viruses

MDPI

Review

From Conventional to Next Generation Sequencing of Epstein-Barr Virus Genomes

Hin Kwok and Alan Kwok Shing Chiang *

Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; hinkwok@hku.hk

* Correspondence: chiangak@hku.hk; Tel.: +852-2255-4091

Academic Editors: Johnson Mak, Peter Walker and Marcus Thomas Gilbert

Received: 23 December 2015; Accepted: 19 February 2016; Published: 24 February 2016

Abstract: Genomic sequences of Epstein–Barr virus (EBV) have been of interest because the virus is associated with cancers, such as nasopharyngeal carcinoma, and conditions such as infectious mononucleosis. The progress of whole-genome EBV sequencing has been limited by the inefficiency and cost of the first-generation sequencing technology. With the advancement of next-generation sequencing (NGS) and target enrichment strategies, increasing number of EBV genomes has been published. These genomes were sequenced using different approaches, either with or without EBV DNA enrichment. This review provides an overview of the EBV genomes published to date, and a description of the sequencing technology and bioinformatic analyses employed in generating these sequences. We further explored ways through which the quality of sequencing data can be improved, such as using DNA oligos for capture hybridization, and longer insert size and read length in the sequencing runs. These advances will enable large-scale genomic sequencing of EBV which will facilitate a better understanding of the genetic variations of EBV in different geographic regions and discovery of potentially pathogenic variants in specific diseases.

Keywords: Epstein-Barr virus; Next-generation sequencing; target capture; genome assembly

1. Introduction

Epstein–Barr virus is a gamma-herpesvirus which infects more than 90% of the world's population. It is associated with cancers such as Hodgkin's lymphoma, Burkitt's lymphoma, gastric cancer, nasopharyngeal carcinoma, and other conditions, such as post-transplant lymphoproliferative diseases and infectious mononucleosis. The geographical distribution of EBV strains and the endemic incidence of EBV-associated diseases have prompted investigations of whether there are distinct strains of EBV that contribute to the disease process. EBV strains have previously been characterized in NPC and other disease at various loci, including EBER-1 and -2, LMP1, BHRF1, BZLF1, and EBNA1 in samples from China, South Asia, and Northern Africa [1–6]. However, these early studies have only been focused on the individual viral genes. Very few studies have been carried out to investigate how genetic variations of EBV on a whole-genome scale might influence infection or pathogenesis of EBV-associated diseases. The large genome size of EBV relative to other viruses renders large-scale sequencing of EBV genomes cost- and time-prohibitive. Advances in next-generation sequencing technology have promoted the determination of EBV genomic sequences and reignited the interest of studying the viral genome and its association to diseases. This review provides an overview of how the sequences of published EBV genomes have been determined, and the bioinformatics analyses involved.

2. Conventional Shotgun Sequencing of EBV Genomes

Conventional shotgun sequencing involved the techniques of digestion of genomic DNA with restriction enzymes, cloning, and dideoxynucleotide (Sanger) sequencing. It has been widely applied to genomic sequencing of organisms, such as humans and Epstein-Barr viruses.

The prototypical strain B95-8 was the first complete EBV genome sequenced. An 833L cell line was obtained by culturing lymphocytes from an individual with infectious mononucleosis (IM), then the EBV released from the 883L line was used to infect marmoset B cells, resulting in the B95-8 line [7]. EcoRI- and BamHI-digested fragments have first been cloned and restriction maps of these clones were obtained [8–10]. The DNA sequence was analyzed by constructing M13 subclone libraries, followed by random sequencing using the dideoxynucleotide method [11]. Due to its early availability, B95.8 has been extensively mapped for transcripts, promoters, open reading frames, and other structural elements, by means of Northern blotting and other methods [12,13]. By filling in the 11 kb deletion in B95-8 with Raji EBV sequence, this chimeric EBV genome serves as the type 1 EBV reference sequence [14] and is widely used in genomic analysis in later studies.

GD1 (Guangdong strain 1) was an EBV genome derived from NPC patients from the Guangdong province of Southern China. It was isolated by infecting umbilical cord mononuclear cells by EBV from the saliva of the patient [15]. The EBV DNA was PCR amplified and sub-cloned, then sequenced by conventional Sanger sequencing. More than 1400 point mutations were identified by comparing against the type 1 reference, some of which were also found in high frequency in NPC biopsies of the same geographical region, hence, suggesting that GD1 is representative of the EBV strains found in Southern Chinese NPC patients.

AG876 was originated from a Ghanaian case of Burkitt's lymphoma and was the first complete type 2 EBV genome published [16]. Sequence analysis was performed by first constructing cosmid libraries through Sau3AI digestion, followed by dideoxynucleotide sequencing. The determination of AG876 sequence has made comparison of whole-genome comparison of type 1 and 2 EBV possible. It had validated that the two major types of EBV are generally very similar outside the divergent regions at EBNA2 and EBNA3 genes.

B95-8, GD1, and AG876 are the products of immense effort and they represent EBVs of different geographical origins (B95-8 as an American strain, GD1 as a Southern Chinese strain, and AG876 as an African strain) and different diseases (B95-8 from IM, GD1 from NPC, and AG876 from Burkitt's lymphoma). A summary of sequencing methods of these three and other published EBV genomes are shown in Table 1.

Table 1. Summary of sequencing methods of published EBV genomes.

Method	Genomes	Accession No.	NGS Read Length *	Ethnicity	Tissue/Fluid of Origin *	Cultured Cell Type *	Bioinformatics Strategy	Software Used	Reference
Shotgun sequencing	B95-8	V01555	n/a	N. American	IM	LCL	n/a	n/a	[11]
	AG876	DQ279927	n/a	African	BL	BL	n/a	n/a	[16]
	GD1	AY961628	n/a	Chinese	NPC saliva	LCL	n/a	n/a	[15]
NGS without EBV enrichment	GD2	HQ020558	PE44	Chinese	NPC biopsy	not applicable	Reference mapping	SOAPdenovo	[17]
	C666-1	KC617875	PE100	Chinese	NPC	mouse xenograft	Reference mapping	BWA, GATK, Samtools	[18]
	K4413-Mi	KC440851	PE175	N. American	PBMC	spLCL	*De novo* assembly	CLC Genomic Workbench	[19]
	K4123-Mi	KC440852		N. American		spLCL			
	NA12878	n/a	PE36	N. American		LCL			
	NA20783	n/a	n/a	European	PBMC	LCL	Reference mapping	BWA, Samtools	[20]
	NA18507			African					
	NA20348			African					
	NA18923			African					
	NA20524			African					
	NA19114			European					
	NA19474			African					
	NA19315			African					
	NA19380			African					
	NA19384			African					
	GC1	KP735248	PE100	Korean	GC	GC	Reference mapping	BWA, Samtools	[21]
	CCH	KP968257	PE250	S. American	BL	not applicable	Reference mapping	CLC Genomic Workbench	[22]
	MP	KP968258		S. American					
	SCL	KP968259		S. American					
	VGO	KP968260		S. American					
	RPF	KR063344		S. American					
	FNR	KR063345		S. American					
	CV-ARG	KR063343		S. American					
	HU11393	KP968261		African					
	H03753A	KR063342		African					
	H018436D	KP968262		African					
	H058015C	KP968263		African					
	H002213	KP968264		African					

