# A survey of RNA secondary structural propensity encoded within human herpesvirus genomes: global comparisons and local motifs

**Ryan J. Andrews** [1], **Collin A. O'Leary** [1], **Walter N. Moss** [Corresp. 1]

[1] The Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, US

Corresponding Author: Walter N. Moss
Email address: wmoss@iastate.edu

There are 9 herpesviruses known to infect humans, of which Epstein–Barr virus (EBV) is the most widely distributed (>90% of adults infected). This ubiquitous virus is implicated in a variety of cancers and autoimmune diseases. Previous analyses of the EBV genome revealed numerous regions with evidence of generating unusually stable and conserved RNA secondary structures and led to the discovery of a novel class of EBV non-coding (nc)RNAs: the stable intronic sequence (sis)RNAs. To gain a better understanding of the roles of RNA structure in EBV biology and pathogenicity, we revisit EBV using recently developed tools for genome-wide motif discovery and RNA structural characterization. This corroborated previous results and revealed novel motifs with potential functionality; one of which has been experimentally validated. Additionally, since many herpesviruses increasingly rival the seroprevalence of EBV (VZV, HHV-6 and HHV-7 being the most notable), analyses were expanded to include all sequenced human Herpesvirus RefSeq genomes, allowing for genomic comparisons. In total 10 genomes were analyzed, for EBV (types 1 and 2), HCMV, HHV-6A, HHV-6B, HHV-7, HSV-1, HSV-2, KSHV, and VZV. All resulting data were archived in the RNAStructuromeDB (
https://structurome.bb.iastate.edu/herpesvirus ) to make them available to a wide array of researchers.

1  **A survey of RNA secondary structural propensity encoded within human herpesvirus**
2  **genomes: global comparisons and local motifs.**

4  Ryan J. Andrews[1], Collin A. O'Leary[1], and Walter N. Moss[1*]
5
6  [1]Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State
7  University, 2437 Pammel Drive, Ames, IA, 50011, USA.
8
9  *Correspondence: wmoss@iastate.edu
10
11 Author information:
12
13 Ryan J. Andrews – randrews@iastate.edu
14 Collin A. O'Leary – caoleary@iastate.edu
15 Walter N. Moss – wmoss@iastate.edu

17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

**Abstract**

There are 9 herpesviruses known to infect humans, of which Epstein–Barr virus (EBV) is the most widely distributed (>90% of adults infected). This ubiquitous virus is implicated in a variety of cancers and autoimmune diseases. Previous analyses of the EBV genome revealed numerous regions with evidence of generating unusually stable and conserved RNA secondary structures and led to the discovery of a novel class of EBV non-coding (nc)RNAs: the stable intronic sequence (sis)RNAs. To gain a better understanding of the roles of RNA structure in EBV biology and pathogenicity, we revisit EBV using recently developed tools for genome-wide motif discovery and RNA structural characterization. This corroborated previous results and revealed novel motifs with potential functionality; one of which has been experimentally validated. Additionally, since many herpesviruses increasingly rival the seroprevalence of EBV (VZV, HHV-6 and HHV-7 being the most notable), analyses were expanded to include all sequenced human Herpesvirus RefSeq genomes, allowing for genomic comparisons. In total 10 genomes were analyzed, for EBV (types 1 and 2), HCMV, HHV-6A, HHV-6B, HHV-7, HSV-1, HSV-2, KSHV, and VZV. All resulting data were archived in the RNAStructuromeDB (https://structurome.bb.iastate.edu/herpesvirus) to make them available to a wide array of researchers.

**Introduction**

The herpesviruses are an ancient family of double-stranded DNA viruses (Order *Herpesvirales*, Family *Herpesviridae*). The three major subfamilies (clades) of mammalian herpesviruses—α (*Alphaherpesvirinae*), β (*Betaherpesvirinae*), and γ (*Gammaherpesvirinae*)—appeared ~180 to 220 mya(McGeoch et al. 1995). Nine herpesviruses are currently known to infect humans: Epstein–Barr virus (EBV), human cytomegalovirus (HCMV), human herpesvirus 6A and 6B (HHV-6A and HHV-6B), human herpesvirus 7 (HHV-7), herpes simplex viruses 1 and 2 (HSV-1 and HSV-2), varicella-zoster virus (VZV), and Kaposi's sarcoma-associated herpesvirus (KSHV). Of these, EBV is most prevalent with >90% of adults being infected (Kieff & Rickinson 2001). EBV has a large (~172 kb) double-stranded DNA genome that recapitulates many of the biological processes important to the host genome (Young & Rickinson 2004) including periods of pervasive transcription (O'Grady et al. 2014). After spreading through the host by lytic replication, EBV establishes latency in B cells via several distinct patterns of coding and noncoding gene expression. Both lytic and latent replication are associated with various *cancers* (e.g. lymphomas and carcinomas) (Elgui de Oliveira et al. 2016) and *autoimmune diseases* (e.g. multiple sclerosis and Lupus) (Draborg et al. 2013); however, molecular mechanisms connecting EBV infection to disease remain elusive—making EBV an important target of study. One route for the regulation of viral/host biology and for mediating host-virus interactions is RNA secondary structure. Many key biological processes of both hosts and their pathogens are affected by RNA structure: e.g. RNA transcription, processing, stability, and expression are all affected by RNA folding (Andrews & Moss 2019a; Bevilacqua et al. 2016).

Although several functional RNA structural motifs have been identified in EBV (Moss et al. 2014), knowledge of the roles of RNA and RNA folding in EBV infection and disease still require additional attention. The first viral noncoding (nc)RNAs, discovered in 1981, were the EBV encoded RNAs (EBERs -1 and -2)(Lerner et al. 1981). Both EBERs are highly-expressed in infected cells (dwarfing endogenous ncRNA expression) and possess extensive secondary structure that mediates their interactions with host and viral biomolecules to promote infection (Lee et al. 2015; Lee et al. 2012; Lee et al. 2016; Moss et al. 2014). Over 20 years later

101  additional viral ncRNAs were discovered: the BART and BHRF EBV-encoded micro (mi)RNAs
102  in 2004 (Pfeffer et al. 2004) and the EBV viral small nucleolar (v-snoRNA)1 in 2009 (Hutzinger
103  et al. 2009).
104
105  A genome-wide bioinformatics and experimental analysis was previously performed in EBV to
106  scan for functional RNA motifs (Moss & Steitz 2013). In the previous analysis, the ncRNA
107  discovery program RNAz (Gruber et al. 2010) was used to analyze sliding windows of aligned
108  EBV genome sequences (plus one closely related lymphocryptovirus from rhesus monkeys:
109  Macacine herpesvirus 4) to predict windows that have unusual thermodynamic stability and
110  conservation of RNA secondary structure—both indicators of functionality (Clote et al. 2005).
111  Approximately 30% of the (~170 kbp) genome was predicted to potentially encode functional
112  RNA structure. In addition to known structures (EBERs, v-snoRNA1, and pre-miRNA hairpins),
113  many novel structural elements were predicted. This analysis discovered a new class of EBV
114  ncRNAs, the stable intronic sequence (sis)RNAs, which consist of abundant, highly structured
115  intronic-sequence RNAs (derived from the EBV W repeats). The sisRNAs consist of a complex
116  mix of excised introns and, more recently discovered, alternatively spliced long (l)ncRNAs (up to
117  ~ 22 kb) (Cao et al. 2015b; Tompkins et al. 2018). The sisRNAs interact with multiple host
118  regulatory proteins (Tompkins et al. 2018) and are essential to EBV's ability to transform human
119  B cells (Bridges et al. 2019). Very long (~500 nt) hairpins were modeled in the latent origin of
120  replication (oriP), which were later shown to occur in bidirectional lytic lncRNAs that are hyper
121  A-to-I edited and promote lytic replication (Cao et al. 2015a). Elsewhere, a structural motif
122  partially spans two exons and a whole intron in the EBV LMP2 gene; structure in this motif
123  mediates interactions with splicing regulatory proteins and, surprisingly, nuclear actin to affect
124  LMP2 splicing (a novel role for actin) (Kumarasinghe & Moss 2019).
125
126  Although this analysis of EBV led to interesting discoveries, there were several limitations to this
127  previous approach: (1) motifs that were not conserved (e.g. motifs unique to EBV) were
128  potentially missed; (2) multiple alternative models were predicted for the same nt in different
129  analysis windows; and (3) subjectively selected regions, comprised of overlapping windows with
130  favorable metrics, were selected for generating final models. To address this, in this current
131  study we used a new computational pipeline for RNA secondary structure analysis, ScanFold,
132  which addresses these limitations and has been successfully applied to the Zika and HIV-1
133  genomes (Andrews et al. 2018) as well as the human MYC mRNA (O'Leary et al. 2019). The
134  ScanFold pipeline divides RNA structural motif discovery into two parts; the first step is
135  accomplished using ScanFold-Scan, where individual sequences are analyzed using a sliding
136  window approach. This generates overlapping windows covering each nt of the sequence. For
137  each window ScanFold-Scan calculates 4 folding metrics: (1) The minimum free energy
138  (MFE) $\Delta G°$, which predicts optimal thermodynamic stability. (2) The $\Delta G°$ z-score, which
139  measures the stability of native vs. random sequence. (3) A p-value for the $\Delta G°$ z-score, which
140  is a quality control metric that calculates the fraction of randomized sequences with greater
141  stability than native: p-values closer to 0 indicate a high confidence z-score. (4) The ED
142  (ensemble diversity), which indicates the structural diversity of Boltzmann-weighted
143  conformations calculated from a partition function. The second step, carried out by the program
144  ScanFold-Fold, builds secondary structural models from base pairs likely to be functionally
145  significant. ScanFold-Fold first compiles and averages ScanFold-Scan metrics for every
146  base pairing arrangement (for every nucleotide) predicted across the sequence. The key metric
147  used for selecting the most favorable bp for a nt is the $\Delta G°$ z-score$_{avg}$, which highlights bp that
148  are *consistently* found in low z-score windows. This process was able to detect (and model with
149  accuracy) all known structured motifs in the HIV-1 and ZIKV genomes (Andrews et al. 2018).
150  Further analysis revealed the entirety of ScanFold-Fold results for HIV, ZIKV, and Hepatitis C

151  genomes were consistent with experimentally derived models – where the models from regions
152  with the most negative z-score$_{avg}$ values matched experimentally derived predictions and
153  models in positive z-score$_{avg}$ regions correlated to unstructured regions (Andrews et al. 2019).
154  These results showed that `ScanFold` can serve not only as a motif discovery pipeline (Andrews
155  et al. 2018), but as a tool to generally characterize RNA structural landscapes (Andrews et al.
156  2019).
157
158  In this current study we revisit EBV using this new approach, finding evidence for additional
159  functional elements in EBV including one that was experimentally characterized. Significantly, in
160  this new study, focus is placed on the more prevalent EBV type 1 RefSeq genome (Smatti et al.
161  2018) (rather than the type 2 genome previously analyzed); this current study, however, also
162  includes data for the type 2 genome. Additionally, all sequenced human Herpesvirus RefSeq
163  genomes have been analyzed to allow for global comparisons between genomes and to
164  facilitate the identification of specific viral RNA motifs for further structure/function analyses. All
165  results for EBV and other human herpesviruses are archived on the RNAStructuromeDB a
166  public resource archiving RNA secondary structural data for humans and our pathogens
167  (Andrews et al. 2017).
168
169  **Methods**
170
171  Human herpesvirus genome sequences were acquired from the NCBI RefSeq database
172  (O'Leary et al. 2016): EBV-1 (NC_007605.1), EBV-2 (NC_009334.1), HCMV (NC_006273.2),
173  HHV-6A (NC_001664.4), HHV-6B (NC_000898.1), HHV-7 (NC_001716.2), HSV-1
174  (NC_001806.2), HSV-2 (NC_001798.2), KSHV (NC_009333.1), VZV (NC_001348.1). Each
175  sequence was analyzed using the `ScanFold` pipeline. First, genome sequences were scanned
176  using a sliding analysis window of 150 nt with a single nt step size using `ScanFold-Scan`. In
177  each analysis window, the minimum Gibbs free energy of folding (ΔG) and its associated
178  secondary structure were predicted, assuming the genome sequence was transcribed into RNA.
179  A ΔG z-score is calculated by comparing the native sequence to the average ΔG of 50
180  randomized sequences (shuffling nucleotides) according to the following equation: z-score =
181  $(\Delta G_{native} - \Delta G_{random, average}) / \sigma$; here, σ is the standard deviation of all calculated ΔGs. Negative z-
182  scores indicate the number of standard deviations more stable the native sequence is vs.
183  random. A p-value is calculated, which is a quality control metric for the z-score that measures
184  the fraction of random sequences more stable than native (ideally, this will be close to zero).
185  The ensemble diversity is calculated from a partition function, which measures the diversity of
186  folds in the possible folding ensemble. Here, low numbers indicate that there is a single
187  dominant conformation, while higher numbers indicate alternative conformations or a lack of
188  stable structure. From the partition function a centroid structure is also calculated, this is the
189  secondary structure model that has the least difference with all other Boltzmann weighted
190  conformations in the ensemble—thus, it can be considered as a structure that represents the
191  ensemble fold. The parameters used for window/step size and randomizations were previously
192  optimized (Andrews et al. 2018).
193
194  The output of `ScanFold-Scan` were then analyzed using the program `ScanFold-Fold`. Here,
195  unique motifs are generated by creating consensus secondary structures across all overlapping
196  scanning windows, where base pairs are weighted by the z-score of the window in which they
197  occur. Motifs are thus generated from the base pairs that most contribute to the unusual
198  structural stability of a sequence. Presumably, the evolved order of these sequences creates
199  the particular stability of these pairs, indicating their potential for function (e.g. evolution working
200  to stabilize such structures). A key component of `ScanFold-Fold` is the decision-making

201  process used to define base pairs when multiple, unusually stable, pairs can be predicted for a
202  residue; here, the algorithm selects the pairing partner that most consistently contributed to low
203  z-score windows throughout the scan (see Methods of (Andrews et al. 2018)). For regions of
204  interest, these unusually stable base pairs were used as constraints for generating global
205  secondary structural models. This was done to "fill in" base pairs that were "missing" due to the
206  generation of the consensus structure (e.g. where multiple, equally stable, pairs were averaged
207  out) and to deduce potential long-range interactions, which spanned regions larger than the
208  (150 nt) windows used. To do this, the program `RNAfold` (Lorenz et al. 2011) was used to
209  predict the minimum free energy secondary structure possible, given the constraint of forming
210  all `ScanFold-Fold` with average z-score bp < -1 or -2 (described as the "Global Refold" option
211  in (Andrews et al. 2019)).

