# An Atlas of the Epstein-Barr Virus Transcriptome and Epigenome Reveals Host-Virus Regulatory Interactions

## Supplementary Material

Aaron Arvey[1], Italo Tempera[2], Kevin Tsai[2], Horng-Shen Chen[2], Nadezhda Tikhmyanova[2], Michael Klichinsky[2] Christina Leslie[1], Paul Lieberman[2]
[1]Memorial Sloan-Kettering Cancer Center, New York, NY
[2]Wistar Institute, Philadelphia, PA

Supplemental Materials Inventory:

| Data | Source | Ref |
|---|---|---|
| ENCODE* | http://genome.ucsc.edu/ENCODE/ | [1] |
| Cheung RNA-seq | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP001563 | [2] |
| Pickrell RNA-seq | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP001540 | [3] |
| Montgomery RNA-seq | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=ERP000101 | [4] |
| Kasowski RNA-seq | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP002126 | [5] |
| EBNA1 ChIP-seq | http://ebv.wistar.upenn.edu | [6] |
| MUTU MNase-seq | http://ebv.wistar.upenn.edu | current study |
| Vitamin D ChIP-seq | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP002673 | [7] |
| Cell Line Markers | http://www.broadinstitute.org/mpg/pubs/hapmap_cell_lines/ | [8] |
| HapMap Cell Lines | http://hapmap.ncbi.nlm.nih.gov/downloads/samples_individuals/ | [9] |
| Pique-Regi DNase | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP004446 | [10] |
| Primate LCLs | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP003637 | [11] |
| E2F4 | http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP002403 | [12] |

Table S1: Data sources. *The ENCODE data was downloaded from a central repository at the UCSC genome browser; however, the experiments were performed by a number of labs.

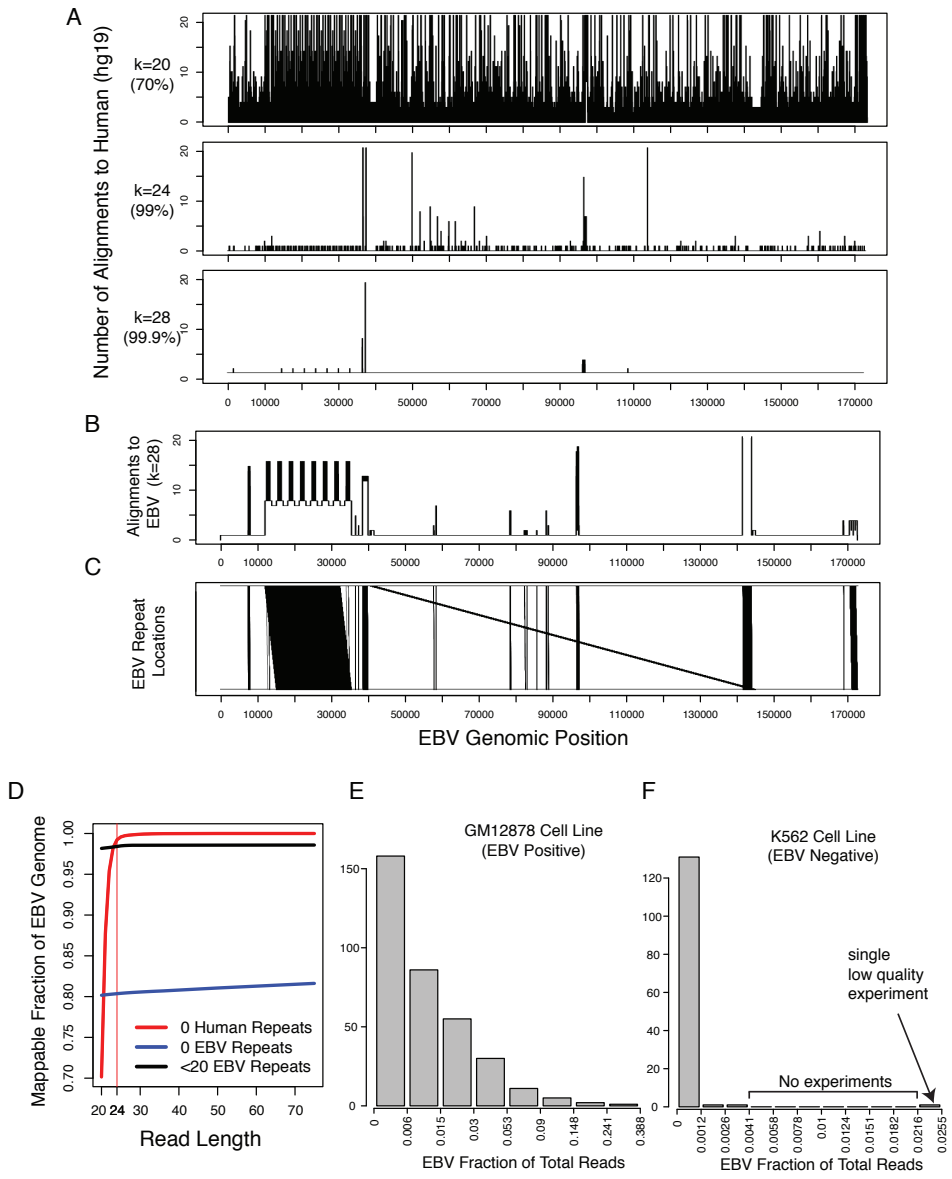
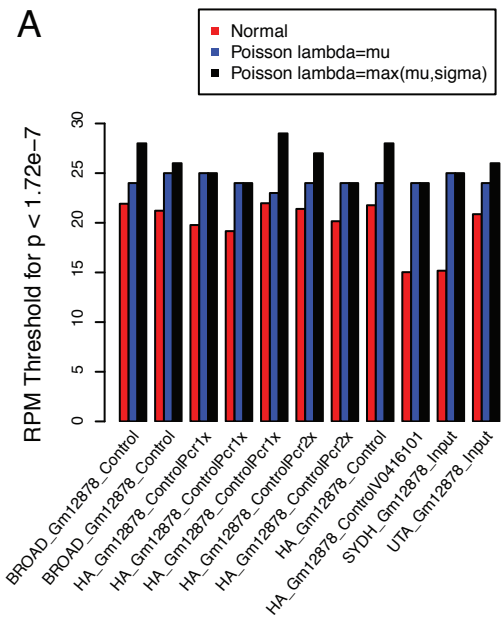
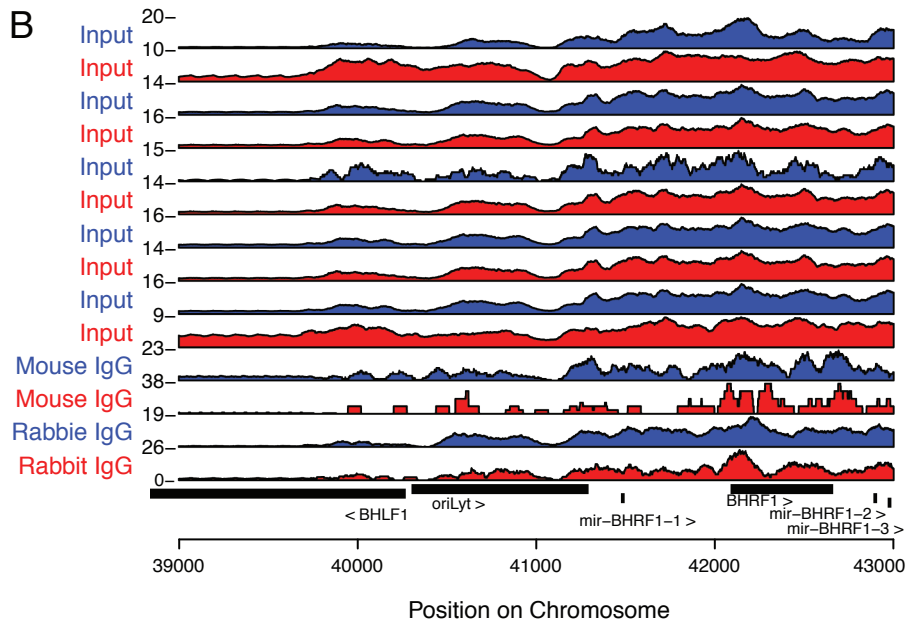
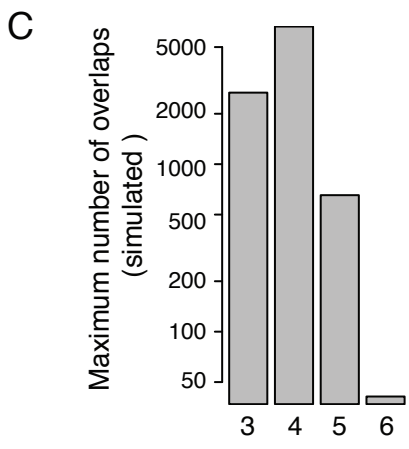
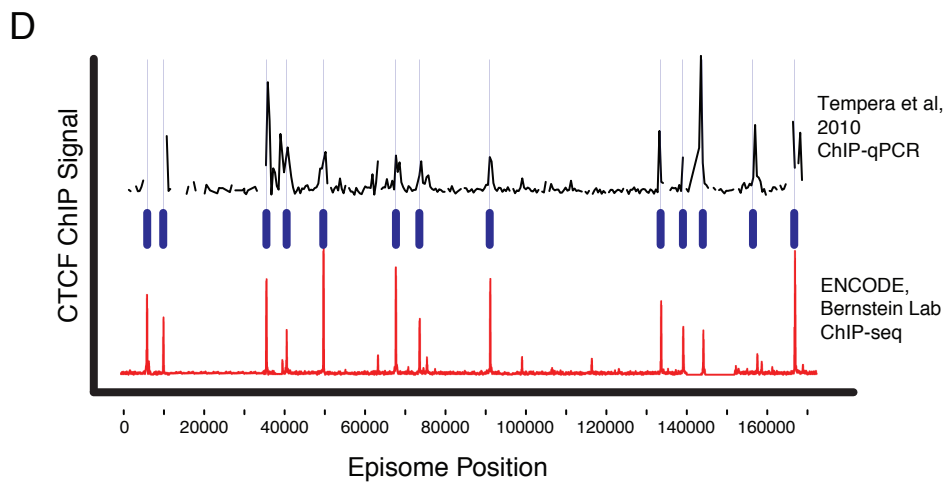


