IN SILICO ANALYSIS OF PATHWAYS TARGETED BY EBV INFECTION AND MALIGNANT TRANSFORMATION

Ramakrishna Rao Sompallae
From the Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

IN SILICO ANALYSIS OF PATHWAYS TARGETED BY
EBV INFECTION AND MALIGNANT TRANSFORMATION

Ramakrishna Rao Sompallae

Stockholm 2009
Cover: Gene expression profiles from EBV negative and EBV positive, malignant and proliferating B-cell lines.

All published articles were reproduced with the permission from the publishers.

Published by Karolinska Institutet.

© Ramakrishna Rao Sompallae, 2009
ISBN 978-91-7409-693-4

Printed by
www.reproprint.se
Gårdsvägen 4, 169 70 Solna
To My Beloved Parents
and my brother and sister
**Abstract**

Epstein-Barr virus (EBV) is a ubiquitous γ-herpes virus with dual cell tropism for human B-lymphocytes and epithelial cells. EBV infection is linked to several malignancies such as Burkitt's lymphoma (BL) and nasopharyngeal carcinoma (NPC). In vitro EBV is a potent transforming virus that converts resting B-lymphocytes into indefinitely proliferating lymphoblastoid cells (LCLs).

The overall aim of this study was to develop and utilize bioinformatics methods to dissect the molecular mechanisms by which EBV modulates the cellular environment. EBV is a large DNA virus that encodes about 100 open reading frames (ORFs) expressed at various times during infection. Using established sequence search methods such as patterns and hidden Markov models (HMM), we have identified catalytic domains of ubiquitin specific proteases (deubiquitinating enzymes, DUBs) in a limited number of EBV ORFs. The DUB activity of three high scoring candidates: BPLF1, BSLF1 and BXLF1, was confirmed by functional and mutational analysis (Paper-I).

EBV establishes distinct programs of viral gene expression in latently infected and malignant cells. The EBV nuclear antigen (EBNA)-1 is expressed in all EBV carrying proliferating cells. EBNA-1 has been associated with the induction of cellular oxidative stress due to the production of reactive oxygen species (ROS) and DNA damage. To search for the cellular genes involved in these effects we used gene expression data sets from public databases. By analyzing the expression of genes involved in ROS metabolism in EBV positive and negative Burkitt’s lymphoma (BL) cell lines and lymphoblastoid cell lines (LCL) we found that the NADPH oxidase (NOX)-2 is induced in cells expressing EBNA-1. Activation of the NOX2 gene by EBNA-1 is associated with the induction and/or maintenance of genomic instability, a critical step in malignant transformation (Paper-II). The expression of EBNA-1 is associated with a global rearrangement of cellular transcription. To investigate the primary and secondary targets of this transcriptional effect we have analyzed the gene expression profiles of stable and inducible EBNA-1 expressing cells. Functional analysis of the regulated transcripts revealed that EBNA-1 influences the expression of genes involved in the maintenance of chromatin architecture. Several subunits of chromatin remodeling complexes were down-regulated on EBNA-1 expression (Paper-III).

Microarray analysis and systems biology approaches were implemented to investigate the cellular pathways modulated by tripeptidyl peptidase II (TPPII) a cellular protein that is highly expressed in BL cells and participates in the induction of the malignant phenotype. Comparison of the gene expression profiles of control and TPPII knock-down BL cells and systems level analysis of differentially regulated genes demonstrated that the MAPK signaling pathway is selectively inactivated by TPPII knockdown (Paper-IV).

In summary, this work demonstrated the potential of bioinformatics tools and high-throughput genomic approaches in identifying the novel strategies of EBV.

**Key words**: Bioinformatics, systems biology, gene expression analysis, Epstein-Barr virus, Burkitt's lymphoma.
LIST OF PUBLICATIONS

This thesis is based on the following articles and manuscript that will be referred to in the text by their roman numerals:

* The authors contributed equally

II  Gruhne B, Sompallae R, Marescotti D, Kamranvar SA, Gastaldello S, M asucci M G. The Epstein–Barr virus nuclear antigen-1 promotes genomic instability via induction of reactive oxygen species. Proc Natl Acad Sci. 2009 Feb 17;106(7):2313-8

III  Sompallae R, Callegari S, Gruhne B, Kamranvar SA, Marescotti D, M asucci M G. Epstein-Barr Virus Nuclear Antigen-1 expression influences transcription of cellular genes involved in chromatin remodeling. Submitted for publication

IV  Sompallae R, Stavropoulou V, Houde M, M asucci M G.
The MAPK Signaling Cascade is a Central Hub in the Regulation of Cell Cycle, Apoptosis and Cytoskeleton Remodeling by Tripeptidyl-Peptidase II. Gene Regulation and Systems Biology 2008:2 253-265
OTHER PUBLICATIONS

- Gruhne B, Sompallae R, Masucci MG. Three Epstein-Barr virus latency proteins independently promote genomic instability by inducing DNA damage, inhibiting DNA repair and inactivating cell cycle checkpoints. Oncogene. 2009 Aug 31 (Epub).
# TABLE OF CONTENTS

1 Synopsis .....................................................................................................................1
2 Introduction ................................................................................................................2
  2.1 Epstein-Barr Virus ..............................................................................................2
      2.1.1 Latent viral gene expression ....................................................................2
      2.1.2 EBNA-1 ..............................................................................................5
      2.1.3 EBNA-2 ..............................................................................................6
      2.1.4 EBNA-3 family of proteins ....................................................................6
      2.1.5 EBNA-LP ............................................................................................7
      2.1.6 LMP genes ..........................................................................................7
  2.2 EBV and the ubiquitin signaling .........................................................................8
      2.2.1 The ubiquitin proteasome system (UPS) .............................................8
      2.2.2 EBV modulates the ubiquitin system .................................................9
  2.3 EBV and genomic instability .............................................................................10
  2.4 Upregulation of TPPII in Burkitt’s lymphomas .............................................11
3 Aims .........................................................................................................................13
4 Methods ....................................................................................................................14
  4.1 Sequence analysis .............................................................................................14
      4.1.1 BLAST ...............................................................................................14
      4.1.2 Sequence pattern ...............................................................................15
      4.1.3 Hidden Markov model (HMM) .............................................................15
  4.2 Gene expression analysis ....................................................................................16
      4.2.1 DNA Microarrays ..............................................................................17
      4.2.2 Next generation Technologies .............................................................18
      4.2.3 Data analysis ......................................................................................20
      4.2.4 Biological interpretation of high-throughput data .............................20
      4.2.5 Promoter analysis ..............................................................................21
      4.2.6 Protein interaction networks ...............................................................21
      4.2.7 Public repositories of microarray data ................................................22
5 Results and Discussion .............................................................................................23
    EBV encodes at least three deubiquitinating enzymes (Paper – I) ................23
    EBNA-1 induces ROS production by activating NOX2 (Paper – II) ...........23
    EBNA-1 regulates chromatin remodeling complexes (Paper – III) ..........24
    TPPII modulates the MAPK pathway (Paper – IV) .....................................25
6 Concluding remarks ..................................................................................................27
7 Acknowledgements ..................................................................................................28
8 References .................................................................................................................30
ABBREVIATIONS

ATP Adenosine triphosphate
BL Burkitt’s Lymphoma
BLAST Basic local alignment search tool
Cy3 Cyanine 3
Cy5 Cyanine 5
DNA Deoxyribonucleic acid
DS Dyad symmetry
DUB Deubuquitinating enzymes
EBV Epstein-Barr Virus
EBNA EBV nuclear antigen
FR Family of repeats
GO Gene Ontology
HECT Homologous to the E6-A P Carboxyl Terminus
HHV Human herpesvirus
HL Hodgkin’s lymphoma
HMM Hidden Markov model
LCL Lymphoblastoid cell lines
LMP Latent membrane protein
KEGG Kyoto Encyclopedia of Genes and Genome
mRNA Messenger ribonucleic acid
MAPK Mitogen activated protein kinase
NADP Nicotinamide adenine dinucleotide phosphate
NCBI National Center for Biotechnology Information
NFκB Nuclear factor kappa B
NPC Nasopharyngeal carcinoma
ORF Open reading frame
OriP Origin of replication
Q-PCR Quantitative real-time polymerase chain reaction
ROS Reactive oxygen species
SAM Significance analysis of microarrays
TPPⅡ Tripeptidyl peptidase II
TR Terminal repeats
UPS Ubiquitin proteasome system
1 SYNOPSIS

Over the course of evolution, viruses have developed highly sophisticated mechanisms for interacting with their host cells. Such interactions involve the modulation of many signaling pathways, which results in global reshaping of the cellular environment that promotes virus persistence or replication. A classic example is provided by Herpesviruses that have co-evolved with their hosts for millions of years thus achieving a balance between viral persistence and immune control. Such viral effects can in some cases transform the host cell and promote tumorigenesis. Therefore, a better understanding of the molecular basis of virus-host interactions is essential for the identification of potential targets for rational drug design.

