Virus Genes (2015) 51:329–337
DOI 10.1007/s11262-015-1248-z

The full-length DNA sequence of Epstein Barr virus from a human gastric carcinoma cell line, SNU-719

Kyung-A Song[1,2] · San-Duk Yang[3] · Jinha Hwang[3] · Jong-Il Kim[3,4,5] · Myung-Soo Kang[1,2]

Received: 17 April 2015 / Accepted: 9 September 2015 / Published online: 12 October 2015

Abstract The consistent presence of Epstein–Barr virus (EBV) in malignant cells of EBV-associated gastric carcinoma (EBVaGC) suggests it plays an important role during the development of EBVaGC. However, the entire genomic sequence of EBV from EBVaGC has yet to be determined. This study first determined, annotated, and analyzed the full genomic sequence of EBV from the naturally infected gastric carcinoma cell line SNU-719 using next-generation sequencing and comparative analyses. In consistent with the notion that EBV sequence isolates better reflect their geographic area than tissue origin, the SNU-719 EBV (named as GC1) was categorized as an East Asian type I EBV. Compared with the prototype B95.8 sequence, SNU-719 EBV contained 1372 variations, with 937 and 435 within coding and non-coding regions, respectively. Of the 937 variations, 465 were non-synonymous changes, while 472 synonymous changes included partial internal deletions in the coding regions of LMP1 and gp350. The RNAseq transcriptome revealed that multiple BART transcripts comprised the majority of EBV RNA reads. The SNU-719 EBV expressed high levels of BART, LF3, BHLF1, and BNLF2. Evidence of RNA editing at multiple sites in the host chromosome was found; however, no evidence of genome integration was seen. The annotated SNU-719 EBV sequence will be a useful reference in future EBVaGC studies.

Keywords Epstein–Barr virus · Gastric carcinoma · EBVaGC · Genome · Next-generation sequencing · RNAseq

Introduction

The Epstein–Barr virus (EBV) is a ubiquitous oncogenic herpesvirus that infects over 90 % of the human population [1, 2]. EBV viral particles have been observed in lymphoma cells from Burkitt's lymphoma (BL) patients and in patients with infectious mononucleosis (IM) [1, 2]. EBV or

Edited by Hartmut Hengel.

Kyung-A Song and San-Duk Yang equally contributed to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11262-015-1248-z) contains supplementary material, which is available to authorized users.

✉ Jong-Il Kim
jongil@snu.ac.kr

✉ Myung-Soo Kang
mkang@skku.edu

[1] Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul, Korea

[2] Samsung Biomedical Research Institute (SBRI), Samsung Electronics Co, Ltd., Seoul, Korea

[3] Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea

[4] Department of Biochemistry, Seoul National University College of Medicine, 103 Daehakno, Jongnogu, Seoul 110-799, Korea

[5] Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul, Korea

its DNA has been found with increasing regularity in diverse malignancies, including nasopharyngeal carcinoma (NPC) [3], T cell lymphoma, Hodgkin's lymphoma [4, 5], and gastric carcinoma (GC) [6].

The incidence of GC is highly variable depending on geography and ethnicity, with a high prevalence in East Asia. Around 10 % of GC patients are diagnosed with EBV-associated GC (EBVaGC). More than 35 and 90 % of patients with post-surgical gastric stump/remnants and lymphoepithelioma-like carcinomas, respectively, have EBVaGC [6, 7]. In EBVaGC, virtually all carcinoma cells contain EBV DNA, with the EBV terminal repeat sequences uniform in length. This would indicate that the tumors might arise from an EBV-infected single cell and that the EBV genome was present during malignant transformation and proliferation [8].The presence of viral genomes in EBV-positive GC tissues [8, 9] [10] strongly suggests that EBV is the causative agent in EBVaGC.

Genetic variations of EBV isolated from different geographic populations are well documented [10–17]. The full-length sequence of the EBV genome has been determined from several cell lines and tissues [18, 19]. The GD1, GD2, and HKNPC isolates were from patients with NPCs from Southern China [20–22]. A sequence from B95.8 cell line is the first and prototype EBV (V01555.2) [20, 23–25]. However, comprehensive genome-wide analyses of EBV in GC have yet to be reported. Therefore, we used next-generation sequencing (NGS) systems [26, 27] with the Illumina genome analyzer to determine the entire sequence of the EBV genome. The virus was isolated from a GC cell line (SNU-719) naturally infected with EBV. Additionally, genome-wide RNAseq analyses revealed restricted viral gene expression in vivo in EBVaGC tissues and cells.

Materials and methods

Whole genome and RNA sequencing

SNU-719 is a GC cell line naturally infected with EBV. Cells were routinely maintained in RPMI media supplemented with 10 % fetal calf serum [28]. Genomic DNA was isolated using a G-spin Genomic DNA extraction kit (iNtRON Biotechnology, Seongnam, Korea). For whole genome sequencing (WGS) of SNU-719 cells, genomic DNA (1 μg) was fragmented using a sonicator (20 % duty, intensity set at 5, 200 cycles per burst for 5 s) (Covaris). A DNA library was generated using a TruSeq DNA sample prep kit v2 according to the manufacturer's protocol (Illumina). The concentration of the library was quantified using a Bioanalyzer (Agilent Technologies), with 6–8 pmol per lane of DNA applied to the flow cell. Paired-end sequencing was performed using the HiSeq 2000 (Illumina) platform, yielding two 100-bp paired-end reads [21]. For RNA sequencing, total RNA from SNU-719 cells and a primary GC tissue (an EBVaGC 086T) were used to generate mRNAseq libraries using the TruSeq RNA Sample Preparation kit (Illumina). The mRNAseq libraries were then sequenced on the HiSeq 2000 platform according to the manufacturer's recommendations, with two 101-bp paired-end reads generated.

