* Cancer letters, 2008. **270**(1): p. 66-76.

208. Hino, R., et al., *Activation of DNA Methyltransferase 1 by EBV Latent Membrane Protein 2A Leads to Promoter Hypermethylation of PTEN Gene in Gastric Carcinoma.* Cancer Research, 2009. **69**(7): p. 2766-2774.

209. Chong, J.-M., et al., *Global and non-random CpG-island methylation in gastric carcinoma associated with Epstein-Barr virus.* Cancer Science, 2003. **94**(1): p. 76-80.

210. Sel, S., et al., *Human renal cell carcinogenesis is accompanied by a coordinate loss of the tissue specific transcription factors HNF4a and HNF1a.* Cancer letters, 1996. **101**(2): p. 205-210.

211. Wu, M., et al., *Epstein-Barr virus-associated gastric carcinomas: Relation to H. pylori infection and genetic alterations.* Gastroenterology, 2000. **118**(6): p. 1031-1038.

212. Leung, S., et al., *p53 overexpression is different in Epstein-Barr virus-associated and Epstein-Barr virus-negative carcinoma.* Histopathology, 1998. **33**(4): p. 311-7.

213. Martin, J., et al., *The Role of Sonic Hedgehog Reemergence During Gastric Cancer.* Digestive Diseases and Sciences, 2010. **55**(6): p. 1516-1524.

214. Martin, S.T., et al., *Aberrant methylation of the human hedgehog interacting protein (HHIP) gene in pancreatic neoplasms.* Cancer Biology & Therapy, 2005. **4**(7): p. 728-733.

215. Götze, S., et al., *Frequent promoter hypermethylation of Wnt pathway inhibitor genes in malignant astrocytic gliomas.* International Journal of Cancer, 2010. **126**(11): p. 2584-2593.

216. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

217. Ostrom, Q.T., et al., *CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012.* Neuro-Oncology, 2015. **17**(suppl 4): p. iv1-iv62.

218. Society, A.C., *Cancer Facts & Figures 2015.* Atlanta: American Cancer Society, 2015.

219. Wen, P.Y. and S. Kesari, *Malignant Gliomas in Adults.* New England Journal of Medicine, 2008. **359**(5): p. 492-507.

220. Cobbs, C.S., et al., *Human Cytomegalovirus Infection and Expression in Human Malignant Glioma.* Cancer Research, 2002. **62**(12): p. 3347-3350.

221. Lau, S.K., et al., *Lack of association of cytomegalovirus with human brain tumors.* Mod Pathol, 2005. **18**(6): p. 838-843.

222. Mitchell, D.A., et al., *Sensitive detection of human cytomegalovirus in tumors and peripheral blood of patients diagnosed with glioblastoma.* Neuro-Oncology, 2008. **10**(1): p. 10-18.

223. Poltermann, S., et al., *Lack of association of herpesviruses with brain tumors.* Journal of Neurovirology, 2006. **12**(2): p. 90-99.

224. Saddawi-Konefka, R. and J. Crawford, *Chronic Viral Infection and Primary Central Nervous System Malignancy.* Journal of Neuroimmune Pharmacology, 2010. **5**(3): p. 387-403.

225. Scheurer, M., et al., *Detection of human cytomegalovirus in different histological types of gliomas.* Acta Neuropathologica, 2008. **116**(1): p. 79-86.

226. Sabatier, J., et al., *Detection of human cytomegalovirus genome and gene products in central nervous system tumours.* British Journal of Cancer, 2005. **92**(4): p. 747-750.

227. Slinger, E., et al., *HCMV-Encoded Chemokine Receptor US28 Mediates Proliferative Signaling Through the IL-6–STAT3 Axis*. Vol. 3. 2010. ra58-ra58.

228. Lucas, K., et al., *The detection of CMV pp65 and IE1 in glioblastoma multiforme.* Journal of Neuro-Oncology, 2011. **103**(2): p. 231-238.

229. Ranganathan, P., et al., *Significant Association of Multiple Human Cytomegalovirus Genomic Loci with Glioblastoma Multiforme Samples.* Journal of Virology, 2012. **86**(2): p. 854-864.

230. Rahbar, A., et al., *Low levels of Human Cytomegalovirus Infection in Glioblastoma multiforme associates with patient survival; -a case-control study.* Herpesviridae, 2012. **3**(1): p. 3.

231. Bhattacharjee, B., N. Renzette, and T.F. Kowalik, *Genetic Analysis of Cytomegalovirus in Malignant Gliomas.* Journal of Virology, 2012. **86**(12): p. 6815-6824.