Table 1. *Cont.*

Method	Genomes	Accession No.	NGS Read Length *	Ethnicity	Tissue/Fluid of Origin *	Cultured Cell Type *	Bioinformatics Strategy	Software Used	Reference
EBV enrichment by lytic induction	Akata Mutu	KC207813 KC207814	PE100	Japanese African	BL	BL	*De novo* assembly	Velvet	[23]
F-factor cloning	M81	KF373730	n/a	Chinese	NPC	LCL	n/a	GS Reference Mapper	[24]
Amplicon sequencing	HKNPC1	JQ009376	PE76	Chinese	NPC biopsy	not applicable	Reference mapping	BWA, Samtools	[25]
	LCL1 LCL3 LCL9 LCL10	n/a	PE150	African	PBMC	spLCL	Reference mapping	Strand NGS	[26]
EBV capture by hybridization	HKNPC2 HKNPC3 HKNPC4 HKNPC5 HKNPC6 HKNPC7 HKNPC8 HKNPC9	KF992564 KF992565 KF992566 KF992567 KF992568 KF992569 KF992570 KF992571	PE76	Chinese	NPC biopsies	not applicable	*De novo* assembly	Velvet	[27]
	71 EBV genomes	See ref.	PE76	Mixed	NPC biopsy, Healthy saliva, HL, BL, PTLD & IM	Mixed	*De novo* assembly	Velvet	[28]
	EBVaGC1 EBVaGC2 EBVaGC3 EBVaGC4 EBVaGC5 EBVaGC6 EBVaGC7 EBVaGC8 EBVaGC9	KT273942 KT273943 KT254013 KT273944 KT273945 KT273946 KT273947 KT273948 KT273949	PE125	Chinese	EBVaGC	not applicable	*De novo* assembly	Velvet	[29]

* PE: Paired-end, followed by the length of reads in base-pair; spLCL: spontaneous lymphoblastoid cell line; LCL: lymphoblastoid cell line infected with external virus source; BL: Burkitt's lymphoma; PTLD: post-transplant lymphoproliferative disease; PBMC: peripheral blood mononuclear cells; GC: gastric carcinoma; IM: infectious mononucleosis; EBVaGC: EBV-associated gastric carcinoma; n/a: information not available.

3. Next-Generation Sequencing of EBV Genomes

3.1. Direct Next-Generation Sequencing without EBV Enrichment

GD2, as with GD1, was an EBV genome from NPC patient of Guangdong province. However, instead of capturing the EBV from saliva of the patient, GD2 was directly sequenced from an NPC tumor [17]. Its genome was obtained as a small subset of sequence data from next-generation sequencing of total DNA sequences derived from the primary NPC tumor, hence, can be acknowledged as the first natural EBV genome determined by next-generation sequencing technology [17]. This study illustrated that it is possible to directly determine EBV genomes in tumor biopsies without enrichment, though only reads 0.0141% of the total reads were mapped to EBV and used to construct GD2 [17].

C666-1 is a sub-clone of C666, an epithelial cell line derived from an NPC xenograft of Southern Chinese origin [30]. This C666-1 NPC cell line is the most representative NPC cell line to date, since it retains the native EBV while other NPC-derived cell lines have lost their EBV through the *in vitro* culture. The C666-1 genome was sequenced in paired-end 100 bases protocol on an Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA). With a total of 251 gb of output data, the EBV coverage reaches 504 folds despite the process not involving PCR amplification or other methods of enrichment. The consensus EBV sequence was constructed through reference mapping and Sanger sequencing [18]. The C666-1 EBV was found to be phylogenetically close to other Chinese NPC sequences GD1, -2, and HKNPC1.

More recently, two EBV genomes in immortalized human B lymphocyte cell lines were sequenced using the Illumina MiSeq platform [19]. These cell lines were derived from peripheral blood of two healthy donors. Sequencing reads from total DNA of the cell lines were mapped to the EBV reference genome and the mapped reads were assembled by CLC Genomic Workbench to give the two EBV genomes, K4413-Mi and K4123-Mi. In the same study, NA12878 EBV genome, which represents the EBV in peripheral blood of a Caucasian subject, was assembled using the data from the 1000 Genome Project [19]. Similarly, ten EBV sequences were constructed from the reads unmapped to human-generated from sequencing the lymphoblastic cell lines from the 1000 Genome Project [20]. These studies have demonstrated that published human genomic data may contain viral sequences useful in EBV studies.

Sequencing expense incurred by the presence of the much more abundant cellular genomic DNA inherent in standard DNA preparations can be prohibitive to large scale sequencing of viral genomes. These studies have shown that it is possible to assemble EBV as a side product of sequencing of cellular DNA, provided there are high copy numbers of EBV and a high throughput of sequencing data.

3.2. Next-Generation Sequencing with EBV Enrichment

3.2.1. Enrichment by Induction of Lytic Viral Replication

Akata and Mutu are Burkitt's lymphoma cell lines which are commonly-used model cell lines. The sequencing of EBV strains in Akata and Mutu cell line was made feasible by taking advantage of the properties of lytic replication of EBV in these lines to increase the viral copy number. Induction of viral lytic replication was performed by incubating the cells in medium containing anti-immunoglobulinA (anti-IgA) and anti-IgG [23]. Subsequently, the relative amount of cellular DNA was reduced by Hirt DNA extraction method. Their EBV genomes were determined by paired-end 100-base sequencing and were *de novo* assembled using reads aligned to EBV reference [23].