212
213  To test predicted models for these select regions, a comparative sequence/structure analysis
214  was performed. The EBV-1 sequence was used in a `BLASTn` analysis to deduce homologous
215  sequence. These were aligned using the program `MAFFT` (Katoh et al. 2009; Katoh & Frith
216  2012; Katoh et al. 2005; Katoh et al. 2002; Katoh et al. 2017; Katoh & Standley 2013; Katoh &
217  Standley 2016; Katoh & Toh 2008a; Katoh & Toh 2008b; Katoh & Toh 2010; Kuraku et al. 2013;
218  Nakamura et al. 2018; Yamada et al. 2016) implementing the G-INS-i strategy, which is
219  optimized for sequences with global homology. The `ScanFold-Fold` constrained secondary
220  structure model for EBV-1 was mapped to this structure and conservation of base pairing was
221  assessed as the percentage of canonical pairing (GC/CG, AU/UA, or GU/UG base pairs)
222  observed across aligned positions. Mutations were tracked across model base pairs to identify
223  structure-preserving mutations that were potentially consistent (single point mutations) or
224  compensatory (double point mutations) with respect to the structure model.

225
226  To experimentally test RNA structural motifs for their effects on post-transcriptional gene
227  regulation, select sequences were inserted into the pIS2 vector downstream of its _Renilla_
228  Luciferase (RL) gene (within the 3' UTR). All conditions were tested in biological triplicate as
229  described in (O'Leary et al. 2019). Briefly, RL constructs, along with _Firefly_ (FF) Luciferase were
230  transfected into HeLa cells and incubated at 37°C and 5% $CO_2$ while maintained in DMEM
231  supplemented with 10% FBS, penicillin, streptomycin, and L-glutamine. After 24 h, transfected
232  cells were split between a 24 well plate (for qPCR) and a 96 well plate (for luciferase assays)
233  and incubated an additional 24 h before analysis. Dual luciferase assays were designed and
234  performed following recommendations of (Van Etten et al. 2013): cells were lysed, and
235  luciferase activity was measured using Promega's Dual Luciferase Reagent Assay kit with a
236  collection time of 10 seconds. Relative response ratios (RRR), calculated as the ratio of RL to
237  FF relative light units (RLUs), were calculated for each sample (Table S2).

238
239  The RT-qPCR analysis was carried out as described in (O'Leary et al. 2019): briefly, whole cell
240  RNA was extracted in TRIzol followed by a Dnase I treatment (NEB) for 2h at 37°C. RNA was
241  further purified with Zymo's RNA Clean and Concentrator kit before reverse transcription which
242  was carried out using 1 μg of purified RNA, random hexamers, and Superscript III
243  (ThermoFisher). Transcript abundance was then measured via qPCR and analyzed using the
244  ΔΔCt method, where the relative abundance of RL was calculated using FF as the reference
245  gene. The results of both RT-qPCR and dual luciferase assays can be seen in Table S2.

246
247
248  **Results**
249
250  _Global results_

251
252 Average MFE ΔG values across each genome ranged from -24.8 kcal/mol (HHV-7) to -64.1
253 kcal/mol (HSV-2; Table 1). This large difference in predicted RNA folding stability, as one would
254 expect, correlates with the GC%, which ranged from 36.2% to 70.4%. The average z-score and
255 ensemble diversity (ED), however, do not follow trends for ΔG or GC%. The z-score quantifies
256 the greater-than-random folding stability of an RNA sequence and is primarily dependent on the
257 sequence *order* and not its composition. Likewise, the ED value indicates the diversity of
258 potential structural conformations in an RNA's folding ensemble, which also appears to be an
259 evolved property of ordered/functional RNA sequences (Moss 2018a). The average z-score
260 ranged from -0.09 (HHV-6B) to -0.63 for EBV-1, while the average ED ranged from 33.7 (HHV-
261 7) to 32.3 (EBV-1). Indeed, all folding metrics suggested functional structured motifs were most
262 prevalent in the EBV type 1 genome.
263
264 Unsurprisingly, the predictions for EBV-1 and EBV-2 were almost identical (Table 1), as the two
265 genomes are only ~5% different in sequence and the majority of this difference clusters within
266 one gene region (EBNA-2) (Tzellos & Farrell 2012). The percentage of the EBV genome
267 spanned by windows with z-scores < -1 (one standard deviation more stable than random) was
268 predicted to be 35.1% and 34.4% for types 1 and 2, respectively. This is consistent with
269 previous predictions using RNAz, which predicted ~30% of the genome to fall within low z-score
270 windows (Moss & Steitz 2013). These values are stronger than any other human herpesvirus
271 genome: e.g. the next highest percentage of windows with z-score < -1 were for KSHV (25%).
272 This also held true for windows with z-scores < -2 (two standard deviations more stable than
273 random), where EBV-1 and EBV-2 had 13.2% and 12.7% of their windows below this stricter
274 cutoff: the next highest percentage was for HSV-1 (8.1%). The virus with the smallest fraction of
275 its genome spanned by low z-score windows was HHV-6B (19.0% and 4.7% for the -1 and -2 z-
276 score cutoffs, respectively).
277
278 In the prediction of the MFE ΔG for each analysis window, a model secondary structure is also
279 generated. Across each genome this resulted in many predicted base pairs (e.g. the HCMV
280 genome has 1,134,919 predicted base pairs; Table 1), where a specific nucleotide can be
281 paired differently across several overlapping windows. As a result, *many potential base pairing*
282 *partners may be predicted per nucleotide*, resulting in the total number of predicted base pairs
283 to often exceed genome length; a confounding factor which has proved to be a challenge in
284 RNA 2D structure modeling (see Introduction). The ScanFold-Fold algorithm confronts this
285 challenge by predicting the single most likely orientation for each nucleotide based on its
286 contributions to low z-score windows (indicating ordered stability and, potentially, function). The
287 viruses that had the greatest percentages of low z-score base pairs were EBV-1 and EBV-2:
288 EBV-1 had 8.8% of its base pairs predicted with average z-score < -1 and EBV-2 had 2.6% of
289 its base pairs predicted with average z-score < -2 (Table 1). Additionally, ScanFold-Fold
290 extracts each discrete structural motif (i.e. single hairpins or multi-branched stem loops)
291 containing at least one base pair with an average z-score < -2. This results in a list of motifs for
292 each genome, where again, EBV-1 and EBV-2 had the greatest number of motifs with 858 and
293 870 respectively.
294
295 In summary, all predictions indicate a particular importance for functional RNA secondary
296 structures encoded within the EBV genome vs. other herpesviruses. The ScanFold-Fold
297 results also give us a means of generating motifs of interest for further analysis.
298
299 *Known structural motifs in EBV*
300

301 The `ScanFold` analysis was able to deduce the location and, in many cases, the secondary
302 structure of known RNA structural motifs in EBV-1. All known EBV-1 miRNA sequences fell
303 within low z-score regions and were contained within hairpins formed by low z-score base pairs.
304 For example, Figure 1 summarizes `ScanFold` results for the EBV-1 BART miRNA gene cluster.
305 All but one miRNA sequence fall within hairpins comprised of base pairs with average z-score <
306 -2 (Fig. 1a; ebv-BART-5 miRNAs occurred in a hairpin comprised of z-score < -1 base pairs).
307 Features of the `ScanFold-Fold` model secondary structures for the EBV miRNAs are
308 consistent with what is known about active precursor (pre)-miRNA hairpins. For example, the
309 ebv-BART3* and ebv-BART3 sequences are annealed to each other, are offset by two
310 nucleotides and contain internal bulge loops (Fig. 1b). A similarly sized hairpin was detected
311 elsewhere in the EBV genome, which corresponds to the EBV viral small nucleolar (v-sno)RNA1
312 (Fig. 2) (Hutzinger et al. 2009). All model pairs for the v-snoRNA are correctly predicted
313 (consistent with C/D box snoRNAs); however, the average z-scores, while negative, were not
314 lower than -1 (Fig. 2a). Thirteen additional base pairs are predicted beyond the annotated v-
315 snoRNA region (orange pairs in Fig. 2b).
316
317 `ScanFold-Fold` was able to detect the terminal hairpin loops of very large hairpin structures
318 found in ncRNAs from two regions. The first hairpin (Fig. S1) occurs in the W repeat region,
319 which generates the stable intronic sequence (sis)RNA-2. This sequence occurs 5-8X in
320 circulating viral strains and the transcribed sisRNAs accumulate to high abundance in infected
321 cells (Moss & Steitz 2013) and appear to be important to EBV-induced transformation of human
322 B cells (Bridges et al. 2019; Szymula et al. 2018). The second region is the latent origin of
323 replication (oriP), which is transcribed bidirectionally in lytically reactivated cells. In both oriP
324 transcripts, tandem repeat sequences anneal to each other to form very long hairpins, which
325 interact with human ADAR (adenosine deaminase acting on RNA), are hyper-edited, localized
326 to the nucleus, interact with paraspeckle assembly factors, and play likely roles in immune
327 modulation (Cao et al. 2015a). `ScanFold-Scan` finds the terminal hairpins of each oriP
328 transcript in windows with less than -5 z-scores, however, the final model reported by the
329 `ScanFold-Fold` algorithm predicts multiple local inter-repeat interactions (Fig. S2). In both the
330 sisRNA-2 and oriP transcripts, long range base pairs in each hairpin cannot be predicted, as
331 they span distances greater than the window size used (150 nt here). This highlights an
332 important limitation of `ScanFold` and, indeed, all scanning window approaches.
333
334 `ScanFold-Fold` was able to successfully predict most of the known structure of the EBV
335 encoded small RNA (EBER)1 (Fig. 3), a hyper-abundant viral ncRNA with implications to
336 oncogenesis (Iwakiri 2016; Moss et al. 2014). Similar to the long hairpins discussed above,
337 EBER1 long-range base pairs were not predicted, however, 61% of the pairs not spanning more
338 than 150 nt were correctly predicted; the majority (14/15) with average z-scores < -2. In EBER2,
339 however, no model base pairs were predicted (Fig. 4) and, indeed, few base pairs with low z-
340 scores were predicted. The 21 base pairs with average z-scores < -1 do not correspond to any
341 known EBER2 structures. This is due to the poorer predicted metrics for windows spanning the
342 annotated EBER2 region; indeed, windows spanning EBER2 had positive z-scores, indicating a
343 sequence that is potentially ordered to be loosely structured (Andrews et al. 2019). Interestingly,
344 the regions of EBER2 that fall within positive z-score regions correspond to an interaction site
345 that binds to nascent transcripts from the terminal repeat (TR) region of the EBV genome. This
346 interaction facilitates the recruitment of the human PAX5 regulatory protein to the TR genomic
347 region to promote lytic replication (Lee et al. 2015). Interestingly, `ScanFold` predictions in the
348 TR region find that the EBER2 interaction sites are *also* spanned by positive z-score windows
349 and `ScanFold-Fold` predictions suggest that these regions are ordered to be
350 unstructured/accessible.

351

352 *New structural motifs in EBV*

353

354 In addition to finding known EBV structures, `ScanFold` predicts many additional motifs with
355 putative functions. We focused our attention on one region which was particularly rich in base
356 pairs predicted to have average z-scores that were exceptionally low (< -2). This region
357 approximately spans EBV bp 48,800 to 50,200 and overlaps several important viral genes (Fig.
358 5). The first cluster of predicted motifs overlaps three lytic genes BFRF1, BFRF2, and BFRF3,
359 which play essential roles in nuclear egress(Farina et al. 2005) (the first step in virion release
360 into infected cells), regulation of late viral gene expression(Aubry et al. 2014), and assist in the
361 assembly of infections particles(Henson et al. 2009), respectively. Each gene overlaps each
362 other, where BFRF2 and BFRF3 both overlap the 3' UTR of BFRF1 and BFRF3 overlaps the 3'
363 UTR of BFRF2. Thus, functional structural motifs can be playing multiple roles in the post-
364 transcriptional control of BFRF1-3. We focused our attention, however, on potential roles in the
365 3' UTR of BFRF1. A 606 nt fragment sequence (bp 48,856 – 49,461), corresponding to the
366 motifs predicted with average z-scores less than -1 in the BFRF1 3' UTR were further analyzed.