232. Fonseca, R.F., et al., *The prevalence of human cytomegalovirus DNA in gliomas of Brazilian patients.* Memórias do Instituto Oswaldo Cruz, 2012. **107**: p. 953-954.

233. Rahbar, A., et al., *Human cytomegalovirus infection levels in glioblastoma multiforme are of prognostic value for survival.* Journal of Clinical Virology, 2013. **57**(1): p. 36-42.

234. Ding, D., et al., *Does the existence of HCMV components predict poor prognosis in glioma?* Journal of Neuro-Oncology, 2014. **116**(3): p. 515-522.

235. dos Santos, C.J., et al., *High prevalence of HCMV and viral load in tumor tissues and peripheral blood of glioblastoma multiforme patients.* Journal of Medical Virology, 2014. **86**(11): p. 1953-1961.

236. Tang, K.-W., K. Hellstrand, and E. Larsson, *Absence of cytomegalovirus in high-coverage DNA sequencing of human glioblastoma multiforme.* International Journal of Cancer, 2015. **136**(4): p. 977-981.

237. Cimino, P.J., et al., *Detection of viral pathogens in high grade gliomas from unmapped next-generation sequencing data.* Experimental and Molecular Pathology, 2014. **96**(3): p. 310-315.

238. Cosset, É., et al., *Comprehensive metagenomic analysis of glioblastoma reveals absence of known virus despite antiviral-like type I interferon gene response.* International Journal of Cancer, 2014. **135**(6): p. 1381-1389.

239. Yamashita, Y., et al., *Lack of presence of the human cytomegalovirus in human glioblastoma.* Mod Pathol, 2014. **27**(7): p. 922-929.

240. Bianchi, E., et al., *Human cytomegalovirus and primary intracranial tumors: frequency of tumor infection and lack of correlation with systemic immune anti-viral responses.* Neuropathology and Applied Neurobiology, 2014: p. n/a-n/a.

241. Baumgarten, P., et al., *Human cytomegalovirus infection in tumor cells of the nervous system is not detectable with standardized pathologico-virological diagnostics.* Neuro-Oncology, 2014. **16**(11): p. 1469-1477.

242. Amirian, E.S., et al., *Presence of Viral DNA in Whole-Genome Sequencing of Brain Tumor Tissues from The Cancer Genome Atlas.* Journal of Virology, 2014. **88**(1): p. 774.

243. Dziurzynski, K., et al., *Consensus on the role of human cytomegalovirus in glioblastoma.* Neuro-Oncology, 2012. **14**(3): p. 246-255.

244. Cao, S., et al., *High-Throughput RNA Sequencing-Based Virome Analysis of 50 Lymphoma Cell Lines from the Cancer Cell Line Encyclopedia Project.* Journal of Virology, 2015. **89**(1): p. 713-729.

245. Strong, M.J., et al., *Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples.* PLoS Pathog, 2014. **10**(11): p. e1004437.

246. Network, C.G.A.R., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-1068.

247. Brennan, C.W., et al., *The Somatic Genomic Landscape of Glioblastoma.* Cell, 2013. **155**(2): p. 462-477.

248. Verhaak, R.G.W., et al., *Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.

249. The Cancer Genome Atlas Research, N., *Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas.* New England Journal of Medicine, 2015. **372**(26): p. 2481-2498.

250. Chen, L.Y., et al., *RNASEQR—a streamlined and accurate RNA-seq sequence analysis program.* Nucleic Acids Research, 2012. **40**(6): p. e42.

251. Lee, J., et al., *Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines.* Cancer Cell, 2006. **9**(5): p. 391-403.

252. Furnari, F.B., et al., *Malignant astrocytic glioma: genetics, biology, and paths to treatment.* Genes & Development, 2007. **21**(21): p. 2683-2710.

253. Gill, B.J., et al., *MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma.* Proceedings of the National Academy of Sciences, 2014. **111**(34): p. 12550-12555.

254. Brastianos, P.K., et al., *Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations.* Nat Genet, 2013. **45**(3): p. 285-289.

255. Stern-Ginossar, N., et al., *Decoding Human Cytomegalovirus.* Science, 2012. **338**(6110): p. 1088-1093.

256. Strong, M.J., et al., *Comprehensive High-Throughput RNA Sequencing Analysis Reveals Contamination of Multiple Nasopharyngeal Carcinoma Cell Lines with HeLa Cell Genomes.* Journal of Virology, 2014. **88**(18): p. 10696-10704.