The modern high-throughput techniques complements the classical single gene studies by allowing investigation of dynamic cellular mechanisms on a global scale and provides systems-level understanding of physiological states. Data analysis is the crucial task of these experiments and requires bioinformatics tools to achieve biologically meaningful conclusions in a short time scale. The revolution in computer technology and memory storage capability has made it possible to organize and model such large data and quicken the analysis. With the growing number of such studies, a huge amount of data is generated and stored in publicly available databases for meta-analysis. Although, these computational methods do not substitute for wet lab techniques they provide an important motivation for experimental work.

This thesis is based on the implementation of various bioinformatics methods combined with the experimental studies to gain new insight into the molecular basis of EBV infection and malignant transformation. A bioinformatics strategy based on the identification of functional domains and optimized stringency levels has been developed to identify homologues of cellular enzymes encoded by viruses and other infectious agents whose proteins share little homology with their infected hosts. In addition, purpose made and publically available gene expression arrays have been mined in order to identify virus-induced changes in cellular gene expression that will help in understanding viral strategies for cellular adaptation and cellular signaling pathways that are critically involved in the regulation of malignant cell traits.

This work illustrates the potential of high content data collection and bioinformatics analysis tools in deciphering the relationship between viral proteins and cellular pathways affected.
2 INTRODUCTION

2.1 EPSTEIN-BARR VIRUS

The Epstein-Barr virus (EBV) (also known as human herpesvirus 4, HHV4) is a lymphotoptic human DNA \( \gamma \)-herpes virus that is widespread in all human populations, with over 90% of the adults being lifelong carriers (Rickinson and Kieff, 2007). Primary EBV infection is common during early childhood and is usually clinically silent. However, when infection is delayed until adolescence or adulthood, it may cause infectious mononucleosis (IM), a self-limiting lymphoproliferative disease (Young and Rickinson, 2004). EBV is known to be involved in the pathogenesis of a broad spectrum of malignancies of lymphoid and epithelial cell origin (Rickinson and Kieff, 2007). The virus was first detected in biopsies of endemic Burkitt’s lymphoma (BL) (Epstein et al., 1964), but was later implicated in a number of diseases including nasopharyngeal carcinomas (NPC), Hodgkin’s lymphomas (HL), NK- and T-cell lymphomas and gastric carcinomas (Thompson and Kurzrock, 2004). Epidemiological and molecular studies link EBV to human malignancies but the cellular mechanisms involving in pathogenesis are poorly understood.

The 172 kb long double-stranded DNA of EBV genome encodes more than 100 open reading frames (ORFs) of which only few are well studied (Figure 1) (Rickinson and Kieff, 2007). The viral genome does not normally integrate into the cellular DNA but persists as circular episome. Circularization is mediated via terminal repeats (TR) of approximately 500 bp present in EBV genome (Figure 2). Like many other viruses that are adapted to persist in the infected host, EBV replicates and hides in different cellular compartments where it follows lytic or latent expression programs, respectively. The lytic program expresses most of the EBV genes and results in the production of infectious viral particles whereas the latent program is characterized by the expression of a limited repertoire of viral proteins. These viral programs result in three different types of latencies (Table 1) that contribute to avoid activation of the immune responses and to promote co-existence with the host (Dolcetti and Masucci, 2003).

2.1.1 Latent viral gene expression

The three different programs of EBV latent gene expression are displayed in associated tumors and in vitro transformed B-cells. The activity of latent EBV promoters C, W and Q promoters of EBV nuclear antigens (EBNAs) and promoter of latent membrane proteins (LMPs), defines the pattern of latent protein expression.
Figure 1. Graphical representation of the linear EBV genome and open reading frames (ORF) (adapted from NCBI genomes). The ORFs are indicated in green according to the scale. Protein coding regions are indicated in red.
The type I latency is characterized by the expression of EBNA-1 alone. The classical example with the “EBNA-1 only” latent gene expression is the endemic BLs (Rowe et al., 1987). In latency II, the expression of viral proteins is limited to EBNA-1 and the LMPs (Rowe et al., 1992). This program is expressed in EBV associated HL, NPC and NK/T-cell lymphomas (Rickinson and Kieff, 2007). In both type-I and -II latency programs EBNA-1 expression is initiated from Qp promoter (Figure 2) (Rickinson and Kieff, 2007).

Figure 2. Episomal form of the EBV genome. The terminal repeats (TR) are formed during the circularization of the linear DNA. The outer line represents the single giant mRNA where all the EBNA's are transcribed from either the Cp or Wp promoter. The inner line represents the EBNA1 transcript originating from the Qp promoter.

Table 1. EBV latent gene expression patterns

<table>
<thead>
<tr>
<th>Latent</th>
<th>Viral gene expression</th>
<th>Malignancies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>EBNA-1</td>
<td>BL</td>
</tr>
<tr>
<td>Type II</td>
<td>EBNA-1, LMP-1, 2A and 2B</td>
<td>NPC, cHL, NK/T cell</td>
</tr>
<tr>
<td>Type III</td>
<td>EBNA-1, 2, 3A-C, LP, LMP-1, 2A and 2B</td>
<td>PTLD, IM</td>
</tr>
</tbody>
</table>

EBV latent protein expression patterns and malignancies associated with the type of infection. (EBNA, Epstein-Barr virus nuclear antigen; LMP, latent membrane protein; BL, Burkitt’s lymphoma; cHL, classical Hodgkin lymphoma; NPC, nasopharyngeal carcinoma; PTLD, post-transplant lymphoproliferate disorder; IM, Infectious mononucleosis)

The type III latency is the least restrictive program and results in the expression of six EBNA's (EBNA-1, 2, 3A-C and LP) and three latent membrane proteins (LMP-1, 2A, 2B). The six nuclear antigens are transcribed as a single giant mRNA from adjacent promoters Cp and Wp (Figure 2) (Rowe, 1999). The long transcript is differential spliced to generate individual mRNAs that encode different EBNA's (Rowe, 1999).
This expression pattern is observed in B-cells that are immortalized by EBV infection in vitro, giving rise to lymphoblastoid cell lines (LCLs) (Young and Rickinson, 2004). The transcription of LMP genes is driven from two distinct promoters located close to the TR (Rickinson and Kieff, 2007). In addition to the latent proteins, EBV expresses non-coding RNAs (EBERs) (Rickinson and Kieff, 2007) and microRNAs (miRNAs) (Pratt et al., 2009) in all latently infected cells.

2.1.2 EBNA-1

EBNA-1 is the only viral protein ubiquitously expressed in all EBV-positive malignancies either alone or together with other viral proteins (Thompson and Kurzrock, 2004). This 641 amino acids long protein is crucial for replication of the viral genome and partitioning of the replicates during host cell mitosis. Due to its association with development of lymphomas and with increasing metastatic ability of cancer cells in mouse models, EBNA-1 has been suggested to be oncogenic (Rowe, 1999; Sheu et al., 1996; Wilson et al., 1996).