367

368 All base pairs predicted by `ScanFold-Fold` to have average z-scores < -1 were constrained to
369 be paired to each other and the remaining sequence was refolded using the `RNAfold` algorithm
370 (the same algorithm used in `ScanFold-Scan` to predict windows(Lorenz et al. 2011)), which
371 predicts the global MFE structure. This fills in base pairs that are "missing" in the `ScanFold-`
372 `Fold` consensus (Fig. 6a), allows for potential long-range interactions to occur and allows for
373 base pairs that do not contribute to the low z-score of this region to form. The resulting global
374 model of this region is shown in Fig. 6b. The overall fold can be divided into eight motifs. In most
375 cases a motif was simply the `ScanFold-Fold` predicted structure. In Motif (M)2 and M6,
376 several base pairs were added to extend helices (Fig. 6b). M4, on the other hand, was predicted
377 to form a multibranch loop structure, which encompasses three `ScanFold-Fold` predicted
378 hairpins: M4.1, M4.3 and M4.5. To test the resulting model, a phylogenetic analysis was
379 undertaken. The EBV-1 sequence was queried against the nt database using `BLASTn` (Altschul
380 et al. 1990), which identified 100 putative homologs, all from other EBV strains. Model base
381 pairs were compared to an alignment of EBV-1 and these 100 homologous sequences (Data
382 S1). Canonical base pairing was conserved 99.1% and, when mutations occurred, they
383 generally preserved structure. Of the 202 base pairs predicted in this region, 35 showed
384 evidence of consistent mutation—a single point mutation which preserves base pairing (Fig. 6b).
385 Only two base pairs showed evidence of compensatory mutation—a double point mutation
386 which preserves base pairing. One putative compensatory mutation occurs in the basal stem of
387 M4.5 and the other in M7 (Fig. 6b).

388

389 A second cluster of predicted motifs overlaps the F/Q exons of EBNA1 (Fig. 5). EBNA1 is a key
390 viral gene involved in the replication and partitioning of the EBV genome (Frappier 2012).
391 During latency program I, EBNA1 is the sole expressed protein and transcription begins at the Q
392 promoter (Qp; Fig. 7a). This results in the production of a short Q exon, which is spliced
393 upstream of the U leader exon in the EBNA1 5' UTR, which was previously found to contain a
394 structured internal ribosomal entry site (IRES) that stimulates non-canonical (cap-independent)
395 translation (Isaksson et al. 2003b). During lytic reactivation transcription from the F promoter
396 (Fp) leads to the incorporation of the longer F exon into the 5' UTR of EBNA1 (includes the Q
397 exon sequence; Fig. 7a). To see if the `ScanFold-Fold` predicted structural motifs would
398 persist in the context of this longer 5' UTR, the entire sequence (plus the start codon), were
399 analyzed using the `ScanFold` pipeline. Three hairpins were predicted (HP1-3; Fig. 7b) which
400 recapitulate those predicted in the EBV genome scan (Fig. 5). The most significant base pairs

401 (average z-scores < -2) all occur within HP1, which begins at the first nucleotide transcribed
402 from Fp.
403
404 In the context of the 5' UTR, the downstream F/Q exon nt are predicted to form a novel hairpin,
405 HP4. No significant base pairing was predicted between the F/Q and U leader exons.
406 Interestingly, no base pairs were predicted in the IRES-containing U leader exon with average
407 z-scores < -1. The negative z-score (< 0 but > -1) base pairs predicted here do not correspond
408 to any found in the EBNA1 IRES model structure (Moss et al. 2014). After constraining
409 `ScanFold-Fold` predicted base pairs (< -1) and re-folding the 5' UTR, however, four additional
410 hairpins (HP5-8) were predicted (Fig. 7c). HP7 and HP8 are found in the current consensus
411 model of the EBNA1 IRES (Rfam ID# RF00448) in the Rfam database (Burge et al. 2013; Daub
412 et al. 2015; Gardner et al. 2009; Griffiths-Jones et al. 2003; Griffiths-Jones et al. 2005; Kalvari et
413 al. 2018a; Kalvari et al. 2018b; Nawrocki et al. 2015). HP6 is a novel predicted configuration of
414 the IRES nucleotides and HP5 is a novel predicted hairpin upstream of the IRES, which partially
415 overlaps (incorporating two nucleotides of) the F/Q exon (Fig. 7c). HP6-8 are predicted to occur
416 within a multibranch loop structure that sits above an extensive basal stem structure.
417
418 To check for potential conservation of the model structure, as well as identify putative consistent
419 and compensatory changes to support model base pairs, a phylogenetic comparison was
420 performed using an alignment of 49 EBV sequences (Data S2). Overall conservation of
421 canonical base pairing was 98.0%. There were 22 putative consistent mutations and 3
422 compensatory mutations identified (Fig. 7c). These putative compensatory mutations are found
423 in HP4, HP6 and HP7. No consistent or compensatory mutations were identified within the long
424 basal stem under HP6-8.
425
426 *Experimental analysis of a novel structured region*
427
428 To test potential functions of the motifs overlapping BFRF1, BFRF2, and BFRF3, four constructs
429 were generated that comprise the entire WT structural region and three fragments (containing
430 M1-2, M3-6, and M7-8; Fig. 6c), which were placed downstream (within the 3' UTR) of *Renilla*
431 Luciferase (RL) and expressed in human cell lines. Two values were measured for each
432 construct, as well as an empty vector control (Fig. 6d): the Luciferase activity fold-change
433 (Luciferase activity is a measure of protein expression levels) and the translational efficiency
434 (which normalizes protein levels using the RNA abundance measured by RT-qPCR); see
435 Methods. Compared to the control (C), the WT construct had reduced Luciferase activity, but a
436 ~7-fold increase in translational efficiency. The M1-2 construct saw marked increases in
437 Luciferase activity (~2-fold) and translational efficiency (~36-fold increase). M3-6 reduced the
438 Luciferase activity to wild type levels, however translational efficiency was roughly cut in half
439 compared to control. M7-8 reduced the Luciferase activity and translational efficiency below WT
440 levels (by roughly 60% in both cases). These results indicate the presence of repressive
441 elements in the region spanned by M3-8 and stimulatory elements in M1-2.
442
443 *Structural motifs beyond EBV*
444
445 Beyond EBV, RNA secondary structures have been previously discovered in KSHV. A recent
446 analysis of the KSHV PAN ncRNA (Sztuba-Solinska et al. 2017) (an RNA transcript that is
447 important to late lytic gene expression and the release of progeny virions (Borah et al. 2011;
448 Sun et al. 1996)) generated a global secondary structure model for PAN based on
449 comprehensive SHAPE probing under *in/ex cellulo*, and *in/ex virio in vitro* conditions.
450 `ScanFold-Fold` results for the KSHV genomic region encoding PAN recapitulated many
451 features of the experimentally-informed global model (Fig. 8). Notable, however, is the lack of

452  highly negative average z-score base pairs from the `ScanFold-Fold` predictions: indeed, the
453  majority of base pairs were predicted with negative average z-scores that did not fall below -1.
454  Base pairs predicted with average z-scores below -1 clustered between PAN nucleotides 137 to
455  327; the only motif from the previous global model with average z-score < -1 fell within this
456  region (H8; Fig. 8).

458  Motif base pairs (after refolding; see Methods) are 99.4% conserved in an alignment of PAN
459  homologs (Data S3) and the little variation observed was consistent with base pairing. 11 out of
460  19 motifs predicted by `ScanFold` appeared in the global model (Fig. 8) (Sztuba-Solinska et al.
461  2017) and, with small variations (e.g. see results of the PAN ENE below), recapitulated the
462  experimentally-guided model base pairs. The nucleotides in the 8 `ScanFold` motifs that were
463  not found in the global model were, in that model, generally found in longer range interactions
464  and/or showed high SHAPE reactivity under various probing conditions (indicating single-
465  stranded or loose structure). These results are consistent with the observation that regions with
466  z-scores greater than -1 generally corresponded to higher SHAPE reactivities (Andrews & Moss
467  2019b) (potentially due to the lack of a predominant secondary structure and/or the presence of
468  primary sequence motifs whose activity is facilitated by being unstructured). For example, the
469  motif predicted for nucleotides 34 to 91 contains the M̲ta r̲esponsive e̲lement (MRE; a protein
470  binding site that stabilizes PAN (Massimelli et al. 2011; Tunnicliffe et al. 2019) partially occluded
471  within a weak stem (comprised of GU and AU base pairs; Fig. 8). In the global model the MRE
472  site is highly reactive to SHAPE probing (under ex virio conditions) and the remaining motif
473  bases are contained within long-range helices (H2 and H3 in Figure 1 of (Sztuba-Solinska et al.
474  2017)). The motif spanning nucleotides 343 to 376, however, is modeled as a small alternative
475  hairpin (H12 in Figure 1 of  (Sztuba-Solinska et al. 2017)) in a looped out region with moderate
476  SHAPE reactivity (under ex virio conditions).

478  The best characterized (in structure and function) motif is the ENE (e̲xpression and n̲uclear
479  r̲etention e̲lement) found in the 3' end of the KSHV PAN ncRNA (Conrad et al. 2006). This
480  element was found to increase the stability/lifetime of intronless transcripts, such as the PAN
481  RNA, via the sequestration of the PAN poly(A) tail in an extensive triple-helix structure with
482  bulged uracil nucleotides in the ENE (Mitton-Fry et al. 2010). In the analysis of the KSHV
483  genome, `ScanFold-Fold` was able deduce the location of the ENE and predict a secondary
484  structure in agreement (Fig. 8) with the crystal structure of the PAN ENE(Mitton-Fry et al. 2010)
485  (crystalized nucleotides highlighted in green) and the SHAPE directed model (of the cytoplasmic
486  and *ex virio* RNA; Figures S4 and S7 of (Sztuba-Solinska et al. 2017)).

488  Beyond the KSHV PAN there are many base pairs predicted with higher significance: there are
489  39906 base pairs that have average z-scores < -1 and 9446 < -2 predicted in the KSHV
490  genome. The values for KSHV are on par with predictions for other herpesviruses (Table 1),
491  indicating significant fractions of each genome likely encode functional RNA secondary
492  structures (all results available on the RNAstructuromeDB (Andrews et al. 2017) and in
493  Supplemental Dataset (https://zenodo.org/record/3964325).

496  **Discussion**

498  Innovations present in the `ScanFold` approach made this current report possible. The initial
499  `ScanFold-Scan` analysis generates metrics of local RNA folding that allow for genome-wide
500  comparisons between human herpesviruses. A key finding of this study is that EBV (types 1 and
501  2) both have evidence of containing more sequences that are ordered to form (likely functional)

