Dissertation zur Erlangung des Doktorgrades der Fakultät für Biologie der
Ludwig-Maximilians-Universität München

IDENTIFICATION OF CELLULAR LONG NON-CODING RNAS REGULATED BY THE EBV NUCLEAR ANTIGEN EBNA2

Simone Antje Daniela Wagner
München, August 2018

Erstgutachter: Prof. Dr. Bettina Kempkes
Zweitgutachter: Prof. Dr. Wolfgang Enard
Tag der Abgabe: 30.08.2018
Tag der mündlichen Prüfung: 27.06.2019

Zusammenfassung

Das Epstein-Barr Virus (EBV) ist das vierthäufigste infektiöse Karzinogen für Menschen. Das Virus etabliert eine lebenslange Latenz in den B-Zellen des Wirts, welche mit verschiedenen B-Zell Tumoren einhergeht. Eine EBV-Infektion von B-Lymphozyten *in vitro* führt zur Transformation in eine lymphoblastoide Zelllinie (LCL) und deren uneingeschränkten Proliferation. Diese Transformation wird unter anderem von den EBV nuklearen Antigenen (EBNAs) gesteuert. Dazu gehören EBNA2 (E2) und EBNA3A (E3A), zwei ko-exprimierte Transkriptionsfaktoren (TFs), die um die Bindung des DNA-Ankers C-Promotor Binding Factor 1 (CBF1) konkurrieren. Unter deren Zielgenen befinden sich auch mehrere lange, nicht-kodierende RNAs (lncRNAs), wichtige Transkriptionsregulatoren, die mit einer Vielfalt von Krankheiten assoziiert wurden. In dieser Arbeit wurde die E2- und E3A-abhängige Genregulation in einer umfassenden Transkriptomanalyse durch RNA-Seq untersucht, wobei wir uns auf E2 Ziel-Gene, mit besonderem Schwerpunkt auf lncRNAs, fokussiert haben. Unsere Ergebnisse zeigen, dass sich E2- und E3A-regulierte Ziel-Gene genomweit in ko-regulierten Genblöcken (CRGBs) gruppieren. Diese CGRBs bestehen sowohl aus proteinkodierenden als auch aus lncRNA Genen. Ausgewählte Ziel-Gene wurden durch RT-qPCR bestätigt und es konnte zudem gezeigt werden, dass sie während der Etablierung der Latenz nach EBV-Infektion reguliert werden. Mehrere dieser lncRNAs sind bereits aus verschiedenen Krebsarten bekannt. Weiterhin konnten wir hervorheben, dass die Regulation von lncRNAs und dem nächstgelegenen proteinkodierenden Gen positiv korreliert, was für die Regulation von entfernt gelegenen proteinkodierenden Gene durch E2 induzierte lncRNAs spricht. Wir entdeckten, dass E2 sowohl im Zellkern als auch im Zytoplasma in öffentlich zugänglichen Genbanken annotierte lncRNAs reguliert. Zusätzlich haben wir intergenisch (zu ENSEMBL annotierten Genen) transkribierte Gene detektiert, die im Zellkern angereichert waren. Diese sind zum Teil auch noch nicht von der umfassenden lncRNA Datenbank LNCat erfasst und können somit als neu definiert werden. Um direkte E2 Ziel-Gene von denen anderer Transkriptionsfaktoren, die durch E2 induziert werden, zu unterscheiden, haben wir auch E2-regulierte Gene nach Blockade der *de novo* Proteinsynthese beschrieben. 3 % der E2-Zielgene wurden unter fehlender *de novo* Proteinsynthese reguliert, einschließlich einiger lncRNAs. Wir fanden, dass 70 % der gemeinsamen Ziel-Gene von E2 und E3A gegenreguliert waren. Schließlich haben wir erste Effekte von E2 und E3A auf das virale Transkriptom untersucht, wo wir eine Induktion lytischer Gene vermuten. In einer unabhängigen Studie konnten wir den Beitrag von Early B Cell Factor 1 (EBF1) zur Chromatin-bindung von E2 an CBF1-unabhängigen Bindungsstellen im humanen Genom funktionell bestätigen. Zusammenfassend zeigen unsere Daten, dass EBV lncRNAs reguliert welche zur Tumorgenese beitragen könnten.

Die Ergebnisse dieser Arbeit erweitern das Wissen über die transkriptionelle Regulation von Wirtszellgenen durch E2 und liefern einen Ausgangspunkt für die Untersuchung der Wirkung von E2 auf das virale Transkriptom.

Abstract

Epstein-Barr virus (EBV), the fourth most common infectious carcinogen for humans, can establish a life-long latency in the host B cells, with such latent infection associated with various B cell malignancies. The infection of B lymphocytes by EBV *in vitro* results in transformation and unlimited proliferation of a so called lymphoblastoide cell line (LCLs). This transformation is driven by EBV nuclear antigens (EBNAs), in addition to other factors. These include EBNA2 (E2) and EBNA3A (E3A), which are co-expressed transcription factors (TFs) competing for binding to a well-studied anchor C promoter binding factor (CBF1). E2 and E3A have been shown to share cellular target genes, including several long non-coding RNAs (lncRNAs). LncRNAs are key transcriptional regulators, associated with a huge variety of diseases including cancer. Herein, we comprehensively assessed the E2 and E3A dependent gene regulation in a transcriptomic analysis by RNA-Seq. We focused on E2 target gene regulation with particular emphasis on known and novel lncRNAs. Our results show that E2 and E3A regulated target genes cluster genome-wide in co-regulated gene blocks (CRGBs). These CRGBs consist of both protein-coding and lncRNA genes. Candidate target genes were confirmed by RT-qPCR and demonstrated to be regulated during the establishment of latency following EBV infection. Several of these lncRNAs were reported to be dysregulated in various cancer types. Further, we demonstrated that the regulation of lncRNAs and the closest protein-coding gene correlates positively, supporting the possibility that E2 induced lncRNAs regulate remote protein-coding genes. We discovered that E2 regulated lncRNAs annotated by public available genome databases in both the nucleus and the cytoplasm. Additionally, we detected intergenic transcribed genes (unannotated by ENSEMBL) enriched in the nucleus. To some extent, these intergenic transcribed genes were not covered by the comprehensive lncRNA database LNCat and can thus be defined as novel. To determine direct E2 target genes and that of other transcription factors induced by E2, we profiled E2 differentially regulated genes with competent and inhibited protein synthesis. 3 % of the E2 target genes were regulated with absent *de novo* protein synthesis, including several lncRNAs largely emanating from enhancer marked chromatin, indicating that 97% of E2 targets are indirect targets. We found, that up to 70 % of the shared target genes of E2 and E3A were counter-regulated. Finally, we assessed first effects of E2 and E3A on the viral transcriptome finding an induction of lytic genes.

In an independent study, we were able to functionally confirm the contribution of early B cell factor 1 EBF1 to chromatin binding of E2 at CBF1 independent binding sites.

In conclusion, our data indicate that EBV regulates lncRNAs which could contribute to tumorigenesis. The findings of this thesis extends the knowledge on the transcriptional regulation by E2 of host cell genes and provide an initial point for the investigation of the impact of E2 on the viral transcriptome.

Success is not final,
failure is not fatal:
it is the courage to continue that counts.
~Winston Churchill~

Table of Contents

Registers

Figures

Tables

List of abbreviations

..... *number*
% ..... *per cent*
% (v/v) ..... volume percent
°C ..... *Degree Celsius*
µl ..... *microliter*
3‘ *3-prime*
3D ..... *three dimensional*
4D ..... *four dimensional*
5‘ *5-prime*
AIDS ..... *aquired immune deficiency syndrome*
AKT ..... *protein kinase B*
AML ..... *acute myeloid leukemia*
Ampl. ..... *amplicon*
ass. ..... *assigned*
ATP ..... *adenosine triphosphate*
b.i. ..... *before infection*
BAC ..... *bacterial artificial chromosome*
BAM ..... *binary alignment map*
BARTs ..... *BamHI A rightward transcripts*
BigWig ..... *wiggle (wig) files in an indexed binary format*
BL ..... *Burkitt's lymphoma*
bp ..... *basepair*
BRD4 ..... *bromodomain-containing protein 4*
CBF1 ..... *C promoter binding factor 1*
CD ..... *Cluster of differentiation*
CDK ..... *cycline dependent kinase*
cDNA ..... *complementary DNA*
ceRNA ..... *competing endogenous RNA*
ChIP ..... *chromatin immunoprecipitation*
CHMP2A ..*charged multivesicular body protein 2a*
ChX ..... *cyclohexmidine*
cm ..... *centimeter*
CNTRL ..... *control*
CRGB ..... *co-regulated gene blocks*
CSS ..... *chromatin state segmentation*
cyto ..... *cytoplasm*
d ..... *day*
DE ..... *differential expression*
depcorr ..... *dependency corrected*
DNA ..... *desoxyribonucleic acid*
dox ..... *doxycycline*
ds ..... *downsampling*
e.g ..... *for example*
E2 ..... *EBNA2*
E3A *EBNA3A*
EBER ..... *Epstein-Barr virus encoded small RNAs*
EBF1 ..... *early B cell factor 1*
EBNA ..... *EBV nuclear antigen*
ECL ..... *enhanced chemiluminescence*
EDTA ..... *Ethylenediaminetetraacetic acid*
eGFP ..... *enhanced green fluorescent protein*
ENCODE ..... *ENCyclopedia Of DNA Elements*
Epstein-Barr virus ..... *Epstein-Barr virus*
ER ..... *estrogen receptor*
eRNA ..... *enhancer RNA*
ESE ..... *EBV super enhancers*
estr. ..... *estrogen*

EtBr ..... *ethidium bromide*
FACS ..... *Fluorescence activated cell sorting*
FC ..... *fold change*
FCS ..... *fetal calf serum*
FDR ..... *false discovery rate*
fwd ..... *forward*
g ..... *gravitational constant*
g.o.i ..... *gene of interest*
GAPDH ..... *Glyceraldehyde 3-phosphate dehydrogenase*
GC ..... *germinal center*
GRO-Seq ..... *Global Run On sequencing*
GST ..... *Glutathione S-transferase*
gusB ..... *beta-glucuronidase*
h ..... *hour*
H ..... *histone*
H_2O ..... *water*
H3K27ac ..... *histone 3 lysine 27 acetylation*
H3K27me3 ..... *histone 3 lysine 27 trimethlyation*
H3K4me3 ..... *histone 3 lysine 4 trimethylation*
H3K79me3 ..... *histone 3 lysine 79 trimethylation*
H3K9me3 ..... *histone 3 lysine 9 trimethylation*
HA ..... *hemagglutinin*
HAT ..... *histone acetyltransferase*
HCl ..... *Hydrogen chloride*
HDAC ..... *histone deacetylase*
HHV 4 ..... *human herpes virus 4*
HKMT ..... *histone lysine methyltransferases*
HL ..... *Hodgkin lymphoma*
HMM ..... *Hidden Markov Model*
hnRNPL ..... *heterogeneous nuclear ribonucleoprotein L*
HRP ..... *horseradish peroxidase*
HRS ..... *Hodgkin and Reed-Sternberg*
ID ..... *identifier*
IgG ..... *immunoglobulin G*
IGV ..... *Integrative Genomics Viewer*
IKK ..... *IκB kinase*
IM ..... *infectious mononucleosis*
IP ..... *immunoprecipitation*
junct ..... *junction*
kb ..... *kilobase*
ko ..... *knock out*
KU ..... *kilounit*
l.t.r. ..... *left to right*
LCL ..... *lymphoblastoide cell line*
LiCl ..... *lithium chloride*
lincRNA ..... *long intergenic non-coding RNA*
LMP ..... *latent membrane protein*
LMU ..... *Ludwig-Maximilians university*
LNCat ..... *long non-coding RNA atlas*
lncRNA ..... *long non-coding RNA*
lt ..... *longterm*
MAPK ..... *(mitogen-activated protein) kinase*
Mb ..... *megabase, megabase*
Med1 ..... *mediator 1*
$MgCl_2$ ..... *magnesium chloride*
min ..... *minute*
miRNA ..... *microRNA*
ml ..... *milliliter*
MLL3 ..... *mixed lineage leukemia 3*
MLL4 ..... *mixed lineage leukemia 4*
mM ..... *millimolar*
mRNA ..... *messenger RNA*
mtDNA ..... *mitochondrial DNA*
n ..... *number of replicates*
NaCl ..... *Sodium chloride*
ncg ..... *non-coding gene*
NF-κB .*nuclear factor kappa-light-chain-enhancer of activated B cells*
ng ..... *nanogram*
NGFR ..... *nerve growth factor receptor*
norm ..... *normalized*
NP-40 ..... *nonyl phenoxypolyethoxylethanol*
nt ..... *nucleotides*
N-terminal ..... *amino-terminal*
nucl ..... *nucleus*
ORF ..... *open reading frame*
p. ..... *page*
p.i. ..... *post infection*
PBS ..... *phosphate buffered saline*
PCA ..... *principle component analysis*
pcg ..... *protein-coding gene*
PFA ..... *paraformaldehyde*
PI3-K ..... *phosphoinositide 3-kinase*
PIC ..... *proteinase inhibitor cocktail*
pmol ..... *picomol*
PolII ..... *RNA polymerase II*
PolyA ..... *polyadenylated*
PPAN ..... *Peter Pan*
PRC1 ..... *polycomb repressive complex 1*
PRC2 ..... *polycomb repressive complex 2*
pRNA ..... *promoter RNA*
PTLD *post transplant lymphoproliferative disease*
PTM ..... *posttranslational modifications*
PVDF ..... *polyvinylidene difluoride*
qPCR ..... quantitative polymeraser chain reaction
Rb ..... *retinoblastoma*
rev ..... *reverse*
RIN ..... *RNA integrity number*
RNA ..... *ribonucleic acid*
RNAi ..... *RNA interference*
rRNA ..... *ribsomal RNA*
RT ..... *Reverse Transcription, room temperature*
RT-qPCR ..... *reverse transcription quantitative polymerase chain reaction*
SD ..... *standard deviation*
SDS ..... *sodium dodecyl sulfate*
SDS-PAGE ..... *SDS Polyacrylamide gel electrophoresis*
sec ..... *seconds*
SEM ..... *standard error of the mean*
Seq ..... *sequencing*
SHM ..... *somatic hypermutation*
shRNA ..... *short hairpin RNA*
siRNA ..... *small interfering RNA*
SMD ..... *STAU1-mediated mRNA decay*
snoRNA ..... *small nucleolar RNA*
ß2m ..... *beta-2 microglobulin*
SSF1 ..... *Suppressor of SW14*
STAT ..... *signal transducer and activator of transcription*

TAD .....*topologically associating domains*
TCGA .....*The Cancer Genome Atlas*
TE ..... *Tris-EDTA*
TF ..... *transcription factors*
TR..... *T cell receptor*
transc. supp. ..... *transcript support*
Tris ..... *Tris(hydroxymethyl)aminomethane*
TRM .....*tripartite terminase*
tRNA.....*transfer RNA*
TSS..... *transcriptional start side*
TT-Seq.....*Transient Transcriptome sequencing*
tv *transcript variant*
TxN .....*transcription*
U .....*units*
UTR.....*untranslated region*
UV..... *ultraviolet*
V .....*volt*
wt ..... *wildtype*
α .....*alpha (anti)*
Δ .....*delta (deletion)*
κ .....*kappa*
µF .....*microfarad*
µg .....*microgram*

1 Introduction

1.1 The Epstein - Barr virus

The Epstein-Barr virus or human herpes virus 4 (HHV 4) is a human pathogenic double stranded DNA virus, with a genome of 172 kb (Baer et al., 1984). It was first discovered by Michael Epstein and Yvonne M. Barr in 1964 (Epstein, Achong, & Barr, 1964). They isolated the virus from B lymphocytes of an African patient suffering from Burkitt's lymphoma. Transmission befalls by body fluids, especially by saliva, in rare cases also by transplantations or blood transfusions (reviewed by Smatti et al., 2018). Ordinarily, the infection with EBV occurs in young childhood and is asymptomatic. However, when occurring during adolescence, the infection can result in infectious mononucleosis (IM; Henle, Henle, & Diehl, 1968). 90 % of the adult population is EBV positive (Chang, Yu, Mbulaiteye, Hildesheim, & Bhatia, 2009). The virus persists lifelong in the body, can be reactivated and usually remains undetected not causing a disease as it is controlled by the immune system. However, when reactivated, the virus is again contagious (reviewed in Smatti et al., 2018, Stanfield & Luftig, 2017, Young, Yap, & Murray, 2016, Thorley-Lawson, 2015, Amon & Farrell, 2005). Two EBV types, type 1 and 2 (or type A, B respectively) are existing, which are characterized by sequence variations in the Epstein-Barr nuclear antigen 2 (EBNA2/E2) and Epstein-Barr nuclear antigen 3 (EBNA3/E3) genes. The type 1 is more often represented worldwide and has a greater transforming potential, as type 1 virus transformed cells yield cell lines much more rapidly than type 2 virus transformants (reviewed in Young et al., 2016).

1.1.1 EBV's life cycle and its oncogenic potential

An EBV infection starts in the oropharyngeal mucosa. There, in the epithelial cells, the virus replicates usually lytically. Primary infection of EBV causes humoral and cellular immune responses, where antibodies against an EBV capsid antigen or an early antigen are produced and EBV-specific cytotoxic T cells tackle the infected cells. Nevertheless, the virus is not completely eliminated, it may finally enter a circulating memory B cell and establish a latency (infection rate 1 in 10^4 to 10^5 memory B cells; reviewed in Geng & Wang, 2015). It is still controversial how the virus enters the memory B cells. There are until now two infection models, one assumes a direct infection of memory B cells by EBV and the other one is the germinal center model, which is more consistent with different independent findings on EBV (reviewed in Thorley-Lawson, 2015; Young et al., 2016; Vockerodt et al., 2015). Contradictory are the findings *in vitro* versus *in vivo*, *in vitro* EBV can "immortalize" B cells by the coordinated expression of viral latent genes ("growth program"=

latency III). Here, EBV expresses all protein coding *EBNA* and latent membrane protein (*LMP*) genes as well as non-coding Epstein-Barr virus encoded small RNAs (EBER) and BamHI A rightward transcripts (BARTs) in order to persist latently in unlimited proliferating lymphoblastoide cells (LCLs) but it does not persist like this *in vivo*. *In vivo*, it persist by downregulating the viral gene expression in a quiescent resting memory B cell (reviewed in Thorley-Lawson, 2015). The germinal center model suggests that EBV first infects naïve B cells and drives B cell proliferation by the expression of the growth program. Following, EBV mimics a germinal center (GC) reaction by the expression of the transient "default program" (= latency II; *EBNA1*, *LMP*s, EBERs and BARTs expressed). From there the cell exits as memory B cell, where the viral gene expression is minimized (latency 0; only EBERs expressed), just during proliferation of the memory B cells, EBNA1 and the non-coding EBERs and BARTs are expressed (latency I) to ensure the replication of the viral DNA, which forms a circle and persists as an episome in the nuclei of infected cells and its distribution to daughter cells (reviewed in Young et al., 2016; Küppers, 2003). GCs are lymphoid structures where critical processes for antigen selection, maturation and selection of immunoglobulin class take place such as somatic hyper-mutation (SHM). No signs of SHM were found in EBV positive B cells isolated from GCs of patients with IM, which among other findings supports the direct infection model. However, in principle EBVs biological behavior is the initiation, establishment and maintenance of a persistent infection by mimicking normal B cell biology (Vockerodt et al., 2015; Figure 1). Its discovery in a tumor and its ability to drive unlimited B cell proliferation hints towards an oncogenic potential. According to recent epidemiologic studies on infection-attributable cancer, the fourth most common infectious cause of cancer of all counted cases at the different time points of the studies and different investigated populations was EBV, which corresponds to 1% of cancer burden (Oh & Weiderpass, 2014; de Martel et al., 2012; Khan & Hashim, 2014). EBV has been associated with a striking variety of cancer types. Thorley-Lawson categorized EBV-associated cancers into three groups i) tumors, for which the results need to be more profound (e.g. breast or hepatocellular cancer), ii) tumors, for which there is strong evidence (e.g. nasopharyngeal or gastric cancers), but a latently infected biological equivalent is missing and iii) tumors which are solidly linked to EBV (Thorley-Lawson, 2015). This last group includes post-transplant lymphoproliferative disease (PTLD), Hodgkin's lymphoma (HL) and Burkitt's lymphoma (BL).

Post-transplant lymphoproliferative disease (PTLD)

EBV-induced transformation is suppressed by an EBV-specific immune response targeting mainly E3 proteins in healthy carriers. In T cell immunosuppressed patients after transplantations or in AIDS patients, the response is impaired and this can lead to the uncontrolled expansion of EBV-transformed B cells. This can result in EBV-positive B cell tumors arising after solid organ or hematopoietic stem cell transplantation. PTLDs are a heterogeneous collection of B cell tumors

(Vockerodt et al., 2015). All the viral genes of the growth program of latency III are expressed. However, there is evidence for SHM in most of these tumors hinting towards a GC or post-GC B cell origin. (Thorley-Lawson, 2015; Küppers, 2003). LCLs are an important model to study EBV-associated cancers and they express type III latency genes. Thus, the viral expression program in immunodeficient patients can be compared to the pattern observed in LCLs (Longnecker & Neipel, 2007).

Classical Hodgkin's lymphoma (HL)

HL accounts for 30 % of lymphoid malignancies. It is marked by the atypical large Hodgkin and Reed-Sternberg (HRS) tumor cells. However, these cells only constitute less than 1 % of the tumor tissue since most of it is comprised of diverse benign blood cells. B cells are almost always the origin of HRS cells. The SHM machinery is silenced in these cells and they acquire mutations. It's assumed that HRS cells derive from pre-apoptotic GC B cells that escaped apoptosis. The latency II default expression program of EBV may play a role in the escape of the pre-apoptotic GC B cells. Up to 40 % of the tumors contain EBV, with higher incidences in children and also in elderly people presumably because of the underdeveloped or senescent immune system (Geng & Wang, 2015; Küppers, 2003; Vockerodt et al., 2015; Thorley-Lawson, 2015).

Burkitt's lymphoma (BL)

This lymphoma was the first EBV associated disease. In the endemic form, which occurs for example in equatorial Africa, almost all cases are EBV positive. At lower incidence, in the sporadic form outside of endemic areas or in an HIV-associated form, EBV is only associated with up to 30 % of the cases. BL is marked by one of three different translocations of the proto-oncogene *MYC* to immunoglobulin genes, all leading to a constitutive activity of this transcription factor (TF). This overexpression is the cause for the high proliferation of BL cells. BL cells are derived from centroblasts, they show active SHM. In this tumor, minimal expression of viral genes (latency I program) exists. The overexpression of *MYC* which leads to uncontrolled proliferation would also induce apoptosis. Therefore, the BL cell has to overcome this by another incident, where EBV might contribute to by the expression of the long non-coding RNAs EBERs (Küppers, 2003; Vockerodt et al., 2015).

Which exact role EBV plays in lymphomagenesis still remains to be unraveled. The high frequency of EBV positive cases of diverse lymphomas despite the low frequency of EBV positive B cells in virus carriers indicates an involvement of EBV in the B cell transformation.

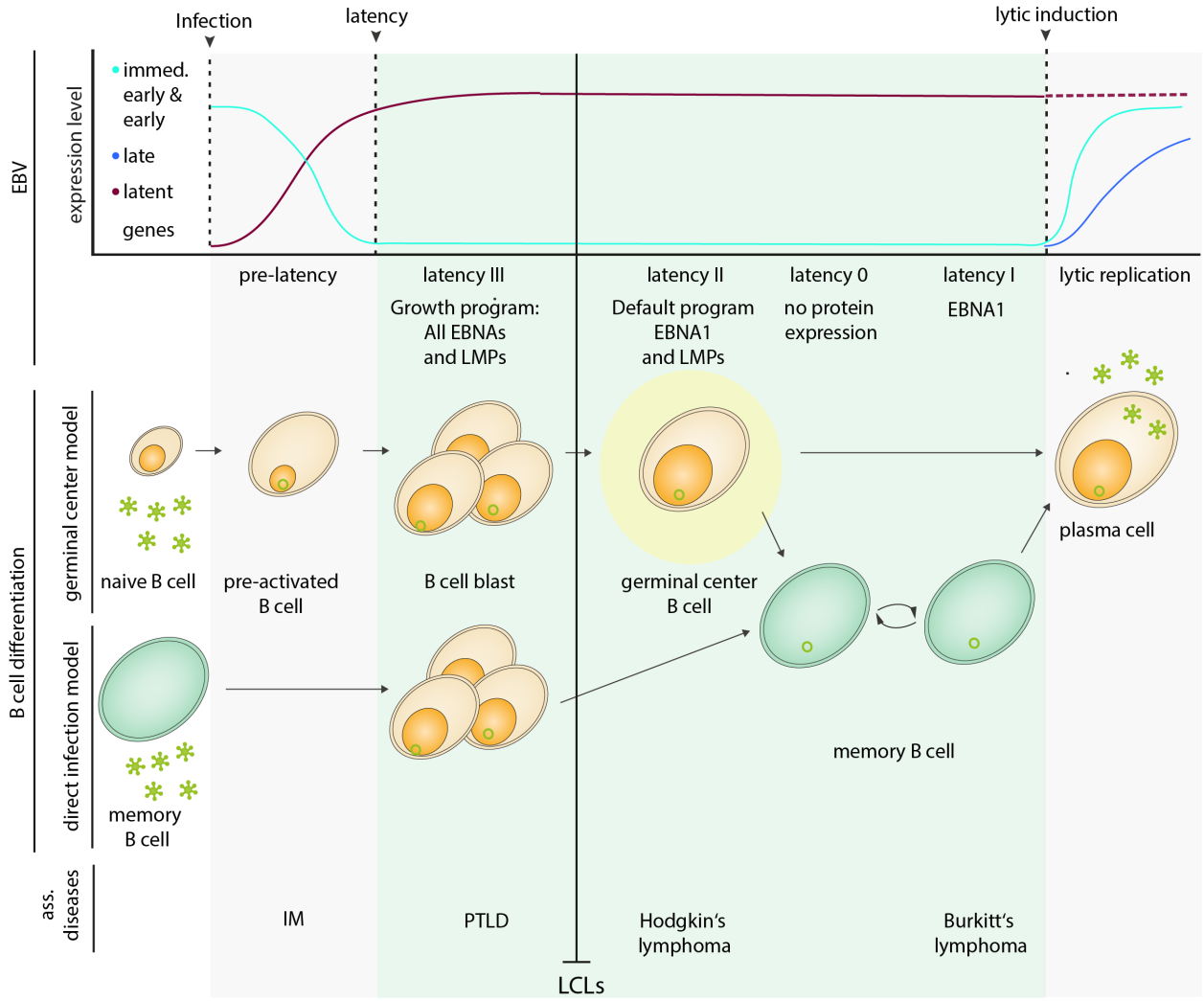


Figure 1: Schematic representation of the EBV life cycle. Upper panel: Dynamic of viral gene expression. The different stages of the life cycle are accompanied by the different expression of certain viral genes. Characteristic for the pre-latent phase is the expression of latent genes (brown line) and a limited subset of lytic genes corresponding to the so-called immediate early and early lytic class (turquoise line). After the pre-latent phase, latency is established. The synthesis of new virus is repressed during the initial latency phase due to the lack of expression of other crucial viral genes. An extrachromosomal plasmid represents the EBV genome during latency in the nucleus of infected cells supported by the expression of latent genes. Exogenous or viral stimuli can trigger the switch to the lytic phase, where again the expression of early and late lytic genes support amplification of viral genomes and the expression of structural proteins (grey line) is accountable for viral progeny (modified from Hammerschmidt, n.d.). Middle panel: Normal persistence. EBV resides in host memory B cells. Two models are proposed to explain the viral entry in memory B cells. The germinal center model suggests that EBV first infects naïve B cells and drives B cell proliferation by the expression of the growth program. Following, the infected B cell mimics a germinal center (GC) reaction by the expression of the transient default program (latency II; EBNA1 and LMPs expressed). From there it exits as memory B cell, where the viral gene expression is minimized (latency 0), just during proliferation of the memory B cells, EBNA1 is expressed (latency I) to ensure the replication of the viral DNA, which forms a circle and persists as extrachromosomal plasmid in the nuclei of infected cells and its distribution to daughter cells. In the direct infection model, EBV directly infects memory B cells, potentially including an intermediate latency III stage. Lower panel: The EBV-associated malignancies. This figure displays at which latency state infectious mononucleosis (IM), post-transplant lymphoproliferative disease (PTLD), Hodgkin's lymphoma or Burkitt's lymphoma occurs. When primary B cells get infected with EBV *in vitro*, they arrest in latency III as LCLs and proliferate unlimitedly (Figure modified from Young, Yap, & Murray, 2016).

1.1.2 EBV latent proteins E2 and E3A and chromatin tethering

By the activation of resting B cells, EBV exploits the cellular transcription and translation machineries. The transformation process is driven by the coordinated expression of nine latent viral genes coding for nuclear and membrane antigens as well as non-coding genes. Among the nuclear antigens are *EBNA2* (*E2*) and *EBNA3A* (*E3A*), two co-expressed key genes. Instant post infection, *E2* together with Epstein-Barr nucleus antigen leader protein (*EBNA-LP*) is the first gene expressed (Alfieri, Birkenbach, & Kieff, 1991), followed by *E3A* and others. The EBNA proteins are TFs of EBV which modulate viral and host gene expression.

1.1.2.1 E2

E2 was discovered, when P3HR1, a laboratory-derived EBV strain with an *E2* deletion was unable to transform B cells *in vitro* (Rabson, Gradoville, Heston, & Miller, 1982). Returning the *E2* gene back into P3HR1 has confirmed the indispensability of *E2* in the B cell transformation (Cohen, Wang, Mannick, & Kieff, 1989; Hammerschmidt, & Sugden, 1989). E2 is encoded by a single exon; the structure of the entire protein has not been solved so far. E2 is not able to bind to DNA itself. It requires cellular adaptor proteins to bind to chromatin. The contact of E2 with different transcription factors and co-activators occurs at an acidic activation domain of E2. E2 regulates transcriptional initiation and elongation to some extend by cyclin-dependent kinase 9 (CDK9) dependent phosphorylation of the C-terminal domain (CTD) of RNA polymerase II (PolII; Palermo, Webb, Gunnell, & West, 2008).

The so far best investigated DNA adaptor for E2 is C-promoter binding protein (CBF1, also known as RBPJκ). In the absence of E2, CBF1 recruits a corepressor complex for the repression of target gene transcription. When bound by E2 and coactivators, the repression is relieved by competition with the corepressor (reviewed in Kempkes & Ling, 2015). CBF1 belongs to the Notch signaling pathway. By binding to CBF1, E2 mimics a constitutively activated NOTCH receptor (Sakai et al., 1998). The Notch pathways play a role in cell fate determination, cell differentiation and developmental pattern formation in Drosophila. E2 might execute Notch-like functions.

Alternative anchors of E2 are discussed. In the recent past, we and others could show that early B cell factor 1 (EBF1) is also an important DNA anchor for E2 (Glaser et al., 2017; Lu et al., 2016; see section 3.1, p. 38). EBF1 is important to activate B cell-specific genes and is claimed to act as pioneer factor, meaning its binding evokes chromatin accessibility, histone modifications and target gene expression (reviewed in Boller, Li, & Grosschedl, 2018).

E2 is a transactivator of viral and cellular promoters. The knowledge about E2 functions is largely based on genetic analysis of E2-responsive elements within viral promoters. E2 activates the viral C promoter, as well as the *LMP1* and *LMP2A* and *LMP2B* promoters. A variety of genome-wide

array-based screens or candidate investigations in EBV-infected B cells or E2 expressing B cell lines were accomplished to study the impact of E2 on target gene expression. A lot of target genes were detected to be upregulated by E2, like *SLAMF1*, *DNase1L3* or *ABHD6* (Sabine Maier et al., *2006*). E2 also induces genes in the absence of *de novo* protein synthesis, like *MYC* (Kaiser et al., 1999). E2 can also be an active repressor of target genes, like *CD79B* (Maier et al., 2006) or *BCL6* (Boccellato et al., 2007).

1.1.2.2 E3A

E3A belongs to the E3 protein family, a group of latency-associated proteins, co-expressed together with E2. It is assumed that this protein family has arisen during the evolution of EBV by multiple gene duplication events because they are similar in sequence and gene structure. Nevertheless there are no hints towards redundant functions. E3s are critical in EBV persistence and for modulation of B cell lymphomagenesis. E3A is essential for *in vitro* B cell transformation (Tomkinson, Robertson, & Kieff, 1993), controversially, researcher were able to establish E3A negative LCLs (Hertle et al., 2009; Skalska et al., 2013; Skalska, White, Franz, Ruhmann, & Allday, 2010). Furthermore, it is important for the efficient proliferation of the B cells since important tumor suppressor pathways are targeted by E3A. E3A is exclusively located in the nucleus (it contains six nuclear localization signals) and is tightly associated with chromatin, but also is not able to bind to DNA directly. E3A is regarded as transcriptional repressor.
CBF1 is also an anchor for all E3 proteins and together, they can interfere with E2-mediated transactivation, since has been shown that E3A binds to the same site on CBF1 as E2 (Robertson, Lin, & Kieff, 1996). Thus, E2 and E3 interaction with CBF1 is presumably mutually exclusive. Microarrays have uncovered that E3A not only represses but also activate host genes. Target genes of E2 and E3A have been sown to overlap, E2 and E3A either act in concert or counteract. For example, E3A antagonizes the *MYC* activation by E2 (Hertle et al., 2009; McClellan et al., 2013). It has further been shown by our laboratory that E2 and E3A directly compete for CBF1 binding (Harth-Hertle et al., 2013) at enhancers resulting in either activation (E2) or repression (E3A) of the enhancer. It still remains to be unraveled how widely E2 and E3A have antagonistic roles in the host gene regulation (reviewed in Allday, Bazot, & White, 2015).

1.2 Epigenetic regulation

Epigenetics is the "study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence" (Wu Ct & Morris, 2001). Epigenetic changes are not based on DNA sequence changes such as mutations or chromosomal recombinations, but modifications which result in phenotypic changes. Epigenetic mechanisms include DNA methylation, covalent and noncovalent chromatin variations and expression of non-coding RNAs. DNA methylation is defined as methylation of cytosine residues of CpG sites. Covalent chromatin variations are histone modifications, while non-covalent mechanisms include chromatin remodeling or the incorporation of special histone variants (Goldberg, Allis, & Bernstein, 2007). Chromatin remodeling is defined as the ATP-dependent change in nucleosome positioning.

1.1.3 Histone modifications

Histones (H) are the components of the nucleosome, which is the basic unit of chromatin. Chromatin is the compact form of DNA, where the helix winds around a histone octamer comprising of two dimers of H2A an H2B and a tetramer of H3 and H4 histone variants. The histone tails stick out of the nucleosome and can be post-translationally modified. Acetylation, phosphorylation, methylation, and ubiquitylation are the most common modifications among many others. Posttranslational modifications (PTM) on histones can be indicatory for transcriptionally active or silent chromatin. These histone modifications can be assessed by chromatin immunoprecipitation (ChIP). High levels of acetylated lysine on the H3 and H4 tails (e.g. H3K27ac), trimethylation of lysine 4 on H3 (H3K4me3) or trimethylation of H3 lysine 79 (H3K79me3) are among others marks for active genes. However, marks for inactive genes include trimethylation of lysine 27 on the H3 (H3K27me3) or trimethylation of H3 lysine 9 (H3K9me3). Sequence-specific TFs, which regulate transcription, can recruit chromatin-modifying enzymes to target sites. Histone acetylation for instance is dynamic and regulated by two contrary acting enzymes, the histone acetyltransferases (HATs; e.g. CBP/ p300) and histone deacetylases (HDACs). Histone methylation mainly occurs at lysines or arginines and is regulated by methyltransferases and demethylases. Among the histone lysine methyltransferases (HKMTs) are mixed lineage leukemia 3 and 4 (MLL3, MLL4 respectively), two H3K4 monomethyltransferases or polycomb repressive complex 1 and 2 (PRC1, PRC2 respectively), two H3K27 trimethyltranferases. Histone modifications regulate chromatin structure by recruitment of remodeling enzymes. As a result, these modifications can influence transcription (Figure 2; reviewd by Zhang, Cooper, & Brockdorff, 2015; Bannister & Kouzarides, 2011).
The so-called histone code hypothesis suggests specified functions for genomic elements according to distinct combinatorial patterns of histone modifications (reviewed by Rando, 2012).

Additionally to histone modifications, the different classes of elements are marked by distinct patterns of TF binding (Heintzman et al., 2007). Therefore, the chromatin can be categorized in different segments according to their histone modifications and other features. Ernst *et al*. published 2011 a chromatin state segmentation (CSS) according to the Hidden Markov Model (HMM) for the nine human cell lines included in the ENCyclopedia Of DNA Elements (ENCODE) project (i.a. GM12878, a EBV-immortalized B cell; Ernst et al., 2011). Focusing on enhancers, they can be divided into strong, weak and poised enhancers according to their chromatin state. They mainly act as *cis*- regulatory elements that are bound by specific TFs to enhance the transcription of corresponding genes, irrespective of their orientation and location relative to the promoters. A promoter is comprised of two elements. The core promoter corresponds to the region around the transcriptional start side (TSS) and is required for the initiation of transcription and the recruitment of RNA polymerase II (PolII). The proximal promoter resides upstream of the TSS and contains several TF-binding sites of the corresponding genes (Lee, Hsiung, Huang, Raj, & Blobel, 2015). Enhancers are believed to be essential for tissue specificity and developmental regulatory gene expression (Bulger & Groudine, 2011; Plank & Dean, 2014). They exhibit epigenetic characteristics: In general, an open chromatin architecture (DNase I hypersensitive sites), binding sites for PolII (RNA Polymerase II) and coactivators like the mediator complex or p300/CBP, and especially histone modifications such as H3K4me1 or H3K27ac (histone 3 Lysine 4 monomethylation, histone 3 lysine 27 acetylation respectively; Ernst et al., 2011).

Recently, regions which are basically comprised of multiple adjacent enhancers and bound by groups of specific TFs were defined as super-enhancers. These enhancers drive genes important for cell function and identity. They show exceptionally broad and high binding signals for H3K27ac, mediator 1 (Med1) and bromodomain-containing protein 4 (BRD4; Whyte et al., 2013; Hnisz et al., 2013).

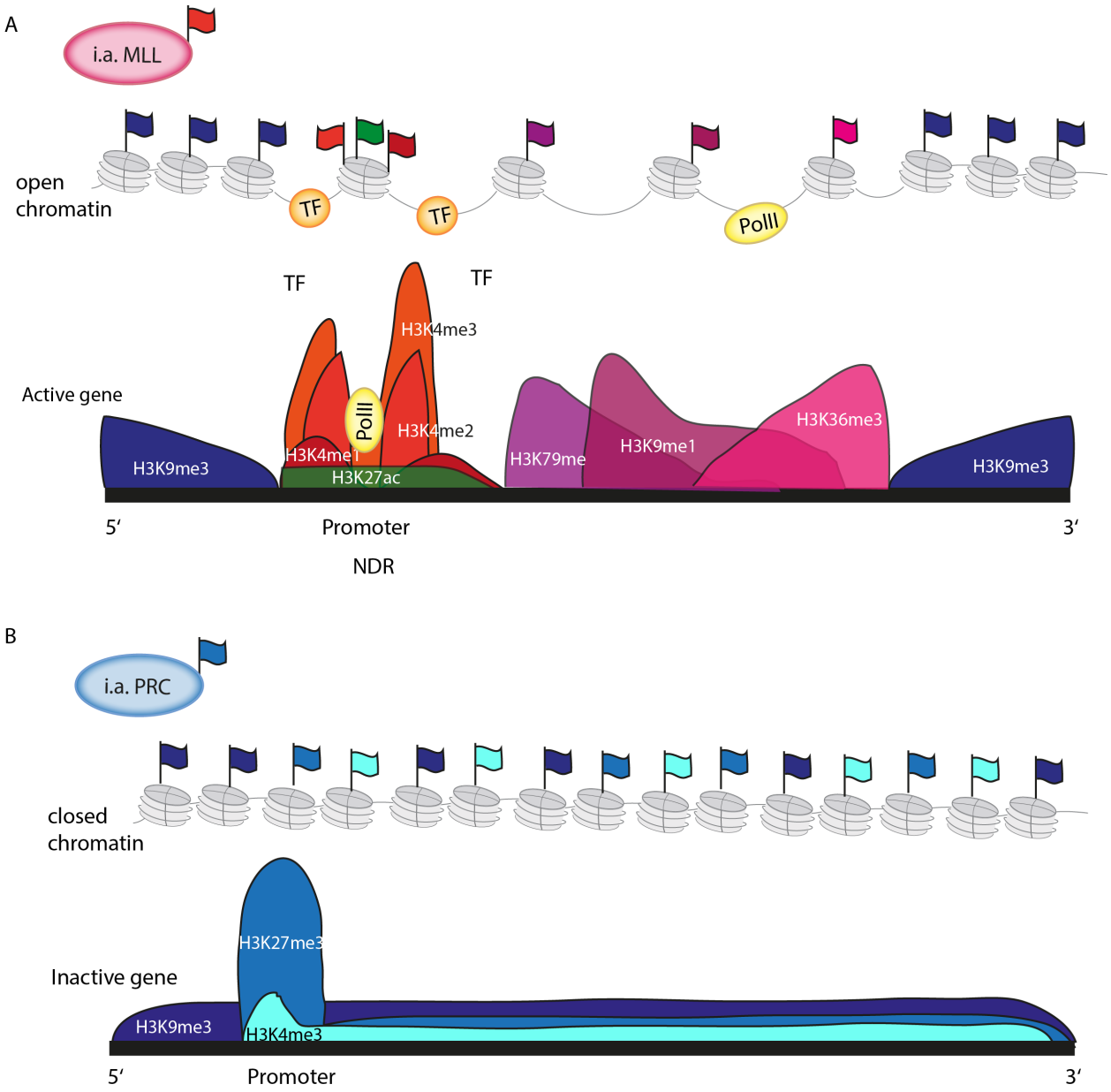


Figure 2: Schematic representation of ChIP-signals of post-translational histone modifications of active (A) and repressed (B) genes. A Exemplified PTMs found at the promoter regions and the gene bodies of actively transcribed genes. A representative histone modifier MLL mediates H3K4 monomethylation. Open chromatin is displayed by a slack nucleosome chain (NDR=nucleosome depleted region). **B** Exemplified PTMs found at the promoter regions and the gene bodies of silenced genes. A representative histone modifier PRC mediates H3K27 trimethylation. Closed chromatin is displayed by a dense nucleosome chain (Figure modified from Barth & Imhof, 2010).

1.1.4 Long-range chromatin interactions and three-dimensional (3D) genome organization

Chromosome conformation capture (3C)-based techniques present opportunities to explore chromatin interactions and 3D-genome organization in an unprecedented scale and resolution. These methods provide resolution on the 3D organization of the human genome, which appears to be essential for gene regulation (Bickmore, 2013; Dekker, Marti-Renom, & Mirny, 2013; Sexton & Cavalli, 2015). The "family of 3C-techniques" aims to detect physical interactions of genomic regions and involve five steps: crosslinking the chromatin at sites of physical interactions by formaldehyde fixation, shearing the chromatin by sonication or restriction enzyme digest, ligation under dilute conditions in order to bias towards a ligation between DNA ends of close proximity, detection of ligation junctions and computational calculation of interaction frequencies (Hakim & Misteli, 2012). Distal regulatory elements such as enhancers need to be physically conjoined with their target genes on DNA level. At a higher level, topologically associating domains (TADs) have been suggested to be a superordinate unit of mammalian genome organization. A TAD is considered as self-interacting genomic region and up to 2 Mb in size. A TAD is defined by applying certain algorithms to Hi-C data (Dixon et al., 2012). DNA sequences within a TAD physically interact more frequently with each other than with sequences outside of the TAD. The protein CTCF and the protein complex cohesin are thought to be important for TAD formation (Pombo & Dillon, 2015). TADs have been reported on the basis of lower-resolution contact maps. With increasing resolution, much smaller (median length= 185 kb) contact domains could be observed, too small to be detected in previous maps. These domains were also conserved across cell types and exhibited patterns of long-range contacts (subcompartments). Detection of TADs involves the detection of domain boundaries, which were not as distinct with higher resolution. Additional boundaries were reported beyond those of previous maps, which were associated with subcompartment transition or looping (Rao et al., 2014). Recently, several different high-throughput technologies based on Chromatin Conformation Capture (3C) have been developed such as Hi-C (simultaneously capturing all genomic interactions as a population-average snapshot; Lieberman-Aiden et al., 2009), ChIA-PET (chromatin interaction analysis by paired-end tag sequencing; G. Li et al., 2010) or Capture Hi-C (hybridization selection to capture interactions of candidate fragments; Mifsud et al., 2015). ChIA-PET combines chromatin immunoprecipitation (ChIP) with a 3C method to enrich for interactions mediated by one TF.

Looping data obtained for the EBV-immortalized B cell GM12878 cells through Capture Hi-C are provided by Mifsud *et al*. Furthermore, CTCF-mediated chromatin conformation was assessed in GM12878 cells by ChIA-PET (Szalaj et al., 2016; Tang et al., 2015b). Using Hi-C, the genomic architecture of the genomes of the nine ENCODE cell lines was investigated (Rao et al., 2014). GM12878 cells showed the densest organization with 4.9 billion contacts.

1.1.5 Non-coding RNAs (ncRNAs)

The fraction of the genome coding for proteins constitutes approximately 1.2 %. Many regulatory elements are transcribed into non-coding RNAs (Human Genome Sequencing Consortium International 2004). This is indicatory for a substantial role of ncRNAs in complex organisms. RNAs are also involved in epigenetic events. Non-coding RNAs include a huge variety of small non-coding RNAs like microRNAs (miRNAs), small interfering RNAs (siRNAs) or small nucleolar RNAs (snoRNAs) and the long non-coding RNAs (lncRNAs). Small RNAs have been shown to induce posttranscriptional and transcriptional RNA interference (RNAi)-related pathways. They collaborate with the DNA methylation machinery or components of the chromatin (Goldberg et al., 2007). LncRNAs also possess the potential to influence epigenetic processes such as DNA methylation, histone modification activity or posttranscriptional regulation, since they exhibit complex structural features (reviewed in C. Wang et al., 2017).

1.3 Long non-coding RNAs (lncRNAs)

The quantity of ncRNAs steadily increases due to identifications through genome-wide human transcriptional studies. NONCODE for example is an integrated knowledge database designed for ncRNAs, despite transfer RNAs (tRNAs) and ribosomal RNA (rRNAs) and especially the number of lncRNAs has increased from the NONCODE version 3.0 to version 4.0 (two years) by almost 3 fold from 73,327 to 210,831 (Xie et al., 2014; for information on databases see below). The expansion of the regulatory potential of ncRNAs might be a reason for the evolution of developmental processes, which could be responsible for the complexity of organisms (Mattick, 2004).

1.3.1 Definition of lncRNAs

LncRNA are distinguished from small ncRNAs by the size. LncRNA transcripts are longer than 200 nt and have no coding potential. Additionally, they are poorly conserved compared to small non-coding RNAs (only a small number is conserved across species like *XIST, NEAT1 or MALAT1*). The bulk of lncRNAs are transcribed by PolII (confirmed by PolII occupancy) like messenger RNAs (mRNAs). They can be post-transcriptionally modified by splicing, capping and polyadenylation. Generally lower expression, a fewer number of exons and a much higher tissue specificity distinguish them from mRNA (Derrien, 2012; Iyer, et al., 2015). For further characterization, they can be divided into subgroups according to their genomic location relative to protein-coding genes, the sequence, the structure and their functional features. Until now, there is no consensus on the

classification and the nomenclature. Types of classifications consider the genomic location, such as intergenic lncRNA (lincRNA) or intronic lncRNA, or the orientation of the product regarding the DNA strand, such as sense or antisense lncRNAs, as well as the association with known chromatin states like enhancer or promoter associated lncRNAs (eRNAs, pRNAs respectively). Especially eRNAs can be product of either unidirectional or bidirectional transcription. The classifications might overlap (Figure 3A). It has to be mentioned here, that no clear definition regarding the kind of association of a lncRNA with a chromatin state exists. For instance, the association might be either an intersection of e.g. the transcription start side or the entire gene body of the lncRNA with a certain chromatin state (Bonasio & Shiekhattar, 2014; Fritah, Niclou, & Azuaje, 2014; Rashid, Shah, & Shan, 2016; Salviano-Silva, Lobo-Alves, Almeida, Malheiros, & Petzl-Erler, 2018; K. C. Wang & Chang, 2011).

1.3.2 Cellular functions of lncRNAs

The location of the lncRNA might occasionally determine the functional context of a lncRNA (reviewed in C. Wang et al., 2017). LncRNAs can exert diverse transcriptional or post-transcriptional functions in the nucleus as well as in the cytoplasm as described in the following section. In some cases it appears that the lncRNA transcription rather than the lncRNA itself is regulatory. The expression of lncRNAs is strictly regulated and cell type-/ tissue-specific, hinting towards a crucial role in physiological mechanisms. Variations of their expression or mutations in their primary sequence have been linked to disorders. Despite a missing conservation at primary sequence level between lncRNAs, parallels can be found in their mode of action, since lncRNAs can bind to other molecules such as proteins, DNA or RNA (Bonasio & Shiekhattar, 2014; Fritah et al., 2014; Rashid et al., 2016; Salviano-Silva et al., 2018; K. C. Wang & Chang, 2011). There is evidence that the majority of nascent RNA generated in the nucleus is rapidly turned over (Lam, Li, Rosenfeld, & Glass, 2014).

1.3.2.1 LncRNAs in the nucleus

Initially, lncRNAs were thought to primarily reside in the nucleus. Wang & Chang reviewed the molecular mechanisms for functions of lncRNAs in 2011 and suggested four archetypes of mechanisms: signals, decoys, guides and scaffolds (Figure 3B). Individual lncRNAs may realize several archetypes. Since 2011, multiple other mechanisms for lncRNAs were discovered (reviewed in Salviano-Silva et al., 2018).

Since the expression of lncRNAs is cell type specific, they could serve as molecular signals in response to certain stimuli, interpret a cellular context or integrate developmental clues. They function as indicators of transcriptional activity (e.g. *HOTAIR*; Rinn et al., 2007) LncRNAs can furthermore positively or negatively regulate transcription. The decoy lncRNA binds and titrates away a protein target, which could be a TF or a chromatin modifier for instance (e.g. *MALAT1*; Tripathi et al., 2010). The guide archetype binds to protein(s) and directs leads the ribonucleoprotein complex to specific targets. Changes in gene expression can be triggered in *cis* (e.g. *XIST*; Wutz, Rasmussen, & Jaenisch, 2002) or in *trans* (e.g. *JPX*; Tian, Sun, & Lee, 2010). Finally, lncRNAs can provide a platform for the assembling of several molecular components. This is a complex class of lncRNAs with different domains binding to distinct effector molecules. By binding multiple different effectors, these molecules are combined in time and space (e.g. *ANRIL*; Kotake et al., 2011; Yap et al., 2010).

<u>eRNAs</u>

Combining the outcomes of several deep sequencing approaches, eRNAs were defined as following. Putative enhancer regions, marked by high levels of H3K4me1 give rise to eRNAs, and their expression is additionally characterized by H3K27ac modification. It was demonstrated that the histone methyltransferases MLL3/4 promotes the synthesis of eRNA (Dorighi et al., 2017). These enhancer regions can be associated with binding of LDTFs (lineage determining TFs), transcriptional co-activators (e.g. Mediator, p300, CBP), PolII (serine 5 phosphorylated) and more. In general, eRNAs exhibit a 5' cap and are predominantly monoexonic. eRNAs which display polyadenylation are linked to unidirectional transcription, whereas the transcripts without polyadenylation are linked to bidirectional transcription. The latter are more common (Lam et al., 2014). eRNAs have a half-life of approximately 2 min compared to the approximately 80 min half-life of mRNAs (lincRNAs similar to mRNA, other lncRNAs approx. 7 min (Schwalb et al., 2016). eRNAs are dynamically regulated upon stimuli. eRNAs are enriched at enhancers which are engaged in chromatin looping, which hints towards a potential functions of eRNAs in looping formation. Enhancer transcription could be simply transcriptional noise at open chromatin, the process of enhancer transcription could be important or the RNA transcript itself is important for enhancer activity. Numerous reports imply a contribution of eRNAs in enhancer mediated

activation of neighboring coding genes. This contribution could be in the facilitation of proper formation of chromosomal looping between enhancers and TSS (reviewed in Lam et al., 2014).

1.3.2.2 LncRNAs in the cytoplasm

Many lncRNAs reside in the cytoplasm and exert their function there. Rashid *et al*. summarized their functions in four groups: modulation of mRNA stability, modulation of translation, competing endogenous RNAs and mediation of protein modifications (Figure 3C).
Several lncRNAs are known to target mRNA transcripts and modulate their stability, some of them increase mRNA stability by e.g. sequestering STAU1 (*TINCR*; Kretz, 2013) and others decrease the stability e.g. by recruiting STAU1 (*½-sbsRNA*; Gong & Maquat, 2011; Kim et al., 2007). Staufen 1 (STAU1)-mediated mRNA decay (SMD) is induced when the 3' untranslated region (UTR) of a mRNA binds to STAU1. Involvement of lncRNAs in translational regulation has also been reported in order to manage complex protein dynamics in a spatio-temporal manner. LncRNAs have been observed to promote translation by activating polysomes for cap-independent translation (*AS Uchl 1*; Carrieri et al., 2012) or inhibit translation by enhancing the translational repressor machinery (*lincRNA-p21*; Yoon et al., 2012). There is a competition about miRNA binding between coding and non-coding RNAs. Competing endogenous RNAs (ceRNAs) can protect coding genes from repression by sequestering miRNAs and therefore hinders the miRNA from binding to its targets (e.g. *HULC*; J. Wang et al., 2010). LncRNAs furthermore can modulate modifications of cytoplasmic proteins like phosphorylation/ dephosphorylation or ubiquitilation/ deubiquitilation. An example for modulation of phosphorylation is the lncRNA *NKILA* which binds directly to IκB and hinders IκB kinase (IKK) from phosphorylating IκB (Liu et al., 2015).

An example for a compartment independent function of lncRNAs is that they can serve as precursors for miRNAs (e.g. *H19*; Dey, Pfeifer, & Dutta, 2014).

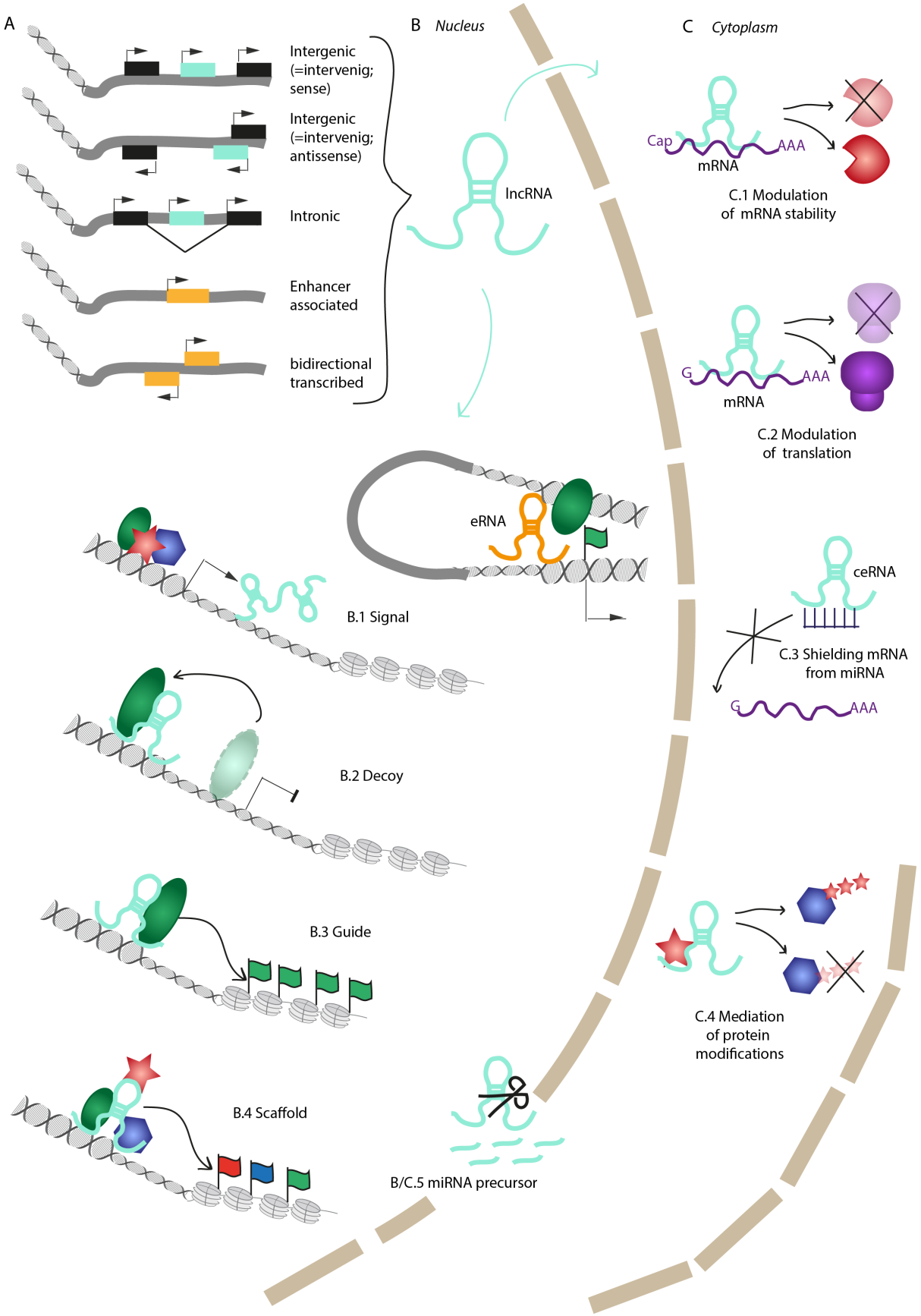


Figure 3: Definition and function of lncRNAs. A Definition for lncRNAs based on their genomic location relative to protein-coding genes lincRNA/intronic lncRNA; the orientation of the product regarding the DNA strand sense or antisense lncRNAs, as well as the association with known chromatin states (eRNAs). **B, C** Regulatory

mechanisms of lncRNAs in the nucleus (B) and the cytoplasm (C). B.1 Molecular signals, activating or silencing gene expression e.g. acting as eRNAs, inducing transcription in cis or in trans B.2 Decoy for regulatory proteins, such as transcription factors and chromatin modifiers B.3 Guiding proteins (in general, chromatin modifiers) to specific target sites, e.g. as eRNAs recruiting proteins such as mediator to format chromatin loops B.4 As scaffolds, binding different proteins and forming RNP complexes C.1 Modulators of mRNA stability C.2 Modulators of translation C.3 CeRNAs as shields for mRNA C.4 Mediators for protein modification; Serving as precursors for miRNA (compartment independent; Figure modified from Salviano-Silva, Lobo-Alves, Almeida, Malheiros, & Petzl-Erler, 2018).

1.3.3 Databases dedicated to lncRNAs

The first evidence for lncRNA transcription was revealed when after cloning no translated products could be detected. Since then, major progress in the detection of lncRNAs was made by tilling arrays and more recently with deep sequencing methods. For 50 out of 700 lncRNAs, a significant effect on cancer cell growth could be observed in a recent large-scale lncRNA knock out screening (Zhu et al., 2016). Although it is evident, that individual lncRNAs exert important functions in diverse biological processes, a large gap exists between the huge amount of detected lncRNAs and an associated proven molecular or cellular function. Still, lncRNAs become increasingly available in public datasets. There were several databases established which differ in quality and data coverage. First, the number of lncRNAs contained in these databases varies from < 10,000 to > 70,000 and second the determination varies from experimentally confirmation to bioinformatical prediction (Fritah et al., 2014). Xu *et al.* developed LNCat (lncRNA atlas, freely available at http://biocc.hrbmu.edu.cn/LNCat/), a comprehensive database for lncRNAs by reviewing 24 lncRNA annotation resources referring to >205,000 lncRNAs in over 50 tissues and cell lines. Furthermore, the resources were characterized with respect to exon structure or expression for instance. This atlas contains three of the largest and best known resources for lncRNAs, GENCODE, LNCipedia and NONCODE (J. Xu et al., 2016).

1.3.4 LncRNAs in human diseases

The combination of an epigenetic function with tissue specificity, variability and plasticity suggest that lncRNAs are crucial factors in disease genesis (C. Wang et al., 2017). Since lncRNAs have fundamental functions in maintaining cellular and organismal homeostasis, dysregulation of lncRNAs was demonstrated in diverse studies to be associated with a huge variety of diseases including cancer (Hu et al., 2018). The lncRNA *DSCAM-AS1* for example is upregulated in breast cancer and mediates tumor progression and tamoxifen resistance by targeting heterogeneous nuclear ribonucleoprotein L (hnRNPL; Niknafs et al., 2016). lncRNAs have already been reviewed to exert functions during malignant hematopoiesis (Alvarez-dominguez & Lodish, 2014; Alvarez-Dominguez & Lodish, 2017). For instance, maternally expressed gene 3 (*MEG3*) was observed to

be down-regulated in acute myeloid leukemia (AML) by hyper-methylation of its promoter (Benetatos et al., 2010) and it might be involved in the regulation of the retinoblastoma (*Rb*) and *p16INK4a* pathway and thus in cell proliferation of many cancer types (Benetatos, Vartholomatos, & Hatzimichael, 2011). In this respect, Yan *et al.* analyzed the Cancer Genome Atlas (TCGA) data regarding alterations at transcriptional, genomic and epigenetic levels and identified potentially clinically relevant noncoding transcripts (Yan et al., 2015). Here, they included lncRNAs of 5,037 human tumor specimens across 13 cancer types. They could show that the dysregulation of expression of lncRNAs is common in cancer, with the majority of lncRNAs being cancer type unique but some alterations are also shared between different cancer types. Furthermore, they suggest that somatic copy number alterations lead to the dysregulation of lncRNAs in cancer as well as epigenetic silencing of lncRNAs. Moreover, their data suggest a determination of tumors by lncRNAs as biomarkers.

1.4 EBV and chromatin conformation regulation of the host genome

By next generation sequencing techniques combining ChIP-Sequencing, RNA-Sequencing, 3C-Sequencing and other sequencing methods, a mass of data regarding genome regulation is available. Comprehensive data sets provided by the ENCODE project on functional DNA elements can be obtained for an EBV-immortalized B cell, GM12878. E2 preferentially binds to active chromatin of enhancers (Glaser, PhD thesis, 2017; Zhao et al., 2011), the same holds true for E3A (Glaser, PhD thesis, 2017; Zhou et al., 2015), indicating that enhancers are frequently occupied target sites for the transcriptional regulation by EBV. The genome can be regulated over a long distance, mediated by chromatin contacts. Enhancers can loop to promoters, two or more genomic regions can be connected with each other (reviewed in Pombo & Dillon, 2015). EBV has been reported to rearrange enhancer-promoter loops. It has been shown, that *MYC* expression can be activated by E2 from distal enhancers 3' of the transcription start side of *MYC* (Wood et al., 2016; Zhao et al., 2011). Furthermore, it was observed that E2 binding sites accumulate in super-enhancers and that these E2 super enhancers were not close to known TSSs. It could be revealed that all EBNA proteins co-occur at the same enhancer sites in the genome (EBV enhancers) and that 10% of them show hallmarks of super-enhancers (EBV super-enhancers, ESE). Furthermore, it could be shown that most of the EBV super-enhancers reside in the same TAD as their corresponding genes (Zhou et al., 2015). Liang *et al.* could report that E2 regulates the eRNAs *MYC*-428 and *MYC*-525 which are derived from ESEs and that these eRNAs regulate *MYC* expression (Liang et al., 2016).

1.5 Objectives

In the Kempkes laboratory, extensive investigations on E2 target genes have already been conducted. More recently, microarrays on E2 (Thumann, PhD thesis, 2016) and E3A targets (Hertle et al., 2009) gave additional insights into transcriptional regulation by both viral TFs. This, and investigations by other research groups revealed that E2 and E3A share target genes with subsets of counter- and co-regulated genes (Glaser, PhD thesis, 2017). Furthermore, it could be published by our group that E2 and E3A compete for CBF1 binding to an enhancer at a model locus (Harth-Hertle et al., 2013). The major anchor for E2 is CBF1, however, our laboratory could observe CBF1 independent binding of E2 to chromatin. This raised the question if alternative cellular TFs can be used by E2 as anchor to DNA as well. Genome-wide binding analyses comparing E2 with 89 other cellular TFs revealed also a high correlation of E2 with the B cell TF EBF1. To functionally investigate this correlation, ChIP experiments followed by quantitative polymerase chain reaction (qPCR) were performed.