EBV sequence assembly, annotation, and transcriptome analyses (see "Results" section for detail)

Phylogeny and comparative analysis

A modified EBV prototype B95.8 genome harboring the RAJI genome sequence inserted into a deleted region of the B95.8 genome were used as the prototype I reference genome in this study (B95.8/RAJI, GenBank NC_007605). Six EBV genomes [the prototype I reference, type II EBV AG876 (DQ279927.1), GD1 (AY961628), GD2 (HQ020558), HKNPC1 (JQ009376.2), and AKATA (KC207813.1) were used for phylogenetic and comparative analyses with a Korean SNU-719 EBV from the SNU-719 cell line in this study (GenBank accession KP735248, named as GC1 genome). Where necessary, the genome in this study was cross-compared with 171,928 bases containing 35693 gaps denoted as N of YCCEL1 EBV genome (GenBank LN827561). The GC1 areas that correspond to N in YCCEL1 EBV was not compared in the pair-wise comparison. We used the maximum likelihood method within the Molecular Evolutionary Genetics Analysis (MEGA) software (v6.0) [29]. The divergence scale, in numbers of substitutions per site, is shown at the foot of each tree. The single nucleotide variations (SNVs), insertions (In), and deletions (Del) in GC1, when compared with B95.8/RAJI, were determined using the cross-match program [30].

Results

HiSeq WGS and assembly of the EBV genome

NGS of SNU-719 genomes generated 1,676,643,550 redundant raw reads, corresponding to 174 GB (the average base per read length was 104 bases). The average coverage was 29-fold (174/6 human diploid genome). These raw reads were preprocessed through a de-duplication program using an in-house de-duplication algorithm. By using BWA [31], the resulting 952,751,016 reads (99.1 GB, coverage 17-fold, 99.1/6) were matched against a human (hg19) and the circular reference of B95.8/RAJI. Use of the multiple match method allowed all redundant reads to match at multiple times (Fig. 1a). The resulting EBV-mapped reads totaled 19.9 MB (0.02 %) and were sorted, aligned, and indexed by Samtool. The coverage depth for each EBV sequence was determined by GATK. SNV/Indel in the draft sequence was detected by Cross match, from which the draft primary sequence with unfilled gaps "N" and wildcards (*) was extracted (Fig. 1a). The gaps were further filled by subsequent polymerase chain reaction (PCR) amplifications and Sanger sequencing. Ambiguous sites were further refined by comparing sequences from the RNAseq data where necessary (Fig. 1b). The highly repetitive regions with nearly identical sequence were left unfilled, with the exception of IR3 where unassembled GA repeats were filled by copying and pasting sequences of the reference. The SNU-719 EBV genome was deposited (SNU-719 EBV in GenBank submission number KP735248). Genes of the SNU-719 EBV were annotated based on the reference genome. The average coverage of EBV was 115-fold (19.9 MB/172 KB of B95.8/RAJI), significantly higher than that for EBV GD2 (17-fold; Table 1) [21].

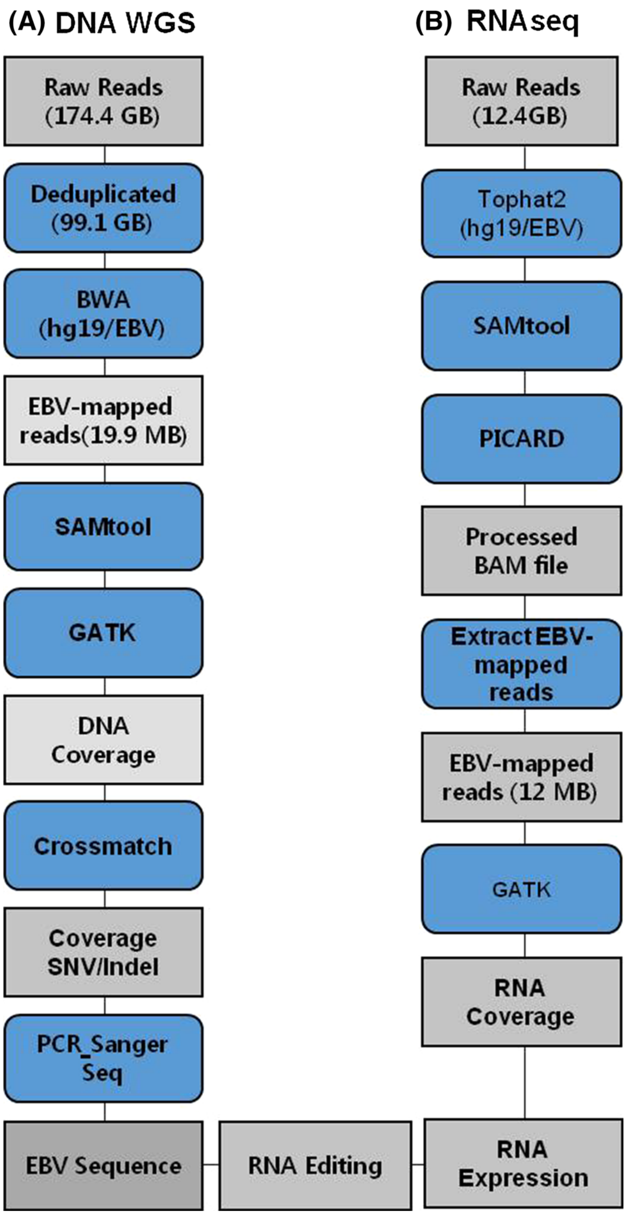

Fig. 1 Workflows for whole genome sequencing (WGS) and RNAseq transcriptome analyses of SNU-719 cells. Data analysis pipelines used for EBV assembly (**a**) and RNAseq (**b**)

SNU-719 EBV analyses

The entire genomic sequence of SNU-719 was aligned to multiple sequences of other known EBV isolates. The phylogenetic tree was constructed by the maxim likelihood and bootstrap analysis using molecular evolutionary genetics analysis (MEGA) software version 6.0 [32]. We observed a consistent clustering of SNU-719 EBV with NPC EBV strains (GD1, AKATA, GD2, HKNPC1; Fig. 2). The SNU-719 EBV was found to be a type I EBV and most similar to the GD1 strain. It was also similar to, although to a lesser extent, the AKATA, GD2, and HKNPC1 isolates, which are also type I EBVs. Similar to GD1 and GD2 from China [21], SNU-719 EBV contained the same mutations in EBNA1 (487V, 499E, 502N, 524I, and 528V) and LMP1 (322N, 334R, 338S, and a 10-amino acid (a. a.) deletion at 343–352). Six SNVs in the BZLF1 coding sequence of SNU-719 EBV occurred at different sites from those in GD1, indicating some divergence from GD1 or GD2 (Fig. 3) [21, 33].