EBNA-1 has specific DNA binding regions that are required for the viral episome replication and persistence in latently infected cells. EBNA-1 binds to the latent origin of replication (oriP) present in the EBV genome. OriP contains multiple copies of EBNA-1 recognition sites clustered into the family of repeats (FR) and the dyad symmetry (DS) elements (Rickinson and Kieff, 2007). FR is composed of 20 copies of high-affinity EBNA-1 binding sites while DS has 4 copies with lower binding affinity than FR (Rawlins et al., 1985; Rickinson and Kieff, 2007). Each of these sites contains 16 bp sequences that are required for EBNA-1 binding. EBNA-1 forms stable homodimers and binds to oriP through DNA-binding domain located in the C-terminal part of the protein (Figure 3). EBNA-1 also contains three Arg-Gly (RGG) rich regions that are interrupted by the long stretch of Gly-Ala repeat (GAr) sequence. The RGG domains are mainly involved in associating the viral episome to host chromosomes (Ambinder et al., 1991; Chen et al., 1993; Hennessy and Kieff, 1983). This key function of EBNA-1 in replication and segregation of viral DNA supports the need for its consistent expression in all EBV carrying proliferating cells.

Figure 3. Schematic representation of DNA binding domains and Gly-Ala repeat (GAr) region present in EBNA-1. The RGG motif regions correspond to Arg-Gly rich sequence through which EBNA-1 binds to cellular DNA. NLS defines nuclear localization signal. The oriP binding and transactivation domain binds to specific sequence in the viral genome.
The binding of EBNA-1 to DNA also plays a role in viral gene transcription both as an activator and as repressor. EBNA-1 through its C-terminal oriP binding and transactivation domain (Figure 3), binds to the FR elements in the viral episome and enhances the transcription of other latent genes from the Cp and LMP-1 promoters (Leight and Sugden, 2000; Rickinson and Kieff, 2007). It also exerts a negative feedback regulation on its own promoter Qp (Rickinson and Kieff, 2007; Schaefer et al., 1995; Tsai et al., 1995). Several studies on cell lines and transgenic mice have elucidated the role of EBNA-1 on cellular gene expression (Flavell et al., 2008; Gruhne et al., 2009; Tsimbouri et al., 2002; Wood et al., 2007). EBNA-1 seems to target wide range of cellular genes involved in key signaling pathways that regulates cell proliferation, death and survival implicating role of EBNA-1 in B cell transformation (Dresang et al., 2009; Flavell et al., 2008; Tsimbouri et al., 2002; Wood et al., 2007). These observations were further supported by the ability of EBNA-1 to bind the promoter sequences of several cellular genes (Dresang et al., 2009).

2.1.3 EBNA-2

EBNA-2 is the first latent gene expressed after primary infection and is essential for EBV-induced immortalization of the B cells (Cohen et al., 1989; Hammerschmidt and Sugden, 1989). This 487 amino acid long nuclear protein is a key regulator of both viral and cellular gene expression. EBNA-2 activates transcription of all EBV latent genes from the major latency Cp promoter and the enhancer element shared by LMP genes (Ghosh and Kieff, 1990; Jin and Speck, 1992). EBNA-2 induces expression of cellular genes such as CD23, CD21, c-fgr and c-myc, which are likely to be crucial for immortalization (Maier et al., 2006). EBNA-2 also interacts with cellular recombination site binding protein (RBP)-Jκ to activate the Notch signaling pathway continuously (Kohlhof et al., 2009; Zimber-Strobl et al., 1994).

2.1.4 EBNA-3 family of proteins

The EBNA-3 family of proteins EBNA-3A, -3B and -3C (also known as EBNA-3, -4 and -6) are encoded by tandem genes and have partial sequence similarities (Rickinson and Kieff, 2007; Yenamandra et al., 2009). They all associate with cellular proteins and may act as transcriptional regulators (Robertson et al., 1996). EBNA-3C is able to activate the LMP1 promoter (Allday et al., 1993). It has been suggested that EBNA-3 proteins interact with cellular cell cycle proteins to disrupt the G2/M checkpoint (Krauer et al., 2004).
2.1.5 EBNA-LP

EBNA-LP (EBNA-leader protein, also known as EBNA-5) is a multi-repeat domain containing protein with the repeats of 22 and 44 amino acids derived from W1, W2 exons and Y1, Y2 exons respectively (Rickinson and Kieff, 2007). The size of the protein varies by varying the number of repeats in the coding region. EBNA-LP forms complexes with a cellular anti-apoptosis protein BCL2 or its EBV counterpart, BHRF1, and thus plays a role in the regulation of cell death by apoptosis in virus-infected cells (Matsuda et al., 2003). There is accumulating evidence of multifunction of EBNA-LP, although the function in EBV biology is not yet fully understood.

2.1.6 LMP genes

LMPs are integral membrane proteins that lack significant extracellular domains and mimics cellular receptors. LMP-1 is a functional homolog of the human CD40 and forms a constitutively active receptor for tumor necrosis factor (TNF) ligands (Bornkamm and Hammerschmidt, 2001). It is expressed in most EBV related malignancies except Burkitt’s lymphoma. LMP-1 is essential for EBV immortalization and may transform cells in vitro (Young and Rickinson, 2004). LMP-1 activates several signaling pathways that up-regulates anti-apoptotic proteins such as Bcl-2, Mcl1, A20 and provides growth signals (Soni et al., 2007). The LMP-2A, mimics active forms of the B-cell receptor (BCR) (Bornkamm and Hammerschmidt, 2001). It inhibits the activation of virus lytic cycle by continuously stimulating the tyrosine kinases of the BCR signal transduction and protects from apoptosis (Bornkamm and Hammerschmidt, 2001). The function of LMP-2B is less clear. It shares the trans-membrane domain with LMP-2A but lacks the N-terminal domain involved in BCR signaling.

Most of the studies related to EBV are focused on the latent viral genes, since they are the most prominently expressed proteins in EBV-associated malignancies. This leaves a large number of EBV ORFs poorly understood in terms of their function. As the information of protein function and structure will aid the understanding of molecular pathways targeted during infection, it is important to predict function and structure of these viral proteins. We have attempted to screen the EBV genome for detecting distant functional homologs of key cellular enzymes (Paper-I). Using various sequence analysis methods and adjusting the cut-offs on statistical scores, we detected functional domains of ubiquitin proteases in ORFs encoded in the EBV genome.
2.2 EBV and the ubiquitin signaling

2.2.1 The ubiquitin proteasome system (UPS)

Protein ubiquitination is a posttranslational modification that can define the protein stability, localization, activity and interactions. It is now clear from various studies that ubiquitination is essential for fundamental cellular processes such as cell cycle progression, signal transduction and apoptosis. Dysregulation of the ubiquitin signaling may also contribute to tumor progression through perturbation of key cellular processes (Hoeller and Dikic, 2009).

Conjugation of ubiquitin to the substrate protein proceeds via a three-step mechanism. Initially, ubiquitin is activated by the ubiquitin-activating enzyme (E1) and then transferred by one of the ubiquitin-conjugating enzymes (E2) to a member of the ubiquitin ligase family (E3) or to the substrate protein with the help of E3 (Pickart and Cohen, 2004). Thus, the E3 enzymes catalyze the covalent attachment of ubiquitin to the substrate. The process can be repeated to build a chain of four or more ubiquitin molecules on specific signaling. Following the attachment of ubiquitin, the substrate proteins are either sorted for the degradation by 26S proteasome or triggered to perform specific function depending on the type of ubiquitination (Finley, 2009; Ravid and Hochstrasser, 2008). Ubiquitin is then released and recycled by the help of specific enzymes known as deubiquitinating enzymes (DUBs). DUBs play important roles both in reversing the signal triggered by ubiquitination and in recycling ubiquitin to maintain the steady-state levels of free ubiquitin (Hussain et al., 2009; Ventii and Wilkinson, 2008).