As could be reported for E2 and E3A, both TFs preferentially bind to enhancer regions (Glaser, PhD thesis, 2017; Zhao et al., 2011; Zhou et al., 2015). Additionally to the competition of E2 and E3A for enhancer binding, it was observed by our group that E3A represses transcription deriving from an intergenic enhancer at a well-studied model locus, *CXCL9/10*, as assessed by RT-qPCR (quantitative reverse transcription polymerase chain reaction; Figure 4). Using ChIP-primer for E2 and E3A binding sites on cDNA at an enhancer locus which is marked with features of active chromatin (high GRO-Seq, H3K27ac and H3K4me1 signals, PolII- and CTCF-binding), we noticed E3A-dependent regulation of transcription of the enhancer and the neighboring genes. These enhancer transcripts could be enriched in the nucleus. We hypothesize that E2 and E3A control the transcription of lncRNAs which are presumably involved in the regulation of neighboring or remote coding genes (Figure 5). The regulation by E2 and E3A could occur on a higher level of chromatin organization. E2 and E3A might counter-regulate shared target genes genome-wide. Our previous work has shown in a microarray analysis that E2 can induce multiple lncRNAs (Glaser et al., 2017; Thumann, PhD thesis, 2016). Many lncRNAs have been reported to have certain relevance in human diseases such as cancer (Hu et al., 2018). Since EBV is associated with several lymphomas and other cancers (Ko, 2015; Pattle & Farrell, 2006), one conceivable mechanism for the challenge of cancer could be the induction or repression of lncRNAs, which again could have an impact on the regulation of protein coding genes.

For the genome-wide detection of E2 and E3A target genes RNA-Sequencing was conducted, comparing the cytoplasmic with the nucleic compartment of established lymphoblastoid cells. The main obstacle was the approach for the analysis and the accompanied parameters, since the bioinformatic thresholds and definitions have a major impact on the outcome of the data. The resulting data were analyzed in context of information collected by our lab and publicly available

information on LCLs (GM12878). ChIP-Seq Data on E2 and E3A were acquired in our lab, as well as a comprehensive cluster analysis for EBNA peaks (pan EBNA, merged E2, E3A and E3C peaks), which was used in the following to describe E2 binding sites relevant for the analysis.

Despite numerous efforts to comprehensively catalog lncRNAs, there are several reasons why the identification of lncRNAs is incomplete, for instance the high tissue- and pathway-specificity of lncRNA expression (Iyer, et al., 2015). The main aim of this thesis was the identification and characterization of already annotated and potentially novel E2-regulated lncRNAs.

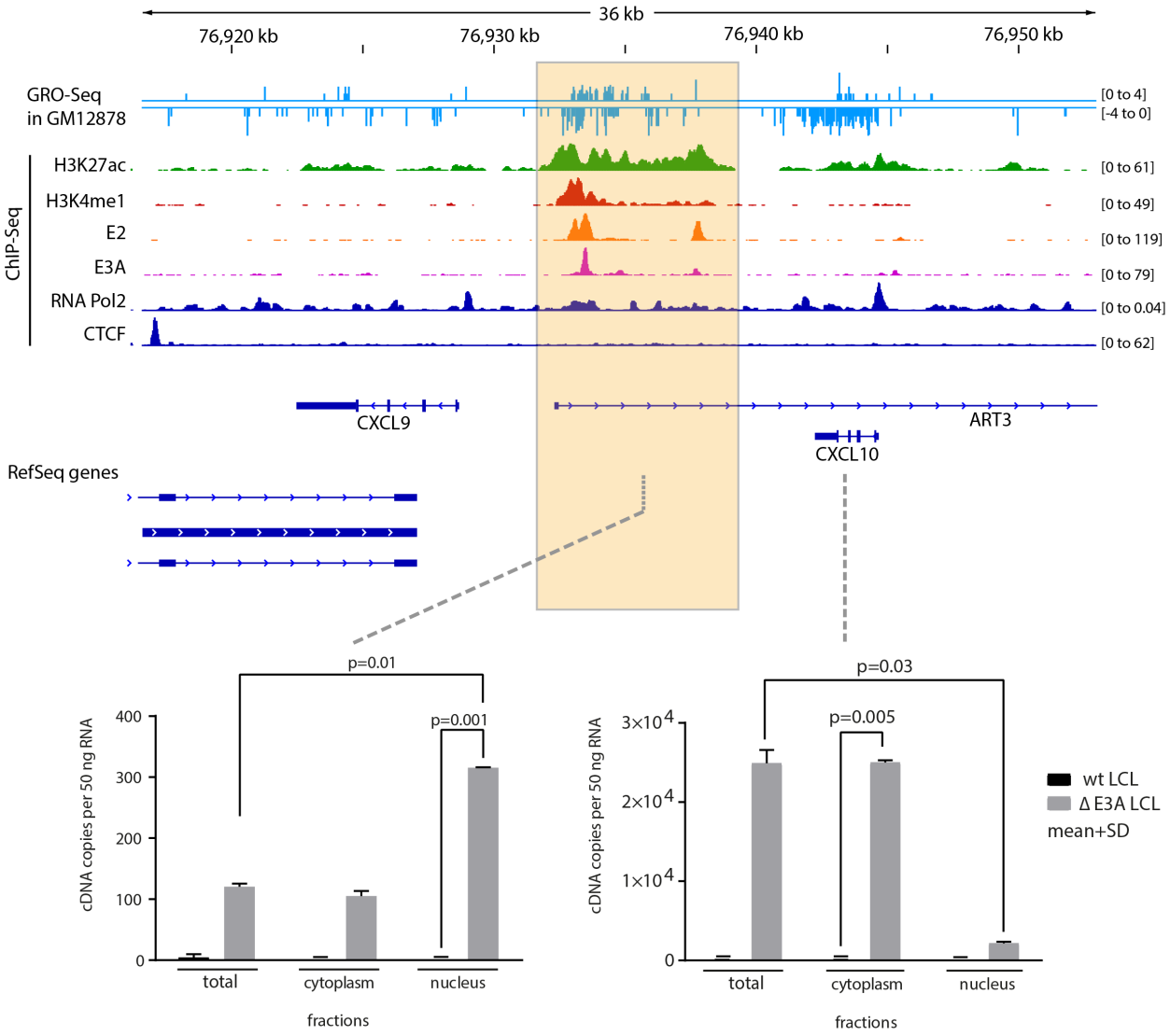


Figure 4: E3A-dependent repression of transcription of intergenic enhancer at model locus. Upper panel: IGV image displaying a GRO-Seq track showing the coverage in GM12878 (Core LJ, et al. 2014; data range set to auto scale), tracks obtained from ChIP-Seq (H3K27ac, H3K4me1, RNA PolII and CTCF sourced from ENCODE for GM12878; E2 and E3A generated in our lab; data range set to auto scale) and a RefSeq annotation track; Lower panel: Bar graphs showing the absolute quantification of RT-qPCR of the intergenic region (left) and the protein coding gene *CXCL10* in different RNA preparations. RNA was isolated from 10^7 wt and ΔE3A LCLs (total) or cell fractions (1.2 x 10^7 cells for cytoplasm and 2 x 10^8 cells for nucleus) and 2 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.}$ =3). Statistical significance is indicated (paired t-test).

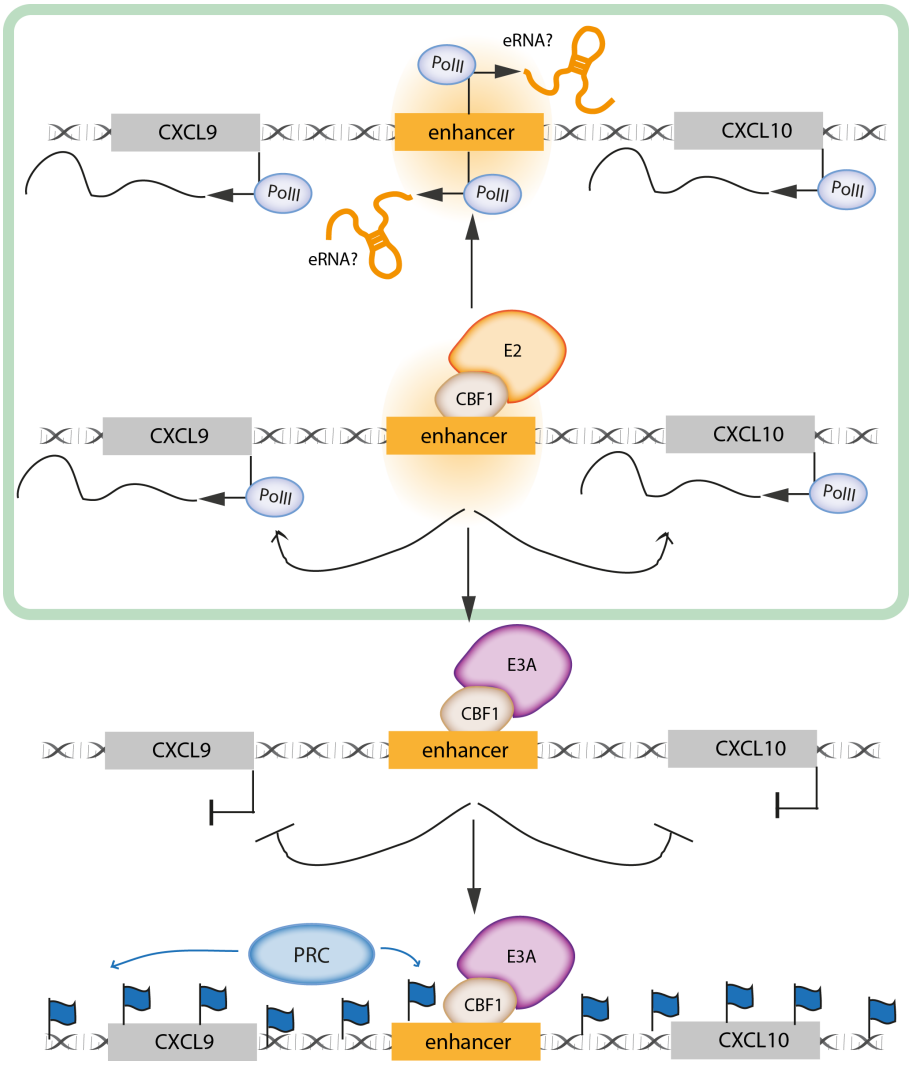


Figure 5: Working hypothesis - Control the transcription of lncRNAs by E2 and E3A which might be involved in the regulation of neighboring/remote coding genes. Extension of the proposed model for E3A by the E2 counter-action. The transactivator E2 binds through CBF1 to intergenic regions (preferably enhancers), positively influencing the transcription of neighboring genes by the induction of a lncRNA (eRNA). E3A displays E2 from CBF1, negatively influencing the transcription of neighboring genes by repression of a lncRNA. The repression of the chromatin domain is sealed by the polycomb repressive mark H3K27me3 (blue flags) mediated by PRC proteins (Figure modified from Harth-Hertle et al., 2013). The regulation by E2 and E3A could occur on a higher level of chromatin organization.

2 Material & methods

2.1 Cell culture related information

2.1.1 Donor samples

Surgically removed adenoids were obtained from anonymous donors from Munich, Germany. The local ethics committee (Ethikkommission bei der LMU Muenchen) approved the use of this human material.

2.1.2 Isolation of human primary cells

Human primary B cells were isolated from adenoids by Ficoll gradient centrifugation with Ficoll-Paque Premium (density 1.077+/-0.001 g/ml; GE Healthcare). T cell numbers were minimized by erythrocyte rosetting. Successful depletion of T cells was verified by staining for the B cell marker CD19 (α-CD19 antibody: APC Mouse Anti-Human CD19, Cat#555415, BD Pharmingen; isotype CNTRL: Mouse IgG1 neg. CNTRL APC, #MCA928APC, Serotec) and for the T cell marker CD3 (α-CD3 antibody: PE Mouse Anti-human CD3, Cat# 555333, BD Parmingen; isotype CNTRL: Mouse IgG1 neg. CNTRL RPe, #MCA928PE, Serotec) and subsequent analyzation by flow cytometry using the FACSCalibur (BD Biosciences) and FlowJo software (BD Biosciences). Primary B cells were seeded in RPMI 1640 medium (Gibco Life Technologies) supplemented with 20 % FCS (fetal calf serum, Bio&Sell), 4 mM L-Glutamine and 0.5 x penicillin/streptomycin (Gibco Life Technologies), 1 mM Pyruvat (Gibco Life Technologies), 100 µM α-TG (SIGMA), 1 % MEM NEAA (Gibco Life Technologies).

2.1.3 Cell Lines and cell culture conditions

DG75 cells are an EBV negative Burkitt's lymphoma cell line. They were maintained as suspension cultures in RPMI 1640 medium (Gibco Life Technologies) supplemented with 10 % FCS (fetal calf serum, Bio&Sell), 4 mM L-Glutamine and 1 x penicillin/streptomycin (Gibco Life Technologies). They were split twice a week according to their density 1:3 to 1:4. The **DG75$^{doxHA\text{-}E2}$/CBF1 wt** (CKR128-34) and the **DG75$^{doxHA\text{-}E2}$/CBF1 ko** (CKR178-10) cell lines carry the Dox inducible HA-E2 expression plasmid (described in Glaser et al., 2017), where eGFP and NGFR are simultaneously expressed together with HA-E2. E2 expression can be induced by doxycycline treatment (1 µg/ml). **Jurkat cells** are immortalized human T lymphocyte cells. They were maintained as suspension cultures in RPMI 1640 medium (Gibco Life Technologies) supplemented with 10 % FCS (Bio&Sell),

4 mM L-Glutamine and 1 x penicillin/streptomycin (Gibco Life Technologies). They were split twice a week 1:5.

The **LCL ER/EB2-5** has been described before (B Kempkes et al., 1995). It is an EBV immortalized B cell line that has been infected with the mini-EBV plasmid 554-4 (E2 expression plasmid flanked by genomic sequences of the B95-8 virus strain) and EBV strain P3HR1 that is deleted for the open reading frame of E2. ER/EB2-5 cells express an E2 open reading frame fused to the estrogen receptor hormone binding domain (ER/E2). For analysis of E2 dependent gene regulation, medium of ER/EB 2-5 cells was depleted for estrogen (3x washing with estrogen free medium) and cells were cultured in estrogen free medium for 4 days. For cycloheximide (ChX) treatment 50 µg/ml of the protein synthesis inhibitor (Sigma) was added after depletion for 1 h. For the reactivation of E2, 1 µM β-estradiol (E2758 Sigma) was added to the estrogen-free cell culture medium and after 6 h cells were harvested for further analysis (Figure S1). They were split twice a week 1:3.

The **LCL EB2-3** has been described before as well (B Kempkes et al., 1995). It was used as a control (CNTRL) cell line for ER/EB2-5. It originates from the same donor as ER/EB2-5 and got infected with P3HR1 and the mini-EBV plasmid 554 (normal E2 protein expression). They were split twice a week 1:3.

The **wt** (CP364-1) and **ΔE3A mutant** (CP364-42; "E3AmtB") **LCLs** were described before (Hertle et al., 2009). Both LCLs originate from primary B cells of the same donor infected with recombinant EBV, either the wt BAC (2089) or a ΔE3A BAC (Be715). They were split twice a week 1:4 to 1:5.

The cell line **ΔE3A-LCL$^{doxHA\text{-}E3A}$** (MH1680-9) was established by transfection of the ΔE3A mutant cell line with respective pRTS-1 derivatives (MH1680-9; Harth-Hertle et al., 2013). Addition of doxycycline (1 µg/µl) leads to the simultaneous expression of N-terminal HA-tagged E3A and nerve growth factor receptor (NGFR). They were cultivated in 1 µg/ml puromycin containing media. They were split twice a week 1:4.

All LCL cells were maintained as suspension cultures in RPMI 1640 medium (Gibco Life Technologies) supplemented with 20 % FCS (Bio&Sell), 4 mM L-Glutamine and 1 x penicillin/ streptomycin (Gibco Life Technologies).

Supernatant from the marmoset B cell line B95.8 was used as the source of wild-type virus to generate LCL from primary B cells. 0.5 µg/ml cyclosporin A (Sigma) was initially added to the cultures to inhibit T cell growth.

2.1.4 Flow cytometry

Inducibility of E2 expression in DG75doxHA-E2/CBF1 wt and ko cell lines was evaluated by monitoring the expression of the eGFP surrogate marker of pCKR74.2 (Figure S2). Cells were induced for 24 h with doxycycline (dox), washed with phosphate buffered saline (PBS) and fixed with 0.5 % paraformaldehyde (PFA) in PBS. Inducibility of E3A expression in LCLdoxHA-E3A (MH1680-9) was evaluated by staining for NGFR expression (Figure S3). Cells were induced for 24 h with dox, washed with PBS and stained with 50 µl α-NGFR (= hybridoma supernatant, HB8737-1, E. Kremmer) for 20 min on ice. Subsequently, cells were washed with PCS and stained with 50 µl 1:400 goat α-mouse Cy5 (Dianova; ~0.7µg/µl; 1:1 in glycerol). Afterwards, cells were washed again with PBS and fixed with 0.5 % PFA in PBS. For quantification of induced cells, FACSCalibur (BD Biosciences) and for analysis FlowJo software (BD Biosciences) were applied (Figure S3).

2.2 RNAi related techniques

2.2.1 Transfection

5x 10^6 DG75 cells were transfected by electroporation at 250 V and 950 µF in 250 µl reduced serum media (Opti-MEM, Gibco Life Technologies; without supplements) using 0.4 cm electrode gap cuvettes (Bio-Rad) and the Bio-Rad Gene Pulser.

2.2.2 siRNA knock down in DG75 cells

5x 10^6 cells were transfected with 100 pmol control siRNA-A (siCNTRL) or EBF1 siRNA (siEBF1; both Santa Cruz Biotechnology, sc-37007 and sc-10695) by electroporation. 24 h after transfection, 1x 10^7 induced, siRNA treated cells were harvested for chromatin isolation and 5x 10^6 cells for protein isolation (Figure S4).

2.3 Immunoblotting

5x 10^6 cells were lysed in 200 µl NP-40 lysis buffer (1% NP-40, 150 mM NaCl, 10 mM Tris- HCL pH 7.4, 1 mM EDTA pH 8.0, 3 % Glycerol) for 2 h on ice. 30 µg of total cell lysate were submitted to SDS-PAGE under reducing conditions. Immunoblotting was performed on polyvinylidene difluoride (PVDF) membranes. Western blots were probed with the following primary antibodies: rat α-E2 (R3; IgG2A; E. Kremmer), rat α-CBF1 (RBP-J 7A11, E.Kremmer), rat α-GST (GST 6G9,

IgG2A, E. Kremmer), mouse α-EBF (Santa Cruz Biotechnology, sc-137065) and α -GAPDH (EMD Millipore MAB374). Horseradish peroxidase (HRP)-coupled secondary antibodies (Santa Cruz Biotechnology) and an enhanced chemiluminescence (ECL) kit (GE Healthcare) were used for visualization. For subsequent quantification of protein levels, exposed films were scanned in transmission mode and protein band intensities were determined by densitometry using ImageJ software (http://rsbweb.nih.gov/ij/; Schneider, Rasband, & Eliceiri, 2012).

2.4 DNA related techniques

2.4.1 Chromatin immunoprecipitation (ChIP)

This ChIP protocol is based on reference (Ciccone, Morshead, & Oettinger, 2003) with minor modifications as indicated below. In brief, 10^7 DG75$^{doxHA\text{-}E2}$ cells were harvested and washed twice in ice cold PBS, resuspended in 20 ml RPMI 1640 (Gibco Life Technologies) and formaldehyde (1 % final) was added for crosslinking. The reaction was stopped by addition of glycine (125 mM final) after 7 min and gentle shaking for 5 min at room temperature (RT). Cells were pelleted and washed twice in ice cold PBS. Nuclei were isolated by washing the cells 3x with 10 ml of ice cold Lysis Buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.5 % NP-40, 1x proteinase inhibitor cocktail (PIC, Roche)) and subsequent centrifugation (300x g for 10 min at 4 °C). Nuclei were resuspended in 1 ml Sonication Buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, pH 8.0, 0.5 % SDS, 1x PIC) and incubated on ice for 10 min. Chromatin was sheared to an average size of 200–300 bp by four rounds of sonication for 10 min (30 sec pulse, 30 sec pause) using a Bioruptor device (Biogenode). Cell debris was separated by centrifugation at maximum speed for 10 min at 4 ˚C and chromatin containing supernatants were stored at -80 ˚C or directly used for IP. To prepare input DNA, 12.5 µl aliquots (1/10 of the amount used per IP) were saved at -80 ˚C. For IPs 250 µl chromatin (equals 5x 10^6 cells) were diluted 1:4 with IP Dilution Buffer (12.5 mM Tri-HCl, pH 8.0, 212.5 mM NaCl, 1.25 % Triton X-100, 1x PIC) and incubated with 100 µl of hybridoma supernatant on a rotating platform at 4 ˚C overnight. A combination of E2 and HA-tag specific antibodies (α-E2 R3 (rat IgG2a), α-E2 1E6 (rat IgG2a), and α-HA R1-3F10 (rat IgG1)) was used to precipitate E2 and an isotype-matched unspecific antibody mixture (α-GST 6G9 (rat IgG2a) and α-CD23 Dog-CD3 (rat IgG1) both by E. Kremmer) was used as isotype control. The EBF antibody (C-8; sc-137065, Santa Cruz Biotechnology) was used to precipitate EBF1 and an antibody specific for ovalbumin (M-Ova 3D2, E. Kremmer) was used as an isotype control. Protein G sepharose (GE Healthcare) was equilibrated with IP Dilution Buffer, added to the lysate and incubated at 4 ˚C for 4 h with constant rotation. Beads were extensively washed with: 2x Wash Buffer I (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, pH 8.0, 1 % Triton X-100, 150 mM NaCl, 0.1 % SDS, 1x PIC), 1x Wash Buffer II (20 mM Tris-HCl, pH 8.0, 2 mM EDTA, pH 8.0, 1 % Triton X-100, 500 mM NaCl, 0.1 % SDS, 1x PIC), 1x Wash Buffer III (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, pH 8.0, 250 mM LiCl, 1 % NP-40,

1 % sodium deoxycholate, 1x PIC) for 5 min under rotation, and 2x with TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, pH 8.0) for 1 min. Protein-DNA complexes were eluted with 2x 150 µl Elution Buffer (25 mM Tris-HCl, pH 7.5, 10 mM EDTA, pH 8.0, 1 % SDS) at 65 ˚C for 15 min. Input samples were adjusted to 300 µl with Elution Buffer. Eluates and input samples were incubated with Proteinase K (1.5 µg/µl final, Roche) for 1 h at 42 ˚C. Cross-linking was reversed by incubation at 65 ˚C overnight. DNA was recovered using QIAquick PCR purification kit (Qiagen).

2.4.2 Chromatin immunoprecipitation quantitative polymerase chain reaction (ChIP-qPCR)

qPCR was performed using LightCycler 480 SYBR Green I Master (Roche) on a LightCycler 480 II instrument (Roche) as described previously (Harth-Hertle et al., 2013). Two technical replicates were analyzed for each biological replicate. All primer were established to fit an annealing temperature of 63 ˚C. To account for differences in amplification efficiencies, a standard curve was generated for each primer pair using serial dilutions of sheared DNA (input) as template. DNA quantities detected in input samples were adjusted to the amount of chromatin used per IP by multiplication with 20. Values obtained from IP samples with unspecific IgG CNTRL were subtracted from the DNA amounts recovered by IP with specific antibody. The percent of input was calculated as (DNA from specific IP corrected for IgG CNTRL background/ DNA input) x 100. To validate the ChIP, qPCR known (ChIP-Seq) positive and negative loci were performed. To adjust to divergent E2 inducibility in wildtype and knock out cells, the percent input was calculated relative to a known negative locus (ChIP-Seq; percent input at tested locus/percent input of known negative locus). To display the change in binding, the mean relative input of the wildtype cells treated with control siRNA was set to one. A paired t-test was performed to assess significance of differences of means. Graphs were created using Graph Pad Prism.

2.4.3 Isolation of genomic DNA and quantification by quantitative polymerase chain reaction (qPCR)

For the isolation of complete genomic DNA QIAamp DNA Mini Kit was applied according to the manufacturers' instructions using 5 x 10^6 cells as input material. DNA was eluted in 100 µl H_2O and concentration was monitored using the Qubit 2.0 Fluorometer. Isolated gDNA was quantified using the Roche LightCycler 480 II instrument and LightCycler 480 SYBR Green I Master (Roche) reagent according to the manufacturers' instructions (see section 3.2.1 p. 44; primers see Table 4 p. 37).

2.4.4 EBV copy number assessment

gDNA was isolated from 5x 10^6 Namalwa, wt LCL and ΔE3A LCL cells and qPCR of a cellular (ß2m) and a viral (BALF5) housekeeping genes was performed (Figure 11B). Absolute quantification was conducted (see 2.5.3) and mean amount of DNA copies/14 ng of DNA for ß2m and BALF5 were obtained (n_{tech}=3). The mean amount of DNA copies over three biological replicates for BALF5 was divided by the amount of DNA copies of ß2m for every cell line and normalized to Namalwa cells, since they are known to carry two EBV copies (Whitaker, 1985).

2.5 RNA related techniques

2.5.1 RNA extraction

For total RNA extraction with the RNeasy Mini Kit (Qiagen), 4 x 10^5 to 1 x 10^7 cells were lysed in RLT buffer (Qiagen) supplemented with β-Mercaptoethanol (Sigma) according to the manufacturer's protocol. The lysed cells were loaded onto a QIAShredder and spun down for 2 min with full speed at RT. The protocol was followed including the optional DNase digest for 30 min instead of 15 min. Nuclear and cytoplasmic fractions were obtained as published (Weil, Boutain, Audibert, & Dautry, 2000) with some modifications. Two-fold approach with each 10^8 cells were harvested and washed twice with ice-cold PBS followed by centrifugation for 5 min at 500x g. 10^8 cells were resuspended in 2 ml Lysis Buffer (0.5 % NP40, 1.5 mM NaCl, 10 mM Tris HCl pH 7.8, 1.5 mM MgCl2, 10 mM EDTA, pH 8, RNase inhibitor 100 U/ml from Promega) and incubated for 10 min on ice. After centrifugation at 500x g for 5 min at 4 °C, the supernatant containing the cytoplasmic fraction was spun down for 1 min with full speed (= 25,000x g) at 4 °C. 200 µl (10^7 cell equivalents) of the supernatant was submitted to RNA isolation (with 700 µl RLT-Buffer). The nuclear pellet was washed once with Lysis Buffer and subsequently submitted to RNA isolation using 2.4 ml RLT buffer. The nuclear fraction was soaked 20x through a 20-gauge needle to destroy precipitates. The lysed pellet of 10^8 cells was split and loaded on three QIA-Shredder. The protocol of the RNeasy Mini Kit was followed including the optional DNase digest with the double amount of DNase (55 KU) for 1 h. After elution, the nucleic RNA of 2 x 10^8 cells was pooled. RNA quantity and quality was monitored by Qubit 2.0 Fluorometer and BioAnalyzer (Agilent), RIN >7 (RNA integrity number; Figure S5 left). Successful fractionation was verified by RT-qPCR of crucial candidate genes for the cytoplasm (GAPDH) and the nucleus (NEAT1; based on De Santa et al., 2010; Figure S6).

2.5.2 Reverse transcription (RT) of RNA into cDNA

RNA was reverse transcribed to obtain cDNA as target for PCR analyses (RT-qPCR). SuperScript IV First-Strand Synthesis Reaction (Thermo Fischer Scientific) was applied according to the manufacturers' instructions using 1 to 4 µg RNA as input. As a standard procedure for each analyzed sample, two reactions were set up, one containing reverse transcriptase and the second without enzyme, serving as negative control for genomic DNA contamination. For qPCR analysis, undiluted or up to 1:8 diluted cDNA (according to 50 ng input RNA per well) was used as template for quantification.

2.5.3 Quantitative polymerase chain reaction (RT-qPCR)

RT-qPCR was performed using LightCycler 480 SYBR Green I Master (Roche; 10 µl reaction volume: 5 µl LightCycler 480 SYBR I Green, 1 µl 5 µM Forward Primer, 1 µl 5 µM Reverse Primer, 1 µl H_2O and 2µl sample) on a LightCycler 480 II instrument (Roche) according to the manufacturer's protocol. Cycling conditions were 10 min at 95 °C and 45 cycles of 3 sec at 95 °C, 10 sec at 63 °C and 20 sec at 72 °C. All Primers were established to fit an annealing temperature of 63 °C (primers see Table 2 p. 34/ Table 3 p. 37). PCR products were analyzed by melting curve analysis and tested for correct size by gel electrophoresis. If possible, primer were designed to generate an amplicon of similar size (200 to 400 bp) to not bias the SYBR green signal intensities of different products. Standard curve generation was performed as described before with minor changes (Harth-Hertle et al., 2013). Absolute quantification of transcripts was performed using cDNA equivalent to 50 ng RNA prepared from 1 to 4 µg RNA as a template and was based on the standard samples of known concentration and the respective PCR efficiency for each primer pair. Analysis was done by the LightCyler® 480 software (SW1.5). Relative quantification was calculated by the software, taking a calibration of each plate into account. As housekeeping genes for the E3A system served GAPDH. Since E2 is inducing cell growth, most of the housekeeping genes are regulated by E2, like GAPDH or gusB (data not shown). Eisenberg et al suggested a new set of housekeeping genes which show constant expression levels across 16 human tissue types (including white blood cells; Eisenberg et al 2013). The charged multivesicular body protein, CHMP2A, was one of the suggested housekeeping genes and showed no substantial regulation by E2 in our RNA-Seq data. Thus, CHMP2A was used as housekeeping gene for E2. U6snRNA served as a non-coding housekeeping gene. It has to be mentioned that U6snRNA is a transcript produced by PolIII, whereas the other E2 regulated targets are presumably PolII produced targets.

2.5.4 Endpoint PCR

To apply a temperature gradient, a diagnostic PCR analysis was conducted using, apart from the temperature, the same cycling conditions as for the LightCycler on the PeqLab Cycler. To stick with the same reaction conditions, the LightCycler 480 SYBR I Green (Roche) was used here as well in a 20 µl reaction volume (10 µl SYBR I Green, 1 µl 5 µM Forward Primer, 1 µl 5 µM Reverse Primer, 3 µl H_2O and 5 µl sample; for primers, see Table 2 p. 34, ◊ only for endpoint PCR). Subsequently, the products were loaded on a 2 % agarose gel containing 0.01 % (v/v) EtBr. Subsequently, fragments were visualized under UV light.

2.5.5 Transcriptome analysis by RNA-Seq

For the complete E3A cell system libraries and the nuclear libraries of the E2 cell system, 30 µg of isolated RNA were applied to a DNase digestion in solution (1 U/µl, 1 U/µg RNA; Promega) for 1 h at 37 °C followed by 10 min at 65 °C for termination and RNA quality was examined using the Agilent BioAnalyzer (Figure S5, left; RNA Nano Kit), which identifies the major abundant RNA fractions contained in the samples (peaks of 18S/28S rRNA). The rRNA was depleted by subjecting 5 µg fractionated, DNase-digested RNA to the RiboZero Magnetic Gold Kit (Epicentre). The RNA quality was examined again using the BioAnalyzer (Figure S5, right; RNA Pico Kit). After rRNA depletion, the major abundant RNA fractions, 18S/28S rRNA, are not detectable any more. With samples of the E3A cell system and the nuclear fractions of the E2 cell system were proceeded as follows: 50 ng of the rRNA depleted RNA were applied as input to the library preparation using the ScriptSec v2 RNA-Seq Library Preparation Kit from Epicentre. The obtained libraries were di-tagged, barcoded libraries. To purify these libraries, the AMPure Bead-System was applied, retaining fragments of 200 to 500 bps. The libraries were sent to external Sequencing Facility LAFUGA (Blum Lab-Genomics, Gene Center Munich), which verified the quality of the libraries for us. The libraries were sequenced in a 100 bp-paired-end, multiplexed fashion on the HiSeq1500 Sequencer from Illumina (6 libraries per lane were loaded on the flow cell). The libraries (of three biological replicates) from E3A experiments was sequenced three times (three technical replicates), the libraries (of three biological replicates) from E2 experiments was sequenced two times (two technical replicates).

For the cytoplasmic libraries of the E2 cell system, an in house service of the Munich Sequencing Alliance (Institute of Human Genetics, Helmholtz Zentrum Muenchen) was used. 20 µl of the DNase digested RNA (~500 µg/ml) were sent to the in house sequencing facility, which verified the RNA quality, depleted for rRNA using the RiboZero Magnetic Gold Kit (Epicentre) and prepared the libraries for us, using the TrueSeq Stranded Total RNA Sample Prep Kit (Illumina). The libraries were sequenced in a 100 bp-paired-end, multiplexed fashion on the HiSeq2500.

2.5.6 Bioinformatic methods

Multiplexed fastqsanger files of self-prepared libraries were provided on the GALAXY platform of LAFUGA (https://blum-galaxy.genzentrum.lmu.de). Files were demultiplexed using the Illumina demultiplex tool (Galaxy Version 1.0.0). Cytoplasmic E2 libraries were provided already demultiplexed.

2.5.6.1 Analysis of the human transcriptome

The analysis of the human transcriptome was performed in collaboration with Gergely Csaba from the group of Ralf Zimmer, Teaching and Research Unit Bioinformatics, LMU. The analysis of RNA-Seq requires a concatenation of different tasks, basically mapping, counting and DE- testing (Figure 6).

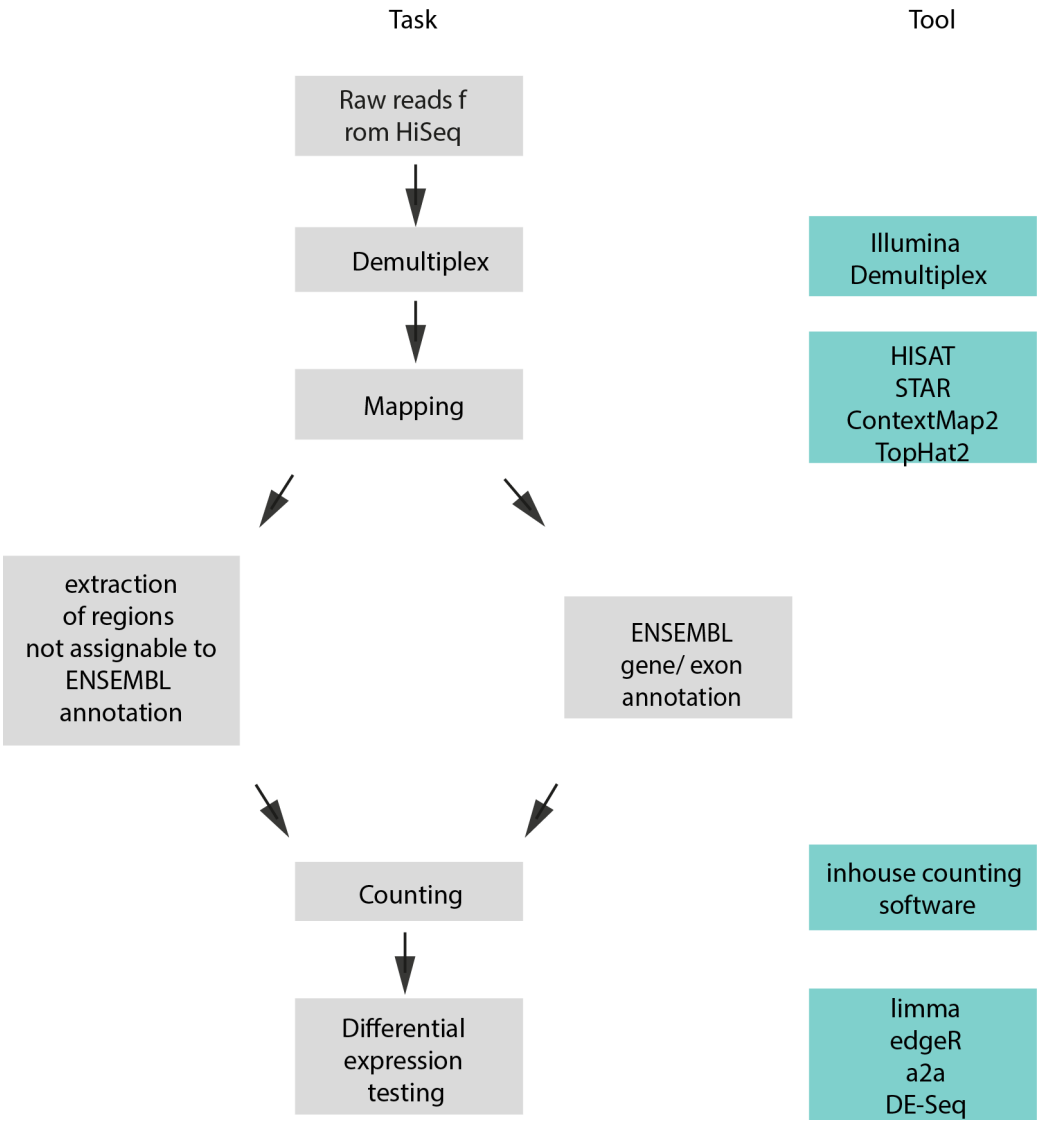


Figure 6: Flow Chart displaying the process of RNA-Seq analysis. Three major tasks have to be completed for identification of regulated genes: mapping, counting and differential expression testing. If libraries are prepared in a multiplexed manner to load more than one sample on one lane, the read files obtained from the sequencer, have to be demulitplexed prior to mapping. For mapping, raw reads are aligned based on their sequence to the sample corresponding genome using the mapper HISAT, STAR, ContextMap2 or TopHat2. For Counting, number of reads aligning to a defined (genomic position by chromosome number, start- and end) feature like gene, transcript, exon or other region are counted based the chosen annotation file (were the features are annotated to a genomic position). After counting the aligned reads in different conditions, the numbers of reads can be compared for a condition pair by using the tools limma, edgeR, a2a and DE-Seq, which perform differential expression testing.

Mapping of RNA-Seq reads

RNA-Seq reads mapping to the human genome (genome built hg19 = GrCh37) was conducted. Technical replicates of the self-prepared libraries were merged. As a consequence, in average 8x 10^7 reads were obtained for every sample. For mapping, four different mappers, HISAT (D. Kim, Langmead, & Salzberg, 2015), ContextMap2 (Bonfert, Kirner, Csaba, Zimmer, & Friedel, 2015), STAR (Dobin et al., 2013) and TopHat2 (D. Kim et al., 2013) were used.
Covered regions consistently found by all mapper (read pair had to reside completely within a gene) were extracted and assigned to a gene or transcript. For gene/transcript/exon annotation, the GRCh37 assembly of ENSEMBL (release 75) was used (= Gencode v19). Regions not identical to any ENSEMBL gene annotation (in a strand specific manner) were extracted and assigned with a new ID for intergenic genes. Regions not overlapping any exons within a gene were extracted and assigned with a new ID for intronic genes.
A general remark has to be made here, library preparation kits can be responsible for reads mapping to the opposite strand of a transcribed locus. The two kits used in this thesis perform different regarding this, resulting in different absolute number of reads mapping to opposite strand regions for the cytoplasmic fraction. These opposite strand reads have to be considered in later analysis of antisense genes.

Counting of aligned reads

For counting, an in house count software (can be obtained from AG Zimmer) was applied to calculate number of aligned reads mapping to a gene. Either read pairs aligning completely to a gene were counted (read origin = gene) or read pairs aligning to a gene were only counted, if the fragment structure is fully consistent with a known transcript corresponding to that gene (read origin = gene with transcript support). In the latter case, the amount of reads was the sum of reads aligning to the gene and reads aligning to the transcripts. Objects for counting were read pairs (= fragment). If multiple read pairs were generated by the same RNA- fragment, two ways of down-sampling were applied to circumvent PCR amplification biases (one fragment could be amplified better than others). Either, the counts were downsampled on fragment level: Under the assumption that multiple detected identical fragments are PCR artefacts, each fragment is just counted once. Or, counts were downscaled in a way that the impact of highly amplified fragments was decreased. Therefore, the ratio of counts per fragment for all samples divided by the maximal detected counts per fragment multiplied with the log2FC of the max. detected counts per fragment was taken into account for each fragment. Counts per gene were the sum of all downscaled fragment counts.

Differential expression (DE) testing

For DE-testing four different DE-Methods were used, DE-Seq (Anders & Huber, 2010), limma (Ritchie et al., 2015), edgeR (Robinson, McCarthy, & Smyth, 2010) or an in house method, a2a. All methods test differential expression based on different models with different statistical significance testing. If a significant change for an intergenic/intronic region is observed, the change might arise as a byproduct of the containing or overlapping gene. In order to find changes directly associated to the intergenic/intronic regions it was confirmed if an observed change significantly differs from the changes of all overlapping objects. The probability of changes is measured as unexpectedness of observed changes with respect to the empirical noise of the changes derived from the data. This also allows to compare two changes (double differential view) as they imply also a change (the change of the changes). This is then applied to the intergenic and intronic differential expression.

2.5.6.2 Analysis of the viral transcriptome

In a second collaboration with the group of Erik Flemington, Cancer Crusaders Bioinformatics Core in New Orleans was started to analyze the RNA-Seq reads mapping to the EBV genome (mapping to Akata genome) was accomplished. This laboratory published the, until now, most complete annotation for the Akata genome by combining different sequencing methods (Lin et al., 2013). The Akata EBV strain is non-defective and represents a wild-type virus.

Workflow and Analysis

Reads were mapped using the STAR mapper. Since EBV genome is of dense organization with multiple genes overlapping, reads are mapping to multiple genes or isoforms and transcript quantification is difficult. Therefore, for counting and differential expression testing, RSEM (B. Li & Dewey, 2011) was applied. This package has the advantage to effectively deal with ambiguously-mapping reads.
Analysis was performed by the Flemington group. Results were provided as normalized read counts. Reads obtained from ChIP-Seq (Glaser, PhD thesis, 2017) were aligned to the Akata genome and peaks were called using MACS2 (Y. Zhang et al., 2008).

2.5.6.3 Visualization of RNA-Seq results

For visualization, coverage tracks were generated for sense and antisense strand separately from mapped reads (bam files; mapper=STAR) and loaded into the Integrative Genomics Viewer (IGV) from the Broad Institute (https://software.broadinstitute.org/software/igv/download).

2.5.6.4 Self-performed computational work

For self-performed computational work and visualization, the GALAXY platform of the Backofen Lab, Chair for Bioinformatics, Institute for Informatics, University of Freiburg was used (https://galaxy.uni-freiburg.de). Heatmaps were generated using the tool "multiBamSummary", which calculated average read coverages for BAM files (mapped reads). For this, the genome was split into bins of 1000 bp. For each bin, the number of reads found in each BAM file was counted. For coverage calculation, bins with zero or large counts were also included, distance between bins = 0 bp. The tool "plotCorrelation" created a heatmap of correlation scores between indicated samples obtained by Spearman correlation.

2.6 Gene ontology analysis

For gene enrichment testing, classical ORA (overrepresentation analysis) was conducted by Gergely Csaba, using a hypergeometric test.
We asked for GO-terms of biological processes, for which an enrichment could be calculated for the subset of counter-regulated genes (741; strength and direction of regulation was not included). The background was defined as all genes, which were detected in any sample/any condition/any compartment with ≥ 1 read count. Hits were not filtered for dependency. Significance was limited to $p < 0.05$. Selected GO terms are shown (target genes per term)/ (total genes per term).

2.7 Primers

2.7.1 Human chromatin primers

Table 1: Primer pairs for qPCR on human chromatin

Internal Designation	Target	Sequence	Ampl. Length in bp	An-notation	Position hg19
SAR105.1fwd SAR105.1rev	RHOH TSS-6.749bp	AGGGAAAGTTAAGACAGGCCTT TAGAGGATTCCAACCCAATGCC	110		chr4:40187701 +40187810
SAR105.2fwd SAR105.2rev	DNAJB12 TSS-56.9kb	CTTGGCACGAACTCTTCCTTC AGATGTGAGAATGATGTGGCCG	117		chr10:74057129 +74057245
SAR105.4fwd SAR105.4rev	SWAP70 TSS-62.2kb	GCTGGGTTTGCGGTTTAATG TTGCTGCTCCCTAAGGTTTG	100		chr11:9623168 +9623267
SAR92.5fwd SAR92.5rev	POU2F2	TGGCCTCAAGGGAGTGAAC GGCCACCTCTCTTGTTGTG	83		chr19:42633923 +42634005
SAR92.6fwd SAR92.6rev	CLEC16A	AGTGCACATTGGCAGAACTC CTCTGAGCTTGGGTCCTCAC	80		chr19:42633923 +42634005
SAR92.9fwd SAR92.9rev	CSRNP1	TAGGCAAGCGTGAGAGAAGC CCCGGAAACATCTGTGAGTC	88		chr3:39189615 +39189702
MH2675fw MH2675rv	ADAMDEC1 TSS -16,65 bp	CTTCATGGCTACAGACTCTTGG CCTATGTCTCGCTTCCTGCT	93	+ CNTRL locus, E2 binding site	chr8:24224967-24225059
ST122_F ST122_R	RPL30 TSS -1,94 bp	CTGGTCTGACGCTCCTGACT CAGTGCCCAGAATTCCAGAT	120	- CNTRL, no E2 binding site	chr8:99055763-99055882

2.7.2 Human cDNA primers

Table 2: Primer pairs for RT-qPCR on cellular transcripts; *Transcript information based on ENSEMBL GrCH75; ◊ only for endpoint PCR

Internal Designation	Target	Sequence	Ampl. Length in bp	Annotation	Position hg19
SAR40.3fwd SAR40.3rv	MYC*	CTCTCAACGACAGCAGCTC CCACAGAAACAACATCGATTTC	210	exon-exon-junct. on 2nd and 3rd exon; targets 3/5 transcripts	chr8:128751101 +128752686
SAR149.3fwd SAR149.3rev	PCAT1 tv1*	AGAGGCAGAGGATGTTGACAC AACGTGTGCATCCCAAGAGG	253	exon-exon-junct. on the 2nd and the 4th exon of tv1	chr8:127889918 +128018828
SAR127.6fwd SAR127.6rev	CASC19*	TCCTTGCCAGTGTCTTCTCC CCCGTAGATTGCAAACTCCTAG	210	exon-exon-junct. on the 1st and	chr8:128200088 +128209816

				the 3rd exon	
SAR127.10fw SAR127.10rev	CASC21* (RP11-382A18.2)	TGAAACCTCTGCTTCCTGGC TGCTGGACGTCTTCTCTTGG	220	exon-exon-junct. on the 2nd and the 3rd exon	chr8:128334369 +128351642
SAR149.5fwd SAR149.5rev	LINC00977* ◊	TGCTGGCCTCTTTCTCATGG AAGCCAGAAATCCAGGACCC	252	exon-exon junct. on 3rd and 2nd exon	chr8:130229460 +130235525
SAR179.3fwd SAR179.3rev	LINC00977*	TGGTGGTGGTTGTATCAGGC GGGCTTGCTTTCATTCTGTGG	321	monoexonic primer on 3rd exon, gDNA same size	chr8:130229187 +130229507
SAR127.11fw SAR127.11rev	CCDC26*	GCAATGCCTCCTCCTCCTTC GGCCTGAGGAGAGAAGACAC	280	exon-exon-junct. on the 3rd and 4th exon; targets 3 /4 transcripts	chr8:130365015 +130382620
SAR127.3fwd SAR127.3rev	CD84*	GTCAGCTCTTGGTCCTCAGG AGACTCAGAAACAGCACCCG	372	exon-exon-junct. on 2nd and 3rd exon; targets 4/5 transcripts	chr1:160523756 +160535384
SAR156.1fwd SAR156.1rev	RP11-528G1.2*◊	GCAGTTCCAGCAAGCATGTC TGACCCTCCTCTCCTTCACC	371	exon-exon junct primer on 1st and 2nd exon	chr1:160506900 +160541021
SAR179.2fwd SAR179.2rev	RP11-528G1.2*	CAGAAATGGCGTGGGTGAAG ACATCAGTGGCAATCTCCTCC	212	mono-exonic primer on 2nd exon, gDNA same size	chr1:160540989 +160541200
SAR127.1fwd SAR127.1rev	SLAMF1*	GTATCAAGGTGCAGGTCCCG GGCAGTTGGGAAGCAAAGTG	375	exon-exon-junct. on the 2nd and the 3rd exon; targets all 5 transcripts	chr1:160604619 +160607286
SAR127.4fwd SAR127.4rev	CD48*	GTGCCATTCTTGCTGCTCAC ACTTGATCCTCAGAGTGGCG	372	exon-exon-junct. on the 2nd and the 3rd exon; targets all 4 transcripts	chr1:160651021 +160654804
SAR156.7fwd	PPAN-P2RY11*	TGCGTGATGTGGTCTCCTC	204	exon-exon-junct. on the 5th and the 6th	chr19:10221676 +10224625

SAR156.7rev		TGTAGTCGATGAGGAGGCAG		exon; targets 4 PPAN transcripts + 2 read-through transcripts	
SAR156.8fwd SAR156.8rev	CTD-2240E14.4*	GTGGGTCAAGGGTCAGGTTC GGACGACCTTATCCCTGCTG	326	mono-exonic transcript; gDNA same size	chr19:10200636 -10200961
SAR179.4fwd SAR179.4rev	ANGPTL6*	CAGCCCAGTAGACACCATCC GTGGAGGCTGGACTGTGATC	200	exon-exon-junct. on the 4th and the 6th exon ; targets both transcripts	chr19:1020330 +10204442
SAR127.13fw SAR127.13rv	CHMP2A*	CCCAGCTCATCCAGAACCTG AAGAAGATGGCCAAGCAAGG	374	exon-exon-junct. on the 2nd and the 5th exon	chr19:5906329 +59065438
JAH31.1Bfwd JAH31.1Arev	CASC8 tv3*	TTGAGTGAGGCTGGAACTGG AGGATCGACGTTCAAGGTGG	201	exon-exon junct. Fwd-primer on exon specific for tv3, rev-Primer on exon 6	chr8:12830216 +128335320
BS688for LG90rev	GAPDH	GAAGGTGAAGGTCGGAGTC TGGGTGGAATCATATTGGAAC		exon-exon junction primer	chr12:6644004 +6645878
MH2333fw MH2333rev	gusB	CGCCCTGCCTATCTGTATTC TCCCCACAGGGAGTGTGTAG	91	exon-exon junction primer	chr7:65439995 -65441028
MH2682fw MH2682rev	U6snRNA	CTCGCTTCGGCAGCACATATAC GGAACGCTTCACGAATTTGCGTG	96	according to Didier Auboeuf	Picks up 24 loci; It is common in vertebrate genomes to find many copies of the U6 snRNA gene or U6-derived pseudogenes
MH2594fw MH2594rev	Neat1	GGGCCATCAGCTTTGAATAA CTTGAAGCAAGGTTCCAAGC	149	mono-exonic transcript; gDNA same size	chr11:6519171 +65191859
MH350 fw MH350rev	ADAMDEC1	GAAGAGCACTGACGGGAAAC ACCAAGGCCACTTGAACATC	231	Exon-exon junction primer	chr8:24254909 +24256421
MH199 fw MH199 rev	BMP4	TCCACAGCACTGGTCTTGAG CGTGTCACATTGTGGTGGAC	337	Exon-exon junction primer	chr14:5441735 -54418645

MH2145fw MH2145rev	CXCL10	TGACTCTAAGTGGCATTCAAGG CCTTTCCTTGCTAACTGCTTTC	239	Exon-exon junction primer	chr4:76943518 -76944544
GS 33.1 for GS 33.1 rev	DNase1L3	AGGACACCACGGTGAAGAAG GTGAAGGCCCTTGAAGACTG	202	Exon-exon junction primer	chr3:58178469 -58183590

2.7.3 Viral cDNA primers

Table 3: Primer pairs for RT-qPCR on viral transcripts

Internal Designation	Target	Sequence	Amplicon Length in bp	Annotation
SAR228.6fwd SAR228.6rev	BGRF1/ BDRF1*	TCAACAAGGAGACCAGCACC ACTTACCACGTTCAGCAGCC	290	exon-exon-junct. on 1st and 2nd exon
SAR228.3 fwd SAR228.3 rev	BHRF1*	TGGGCATGTGTTGGAATTGC CAGTGTCCTCTGGCGAAAGG	224	monoexonic isoform; viral DNA same size! Fwd- primer in intron of late splice variant
SAR228.5fwd SAR228.5rev	BNRF1*	TACAGGACCATCAACGCCAC GCCTGTGCCCATGAACTTTC	206	monoexonic transcript; viral DNA same size!
SAR228.10fwd SAR228.10rev	LMP2A*	GGGTCCCTAGAAATGGTGCC GTGGGTCCTCAATCCTCCATG	391	exon-exon-junct. on 1st and 3rd exon
MH280fw MH280rev	LMP1	TCCTCCTCTTGGCGCTACTG TCATCACTGTGTCGTTGTCC	490	from Yoshioka et al., 2001; according to Imai et al., 1996
LG521 fw LG521 rv	E2	TCTGCTATGCGAATGCTTTG CACCGTTAGTGTTGCAGGTG	255	within one exon

* Primers were designed using the Akata genome annotation provided by the Flemington Lab and the P3HR1 sequence.

2.7.4 gDNA primers

Table 4: Primer pairs for qPCR on genomic DNA

Internal Designation	Target	Sequence	Amplicon Length in bp	Position hg19
LG171for LG171rev	BALF5	CATGCTCTACGCCTTCTTCC ATGCACATCCTCCTTCTTGG	343	-
LG172.Bfor LG172.Brev	ß2m	ATTGGGATTGTCAGGGAATG GGATGCTAGGACAGCAGGAC	266	chr15:45008038+45008303

3 Results

The following results section is divided into two parts. The first part will focus on the accession of E2 to DNA while in the second part, E2 mediated gene regulation is investigated.

3.1 Accession of E2 to DNA: E2 requires EBF1 to bind to its CBF1 independent binding sites in the human genome

Parts of this work are published in Glaser LV*, Rieger S*, Thumann S, Beer S, Kuklik-Roos C, Martin DE, et al. (2017) EBF1 binds to EBNA2 and promotes the assembly of EBNA2 chromatin complexes in B cells. PLoS Pathog 13(10): e1006664. https://doi.org/10.1371/journal.ppat.1006664

For this study, a CBF1 deficient human B cell line was used. This cell line was generated by homologous recombination in the somatic B cell line DG75 (EBV negative Burkitt's lymphoma cell line), to screen for CBF1 independent functions of E2. The DG75$^{doxHA\text{-}E2}$/CBF1 wt and the DG75$^{doxHA\text{-}E2}$/CBF1 ko cell lines carry the Dox inducible HA-E2 expression plasmid, thus E2 expression can be induced by doxycycline treatment (Figure S2).

3.1.1 Peak selection and characterization

E2 ChIP-Seq experiments conducted in DG75$^{doxHA\text{-}E2}$ CBF1 wt or CBF1 ko B cells in the Kempkes laboratory revealed that E2 also binds to chromatin independently of CBF1, therefore the question about an alternative DNA anchor that might be used by E2 to bind chromatin arose. A genome wide correlation analysis of TF binding patterns uncovered that E2 binding sites highly correlate with EBF1 binding sites (Glaser, PhD thesis, 2017). Furthermore, EBF1 appeared to be the only significant enriched TF signal at CBF1 independent E2 binding sites.
Thus, we investigated the impact of EBF1 on the binding pattern of E2 by siRNA-mediated knock down of EBF1 and subsequent ChIP-qPCR of E2 and EBF1. CBF1 dependent and independent E2 binding sites with a moderate signal (50 % < mean signal > 90 % percentile), which are also present in LCLs and positive for EBF1 binding in LCLs were selected for qPCR (Figure 7A). To characterize CBF1 dependent and independent E2 peaks which meet the described requirements, an intersection with predefined clusters of EBNA peaks by correlation was conducted. Cluster

analysis for EBNA peaks identified eight distinct clusters of combinations for E2, CBF1, EBF1 and CUX1 (Figure S7). These four TFs were chosen based on a preceding genome-wide TF binding pattern correlation analysis using 89 TFs. Cluster I for instance, positive for all four investigated TFs (Figure S7A/B), shows the strongest enrichment for H3K4me1 as well as for H3K27ac (Figure S7C), indicating an association of these binding sites with active enhancers. The majority of E2 peaks from this cluster reside at strong enhancers (66.7 %) according to CSS by ENCODE in GM12878 (Figure S7D). In cluster III to VI, CBF1 is missing, in cluster V to VIII EBF1 is missing and in cluster IV, V and VIII, CUX1 is missing. Cluster V is positive for E2 only. We hypothesized that CBF1 independent E2 binding sites differ from CBF1 dependent binding sites in their properties. Contrarily, CBF1 dependent and independent E2 binding sites only slightly differ in their cluster composition (Figure 7B). CBF1 independent compared to dependent peaks are enriched for cluster I (from 60% to 70%) and cluster V-VIII are depleted (from ~20% to 8%). Cluster V-VIII are clusters without EBF1, while cluster I is EBF1 positive and is accompanied by marks for open chromatin.

The comparison of CBF1 dependent with CBF1 independent E2 peaks in LCLs regarding their cluster distribution did not indicate substantial differences. Thus, these binding sites may be similar regarding their TF binding pattern and their accompanied chromatin states. However these data were obtained in LCLs and the chromatin landscape could differ in DG75.

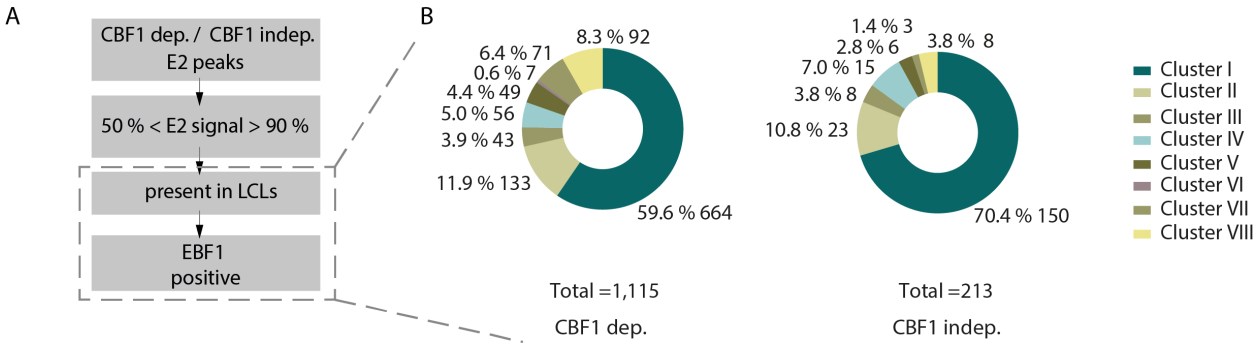


Figure 7: Cluster Correlation of E2 peaks selected for EBF1 KO analysis show an enrichment for Cluster I and a depletion of Cluster VIII for CBF1 independent E2 peaks. A Scheme of E2 peak selection for ChIP-qPCR (from bottom to top): All CBF1 dependent and independent E2 peaks with a 50 % < mean signal > 90 % percentile, which were also present in LCLs and EBF1 positive in LCLs were selected for ChIP-qPCR upon EBF1 KD. **B** Donut plots of cluster correlation of selected E2 peaks: intersection of the selected E2 peaks with the defined clusters (cluster definition from 7; see Figure S7).

3.1.2 Confirmation of knock down and ChIP strategy

The aim was to establish a siRNA-mediated knock down strategy. Furthermore, the established ChIP protocol needed to be adapted to the EBF1 antibody.
The role of EBF1 for CBF1 independent EBNA2 binding events was investigated applying knock down experiments. DG75$^{doxHA\text{-}E2}$ CBF1 wt or CBF1 ko B cells were transfected with non-targeting siRNAs (siCNTRL) or EBF1 specific siRNAs (siEBF1). 8 h post transfection, E2 transcription was induced by 1 µM doxycycline. 24 h post transfection, cells were harvested and analyzed by immunoblots and ChIP-qPCR (Figure S4).
siRNA mediated knock down of EBF1 successfully depleted EBF1 to approximately 35 % of the expression in siCNTRL transfected cells (Figure 8A, B). In order to verify the established ChIP protocol and to later account for differences in the two cell lines (e.g. much higher E2 protein expression level in CBF1 ko cells compared to wt cells), a locus positive and a locus negative for E2 binding were investigated (Figure 8C). Both loci were confirmed to be either positive or negative for E2 binding by ChIP-Seq. E2 could be precipitated well from the positive locus in CBF1 wt cells and it still binds weakly to the positive locus in absence of CBF1. At the negative locus, E2 binding is detected on background level. The background level of E2 binding in siCNTRL transfected cells in CBF1 wt and ko cells at this locus was used for normalization for the following ChIP-qPCR results.

A siRNA-mediated knock down for EBF1 could successfully be established. The existing ChIP protocol could be adapted to the EBF1 antibody and E2 binding could be verified at an established positive locus.

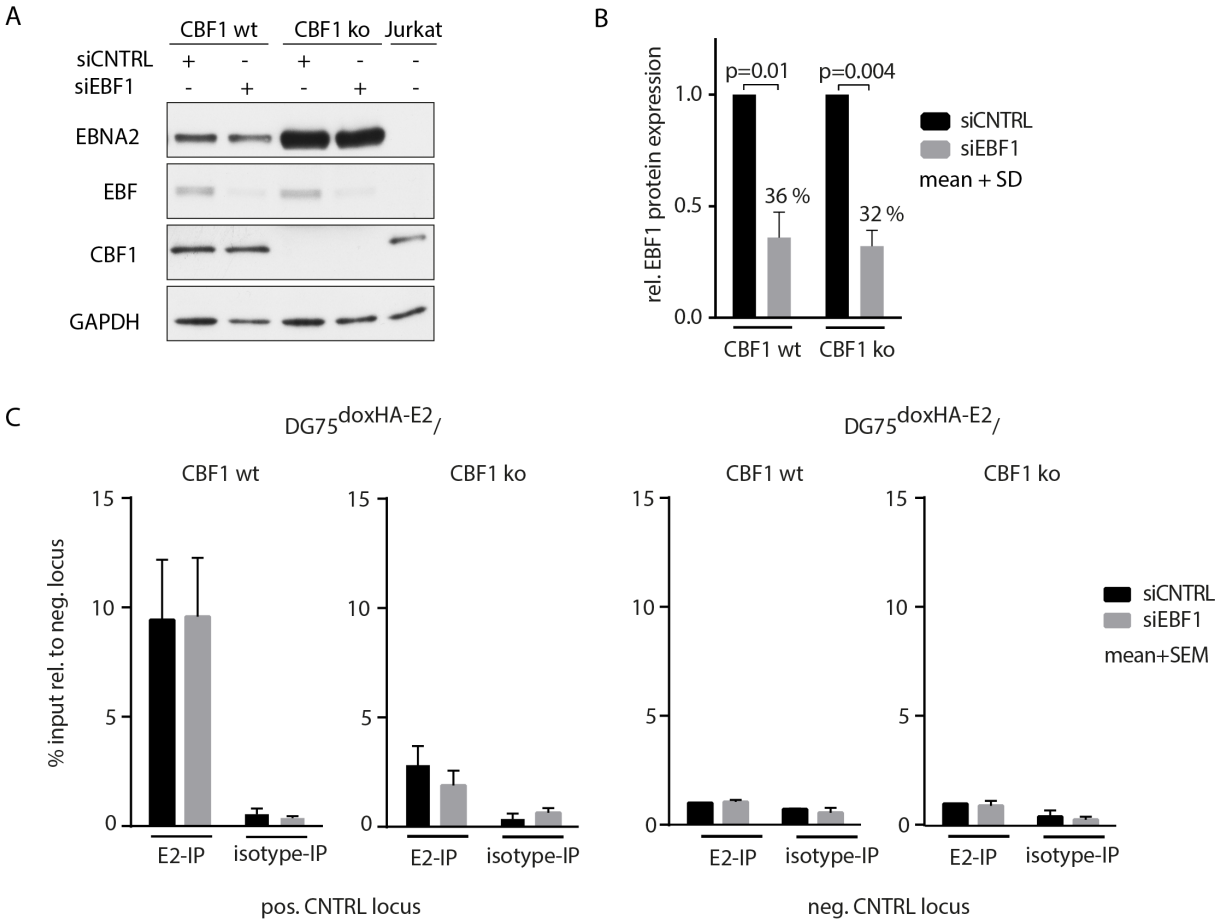


Figure 8: Confirmation of EBF1 knock down by immunoblotting and ChIP-qPCR of CNTRL loci confirming the established ChIP. **A,B** DG75$^{doxHA\text{-}E2}$ CBF1 wt or CBF1 ko B cells were transfected with a mixture of scrambled non-targeting siRNAs (siCNTRL) or EBF1 specific siRNAs (siEBF1). 8 h post transfection, E2 transcription was induced. 24 h post transfection, cells were harvested and analyzed by immunoblots and ChIP-qPCR. **A** Representative immunoblots showing expression levels of E2, EBF1, CBF1, and GAPDH before and after knock down (n=3). EBF1 negative Jurkat cell lysate served as a negative control. **B** Protein band intensities were quantified by densitometry. The mean change of EBF1 protein expression in siRNA (siEBF1) treated compared to control (CNTRL) treated cells is significant according to paired t-test as indicated. Standard deviations (SD) are indicated ($n_{biol.}$ =2). **C** ChIP-qPCR results for E2 binding to chromatin at a laboratory internal established E2 positive (ADAMDEC1 TSS-16,646 bp) and an E2 negative locus (RPL30 TSS-1,936 bp). % input was calculated relative to negative locus to adjust to divergent E2 inducibility in wildtype and knock out cells (ChIP-Seq; percent input at tested locus/percent input of known negative locus). Standard errors are indicated ($n_{tech.}$ = 2, $n_{biol.}$ = 3).