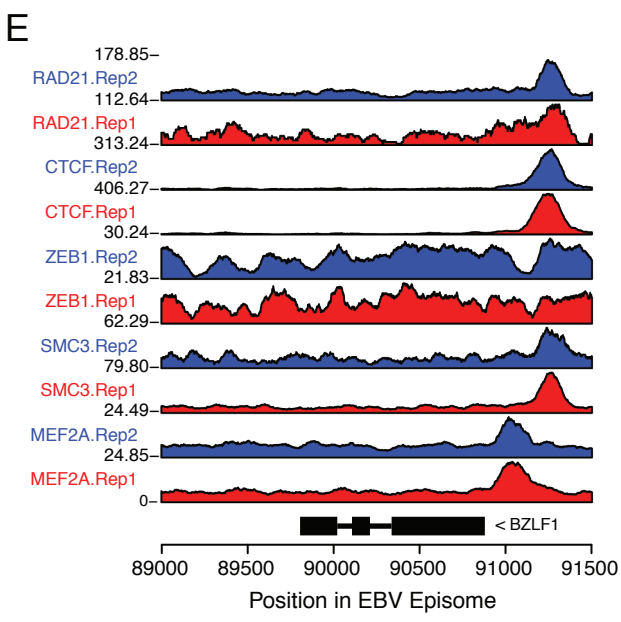

Figure S1: Mappability of the EBV genome. (A) The human alignment counts of reads sampled from the EBV genome of a given length k allowing for 1 mismatch. For experiments with reads of length k=20nt, only 70% of the genome is mappable, whereas increasing read length to 24nt increases EBV-genome coverage to 99%. For reads longer than 32nt, only the trivial poly-proline and poly-alanine-glycine repeat elements at 36kbp and 96kbp are unmappable. (B) The EBV alignment counts of reads sampled from the EBV genome. Only the region at 143kbp has more than 20 repeats, rendering it unmappable. This region is deleted in the B95.8 strain, which is the strain on which the majority of experiments were performed. (C) The location of EBV repetitive elements are shown at lines connecting identical (allowing for one mismatch) elements. (D) Mappability increases with respect to read length and 24nt is sufficient to accurately map reads to the EBV genome. (E) ENCODE experiments performed in an EBV positive cell line have more reads mapping to the EBV genome than an EBV negative cell line. The x-axis shows percentage of reads mapping to EBV in experiments on the GM12878 EBV-positive lymphoblastoid cell line. (F) Percentage of reads mapping to EBV in experiments on the K562 EBV-negative erythroleukemic cell line. The x-axis has a much smaller range than in E.

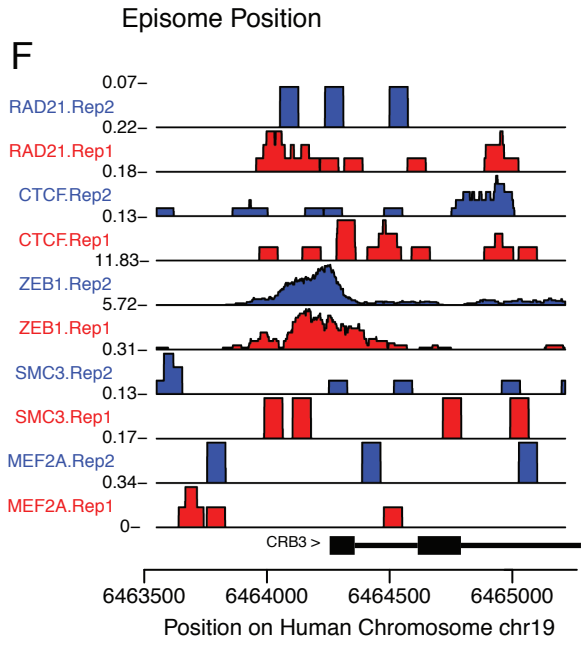

A
Normal
Poisson lambda=mu
Poisson lambda=max(mu,sigma)
RPM Threshold for p < 1.72e−7
BROAD_Gm12878_Control
BROAD_Gm12878_Control
HA_Gm12878_ControlPcr1x
HA_Gm12878_ControlPcr1x
HA_Gm12878_ControlPcr1x
HA_Gm12878_ControlPcr2x
HA_Gm12878_ControlPcr2x
HA_Gm12878_Control
HA_Gm12878_ControlV0416101
SYDH_Gm12878_Input
UTA_Gm12878_Input

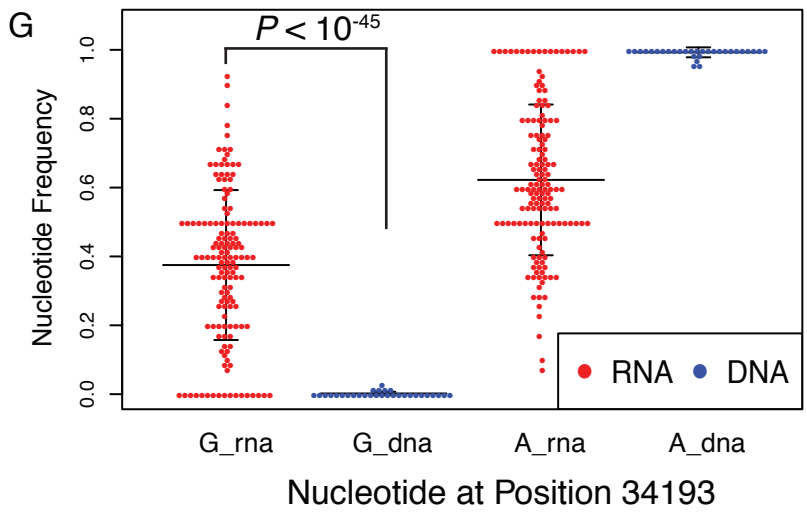

B
Input
Mouse IgG
Rabbie IgG
Rabbit IgG
< BHLF1
oriLyt >
mir−BHRF1−1 >
BHRF1 >
mir−BHRF1−2 >
mir−BHRF1−3 >
Position on Chromosome