The human genome contains ~95 DUBs classified into five subfamilies (Nijman et al., 2005). Four of these families: ubiquitin specific protease (USP), ubiquitin C-terminal hydrolase (UCH), ovarian tumor-related protease (OTU) and Josephin-domain protease (JD) belong to cysteine proteases and contain the catalytic residues cysteine and histidine. While, the fifth family containing Jab1/MPN/Mov34 metalloenzymes (JAMM s) need Zinc ion for catalysis (Nijman et al., 2005). One of the classical examples to illustrate the significance of DUBs as regulators of protein stability, is USP7/Herpes virus associated USP (HAUSP). HAUSP regulates the turnover of p53 and provides a means to regulate cell growth (Li et al., 2002). The DUB activity has been demonstrated in proteins encoded by several viruses and pathogenic bacteria that lack an UPS suggesting that these enzymes play specific roles in the regulation of both viral and bacterial infection (Isaacson and Ploegh, 2009; Lindner, 2007; Rytkonen and Holden, 2007). Prediction of these enzymes is complicated and challenging due to the wide sequence variation present among members of different families. We have developed a combinatorial bioinformatics
approach using catalytic domain sequences from different subfamilies to identify
distal homologs across viral genomes (Paper-I). Such methods will identify the
functions of viral proteins and may help in understanding viral strategies.

Normally, substrate proteins with polyubiquitin chains are degraded by 26S
proteasome in an ATP-dependent manner. The proteasome is a large multisubunit
complex consisting of a 20S core particle (CP) and two 19S regulatory particles (RP)
(Finley, 2009; Voges et al., 1999). The proteasome contains three catalytic activities,
the trypsin-, the chymotrypsin- and the post glutamyl peptidyl hydrolytic- (or the
caspase) -like that are secured inside the 20S core particle. The 19S regulatory
subunit covers both sides of the 20S CP and mediates key functions such as
recognition and unfolding of the target proteins to make them access to the catalytic
core (Finley, 2009).

2.2.2 EBV modulates the ubiquitin system

Considering the broad range of substrates and processes in which ubiquitination is
involved, it is not surprising that viruses take advantage of this pathway in various
ways. Many viruses have shown to encode proteins that mimic, block or redirect the
activity of cellular enzymes of the ubiquitin pathway. The first evidence of such viral
protein intrusion in ubiquitin signaling was demonstrated by E6 oncoprotein of
human papilloma virus (HPV). The E6 protein forms a complex with the cellular E3
ligase E6-AP and specifically promotes p53 ubiquitination a feature important for
cellular transformation (Scheffner et al., 1990). Several EBV proteins interfere with,
or use, the ubiquitin pathway to avoid recognition by the immune system and to
stimulate proliferation (Masucci, 2004). The EBV nuclear protein EBNA-1 contains a
long Gly-Ala repeat (GAR) domain that can protect against proteasomal degradation
(Levitskaya et al., 1997). This inhibition of EBNA-1 degradation helps the virus to
escape immune recognition in latent infections (Blake et al., 1997; Levitskaya et al.,
1995). LMP-1 and LMP-2A are also involved in the modulation of host cell
ubiquitination. LMP-1 upregulates several subunits of the proteasome and results in
altered enzymatic activity (Frisan et al., 1998), which promotes antigenicity of EBV
infected cells. LMP-2A regulates E3 ligases of the neural precursor cell expressed,
developmentally down-regulated (Nedd) 4 family and promotes ubiquitination and
degradation of the effector kinases of BCR (Ikeda et al., 2000; Winberg et al., 2000).
LMP-2A thus blocks BCR signaling and inhibits lytic cycle activation. It is now
evident from several other studies that viruses encode one or more enzymes of
 cellular ubiquitin proteasome pathway (Chen and Gerlier, 2006; Randow and Lehner,
2009). Our systematic screening of the EBV genome for the functional domains of
DUBs has identified three EBV ORFs with bona-fide deconjugase activity (Paper-I). Thus, several of the EBV encoded proteins are involved in the regulation of ubiquitin pathway to help the virus to survive, replicate and promote tumorigenesis.

2.3 EBV AND GENOMIC INSTABILITY

Maintaining genome stability is crucial for cell growth and cell survival. Different genetic disorders, including most human cancers, are associated with different types of genomic instability. In a tumor cell, genomic instability could result from high incidence of DNA damage overrunning the ability of the normal repair systems to restore, or from the inactivation of one or more DNA repair pathways (Schmutte and Fishel, 1999). Chromosome instability (CIN) is characterized by loss or rearrangement of chromosomes, resulting in aneuploidy, translocations and loss of heterozygosity (LoH) (Honma, 2005). Chromosomal translocations are the most common genomic alterations observed in malignant cells. Certain chromosomal translocations can lead to the creation of a fusion gene, encoding a chimeric protein (Aman, 2005). Some of these fusion genes have proven to be specific for different tumors type (Aman, 2005) and can therefore be used as diagnostic markers as well as possible therapeutic targets.

Chromosomal aberrations can be either lethal and are generated de novo at each cell cycle (Kadhim et al., 1992) or non-lethal and are transmitted to the progeny (Heng et al., 2006). Both types of chromosomal abnormalities are prominently observed in tumor cells (Heng et al., 2006). Chromosome aberrations and genome instability has been observed in EBV-carrying cells and associated tumors (Bernheim et al., 1980; Shao et al., 2001; Weniger et al., 2006; Zech et al., 1976). Recent reports have showed that EBV-infection is associated with 3-10-fold increase of chromosomal aberrations such as dicentric chromosomes, chromosome fragments and gaps (Figure 4) (Kamranvar et al., 2007). Interestingly, these chromosomal aberrations appeared in cell lines expressing different forms of latencies. A significant number of chromosomal aberrations observed in cell lines expressing latency type I suggests that EBNA-1 alone can contribute to this phenotype (Kamranvar et al., 2007). However, the contribution of the virus or viral proteins to the molecular events that leads to this malignant cell phenotype is poorly understood.

In cells with latency III EBV gene expression, it was suggested that EBNA-3C (Parker et al., 2000) and LMP-1 (Chen et al., 2008) might promote genomic instability through inhibition of DNA repair or inactivation of cell cycle checkpoints, which would allow the observed propagation of DNA damage. As a part of this work,
we have elucidated the capacity of EBNA-1 to induce the mutator phenotype of genomic instability (Paper -II).

Figure 4. Chromosomal aberrations detected in EBV negative and carrying cells. Expression of EBNA-1 only is sufficient for induction of genomic instability. (adapted from kamranvar et al, 2007).

2.4 UPREGULATION OF TPPII IN BURKITT’S LYMPHOMAS

BL is a highly malignant B-cell tumor that occurs with high frequency in tropical areas of Africa and New Guinea (endemic BL) where it is consistently EBV-associated, and with lower frequency all over the world (sporadic BL) where its EBV-association is less strong (Klein, 1994). Regardless of its geographical origin, the tumor invariably carries a chromosomal translocation involving the c-myc gene on chromosome 8 and one of the immunoglobulin genes heavy or light chain loci on chromosomes 14, 2 or 22, respectively. BL is regarded as the prime example of a latency I associated malignancy with only EBNA-1 expression. The rare occurrence of EBV-negative variant of the tumor suggests that EBV infection acts as a cofactor in the pathogenesis of this malignancy.

Previous studies on cell lines that exhibit in vitro BL cell phenotype showed that overexpression of c-myc is directly correlated with alterations of ubiquitin-dependent proteolysis (Gavioli et al., 2001). The enzymatic activity of the proteasome is significantly reduced in BL cells but these cells are nevertheless resistant to high doses of proteasome inhibitors that are readily toxic for EBV immortalized B-lymphoblasts (Frisan et al., 1998; Gavioli et al., 2001). Furthermore, proteolysis proceeds virtually undisturbed in the presence of high doses of proteasome inhibitors, suggesting that other enzymatic partners may contribute to maintain protein turnover. The upregulation of TPPII and several deubiquitinating enzymes in BLs and in c-Myc...
overexpressing cells suggested a possible role for these enzymes in proteolysis (Gavioli et al., 2001). Recent studies have demonstrated that the c-Myc protein is stabilized in a number of BL cell lines, suggesting that defective UPS-dependent c-Myc proteolysis may play an important role in the tumorigenesis (Gregory and Hann, 2000).