3.1.3 siRNA-mediated knock down of EBF1 impairs E2 binding at CBF1 independent sites

We sought to investigate the contribution of EBF1 to E2 binding to CBF1 dependent and independent binding sites.
E2 and EBF1 were immunoprecipitated from chromatin in DG75$^{doxHA\text{-}E2}$ CBF1 wt or CBF1 ko B cells and three CBF1 independent (Figure 9A) and CBF1 dependent (Figure 9B) E2 binding sites were analyzed by qPCR. The decrease in E2 binding in CBF1 ko cells compared to CBF1 wt cells can be observed at all investigated loci. EBF1 binding is decreased in a similar manner in CBF1 ko cells. The siRNA mediated knock down of EBF1 leads to a decrease in EBF1 binding at all probed loci, while E2 binding is decreased only at the CBF1 independent sites.

These data clearly shows EBF1 dependent E2 binding at CBF1 independent binding sites. Thus, EBF1 is used as an alternative anchor for E2 and this complex formation is important at CBF1 independent binding sites.

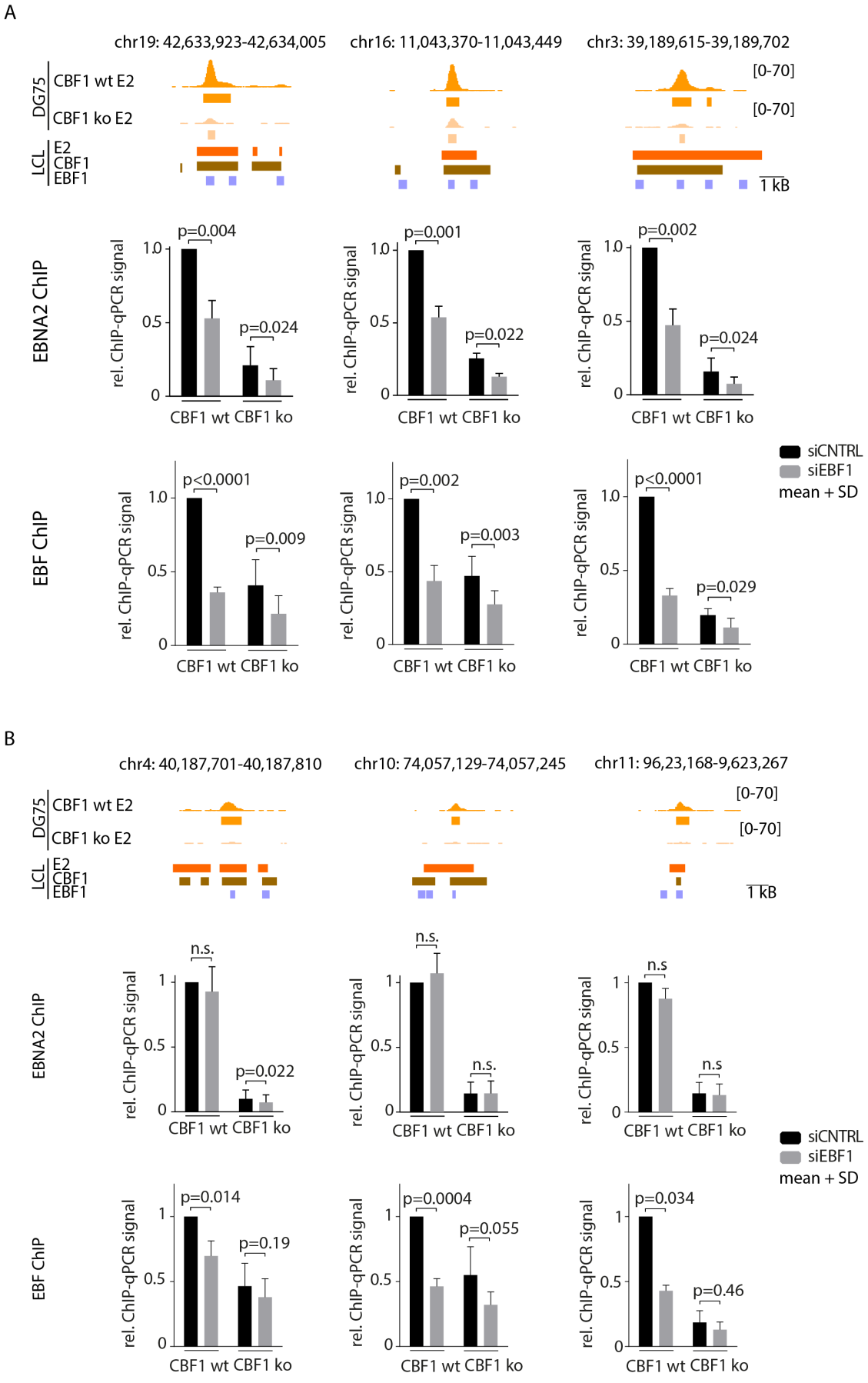

A
chr19: 42,633,923-42,634,005
chr16: 11,043,370-11,043,449
chr3: 39,189,615-39,189,702
DG75
CBF1 wt E2
CBF1 ko E2
LCL
E2
CBF1
EBF1
[0-70]
[0-70]
1 kB
EBNA2 ChIP
rel. ChIP-qPCR signal
p=0.004
p=0.024
p=0.001
p=0.022
p=0.002
p=0.024
CBF1 wt CBF1 ko
EBF ChIP
p<0.0001
p=0.009
p=0.002
p=0.003
p<0.0001
p=0.029
siCNTRL
siEBF1
mean + SD
B
chr4: 40,187,701-40,187,810
chr10: 74,057,129-74,057,245
chr11: 96,23,168-9,623,267
n.s.
p=0.022
n.s.
n.s.
n.s
n.s
p=0.014
p=0.19
p=0.0004
p=0.055
p=0.034
p=0.46

Figure 9: E2 requires EBF1 to bind to its CBF1 independent binding sites. E2 binding signals and peak tracks as obtained in DG75$^{doxHA\text{-}E2}$ (DG75) and E2, CBF1 and EBF1 binding peaks tracks in LCLs are shown for **A** three CBF1 independent and **B** three CBF1 dependent E2 binding sites. ChIP-qPCR results for EBNA2 binding to chromatin before and after EBF1 knock down are shown below the chromatin profiles. % input was calculated relative to a known negative locus. Mean relative input of the wildtype cells treated with siCNTRL was set to one. Standard deviation and p-values, based on Student's paired t-test, are indicated ($n_{tech.}$ = 2, $n_{biol.}$ = 2).

3.2 Analyses of cellular and viral genes regulated by E2 and E3A

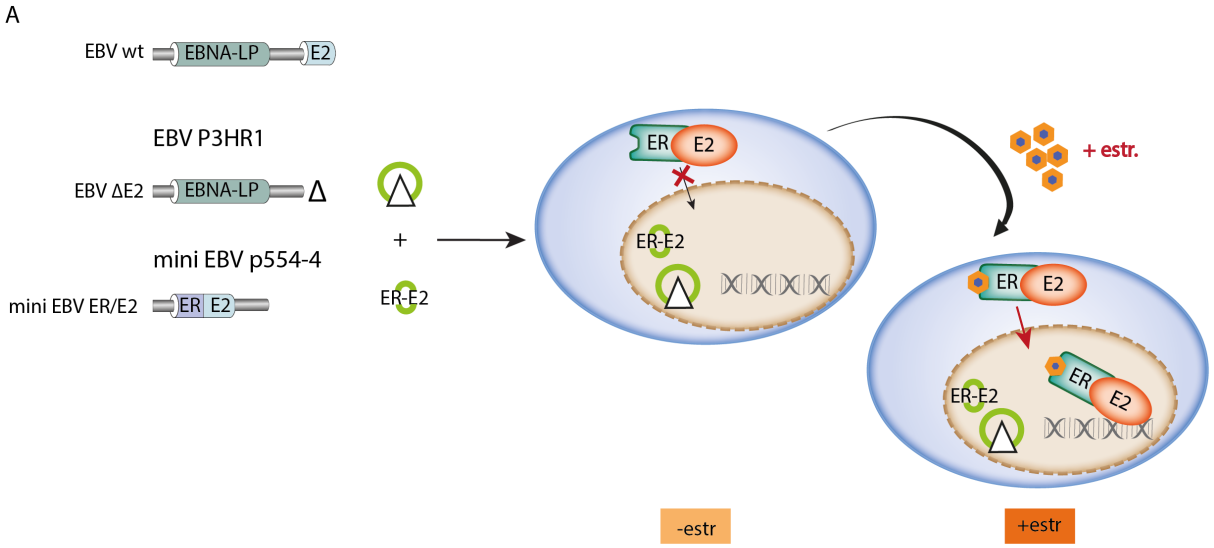
3.2.1 The cell systems: conditional ER/EB2-5 cells and wt versus ΔE3A LCLs

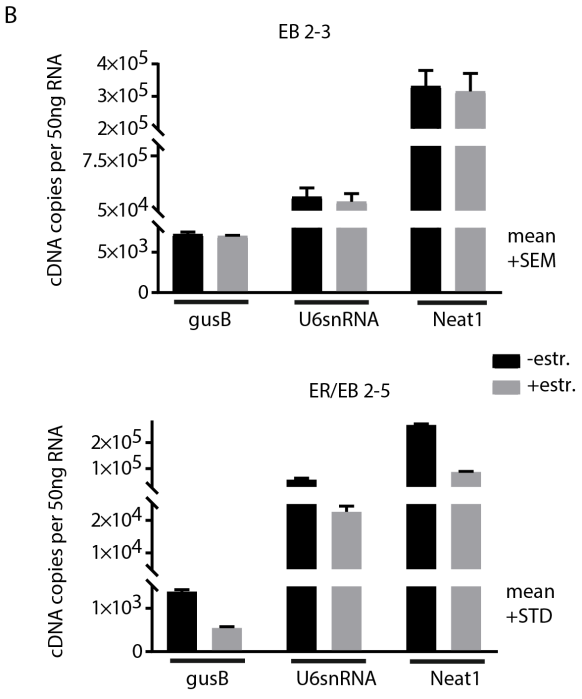
Previously, in the Kempkes laboratory it could be shown that intergenic genomic regions with high chromatin accessibility displayed E3A dependent transcription. As lncRNAs play an important role in development and pathogenesis of multiple diseases including cancer (see section 1.3.4, p. 16), we examined the capability of EBV to regulate lncRNAs genome wide for the transcription factor E3A and the co-expressed master regulator of transcription E2. Therefore, two different cell systems were used to study regulation of transcription by E2 and E3A.

ER/EB2-5 cells express an E2 open reading frame fused to the estrogen receptor (ER) hormone binding domain (ER/E2). Estrogen in the cell culture medium of ER/EB2-5 induces nuclear localization of ER/E2 and hence E2 can act as transcriptional activator and ensures cell proliferation (Figure 10A; see section 2.1.3, p. 21). Estrogen depletion of the cell culture medium inactivates ER/E2 and causes cell cycle arrest. Re-addition of estrogen causes rapid re-induction of E2 target genes. Cycloheximide (ChX) prevents protein translation and is added to the cells 1 h prior to estrogen treatment in order to investigate E2 targets independent of translation (Figure S1). E2 targets regulated in the presence of ChX were once operationally defined as direct E2 target genes (Kaiser et al., 1999). Since LCLs do not express the estrogen receptor endogenously, estrogen treatment does not alter gene expression in general (Figure 10B) proved by RT-qPCR at three different genes in the CNTRL LCL EB2-3. In contrast, the ER/EB2-5 cell line shows changes of expression of the housekeeping genes upon 6 h of estrogen treatment most likely due to the impact of E2. We aimed to assess the time point of highest target RNA abundance. After depletion, re-activation of ER/EB2-5 cells with estrogen induced a peak in RNA abundance of candidate target genes, two cellular (*MYC, DNase1L3*) and a viral target (*LMP1*), at 6 h post reactivation of ER/E2 by estrogen (Figure 10C).

The ER/EB2-5 system was selected as an appropriate system to study E2 dependent gene regulation, especially in combination with the treatment of ChX. Side effects of estrogen treatment

could be excluded and E2 target genes could be induced with a peak abundance at 6 h post re-activation.





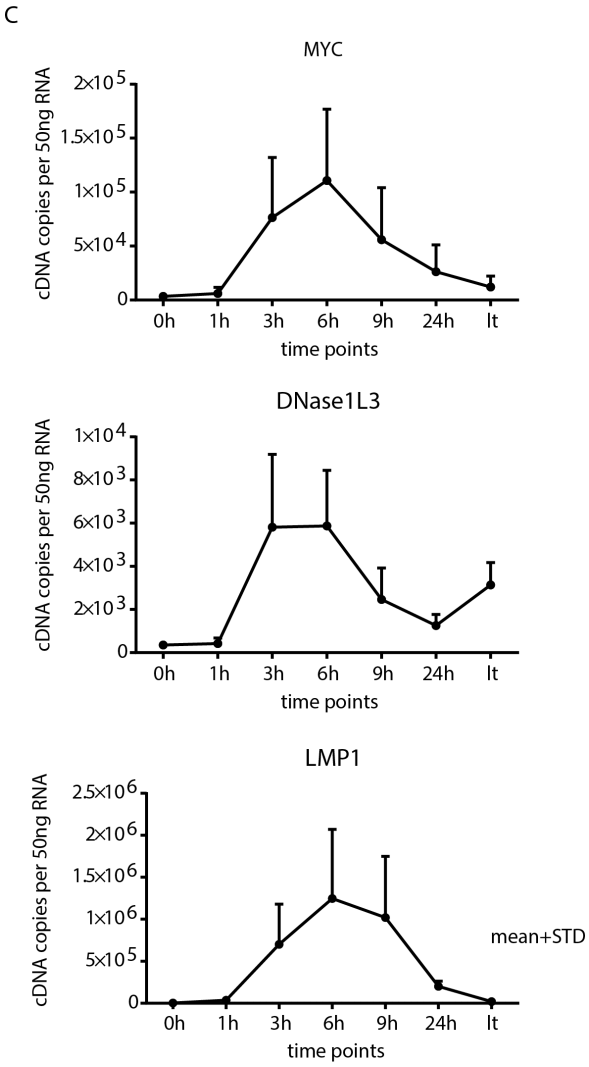
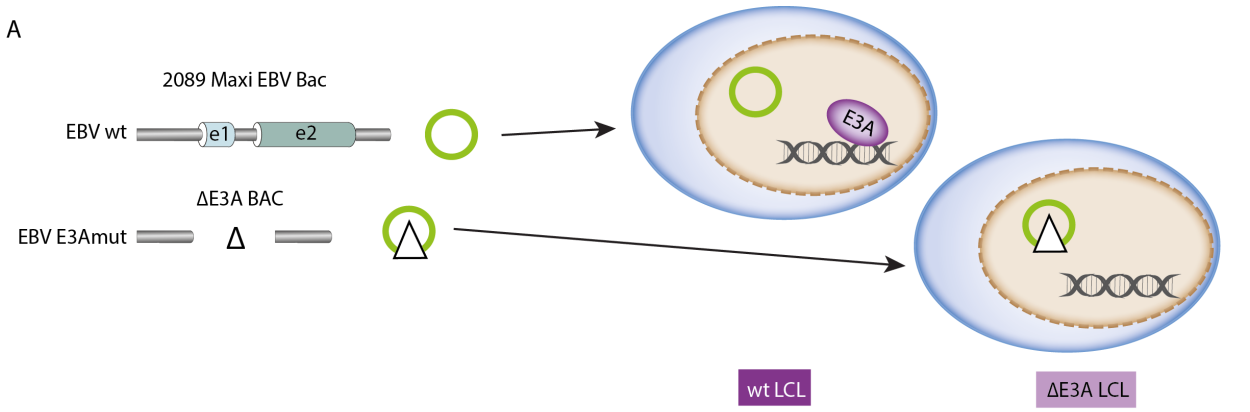
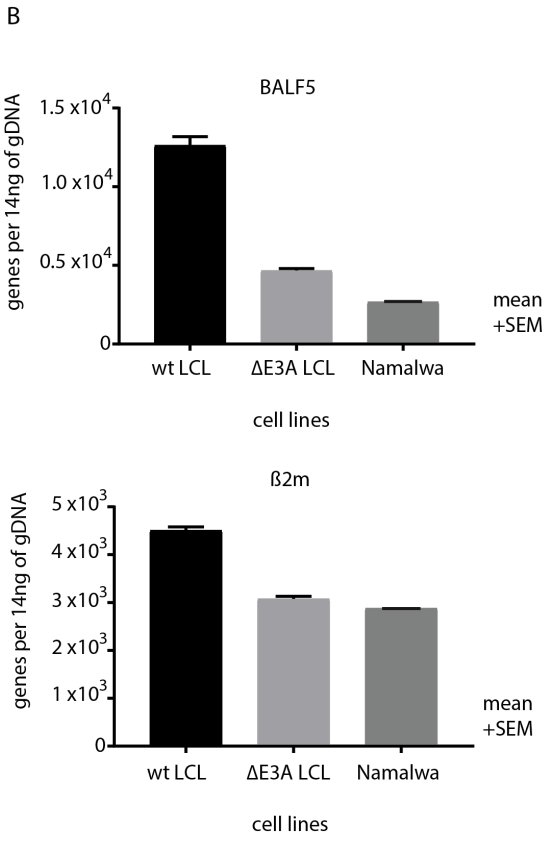


Figure 10: Concept of E2 cell system. A Schematic representation of the ER/EB2-5 system. ER/EB2-5 is an EBV immortalized B cell line infected with the mini-EBV plasmid 554-4 and the ΔE2 EBV strain P3HR1. ER/EB2-5 cells express ER/E2, where E2 is fused to the hormone-binding domain of the estrogen receptor (ER). In the absence of estrogen (estr.), E2 shuttles to the cytoplasm and is virtually inactive. In the presence of estrogen, E2 locates in the nucleus and is active as a transcription factor. **B** Estr. shows effects on ER/EB2-5 (bottom; $n_{tech.}$ = 3) but no effects on CNTRL cell line EB2-3 (top; $n_{tech.}$ = 3, $n_{biol.}$ = 3). Quantification of gene expression by RT-qPCR of the transcripts of two established housekeeping genes (coding: *gusB*; non-coding *U6snRNA*) and lncRNA *Neat1* is displayed. EB2-3 cells were treated with 1 µM ß-estradiol for 0 h and 6 h, ER/EB2-5 cells were depleted for estr. and reactivated for 0 h and 6 h. RNA was isolated from 10^7 cells and 1 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA was obtained by absolute quantification. **C** ER/EB2-5 cells depleted for estr. and reactivated for the indicated hours show a peak in RNA abundance of known E2 target genes at 6 h post reactivation of ER/E2 by estrogen. ER/EB2-5 cells were depleted for estr. and reactivated for the indicated hours (lt= long term cultivation under estrogen). RNA was isolated from 10^7 cells and 2 µg RNA were reverse transcribed to cDNA.

To study the regulation of E3A target genes, a wt LCL was compared to a ΔE3A LCL. Both LCLs originate from primary B cells of the same donor infected with recombinant EBV, either the wt EBV BAC (2089) or a ΔE3A EBV BAC respectively (Be715; Figure 11A). In order to account for eventual differences in EBV copy numbers in downstream transcriptome analysis, the copy number of both cells was assessed compared to the Namalwa cell line, which is known to carry two EBV genome copies per cell (Whitaker, 1985; Figure 11B). As the wt LCL only carries 1.85-fold more EBV copies compared to the ΔE3A LCL, this difference was not taken into account for further analyses. Since transcriptome variations can accumulate in a mutant cell line due to adaption of the cell, we also compared the regulation of candidate genes in this setting (wt versus ΔE3A LCL) to the regulation in a dox-inducible E3A cell system by RT-qPCR (Figure 11C). *ADAMDEC1* was repressed 1.7 - fold in the dox-inducible system, while it was repressed 243-fold in the wt vs. ΔE3A system (Affymetrix Array: 44.8 - fold repression), *BMP4* was repressed 2.4-fold in the dox-inducible system, while in the wt vs. ΔE3A system, it could not be assessed because repression was beyond detection limit (Affymetrix Array: 17. 5 - fold repression) and *CXCL10* was repressed 5.4 - fold in the dox-inducible system, while due to a repression beyond detection limit in the wt vs. ΔE3A system, it could not be assessed (Affymetrix Array: 30.2 - fold repression; Affymetrix Data from Hertle et al., 2009). Thus, the wt vs. ΔE3A system shows the same trend of regulation than the one observed in the inducible system, albeit the effect of repression is much stronger.

Comparing the wildtype with an E3A mutant cell lines was suitable to investigate E3A dependent gene regulation. Both cell lines did not differ substantially in their EBV copy numbers and a strong trend of regulation can be observed, which could be confirmed in an E3A-inducible cell line.





EBV copy numbers

	wt LCL	ΔE3A LCL	Namalwa
viral BALF5/ cellular ß2m	2.80	1.52	0.93
norm. to Namalwa	6.03	3.27	2

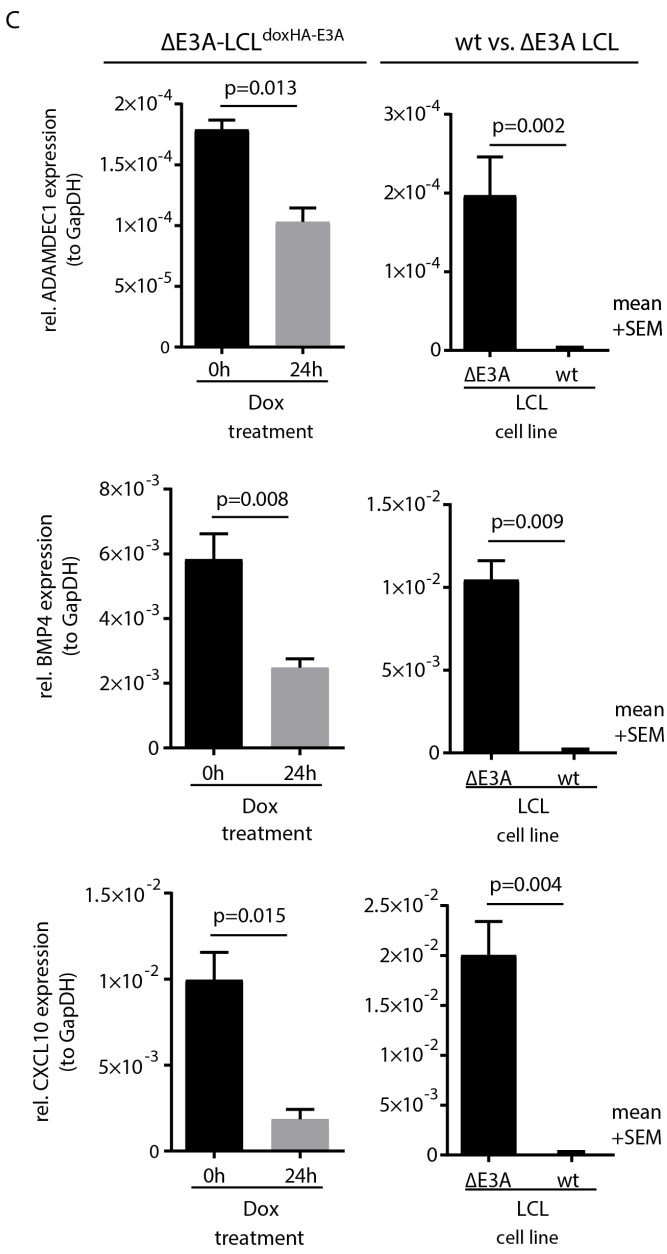


Figure 11: E3A cell system characterization. A Schematic representation of the E3A cell system. Both LCLs originate from primary B cells of the same donor infected with recombinant EBV, either the wt BAC (2089) or a ΔE3A BAC (Be715. **B** LCLs appear to have ~ 1.85-fold more EBV copies than ΔE3A LCLs. gDNA was isolated from 5 x 10^6 cells and qPCR of a viral (*BALF5*; upper panel) and a cellular (*ß2m*; middle panel)

housekeeping gene was performed. Mean and SEM obtained by absolute quantification per 14 ng gDNA ($n_{biol.}$ =3) are displayed. Copy number calculation for wt and mut cell line was normalized to Namalwa cell line (bottom panel). **C** The applied E3A cell system (wt vs ΔE3A LCL) shows same trend of regulation of E3A targets as a dox-inducible E3A system ΔE3A-LCL$^{doxHA\text{-}E3A}$. Quantification of expression level of transcripts of three known E3A target genes *ADAMDEC1, BMP4* and *CXCL10* relative to the expression level of *GAPDH* is shown for two different cell systems, the doxHA-E3A inducible LCL (MH1680-9; left panel) and a wt LCL compared to a ΔE3A LCL of the same donor. The doxHA-E3A inducible LCLs were treated for 0 h and 24 h with 1 µM dox. RNA was isolated from cell pellets and 2 µg RNA was processed for RT-qPCR. Mean and SEM obtained by relative quantification (to *GAPDH*) is displayed. P-values received by paired t-test as indicated.

3.2.2 Transcriptome analysis by RNA-Seq

To examine the genome-wide transcriptional regulation by E2 and E3A, RNA-Sequencing was conducted. The RNA of three biological replicates of several conditions was sequenced:

i) The nucleic and the cytoplasmic compartment of ER/EB2-5 cells depleted from estrogen for 4 d and reactivated for 0 h (-estr.) and 6 h (+estr.) or reactivated under the treatment of ChX in order to block translation (ChX-estr./ ChX+estr.). The cytoplasmic compartment of the ER/EB2-5 cells with absent *de novo* protein synthesis was not analyzed and discussed in the frame of this work due to high variations between the biological replicates.

ii) The nucleic and cytoplasmic compartments of wt LCLs and ΔE3A LCLs.

The cytoplasmic fraction was separated from the nucleic fraction from the same cell batch.

3.2.2.1 78 % of EBVs genes can be differentially expressed by E2 in the ER/EB2-5 system

Reads obtained from RNA-Sequencing, which didn't align to the human genome were mapped to the viral genome of Akata using the STAR mapper. The Akata EBV strain is non-defective and represents a wild-type virus. It has to be mentioned that mapping to this genome is neither perfect for the E2 cell system (P3HR1 genome+ p554-4) nor for the E3A cell system (2089 BAC/Be715 BAC). Up to now, the Akata genome is the most complete annotation for EBV which is available. Aiming for high precision in the detection of regulated viral genes, for counting and differential expression testing, the RSEM package was applied (B. Li & Dewey, 2011), which effectively deals with ambiguously-mapping reads.

In total, 172 elements were annotated by Lin *et al*. for the Akata genome, including 92 genes, 43 miRNAs and 37 regulatory regions (Lin et al., 2013). Upfront, it has to be mentioned that the nomenclature for EBV open reading frames (ORF) originates from BamH1-restriction fragments

map. The fragments were ordered based on their sizes in descending order (A to Z). The fragment on which the ORFs were found and the direction in which they extend was eponymous. The numerous ORFs are further divided into latent and lytic genes. EBV genes are consecutively expressed in three different phases upon lytic reactivation, the immediate-early, early (intermediate) and late phase. The EBV lytic cycle is triggered by the expression of the immediate early genes, *BZLF1* and *BRLF1*. They induce the expression of genes encoding the viral polymerases and the replication machinery. Late genes encode for structural proteins necessary for packaging viral DNA into capsids and the release of infectious virions (reviewed by McKenzie & El-Guindy, 2015).

We continued with genes covered with > 20 reads per gene in either of the conditions (-/ +estr., ChX-estr./ ChX+estr). The significantly (FDR < 0.05) regulated genes were further analyzed for E2 regulation in the different subcellular compartments/different conditions. 75 genes were detected to be significantly regulated (log2FC > 1 or < -1) by E2 in the cytoplasm, 73 upregulated (log2FCs range from 1.46 to 9.85, Table S3) and two downregulated. 47 genes were detected to be significantly induced (log2FC > 1) by E2 in the nucleus (log2FCs range from 1.13 to 5.48, Table S3) and 68 were detected as significantly induced in the nucleus in absence of *de novo* protein synthesis (+ChX; log2FCs range from 1.02 to 10.53, Table S3). Genes only regulated in absence of *de novo* protein synthesis (+ChX) in the nucleus were ignored in order to exclude false positives. 46 genes were induced by E2 in both (-ChX/+ChX) conditions. 39 genes were induced by E2 in all three states (cytoplasm/ nucleus/ nucleus +ChX). Surprisingly, most of the regulated genes are lytic genes. It can be observed that a greater number of genes are significantly regulated in the cytoplasm compared to nucleus. Furthermore, the regulation under ChX seems for some genes to amplify the induction.
Comparing E2 regulation of viral genes with E3A regulation, it can be observed that less genes are regulated by E3A. Ignoring the genes covered with < 20 reads in either of the conditions (wt/ΔE3A), 42 genes are significantly (FDR < 0.05) regulated (log2FC > 1 or < -1) in the cytoplasm, 41 upregulated (log2FCs range from 1.08 to 10.82, Table S3) and one gene, the lncRNA *EBER1* downregulated (-1.81). 23 genes were upregulated in the nucleus (log2FCs range from 1.0 to 11.44, Table S3) and three genes downregulated. The wt versus ΔE3A mutant ratio leads to the highest log2FC values for E3A. RPMS1 was regulated by E3A in both compartments. Ten genes, including *BNRF1* were induced by both TFs in all conditions/compartments. 42 genes were induced by both TFs in one of the compartments.

To conclude, E2 induces a great number of EBV genes, including lytic genes. Under the assumption that the ER/EB2-5 system resembles the processes during infection, one could suggest that E2

induces a lytic cycle. Whether this induction leads to an abortive or a productive lytic cycle needs further investigations. E3A regulated most of its target genes in concert with E2.

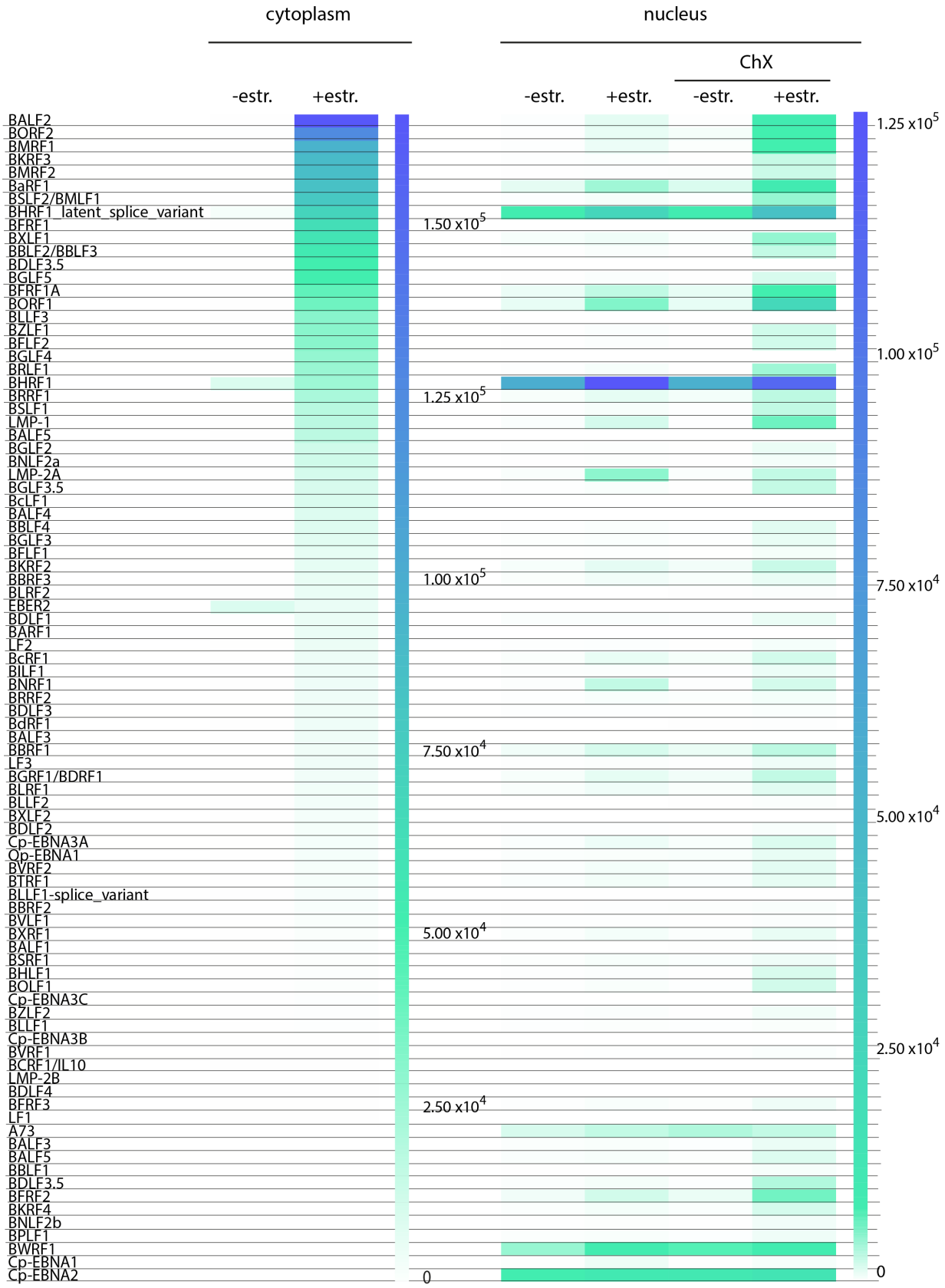


Figure 12: Heatmap displaying the mean normalized read counts of all significantly ($p \leq 0.05$) differentially expressed genes by E2 in the cytoplasm or the nucleus (- /+ ChX) detected by RSEM. Each horizontal row represents a viral gene and every vertical row the mean normalized read counts ($n_{biol.}$ =3) as indicated. All genes sign. differentially expressed by E2 (no log2FC cutoff) in any compartment/condition are listed and sorted in descendant manner according to the mean of norm. read counts in the cytoplasmic +estr. condition. Genes with < 20 reads in both conditions were excluded. Two different scales were applied for the cytoplasmic and the nucleic compartment. Cytoplasm: smallest value= white= 0 read counts, baseline value=green=$5x\ 10^4$ read counts, largest value= blue= $1.65x\ 10^5$ read counts. Nucleus: smallest value= white= 0 read counts, baseline value= green= $5x\ 10^3$ read counts, largest value=blue=$1.26\ x\ 10^5$ read counts. Graph Pad Prism was used for plotting.

In order to confirm candidate target genes of E2 detected by RNA-Seq, *BGRF1/BDRF1* (Figure 13), *BHRF1* (Figure 14) and *BNRF1* in the intron of *LMP2A* (Figure 16) were investigated by RT-qPCR. All three candidate loci are characterized by being induced by E2 in cytoplasm, the nucleus, or both and being regulated in absence of *de novo* protein synthesis. For two out of three loci, E2 binding sites reside in close genomic proximity.

The early gene *BGRF1/BDRF1* encodes most probably (by sequence homology) for the tripartite terminase (TRM) subunit 3 or DNA packaging subunit 3 protein, which belongs to the herpesviridae TRM protein family (Figure 13). The three HHV's terminase subunits are conserved across the HHV family. Together they form a complex in the host cytoplasm, the third subunit is relevant for the nuclear localization. The DNA packaging complex is responsible for the translocation of the viral DNA in empty capsids. TRM3 exerts RNase H-like nuclease activity with which concatemeric viral DNA is cleaved into genomes of equal length (https://www.viprbrc.org, https://www.uniprot.org). The regulation of a spliced transcript of *BGRF1/BDRF1* could be confirmed by RT-qPCR (Figure 15, upper panel). The transcript is enriched in the cytoplasm. The log2FC in RNA-Seq under ChX is 2.14, thus, this target is regulated in the absence of *de novo* protein synthesis.

BHRF1 encodes for a homolog of the mammalian B cell leukemia/lymphoma Bcl-2 protein (Figure 14). BHRF1 might counteract the MYC-induced apoptosis. It binds to a number of pro-apoptotic proteins expressed by host cells. *BHRF1* expression leads to a resistance to a range of cytotoxic agents (Kvansakul et al., 2010). Two isoforms are annotated for *BHRF1*, a monoexonic transcript which is expressed in the early lytic phase and a latent splice variant which consists of two exons. Specific primers for the distinct detection of the early lytic variant could be established for RT-qPCR, the regulation by E2 could be confirmed (Figure 15, lower panel). The transcript is enriched in the nucleus. The log2FC in RNA-Seq under ChX is 1.02, thus, this target is weakly regulated in the absence of *de novo* protein synthesis.

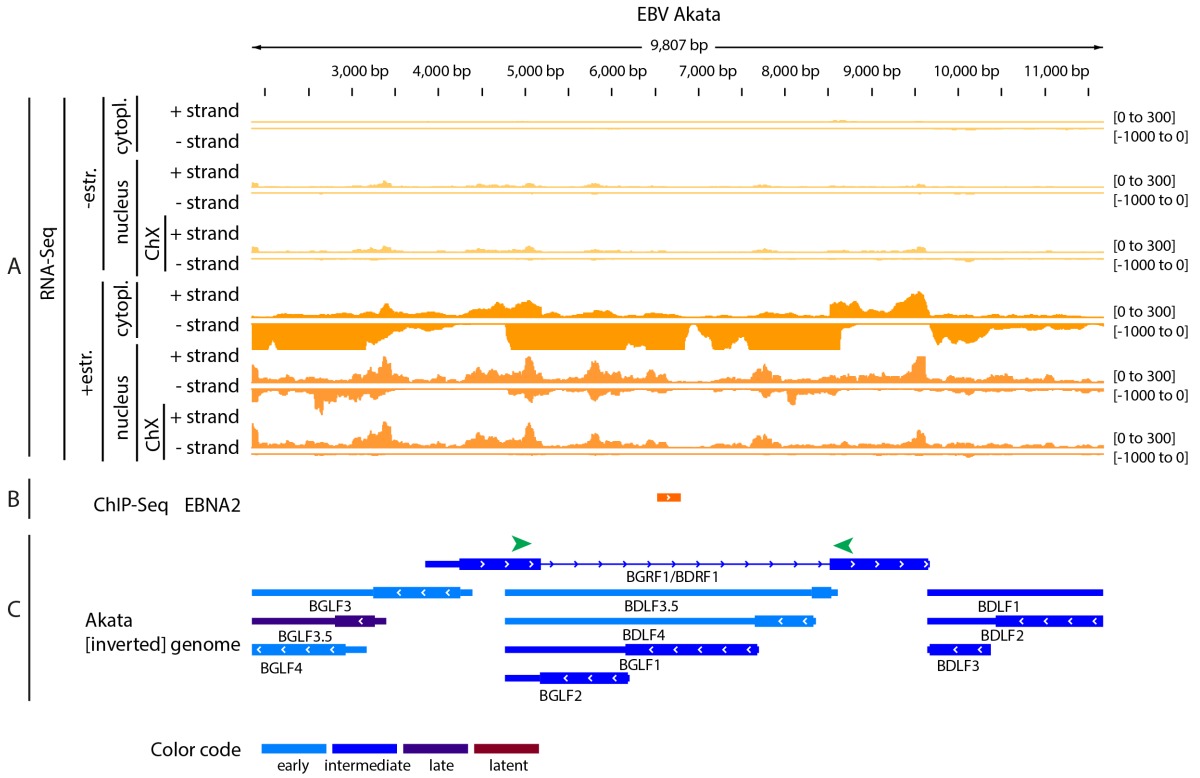


Figure 13: Overview over *BGRF1/BDRF1* locus with pictured E2 dependent induction of transcription in ER/EB2-5. Schematic map depicting **A** Expression based RNA-Seq tracks displaying the coverage in the different conditions/fractions as indicated (tracks were set to the stated data range). **B** Peak track obtained from E2 ChIP-Seq (Laura Glaser; mapped to Akata). **C** Annotation track provided by Flemington Lab. Green arrows indicate RT-qPCR primer positions; underneath: color code for annotation track.

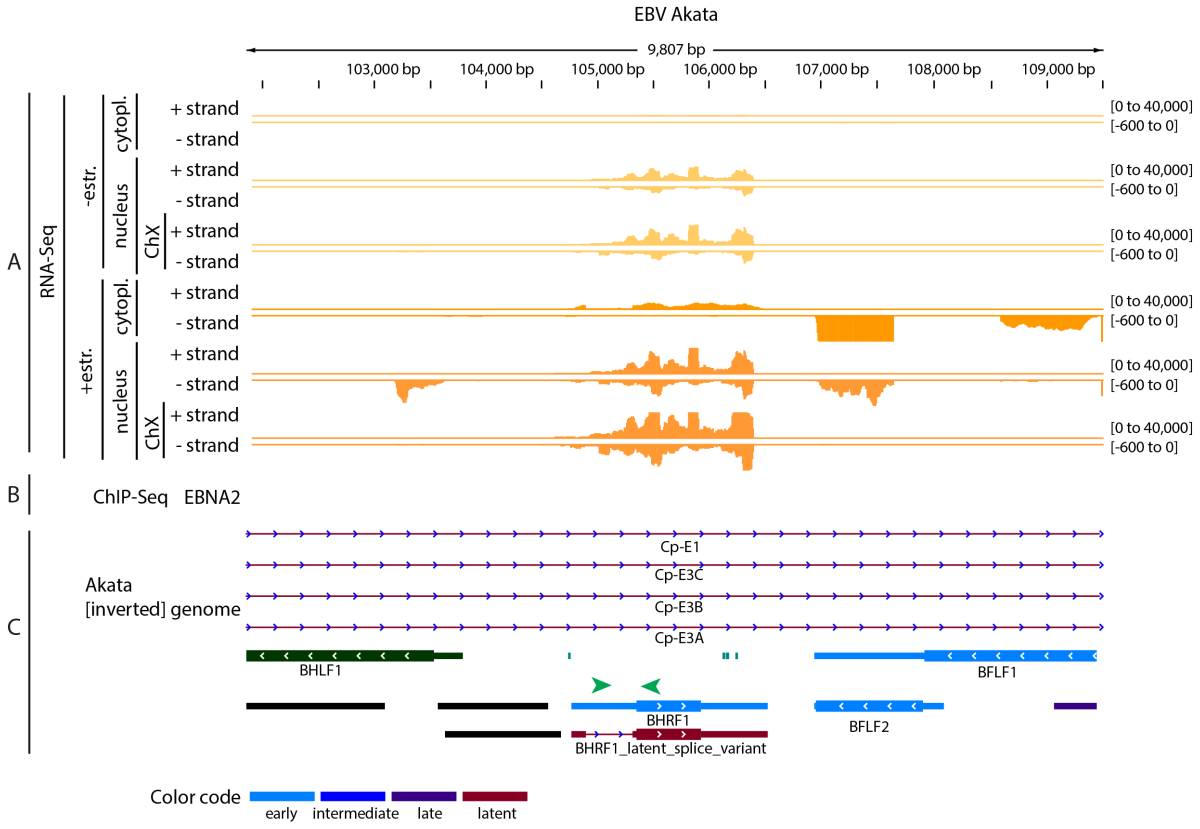


Figure 14: Overview over *BHRF1* locus with pictured E2 dependent induction of transcription in ER/EB2-5. Schematic map depicting **A** Expression based RNA-Seq tracks displaying the coverage in the different conditions/fractions as indicated (tracks were set to the stated data range). **B** Peak track obtained from E2 ChIP-Seq (Laura Glaser; mapped to Akata). **C** Annotation track provided by Flemington Lab. Green arrows indicate RT-qPCR primer positions; underneath: color code for annotation track.

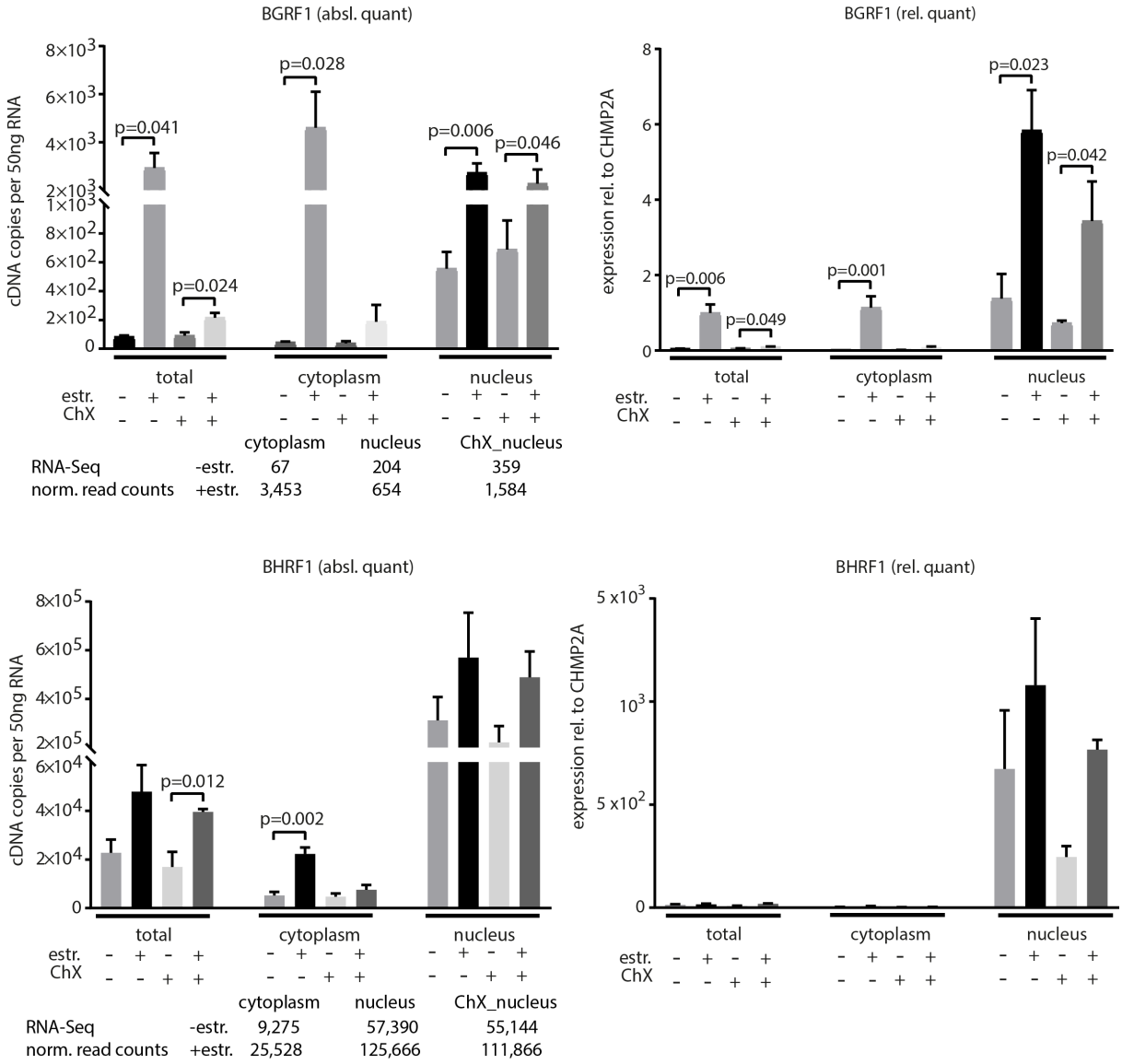


Figure 15: RT-qPCR confirmation of transcripts of the viral E2 target genes *BGRF1/BDRF1* and *BHRF1*. Bar graphs showing the absolute (left) and the relative (right; rel. to *CHMP2A*) quantification by RT-qPCR of the transcripts two E2 target genes *BGRF1/BDRF1* and *BHRF1* in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h partially under ChX treatment. RNA was isolated from 10^7 cells (total) or cell fractions (1.2x 10^7 cells for cytoplasm and 2x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.}$ =3, $n_{biol.}$ =3). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq are displayed. Graph Pad Prism was used for plotting.

The early gene *BNRF1* encodes for the major EBV tegument protein (Figure 16 bottom). BNRF1 is important for the EBV invasion of B lymphocytes (López et al., 2005). BNRF1 was reported to have a function in the viral transport from the endosomes to the nucleus (Feederle et al., 2006). Furthermore, it targets host-cell intrinsic defense proteins and promotes viral early gene activation (Tsai, Thikmyanova, Wojcechowskyj, Delecluse, & Lieberman, 2011). The monoexonic transcript overlaps with the introns of *LMP2A* and *LMP2B*. Both genes are transcribed from the sense strand. The primers are specific for BNRF1 transcript but still could detect unspliced intronic *LMP2A/B*. The regulation could be confirmed by RT-qPCR (Figure 17, upper panel). The transcript can be detected in the cytoplasm, which supports the identification of *BNRF1*. Nevertheless, the high level of abundance in the nucleus detected by RT-qPCR compared to the cytoplasm could be due to the detection of unspliced LMP2A.
LMP2A is a known direct target gene of E2 (reviewed in Kempkes & Ling, 2015; Figure 16 top). It is a latent transmembrane protein that negatively regulates B cell receptor signaling and has been furthermore reported to activate (MAPK-, PI3-K/AKT- signaling) or down-regulate (NF-κB- and STAT-signaling) several other pathways important for different fundamental events such as proliferation, differentiation or transformation (El-Sharkawy, Al Zaidan, & Malki, 2018). Here it is shown as a control for E2 regulation. The gene can be identified to be strongly regulated in the cytoplasm. Regulation can be confirmed by RT-qPCR (Figure 17, lower panel). Since the housekeeping gene *CHMP2A* was observed to be regulated in these experiments (Figure 18), the relative quantifications are also shown. However, since *CHMP2A* is a protein coding gene enriched in the cytoplasm, this normalization leads to an amplification of the signal in the nucleus. The E2 dependent upregulation of candidate lytic genes detected by RNA-Seq could also be confirmed by immunoblotting (data not shown).

In summary, we could observe an E2-dependent regulation of viral genes of the lytic and the latent phase of EBV in the ER/EB2-5 cells. Included were genes, which play a role in DNA packaging in viral capsids, in the EBV invasion or which have anti-apoptotic capacity. However, the regulation of viral target genes by E2 and the importance for EBV biology requires further investigation.

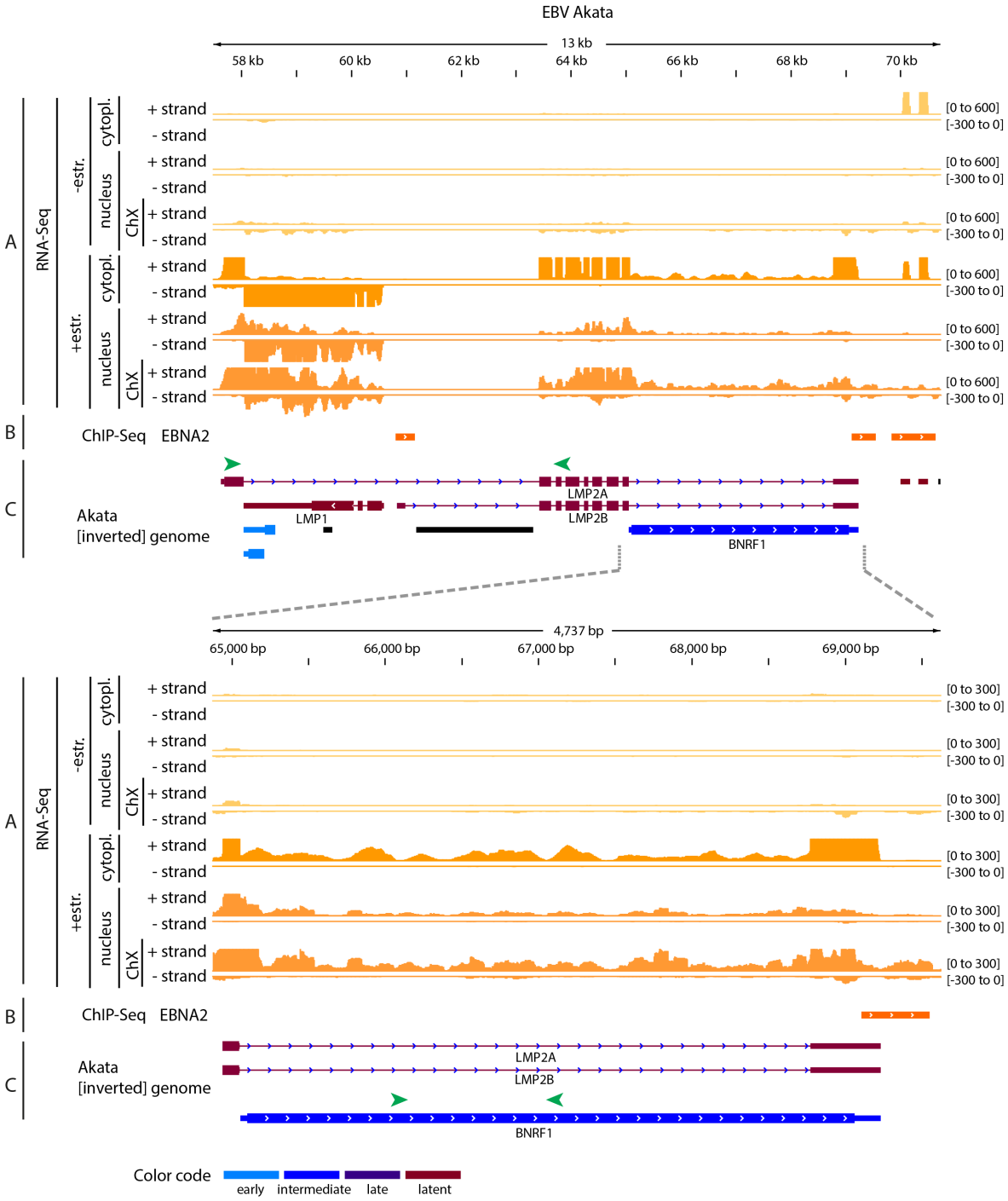


Figure 16: Overview over *LMP2A* locus with pictured E2 dependent induction of transcription in ER/EB2-5. Upper panel: *LMP2A* overview; lower panel: a magnification of the upper panel zooming in on *BNRF1*. Schematic map depicting **A** Expression based RNA-Seq tracks displaying the coverage in the different conditions/fractions as indicated (tracks were set to the stated data range). **B** Peak track obtained from E2 ChIP-Seq (Laura Glaser; mapped to Akata). **C** Annotation track provided by Flemington Lab; underneath: color code for annotation track

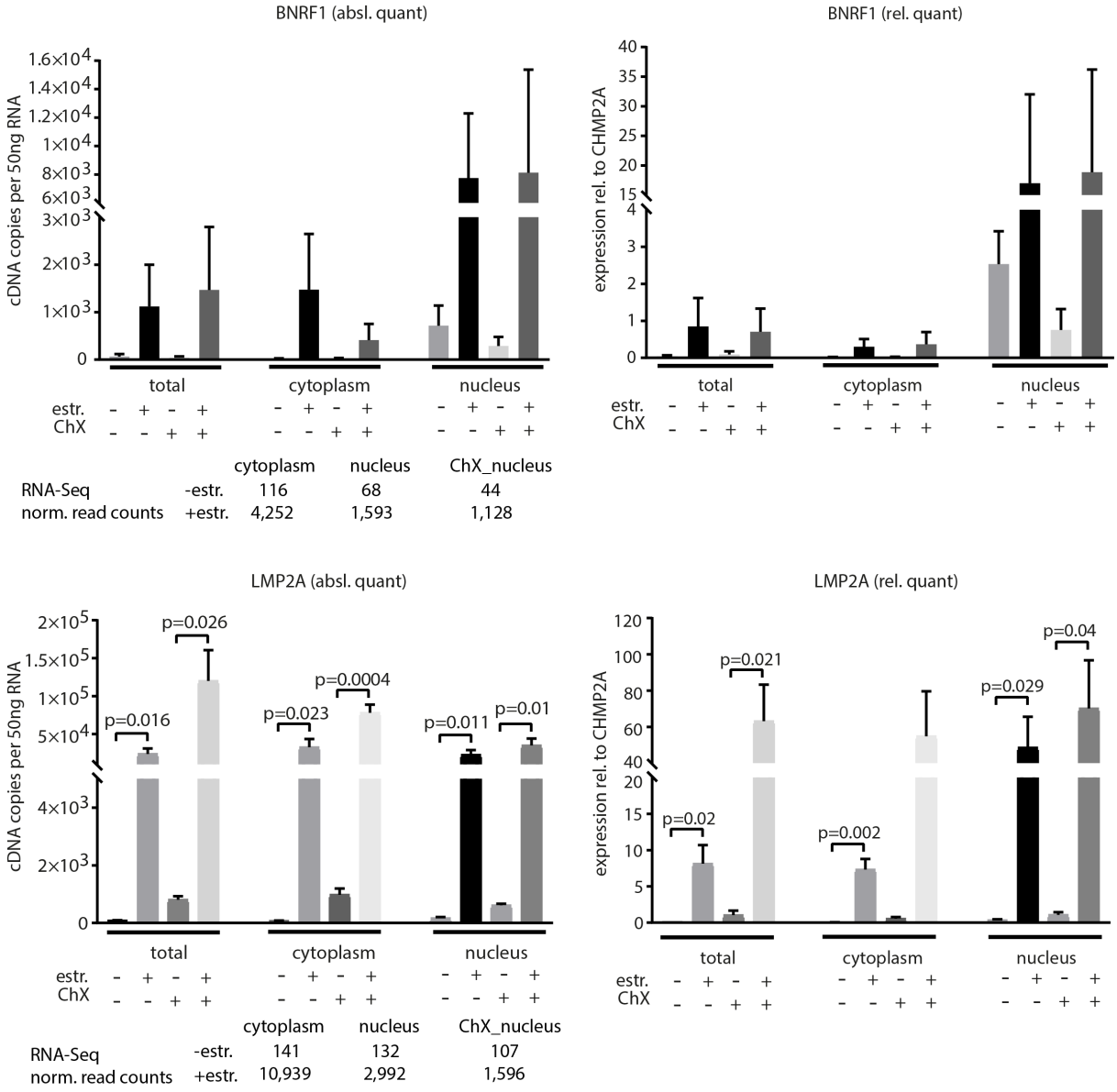


Figure 17: RT-qPCR confirmation of transcripts of the viral E2 target genes *BNRF1* and *LMP2A.* Bar graphs showing the absolute (left) and the relative (right; rel. to *CHMP2A*) quantification by RT-qPCR of the transcripts of two E2 targets *BNRF1* and *LMP2A* in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h partially under ChX treatment. RNA was isolated from 10^7 cells (total) or cell fractions (1.2x 10^7 cells for cytoplasm and 2x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.}$ =3, $n_{biol.}$ =3). Mean +SEM is displayed. P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq are displayed. Graph Pad Prism was used for plotting.

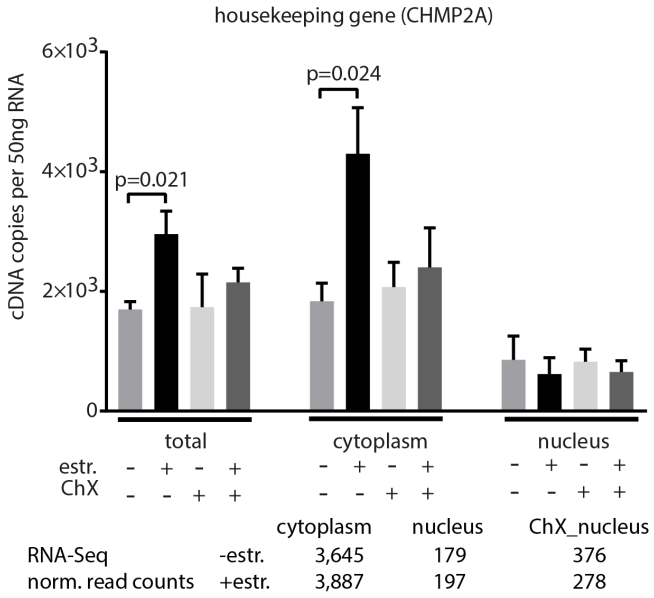


Figure 18: Housekeeping gene as control for RT-qPCR. Bar graphs showing the absolute quantification of RT-qPCR of a transcript of housekeeping gene *CHMP2A* in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h partially under ChX treatment. RNA was isolated from 10^7 cells (total) or cell fractions (1.2x 10^7 cells for cytoplasm and 2x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.}$ =3, $n_{biol.}$ =3). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq are displayed. Graph Pad Prism was used for plotting.

3.2.2.2 E2 manipulates the host transcriptome

3.2.2.2.1 Efficiency of read alignment is compartment and mapper dependent

In order to assess the coverages of our samples, reads obtained from RNA-Seq were aligned to the human genome hg19.

We aimed to compare four different mapping algorithms, HISAT (D. Kim et al., 2015), STAR (Dobin et al., 2013), ContextMap2 (Bonfert et al., 2015), and TopHat2 (D. Kim et al., 2013). Mapping was conducted by Gergely Csaba. Investigating the whole genome, transcriptome, intergenic and intronic regions and known and novel junctions, the alignment performance varied. As expected, in average ~80% of all reads aligned to hg19 (Figure S8). HISAT performed poorly for the samples derived from the E3A system (Figure S8B) and mapped incomparably more reads of the cytoplasm to intergenic junctions (Figure S13). Therefore, this mapper was excluded from all further analyses. More reads from the cytoplasmic compartment could be mapped to annotated transcripts (ENSEMBL GRCh37.75; Figure S9), known (Figure S12) and intergenic (novel) junctions (Figure S13). As expected, more reads from the nuclear compartment could be mapped to intergenic (Figure S10) and intronic (Figure S11) regions.

Taken together, the sequencing was successful (the majority of reads could be mapped) and the quantity of mapped reads is compartment and mapper dependent.

3.2.2.2.2 Strategy for identification of unannotated intergenic and intronic genes

In order to define potential novel genes, regions not overlapping any ENSEMBL annotation were subjected to further analysis. The analysis of nucleic RNAs was difficult due to the nature of immature transcripts. It was not possible to *de novo* assemble transcripts. Since the application of *de novo* assembly tools like StringTie were not feasible for the nucleic fraction of the cells, Gergely Csaba sought to develop a bypassing strategy.
The determination of detected intergenic/intronic transcribed genes consist of multiple steps (Figure 19). First, candidate intergenic/intronic regions were derived from each sample independently. These regions were given by proximal (≤ 50 bp from closest fragment ends) or overlapping fragments (read pairs) not mapping to any ENSEMBL gene (intergenic) or exon (intronic). Since sequencing was conducted strand specific, also regions antisense to overlapping known genes were collected. The second step was, to combine these “raw“ intergenic/ intronic regions to replicate-consistent ones. To this end, the derived regions from all mappers and all replicates for each condition were used in a joined analysis: all intergenic/intronic candidates having ≥ 50 supporting reads were merged and the part of the intergenic/intronic region consistently detected by all mappers in all biological replicates for the same condition was extracted. The discovered intergenic genes were classified by their genomic overlap with annotated ENSEMBL genes (intergenic if no overlap, a combination of sense/antisense genebody/3_PRIME/5_PRIME/FULL otherwise). As a result, a purified set of observed non-annotated (ENSEMBL) regions was obtained. As these regions could still be a byproduct of transcription of gene-overlapping regions, we corrected for noise by comparing the changes to the changes of the overlapped gene in the differential analysis.

Thus, a different strategy to investigate unannotated differentially expressed genes could be developed and was used in the following analyses.