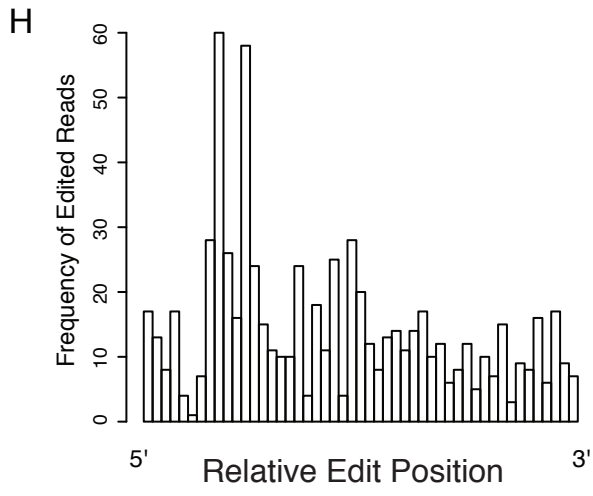

C
Maximum number of overlaps (simulated )

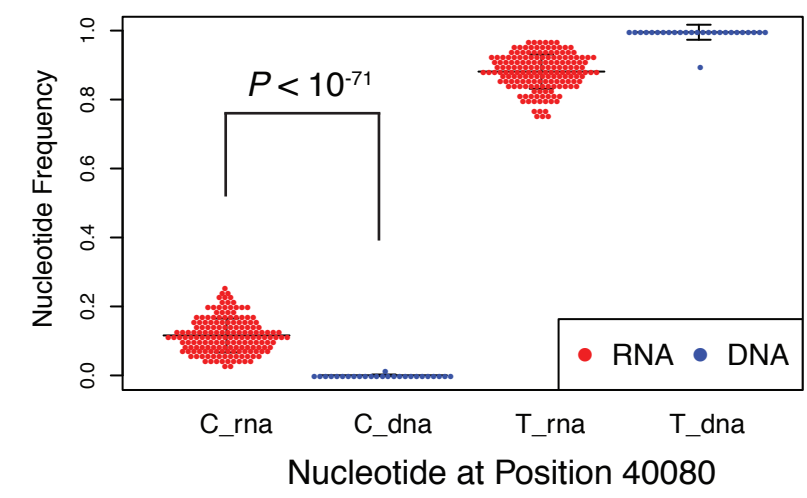

D
CTCF ChIP Signal
Tempera et al, 2010 ChIP-qPCR
ENCODE, Bernstein Lab ChIP-seq
Episome Position

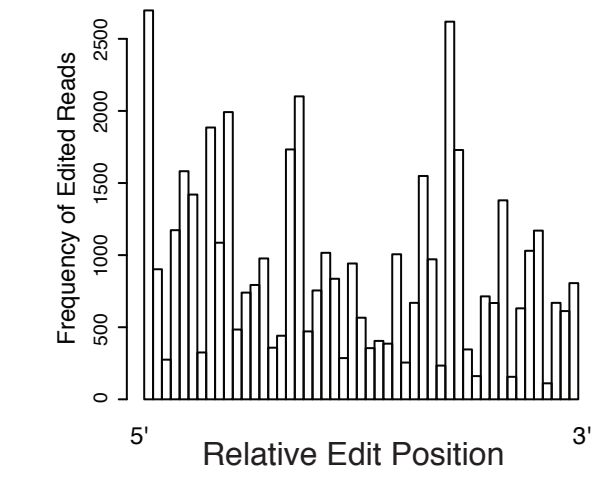

E
RAD21.Rep2
RAD21.Rep1
CTCF.Rep2
CTCF.Rep1
ZEB1.Rep2
ZEB1.Rep1
SMC3.Rep2
SMC3.Rep1
MEF2A.Rep2
MEF2A.Rep1
< BZLF1
Position in EBV Episome

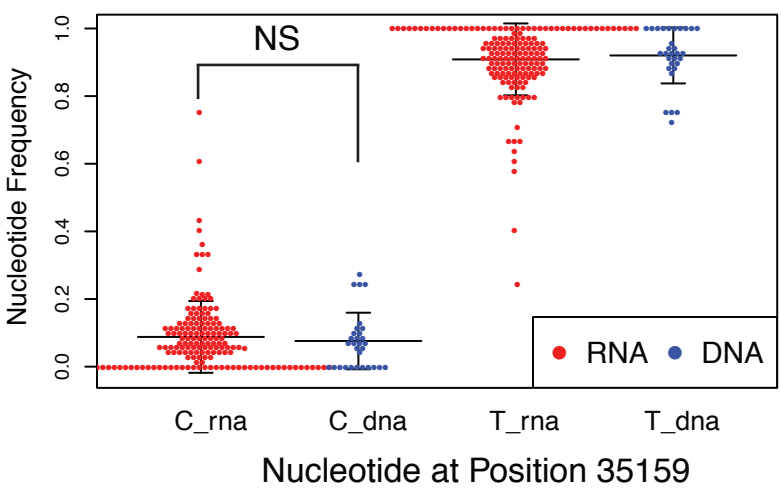

F
RAD21.Rep2
RAD21.Rep1
CTCF.Rep2
CTCF.Rep1
ZEB1.Rep2
ZEB1.Rep1
SMC3.Rep2
SMC3.Rep1
MEF2A.Rep2
MEF2A.Rep1
CRB3 >
Position on Human Chromosome chr19

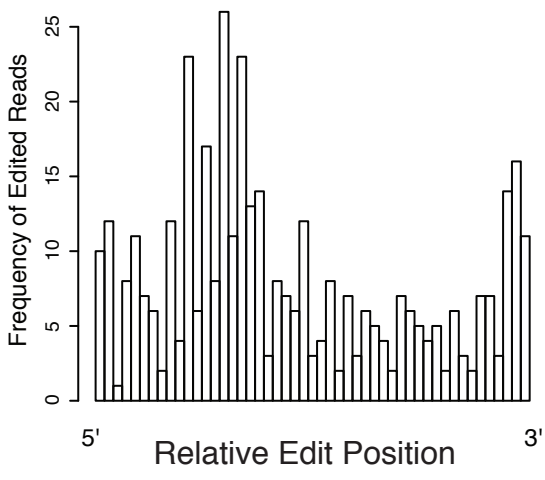

G
P < 10-45
Nucleotide Frequency
RNA
DNA
G_rna
G_dna
A_rna
A_dna
Nucleotide at Position 34193

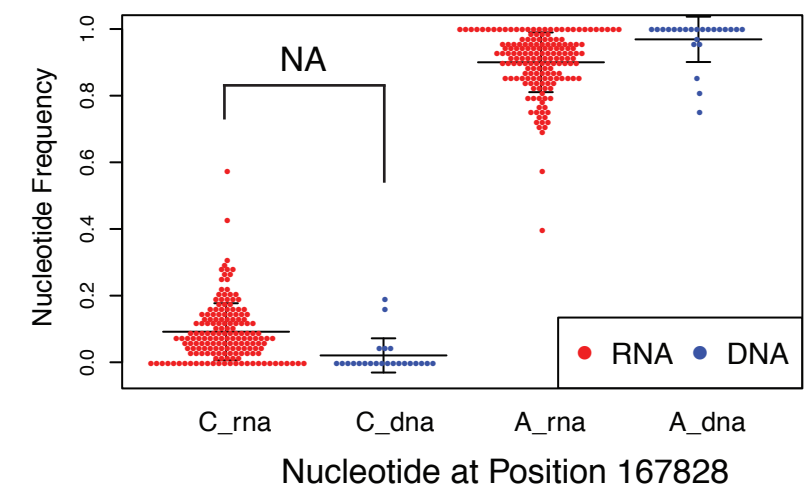

H
Frequency of Edited Reads
5'
3'
Relative Edit Position


P < 10-71
Nucleotide Frequency
RNA
DNA
C_rna
C_dna
T_rna
T_dna
Nucleotide at Position 40080