TPPII is a serine peptidase with endoproteolytic activity that removes tripeptides from free N-terminus of oligopeptides (Tomkinson and Jonsson, 1991). This 138 kDa protein forms oligomeric complexes that are larger than the 26S proteasome (Balow et al., 1983). It has been proposed that TPPII may act downstream of the proteasome and accelerate the production of free amino acids from longer precursors with the help of other exopeptidases (Wang et al., 2000). The role of TPPII in antigen processing was suggested by its involvement in the production of antigenic peptides in cells with impaired proteasome activity (Seifert et al., 2003). TPPII plays a more specific role as a cholecystokin-inactivating enzyme in rat brains (Rose et al., 1996) and in apoptotic pathways (Chateau et al., 2001). However, cells with high TPPII activity and low proteasome activity seems to provide growth advantage to cells by altering the degradation of cellular proteins mainly inhibitors of apoptotic proteins (IAPs) (Hong et al., 2003). Overexpression of TPPII in HEK 293 cells enhanced growth rates and promoted a resistance to apoptosis (Stavropoulou et al., 2005). These cells also evaded mitotic arrest induced by spindle poisons and accumulated chromosomal aberrations (Stavropoulou et al., 2006). On the other hand the knock-down of TPPII delayed cell growth and induced apoptosis (Stavropoulou et al., 2006). The role for TPPII in proliferation and survival was also confirmed by the studies involving knock-out mice (Firat et al., 2007). Taken together these findings suggest that TPPII may contribute to tumor cell proliferation by inhibiting apoptosis and inducing cell growth in BLs. It remains unclear how TPPII can coordinate these functions in tumor cells. In this thesis we have used microarray based gene expression profiles to identify the cellular pathways modulated after TPPII knock-down in BL cells that generally have high levels of TPPII expression (Paper-IV).
3 AIMS

The general aim of the work presented in this thesis was to apply bioinformatics methods for investigation of the cellular pathways targeted by EBV during infection and malignant transformation.

More specifically, we aimed to

- develop bioinformatics sequence analysis method for predicting distant homologs of cellular enzymes.
- investigate the gene expression data for identifying the cellular processes affected by the viral proteins.
- investigate the signaling pathways modulated during malignant transformation.
4 METHODS

The full details of experimental methods and analysis are available in the published papers and manuscripts. Here I describe the general bioinformatics tools used in analyzing large datasets.

4.1 SEQUENCE ANALYSIS

The advancement of sequencing technologies in the past decade have made it possible to rapidly sequence hundreds of genomes from viruses, bacteria to animals, including humans. Much of this sequence data is uncharacterized, and most encode proteins with unknown function. One of the most effective means of inferring the function of an unidentified sequence is to find the homologous proteins or domains that are conserved for important structural and functional reasons. At the protein level, functional domains have highly conserved sequences within species subgroups and less across other species. Many methods have been proposed to predict functional domains that confer function to a novel protein. These methods are helpful to detect homologs between less evolved species that have low evolutionary substitution rates and fail to detect distant homologs across viruses and other microorganisms that evolve with higher frequency of sequence variation (Pybus and Rambaut, 2009).

Although there has been lot of efforts made to detect and characterize the products of viruses, there are still many viral ORFs with unknown function. The most widely used means of inferring function involves performing a sequence comparison between a query sequence and a functionally characterized sequence in a protein database. Since homology inferences are based upon statistical scores, it becomes uncertain to detect the distal homologs when the sequence identity is weak. In order to discover such remote homologs we need to optimize the stringency levels and restrict searches to the functional domains.

There exist two general classes of techniques used in searches for protein homologs, mainly pairwise sequence comparisons such as i). Basic Local Alignment Search Tool (BLAST), and family based comparisons such as ii). Motif or pattern searches and iii). Hidden Markov model (HMM) profile analysis.

4.1.1 BLAST

For homology detection using pairwise sequence comparison, BLAST is the commonly used program. The BLAST has been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a
BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits (Altschul et al., 1990).

### 4.1.2 Sequence pattern

The family based methods leverage the information contained in a set of proteins that are known to be homologous. Patterns are family based regular expressions generated from highly conserved sequence regions (Bairoch, 1992). They are derived based on the occurrence of particular amino acids that were conserved throughout the members of the protein families. Such sequence patterns can uncover exactly or highly conserved regions with fixed or variable spacing. However, because patterns are derived from existing sequences this technique might have limitation to detect remote homologs.

### 4.1.3 Hidden Markov model (HMM)

Profile based methods such as HMM are more sensitive than pairwise methods and involves two steps, first building a statistical model from the set of aligned sequences and then comparing that model to query sequences. HMMs are more sophisticated than other profile based sequence search methods and have been demonstrated to be very effective in detecting remote homologs (Eddy, 1996). They make use of position-specific scoring matrices that represents the distribution of amino acids at each position in the domains of protein families that were used for training HMM (Eddy, 1998). After training with a set of proteins, HMMs can identify positions of amino acids which describe conserved domains and can discriminate the query sequence with assigning a statistical score (Schuster-Bockler and Bateman, 2007).

The typical HMM is a chain of match (square), insert (diamond) and delete (circle) states, each of which corresponds roughly to a position (column) in the alignment from which it was built (Figure 5). A profile HMM has several types of probabilities, transition probabilities from one state to next, emissions probabilities trained from the occurrence of a given residue existing at that position in the alignment (Eddy, 1998).

![Figure 5](image.png)
along that path. Query sequence is then compared to the model and assigned a score; the ability of these scores is to separate true from false homologs, and to accurately predict the level of certainty.

The HMM bit score is a log-odds score in log base two. Specifically, it is:

$$S = \log_2 \frac{P(seq|HMM)}{P(seq|null)}$$

$P(seq|HMM)$ is the probability of the query sequence according to the HMM generated from a protein family while $P(seq|null)$ is the probability of the query sequence given a “null hypothesis” model of the random sequence. Thus, a positive score means that the HMM is a better model of the query sequence than the null model is.

In this thesis we described a combinatorial approach using established sequence analysis methods (Figure 6) to screen the EBV ORFs for remote homologs of cellular enzymes.

![Figure 6](image)

**Figure 6.** A schematic diagram showing the three major sequence search methods used to screen EBV ORFs for remote homologs of deubiquitinating enzymes. Candidates identified using sequence search methods were then filtered depending on the domain location, gene expression and localization.

### 4.2 GENE EXPRESSION ANALYSIS

Gene expression is a dynamic process by which the DNA is transcribed into messengerRNA (mRNA), which then is translated into proteins. Transcription not only depends on genetic factors such as DNA sequences but is also influenced by epigenetic factors that include chromatin remodeling complexes (Wang et al., 2007). The cell phenotype is mainly characterized by the global gene expression. As the
conditions change in a cell, the transcription and translation of literally hundreds of genes may be altered. The total number of possible RNA transcripts that expressed in a cell is defined as transcriptome. The high-throughput experimental methods allow “Genome-wide” analysis of the transcriptome at a specific time to give a global picture of all molecular events.

The classical low-throughput techniques for quantifying the products of gene transcription include northern blotting and polymerase chain reaction (PCR) (Saiki et al., 1989). In the mid-1990s, the high-throughput technologies have been developed to quantify several genes at the same time. These techniques mainly belong to either hybridization or sequencing methods. The DNA microarrays belong to the hybridization methods (Schena et al., 1995) while the latter includes serial analysis of gene expression (SAGE) (Velculescu et al., 1995) and the next generation sequencing platforms (Ansorge, 2009).

