The transcriptomic landscape of Epstein-Barr virus associated tumors at cellular and single-molecule level

Yarong Tian

Department of Infectious Diseases

Institute of Biomedicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2022

The transcriptomic landscape of Epstein-Barr virus associated tumors at cellular and single-molecule level

yarong.tian@gu.se

ISBN 978-91-8009-835-9 (PRINT)
ISBN 978-91-8009-836-6 (PDF)

Printed in Borås, Sweden 2022
Printed by Stema Specialtryck AB

Cell the unit,

Good at first.

The same nature,

Varies on nurture.

- Three Character Classic

To my family, biological and scientific

Shaanxi – Fujian – Gothenburg

The transcriptomic landscape of Epstein-Barr virus associated tumors at cellular and single-molecule level

Yarong Tian

Department of Infectious Diseases, Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Epstein-Barr virus (EBV) was the first oncovirus found in humans. Almost all adults worldwide are asymptomatic carriers of EBV. The latent EBV-infection malignifies in approximately 200,000 individuals each year. The risk of developing certain types of EBV-associated cancer is high in specific regions, for example nasopharyngeal carcinoma in Southeast Asia and Burkitt's lymphoma in Africa. The overall aim of this thesis was to characterize the EBV gene expression patterns in biopsies and elucidate the function of the expressed viral genes.

Bulk transcriptome datasets of 615 tumors from four types of known EBV-associated neoplasms and single-cell transcriptome data from 63 nasopharyngeal samples were screened for EBV expression. The most abundant EBV RNA found at both tissue and single-cell levels, were RPMS1 and the novel co-terminating transcripts which we named BAREs. LMP1/BNLF2a/b and LMP2A/B/BNRF1 were expressed to a lesser extent and large differences were observed between individuals. Single-cell sequencing of B-lymphocytes isolated from the peripheral blood of a patient with a high

EBV DNA load showed a similar EBV expression profile as the EBV-positive tumors. Moreover, the highly expressed EBV genes RPMS1 and BAREs were subjected to full-length single-molecule sequencing and all isoforms were characterized using our newly developed bioinformatics tool FLAME.

Our results show that available EBV cell models inadequately portray primary tumors with regard to the viral gene expression and/or the propensity for reactivation. We developed an *in vitro* nasopharyngeal pseudostratified epithelium model which could mimic an EBV infection in the nasopharynx. A donor-dependent susceptibility for EBV infection was observed and both latent and lytic EBV expression patterns were detected in cells from a single donor. Single-cell sequencing data analysis could further distinguish that cells in late lytic stage with virus host shutoff were found amongst the suprabasal cells.

The single-cell data from peripheral EBV-transformed B-lymphocytes identified that EBV induces proliferative pathways. In nasopharyngeal carcinoma tissue the EBV-transformed epithelial cells exists in a microenvironment with lymphocytic infiltration and interferon. Single-cell characterization of the nasopharyngeal cancer cells identified that the EBV expression of RPMS1 along with the miR-BARTs encoded in the introns promotes immune evasion by downregulation of interferon responsive genes. The findings suggest that EBV contributes to tumorigenesis in two ways, the first is by host cell reprogramming and induction of proliferation by EBNAs and LMP1, and the second is by immune evasion and escape by RPMS1 and BNLF2a.

Keywords: Epstein-Barr virus, tumor, RPMS1, miRNA, single-cell sequencing, immune evasion

ISBN 978-91-8009-835-9 (PRINT)
ISBN 978-91-8009-836-6 (PDF)

Sammanfattning på svenska

Epstein-Barr-virus (EBV) var det första tumörviruset som hittades hos människor. Nästan alla i den vuxna befolkningen världen över är asymtomatiska bärare av EBV. Den latenta EBV-infektionen malignifierar hos cirka 200’000 individer varje år. Risken att utveckla särskilda former av EBV-associerad cancer är hög i vissa regioner, t.ex. nasofarynxcancer i Sydostasien och Burkitts lymfom i Afrika. Det övergripande syftet med denna avhandling var att skärskåda det virala genuttrycket i tumörbiopsier och klarlägga funktionerna av de uttryckta generna.

Transkriptomdata från 615 tumörer med känd EBV-koppling och encellstranskriptomik av 63 nasofarynxprover undersöktes med avseende på EBV-RNA. De virala gener som visades vara uttryckta i störst omfattning på såväl vävnads- som enskild cellnivå var RPMS1 samt nyupptäckta transkript med samma slutsekvens som RPMS1, vilka vi valt att benämna BAREs. Andra virala gener, främst LMP1/BNLF2a/b och LMP2A/B/BNRF1, var uttryckta i mindre utsträckning och stora skillnader kunde observeras mellan individer. Encellssekvensering av B-lymfocyter isolerade från perifert blod taget från en patient med höga nivåer av EBV-DNA visade på ett mönster av viralt genuttryck som liknade det i EBV-positiva tumörer. De i tumörvävnad högst uttryckta EBV-generna RPMS1 och BAREs genomgick dessutom fullängdssekvensering på enskild molekylnivå varvid alla isoformer karaktäriserades med vårt nyutvecklade bioinformatiska verktyg FLAME.

Våra resultat visar att tillgängliga cellmodeller inte avspeglar primära tumörer beträffande viralt genuttryck och benägenhet till reaktivering. Vi utvecklade en *in vitro* pseudostratifierad nasofaryngeal epitelmodell i syfte att efterlikna en EBV-infektion i nasofarynx. En donatorberoende känslighet för EBV-infektion observerades och såväl latent som lytiskt genuttryck kunde påvisas

hos en enskild donator. Vidare kunde vi med encellssekvensering urskilja fullskalig lytisk EBV-infektion och därtill nedreglering av den cellulära transkriptionen i suprabasala celler.

Encellssekvensering av perifera EBV-transformerade B-lymfocyter indikerade att EBV inducerar proliferativa signalvägar. I vävnad från nasofarynxcancer existerar de EBV-transformerade epitelcellerna i en mikromiljö med lymfocytinfiltration och interferon. Karaktärisering av nasofaryngeala celler på enskild cellnivå visade att det virala uttrycket av RPMS1 och miR-BARTs som kodas i intronen främjar immunflykt genom nedreglering av interferonstimulerade gener. Detta fynd tyder på att EBV bidrar till tumöruppkomst på två principiellt olika sätt, dels genom omprogrammering av värdcellen och induktion av proliferation via EBNAs och LMP1, dels genom immunflykt förmedlad av RPMS1 och BNLF2a.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. **Yarong Tian***, Guojiang Xie*, Isak Holmqvist, Alan Bäckerholm, Sanna Abrahamsson, Jonas Carlsten, Ka-Wei Tang. **The landscape of Epstein-Barr virus expression in human cancer.**

Manuscript

II. Holmqvist I*, Bäckerholm A*, **Tian Y**, Xie G, Thorell K, Tang KW. **FLAME: long-read bioinformatics tool for comprehensive spliceome characterization.**

RNA. 2021;27(10):1127-1139.

III. Ziegler P, **Tian Y**, Bai Y, Abrahamsson S, Bäckerholm A, Reznik AS, Green A, Moore JA, Lee SE, Myerburg MM, Park HJ, Tang KW, Shair KHY. **A primary nasopharyngeal three-dimensional air-liquid interface cell culture model of the pseudostratified epithelium reveals differential donor- and cell type-specific susceptibility to Epstein-Barr virus infection.**

PLoS Pathogens. 2021; 17(4): e1009041.

IV. Alan Bäckerholm*, **Yarong Tian***, Isak Holmqvist, Guojiang Xie, Diana Vracar, Sanna Abrahamsson, Ka-Wei Tang. **Detection of latent Epstein-Barr virus gene expression in single-cell sequencing of peripheral blood mononuclear cells.**

Manuscript

CONTENT

ABBREVIATIONS

BART	BamHI-A region Rightward Transcript
BCR	B Cell Receptor
BL	Burkitt's Lymphoma
cpm	Counts Per Million
eBL	endemic Burkitt's Lymphoma
EBV	Epstein Barr Virus
GAC	Gastric Adenocarcinoma
HL	Hodgkin's Lymphoma
ISH	In situ hybridization
kb	kilobase
LCL	Lymphoblastoid Cell Line
lncRNA	long non-coding RNA
miR-BART	microRNA in BamHI A Rightward Transcript
miRNA	microRNA
NGS	Next Generation Sequencing
NPC	Nasopharyngeal Carcinoma
PBMC	Peripheral Blood Mononuclear Cell
PDX	Patient-derived xenograft
ppm	Parts Per Million
PTLD	Post Transplant Lymphoproliferative Disorder
RT-qPCR	Reverse Transcription quantitative PCR
sBL	sporadic Burkitt's Lymphoma
TCR	T Cell Receptor
tpm	Transcripts Per Million

1 INTRODUCTION

Epstein-Barr virus (EBV), a member of the Herpesviridae family, also called Human gammaherpesvirus 4 (HHV4), is transmitted through saliva. EBV was first discovered by electron microscope in a cell line of cultured Burkitt's lymphoblasts [1]. An infection with EBV during childhood is usually asymptomatic or mild. A delayed primary infection to adolescence or adulthood may result in infectious mononucleosis (IM). Following the primary infection – no matter if it passed unnoticed or with characteristic clinical manifestations – the virus establishes a latency stage in B-lymphocytes. The vast majority of adults around the world carry EBV asymptomatically, however, a small fraction of infected people develop different types of EBV –associated diseases under certain conditions. EBV has been shown to be associated with several types of human tumors originating from lymphocytes or epithelial cells (Table 1). It is also suggested that EBV infection increases the risk of development of multiple sclerosis (MS) [2, 3] and chronic fatigue syndrome (CFS) [4]. There is currently no specific treatment against EBV-associated diseases available.

Table 1. *EBV and cancer*

Cancer	EBV prevalence	Reference
Burkitt's lymphoma (BL)	95-100% in endemic (eBL) 15-85% in sporadic (sBL)	[5] Howley et al. 2021
Hodgkin's lymphoma (HL)	20-80%	
post-transplant lymphoproliferative disorder (PTLD)	>90% tumors developing within one year after transplantation	
gastric adenocarcinoma (GAC)	10%	[6] TCGA-STAD 2014
nasopharyngeal carcinoma (NPC)	100%	[7] Chen et al. 2019

New cancer cases attributable to EBV infection yearly were estimated to be 160,000 in a worldwide incidence analysis in the 2018 GLOBOCAN project [8]. The association was further confirmed based on next generation sequencing (NGS) datasets from tumor biopsies (RNA; PCAWG, DNA) [9, 10]. More than 55 years of effort have been put into the biological life cycle of EBV and its role in the causation of cancer. The coevolution of EBV with its host and their intricate and complex interactions during its life-long persistence complicate the matter.

Although many types of therapies are under development [11], specific and effective therapeutics is still unavailable. In recent years, a considerable amount of work on vaccines has been done, including viral glycoproteins (gp350, gp42, gH/gL, Appendix B) as preventive measures for PTLD, infectious monnucleosis, and multiple sclerosis. At the end of 2021, the clinical trial NCT05164094 on mRNA EBV vaccine (mRNA-1189) was initiated [12]. Also, therapeutic vaccines for NPC expressing EBNA1, LMP1 and LMP2A, and T-cell therapy for EBV-associated diseases are in development [13].

This thesis will attempt to comprehensively elucidate the EBV transcriptome in different samples, including bulk tissues and single-cells, from *in vitro* to *in vivo*. The data was generated from multiple sequencing platforms, traditional bulk RNA sequencing (RNA-seq), single-cell RNA sequencing (scRNA-seq) and single-molecule sequencing. Based on the detailed landscape of viral expression in cancer cells, the viral genes may in the future serve as drug targets for the treatment of EBV-associated diseases.

1.1 THE EBV GENOME

The EBV genome, double-stranded DNA, approximately 170 kilobases (kb) in length, is packed into capsid proteins in a linear form in the virion, and upon entry into the host nucleus, it is circularized and chromatinized as an episome in latency phase [14]. The EBV genome rarely integrates into the host genome, but such events have been described in tumors [15, 16]. The EBV reference genome NC_007605.1 (NCBI RefSeq) was updated in 2018 and is a hybrid between the B95-8 and Raji strains (Figure 1). Besides the repeat sequences, the EBV genome also harbors a wealth of genes involved in the different life cycle stages of the virus. A detailed gene list can be found in Appendix A. Forty-three genes are shared with other viruses in the Herpesviridae family [17, 18].

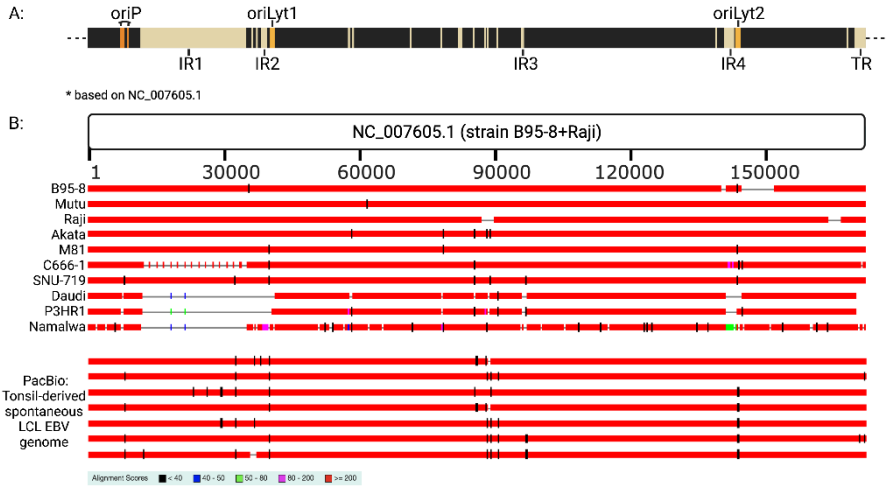


Figure 1. *Illustration of the EBV genome architecture. A: The bars in beige are repeat regions, IR1-4 are four main internal repeats, and TR stands for terminal repeat. The origins of replication are highlighted in orange, oriP is considered as the latent (plasmid/episome) replication origin, and oriLyts are used in the lytic life cycle. B: Comparison of EBV genomes from different cell lines by multi-alignment (BLASTN). The bottom seven genomes were generated by long-read sequencing. The horizontal bars are colored and coded by alignment score.*

EBV genomic variations have been suggested to be correlated with the incidence of some diseases [19-21]. One mutation in RPMS1 and two in BALF2 were found to be correlated to NPC cases in southern China

[22, 23]. EBV genetic variations in different populations have recently been cloned and sequences from tonsil-derived lymphoblastoid cell line (LCL) have been generated by using long-read sequencing [24]. However, there are genomic variations within the same individual, and it has been shown that EBV strains in saliva differ from the clones in tumor lesions in gastric cancer [25]. A multi-alignment of frequently used EBV strains in different cell lines compared to the NC_007605.1 reference using BLASTN (default settings) is illustrated in Figure 1B. Some of the EBV strains have been sequenced several times [26]. The EBV strain is one of the main factors that needs to be taken into consideration for the design of vaccines and precision medical treatment, as well as for the choice of model system for research.