Frequency of Edited Reads
5'
3'
Relative Edit Position


NS
Nucleotide Frequency
RNA
DNA
C_rna
C_dna
T_rna
T_dna
Nucleotide at Position 35159


Frequency of Edited Reads
5'
3'
Relative Edit Position


NA
Nucleotide Frequency
RNA
DNA
C_rna
C_dna
A_rna
A_dna
Nucleotide at Position 167828

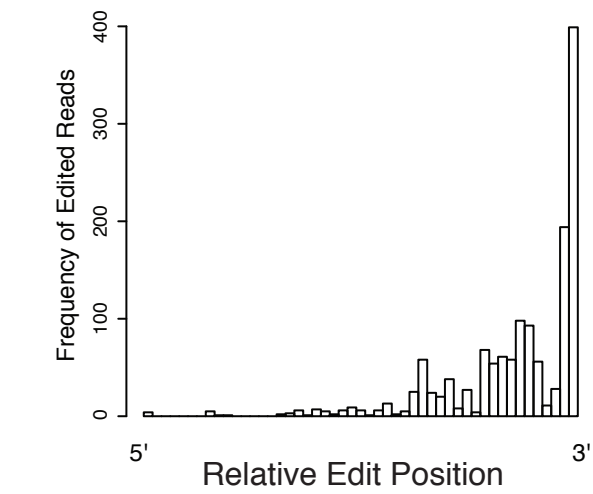
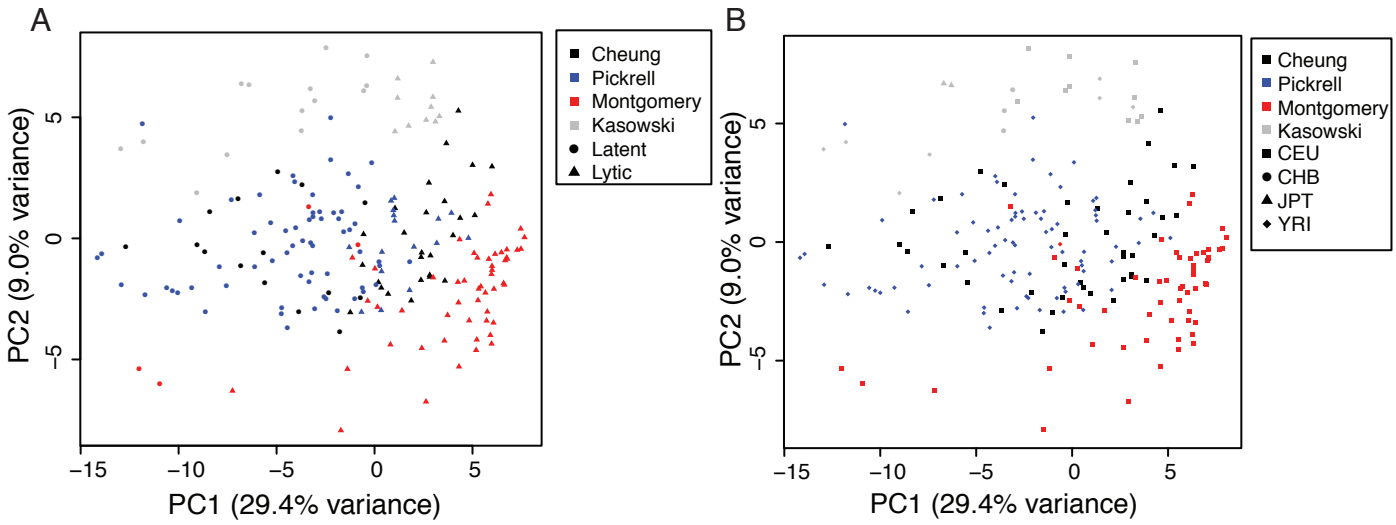

Frequency of Edited Reads
5'
3'
Relative Edit Position

Figure S2: Analysis of ChIP-seq data peak calling, clustering, and reproducibility. (A) ChIP-seq peaks were called using an empirically derived $p$-value threshold of $1.72 \times 10^{-7}$ ($p < 0.01$ genome-wide) from input control experiments. ChIP-seq RPM was sampled across the viral episome (excluding deleted and unmappable regions) at 100bp intervals to determine the mean and variance. These parameters were then used in Normal and Poisson distributions to determine an appropriate $p$-value threshold. As confirmation that we had a low false positive rate, the max RPM value found in input control experiments was always less then the derived $p$-value cutoff value. (B) Input and IgG ChIP controls show no signal at the high-occupancy region in oriLyt. (C) Transcription factor clustering in the viral genome (as shown in Figure 2 in the main text) is statistically significant. Transcription factor binding sites were permuted 10,000 times and maximum overlap was determined in each iteration. This leads to an empirical p-value of $p < 0.005$ for a region with 6 or more binding sites. The viral genome contains multiple loci with more than 8 TF binding sites. (D) CTCF binding is consistent between labs, experimental techniques, and EBV strains in type III latent cell lines. The top ChIP was performed at the Wistar Institute using by tiling qPCR primers on the MUTUIII type III latent cell line. The bottom ChIP was performed at the Broad Institute using ChIP-seq on lymphoblastoid cell line GM12878. The blue bars connected with dashed black lines show peak location correspondence. (E) The promoter of BZLF1 (Zp) is bound by MEF2A and Cohesin-associated factors, but not bound by Zeb1 as predicted by previous studies. There are sufficient reads to detect Zeb1 peaks in EBV (>400K alignments to EBV in both replicates) and mappability at Zp is comparable for all experiments (read length is 36nt). (F) Zeb1 binds a positive control region (CRB3 promoter). (G) RNA-DNA differences in the EBV transcriptome and controls. The top two rows show likely RNA-DNA differences while the bottom two rows illustrate the necessity for the additional controls. Frequency of a given nucleotide at a given position in both the transcriptome and the genome (signficance given by t-test; NS: not significant; NA: not applicable due to quality filtering). Bars show mean $\pm$ standard deviation. Note that the nucleotides shown are for the forward strand, even though the transcripts may be templated on the negative strand (e.g. position 40080, which is in BHLF1). (H) The location in reads that contain the RNA-DNA difference. The x-axis is percentage distance from $5'$ to $3'$ since multiple length reads were included in the analysis.