### 4.2.1 DNA Microarrays

The high-density DNA microarrays have become most popular and are widely used in many laboratories (Auer et al., 2009). This increasing use and acceptance of microarrays to study various genetic and cellular processes is clearly demonstrated by the growing number of citations appearing in the published literature. The two principle platforms that are widely used are the spotted arrays and the in situ synthesized arrays. The spotted arrays contain typically long sequences (complete cDNA clones) chemically stabilized onto the glass slides (Schena et al., 1995). They can be customized with the gene sets and can be produced in-house. The in situ synthesized arrays are commercially manufactured by fabricating short oligonucleotides directly onto the glass slide (Lockhart et al., 1996). Probes are designed to give optimal hybridization conditions considering parameters such as melting temperature and sequence uniqueness (Hardiman, 2004). Agilent and Affymetrix are the two main providers of the in situ synthesized arrays. In this thesis, we have used both Agilent and Affymetrix platforms in order to understand the ‘systems-level’ properties of viral and cellular proteins (Paper-III and -IV).

**Agilent microarray technology:** The Agilent produces microarrays by in situ printing of 60 nucleotides probes (Hughes et al., 2001). The probe design relies on multiple up-to-date and publicly available sequence databases. The Agilent sample preparation protocol relies on direct labeling; one (Cy3-labeled) or two (Cy3- and Cy5- labeled) samples are usually hybridized at a time (Wolber et al., 2006). Alternatively, indirect labeling techniques can also be successfully used. The electronic images produced during the scanning can be analyzed by the use of
different algorithms and software. Agilent feature extraction methods aim at quantifying the feature signals and the background, performing the background subtraction, normalizing the dye effect, and computing the log ratios and their error estimates. More recently, Agilent has also introduced the multiplex technology, where multiple sets of probes printed onto the same slide can be independently assayed (Wolber et al., 2006).

**Affymetrix GeneChip technology:** In the Affymetrix GeneChip technology, 25mer oligonucleotide probes are directly synthesized on the surface of the arrays by the use of photolithography technology (Lockhart et al., 1996). Affymetrix GeneChips are the most frequently used microarrays for expression profiling. Multiple independent oligonucleotides (20, 16, or 11 couples according to the chipset) are designed in silico, from available sequence databases, to hybridize to different regions of the same transcript. In addition, to each perfect match (PM) probes; mismatch (MM) probes having a different base in the 13th position are also designed. This MM probes serve as controls for specific hybridization and they should facilitate the direct subtraction of background and cross-hybridization signals. All the probe pairs, constituted by a PM probe with its own MM partner probes for each transcript are referred to as probe set. Affymetrix is now providing exon arrays to enable the gene expression analysis together with the alternative splicing.

### 4.2.2 Next generation Technologies

Hybridization methods introduce a systematic variability, which decreases the precision of expression measures. The other drawback is that even the most recent arrays, cannot capture the whole complexity of the genome or transcriptome. Therefore, novel methods are being developed to monitor gene expression with increased sensitivity and specificity. The new class of unbiased methods called as “Massively parallel sequencing”, promises to better handle the ambiguity of complex gene expression (Tucker et al., 2009). This massively parallel sequencing (MPS) technique will provide increasingly high-resolution analyses of sequences compared to the DNA microarrays. Currently, four different MPS platforms are available from different companies; the most widely used being 454 sequencing and Illumina sequencing. The common technological feature is to sequence a huge number of DNA fragments that are spatially separated in a flow cell.

**The 454 platform:** The 454 Life Sciences platform is based on sequencing-by-synthesis with pyrosequencing technology (Margulies et al., 2005). It was originally developed to speed up the whole genome sequencing. Sequencing based gene expression analysis in general has the advantage that novel transcripts can be
discovered in addition to providing a high specificity. Briefly, the methodology includes linker ligation and immobilization of fragments onto microbeads, followed by PCR amplification of individual transcripts in emulsions of water in oil, where each droplet is deposited in its own well in a “picotiter plate” flowcell containing <1.6 million 75-picoliter wells. Sequence is obtained by iterative pyrosequencing (Nyren et al., 1993), whereby wells are loaded once with bead-tethered sequencing enzymes (polymerase, sulfurylase, and luciferase), and buffer containing one of four dNTPs is passed horizontally over the wells. If there is a match to the primed template, polymerase incorporates the nucleotide and releases a pyrophosphate molecule which, when converted to ATP by sulfurylase, generates a luciferase-catalyzed luminometric signal (Ronaghi et al., 1996; Ronaghi et al., 1998). After washout of residual nucleotides, the cycle is repeated with the next dNTP. The 454 sequencing enables the collection of around 1 million sequence reads of up to 400 bases with an accuracy of 99.96%.

**Illumina sequencing:** This is the first of the massively parallel short-read platforms developed by Illumina. In this platform, picotiter plate is replaced with a glass slide coated with oligonucleotide anchors (Michael et al., 1998). DNA to be sequenced is applied to the surface and allowed to hybridize to the anchors, and a first round of sequencing is performed. Then, amplification is performed by the replicated (attached) sequence “bridging” over to adjacent anchors in successive rounds of replication forming a “cluster” of replicated sequences around the original one. Illumina sequencing produces many more reads than 454 sequencing, and has considerably lower cost. The method provides read lengths of less than 50 bases making whole-genome sequencing and assembly inefficient. For experiments where the composition of a selection of sequences from a known genome is under analysis (for example when sequencing chromatin immunoprecipitated together with a transcription factor), the method is ideal - especially when strategies to amplify the ends of a sequence “paired-end sequencing” are used (Korbel et al., 2007).

Of the platforms that are presently commercially available, the two latest addition are the SOLiD (Supported Oligonucleotide Ligation and Detection) instrument from Applied Biosystems (Tucker et al., 2009) and the Helicos platform (Braslavsky et al., 2003). The target amplification in SOLiD is similar to 454 but the sequencing interrogates dinucleotide combinations according to a more complicated scheme. Helicos promises single-molecule sequencing without amplification (Braslavsky et al., 2003). The read lengths obtained from both these techniques are less than 50 bases (Tucker et al., 2009).
4.2.3 Data analysis

A main issue in microarray and any high-throughput studies is how to analyze and retrieve valuable information from the enormous amount of data generated. The data analysis includes key processes such as quality control, background correction, normalization and assessment of differential expression (Quackenbush, 2006). Background correction is essential, as part of the measured probe intensities is due to non-specific hybridization and the noise in the optical detection system. The observed signal intensities need to be normalized to give accurate measurements of specific hybridizations. A number of normalization strategies have been suggested to eliminate different systematic bias from the data. Without proper normalization, it is impossible to compare measurements from different arrays due to many sources of variation. The source of variation includes sampling, different efficiencies of reverse transcription, labeling, hybridization reactions, physical problems of the arrays.

The core objective of the data analysis is to identify genes differentially expressed in various biological conditions. The fold changes calculated are mainly used for inferring the differential expression assuming that the reliability and the significance would increase together with the magnitude of fold change. The wide spread use of different microarray platforms has laid the ground to a completely new scientific area of developing new statistical strategies for microarray data. Most of these approaches are modified t-tests and one of the popular methods is significance analysis of microarrays (SAM) (Tusher et al., 2001).

There is a large list of bioinformatics software available from commercial organizations and open-source networks that provide specific tools for analysis. The R platform of statistical analysis has been increasingly popular for analyzing the microarray data. It has a specialized microarray analysis project called Bioconductor (Gentleman et al., 2004) and has many integrated statistical methods available. Many packages have been added in the last few years increasing the range of possible data analysis.

4.2.4 Biological interpretation of high-throughput data

Microarray analysis always results in a list of interesting genes that are differentially expressed. The long gene lists, however, cannot be considered the end point of the analysis. Rather, they form the starting point of a more meaningful interpretation, whereby biological patterns are typically highlighted. To infer the overall changes in terms of functions and processes, functional annotation terms are usually obtained from libraries such as Gene Ontology (GO) (Ashburner et al., 2000) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). The GO
database provides ontology of defined terms representing gene products, while the KEGG provides biochemical pathway information of the gene (Kanehisa and Goto, 2000; Kanehisa et al., 2006). This essentially shifts the level of analysis from individual genes to sets of biologically related genes.