Gene annotation, identification of SNV and indels

Comparison with the reference genome revealed that SNU-719 EBV had at least 109 genes, of which 86 encoded a protein and 23 were able to transcribe RNA (see GenBank KP735248 for annotations and sequences). Compared with the reference sequence, there were changes at 1372 sites at the DNA level. This included 1288 SNV, 36 base-insertions at 22 sites, and 1469 deleted bases across 6 sites. Among the SNVs, 77 % (996/1288) exhibited homozygous changes (defined as >90 % coverage at the indicated site with an altered base) and 33 % (292/1291) were heterozygous (Table S1). Examination of amino acid changes showed that 68 % (937/1372) were located in coding sequence regions, with 465 non-synonymous and 472 synonymous SNV changes identified (Table S1).

We found that SNVs occurred in 82 protein-coding regions, including BCRF1, LMP1, EBNA3A, 3B, 3C, EBNA2, LF3, RTA, ZTA, BDLF3 (gp85), BPLF1

Table 1 Summary of reads obtained from NGS step in this study

Sequencing	Steps of processing	No. of read[1]	Size[2] (base)	[3] % of read	[4]Coverage	Assembled (base)
DNA WGS	Total raw reads	1,676,643,550	174.4 G	na	29.1^{h}	na
	Post-deduplicated reads	952,751,016	99.1 G	100	16.5^{h}	na
	Human reads	932,892,424	97.0 G	97.9	16.2^{h}	na
	EBV reads[5]	190,948	19.9 M	0.02	115.6^{e}	^{D}170,390
RNAseq	Total raw reads	119,539,924	12.4 G	na	2.1	na
	Post-deduplicated reads	88,620,752	9.2 G	100	1.5	na
	Human reads	88,522,955	9.2 G	99.9	1.5	na
	EBV reads	119,150	12 M	0.13	72.1^{e}	^{R}128,062

[1] average base length per read is 104 base

[2] No. of read × 104 base

[3] % relative to reads without PCR duplicates

[4] Coverage to 6 GB of human dipliod or 171,823 base of reference 95.8/RAJI EBV(NC_007605); ^{D}base composition of SNU-719 EBV DNA: A 33739, C 507141, G 50483, T 35428, ^{D}1468 deletion, and 41 insertion compared to reference genome;^{R}total sum of transcipts

[5] Mapped to cicrular EBV in multiple match method

(tegument), BOLF1(capsid assembly protein), and BLLF1 (gp350). Notable changes included apparent deletions in the EBNA2 proline-rich domain ($_{67}$PPPPPPPPPPPPPPPPPPPPPPSPPPPP$_{94}$ to $_{67}$TTTPPT$_{72}$); in the LMP1 (10 a. a. of $_{343}$GGHSHDSGHG$_{352}$ at C-terminus; a 43 base-deletion in the promoter); in the glycoprotein 350 (9 amino acids deleted $_{676}$LSPSTSDNS$_{684}$); and a 68 amino acid substitution at the C-terminus of LF3 (Table S1). Frequent internal deletions in EBNA1 Gly-Ala repeats are thought to be due to sequencing difficulty or errors in the IR3 repeat-rich region. Deletions in the LMP1 promoter should result in a lack of expression.

Transcriptome analyses

RNA sequencings for total RNAs from SNU-719 cells and a primary EBVaGC tissue (086T) were conducted to understand viral transcription. Total raw reads (12.4 GB) were aligned to the human (hg19) and reference EBV sequences using TOPHAT2 v2.0.12. Aligned BAM files were indexed, sorted, and deduplicated by SAMTOOLS v0.1.19 and PICARD v1.86. The EBV-mapped reads were selected from processed BAM files and abundance of EBV transcripts was calculated by depth of coverage function in Genome Analysis Toolkit (GATK) v2.7.2 (Fig. 1B). The average coverage of EBV was 72.1-fold (Table 1).Genome-wide RNAseq transcriptome analyses revealed viral transcription from at least 28 % of viral genomes when a summed coverage cutoff greater than 10 was assumed to be positive expression at the detectable level (Fig. S1). The genes expressed at the highest levels were BARTs (A73, RPMS1, BARF0, BALF3, BALF5, BALF4, LF2), with an average coverage depth greater than 300 (Fig. 3). Genes that were expressed at moderate or high levels with average coverage depth ≥10 included BNLF2a/b, LF1, LF3, BHLF1 BILF1, BdRF1, miR-BART-2, miR-BART15, BORF2, gp350, gp L, gp85, capsid proteins of VP23, VP19C, and VP26 (Fig. 3; Table S2). Most EBV-encoded miRNAs were not expressed at the detectable level except for miR-BHRF1-2 (average coverage depth ~8), miR-BHRF1-3 (~9), miR-BART-2 (~35), and miR-BART15 (~14). Expression of EBNA1 and LMP2A, which are best known as constitutively expressed viral transcripts in EBV-infected cells, were, however, relatively weak with coverage of 5.9 and 4.4, respectively. Two viral trans-activators for lytic activation, BZLF1 (ZTA) and BRRLF1, were also weakly expressed. Any genes with coverage less than 4.4, corresponding to LMP2A, were assumed to not be expressed. These genes included LMP-1, the EBNA3 family, EBNA-2, and EBNA-LP (Table S2). LF3 was moderately expressed in the cell line, with a mean coverage depth of 55.1, significantly higher than that for EBNA1 and LMP2A (Table S2). In addition, a short RNA transcript (nt 145,850–145,951) that corresponds to nucleotides (nt) 146,233–146,334 of the reference sequence was expressed at high level in both SNU-719 cell and EBVaGC tissue sample (086T) (Fig. S1). This region has been thought to be located within RPMS1 intron 1 (nt 138,481–149,580) with no expression. However, the high abundance of RNAs transcribed from this region in this study may indicate the presence of GC-specific RPMS1 isoform that harbors a region with an additional or alternative exon in GC1. In keeping with this assumption, this short transcript with high coverage was flanked by conserved consensus splice acceptor (AG) and donor sequence (GT) at before (at nt 145,848–145,849) and after (at nt 145,952–145,953) the indicated exon, respectively. RNAseq coverage at each transcription site was compared with that from whole genome sequencing to identify possible changes