C

| | Lat | Lyt |
|---|---|---|
| Female | 57 | 56 |
| Male | 40 | 46 |

$P < 0.64$ (NS)

D

| | Lat | Lyt |
|---|---|---|
| Lat | 62±6 | 11±10 |
| Lyt | 5±6 | 61±10 |

$P < 10^{-21}$

E

| | Lat | Lyt |
|---|---|---|
| Lat | 87±10 | 18±15 |
| Lyt | 10±10 | 86±15 |

$P < 10^{-26}$

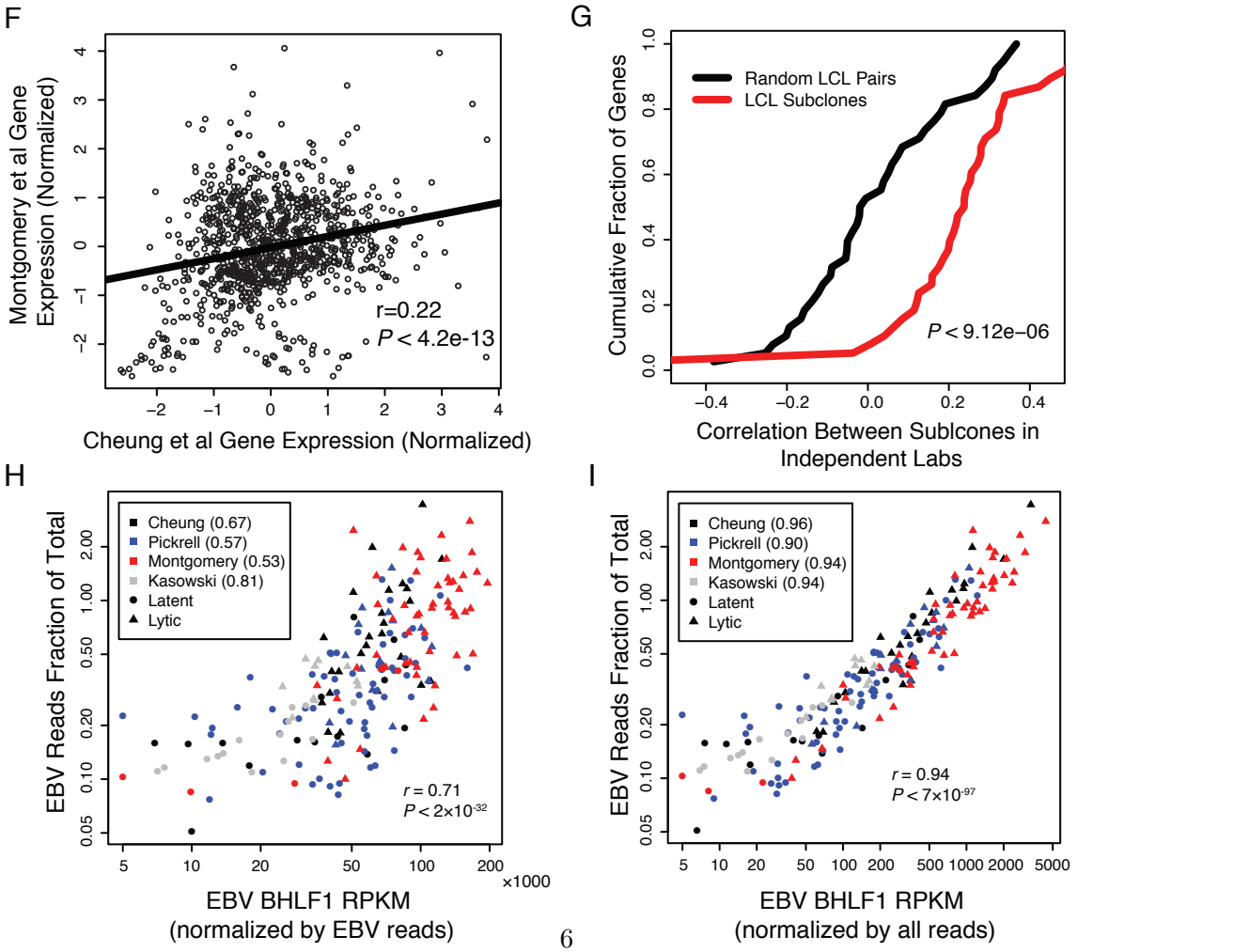


Figure S3: (A) PCA analysis of viral gene expression indicates that lytic reactivation (as determined by clustering analysis) captures $\sim 30\%$ of expression variance. The second axis of variance is best captured by lab-specific source and protocol. (B) The nationality of donor is not a substantial contributor to EBV gene expression. (C) Sex is not correlated with lytic reactivation clustering. (D) The clustering of EBV samples by RNA-seq is stable. The mean frequency of co-clustering after subsampling 70% of cell lines, is used to determine significance of stability (Fisher's exact). The subsampling and clustering was repeated 10,000 times and always had significant overlap with initial clustering ($p < 10^{-4}$ each iteration) and $\pm$ standard deviation is shown. (E) Same as D, but stability is relative to gene clustering instead of cell line clustering. (F) The same donor cell line sampled by different labs has more consistent gene expression than different donor cell lines. The axes show the relative expression (normalized by mean and standard deviation). A background distribution was generated by randomly matching LCLs 100,000 times to derive empirically that the found correlation is not obtainable by chance. (G) The correlation between expressed genes (n=38) across LCLs (n=26) are more similar in cell line subclones from the same donor (red). An empirical background calculated for different donor cell lines is significantly less correlated (shown in black). (H) BHLF1 gene expression (normlized to reads per million aligned to EBV) correlates with the fraction of reads coming from EBV (relative to uniquely mapping reads to the human genome). This correlation is consistent in all labs that performed RNA-seq (correlation in lab is shown in legend). The x-axis was incremented by 5 to improve dynamic range. (I) Same as H, except BHLF1 gene expression is normlized to reads per million aligned to EBV or human.