502 RNA structures than any other herpesviruses: e.g. the average z-score for EBV is ~2X that of
503 the next lowest (KSHV; Table 1). It should be noted, however, that even for the genome with the
504 least negative average z-score, HHV-6B (-0.09; Table 1), significant percentages of the genome
505 are spanned by low z-score windows. These give rise to many predicted base pairs with likely
506 functionality: e.g. for HHV-6B 5.8% and 0.8% of the 607,730 base pairs predicted by
507 `ScanFold-Fold` have average z-scores < -1 (28,525 bp) and < -2 (4,780 bp), respectively.
508 Thus, although RNA secondary structure is predicted to play greater roles in EBV, all human
509 herpesviruses are likely to utilize regulatory RNA structure to a larger extent than previously
510 appreciated. The results of these analyses are made available on the RNAstructuromeDB
511 (Andrews et al. 2017) to facilitate their usage by a wide array of researchers interested in
512 human herpesviruses. Here, results can be browsed alongside genomic annotations (which
513 have been loaded from NCBI for each genome) or any JBrowse compatible file (Buels et al.
514 2016) a user has available. As an example, for EBV-1 we include additional tracks alongside
515 `ScanFold` results to help identify regions of interest; the McIntosh lab has generated RNA
516 sequencing data for the EBV-1 genome as it transitions from a latent to a lytically active state
517 (Frey et al. 2020). The sequencing data from this study can now be seen as a coverage tracks
518 and allows users to quickly assess whether regions highlighted by `ScanFold` fall within actively
519 transcribed regions.
520
521 In addition to facilitating global comparisons, the `ScanFold` approach defines specific motifs of
522 likely function. All predicted motifs containing at least one base pair with an average z-score < -
523 2 have been modeled, annotated, and compiled into individual PDF files for easy viewing
524 (Supplemental Dataset - https://zenodo.org/record/3964325). In addition to presenting the
525 thermodynamic properties for each motif, the NCBI gene features overlapping the motif have
526 been listed. In EBV-1, `ScanFold` was able to identify the locations of and recapitulate elements
527 of known functional structured RNAs (Figs. 1-3 and Figs. S1-S2). For example, although
528 `ScanFold` is not explicitly a miRNA discovery program, it does a strikingly good job of deducing
529 pre-miRNA hairpins. All known EBV miRNAs are contained within hairpins formed by base pairs
530 with z-averages < -1 and the majority have base pairs < -2 (RNAstructuromeDB and
531 Supplemental Dataset - https://zenodo.org/record/3964325): e.g. the BART miRNA cluster
532 shown in Fig. 1. The exceptionally low average z-scores of pre-miRNA base pairs indicates a
533 particular bias in their sequence order, highlighting the strict structural requirements of miRNA
534 maturation (Lee et al. 2002; Zeng & Cullen 2003).
535
536 `ScanFold` results may also provide information on the processing of another EBV ncRNA, v-
537 snoRNA1. In addition to predicting a C/D box snoRNA-like hairpin in the annotated v-snoRNA1
538 region, additional base pairs are predicted that lengthen the hairpin (Fig. 2). Interestingly, these
539 additional predicted base pairs have lower average z-scores than those of the core v-snoRNA1
540 structure. The extended hairpin predicted here may be a precursor structure, required to mature
541 the v-snoRNA1 (Hutzinger et al. 2009; Matera et al. 2007). It is also interesting that, while being
542 negative, the majority of core base pairs are not predicted to have average z-scores below -1.
543 This may be to facilitate intermolecular RNA-RNA interactions with target sites on the host
544 RNAs: e.g. utilizing the D' and D sites(Hutzinger et al. 2009) annotated on Fig. 2.
545
546 Similar results were found in the EBER2 and TR regions of EBV, which form biologically
547 essential intermolecular RNA-RNA interactions with each other (Fig. 4) (Lee et al. 2015). But for
548 several isolated base pairs, the EBER2 TR interaction region is predicted to have unusually
549 stable base pairs (Fig. 4a) as are the two interaction sites in the TR RNA (Fig. 4b), which occur
550 within a region of highly positive predicted z-score—indicating they may be ordered to be
551 unusually unstable (presumably to facilitate intermolecular base pairing). None of the EBER2

552   reference structure (Rfam ID# RF02712) base pairs are predicted by `ScanFold-Fold`(Fig.
553   4c); emphasizing its loose secondary structure, which is also supported by previous
554   experimental analyses (Lee et al. 2015).
555
556   In contrast to EBER2, EBER1 is predicted to contain base pairs with average z-scores < -2,
557   indicating a high degree of ordered structure, which is corroborated by experimental
558   analyses(Glickman et al. 1988; Lee et al. 2015). These low z-score base pairs form three
559   hairpins in the reference structure (Rfam ID# RF01789), with the two remaining reference
560   structure hairpins being predicted by base pairs with average z-score < -1 (Fig. 3). The
561   remaining 46% of base pairs in the reference structure that were not predicted by `ScanFold-`
562   `Fold` highlight an important limitation of `ScanFold`: it cannot predict longer range base pairs
563   (due to the limitation of the window size used). To predict the reference structure, however,
564   `ScanFold-Fold` results can aid modeling by providing base pairs to use as constraints (e.g.
565   Fig. 3c). *A priori* knowledge of the final transcript sequence is, however, essential here. This
566   highlights another important limitation of whole genome scans: defining the boundaries of a
567   structured domain or ncRNA. In EBER1, for example, 10 base pairs with average z-scores < -2
568   are predicted between EBER1 and upstream EBV sequences; as is a single base pair < -1
569   predicted to occur downstream (Fig. 3a). The EBER1 sequence can be co-transcribed as part of
570   intronic sequences for l̲atent m̲embrane p̲rotein (LMP)2 and these structures between EBER2
571   and flanking sequence could play some role in splicing; however, additional work is needed to
572   test this speculative hypothesis.
573
574   Beyond previously described structures in EBV, `ScanFold` uncovered many novel motifs; we
575   focused on structures predicted in a particular "hot spot" with base pairs having average z-
576   scores < -2 (indicating highly ordered sequences/structures; Fig. 5). The motifs overlapping
577   BFRF1-3 do so primarily in 3' UTR sequences (exclusively in BFRF1 and partially in BFRF2),
578   thus their hypothesized roles in regulation of gene expression; as 3' UTRs are particularly rich
579   areas for post-transcriptional control elements (Mayr 2017). Indeed, we found that inclusion of
580   motifs M1 and M2 into a reporter construct stimulated expression of luciferase, while motifs M3
581   to M6 and M7 to M8 both suppressed expression (Fig. 6b). Many factors can be playing roles in
582   the observed effects of these motifs. For example, multiple host regulatory proteins are
583   predicted to target M1-8 (Data S5) as are both host and viral miRNAs (Table S1). Motif structure
584   can play many roles here: by forming specific recognizable motifs; occluding or presenting
585   primary sequence motifs, or by altering the relative positioning of regulatory motifs. Much
586   additional work is required to validate predicted interactions in this region and to define the
587   exact roles of the conserved and unusually stable secondary structures discovered here: e.g.
588   introducing minimal mutations to modulate predicted RNA secondary structures can help
589   determine to what extent RNA structure plays a role in the observed results.
590
591
592   Similarly, we identified four highly stable/conserved hairpins in the F/Q exon of EBNA1, which
593   are, essentially, appended as a single domain to the 5' UTR of this critical viral gene during lytic
594   replication. An interesting observation was that these novel elements had base pairs that were
595   predicted with average z-scores that were much lower than those of the known structures in the
596   downstream U leader exon (Isaksson et al. 2003a) (Fig. 7). Indeed, the EBNA1 IRES base pairs
597   could only be predicted after refolding the 5' UTR with upstream sequences constrained based
598   on the `ScanFold-Fold` predicted base pairs. Phylogenic support is found for all model
599   structures, indicating that all have likely importance; however, the IRES shows much less
600   evidence of unusual stability. Perhaps, as in the previously discussed examples, some flexibility
601   is required in the IRES for its function. What potential roles then, can be proposed for the rigidly

602  defined structures in the F/Q exon? One interesting possibility is that structure here is inhibiting
603  cap-dependent translation (e.g. by sequestering the 5' end of the mRNA in the exceptionally
604  stable HP1; Fig 7), potentially driving the use of the EBNA1 IRES in non-canonical translation.
605  This hypothesis is supported by data from a previous study (Isaksson et al. 2003b), which
606  showed that inclusion of this structured upstream sequence reduced relative luciferase activities
607  of all tested constructs. Furthermore, when the IRES domain was not present, constructs
608  containing the upstream structures had the lowest relative luciferase observed levels.
609
610  **Conclusion**
611
612  A significant outcome of this study was the prediction of extensive functional RNA structures
613  throughout other human herpesviruses. It is likely that, similar to EBV, these motifs are playing
614  important roles in the regulation of herpesvirus biology, infection and disease. Additional studies
615  of these motifs will provide many insights, which are facilitated by making all results available on
616  the RNAstructuromeDB. Similarly, a web server is available (Andrews et al. 2019; Andrews &
617  Moss 2019b) for running `ScanFold` calculations and for aiding in mutational design of
618  constructs used in experimental assays (e.g. using `RNA2DMut`(Moss 2018b)).
619

620  **Acknowledgements**
621  This research was supported by NIH/NIGMS grants R00GM112877 and R01GM133810, as well
622  as by startup funds from the Roy J. Carver Charitable Trust. Thanks also to Nuwanthika
623  Kumarasinghe for her assistance.
624
625  **Competing Interests**
626
627  The authors declare no competing interests.
628
629  **Figure Captions**
630
631  **Figure 1.** Results for the EBV-BART miRNA cluster. (**a**) IGV visualization of results. EBV-1
632  genome coordinates are followed by `ScanFold-Fold` predicted base pairs represented as arcs
633  (colored blue and green for bp with average z-score < -2 and -1, respectively, and yellow for bp
634  with negative average z-scores > -1), and the genome sequence (A, C, G, and T are in green,
635  blue, orange and red, respectively). Below this is a bar-graph showing the window z-score
636  values predicted by `ScanFold-Scan` (each bar sits at the 1$^{st}$ nt of the window and is colored
637  red or blue for negative/positive values, respectively; the range of values are indicated in
638  brackets). A cartoon of the gene structure of EBV-1 is shown in blue with the location of BART
639  miRNAs annotated. (**b**) At the bottom is the `ScanFold-Fold` secondary structure model for the
640  EBV-BART3 pre-miRNA hairpin. Mature sequences are annotated in blue and red.

641  **Figure 2.** Results for the v-snoRNA1. (**a**) IGV visualization of results.  EBV-1 genome
642  coordinates are followed by `ScanFold-Fold` predicted base pairs represented as arcs (colored
643  green for bp with average z-score < -1 and yellow for bp with negative average z-scores > -1),
644  the genome sequence (A, C, G, and T are in green, blue, orange and red, respectively). (**b**) To
645  the right is the `ScanFold-Fold` secondary structure model for v-snoRNA1 where additional
646  sequences, beyond the annotated v-snoRNA1 sequence is colored orange. Proposed D' and D
647  interaction sites with host rRNA sequences are indicated in pink and blue, respectively.

648    **Figure 3.** Results for EBER1. (**a**) IGV visualization of results. At the top of IGV, EBV-1 genome
649    coordinates are shown, followed by `ScanFold-Fold` predicted base pairs represented as arcs
650    and the location of EBER1 is indicated by the blue cartoon. Base pairs in the reference EBER1
651    structure correctly predicted by `ScanFold-Fold` are highlighted in green on the secondary
652    structure in (**b**). (**c**) indicates the correctly predicted pairs after refolding the EBER1 sequence
653    using `ScanFold-Fold` bp with average z-scores < -2 as constraints.

654    **Figure 4.** Results for EBER2 and its interaction sites on the terminal repeat RNA sequence. (**a**)
655    EBV-1 genome coordinates are followed by `ScanFold-Fold` predicted base pairs represented
656    as arcs (colored blue and green for bp with average z-score < -2 and -1, respectively, and
657    yellow for bp with negative average z-scores > -1), the genome sequence (A, C, G, and T are in
658    green, blue, orange and red, respectively). The EBER2 encoding region is indicated with a blue
659    cartoon, this includes an annotation of the terminal repeat (TR) interacting nucleotides in
660    EBER2. (**b**) `ScanFold-Fold` predicted base pairs for the TR region, the corresponding binding
661    region to EBER2, represented in panel A, with the addition of annotations for the two EBER2
662    interaction regions (sites A and B) and a bar-graph showing the window z-score values
663    predicted by `ScanFold-Scan` (each bar sits at the 1st nt of the window and is colored red or
664    blue for negative/positive values, respectively; the range of values are indicated in brackets). (**c**)
665    Secondary structure model of EBER2 with the TR interacting nucleotides annotated in orange
666    and red for Site A and B, respectively. Model duplexes for each interaction are shown to the
667    right.

668    **Figure 5.** Results for the EBV-1 genome region partially encoding BFRF1-3 and part EBNA1.
669    EBV-1 genome coordinates are followed by `ScanFold-Fold` predicted base pairs represented
670    as colored arcs and the location of each gene is indicated by the blue cartoon. The light blue
671    highlighted region is the region used for subsequent experimental analysis.

672    **Figure 6.** Results for the region partially overlapping BFRF1-3. (**a**) IGV visualization of results.
673    At the top of IGV, EBV-1 genome coordinates are shown, followed by `ScanFold-Fold`
674    predicted base pairs represented as colored arcs. (**b**) Secondary structure models predicted
675    after energy minimization using `ScanFold-Fold` base pairs (average z-score < -1) as
676    constraints. Eight motifs (M1-8) were defined (labeled in both the (**a**) and (**b**)). Nucleotides with
677    mutations preserving the shown base pair are circled on the structure models. (**c**) Cartoon
678    showing the locations of three regions (colored green, purple, and red) that were added to
679    downstream of *Renilla* luciferase (RL; diagrams of constructs are shown with RL in yellow and
680    the added fragments of EBV-1 in colors corresponding to the cartoon). (**d**) Results of luciferase
681    assays for each construct; the translational efficiency and luciferase mRNA and protein fold-
682    change (vs. empty vector control [C]) are plotted for C, wild type (WT) and three fragments of
683    the structured region. The translational efficiencies reported here were calculated by dividing the
684    relative protein abundance of RL (RRR values) by mRNA levels of RL ($2^{-\Delta\Delta CT}$ values); see
685    methods for detail and Table S2 for raw values.

686    **Figure 7.** Results for the EBNA1 5' UTR. (**a**) Cartoon showing the EBV-1 genome region with
687    the Fp (used in lytic replication) and Qp (used in latency type I and II) promoters annotated
688    alongside the F/Q, U, and K exons (with start codon colored green).(**b**) `ScanFold-Fold`
689    predicted base pairs represented as colored arcs above the sequence of the EBNA1 5' UTR
690    formed after transcription from Fp. Spliced exon sequences are highlighted and labeled below
691    the arc diagram structure (with highlights colored as in the cartoon). (**c**) Secondary structure

692    model of the EBNA1 5' UTR based on energy minimization using `ScanFold-Fold` base pairs
693    with average z-scores < -1 as constraints. The locations of exon starts and the start codon are
694    annotated; as well, nt that show evidence of undergoing structure-preserving mutations are
695    circled.