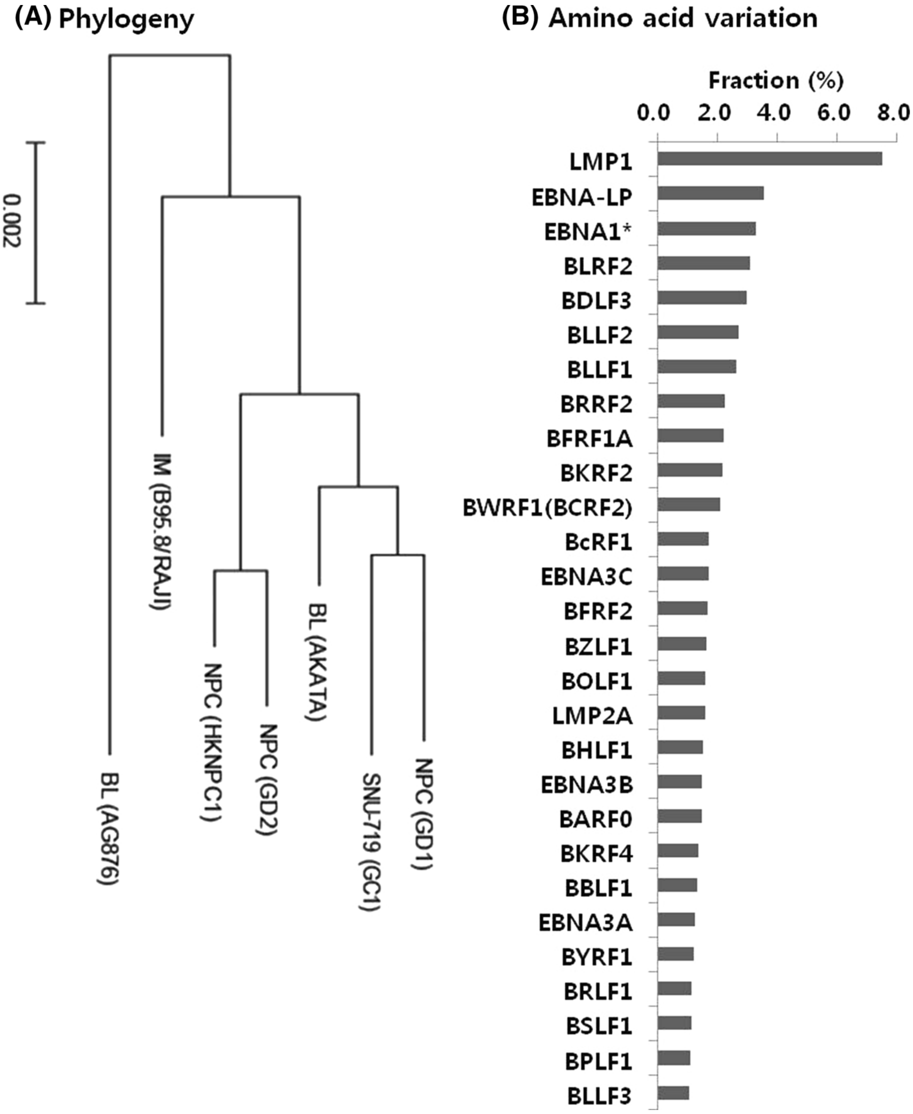

Fig. 2 Phylogenetic and comparative analyses of the SNU-719 EBV from this study with known viral genomes. **a** Phylogenetic divergence among type I Chinese NPC EBV (GD1, GD2, and HKNPC1), type I Japanese BL EBV (AKATA), type I African IM EBV (B95.8/RAJI), type II African BL (AG876), and Korean GC EBV (SNU-719 in this study) were compared at the whole genome level. The divergence scale (node height showing number of substitutions per site) is indicated at the foot of each tree. **b** Percentage (%) of non-synonymous variations in the coding sequence compared to the reference

in transcript at the RNA level. We found strong evidence of RNA editing around at least 14 sites; one example of RNA editing was that which occurred in the putative novel RPMS1 exon as described above (Table 2).

The assembled sequence and annotation were submitted to GenBank (KP 735248). The raw data of DNA seq and RNAseq were deposited as SRX959119 and SRX960421 (also GSE60873 in GEO), respectively, in SRA of NCBI.