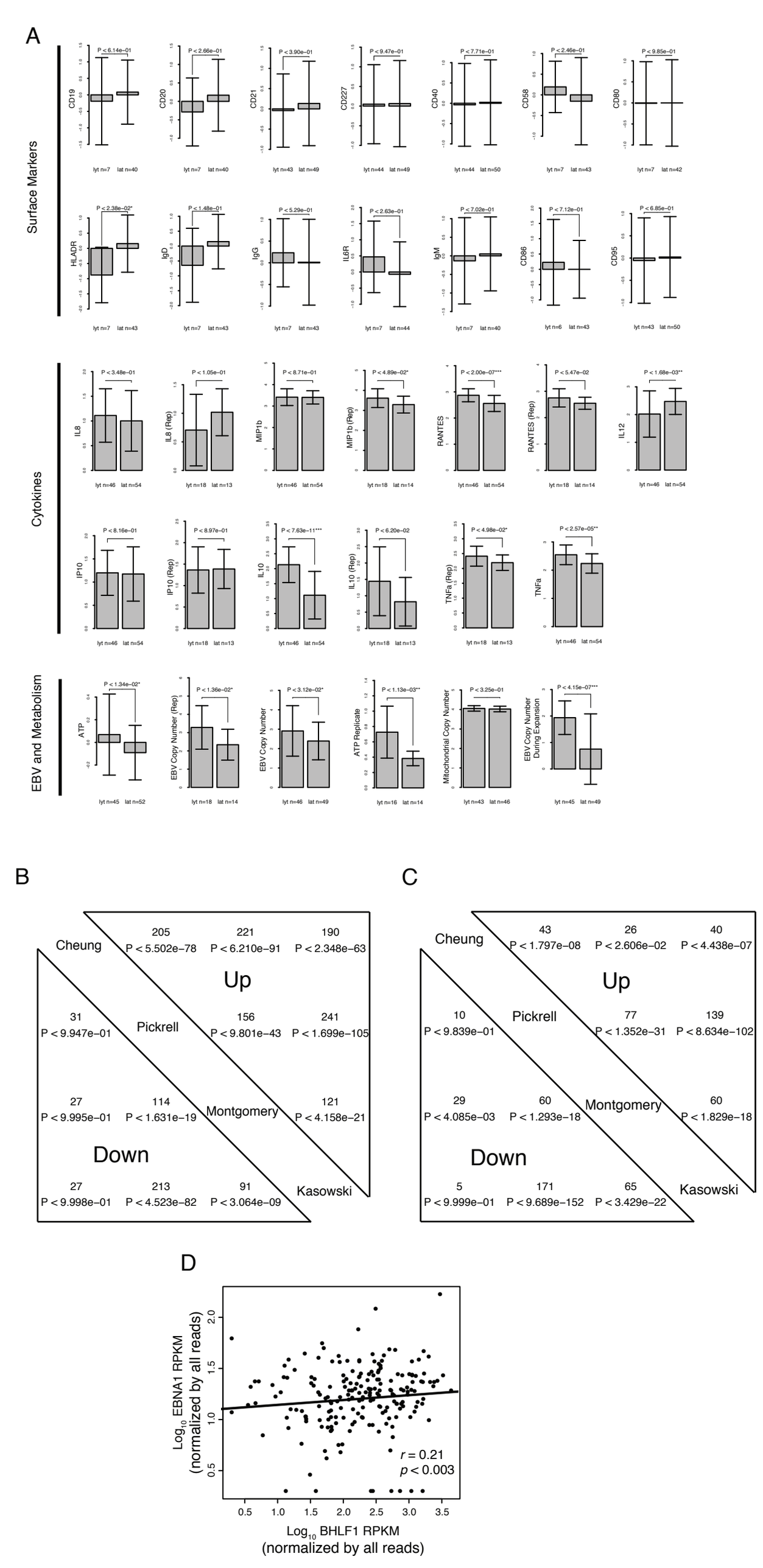

A
Surface Markers
CD19
P < 6.14e−01
lyt n=7 lat n=40
CD20
P < 2.66e−01
lyt n=7 lat n=40
CD21
P < 3.90e−01
lyt n=43 lat n=49
CD227
P < 9.47e−01
lyt n=44 lat n=49
CD40
P < 7.71e−01
lyt n=44 lat n=50
CD58
P < 2.46e−01
lyt n=7 lat n=43
CD80
P < 9.85e−01
lyt n=7 lat n=42
HLADR
P < 2.38e−02*
lyt n=7 lat n=43
IgD
P < 1.48e−01
lyt n=7 lat n=43
IgG
P < 5.29e−01
lyt n=7 lat n=43
IL6R
P < 2.63e−01
lyt n=7 lat n=44
IgM
P < 7.02e−01
lyt n=7 lat n=40
CD86
P < 7.12e−01
lyt n=6 lat n=43
CD95
P < 6.85e−01
lyt n=43 lat n=50
Cytokines
IL8
P < 3.48e−01
lyt n=46 lat n=54
IL8 (Rep)
P < 1.05e−01
lyt n=18 lat n=13
MIP1b
P < 8.71e−01
lyt n=46 lat n=54
MIP1b (Rep)
P < 4.89e−02*
lyt n=18 lat n=14
RANTES
P < 2.00e−07***
lyt n=46 lat n=54
RANTES (Rep)
P < 5.47e−02
lyt n=18 lat n=14
IL12
P < 1.68e−03**
lyt n=46 lat n=54
IP10
P < 8.16e−01
lyt n=46 lat n=54
IP10 (Rep)
P < 8.97e−01
lyt n=18 lat n=13
IL10
P < 7.63e−11***
lyt n=46 lat n=54
IL10 (Rep)
P < 6.20e−02
lyt n=18 lat n=13
TNFa (Rep)
P < 4.98e−02*
lyt n=18 lat n=13
TNFa
P < 2.57e−05**
lyt n=46 lat n=54
EBV and Metabolism
ATP
P < 1.34e−02*
lyt n=45 lat n=52
EBV Copy Number (Rep)
P < 1.36e−02*
lyt n=18 lat n=14
EBV Copy Number
P < 3.12e−02*
lyt n=46 lat n=49
ATP Replicate
P < 1.13e−03**
lyt n=16 lat n=14
Mitochondrial Copy Number
P < 3.25e−01
lyt n=43 lat n=46
EBV Copy Number During Expansion
P < 4.15e−07***
lyt n=45 lat n=49
B
Cheung
205
P < 5.502e−78
221
P < 6.210e−91
190
P < 2.348e−63
Up
Pickrell
31
P < 9.947e−01
156
P < 9.801e−43
241
P < 1.699e−105
Montgomery
27
P < 9.995e−01
114
P < 1.631e−19
121
P < 4.158e−21
Down
Kasowski
27
P < 9.998e−01
213
P < 4.523e−82
91
P < 3.064e−09
C
Cheung
43
P < 1.797e−08
26
P < 2.606e−02
40
P < 4.438e−07
Up
Pickrell
10
P < 9.839e−01
77
P < 1.352e−31
139
P < 8.634e−102
Montgomery
29
P < 4.085e−03
60
P < 1.293e−18
60
P < 1.829e−18
Down
Kasowski
5
P < 9.999e−01
171
P < 9.689e−152
65
P < 3.429e−22
D
Log10 EBNA1 RPKM
(normalized by all reads)
0.5
1.0
1.5
2.0
r = 0.21
p < 0.003
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Log10 BHLF1 RPKM
(normalized by all reads)

Figure S4: (A) Quantitative assay of cell line characteristics correlate with lytic vs. latent cell lines. One of the most significant correlations is the amount of EBV episome during LCL expansion. We also note that EBV episome count and ATP weakly correlate in independent replicates. These assays were performed by an independent lab than the RNA-seq experiments which are used to assign cell lines to more lytic and more latent subtypes [8]. P-values shown are nominal; * $p < 0.05$ (nominal); ** $p < 0.05$ and *** $p < 0.0005$ (Holm corrected significance). Error bars show standard deviation. (B) Human genes associated with viral lytic gene expression are consistent across labs, cell lines, and donor populations. The overlap in the top 1000 human genes that are upregulated with viral lytic genes (upper triangle) or downregulated (lower triangle). It is worth noting that downregulated genes are less consistent than upregulated genes. (C) The pathways of human genes associated with viral lytic reactivation are consistent across labs, cell lines, and donor populations. The overlap in the top 300 human pathways that are upregulated with viral lytic genes (upper triangle) or downregulated (lower triangle). (D) Upregulation of EBV transcription factor EBNA1 is associated with increased lytic reactivation (BHLF1 expression), suggesting that the upregulation of human EBNA1 targets is a result of increased regulator expression.

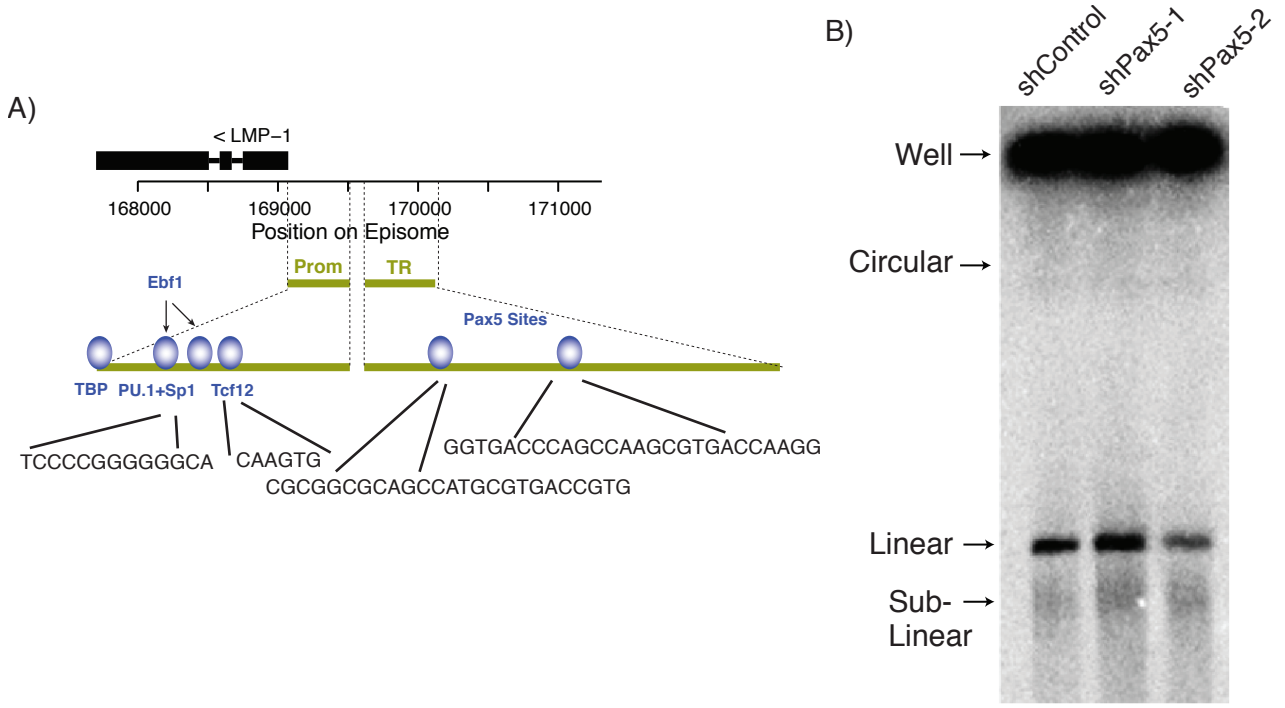



Figure S5: (A) DNA motifs for multiple Pax5 sites and other known and novel binding sites at the promoter of LMP1. Ebf1 motifs and ChIP-seq signal suggst multiple binding sites. (B) EBV episome structure after knockdown of Pax5 assayed by PFGE. The 'well' fragment likely includes circular episomes that are associated with host chromosomes that disallow migration. The fraction of viral genome that lies in the linear and sublinear fractions is small relative to the non-migrating ('well') fraction.