Several tools are now available to assess whether a given gene set is overrepresented in a functional category (Khatri and Draghici, 2005). These methods statistically verify the significance of enrichment by using various models including hypergeometric, binomial, $\chi^2$ (chi-square) and Fisher’s exact test (Khatri and Draghici, 2005). In paper III, the methods implemented in the Database for annotation, visualization and integrated discovery (DAVID) gene annotation system were utilized for selecting significant families with default parameters (Huang da et al., 2007). In paper IV, GO tree machine (GOTM) with Fisher’s exact test was used for screening the over-representation of gene ontology categories with differentially regulated genes.

GO and pathway analysis plays an important role in understanding virus-host interactions by identifying the biological processes and signaling pathways affected during infection. Similarly, one can test whether the expression of genes sitting in specific portions or entire chromosome are involved in certain experimental conditions.

### 4.2.5 Promoter analysis

Transcription factors (TFs) bind to DNA in a sequence-specific manner and initiates transcription. Co-regulation of mammalian genes usually depends on sets of transcription factors that coordinately bind the promoter sequences and interact with each other (Werner, 2007). The identification of TF binding sites in the promoter sequences is useful for understanding gene expression in response to internal and external signals. Therefore, the promoter sequences of coexpressed genes are scanned for such regulatory motifs. These approaches aim to link gene expression data to the activity of transcription factors in cause-effect models and infers the transcriptional regulatory networks (Babu, 2008; Goutsias and Lee, 2007; Lee, 2005; Sivachenko et al., 2007). In this work, we have described an HMM approach to screen promoters of coregulated genes for specific binding sites (Paper-III). Moreover, there exists other approaches such as Boolean networks and Bayesian networks to infer gene regulatory networks.

### 4.2.6 Protein interaction networks

Genes co-expressed on signal transduction usually encode interacting proteins, often members of certain pathways. Through analyzing protein-protein interactions of
regulated genes, one gains understanding of how these molecules interact with each other, as well as their functional roles at the systems level (Ideker et al., 2002). Protein interaction data is represented as a network, with nodes corresponding to genes or gene products, and edges corresponding to physical interactions between genes. The identified subnetworks with differentially expressed genes can be considered as pathways affected in response to physiological conditions. A number of interaction databases, including BIND (Gilbert, 2005), IntAct (Kerrien et al., 2007), MINT (Zanzoni et al., 2002), and HPRD (Prasad et al., 2009), provide a variety of datasets and analysis tools. Further, several scoring methods have been proposed to rank the networks derived based on interactome and gene expression data. In paper III and IV, protein interaction data of regulated genes was used to identify protein complexes or pathways affected.

One fundamental problem with these knowledge-based methods is that the information is collected from data gathered using sequential discovery. This type of analysis considering known information will be biased towards certain topics, and disregards original data. Common solutions to this problem involve using unbiased measures to create groups, such as co-regulation or recurring motifs (Xie et al., 2005). Such solutions, however, might remove the bias but in return sacrifices the ease of interpretation, as these categories usually do not correspond to a specific hypothesis. Since improved interpretability was one of our initial goals, another solution is needed.

4.2.7 Public repositories of microarray data

The global gene expression data generated in a microarray study often exceeds the area of the particular research group who produced it. To accelerate the research efficiency it is therefore practical to have data publicly available for others to explore. In the beginning this was primarily available on the website of each research group or by the journal publishing the study. This spread of data made it rather a time consuming task for other researchers to find and collect the information. In response to the exponentially increasing number of microarray data sets published and the urge to be able to do meta analysis of data from several data sets, several public repository data bases for storage of microarray data have been launched. The largest one is the Gene Expression Omnibus (GEO) at NCBI hosting 348960 samples using 6404 platforms (Sept. 2009, http://ncbi.nlm.nih.gov/geo/) (Barrett and Edgar, 2006; Barrett et al., 2007). In Europe, the most popular database is ArrayExpress, is accommodated by the European Bioinformatics Institute. It contains microarray data from 1134 studies with 31275 hybridizations (Sept. 2009, http://www.ebi.ac.uk/arrayexpress/) (Parkinson et al., 2005).
5 RESULTS AND DISCUSSION

EBV encodes at least three deubiquitinating enzymes (Paper - I)

EBV, as previously discussed, encodes more than 100 ORFs and only few of these are well characterized. Identification of the protein function related to key cellular pathways would help in understanding viral strategies during infection. In this study, we have analyzed all EBV ORFs to predict distant functional homologs of human ubiquitin deconjugases using sequence analysis methods. Restricted search with the conserved catalytic domains of known DUB families has allowed us to overcome the sequence complexity existing in the rest of the sequence.

Our sequence analysis approach using several search methods and scoring the candidates based on their prediction in number of searches (see Paper-I for more details on scoring) proved successful in listing 16 candidates including previously identified viral DUB, BPLF1 that encodes large tegument protein (Kattenhorn et al., 2005; Schlieker et al., 2005). Functional assays have confirmed the catalytic activity of the EBV primase, BSLF1, and the thymidine kinase, BXLF1, in addition to BPLF1. Like cellular DUBs, the deconjugase activity of these novel viral DUBs was inhibited by treatment with cysteine protease inhibitors.

One important observation was that the enzymatic activity of BSLF1 and BXLF1 ORFs expressed in bacteria was weaker than that expressed in eukaryotic cells suggesting that there may be a regulatory mechanism involved. Further sequence analysis with the homologs of new viral DUBs revealed that BPLF1 and BSLF1 are conserved across the members of herpesvirus family while the BXLF1 homologs showed least conservation. Future studies on these viral proteins will allow us to understand their role in the virus life cycle and their effect on specific cellular functions.

These results indicate that such bioinformatics tools are very helpful in predicting novel functions of viral proteins and identifying viral strategies used during infection and malignant transformation.

EBNA-1 induces ROS production by activating NOX2 (Paper - II)

In this study, we have demonstrated the use of gene expression profiles available in public databases. To understand the possible mechanism by which EBNA-1 induces ROS production, we collected microarray datasets from EBV positive and negative cell
lines that are available at NCBI GEO. Data normalization and extraction of significantly regulated genes is critical in this type of analysis. Using SAM method, we identified four genes of ROS pathway that are differentially regulated in EBV carrying cells. Two genes involved in ROS production, Acyl-CoA oxigenase-1 (ACOX1) and Cytochrome b-245 heavy chain (CYBB, NOX2) and the ROS scavenger glutathione peroxidase-1 (GPX1), were significantly upregulated while lysyl oxidase like-2 (LOXL2) was downregulated in EBV positive cells. NOX2 is the catalytic subunit of the nicotinamide adenine dinucleotide phosphate (NADPH) oxidase complex that produces ROS. The overexpression of NOX2 was confirmed by detection of the protein only in EBV carrying Ramos and BJAB but not in their EBV negative parentals. Further, higher levels of NOX2 mRNA and protein were detected in BJAB expressing stable or inducible EBNA-1, while LMP-1, a diagnostic marker of EBV latency III, had no effect.

Based on the sequence analysis of NOX2 promoter, we have identified the sequence region similar to the Qp promoter of EBV. The luciferase reporter assays with NOX2 promoter resulted in EBNA-1 dose-dependent increase of luciferase activity. These observations suggest that EBNA-1 may be directly involved in the induction of NOX2. Thus, EBNA-1 could participate in B-cell growth transformation by inducing ROS through NOX2 and initiating many signaling pathways promoting genomic instability.

**EBNA-1 regulates chromatin remodeling complexes (Paper - III)**

The role of EBNA-1 on cellular transcription has been a curious link between EBV and associated malignancies. The results presented in Paper-II demonstrated the ability of EBNA-1 to transcriptionally regulate the cellular genes. In this study we tried to uncover the global effect of EBNA-1 expression on cellular transcription. We used both stable and inducible transfectants of EBNA-1 to discriminate early and late affects. Three hundred and nineteen cellular genes were regulated shortly after induction of EBNA-1 in a conditional transfectant while a ten fold higher number, 5921 genes in continuous EBNA-1 expressing and 4383 genes in stable transfectants was found regulated.