1.2 HOST PERTURBATIONS IN EBV POSITIVE TUMORS

Host epigenomic, genomic and transcriptomic perturbations can be found in the different types of EBV-associated malignancies (Table 2). Somatic mutations in the most common EBV-associated neoplasms NPC and GAC [27], eBL and sBL [28], have been well characterized by NGS (whole genome sequencing and whole exome sequencing). Tumor cell microdissection enrichment followed by whole genome sequencing of NPC samples further uncovered that the host changes of immune pathways were enriched in the EBV-positive tumors in comparison with blood samples [29]. This indicates that EBV might be able to establish latency in cells with a perturbed immune response, or that EBV induces these changes. Generally perturbed pathways can be observed in the different cancers. EBV appears to be able to downregulate hurdles for proliferation (e.g. TP53), but still needs to overcome specific compensatory mechanisms (e.g. CDKN2A) in order to malignify.

Table 2. *The landscape of host perturbations in EBV-positive tumors compared with EBV-negative tumors in each cancer type*

	EBV+ NPC	**EBV+ GAC**	**EBV+ BL (eBL/sBL)**
Epi	CIMP*; 3p21.3**; 6p21.3 (immune genes); 9p21; Tumor suppressor: CDKN2A (p16, chr9p21.3) methylation; CDH1; PTEN	CIMP; CDKN2A promoter methylation; lack of MLH1 hypermethylation	
DNA	[**gain**] 1q, 11q, 12p, 12q, 17q; CCND1(11q13); MYC(8q24) [**loss**] 3p, 9p, 11q, 13q,14q,16q; CDKN2A; MTAP;	[**gain**] 9p24.1, 3q, 7, 20 [**loss**] 18q	IG-MYC translocation t(8:14), t(8:17) 80/20: IGH (chr14) /IGL (22) or IGK (2)

	TGFBR2; Type I IFNs [**mutations**] relative low mutation load; ERBB2/3; PIK3CA; PTEN; N/KRAS ARID1A; MHC-I NF-kB negative regulators (TRAF3, CYLD, NFKBIA, NLRC5) TP53 (low TP53 mutations compared to other head and neck cancers)	[**mutations**] PIK3CA; PTEN ARID1A; BCOR; JAK2; NOTCH1 less TP53 mutation	[**mutations**] somatic hypermutation; fewer driver mutation; TCF3; ID3; CCND3 ARID1A; CDKN2A; PVT1 promoter; PAX5; less TP53 mutation
RNA	CD274 (PD-L1)↑; NF-kB pathway↑; cyto-/chemo-kines↑; IFN-induced genes↑; CDKN2A↓; TGFBR2↓; MHC-1↓;	PD-L1/2↑; Immune cell signaling↑; DNMT1↑; PAX4↓; EGF↓;	AICDA (eBL)↑; MDK↓;

** CIMP: CpG island methylator phenotype; ** Human chromosomes.*

↑ increase/activation; ↓ decrease;

This table is based on these references: Han et al. 2021 [27] (review on NPC and GAC), Bruce et al. 2021 [29] (NPC, China), Lin et al. 2014 [30] (NPC, Singapore), Wong et al. 2021 [31] (review on NPC), TCGA-STAD 2014 [6] (GAC, USA), Chen et al. 2021 [25] (GAC, China), Grande et al. 2019 [28] (BLGSP), Schmitz et al. 2012 (BL) [32].

1.3 EBV GENES

EBV encodes more than 100 genes across the entire genome [33]. The genes in the first fragment till IR2 in Figure 2 only encode rightward transcripts, the other regions contain bidirectional genes. Nomenclature of EBV genes is based on BamHI-digested fragments, and the classification is dependent on the sequential activation post-infection within hours [34], and grouped into immediate early (IE, alpha), early (EL, beta) and late (LL, gamma) lytic genes (color-coded in Appendix A).

EBV can also transcribe several different types of non-coding RNAs (Figure 2). The viral long non-coding RNA RPMS1 is a 4kb polyadenylated transcript [35]. A potential long non-coding transcript from BHLF1 has also been suggested [36]. Two EBV-encoded small non-coding RNAs (EBER1 and EBER2) are transcribed by host cellular RNA polymerase III, and play roles in EBV latency maintenance [37]. Other small non-coding RNAs are also found in the EBV genome, four microRNAs (miRNAs) in the BamHI H rightward frame (miR-BHRF1) and 40 miRNAs in the BamHI A region (miR-BART), all verified and recorded in miRBase [38]. In addition, circular RNAs have been detected in latency (EBNA/RPMS1) and during reactivation (BHLF1/LMP2) [39-41]. It was suggested that miR-BART11 and 17-3p [42], circBART2.2 [39] from the RPMS1 gene could upregulate PD-L1 to facilitate immune escape, while miR-BHRF1-2 could reduce PD-L1/L2 [43]. Interestingly, both the lncRNAs, miRNAs and three circular RNAs are adjacent to the origins of viral DNA replication (Figure 2). Other herpesviruses also express spliced transcripts and non-coding RNAs during latency (miRNA [44], circRNA [40]).

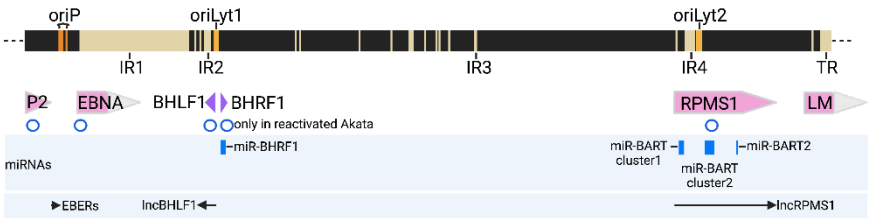


Figure 2. *The distribution of non-coding RNAs found in EBV. EBV NC_007605.1 annotation was used as the reference for plotting. Blue circle: circRNA; Blue rectangle:*

microRNA; Pink arrows represent latency genes, purple unknown. The direction of the transcripts is indicated by the arrows.

In order to uncover how EBV contributes to oncogenesis in the host cells, it is necessary to characterize the expressed viral transcripts, especially in cancer cells. All types of EBV positive tumors contain mainly latent EBV infection. EBV latency programs are divided into (1) latency 0: only non-coding RNAs; (2) latency I: genes in latency 0 with the addition of EBNA1; (3) latency IIa: addition of LMP1 and LMP2A/B; IIb; addition of EBNAs and BHRF1; (4) latency III: full range expression of LMPs, EBNAs, and BHRF1 [45]. BL was classified as latency I in the literature, and NPC was latency II(a). Our findings from primary tissues in **Paper I** do not support the proposed latency programs.

The role of EBV in oncogenesis has been studied *in vitro* using cell models originating from tumor-derived cell lines or EBV-transformed LCL. These models have been essential for the study of virus-host interactions during transformation and reactivation. In recent years, an increased amount of omics-data has been generated from primary EBV-infection of lymphocytes (transcriptome [46], proteome [47], epigenome/transcriptome/metabolome [48]) and reactivation (epigenome [49]). The following two parts aim to summarize the dynamic EBV expression changes during LCL establishment and EBV positive cell line reactivation.

1.3.1 EBV LIFE CYCLE

Once EBV attaches and enters the host cell, it could undergo a lytic cycle, by sequentially activating the viral lytic genes (immediate early, early, late lytic) and repurpose the host machinery for viral production. The major events include: (1) The EBV envelope is fused with the cellular membrane by using the viral glycoproteins for binding with receptors on the host cell. After fusion the capsid is released into the cytosol (Figure 3); (2) The EBV capsid is transported to the nucleus and injects the viral genomic DNA; (3) Both new viral capsid assembly and EBV genome encapsidation occurs in the nucleus; (4) After nuclear egress, the capsid and associated tegument proteins are enveloped with the help from endoplasmic reticulum (ER), Golgi apparatus; (5) Mature EBV virion is transported to the plasma membrane for release [45].

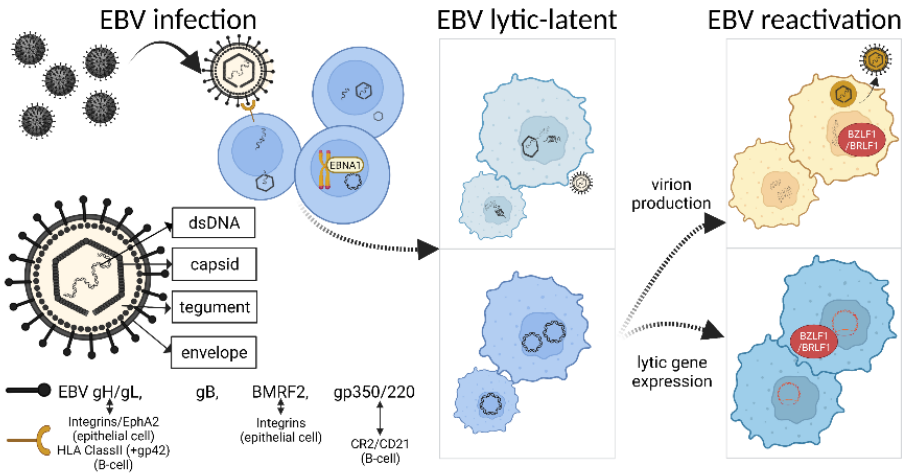


Figure 3. *EBV structure and its life cycle. EBV interacts with its host cell through different glycoproteins(black) on its envelope binding with corresponding host receptors (Left bottom, yellow). After the viral capsid enters the host cell cytosol, EBV could undergo lytic cycle in some cell, or establish latent stage under intracellular factors restriction (Middle). EBV latency could be reactivated to express some lytic gene or fully lytic to virion secretion with the help of the immediate early gene BZLF1 (Right).*

EBV could also enter the transcriptionally latent stage in the host cell with only a few viral protein-coding genes being expressed. In latency phase, the EBV genomic DNA is highly methylated and maintained in the nucleus as an extra chromosome. During latency, the DNA

replication of the EBV genome is in synchrony with the host genome assisted by EBNA1 [50]. EBV positive cell lines can reactivate the EBV lytic cycle as a response to multiple types of chemicals [35, 54, 69]. EBV reactivation is frequently observed in EBV associated tumors. For example, the EBV DNA load in the plasma samples of NPC and PTLD patients is markedly increased [51-53], and antibody titers against EBV lytic protein are higher, for example, EA IgG [54].

1.3.2 KINETICS OF EBV EXPRESSION DURING PRIMARY INFECTION

To dissect the EBV gene expression in cell cultures at the transcriptomic level, several techniques could be utilized, such as quantitative reverse transcription PCR (RT-qPCR), EBV probe-based array and RNA-seq. EBV transcriptomics in CD19+ B-lymphocytes have been well deciphered by RNA-seq [55, 56] at different time points post after *de novo* infection (day 0/2/4/7/14 etc.). Naïve B-lymphocyte EBV-infection has also been monitored during one week using RNA-seq, and transcriptional perturbation was seen within 24h [57]. Similarly, mass spectrometry proteomics have been used to follow the EBV-infection [47]. The relative abundance of EBV genes at protein (left panel) and RNA (middle) levels are presented in Figure 4. Multiple EBV lytic genes were detectable. The lytic EBV expression pattern in LCL has also been addressed in **Paper I** and **IV**. However, LCL is conventionally classified as latency III.

B-lymphocyte heterogeneity is an issue when studying the role of EBV in LCL. EBV-positive B lymphocytes *in vivo* have been shown to be primarily memory B lymphocytes (**Paper IV**), however, LCLs are mainly generated from PBMCs (peripheral blood mononuclear cells) or CD19+ B lymphocytes. During the first week of the generation of LCLs from primary human CD19+ B lymphocytes, EBV-infected B-lymphocytes first undergo hyperproliferation, and then growth arrest or continuous proliferation [58]. LCLs that continuously proliferate have normal karyotypes, but before 150 passages the cells either undergo proliferative crisis and die or immortalizes with a strong telomerase activity and become aneuploid [59]. LCL-transformation also circumvents the immune response mediated by CD8+ and/or CD4+ T-lymphocytes [60, 61].