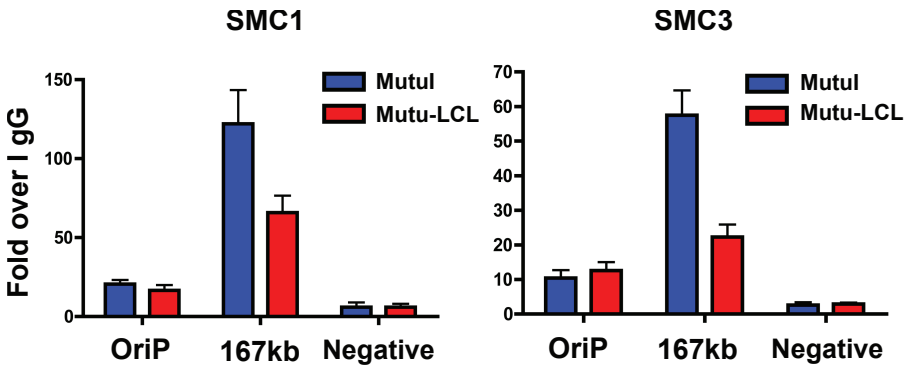



Figure S6: Validation of Cohesin subunits SMC1 and SMC3 binding to EBV genome. ChIP assays with SMC1 (left) or SMC3 (right) antibody were assayed by quantitative real-time PCR with primers specific for EBV regions at OriP, the 167kb CTCF-Cohesin peak, or a control region of EBV at position 32kb (Negative). These regions were assayed in MutuI and Mutu-LCL cells, confirming that Cohesin subunits bind consistently in multiple latency types and EBV strains.

## Aligning high-throughput sequencing experiments to EBV

We obtained raw data from a variety of sources (shown in Table S1) and aligned the data to EBV. All EBV alignments and processed data are available at `http://ebv.wistar.upenn.edu`.

### Mappability of the EBV genome

There are two primary obstacles in aligning reads to the EBV genome: (1) subtracting reads that map to the human genome can leave parts of the EBV genome unmappable and (2) EBV has several large, functionally relevant repetitive elements. We consider a region of EBV to be "mappable" if there is no homologous region in the human genome and the region occurs fewer than 20 times in the EBV genome. Depending on the read length $k$, the fraction of the EBV genome that is mappable varies. For instance, only 70% of the genome is mappable with 20nt reads. This increases to 99% with reads of length 24 and 99.9% using reads of length 32 (Figure S1). The only regions that are unmappable with 32nt reads are two simple repeat regions consisting of poly-proline and poly-alanine-glycine tracts.

The repetitive elements in EBV make exact read mapping challenging. Using reads of length 24-75nt, 18.3-19.7% of all reads fall into EBV-repetitive regions (Figure S1). Many of the repetitive loci have interesting RNA and regulatory elements. An example includes the W repeats, which are composed of tandem 3071bp regions each of which has a nested repeat element.

To better understand all the functional elements of EBV, we wish to maximize the regions of the genome that we can assay. To this end, we allow each read to have up to 20 alignments in EBV (Figure S1). We select 20 alignments as our cutoff as it drastically increases the mappability of the genome (Figure S1) and introduces minimal artifacts since many of the repetitive elements are tandem repeats. As the repeats are in tandem, the functional element of interest is at the least in one (or all) of the proximal repeats. The one functional element that is not tandem is the duplication of oriLyt; however, nearly all of our data from experiments performed in B95.8, which has a large deletion encompassing the oriLyt (R) at position 144kbp in the EBV complete genome.

### Read processing

Reads in some experiments required additional processing prior to alignment. RNA-seq reads from [2, 4] suffered from lower quality at the $3'$ end. Quality trimming was applied to experiments having fewer than 30% of reads aligned to the host genome and suffering from low quality at the $5'$ (left) and/or $3'$ (right) ends of reads. We trimmed by quality thresholding at the (equivalent) phred score $> 30$, accepting reads that phred score $> 30$ at 85% of their nucleotides and were not trivial repetitive reads. This was implemented by a slightly altered version of the fastx package, which considers $5'$ as well as $3'$ quality filtering [13]. If adaptor sequences were ligated to DNA fragments, such as in small RNA-seq, adaptor sequence was clipped off using the fastx package [13].

### Quantifying read overlap counts

After processing and aligning the raw reads, we computed read overlap counts in reads per million (RPM) reads aligned. DNase and RNA-based sequencing assays were quantified by overlapping reads, whereas ChIP-seq, MNase-seq, and similar assays were quantified by 150nt read extension with only the middle 75nt extracted for overlap counts. For multi-mapping reads, we divided read counts over the number of

possible mapping location. If a TF binds only at a single repetitive element, the reads will be distributed uniformly across all repeat elements. We also note that the copy number of repetitive elements, while relatively stable in the B95.8 EBV strain, may vary across LCLs and other EBV strains.

## Transcription factor binding

### Clusters of transcription factors

We identified several spatial clusters of transcription factor binding events. In all clusters, there are many negative controls (IgG, input DNA, and many non-binding TFs). To determine if a cluster contains significantly more TF binding events than would be expected by chance, we permuted all binding sites 10,000 times and determined the maximum overlap (Figure S2). We found a maximum of 6 or more overlapping TFs occurred in 50/10000 iterations, giving an empirical $p$-value of $p < 0.005$.

### Consistency with prior studies

CTCF ChIP-seq findings were comparable with previous findings in a different lab and using a different EBV strain [14] (Figure S2). We also confirmed known binding sites for Yy1 (Wp), Pax5 (Wp), Ebf1 (LMP1p), PolIII (EBERp), and others. In contrast, a notable difference of our data with previous findings is that Zeb1 ChIP was not enriched at Zp (antibody H-102, SC-25388, as previously used [15, 16]). As a positive control we find Zeb1 binds the host CRB3 promoter and other loci bound in previous studies [17]. It is worth noting that our data is from type III latent lymphoblastoid cell lines whereas Mertz and colleagues have conducted ChIP in HEK293 transfected cells [15] and mutant expression plasmid studies in a type III latent (MUTU-derived) cell line [16]. It is also possible that Zeb1 may interact with Zp but is not detectable by ChIP-seq due to diffuse, non-punctate binding.

## RNA-DNA differences

We determined nucleotides in the transcriptome (as assayed by RNA-seq) that differ from nucleotides in the reference genome (confirmed by DNA sequencing and IgG controls). We aligned reads allowing for one mismatch and determined nucleotides with greater than $>$ 10x coverage and have $>$15% of RNA-seq reads containing the non-reference genome nucleotide. We then examined 170 RNA-seq experiments with high coverage ($>$10x) across multiple labs to confirm that the edit was reproducible and not cell line specific. We required that each putative edit occur in at least 10 independent experiments. We also required that the edit occur uniformly across the read, since if the edit only occurs at the beginning or end of a read, it is possible that further quality filtering was necessary. Finally, we determined that there was significantly higher fraction of the RNA-DNA difference in RNA-seq experiments than in DNA-seq experiments ($>$5x DNA-seq coverage) by a two-sided Student's t-test.