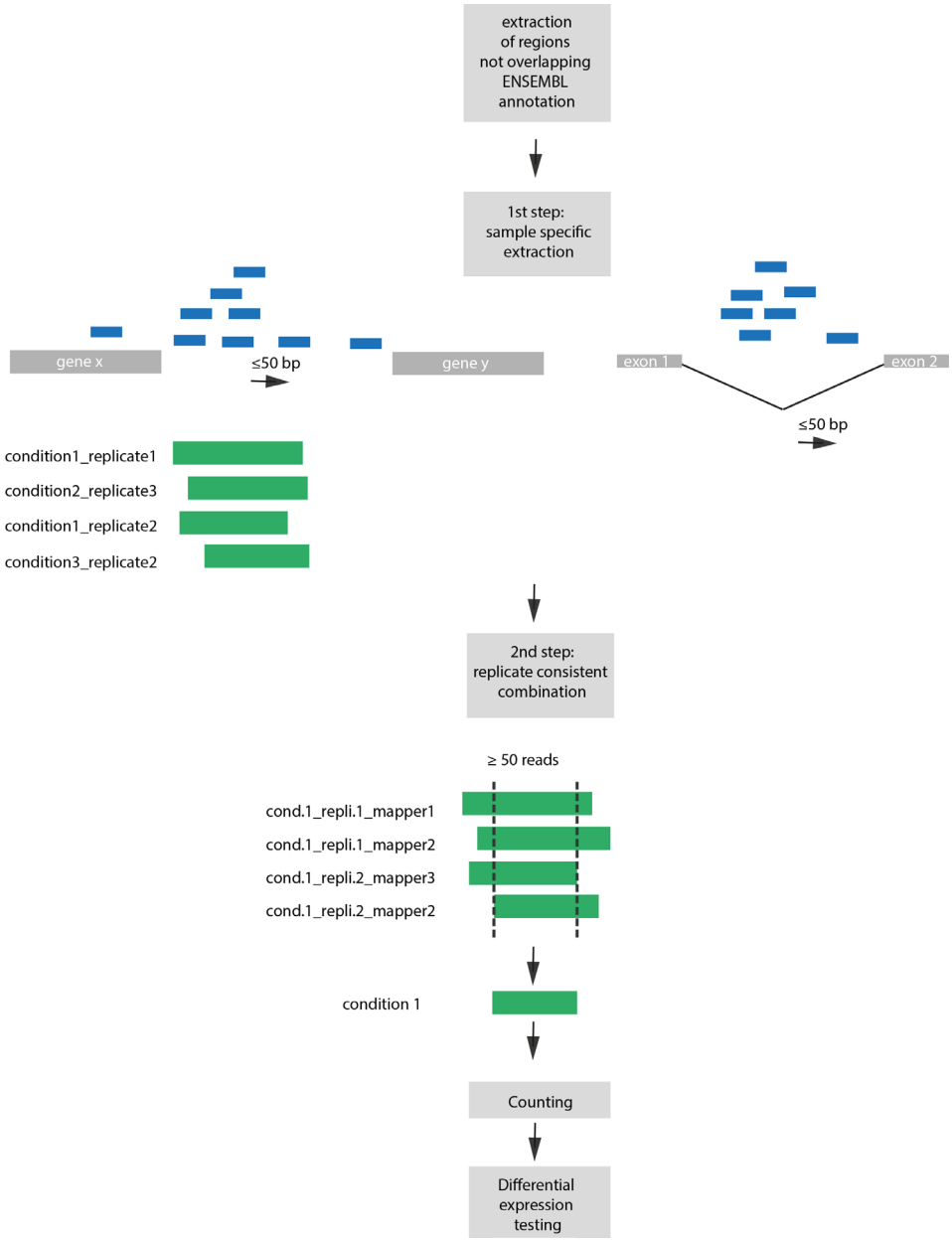


Figure 19: Simplified scheme on the inference of intergenic and intronic genes from extracted read covered regions not consistent with ENSEMBL gene or exon annotation. Two major steps were performed: intergenic (between two genes annotated by ENSEMBL) and intronic (between two exons annotated by ENSEMBL) regions, which were covered with reads were extracted from each sample (condition/replicate). The extracted regions covered with ≤ 50 reads are joined for each condition separately and the common region detected in all biological replicates of the same condition by all used mappers was defined as a new intergenic or intronic gene. Blue boxes= simplified read pairs (=fragments). Green boxes= extracted regions.

3.2.2.2.3 Biological replicates show high similarity in RNA-Sequencing

In order to validate the quality of the biological replicates, two quality controls were conducted. On the one hand, the raw read counts of the replicates were compared with each other after mapping (conducted by Gergely Csaba). The samples of cytoplasmic (Figure S14) and nuclear compartment (Figure S15) of ER/EB2-5 cells - / + estr. were similar as they lined up on the diagonal and the samples of cytoplasmic (Figure S17) and nuclear (Figure S18) compartment of wt/ΔE3A cells were highly similar as well as. In contrast, the samples of the nuclear compartment of ER/EB2-5 cells ChX-estr./ ChX+estr. show high variations compared to each other also at high read counts (Figure S16). On the other hand, a Spearman correlation of average read coverages of the replicates was calculated (Figure 20 to Figure 22) proving, that overall biological replicates of same treatment or condition and same compartment show high similarities and form clusters. In general, all prepared samples (not only replicates) show a high correlation (r_{min}= 0.75). Interestingly, in the E2 system, the samples corresponding to the same compartments form a cluster of high correlation while the samples deriving from the same treatment form a higher order cluster (Figure 20). In contrast, in the E3A system, the samples of the same cell line form a high correlation cluster while the samples corresponding to the same compartment show a higher order cluster (Figure 22), demonstrating the stronger diversity between +/- E3A as compared to +/- E2. All the samples of the nuclear compartment of ER/EB2-5 cells ChX-estr./ ChX+estr. show a high correlation (r= 0.83-0.92; Figure 21). The clustering here is a bit diffuse probably due to the high variations between the biological replicates of the same condition (Figure S16). ChX-treated samples form a higher order cluster independent of estrogen treatment, indicating a high impact of ChX on the transcription, only exception is the first replicate of ChX+estr.

Taken together, the biological replicates used in this study exhibited acceptable similarity. However, ChX-treated replicates show higher variations which have to be considered in further analysis.

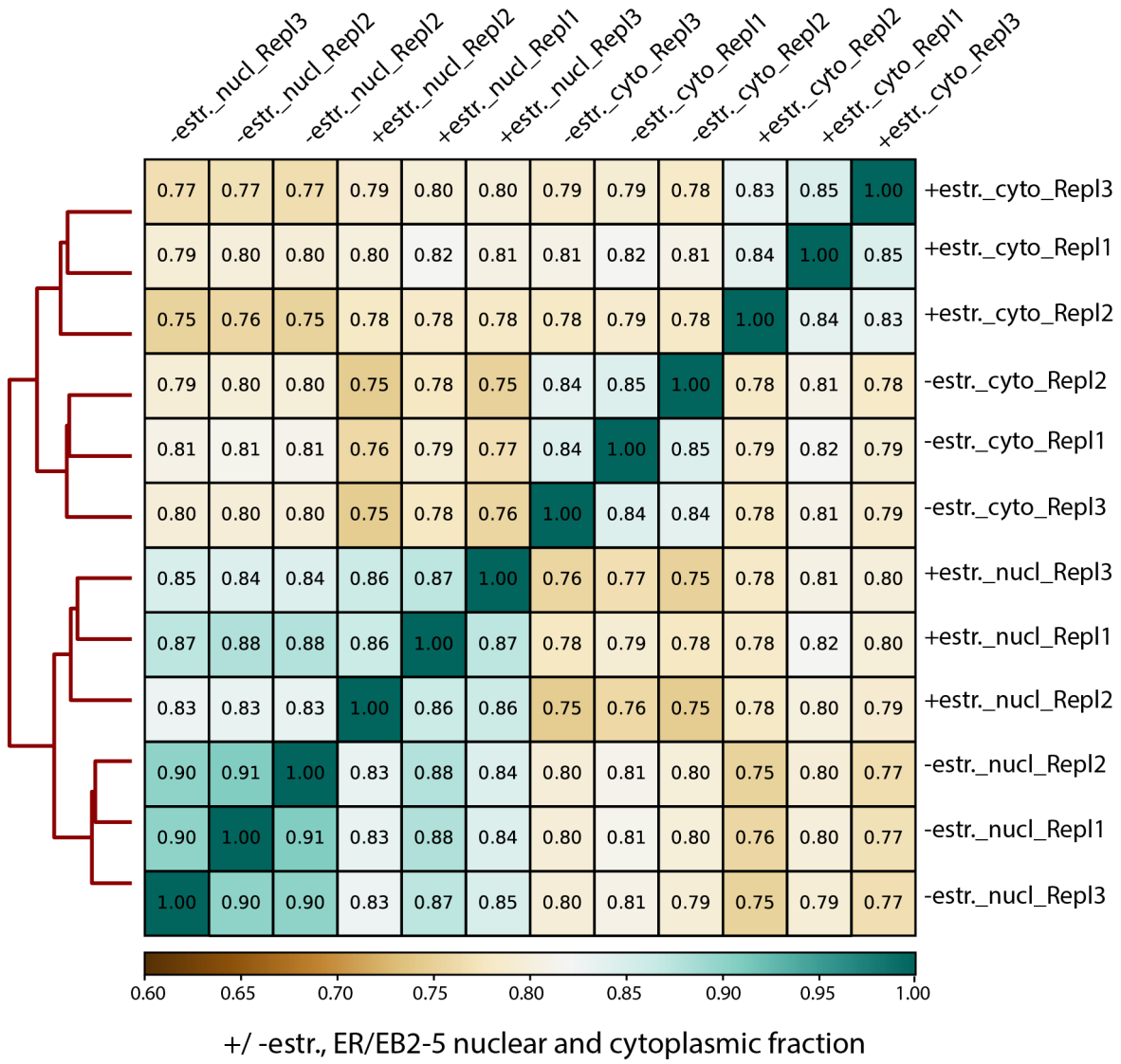


Figure 20: Biological replicates of the same condition and the same subcellular location cluster by correlation analysis. Average read coverages for BAM files (mapped reads) were calculated. The genome was split into bins of 1000 bp. For each bin, the number of reads found in each BAM file was counted. Unsupervised hierarchical clustering was performed by Spearman correlation and visualized in a heatmap; scale from 0.6 to 1.0. GALAXY platform was used for computation and visualization.

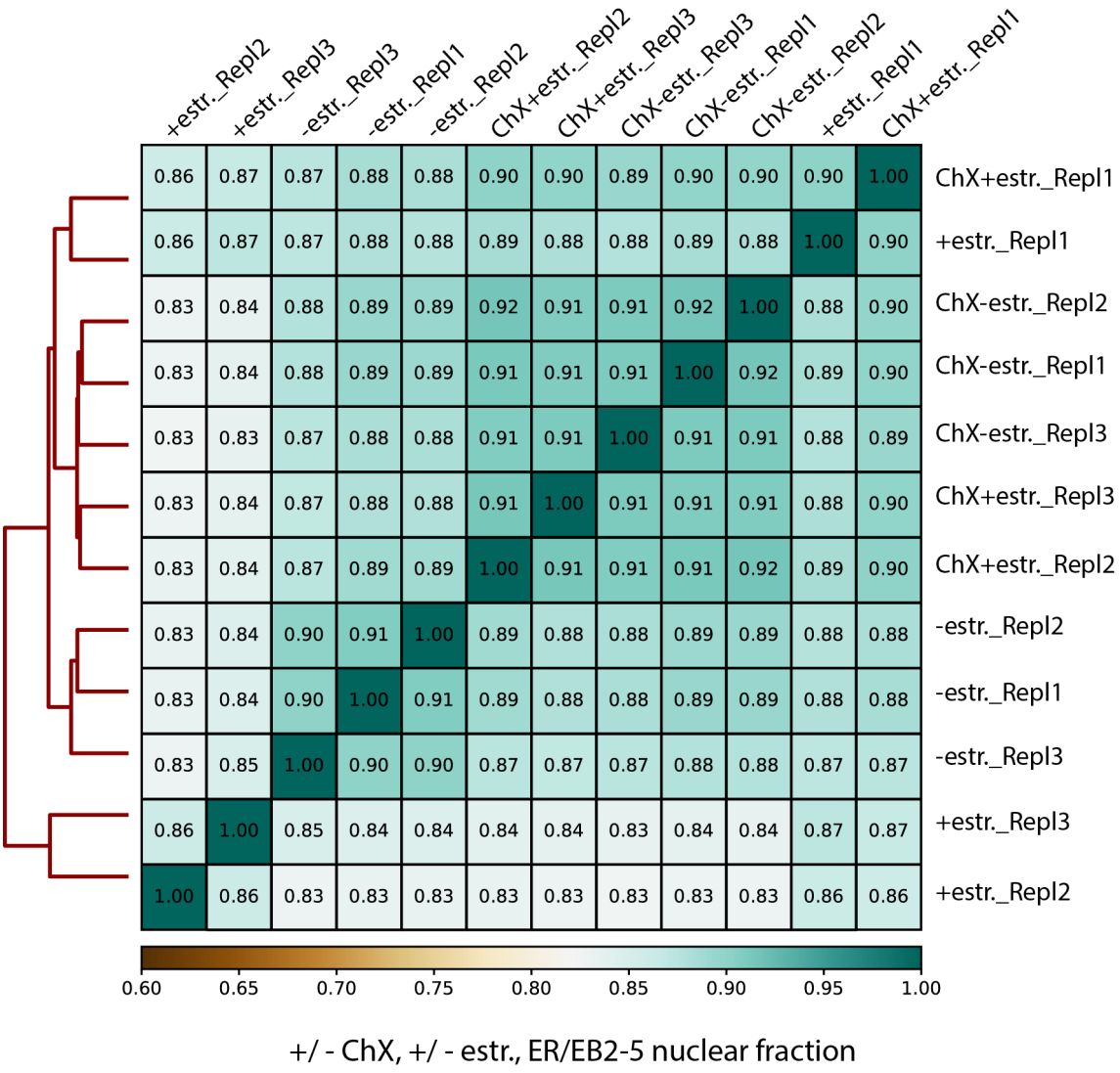


Figure 21: Excluding replicate 1 +estr. , biological replicates of the same condition cluster by correlation analysis. Average read coverages for BAM files (mapped reads) were calculated. The genome was split into bins of 1000 bp. For each bin, the number of reads found in each BAM file was counted. Unsupervised hierarchical clustering was performed by Spearman correlation and visualized in a heatmap; scale from 0.6 to 1.0. ChX- samples were reactivated for E2 for 0 h and 6 h by estrogen (-estr. and +estr., respectively) under the treatment of ChX (1 h before estrogen addition) leading to the inhibition of *de novo* protein synthesis. GALAXY platform was used for computation and visualization.

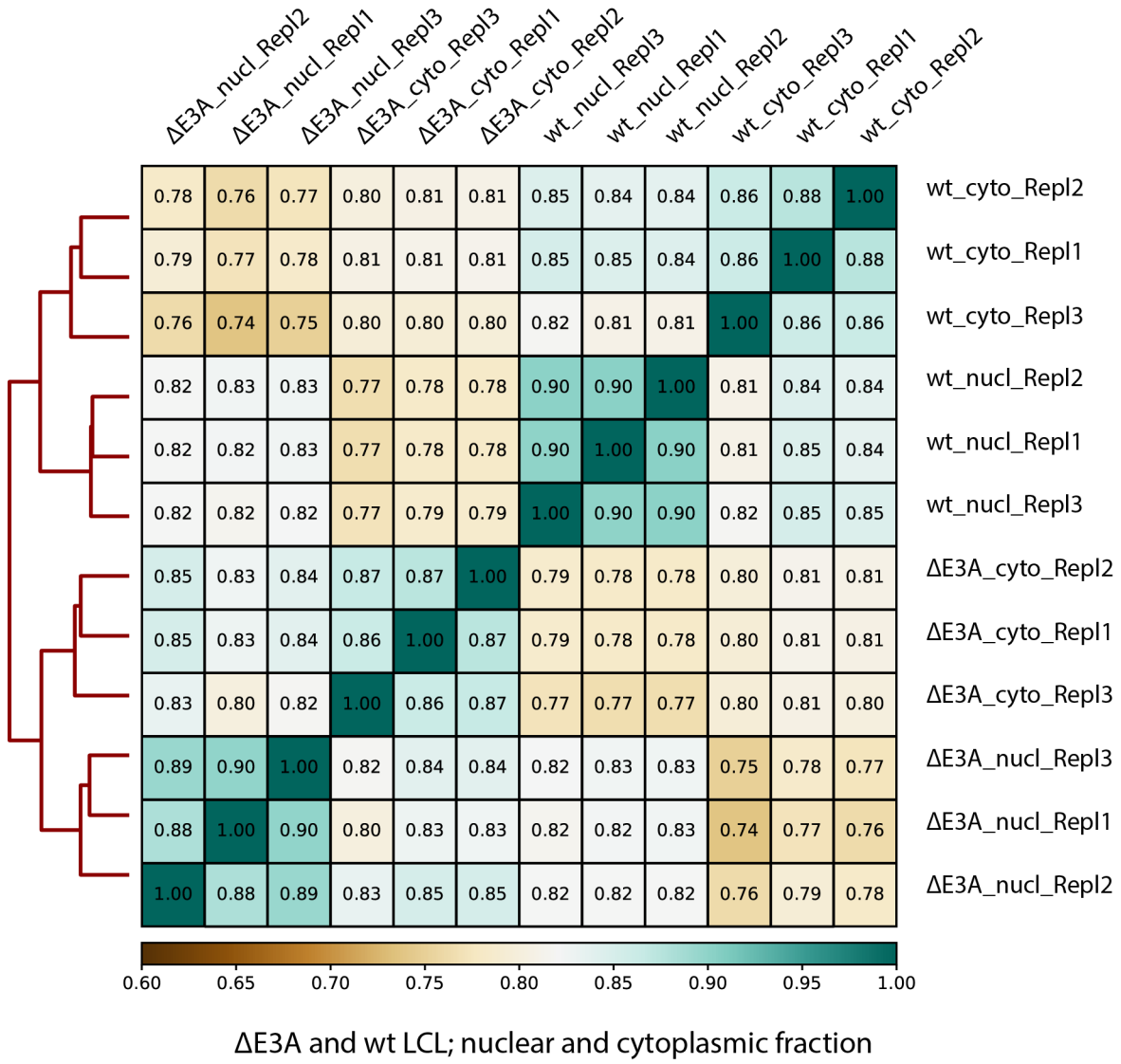


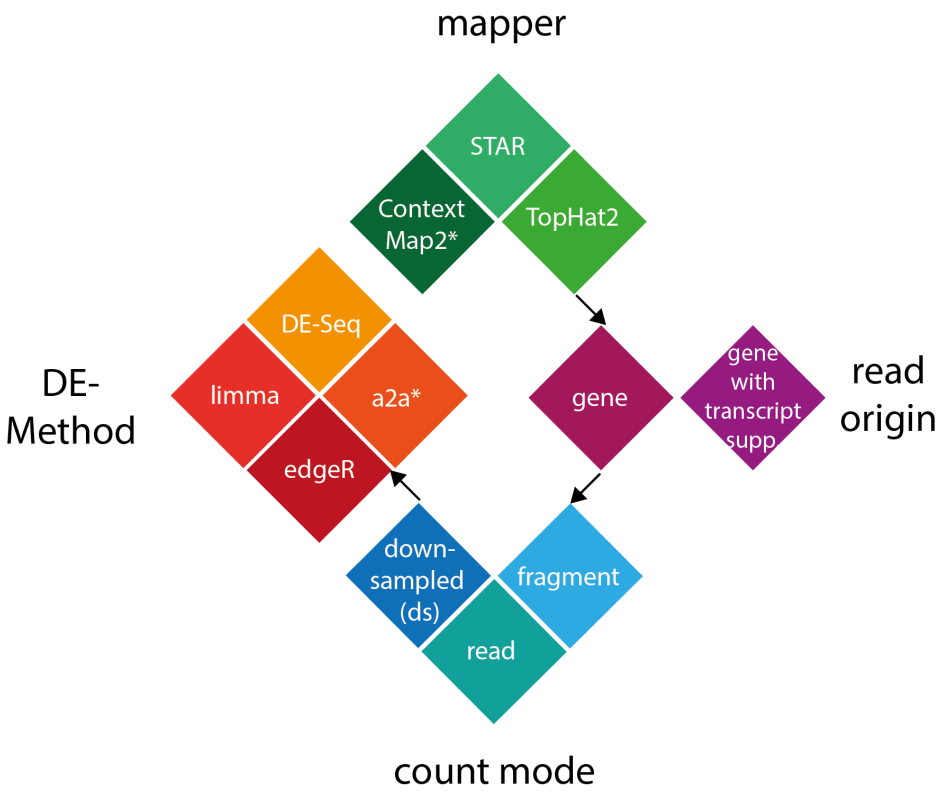
Figure 22: Biological replicates of the same cell line, +/- E3A, and same subcellular location cluster by correlation analysis. Average read coverages for BAM files (mapped reads) were calculated. The genome was split into bins of 1000 bp. For each bin, the number of reads found in each BAM file was counted. Unsupervised hierarchical clustering was performed by Spearman correlation and visualized in a heatmap; scale from 0.6 to 1.0. GALAXY platform was used for computation and visualization.

3.2.2.2.4 Four dimensional (4D)- combinatorics promise a high reliability in detection of regulated genes

Following mapping, aligned read pairs (fragments) had to be counted on the chosen count feature (genomic element). Subsequently, the features were tested for differential expression in a compared condition pair.

Aiming for high reliability in the detection of regulated genes and to circumvent biases by the selection of one certain tool setup, combinatorics using four different dimensions were applied (Figure 23). Three different mapper (STAR, ContextMap2, TopHat2), combined with two different read origins (gene, gene with transcript supporting reads), combined with three different count modes and four different DE-Methods resulted in 72 possible combinations (=setups). For DE-testing four different DE-Methods were used, DE-Seq (Anders & Huber, 2010), limma (Ritchie et al., 2015), edgeR (Robinson et al., 2010) or an in house method, a2a. Regarding the detected amount of significant regulated genes, all combinations using ContextMap2 (*) as a mapper were not advantageous for the samples of the cytoplasmic compartment of ER/EB2-5 cells - /+ estr. (Figure 24A). All combinations using a2a (*) as DE-method yielded slightly better results for the samples of the nuclear compartment of ER/EB2-5 cells ChX-estr./ ChX+estr (Figure 24C). In contrast, the entire setups result in similar outputs for the samples of the nuclear compartment of ER/EB2-5 cells - /+ estr. (Figure 24B), and for the samples of both compartments of wt/ΔE3A cells (Figure 25). We determined to use the common amount of genes consistently detected by the most beneficial setups for the downstream analysis. As a consequence, the combinations using ContextMap2 for the cytoplasmic samples of ER/EB2-5 cells - /+ estr. were disregarded for the downstream analysis. Furthermore, only combinations using a2a for nuclear samples of ER/EB2-5 cells ChX-estr./ ChX+estr were included in downstream analysis. For the nuclear samples of ER/EB2-5 cells ChX-estr./ ChX+estr, our aim was to detect in the most sensitive way, for all other samples our aim was to be most specific.

The detection of differentially expressed genes was not dependent on a certain tool set. We continued to analyze the consistently detected genes by multiple setups which promised a high reliability in the resulting data.



3 x 2 x 3 x 4= 72 possible combinations (setups)

Figure 23: 4D- matrix of mapper, count type, sampling and DE-method combinatorics. Scheme of combinatorics used to circumvent biases in the detection of differentially expressed genes: Three different mapper combined with two different count types combined with three different sampling styles and four different DE-Methods resulting in 72 possible combinations (=setups); ContextMap2, STAR and TopHat2 were applied as mapper. Read origin: Either read pairs aligning completely to a gene were considered (read origin = gene) or read pairs aligning to a gene were only considered, if any gene derived transcript was covered by reads (read origin = gene with transcript support). Different count modes were performed: i) reads were counted ii) reads were downsampled on fragment level iii) downsampling (ds) to decrease impact of highly amplified fragments (for detailed explanation please see section 2.5.6.1, p. 29). *ContextMap2 was excluded for cytoplasmic samples; *a2a was the only DE-Method used for ChX-treated samples

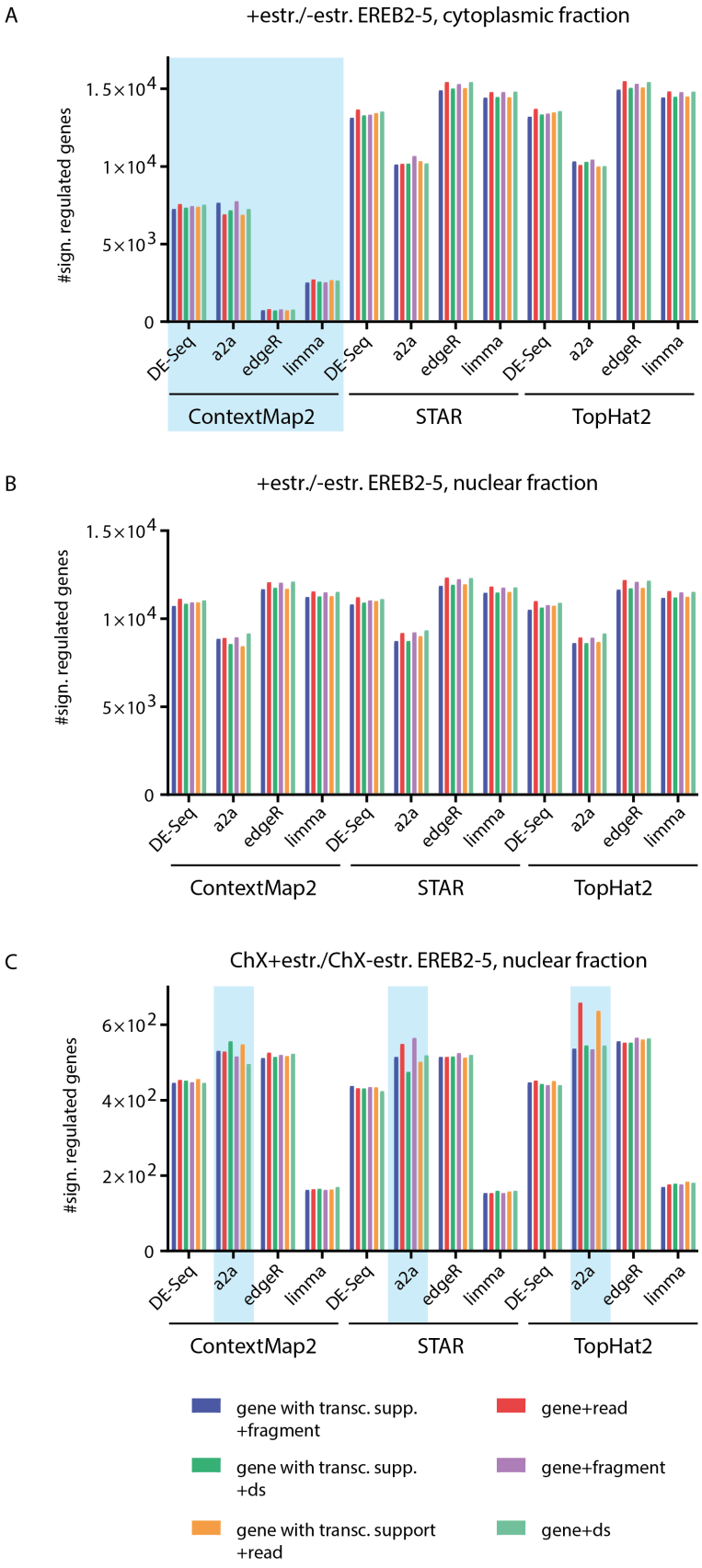


Figure 24: Different setups leading to different (A, C) or similar (B) results regarding the number (#) of significantly regulated genes in the E2 cell system. Bar graphs displaying the number of detected significantly regulated genes in the different setups for **A** regulation by E2 (+estr./-estr.) in the cytoplasm (highlighted in blue are the worst cases, the combinations with ContextMap2as mapper), **B** regulation by E2 (+estr./-estr.)

in the nucleus and **C** regulation by E2 (+estr./-estr.) in the nucleus in absence of *de novo* protein synthesis (ChX; highlighted in blue are the best cases, the combinations with a2a as DE-Method). Gene with transc. supp.: gene with transcript support. Ds: downsampling (for detailed explanation please see section 2.5.6.1, p. 29). Graph Pad Prism was used for plotting.

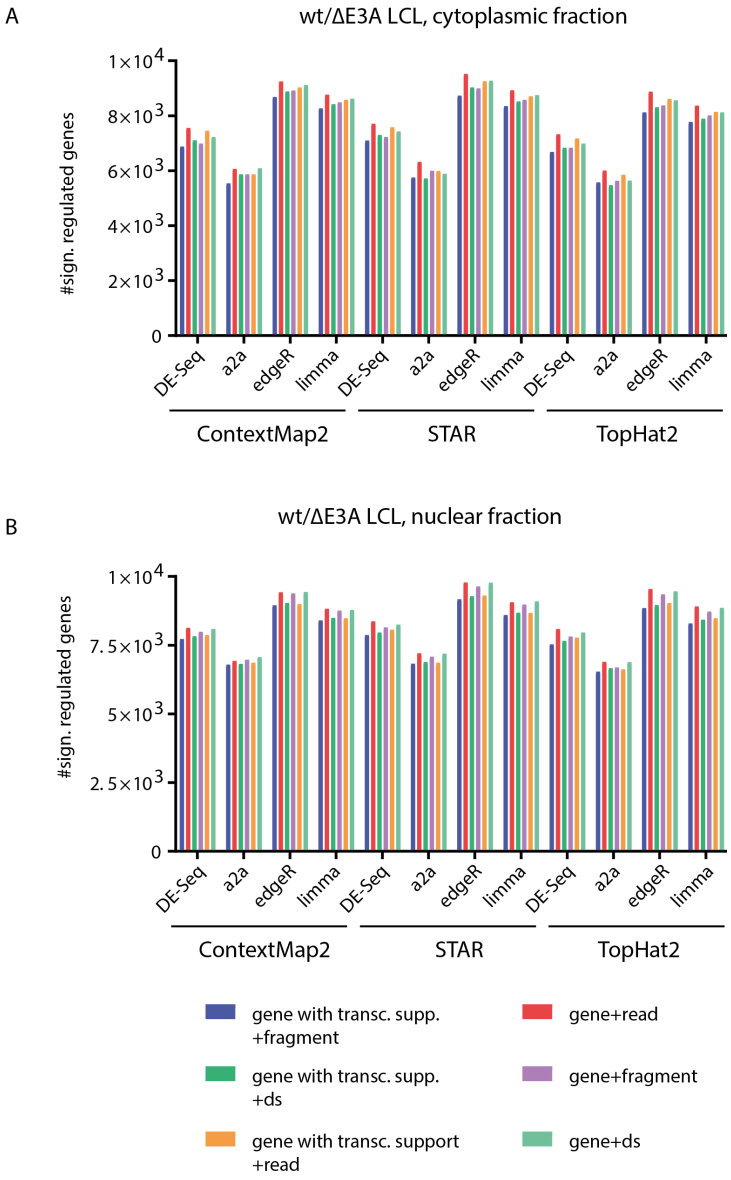


Figure 25: Different setups leading to similar results regarding the number (#) of significant regulated genes in E3A cell system. Bar graphs displaying the number of significantly regulated genes in the different setups for **A** regulation by E3A (wt/ΔE3A LCL) in the cytoplasm **B** regulation by E3A (wt/ΔE3A LCL) in the nucleus. Gene with transc. supp.: gene with transcript support. Ds: downsampling (for detailed explanation, see section 2.5.6.1, p. 29). Graph Pad Prism was used for plotting.

3.2.2.2.5 Protein coding and non-coding genes are regulated by E2 and E3A

In the following, the common set of genes, consistently detected by the most beneficial setups, are described. The number of detected genes is dependent on the cutoffs/thresholds for read coverage, fold change (FC) and significance (FDR).

To further increase the reliability in the detection of differential expressed genes, we agreed to choose the worst FDR and the worst fold change detected by one of the setups for all following analyses. The worst FDR corresponds to the FDR, which is the worst called FDR by one of the setups and the worst log2FC corresponds to the log2FC closest to zero detected by one of the setups. The result of the selection was exemplified for the ENSEMBL genes. Using the cutoffs FDR ≤ 0.05 and log2FC ≥ 1 or ≤ -1, 48 setups for the cytoplasmic ER/EB2-5 samples - /+ estr. resulted in the consistent detection of 6,524 ENSEMBL genes (Figure 26A), all 72 setups for the nuclear ER/EB2-5 samples - /+ estr. led to the consistent detection of 5,360 ENSEMBL genes (Figure 26B) and 18 setups for nuclear samples of ER/EB2-5 cells ChX-estr. / ChX+estr led to the consistent detection of 262 ENSEMBL genes (Figure 26C). An additional cutoff was introduced, only genes, which were covered with > 20 reads in one of the compared conditions were considered. This cutoff selected genes which were sufficiently covered to study regulation. The number of reads chosen for this cutoff was the result of RT-qPCR confirmation of candidate genes. Genes with 20 reads could be repeatedly detected by RT-qPCR.

The ENSEMBL annotation categorizes genes and transcripts into different biotypes, which can be grouped into protein coding, pseudogene, long non-coding and short non-coding. Long non-coding genes can be subclassified into, for instance, 3' overlapping ncRNA, antisense RNA, lincRNA, sense intronic RNA or sense overlapping RNA. In the following, these last-mentioned biotypes were filtered and termed lncRNAs, the remaining biotypes were pooled (residual biotypes). RNAs smaller than 200 bp were lost to a great extent during the RNA preparations (kit retains only RNAs > 200bp according to the manufacturer), so we excluded small RNAs (< 200 bp) from downstream analyses. In the cytoplasm, application of strengthened cutoffs (read counts > 20, FDR < 0.05 and log2FC >1 or < -1) resulted in the isolation of 5,578 genes that are significantly differentially expressed, including 999 lncRNAs. In the nucleus, 4,562 genes are significantly differentially expressed, including 646 lncRNAs. 226 genes are significantly differentially expressed in the nucleus when *de novo* protein synthesis is blocked, including 55 lncRNAs (Figure 27). For further analysis, the cutoff for the log2FC of ENSEMBL genes was set to > 0.85 or < -0.85 in ER/EB2-5 samples, since the variance was minor and the majority of known target genes of E2 could be detected with these cutoffs.

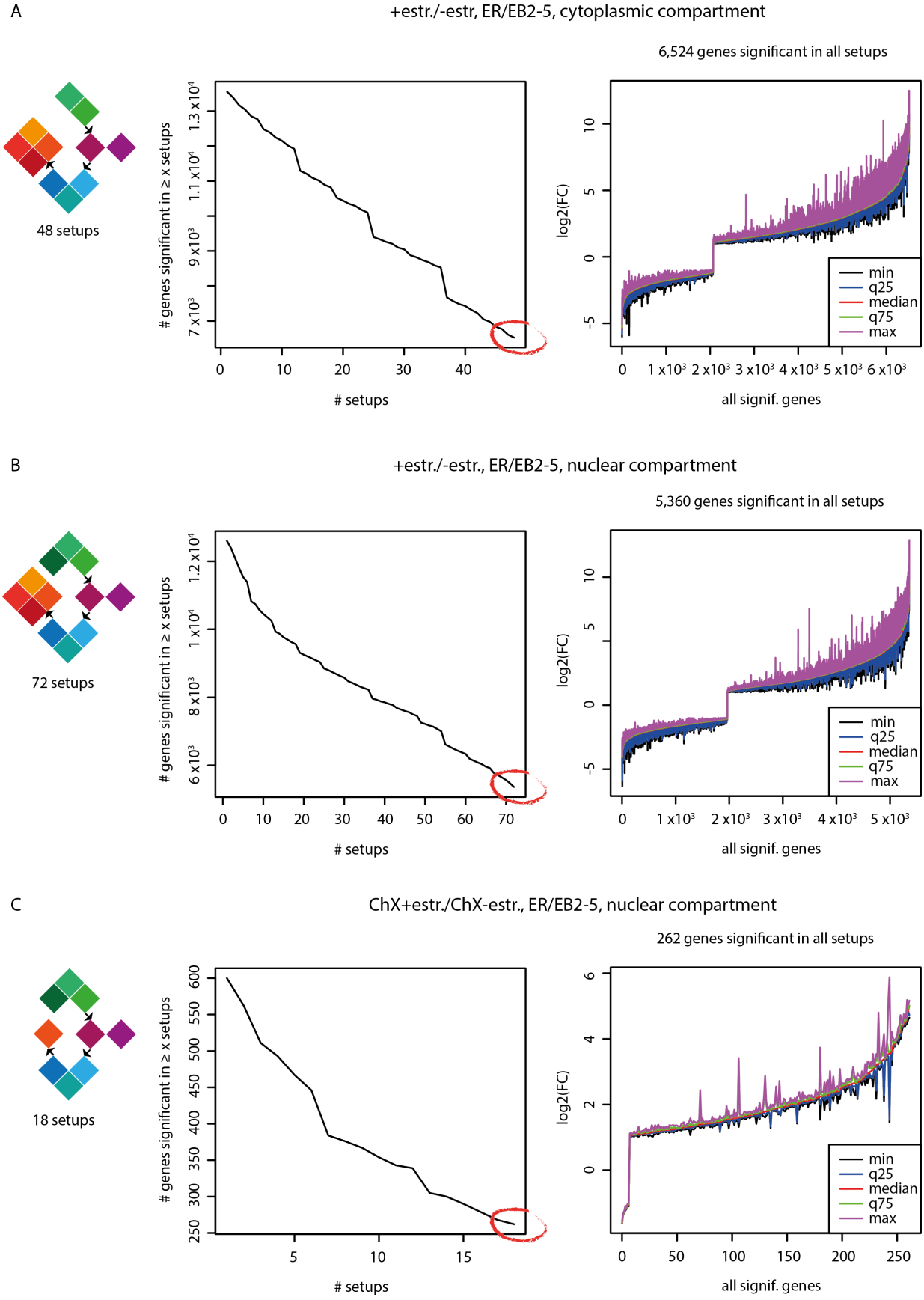


Figure 26: Downstream, consistently by the different setups detected E2-dependently regulated ENSEMBL genes with were analyzed. The miniaturized 4D-matrix indicates the combinatorics for the particular regulation analysis. The left panel shows a cumulative plot with the number of significantly (FDR ≤ 0.05)

regulated (log2FC ≥ 1 or ≤ -1) genes in ≥ x setups and the right panel shows a plot with the log2FCs of the amount of genes which were consistently detected by all setups. **A** Regulation by E2 (+estr. /-estr.) in the cytoplasmic compartment of ER/EB2-5: 6,524 genes were consistently detected by 48 setups. **B** Regulation by E2 (+estr. /-estr.) in the nucleic compartment of ER/EB2-5: 5,360 genes were consistently detected by 72 setups. **C** Regulation by E2 (+estr./-estr.) in the nucleic compartment of ER/EB2-5 in absence of *de novo* protein synthesis: 262 genes were consistently detected by 18 setups (plots were created by Gergely Csaba).

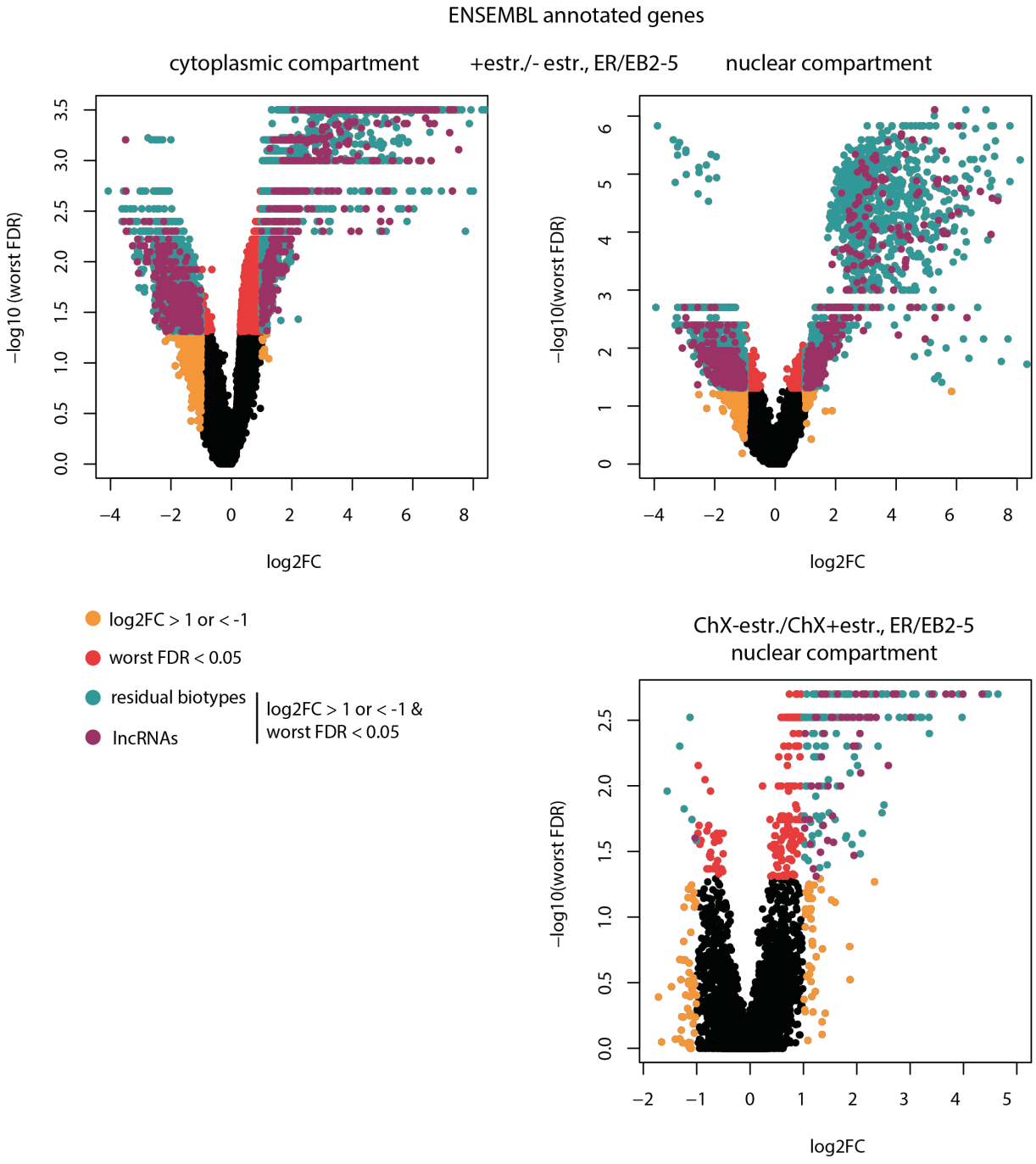


Figure 27: E2-dependent regulation of ENSEMBL annotated genes. Volcano plots displaying the log2FC vs the worst common FDR detected by all reasonable (48 for cytoplasmic samples, 72 for nucleic samples and 18 for nucleic ChX-treated samples) setups for E2 regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all biotypes without lncRNAs with the cutoff for FDR < 0.05 and log2FC > 1 or < -1, purple= all lncRNAs with the cutoff for FDR< 0.05 and log2FC > 1 or < -1. This definition resulted in 5,578 regulated genes in the cytoplasm including 999 lncRNA genes (upper left), 4,562 regulated genes in the nucleus including 656 lncRNA genes (upper right) and 226 regulated

genes in the nucleus in absence of *de novo* protein synthesis including 55 lncRNA genes (lower right). For plotting, filtered gene lists were input for R.

In addition to the ENSEMBL annotated genes, according to our strategy, differentially regulated genes were identified in intergenic (Figure 28) and intronic (Figure 29) regions. Similarly, the common amount of genes consistently detected by all 72 setups was further analyzed in the following (process not shown). As these regions could still be a byproduct of transcription of gene-overlapping regions, we corrected for noise by comparing the changes to the changes of the overlapped gene in the differential analysis (=depcorr.). The same cutoffs were applied as for the ENSEMBL genes. In the cytoplasm, these thresholds resulted in 133 intergenic and no intronic genes regulated significantly. In the nucleus, 483 intergenic and 8 intronic genes are regulated significantly. 10 intergenic and no intronic genes are regulated significantly in the nucleus when *de novo* protein synthesis is blocked.

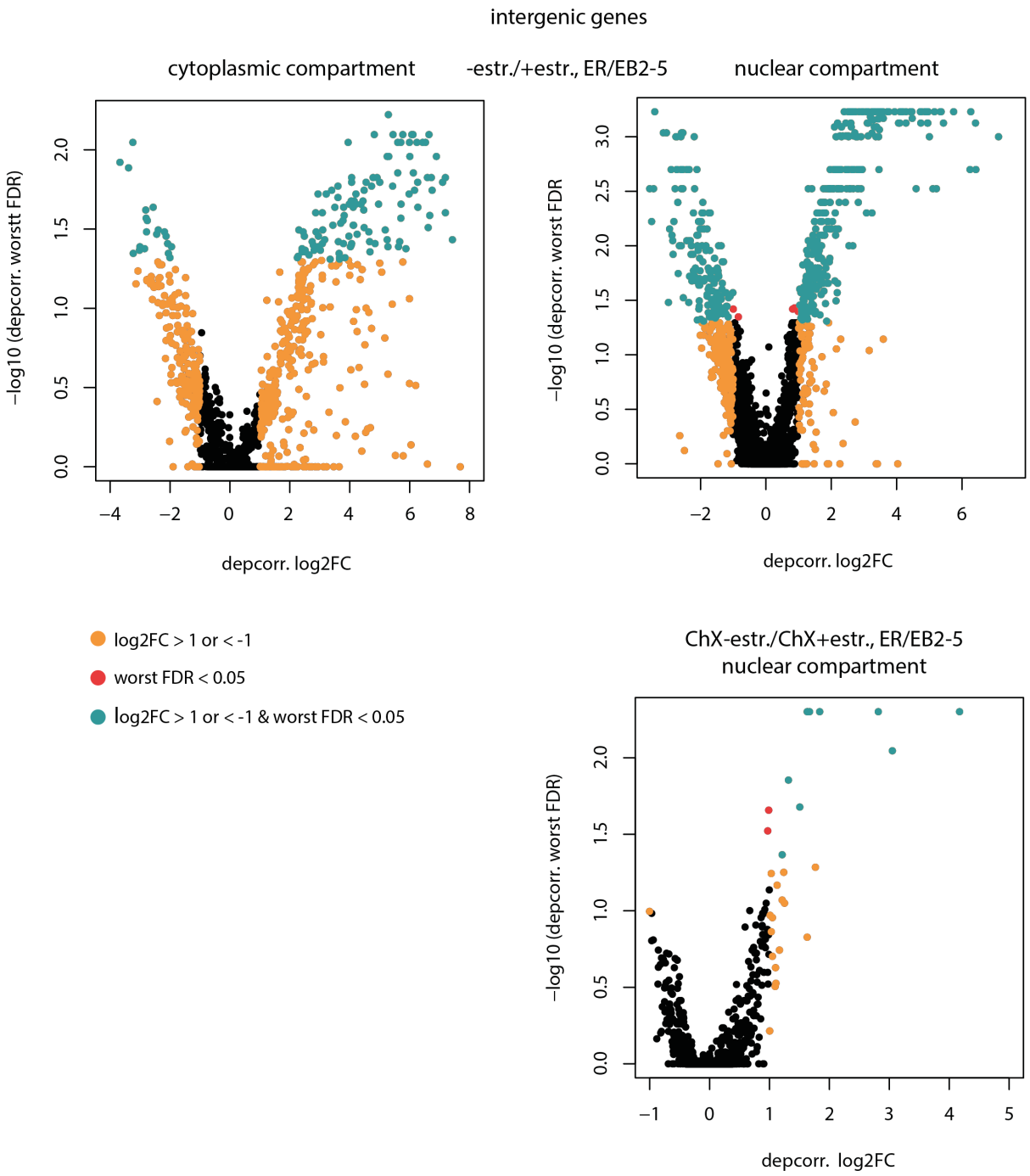


Figure 28: E2-dependent regulation of intergenic transcription. Volcano plots displaying the corrected log2FC (depcorr. = noise corrected log2FC) vs the worst common FDR detected by all setups for E2 regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all genes with the cutoff for FDR < 0.05 and log2FC > 1 or < -1. This definition resulted in 133 regulated genes in the cytoplasm (upper left), 483 regulated genes in the nucleus (upper right) and ten regulated genes in the nucleus in absence of *de novo* protein synthesis (lower right). For plotting, filtered gene lists were input for R.

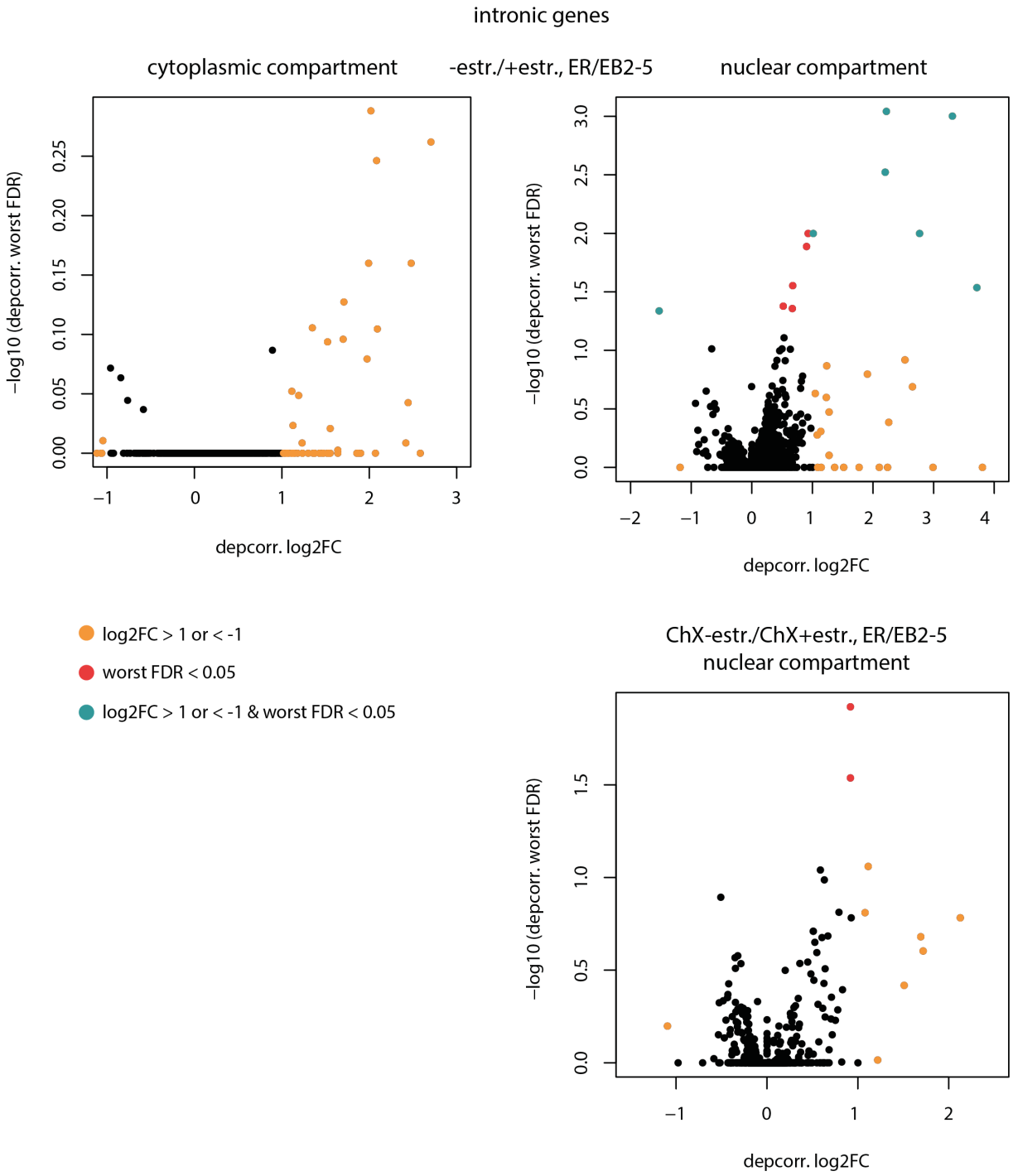


Figure 29: E2-dependent regulation of intronic transcription. Volcano plots displaying the corrected log2FC (depcorr. = noise corrected log2FC) vs the worst common FDR detected by all setups for E2 regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all genes with the cutoff for FDR < 0.05 and log2FC > 1 or < -1. This definition resulted in zero regulated genes in the cytoplasm (upper left), eight regulated genes in the nucleus (upper right) and zero regulated genes in the nucleus in absence of *de novo* protein synthesis (lower right). For plotting, filtered gene lists were input for R.

The result of the selection on the E3A cell system was exemplified for the ENSEMBL genes. Using the cutoffs FDR ≤ 0.05 and log2FC ≥ 1 or ≤ -1, all 72 setups for the cytoplasmic wt/ΔE3A cell samples resulted in the consistent detection of 3,200 ENSEMBL genes (Figure 31A) and all 72 setups for the nuclear wt/ΔE3A cell samples lead to the consistent detection of 3,783 ENSEMBL genes (Figure 30B). In the cytoplasm, application of strengthen cutoffs (read counts > 20, FDR < 0.05 and log2FC >1 or < -1) resulted in 2,788 significantly regulated genes, this includes 518 lncRNAs. In the nucleus, 3,367 genes are regulated significantly, comprising 785 lncRNAs (Figure 31).

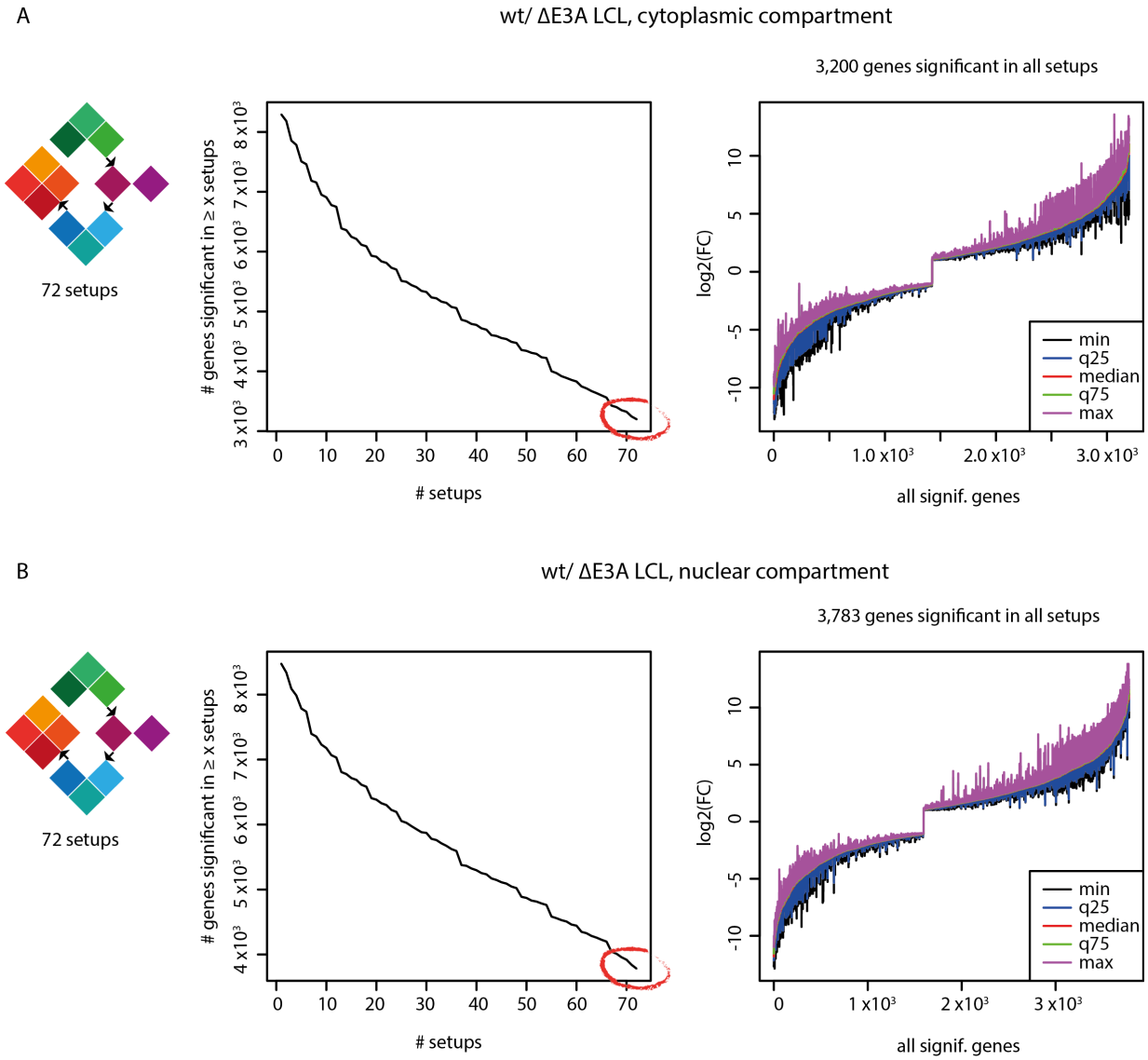


Figure 30: Downstream, consistently by the different setups detected E3A-dependently regulated ENSEMBL genes were analyzed. The miniaturized 4D-matrix indicates the combinatorics for the particular regulation analysis. The left panel shows a cumulative plot with the number of significantly (FDR ≤ 0.05) regulated (log2FC ≥ 1 or ≤ -1) genes in ≥ x setups and the right panel shows a plot with the log2FCs of the amount of genes which were consistently detected by all setups. **A** Regulation by E3A (wt/ΔE3A LCL) in the cytoplasmic compartment of LCLs: 3,200 genes were consistently detected by 72 setups. **B** Regulation by E3A (wt/ΔE3A

LCL) in the nucleic compartment of LCLs: 3,783 genes were consistently detected by 72 setups (plots were created by Gergely Csaba).

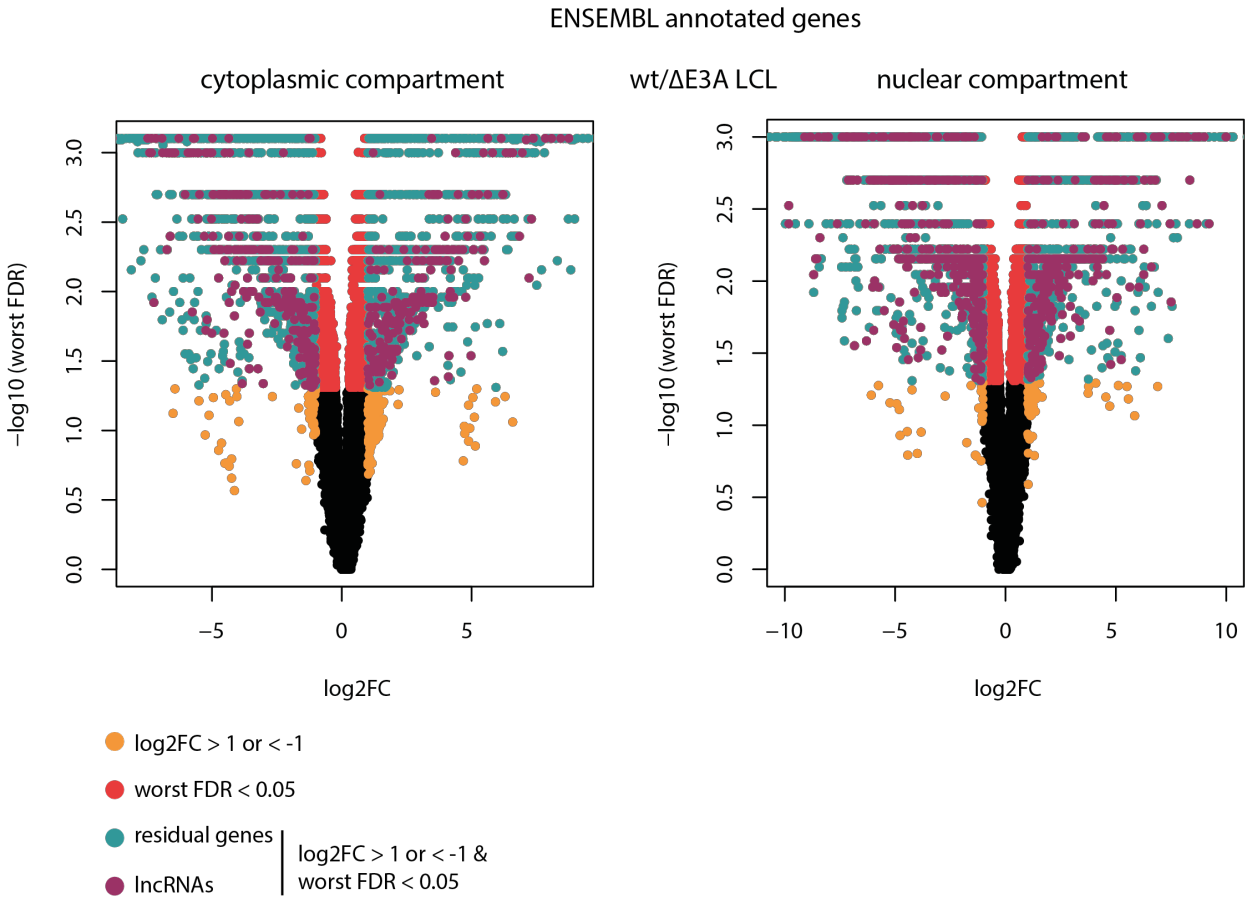


Figure 31: E3A-dependent regulation of ENSEMBL annotated genes. Volcano plots displaying the log2FC vs the worst common FDR detected by all setups for E3A regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all biotypes without lncRNAs with the cutoff for FDR < 0.05 and log2FC > 1 or < -1, purple= all lncRNAs with the cutoff for FDR < 0.05 and log2FC > 1 or < -1. This definition resulted in 2,788 regulated genes in the cytoplasm including 518 lncRNA genes (left) and 3,367 regulated genes in the nucleus including 785 lncRNA genes (right). For plotting, filtered gene lists were input for R.

Equally to the E2 regulation, differentially regulated genes were identified in intergenic (Figure 32) and intronic (Figure 33) regions. In the cytoplasm, chosen thresholds resulted in 478 intergenic and 347 intronic significantly regulated genes. In the nucleus, 1,239 intergenic and 369 intronic genes are regulated significantly.

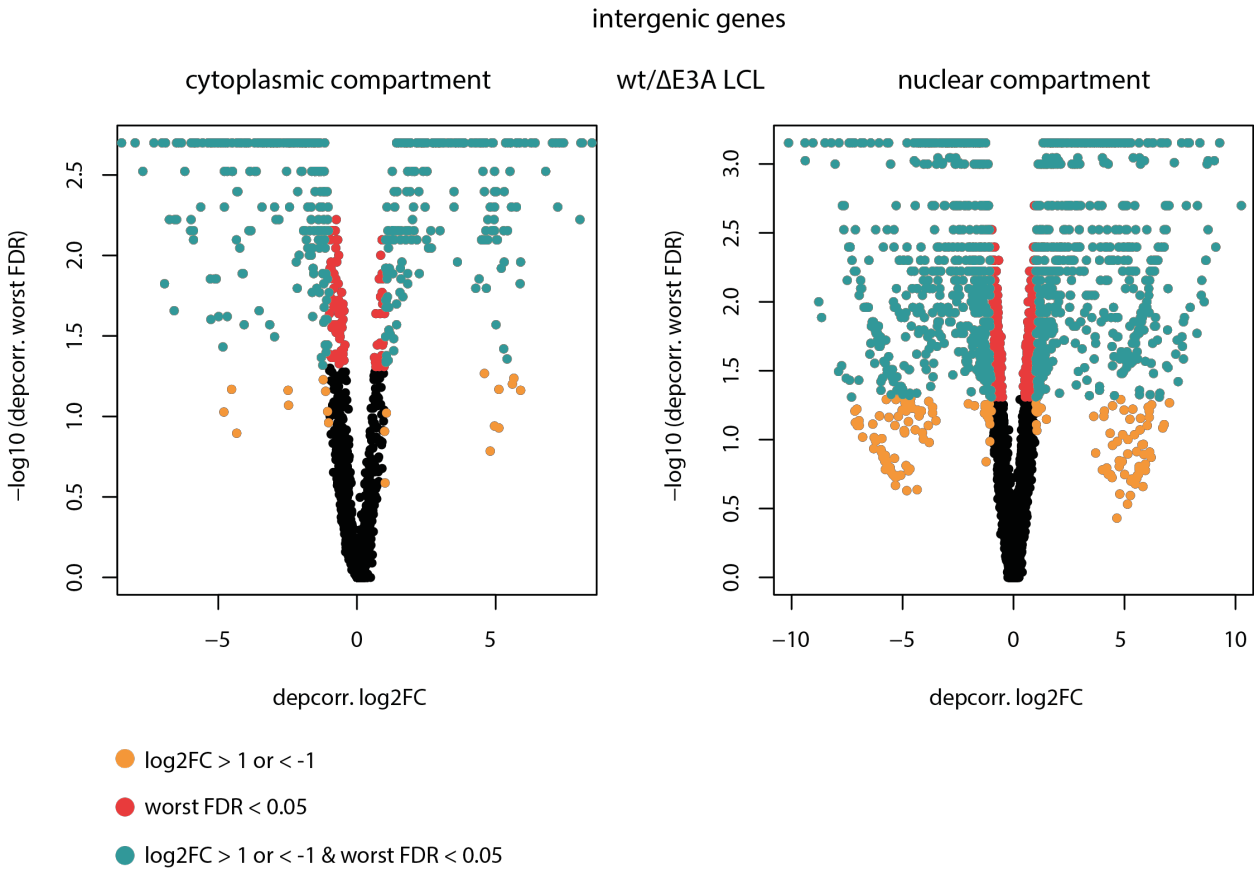


Figure 32: E3A-dependent regulation of intergenic transcription. Volcano plots displaying the corrected log2FC (depcorr. = noise corrected log2FC) vs the worst common FDR detected by all setups for E3A regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all genes with the cutoff for FDR < 0.05 and log2FC > 1 or < -1. This definition resulted in 478 regulated genes in the cytoplasm (left) and 1,239 regulated genes in the nucleus (right). For plotting, filtered gene lists were input for R.