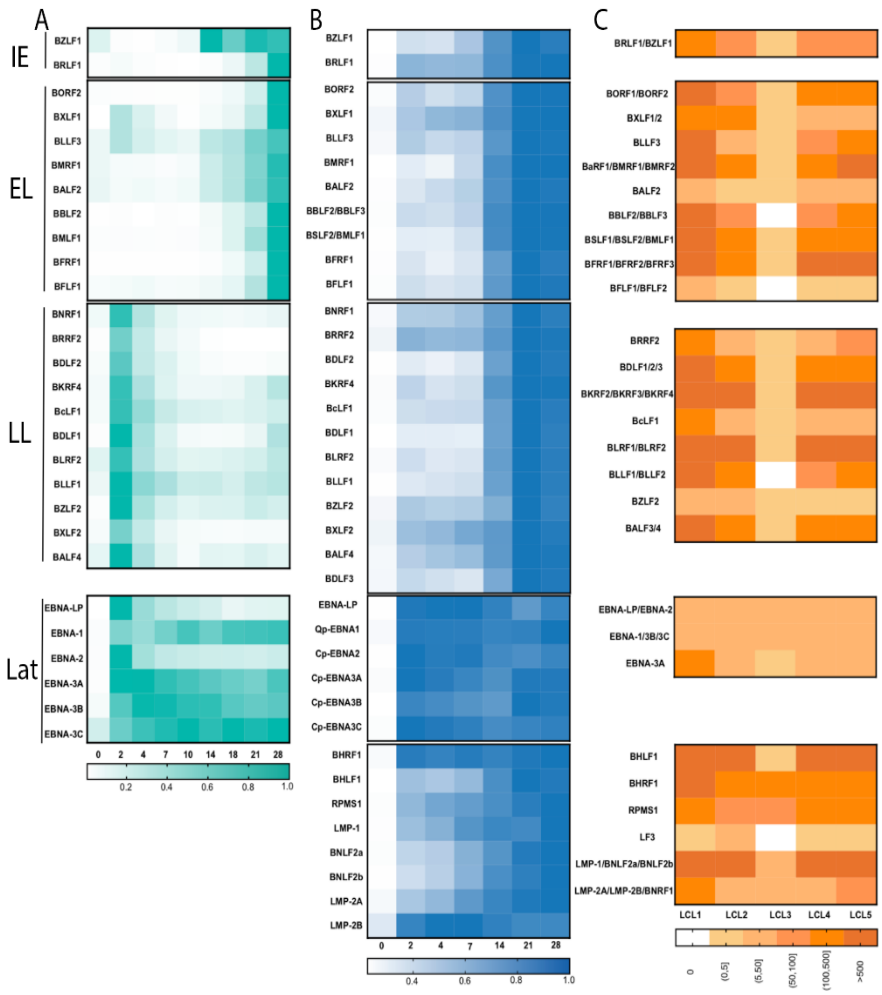


Figure 4. *EBV gene expression during primary in vitro infection. A: Proteomics of EBV proteins in different time points after primary infection (days), genes were grouped and ordered by IE (immediately early), EL (early lytic), LL (late lytic), Latent (Lat) and other genes of interest. The relative abundance of each protein for the different days post-infection is shown ([47] Wang et al. 2019 Cell Metab.). B: The relative abundance of EBV gene RNA levels. C: Cpm-values of EBV genes in LCL single-cell sequencing datasets processed as bulk.*

1.3.3 EBV TRANSCRIPTION IN REACTIVATION

The latent EBV can reactivate and enter the lytic cycle and produce virus particles. Reactivation can also be aborted (abortive lytic), in this case the immediate-early genes BZLF1/BRLF1 are induced, but the subsequent transcription of early and late lytic genes are not initiated and no virus particles are produced (Figure 3) [62]. EBV-positive cell lines have a variable ability to be reactivated and enter the lytic cycle. For example, the BL cell line, Raji, lacks the viral DNA binding protein BALF2 and cannot produce viral particles. By adding BALF2 to Raji cells a complete reactivation can be induced [63]. Although some cell lines are not able to enter the viral productive stage, many lytic genes can be triggered to be transcribed *in vitro* [64]. EBV reactivation has been considered as a potential therapeutic approach for the treatment of both NPC [68], and EBV-positive lymphoma. Various drugs as well as T-lymphocytes have been tested to induce a “kick and kill” scenario for anti-viral treatment [69].

The path from latency to reactivation and viral production is complicated. A CRISPR/Cas9 screen for genes essential for EBV reactivation in P3HR1 identified MYC, which is also essential for the development of endemic BL [65]. Here are some examples of how EBV reactivation has been investigated. EBV gene expression during reactivation has been profiled in various cell lines including JSC-1, Akata and Raji [33]. Akata EBV transcription in IgG-induced reactivation has also been profiled using long-read sequencing [35], and the early-late genes sequential activation has been described before [34]. Also, the global Akata EBV bidirectional transcription during reactivation was descirbed based on CAGE-seq [66]. Furthermore, RNA polymerase mapping in Mutu I during latency and reactivation was characterized using Precision nuclear Run On followed by deep Sequencing (PRO-seq), which indicated that RNA polymerase paused at CTCF binding sites on the EBV genome during reactivation [49]. The EBV genome template for transcription switched from latency to reactivation has been reviewed [67]. However, the conclusions of these studies are not suitable to be generalized on the reactivation mechanism because (1) various starting cell lines and EBV strains were investigated; (2) different chemicals were utilized for inducing lytic reactivation, which might cause cascade reactions of different pathways affecting lytic EBV gene expression; (3) RNA-seq data on different time point post inducing are not available.

1.4 MODELS FOR EBV RESEARCH

Cell lines are the most frequently used *in vitro* models for EBV research, of which the BL-derived lymphocytes Namalwa, Daudi, P3HR1 and Raji were included in the CCLE project (DepMap) and well characterized using multi-omics. In Table 3, a collection of EBV-positive cell lines is listed. B95-8 is an EBV positive cell line constructed in marmoset B lymphocytes using IM-derived EBV, and its genome sequence was the first assembled and is frequently used as the EBV reference genome. M81 is an NPC-derived EBV-infected marmoset lymphocyte cell line and is considered to harbor an EBV-strain more prone to reactivate than the B95-8 strain. C666-1 is one of the few NPC cell lines available. It is derived from xenograft in nude mice and cannot undergo lytic reactivation.

Compared to EBV positive BL cell lines, the EBV-positive NPC cell lines are not easy to establish directly from the biopsy [70] and requires a selection step in patient-derived xenografts (PDXs). Both the latent EBV cell line C666-1 [71] and the first reactivable C17 were derived from EBV-positive PDXs in nude mice [70]. NPC43, harboring variably EBV-copies at different passages, was cultured directly from an NPC biopsy using ROCK inhibitor in order to inhibit lytic reactivation [64]. The gastric cell line SNU719 EBV expression patterns is like most BL cell lines different from tumor biopsies (**Paper I**) [72]. Thus, most of the EBV-positive cell lines cannot accurately portray *in vivo* conditions, which is a major challenge for the research field.

Table 3. *The well-characterized cell lines available for EBV research*

Name	Cell/ EBV Type	EBV copy	ID of EBV genome	Length of EBV genome	Cellosaurus No.	Disease
B95-8	marmoset B*/ I	>800	V01555.2**	172281	CVCL_1953	IM
Namalwa	B/ I	1-2	LR813082.1	152359	CVCL_0067	BL
Daudi	B	~100	LN827545.1	172089	CVCL_	BL

(negative MHC I)					0008	
P3HR1 (non-transforming)	B/ II	>800	LN827548.2	172083	CVCL_2676	BL
Raji	B/ I	~60/500	KF717093.1	166182	CVCL_0511	BL
Akata	B/ I	~20	KC207813.1	171323	CVCL_0148	BL
Mutu	B		KC207814.1	171687	CVCL_7202	BL, t(8:14)
M81	marmoset B/I		KF373730.1	176041	/	NPC
C666-1	E*	~12	KC617875.1	171317	CVCL_7949	NPC
SNU719	E/I	~800	AP015015.1	169425	CVCL_5086	GAC

*B stands for lymphocyte and E for epithelial cell [73]. **The ID of the EBV genome is the accession number in NCBI. The genome ID prefix stands for the source of the storage database.

EBV-transformed LCL has been utilized for generating donor specific genetic materials, because of its low propensity for acquiring mutations. LCL can be grouped into spontaneous LCL from peripheral blood or tissue, and *in vitro* transformed LCL according to the procedure for generation [74]. The LCLs generated for EBV research were mainly derived from enriched CD19+ B lymphocytes, rather than PBMC. The EBV strain used for transformation often comes from the B95-8, Akata or Mutu cell lines. However, the EBV expression in LCLs are quite different from EBV positive B lymphocytes *in vivo* (**Paper I and IV**). The population doubling rate of LCLs is dependent on the accumulation of host changes (such as telomerase activity and karyotype), which are the key factors to distinguish them between “transformed” and “immortalized” [59].

For epithelial cells, human adult stem cells-derived airway organoids have been widely used for the study of respiratory viruses [77]. EBV-associated nasopharyngeal carcinoma most often originates from a specific location of the nasopharyngeal lateral wall, where the epithelium is pseudostratified and lymphocyte-rich (**Paper III**). EBV has been shown to preferentially induce the lytic cycle when infecting stratified differentiated epithelial cells [78]. Therefore, to emulate the EBV-infection that could develop into NPC, the air-liquid interface culture model of the human airway epithelial cell has been developed to study the interaction between EBV and the host [79, 80].

More and more evidence indicate that the role of EBV in cancer might be through its contribution to immune evasion, which points out the need for an *in vivo* model for EBV research. Some NPC patient-derived xenografts have been constructed for such studies including x666, C15/17/18 and other four new PDXs [64].

Also, the relationship of immune control and the EBV infection have been studied in humanized mice models [81, 82]. It has been suggested that EBV miRNA attenuated T-cell mediated immune control through dampening antigen recognition [83]. EBV primary infection was also modeled in other animals, such as tree shrews [84] and rabbits [85]. Moreover, many researchers have analyzed primary biopsies, including liquid biopsies and solid tissues, which better reflect the *in vivo* physiological events of EBV in its natural host.

1.5 RNA SPECIES IN MAMMALIAN CELLS AND THEIR PROPERTIES

There are multiple species of RNA in the mammalian cells transcribed by different types of RNA polymerases (Figure 5). The majority of the RNAs in human cells are non-coding transcripts. Protein-coding regions of the genome only account for approximately 2% of the human genome and only around 15% of the protein coding genes are druggable [86]. The different RNA species possess different half-lives [87], and genome-wide individual RNA half-life (decay) was deciphered in human K562 and HeLa cells by combining PRO-seq and RNA-seq [88]. The non-coding RNAs have been demonstrated to play roles in transcription and post-transcriptional regulation as well as scaffolding and condensates [89].

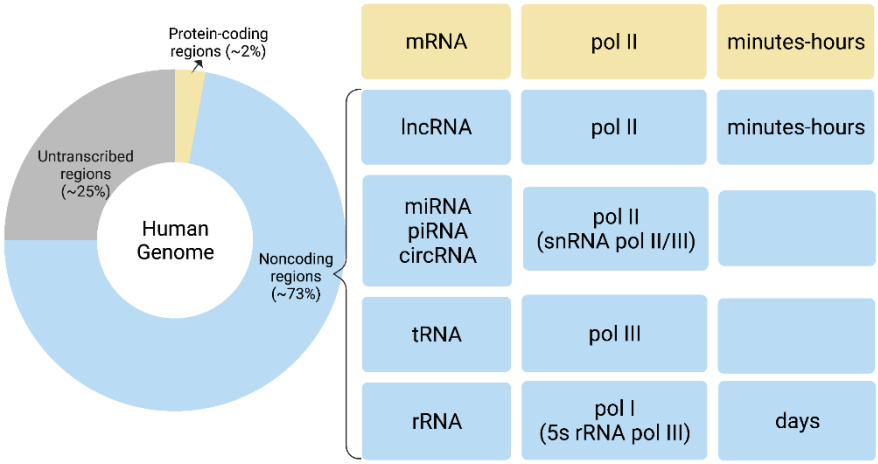


Figure 5. *RNA species in a human cell. Left circle: RNA in the human genome; Right columns: RNA species, RNA polymerases, RNA half-life. RNA: ribonucleic acid; mRNA: messenger RNA; lncRNA: long non-coding RNA; miRNA: microRNA [90]; piRNA: PIWI-interacting RNA; circRNA: circular RNA; snRNA: small nuclear RNA; tRNA: transfer RNA; rRNA: ribosomal RNA.*

1.6 LIBRARY CONSTRUCTION AND SEQUENCING TECHNOLOGY

Most transcriptomic datasets have been generated from oligo(dT)-enriched libraries and thus limited to the polyadenylated RNA. Viral genomes are compact and contains multiple intersecting and co-terminating genes. The direction and coverage are therefore important for annotating the short reads correctly to the EBV genome. Therefore, stranded libraries containing the information on transcript direction would be better suited for viral transcriptome analysis. More recently, the rRNA depleted library preparation protocols have become more common, containing both mRNAs and small RNAs. Furthermore, specific workflows have been developed and utilized for miRNA and circRNA sequencing (Figure 6). However, the biogenesis and turnover of the RNAs are dynamically changing [91]. The transcription initiation, elongation and cleavage should be taken into consideration to get an accurate quantification [92], as well as the stochastic nature of transcription [93].

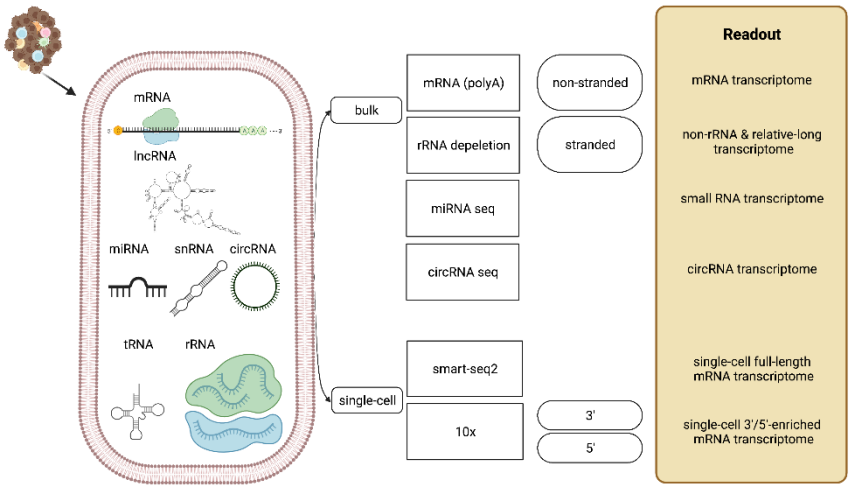


Figure 6. *Transcriptome profiling techniques and corresponding readouts.*

1.7 SINGLE-CELL RNA SEQUENCING

Due to the variable level of non-malignant cells in tumor lesions, the specific properties of the malignant cells are difficult to discern in the bulk RNA-seq data of tumor biopsies. Single-cell RNA sequencing (scRNA-seq) technology can disentangle the tumor microenvironment at a single-cell resolution. In recent years, there have been multiple studies published using single-cell sequencing of EBV-associated samples, however the EBV gene expression was not reported (Table 4). We analyzed NPC scRNA-seq dataset, and compared EBV-positive malignant cells with healthy epithelial cells (**Paper I**). The malignant cells were identified using inferred copy number variation analysis based on the single-cell transcriptome comparison [94, 95]. Moreover, the differentiation of the cells could be predicted using the RNA velocity [96] combined with the trajectory inference [97]. All the scRNA-seq workflows rely on reverse transcription of oligo(dT) (indexed for 3' transcriptome) captured mRNA followed by a template switch using an indexed 5' primer. Smart-seq is similar, but the whole transcript is sequenced. However, the capture rate of single cell transcriptome technologies was limited to 20-40% compared to bulk RNA-seq [98], which means that the scRNA-seq only depict the most abundant mRNA in each cell.