After applying these filters and compressing repetitive regions of the genome, we identified 7 candidate sites that we examined manually. We flagged 2 positions as possible RNA-DNA differences (Figure S2). The other 5 putative sites are likely either templated by DNA (though perhaps in only one of many repetative regions) or artifactual (examples shown in bottom two rows of Figure S2). For instance, the position 35159 is in the W repeats, which are estimated to occur 7 times in tandem in the EBV genome, one of which may harbor the observed T-to-C differences in both RNA and template DNA.

Another control that proved critical was filtering by edit position in read, as shown in the bottom panel of Figure S2.

## Lytic cycle reactivation

We observed that lymphoblastoid cell lines tend toward one of two discrete expression programs. These two programs are representative of traditional type III latency and abortive lytic reactivation (see main text and Figure 3).

### Clustering samples into latent and lytic phase

The clustering of EBV gene expression profiles (as seen in Figure 3 in the main text) is performed by hierarchical clustering. Prior to clustering, we computed the log ratio of each gene with respect to its median expression value. We then used the Spearman correlation distance ($1-\rho$ where $\rho$ is the Spearman correlation) between all pairs of genes and all pairs of cell lines. The hierarchical clustering was performed by `hclust` in R and genes were ordered within clusters by coefficients of the first principal component.

The first principal component captures nearly 30% of the variance and represents a direction along which samples can be discriminated as latent or lytic (Figure S3). In contrast, the 2nd principal component best discriminates between multiple labs, suggesting that there may be lab-specific bias to the data, which is not surprising given the differences in RNA extractions, library preparation methods, and sequencing instruments. However, this bias is not the primary determinant of gene expression as it accounts for less than 10% of total variance, whereas differences in lytic and latent cycle gene expression account for nearly 30%. Furthermore, LCL viral gene expression sampled in two independent labs was significantly correlated ($p <$ 9.12e-6; Figure S3).

We also found that the clustering is highly statistically stable (Figure S3). We find that removing 30% of the samples and reclustering results in the majority of samples remaining in their original cluster. We performed 10000 iterations of reclustering and always found significant overlap with original clustering ($p < 0.01$ by Fisher's exact test, resulting in an empirical $p < 10^{-5}$). We performed an identical simulation, but removing 30% of genes and found that the clustering is robust to the gene expression features (Figure S3).

### Correlation with host gene expression

Host gene expression correlates with viral gene expression (see main text). We correlated lytic marker BHLF1 expression with every expressed human gene. BHLF1 was selected since it is a known lytic-associated transcript, highly correlated with percentage of reads aligning to EBV (Figure S3), and expressed at a sufficiently high level to obtain a robust estimate of gene expression in all examined samples.

Since the RNA-seq data comes from multiple labs, which contributes a significant portion of the expression variance (Figures S3, S4), we analyzed each lab's data independently. Pairwise comparisons between each lab show that the human genes upregulated in the samples that are classified as lytic (based on viral gene expression clustering) overlap significantly (Figure S4).

We note that even though lytic reactivation consistently correlates with many host genes, it does not seem to be a primary contributor to host variation. Rather, it seems that lab-specificity is the dominant source of variation in human gene expression (Figure S4).

### Pathways regulated in latent phase and lytic reactivation

To determine which human pathways are involved in lytic reactivation, we searched for sets of genes whose correlation with BHLF1 is significantly higher than expected by chance. Pathways were found using the Gene Set Enrichment Analysis (GSEA) tool from the Broad Institute [18]. BHLF1 correlations performed in each lab independently were used as the rankings input to the "leading edge" analysis.

## References


[1] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

[2] Vivian G. Cheung, Renuka R. Nayak, Isabel X. Wang, Susannah Elwyn, Sarah M. Cousins, Michael Morley, and Richard S. Spielman. Polymorphic Cis- and Trans-Regulation of Human Gene Expression. *PLoS Biology*, 8(9):e1000480+, September 2010.

[3] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, March 2010.

[4] Stephen B. Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P. Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, March 2010.

[5] Maya Kasowski, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M. Waszak, Lukas Habegger, Joel Rozowsky, Minyi Shi, Alexander E. Urban, Mi-Young Y. Hong, Konrad J. Karczewski, Wolfgang Huber, Sherman M. Weissman, Mark B. Gerstein, Jan O. Korbel, and Michael Snyder. Variation in transcription factor binding among humans. *Science*, 328(5975):232–235, April 2010.

[6] Fang Lu, Priyankara Wikramasinghe, Julie Norseen, Kevin Tsai, Pu Wang, Louise Showe, Ramana V. Davuluri, and Paul M. Lieberman. Genome-wide analysis of host-chromosome binding sites for Epstein-Barr Virus Nuclear Antigen 1 (EBNA1). *Virology Journal*, 7(1):262+, October 2010.

[7] Sreeram V. Ramagopalan, Andreas Heger, Antonio J. Berlanga, Narelle J. Maugeri, Matthew R. Lincoln, Amy Burrell, Lahiru Handunnetthi, Adam E. Handel, Giulio Disanto, Sarah-Michelle Orton, Corey T. Watson, Julia M. Morahan, Gavin Giovannoni, Chris P. Ponting, George C. Ebers, and Julian C. Knight. A ChIP-seq defined genome-wide map of vitamin D receptor binding: Associations with disease and evolution. *Genome Research*, 20(10):1352–1360, October 2010.

[8] Edwin Choy, Roman Yelensky, Sasha Bonakdar, Robert M. Plenge, Richa Saxena, Philip L. De Jager, Stanley Y. Shaw, Cara S. Wolfish, Jacqueline M. Slavik, Chris Cotsapas, Manuel Rivas, Emmanouil T. Dermitzakis, Ellen Cahir-McFarland, Elliott Kieff, David Hafler, Mark J. Daly, and David Altshuler. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genetics*, 4(11):e1000287+, November 2008.

[9] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, September 2010.

[10] Roger Pique-Regi, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, March 2011.

[11] Carolyn E. Cain, Ran Blekhman, John C. Marioni, and Yoav Gilad. Gene Expression Differences Among Primates are Associated With Changes in a Histone Epigenetic Modification. *Genetics*, 187(4):e126177+, February 2011.

[12] Bum-Kyu Lee, Akshay A. Bhinge, and Vishwanath R. Iyer. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Research*, 39(9):3558–73, January 2011.

[13] Hannon Lab. FASTX Toolkit.

[14] Italo Tempera, Andreas Wiedmer, Jayaraju Dheekollu, and Paul M. Lieberman. CTCF prevents the epigenetic drift of EBV latency promoter Qp. *PLoS Pathogens*, 6(8):e1001048+, August 2010.

[15] Xianming Yu, Zhenxun Wang, and Janet E. Mertz. ZEB1 Regulates the Latent-Lytic Switch in Infection by Epstein-Barr Virus. *PLoS Pathogens*, 3(12):e194+, December 2007.

[16] Amy L. Ellis, Zhenxun Wang, Xianming Yu, and Janet E. Mertz. Either ZEB1 or ZEB2/SIP1 Can Play a Central Role in Regulating the Epstein-Barr Virus Latent-Lytic Switch in a Cell-Type-Specific Manner. *Journal of Virology*, 84(12):6139–6152, June 2010.

[17] K. Aigner, B. Dampier, L. Descovich, M. Mikula, A. Sultan, M. Schreiber, W. Mikulits, T. Brabletz, D. Strand, P. Obrist, W. Sommergruber, N. Schweifer, A. Wernitznig, H. Beug, R. Foisner, and A. Eger. The transcription factor ZEB1 ($\delta$EF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity. *Oncogene*, 26(49):6979–6988, May 2007.

[18] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.