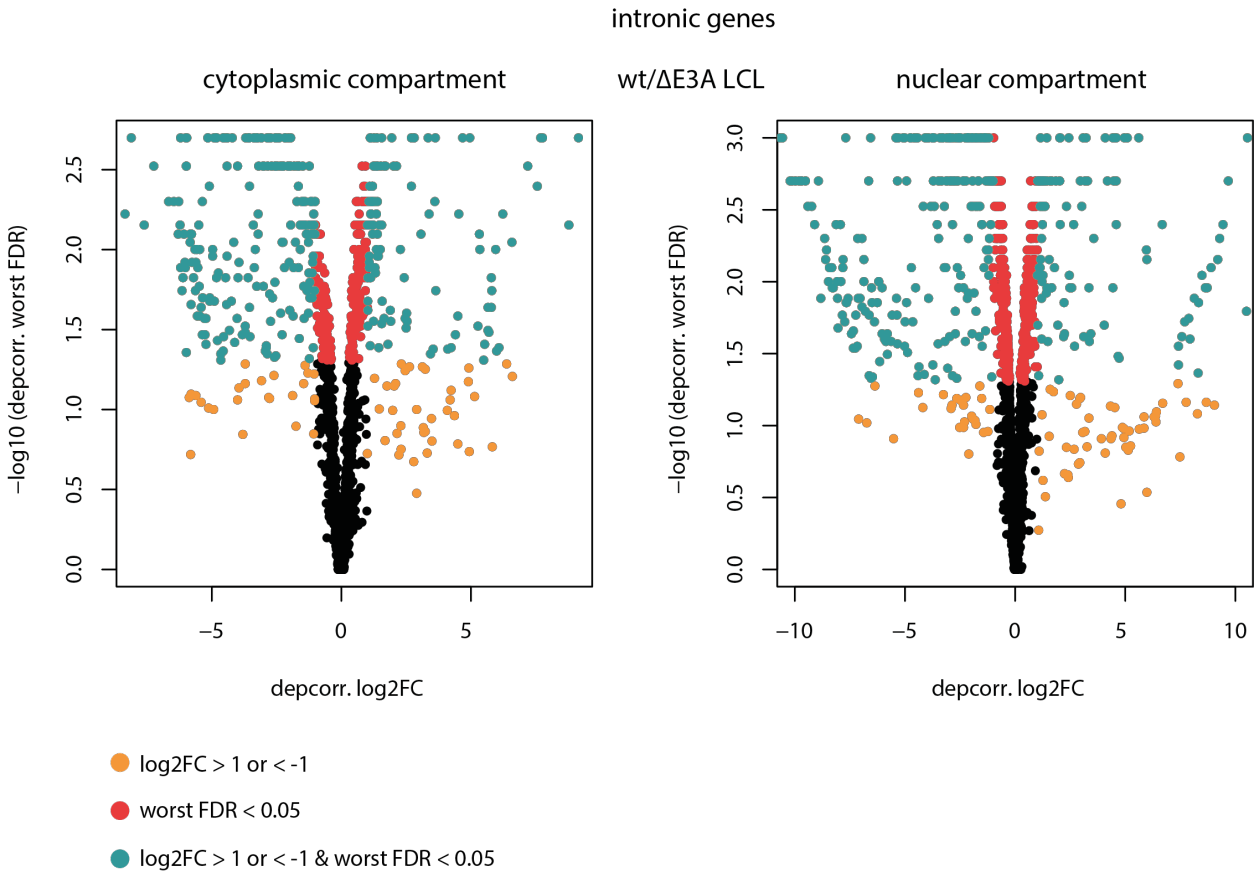


Figure 33: E3A regulation of intronic transcription. Volcano plots displaying the corrected log2FC (depcorr. = noise corrected log2FC) vs the worst common FDR detected by all setups for E3A regulated genes. Black= all genes with > 20 reads, orange= log2FC > 1 or < -1, red= cutoff for FDR < 0.05, green= all genes with the cutoff for FDR < 0.05 and log2FC > 1 or < -1. This definition resulted in 347 regulated genes in the cytoplasm (left) and 369 regulated genes in the nucleus (right). For plotting, filtered gene lists were input for R.

Taking the 20 best significant regulated genes in each condition pair (10 conditions, 45 condition pairs, 380 significant regulated genes) and performing a PCA (Principle Component Analysis; conducted by Gergely Csaba), it could be observed that the condition pairs cluster to some extent (Figure 34) together, despite the fact that the biological replicates show a quite high variation. Samples of same estrogen treatment cluster together (Figure 34A) and samples of same estrogen and ChX treatment cluster together (Figure 34B). ER/EB2-5 cells +estr. cluster to wt LCLs, confirming, that these cell lines express similar genes and cells depleted for E2 separate in opposite direction to cells depleted for E3A (Figure 34C, D). This result demonstrates that samples can be separated based on their regulated genes and cluster according to their biological affiliation.

In conclusion, we were able to detect differentially expressed annotated protein-coding and lncRNA genes as well as genes, so far not incorporated in the ENSEMBL annotation which were regulated dependently on E2 and E3A.

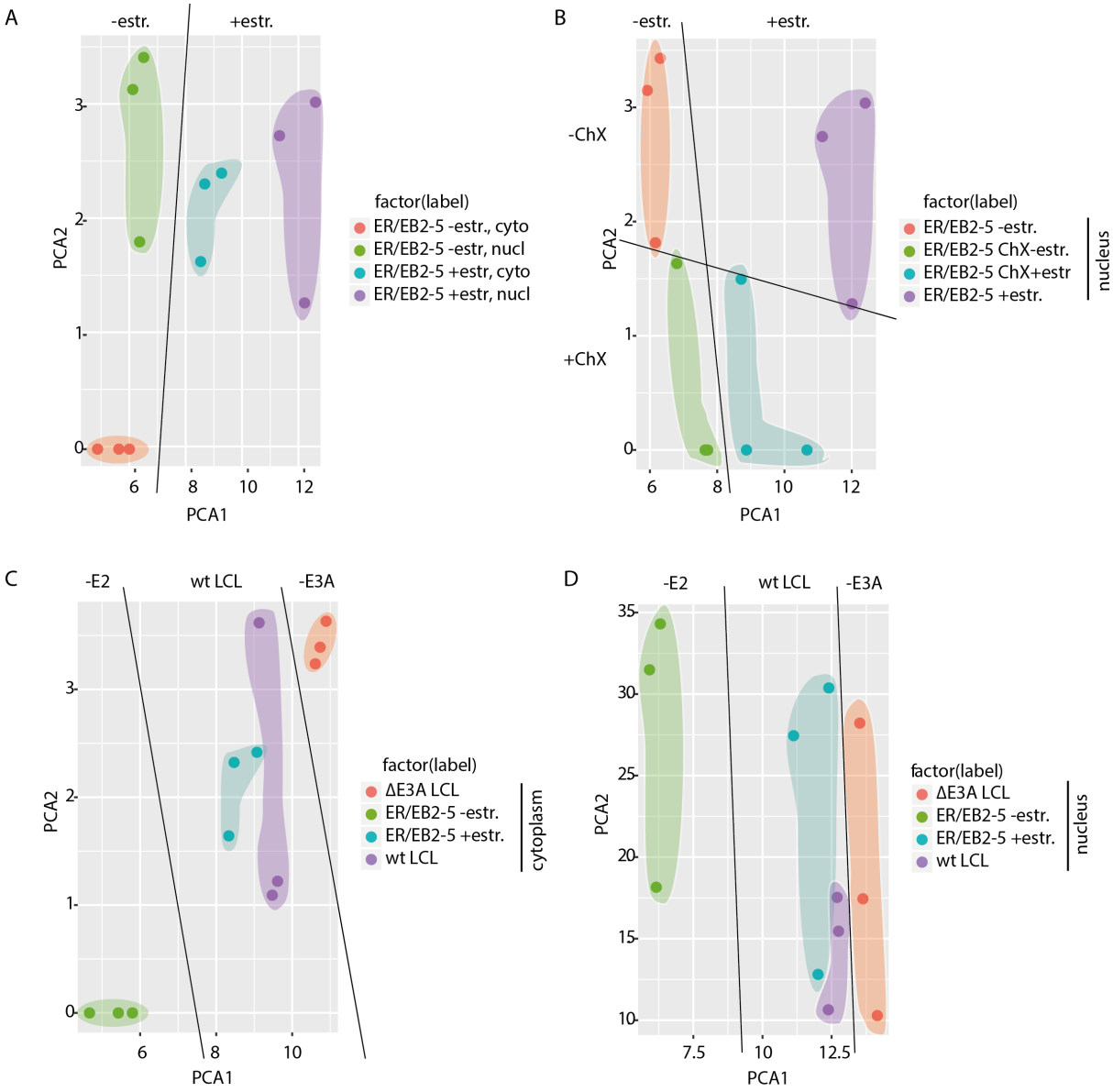


Figure 34: PCA based on the 20 best (highest log2FCs) significantly (FDR ≤ 0.05) regulated genes in each condition pair showing how the conditions behave towards each other. A Conditions without estrogen treatment group together on the left and conditions with estrogen treatment group together on the right; the compartment has no effect. **B** Conditions without estrogen treatment group on the left and conditions with estrogen treatment group on the right, while the ChX treated samples group at the bottom (nucleic compartment). **C** Samples of wt LCLs group together with samples of ER/EB2-5 LCLs while samples of cells without active E2 group on the left bottom and samples of cells without E3A group on the right top in the cytoplasm. **D** Samples of wt LCLs group together with samples of ER/EB2-5 LCLs while samples of cells without active E2 group on the left top and samples of cells without E3A group on the right bottom in the nucleus (PCA was performed by Gergely Csaba; for plotting, results were input for R; black lines and color fields without mathematical importance, only for visual isolation of groups).

3.2.2.2.6 EBV regulates its target genes in gene blocks

Genes within a contact domain or TAD are co-regulated, either they are active or repressed. We sought to recognize co-regulated gene blocks by E2 or E3A.

We noticed that the set of identified cellular target genes of E2 contained to a great extend co-regulated genes which were in physical proximity regarding their genomic position. To investigate, whether this is coincidence or can be observed genome-wide, Gergely Csaba tested all significantly (FDR ≤ 0.05) E2 (- ChX) regulated (log2FC ≥ 0.85) genes versus randomly selected regulated genes for clustering in blocks (Figure 35). A co-regulated gene block (CRGB) was defined by a genomic region of genes regulated in the same direction (or not regulated) and had to contain ≥ 2 genes. The borders of a CRGB were defined by a gene regulated in the opposite direction. There were no genomic size limitations for the CRGB. We excluded mitochondrial genes, chromosomes with less than 5 regulated genes from the analysis and EBV genes. Investigating the length of CRGBs, we found that 80 % of all E2 regulated CRGB were ≤ 2 Mb (in both, cytoplasm and nucleus) and 70 % of all E3A regulated blocks were ≤ 2 Mb (in both, cytoplasm and nucleus; Fig. S19). TADs are defined to be up to 2 Mb in size, this was used as a benchmark (Dixon et al., 2012). For the cytoplasmic compartment for example, around 75% of the involved 6,132 E2 target genes can be found in blocks of up to ten genes (Figure 35 upper left). More genes were clustered in blocks in the cytoplasm compared to the nucleus (Figure 35 upper panel). For E3A block regulation can also be observed (Figure 35 lower panel). Rao *et al*. published 2014 so called contact domains for GM12878. They claimed to investigate with the highest resolution, however, in their investigation TAD-defined boundaries were not detected. They came up with smaller contact domains than TADs (Rao et al., 2014; see section 1.1.4, p. 10). Intersecting 2,038 CRGB regulated by E2 in the nucleus with the 9,273 contact domains published, we found 33 % (673) of CRGBs completely overlapping with the published contact domains (not shown). Thus, the majority of CRGB are only partially overlapping with contact domains.

These data indicate that EBV regulates its target genes in blocks of various sizes of consecutive similar regulated genes.

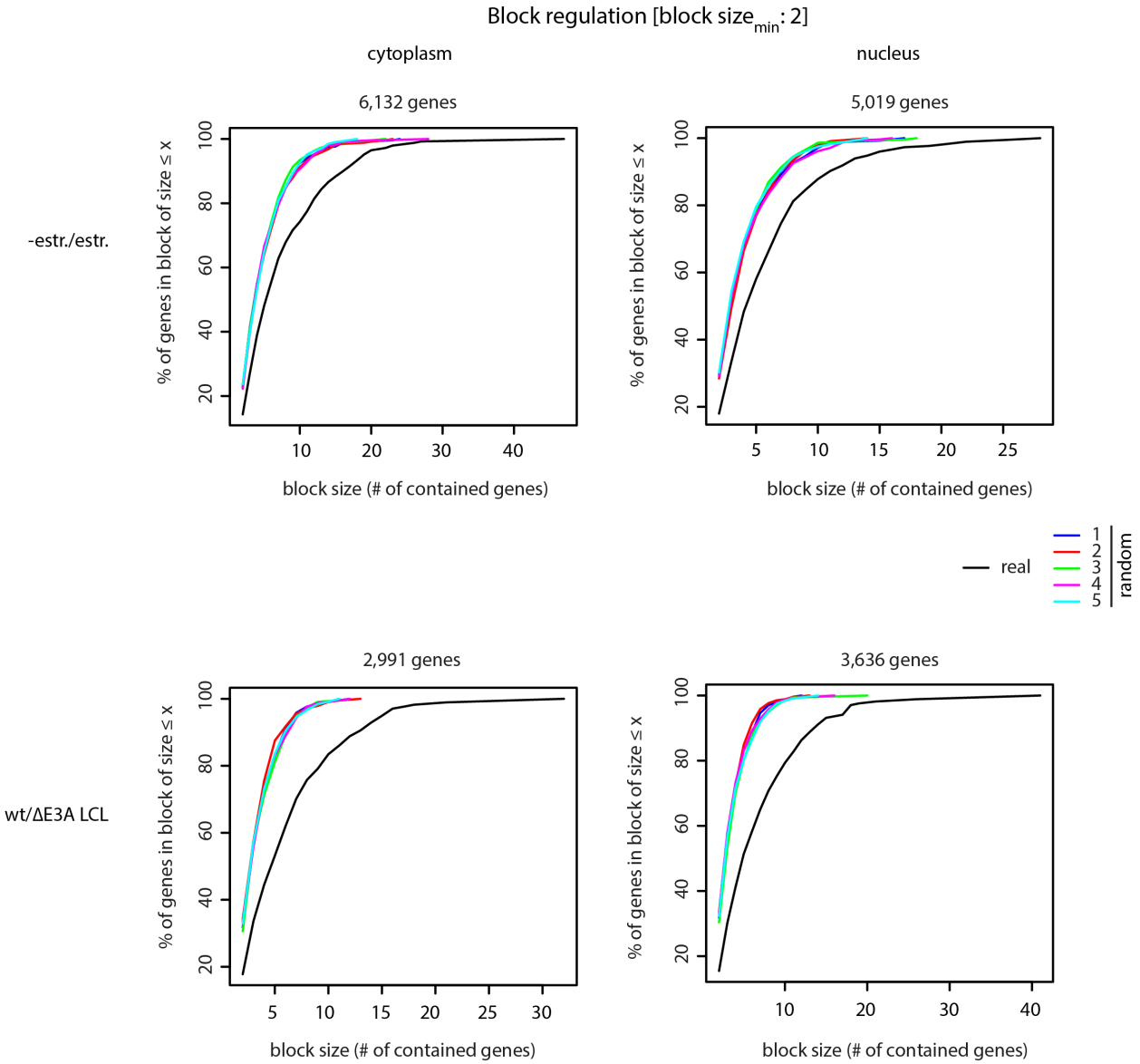


Figure 35: E2 regulates its target genes block wise. Cumulative plots displaying the % of genes in block size of ≤ x, minimum genes per block = 2 compared to 5 runs of randomly selected genes. 6,132 genes were involved in the block regulation by E2 in the cytoplasm (upper left), 5,019 genes were involved in the block regulation by E2 in the nucleus (upper right), 2,991 genes were involved in the block regulation by E3A in the cytoplasm (lower left) and 3,636 genes were involved in the block regulation by E3A in the nucleus (lower right); mtDNA (mitochondrial DNA) and chromosomes containing < 5 regulated genes are excluded. For all comparisons of real to random, p-values (by Kolmogorov–Smirnov test) = 0 (plots were created by Gergely Csaba using genes with FDR ≤ 0.05 and log2FC ≥ 0.85).

3.2.2.2.7 Regulated blocks of cellular target genes consist of protein coding and non-coding genes with cancer links

Next, we aimed to confirm candidate target blocks by RT-qPCR and intended to characterize these blocks by integration of public available data on LCLs.

Three of these E2 target blocks in the neighborhood of *MYC* (Figure 36, Figure 37), *SLAMF1* (Figure 40) and *PPAN* (Figure 42) were investigated further by integrating public available data for the LCL GM12878 included in the ENCODE project and confirmed by RT-qPCR. These loci were chosen since they all contain reviewed protein coding E2 target genes (described in the following). All three candidate loci are characterized by the presence of at least two genes being induced by E2 in cytoplasm, the nucleus, or both and being regulated in absence of *de novo* protein synthesis, implicatory for a "direct" induction. Furthermore, the blocks are characterized by locally increased GRO-Seq signals indicating ongoing transcription, active chromatin marks like H3K27ac, H3K4me1, increased PolII binding indicatory for active transcription and local CTCF binding sites signaling possible TAD boundaries (coverage tracks obtained by ChIP-Seq data for H3K27ac, H3K4me1, RNA PolII and CTCF were sourced as BigWig files from ENCODE). In fact, TADs were described for these loci (as indicated), indicating that EBV potentially regulates its targets TAD-wide. Additionally, DNA-loops were published for GM12878 obtained by capture Hi-C experiments, connecting either two fragments containing each a promoter (promoter of genes annotated by Ref-Seq gene annotation) or a promoter-fragment with a non-promoter-fragment (="other"; Mifsud et al., 2015). Worthy of note here is that there are loops generated in GM12878 at these loci and all/most of the loops are connecting fragments of the surrounding with *MYC, SLAMF1* or *PPAN*. Furthermore, all these loci contain at least one annotated lncRNA that is co-regulated together with protein coding genes. Intriguingly, these lncRNAs identified by RNA-Seq as E2 targets can all be linked to cancer as discussed in the following section.

MYC

MYC was defined as a direct target of E2 (Kaiser et al., 1999; Figure 36, Figure 37). *MYC* (*C-MYC*) is a proto-oncogene which encodes a transcription factor. MYC responsive genes are involved in almost every important cellular function. Most importantly, MYC is involved in the regulation of various growth-promoting signal transduction pathways. Interestingly, almost all cancer-associated genetic changes in *MYC* are linked to non-coding regulatory regions (Miller, Thomas, Islam, Muench, & Sedoris, 2012). E2 is published to activate multiple *MYC* enhancers and to rearrange the *MYC* locus by hijacking long-range enhancer-promoter loops (Wood et al., 2016; Figure 36). Moreover, two bidirectionally transcribed eRNAs were already described to be targets of E2 and to influence *MYC* transcription (Liang et al., 2016; Fig 29. at the bottom). This group used short hairpin RNA (shRNA)-mediated knock down to investigate the impact of those eRNAs on *MYC.* shRNA target sites and RT-qPCR primer products reside in *CASC19/21*. In the environment of *MYC*, several non-coding RNA genes were detected as E2 targets by RNA-Seq in this study, four of them at the 3' of the TSS of *MYC, PCAT1* (two transcript variants (tvs) according to GRCh37.75; both multiexonic; sense transcripts), *CASC8* (four tvs according to GRCh37.75; all multiexonic; antisense transcripts), *CASC19* (one transcript according to GRCh37.75; multiexonic; antisense transcript) and *CASC21* (= *RP11-382A18.2*; one transcript according to GRCh37.75; multiexonic; sense transcript), and two of them 5' of the TSS of *MYC, LINC00977* (one transcript according to GRCh37.75; multiexonic; antisense transcript) and *CCDC26* (four tvs according to GRCh37.75; all multiexonic; antisense transcripts). The eRNAs published by Liang *et al*. reside in the introns of *CASC19* and *CASC21* (Figure 37). The CASC (CAncer suSCeptibility) genes are related to pancreatic cancer (Wolpin et al., 2014). *PCAT1* is implicated in disease progression of prostate cancer and is claimed to control HRR (homologous recombination repair) in all sorts of cancer (Prensner et al., 2011). *CCDC26* contributes to tumorigenesis in pancreatic cancer (Peng & Jiang, 2016) and is dysregulated in AML (Hirano et al., 2015). *LINC00977* is also linked to pancreatic cancer (Wolpin et al., 2014). In fact, the region 5' of *MYC* was identified to contain multiple cancer susceptibility loci (Grisanzio & Freedman, 2010; Huppi, Pitt, Wahlberg, & Caplen, 2012; Wolpin et al., 2014). All of these non-coding RNA genes are co-regulated together with *MYC* by E2 as confirmed by RT-qPCR (Figure 38, Figure 39). Their spliced transcripts could be detected in the total, cytoplasmic or nucleic RNA preparation, - /+ ChX. For the transcript of the non-coding gene LINC00977, unfortunately no exon-exon-junction primer could be established for RT-qPCR. Nevertheless, the spliced transcript could be repeatedly detected by RT-PCR from cDNA templates of different RNA preparations (Figure S20). The log2FCS in RNA-Seq under ChX for this locus ranges from 2.03 to 5.01, thus, the targets are well regulated when *de novo* protein synthesis is blocked.

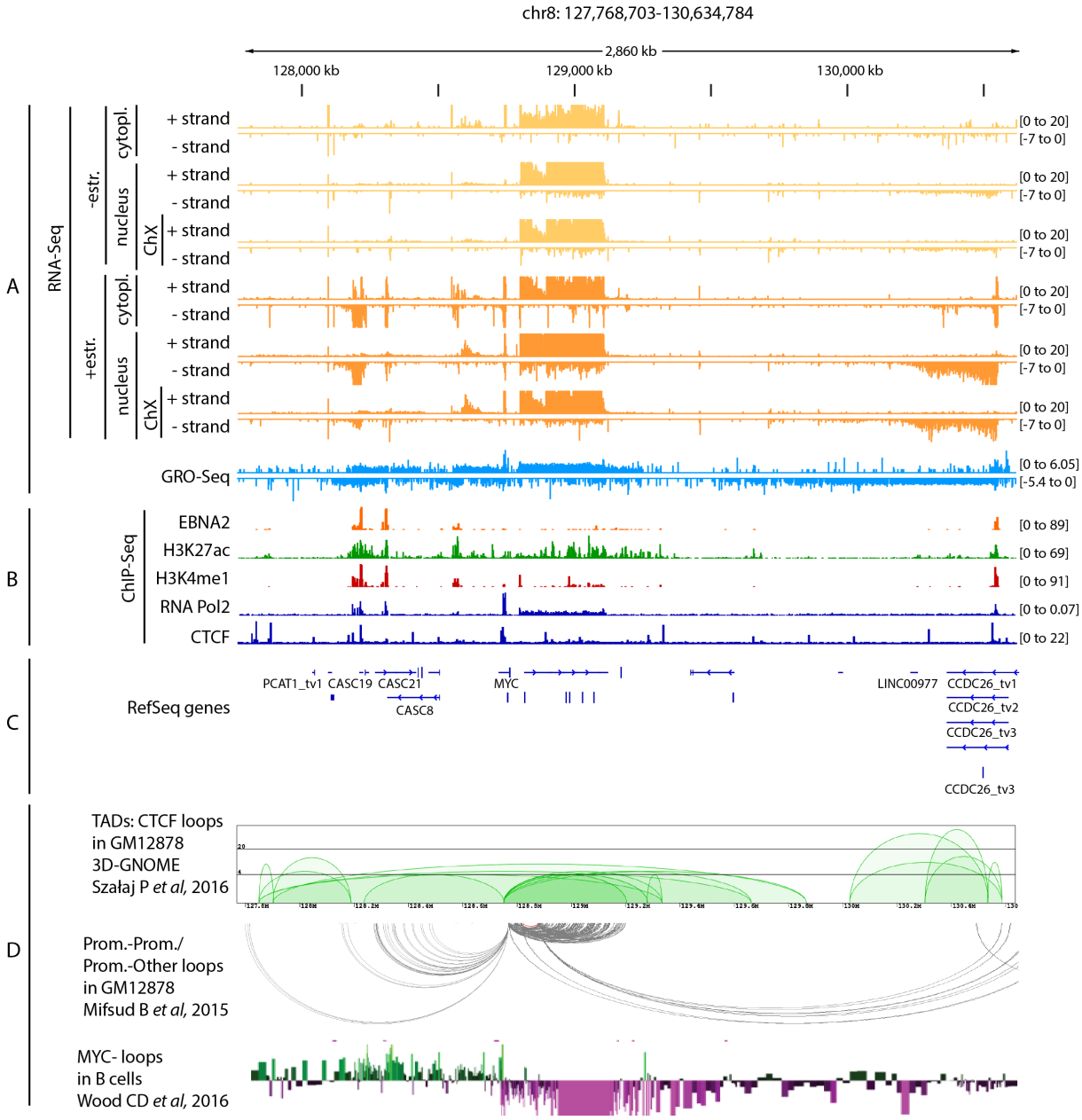


Figure 36: Overview of the *MYC* gene locus with pictured E2 dependent induction of transcription in ER/EB2-5 and references for active chromatin and looping activity in GM12878. Schematic map depicting **A** Expression based RNA-Seq tracks displaying the coverage in the different conditions/fractions as indicated (coverage encoded in bigwig files; converted from BAM files, mapping by STAR mapper; set to the indicated data range) and a GRO-Seq track showing the coverage in GM12878 (Core LJ, et al. 2014; data range set to auto scale). **B** Tracks obtained from ChIP-Seq (H3K27ac, H3K4me1, RNA PolII and CTCF sourced from ENCODE for GM12878; data range set to auto scale). **C** Annotation track (only names of relevant genes and different tvs specified). **D** Additionally displayed are CTCF-mediated TADs and promoter-to-promoter loops (grey= loops directing to E2 target genes) in GM12878 as well as MYC-loops in B cells, where -E2 4C-Seq reads were subtracted from +E2 4C-Seq reads (green= E2 induced loops, purple= E2 repressed loops).

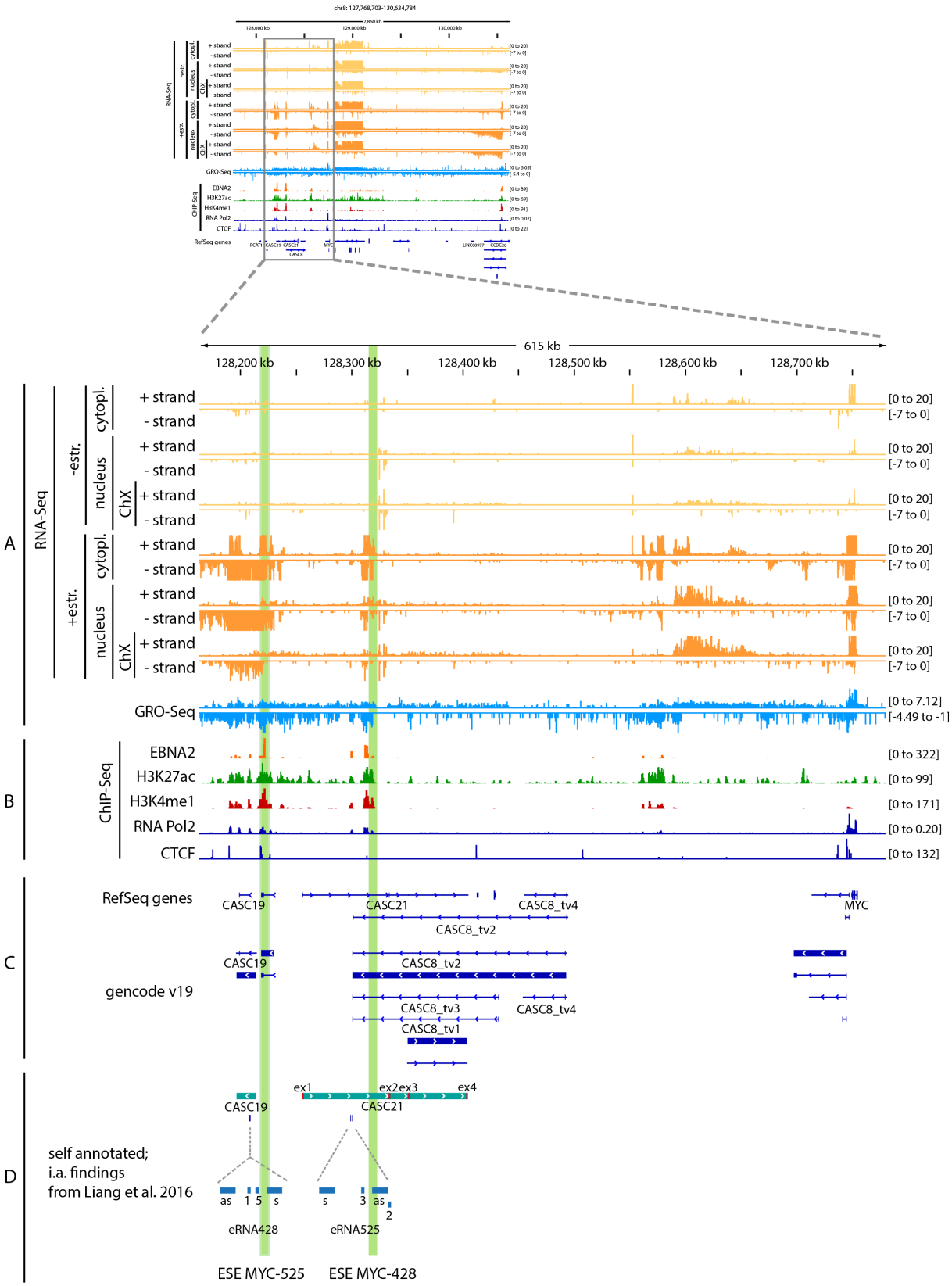


Figure 37: Overview of the region 3' of the TSS of *MYC* with pictured E2 dependent induction of transcription in ER/EB2-5 and references for active chromatin in GM12878. Schematic map depicting **A** Expression based RNA-Seq tracks (1-12) displaying the coverage in the different conditions/fractions as indicated (coverage encoded in bigwig files; converted from BAM files, mapping by STAR mapper; set to the indicated data range) and a GRO-Seq track showing the coverage in GM12878 (Core LJ, et al. 2014; data range set to auto scale). **B** Tracks obtained from ChIP-Seq (H3K27ac, H3K4me1, RNA PolII and CTCF sourced from ENCODE for

GM12878; data range set to auto scale). **C** Annotation tracks (only names of relevant genes and different tvs specified). **D** Annotation of published eRNAs and used shRNAs for knock down confirmation of these eRNAs (1, 5= shRNAs for knock down of eRNA428; 2, 3= shRNA for knock down of eRNA525; s= "sense" amplicon investigated by RT-qPCR; as= "antisense" amplicon investigated by RT-qPCR.

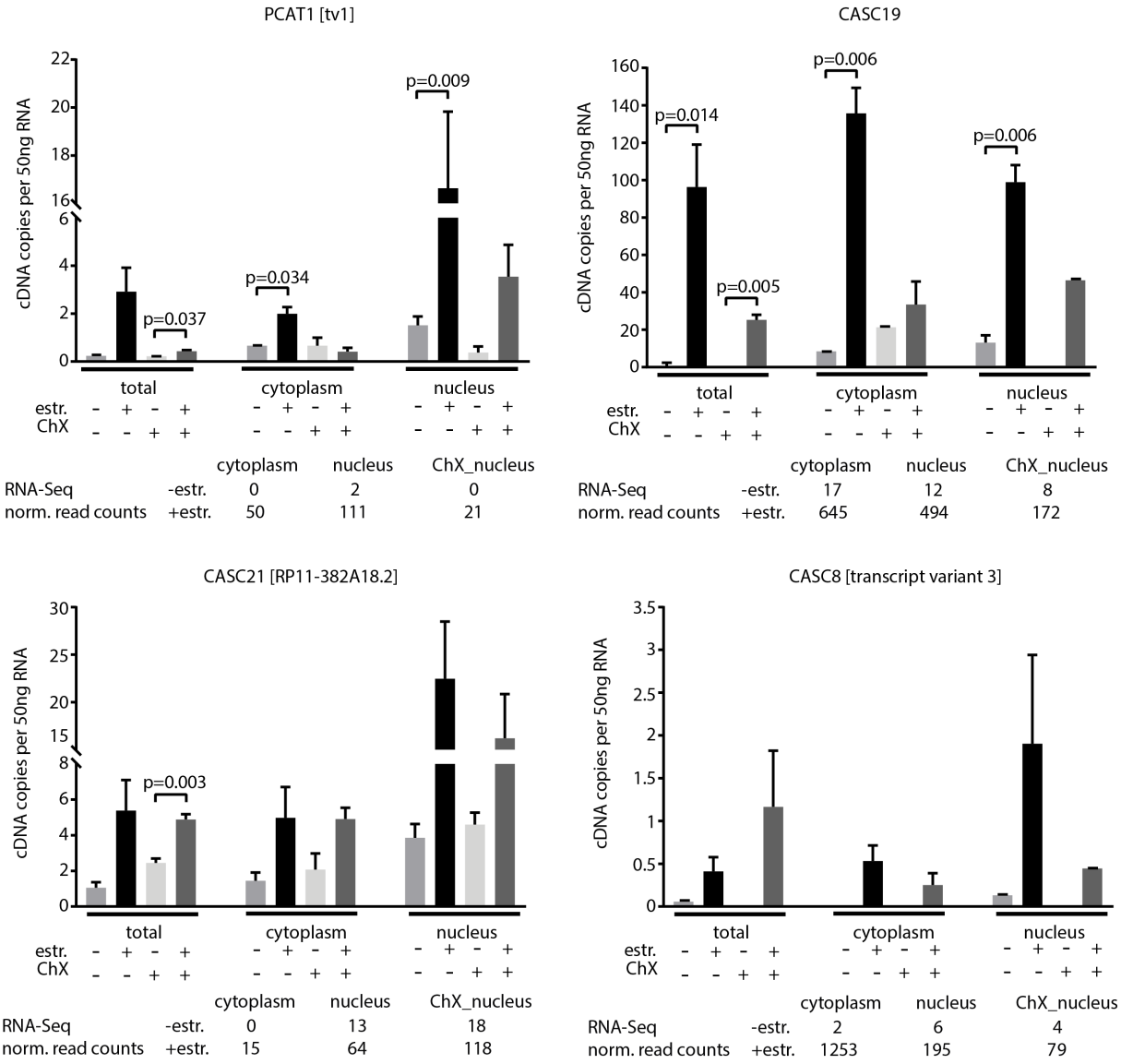


Figure 38: RT-qPCR confirmation of E2 target genes in the *MYC* neighborhood 3' of the TSS of *MYC*. Bar graphs showing the absolute quantification by RT-qPCR of four non-coding transcripts, PCAT1 (tv1), CASC19, CASC21 and CASC8 in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h, with a subpopulation under ChX treatment. RNA was isolated from 10^7 cells (total) or subcellular fractions (1.2x 10^7 cells for cytoplasm and 2x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.} = 3$, $n_{biol.} = 3$). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq aligned to ENSEMBL genes are displayed. Graph Pad Prism was used for plotting.

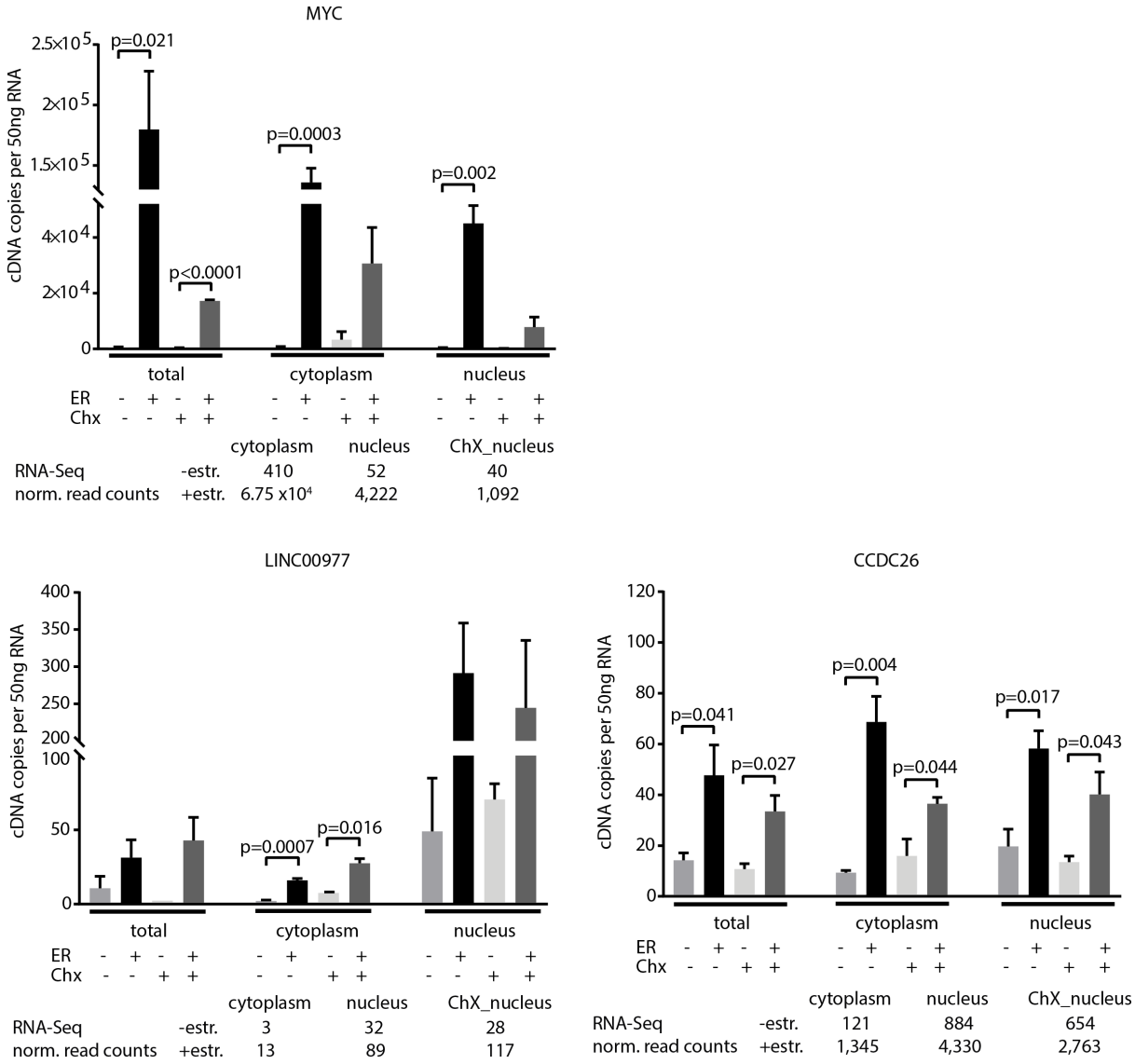


Figure 39: RT-qPCR confirmation of E2 target genes in the *MYC* neighborhood 5' of the TSS of *MYC*. Bar graphs showing the absolute quantification by RT-qPCR of MYC transcripts and the two non-coding transcripts LINC00977 and CCDC26 in different RNA preparations. The primer pair for CCDC26 detects three of four tvs. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h with a subpopulation under ChX treatment. RNA was isolated from 10^7 cells (total) or subcellular fractions (1.2 x 10^7 cells for cytoplasm and 2 x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.} = 3$, $n_{biol.} = 3$). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq aligned to ENSEMBL genes are displayed. Graph Pad Prism was used for plotting.

SLAMF1

SLAMF1 is also an already known target of E2 (Maier et al., 2006; Figure 40). SLAM-family receptors are well described in the literature to be expressed in diverse immune cell types. They contain two immunoglobulin-like domains in their extracellular region and exert important functions in immune cells (Veillette, 2006). SLAMF1 (CD150) cell surface molecule is upregulated in B cell lymphoma cell lines with type III EBV latency (Takeda, Kanbayashi, Kurata, Yoshiyama, & Komano, 2014). In SLAMF1's environment, the genes of two other cell surface molecules, *CD48* and *CD84*, belonging to the SLAM-family receptors (SLAMF2 and SLAMF5, respectively) and a non-coding RNA gene, *RP11-528G1.2*, were detected as E2 targets by RNA-Seq in this study. *RP11-528G1.2* (one transcript according to GRCh37.75; multiexonic; sense transcript) is transcribed antisense to CD84. It was observed to be misregulated in gastric cancer (Song et al., 2016). This non-coding RNA gene is co-regulated together with the three SLAM-family receptors by E2, confirmed by RT-qPCR (Figure 41). The spliced transcripts of the protein coding genes could be detected in the total, the cytoplasmic or the nucleic RNA preparation, - /+ ChX. For the transcript of the non-coding RP11-528G1.2, unfortunately no exon-exon-junction primer could be established for RT-qPCR. Nevertheless, the spliced transcript could be detected by RT-PCR from cDNA templates of different RNA preparations (Figure S20). The log2FCS in RNA-Seq under ChX for this locus ranges from 1.21 to 1.75, thus, the targets are weakly regulated when *de novo* protein synthesis is blocked.

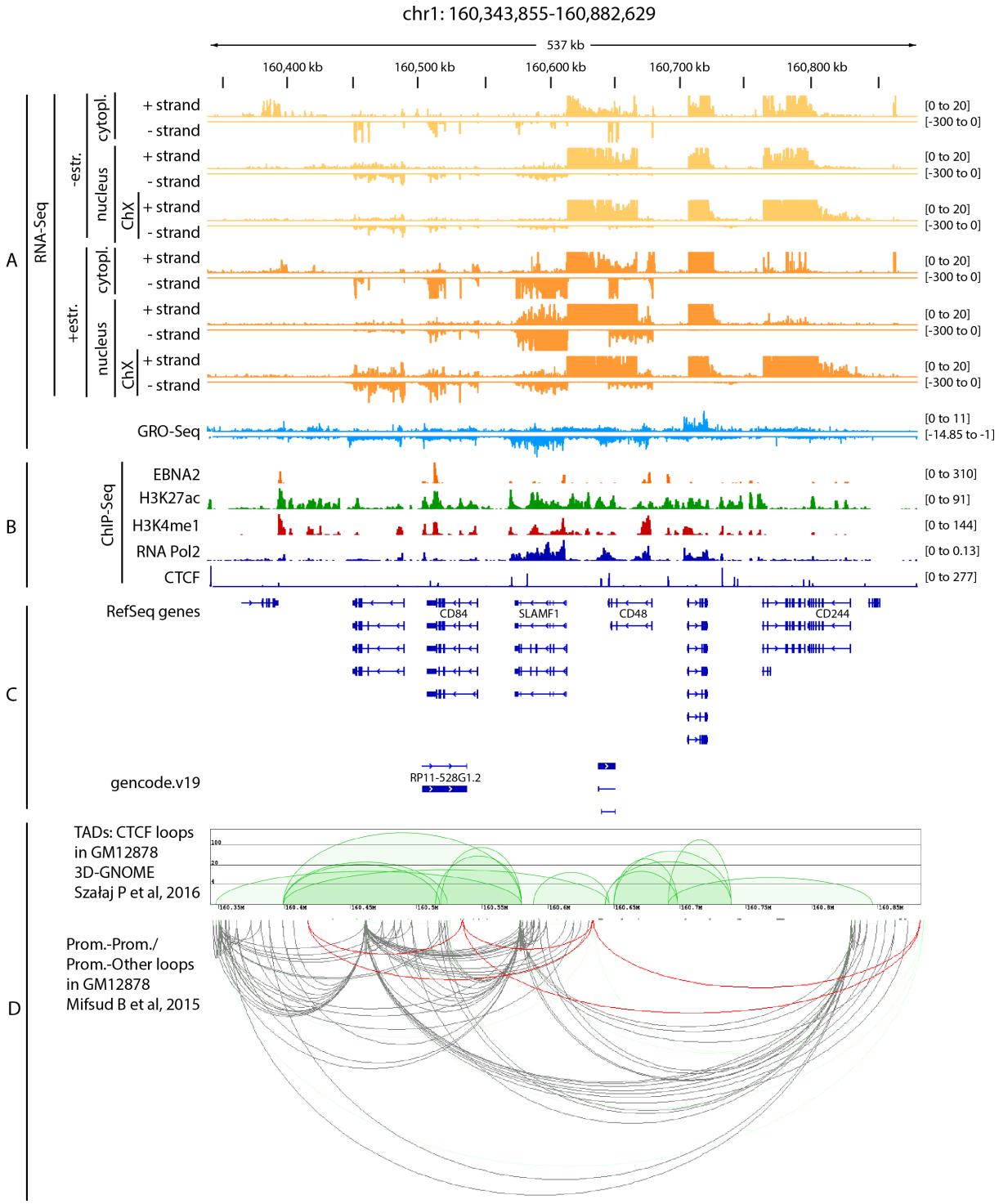


Figure 40: Overview of the *SLAMF* gene locus with pictured E2 dependent induction of transcription in ER/EB2-5 and references for active chromatin and looping activity in GM12878. Schematic map depicting **A** Expression based RNA-Seq tracks (1-12) displaying the coverage in the different conditions/fractions as indicated (coverage encoded in bigwig files; converted from BAM files, mapping by STAR mapper; set to the indicated data range) and a GRO-Seq track showing the coverage in GM12878 (Core LJ, et al. 2014; data range set to auto scale). **B** Tracks obtained from ChIP-Seq (H3K27ac, H3K4me1, RNA PolII and CTCF sourced from ENCODE for GM12878; data range set to auto scale). **C** Annotation tracks (only names of relevant genes and different transcript variants specified). **D** Additionally displayed are CTCF-mediated TADs and promoter-to-promoter loops (grey=loops directing to E2 target genes)/promoter-to-other loops (red= loops directing to E2 target gene) in GM12878.

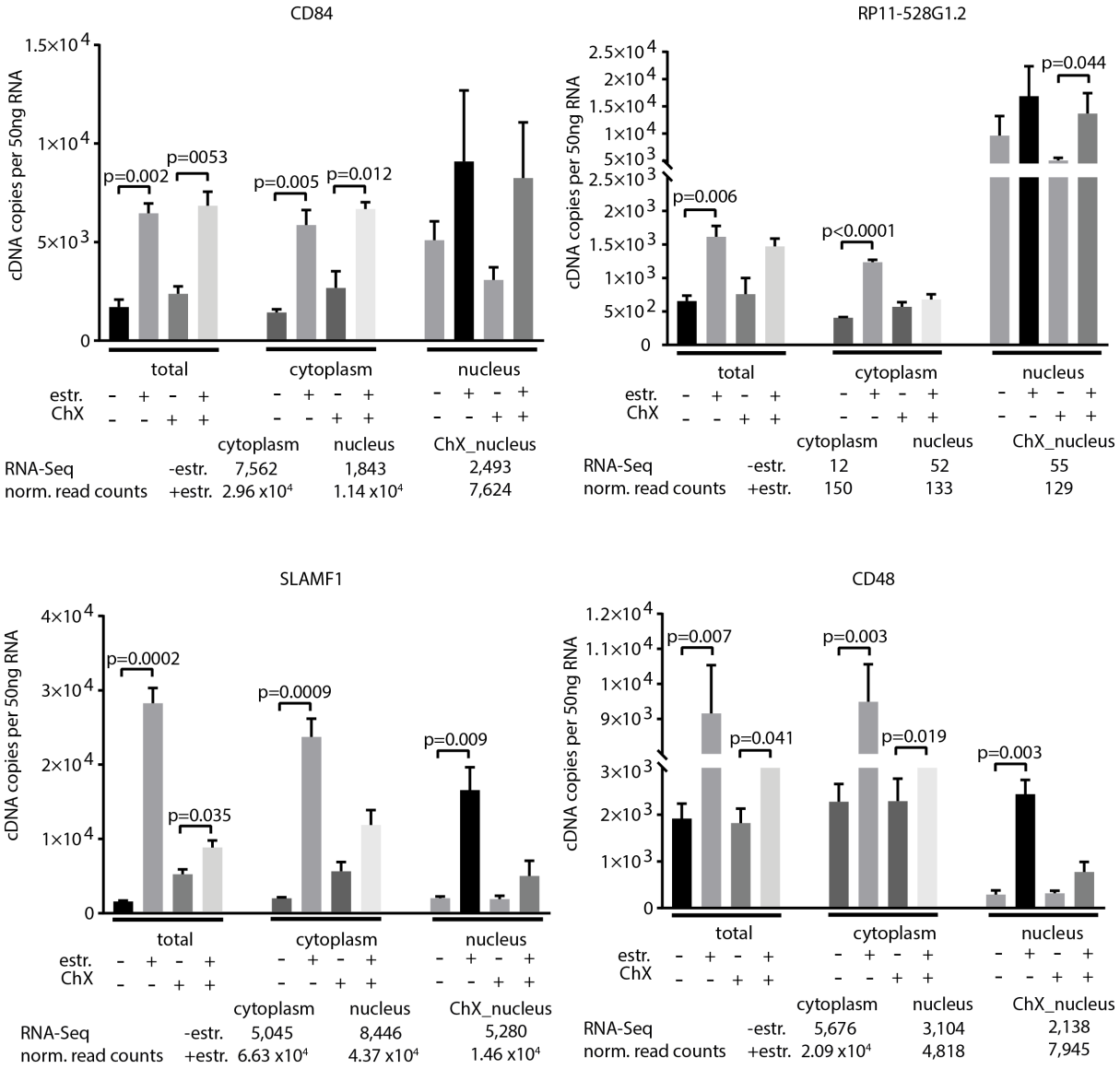


Figure 41: RT-qPCR confirmation of E2 target genes in the *SLAMF* neighborhood. Bar graphs showing the absolute quantification by RT-qPCR of transcripts of *CD48, SLAMF1, CD84* and the non-coding transcript RP11-528G1.2 in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h, with a subpopulation under ChX treatment. RNA was isolated from 10^7 cells (total) or subcellular fractions (1.2 x 10^7 cells for cytoplasm and 2 x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.} = 3$, $n_{biol.} = 3$). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq aligned to ENSEMBL genes are displayed. Graph Pad Prism was used for plotting.

PPAN

Peter Pan (PPAN) is a reported target of E2 (Spender et al., 2006; Figure 42). The neighborhood of *PPAN* contains genes with diverse functions in proliferation and immune response. *PPAN* encodes the Suppressor of SW14 1 (SSF1) homolog, which is a conserved protein (SSF1 in yeast and PPAN in drosophila]. It is published to be essential for cell growth and proliferation (Welch et al., 2000). *P2RY11* encodes a purinergic receptor and among other roles it is described to exert a role in immune response with cell type-specific effects (Vitiello, Gorini, Rosano, & la Sala, 2012). PPAN-P2RY11 is a read-through product and the fusion protein shares sequence identity with each individual gene product (provided by RefSeq, Nov 2010; https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=692312). In the neighborhood resides ANGPTL6 which functions as a chemokine enhancing proliferation in blood cells ($CD34^+$ cord blood cells; Fatrai et al., 2011). These protein coding genes were detected as E2 targets by RNA-Seq in this study together with a non-coding RNA gene *CTD-2240E14.4. CTD-2240E14.4* (one transcript according to GRCh37.75; monoexonic; antisense transcript) was found to be misregulated in 13 different cancer types (Yan et al., 2015). The read-through transcript PPAN-P2RY11, ANGPTL6 and the non-coding RNA CTD-2240E14.4 are co-regulated by E2 as confirmed by RT-qPCR (Figure 43). Specific primers for the detection of distinct transcripts of only P2RY11 or only the read-through transcript PPAN-P2RY11 could not be established for RT-qPCR. Specific primers for only PPAN transcripts are not feasible. The exon-exon junction primers used detect the spliced transcripts of both, PPAN and PPAN-P2RY11. The spliced transcripts of ANGPTL6 and the monoexonic transcript of CTD-2240E14.4 could be detected in the total, the cytoplasmic and the nucleic RNA preparation, - /+ ChX. However, the regulation of *ANGPTL6* observed by RT-qPCR appears to be not so convincing, particularly because it seems to be regulated by ChX. The log2FCS in RNA-Seq under ChX for this locus ranges from 1.12 to 1.45, thus, the targets are weakly regulated when *de novo* protein synthesis is blocked.

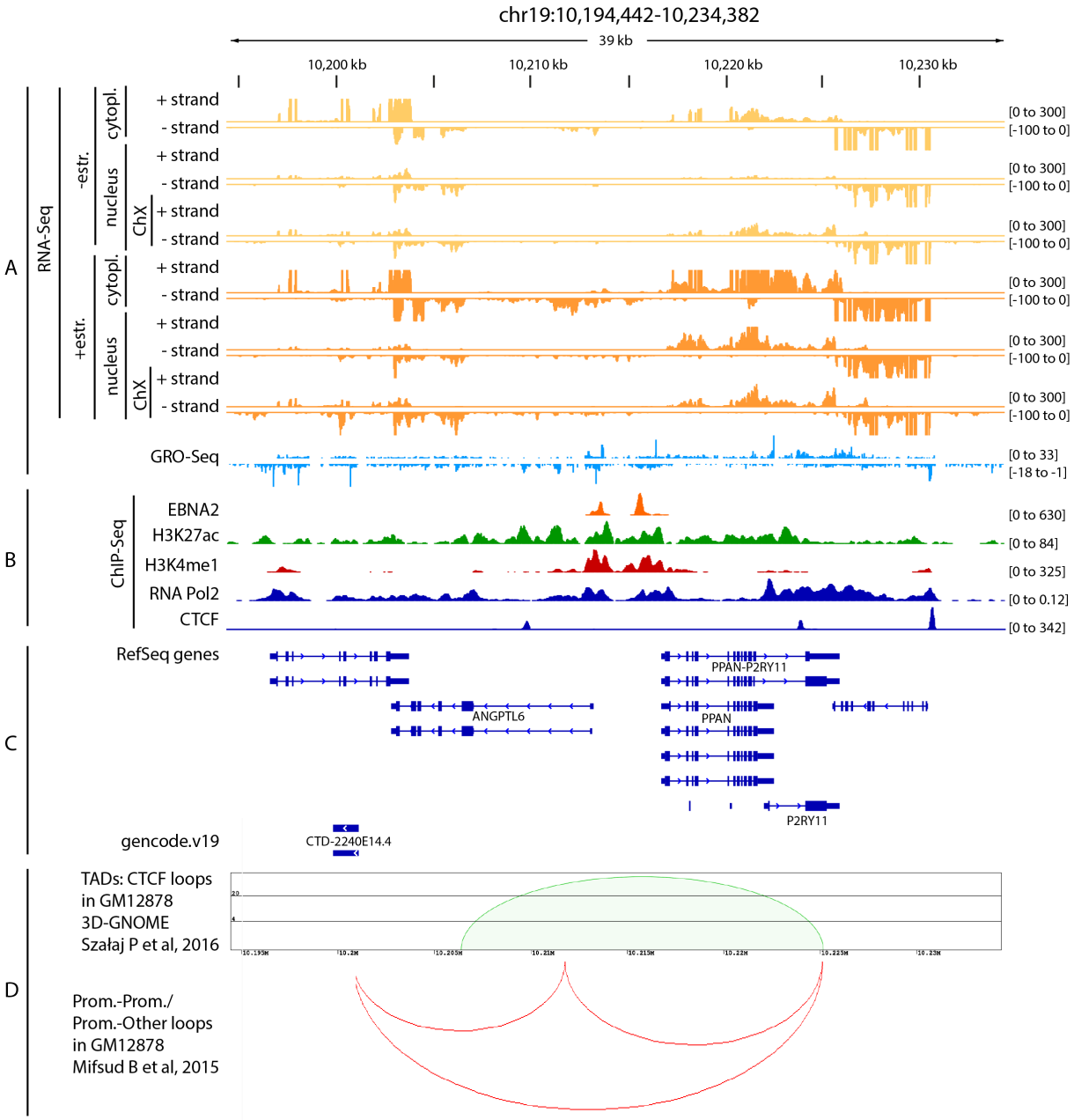


Figure 42: Overview of the *PPAN* gene locus with pictured E2 dependent induction of transcription in ER/EB2-5 and references for active chromatin and looping activity in GM12878. Schematic map depicting **A** Expression based RNA-Seq tracks (1-12) displaying the coverage in the different conditions/fractions as indicated (coverage encoded in bigwig files; converted from BAM files, mapping by STAR mapper; set to the indicated data range) and a GRO-Seq track showing the coverage in GM12878 (Core LJ, et al. 2014; data range set to auto scale). **B** Tracks obtained from ChIP-Seq (H3K27ac, H3K4me1, RNA PolII and CTCF sourced from ENCODE for GM12878; data range set to auto scale). **C** Annotation tracks (only names of relevant genes and different transcript variants specified). **D** Additionally displayed are CTCF-mediated TADs and promoter-to-other loops (red= loops directing to E2 target gene) in GM12878.

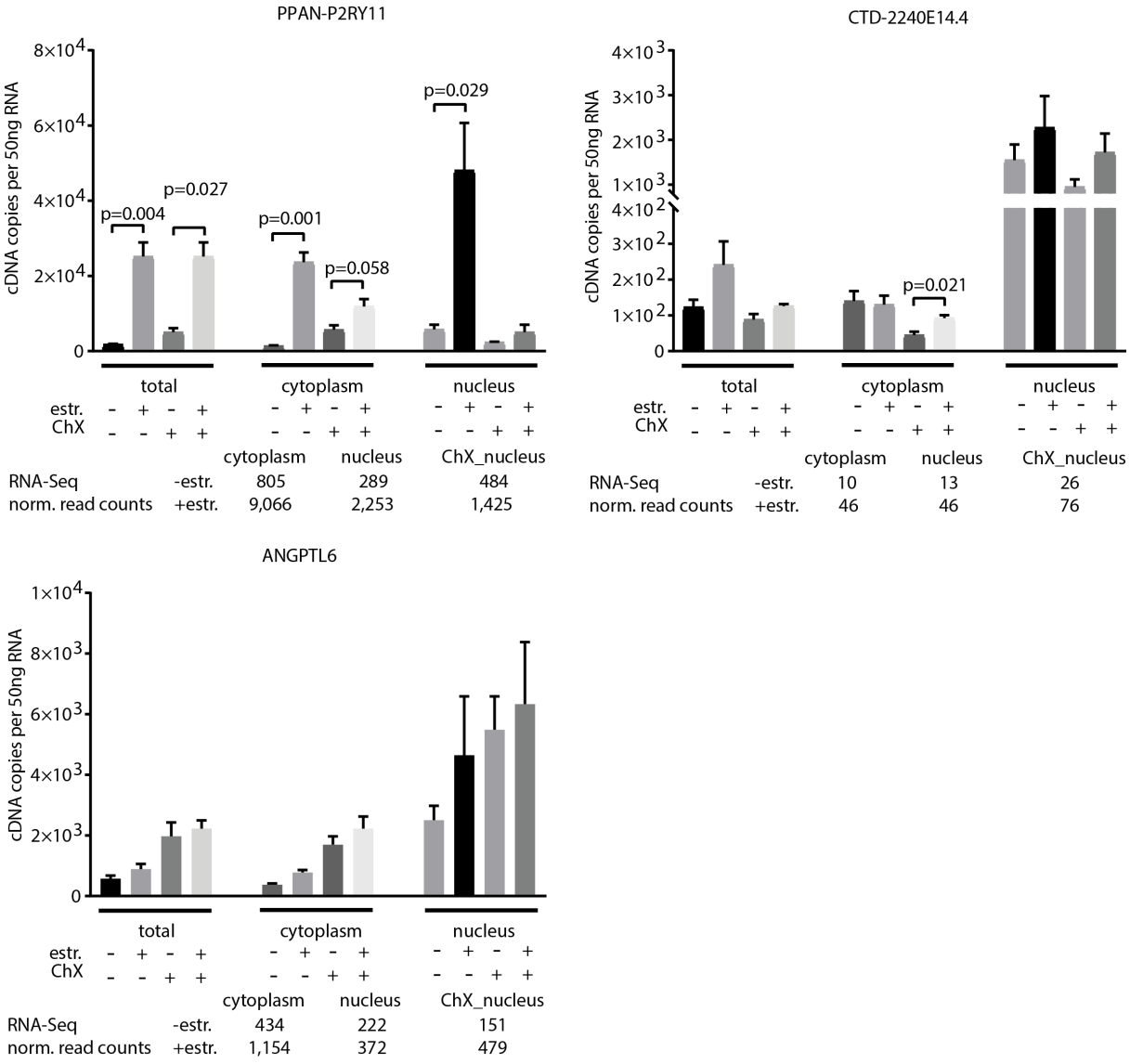


Figure 43: RT-qPCR confirmation of E2 target genes in the PPAN neighborhood. Bar graphs showing the absolute quantification by RT-qPCR of the PPAN-P2RY11 readthrough transcript, ANGPTL6 and the non-coding transcript CTD-2240E14.4 in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h, with a subpopulation under ChX treatment. RNA was isolated from 10^7 cells (total) or subcellular fractions (1.2 x 10^7 cells for cytoplasm and 2 x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.}$ =3, $n_{biol.}$ =3). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq aligned to ENSEMBL genes are displayed. Graph Pad Prism was used for plotting.

It was not necessary to perform a relative quantification, since a housekeeping (non-coding) gene did not show major regulation. However, the protein-coding housekeeping gene showed a slight regulation in the cytoplasm upon E2 reactivation, in contrast to the results of RNA-Sequencing (Figure 44).

Together, these RT-qPCR data support the RNA-Sequencing results on gene blocks regulated by E2 in presence and absence of *de novo* protein synthesis.

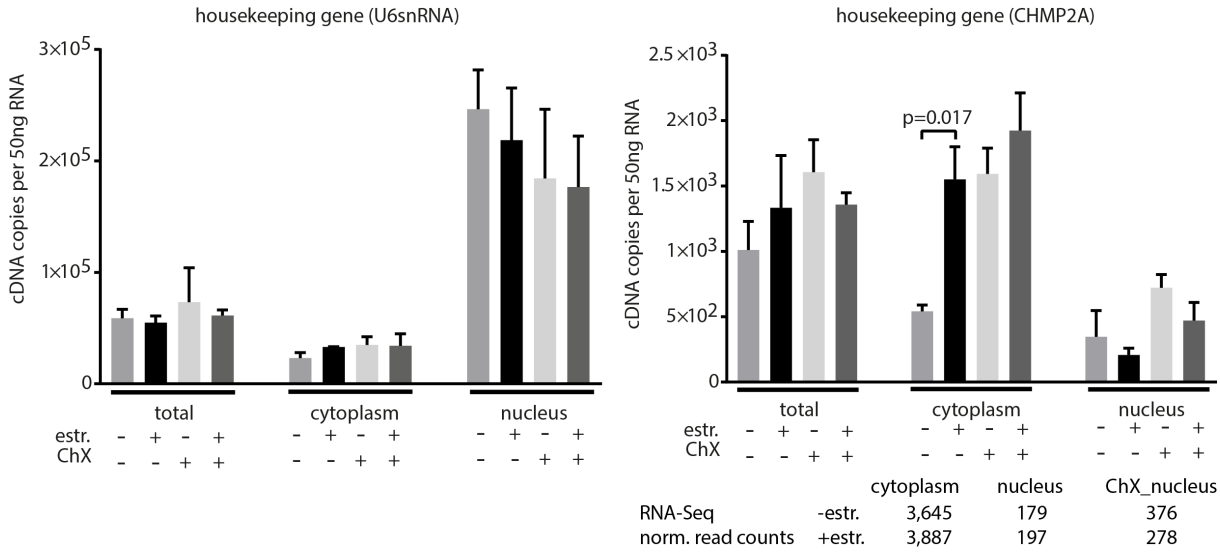


Figure 44: Transcripts of housekeeping genes as control for RT-qPCR. Bar graphs showing the absolute quantification by RT-qPCR of transcripts of the non-coding gene *U6snRNA* and the protein-coding gene *CHMP2A* in 4 in different RNA preparations. ER/EB2-5 cells were depleted for estrogen and reactivated for 0 h and 6 h with a subpopulation under ChX treatment. RNA was isolated from 10^7 cells (total) or subcellular fractions (1.2 x 10^7 cells for cytoplasm and 2 x 10^8 cells for nucleus) and 4 µg RNA was reverse transcribed to cDNA. Concentration of cDNA copies per 50 ng RNA as indicated ($n_{tech.} = 3$, $n_{biol.} = 3$). P-values obtained from unpaired t-test indicated if significant ($p < 0.05$). Underneath the bar graph, the raw read counts obtained from RNA-Seq aligned to ENSEMBL genes are displayed. Graph Pad Prism was used for plotting.

3.2.2.2.8 Protein coding and non-coding targets are regulated by E2 during establishment of latency

Post infection, latency is established (Fig. 1). E2 is one of the first viral genes expressed post infection. Since the transcriptional analysis described so far is conducted in established cell systems, we wondered whether the ER/EB2-5 system recapitulates the physiological events during the establishment of latency and whether the protein coding and non-coding target genes of E2 are regulated in a similar or different manner.