3.2.2. Enrichment by PCR

Induction of lytic replication can only be applied to certain cell lines. One way to increase the abundance of viral DNA in non-inducible lines or primary specimens is by PCR amplification. EBV from a primary NPC biopsy, HKNPC1, was PCR-amplified using 60 primer pairs and the Illumina Genome Analyzer II platform was used to determine its sequence [25]. The HKNPC1 EBV genome

was generated by mapping the reads to the wild type EBV and calling the consensus sequence from the alignment. As much as 90% of total usable reads were mapped to type 1 EBV reference, hence, signified how PCR enrichment can greatly increase cost-effectiveness in determination of the viral sequence. Similarly, the EBVs of spontaneous LCLs established from peripheral blood mononuclear cells (PBMC) of four Kenyan children were PCR-amplified by using 59 pairs of primer [26]. The pooled PCR products were sequenced by Illumina MiSeq platform, and there were more than 90% of reads in each sample could be aligned to reference EBV genome.

3.2.3. Cloning into F-Factor Plasmids

The M81 EBV was cloned into the F-factor-based replicon in *E. coli* [31], at the EBV terminal repeats [24]. The genome of M81 was then determined by Illumina HiSeq 2000 sequencing and assembled by GS Reference Mapper software [24]. Standard Sanger sequencing was used to confirm ambiguous regions and repeat regions. The study showed that the M81 EBV exhibits a higher tissue tropism towards epithelial cells than B95-8 [24]. The virus sequence was also found to be highly similar to viruses isolated from Chinese NPCs. Polymorphisms in BZLF1 gene have been pointed by recombinant virus assays to have contributed to this phenomenon. It proved the value of genomic sequencing of EBV strains from different geographic localities and disease origin.

3.2.4. Target DNA Capture by Hybridization

Amplicon sequencing of HKNPC1 has signified how target enrichment can significantly increase the proportion of viral DNA to host cellular DNA and, hence, greatly increase efficiency in utilizing the capacity of next-generation sequencing technology. However, the success in sequencing of HKNPC1 is not sufficient to drive projects of large-scale sequencing of EBV genomes. Amplicon sequencing works well for sequencing of single samples, but the procedures of PCR, following by purification and normalization of PCR products, can be time-consuming and labor-intensive for large numbers of samples. Moreover, PCR amplification can be a source of sequence errors, even if high-fidelity polymerases are used [32]. It is also difficult to control for the pooling of PCR products into equal molar concentration across the whole genome.

The technology of target DNA enrichment by hybridization can be traced back to traditional techniques for enriching nucleic acids with biotinylated DNA bait [33]. Modifications of this technique have been made to serve different purpose, for example, to detecting low-frequency variants and recombinant DNA molecules from a DNA pool [34]. Exome sequencing, which also known as targeted exome capture, utilized the enrichment-by-hybridization approach to enrich exonic DNA, which is approximately 1% of the entire human genome [35]. The same principle has been applied to capture the target DNA of interest, in our case, the EBV DNA.

RNA Bait

The SureSelect target enrichment system has been used to successfully enrich and sequence EBV and KSHV DNA [36]. RNA probes of 120 bp in length, overlapped at 5x, have been designed to cover the EBV genome. Three published EBV (B95-8, C666-1, and HKNPC1) genomes were first re-sequenced to validate the sequencing workflow. The whole sequences of eight NPC biopsy-derived EBV (NPC-EBV) genomes, designated HKNPC2 to -9, were then determined [27].

The same hybridization capture system has been utilized in a study to determine the genomic sequences of 71 geographically-distinct EBV strains from cell lines, multiple types of primary tumors such as NPC, Burkitt's lymphoma and Hodgkin's lymphoma, blood samples, and a saliva sample [28]. The study greatly enriched the genomic data of EBV strains, in particular those from East Africa and Australia. It is also the most comprehensive study on geographic distribution of strains defined on the level of whole EBV genome.

DNA Probe

Target capture through hybridization enabled the sequencing of EBV with greater time- and cost-effectiveness than PCR-based enrichment. The higher coverage uniformity also improved the quality of contigs from *de novo* assembly. However, the method still leaves us the task of joining the contigs to form complete genomes. The more fragmented the contigs are, the greater effort is needed to join them. With the aim of improving the capture system from sequence uniformity to contig quality, we utilized individually-synthesized DNA oligos as probes for EBV capture. We also increased the read length from paired-end 100 bp to 300 bp. The experimental detail is described in section five.

4. Construction of EBV Genomes Using NGS Data

4.1. Reference Mapping and Consensus Calling

Bioinformatics analysis of next-generation sequencing data is crucial to convert the raw output data to something of biological interest. Two major approaches are reference mapping and *de novo* assembly. With reference mapping each read is aligned to the reference genome. This is common in re-sequencing projects which the genomic sequence of the organism is already known and one is only interested in the variations among different strains of the same species. The commonest short-read alignment programs have been developed based on the Burrow-Wheeler transform (BWT) algorithm. Examples of programs using this algorithm include BOWTIE [37], SOAP2 [38] and the one used in this study, namely Burrow-Wheeler Aligner (BWA) [39]. This algorithm first creates an efficient index of the reference genome in order to facilitate rapid searching under limited system memory. The trade-off for this fast BWT method is the sensitivity of finding alignments. BWA is only able to find alignments within a certain "edit distance" to the reference genome sequence [39]. Edit distance can be described as the number of operations required to transform one sequence to another, which affected by the number of gaps and mismatch between the reads and the reference genome. As a result of this limitation of edit distance, large insertions which exceeded the size of the allowed mismatch will not be detected. Moreover, mismatches occur more frequently and polymorphic regions, such as EBNA-2, -3, and LMP-1 and -2, where sequence reads most probably differ from the reference, and hence, reduce the validity of the consensus sequence generated from reference mapping. Repeat regions are also prone to misalignment and very often result in a high number of reported variations in the consensus sequence. A great effort for validation is required in these regions of the EBV genome.

For data generated from sequencing protocol without EBV enrichment, it is encouraged to first align the reads to the human genome and use the unmapped reads for further EBV analyses. The EBV genome contains genes that are homologous to human genes. For example, BHRF1 protein from EBV is a homolog of human Bcl-2 [40], while BCLF1 is a homolog to human interleukin-10 [41]. Removing human reads reduces the possibility of mistaking human sequences as EBV sequences. It also greatly reduces the file size and the demand of computing power in downstream analyses.