Discussion

Both strands in the DNA and RNA transcript were sequence and assembled. Even if the entire DNA was sequenced, mutations in the cellular genome were not cataloged as this study specifically intended to deduce primary viral sequences. Considering 17- and 115-fold coverage for host and viral genomes, respectively, in this study, this would suggest there are an average of seven copies of EBV (115/17 = 7) in a single tumor cell. This is consistent with previous results; an undifferentiated NPC tumor was found to harbor multiple EBV genomes [21]. The average coverage across viral genome positions was 137 (median 119). The vast majority of sites (97.6 %) were covered more than 10 times; around 1.7 % of total bases were covered less than five times. This is an indicative of a possible deletion in a subset of multiple genomes, of which 0.8 % were covered once or less, reflecting some of them may have a deletion in these low coverage areas. (Table S3). Reads that span this putative deleted region were found, making this claim. On the other hand, it is also possible that these are just regions of the genome that are hard to sequence through. This contrast in extremely low or high coverage depth depending on region might have

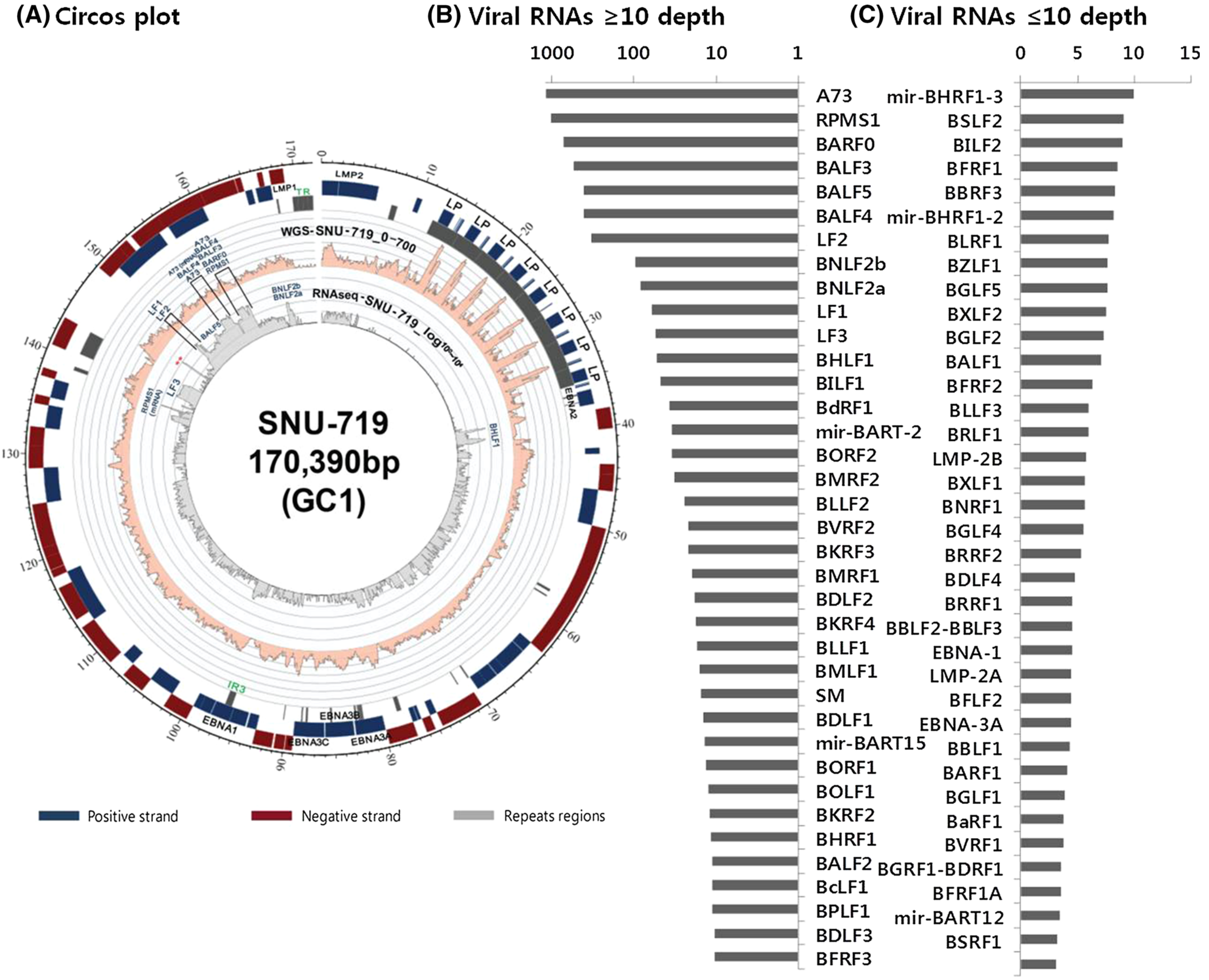

Fig. 3 Genome map and transcriptome assembly results for SNU-719 EBV and highly expressed viral transcripts. **a** Coverage depth of NGS (*pink*, linear scale) and RNAseq (*gray*, log scale). *Vertical bars* indicate the variations in GC1, GD1, and AKATA from the reference EBV isolate. High variation densities are *shaded*. Representative annotated areas, transcription direction, repeat or regulatory regions are also shown. **b**, **c** Viral RNAs ≥10 and ≤10 mean depth coverage (see Table 2 for details) (Color figure online)

arisen from the sequencing of multiple different genome species with frequent deletions at different sites. This is indicative of multiple species of viral genomes per cell population and/or possible heterogeneity in genome populations. Of the 1372 alterations, many were located within repeat regions which are prone to sequencing errors. Alterations in the non-repeat regions are suggestive of intergenome differences. The apparent heterogeneity at these positions could be ascribed to both sequencing/assembly errors, and true heterogeneity comprising multiple genomes with variations in length (often small deletions) and differing compositions at specific sites. Taken together, these data suggest the possibility that low-level genomic evolution occurs during long-term cell culture.