Table 4. *The studies using scRNA-seq of EBV-positive samples*

No.	Sample type	scRNA-seq	ID of Dataset	Reference
1	LCL (CD19+ B lymphocytes)	10x 3' v2	GSE158275	[99] SoRelle et al. 2021
2	LCL	10x 3' v2	PRJNA508890/ PRJNA521545	[100] Osorio et al. 2019
3	NPC	10x 3' v2	CNP0000428	[94] Chen et al. 2020
4	NPC	10x 3' v2	HRA000087	[101] Jin et al. 2020
5	NPC/PBMC	10x 5' TCR VDJ	HRA000159/GSE 162025	[102] Liu et al. 2021
6	NPC	10x 5' TCR/BCR VDJ	GSE150825	[103] Gong et al. 2021
7	NPC	10x 3' v2	-	[104] Zhao et al. 2020
8	GAC	10x	HRA000051	[105] Zhang et al. 2021
9	LCL	10x 3'	PRJNA794826	[106] Bristol et al. 2022

* TCR/BCR: T/B cell receptor.

1.8 EBV MIR-BARTS

EBV miRNAs include miR-BHRF1s and miR-BARTs (Figure 2 and 7), and the transcription of respective pri-mRNAs are in close proximity to oriLyt1 and oriLyt2 respectively. The miRNA transcriptome analysis based on miRNA sequencing in EBV-associated tumors have been reported in GAC [6] and BL [28]. The viral miRNAs miR-BART6-3p, miR-BART7-3p, and miR-BART10-3p were in the most abundantly expressed EBV miRNAs within both tissue types [107]. EBV miRNAs in NPC biopsies and NPC cell line C666-1 were demonstrated using multiplexed stem-loop RT-PCR, and found all miR-BARTs were expressed at high levels, but not miR-BHRF1s [108]. The same miR-BARTs were detected in the plasma of patients with GAC, BL and NPC [109], NPC tissues [110], and NPC PDXs [111].

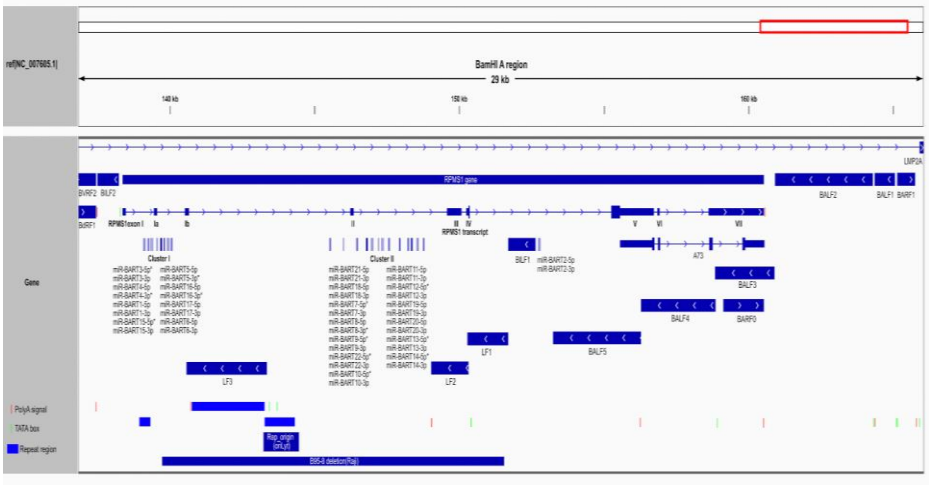


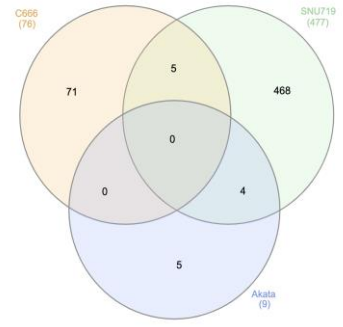
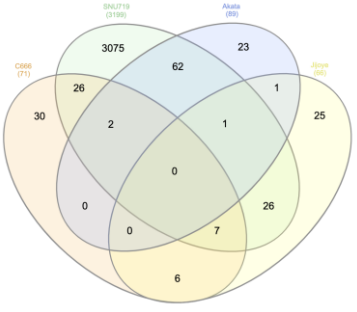
Figure 7. *Illustration of BamHI A region (IGV). CDS, white arrows.*

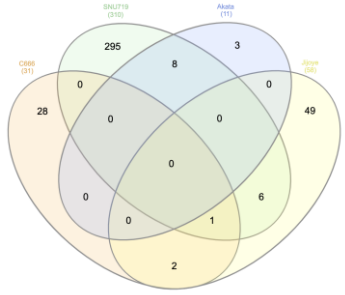
EBV miRNAs interact with their target mRNA primarily through Ago2. The miR-BARTs and their targets have been investigated using various techniques by Ago enrichment [112]. However, because of the diverse EBV expression in the cell models as well the EBV strain and host background very few target genes overlapped between the four published studies (Table 5 and 6).

Table 5. *EBV miRNA target screening*

Reference	Method	Cell and antibodies	Number of targets
[113] EMBO, Riley et al. 2012	HITS-CLIP	Jijoye (BL) anti-Ago mAb 2A8 or 11A9	925 Abundant EBV miRNAs: BART10-3p, BART13-3p, BART4-5p, BART7-3p
[114] Plos Path., Skalsky et al. 2012	PAR-CLIP	EBV-B95-8 LCLs: EF3D-AGO2, a FLAG-tagged Ago2 LCL35, and LCL-BAC LCL-BAC-D1 (miR-BHRF1-1 ko) LCL-BAC-D3 (miR-BHRF1-3 ko) mAb: anti-FLAG and anti-Ago2	540
[115] Plos Path., Kang et al. 2015	PAR-CLIP	C666-1 (NPC), Ab against all four human Agos	1254 Abundant EBV miRNAs: BART2-5p, BART9-3p, BART19-3p BART22
[107] Plos Path., Ungerleider et al. 2021	CLASH	Akata (BL), SNU719 (GAC) pan-AGO antibody	SNU719: 16976, Akata: 534,

Table 6. *Potential target mRNAs of the miR-BART6/7/10*

miR-BART	Found in two cell lines	Found in three cell lines
miR-BART6  C666(76) SNU719(477) Akata(9)	[C666] and [SNU719]: STEAP3, HSPH1, TNPO2, KIAA1217, CTNNB1 [SNU719] and [Akata]: CD74, ZMAT3, DDX58, GIGYF1	
miR-BART7  C666(71) SNU719(3199) Akata(89) Jijoye(66)	[C666] and [SNU719]: KIAA1522, MSI2, IL23A, IL27RA, POLR1B, PPIL3, STK35, UBXN7, UBA6, P4HA2, JARID2, NONO, SESN2, POGZ, CDC123, GSTO1, PPP1CC, IFT88, RBBP6, RALY, SHROOM3, G3BP1, PRLR, IARS, ZFP91, HIPK2 [Akata] and [Jijoye]: RASGRP3	[C666] and [SNU719] and [Akata]: COX5A, LRRC58 [SNU719] and [Akata] and [Jijoye]: YWHAG [C666] and [SNU719] and [Jijoye]: RPL5, CSDE1, ZNF180, PATZ1, CLCN3, ING3, ZCCHC7

miR-BART10  C666(31) SNU719(310) Akata(11) Jijoye(58)	[SNU719] and [Akata]: ATP6V0E1, TPR, FAM111A, RSPRY1, TMED8, SNU13, DARS, CREBBP [SNU719] and [Jijoye]: SMCHD1, AP1S3, TNFRSF10B, PPP6C, RBM26, TFRC [C666] and [Jijoye]: ANKRD11, ACSL4	[C666] and [SNU719] and [Jijoye]: MEX3C

2 AIMS

The aim of this thesis was to comprehensively characterize the EBV transcriptome in primary EBV-positive tissues and provide new viral targets for future therapy.

Paper I - Determining the EBV expression in tumor biopsies at both bulk and single-cell level and to understand the function of the EBV RNA in tumors.

Paper II - Characterization of the splice variants of RPMS1 using single-molecule long-read sequencing.

Paper III - Establishing a *de novo* infection in pseudostratified nasopharyngeal epithelial cells and characterizing the infected cells at single cell level.

Paper IV - Establishing single-cell bioinformatic methods and workflow for identifying and characterizing EBV in B-lymphocytes from peripheral blood.

3 MATERIALS AND METHODS

3.1 Cells and patient samples

The Burkitt's lymphoma B-cell lines Namalwa, Daudi and Akata, the NPC cell lines C666-1 and HK1, were all maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum and cultured at 37^{o}C with 5% CO_2. The other Namalwa-derived cell lines were kept in the same medium with additional G418 (and puromycin).

In **Paper I**, there were three Namalwa-derived cell lines created using Tet-On 3G system (Takara), they were all based on a stable Tet3G expressing cell line with pCMV-Tet3G selected by G418 screening; (1) pTRE is a cell line with empty pTRE3G-BI-mCherry, in which mCherry can be induced by doxycycline (Dox); (2) pTRE-RPMS1 (pTRE3G-BI-mCherry-RPMS1lncRNA) is the second cell line with a bidirectional TRE3G promoter encoding both the long non-coding RNA RPMS1 and mCherry which can be induced in response to Dox; (3) Pro-Re is the cell line in which the promoter of RPMS1 gene was replaced by the bidirectional promoter TRE3G-BI-mCherry using CRISPR/Cas9. The puromycin gene was inserted in these cell lines by co-transfection of a 1.8kb linear dsDNA fragment. Therefore, these cell lines were kept in RPMI medium with both 100 μg/ml G418 and 2 μg/mL puromycin.

The CRISPR vector px458-RS2-29 used for the establishment of the Pro-Re cell line contained two guide gRNAs [116]. The guide RNAs were designed at 138265-138284, 138332-138351 of the EBV genome NC_007605.1. The Namalwa cells were chosen because they contain an integrated EBV genome. The RPMS1 promoter in these cells were replaced with the TRE3G-BI-mCherry construct by homology-directed repair. This construct contained a bidirectional promoter encoding mCherry flanked by a 300-500 fragment containing the complementary sequence of the insert site in the EBV genome.

Paper II presented the single-molecule long-read sequencing characterization of RPMS1 splice variants in the NPC cell line C666-1, the Burkitt's lymphoma cell line Daudi, and a GAC biopsy. The GAC

samples were punch biopsies taken directly after resection and subsequently snap frozen.

The cell cultures used in **Paper III** included HK1 and Akata. HK1-EBV and rAkata harbor recombinant EBV with pSV40-EGFP inserted into the BXLF1 gene. The recombinant Akata EBV strain was modified by inserting EGFP cassette (under SV40 promoter) into BXLF1 loci [46, 56, 75, 76]. The pseudostratified nasopharyngeal epithelium (pseudo-ALI) cultures were generated from primary nasopharyngeal cells. The primary cells were collected from cytobrush scraping of the donors' nasopharynx. The cells were cultured, expanded and reprogrammed on 3T3-J2 feeder fibroblasts with the presence of ROCK inhibitor, and differentiated into pseudo-ALI in an air-liquid interface.

Blood samples used in **Paper IV** were collected at the Sahlgrenska University hospital. For one sample, the PBMC was isolated from whole blood using Lymphoprep™ (Stemcell). The primary B-lymphocytes were enriched using Dynabeads™ Untouched™ Human B Cells Kit (Invitrogen) and subsequently prepared for scRNA-seq according to the protocol CG00039 Rev C. Five published LCL scRNA-seq datasets were included for reanalysis (Table 7).

Table 7. *The details of LCLs in single cell RNA sequencing studies*

Sample	Donor	EBV strain	B lymphocytes	Culture time (month)
LCL1	1	B95-8 (Type 1)	CD19+ B	1
LCL2	1	M81 (Type 2)	CD19+ B	1
LCL3	2	B95-8	CD19+ B	6
LCL4	3 (GM18502)	B95-8	/	/
LCL5	4 (GM12878)	B95-8	/	/

* All datasets were generated using 10x 3' chemistry.

3.2 RNA quantification

RNA was extracted using TRIzol reagent (Life Technologies) followed by quantification by NanoDrop 2000. DNA was removed by DNase treatment (TURBO DNA-free™ Kit, Thermo Fisher Scientific).

(1) RT-qPCR

The DNA-free RNA samples were the starting materials for RT-qPCR. Both SuperScript III Platinum One-Step RT-qPCR Kit and two-step workflow were used. Reverse transcription was performed using High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher) in **Paper I**, and TATAA GrandScript cDNA FreePrime Kit (TATAA Biocenter) in **Paper IV**. The primers used for the cDNA synthesis were random hexamers and oligo(dT)20 primers. The reaction mixture was incubated at 25ºC for 10min, 42ºC 45 min, 85ºC 5 min and held at 4ºC. The plasmid 17ADVGAP contained the entire RPMS1 lncRNA cDNA, which was ordered from GeneArt. A T7 promoter was added upstream of RPMS1 using annealed double strand oligonucleotides, which was then used as the template for preparing RPMS1 RNA standard using *in vitro* transcription (MEGAscript™ T7 Transcription Kit). The synthesized RPMS1 RNA was used as a spike-in control for RNA extraction and RT-qPCR in **Paper I**. For enhanced RT-qPCR mentioned in **Paper IV**, the cDNA was preamplified using mixed target-specific primer pairs for multiplex EBV genes [117].