With reference-mapping one assumes that the new genome sequenced is highly similar to the reference sequence. There are two NCBI reference EBV genomes, one for each major type of EBV: type 1 (Accession no. NC_007605.1) and type 2 (NC_009334.1). Mapping the reads to one of the references will normally suffice to identify the type of EBV sequenced, since gaps of no reads will be observed at type-specific genes EBNA-2 and -3 when data of type 1 EBV is mapped to the type 2 reference, and *vice versa*. However, it is recommended to map the reads to both EBV references since, in some uncommon scenarios, both types of the virus may co-exist in a specimen. Future publications should also report exactly which of the reference sequences to be used.

Consensus calling constructs EBV genomes using mapped data. The dominant nucleotide species from the reads mapped along the reference will be called at every single position and considered as the consensus sequence. The minor nucleotide species will be either be regarded as sequencing or mapping error, or inferred that a coinfection of minor viral strain is present. In the study of HKNPC1 amplicon sequencing, we defined a position to be homogeneous if the variant frequency is $\geqslant$95% and a position

to be heterogeneous if the variant frequency is between 20% and 94% [25]. Multiple nucleotide species is very commonly observed in repeat regions due to misalignment of reads, therefore meticulous validation should be conducted by pyrosequencing or other techniques. Nucleotide positions with read depth less than five were classified as ambiguous sites as there is insufficient depth to make a high confidence call. Since these cut-offs are arbitrary and are subjected to adjustment in different sequencing protocols, it is advised to determine the error rate and percentage cut-off for genuine minor variants experimentally through manual mixing two different strains of the virus.

4.2. De Novo *Assembly*

De novo sequencing involves sequencing a novel genome without the aid of external data. It does not depend on existing reference genomes and, hence, can reveal large genetic variations or regions differing significantly from the references. The general prerequisites for high-quality assemblies include high base quality of reads and uniform coverage.

In comparison to reference mapping, where sequence error is adjusted by cut-offs during the variants and consensus calling, it is necessary to ensure a good quality of input sequence data in *de novo* assembly. Each base in a read is assigned a quality score by the Illumina sequencing platform using a phred-like algorithm [42,43], the distribution of mean quality scores is a rough estimation of the quality of the sequencing run. The per-sequence quality scores module of the FastQC software reports the distribution of mean quality scores of individual reads and this serves as the basis for trimming. There are at least two main approaches in removing low quality bases from reads. The first is to select a single cut-off for all reads, based on a summary of base quality data. The choice of trimming length is a compromise between base quality and desired read length for assembly. In a typical 300 bp MiSeq run the lower quartile of quality score per base penetrated Q20 at approximately 230 bp from 5' end. Removing the last 70 bases from 3' end ensures only high quality bases in a read are retained. In a paired-end 76 bp run where the overall base quality is generally higher, a more stringent trimming regime might be tolerated and only the last six bases are trimmed (Figure A1).

The second main approach is to trim base on the sequence quality for individual reads. The length of reads after trimming varies because of the variable base quality in each reads. A number of software is available for this purpose and is reviewed by del Fabbro and colleagues [44]. The best algorithm is sample-dependent and requires trial-and-error.

Algorithms of most genome assembly software make use of information on expected coverage to infer genome assemblies and copy number variations. It directly affects the quality of *de novo* assemblies. The highly-uneven coverage in HKNPC1 obtained by the amplicon sequencing approach resulted in more fragmented contigs, with 25 contigs alignable to the reference EBV, compared with 15 contigs from the data of target capture sequencing (Figure 1). Uneven coverage poses problems mainly in repeats. The amplicon sequenced HKNPC1 failed to be assembled into contiguous sequence in almost all major and minor repeats of EBV. Read uniformity is particularly important at low sequencing depth. This is often the case for sequencing samples of low EBV viral load.

The assembler used in the study of sequencing HKNPC2 to -9 [32], namely Velvet [45], was developed based on the de Bruijn graphs method to resolve the assembly of high depth and short read data generated from NGS technology. This algorithm was originally developed for problems in combinatorial mathematics [46]. It is based on graphs of small, fixed-length subsequences know as k-mer, where k equals to the subsequence length. The assembly result strongly depends on this value of k and other parameters, such as expected average k-mer coverage (exp_cov) and the lowest coverage of nodes to be excluded (cov_cutoff). These parameters, particularly the value of k, are difficult to estimate without running through different values of these parameters and scrutinize the result manually. Case-by-case optimization is required even for assembling genomes of the same species and data from the same sequencing protocol.

Sanger sequence remains, to date, to be the gold standard for bridging the contigs and validation of variations. With our current target capture platform, contigs of most of the samples do share the

similar positions of breaks at repeat regions (Figure 2e) of the EBV genome. Therefore. the same set of primers could be applied to bridge the gaps in most of the samples. This will greatly increase the efficiency of constructing complete EBV genomes when the study is scaled up to include hundreds of samples. With this aim of reducing contig gaps in mind, we tested another hybridization system for EBV capture.

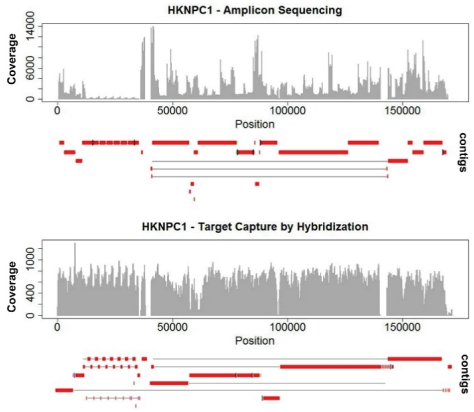

Figure 1. Coverage and alignment of contigs of HKNPC1 EBV by amplicon sequencing and target capture. Uneven read coverage is observed in amplicon sequencing of HKNPC1. More uniform coverage is observed in HKNPC1 enriched by target capture through hybridization using Agilent RNA-bait. The contigs assembled in amplicon sequencing are more fragmented than that from target capture.