Certain viral transcripts in SNU-719 cells were highly abundant (A73, RPMS1, BARF0, BALF3, BALF5, BALF4, LF2, LF1, LF3). Other early and late genes (BNLF2b, BNLF2a, BHLF1, BILF1, BdRF1, mir-BART-2, miR-BART15, BORF2, BMRF2, BLLF2, BVRF2, BKRF3, BMRF1, BDLF2, BKRF4, BLLF1) were also present in SNU-719 cells but at moderate levels. Additionally, low-level transcription of BZLF1, BRLF1, and EBNA1 was seen in other EBVaGC tissues from this study and in a previous report [34]. Despite the detection of two lytic trans-activators BZLF1 and BRLF1, the lack of expression for most other downstream lytic genes likely reflects incomplete or abortive lytic replication in vivo. Moreover, despite high abundance of BART transcripts, most miRNAs arrayed in its introns were not detected. This could be either due to no expression or deselection. Small size (<90 base) of non-polyadenylated BART miRNAs under the size limit of RNAseq selection (~100base)

Table 2 RNA editing site with high probability in the SNU-719 EBV (GC1) transcriptome

GC1[a]	REF[b]	WGS coverage[c]					Assembly[d]	RNAseq coverage[e]					Transcribe[f]	Max RNA[g]	Remarks[h]
Position	Position	A	C	G	T	ΣWGS	C1 DNA	A	C	G	T	RNase	C1 RNA	Editing ratio	
38,263	38,337	11	23	4	0	38	C	57	4	0	0	61	A	0.93	BHLF1,DAUDI deletion, P3HR1 deletion
137,970	138,359	166	2	0	1	169	A	33	0	0	26	59	W	0.56	RPMS1 mRNA
138,092	138,481	26	1	118	1	146	G	19	0	12	0	31	R	0.62	
138,094	138,483	129	0	1	0	130	A	2	9	15	0	26	S	0.58	
138,095	138,484	145	0	1	2	148	A	2	3	12	0	5	M	0.6	
142,846	143,232	7	1	35	0	43	G	30	0	0	0	57	R	0.53	LF3,B95.8 deletion, OriLyt
145,845	146,228	46	0	2	194	242	T	0	0	27	3	40	G	0.93	Putative novel RPMS1 exon, B95.8 deletion
145,847	146,230	0	187	45	0	232	C	0	101	37	0	49	G	0.94	Putative novel RPMS1 exon, B95.9 deletion
145,954	146,337	0	0	162	2	164	G	0	68	46	0	102	V	0.99	Putative novel RPMS1 exon, B95.10 deletion
145,955	146,338	0	0	163	1	164	G	0	0	126	0	94	S	0.72	Putative novel RPMS1 exon, B95.11 deletion
149,194	149,577	19	0	1	78	98	T	32	0	0	14	46	W	0.69	LF2,B95.8 deletion
160,150	160,533	2	1	0	142	145	T	21	0	0	5	26	A	0.81	BALF3_BARF0
160,151	160,534	29	1	0	113	143	T	19	0	0	5	24	A	0.79	BALF3_BARF0
160,156	160,539	0	2	0	126	128	T	3	0	0	1	4	A	0.75	BALF3

[a] Base site in GC1 EBV from SNU-719 in this study

[b] Corresponding position in the REF B95.8/RAJI(NC_007605)

[c,e] Raw coverage depth assigned to indicated base by WGS and RNAseq

[d,f] Sequence of DNA and RNA assembled from WGS and RNAseq

[g] Editing ratio to the base with of maximum coverage

[h] Annoated based on REF

would not have been extracted during cDNA processing for RNAseq. The same may apply to the EBER transcripts, which usually are by far the most abundant RNA species in EBV-infected cell lines. Yet these EBERs were not absent from viral transcriptomes determined by RNAseq (Table S2).

The clustering of the SNU-719 EBV in this study with Asian type I EBVs such as GD1, GD2, AKATA, and HKNPC1, and separation from African EBVs is consistent with previous results. This solidifies the notion that the relationship between EBVs has better correlation with geography rather than tissue origin as previously reported [32, 35, 36]. In previous studies, the majority of EBVaGCs [37, 38] and NPCs [16, 39] have a 30 base-deletion of the LMP1 coding exon 3. The SNU-719 EBV in this study exhibits the same deletion. A unique deletion in the LMP1 promoter region of SNU-719 EBV is not seen in the GD1 or GD2 isolates, and accounts for the consistent lack of LMP1 expression in EBVaGC [28, 40, 41]. Among 1363 variations found in the SNU-719 EBV genome, SNVs in the CDS region were clustered in certain regions (BPLF1, BWRF1, BOLF1, BLLF1, EBNA3C, BcLF1, BKRF1, BcRF1, LMP1, EBNA3A,BORF2, BRRF2, EBNA3B and BPLF1; Tables S1, S2); however, further investigation is necessary to uncover whether these alterations are linked to the development of GC.

The RNA editing process can modify RNA post-transcriptionally catalyzed by member of the adenosine deaminase acting on RNA (ADAR) family [42]. The apolipoprotein B mRNA editing enzyme and catalytic polypeptide-like (APOBEC) family possess cytosine deaminase activity on both DNA and RNA resulting in C to

G/T mutation. APOBEC3 (A3) hypermutates viral genomes and acts as a viral restriction factor for a number of viruses [43, 44]. APOBEC-mediated cytosine deamination is responsible for mutation of PIK3CA helical domain in across multiple cancers including human papillomavirus-driven tumor [45]. Recently, APOBEC3G has been known to be over expressed in EBVaGC by our group [46], and this is likely responsible for the RNA editing and/or DNA mutation in the SNU-719 EBV of this study. Genome-wide RNAseq for SNU-719 EBV also verified the overexpression of the APOBEC3 family (data not shown). PI3 K signaling network plays roles in receptor-mediated endocytosis and clathrin-independent endocytosis as the alphaVbeta5 integrin-mediated endocytosis of adeno-associated virus-2 (AAV-2) occurs via a Rac1 and PI3K activation cascade [47]. Given that PIK3CA activation mutations are very frequent in EBVaGC but rare in EBVnGC [48] and EBV infection into epithelial cells occurs via EBV gH binding to integrin alphaVbeta5/8 on target cells [49], the overexpression of APOBEC family proteins in EBVaGC likely induces alteration on DNA (such as PIK3CA), which may ultimately accelerate an EBV-induced epithelial transformation. Alternatively, virus may utilize APOBEC-mediated RNA editing in RNA transcripts or DNA editing in DNA genome as a mean used to evade host restriction activity.