Therefore, primary B cells were isolated from adenoids of three individual donors and infected with the EBV strain B95.8. At different time points post infection, cells were harvested, RNA was isolated and specific E2 targets were quantified by RT-qPCR. Different strategies for quantification of transcript levels were performed. Usually, absolute quantification is based on the RNA amount which is input in reverse transcription. Since the RNA content per cell differs dramatically during the time course of infection due to cell growth (Table S1), an absolute quantification based on the cell number submitted to reverse transcription was performed in addition. Because of changes of the protein coding and non-coding housekeeping gene RNA levels (Figure 45), the absolutely quantified copy numbers per cell number of the candidate E2 targets were normalized to the housekeeping gene of same biotype. It can be appreciated that *E2* itself increases dramatically during infection, peaks at day three post infection and then decreases slightly (Figure 46, bottom right). *MYC* increases after infection, peaks at day three post infection and decreases again until it almost reaches its pre-infection levels (Figure 47, bottom right). One of the non-coding targets in the *MYC* neighborhood, *CASC21,* increases after infection with a peak at day 6 post infection (Fig. Figure 48, bottom right). *SLAMF1* quickly increases in the first hours post infection, reached a plateau with a minimal local peak at 24 h post infection and then decreases slightly (Figure 49, bottom right). The non-coding target in the *SLAMF1* neighborhood, *RP11-528G1.2*, increases after infection with a peak at day 6 post infection (Figure 50, bottom right). *PPAN-P2RY11* slowly increases after infection, peaks at day three post infection and then decreases (Figure 51, bottom right). The non-coding targets in the *PPAN* neighborhood, *CTD-2240E14.4*, increases after infection with a peak at day 6 post infection (Figure 52, bottom right).

Summing up, E2 coding and non-coding targets are already regulated during establishment of latency, indicating that the ER/EB2-5 system recapitulates the infection and latency establishment process.

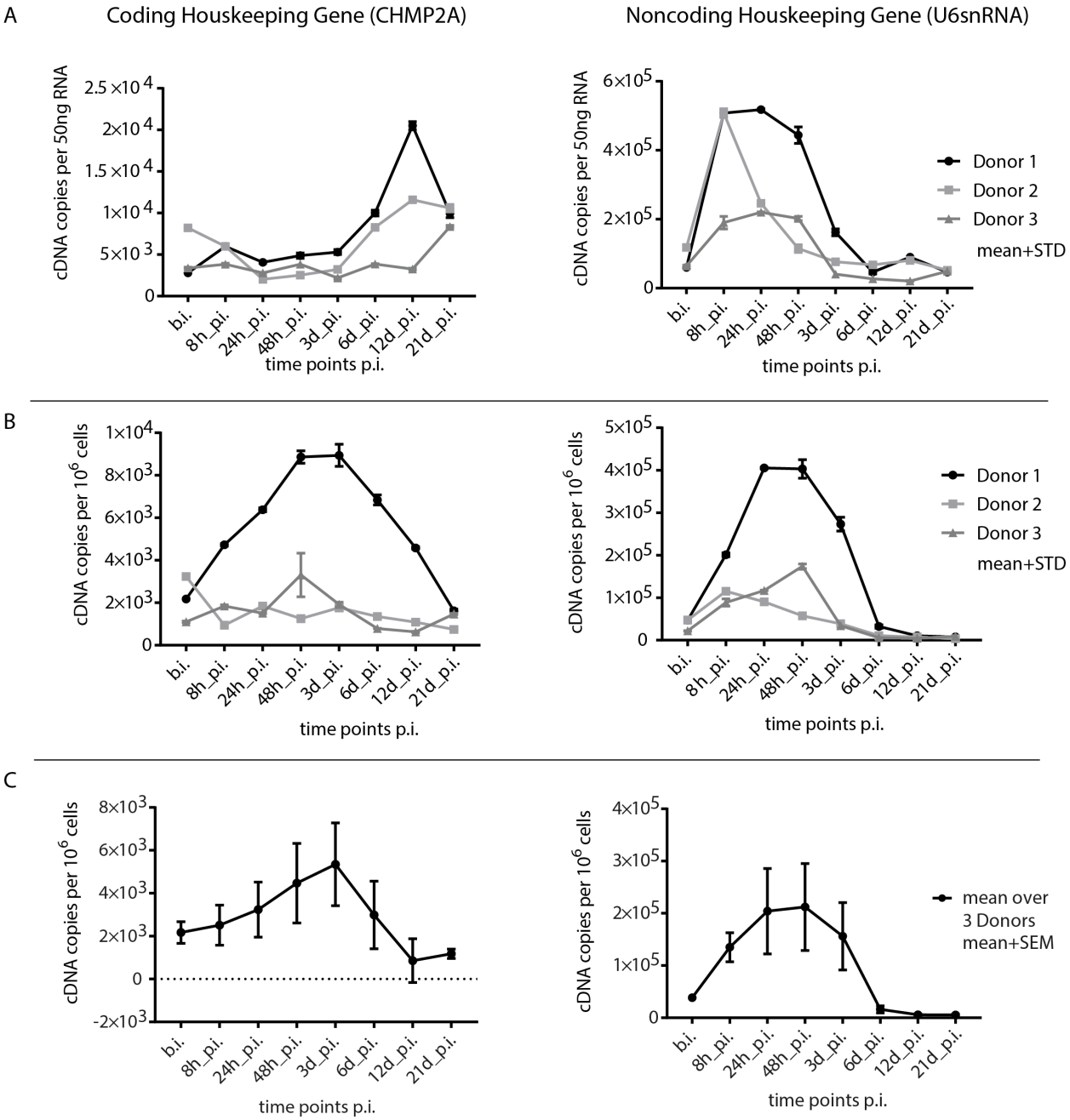


Figure 45: RT-qPCR of housekeeping genes during the establishment of latency. Line plots showing different analyses of RT-qPCR of transcripts of the housekeeping genes *CHMP2A* (protein coding) and *U6snRNA* (non-coding) at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount used for cDNA preparation for both genes, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number of both genes, concentration of cDNA copies per 10^6 cells as indicated. **C** Mean over three donors (of quantification shown in B) of both genes, concentration of cDNA copies per 10^6 cells as indicated. Graph Pad Prism was used for plotting.

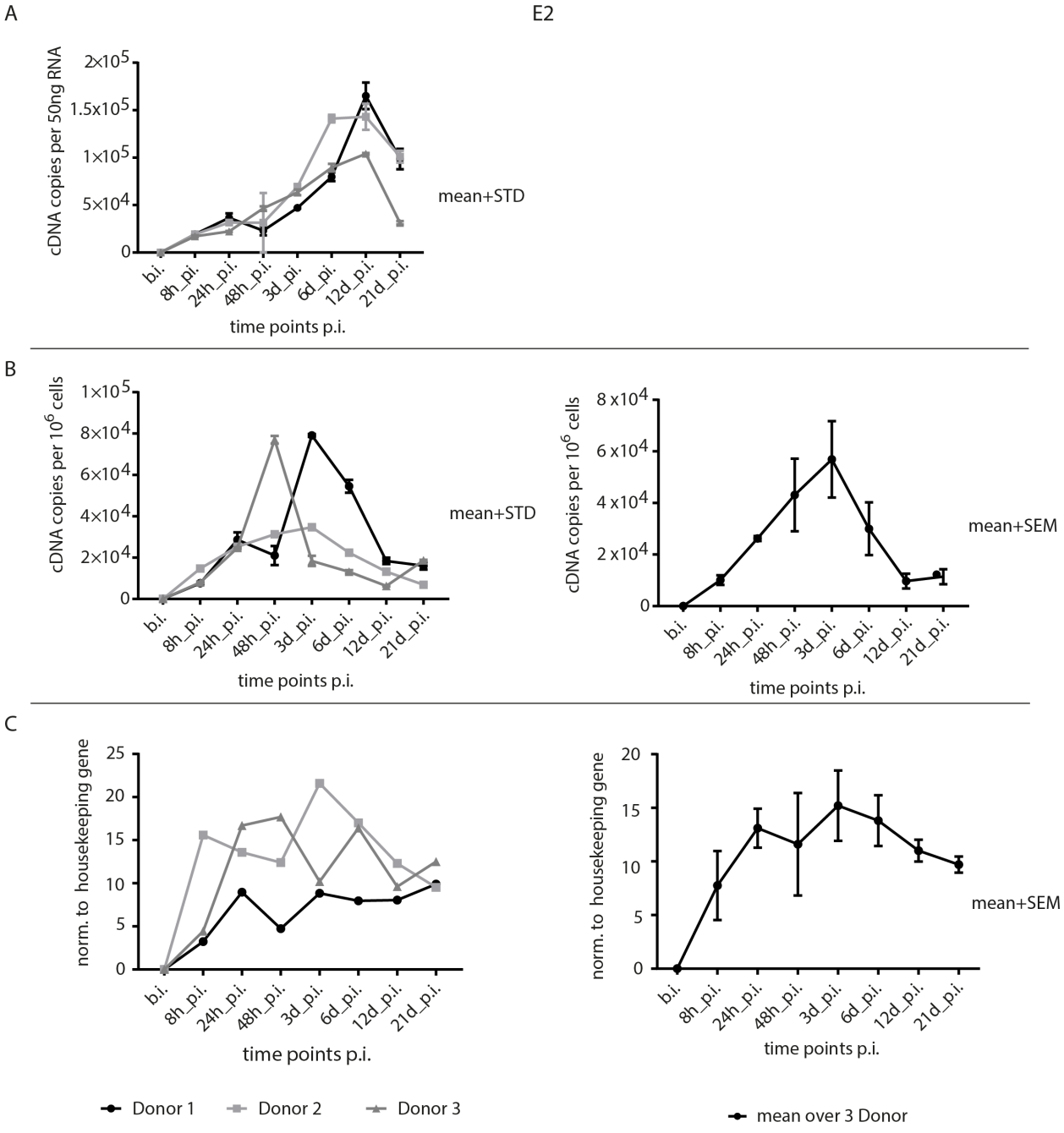


Figure 46: RT-qPCR of *E2* during the time course of infection showing a peak in abundance at 3 d p.i. and a steady increase of RNA abundance. Line plots showing different analyses of RT-qPCR of *E2* at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

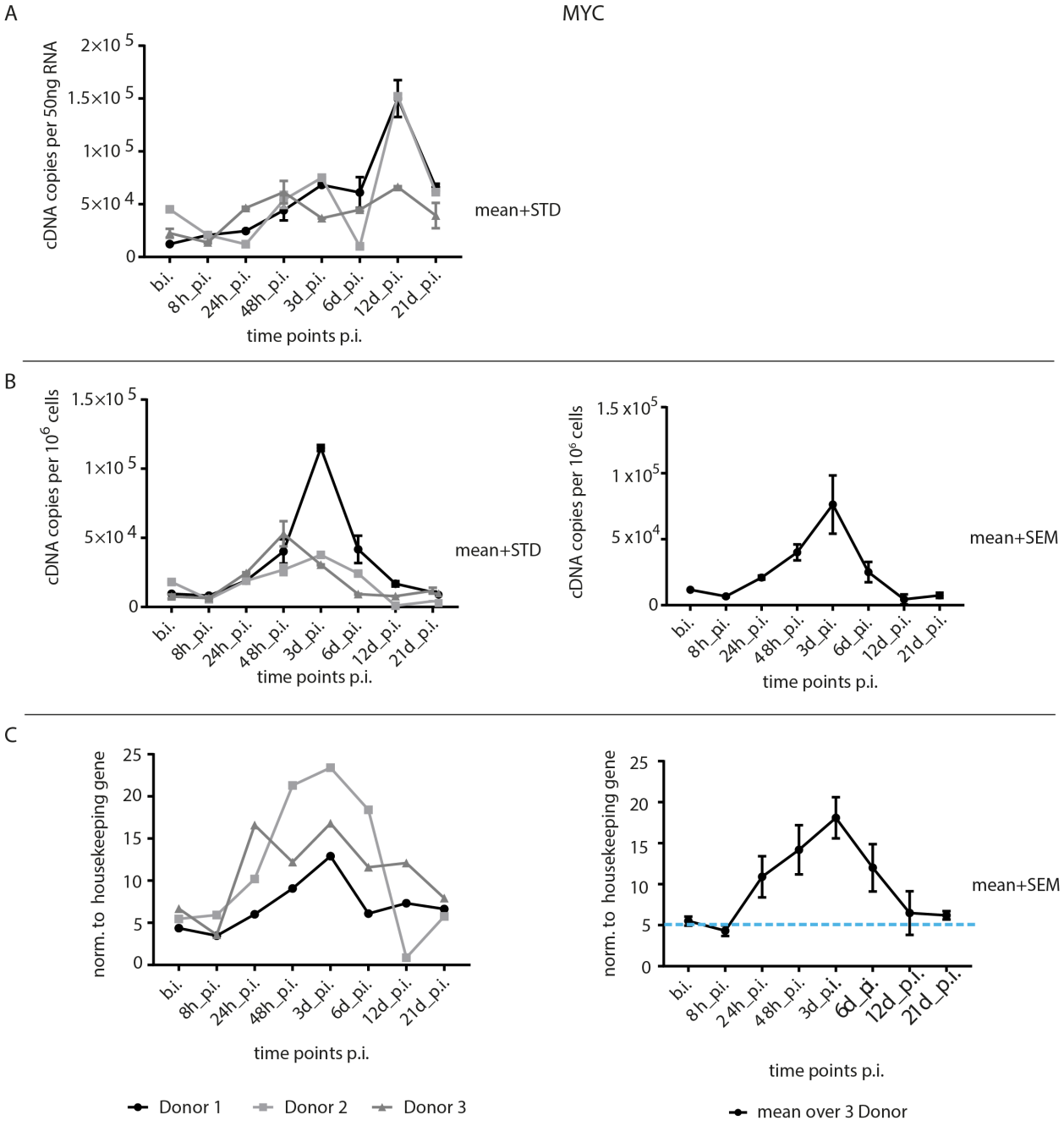


Figure 47: RT-qPCR of *MYC* during the time course of infection showing a peak in abundance at 3 d p.i.. Line plots showing different analyses of RT-qPCR of *MYC* at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

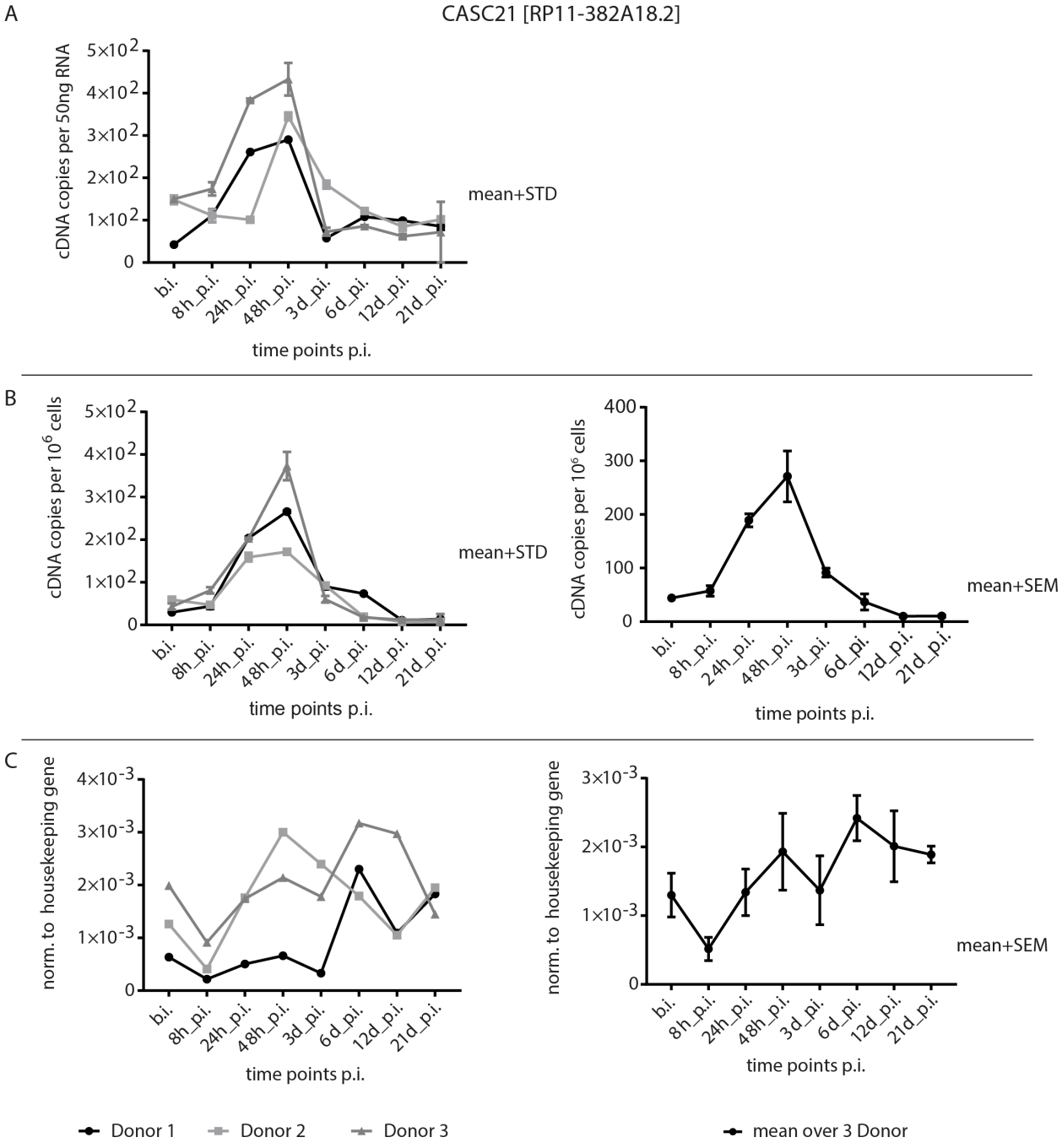


Figure 48: RT-qPCR of non-coding *CASC21* at *MYC* locus during the time course of infection showing a peak 6 d p.i. and a steady increase of RNA abundance. Line plots showing different analyses of RT-qPCR of *CASC21* at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

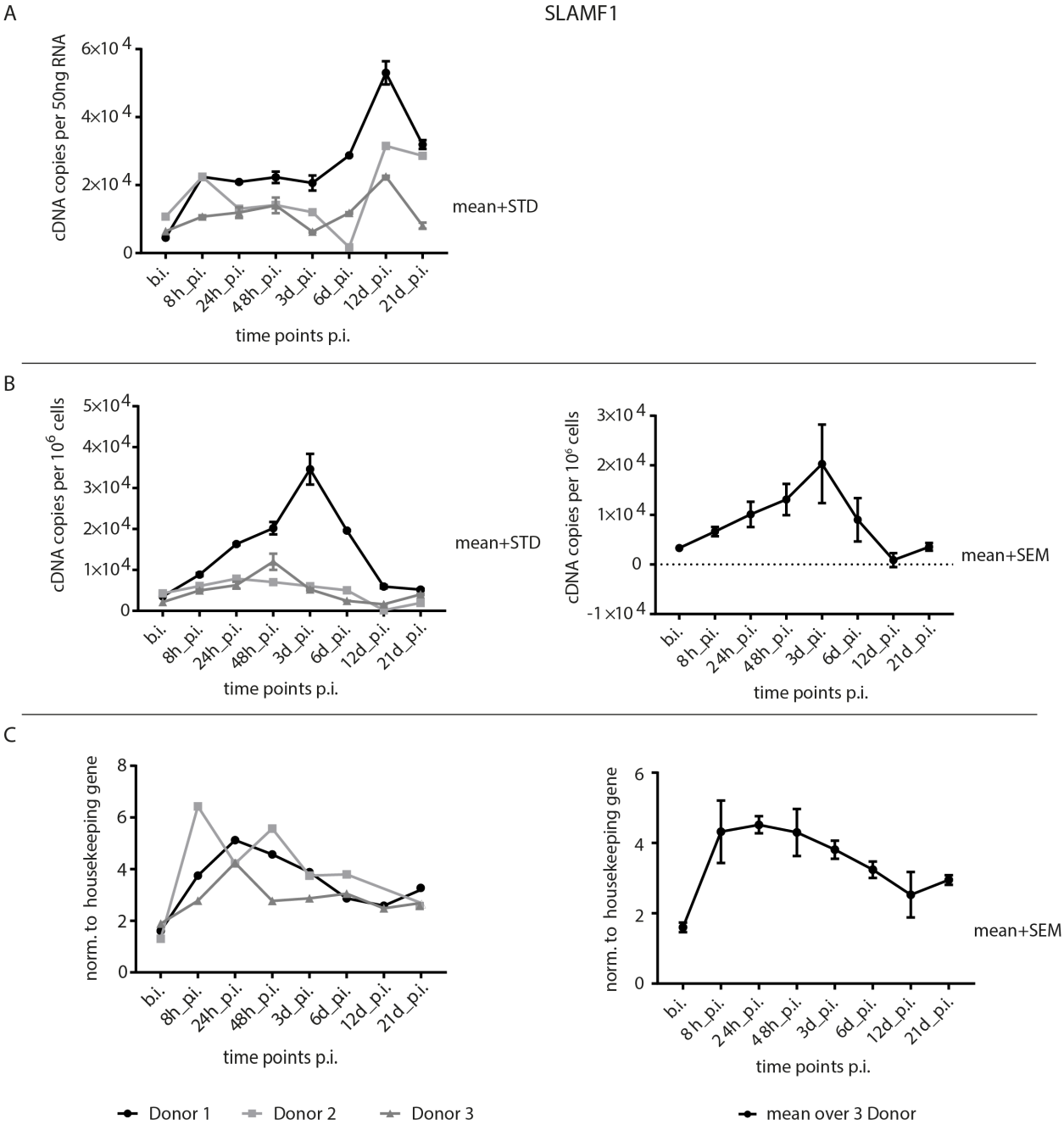


Figure 49: RT-qPCR of *SLAMF1* during the time course of infection showing a peak in abundance at 24 h p.i.. Line plots showing different analyses of RT-qPCR of *SLAMF1* at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

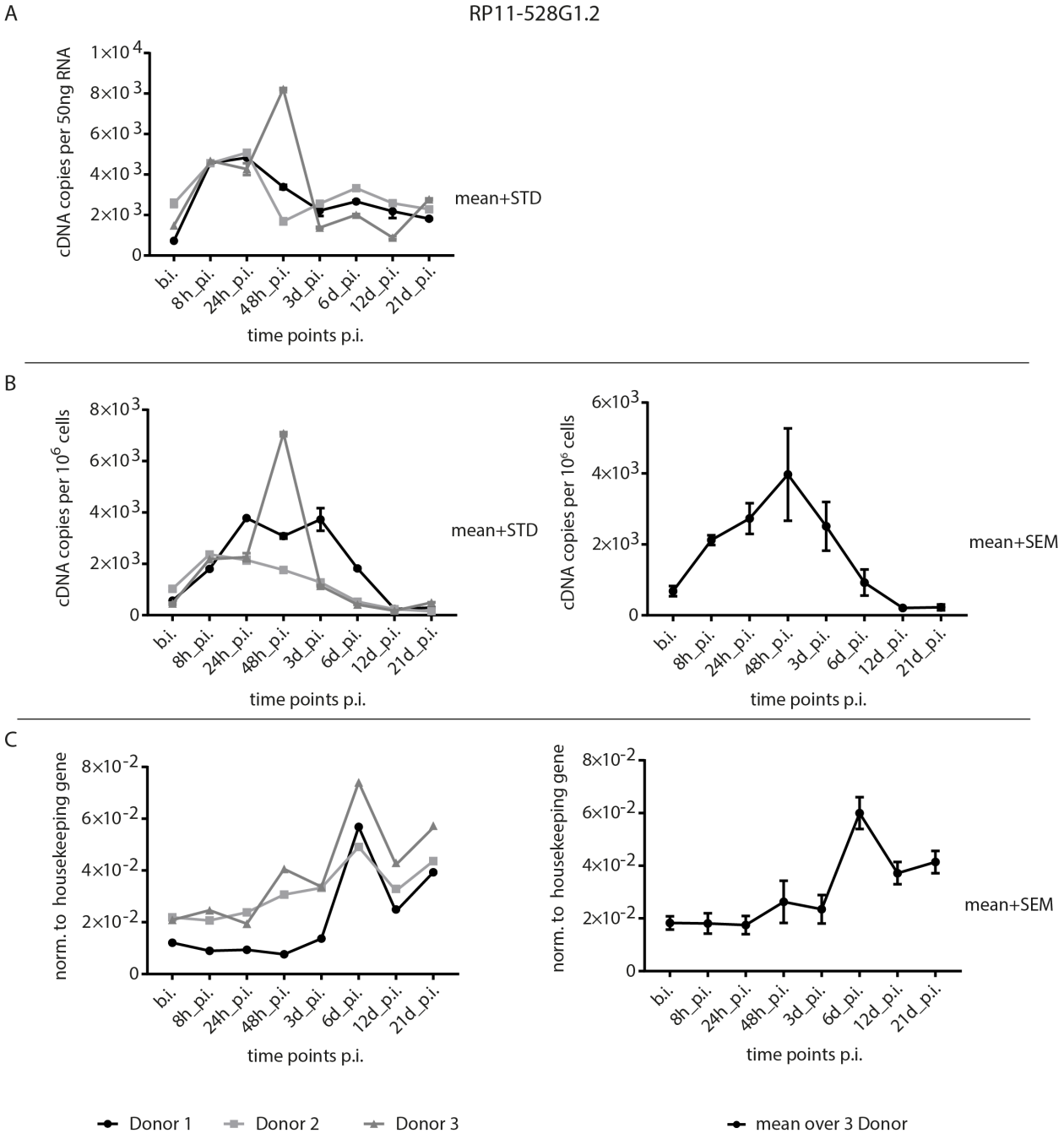


Figure 50: RT-qPCR of non-coding *RP11-528G1.2* at *SLAMF* locus during the time course of infection showing a peak 6 d p.i. and a steady increase of RNA abundance. Line plots showing different analyses of RT-qPCR of RP11-528G1.2 at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

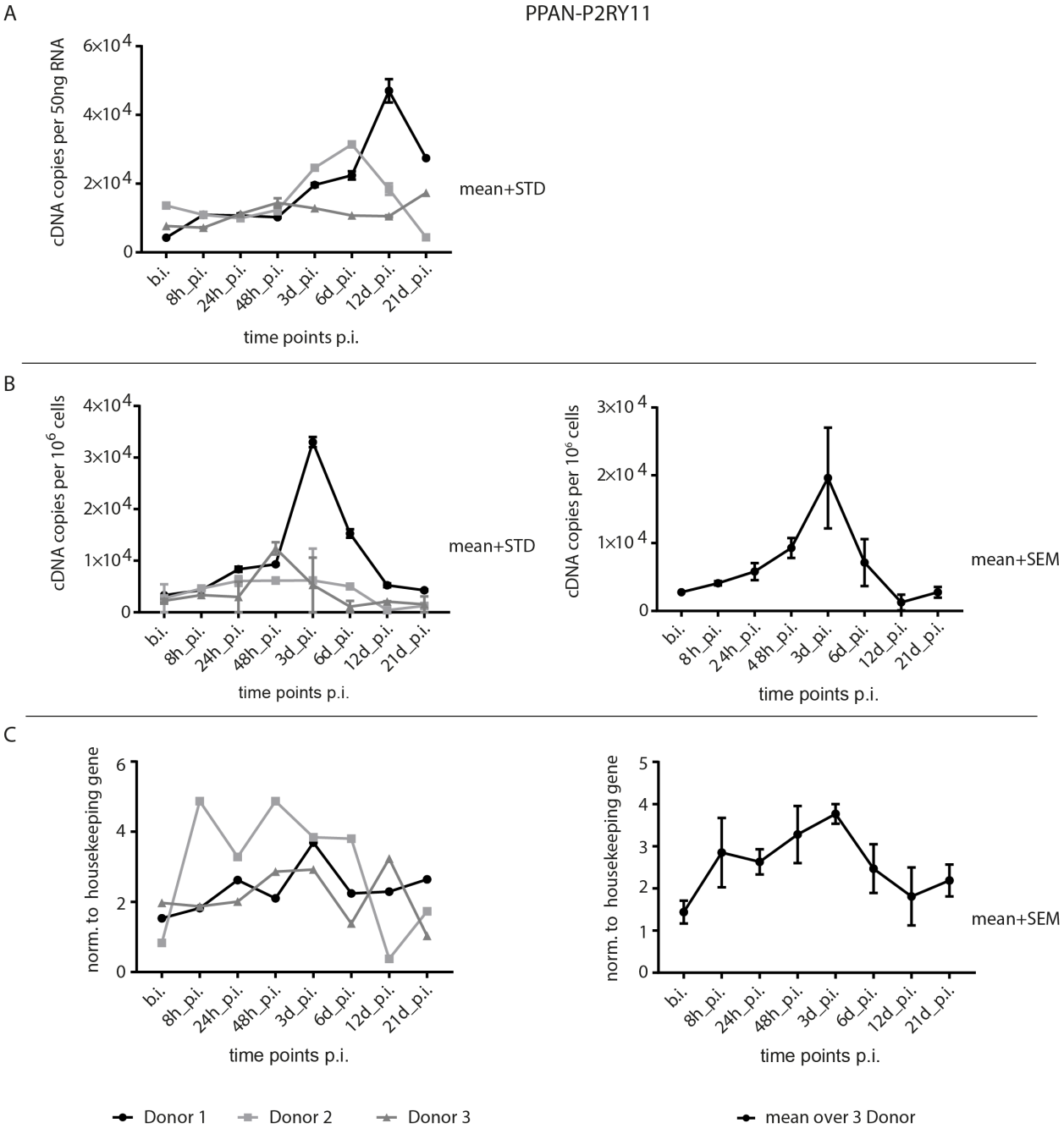


Figure 51: RT-qPCR of PPAN-P2RY11 during the time course of infection showing a peak in abundance at 3 d p.i.. Line plots showing different analyses of RT-qPCR of SLAMF1 at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

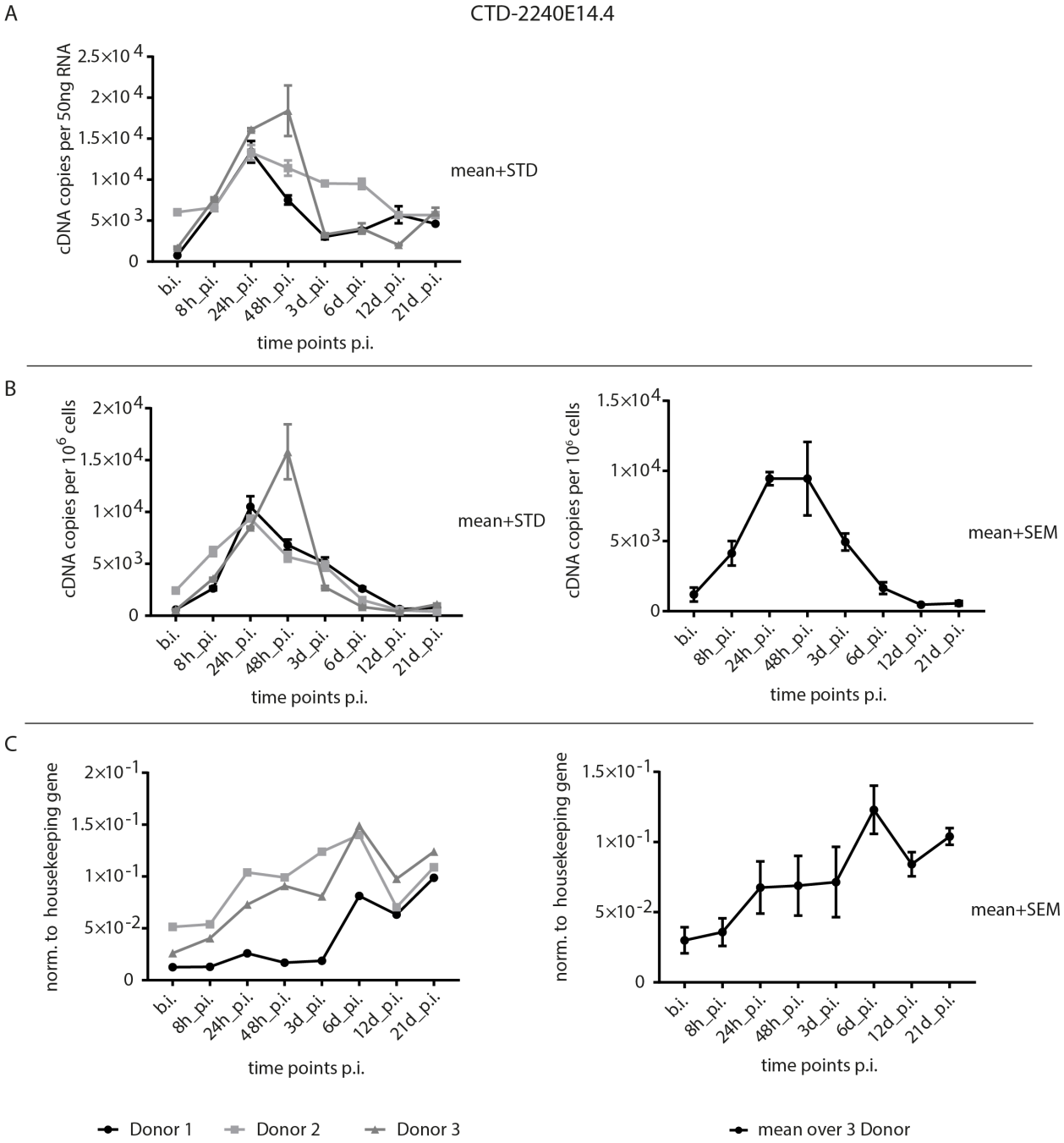


Figure 52: RT-qPCR of non-coding *CTD-2240E14.4* at *PPAN* locus during the time course of infection showing a peak 6d p.i. and a steady increase of RNA abundance. Line plots showing different analyses of RT-qPCR of CTD-2240E14.4 at different time points post infection. Primary B cells were isolated from adenoids of three different donors and infected with B95.8 viral supernatant. RNA was isolated from 5 x 10^6 or 10^7 cells and 1 µg of RNA was reverse transcribed to cDNA ($n_{tech.}$ = 2). **A** Absolute quantification based on RNA amount, concentration of cDNA copies per 50 ng RNA as indicated. **B** Absolute quantification based on cell number, concentration of cDNA copies per 10^6 cells as indicated for all three donors (left) and the mean over three donors (right). **C** cDNA copies (quantification shown in B) normalized to the coding housekeeping gene CHMP2A for all three donors (left) and the mean over three donors (right). Graph Pad Prism was used for plotting.

3.2.2.2.9 Regulation of non-coding genes correlates positively with the regulation of the neighboring protein coding genes

A mark of an active enhancer is its transcription, and an active enhancer is associated with the activation of transcription of remote gene loci. eRNAs were once claimed to rather act in *cis* in contrast to lincRNAs (W. Li et al., 2013). Next, we questioned whether the detected non-coding genes (ncg) were regulated alongside with protein coding genes (pcg). Furthermore, we investigated whether an enhancer chromatin state of the lncRNA has a substantial impact on activation of remote genes and whether the assignment of *trans-/cis-* mode of action for eRNAs/lincRNAs is obsolete.

For eRNA assignment, the chromatin state segmentation published by Ernst *et al.* in the context of the ENCODE project was used (Ernst et al., 2011). If the TSS flanking region (TSS -1000 bp +100 bp) of the lncRNAs intersected (≥ 1 bp) with one of the four enhancer chromatin states (two weak enhancer and two strong enhancer states), the lncRNAs was considered as eRNA in the following investigations. The lncRNA could have originally another classification. The classifications are not mutually exclusive (see section 1.3.2.1, p. 13).

We calculated the frequencies of distances of an expressed and significantly regulated eRNA to the closest expressed and significantly regulated pcg and compared this setting with the distances of other lncRNAs to their pcg partner. We observed that surprisingly, other lncRNAs were closer in proximity to pcgs than eRNAs (Figure 53A, left). This finding was intensified by the limitation of the distance to 2000 kb and the selection of only pairs with no interjacent unregulated gene (Figure 53A, right). TADs are claimed to be up to 2 Mb in size (Dixon et al., 2012), this was used as a comparison. We expected shorter distances between the regulated eRNAs to the closest pcg. In order to test, whether eRNAs compared to other lncRNAs were co-regulated with pcgs in their neighborhood, the correlation of log2FCs between eRNAs and other lncRNAs with their closest pcg partner was assessed (Figure 53B). As a control, randomly selected regulated RNAs were chosen. Since we were concerned to bias the result by selecting a specific subcellular compartment (pcgs are enriched in the cytoplasm and lncRNAs might be enriched in the nucleus) the analysis was carried out for lncRNAs in the cytoplasm versus pcgs in the cytoplasm (Figure 53B, right) for lncRNAs in the nucleus versus pcgs in the cytoplasm (Figure 53B, top right) and for lncRNAs in the nucleus versus pcgs in the nucleus (Figure 53B, bottom right). While the log2FCs of the control group of RNAs did not correlate with the log2FCs of their closest pcg neighbor, the log2FCs of eRNAs correlated with the log2FCs of their closest pcg neighbor. The “other lncRNAs” correlated slightly better (r=0.75 versus r=0.8) than eRNAs. The selection of a specific subcellular compartment for the biotypes did not result in a substantial difference in the correlation.

These data show that despite a long range distance of regulated lncRNA to the closest regulated pcg, their regulation strength and direction correlated positively. The enhancer chromatin state at the TSS flanking region of the transcripts of the defined eRNA genes did increase neither the proximity nor the correlation of the lncRNA-pcg pairs.

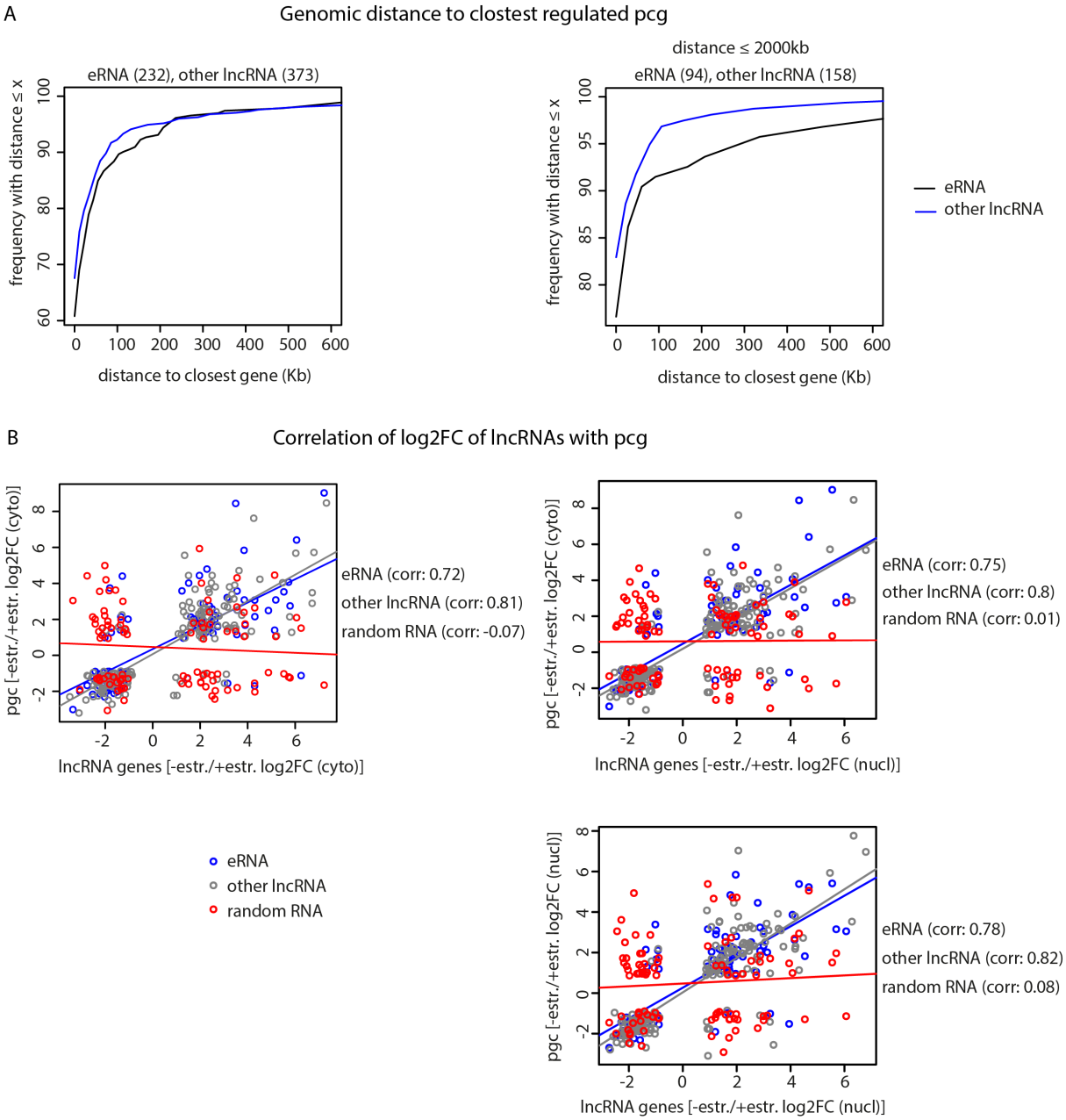


Figure 53: E2 significantly (FDR < 0.05) regulated (log2FC > 1 or <-1) lncRNAs and pcgs are not in next proximity, but the regulation of these target genes correlates positively. 4,480 genes which are similarly regulated by E2 in both subcellular compartments and are covered with > 20 reads in both compartments (= expr+ reg.) were selected for this analysis (3,653 pcgs, 232 eRNAs and 373 other lncRNAs) **A** Cumulative plots displaying the frequency of distances from lncRNA to pcg (left) and the frequency of distances restricted to max. 2,000 kb (right). **B** The log2FCs of eRNA and other lncRNAs correlate positively with the log2FCs of pcgs in comparison to random selected RNAs. Pearson correlation on the log2FCs of two compared groups (given by x- and y-axes); line displays result of linear regression on these groups (selection of gene subsets was conducted by me; plots were created by Gergely Csaba).

3.2.2.2.10 Genome-wide characterization of E2 regulated genes

Aiming to comprehensively characterize E2 regulated genes with focus on lncRNAs in detail, several existing data on LCLs were integrated to delineate different subsets of E2 regulated genes. In the following, two groups of genes were subject to the investigations

i) ENSEMBL genes and
ii) intergenic genes (with regards to ENSEMBL gene annotation).

ENSEMBL annotated genes for hg19/GrCh37 from the release 75 (Feb 2014, latest update available for hg19/GrCh37 at beginning of this thesis; 63,677 annotated genes), which resembles the Gencode V19 annotations (GENCODE consortium within the framework of the ENCODE project) were further analyzed. All for our cells relevant genes, for which ≥ 1 read was detected in any sample, where input for a giant "filter tree". This resulted in 31,192 human genes. According to the ENSEMBL biotype categorization (see section 3.2.2.2.5 p. 68) these genes consisted of 17,611 protein coding genes (pcgs), 7,676 lncRNAs and 5,905 genes of heterogeneous biotypes (pseudogenes, polymorphic_pseudogenes, IG_C_gene, IG_V_gene IG_V_pseudogene, IG_C_pseudogene TR_V_gene, TR_C_gene, TR_V_pseudogene, processed_transcript, misc_RNA, snoRNA, snRNA, miRNA). The assigned biotype of a gene corresponds presumably to the biotype of the first detected transcript for this gene, there might be more transcripts belonging to a gene assigned with different biotypes. It has to be mentioned, that in this analysis, of all the 17,611 protein coding genes there are ~700 genes, for which the biotype of the transcript with highest F_{rank} (explanation see below) is a non-coding biotype.

The second group covers differentially expressed regions, which have not been annotated by ENSEMBL (for determination see section 3.2.2.2.2, p. 58). The discovered 8,918 regions were designated as genes. Assuming that all protein coding genes are already uncovered, these genes might all be lncRNAs genes. Regulated intronic regions were not further analyzed due to the uncertainty of their genesis.
Where genes regulated in the same direction after comparing two different datasets were encountered, we designated them as co-regulated.

<u>E2 targeted lncRNA genes are found in nucleus and cytoplasm</u>

First, the question arose in which subcellular compartment, nucleus (nucl) or cytoplasm (cyto), E2 regulated RNAs are predominantly located. Particularly interesting for us were E2 regulated lncRNAs and their location with regard to their function. In the following, genes regulated by E2 with regard to the two compartments are analyzed.

ENSEMBL genes

The 31,192 ENSEMBL genes were filtered for their read coverage (Figure 54A). 19,154 genes, covered with > 20 reads per gene in either of the two compartments (nucleus or cytoplasm) or in either of the two conditions (-/+estr.) were considered as expressed. The remaining genes were subsequently selected for significant (worst FDR < 0.05) regulation (log2FC > 0.85 or < - 0.85) in either of the two subcellular compartments (decision [1] in the filter tree, Figure 54A). 7,599 genes remained, which were then screened for genes which were regulated in the same direction (co-regulated) by E2 in the cytoplasm and the nucleus, leaving 5,104 genes (decision [2] in the filter tree, Figure 54A, Figure 54B). Among those are 797 lncRNAs (Figure 54B). 2,495 genes are not significantly co-regulated. The 5,104 genes can be significantly co-regulated in both compartments, but the threshold of the read coverage at the very beginning was set in regard to one of the four samples. Demanding > 20 reads in the nucleic samples left 4,620 genes considered expressed here, hence, 484 genes are enriched solely in the cytoplasm (Figure 54C). Demanding > 20 reads in the cytoplasmic samples remained 4,964 genes considered expressed here, consequently are 140 genes enriched solely in the nucleus. 4,480 genes were covered with > 20 reads in both compartments.

Wondering, whether the nucleic compartment is enriched for lncRNA genes, the biotypes were demanded for the genes co-regulated in the nucleus and expressed or enriched there (Figure 54Figure 54D, upper panel first two graphs l.t.r) and for the genes being regulated solely in the nucleus (with > 20 reads; Figure 54D, lower panel left). The biotype of lncRNAs was further split up in lncRNAs, which emanate from chromatin with enhancer state (> 1 overlap with enhancer state according to CSS HMM; Ernst et al., 2011) and are for this thesis defined as eRNAs and other lncRNAs (all other ENSEMBL annotated lncRNA, whose TSS flanking region did not overlap with an enhancer state). Since eRNAs are thought to have mainly gene regulatory functions, they were expected to reside in the nucleus. Comparing the nucleic lncRNAs with their cytoplasmic counterparts, it can be observed that there is quasi no difference in biotype composition regarding the genes which are expressed in nucleus and cytoplasm (Figure 54D, upper panel, middle), most likely because of the huge overlap of genes residing in both subcellular compartments (Figure 54C). Strikingly, the set of genes enriched in both the cytoplasm and nucleus contains more lncRNAs compared to the total amount of expressed genes. Furthermore, more co-regulated lncRNAs reside in the nucleus than in the cytoplasm (Figure 54D, upper panel, edges). Against all expectations, eRNAs were augmented in the cytoplasm enriched subset. Equally, examining the specifically in the nucleus and cytoplasm regulated genes, eRNAs and other lncRNAs are augmented in the cytoplasm (Figure 54D, lower panel). Summing up, our data indicate that E2

regulated lncRNAs and especially eRNAs reside in both compartments and are according to our methods augmented in the cytoplasm of ER/EB2-5 cells.

Next, we investigated if the lncRNAs in the different gene subsets of the two subcellular compartments differed in length or splice status. Assuming that eRNAs are rather short and monoexonic transcripts, which are not spliced and transported out of the nucleus, we expected the whole population of lncRNAs in the nucleus to be rather short and monoexonic, whereas the longer, multiexonic transcripts could reside in both compartments according to our assumptions. For this analysis, we examined the underlying transcripts of the different gene subsets. In order to determine the most likely differentially expressed transcripts for each gene, ranking analysis by F-statistics was conducted. For all transcripts belonging to a gene, a F-measure was assessed, the transcripts were ranked according to their F measure and as a cutoff, only transcripts with $F_{rank} \leq 2$ were considered for further analysis. Investigating the median length of the transcripts of lncRNAs in the different gene subsets with this as requirement, we observed that in general the transcripts of all expressed lncRNAs (mean ~1.0 to 1.5 kb) differed in length from the transcripts of all expressed pcgs (mean= 3.4 kb; Figure 54E). However, the length of the transcripts deriving from lncRNA genes is similar comparing subsets in the nucleus to cytoplasm. eRNAs were augmented in the cytoplasm enriched genes (Figure 54D), which does not impact the transcript length in this subset. Transcripts of lncRNA genes enriched in the nucleus occur to be shorter (mean= 960 bp) compared to the transcripts of all expressed lncRNA genes (mean= 1.3 kb). One has to bear in mind that the subset of genes enriched in the nucleus is the smallest. The length of transcripts of lncRNA genes specifically regulated in the nucleus or cytoplasm is similar.

In order to study the splice status of the transcripts, the number of exons of the transcripts ($F_{rank} \leq 2$) per regulated lncRNA gene was assessed. We determined that overall, the transcripts of all expressed lncRNAs have less exons (mean= 2.5 exons) than transcripts of all expressed pcgs (mean= 10.4 exons; Figure 54F). Nevertheless, the number of exons of transcripts deriving from lncRNA genes does not differ significantly comparing subsets in the nucleus to cytoplasm. eRNAs were augmented in the cytoplasm enriched genes, which has no impact on the number of exons in this subset compared to the same subset in the nucleus. The transcripts of lncRNA genes specifically regulated in the cytoplasm appear to have less exons. In conclusion, the lncRNAs in the different gene subsets of the two compartments do not differed in length or splice status.

We found that E2 regulated lncRNAs including eRNAs reside in both subcellular compartments, the nucleus and the cytoplasm and the lncRNAs in both compartments are largely similar regarding their transcript length and splice status.

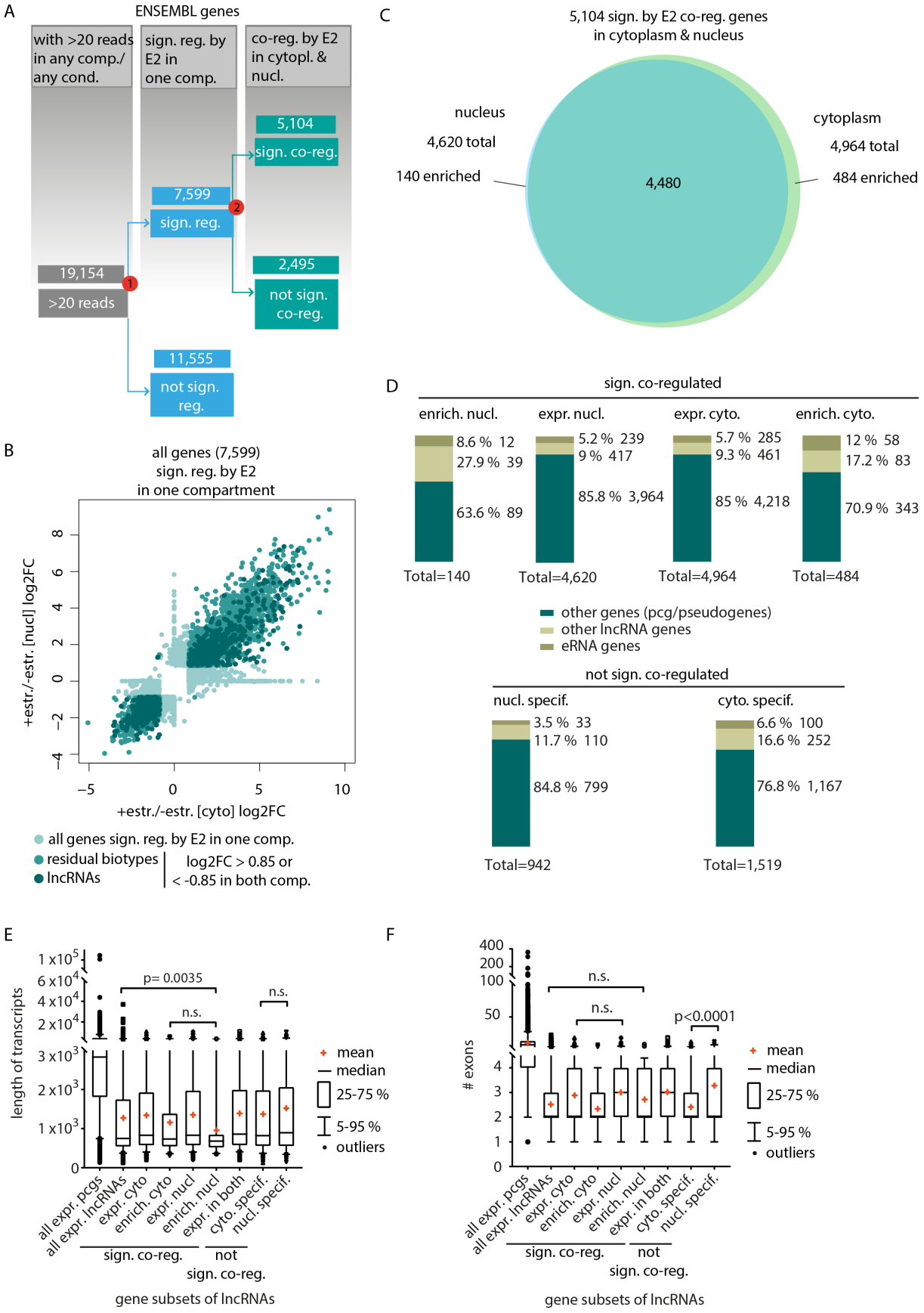


Figure 54: Characterization of E2 regulated ENSEMBL genes. A Decision tree on subsets of genes to be further analyzed: All genes with > 20 reads in any compartment or any condition were considered to be expressed and filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 0.85 or <- 0.85) by E2 in

any compartment and 2. significant (worst FDR < 0.05) regulation (log2FC > 0.85 or <- 0.85) in the same direction (co-regulation) in both subcellular compartments. **B 67 % of E2 target genes in one of the compartments are co-regulated in both subcellular compartments**. Scatterplot showing the positive correlation of log2FCs of the E2 co-regulated (5,104 genes out of the 7,599) genes in the cytoplasm vs. nucleus (not filtered for significance in both compartments for plotting). Log2FCs of filtered genes were plotted using R. **C 88 % of co-regulated genes (4,480) are expressed (>20 reads) in both compartments**. Venn diagram displaying the localization of regulated genes. The intersection of co-regulated genes expressed in both compartments was plotted using R. **D The number of lncRNAs is increased in the subset of genes enriched in the nucleus**. Vertical slices showing the fraction of biotypes (eRNAs defined as RNAs with ≥ 1 bp overlap with enhancer signature -1000 bp +100 bp around TSS) of co-regulated genes (upper panel) and not co-regulated genes (lower panel) in the cytoplasm and the nucleus (> 20 reads). Plots were created using Graph Pad Prism. **E The lncRNA transcripts of most gene subsets don't differ in their length**. Box plots showing the median length of all transcripts ($F_{rank} \leq 2$) derived from E2 regulated genes in the different gene subsets (mean indicated); p-values of interesting comparisons indicated, obtained from unpaired t-test with Welch's correction. Plots were created using Graph Pad Prism. **F The lncRNA transcripts of most gene subsets don't differ in their number of exons**. Box plots showing the average number of exons of transcripts ($F_{rank} \leq 2$) regulated by E2 in the different gene subsets; mean indicated; p-values of interesting comparisons indicated, obtained from unpaired t-test with Welch's correction. Plots were created using Graph Pad Prism.

Intergenic genes

As before, the 8,918 intergenic (interjacent to ENSEMBL annotated genes) transcribed genes were filtered for their read coverage (Figure 55A). 4,772 genes, covered with >20 reads per gene in either of the two compartments (nucleus or cytoplasm) in either of the two conditions (-/+estr.) were considered as expressed. The remaining genes were subsequently selected for significant (worst FDR <0.05) regulation (log2FC > 1 or < - 1) in either of the two compartments (decision [1] in the filter tree, Figure 55A). 603 genes remained which were then screened for genes, which were co-regulated by E2 in the cytoplasm and the nucleus, leaving 360 genes (decision [2] in the filter tree, Figure 55A, Figure 55B). 243 genes are not significantly co-regulated. In the nucleic samples, 288 genes remained after demanding a gene coverage of > 20 reads (=considered expressed in the nucleus), hence, 72 genes are enriched in the cytoplasm only (Figure 55C). In the cytoplasmic samples, 233 genes remained after demanding gene coverage of > 20 reads (considered as expressed in the cytoplasm), consequently, 127 genes are only enriched in the nucleus. 161 genes were covered with > 20 reads per gene in both compartments. It can be appreciated, that more intergenic genes are located in the nucleus than in the cytoplasm.

We questioned whether novel lncRNAs can be identified in the different gene subsets. For this, we intersected the E2 regulated, intergenic genes detected in ER/EB2-5 with the entries of the lncRNA Atlas LNCat (Figure 55D). This comprehensive database combines lncRNAs from 24h resources including GENCODE, LNCipedia and NONCODE. If the smaller of the two compared genomic intervals (gene entry LNCat versus detected intergenic gene) overlapped > 50% with the other data set, the intergenic gene was considered as already known. If the smaller of the two compared intervals overlapped < 50 % with the other data set, the intergenic gene was considered as potentially novel. If there was no overlap at all, the intergenic gene was considered as novel. The

request for an overlap for the genes co-regulated and enriched or expressed in the nucleus or in the cytoplasm (Figure 55D, upper panel) and for the genes being regulated solely in the nucleus or cytoplasm (with >20 reads; Figure 55D, lower panel) resulted in remarkable 20-35 % of novel intergenic genes in the different subsets. Up to 10 % of the intergenic genes are already recorded and 65-70 % of the genes partly overlap with entries of LNCat. Altogether, there is evidence for identification of novel lncRNA genes. This would need to be confirmed by the search for ORFs.
Furthermore, we investigated the length of all E2 regulated intergenic genes in the different gene subsets of the two compartments (Figure 55E). Due to the enrichment of immature transcripts in the nucleus, the outer borders of a gene were difficult to infer. The average length of all detected intergenic genes is 3 kb, intergenic genes enriched in the cytoplasm (mean= 1.7 kb) and specific for the cytoplasm (mean= 1.4 kb) were observed to be significantly shorter than their nucleic counterparts (mean= 2.4 and 2.5 kB, respectively). The intergenic genes of the remaining subsets don't differ substantially. The intron-exon structure of the intergenic genes could not be reliably predicted, also due to the nature of the data (many intronic reads; Figure S22). Introns could not be reliably inferred even with increasing sensitivity. Furthermore, no convincing data could be collected to infer the required number of fragments as evidence for a junction, since increasing the number of fragments required for the acceptance of a junction did not substantially alter the number of inferred introns.
Last, we wanted to determine the chromatin state at the flanking region (-1000 bp, +100 bp) of the transcription start side (TSS) of the intergenic genes (Figure 55F; > 1 bp overlap of TSS flanking region with enhancer state according to CSS HMM; Ernst et al., 2011) in order to investigate, whether E2 predominantly activates eRNAs. Most of the intergenic genes had no chromatin state assigned (55-77 %). The majority of the detected intergenic genes with an assigned chromatin state harbor marks of transcription (TxN; Figure 55F, top left). The distribution of chromatin states is similar in all gene subsets except for the subsets of genes which are specifically regulated in the cytoplasm, most of the genes are not assigned with a chromatin state. H3K36me3 and H4K20me1 are marks associated with transcriptional activation.

Taken together, novel (with regards to the ENSEMBL annotation) intergenic transcribed genes can be enriched in the nucleus. Up to 30 % of the intergenic genes seem to be entirely novel since they do not overlap with genes of the comprehensive database LNCat. Intergenic transcribed genes in the cytoplasm were found to be shorter than the intergenic transcribed genes in the nucleus. As expected, intergenic transcribed genes were shown to exhibit marks of active transcription, however they were not enriched for enhancer marks.

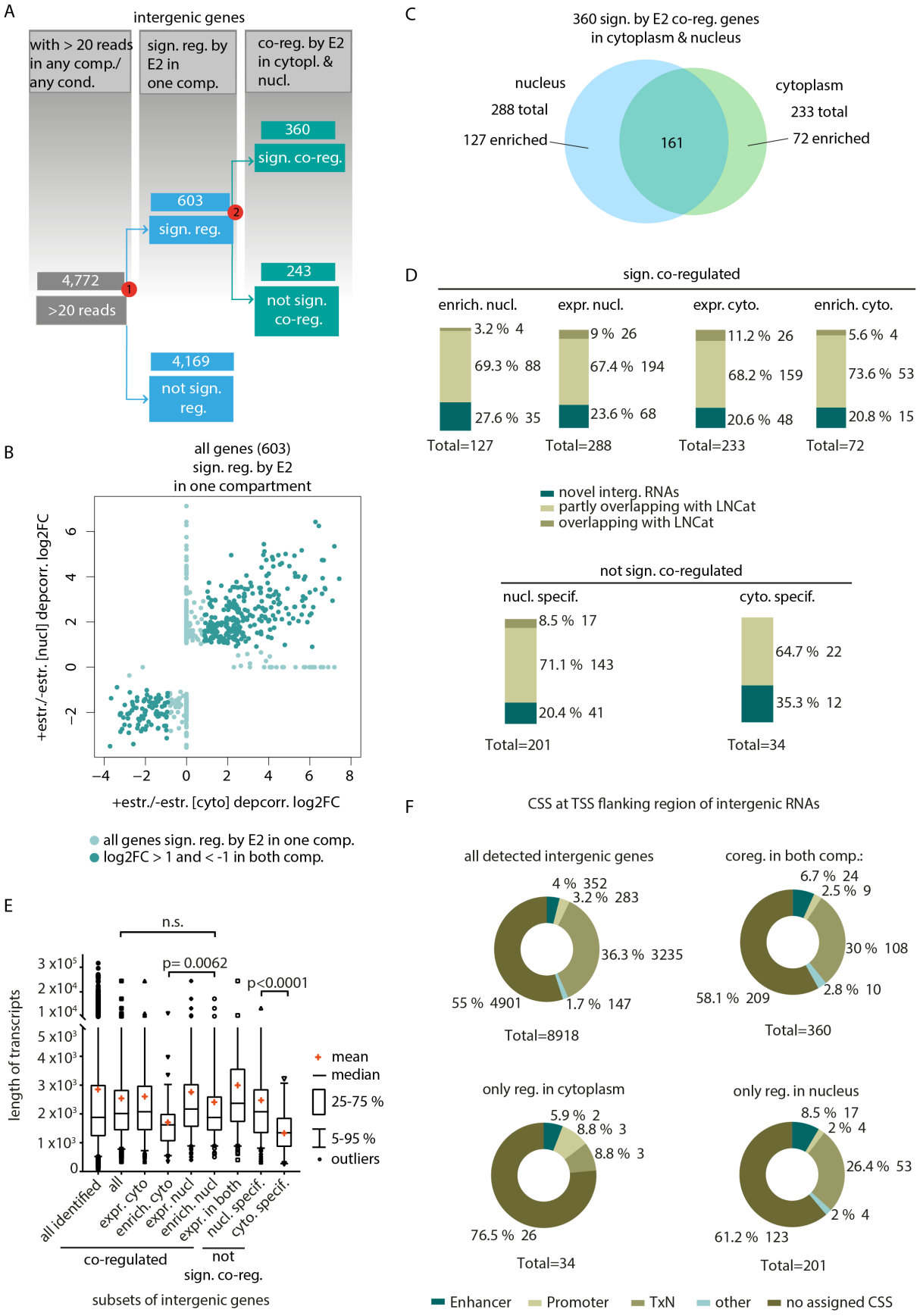


Figure 55: Characterization of E2 regulated intergenic genes. A Decision tree on subsets of genes to be further analyzed: All genes with > 20 reads in any compartment or any condition were filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 1 or < - 1) by E2 in any compartment and 2. significant

regulation (worst FDR < 0.05) regulation (log2FC > 1 or < - 1) in the same direction (co-regulation) in both compartments **B 60 % of E2 target genes in one of the compartments are similarly regulated (co-regulated) in both subcellular compartments**. Scatterplot showing the positive correlation of log2FCs of the E2 co-regulated (360 genes out of the 603) genes in the cytoplasm vs. nucleus (not filtered for significance in both compartments for plotting). Log2FCs of filtered genes were plotted using R. **C 45 % of co-regulated genes (161) can be found in both compartments.** Venn diagram displaying the localization of regulated genes. The intersection of co-regulated genes expressed in both compartments was plotted using R**. D The number of novel intergenic RNAs is increased in the subset of genes enriched in the nucleus.** Vertical slices showing the fraction of intergenic genes not, partly (< 50 % of the smaller compared interval) or entirely (> 50 % of the smaller compared interval) overlapping with entries in the LNCat (upper panel= co-regulated genes, lower panel= not co-regulated genes) in the cytoplasm and the nucleus (> 20 reads). Plots were created using Graph Pad Prism. **E The genes of most subsets do not significantly differ in their median length**. Box plots showing the median length of genes regulated by E2 in the different gene subsets. Plots were created using Graph Pad Prism. **F 55-76 % of the genes don't hold an assigned CSS, genes specifically regulated in the nucleus are enriched for marks for transcription (TxN)**. Donut plots showing the fractions of genes with different CSS at their TSS flanking region (-1000 bp +100 bp; ≥ 1 bp overlap of TSS flanking region with chromatin state). CSS in GM12878 based on the definition by Ernst et al, 2011 (ENCODE project; other= insulator, polycomb repressed, heterochromatin, repetitive/CNV). Since multiple chromatin states can be associated per transcript, an exclusive classification was conducted (enhancer>promoter>TxN>other). Plots were created using Graph Pad Prism.