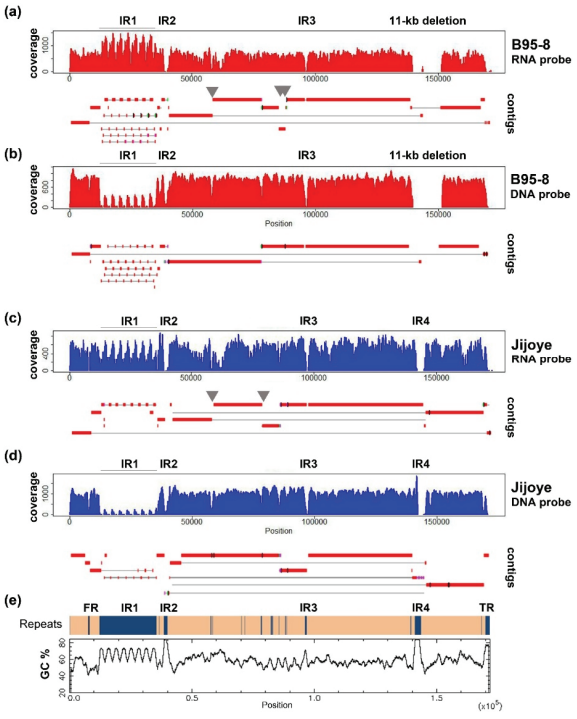

Figure 2. Coverage and alignment of contigs of B95-8 and Jijoye EBV by different capture strategies and sequencing protocols. (**a**) B95-8 EBV captured by RNA bait and sequenced by MiSeq PE150; (**b**) B95-8 EBV captured by DNA probe and sequenced by MiSeq PE300; (**c**) Jijoye EBV captured by RNA-bait and sequenced by MiSeq PE150; (**d**) Jijoye EBV captured by DNA probe and sequenced by MiSeq PE300; and (**e**) distribution of repeat regions and GC percentage across the EBV genome. The GC percentage plot is created by EMBOSS Cpgplot.

5. Resequencing of Prototype Type 1 (B95-8) and Type 2 (Jijoye) EBV—A Comparison of DNA and RNA Probes

5.1. Cell Lines

B95-8 has been re-sequenced in multiple studies and, hence, served as a gold standard for sequence accuracy. Jijoye is re-sequenced in this study to validate the capability of the system to capture type 2 EBV.

5.2. DNA Probes

Customized Integrated DNA Technology (IDT) xGen® Lockdown® probes are 120 bp DNA oligos designed across the whole genome of type 1 EBV and selected regions of type 2 EBV. These oligos covered the genome at end-to-end (1x) coverage. The oligo pools are re-suspended and mixed according to the manufacturer's protocol.

5.3. Library Preparation, Target Capture, and Sequencing Analysis

The MiSeq platform was used to analyze the EBV genomes in B95.8 and Jijoye cell lines. NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, MA, USA) and SeqCap® EZ Hybridization and Wash Kits (Roche Nimblegen, Madison, WI, USA) were used for library preparation and target capture. DNA oligo probes were synthesized as a pool of custom oligos by Integrated DNA Technologies (IDT). All library preparation, hybridization, and post-reaction clean-up steps were performed according to the Rapid Protocol for DNA Probe Hybridization and Target Capture Using an Illumina TruSeq (Version 2.0) observing all recommended quality control steps. Denatured DNA libraries were mixed with a PhiX control library. Cluster generation and 300-bp pair-ended sequencing were performed in succession using the MiSeq platform, according to manufacturer's protocol.

5.4. De Novo *Assembly of EBV Genomes*

Sequencing reads from MiSeq Personal Sequencer were demultiplexed into individual samples by allowing one mismatch in the index sequence. Quality assessment and filtering on the raw reads were carried out to remove reads containing adaptor sequences. Coverage of reads was assessed by mapping untrimmed reads of each sample to the type 1 (NC_007605) and type 2 (DQ279927) reference EBV genomes by BWA software. Bases of poor quality at the end of sequence reads were trimmed. These reads were assembled using a de Bruijn graph assembler Velvet [45]. Location and orientation of contigs were evaluated by pairwise aligning of the contigs to the reference EBV genomes using NCBI alignment tools.

5.5. Results

EBV DNA libraries captured by DNA probes and sequenced in pair-ended 300 bp run were trimmed to 230 bp and utilized in assembling the contigs. Using a k-mer length of 55, an expected k-mer coverage of 650, and a k-mer coverage cut-off of 50, the output graph for B95-8 has 94 nodes. The maximum contig size is 38,247 bp and N50 equals 42,581. Using a k-length of 45, an expected k-mer coverage of 900, and a k-coverage cut-off of 200, the output graph for B95-8 has 91 nodes. The maximum contig size is 42,247 bp and N50 equals 39,849 bp. The coverage profile and alignment of contigs to the reference sequences are illustrated in Figure 2.

We compare the data to that produced from different reagents and sequencing protocols, in which EBV DNA libraries were captured by RNA bait and sequenced in pair-ended 150 bp run. The method was described in details in a previous publication [27]. The output reads were trimmed to 100 bp, and then utilized in assembling the contigs. Using a k-mer length of 37 and a k-mer coverage cut-off of 50, the output graph for B95-8 has 70 nodes. The maximum contig size is 50,083 bp and N50 equals 17,546. Using a k-mer length of 37, an expected k-mer coverage of 400, and a k-coverage cut-off of

50, the output graph for Jijoye has 45 nodes. The maximum contig size is 44,262 bp and N50 equals 19,989 bp.

Most of the non-repeats regions are assembled reasonably well in both approaches. Increasing the read length to paired-end 300 bp helped significantly to resolve minor repeats. Repeats in BPLF1 (NC_007605 coordinates 57,396–57,642; 58,099–58,233), EBNA3A (81,920–82,781), and -3B (85,234–85,410) genes (grey arrows in Figure 2), which were broken in contigs in 150 bp run, were bridged in contigs of the 300 bp run. However, increasing read length does not help to close the gaps at internal repeats 1, -2, -3, and -4, and repeats at the OriP region, since they either have a large repeating unit, as in internal repeat 1 (IR1), or the entire repeat length is too large to be covered in a single read (e.g., IR4). Some of these difficult regions are of particularly high GC content (over 80% in IR2 and -4, shown in Figure 2e) which poses difficulties in generating high-quality reads and assembly.

6. Future Development

One of the bioinformatics strategies used to improve the quality of assemblies is by utilizing variable sizes of k-mer to building of de Bruijn graphs, and merges the results from different inputs of k. This was suggested to improve quality and length of contigs in transcriptome analysis [47] and metagenomic analysis [48]. Application of this strategy to EBV genome assembly might further improve contig quality by bridging some of the gaps due to minor repeats.