Meanwhile, another EBV genome (LN827561) from YCCEL1, another GC cell line with natural EBV infection, was uploaded in NCBI and the literature [50]. Considering that both SNU-719 and YCCEL1 are EBV-infected gastric cancer cell lines established from Korean patients, the SNU-719 EBV (GC1) may also have the same deletion or insertion—if any—that YCCEL1 EBV genome might have. In comparison to reference genome, GC1 has 1288 SNVs, 36 base-insertion at 22 sites and 1469 base-deletion at six sites; 60 % of same SNVs (764/1288, 60 %) and 75 % of same insertion (27/36 bases) occurred also in YCCEL1 EBV. The signature deletion of 30 bases in LMP1 coding region, unique feature in EBV of GC and NPC from East Asian area, was also found in GC1 and YCCEL1 EBV genome. Extensive number of unfilled gaps in YCCEL1 hindered pair-wise comparison. While YCCEL1 had heterogeneous 66 base-insertions in LMP1 coding area, GC1 lacks the same insertion as evidenced by previous and current Sanger sequencings. Instead, GC1 had 43 base-deletions in LMP1 promoter area, leading to the absence of LMP1 expression. In overall, cross comparison of SNU-719 EBV with the YCCEL1 EBV showed the overall sequence homology by 98.59 %.

In conclusion, we have described the entire genomic sequence of an EBV isolate that naturally infects a GC cell line. We believe this EBV isolate, GC1, will be useful for future studies regarding EBVaGC carcinogenesis as it can act as a reference sequence.

Acknowledgments This study was supported by a grant from the National Research Foundation of Korea, NRF-2015R1A2A2A01006459, a grant of the Korea Healthcare Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), and funded by the Ministry for Health & Welfare, Republic of Korea. (HI09C1552).

Author contributions KAS initiated and performed parts of research, collected and analyzed all data. SDY collected, analyzed, and deposited data. JH collected, analyzed, and deposited RNAseq data. JIK and MSK conceived, designed, supervised study, made figure, and wrote the manuscript.

Compliance with ethical statements

Conflict of interest The authors declare no competing financial interests.

References

1. M.A. Epstein, B.G. Achong, Y.M. Barr, Lancet **1**, 702–703 (1964)
2. M.A. Epstein, Y.M. Barr, Lancet **1**, 252–253 (1964)
3. H. zur Hausen, H. Schulte-Holthausen, G. Klein, W. Henle, G. Henle, P. Clifford, L. Santesson, Nature **228**, 1056–1058 (1970)
4. L.M. Weiss, J.G. Strickler, R.A. Warnke, D.T. Purtilo, J. Sklar, Am. J. Pathol. **129**, 86–91 (1987)
5. J.F. Jones, S. Shurin, C. Abramowsky, R.R. Tubbs, C.G. Sciotto, R. Wahl, J. Sands, D. Gottman, B.Z. Katz, J. Sklar, N. Engl. J. Med. **318**, 733–741 (1988)
6. G. Murphy, R. Pfeiffer, M.C. Camargo, C.S. Rabkin, Gastroenterology **137**, 824–833 (2009)
7. D.E. Burgess, C.B. Woodman, K.J. Flavell, D.C. Rowlands, J. Crocker, K. Scott, J.P. Biddulph, L.S. Young, P.G. Murray, Br. J. Cancer **86**, 702–704 (2002)
8. M. Fukayama, Y. Hayashi, Y. Iwasaki, J. Chong, T. Ooba, T. Takizawa, M. Koike, S. Mizutani, M. Miyaki, K. Hirai, Lab. Investig. J. Tech. Methods Pathol. **71**, 73–81 (1994)
9. K. Takada, Mol. Pathol. **53**, 255–261 (2000)
10. P. Busson, C. Keryer, T. Ooka, M. Corbex, Trends Microbiol. **12**, 356–360 (2004)
11. R.H. Edwards, F. Seillier-Moiseiwitsch, N. Raab-Traub, Virology **261**, 79–95 (1999)
12. M.I. Gutierrez, G. Spangler, D. Kingma, M. Raffeld, I. Guerrero, O. Misad, E.S. Jaffe, I.T. Magrath, K. Bhatia, Blood **92**, 600–606 (1998)
13. Y.Z. Jing, Y. Wang, Y.P. Jia, B. Luo, Chin. J. Cancer **29**, 1000–1005 (2010)
14. N.S. Sung, R.H. Edwards, F. Seillier-Moiseiwitsch, A.G. Perkins, Y. Zeng, N. Raab-Traub, Int. J. Cancer **76**, 207–215 (1998)
15. L.L. Zhang, D.J. Li, Z.H. Li, X.S. Zhang, R.H. Zhang, X.J. Yu, L.Z. Chen, Q.S. Feng, Y.X. Zeng, W.H. Jia, Aizheng, Chin. J. Cancer **26**, 1047–1051 (2007)
16. X.S. Zhang, K.H. Song, H.Q. Mai, W.H. Jia, B.J. Feng, J.C. Xia, R.H. Zhang, L.X. Huang, X.J. Yu, Q.S. Feng, P. Huang, J.J. Chen, Y.X. Zeng, Cancer Lett. **176**, 65–73 (2002)
17. X.S. Zhang, H.H. Wang, L.F. Hu, A. Li, R.H. Zhang, H.Q. Mai, J.C. Xia, L.Z. Chen, Y.X. Zeng, Cancer Lett. **211**, 11–18 (2004)
18. G. Miller, T. Shope, H. Lisco, D. Stitt, M. Lipman, Proc. Natl. Acad. Sci. U.S.A. **69**, 383–387 (1972)