257. Bergallo, M., et al., *Evaluation of Two Set of Primers for Detection of Immediate Early Gene UL123 of Human Cytomegalovirus (HCMV).* Molecular Biotechnology, 2008. **38**(1): p. 65-70.

258. Strong, M.J., et al., *Latency expression of the Epstein-Barr virus-encoded MHC class I TAP inhibitor, BNLF2a in EBV-positive gastric carcinomas.* Journal of Virology, 2015.

259. The Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-209.

260. Fornara, O., et al., *Cytomegalovirus infection induces a stem cell phenotype in human primary glioblastoma cells: prognostic significance and biological impact.* Cell Death Differ, 2016. **23**(2): p. 261-269.

261. Soroceanu, L., et al., *Cytomegalovirus Immediate-Early Proteins Promote Stemness Properties in Glioblastoma.* Cancer Research, 2015. **75**(15): p. 3065-3076.

262. Dziurzynski, K., et al., *Glioma-associated Cytomegalovirus Mediates Subversion of the Monocyte Lineage to a Tumor Propagating Phenotype.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2011. **17**(14): p. 4642-4649.

263. Gao, J., et al., *Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal.* Science Signaling, 2013. **6**(269): p. pl1-pl1.

264. Cerami, E., et al., *The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.* Cancer Discovery, 2012. **2**(5): p. 401-404.

265. Münger, K., et al., *Mechanisms of Human Papillomavirus-Induced Oncogenesis.* Journal of Virology, 2004. **78**(21): p. 11451-11460.

266. Wu, E.-q., et al., *Profile of physical status and gene variation of human papillomavirus 58 genome in cervical cancer.* Journal of General Virology, 2009. **90**(5): p. 1229-1237.

267. Feitelson, M.A., et al., *Genetic mechanisms of hepatocarcinogenesis.* Oncogene, 2002. **21**(16): p. 2593-604.

268. Zhang, X.-D., Y. Wang, and L.-H. Ye, *Hepatitis B virus X protein accelerates the development of hepatoma.* Cancer Biology & Medicine, 2014. **11**(3): p. 182-190.

269. Strong, M.J., Z. Lin, and E.K. Flemington, *Expanding the Conversation on High-Throughput Virome Sequencing Standards To Include Consideration of Microbial Contamination Sources.* mBio, 2014. **5**(6).

270. Stragliotto, G., et al., *Effects of valganciclovir as an add-on therapy in patients with cytomegalovirus-positive glioblastoma: A randomized, double-blind, hypothesis-generating study.* International Journal of Cancer, 2013. **133**(5): p. 1204-1213.

271. Soderberg-Naucler, C., A. Rahbar, and G. Stragliotto, *Survival in Patients with Glioblastoma Receiving Valganciclovir.* New England Journal of Medicine, 2013. **369**(10): p. 985-986.

272. Wick, W., A. Wick, and M. Platten, *Good maths is needed to understand CMV data in glioblastoma.* International Journal of Cancer, 2014. **134**: p. 2991-2992.

273. Wick, W. and M. Platten, *CMV infection and glioma, a highly controversial concept struggling in the clinical arena.* Neuro-Oncology, 2014. **16**(3): p. 332-333.

274. Ok, C.Y., et al., *Prevalence and Clinical Implications of Epstein–Barr Virus Infection in De Novo Diffuse Large B-Cell Lymphoma in Western Countries.* Clinical Cancer Research, 2014. **20**(9): p. 2338-2349.

275. Ok, C.Y., et al., *EBV-positive diffuse large B-cell lymphoma of the elderly*. Vol. 122. 2013. 328-340.

276. Hoeller, S., et al., *Epstein-Barr virus–positive diffuse large B-cell lymphoma in elderly patients is rare in Western populations.* Human Pathology, 2009. **41**(3): p. 352-357.

277. Strong, M.J., et al., *A comprehensive next generation sequencing-based virome assessment in brain tissue suggests no major virus - tumor association.* Acta Neuropathologica Communications, 2016. **4**(1): p. 1-10.

BIOGRAPHY

Michael James Strong was born on March 10, 1985 in Flint, Michigan, where he grew up with his parents and brother. He received his Bachelor of Arts degree from Kalamazoo College in Kalamazoo, Michigan and Master of Science and Master of Public Health degrees from Tufts University in Boston, Massachusetts. During his training in Boston he met his future wife Amy. He and Amy both entered the combined MD/PhD program at Tulane University in New Orleans, Louisiana. His dissertation work was under the mentorship of Dr. Erik Flemington and utilized next generation sequencing and bioinformatics to investigate oncogenic pathogens. He plans to pursue a career as a physician scientist.