The major repeats, including internal repeats (IR), terminal repeats (TR), and family of repeats (FR) in the OriP region remain to be an obstacle for completing the EBV genome. Single molecule real-time sequencing (SMRT) technology is able to sequence ultra-long reads of on average 10,000 bp or longer. The platform has been applied to sequence microbial genome and was able to obtain almost gapless contigs upon assembly [49]. Not being vulnerable to GC bias, it can potentially sequence through difficult regions such as IR2 and -4. When target capture and sample preparation for SMRT sequencing is better established and the cost become less prohibitive, it might greatly facilitate construction of EBV genomes in large scale sequencing projects.

With growing number of EBV genomes available, it becomes increasing important to design studies that can reveal biological significance of the genomic variations of the virus. Consideration of disease associations, for example, must include analyses of appropriate control sequences. EBV that persists naturally in immunocompetent hosts without noticeable symptoms from the same geographic region and ethnic group serve as useful controls since it minimizes the effect of geographical variations among EBV genomes. Although little is known about the correlation of age and sex with viral strains, it is of common practice to include age- and sex-matched control to ensure only the condition of interest is interrogated. Since EBV is known to be tropic to certain cell types, in particular B-cells and epithelial cells, specimens from different body compartments (peripheral blood mononuclear cells, plasma, saliva, nasopharyngeal biopsies, *etc.*) of the same individual, or viruses from the same compartment of different individuals, will shed light on the cell-type-specific effect of the virus.

This review provides an overview of the development of EBV genomic sequencing since the first EBV genome obtained. It also described recent advances and potential future directions for EBV sequencing projects. On the basis of these advances, future work of EBV genomic sequencing will facilitate a better understanding of the genetic variations of EBV and discovery of potentially-pathogenic variants.

Acknowledgments: This study is funded by EBV Research grant 20004525, HMRF grant (Project 02131706) and NPC Area of Excellence (AoE/M 06/08 Centre of Nasopharyngeal Carcinoma Research) of AKSC. The above funds cover the publication costs.

Author Contributions: Hin Kwok and Alan Kwok Shing Chiang conceived and designed the experiments; Hin Kwok performed the experiments; Hin Kwok and Alan Kwok Shing Chiang analyzed the data; Alan Kwok Shing Chiang contributed reagents/materials/analysis tools; Hin Kwok and Alan Kwok Shing Chiang wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A

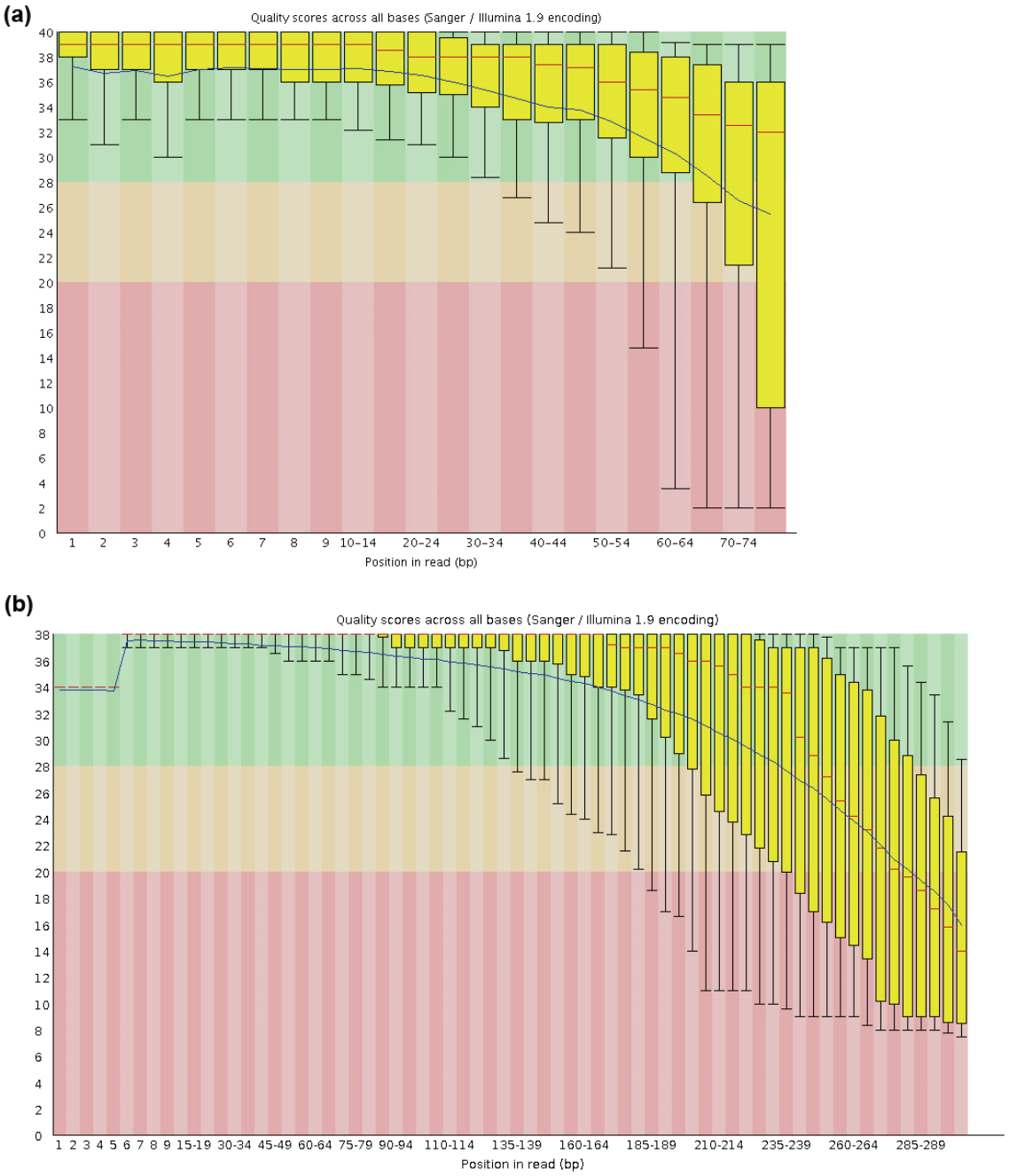

Figure A1. Per base sequence quality of capture sequencing of B95-8 EBV on Genome Analyzer IIx 76 bp protocol (**a**) and MiSeq 300 bp protocol (**b**). These quality profiles are representative of other samples in the same run.

