Quantitative and qualitative RNA-seq based evaluation of Epstein-Barr virus transcription in type I latency Burkitt’s lymphoma cells.

Short Title: RNA-seq analysis of EBV transcriptome

#Zhen Lin1, #Guorong Xu2, Nan Deng2, Christopher Taylor2,3, *Dongxiao Zhu2,3, *Erik K Flemington1

Co-First Authors:

*Co-PI/Corresponding Authors

Erik Flemington    Dongxiao Zhu
504-988-1167     504-280-2406
eflemin@tulane.edu  dzhu@cs.uno.edu

1Department of Pathology
Tulane University Health Sciences Center and Tulane Cancer Center
1430 Tulane Avenue
New Orleans, LA 70112

2Department of Computer Science
University of New Orleans, Lakefront
2000 Lakeshore Drive
New Orleans, LA 70148

3Research Institute for Children
Children’s Hospital, New Orleans
200 Henry Clay Ave
New Orleans, LA 70118

Word counts

Abstract: 97
Text: 1873 (title page through acknowledgements)
Abstract: RNA-seq provides a rich source of transcriptome information with high qualitative and quantitative value. Here we provide a pipeline for EBV transcriptome analysis using RNA-seq and we apply it to two type I latency cell lines, Mutu I and Akata. This analysis revealed substantial average expression of many lytic genes in predominantly latent cell populations. The lytic transcripts, BHLF1 and LF3, were expressed at levels greater than 98% of all cellular polyadenylated transcripts. Exon junction mapping accurately identified the Qp derived EBNA1 splicing pattern, lytic gene splicing, and a complex splicing pattern within the BamHI A region.
Microarrays have been used to assess the levels of EBV gene expression in experimental and clinical settings (1, 7, 16-18). Nevertheless, this analysis typically requires the use of custom arrays with user specified probes against each EBV gene of interest. Despite their utility, microarrays have a limited dynamic range, being limited at the low end by the level of background and limited at the high end by signal saturation. Further, the accuracy of microarray data can be a concern because of chip defects, cross hybridization, and the analog nature of the approach. Accordingly, back-to-back comparisons of microarray and RNA-seq data have demonstrated the enhanced performance of RNA-seq in the quantitative assessment of cellular transcripts (8, 9, 14). Using RNA-seq, transcript structure information can also be deduced from a relatively unbiased data set whereas transcript structure information derived from tiling microarrays is dependent on the probe design and is therefore subject to investigator biases.

Due to the perceived potential of RNA-seq in transcriptome analysis, there has been intense interest in the development of informatics approaches to analyze cellular transcriptomes (2). For the most part, these approaches should be directly applicable to the analysis of viral transcriptomes. Nevertheless, the appropriately formatted annotation files for viruses or other ectopic organisms and the incorporation of this annotation information into existing pipelines have been lacking. We have created the necessary annotation files for EBV and merged them with annotation files for the human cellular genome so that EBV specific transcript data can be generated in the context of cellular data. This pipeline allows for the simultaneous analysis of cellular and viral transcriptomes, the digital quantification of EBV transcripts, and the visualization of EBV specific reads and splice junctions on a genome browser (see...
supplemental file S1 for pipeline details). Using this approach, we have analyzed EBV transcriptomes for the EBV-positive Burkitt’s lymphoma type I latency cell lines, Mutu I and Akata.

**Analysis of EBV gene expression in Mutu I and Akata cells.** Sequencing data used for the Mutu I analysis were control samples from a previous study where we assessed miR-155 mediated cellular transcriptome changes (14) (available from the National Center for Biotechnology Information Sequence Read Archive (SRA011001)). For this study, two separate control RNA preps were generated and single-end 50 base technical replicates were run for each poly(A)+ selected RNA. Akata sequencing data was generated anew from whole cell RNA prepared using a miRNeasy kit (Qiagen) according to the vendor’s protocol. Akata sequencing libraries were generated using the Illumina RNA-seq kit (Part #1004898) and run on a GA2x machine for single-end 74 base extensions (deposition to National Center for Biotechnology Information Sequence Read Archive is in progress). Sequences were simultaneously aligned to all human chromosomes plus the EBV genome (AG876 strain (3), GenBank DQ279927) (see supplemental file S1 for general and detailed pipeline information).