(2) Bulk RNA sequencing in **Paper I**

The Namalwa derived cell lines pTRE, pTRE-RPMS1 and Pro-Re were prepared in triplicates with different DOX concentrations, pTRE: 0, 50 and 1000 ng/mL; pTRE-RPMS1: 0, 50 ng/mL; Pro-Re: 0,1000 ng/mL. The 21 total RNA samples were sent to Genewiz (Germany) in dry ice for rRNA depleted stranded RNA sequencing.

(3) Single-cell RNA sequencing

In **Paper III**, the EBV infected pseudoALI was washed and dissociated into single-cell suspension followed by 10x 3' v3 library preparation. For

Paper IV the single-cell gene expression libraries and single-cell BCR V(D)J libraries were prepared following the manufacturer's user guide: Chromium NextGEMSingleCell5'v2 User Guide RevC (CG000331). More than ten thousand cells were loaded per reaction. Both the patient and the control B-lymphocytes were loaded into two sequencing lanes, the library pool was subjected to high-depth sequencing aiming at approximately 5,000 reads/cell for the VDJ library and 27,500 reads/cell for the transcriptome library. The quality control of the library was checked using TapeStation (Agilent), and the pool of indexed libraries was sequenced by NGI Sweden.

Datasets from four NPC scRNA-seq studies were collected and analyzed in **Paper I**. The design of the four studies and the available data types were shown in Figure 8.

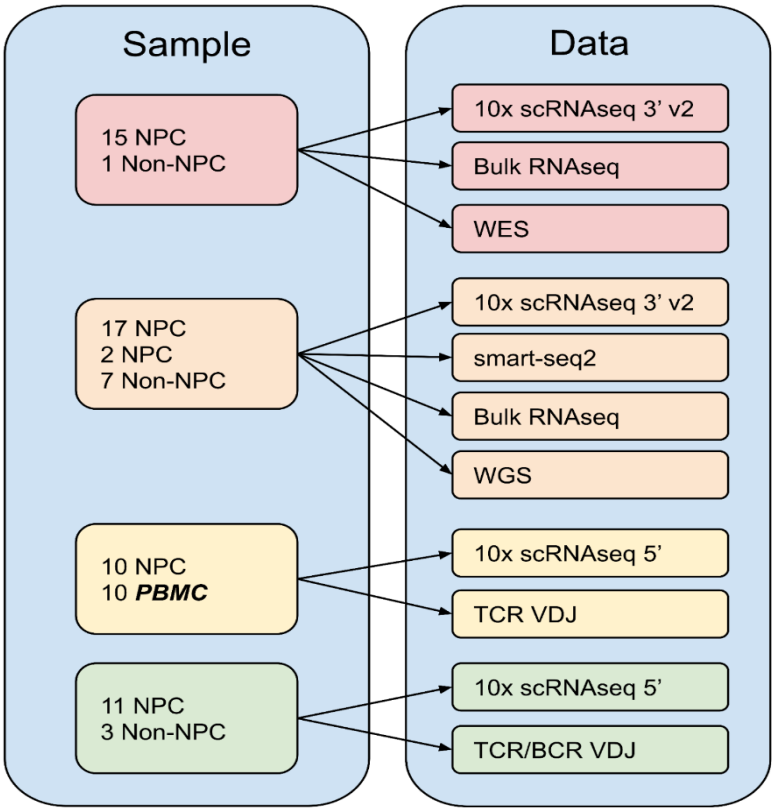


Figure 8. *The data types of four NPC studies.*

(4) Single-molecule long-read sequencing in **Paper I and Paper II**

The total RNA was extracted from the sample using TRIzol and subjected to DNA removal using DNase I. RPMS1 reverse primer (5'-TTGCATGTCTCACACCATGG-3') was used for gene specific reverse transcription according to the Nanopore protocol SQK-DCS109. The cDNA library was built using Maxima H Minus Reverse Transcriptase (Thermo).

Specific forward primers at the start of RPMS1 exon 1 and BARE1-3 (BamHI A rightward element) were used for PCR amplification. The PCR was performed using Q5 High-Fidelity 2x Master Mix (New England Biolabs). Subsequent library preparation steps using NEBNext Ultra II End repair/dA-tailing Module and Blunt/TA Ligase Master Mix created libraries with covalently attached sequencing adaptors. The purification steps were performed using Agencourt AMPure XP beads. The final purified libraries were mixed with the beads, loaded onto the flow cell and sequenced on a Nanopore MinION device.

3.3 Sequencing data analysis

(1) Bulk RNA-seq analysis

The raw reads were quality filtered, and the adapters were removed using prinseq and TrimGalore (Figure 9). All reads were trimmed from the 3'-end, and the reads with a mean quality score below 20, length shorter than 30bp or percentages of Ns higher than 10 were removed. The trimmed sequences were mapped to human genome reference GRCh38 and EBV reference separately using STAR. The EBV reference was modified from NC_007605.1 for stranded and non-stranded datasets. The alignment results were filtered to allow 10 multi-mapped reads, maximum of 3 mismatches and a minimum of 40 nucleotides alignment length. The relative EBV load for each sample was calculated as the number of EBV reads per million total reads (parts per million, ppm). The reads from each annotated gene were recorded with featureCounts from subread. To calculate the amount of a specific mRNA per million RNA molecules, the gene length and the sequencing

depth were normalized sequentially to get the relative transcripts per million (tpm) values.

The coverage plots were based on the alignment files and created using in-house R scripts. The depth of each position on the EBV genome was calculated using bedtools, 100 bins were used to segment the BamHI-A region, and 15% of the highest peak was the threshold of the reported bins. Reads with the polyA signal sequence AATAAA were extracted from the filtered fastq file and mapped towards NC_007605.1 using BWA. The downstream softclipped polyA tails were counted and summarized with the polyA signal location on the EBV NC_007605.1 reference. The junction reads (>5) within the BamHI-A region of the EBV positive samples (ppm>10) were plotted using GVIZ in R.

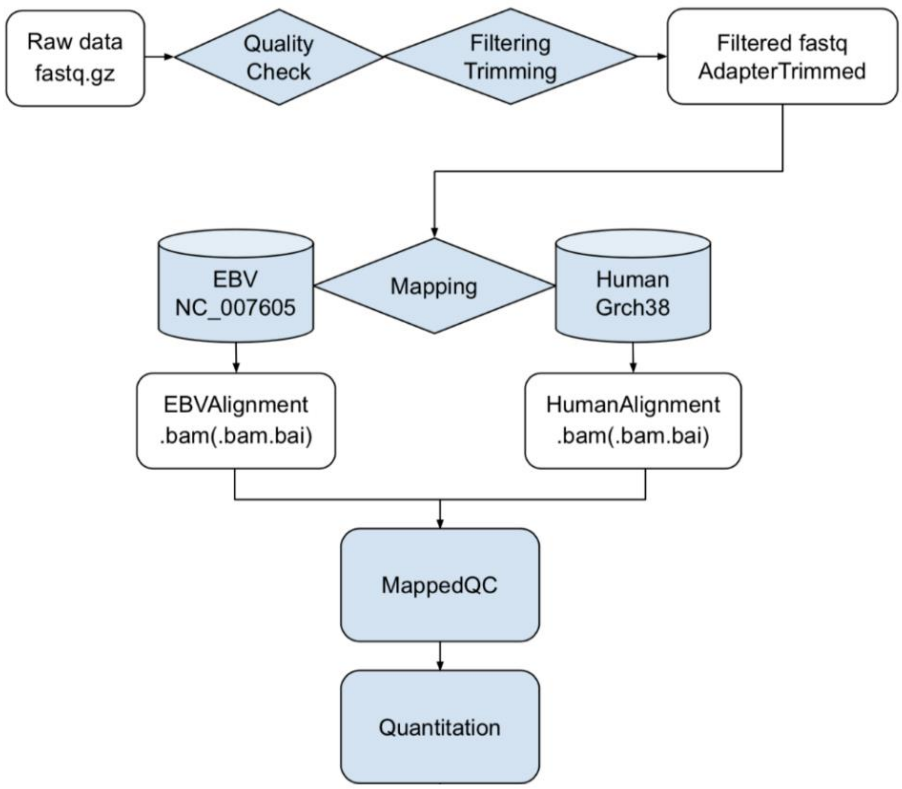


Figure 9. *The workflow for bulk RNA-seq analysis.*

(2) Single-cell RNA-seq analysis

To get a comprehensive landscape of EBV gene expression in scRNA-seq, multiple EBV references based on NC_007605.1 (B95-8 and Raji) and/or Akata EBV genome KC207813 were used: 1) the entire EBV reference genome as an exon, 2) original reference with CDS/gene changed to exon, 3) the modified reference genome with fused annotations. For the alignment and read counting, cellranger was used to index the merged genomes of human and EBV DNA. The output raw and filtered matrix were compared and only subtle variations in the number of EBV positive cells were observed in most samples. Therefore, the filtered matrix was used for the subsequent analyses. The cells with a minimum of 200, maximum of 9,000 genes were selected. Every sample was processed separately using Seurat in R, and all genes of the selected cells were used for clustering after sctransfrom.

In **Paper I**, the cell types were annotated using singleR and curated cell markers. The EBV expression profile was plotted in a heatmap using UMIs of EBV genes. The fraction of EBV-positive cells within each cell type and the fraction of EBV gene amongst the positive cells was calculated. To enable a comparison of EBV expression levels in bulk RNA-seq with scRNA-seq, the EBV gene cpm in all cells, epithelial cells, and only EBV positive epithelial cells were calculated and shown in heatmaps, generated using Prism 8. The epithelial cells in each sample were extracted as separated Seurat objects and then renormalized and regrouped (Figure 10).

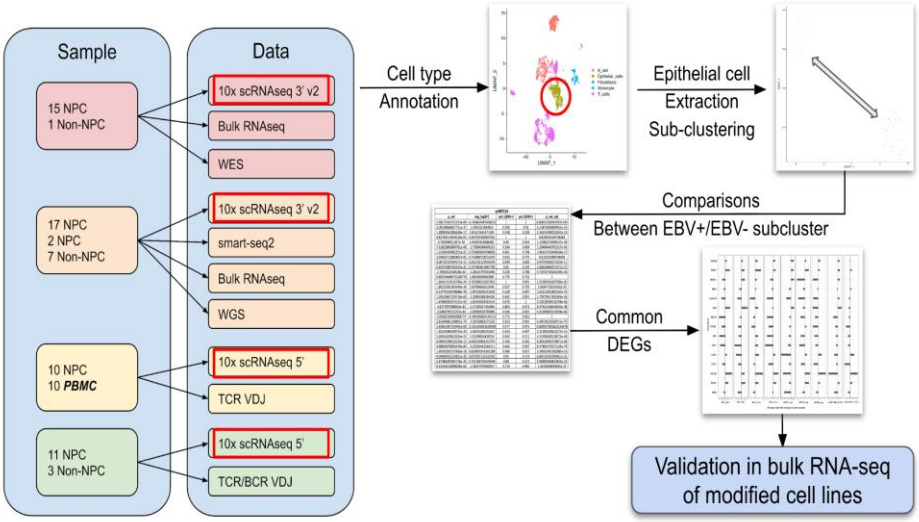


Figure 10. *The workflow for scRNA-seq analysis in* ***Paper I***

Ten samples with subclusters that were distinctively EBV-positive or negative were further analyzed. By comparing EBV-positive epithelial subclusters with the EBV-negative subclusters, a wide span of number of differentially expressed genes were found between samples. The pathway enrichment analysis of the differentially expressed genes was performed in GSEA/ GO/ clusterProfiler 4.0. Twelve genes were found to be perturbed in all ten samples.

(3) long-read sequencing analysis

After the fast5 files were generated, the bases were called using Guppy with default settings to generate fastq files. The alignment of the long reads was conducted using minimap2, and secondary alignment was excluded. The sam files from minimap2 were converted into bam files using samtools. Upon manual inspection, a large proportion of the reads were considered as artifacts, therefore the following criteria were used to filter the reads: 1) the read should not be shorter than 1,000 bases, 2) the read needs to contain the forward primer sequence for RPMS1/BAREs. The filtered bam files were then converted into BED12 format using bedtools, which were the input files for the newly developed tool FLAME (full-length adjacency matrix and exon enumeration). The reads that aligned to the reference annotation were assigned as annotated

while the reads that did not map were classified as incongruent. The incongruent reads were further analyzed and the differences from the local reference annotation in terms of exon start position, end position or length were used for complementing the reference annotation. The reads with the same exon patterns were collapsed and counted, and the consecutive exon linkage was listed in a weighted adjacency matrix. Three aspects were taken into consideration for novel exon definition: 1) the usage frequency of the splice site, 2) splicing donor and acceptor GU-AG signal at the intron-exon junction, and 3) supportive evidence of junction reads from short read sequencing.

4 RESULTS AND DISCUSSION

Paper I

In this project we characterized the EBV polyadenylated RNA in tumor biopsies at tissue and cellular level. In total, 676 bulk datasets from four types of tumors and associated cell lines were screened for EBV, 156 samples had an EBV load of more than 10 ppm. At single-cell level, 63 nasopharyngeal tissues from four studies were analyzed, of which 52 datasets are from NPC.

In the EBV positive bulk RNA-seq datasets from tumor biopsies, more than 85% of EBV reads were aligned to the BamHI A region. However, the conclusions regarding the EBV gene expression profile in the former studies were contradictory to our results due to (1) most RNA-seq datasets were generated with non-stranded sequencing protocols, and (2) EBV BamHI-A region contains intersecting genes, including both rightward genes RPMS1 and seven leftward genes. By investigating the reads mapped in this region, we concluded that the most highly expressed EBV gene was RPMS1 in primary tumors. The reasons included (1) the reads mapped to RPMS1 unique exons; (2) the coverage peaks matched to the RPMS1 exon coordinates; (3) the junction-spanning reads correlated well with the RPMS1 exon-exon connections; (4) the location of the reads with polyadenylation signal (AAUAAA) and tail (A-stretch) indicated rightward termination. Stranded sequenced datasets of eBLs further support the conclusion that the majority of read were transcribed in a rightward direction. Also, peaks in the RPMS1 introns could be characterized using single-molecule sequencing of gene-specific amplicons and classified into the new genes which we named BAREs. Moreover, for the overlapped 3'-end of three leftward genes LMP1, BNLF2a/b, we concluded that the dominant transcripts were BNLF2a/b because the tpm-value of the 2 kb LMP1 unique region was less than 5 in the majority of samples.