References

1. Grunewald, V.; Bonnet, M.; Boutin, S.; Yip, T.; Louzir, H.; Levrero, M.; Seigneurin, J.M.; Raphael, M.; Touitou, R.; Martel-Renoir, D.; *et al*. Amino-acid change in the Epstein-Barr-virus zebra protein in undifferentiated nasopharyngeal carcinomas from europe and North Africa. *Int. J. Cancer* **1998**, *75*, 497–503. [CrossRef]
2. Sacaze, C.; Henry, S.; Icart, J.; Mariame, B. Tissue specific distribution of Epstein-Barr virus (EBV) BZLF1 gene variants in nasopharyngeal carcinoma (NPC) bearing patients. *Virus Res.* **2001**, *81*, 133–142. [CrossRef]
3. Dardari, R.; Khyatti, M.; Cordeiro, P.; Odda, M.; ElGueddari, B.; Hassar, M.; Menezes, J. High frequency of latent membrane protein-1 30-bp deletion variant with specific single mutations in Epstein-Barr virus-associated nasopharyngeal carcinoma in Moroccan patients. *Int. J. Cancer* **2006**, *118*, 1977–1983. [CrossRef] [PubMed]
4. Chang, K.P.; Hao, S.P.; Lin, S.Y.; Ueng, S.H.; Pai, P.C.; Tseng, C.K.; Hsueh, C.; Hsieh, M.S.; Yu, J.S.; Tsang, N.M. The 30-bp deletion of Epstein-Barr virus latent membrane protein-1 gene has no effect in nasopharyngeal carcinoma. *Laryngoscope* **2006**, *116*, 541–546. [CrossRef] [PubMed]
5. Nguyen-Van, D.; Ernberg, I.; Phan-Thi Phi, P.; Tran-Thi, C.; Hu, L. Epstein-Barr virus genetic variation in Vietnamese patients with nasopharyngeal carcinoma: Full-length analysis of LMP1. *Virus Genes* **2008**, *37*, 273–281. [CrossRef] [PubMed]
6. See, H.S.; Yap, Y.Y.; Yip, W.K.; Seow, H.F. Epstein-Barr virus latent membrane protein-1 (LMP-1) 30-bp deletion and Xho I-loss is associated with type III nasopharyngeal carcinoma in Malaysia. *World J. Surg. Oncol.* **2008**, *6*. [CrossRef] [PubMed]
7. Farrell, P.J. Epstein-Barr virus. The b95–8 strain map. *Methods Mol. Biol.* **2001**, *174*, 3–12. [PubMed]
8. Dambaugh, T.; Beisel, C.; Hummel, M.; King, W.; Fennewald, S.; Cheung, A.; Heller, M.; Raab-Traub, N.; Kieff, E. Epstein-Barr virus (B95–8) DNA VII: Molecular cloning and detailed mapping. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 2999–3003. [CrossRef] [PubMed]
9. Skare, J.; Strominger, J.L. Cloning and mapping of BamHi endonuclease fragments of DNA from the transforming B95–8 strain of Epstein-Barr virus. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 3860–3864. [CrossRef] [PubMed]
10. Arrand, J.R.; Rymo, L.; Walsh, J.E.; Bjorck, E.; Lindahl, T.; Griffin, B.E. Molecular cloning of the complete Epstein-Barr virus genome as a set of overlapping restriction endonuclease fragments. *Nucleic Acids Res.* **1981**, *9*, 2999–3014. [CrossRef] [PubMed]
11. Baer, R.; Bankier, A.T.; Biggin, M.D.; Deininger, P.L.; Farrell, P.J.; Gibson, T.J.; Hatfull, G.; Hudson, G.S.; Satchwell, S.C.; Seguin, C.; *et al*. DNA sequence and expression of the b95–8 Epstein-Barr virus genome. *Nature* **1984**, *310*, 207–211. [CrossRef] [PubMed]
12. Hummel, M.; Kieff, E. Epstein-Barr virus RNA. VIII. Viral RNA in permissively infected B95–8 cells. *J. Virol.* **1982**, *43*, 262–272. [PubMed]
13. Weigel, R.; Miller, G. Major EB virus-specific cytoplasmic transcripts in a cellular clone of the HR-1 burkitt lymphoma line during latency and after induction of viral replicative cycle by phorbol esters. *Virology* **1983**, *125*, 287–298. [CrossRef]
14. De Jesus, O.; Smith, P.R.; Spender, L.C.; Elgueta Karstegl, C.; Niller, H.H.; Huang, D.; Farrell, P.J. Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the bart (CST, BARF0) RNAs of EBV. *J. Gen. Virol.* **2003**, *84*, 1443–1450. [CrossRef] [PubMed]
15. Zeng, M.S.; Li, D.J.; Liu, Q.L.; Song, L.B.; Li, M.Z.; Zhang, R.H.; Yu, X.J.; Wang, H.M.; Ernberg, I.; Zeng, Y.X. Genomic sequence analysis of Epstein-Barr virus strain GD1 from a nasopharyngeal carcinoma patient. *J. Virol.* **2005**, *79*, 15323–15330. [CrossRef] [PubMed]
16. Dolan, A.; Addison, C.; Gatherer, D.; Davison, A.J.; McGeoch, D.J. The genome of Epstein-Barr virus type 2 strain AG876. *Virology* **2006**, *350*, 164–170. [CrossRef] [PubMed]
17. Liu, P.; Fang, X.; Feng, Z.; Guo, Y.M.; Peng, R.J.; Liu, T.; Huang, Z.; Feng, Y.; Sun, X.; Xiong, Z.; *et al*. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue using next-generation sequencing technology. *J. Virol.* **2011**, *85*, 11291–11299. [CrossRef] [PubMed]
18. Tso, K.K.; Yip, K.Y.; Mak, C.K.; Chung, G.T.; Lee, S.D.; Cheung, S.T.; To, K.F.; Lo, K.W. Complete genomic sequence of Epstein-Barr virus in nasopharyngeal carcinoma cell line C666-1. *Infect. Agent Cancer* **2013**, *8*. [CrossRef] [PubMed]