Reads per kilobase of exon per million mapped reads (RPKM – a measure of relative gene expression) values for all genes were calculated using SAMMate (http://sammate.sourceforge.net).

Fig. 1A shows the sequence read distribution across the entire EBV genome for Mutu I and Akata RNAs (an expanded/high resolution view can be seen in supplemental file S2). Ample read evidence is observed across the majority of the EBV genome. Despite this, however, the
relatively few intergenic regions that exist within the EBV genome tend to lack reads (for example, see Fig. 1B), supporting the contention that possible contaminating DNA does not represent a major source of read evidence. The abundance of reads across all latency genes were relatively low and consistent with these cell lines exhibiting type I latency, no reads mapped to the EBNA2 open reading frame (Fig. 1C and Fig. 2).

In contrast to the low levels of latency gene expression observed, we were surprised by the robust levels of many of the lytic genes in both Mutu I and Akata cells (Figs. 1 and 2). Many of these lytic genes show expression that is well above the median for all expressed cellular genes (median RPKM = 14.1 (Mutu I) and = 10.9 (Akata) - calculated as the median RPKM of genes with greater than 1 RPKM (1 RPKM typically represents approximately 1 transcript per cell (6)) (Fig. 2A and B). Strikingly, the BHLF1 and LF3 transcripts are represented at such high levels that only between 0.66 to 2% of all annotated poly(A)+ cellular genes are expressed at higher levels in Mutu I and Akata cells (Fig. 2C). The expression values observed here for EBV genes are not due to background since we ran RNA-seq data from the EBV negative cell lines, A549 (10) and MCF7 (13) through our pipeline and obtained no alignments to the EBV genome (Fig. 3). The substantial average expression levels observed here for some lytic genes could arise from either their expression in latency and/or from very high expression in a small proportion of cells that are actively undergoing lytic replication. The latter scenario most likely explains sequences obtained for most of these lytic genes. Nevertheless, it is intriguing to speculate that the former scenario may account for at least some of these genes. For example, BHLF1 and LF3 transcripts have been shown to be derived from multiple promoters, some of which are induced upon reactivation and others of
which are constitutive (5, 15). The high transcript levels that we observed under non-induced conditions suggest that these genes may play a role in the latent phase of the EBV life cycle. Overall, these data illustrate the sensitivity of RNA-seq for assessing transcript levels. Further, the BHLF1 and LF3 examples described here illustrate how the digital nature of RNA-seq allows the user to compare the abundance of transcripts from one gene with the abundance of transcripts of other genes within the transcriptome.

Notably, despite carrying out poly(A)+ RNA selection prior to sequencing, we still detect the expression of non-polyadenylated transcripts such as the EBERs in Mutu I cells (Fig. 2). However, we note that the error for non-polyadenylated transcripts tend to be high, probably due to differences in the efficacy of poly(A)+ RNA selection between the two biological replicates in Mutu I cells. Only low levels of EBERs were detected in Akata cells indicating that the poly(A)+ RNA selection was more effective in our newest RNA-seq experiment.

**Splicing evidence in Mutu I and Akata.** While RNA-seq can provide digital quantification of gene expression, reads that span exon junctions can provide information about gene isoform usage. We used the junction mapper, Tophat (12), to identify junction mapped reads throughout the EBV genome (see supplemental file S1) for Mutu I and Akata. While no evidence of Cp or Wp derived EBNA1 transcripts was found, evidence for Qp derived EBNA1 splice junctions was observed in both Mutu I and Akata cells (Fig. 4A). Junction reads were also detected for EBV lytic genes in both Mutu I and Akata cells including junction reads for both BZLF1 (Fig. 4B) and BSLF2/BMLF1 (supplemental file S3). Further, evidence for multiple isoform expression (i.e. alternative splicing events) was detected for many genes such as
BLLF1/BLLF2 (Fig. 4C) as well as the complex BamHI A region (4, 11) (supplemental file S4). Within the BamHI A region (supplemental file S4), for example, there is evidence for alternative splicing at the A73 gene in both Akata and in Mutu I with JUNC00000180 from Mutu I cells providing evidence of exon skipping (skipping of exons 2 and 3). Within the genomic regions spanning the two BART microRNA clusters, there are very few reads, consistent with these microRNAs being produced from excised introns that are presumably unstable and non-polyadenylated (and therefore not enriched during our poly(A)+ selection procedure). In both Mutu I and Akata, there is evidence for two large introns that span the entire region of both of these clusters of microRNAs (JUNC00000094 and JUNC00000178 in Mutu I and JUNC00000053 and JUNC00000084 in Akata). Consistent with this junction evidence, there are pronounced read spikes in Akata cells immediately upstream from the first junction (centered at position 139,270), between these two junctions (centered at position 147,770), and immediately downstream from the second junction (centered at position 151,115) (supplemental file S4) supporting the idea that a stable, poly(A)+ spliced transcript is generated from this transcription unit. The two introns excised from this transcript can conceivably give rise to all BART microRNAs within these two clusters.