696    **Figure 8.** Results for KSHV PAN. (**a**) IGV visualization of results with KSHV genome
697    coordinates followed by `ScanFold-Fold` predicted base pairs represented as arcs (colored
698    green for bp with average z-score < -1, yellow for bp with negative average z-scores > -1, and
699    grey for > 0), the genome sequence (A, C, G, and T are in green, blue, orange and red,
700    respectively). (**b**) The bottom panel contains motif secondary structures predicted by
701    `ScanFold-Fold` depicted using VARNA (Darty et al. 2009). The MRE is represented as an
702    orange highlight on the arc diagrams and as orange bps in the secondary structures. The
703    nucleotides from the ENE motif which have been crystalized(Mitton-Fry et al. 2010) are
704    highlighted in green.

705

706    **Supplemental Information**

707    **Supplemental Dataset.** `ScanFold` output for all genomes. Raw output alongside files used to
708    visualize data on IGV or other genome viewers. Please see README for descriptions of each
709    file. Additional information on using `ScanFold` and analyzing output files is available in
710    (Andrews et al. 2019). Dataset available as a Zenodo repository at:
711    https://zenodo.org/record/3964325 or from the RNAStructuromeDB at
712    https://structurome.bb.iastate.edu/herpesvirus.

713

714    **Supplemental Data S1.** MAFFT alignment of the BFRF1 3'UTR using 100 BLASTn deduced
715    homologous sequences. Closely related sequence alignments can show conservation at both
716    the primary sequence level and secondary structure level. Mutations which conserve base
717    pairing of the secondary structure (i.e. point and compensatory mutations) are evidence of
718    potentially functional structures as evolution is favoring the conservation of the structure across
719    species. Significant `ScanFold` predicted structures can be queried for primary and secondary
720    conservation to further home in on functional regions. The EBV BFRF1 3' UTR is the first
721    sequence in the file, followed by the homologous sequences.

722    **Supplemental Data S2.** MAFFT alignment of the EBNA1 3' UTR using 48 BLASTn deduced
723    homologous sequences. The purpose of MAFFT alignments and importance of resulting data is
724    described above (Supplemental Data S1). The EBNA1 3' UTR is the first sequence in the file,
725    followed by the homologous sequences.

726    **Supplemental Data S3.** MAFFT alignment of the PAN RNA using 33 BLASTn deduced
727    homologous sequences. The purpose of MAFFT alignments and importance of resulting data is
728    described above (Supplemental Data S1). The KSHV PAN RNA is the first sequence in the file,
729    followed by the homologous sequences.

730    **Supplemental Data S4.** RBPMap predictions for BFRF1 3' UTR. File contains predicted binding
731    site location, motif sequence, k-mer sequence, p-value, and z-score for each prediction. RBP

732 binding regions can be cross referenced to ScanFold predicted structures and to miRNA sites,
733 which facilitates hypothesis generation regarding the function of a motif and/or region.

734 **Supplemental Table S1.** A list of `miRDuplex` predictions for BFRF1 3' UTR. Predicted
735 miRNAs are listed along with sequence and positional information of the corresponding target
736 (BFRF1 3' UTR). miRNA binding regions can be cross referenced to ScanFold predicted
737 structures and to RBP sites, which facilitates hypothesis generation regarding the function of a
738 motif and/or region.

739 **Supplemental Table S2.** Results of dual luciferase assays and RT-qPCR for motifs in the
740 overlapping BFRF1-3 region. These values were used to generate the bar/line graph in Fig. 6d.

741 **Supplemental Figure S1.** `ScanFold-Fold` results for the ebv-sisRNA-2 Terminal Hairpin.

742 **Supplemental Figure S2.** `ScanFold` results for the oriP Region of the EBV genome.

743
744 **References**
745

746 Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *J Mol*
747   *Biol* 215:403-410. 10.1016/S0022-2836(05)80360-2
748 Andrews RJ, Baber L, and Moss WN. 2017. RNAStructuromeDB: A genome-wide database for RNA
749   structural inference. *Sci Rep* 7:17269. 10.1038/s41598-017-17510-y
750 Andrews RJ, Baber L, and Moss WN. 2019. Mapping the RNA structural landscape of viral genomes.
751   *Methods*. 10.1016/j.ymeth.2019.11.001
752 Andrews RJ, and Moss WN. 2019a. Computational approaches for the discovery of splicing regulatory
753   RNA structures. *Biochim Biophys Acta Gene Regul Mech*. 10.1016/j.bbagrm.2019.04.007
754 Andrews RJ, and Moss WN. 2019b. Mapping the RNA structural landscape of viral genomes. *Methods*.
755   10.1016/j.ymeth.2019.11.001
756 Andrews RJ, Roche J, and Moss WN. 2018. ScanFold: an approach for genome-wide discovery of local
757   RNA structural elements—applications to Zika virus and HIV. *PeerJ* 6:e6136. 10.7717/peerj.6136
758 Aubry V, Mure F, Mariame B, Deschamps T, Wyrwicz LS, Manet E, and Gruffat H. 2014. Epstein-Barr virus
759   late gene transcription depends on the assembly of a virus-specific preinitiation complex. *J Virol*
760   88:12825-12838. 10.1128/JVI.02139-14
761 Bevilacqua PC, Ritchey LE, Su Z, and Assmann SM. 2016. Genome-Wide Analysis of RNA Secondary
762   Structure. *Annu Rev Genet* 50:235-266. 10.1146/annurev-genet-120215-035034
763 Borah S, Darricarrere N, Darnell A, Myoung J, and Steitz JA. 2011. A viral nuclear noncoding RNA binds
764   re-localized poly(A) binding protein and is required for late KSHV gene expression. *PLoS Pathog*
765   7:e1002300. 10.1371/journal.ppat.1002300
766 Bridges R, Correia S, Wegner F, Venturini C, Palser A, White RE, Kellam P, Breuer J, and Farrell PJ. 2019.
767   Essential role of inverted repeat in Epstein-Barr virus IR-1 in B cell transformation; geographical
768   variation of the viral genome. *Philos Trans R Soc Lond B Biol Sci* 374:20180299.
769   10.1098/rstb.2018.0299
770 Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L,
771   and Holmes IH. 2016. JBrowse: a dynamic web platform for genome visualization and analysis.
772   *Genome Biol* 17:66. 10.1186/s13059-016-0924-1
773 Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, and Bateman A.
774   2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 41:D226-232. 10.1093/nar/gks1005
775 Cao S, Moss W, O'Grady T, Concha M, Strong MJ, Wang X, Yu Y, Baddoo M, Zhang K, Fewell C, Lin Z,
776   Dong Y, and Flemington EK. 2015a. New Noncoding Lytic Transcripts Derived from the Epstein-

777    Barr Virus Latency Origin of Replication, oriP, Are Hyperedited, Bind the Paraspeckle Protein,
778    NONO/p54nrb, and Support Viral Lytic Transcription. *J Virol* 89:7120-7132. 10.1128/JVI.00608-
779    15

780  Cao S, Strong MJ, Wang X, Moss WN, Concha M, Lin Z, O'Grady T, Baddoo M, Fewell C, Renne R, and
781    Flemington EK. 2015b. High-throughput RNA sequencing-based virome analysis of 50 lymphoma
782    cell lines from the Cancer Cell Line Encyclopedia project. *J Virol* 89:713-729. 10.1128/JVI.02570-
783    14

784  Clote P, Ferre F, Kranakis E, and Krizanc D. 2005. Structural RNA has lower folding energy than random
785    RNA of the same dinucleotide frequency. *RNA* 11:578-591. 10.1261/rna.7220505

786  Conrad NK, Mili S, Marshall EL, Shu MD, and Steitz JA. 2006. Identification of a rapid mammalian
787    deadenylation-dependent decay pathway and its inhibition by a viral RNA element. *Mol Cell*
788    24:943-953. 10.1016/j.molcel.2006.10.029

789  Darty K, Denise A, and Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary
790    structure. *Bioinformatics* 25:1974-1975. 10.1093/bioinformatics/btp250

791  Daub J, Eberhardt RY, Tate JG, and Burge SW. 2015. Rfam: annotating families of non-coding RNA
792    sequences. *Methods Mol Biol* 1269:349-363. 10.1007/978-1-4939-2291-8_22

793  Draborg AH, Duus K, and Houen G. 2013. Epstein-Barr virus in systemic autoimmune diseases. *Clin Dev
794    Immunol* 2013:535738. 10.1155/2013/535738

795  Elgui de Oliveira D, Muller-Coan BG, and Pagano JS. 2016. Viral Carcinogenesis Beyond Malignant
796    Transformation: EBV in the Progression of Human Cancers. *Trends Microbiol* 24:649-664.
797    10.1016/j.tim.2016.03.008

798  Farina A, Feederle R, Raffa S, Gonnella R, Santarelli R, Frati L, Angeloni A, Torrisi MR, Faggioni A, and
799    Delecluse HJ. 2005. BFRF1 of Epstein-Barr virus is essential for efficient primary viral
800    envelopment and egress. *J Virol* 79:3703-3712. 10.1128/JVI.79.6.3703-3712.2005

801  Frappier L. 2012. The Epstein-Barr Virus EBNA1 Protein. *Scientifica (Cairo)* 2012:438204.
802    10.6064/2012/438204

803  Frey TR, Brathwaite J, Li X, Burgula S, Akinyemi IA, Agarwal S, Burton EM, Ljungman M, McIntosh MT,
804    and Bhaduri-McIntosh S. 2020. Nascent Transcriptomics Reveal Cellular Prolytic Factors
805    Upregulated Upstream of the Latent-to-Lytic Switch Protein of Epstein-Barr Virus. *J Virol* 94.
806    10.1128/JVI.01966-19

807  Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones
808    S, Eddy SR, and Bateman A. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res*
809    37:D136-140. 10.1093/nar/gkn766

810  Glickman JN, Howe JG, and Steitz JA. 1988. Structural-Analyses of Eber1 and Eber2 Ribonucleoprotein-
811    Particles Present in Epstein-Barr Virus-Infected Cells. *Journal of Virology* 62:902-911.

812  Griffiths-Jones S, Bateman A, Marshall M, Khanna A, and Eddy SR. 2003. Rfam: an RNA family database.
813    *Nucleic Acids Res* 31:439-441.

814  Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, and Bateman A. 2005. Rfam: annotating
815    non-coding RNAs in complete genomes. *Nucleic Acids Res* 33:D121-124. 10.1093/nar/gki081

816  Gruber AR, Findeiss S, Washietl S, Hofacker IL, and Stadler PF. 2010. RNAz 2.0: improved noncoding RNA
817    detection. *Pac Symp Biocomput*:69-79.

818  Henson BW, Perkins EM, Cothran JE, and Desai P. 2009. Self-assembly of Epstein-Barr virus capsids. *J
819    Virol* 83:3877-3890. 10.1128/JVI.01733-08

820  Hutzinger R, Feederle R, Mrazek J, Schiefermeier N, Balwierz PJ, Zavolan M, Polacek N, Delecluse HJ, and
821    Huttenhofer A. 2009. Expression and processing of a small nucleolar RNA from the Epstein-Barr
822    virus genome. *PLoS Pathog* 5:e1000547. 10.1371/journal.ppat.1000547

823  Isaksson A, Berggren M, and Ricksten A. 2003a. Epstein-Barr virus U leader exon contains an internal
824    ribosome entry site. *Oncogene* 22:572-581. 10.1038/sj.onc.1206149