In accordance with the bulk sequencing data, RPMS1 was ubiquitously detected in malignant cells in scRNA-seq datasets from NPC. The data from the first two studies were generated using a 3' library preparation method and the other two 5'. The fourth study captured only a few

epithelial cells and in the second study, the tumor biopsies were dissociated and FACS-sorted for epithelial cells. The epithelial cells were then remixed with the stromal cells and the enriched epithelial content can be observed for these samples, three datasets contained approximately 50% epithelial cells.

In a few samples the expression of LMP1/BNLF2a/b was dominant. In these tumors EBV reads were found in all types of stromal cells. Also, the fraction of EBV-positive cells was similar for the different cell types. Moreover, EBV expression in the stromal cells mirrored the EBV expression in the epithelial cells, which argues that the EBV RNA originates from contaminants, exosomes or apoptotic bodies. However, in NPC46, a large proportion of the cells expressed BRLF1/BZLF1 indicating an ongoing reactivation, and this pattern was not observed in the epithelial cells.

To understand the role of EBV in the malignant cells, the epithelial cells of each sample were extracted as a separated seurat object. Because the number of epithelial cells varied amongst the samples, we focused on the samples with more than 100 epithelial cells. Normalization and clustering of the epithelial cells were reperformed, and differentially expressed genes between EBV positive and negative clusters were found. The number of genes varied, and the significantly enriched pathways of each sample based on the FoldChange and adjusted p value selected genes were mainly TNFA signaling via NFKB, MYC targets, immune response, G2M checkpoint, E2F targets, apoptosis, etc. Twelve commonly changed genes were found in the gene lists without filtering. Interestingly, most of them were target genes for interferon response. We could validate this finding in a Burkitt's lymphoma cell line in which expression of RPMS1/miR-BARTs were induced, as well as the LCLs from the same donor transformed with different EBV strains B95-8 (miR-BARTs deleted) and M81. Interferon gamma were found to be expressed at higher levels in the EBV-positive NPC epithelial cells compared with EBV-negative tumors/biopsies.

Paper II

Full length EBV lncRNA RPMS1 splice variants from a GAC biopsy and the NPC cell line C666-1 were characterized using nanopore long read sequencing. Analysis of the entire spliceome was initially found to be restricted by current computational tools which rely on which rely on the existing exon annotation. To solve this issue, we developed a Python-based tool called FLAME, which efficiently retrieved and annotated transcript variants from long read sequencing with the complementation of short read sequencing data. All RPMS1 isoforms reported before were detected using FLAME, including the potential ORFs. In addition, 32 novel exons were found within RPMS1, and the three large exons III/V/VII contained abundant intraexonic splicing events. The different usage frequency of the exon Ib, II, and VIa was different between C666-1 and GAC. Overall, the pipeline FLAME is suitable for accurate identification of novel exons. The comprehensive characterization of the alternative splicing variant of RPMS1 could be utilized as a reference for multiple purposes, such as, potential protein screening, secondary structures prediction, and ASO design.

Paper III

The maintenance of EBV latency and the virus propensity to switch into lytic replication is highly influenced by the method by which the cells are cultured and cell differentiation. The EBV genome in the HK1-EBV cell line is kept in a latent stage when the cells are grown in two-dimensional setting and lytic reactivation is observed when the cells are cultured as stratified epithelium. Three-dimensional pseudostratified air-liquid interface (pseudo-ALI) of reprogrammed nasal epithelium can be used as a model for investigating EBV *de novo* infection. This infection model would thus more accurately portray the initial infection/transformation process leading to the development of EBV-associated NPC. *In situ* hybridization and immunostaining demonstrated that all cell types of the nasopharyngeal pseudostratified epithelium were represented in our cultures.

We present evidence that EBV can infect nasopharyngeal pseudo-ALI cultures and can be maintained in 3 out of 9 donors, of which two

samples exhibited latent EBV infection markers (EBER+, BZLF1-, gp350-). Cultures derived from donor 4 expressed EBV lytic genes (BZLF+, gp350+, LMP1+). This sample was subsequently subjected to scRNA-seq.

Based on the EBV gene expression of individual cells the different stages of the virus life cycle could be identified including lytic cells in which the host expression had been shut off. Furthermore, scRNA-seq confirmed that every cell type including basal, suprabasal, mucosecretory and ciliated cells were susceptible to EBV infection. The suprabasal cells had the highest proportion of cells in which EBV was lytic.

Paper IV

In this study, we screened single cell RNA sequencing datasets from peripheral blood cells for latent viral content. In a human immunodeficiency virus-infected (HIV) patient lymphocyte we found EBV RNA from RPMS1. Further investigations of blood from immunosuppressed patients using an enhanced sensitivity RT-qPCR showed that the EBV non-coding RNAs (RPMS1 and EBERs) were most abundantly expressed. Also, the number of EBV genes detected was positively correlated with the EBV DNA load in the blood sample.

In order to characterize the EBV-transformed cells we enriched B-lymphocytes from a splenectomized patient with a high EBV DNA load in blood, >5,000,000 EBV copies per milliliter. EBV was detected in approximately 30% of the B-lymphocytes based on EBER-ISH. The B-lymphocytes were subjected to 5' single cell transcriptome and B-cell receptor VDJ sequencing. Considering the low capture rate associated with the single cell sequencing technology we expected to detect significantly fewer EBV-positive cells compared with EBER-ISH. EBV RNA was only found in 645 out of 25,218 cells (2.6%). By inferring the positive EBV-status for cells with the same B-cell receptor VDJ arrangement we could further increase the EBV-positive fraction to 6.6%. The EBV expression pattern for this patient showed a high similarity with the tumors (**Paper I**) with more than 70% of the EBV transcripts originating from the BamHI-A region. RPMS1 and LMP1

were the dominant EBV mRNA in the B-lymphocytes. The transcriptome of the EBV-positive cells *in vivo* was different from *in vitro* EBV-transformed B-lymphocytes in terms of number of RNA molecules and number of EBV genes expressed per cell.

We also reanalyzed the single-cell transcriptomic data of five LCLs, and a subset of cells with abundant EBV gene species (lytic) and low total RNA content were observed. This population of cells will affect the EBV expression pattern in bulk RNA-seq analysis. Four out of the five LCLs showed lytic EBV gene expression pattern (Figure 4C). LCL3 was in latency III with more than 5 counts per million (cpm) of the latency genes. The reasons for the different EBV expression pattern formation could be due to host cell factors and the number of passages.

5 CONCLUSION AND FUTURE PERSPECTIVES

To accurately depict the EBV expression landscape, the biases and limitations of different transcriptomic profiling techniques should be considered. Based on the results shown in our studies, the constitutively expressed EBV genes in tumor cells *in vivo* are non-coding RNAs EBERs, RPMS1/BAREs and EBV microRNAs. The most abundant RPMS1/BAREs are highly spliced genes (**Paper I** and **Paper II**). In some individuals, protein coding genes BNLF2a/b, LMP1 and LMP2 were also abundant (**Paper I** and **Paper IV**). All EBV-positive samples contained low levels of lytic EBV transcript in bulk sequencing datasets indicating that a few cells in the samples were undergoing lytic reactivation. This low background signal was significantly reduced in the scRNA-seq dataset, but a few cells in most tumors still expressed lytic transcripts. In a recent study the transcriptome of microdissected NPC tissues showed that IFN related genes were upregulated at the normal adjacent compared with the dysplastic epithelium [118][118]. However, scRNA-seq of NPC in **Paper I** suggested that although EBV-positive NPC epithelium had higher levels of interferon, the IFN stimulated genes in EBV-positive malignant cells were downregulated by miR-BARTs. Therefore, the EBV-positive cells most likely have a higher tolerance for stress signals and through expression of the miR-BARTs likely increases the fitness of the cell for that particular microenvironment.

EBV expression pattern varies between *in vivo* and *in vitro* tissues. Firstly, primary tumor biopsies contained different types of EBV transcripts compared to *in vitro* cultures based on the bulk RNA-seq analysis (**Paper I**). Secondly, the total amount of RNAs per cell in EBV transformed LCL was much higher than primary EBV-positive B-lymphocytes (**Paper IV**). Finally, the enriched EBV-positive primary B -lymphocytes displayed the same EBV expression pattern as primary tumors (**Paper IV**). This argues that the EBV latency program in primary tissue is ubiquitous for all EBV-infected cells and that the minor differences observed between patients is mainly due to restrictions by the host immune system. In order to understand the EBV transformation

process a comprehensive approach using, (1) 3-D models with different stages of EBV (**Paper III**), (2) co-culturing EBV infected cells with stromal cells, (3) *in vivo* models, or (4) primary EBV-positive cells (**Paper IV**) should be utilized.

Due to the limitation of the different RNA sequencing protocols, not all types of RNAs, including circular RNA and short non-coding RNA, could be quantified and compared simultaneously. However, there are some bioinformatic tools emerging for direct comparing linear and circular RNA [119, 120]. Moreover, customized viral oligos compatible with existing platforms could provide a strategy for comprehensive profiling. Furthermore, the ratio of EBV RNA and protein could be affected by the half-life of the viral RNA and the turn-over of the viral protein as well as translational perturbations caused by viral elements [121, 122]. Hence, the interaction of EBV with its host on the temporal and spatial level needs to be further investigated in order to get more detailed information on the *in vivo* dynamic co-development of EBV immune evasion and transformation.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor **Ka-Wei Tang** for giving me the opportunity to join your group and to work on tumor virology using new techniques. It has been a great life experience to conduct my PhD study at University of Gothenburg with your support. Thanks for your intellectual input in every discussion and your continuous encouragement! The learning environment you created was extremely helpful for me to grow. With your help issues became simpler, and I became more confident and braver.

Many thanks to my co-supervisors, **Kristoffer Hellstrand, Anna Martner, and Rickard Nordén,** for your generous help and support. **Fredrik Bergh Thorén** for creating a nice environment. Thanks to our collaborators **Kathy Shair, Kaisa Thorell**, **Carolina Guibentif** for your input and time.

Many thanks to my peers in the Tang group: **Diana** for your positive feedback and energetic mood. Growing as a graduate student together with you is a precious memory in my life. Thank you for teaching me medical knowledge and reading the BL article together in the mornings! **Sanna** for helping me with both analysis and bioinformatics. I could not image how it would have been without your help. Thank you for your patience! **Guojiang** for helping me with experiments and data reanalysis. Thanks for your support and “kao-pu”! **Isak** for your care and new ideas. I really appreciate your work, thoughts, figures and the discussions you initiated. Thanks for helping me with the thesis! **Alan** for helping me to organize and analyze the data. Your expertise with computers made it much easier for me to learn. **Brwa** for always being very helpful and supportive. Thank you for your comments on my thesis!

Former group members of the Tang group: **Jonas** for helping me with the project plan. Thank you for your kind care when I broke my computer screen in the office! **Joanna** I really enjoy working with you. Thank you for teaching me how to talk with PIs! **Harsha** for help with UPPMAX and data transfer. **Manuela** for the figure of the LNA experiment. **Sofia** for helping me with the work at virology and NGS experiments. **Torun** for introducing me to grants.

It is of pleasure to acknowledge my fellows on the same floor: **Hana, Linnea, Sanchari, Malin, Roberta, Elin, Junko, Mike, Alexander, Hanna, Belson, Anne, Nuttida, Chiara, Ali, Mohammad, Veronika, Andreas, Ebru;** and **Johan, Kasthuri, Ebba** for creating friendly and supportive environment.

I would also like to thank my friends**: Li Liu, Yinghui, Hao, Zhicheng, Guoqiang, Shuwen, Lijuan, Yongjin, Haixia, Shan Jiang, Oi-Kuan, Xiaojing, Yan, Mi, Jian, Junrui, Haoyu, Men, Xia, Hang, Shan Shu, Yunyun, Yufang, Aifang, Xin,** for your care and assistance over the years!

Thanks to my family for your endless support!