The promoter analysis of regulated genes showed that a significant number of genes contained EBNA-1 binding sites and several with multiple sites suggesting as direct targets. In addition, EBNA-1 also seems to regulate several genes that are annotated as transcription regulators (TR). Enrichment of TRs was found among the EBNA-1 regulated genes compared to genome wide occurrence accounting for the indirect regulatory activity of EBNA-1.
The GO analysis of commonly regulated genes from all EBNA-1 expressing conditions found “chromatin maintenance” as affected process. The protein products of the genes enriched in this category were histone H2B variants, all upregulated, and the DNA binding proteins SWI\textit{t}ch/Sucrose nonfermentable (SWI/SNF) related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1 (SMARCB1, also known as hSNF5) and immunoglobulin mu binding protein 2 (IGHMBP2), both downregulated. Interestingly, the gene numbers in each category increased depending on the duration of EBNA-1 expression. This observation of SMARCB1 regulation was supported by the analysis of publicly available gene expression microarrays (NCBI gene expression data, GSE2350) that revealed two-fold reduction of SMARCB1 transcripts in BLs expressing latency I compared to EBV negative BLs.

The protein interaction analysis of genes related to chromatin maintenance revealed that SMARCB1, SMARCA4 and SMARCD2 subunits of SWI/SNF chromatin remodeling complex were downregulated. Protein network analysis has also identified several subunits of the NuRD and PRC1 complexes that were downregulated. Involvement of these complexes in transcriptional repression indicates that EBNA-1 may promote the transcriptional activation through downregulation of chromatin remodelers.

**TPPII modulates the MAPK pathway (Paper - IV)**

The motivation for this study was the observation of TPPII up-regulation in BL and BL-like cells that have impaired proteasome activity. To gain insights on the cellular functions that are regulated by TPPII gene expression profiles were analyzed from the BL line Namalwa that spontaneously expresses high levels of TPPII, and in Namalwa cells where TPPII expression was silenced by shRNA.

One hundred and eighty seven genes were modulated in the TPPII knockdown cells, of which 137 were down regulated and the remaining 50 were up regulated. Surprisingly, GO and KEGG pathway analysis of differentially regulated genes identified several partners of the MAPK pathway as affected. The MAPK is a key signaling pathway in the cell that is activated by a variety of extracellular stimuli and regulates a broad array of biological processes; including Focal adhesion and TGF-\(\beta\) signaling that were significantly enriched with altered genes by TPPII knockdown. Systems biology approach using protein-protein interactions of differentially regulated genes confirmed that MAPKs are central hubs in the molecular interactions regulated by TPPII (see Paper-IV). Pathway analysis combined with protein interaction network reveals that the regulated genes cooperate to inactivate the MAPK cascade by down regulating several
of the kinases and upregulating the inhibitors. From the protein interaction analysis, several up and downstream partners of MAP kinases were also found affected by TPPII silencing. Thus, our findings suggest that the pleiotropic effect of TPPII is mediated by regulation of MAPK signaling.

Taken together, these findings add to our current knowledge of cellular processes modulated during EBV infection and in malignant transformation.
6 CONCLUDING REMARKS

This thesis illustrates the use of bioinformatics methods and gene expression profiles to explore molecular pathways modulated during EBV infection and in malignant cells. The implementation and further development of such methods will provide new tools for identifying key cellular pathways modulated by viruses.

By combining different sequence analysis methods, we could predict distal homologues of cellular DUBs encoded in the EBV genome. Future studies on these viral DUBs should focus on understanding their potential contribution to EBV infection and malignant transformation. The success of our strategy in predicting EBV encoded DUBs suggest that a similar approach may be possible to identify similar functional homologues encoded by other pathogens.

Gene expression and regulation defines the type and state of the cell. Our work shows that meta-analysis of publicly available gene expression arrays is a powerful strategy for mining biologically relevant information while saving time and financial resources. Several tools are now being developed in order to make such analysis easy and comprehensive (French et al., 2009; Ivliev et al., 2008). The finding that EBNA-1 regulates the production of ROS by transcriptional activation of NOX2 identifies one of few cellular genes whose expression is directly affected by EBNA-1. Future work should address the molecular interactions involved in this transcriptional effect including direct binding of EBNA-1 to the NOX2 promoter and the possible involvement of cellular transcription factors. Transcription profiling of inducible and stable EBNA-1 expressing cells has suggested both direct and indirect mode of EBNA-1 induced regulation. The regulation of chromatin remodeling complexes is a novel effect of EBNA-1 that should be addressed in detail in future studies by for example comparing the pattern of euchromatin and heterochromatin in EBNA-1 expressing cells. Similarly, the analysis of global gene expression profiles in control and TPPII knockdown BLs suggested an unexpected involvement of TPPII in the regulation of MAPK pathway. This prediction should be experimentally validated in order to understand the role of TPPII in malignant transformation.

In conclusion, the work described in this thesis has generated new knowledge and testable hypothesis on the mechanisms by which EBV manipulates the cellular environment. I hope that the results of this effort will prove useful in the development of new therapeutic approaches for the treatment of EBV associated pathologies.
7 ACKNOWLEDGEMENTS

Making a Ph. D. thesis is never a lonely work it relies on the help and encouragement of many; I tried to mention some of the names here.

First and foremost, I would like to thank my advisor, **Prof. Maria G. Masucci**, for accepting me as a graduate student and introducing me into the EBV field. Maria always gave me the necessary freedom to pursue my interests and consistently supported me with her invaluable advice, without her support, it simply would not have been possible. Thanks again for believing me and giving me suggestions to improve my scientific skills.

I wish to thank my co-supervisor, **Kristina Lindsten**, for all her support and for boosting my enthusiasm from time to time. Thanks for teaching me how to culture cells and how to run my first Western.

My heartfelt thanks to **Siamak**, my colleague, friend and brother for his support through all these years. Thanks a lot for all those long discussions in and about science. Also thanks to **Laia** and **Sohail** for all your affection.

Special thanks to **Teresa**, first, for bringing her sweet and nice group to CMB and second for always having time for me. Thanks for all the support.

I am thankful to **Stefano** for always believing in my predications and validating them. My best thanks to **Pino**, who is a star in the lab and makes everyone smile at any time of the day. Thank you for bringing all the joy and happiness to the lab. I wish to thank **Javier**, **Omid**, **Steffan** and **Ximena** for all the conversations during lunch and in the corridor.

Thanks to **Bettina**, **Diego**, **Eugenie**, **Lina** and **Sebastian** for all the fun that we shared during these years.


Thanks to **Helena**, the best lab manager.
Special thanks to Yvonne and Linda for your support and administrative assistance through all these years. Tack så mycket!

My thanks to Prof. Bengt Persson for helping me with the parts of this thesis and for giving me time with short meetings. Rickard Sandberg for all our discussions and for introducing me to advanced python programming.

Anna, Fabio, Francesco, Ingrid and Wiebke many many thanks; my life abroad would have been boring with out you guys.

I would like to thank my friend Rajendar and other friends from India for organizing pleasant weekends and keeping my taste buds active to Indian spices.

I am forever grateful to my beloved parents, Sri. Siva Nagireddy and Smt. Sivalelamma for the continuous encouragement and endless love, my brother Jayakrishna, sister Madhavi, brother-in-law Nagendraphani and our new little star Madhav for all the love that has made my success possible. Without their vital help, I doubt I would ever manage. Many thanks to my wife and friend, Krishnaveni, for bringing new light and joy in my life.

This study has been initially started at Dept of Microbiology, Tumor and Cell biology and later carried out at Dept of Cell and Molecular biology of Karolinska Institute with the support from the European Community Integrated Project on Infection and Cancer, INCA, Project no. LSHC-CT-2005-018704.
8 REFERENCES


Tsimbouri, P., Drotr, M.E., Coy, J.L., and Wilson, J.B. (2002). bcl-xL and RAG genes are induced and the response to IL-2 enhanced in EmuEBNA-1 transgenic mouse lymphocytes. Oncogene 21, 5182-5187.