825 Isaksson A, Berggren M, and Ricksten A. 2003b. Epstein-Barr virus U leader exon contains an internal
826       ribosome entry site. *Oncogene* 22:572-581. 10.1038/sj.onc.1206149
827 Iwakiri D. 2016. Multifunctional non-coding Epstein-Barr virus encoded RNAs (EBERs) contribute to viral
828       pathogenesis. *Virus Res* 212:30-38. 10.1016/j.virusres.2015.08.007
829 Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, and
830       Petrov AI. 2018a. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families.
831       *Nucleic Acids Res* 46:D335-D342. 10.1093/nar/gkx1038
832 Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, and Petrov AI. 2018b. Non-
833       Coding RNA Analysis Using the Rfam Database. *Curr Protoc Bioinformatics* 62:e51.
834       10.1002/cpbi.51
835 Katoh K, Asimenos G, and Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods Mol*
836       *Biol* 537:39-64. 10.1007/978-1-59745-251-9_3
837 Katoh K, and Frith MC. 2012. Adding unaligned sequences into an existing alignment using MAFFT and
838       LAST. *Bioinformatics* 28:3144-3146. 10.1093/bioinformatics/bts578
839 Katoh K, Kuma K, Miyata T, and Toh H. 2005. Improvement in the accuracy of multiple sequence
840       alignment program MAFFT. *Genome Inform* 16:22-33.
841 Katoh K, Misawa K, Kuma K, and Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence
842       alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059-3066.
843 Katoh K, Rozewicki J, and Yamada KD. 2017. MAFFT online service: multiple sequence alignment,
844       interactive sequence choice and visualization. *Briefings in Bioinformatics*:bbx108-bbx108.
845       10.1093/bib/bbx108
846 Katoh K, and Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
847       improvements in performance and usability. *Mol Biol Evol* 30:772-780. 10.1093/molbev/mst010
848 Katoh K, and Standley DM. 2016. A simple method to control over-alignment in the MAFFT multiple
849       sequence alignment program. *Bioinformatics* 32:1933-1942. 10.1093/bioinformatics/btw108
850 Katoh K, and Toh H. 2008a. Improved accuracy of multiple ncRNA alignment by incorporating structural
851       information into a MAFFT-based framework. *BMC Bioinformatics* 9:212. 10.1186/1471-2105-9-
852       212
853 Katoh K, and Toh H. 2008b. Recent developments in the MAFFT multiple sequence alignment program.
854       *Brief Bioinform* 9:286-298. 10.1093/bib/bbn013
855 Katoh K, and Toh H. 2010. Parallelization of the MAFFT multiple sequence alignment program.
856       *Bioinformatics* 26:1899-1900. 10.1093/bioinformatics/btq224
857 Kieff E, and Rickinson AJe. 2001. Fields virology.2511-2574.
858 Kumarasinghe N, and Moss WN. 2019. Analysis of a structured intronic region of the LMP2 pre-mRNA
859       from EBV reveals associations with human regulatory proteins and nuclear actin. *BMC Res Notes*
860       12:33. 10.1186/s13104-019-4070-1
861 Kuraku S, Zmasek CM, Nishimura O, and Katoh K. 2013. aLeaves facilitates on-demand exploration of
862       metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity.
863       *Nucleic Acids Res* 41:W22-28. 10.1093/nar/gkt389
864 Lee N, Moss WN, Yario TA, and Steitz JA. 2015. EBV noncoding RNA binds nascent RNA to drive host
865       PAX5 to viral DNA. *Cell* 160:607-618. 10.1016/j.cell.2015.01.015
866 Lee N, Pimienta G, and Steitz JA. 2012. AUF1/hnRNP D is a novel protein partner of the EBER1 noncoding
867       RNA of Epstein-Barr virus. *RNA* 18:2073-2082. 10.1261/rna.034900.112
868 Lee N, Yario TA, Gao JS, and Steitz JA. 2016. EBV noncoding RNA EBER2 interacts with host RNA-binding
869       proteins to regulate viral gene expression. *Proc Natl Acad Sci U S A* 113:3221-3226.
870       10.1073/pnas.1601773113
871 Lee Y, Jeon K, Lee JT, Kim S, and Kim VN. 2002. MicroRNA maturation: stepwise processing and
872       subcellular localization. *EMBO J* 21:4663-4670. 10.1093/emboj/cdf476

873    Lerner MR, Andrews NC, Miller G, and Steitz JA. 1981. Two small RNAs encoded by Epstein-Barr virus
874        and complexed with protein are precipitated by antibodies from patients with systemic lupus
875        erythematosus. *Proc Natl Acad Sci U S A* 78:805-809.
876    Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, and Hofacker IL. 2011.
877        ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26. 10.1186/1748-7188-6-26
878    Massimelli MJ, Kang JG, Majerciak V, Le SY, Liewehr DJ, Steinberg SM, and Zheng ZM. 2011. Stability of a
879        Long Noncoding Viral RNA Depends on a 9-nt Core Element at the RNA 5 ' End to Interact with
880        Viral ORF57 and Cellular PABPC1. *International Journal of Biological Sciences* 7:1145-1160. DOI
881        10.7150/ijbs.7.1145
882    Matera AG, Terns RM, and Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small
883        nucleolar RNAs. *Nat Rev Mol Cell Biol* 8:209-220. 10.1038/nrm2124
884    Mayr C. 2017. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* 51:171-194. 10.1146/annurev-
885        genet-120116-024704
886    McGeoch DJ, Cook S, Dolan A, Jamieson FE, and Telford EA. 1995. Molecular phylogeny and evolutionary
887        timescale for the family of mammalian herpesviruses. *J Mol Biol* 247:443-458.
888        10.1006/jmbi.1995.0152
889    Mitton-Fry RM, DeGregorio SJ, Wang J, Steitz TA, and Steitz JA. 2010. Poly(A) tail recognition by a viral
890        RNA element through assembly of a triple helix. *Science* 330:1244-1247.
891        10.1126/science.1195858
892    Moss WN. 2018a. The ensemble diversity of non-coding RNA structure is lower than random sequence.
893        *Noncoding RNA Res* 3:100-107. 10.1016/j.ncrna.2018.04.005
894    Moss WN. 2018b. RNA2DMut: a web tool for the design and analysis of RNA structure mutations. *RNA*
895        24:273-286. 10.1261/rna.063933.117
896    Moss WN, Lee N, Pimienta G, and Steitz JA. 2014. RNA families in Epstein-Barr virus. *RNA Biol* 11:10-17.
897        10.4161/rna.27488
898    Moss WN, and Steitz JA. 2013. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA
899        structures and a novel stable intronic sequence RNA. *BMC Genomics* 14:543. 10.1186/1471-
900        2164-14-543
901    Nakamura T, Yamada KD, Tomii K, and Katoh K. 2018. Parallelization of MAFFT for large-scale multiple
902        sequence alignments. *Bioinformatics* 34:2490-2492. 10.1093/bioinformatics/bty121
903    Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA,
904        Tate J, and Finn RD. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*
905        43:D130-137. 10.1093/nar/gku1063
906    O'Grady T, Cao S, Strong MJ, Concha M, Wang X, Splinter Bondurant S, Adams M, Baddoo M, Srivastav
907        SK, Lin Z, Fewell C, Yin Q, and Flemington EK. 2014. Global bidirectional transcription of the
908        Epstein-Barr virus genome during reactivation. *J Virol* 88:1604-1616. 10.1128/JVI.02989-13
909    O'Leary CA, Andrews RJ, Tompkins VS, Chen JL, Childs-Disney JL, Disney MD, and Moss WN. 2019. RNA
910        structural analysis of the MYC mRNA reveals conserved motifs that affect gene expression. *PLoS*
911        *One* 14:e0213758. 10.1371/journal.pone.0213758
912    O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White
913        B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E,
914        Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali
915        VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH,
916        Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE,
917        Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P,
918        Murphy TD, and Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status,
919        taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733-745.
920        10.1093/nar/gkv1189

921   Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, and
922         Tuschl T. 2004. Identification of virus-encoded microRNAs. *Science* 304:734-736.
923         10.1126/science.1096781
924   Smatti MK, Al-Sadeq DW, Ali NH, Pintus G, Abou-Saleh H, and Nasrallah GK. 2018. Epstein-Barr Virus
925         Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population:
926         An Update. *Front Oncol* 8:211. 10.3389/fonc.2018.00211
927   Sun R, Lin SF, Gradoville L, and Miller G. 1996. Polyadenylylated nuclear RNA encoded by Kaposi
928         sarcoma-associated herpesvirus. *Proc Natl Acad Sci U S A* 93:11883-11888.
929         10.1073/pnas.93.21.11883
930   Sztuba-Solinska J, Rausch JW, Smith R, Miller JT, Whitby D, and Le Grice SFJ. 2017. Kaposi's sarcoma-
931         associated herpesvirus polyadenylated nuclear RNA: a structural scaffold for nuclear,
932         cytoplasmic and viral proteins. *Nucleic Acids Res* 45:6805-6821. 10.1093/nar/gkx241
933   Szymula A, Palermo RD, Bayoumy A, Groves IJ, Ba Abdullah M, Holder B, and White RE. 2018. Epstein-
934         Barr virus nuclear antigen EBNA-LP is essential for transforming naive B cells, and facilitates
935         recruitment of transcription factors to the viral genome. *PLoS Pathog* 14:e1006890.
936         10.1371/journal.ppat.1006890
937   Tompkins VS, Valverde DP, and Moss WN. 2018. Human regulatory proteins associate with non-coding
938         RNAs from the EBV IR1 region. *BMC Res Notes* 11:139. 10.1186/s13104-018-3250-8
939   Tunnicliffe RB, Levy C, Ruiz Nivia HD, Sandri-Goldin RM, and Golovanov AP. 2019. Structural
940         identification of conserved RNA binding sites in herpesvirus ORF57 homologs: implications for
941         PAN RNA recognition. *Nucleic Acids Res* 47:1987-2001. 10.1093/nar/gky1181
942   Tzellos S, and Farrell PJ. 2012. Epstein-barr virus sequence variation-biology and disease. *Pathogens*
943         1:156-174. 10.3390/pathogens1020156
944   Van Etten J, Schagat TL, and Goldstrohm AC. 2013. A guide to design and optimization of reporter assays
945         for 3' untranslated region mediated regulation of mammalian messenger RNAs. *Methods*
946         63:110-118. 10.1016/j.ymeth.2013.04.020
947   Yamada KD, Tomii K, and Katoh K. 2016. Application of the MAFFT sequence alignment program to large
948         data-reexamination of the usefulness of chained guide trees. *Bioinformatics* 32:3246-3251.
949         10.1093/bioinformatics/btw412
950   Young LS, and Rickinson AB. 2004. Epstein-Barr virus: 40 years on. *Nat Rev Cancer* 4:757-768.
951         10.1038/nrc1452
952   Zeng Y, and Cullen BR. 2003. Sequence requirements for micro RNA processing and function in human
953         cells. *RNA* 9:112-123. 10.1261/rna.2780503

954

# Figure 1

Figure 1. Results for the EBV-BART miRNA cluster.

(**a**) IGV visualization of results. EBV-1 genome coordinates are followed by ScanFold-Fold predicted base pairs represented as arcs (colored blue and green for bp with average z-score < -2 and -1, respectively, and yellow for bp with negative average z-scores > -1), and the genome sequence (A, C, G, and T are in green, blue, orange and red, respectively). Below this is a bar-graph showing the window z-score values predicted by ScanFold-Scan (each bar sits at the 1$^{st}$ nt of the window and is colored red or blue for negative/positive values, respectively; the range of values are indicated in brackets). A cartoon of the gene structure of EBV-1 is shown in blue with the location of BART miRNAs annotated. (**b**) At the bottom is the ScanFold-Fold secondary structure model for the EBV-BART3 pre-miRNA hairpin. Mature sequences are annotated in blue and red.

# Figure 2

Figure 2. Results for the v-snoRNA1.

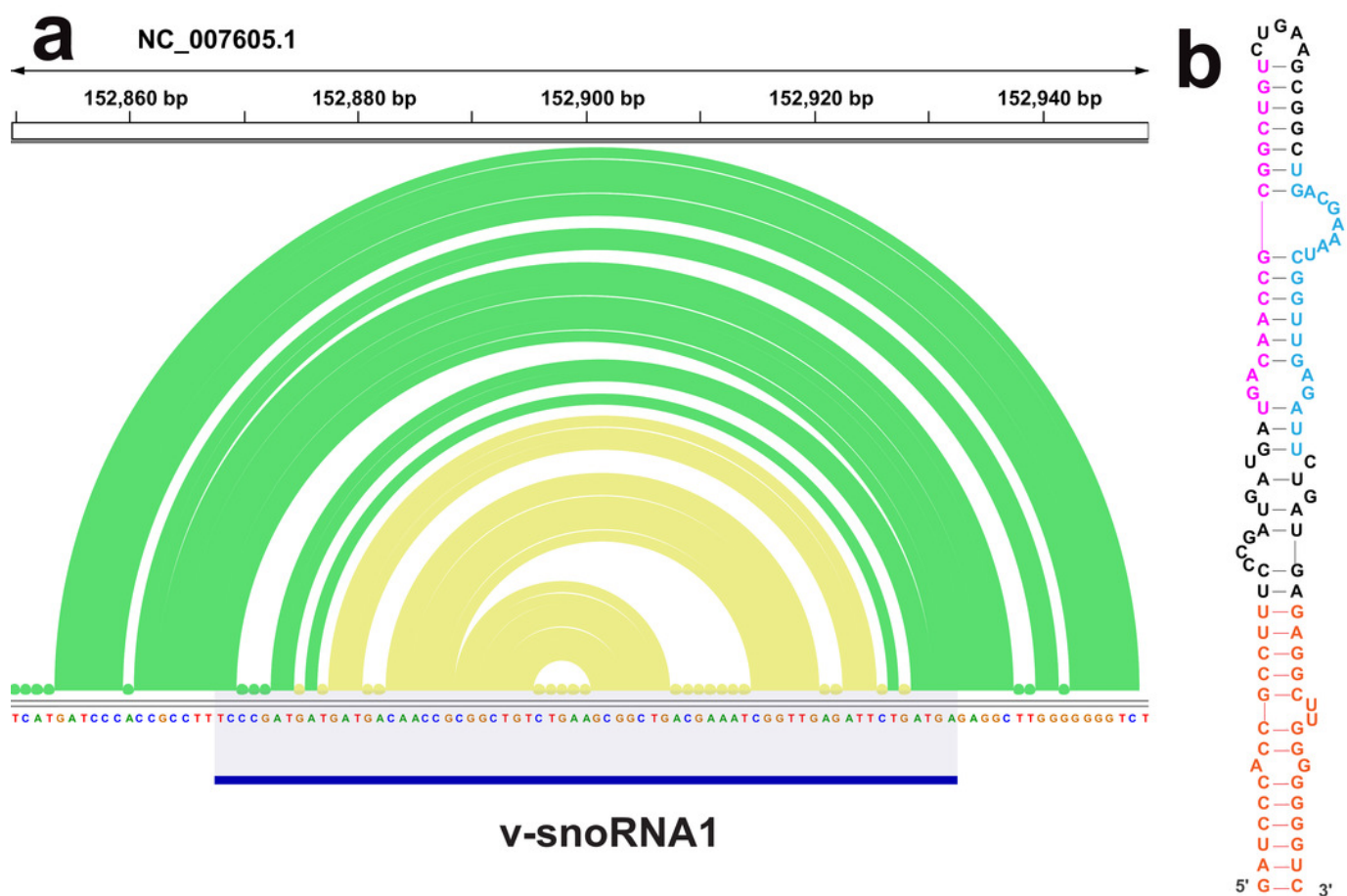
(**a**) IGV visualization of results. EBV-1 genome coordinates are followed by ScanFold-Fold predicted base pairs represented as arcs (colored green for bp with average z-score < -1 and yellow for bp with negative average z-scores > -1), the genome sequence (A, C, G, and T are in green, blue, orange and red, respectively). (**b**) To the right is the ScanFold-Fold secondary structure model for v-snoRNA1 where additional sequences, beyond the annotated v-snoRNA1 sequence is colored orange. Proposed D' and D interaction sites with host rRNA sequences are indicated in pink and blue, respectively.