19. Lei, H.; Li, T.; Hung, G.C.; Li, B.; Tsai, S.; Lo, S.C. Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. *BMC Genomics* **2013**, *14*. [CrossRef] [PubMed]
20. Santpere, G.; Darre, F.; Blanco, S.; Alcami, A.; Villoslada, P.; Mar Alba, M.; Navarro, A. Genome-wide analysis of wild-type Epstein-Barr virus genomes derived from healthy individuals of the 1000 genomes project. *Genome Biol. Evol.* **2014**, *6*, 846–860. [CrossRef] [PubMed]
21. Song, K.A.; Yang, S.D.; Hwang, J.; Kim, J.I.; Kang, M.S. The full-length DNA sequence of epstein barr virus from a human gastric carcinoma cell line, SNU-719. *Virus Genes* **2015**, *51*, 329–337. [CrossRef] [PubMed]
22. Lei, H.; Li, T.; Li, B.; Tsai, S.; Biggar, R.J.; Nkrumah, F.; Neequaye, J.; Gutierrez, M.; Epelman, S.; Mbulaiteye, S.M.; *et al*. Epstein-Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations. *Sci. Rep.* **2015**, *5*. [CrossRef] [PubMed]
23. Lin, Z.; Wang, X.; Strong, M.J.; Concha, M.; Baddoo, M.; Xu, G.; Baribault, C.; Fewell, C.; Hulme, W.; Hedges, D.; *et al*. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *J. Virol.* **2013**, *87*, 1172–1182. [CrossRef] [PubMed]
24. Tsai, M.H.; Raykova, A.; Klinke, O.; Bernhardt, K.; Gartner, K.; Leung, C.S.; Geletneky, K.; Sertel, S.; Munz, C.; Feederle, R.; *et al*. Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. *Cell Rep.* **2013**, *5*, 458–470. [CrossRef] [PubMed]
25. Kwok, H.; Tong, A.H.; Lin, C.H.; Lok, S.; Farrell, P.J.; Kwong, D.L.; Chiang, A.K. Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE* **2012**, *7*, e36939. [CrossRef] [PubMed]
26. Simbiri, K.O.; Smith, N.A.; Otieno, R.; Wohlford, E.E.; Daud, I.I.; Odada, S.P.; Middleton, F.; Rochford, R. Epstein-Barr virus genetic variation in lymphoblastoid cell lines derived from Kenyan pediatric population. *PLoS ONE* **2015**, *10*, e0125420. [CrossRef] [PubMed]
27. Kwok, H.; Wu, C.W.; Palser, A.L.; Kellam, P.; Sham, P.C.; Kwong, D.L.; Chiang, A.K. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. *J. Virol.* **2014**, *88*, 10662–10672. [CrossRef] [PubMed]
28. Palser, A.L.; Grayson, N.E.; White, R.E.; Corton, C.; Correia, S.; Ba Abdullah, M.M.; Watson, S.J.; Cotten, M.; Arrand, J.R.; Murray, P.G.; *et al*. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J. Virol.* **2015**, *89*, 5222–5237. [CrossRef] [PubMed]
29. Liu, Y.; Yang, W.; Pan, Y.; Ji, J.; Lu, Z.; Ke, Y. Genome-wide analysis of Epstein-Barr virus (EBV) isolated from EBV-associated gastric carcinoma (EBVaGC). *Oncotarget* **2015**. [CrossRef]
30. Cheung, S.T.; Huang, D.P.; Hui, A.B.; Lo, K.W.; Ko, C.W.; Tsang, Y.S.; Wong, N.; Whitney, B.M.; Lee, J.C. Nasopharyngeal carcinoma cell line (C666-1) consistently harbouring Epstein-Barr virus. *Int. J. Cancer* **1999**, *83*, 121–126. [CrossRef]
31. Delecluse, H.J.; Hilsendegen, T.; Pich, D.; Zeidler, R.; Hammerschmidt, W. Propagation and recovery of intact, infectious Epstein-Barr virus from prokaryotic to human cells. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 8245–8250. [CrossRef] [PubMed]
32. Eckert, K.A.; Kunkel, T.A. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.* **1991**, *1*, 17–24. [CrossRef] [PubMed]
33. Abe, K. Rapid isolation of desired sequences from lone linker PCR amplified cDNA mixtures: Application to identification and recovery of expressed sequences in cloned genomic DNA. *Mamm. Genome* **1992**, *2*, 252–259. [CrossRef] [PubMed]
34. Jeffreys, A.J.; May, C.A. DNA enrichment by allele-specific hybridization (DEASH): A novel method for haplotyping and for detecting low-frequency base substitutional variants and recombinant DNA molecules. *Genome Res.* **2003**, *13*, 2316–2324. [CrossRef] [PubMed]
35. Bamshad, M.J.; Ng, S.B.; Bigham, A.W.; Tabor, H.K.; Emond, M.J.; Nickerson, D.A.; Shendure, J. Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.* **2011**, *12*, 745–755. [CrossRef] [PubMed]
36. Depledge, D.P.; Palser, A.L.; Watson, S.J.; Lai, I.Y.; Gray, E.R.; Grant, P.; Kanda, R.K.; Leproust, E.; Kellam, P.; Breuer, J. Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* **2011**, *6*, e27805. [CrossRef] [PubMed]
37. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*. [CrossRef] [PubMed]

38. Li, R.; Yu, C.; Li, Y.; Lam, T.W.; Yiu, S.M.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967. [CrossRef] [PubMed]
39. Li, H.; Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
40. Huang, Q.; Petros, A.M.; Virgin, H.W.; Fesik, S.W.; Olejniczak, E.T. Solution structure of the BHRF1 protein from Epstein-Barr virus, a homolog of human BCL-2. *J. Mol. Biol.* **2003**, *332*, 1123–1130. [CrossRef] [PubMed]
41. Stewart, J.P.; Rooney, C.M. The interleukin-10 homolog encoded by Epstein-Barr virus enhances the reactivation of virus-specific cytotoxic T cell and HLA-unrestricted killer cell responses. *Virology* **1992**, *191*, 773–782. [CrossRef]
42. Ewing, B.; Hillier, L.; Wendl, M.C.; Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **1998**, *8*, 175–185. [CrossRef] [PubMed]
43. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194. [CrossRef] [PubMed]
44. Del Fabbro, C.; Scalabrin, S.; Morgante, M.; Giorgi, F.M. An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE* **2013**, *8*, e85024. [CrossRef] [PubMed]
45. Zerbino, D.R.; Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef] [PubMed]
46. Idury, R.M.; Waterman, M.S. A new algorithm for DNA sequence assembly. *J. Comput. Biol.* **1995**, *2*, 291–306. [CrossRef] [PubMed]
47. Surget-Groba, Y.; Montoya-Burgos, J.I. Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Res.* **2010**, *20*, 1432–1440. [CrossRef] [PubMed]
48. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef] [PubMed]
49. Koren, S.; Phillippy, A.M. One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **2015**, *23*, 110–120. [CrossRef] [PubMed]