E2 regulates 174 genes in the absence of *de novo* protein synthesis

To identify direct target genes of E2, the translation inhibitor ChX was applied. However, it has to be pointed out that this strategy only targets protein coding target genes. Here, the inhibition of the following translation prevents any downstream effect of the gene.

ENSEMBL genes

31,192 ENSEMBL genes were again filtered for their read coverage in the samples +/- estr. and ChX+estr/ ChX- estr. (Figure 56A). 16,227 genes were covered with > 20 reads per gene in either of the conditions. The remaining genes were subsequently selected for significant (worst FDR < 0.05) regulation (log2FC > 0.85 or < - 0.85) in the nucleus in the absence of ChX (decision [1] in the filtertree, Figure 56A). 4,679 genes remained, which were then screened for regulation in the presence of ChX (decision [2] in the filter tree, Figure 56A, Figure 56B). Among the remaining 178 genes are already described direct E2 targets like MYC, DNase1L3 and CR2 (Figure 56B). 4,501 genes are not significantly regulated under ChX (Figure 56C). Of the 178 genes regulated by E2 in both ChX conditions, 174 are co-regulated. We continued only with genes which are regulated by E2 -/+ChX in the same direction (either up- or downregulation).

Examining the biotype composition of the E2 regulated genes -/+ ChX, we observed that 38 of the co-regulated genes are lncRNAs, among them also 11 eRNAs (Figure 56D, right). There is a slight enrichment for lncRNAs under ChX (from 12.2 % to 21.8 %). Since E2 is known to preferably bind to open chromatin, especially to enhancers, we speculated whether this is mirrored in the chromatin state of the "directly" regulated genes. We aimed to determine the chromatin state at the TSS flanking region of the transcripts belonging to the E2 regulated ENSEMBL genes in

presence and absence of translation (Figure 56E; > 1 bp overlap of TSS flanking region with enhancer chromatin state according to CSS HMM). With or without the treatment of ChX, the enhancer chromatin state at the TSS flanking regions of transcripts of E2 regulated genes is increased compared to the general distribution of CSS of the transcripts of all genes detected in our data sets. Nevertheless, the TSS flanking regions showed enriched promoter CSS for cells without ChX treatment compared to ChX treated cells, whereas genes with “other” or no assigned CSS at the TSS flanking region of their derived transcripts are enriched with ChX (other= insulator, polycomb repressed, heterochromatin, repetitive/CNV). However, these data have to be interpreted cautiously, since the CSS of GM12878 might differ compared to the CSS of ER/EB2-5 cells + estr. as well as –estr.

These data show, that in presence and absence of *de novo* protein synthesis, E2 regulates a substantial amount of genes, including lncRNAs, which largely emanate from enhancer marked chromatin.

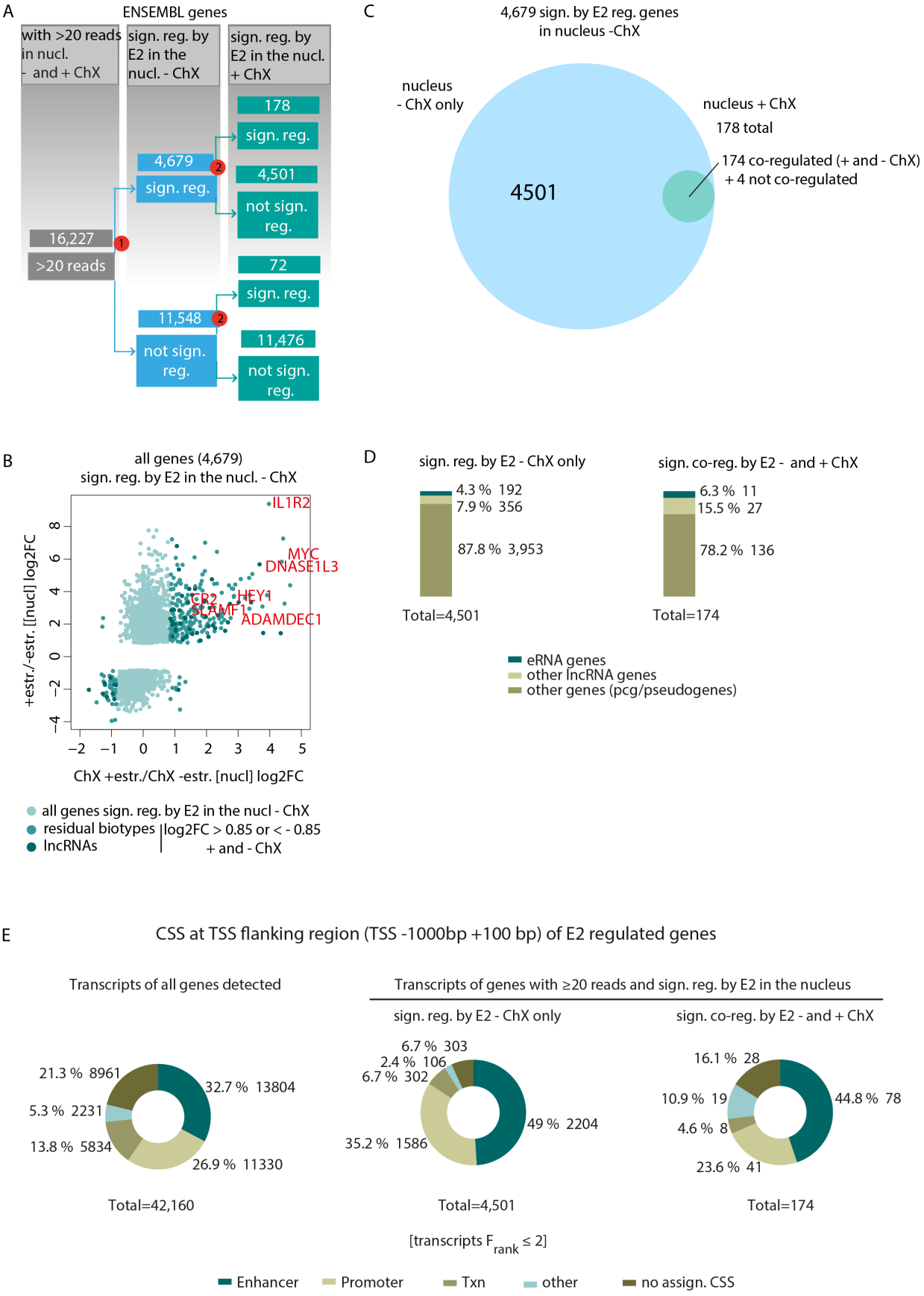


Figure 56: Characterization of E2 regulated ENSEMBL genes in absence of *de novo* protein synthesis
A Decision tree on subsets of genes to be further analyzed: All genes with > 20 reads in both conditions (- and + ChX) were filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 0.85 or <- 0.85) by E2 in

the nucleus in presence of *de novo* protein synthesis (- ChX) and 2. significant regulation (worst FDR < 0.05) regulation (log2FC > 0.85 or <- 0.85) by E2 in absence of *de novo* protein synthesis in the nucleus (+ ChX) **B 178 of E2 target genes are significantly regulated in presence and absence of *de novo* protein synthesis**. Scatterplot showing the positive correlation of log2FCs of the E2 co-regulated (among them the 178 significantly regulated genes +ChX out of the 4,679) genes in the cytoplasm vs. nucleus (not filtered for significance in the samples + ChX for plotting). Log2FCs of filtered genes were plotted using R**. C Only 4 % (178) of E2 target genes are also regulated in absence of *de novo* protein synthesis.** 174 genes are regulated in absence of *de novo* protein synthesis in the same direction (co-regulated) as with normal translation. Venn diagram showing the proportion of co-regulated genes. The intersection of regulated genes expressed in presence and absence of *de novo* protein synthesis was plotted using R. **D The number of lncRNAs is increased in the subset of genes which are co-regulated**. Vertical slices showing the fraction of biotypes (eRNAs defined as RNAs with ≥ 1 bp overlap with enhancer signature -1000 bp +100 bp around TSS) of regulated genes in presence of translation (left) and co-regulated genes in presence and absence of translation (right). Plots were created using Graph Pad Prism. **E E2 regulated genes are enriched for enhancer marks independent of ChX**. Donut plots showing the fractions of genes with different CSS at the TSS flanking region of their derived transcripts (-1000 bp +100 bp; transcripts with $F_{rank} \leq 2$; ≥ 1 bp overlap of TSS flanking region with chromatin state). Genes not regulated by E2 in the nucleus (left) compared to genes regulated by E2 in the nucleus (right). CSS in GM12878 based on the definition by Ernst et al, 2011 (ENCODE project; other= insulator, polycomb repressed, heterochromatin, repetitive/CNV). Since multiple chromatin states can be associated per transcript, an exclusive classification was conducted (enhancer>promoter>TxN>other). Plots were created using Graph Pad Prism.

Next, we investigated whether the E2 regulated genes can be associated with E2 binding sites (computationally called peaks based on enrichment of aligned ChIP-Seq reads) and in which TF clusters these peaks reside. In order to do so, we intersected the E2 peaks obtained by ChIP-Seq (Glaser, PhD thesis, 2017) with i) the TSS flanking region [1], ii) the genebody [2], iii) a fragment looping to the TSS flanking region [3], iv) a TSS flanking region of another gene within the same TAD [4] or v) a related TAD [5] of the transcripts ($F_{rank} \leq 2$) derived from the E2 regulated genes (Figure 57A). The looping data is sourced from Mifsud *et al.* (Mifsud et al., 2015). It has to be mentioned that the promoter and "other" (not promoter) fragments are very large (Figure S22), more than the half are >10kb. This implies that the chance of a positive intersection of one of these fragments with an interval (genomic region) of choice is increased. Since possibly multiple E2 peaks exist per transcript, classification was exclusive (i>ii>iii>iv>v). Two control subsets of genes were examined, genes which are not covered with > 20 reads in the nucleus in presence of translation (Figure 57A top, 14,335 genes, not shown), and genes not significantly regulated by E2 in the nucleus in absence and presence of translation (Figure 57A bottom left). It can be observed that the number of genes not harboring a peak associated to their related transcripts decreases and the number of genes harboring a peak associated to the TSS flanking region or the genebody of their related transcripts increases. Comparing this to the genes significantly regulated by E2 in presence (Figure 57A, bottom middle) or presence/absence (Figure 57A, bottom right) of translation, it can be noticed that the proportion of genes not harboring a peak associated to their related transcripts further decreases to a minimum of 16 % for the co-regulated genes and the proportion of genes harboring a peak at the TSS flanking region or at the genebody of their related

transcripts further increased to over 60 % for the co-regulated genes -/+ ChX. Thus, E2 regulation can be linked to E2 binding in presence and even better in absence of *de novo* protein synthesis. Intersecting the E2-peaks associated with the transcripts ($F_{rank} \leq 2$) of E2 regulated genes with the cluster-assigned panEBNA peaks (E2, E3A and E3C peaks), we intended to reveal, which anchors E2 possibly utilizes to regulate their target genes (Figure 57B). The cluster distribution of all peaks associated in one of the investigated modes (i-v) with transcripts of the genes in our data sets shows that 26 % of the peaks reside in cluster I, 16 % in cluster V and 13 % in cluster VIII to name the three biggest clusters (Figure 57B, top left). 11 % of the peaks associated with the related transcripts were not assigned to a cluster. The cluster allocation was done for the peaks associated with transcripts derived from E2 regulated genes –ChX (Figure 57B, middle panel) or -/+ ChX (Figure 57B, right panel) for all genes (top) for other lncRNAs (middle) and eRNAs (bottom; eRNAs= lncRNAs with enhancer chromatin state at their TSS flanking region). “No ass. cluster” (no assigned cluster) means no associated peaks which could be assigned to a cluster. Tracking cluster I across the different subsets, it can be recognized that for most of the genes (35-73 %) the peaks associated with the related transcripts can be allocated to this cluster. The treatment with ChX enhances this occurrence as well as the selection for eRNAs in both ChX conditions. From these data we can speculate that for the regulation of most of the “direct” target genes, especially the eRNAs, E2 occupies binding sites of cluster I, positive for CBF1, EBF1 and CUX1 and high for H3K4me1 and H3K27ac.

From these data we can conclude that only a minority (4 %) of E2 target genes can be regulated with absent *de novo* protein synthesis. Among the potentially directly regulated genes are also lncRNAs, which largely emanate from enhancer marked chromatin. E2 regulation can be linked to E2 binding in presence and, even better, in absence of *de novo* protein synthesis. As expected, E2 occupied binding sites of cluster I, positive for CBF1, EBF1 and CUX1 and high for H3K4me1 and H3K27ac, may be responsible for the regulation of E2 target genes.

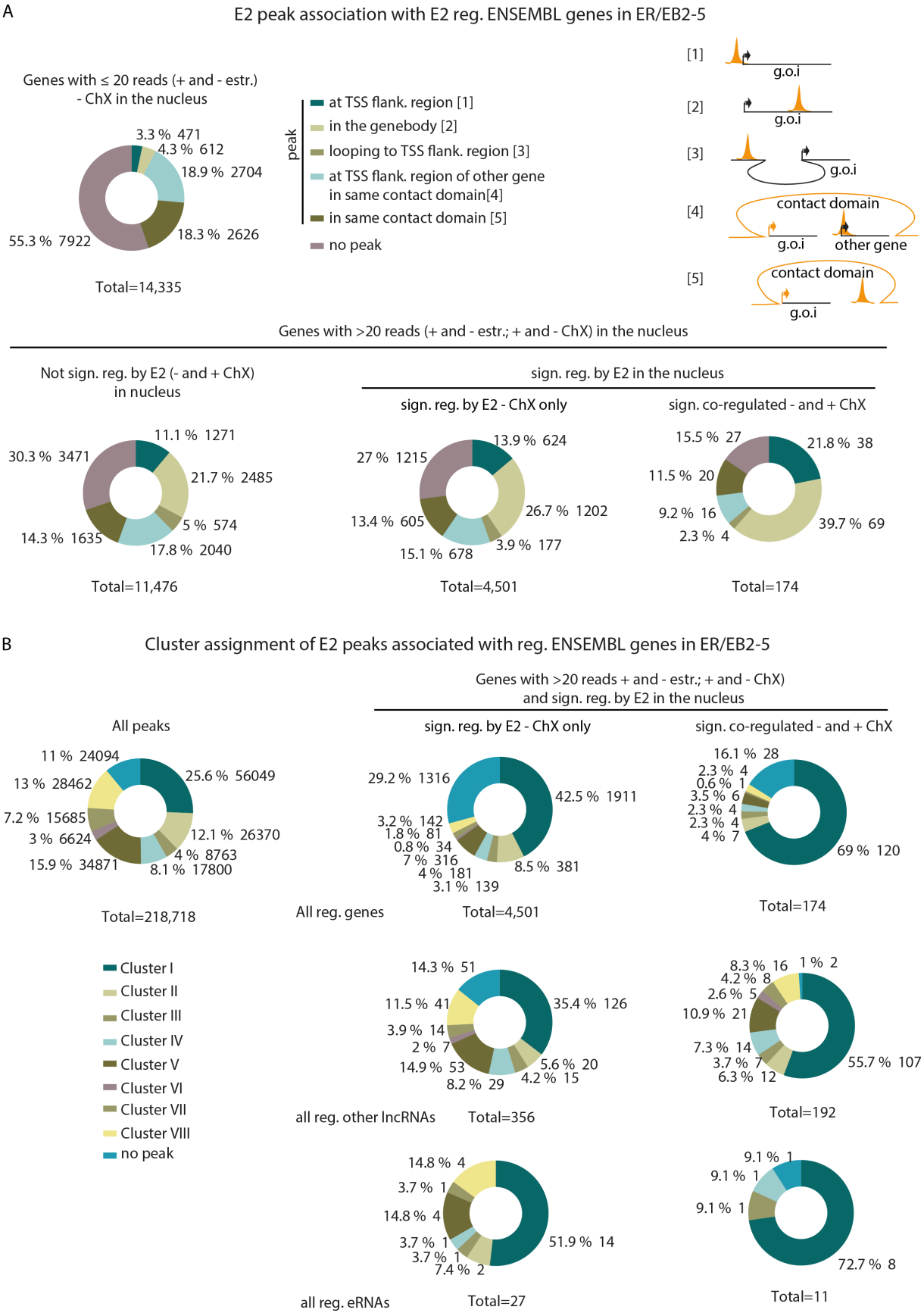


Figure 57: Characterization of E2 regulated ENSEMBL genes in absence of *de novo* protein synthesis
A Transcripts of genes co-regulated in presence and absence of *de novo* protein synthesis are enriched for E2 binding (at TSS flanking region or in genebody). Donut plots showing the fraction of genes with different

associated E2 binding sites at their derived transcripts in different gene subsets. Genes with ≤ 20 reads or not regulated by E2 in the nucleus (left) compared to genes regulated by E2 in the nucleus (right). For a positive connection, E2 peaks had to overlap ≥ 1 bp with the investigated genomic positions clarified in the following: We differentiated between [1] genes with a peak in the TSS flanking region of the derived transcripts ($F_{rank} \leq 2$), [2] genes with a peak in the genebody of the derived transcripts ($F_{rank} \leq 2$), [3] genes with a peak on a fragment looping to the promoter of the derived transcripts ($F_{rank} \leq 2$; Mifsud et al., 2015), [4] genes located in the same contact domain with a transcript occupied by a peak at the TSS flanking region and [5] genes located in the same contact domain with a peak. Since multiple peaks can possibly be associated per transcript, an exclusive classification was conducted ([1]>[2]>[3]>[4]>[5]; g.o.i= gene of interest). **B E2 peaks associated with genes co-regulated in presence and absence of *de novo* protein synthesis are enriched for Cluster I**. Donut plots displaying the cluster distribution of E2 peaks associated with transcripts ($F_{rank} \leq 2$) corresponding to genes in the different gene subsets. Intersection of E2 peaks connected to E2 regulated genes in (A) with EBNA peaks in cluster analysis (Glaser, PhD thesis, 2017). Cluster distribution of all E2 peaks connected to transcripts derived from genes covered with ≥ 1 peak in our data set (left) compared to peaks associated with regulated gene subsets (right; upper panel= all genes, middle panel= all other lncRNAs and lower panel= eRNAs). E2 peaks were identified by ChIP-Seq conducted in our lab (Glaser, PhD thesis 2017). Graph Pad Prism was used for plotting.

Intergenic genes

For the intergenic genes, the 8,918 intergenic genes were filtered for their read coverage in the samples +/- estr. and ChX+estr/ChX- estr. (Figure 58A). 3,623 genes were covered with >20 reads per gene in either one of both conditions. The remaining genes were subsequently selected for significant (worst FDR < 0.05) regulation (log2FC > 1 or < -1) in the nucleus in the absence of ChX (decision [1] in the filter tree, Figure 58A). 372 genes remained, which were then screened for regulation in the presence of ChX (decision [2] in the filter tree, Figure 58A, Figure 58B). Only 9 genes withstood the filtering, all of them are co-regulated.

Also for these genes, we investigated whether the E2 regulated genes can be associated with E2 binding sites (Figure 58C). For the control subset of genes, genes which are not covered with > 20 peaks in the nucleus with present *de novo* protein synthesis (Figure 58C top, 14,335 genes, not shown), and genes not significantly regulated by E2 in the nucleus (Figure 58C bottom left), it can be observed that the distribution of classes of genes with peaks associated to their related transcripts looks similar. Comparing this to the genes significantly regulated by E2 in presence (Figure 58C, bottom middle) or presence/absence (Figure 58C, bottom right) of translation, it can be noticed that the proportion of genes not harboring a peak associated to their related transcripts decreased slightly for the regulated genes and the proportion of genes harboring a peak at the TSS flanking region slightly increased. For the nine co-regulated genes, the transcripts of all of the genes can be linked with a peak, six of them via the same TAD. The chromatin state at the TSS flanking region (-1000 bp, +100 bp) of the transcripts belonging to the E2 regulated intergenic genes in presence and absence of translation was also determined (data not shown), however, the majority of genes were not assigned with a chromatin state, the other genes were enriched for the TxN chromatin state as the ENSEMBL genes.

Using these stringent cutoffs, we detected only nice intergenic transcribed genes regulated by E2 in the presence of ChX. E2 regulation can be linked to E2 binding in presence and even better in absence of *de novo* protein synthesis.

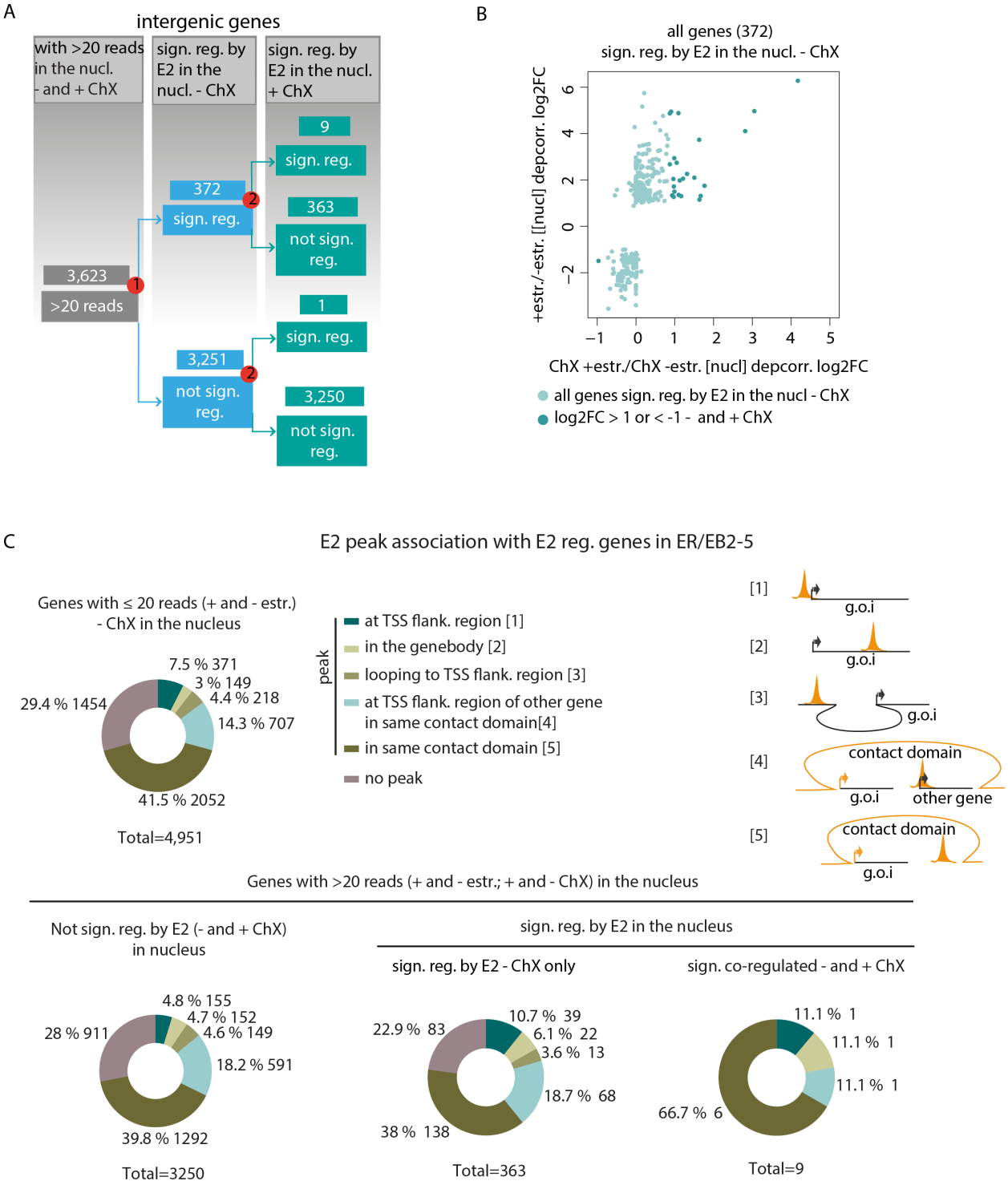


Figure 58: Characterization of E2 regulated intergenic genes in absence of *de novo* protein synthesis
A Decision tree on subsets of genes to be further analyzed: All genes with >20 reads in both conditions (- +ChX) were filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 1 < - 1) by E2 in the nucleus

and 2. significant (worst FDR < 0.05) regulation (log2FC > 1 < - 1) by E2 in absence of *de novo* protein synthesis in the nucleus **B Nine (of 24 regulated) E2 target genes are significantly regulated in presence and absence of *de novo* protein synthesis**. Scatterplot showing the positive correlation of log2FCs of the E2 co-regulated (among them the 9 significantly regulated genes +ChX out of the 4,679) genes in the cytoplasm vs. nucleus (not filtered for significance in the samples + ChX for plotting). Log2FCs of filtered genes were plotted using R. **C Genes co-regulated in presence and absence of *de novo* protein synthesis are slightly enriched for E2 binding.** Donut plots showing the fraction of genes with different associated E2 binding sites in different gene subsets. Genes with < 20 reads or not regulated by E2 in the nucleus (left) compared to genes regulated by E2 in the nucleus (right). For a positive connection, E2 peaks had to overlap ≥ 1 bp with the investigated genomic positions clarified in the following: We differentiated between [1] genes with a peak in the TSS flanking region [2] genes with a peak in the genebody, [3] genes with a peak on a fragment looping to their promoter (Mifsud et al., 2015), [4] genes located in the same contact domain with a transcript occupied by a peak at the TSS flanking region and [5] genes located in the same contact domain with a peak. Since multiple peaks can possibly be associated per transcript, an exclusive classification was conducted ([1]>[2]>[3]>[4]>[5]; g.o.i.= gene of interest).

The majority of E2 and E3A regulated genes are counter-regulated

Since E2 is a well characterized transactivator of various genes and E3A is known to antagonize E2, we aimed to investigate this incidence genome wide. We want to emphasize that we are comparing a conditional cell system with a static cell system which can eventually lead to misinterpretations. Data for the nuclear compartment are reported in the following, the analysis of the cytoplasmic compartment lead to similar results (not shown).

ENSEMBL genes

For this analysis, again the 31,192 ENSEMBL genes were filtered for their read coverage in both systems, E2/E3A presence/absence (Figure 59A). 15,751 genes were covered with > 20 reads per gene in the nucleus of either of the two systems (+/- estr. or wt/mut). The remaining genes were subsequently selected for significant (worst FDR < 0.05) regulation (log2FC > 1 or < -1) by E2 in the nucleus (decision [1] in the filter tree, Figure 59A). 4,071 genes remained, which were then screened for regulation by E3A (decision [2] in the filter tree, Figure 59A). 1,058 genes are significantly regulated by E3A, 3,013 are not. Of the 1,058 genes, which are regulated by both TFs, 741 are counter-regulated, that are remarkable 70 % of shared targets (Figure 59B, C). The strongest by E2 induced gene (log2FC= 9.4) counter-regulated by E3A (log2FC= - 5.2) encodes the interleukin 1 receptor, type II (IL1R2). As expected, among the strongest E3A repressed (log2FC= - 8.2), counter-regulated by E2 (log2FC= 3.4) encodes for the chemokine CXCL10 (well-studied by the Kempkes laboratory). The 741 significantly counter-regulated genes can be separated in 118 lncRNAs and 623 other genes (mostly pcgs). Of the 118 lncRNAs, 99 genes as also expressed (> 20 reads) in the cytoplasm. Of the 99 genes residing in both compartments, 61 are also significantly counter-regulated in both compartments, 38 are nucleus specific (Figure 59B). To re-emphasize a special lncRNA in the MYC neighborhood, CCDC26 is one of the counter-regulated

lncRNAs. It was reported to play a role in different cancer types (Hirano et al., 2015; Peng & Jiang, 2016).

E2 is expressed early during infection, while E3A expression follows later, potentially repressing genes activated by E2. Under this assumption, the co-regulated genes should be all expressed (> 20 reads) in wt LCLs. Furthermore, if E2 and E3A operate in a competitive way and the genes are expressed (> 20 reads) in wt LCLs, it can be assumed that E2 prevailed over E3A (or both TFs exert an activating function on these genes). In contrast, if E3A showed a stronger repressive function than E2 activation (or both TFs exert a repressive function on the genes) genes are not expressed (< 20 reads) in wt LCL. Thus, most interesting for us was the question, if the detected co- and counter-regulated genes are expressed in wt LCLs. Of the not significantly counter-regulated genes (genes can be co-regulated or not regulated by one of the two TFs), 260 genes are indeed expressed in wt LCLs. The remaining 57 genes are repressed by E2 and even stronger repressed by E3A, that might be why they are not expressed in wt LCLs. For the 741 counter-regulated genes, 559 genes are still expressed in wt LCLs, while 182 are not expressed. We reasoned that for the 559 genes, E2's activation prevailed over E3A's repression and for 182 genes, E3A's repression prevailed E2's activation.

We wondered, whether the result of the 559 counter-regulated genes present in LCLs (genes regulated by both TFs, but E2 "won") can be mirrored in the binding behavior of both TFs, we tested the transcripts of those genes for E2 and E3A peaks (Figure 59D). Ideally, it should be possible to associate the 559 counter-regulated genes present in LCLs with both TFs. For 367 of the 559 genes, the transcript can be associated (at TSS flanking region, etc.) with at least one of both TFs, which equals 66 %. The transcripts of 215 genes are linked to an E2 binding site and 152 are linked to an E3A binding site, for 116 genes, the transcripts can be associated to both TFs. However, only at 27 of those 116 genes, both TFs reside in the TSS flanking region of the transcripts. The 741 counter-regulated genes were subjected to a GO-Analysis (Gergely Csaba), where we tested for an enrichment of biological processes compared to a background set of genes. Interestingly, we observed a substantial amount of GO terms linked to development or morphogenesis as well as differentiation and proliferation (Table S2). Only a small amount of terms are linked to immune response.

Taken together, we found 70 % (741) of the E2 and E3A shared target genes to be counter-regulated, which includes 118 lncRNAs. 75 % of the counter-regulated genes are expressed in wt LCLs, which could indicate that E2's activation prevailed over E3A's repression. For 66 % of the counter-regulated genes expressed in wt LCLs, a transcript can be associated with at least one of both TFs. The counter-regulated genes were enriched for processes such as development or morphogenesis as well as differentiation and proliferation.

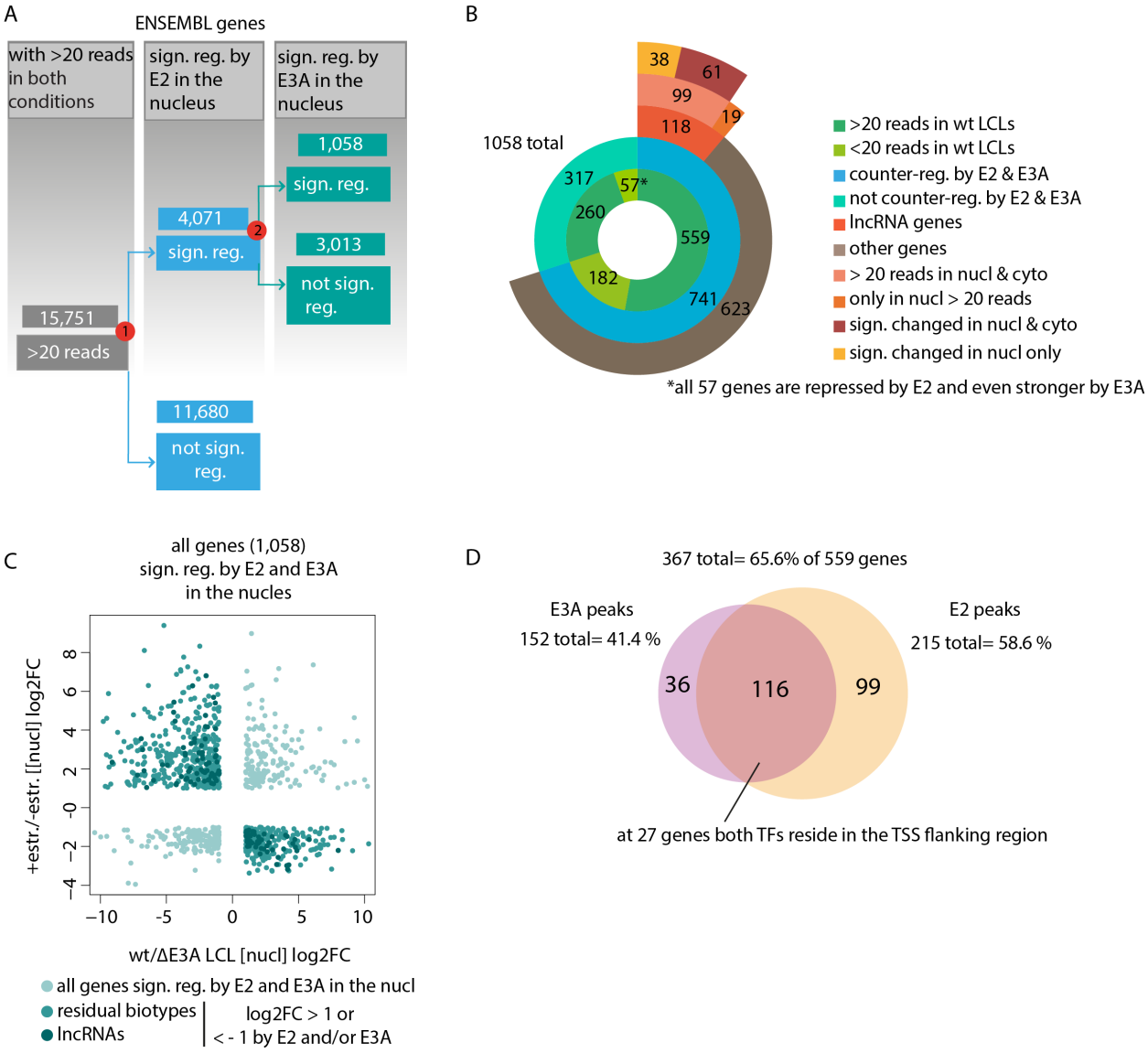


Figure 59: Characterization of E2 and E3A counter-regulated genes in the nucleus. A Decision tree on subsets of genes to be further analyzed: All genes with >20 reads in any condition were filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 1 or < -1) by E2 in any compartment and 2. significant (worst FDR < 0.05) regulation (log2FC > 1 or < -1) by E3A in the nucleus. **B 70% of shared E2 and E3A target genes are counter-regulated and the majority are expressed (> 20 reads) in wt LCLs.** Sunburst plot displaying different gene subsets indicated in the legend. **C** Scatterplots showing the anti-correlation of log2FCs of the shared E2 and E3A target genes (741 genes out of 1,058) in the nucleus (not filtered for significance) Log2FCs of filtered genes were plotted using R. **D 30% (116) of in wt LCL expressed, counter-regulated genes associated with a binding site (367) are occupied by both TFs.** Venn diagram displaying the occupation of counter-regulated genes, expressed (>20 reads) in wt, by E2 and E3A. 559 of 741 counter-regulated Genes are covered with > 20 reads in wt LCLs, for 65% binding sites were detected, 59% of them are E2 binding sites an 41% are E3A binding sites.

Intergenic genes

The 8,918 intergenic genes were also tested for counter-regulation. They were also filtered for their read coverage in both systems, E2/E3A presence/absence (Figure 60A). 3,084 genes were covered with > 20 reads per gene in either of the conditions. The remaining genes were subsequently selected for significant (worst FDR < 0.05) regulation (log2FC > 1 or < - 1) by E2 in the nucleus (decision [1] in the filter tree, Figure 60A). 309 genes remained, which were in the last step screened for regulation by E3A (decision [2] in the filter tree, Figure 60A). 64 genes are significantly regulated by E3A (Figure 60C). Of the 64 genes which are regulated by both TFs, 44 (~70 %) are counter-regulated. Of the 44 counter-regulated genes, 38 have transcripts linked to an E2 peak and 37 have an E3A peak linked to their transcripts (data not shown).

To conclude, also intergenic (towards ENSEMBL annotation) transcribed genes were observed to be counter-regulated by E2 and E3A up to 70 % of their shared target genes. Almost 90 % of the counter-regulated genes are occupied by E2 and E3A.

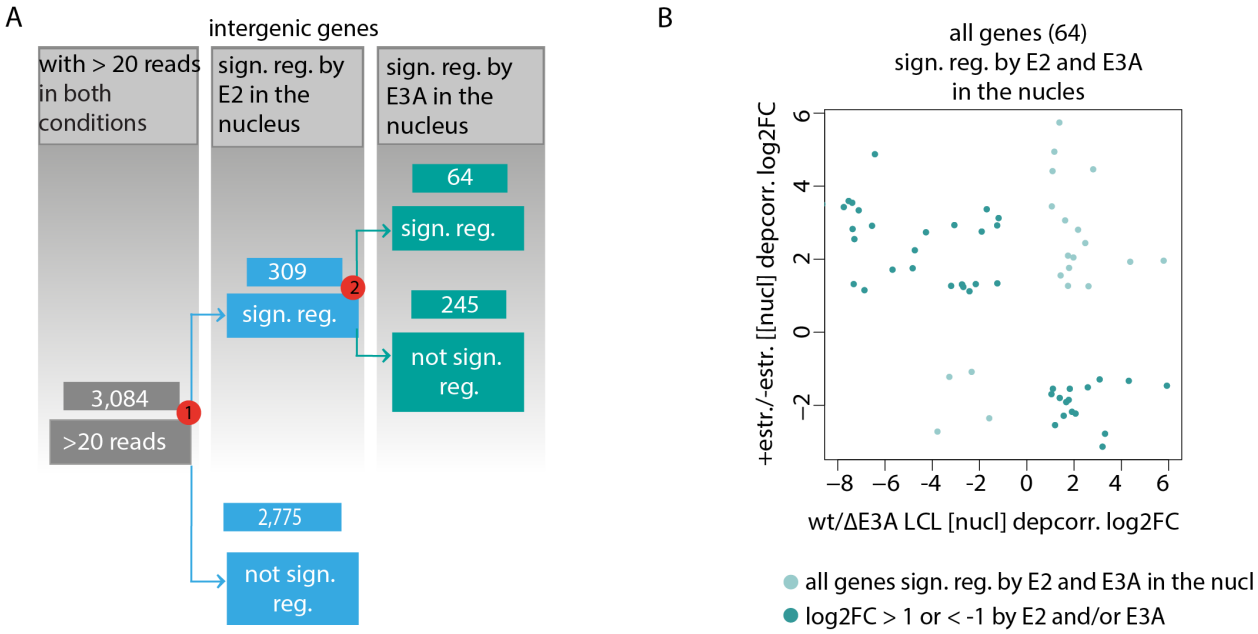


Figure 60: Characterization of E2 and E3A counter-regulated intergenic genes in the nucleus A Decision tree on subsets of genes to be further analyzed: All genes with >20 reads in any condition were filtered for 1. significant (worst FDR < 0.05) regulation (log2FC > 1 and < - 1) by E2 in any compartment and 2. significant (worst FDR < 0.05) regulation (log2FC > 1 and < - 1) by E3A in the nucleus. **B 70 % of the shared intergenic target genes are counter-regulated by E2 and E3A**. Scatterplots showing the anti-correlation of log2FCs of E2 regulated genes vs. E3A regulated genes in the nucleus. Log2FCs of filtered genes were plotted using R.

4 Discussion

4.1 E2 requires EBF1 to bind to its CBF1 independent binding sites

Microarray analysis revealed that E2 regulates cellular gene expression in absence of CBF1. ChIP-Seq investigations in DG75$^{doxHA\text{-}E2}$/CBF1 wt and CBF1 ko cells demonstrated that E2 binds to chromatin in absence of CBF1 (CBF1 independent binding). CBF1 is the most well studied binding anchor of E2. Bioinformatics predict EBF1 to act as an alternative E2 anchor to facilitate its CBF1 independent binding to chromatin. To confirm the computational predication we sought to determine the biochemical underpinnings. Molecular characterization of the contribution of EBF1 to CBF1 independent E2 binding will shed considerable light on E2 mediated transcriptional regulation.

To this end, we performed a siRNA-mediated knock down of EBF1 and immunoprecipitated EBF1 and E2 from isolated chromatin obtained from DG75$^{doxHA\text{-}E2}$ CBF1 wt or CBF1 ko B cells. Subsequently, the recovered genomic DNA was analyzed by RT-qPCR and three CBF1 dependent and independent E2 binding sites with certain requirements were investigated.
We observed EBF1 to be decreased at all investigated loci upon EBF1 knock down, while E2 binding was decreased only at CBF1 independent sites. This indicates that E2 requires EBF1 for chromatin interaction at CBF1 independent sites. More general, EBF1 binding is decreased in CBF1 ko cells compared to wt cells (E2 binding is decreased in a similar manner). Furthermore, the decreased binding of EBF1 upon EBF1 knock down was more profound in CBF1 wt cells compared to ko cells (Figure 9). As positive control, we confirmed the decrease of EBF1 binding upon its knock down (Figure 9A/B). E2 binding is diminished in CBF1 ko cells, since CBF1 is the key adaptor for E2 binding to chromatin. Our observation that in CBF1 deficient cells EBF1 chromatin binding was reduced at all tested loci indicates that CBF1 could be important in EBF1-E2 complex formation. The decrease in E2 binding of CBF1 independent sites upon EBF1 knock down suggest, that EBF1 is required for E2 binding to the CBF1 independent binding sites, thus, E2 binding to CBF1 independent sites appears partially EBF1 dependent. However, E2 binding could not be depleted completely upon EBF1 knock down at CBF1 independent binding sites. The incomplete knock down of EBF1, as seen in Figure 8A/B, may account for some of the E2 binding to CBF1 independent sites (Figure 9). Another reason could be an alternative TF. Unsurprisingly, the CBF1 independent E2 binding sites are depleted in clusters V to VII, since EBF1 binding is absent in clusters V to VII (Figure 7B, Figure S7). However, CBF1 independent E2 binding sites are enriched for cluster I,

which contains all correlated TFs (CBF1, EBF1 and CUX1) represented (Figure S7). Thus, CUX1 could be an alternative TF mediating E2 binding to CBF1 independent sights following EBF1 knock down. Furthermore this cluster was characterized as highly associated with the chromatin state of an enhancer in LCLs, as it shows high signals for H3K4me1 and H3K27ac. This may indicate that the CBF1 independent/EBF1 dependent binding of E2 to chromatin leads to increased enhancer binding (Figure S7).

It was revealed by our group that E2 binding to chromatin is strongly dependent on CBF1, since 86.4% of E2 binding sites are lost in CBF1 ko cells compared to the wt cells. Indeed, E2 signal strengths are much stronger at E2 binding sites in CBF1 wt than in ko cells (Glaser et al., 2017). Our observation of E2 binding loss in CBF1 ko cells is in line with previous findings indicating that CBF1 is the major anchor for E2 (reviewed in Kempkes & Ling, 2015). At CBF1 independent peaks, EBF1 was the only significant motif enriched as discovered by motif search (Glaser et al., 2017). Hence, we investigated E2 chromatin binding upon EBF1 knock down. The result that EBF1 promotes E2 chromatin assembly and that CBF1 might contribute to this assembly is consistent with a recent study (Lu et al., 2016), which found that the formation of EBF1-CBF1 co-occupied binding sites are promoted by E2 and that E2 recruits CBF1 and EBF1 to its target sites. Furthermore, EBF1 has been implicated as one of several cellular factors associated with the maintenance of latency of EVB (Davies et al., 2010). EBF1 is a crucial TF in defining B cell lineage specificity during differentiation (reviewed in Boller, Li, & Grosschedl, 2018). EBF1 was found to regulate genes associated to B cell function and most recently, it was shown to act as pioneer factor establishing chromatin accessibility (Boller et al., 2016). These findings support our hypothesis that E2 uses EBF1 to establish access to B cell enhancers to drive B cell activation.

We were not able to connect E2 target genes to E2 binding sites. E2 activates gene expression predominantly by binding to enhancers and not to the promoters (Glaser, PhD thesis, 2017; Zhao et al., 2011). Enhancers can influence transcription through long-range chromatin interactions (reviewed in Plank & Dean, 2014). Therefore, association of E2 target genes with E2 binding sites is highly challenging. The data for peak selection and cluster analysis were obtained from experiments in LCLs and we transferred this information to the DG75 cell line. The differences in E2 binding patterns observed between LCL and DG75 cell lines may be in part explained by differences in the chromatin landscape between cell lines. Since we lacked EBF1 ChIP-Seq data in DG75 cells, we only included binding sites which were positive for E2 and EBF1 in LCLs, assuming a similar chromatin landscape in DG75.

The correlation of genes regulated by CBF1 independent binding of E2 needs to be investigated by chromosome conformation capture techniques. Such would aide in revealing CBF1 independent biological pathways which E2 might regulate. Furthermore, although we could functionally link EBF1 to E2 binding we could not answer the question about a co-occupation of CBF1 and EBF1 by

E2. However, the role of E2 in CBF1 and EBF1 co-occupation has been explored by Lu *et al.* (Lu et al., 2016). The binding order of the involved TFs remains to be unraveled.

While, E2 has been previously identified to drive the genome-wide cooperative binding of EBF1 and CBF1 (Lu et al., 2016), this work demonstrates the major contribution of EBF1 at sites of CBF1 independent E2 binding. Our work has uncovered a previously unreported and significant role for EBF1 in E2 transcriptional regulation.

4.2 Analyses of cellular and viral genes regulated by E2 and E3A

The recent developments in transcriptional analysis by high-throughput RNA sequencing has provided extensive and rapid expanding catalogues of lncRNA. While significant advances have been made recently, much remains unknown about the expression and functional activity of lncRNAs. The highly tissue specific expression of lncRNA along with the ongoing co-evolution of high-throughput sequencing and bioinformatics algorithms are powering this rapid expansion of knowledge in the area. Microarray analysis revealed E2 dependent regulation of lncRNAs. The genome-wide regulation of lncRNAs by E2 remains an understudied area. E2 is the master regulator of transformation and maintenance of latency following EBV infection. Developing an understanding of the genome-wide E2 regulation of lncRNAs would extend our understanding of the regulatory network of E2 considerably.

The overall aim of the second part of this thesis was the genome-wide identification of E2 and E3A viral and cellular target genes by comparing the transcriptomes of E2 or E3A proficient and deficient human B cell lines. We have shown E2 to regulate the expression of several lncRNAs. To determine the sub-cellular localization of E2 regulated lncRNAs and to enrich for lncRNAs in the nucleoplasm, harvested cells were fractionated into cytoplasm and nucleoplasm. E2 is renowned to induce targets which are themselves strong transcriptional regulators, which adds complexity to transcriptomic analysis of E2 induced gene expression. To address this, we conduced RNA-Seq of cells with induced E2 in the presence or absence of the translational inhibitor cycloheximide (ChX), allowing us to distinguish between E2 targets requiring competent *de novo* protein synthesis for induction. RNA was isolated from the desired cell line and conditions and cDNA libraries were prepared. Transcriptional changes were investigated by RNA-Seq experiments with subsequent confirmation of candidate targets by RT-qPCR.

4.2.1 78 % of EBVs genes can be regulated by E2 in the ER/EB2-5 system

An additional benefit of the transcriptome analysis of LCLs is the integrated EBV transcriptome. Thus, we briefly identified all viral target genes especially regulated by E2 with RNA-Seq. The analysis of the data provided by RNA-Seq regarding viral target gene regulation was conducted in collaboration with the Cancer Crusaders Next Generation Sequence Analysis Core, Tulane University, New Orleans, USA.

In the ER/EB2-5 system, the proliferation of the cells was blocked for 4 days and reinduced again. For E3A regulation, two static systems were compared. But yet under these unphysiological conditions, we sought to identify E2 and E3A regulated viral genes. Therefore, reads were aligned to the Akata genome using the STAR mapper. For counting and DE-testing, the RSEM package was utilized.

We found 73 out of 92 genes upregulated by E2 in the cytoplasm, 47 in the nucleus. 46 genes were regulated with blocked *de novo* protein synthesis. Intriguingly, regulated genes included lytic genes, which are usually not expressed during latency. Surprisingly, almost 80 % of the viral transcriptome was modified by E2 (Figure 12). Provided that the ER/EB2-5 system resembles the process during infection, it may be that E2 is able to induce the lytic cycle, since the genes, responsible for the lytic switch, BZLF1 and BRLF1 are also induced by E2 (Figure 12, Table S3). Preliminary data indicate E2-dependent induction of these genes as confirmed by Western Blot analysis on protein level (data not shown; n=1). Attempts to infect and transform primary B cells with ER/EB2-5 supernatants failed, which would have supported the assumption of a productive lytic cycle (data not shown). This may indicate that E2 induces an abortive lytic cycle. A lytic cycle is considered as abortive, if the cycle was not completed resulting in induced cell lysis and the release of novel infective viral particles. In order to confirm the detection of regulated viral genes, three genes induced by E2 were investigated by RT-qPCR. RNA was prepared of total and fractionated cell lysates and reverse transcribed. We confirmed that all genes were regulated by E2, independent of an E2 binding site in proximity, both with present or absent *de novo* protein synthesis (Figure 13 to Figure 18). While BGRF1/BDRF1 (Figure 15 upper panel) and LMP2A (Figure 17 lower panel) transcripts are enriched in cytoplasm, BHRF1 is enriched in nucleus (Figure 15, lower panel), the localization for BNRF1 determined by RNA-Seq and RT-qPCR is contradictory (Figure 17 upper panel). The involved RNAs are linked to important processes throughout the entire life cycle of EBV. These data support the results obtained by RNA-Seq, where we observed immediate early, early, intermediate, late and latent genes to be regulated by E2. Here, we confirm the upregulation of genes involved in EBV invasion, virus production, prevention of apoptosis and mimicry of B cell signaling. LCLs are known to express lytic genes (see above), but it is unknown whether this expression is directly induced by E2. Co- and counter-regulation of viral genes by E2

and E3A was determined by comparing the datasets obtained by RNA-Seq analysis of both TFs. We observed that E3A acts as a transcriptional activator of viral genes. 42 genes were induced by both TFs, thus, more than 50 % of E2 targets are additional upregulated by E3A (Table S3). It appears that the induction of lytic genes by E2 is amplified by E3A. E2 induces E3A (log2FC in cytoplasm 6.4) and they regulate lytic genes, like BZLF1, BRLF1 or BNRF1 in concert. Thus, both latency associated TFs are involved in regulation of genes which could promote tumorigenesis. E3A is considered an oncogene and might play a role in cell cycle regulation. It is controversial, if E3A is essential for B cell transformation, since E3A mutants can still give rise to LCLs. However, it is important to the transformation process (Allday et al., 2015) and we observe a strong upregulation by E2.

Recently, evidence has emerged that the lytic cycle can contribute to EBV-induced oncogenesis. It was reported that tumor growth could be promoted by low numbers of lytically infected cells. Additionally, BZLF1 can provoke genomic instability. Furthermore, several lytic genes, such as BILF1, BALF4 and BHLF1, have been shown to be expressed at a high level in cell lines and tumor biopsies (reviewed in Young et al., 2016). We found all three genes to be induced by E2. A recent review suggests, that “an abortive lytic cycle would reconcile the need of lytic expression for viral tumorigenesis” (Morales-Sánchez & Fuentes-Panana, 2018). Nevertheless, LCLs can also undergo some level of spontaneous lytic reactivation, with rates varying among LCLs (Adhikary et al., 2007). Furthermore, it has been observed that LCLs express detectable levels of transcripts for early lytic genes. Davies *et al.* propose that each line of Latency III-transformed B cells has a characteristic frequency of EBV lytic reactivation (Davies et al., 2010), which would be in line with a productive lytic cycle. It does not, however, mean that E2 triggers this reactivation. Whether E2 induces a productive lytic cycle or whether the cycle is abortive remains to be determined.

This may indicate that by upregulation of lytic genes during establishment and maintenance of latency, EBV promotes a tumorigenesis. Alternatively, the transient expression of viral lytic genes upon infection could be initiated by E2.

4.2.2 E2 regulates lncRNA which are contained in co-regulated gene blocks (CRGB), are associated with malignancies and could regulate protein coding and non-coding genes.

Microarray analysis of E2 regulation and E3A regulation was conducted and revealed numerous cellular target genes of E2 and E3A. Furthermore, shared gene subsets of E2 and E3A were already uncovered. Nevertheless, the identification of all cellular coding and non-coding target genes of E3A and especially E2 by deep sequencing is still pending.

4.2.2.1 Analytic approach is fundamental for the identification of differentially expressed genes

The analysis of the cellular target gene regulation data generated by RNA-Seq was performed in collaboration with the Teaching and Research Unit Bioinformatics of LMU University Munich. Analysis of RNA-Seq requires a concatenation of different tasks, basically mapping, counting and DE- testing. In order to align the obtained reads with their corresponding genomic origin, alignment was performed using four different splice aware mapper ContextMap2, STAR, HISAT, and TopHat2, all of which allow inference of introns. The primary RNA-Seq data obtained for all technical replicates displayed good quality, mirrored by the percentages of mapped reads in general aligned with the human genome hg19 (Figure S8). We observed variations in alignment of reads towards different genomic regions between nucleic and cytoplasmic regions, as well as variations between the mapper for some samples (Figure S8- Figure S13). In keeping with our expectations, the nucleoplasm was enriched for immature unspliced transcripts and the cytoplasm is enriched for mature, spliced transcripts which is reflected in the percentages of reads mapping to complete transcripts (Figure S9) and junctions (Figure S12/ Figure S13) or intronic and intergenic regions (Figure S10/ Figure S11). The variation between mappers is most likely due to differences in their algorithms and underpinning assumptions.

The comparison of biological replicates on read aligned level showed high similarity, indicating that the cells were not altered substantially between 0 h and 6 h post E2 reactivation or by E3A expression (Figure 20- Figure 22). Focusing on the E2 conditions, nuclear samples were highly correlated (r~0.8) between 0 h and 6 h estrogen. This was particularly profound in samples 0 h and 6 h post E2 reactivation where *de novo* protein synthesis was inhibited (r~0.9; Figure 21), indicating that fewer changes occur. Furthermore, the disorganized clustering of the ChX- treated samples may indicate a substantial impact of ChX on the samples. However, clustering was observed dependent on E2 and E3A proficiency and according to the sub-cellular compartments. Interestingly, in the E2 system diversity between the compartments is responsible for the major clusters, while for E3A diversity between the cell lines is responsible for the major cluster. This may

be explained in one or more ways. For the E3A system, two different cell lines were compared, which could result in E3A independent differences. Furthermore, selection following E3A knock out may have resulted in a cell population transcriptionally adapted to the loss of E3A. In contrast, treatment with estrogen for 6 h did not lead to dramatic changes. The ER/EB2-5 cells may reflect a heterogeneous population of not entirely synchronized cells which could further impact clustering.

To assess the number of reads deriving from each gene, reads were counted using an in-house tool. Genes served as counting features in two ways: either, all gene aligned reads were simply counted or reads were only counted for genes which are supported by reads aligning to a derived transcript. This is a very sensitive approach and bears advantages and disadvantages as discussed below. The read counts were down-sampled to account for PCR amplification. By PCR amplification, multiple identical copies (duplicates) can arise. For example, smaller fragments are easier to PCR amplify and end up over-represented without biological meaning. Therefore, the fraction of PCR duplicates were kept low. When comparing the biological replicates by read count level (Fig, S14-S18) we again observed high read count variations between the biological replicates for the ChX-treated samples.

Initially, we expected to dissect direct and indirect E2 targets by the treatment of the cells with the translation blocker ChX. Since non-coding targets do not undergo translation, their functions are not affected by ChX and they still can mediate the transcription of secondary targets. Hence the clear dissection of direct (primary) and indirect (secondary) targets by ChX has become unfeasible. Additionally, ChX seems to have a major impact on RNA metabolism, leading to high variations between the biological replicates.

ChX, originally isolated from *Streptomyces griseus*, is the most common used chemical reagent to inhibit protein synthesis. It is reported to block the elongation phase of eukaryotic translation by binding to the "E site" of the 60 S ribosome and prevents eEF2-mediated translocation (Schneider-Poetsch et al., 2010). Usually, mRNA is post-translationally degraded after deadenlyation of the PolyA-tail. Once the tail is shortened, mRNA is subject to decay by exosome (3' → 5') or decapping and (5' → 3') Xrn1p digestion. Reduction in mRNA translation upon intrinsic stimuli normally potently induces mRNA decay. However, extrinsic inhibition of translation due to stress can result in mRNA stabilization. Numerous studies reported reduced mRNA decay following ChX treatment, suggesting that ribosomal association of mRNA inhibited mRNA degradation. ChX affects mRNA decay by indirectly preventing mRNA decapping (reviewed in Huch & Nissan, 2014). The variation between the cytoplasmic samples were so high we excluded them from this analysis. The variation of samples from the nuclear compartment was acceptable. However, we observed stabilization of certain mRNAs upon ChX treatment in the nucleus in the RNA-Seq results. One could imagine backlog, where transcripts accumulate also in the nucleus since they are not exported due to blocked processing. Critically, this could also hold true for cytoplasmic lncRNAs. Very recently it

was shown that lncRNAs are associated with active ribosomes. Approximately half of all expressed lncRNAs were detected in the cytoplasm with most of them associated with ribosomes. These lncRNAs have further be shown to be stabilized in response to ChX, in a similar fashion to mRNAs (Carlevaro-Fita, Rahim, Guigó, Vardy, & Johnson, 2016). The ER/EB cells may not be exactly synchronized. Thus, stabilization of different transcripts at different levels by ChX could be the reason for variations between the replicates. Furthermore, it has to be considered that slight variances in the treatment of the cells of the different biological replicates can be responsible for variances in the transcriptome.

DE- testing was conducted by applying four different DE-methods: edgeR, limma, DE-Seq and a2a, each using different normalizing steps or different statistical models. For each required task of RNA-Seq analysis, a variety of tools exist, which all have different approaches and therefore different outputs. The detection of a set of differentially expressed genes could thus be dependent on the selected tools (reviewed by Costa-Silva, Domingues, & Lopes, 2017). All methods are tested using modelling, simulations and benchmarks (default values). Every individual experiment needs adjustments, there are no clear instructions to find the "optimal" method. One approach is to analyze the results in depth and make decisions one-by-one (here gene-by-gene) or - if one looks for the clearest signals - one can take the "robust" approach and combining multiple methods and taking only results if proposed by all (or all reasonable) combinations. We used the consensus from 72 combinations (Figure 23) in order to identify differentially expressed genes tool-independently. Only for the cytoplasmic samples of the ER/EB2-5 cells, use of the mapper ContextMap2 was disregarded because it resulted in a huge decrease in detected genes (48 combinations), we have no explanation for this (Figure 24A). The DE-method a2a was most effective in addressing the large variation between the ChX-treated samples, thus we relied only on this DE-method for this condition (18 combinations; Figure 24C. By applying this consensus-technique, we robustly detected E2 and E3A regulated protein coding and non- coding genes as displayed in volcano plots (Figure 27/Figure 31). Furthermore we were able to extract differentially expressed genes in intergenic and intronic transcribed regions of ENSEMBL annotated genes after mapping (Figure 28/Figure 29, Figure 32/Figure 33). These were also submitted to DE-testing. To exclude reads from any other origin, correction for background and proximal transcribed genes was performed as described in section 3.2.2.2.2 p. 58. Many more transcripts were upregulated than downregulated by E2, in line with results from microarray analysis conducted in estrogen dependent E2- inducible DG75 cell lines (Glaser et al., 2017). In line with results from microarray analysis performed using the same cell lines as in this study, the numbers of up- and downregulated E3A target genes were similar (Hertle et al., 2009). However, upregulation of cellular genes by E3A is not well studied.

Usually, one would use *de novo* synthesis algorithms to identify novel transcripts. These tools predict novel transcripts based on the read coverage. Applying these tools was not possible for the nucleic samples due to the large number of immature transcripts which leads to the prediction of

myriad isoforms. This is why we came up with an alternative strategy to infer unannotated differentially expressed genes as mentioned in section 3.2.2.2.2, p. 58.
Detection of regulated genes is generally strictly dependent on applied thresholds and definitions. During analysis we encountered two major definition related difficulties, i) the definition of read origins and ii) the thresholds for detection of novel regulated loci.

i) First, one has to define the reads which shall be counted for a specific feature. Both, read pairs mapping entirely to the region of a gene and read pairs mapping to a gene plus read pairs mapping to the corresponding transcripts (if any transcript with supporting reads is available) were counted. Taking all reads into account, the detection achieved is sensitive and comprehensive. Generally, introns and exons of the same transcript are transcriptionally well correlated and since introns are typically much longer, many reads may be distributed over the whole intron. Thus, introns count can be utilized as an approach for gene coverage estimation where transcribed introns are the byproduct of gene transcription. Nevertheless, if intronic regions are transcribed without exonic regions, these reads would be lead to false positive gene transcript determination. Indeed, it is likely that to some degree such intronic transcription was regulated by E2 and E3A at certain loci in our studies. Such transcription lead to over-estimation of transcript levels, e.g. ART3 shows intronic transcription with few reads aligning to the transcript, leading to over-estimation of expression. This could be corrected by only selecting those genes which reach a certain threshold of reads on the gene and on the transcript level for further analysis.

ii) Since the thresholds for detection of novel intergenic or intronic genes were tremendously conservative, our false positive detection rate of novel loci will be very low. However, it is likely that a significant false negative rate was apparent in our analysis. This is exemplified in the missing detection of novel genes from chromosome four to nine, since no region was detected consistently differentially expressed by all tool combinations.
We did the following, as described in section 3.2.2.2.2 obenp.on page 58:
First, candidate intergenic regions were derived from each sample independently. These regions were defined by proximal (≤ 50 bp from closest fragment ends) or partially overlapping fragments not mapping to any ENSEMBL gene. Since strand specific sequencing was conducted, regions antisense to overlapping known genes were also collected. The second step was to combine these „raw“ intergenic regions to build a replicate-consistent panel. All intergenic candidates having ≥ 50 supporting reads were merged and the part of the intergenic region consistently detected by all

mappers in all biological replicates for the same condition was extracted. If two regions from different conditions overlapped (≥ 1 bp), one of them was discarded defining them as not sufficient different. One of the mappers used did not align the required ≥ 50 reads to any region of chromosomes four to nine.

Furthermore, we corrected for transcriptional regulation of other known genomic elements by comparing the changes of the novel intergenic/intronic regions to the changes of the overlapped gene in the differential analysis. Therefore, we dismiss alternative explanation for regulation, concluding that the expression changes we detect in intergenic/intronic regions were indeed specifically induced by E2 or E3A. In this thesis, we corrected for background based on gene level. This is more stringent compared to transcript level correction. On the gene level, reads also mapping to intronic regions of the gene were counted. For example, the ART3 gene appears to be regulated, since its intron is regulated. Over estimation of gene expression by determinations utilizing intronic counts has been discussed in i). Since an intronic region was corrected for background transcription on gene level and the underpinning gene is similarly regulated, the significant regulation of the intronic region was annihilated. Enhancer derived intronic (ART3) transcription was readily detected by RT-qPCR at the model locus CXCL9/10 (see Introduction), thus the stringency has to be relaxed for future analysis. Using a dependency correction on transcript level data it should be possible to detect the intergenic (CXCL9/10) and intronic (ART3) transcripts at our model locus of CXCL9/10. For future analysis we recommend the use of less stringent thresholds for read coverage and correction on transcript level on transcript level.

In summary, the thresholds applied in this work resulted in a very stringent and conservative determination of intergenic and intronic transcripts, which lead to false negative rate for novel transcripts but at the same time to a high reliability for the differentially detected genes. The assumption, that intronic transcription has to be derived from already confirmed genes is misleading and has to be reassessed for single genes on transcript level.

4.2.2.2 Discovery of co-regulated gene blocks (CRGBs)

Higher order chromatin structures can result in domain-wide transcriptional activation of genes. There is evidence that E2 binds superenhancers residing within TADs. Furthermore, it was reported, that EBV has local effects on the chromatin organization. Whether EBV exploits other higher order chromatin structures to regulate its target genes remains to be elucidated.