# Figure 3

Figure 3. Results for EBER1.

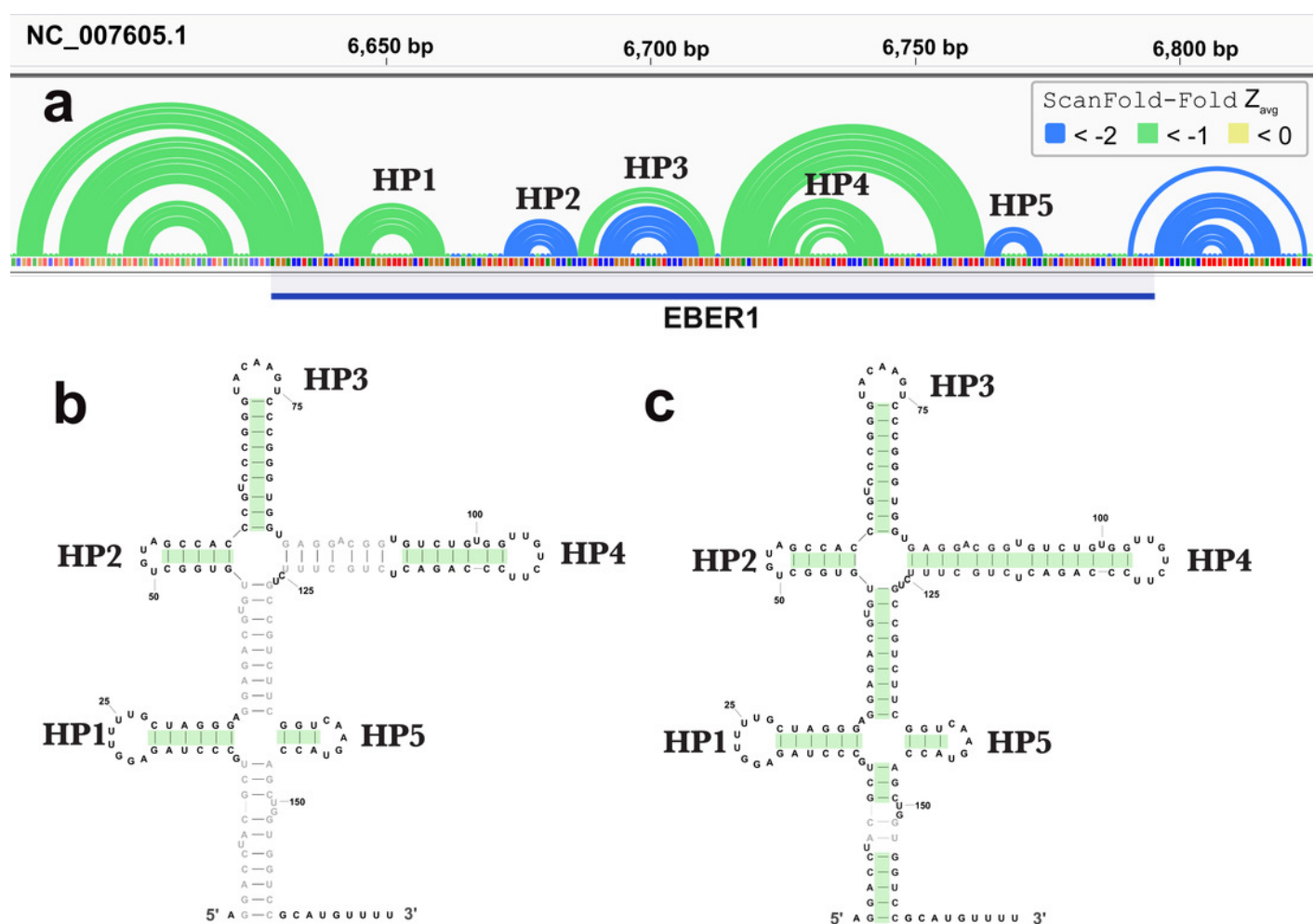
(**a**) IGV visualization of results. At the top of IGV, EBV-1 genome coordinates are shown, followed by ScanFold-Fold predicted base pairs represented as arcs and the location of EBER1 is indicated by the blue cartoon. Base pairs in the reference EBER1 structure correctly predicted by ScanFold-Fold are highlighted in green on the secondary structure in (**b**). (**c**) indicates the correctly predicted pairs after refolding the EBER1 sequence using ScanFold-Fold bp with average z-scores < -2 as constraints.

# Figure 4

Figure 4. Results for EBER2 and its interaction sites on the terminal repeat RNA sequence.

(**a**) EBV-1 genome coordinates are followed by ScanFold-Fold predicted base pairs represented as arcs (colored blue and green for bp with average z-score < -2 and -1, respectively, and yellow for bp with negative average z-scores > -1), the genome sequence (A, C, G, and T are in green, blue, orange and red, respectively). The EBER2 encoding region is indicated with a blue cartoon, this includes an annotation of the terminal repeat (TR) interacting nucleotides in EBER2. (**b**) ScanFold-Fold predicted base pairs for the TR region, the corresponding binding region to EBER2, represented in panel A, with the addition of annotations for the two EBER2 interaction regions (sites A and B) and a bar-graph showing

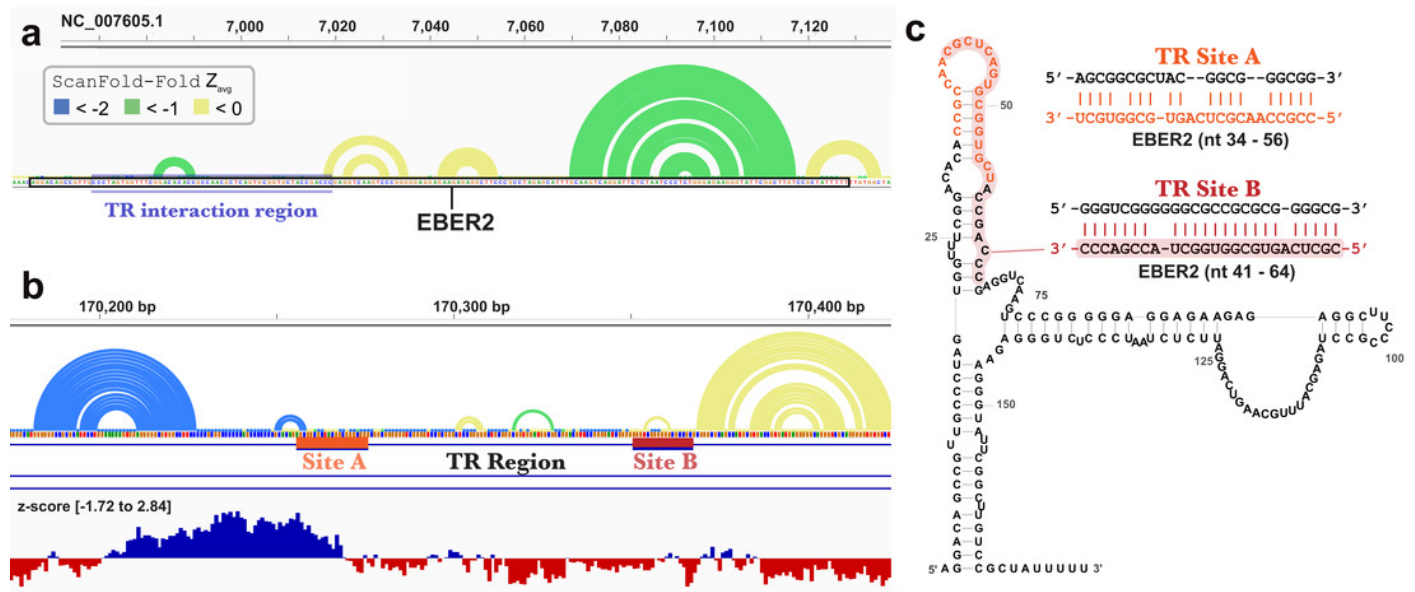
the window z-score values predicted by ScanFold-Scan (each bar sits at the $1^{st}$ nt of the window and is colored red or blue for negative/positive values, respectively; the range of values are indicated in brackets). (**c**) Secondary structure model of EBER2 with the TR interacting nucleotides annotated in orange and red for Site A and B, respectively. Model duplexes for each interaction are shown to the right.

# Figure 5

Figure 5. Results for the EBV-1 genome region partially encoding BFRF1-3 and part EBNA1.

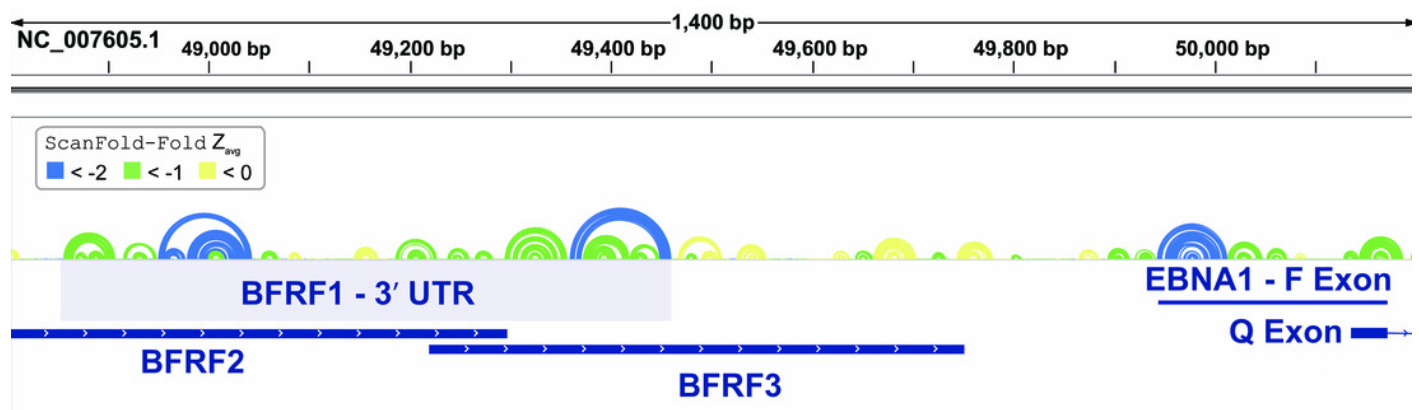
EBV-1 genome coordinates are followed by ScanFold-Fold predicted base pairs represented as colored arcs and the location of each gene is indicated by the blue cartoon. The light blue highlighted region is the region used for subsequent experimental analysis.

# Figure 6

Figure 6. Results for the region partially overlapping BFRF1-3.