REFERENCES

1. Epstein, M.A., B.G. Achong, and Y.M. Barr, *Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma.* Lancet, 1964. **1**(7335): p. 702-3.
2. Guan, Y., et al., *The role of Epstein-Barr virus in multiple sclerosis: from molecular pathophysiology to in vivo imaging.* Neural Regen Res, 2019. **14**(3): p. 373-386.
3. Bjornevik, K., et al., *Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis.* Science, 2022. **375**(6578): p. 296-301.
4. Kerr, J.R., *Epstein-Barr virus (EBV) reactivation and therapeutic inhibitors.* J Clin Pathol, 2019. **72**(10): p. 651-658.
5. Howley, P.M., et al., *Fields Virology: DNA Viruses*. 7th ed. 2021: Wolters Kluwer Health.
6. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-9.
7. Chen, Y.P., et al., *Nasopharyngeal carcinoma.* Lancet, 2019. **394**(10192): p. 64-80.
8. de Martel, C., et al., *Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis.* Lancet Glob Health, 2020. **8**(2): p. e180-e190.
9. Tang, K.W., et al., *The landscape of viral expression and host gene fusion and adaptation in human cancer.* Nat Commun, 2013. **4**: p. 2513.
10. Zapatka, M., et al., *The landscape of viral associations in human cancers.* Nat Genet, 2020. **52**(3): p. 320-330.
11. Young, L.S., L.F. Yap, and P.G. Murray, *Epstein-Barr virus: more than 50 years old and still providing surprises.* Nature Reviews Cancer, 2016. **16**(12): p. 789-802.
12. Sun, C., et al., *The Status and Prospects of Epstein-Barr Virus Prophylactic Vaccine Development.* Front Immunol, 2021. **12**: p. 677027.
13. Cui, X. and C.M. Snapper, *Epstein Barr Virus: Development of Vaccines and Immune Cell Therapy for EBV-Associated Diseases.* Front Immunol, 2021. **12**: p. 734471.
14. Tang, D., et al., *VISDB: a manually curated database of viral integration sites in the human genome.* Nucleic Acids Res, 2020. **48**(D1): p. D633-D641.
15. Ohshima, K., et al., *Integrated and episomal forms of Epstein-Barr virus (EBV) in EBV associated disease.* Cancer Lett, 1998. **122**(1-2): p. 43-50.
16. Chakravorty, S., et al., *Integrated Pan-Cancer Map of EBV-Associated Neoplasms Reveals Functional Host-Virus Interactions.* Cancer Res, 2019. **79**(23): p. 6010-6023.
17. Gatherer, D., et al., *ICTV Virus Taxonomy Profile: Herpesviridae 2021.* J Gen Virol, 2021. **102**(10).
18. McGeoch, D.J., F.J. Rixon, and A.J. Davison, *Topics in herpesvirus genomics and evolution.* Virus Res, 2006. **117**(1): p. 90-104.
19. Bridges, R., et al., *Essential role of inverted repeat in Epstein-Barr virus IR-1 in B cell transformation; geographical variation of the viral genome.* Philos Trans R Soc Lond B Biol Sci, 2019. **374**(1773): p. 20180299.
20. Correia, S., et al., *Sequence Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases.* J Virol, 2018. **92**(22).
21. Farrell, P.J. and R.E. White, *Do Epstein-Barr Virus Mutations and Natural Genome Sequence Variations Contribute to Disease?* Biomolecules, 2021. **12**(1).
22. Feng, F.T., et al., *A single nucleotide polymorphism in the Epstein-Barr virus genome is strongly associated with a high risk of nasopharyngeal carcinoma.* Chin J Cancer, 2015. **34**(12): p. 563-72.

23. Xu, M., et al., *Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma.* Nat Genet, 2019. **51**(7): p. 1131-1136.
24. Yajima, M., et al., *A global phylogenetic analysis of Japanese tonsil-derived Epstein-Barr virus strains using viral whole-genome cloning and long-read sequencing.* J Gen Virol, 2021. **102**(3).
25. Chen, Z.H., et al., *The genomic architecture of EBV and infected gastric tissue from precursor lesions to carcinoma.* Genome Med, 2021. **13**(1): p. 146.
26. Palser, A.L., et al., *Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection.* J Virol, 2015. **89**(10): p. 5222-37.
27. Han, S., et al., *Epstein-Barr Virus Epithelial Cancers-A Comprehensive Understanding to Drive Novel Therapies.* Front Immunol, 2021. **12**: p. 734293.
28. Grande, B.M., et al., *Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma.* Blood, 2019. **133**(12): p. 1313-1324.
29. Bruce, J.P., et al., *Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-kappaB activation and immune escape.* Nat Commun, 2021. **12**(1): p. 4193.
30. Lin, D.C., et al., *The genomic landscape of nasopharyngeal carcinoma.* Nat Genet, 2014. **46**(8): p. 866-71.
31. Wong, K.C.W., et al., *Nasopharyngeal carcinoma: an evolving paradigm.* Nat Rev Clin Oncol, 2021. **18**(11): p. 679-695.
32. Schmitz, R., et al., *Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics.* Nature, 2012. **490**(7418): p. 116-20.
33. Majerciak, V., et al., *A Genome-Wide Epstein-Barr Virus Polyadenylation Map and Its Antisense RNA to EBNA.* J Virol, 2019. **93**(2).
34. Yuan, J., et al., *Virus and cell RNAs expressed during Epstein-Barr virus replication.* J Virol, 2006. **80**(5): p. 2548-65.
35. O'Grady, T., et al., *Global transcript structure resolution of high gene density genomes through multi-platform data integration.* Nucleic Acids Res, 2016. **44**(18): p. e145.
36. Yetming, K.D., et al., *The BHLF1 Locus of Epstein-Barr Virus Contributes to Viral Latency and B-Cell Immortalization.* J Virol, 2020. **94**(17).
37. Moss, W.N., et al., *RNA families in Epstein-Barr virus.* RNA Biol, 2014. **11**(1): p. 10-7.
38. Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones, *miRBase: from microRNA sequences to function.* Nucleic Acids Res, 2019. **47**(D1): p. D155-D162.
39. Ge, J., et al., *Epstein-Barr Virus-Encoded Circular RNA CircBART2.2 Promotes Immune Escape of Nasopharyngeal Carcinoma by Regulating PD-L1.* Cancer Res, 2021. **81**(19): p. 5074-5088.
40. Toptan, T., et al., *Circular DNA tumor viruses make circular RNAs.* Proc Natl Acad Sci U S A, 2018. **115**(37): p. E8737-E8745.
41. Ungerleider, N., et al., *The Epstein Barr virus circRNAome.* PLoS Pathog, 2018. **14**(8): p. e1007206.
42. Wang, J., et al., *EBV miRNAs BART11 and BART17-3p promote immune escape through the enhancer-mediated transcription of PD-L1.* Nat Commun, 2022. **13**(1): p. 866.
43. Cristino, A.S., et al., *EBV microRNA-BHRF1-2-5p targets the 3'UTR of immune checkpoint ligands PD-L1 and PD-L2.* Blood, 2019. **134**(25): p. 2261-2270.

44. Arias, C., et al., *KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features.* PLoS Pathog, 2014. **10**(1): p. e1003847.
45. Murata, T., et al., *Molecular Basis of Epstein-Barr Virus Latency Establishment and Lytic Reactivation.* Viruses, 2021. **13**(12).
46. Inagaki, T., et al., *Direct Evidence of Abortive Lytic Infection-Mediated Establishment of Epstein-Barr Virus Latency During B-Cell Infection.* Front Microbiol, 2020. **11**: p. 575255.
47. Wang, L.W., et al., *Epstein-Barr-Virus-Induced One-Carbon Metabolism Drives B Cell Transformation.* Cell Metab, 2019. **30**(3): p. 539-555 e11.
48. Lamontagne, R.J., et al., *A multi-omics approach to Epstein-Barr virus immortalization of B-cells reveals EBNA1 chromatin pioneering activities targeting nucleotide metabolism.* PLoS Pathog, 2021. **17**(1): p. e1009208.
49. Dunn, L.E.M., Lu, F., Lieberman, P. M., & Baines, J. D., *Increased RNA Polymerase Activity and Pausing at CTCF binding sites on the Epstein Barr Virus Genome During Reactivation from Latency.* bioRxiv, 2021.
50. Rueger, S., et al., *The influence of human genetic variation on Epstein-Barr virus sequence diversity.* Sci Rep, 2021. **11**(1): p. 4586.
51. Gotoh, K., et al., *Immunologic and virologic analyses in pediatric liver transplant recipients with chronic high Epstein-Barr virus loads.* J Infect Dis, 2010. **202**(3): p. 461-9.
52. Lam, W.K.J., K.C.A. Chan, and Y.M.D. Lo, *Plasma Epstein-Barr virus DNA as an archetypal circulating tumour DNA marker.* J Pathol, 2019. **247**(5): p. 641-649.
53. Nilsson, J.S., et al., *Intralesional EBV-DNA load as marker of prognosis for nasopharyngeal cancer.* Sci Rep, 2019. **9**(1): p. 15432.
54. McKenzie, J. and A. El-Guindy, *Epstein-Barr Virus Lytic Cycle Reactivation.* Curr Top Microbiol Immunol, 2015. **391**: p. 237-61.
55. Wang, C., et al., *RNA Sequencing Analyses of Gene Expression during Epstein-Barr Virus Infection of Primary B Lymphocytes.* J Virol, 2019. **93**(13).
56. Yanagi, Y., et al., *RNAseq analysis identifies involvement of EBNA2 in PD-L1 induction during Epstein-Barr virus infection of primary B cells.* Virology, 2021. **557**: p. 44-54.
57. Mrozek-Gorska, P., et al., *Epstein-Barr virus reprograms human B lymphocytes immediately in the prelatent phase of infection.* Proc Natl Acad Sci U S A, 2019. **116**(32): p. 16046-16055.
58. McFadden, K., et al., *Metabolic stress is a barrier to Epstein-Barr virus-mediated B-cell immortalization.* Proc Natl Acad Sci U S A, 2016. **113**(6): p. E782-90.
59. Sugimoto, M., et al., *Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein-Barr virus.* Cancer Res, 2004. **64**(10): p. 3361-4.
60. Forrest, C., et al., *Proteome-wide analysis of CD8+ T cell responses to EBV reveals differences between primary and persistent infection.* PLoS Pathog, 2018. **14**(9): p. e1007110.
61. Taylor, G.S., et al., *The immunology of Epstein-Barr virus-induced disease.* Annu Rev Immunol, 2015. **33**: p. 787-821.
62. Ye, J., et al., *De novo protein synthesis is required for lytic cycle reactivation of Epstein-Barr virus, but not Kaposi's sarcoma-associated herpesvirus, in response to histone deacetylase inhibitors and protein kinase C agonists.* J Virol, 2007. **81**(17): p. 9279-91.

63. Decaussin, G., V. Leclerc, and T. Ooka, *The lytic cycle of Epstein-Barr virus in the nonproducer Raji line can be rescued by the expression of a 135-kilodalton protein encoded by the BALF2 open reading frame.* J Virol, 1995. **69**(11): p. 7309-14.
64. Lin, W., et al., *Establishment and characterization of new tumor xenografts and cancer cell lines from EBV-positive nasopharyngeal carcinoma.* Nat Commun, 2018. **9**(1): p. 4663.
65. Guo, R., et al., *MYC Controls the Epstein-Barr Virus Lytic Switch.* Mol Cell, 2020. **78**(4): p. 653-669 e8.
66. O'Grady, T., et al., *Global bidirectional transcription of the Epstein-Barr virus genome during reactivation.* J Virol, 2014. **88**(3): p. 1604-16.
67. Chakravorty, A., B. Sugden, and E.C. Johannsen, *An Epigenetic Journey: Epstein-Barr Virus Transcribes Chromatinized and Subsequently Unchromatinized Templates during Its Lytic Cycle.* J Virol, 2019. **93**(8).
68. Hau, P.M., et al., *Targeting Epstein-Barr Virus in Nasopharyngeal Carcinoma.* Front Oncol, 2020. **10**: p. 600.
69. Dalton, T., et al., *Epigenetic reprogramming sensitizes immunologically silent EBV+ lymphomas to virus-directed immunotherapy.* Blood, 2020. **135**(21): p. 1870-1881.
70. Yip, Y.L., et al., *Establishment of a nasopharyngeal carcinoma cell line capable of undergoing lytic Epstein-Barr virus reactivation.* Lab Invest, 2018. **98**(8): p. 1093-1104.
71. Cheung, S.T., et al., *Nasopharyngeal carcinoma cell line (C666-1) consistently harbouring Epstein-Barr virus.* Int J Cancer, 1999. **83**(1): p. 121-6.
72. Park, J.G., et al., *Establishment and characterization of human gastric carcinoma cell lines.* Int J Cancer, 1997. **70**(4): p. 443-9.
73. Kim, D.N., et al., *Characterization of naturally Epstein-Barr virus-infected gastric carcinoma cell line YCCEL1.* J Gen Virol, 2013. **94**(Pt 3): p. 497-506.
74. Frisan, T., V. Levitsky, and M. Masucci, *Generation of lymphoblastoid cell lines (LCLs).* Methods Mol Biol, 2001. **174**: p. 125-7.
75. Maruo, S., L. Yang, and K. Takada, *Roles of Epstein-Barr virus glycoproteins gp350 and gp25 in the infection of human epithelial cells.* J Gen Virol, 2001. **82**(Pt 10): p. 2373-2383.
76. Katsumura, K.R., et al., *Quantitative evaluation of the role of Epstein-Barr virus immediate-early protein BZLF1 in B-cell transformation.* J Gen Virol, 2009. **90**(Pt 10): p. 2331-2341.
77. Kim, J., B.K. Koo, and J.A. Knoblich, *Human organoids: model systems for human biology and medicine.* Nat Rev Mol Cell Biol, 2020. **21**(10): p. 571-584.
78. Hutt-Fletcher, L.M., *The Long and Complicated Relationship between Epstein-Barr Virus and Epithelial Cells.* J Virol, 2017. **91**(1).
79. Bukowy-Bieryllo, Z., *Long-term differentiating primary human airway epithelial cell cultures: how far are we?* Cell Commun Signal, 2021. **19**(1): p. 63.
80. Caves, E.A., et al., *Air-Liquid Interface Method To Study Epstein-Barr Virus Pathogenesis in Nasopharyngeal Epithelial Cells.* mSphere, 2018. **3**(4).
81. Munz, C., *Humanized mouse models for Epstein Barr virus infection.* Curr Opin Virol, 2017. **25**: p. 113-118.
82. Schuhmachers, P. and C. Munz, *Modification of EBV Associated Lymphomagenesis and Its Immune Control by Co-Infections and Genetics in Humanized Mice.* Front Immunol, 2021. **12**: p. 640918.
83. Murer, A., et al., *MicroRNAs of Epstein-Barr Virus Attenuate T-Cell-Mediated Immune Control In Vivo.* mBio, 2019. **10**(1).