Conclusions. Our results show robust detection of EBV derived transcripts by RNA-seq using the pipeline outlined here (supplemental file S1). From a quantitative standpoint, several studies have shown this approach to outperform microarrays since it is more accurate (8, 9, 14) and since there is an inherently broad dynamic range. For example, a previous report (1) documented the difficulty in attaining confidence in detecting most EBV derived transcripts
using microarrays because of low signal-to-noise ratios. Here, we show excellent coverage of
the bulk of EBV genes (including lytic genes) in predominantly latently infected cell lines while
at the same time detecting no EBV specific reads in two EBV-negative cell lines. The digital
nature of RNA-seq allows the user to better compare the relative expression of distinct genes
through the calculation of RPKMs. This allowed us to determine that BHLF1 and LF3 are
among the most abundant genes expressed even in predominantly latently infected cell
populations. Lastly, RNA-seq inherently contains splice junction information that can be
readily exploited to garner viral isoform expression patterns.

Our approach can also be readily applied to other viruses by manual conversion of the
respective annotation information (generally available at the National Center for Biotechnology
Information database) to the appropriate format and its subsequent conjugation to cellular
annotation files. This should result in an improvement over microarrays in the analysis of
virus-associated transcriptomes not only for EBV but for other viruses.

**Acknowledgements** This work was supported by the NIH grants CA124311, CA130752, and
CA138268 to E.K.F.; LM010137 to D.Z.; NIH ARRA administrative supplement providing
summer research experiences for students and science educators (CA130752-S1 to E.K.F.);
and an NIH COBRE (P20 RR020152 to the program director, Prescott Deininger).
**Figure legends**

FIG 1. Visualization of RNA-seq coverage across the EBV genome. Coverage (Wiggle) files generated from SAMMate and the EBV annotation file were loaded onto the Integrated Genome Viewer (IGV - developed at the Broad Institute (www.broadinstitute.org)). The Y axis shows the number of reads mapping to each location of the genome. Panel A shows the whole genome view, panel B shows a zoomed view of the intergenic region between the BMRF2 and BSLF2 genes, and panel C shows the lack of reads corresponding to the EBNA2 locus. Data range for coverage data was set to 20 (for Mutu I) and 30 (for Akata) meaning that maximal peaks represent genomic positions where there were at least 20 or 30 reads that crossed that genomic position.

FIG 2. RPKM values for EBV genes in Mutu I (A) and Akata (B) cells. Mutu I results are the average of two technical replicates (TR) from each of two separate RNA preps. Error bars indicate standard deviation for each gene. C) The number and percentage of genes showing higher RPKM values than BHLF1 and LF3 in Mutu I and Akata cells out of a total of 22,803 annotated cellular and viral genes.

FIG 3. Illustration of specificity for RNA-seq in assessing EBV transcriptomes. The total number of reads that mapped to the EBV genome per 10 million mapped reads from the EBV positive cell lines, Akata and Mutu I, and the EBV negative cell lines, A549 and MCF7. No EBV specific reads were identified in either of the EBV negative cell lines. RNA1 and RNA2
refer to biological replicate RNA samples from Mutu I cells. TR1 and TR2 refer to technical sequencing replicates.

FIG 4. Visualization of junction evidence for EBNA1 (A), BZLF1 (B), and BLLF1/BLLF2 (C).

Junction (BED) files were generated by the junction mapper, Tophat as outlined in supplemental file S1.
REFERENCES