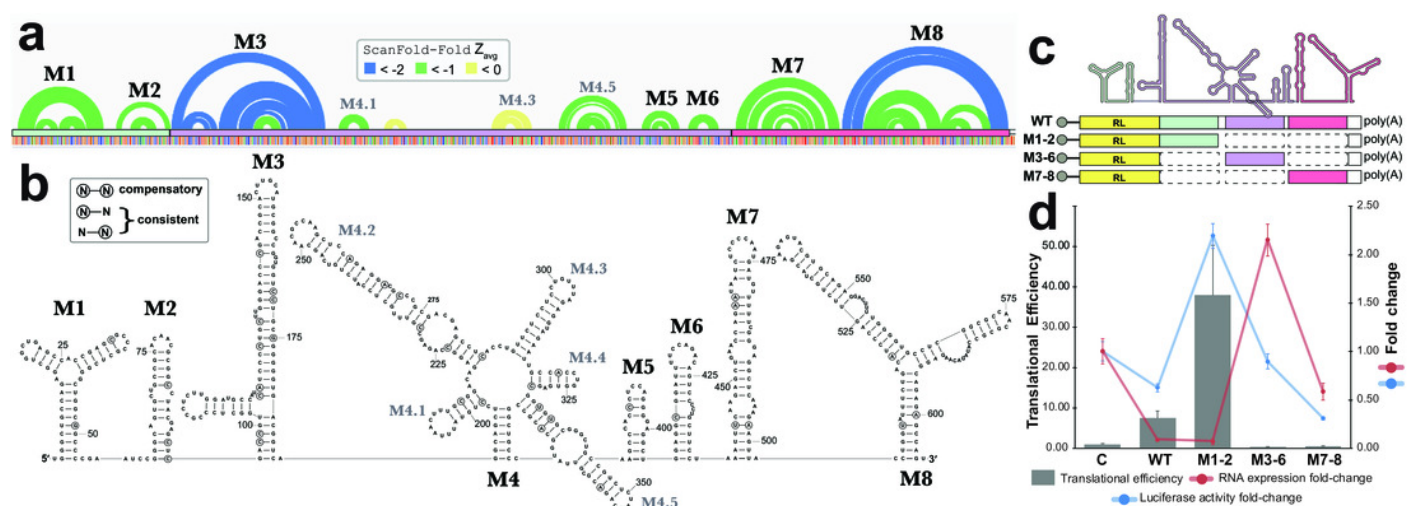
(**a**) The top of this panel shows ScanFold-Fold predicted base pairs represented as colored arcs. (**b**) Secondary structure models predicted after energy minimization using ScanFold-Fold base pairs (average z-score < -1) as constraints. Eight motifs (M1-8) were defined (labeled in both the (**a**) and (**b**)). Nucleotides with mutations preserving the shown base pair are circled on the structure models. (**c**) Cartoon showing the locations of three regions (colored green, purple, and red) that were added to downstream of *Renilla* luciferase (RL; diagrams of constructs are shown with RL in yellow and the added fragments of EBV-1 in colors corresponding to the cartoon). (**d**) Results of luciferase assays for each construct; the translational efficiency and luciferase mRNA and protein fold-change (vs. empty vector c ontrol [C]) are plotted for C, w ild t ype (WT) and three fragments of the structured region. The translational efficiencies reported here were calculated by dividing the relative protein abundance of RL (RRR values) by mRNA levels of RL ($2^{-\Delta DCT}$ values); see methods for detail and Table S2 for raw values.

# Figure 7

Figure 7. Results for the EBNA1 5' UTR.

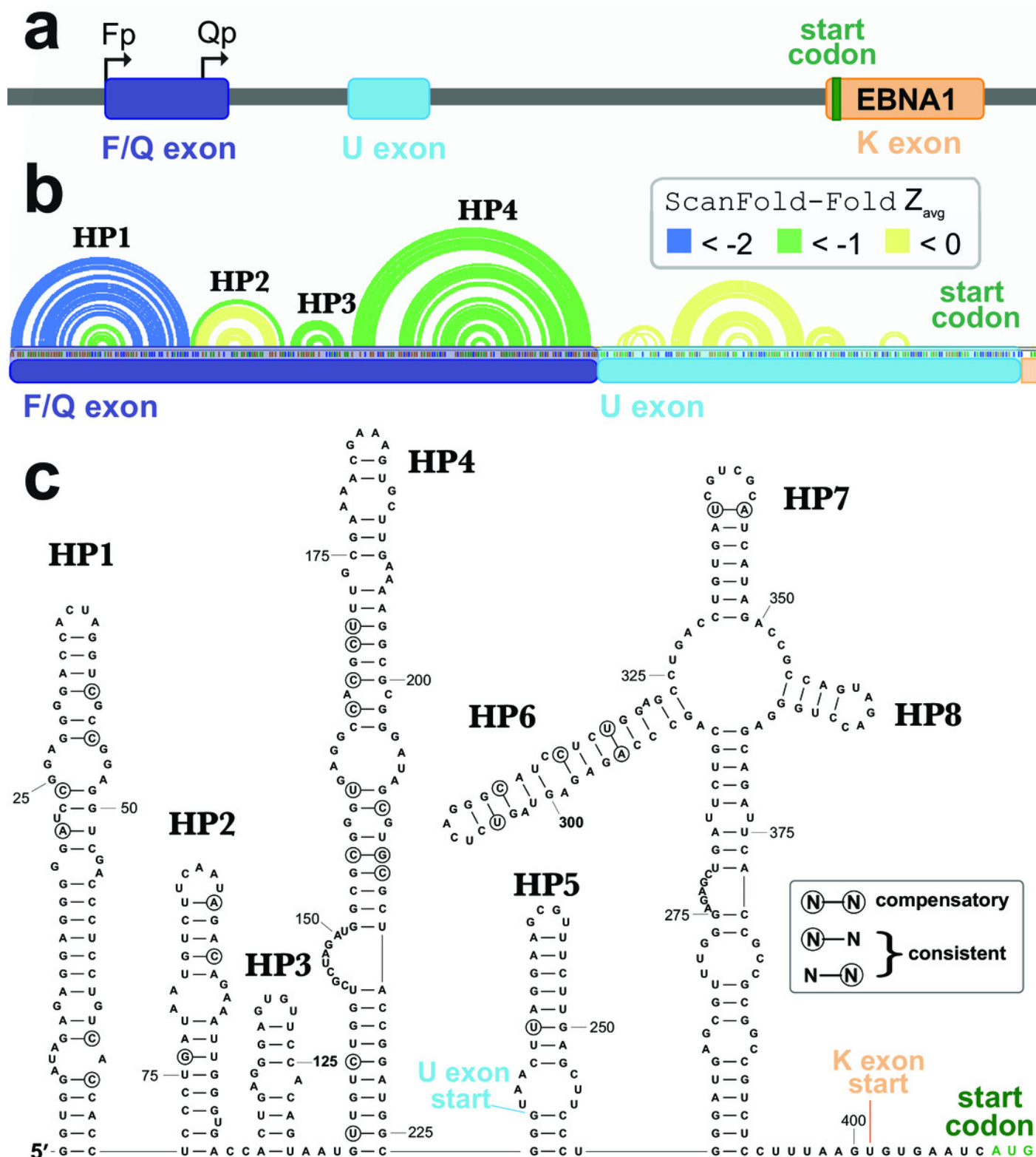
(**a**) Cartoon showing the EBV-1 genome region with the Fp (used in lytic replication) and Qp (used in latency type I and II) promoters annotated alongside the F/Q, U, and K exons (with start codon colored green).(**b**) ScanFold-Fold predicted base pairs represented as colored arcs above the sequence of the EBNA1 5' UTR formed after transcription from Fp. Spliced exon sequences are highlighted and labeled below the arc diagram structure (with highlights colored as in the cartoon). (**c**) Secondary structure model of the EBNA1 5' UTR based on energy minimization using ScanFold-Fold base pairs with average z-scores < -1 as constraints. The locations of exon starts and the start codon are annotated; as well, nt that show evidence of undergoing structure-preserving mutations are circled.

# Figure 8

Figure 8. Results for KSHV PAN.

(**a**) IGV visualization of results with KSHV genome coordinates followed by ScanFold-Fold predicted base pairs represented as arcs (colored green for bp with average z-score < -1, yellow for bp with negative average z-scores > -1, and grey for > 0), the genome sequence (A, C, G, and T are in green, blue, orange and red, respectively). (**b**) The bottom panel contains motif secondary structures predicted by ScanFold-Fold depicted using VARNA (Darty et al. 2009) . The MRE is represented as an orange highlight on the arc diagrams and as orange bps in the secondary structures. The nucleotides from the ENE motif which have been crystalized (Mitton-Fry et al. 2010) are highlighted in green.

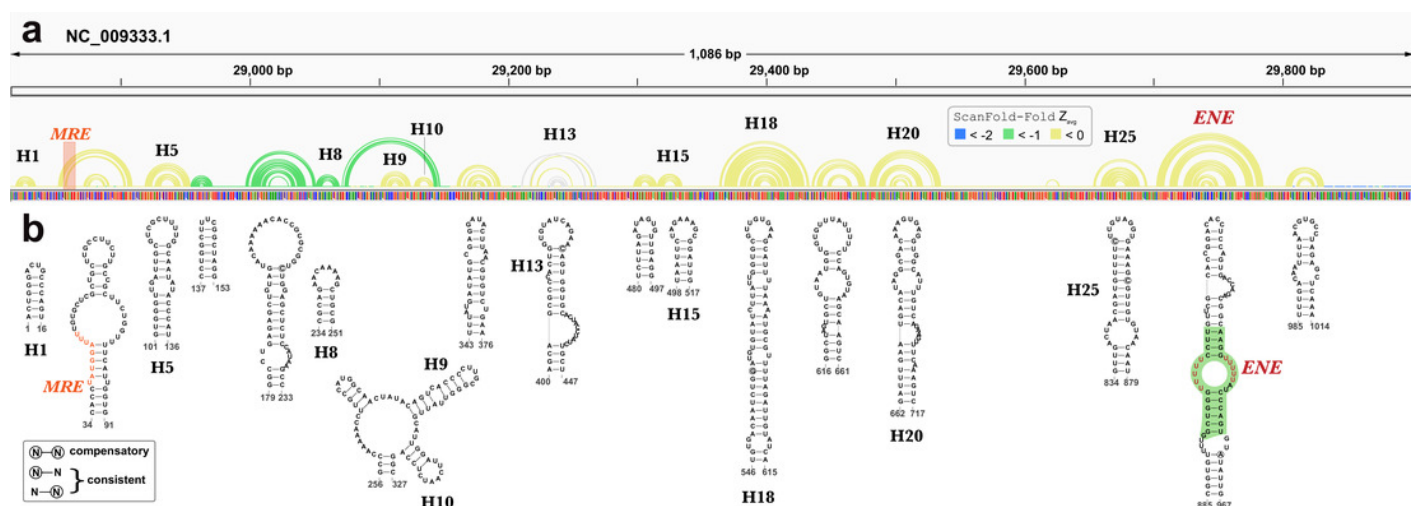



Table 1. ScanFold metrics of the human herpesviruses.

ED (ensemble diversity) is a measure of conformational diversity of RNA. Average Genome ZS (z-score) gives the ΔG z-score averages across all windows spanning each genome. ZS < -1 and < -2 give the percentage of predictions windows with ΔG z-score below each cutoff. Total base pairs gives the total number of stable pairs predicted in each genome (counting both strands and all potential base pairs per nucleotide) and the bp ZS < -1 and < -2 give the percentages of bp below each z-score threshold. Motifs reports the number of individual discrete RNA structures (defined as having a single base helix) containing at least one bp with ZS <-2.

| Virus | GC% | ΔG (kcal/mol) | ED | Average ZS | Windows | ZS < -1 | ZS < -2 | Total bp | bp ZS < -1 | bp ZS < -2 | Motifs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EBV-1 | 59.4% | -50.1 | **32.2** | **-0.63** | 343,348 | **35.1%** | **13.2%** | 835,324 | **8.8%** | 2.5% | 858 |
| EBV-2 | 59.6% | -50.3 | 32.3 | -0.62 | 345,230 | 34.4% | 12.7% | 850,827 | 8.4% | **2.6%** | **870** |
| KSHV | 53.7% | -42.6 | 32.6 | -0.31 | 275,640 | 25.0% | 7.9% | 628,314 | 6.4% | 1.5% | 381 |
| HSV-1 | 68.3% | -60.8 | 32.7 | -0.26 | 304,146 | 24.4% | 8.1% | 856,374 | 5.4% | 1.4% | 436 |
| HHV-7 | 36.2% | -24.8 | 33.7 | -0.26 | 305,862 | 22.8% | 7.6% | 536,275 | 6.8% | 1.6% | 320 |
| HSV-2 | **70.4%** | **-64.1** | 32.7 | -0.20 | 309,052 | 22.8% | 7.5% | 905,657 | 5.0% | 1.2% | 648 |
| HCMV | 57.4% | -46.8 | 32.9 | -0.19 | **470,994** | 22.9% | 6.9% | **1,143,973** | 2.9% | 0.6% | 562 |
| VZV | 46.0% | -34.5 | 32.9 | -0.13 | 249,470 | 20.0% | 5.1% | 498,533 | 5.7% | 1.0% | 226 |
| HHV-6A | 42.4% | -29.7 | 32.7 | -0.12 | 318,458 | 19.6% | 5.4% | 597,872 | 5.9% | 1.2% | 274 |
| HHV-6B | 42.8% | -29.7 | 32.9 | -0.09 | 323,930 | 19.0% | 4.7% | 607,730 | 5.8% | 0.8% | 268 |

2 **Table 1.** ScanFold metrics of the human herpesviruses. ED (ensemble diversity) is a measure
3 of conformational diversity of RNA. Average Genome ZS (z-score) gives the ΔG z-score
4 averages across all windows spanning each genome. ZS < -1 and < -2 give the percentage of
5 predictions windows with ΔG z-score below each cutoff. Total base pairs gives the total number
6 of stable pairs predicted in each genome (counting both strands and all potential base pairs per
7 nucleotide) and the bp ZS < -1 and < -2 give the percentages of bp below each z-score
8 threshold. Motifs reports the number of individual discrete RNA structures (defined as having a
9 single base helix) containing at least one bp with ZS <-2.