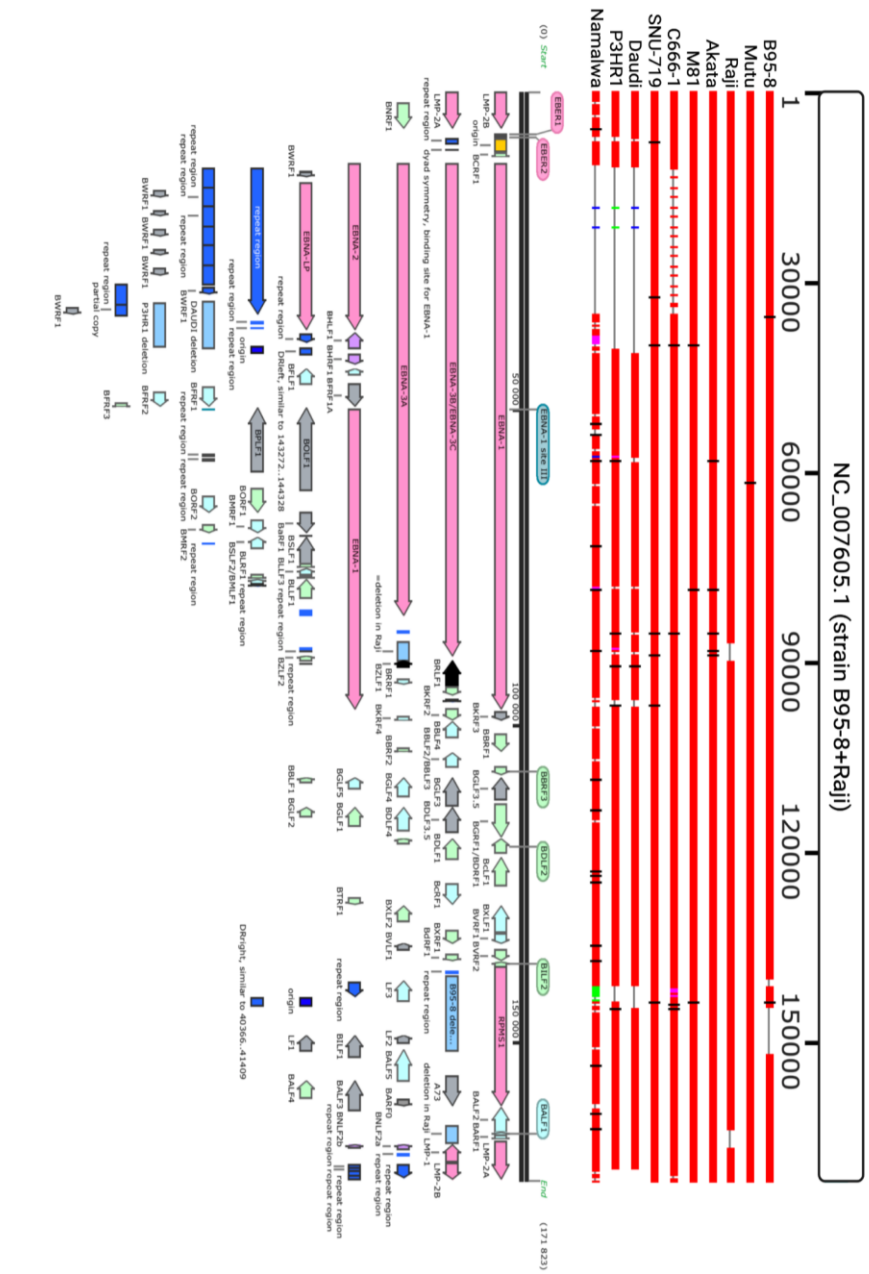
84. Xia, W., et al., *Tree Shrew Is a Suitable Animal Model for the Study of Epstein Barr Virus.* Front Immunol, 2021. **12**: p. 789604.
85. Reguraman, N., et al., *Uncovering early events in primary Epstein-Barr virus infection using a rabbit model.* Sci Rep, 2021. **11**(1): p. 21220.
86. Zhao, R., et al., *Designing strategies of small-molecule compounds for modulating non-coding RNAs in cancer therapy.* J Hematol Oncol, 2022. **15**(1): p. 14.
87. Sharp, P.A., et al., *RNA in formation and regulation of transcriptional condensates.* RNA, 2022. **28**(1): p. 52-57.
88. Blumberg, A., et al., *Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data.* BMC Biol, 2021. **19**(1): p. 30.
89. Statello, L., et al., *Gene regulation by long non-coding RNAs and its biological functions.* Nat Rev Mol Cell Biol, 2021. **22**(2): p. 96-118.
90. Bartel, D.P., *Metazoan MicroRNAs.* Cell, 2018. **173**(1): p. 20-51.
91. Liu, J., et al., *Real-time single-cell characterization of the eukaryotic transcription cycle reveals correlations between RNA initiation, elongation, and cleavage.* PLoS Comput Biol, 2021. **17**(5): p. e1008999.
92. Pichon, X., et al., *A Growing Toolbox to Image Gene Expression in Single Cells: Sensitive Approaches for Demanding Challenges.* Mol Cell, 2018. **71**(3): p. 468-480.
93. Rodriguez, J. and D.R. Larson, *Transcription in Living Cells: Molecular Mechanisms of Bursting.* Annu Rev Biochem, 2020. **89**: p. 189-212.
94. Chen, Y.P., et al., *Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma.* Cell Res, 2020. **30**(11): p. 1024-1042.
95. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
96. La Manno, G., et al., *RNA velocity of single cells.* Nature, 2018. **560**(7719): p. 494-498.
97. Lange, M., et al., *CellRank for directed single-cell fate mapping.* Nat Methods, 2022. **19**(2): p. 159-170.
98. Svensson, V., et al., *Power analysis of single-cell RNA-sequencing experiments.* Nat Methods, 2017. **14**(4): p. 381-387.
99. SoRelle, E.D., et al., *Single-cell RNA-seq reveals transcriptomic heterogeneity mediated by host-pathogen dynamics in lymphoblastoid cell lines.* Elife, 2021. **10**.
100. Osorio, D., et al., *Single-cell RNA sequencing of a European and an African lymphoblastoid cell line.* Sci Data, 2019. **6**(1): p. 112.
101. Jin, S., et al., *Single-cell transcriptomic analysis defines the interplay between tumor cells, viral infection, and the microenvironment in nasopharyngeal carcinoma.* Cell Res, 2020. **30**(11): p. 950-965.
102. Liu, Y., et al., *Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution.* Nat Commun, 2021. **12**(1): p. 741.
103. Gong, L., et al., *Comprehensive single-cell sequencing reveals the stromal dynamics and tumor-specific characteristics in the microenvironment of nasopharyngeal carcinoma.* Nat Commun, 2021. **12**(1): p. 1540.
104. Zhao, J., et al., *Single cell RNA-seq reveals the landscape of tumor and infiltrating immune cells in nasopharyngeal carcinoma.* Cancer Lett, 2020. **477**: p. 131-143.
105. Zhang, M., et al., *Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing.* Gut, 2021. **70**(3): p. 464-475.
106. Bristol, J.A., et al., *Reduced IRF4 expression promotes lytic phenotype in Type 2 EBV-infected B cells.* PLoS Pathog, 2022. **18**(4): p. e1010453.

107. Ungerleider, N., et al., *EBV miRNAs are potent effectors of tumor cell transcriptome remodeling in promoting immune escape.* PLoS Pathog, 2021. **17**(5): p. e1009217.
108. Cosmopoulos, K., et al., *Comprehensive profiling of Epstein-Barr virus microRNAs in nasopharyngeal carcinoma.* J Virol, 2009. **83**(5): p. 2357-67.
109. Gao, W., et al., *Detection of Epstein-Barr virus (EBV)-encoded microRNAs in plasma of patients with nasopharyngeal carcinoma.* Head Neck, 2019. **41**(3): p. 780-792.
110. Chen, S.J., et al., *Characterization of Epstein-Barr virus miRNAome in nasopharyngeal carcinoma by deep sequencing.* PLoS One, 2010. **5**(9).
111. Lung, R.W., et al., *EBV-encoded miRNAs target ATM-mediated response in nasopharyngeal carcinoma.* J Pathol, 2018. **244**(4): p. 394-407.
112. Hafner, M.e.a., *CLIP and complementary methods.* Nat Rev Methods Primers, 2021(1): p. 1-23.
113. Riley, K.J., et al., *EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency.* EMBO J, 2012. **31**(9): p. 2207-21.
114. Skalsky, R.L., et al., *The viral and cellular microRNA targetome in lymphoblastoid cell lines.* PLoS Pathog, 2012. **8**(1): p. e1002484.
115. Kang, D., R.L. Skalsky, and B.R. Cullen, *EBV BART MicroRNAs Target Multiple Pro-apoptotic Cellular Genes to Promote Epithelial Cell Survival.* PLoS Pathog, 2015. **11**(6): p. e1004979.
116. Cao, J., et al., *An easy and efficient inducible CRISPR/Cas9 platform with improved specificity for multiple gene targeting.* Nucleic Acids Res, 2016. **44**(19): p. e149.
117. Andersson, D., et al., *Properties of targeted preamplification in DNA and cDNA quantification.* Expert Rev Mol Diagn, 2015. **15**(8): p. 1085-100.
118. !!! INVALID CITATION !!! .
119. Ma, X.K., et al., *CIRCexplorer3: A CLEAR Pipeline for Direct Comparison of Circular and Linear RNA Expression.* Genomics Proteomics Bioinformatics, 2019. **17**(5): p. 511-521.
120. Ungerleider, N. and E. Flemington, *SpliceV: analysis and publication quality printing of linear and circular RNA splicing, expression and regulation.* BMC Bioinformatics, 2019. **20**(1): p. 231.
121. Gain, C., et al., *Proteasomal inhibition triggers viral oncoprotein degradation via autophagy-lysosomal pathway.* PLoS Pathog, 2020. **16**(2): p. e1008105.
122. Touitou, R., et al., *Epstein-Barr virus EBNA3 proteins bind to the C8/alpha7 subunit of the 20S proteasome and are degraded by 20S proteasomes in vitro, but are very stable in latently infected B cells.* J Gen Virol, 2005. **86**(Pt 5): p. 1269-1277.

APPENDIX

Appendix A. Map of EBV genome

Latent, IE, EL,LL



Appendix B. List of EBV genes

The genes were ordered firstly by EBV stage, then the alphabetical order of gene name. The underlined genes are the EBV antigens selected for animal trials, * are for EBV vaccine clinical trials, ** mRNA vaccine.

Gene Name	Description	EBV stage
EBER1/2	non-coding	Latent
RPMS1	non-coding	Latent
EBNA-LP	Epstein-Barr nuclear antigen leader protein EBNA-LP	Latent
EBNA-1*	Epstein-Barr nuclear antigen 1	Latent
EBNA-2A	Epstein-Barr nuclear antigen 2	Latent
EBNA-3A*	Epstein-Barr nuclear antigen 3	Latent
EBNA-3B*	Epstein-Barr nuclear antigen 4	Latent
EBNA-3C*	Epstein-Barr nuclear antigen 6	Latent
LMP-1*	Latent membrane protein 1	Latent
LMP-2A*	Latent membrane protein 2A	Latent
LMP-2B*	Latent membrane protein 2B	Latent
BRLF1	Rta	IE
BZLF1	Trans-activator protein, XEBRA/ZEBRA/Zta/EB1	IE
BALF1	Viral-bcl-2 antagonist	EL
BALF2	Major DNA-binding protein (EA)	EL
BALF5	DNA polymerase	EL
BaRF1	Ribonucleotide reductase small subunit	EL
BARF1	Transformation-associated?	EL
BBLF2/3	Primase accessory protein	EL

BBLF4	Helicase	EL
BcRF1	Unknown	EL
BDLF4	gp115	EL
BFLF1	(DNA) Packaging protein UL32 homolog	EL
BFLF2	Nuclear membrane protein	EL
BFRF1	Virion egress (membrane/viron) protein UL34 homolog, 37kDa	EL
BFRF2	Virion egress	EL
BGLF4	Protein kinase	EL
BGLF5	Alkaline exonuclease	EL
BHLF1	Unknown	EL
BHRF1	bcl-2 homologue, vBcl2	EL
BKRF3	uracil DNA glycosylase	EL
BLLF2	Unknown	EL
BLLF3	dUTPase	EL
BMLF1	mRNA export factor ICP27 homolog	EL
BMRF1	DNA polymerase processivity factor (EA), accessory protein	EL
BNLF2a	Immune evasion	EL
BNLF2b	Potential gp141	EL
BORF2	Ribonucleoside-diphosphate reductase large subunit	EL
BRRF1	Transcription factor	EL
BSLF1	Primase	EL
BSLF2	SM, mRNA export factor	EL
BXLF1	Thymidine kinase	EL
LF3	Unknown	EL

BALF3	Glycoprotein transport?	LL
BALF4**	Envelope glycoprotein B gB/gp110/gp125	LL
BBLF1	Tegument protein	LL
BBRF1	Capsid protein	LL
BBRF2	Unknown	LL
BBRF3	Glycoprotein gM	LL
BcLF1	Major capsid protein	LL
BCRF1	viral IL-10, vIL10	LL
BDLF1	Triplex (/minor) capsid protein VP23 homolog	LL
BDLF2	Tegument protein	LL
BDLF3	Probable membrane antigen gp85/gp150 BDLF3	LL
BdRF1	Capsid protein p40	LL
BFRF3	Capsid protein p18	LL
BGLF1	Tegument protein	LL
BGLF2	38Kd protein	LL
BGRF1	Packaging protein	LL
BILF1	Glycoprotein gp60	LL
BILF2	Glycoprotein gp78/55	LL
BKRF2	Glycoprotein gL (gp25)	LL
BKRF4	Tegument protein	LL
BLLF1**	Envelope glycoprotein gp350/220 BLLF1	LL
BLRF1	Glycoprotein gN	LL
BLRF2	Tegument protein, capsid protein p23 (VCA)	LL
BMRF2	53/55Kd membrane protein	LL

BNRF1	Major tegument protein p143 (VCA)	LL
BOLF1	Tegument protein	LL
BORF1	may be needed for capsid assembly	LL
BPLF1	Large tegument protein	LL
BRRF2	Tegument protein	LL
BSRF1	Tegument protein	LL
BTRF1	Capsid maturation	LL
BVRF1	Tegument protein	LL
BVRF2	Protease	LL
BXLF2**	Envelope glycoprotein H gH/gp85	LL
BXRF1	Basic core protein	LL
BZLF2**	Glycoprotein 42 gp42	LL
BDLF3.5	Unknown	Unknown
BFRF1a	DNA packaging	Unknown
BGLF3	Unknown	Unknown
BGLF3.5	Tegument protein	Unknown
BVLF1	Unknown	Unknown
BWRF1	Hypothetical protein	Unknown
LF1	Unknown	Unknown
LF2	Unknown	Unknown