19. J. Skare, C. Edson, J. Farley, J.L. Strominger, J. Virol. **44**, 1088–1091 (1982)
20. M.S. Zeng, D.J. Li, Q.L. Liu, L.B. Song, M.Z. Li, R.H. Zhang, X.J. Yu, H.M. Wang, I. Ernberg, Y.X. Zeng, J. Virol. **79**, 15323–15330 (2005)
21. P. Liu, X. Fang, Z. Feng, Y.M. Guo, R.J. Peng, T. Liu, Z. Huang, Y. Feng, X. Sun, Z. Xiong, X. Guo, S.S. Pang, B. Wang, X. Lv, F.T. Feng, D.J. Li, L.Z. Chen, Q.S. Feng, W.L. Huang, M.S. Zeng, J.X. Bei, Y. Zhang, Y.X. Zeng, J. Virol. **85**, 11291–11299 (2011)
22. H. Kwok, A.H. Tong, C.H. Lin, S. Lok, P.J. Farrell, D.L. Kwong, A.K. Chiang, PLoS One **7**, e36939 (2012)
23. N. Raab-Traub, T. Dambaugh, E. Kieff, Cell **22**, 257–267 (1980)
24. R. Baer, A.T. Bankier, M.D. Biggin, P.L. Deininger, P.J. Farrell, T.J. Gibson, G. Hatfull, G.S. Hudson, S.C. Satchwell, C. Seguin et al., Nature **310**, 207–211 (1984)
25. A. Dolan, C. Addison, D. Gatherer, A.J. Davison, D.J. McGeoch, Virology **350**, 164–170 (2006)
26. W.J. Ansorge, New Biotechnol. **25**, 195–203 (2009)
27. J. Shendure, H. Ji, Nat. Biotechnol. **26**, 1135–1145 (2008)
28. S.T. Oh, J.S. Seo, U.Y. Moon, K.H. Kang, D.J. Shin, S.K. Yoon, W.H. Kim, J.G. Park, S.K. Lee, Virology **320**, 330–336 (2004)
29. K. Tamura, J. Dudley, M. Nei, S. Kumar, Mol. Biol. Evol. **24**, 1596–1599 (2007)
30. D. Gordon, C. Abajian, P. Green, Genome Res. **8**, 195–202 (1998)
31. L.S. Chou, C.S. Liu, B. Boese, X. Zhang, R. Mao, Clin. Chem. **56**, 62–72 (2010)
32. K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, Mol. Biol. Evol. **30**, 2725–2729 (2013)
33. Y. Wang, X. Liu, X. Xing, Y. Cui, C. Zhao, B. Luo, Virus Res. **147**, 258–264 (2010)
34. M.J. Strong, G. Xu, J. Coco, C. Baribault, D.S. Vinay, M.R. Lacey, A.L. Strong, T.A. Lehman, M.B. Seddon, Z. Lin, M. Concha, M. Baddoo, M. Ferris, K.F. Swan, D.E. Sullivan, M.E. Burow, C.M. Taylor, E.K. Flemington, PLoS Pathog. **9**, e1003341 (2013)
35. Z. Lin, X. Wang, M.J. Strong, M. Concha, M. Baddoo, G. Xu, C. Baribault, C. Fewell, W. Hulme, D. Hedges, C.M. Taylor, E.K. Flemington, J. Virol. **87**, 1172–1182 (2013)
36. G. Santpere, F. Darre, S. Blanco, A. Alcami, P. Villoslada, Mar Alba M., and Navarro A. Genome Biol. Evol. **6**, 846–860 (2014)
37. K. Hayashi, W.G. Chen, Y.Y. Chen, I. Murakami, H.L. Chen, N. Ohara, S. Nose, K. Hamaya, S. Matsui, M.M. Bacchi, C.E. Bacchi, K.L. Chang, L.M. Weiss, Am. J. Pathol. **152**, 191–198 (1998)
38. H.S. Lee, M.S. Chang, H.K. Yang, B.L. Lee, W.H. Kim, Clin. Cancer Res. **10**, 1698–1705 (2004)
39. S.Y. Leung, S.T. Yuen, L.P. Chung, A.S. Chan, M.P. Wong, Int. J. Cancer **72**, 225–230 (1997)
40. B. Luo, Y. Wang, X.F. Wang, H. Liang, L.P. Yan, B.H. Huang, P. Zhao, World J. Gastroenterol. **11**, 629–633 (2005)
41. A. zur Hausen, A.A. Brink, M.E. Craanen, J.M. Middeldorp, C.J. Meijer, A.J. van den Brule, Cancer Res. **60**, 2745–2748 (2000)
42. D. Dominissini, S. Moshitch-Moshkovitz, N. Amariglio, G. Rechavi, Carcinogenesis **32**, 1569–1577 (2011)
43. M. Bonvin, F. Achermann, I. Greeve, D. Stroka, A. Keogh, D. Inderbitzin, D. Candinas, P. Sommer, S. Wain-Hobson, J.P. Vartanian, J. Greeve, Hepatology (Baltimore, Md) **43**, 1364–1374 (2006)
44. K.N. Bishop, R.K. Holmes, A.M. Sheehy, M.H. Malim, Science (New York, NY) **305**, 645 (2004)
45. S. Henderson, A. Chakravarthy, X. Su, C. Boshoff, T.R. Fenton, Cell Rep. **7**, 1833–1841 (2014)
46. S.Y. Kim, C. Park, H.J. Kim, J. Park, J. Hwang, J.I. Kim, M.G. Choi, S. Kim, K.M. Kim, M.S. Kang, Gastroenterology **148**, 137–147 (2015)
47. S. Sanlioglu, P.K. Benson, J. Yang, E.M. Atkinson, T. Reynolds, J.F. Engelhardt, J. Virol. **74**, 9184–9196 (2000)
48. Cancer Genome Atlas Research Network, Nature **513**, 202–209 (2014)
49. L.S. Chesnokova, S.L. Nishimura, L.M. Hutt-Fletcher, Proc. Natl. Acad. Sci. U.S.A. **106**, 20464–20469 (2009)
50. A.L. Palser, N.E. Grayson, R.E. White, C. Corton, S. Correia, M.M. Ba Abdullah, S.J. Watson, M. Cotten, J.R. Arrand, P.G. Murray, M.J. Allday, A.B. Rickinson, L.S. Young, P.J. Farrell, P. Kellam, J. Virol. **89**, 5222–5237 (2015)