To investigate the potential global regulation of whole chromatin domains block-like by EBV, all E2 and E3A significantly regulated genes were screened *in silico* for clustering. We observed that approximately 80 % of E2 and E3A regulated genes are organized in blocks of up to ten genes (Figure 35). Approximately 80 % of the blocks regulated by E2 and approximately 70 % of gene blocks regulated by E3A genes are smaller than 2 Mb (Fig. S19). These findings may indicate that E2 and E3A regulate genes on a higher level of genome organization. This would mean that EBV makes use of higher order chromatin units by binding there to open chromatin and regulating whole associated chromatin domains genome-wide. E2 local action has been shown to activate a chromatin domain in the neighborhood of MYC by rearranging enhancer-promoter loops (Wood et al., 2016). To determine the role of E2 in regulating gene expression via chromatin rearrangements, a study could map intergenic E2 binding sites to E2 up-regulated genes using the HiC-technique (Zhao et al., 2011). It has also been reported for E3A that it can repress a whole chromatin domain in the neighborhood of CXCL10 (Harth-Hertle et al., 2013). However, when intersecting the detected E2 regulated CRGB with the published contact domains by Rao *et al*., we observed only 30 % of the CRGB residing entirely in contact domains (data not shown). This indicates that CRGB and contact domains may differ in their nature. The hierarchy of genome organization contains from large cell type specific chromosome territories, intermediate tissue invariant TADs, down to chromatin looping. In contrast to the tissue invariant TADs, the chromatin states within the TADs are highly diverse in different cell types or conditions, leading to a modulation of activity and compartment association, as well as differential chromatin loop formation (Dekker & Heard, 2015). The size of the detected by E2 and E3A regulated blocks fits the published size of TADs ranging between 40 kb and 2 Mb when they were first discovered in embryonic stem cells (Dixon et al., 2012), but the genomic intervals are different to published contact domains in GM12878 (Rao et al., 2014). However, contact domains observed by Rao *et al*. do not resemble reported TAD size. Different data on architectural structure exists for LCLs (GM12878) on different hierarchical level. This includes the CTCF-mediated architecture obtained by ChIA-PET (Tang et al., 2015a), “chromosome territories” detected by Hi-C (Selvaraj, R Dixon, Bansal, & Ren, 2013) “contact domains” (median length 185 kb; Rao et al., 2014) and chromatin looping between promoters and enhancers (Mifsud et al., 2015). Regarding EBV it has been shown that EBNA occupied super-enhancers reside with their corresponding gene in the same TAD. They used the low-resolution Hi-C data from Selvaraj *et al.*

We compared the EBNA regulated genes to randomly selected genes and found that they cluster in groups, which we named co-regulated gene blocks (CRGB). CRGBs are defined by gene regulation while TADs or contact domains are defined on a structural level. However, this observation could be a general mode of action of TFs and would be needed to compare to other TFs to resolve this question. Furthermore, one could investigate whether each block contains a "direct" target gene (regulated with absent *de novo* protein synthesis) or intriguingly, a lncRNA which could induce looping and activation of the whole CRGB.

Whether our block regulation data correlates with any of these datasets remains to be determined. Datasets were obtained regarding 3D genome organization using different techniques resulting in different interpretations. A most informative approach to the question of whether EBV exploits any kind of higher level chromatin order would certainly be provided by applying a Hi-C method in different LCLs, proficient and deficient for defined viral TFs.

4.2.2.3 Potential role for E2 regulated lncRNAs in the establishment of lymphomas and other cancer types

Differentially expressed genes detected by RNA-Seq and subsequent analysis required confirmation of candidate genes.

To confirm the detection of regulated protein coding and non-coding genes, three chromatin blocks of co-regulated genes were investigated by RT-qPCR. RNA was prepared of total and fractionated (cytoplasm and nucleoplasm) cell lysates and reverse transcribed. We could confirm that all investigated blocks contained genes regulated by E2, both with present or absent *de novo* protein synthesis (Figure 36- Figure 44). Studying the chromatin features, we noticed locally enriched marks for open chromatin and nascent RNA transcription (Figure 36, Figure 37, Figure 40, Figure 42). All the involved genes reside in the same or neighboring CTCF-mediated TADs or are linked by loops. The regulated blocks are rich in E2 binding sites. The involved lncRNAs are linked to cancer studies in the literature.

We picked these loci as representatives of target genes, co-regulated in blocks containing well-known E2 target protein coding genes and so far unknown non-coding targets. *MYC* appears to reside in a "desert" of non-coding transcriptional activity (Huppi et al., 2012). E2 regulates together with MYC several of the 3' and 5' of the TSS transcribed non-coding genes. Some of them were enriched in the nucleus (*PCAT1, CASC21, CASC8, LINC00977*), while a few were enriched in the cytoplasm (*CASC19, CCDC26*). This holds true for the other two investigated loci, where both lncRNA transcripts are enriched in the nucleus. Liang *et al.* described E2 regulated eRNAs 5' of the *MYC* TSS, eRNAs *MYC-428* and *MYC-525* which are derived from ESEs and regulate MYC expression (Liang et al., 2016). These eRNAs are in close proximity to *CASC19* TSS (*eRNA-428*,

which is -525 bps away from *MYC* TSS) and sense/antisense overlapping of the intron of *CASC21* (*eRNA525*, which is -428 bps away from *MYC* TSS; Figure 37). The sites for their shRNA-mediated knock down seem not to target their eRNAs but rather the transcripts of *CASC19/21*. Thus, it may be that the described eRNAs and *CASC19/21* exert a similar or identical function. The fold change profile obtained by RNA-Seq was recapitulated by RT-qPCR generally, although one has to bear in mind that the two methods cannot be compared directly with each other, since in library preparation there are many more steps involved which can affect the outcome. Moreover, the RNAs for RT-qPCR were not depleted for rRNAs. Furthermore, it must be remembered that despite the identical amounts of RNA subjected to both, reverse transcription or library preparation, the RNA from the nuclear fraction is obtained from 20 times more cells than the RNA of the cytoplasmic fraction in order to detect low abundance transcripts in the nucleus. Nevertheless, the enrichment in the nucleus or cytoplasm could reveal information about the primary localization of the lncRNAs and give hints towards their possible functional implications. Indeed, it has been reported that the function of lncRNAs can be associated with their unique subcellular localization (L.-L. Chen, 2016). All the regulated protein coding genes have important roles in different biological processes which can be associated with infection (SLAM-family receptors/P2RY11 in immune response) and transformation (MYC and PPAN for proliferation); the two major events describing EBVs life cycle. This chapter investigated lncRNAs (*PCAT1, CASC8, CASC19, CASC21, LINC00977, CCDC26, RP11-528G1.2 and CTD-2240E14.4*) regulated by E2 in blocks together with protein coding genes linked to cancer (see section 3.2.2.2.6, p. 79). Many lncRNAs have been reported to be uniquely expressed in differentiated tissues or specific cancer types (Iyer et al., 2015). Dysregulation is common in cancer and despite most lncRNAs showing a dysregulated expression are a cancer type specific, some are also shared among different cancer types (Yan et al., 2015). Intriguingly, we did not observe EBV to dysregulate lncRNAs with known roles in the hematopoietic system. However, browsing the Leukemia Atlas (expression profiles of 12 distinct blood cell populations) from the public available database www.lncScape.de, which represents a comprehensive resource for the non-coding RNA landscape in the human hematopoietic system (Schwarzer et al., 2017), most of the lncRNAs associated with the three loci *MYC, SLAMF* or *PPAN* are recorded. All of these lncRNAs were differentially expressed between the different blood cell populations, suggesting they may play important roles in immune cell function or differentiation. This is most intriguing given the fact that EBV regulates lncRNAs linked to cancer indicating a potential role for these lncRNAs in the establishment of lymphomas and other cancer types. Further, EBV is also associated with epithelial cancers, such as gastric (Iizasa, Nanbo, Nishikawa, Jinushi, & Yoshiyama, 2012) and pancreatic cancer (Samdani, Hechtman, O'Reilly, DeMatteo, & Sigel, 2015).

Interestingly, lncRNAs uncovered in our analysis as regulated by EBV have been linked with these types of cancer, indicating a potential mechanism through which EBV may promote these cancers. Intrinsically, EBV solely aims to initiate, establish and maintain persistent infection and not induce

transformation (Thorley-Lawson, 2015). Thus, despite the association of EBV with a variety of cancers, the fact that the majority of human adults are asymptomatically infected implies that EBV is not a typically oncogenic virus.

Hence, our work in uncovering cancer associated lncRNAs regulated by EBV may provide important clues as to how EBV virus transitions from asymptomatic to oncogenic in some subjects.

4.2.2.4 E2 induced pcgs and ncgs are regulated during the establishment of latency

E2 is one of the first viral coding genes expressed post infection and acts as a potent transcriptional regulator. We identified coding and non-coding target genes of E2 using the ER/EB2-5 system. We sought to address, whether the physiological events during the establishment of latency *in vitro* are recapitulated in the ER/EB2-5 cell system. Furthermore, the fact that EBV regulates lncRNAs linked to cancer could indicate that these lncRNAs contribute to the transformation of primary B cells. Besides, we wanted to assess whether protein coding and non-coding target genes are regulated similar by E2.

To test this, during the establishment of LCLs from primary B cells infected with strain B95.8 EBV, at several time points cells were harvested. Coding and non-coding candidate E2 targets (*MYC/CASC21, SLAMF1/ RP11-528G1.2* and *PPAN/ CTD-2240E14.4*) were quantified by RT-qPCR based on the cell number harvested at each time point. In order to account for changes in cell growth, we normalized to coding and non-coding housekeeping genes. We observed that E2 shows a peak in transcript abundance 72 h p.i, (Figure 46). Protein coding genes are regulated earlier during the establishment of latency than non-coding genes, peaking between 24 h and 72 h p.i. (Figure 47, Figure 49, and Figure 51). The lncRNA transcripts however increase progressively during the establishment of latency, peaking 6 d p.i. (Figure 48, Figure 50, and Figure 52). The protein coding genes are closely regulated following E2 induction. *SLAMF1* peaks earlier than E2 and the other protein-coding genes, potentially because it is involved in immune response and thus required earlier. Protein coding genes are regulated earlier during establishment of latency than non-coding genes. A possible reason for the late transcription may be that the transcription of the lncRNA appear later as a consequence of the activation of chromatin domains (Figure 47 to Figure 52). A hallmark of herpesvirus infection is the establishment of latency, with several strategies adopted to evade recognition by the host immune system during this period. Many reports exist detailing immune-related lncRNAs regulated by herpesviruses to affect immune response mechanisms in the cell (reviewed in Ahmed & Liu, 2018). It is possible that E2 regulated lncRNAs identified in our studies may play a role in such processes during progression towards latency.

Comparison of the transcriptomes of infected versus control cells revealed distinct changes of lncRNAs, thus the expression of the viral proteins altered expression of lncRNAs (Fortes & Morris, 2016). Additionally, the regulation of lncRNA transcription could provide signals of malignant transformation (Schmitt & Chang, 2016). Alternatively, the transcription of lncRNAs may be a host response to pathogen infection, as critical functions for lncRNAs have been identified in the innate immune response (Y. G. Chen, Satpathy, & Chang, 2017).

Investigation of additional regulated genes would confirm this pattern of regulation. Upon infection, the primary B cell transforms into a LCL and increase in size, which is accompanied by transcriptional changes adding complexity to the analysis. The distinction of direct E2 driven changes and changes which result from cell growth is intricate. The induced cell growth is accompanied with increased protein and RNA content per cell (Table S1). Hence, the RNA quantification was based on the cell number and in parallel based on the amount of RNA how it is commonly performed. The differences arising between both analyses are displayed (Figure 45 to Figure 52). As shown, the housekeeping genes undergo strong regulation. In order to correct for the changes arising from cell growth and increasing RNA content, and to determine the specific changes induced by E2 relative quantification was done. Of note, protein coding transcripts of interest were related to a coding housekeeping transcript, and non-coding transcripts of interest were related to a non-coding housekeeping gene. These housekeeping transcripts were not transcribed by the same Polymerase (PolIII instead of PolII) and were transcribed from multiple loci in the genome. Variations between the donors are observable, which is anticipated. All these circumstances may affect the results and subsequently the interpretation; hence, our interpretation must be treated with caution.

4.2.2.5 E2 induced lncRNAs may regulate remote protein-coding genes

The question arouse whether lncRNAs are regulated together with their protein coding neighbors. To address this question, we calculated the frequencies of distances of an expressed significantly regulated eRNA or lncRNA to the closest expressed significantly regulated pcg and a Pearson correlation was calculated for the log2FCs of a neighboring eRNA/lncRNA-pcg pair.
90 % of the distances between the pairs were ≤ 100 kb (Figure 53A). Additionally, eRNAs (by definition) were not observed closer in genomic proximity than other lncRNAs. The fold changes of the lncRNAs and pcg partner correlated positively (ranging from r= 0.72- 0.82), independent of the chromatin state at the TSS flanking region (Figure 53B). Here, we have shown that the lncRNAs appear to be co-regulated by E2 with protein coding genes, since their fold changes correlate positively, despite a huge distance between these pairs. The huge distance implies long-range interactions play an important role in this co-regulation. Since enhancer–promoter loops are

thought to not cross TAD borders (Dekker & Heard, 2015), this genomic limitation might be a hint that in some cases we did not detect enhancer promoter interactions here. Furthermore, enhancers are reported to directly affecting genes in physical proximity. The enhancer chromatin state neither promoted proximity nor increased the correlation, suggesting that the enhancer chromatin state at the TSS flanking region is not sufficient for the definition of an enhancer derived RNA (as discussed below). A similar investigation was conducted by Ilott *et al*., they too did not find a difference in distances between eRNAs and other lncRNAs to their pcg partner. However, they observed a much stronger correlation between the fold changes of eRNAs with the closest pcg compared to other lncRNA. Furthermore, their observed distances between A and B ranged from 0 to 100 kb, while for a certain amount of pairs the observed distances exceeded 2000 kb (Ilott et al., 2014). Long-distance intra- and inter-chromosomal interactions could be the result of co-localization of unlinked genomic regions to a common nuclear compartment (Bateman, Johnson, & Locke, 2012). It has been hypothesized that ncRNAs and coding genes belonging to the same biological pathways are coordinately regulated („guilt-by-association“ approach Schwarzer et al., 2017; Guttman et al., 2009). The correlation of fold changes could further be analyzed following the exclusion of all ncg-pcg relationships exceeding distances of 2000 kb in order to enhance investigations regarding TAD-related gene regulation.

4.2.2.6 E2 regulated lncRNAs: cellular localization, impact of blocked *de novo* protein synthesis and counter-regulation by E3A

4.2.2.6.1 E2 regulated lncRNAs, located in both nucleus and cytoplasm

Further examination of E2 target genes was conducted by filtering all genes with a detected read coverage of >1 read using a decision based filtering application (implemented by Gergely Csaba) for properties regarding coverage, significance and strength of regulation. E2 target genes were then assigned to subgroups which could be characterized more in detail based on localization, biotype, length, splice status, novelty, chromatin state, TF binding cluster and E2 binding relation. Where genes regulated in the same direction after comparing two different datasets were encountered, we designated them as co-regulated. If genes were regulated in opposing directions, we designated them as counter-regulated regarding the underlying investigation.
As discussed above, enrichment in the nucleus or cytoplasm may indicate the primary localization of the lncRNAs and hints towards their possible functional implications. Generalization of this logic may be transferred to different biotypes. The localization of lncRNAs was addressed by the fractionation of the cells in their compartments before RNA isolation. Two classes of genes were further characterized, the ENSEMBL annotated genes and novel detected intergenic lncRNAs.

Then, by RNA-Seq analysis, genes were filtered according to indicated thresholds. This resulted in two subgroups (Figure 54A). The majority of genes were similarly regulated in both compartments with 5,104 significantly co-regulated genes (Figure 54B). A minority (2,495) were not significantly co-regulated, and thus were compartment specifically regulated (most of them are only regulated in one compartment), 1,519 in the cytoplasm and 942 in the nucleus. The former group was further divided into four subsets according to their read coverage, 4,620 genes were sufficiently covered in the nucleus (nucleus expressed), 484 genes were not sufficiently covered in the nucleus (cytoplasm enriched), 4,964 genes which were sufficiently covered in the cytoplasm (cytoplasm expressed) and 140 genes which were not sufficiently covered in the cytoplasm (nucleus enriched; Figure 54C). For these six subgroups, we observed a lncRNA distribution of 15.2 % specifically in the nucleus, 23.2 % specifically in the cytoplasm, 29.2 % enriched in the cytoplasm and 36.5 % enriched in the nucleus (Figure 54D). The highest number of eRNAs according to our definition of an eRNA was found in the subset of genes enriched in the cytoplasm. We noticed no major difference in length of the lncRNAs in the different subgroups (Figure 54E). Thus, we could not confirm compartment specific distribution of lncRNA classes. Furthermore, we observed no difference between the subgroups of co-regulated genes with respect to the splice status (Figure 54F). In the compartment specific subsets we found that the lncRNAs specifically regulated in the cytoplasm appear to have fewer exons than lncRNAs specifically regulated in the nucleus.

Here we show that most of the target genes of E2 are found in both compartments, with varying read coverages in nucleus and cytoplasm. Furthermore, we found genes solely regulated in one of the compartments. The greatest number of E2 regulated lncRNAs was found in the nucleus enriched fraction. This is particularly important as much of the described activity of lncRNAs occurs in the nucleus, e.g. lincRNAs chromatin association (Khalil et al., 2009). Hence, we anticipated that the majority of lncRNA would reside in the nucleus. The FANTOM project reported lncRNAs to bear 5' capping and polyadenylation (Carninci et al., 2005; Hon et al., 2017). 5'caps and PolyA-signals are responsible for RNA export from the nucleus in the cytoplasm and furthermore for the stability of the RNAs (reviewed in Lewis & Izaurflde, 1997; Wickens, Anderson, & Jackson, 1997; Sachs, Sarnow, & Hentze, 1997). It may be that lncRNAs are first exported to the cytoplasm before they exert their function in the nucleus. Moreover, the majority of lncRNAs have been shown to interact with ribosomes in the cytoplasm (Carlevaro-Fita et al., 2016). Thus, it is not unexpected that despite a nuclear function, lncRNAs can be detected in both compartments. Similarly, despite a cytoplasmic function, the lncRNA are also detected in the nucleus, as the locus of origin. Additionally, the biochemical fractionation can also be not completely clean. Thus, a clear distinction is not feasible. One could further enrich for chromatin associated RNAs using special chromatin associated techniques like ChIRP (Chu, Qu, Zhong, Artandi, & Chang, 2011) or ChAR-Seq (Bell et al., 2018). With ChIRP (Chromatin Isolation by RNA Purification), tiling oligonucleotides capture specific lncRNAs with bound protein and DNA sequences, which subsequently can be

sequenced. ChAR-Seq (Chromatin-associated RNA sequencing) maps all RNA-to-DNA contacts across the genome.

E2 preferentially binds at enhancers (Glaser, PhD thesis, 2017; B. Zhao et al., 2011). Enhancer activation has been shown to involve PolII binding and eRNA synthesis (T.-K. Kim et al., 2010). As mentioned above, the definition of an eRNA might be inadequate. As discussed by Laurent *et al.*, classification of lncRNAs remains challenging, since the transcription of a lncRNA could start at an enhancer element or start distal to an enhancer element and merely overlap it, yet both would be classified as eRNAs (St Laurent, Wahlestedt, & Kapranov, 2015). Moreover, it is possible that lncRNA association with an enhancer element is not sufficient for eRNA classification. Additionally, our eRNA definition included all lncRNA genes with a TSS (of corresponding transcripts) overlapping with at least 1 bp, which might be a too tolerant cutoff. Most likely, we were not able to detect eRNAs in general. eRNAs are extremely unstable, at the moment of their generation they begin to be degraded by exosomes. Hence, special methods are required for their detection such as TT-Seq (transient transcriptome; Schwalb et al., 2016), GRO-Seq (global run on; Core, Waterfall, & Lis, 2008) or RNA-Seq with prior exosome depletion (Pefanis et al., 2015). All these methods enrich for the transient, nascent transcriptome and the chance to detect eRNAs would be increased. However, it may be that we detected fractional amounts of eRNAs which were somehow stabilized our accumulated in our cell systems. Since eRNAs are so unstable, they should be only detectable in the nucleus. The fact, that under our definition, we detect the majority of eRNAs in the cytoplasm enriched fractions indicates an erroneous assumption in our analysis. Finally, the enhancer chromatin state assignment is derived from the CSS for the LCL GM12878 of the ENCODE project. In this cell line, E2 and E3A are constitutively expressed and a chromatin signature is well established. The chromatin state of a “LCL” with no active E2 or E2 active for 6 h might be different given that E2 can rearrange the chromatin architecture as discussed above. Therefore, it may be that enhancer marks are already lost in GM12878, because for example E3A has repressed the enhancer locus.

We anticipated, that the lncRNAs residing in the nucleus would be shorter in average length and mostly monoexonic, since they should include most of the eRNAs. Most eRNAs are thought to be bidirectionally transcribed and neither spliced nor polyadenylated (and shorter than 2kb), however unidirectional transcribed eRNAs can be spliced and polyadenylated (and longer than 4kb; reviewed by Lam, Li, Rosenfeld, & Glass, 2014). We found, that the regulated lncRNAs have fewer exons than mRNA as already described (Derrien et al., 2012; Iyer et al., 2015). Further, lncRNAs are shorter in length than mRNA. In our study, most eRNAs were found in the cytoplasm enriched fraction and the transcripts in this subset were neither shorter nor less spliced than transcripts of other subsets. We expected to find shorter, monoexonic transcripts in the nucleic subsets, however they do not differ from the other subsets.

As with the ENSEMBL annotated genes, covered intergenic transcribed genes were filtered according to indicated thresholds. This resulted in two subgroups (Figure 55A). For the intergenic genes, also the majority is similarly regulated in both compartments as we found 360 significantly co-regulated genes (Figure 55B). A minority (243) not significantly co-regulated genes, were genes which are compartment specifically regulated, 201 in the nucleus, and 34 in the cytoplasm. The former group was further divided into four subsets according to their read coverage, 288 genes were sufficiently covered in the nucleus (nucleus expressed), 72 genes which were not sufficiently covered in the nucleus (cyto enriched), 233 genes which were sufficiently covered in the cytoplasm (cytoplasm expressed) and 127 genes which were not sufficiently covered in the cytoplasm (nucleus enriched; Fig 48 C). When intersecting the intergenic genes (no overlap with the ENSEMBL annotation) with the comprehensive lncRNA database LNCat, we identified unannotated genes (Figure 55D). The highest number of unannotated genes were located in the nucleus and are enriched there. Further, the specific nuclear subset was also enriched for unannotated genes compared to the cytoplasmic subset. The transcripts in the cytoplasm were shorter than those in the nucleus (Figure 55E). For the majority of genes no CSS was assigned to the TSS flanking region of their corresponding transcripts by Ernst *et al.* (Ernst et al., 2011), for the rest of the genes, a chromatin state of active transcription was principally assigned (Figure 55F). For the genes only regulated in the nucleus, an enrichment for the enhancer state can be noticed.

With these correlative data we identified novel intergenic transcribed genes regulated by E2. It is immediately obvious that regulated intergenic genes are more compartment specific than ENSEMBL annotated genes. Following the assumption that all protein coding genes have been discovered already, these genes might all be lncRNAs. Most of these lncRNAs have not been annotated until now. Thus, these lncRNAs might be specific for E2 regulation. As anticipated, they are preferentially located in the nucleus, which may indicate more chromatin related functions. One possible reason, why the detected genes in the cytoplasm were shorter could be the immature nature of the transcripts in the nucleus. Thus, they may not be readily spliced and lead to an increased length. The CSS assignment is less informative for the novel lncRNAs. As discussed above, the CSS is derived for a wt LCL with constitutively expressed E2. Most likely, it does not recapitulate the chromatin state immediately following E2 activation.

LncRNAs were suggested to contribute to cell identity as their expression is more cell type specific or tissue specific than is the case for protein isoforms (Cabili et al. 2011). This would support the observation of E2 specific regulation of these transcripts. The reason why they have not been detected in GM12878 cells so far might be because their transcriptional abundance has decreased already. Moreover, particularly the nucleic RNA has not been investigated for GM12878 so far. It could be possible that we detected a small portion of novel eRNAs which were somehow stabilized in our cell systems, seen as the lncRNAs only regulated in the nucleus.

These potential novel lncRNA transcripts should be further investigated for their coding potential. Moreover, one could investigate the hypothetical eRNAs for bidirectional transcription. Bidirectional transcription is characteristic for some eRNAs. However, transcription is always bidirectional in the first place and can be monitored at both, enhancers and promoters, since PolII function is not strand directed, only cryptic polyA sites or dense nucleosomes lead it in one direction (Seila et al., 2008). Thus, bidirectional transcription is not an enhancer unique feature.

4.2.2.6.2 E2 regulated lncRNAs are also partly regulated in the absence of *de novo* protein synthesis

E2 regulates primary/direct targets, some of which are themselves involved in gene regulation, such as MYC or LMP1. Genes regulated by direct targets of E2 can be considered secondary/indirect E2 targets. In order to dissect these secondary targets, ChX treatment was employed. However, this approach is only applicable to protein-coding genes. As already mentioned, treatment with ChX resulted in high variations between the biological replicates, most likely because it impacts RNA stabilization. This variation was too high in the cytoplasm for analysis to proceed, thus the analysis was only completed for the nucleus. Given the fact that ChX has an impact on RNA metabolism, we only analyzed genes regulated by E2 with and without ChX treatment. This analysis was applied to ENSEMBL annotated genes and novel detected intergenic lncRNAs.

Filtering of covered genes resulted in 174 genes which are similarly regulated with present and absent *de novo* protein synthesis and considered significantly co-regulated (defined as “direct” targets). While 4,501 genes were solely significantly regulated with present *de novo* protein synthesis (defined as “secondary” targets; Figure 56A-C). 20 % of the potential direct targets are lncRNAs, 6 % are eRNAs by definition (Figure 56D). The CSS at the TSS flanking region of transcripts derived from E2 regulated genes was enriched for open chromatin of enhancers and promoters for all E2 targets, which decreases again modestly with absent *de novo* protein synthesis (Figure 56E). When assigning E2 peaks to the transcripts corresponding to E2 regulated targets in order to reveal possible connections between binding and regulation, we observed a strong increase of E2 binding somehow linked by proximity to -/+ ChX regulated genes (Figure 57A). Finally, we intersected E2 binding sites linked to transcripts corresponding to genes of different subsets with defined E2 peak clusters (Figure 57B). We found a strong enrichment towards cluster I for both subgroups, the genes regulated by E2 with present and with absent *de novo* protein synthesis. Filtering of covered genes resulted in only nine intergenic transcribed genes which are significantly co-regulated by E2 with present and absent *de novo* protein synthesis, while 372 intergenic transcribed genes were solely significantly regulated with present *de novo* protein synthesis (Figure 58A/B). When assigning E2 peaks to the transcripts corresponding to E2

regulated intergenic transcribed genes in order to reveal a possible connection between binding and regulation, we observed mostly peaks in the common TAD with co-regulated genes (Figure 58C).

These data correlations indicated some target genes of E2, which are most likely direct targets of E2, which compromise only 4 % of all E2 targets. This would support the fact that E2 has numerous targets which themselves initiate transcription. Thus, E2 triggers a huge cascade of transcriptional initiations. Among the targets following blockade of *de novo* protein synthesis are a substantial number of lncRNAs which could support a role for E2 mediated chromatin changes in target gene regulation. The fact that 61.5 % of the genes regulated under blocked *de novo* protein synthesis can be linked to E2 peaks in the TSS flanking region or the genebody of their derived transcripts suggests a direct regulation (Figure 57A). Transcripts of the majority of regulated genes could be linked to peaks predominantely residing in cluster I. Cluster I is characterized by correlation with EBF1, CBF1 and CUX1 binding, as well as high signals for open chromatin. It is well known, that E2 binds preferentially to open chromatin. Our data proposes a link between those binding sites and regulated genes including lncRNAs and eRNAs. Only nine intergenic transcribed genes could be detected as co-regulated and could be considered as direct target. These nine transcripts could be regulated by E2 peaks residing in the common TAD.

We could provide data on “direct” E2 targets, and detected more lncRNAs which may be involved in gene regulation such as eRNAs detected by Liang *et al.* (Liang et al., 2016). To confirm this, knock downs/ knock outs of our detected E2 regulated lncRNA studies would be required. We associated peaks looping to the TSS flanking regions using data provided by Mifsud *et al.* (Mifsud et al., 2015). Zhou *et al.* detected E2 bound EBV super-enhancers and that these super-enhancers were not in genomic proximity to TSSs, but that most of these super-enhancers can be found in the same TAD as their corresponding genes (Zhou et al., 2015). We paired up peaks residing in the same TAD and regulated genes. The information on TADs, or contact domain respectively were provided by Rao *et al.* (Rao et al., 2014). These contact domains are reported to be smaller than TADs and not defined by specific borders, as such, they are not TADs. As already discussed, the CSS could not recapitulate the chromatin landscape at 6 h post E2 activation. Same can be transferred to the identified binding sites. The E2 peaks were obtained from ChIP-Seq data in wt LCLs. The binding sites may be different at 6 h post E2 activation. Indeed, there may be much more binding, since target gene transcription is most abundant at 6 h post infection. The cluster analysis was conducted for features which were all obtained in wt LCLs. The determination of novel intergenic and intronic transcribed genes was very conservative and one might detect more direct intergenic and intronic transcribed targets with loosened thresholds. E2 may be involved in chromatin remodeling, the whole analysis was conducted disregarding this circumstance. A well-characterized mechanism by which lncRNAs modulate gene expression both in *cis* and in *trans*

requires an interaction with chromatin to facilitate histone modification (Khalil et al., 2009). Thus, E2 could utilize lncRNAs to achieve changes in the chromatin environment.
In order to associate binding with regulation one would need to conduct Hi-C experiments or E2 binding sites knock out experiments with subsequent analysis of transcriptional changes.

4.2.2.6.3 E2 regulated lncRNAs are also partly counter-regulated by E3A

E3A is known to antagonize the activation of E2 at various promoters. We sought to determine the genome-wide antagonism of E2 gene expression by E3A. E2 and E3A ENSEMBL and intergenic transcribed target genes were compared by examining genes significantly regulated by E2 in the nucleus and E3A regulation in the nucleus.
70 % of ENSEMBL genes significantly regulated by both TFs were significantly counter-regulated (Figure 59A/C). 16 % of them are lncRNAs, with most of them found in both compartments. Not all lncRNAs are counter-regulated in both compartments, to some extent they are nucleus specifically regulated (Figure 59B). Defining genes with a read coverage of > 20 reads as expressed, 75 % of the counter-regulated genes are still expressed in wt LCLs. The co-regulated genes are almost all expressed in wt LCLs, except for the genes co-repressed by E2 and E3A (Figure 59B). Competition for gene regulation can only partly be mirrored in a binding pattern when attempting to determine the counter-regulation of genes between E2 and E3A (Figure 59D). Strikingly, when testing the counter-regulated genes for enrichment in biological processes, we found a most of the enriched processes to be connected to development, genesis, proliferation and differentiation (Table S2).
A similar picture emerged for the intergenic transcribed genes. 70 % of intergenic transcribed genes significantly regulated by both TFs were significantly counter-regulated by E3A (Figure 60A). Defining genes with a read coverage of > 20 reads as expressed, half of the counter-regulated genes are still expressed in wt LCLs. The co-regulated intergenic transcribed genes are almost all expressed in wt LCLs (data not shown). For over 80 % of transcripts counter-regulated between E2 and E3A both TFs reside in a position supporting regulation (data not shown).
This analysis proved that antagonism of E3A of E2 regulated transcripts is present genome-wide. Taking the model of competition for binding at enhancer sites into account, we observed, that the capacity of E2 for regulation prevailed over E3A for the majority of genes. All co-regulated genes were expected to be expressed in wt LCLs. This was the case for the induced genes, however, E2 and E3A also can coordinately repress genes, which appear to be not expressed in wt LCLs. The counter-regulated genes can be partly linked to binding of both TFs which could explain their regulation. GO-terms related to genesis and proliferation were found to be counter-regulated, which may indicate a capacity of E3A to dampen processes which could be involved in neoplastic activities.
It bares repetition that the determination of novel intergenic and intronic transcribed genes was

very conservative and one might detect more direct intergenic and intronic transcribed targets with loosened thresholds. Glaser studied binding patterns of E2 and E3A in a comprehensive manner, and found strong antagonistic binding genome-wide which could explain the antagonistic gene regulation observed in this study (Glaser, PhD thesis, 2017). Furthermore, comparing the results of microarray analysis of E2 and E3A regulation, shared subsets of target genes could be monitored (data of Harth-Hertle reviewed in Glaser, PhD thesis 2017). Until now it was not known how common E2 and E3A antagonistic regulation of host genes was (Allday et al., 2015). This analysis indicates that the majority of the shared target genes are counter-regulated in the nucleus. Similar results were obtained by the examination of cytoplasmic genes.

One has to bear in mind that similar as for E2 binding sites, E3A binding sites were obtained for wt LCLs and the binding pattern might be different in ER/EB2-5 cells, 6 h post E2 reinduction. However, E3A is constitutively expressed in ER/EB2-5 and constitutively exerting its transcriptional activity. Alterations in E2 activity could have consequences for E3A activity, especially as E2 induces E3A transcription as discussed below. E3A/C are thought to be onco-proteins, while E3B is described as a tumor-suppressor (Allday et al., 2015). E3A is involved in the repression of important tumor suppressive pathways (Styles, Paschos, White, & Farrell, 2018) contrasting with our finding where E3A counter-regulated neoplastic activities. Possibly, E3A could exert a dual function with oncogenic and tumor-suppressor capabilities (Shen, Shi, & Wang, 2018).

It may be that a strong permanent repression by E3A might dampen E2's gene activation in the ER/EB system. As already discussed, EBV most likely did not evolve to be a harmful oncogenic virus, thus, E3A may act to limit the pro-proliferative capacity of E2 which is required to establish latency. The counter-regulation of the lncRNA CCDC26 would support this hypothesis, as CCDC26 is thought to be involved in leukemic cell growth (Hirano et al., 2015).

It would be interesting to submit E2 and E3A regulated lncRNAs to a GO enrichment or KEGG pathway analysis. In general, functional annotation of the regulated and especially the counter-regulated lncRNAs would be informative in assessment of their role during establishment of latency. There exists already several online tools such as Co-LncRNA or LncADeep which conduct various analysis based on protein-coding genes proximal to lncRNA signature, mostly independent of their transcriptional direction (Yang et al., 2018; Z. Zhao et al., 2015), which could be useful. Even specialized analysis tools for cancer risk associated lncRNAs exists (Y. Xu et al., 2017). However, all of these tools assume a genomic proximal protein coding and non-coding genes to be involved in the same pathways, which does not account for the long-distance regulatory functions of some lncRNAs. Similar pathway association studies as for annotated lncRNAs could be performed for novel detected genes in order to assign a potential function.

In conclusion, this study provides extended information on the transcriptional regulation network of the EBV nuclear antigens E2 and E3A. In general, E2 regulates its target genes in CRGB, broad blocks of genes were observed to be co-regulated. E2 regulates numerous annotated lncRNAs as well novel intergenic transcribed genes, which might impact on neighboring gene regulation, since regulation of lncRNAs and distal protein coding genes correlates positively. The identification of target genes is strictly dependent on chosen thresholds for read coverage, significance and fold change. This study applied very strict thresholds for the identification of novel genes which will need further fine-tuning. We did however detect up to 1000 annotated and hundreds of potential novel lncRNAs regulated by E2 and E3A, which we further characterized according to subcellular localization, E2 regulatory dependence on *de novo* protein synthesis, chromatin state, E2 binding sites and E3A counter-regulation. Thus, EBV could exploit lncRNAs to achieve transcriptional changes in the cellular genome. Results on the viral transcriptome need further investigation and data acquisition, but provide a good initial starting point. The sum of changes regarding the transcriptome of the host and the virus could promote tumorigenesis.

5 References

Adhikary, D., Behrends, U., Boerschmann, H., Pfünder, A., Burdach, S., Moosmann, A., … Mautner, J. (2007). Immunodominance of lytic cycle antigens in Epstein-Barr virus-specific CD4+ T cell preparations for therapy. *PloS One*, *2*(7), e583. https://doi.org/10.1371/journal.pone.0000583

Ahmed, W., & Liu, Z.-F. (2018). Long Non-Coding RNAs: Novel Players in Regulation of Immune Response Upon Herpesvirus Infection. *Frontiers in Immunology*, *9*, 761. https://doi.org/10.3389/fimmu.2018.00761

Alfieri, C., Birkenbach, M., & Kieff, E. (1991). Early events in Epstein-Barr virus infection of human B lymphocytes. *Virology*, *181*, 595–608. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1849678

Allday, M. J., Bazot, Q., & White, R. E. (2015). The EBNA3 Family: Two Oncoproteins and a Tumour Suppressor that Are Central to the Biology of EBV in B Cells. In *Current topics in microbiology and immunology* (Vol. 391, pp. 61–117). https://doi.org/10.1007/978-3-319-22834-1_3

Alvarez-dominguez, J. R., & Lodish, H. F. (2014). Long noncoding RNAs during normal and malignant hematopoiesis, *99*(5), 531–541. https://doi.org/10.1007/s12185-014-1552-8.Long

Alvarez-Dominguez, J. R., & Lodish, H. F. (2017). Emerging mechanisms of long noncoding RNA function during normal and malignant hematopoiesis. *Blood*, *130*(18), 1965–1975. https://doi.org/10.1182/blood-2017-06-788695

Amon, W., & Farrell, P. J. (2005). Reactivation of Epstein-Barr virus from latency. *Reviews in Medical Virology*. https://doi.org/10.1002/rmv.456

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., … Barrell, B. G. (1984). DNA sequence and expression of the B95-8 Epstein—Barr virus genome. *Nature*, *310*(5974), 207–211. https://doi.org/10.1038/310207a0

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. https://doi.org/10.1038/cr.2011.22

Barth, T. K., & Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical Sciences*, *35*(11), 618–626. https://doi.org/10.1016/j.tibs.2010.05.006

Bateman, J. R., Johnson, J. E., & Locke, M. N. (2012). Comparing enhancer action in cis and in trans. *Genetics*, *191*(4), 1143–1155. https://doi.org/10.1534/genetics.112.140954

Benetatos, L., Hatzimichael, E., Dasoula, A., Dranitsaris, G., Tsiara, S., Syrrou, M., … Bourantas, K. L. (2010). CpG methylation analysis of the MEG3 and SNRPN imprinted genes in acute myeloid leukemia and myelodysplastic syndromes. *Leukemia Research*, *34*(2), 148–153. https://doi.org/10.1016/j.leukres.2009.06.019

Benetatos, L., Vartholomatos, G., & Hatzimichael, E. (2011). MEG3 imprinted gene contribution in tumorigenesis. *International Journal of Cancer*, *129*(4), 773–779. https://doi.org/10.1002/ijc.26052

Bickmore, W. A. (2013). The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics*, *14*(1), 67–84. https://doi.org/10.1146/annurev-genom-091212-153515

Boccellato, F., Anastasiadou, E., Rosato, P., Kempkes, B., Frati, L., Faggioni, A., & Trivedi, P. (2007). EBNA2 interferes with the germinal center phenotype by downregulating BCL6 and TCL1 in non-Hodgkin's lymphoma cells. *Journal of Virology*, *81*(5), 2274–2282. https://doi.org/10.1128/JVI.01822-06

Boller, S., Li, R., & Grosschedl, R. (2018). Defining B Cell Chromatin: Lessons from EBF1. *Trends in Genetics : TIG*, *34*(4), 257–269. https://doi.org/10.1016/j.tig.2017.12.014

Boller, S., Ramamoorthy, S., Akbas, D., Nechanitzky, R., Burger, L., Murr, R., … Grosschedl, R.

(2016). Pioneering Activity of the C-Terminal Domain of EBF1 Shapes the Chromatin Landscape for B Cell Programming. *Immunity*, *44*(3), 527–541. https://doi.org/10.1016/j.immuni.2016.02.021

Bonasio, R., & Shiekhattar, R. (2014). Regulation of Transcription by Long Noncoding RNAs. *Annual Review of Genetics*, *48*, 433–455. https://doi.org/10.1146/annurev-genet-120213-092323

Bonfert, T., Kirner, E., Csaba, G., Zimmer, R., & Friedel, C. C. (2015). ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, *16*(1), 122. https://doi.org/10.1186/s12859-015-0557-5

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, *144*(3), 327–339. https://doi.org/10.1016/j.cell.2011.01.024

Cabili, M., Trapnell C., Goff L., Koziol M., Tazon-Vega B., …Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.*Genes & development, 25 (18)*, 1915-27. https:// doi: 10.1101/gad.17446611

Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L. A., & Johnson, R. (2016). Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA (New York, N.Y.)*, *22*(6), 867–882. https://doi.org/10.1261/rna.053561.115

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., … RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, *309*(5740), 1559–1563. https://doi.org/10.1126/science.1112014

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., … Gustincich, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, *491*(7424), 454–457. https://doi.org/10.1038/nature11508

Chang, C. M., Yu, K. J., Mbulaiteye, S. M., Hildesheim, A., & Bhatia, K. (2009). The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: A need for reappraisal. *Virus Research*, *143*(2), 209–221. https://doi.org/10.1016/J.VIRUSRES.2009.07.005

Chen, L.-L. (2016). Linking Long Noncoding RNA Localization and Function. *Trends in Biochemical Sciences*, *41*(9), 761–772. https://doi.org/10.1016/j.tibs.2016.07.003

Chen, Y. G., Satpathy, A. T., & Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nature Immunology*, *18*(9), 962–972. https://doi.org/10.1038/ni.3771

Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., & Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular Cell*, *44*(4), 667–678. https://doi.org/10.1016/j.molcel.2011.08.027

Ciccone, D. N., Morshead, K. B., & Oettinger, M. A. (2003). Chromatin Immunoprecipitation in the Analysis of Large Chromatin Domains Across Murine Antigen Receptor Loci. *Methods in Enzymology*, *376*, 334–348. https://doi.org/10.1016/S0076-6879(03)76022-4

Cohen, J. I., Wang, F., Mannick, J., & Kieff, E. (1989). Epstein-Barr virus nuclear protein 2 is a key determinant of lymphocyte transformation. *Proc Natl Acad Sci U S A*, *86*, 9558–9562.

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)*, *322*(5909), 1845–1848. https://doi.org/10.1126/science.1162228

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, *12*(12), e0190152. https://doi.org/10.1371/journal.pone.0190152

Davies, M. L., Xu, S., Lyons-Weiler, J., Rosendorff, A., Webber, S. A., Wasil, L. R., … Rowe, D. T. (2010). Cellular factors associated with latency and spontaneous Epstein–Barr virus reactivation in B-lymphoblastoid cell lines. *Virology*, *400*(1), 53–67. https://doi.org/10.1016/J.VIROL.2010.01.002

de Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D., & Plummer, M. (2012). Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The Lancet. Oncology*, *13*(6), 607–615. https://doi.org/10.1016/S1470-2045(12)70137-7

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., … Natoli, G. (2010). A

large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biology*, *8*(5), e1000384. https://doi.org/10.1371/journal.pbio.1000384

Dekker, J., & Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. *FEBS Letters*, *589*(20 Pt A), 2877–2884. https://doi.org/10.1016/j.febslet.2015.08.044

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews. Genetics*, *14*(6), 390–403. https://doi.org/10.1038/nrg3454

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs : Analysis of their gene structure , evolution , and expression, 1775–1789. https://doi.org/10.1101/gr.132159.111.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Dey, B. K., Pfeifer, K., & Dutta, A. (2014). The H19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. *Genes & Development*, *28*(5), 491–501. https://doi.org/10.1101/gad.234419.113

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082.Topological

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., … Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*, *66*(4), 568–576.e4. https://doi.org/10.1016/j.molcel.2017.04.018

El-Sharkawy, A., Al Zaidan, L., & Malki, A. (2018). Epstein-Barr Virus-Associated Malignancies: Roles of Viral Oncoproteins in Carcinogenesis. *Frontiers in Oncology*, *8*, 265. https://doi.org/10.3389/fonc.2018.00265

Epstein, M. A., Achong, B. G., & Barr, Y. M. (1964). Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet*, *1*, 702–703.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., … Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906

Fatrai, S., van Gosliga, D., Han, L., Daenen, S. M. G. J., Vellenga, E., & Schuringa, J. J. (2011). KRAS(G12V) enhances proliferation and initiates myelomonocytic differentiation in human stem/progenitor cells via intrinsic and extrinsic pathways. *The Journal of Biological Chemistry*, *286*(8), 6061–6070. https://doi.org/10.1074/jbc.M110.201848

Feederle, R., Neuhierl, B., Baldwin, G., Bannert, H., Hub, B., Mautner, J., … Delecluse, H. J. (2006). Epstein-Barr virus BNRF1 protein allows efficient transfer from the endosomal compartment to the nucleus of primary B lymphocytes. *Journal of Virology*, *80*(19), 9435–9443. https://doi.org/10.1128/JVI.00473-06

Fortes, P., & Morris, K. V. (2016). Long noncoding RNAs in viral infections. *Virus Research*, *212*, 1–11. https://doi.org/10.1016/j.virusres.2015.10.002

Fritah, S., Niclou, S. P., & Azuaje, F. (2014). Databases for lncRNAs: a comparative evaluation of emerging tools. *RNA (New York, N.Y.)*, *20*(11), 1655–1665. https://doi.org/10.1261/rna.044040.113

Geng, L., & Wang, X. (2015). Epstein-Barr Virus-associated lymphoproliferative disorders: experimental and clinical developments. *International Journal of Clinical and Experimental Medicine*, *8*(9), 14656–14671. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26628948

Glaser, L. (2016). PhD Thesis. Munich: LMU.

Glaser, L. V., Rieger, S., Thumann, S., Beer, S., Kuklik-Roos, C., Martin, D. E., … Kempkes, B.

(2017). EBF1 binds to EBNA2 and promotes the assembly of EBNA2 chromatin complexes in B cells. *PLoS Pathogens*, *13*(10), 1–30. https://doi.org/10.1371/journal.ppat.1006664
Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, *128*(4), 635–638. https://doi.org/10.1016/j.cell.2007.02.006
Gong, C., & Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature*, *470*(7333), 284–288. https://doi.org/10.1038/nature09701
Grisanzio, C., & Freedman, M. L. (2010). Chromosome 8q24-Associated Cancers and MYC. *Genes & Cancer*, *1*(6), 555–559. https://doi.org/10.1177/1947601910381380
Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., … Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–227. https://doi.org/10.1038/nature07672
Hakim, O., & Misteli, T. (2012). SnapShot: Chromosome conformation capture. *Cell*, *148*(5), 16–18. https://doi.org/10.1016/j.cell.2012.02.019
Hammerschmidt, W. (n.d.). Epstein-Barr-Virus / Basic Research. Retrieved August 6, 2018, from https://www.helmholtz-muenchen.de/agv/forschung/forschungsgebiete/epstein-barr-virus/index.html
Hammerschmidt, W. & Sugden, B. (1989). Genetic analysis of immortalizing functions of EpsteinBarr virus in human B lymphocytes. *Nature, 340*, 393-397. https://doi.org/10.1038/340393a0
Harth-Hertle, M. L., Scholz, B. A., Erhard, F., Glaser, L. V., Dölken, L., Zimmer, R., & Kempkes, B. (2013). Inactivation of Intergenic Enhancers by EBNA3A Initiates and Maintains Polycomb Signatures across a Chromatin Domain Encoding CXCL10 and CXCL9. *PLoS Pathogens*, *9*(9). https://doi.org/10.1371/journal.ppat.1003638
Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. https://doi.org/10.1038/ng1966
Henle, G., Henle, W., & Diehl, V. (1968). Relation of Burkitt's tumor-associated herpes-ytpe virus to infectious mononucleosis. *Proc Natl Acad Sci U S A*, *59*, 94–101.
Hertle, M. L., Popp, C., Petermann, S., Maier, S., Kremmer, E., Lang, R., … Kempkes, B. (2009). Differential Gene Expression Patterns of EBV Infected EBNA-3A Positive and Negative Human B Lymphocytes. *PLoS Pathogens*, *5*(7), e1000506. https://doi.org/10.1371/journal.ppat.1000506
Hirano, T., Yoshikawa, R., Harada, H., Harada, Y., Ishida, A., & Yamazaki, T. (2015). Long noncoding RNA, CCDC26, controls myeloid leukemia cell growth through regulation of KIT expression. *Molecular Cancer*, *14*(1). https://doi.org/10.1186/s12943-015-0364-7
Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., … Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934–947. https://doi.org/10.1016/j.cell.2013.09.053
Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., … Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, *543*(7644), 199–204. https://doi.org/10.1038/nature21374
Hu, G., Niu, F., Humburg, B. A., Liao, K., Bendi, S., Callen, S., … Buch, S. (2018). Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. *Oncotarget*, *9*(26), 18648–18663. https://doi.org/10.18632/oncotarget.24307
Huch, S., & Nissan, T. (2014). Interrelations between translation and general mRNA degradation in yeast. *Wiley Interdisciplinary Reviews. RNA*, *5*(6), 747–763. https://doi.org/10.1002/wrna.1244
Huppi, K., Pitt, J. J., Wahlberg, B. M., & Caplen, N. J. (2012). The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Frontiers in Genetics*, *3*, 69. https://doi.org/10.3389/fgene.2012.00069
Iizasa, H., Nanbo, A., Nishikawa, J., Jinushi, M., & Yoshiyama, H. (2012). Epstein-Barr Virus (EBV)-associated gastric carcinoma. *Viruses*, *4*(12), 3420–3439. https://doi.org/10.3390/V4123420
Ilott, N. E., Heward, J. A., Roux, B., Tsitsiou, E., Fenwick, P. S., Lenzi, L., … Lindsay, M. A. (2014).

Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. *Nature Communications*, *5*(1), 3979. https://doi.org/10.1038/ncomms4979

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., … Chinnnaiyan, A. M. (2015). The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nat Genet.*, *47*(3), 199–208. https://doi.org/10.1038/ng.3192.The

Kaiser, C., Laux, G., Eick, D., Jochner, N., Bornkamm, G. W., & Kempkes, B. (1999). The proto-oncogene c-myc is a direct target gene of Epstein-Barr virus nuclear antigen 2. *J Virol*, *73*, 4481–4. Retrieved from http://jvi.asm.org/cgi/content/full/73/5/4481

Kempkes, B., & Ling, P. D. (2015). EBNA2 and Its Coactivator EBNA-LP. In *Current topics in microbiology and immunology* (Vol. 391, pp. 35–59). https://doi.org/10.1007/978-3-319-22834-1_2

Kempkes, B., Spitkovsky, D., Jansen-Durr, P., Ellwart, J. W., Kremmer, E., Delecluse, H. J., … Hammerschmidt, W. (1995). B-cell proliferation and induction of early G1-regulating proteins by Epstein-Barr virus mutants conditional for EBNA2. *Embo J*, *14*, 88–96. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7828599

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., … Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, *106*(28), 11667–11672. https://doi.org/10.1073/pnas.0904715106

Khan, G., & Hashim, M. J. (2014). Global burden of deaths from Epstein-Barr virus attributable malignancies 1990-2010. *Infectious Agents and Cancer*, *9*(1), 38. https://doi.org/10.1186/1750-9378-9-38

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

Kim, Y. K., Furic, L., Parisien, M., Major, F., DesGroseillers, L., & Maquat, L. E. (2007). Staufen1 regulates diverse classes of mammalian transcripts. *The EMBO Journal*, *26*(11), 2670–2681. https://doi.org/10.1038/sj.emboj.7601712

Ko, Y.-H. (2015). EBV and human cancer. *Experimental & Molecular Medicine*, *47*(1), e130. https://doi.org/10.1038/emm.2014.109

Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., & Xiong, Y. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene*, *30*(16), 1956–1962. https://doi.org/10.1038/onc.2010.568

Kretz, M. (2013). TINCR, staufen1, and cellular differentiation. *RNA Biology*, *10*(10), 1597–1601. https://doi.org/10.4161/rna.26249

Küppers, R. (2003). B cells under influence: Transformation of B cells by Epstein-Barr virus. *Nature Reviews Immunology*, *3*(10), 801–812. https://doi.org/10.1038/nri1201

Kvansakul, M., Wei, A. H., Fletcher, J. I., Willis, S. N., Chen, L., Roberts, A. W., … Colman, P. M. (2010). Structural Basis for Apoptosis Inhibition by Epstein-Barr Virus BHRF1. *PLoS Pathogens*, *6*(12), e1001236. https://doi.org/10.1371/journal.ppat.1001236

Lam, M. T. Y., Li, W., Rosenfeld, M. G., & Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends in Biochemical Sciences*, *39*(4), 170–182. https://doi.org/10.1016/j.tibs.2014.02.007

Lee, K., Hsiung, C. C.-S., Huang, P., Raj, A., & Blobel, G. A. (2015). Dynamic enhancer-gene body contacts during transcription elongation. *Genes & Development*, *29*(19), 1992–1997. https://doi.org/10.1101/gad.255265.114

Lewis, J. D., & Izaurflde, E. (1997). The Role of the Cap Structure in RNA Processing and Nuclear Export. *European Journal of Biochemistry*, *247*(2), 461–469. https://doi.org/10.1111/j.1432-1033.1997.00461.x

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., … Sung, W.-K. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology*, *11*(2), R22. https://doi.org/10.1186/gb-2010-11-2-r22

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., … Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, *498*(7455), 516–520. https://doi.org/10.1038/nature12210

Liang, J., Zhou, H., Gerdt, C., Tan, M., Colson, T., Kaye, K. M., … Zhao, B. (2016). Epstein–Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation. *Proceedings of the National Academy of Sciences*, *113*(49), 14121–14126. https://doi.org/10.1073/pnas.1616697113

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Lin, Z., Wang, X., Strong, M. J., Concha, M., Baddoo, M., Xu, G., … Flemington, E. K. (2013). Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *Journal of Virology*, *87*(2), 1172–1182. https://doi.org/10.1128/JVI.02517-12

Liu, B., Sun, L., Liu, Q., Gong, C., Yao, Y., Lv, X., … Song, E. (2015). A Cytoplasmic NF-κB Interacting Long Noncoding RNA Blocks IκB Phosphorylation and Suppresses Breast Cancer Metastasis. *Cancer Cell*, *27*(3), 370–381. https://doi.org/10.1016/J.CCELL.2015.02.004

Longnecker, R., & Neipel, F. (2007). Introduction to the human gamma-herpesviruses. In A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley, & K. Yamanishi (Eds.), *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. Cambridge. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21348088

López, R., Urquiza, M., Patino, H., Suárez, J., Reyes, C., Patarroyo, M. A., & Patarroyo, M. E. (2005). A B-lymphocyte binding peptide from BNRF1 induced antibodies inhibiting EBV-invasion of B-lymphocytes. *Biochimie*, *87*(11), 985–992. https://doi.org/10.1016/J.BIOCHI.2005.04.009

Lu, F., Chen, H. S., Kossenkov, A. V., DeWispeleare, K., Won, K. J., & Lieberman, P. M. (2016). EBNA2 Drives Formation of New Chromosome Binding Sites and Target Genes for B-Cell Master Regulatory Transcription Factors RBP-jκ and EBF1. *PLoS Pathogens*, *12*(1), 1–24. https://doi.org/10.1371/journal.ppat.1005339

Maier, S., Staffler, G., Hartmann, A., Höck, J., Henning, K., Grabusic, K., … Kempkes, B. (2006). Cellular target genes of Epstein-Barr virus nuclear antigen 2. *Journal of Virology*, *80*(19), 9761–9771. https://doi.org/10.1128/JVI.00665-06

Mattick, J. S. (2004). RNA regulation: a new genetics? *Nature Reviews Genetics*, *5*(4), 316–323. https://doi.org/10.1038/nrg1321

McClellan, M. J., Wood, C. D., Ojeniyi, O., Cooper, T. J., Kanhere, A., Arvey, A., … West, M. J. (2013). Modulation of Enhancer Looping and Differential Gene Targeting by Epstein-Barr Virus Transcription Factors Directs Cellular Reprogramming. *PLoS Pathogens*, *9*(9). https://doi.org/10.1371/journal.ppat.1003636

McKenzie, J., & El-Guindy, A. (2015). Epstein-Barr Virus Lytic Cycle Reactivation (pp. 237–261). Springer, Cham. https://doi.org/10.1007/978-3-319-22834-1_8

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., … Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), 598–606. https://doi.org/10.1038/ng.3286

Miller, D. M., Thomas, S. D., Islam, A., Muench, D., & Sedoris, K. (2012). c-Myc and cancer metabolism. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *18*(20), 5546–5553. https://doi.org/10.1158/1078-0432.CCR-12-0977

Morales-Sánchez, A., & Fuentes-Panana, E. M. (2018). The Immunomodulatory Capacity of an Epstein-Barr Virus Abortive Lytic Cycle: Potential Contribution to Viral Tumorigenesis. *Cancers*, *10*(4). https://doi.org/10.3390/cancers10040098

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., … Feng, F. Y. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nature Communications*, *7*, 12791. https://doi.org/10.1038/ncomms12791

Oh, J. K., & Weiderpass, E. (2014). Infection and cancer: Global distribution and burden of diseases. *Annals of Global Health*, *80*(5), 384–392. https://doi.org/10.1016/j.aogh.2014.09.013

Palermo, R. D., Webb, H. M., Gunnell, A., & West, M. J. (2008). Regulation of transcription by the Epstein-Barr virus nuclear antigen EBNA 2. *Biochemical Society Transactions*, *36*(Pt 4), 625–628. https://doi.org/10.1042/BST0360625

Pattle, S. B., & Farrell, P. J. (2006). The role of Epstein–Barr virus in cancer. *Expert Opinion on Biological Therapy*, *6*(11), 1193–1205. https://doi.org/10.1517/14712598.6.11.1193

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., … Basu, U. (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell*, *161*(4), 774–789. https://doi.org/10.1016/j.cell.2015.04.034

Peng, W., & Jiang, A. (2016). Long noncoding RNA CCDC26 as a potential predictor biomarker contributes to tumorigenesis in pancreatic cancer. *Biomedicine & Pharmacotherapy*, *83*, 712–717. https://doi.org/https://doi.org/10.1016/j.biopha.2016.06.059

Plank, J. L., & Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. *Molecular Cell*, *55*(1), 5–14. https://doi.org/10.1016/j.molcel.2014.06.015

Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology 2015 16:4*, *16*(4), 245. https://doi.org/10.1038/nrm3965

Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., … Chinnaiyan, A. M. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology*, *29*(8), 742–749. https://doi.org/10.1038/nbt.1914

Rabson, M., Gradoville, L., Heston, L., & Miller, G. (1982). Non-immortalizing P3J-HR-1 Epstein-Barr virus: a deletion mutant of its transforming parent, Jijoye. *Journal of Virology*, *44*(3), 834–844. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6294333

Rando, O. J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Current Opinion in Genetics & Development*, *22*(2), 148–155. https://doi.org/10.1016/j.gde.2012.02.013

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Rashid, F., Shah, A., & Shan, G. (2016). Long Non-coding RNAs in the Cytoplasm. *Genomics, Proteomics & Bioinformatics*, *14*(2), 73–80. https://doi.org/10.1016/J.GPB.2016.03.005

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., … Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–1323. https://doi.org/10.1016/j.cell.2007.05.022

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robertson, E. S., Lin, J., & Kieff, E. (1996). The amino-terminal domains of Epstein-Barr virus nuclear proteins 3A, 3B, and 3C interact with RBPJ(kappa). *J Virol*, *70*, 3068–3074.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Sachs, A. B., Sarnow, P., & Hentze, M. W. (1997). Starting at the beginning, middle, and end: translation initiation in eukaryotes. *Cell*, *89*(6), 831–838. https://doi.org/10.1016/S0092-8674(00)80268-8

Sakai, T., Taniguchi, Y., Tamura, K., Minoguchi, S., Fukuhara, T., Strobl, L. J., … Honjo, T. (1998). Functional replacement of the intracellular region of the Notch1 receptor by Epstein-Barr virus nuclear antigen 2. *J Virol*, *72*, 6034–6039.

Salviano-Silva, A., Lobo-Alves, S. C., Almeida, R. C. de, Malheiros, D., & Petzl-Erler, M. L. (2018). Besides Pathology: Long Non-Coding RNA in Cell and Tissue Homeostasis. *Non-Coding RNA*, *4*(1). https://doi.org/10.3390/ncrna4010003

Samdani, R. T., Hechtman, J. F., O'Reilly, E., DeMatteo, R., & Sigel, C. S. (2015). EBV-associated lymphoepithelioma-like carcinoma of the pancreas: Case report with targeted sequencing analysis. *Pancreatology : Official Journal of the International Association of Pancreatology (IAP) … [et Al.]*, *15*(3), 302–304. https://doi.org/10.1016/j.pan.2015.03.016

Santpere, G., Darre, F., Blanco, S., Alcami, A., Villoslada, P., Mar Albà, M., & Navarro, A. (2014). Genome-Wide Analysis of Wild-Type Epstein–Barr Virus Genomes Derived from Healthy Individuals of the 1000 Genomes Project. *Genome Biology and Evolution*, *6*(4), 846–860. https://doi.org/10.1093/gbe/evu054

Schmitt, A. M., & Chang, H. Y. (2016). Long Noncoding RNAs in Cancer Pathways. *Cancer Cell*, *29*(4), 452–463. https://doi.org/10.1016/j.ccell.2016.03.010

Schneider-Poetsch, T., Ju, J., Eyler, D. E., Dang, Y., Bhat, S., Merrick, W. C., … Liu, J. O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nature Chemical Biology*, *6*(3), 209–217. https://doi.org/10.1038/nchembio.304

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22930834

Schwalb, B., Michel, M., Zacher, B., Hauf, K. F., Demel, C., Tresch, A., … Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*, *352*(6290), 1225–1228. https://doi.org/10.1126/science.aad9841

Schwarzer, A., Emmrich, S., Schmidt, F., Beck, D., Ng, M., Reimer, C., … Klusmann, J. H. (2017). The non-coding RNA landscape of human hematopoiesis and leukemia. *Nature Communications*, *8*(1), 1–16. https://doi.org/10.1038/s41467-017-00212-4

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., … Sharp, P. A. (2008). Divergent transcription from active promoters. *Science (New York, N.Y.)*, *322*(5909), 1849–1851. https://doi.org/10.1126/science.1162253

Selvaraj, S., R Dixon, J., Bansal, V., & Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, *31*(12), 1111–1118. https://doi.org/10.1038/nbt.2728

Sexton, T., & Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, *160*(6), 1049–1059. https://doi.org/10.1016/j.cell.2015.02.040

Shen, L., Shi, Q., & Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis*, *7*(3), 25. https://doi.org/10.1038/s41389-018-0034-x

Skalska, L., White, R. E., Franz, M., Ruhmann, M., & Allday, M. J. (2010). Epigenetic Repression of p16INK4A by Latent Epstein-Barr Virus Requires the Interaction of EBNA3A and EBNA3C with CtBP. *PLoS Pathogens*, *6*(6), e1000951. https://doi.org/10.1371/journal.ppat.1000951

Skalska, L., White, R. E., Parker, G. A., Sinclair, A. J., Paschos, K., & Allday, M. J. (2013). Induction of p16INK4a Is the Major Barrier to Proliferation when Epstein-Barr Virus (EBV) Transforms Primary B Cells into Lymphoblastoid Cell Lines. *PLoS Pathogens*, *9*(2), e1003187. https://doi.org/10.1371/journal.ppat.1003187

Smatti, M. K., Al-Sadeq, D. W., Ali, N. H., Pintus, G., Abou-Saleh, H., & Nasrallah, G. K. (2018). Epstein-Barr Virus Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population: An Update. *Frontiers in Oncology*, *8*, 211. https://doi.org/10.3389/fonc.2018.00211

Song, W., Liu, Y., Peng, J., Liang, H., Chen, H., Chen, J., … He, Y. (2016). Identification of differentially expressed signatures of long non-coding RNAs associated with different metastatic potentials in gastric cancer. *Journal of Gastroenterology*, *51*(2), 119–129. https://doi.org/10.1007/s00535-015-1091-y

Spender, L. C., Lucchesi, W., Bodelon, G., Bilancio, A., Elgueta Karstegl, C., Asano, T., … Farrell, P.

J. (2006). Cell target genes of Epstein-Barr virus transcription factor EBNA-2: Induction of the p55α regulatory subunit of PI3-kinase and its role in survival of EREB2.5 cells. *Journal of General Virology*, *87*(10), 2859–2867. https://doi.org/10.1099/vir.0.82128-0

St Laurent, G., Wahlestedt, C., & Kapranov, P. (2015). The Landscape of long noncoding RNA classification. *Trends in Genetics : TIG*, *31*(5), 239–251. https://doi.org/10.1016/j.tig.2015.03.007

Stanfield, B. A., & Luftig, M. A. (2017). Recent advances in understanding Epstein-Barr virus. *F1000Research*, *6*, 386. https://doi.org/10.12688/f1000research.10591.1

Styles, C., Paschos, K., White, R., & Farrell, P. (2018). The Cooperative Functions of the EBNA3 Proteins Are Central to EBV Persistence and Latency. *Pathogens*, *7*(1), 31. https://doi.org/10.3390/pathogens7010031

Szalaj, P., Tang, Z., Michalski, P., Pietal, M. J., Luo, O. J., Sadowski, M., … Plewczynski, D. (2016). An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. *Genome Research*, *26*(12), 1697–1709. https://doi.org/10.1101/gr.205062.116

Takeda, S., Kanbayashi, D., Kurata, T., Yoshiyama, H., & Komano, J. (2014). Enhanced susceptibility of B lymphoma cells to measles virus by Epstein-Barr virus type III latency that upregulates CD150/signaling lymphocytic activation molecule. *Cancer Science*, *105*(2), 211–218. https://doi.org/10.1111/cas.12324

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., … Ruan, Y. (2015a). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, *163*(7), 1611–1627. https://doi.org/10.1016/j.cell.2015.11.024

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., … Ruan, Y. (2015b). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, *163*(7), 1611–1627. https://doi.org/10.1016/j.cell.2015.11.024

Thorley-Lawson, D. A. (2015). EBV Persistence--Introducing the Virus. *Current Topics in Microbiology and Immunology*, *390*(Pt 1), 151–209. https://doi.org/10.1007/978-3-319-22822-8_8

Thumann, S. (2016). Transkriptionsregulation durch das EBV-nukleäre Antigen 2 - Abhängigkeit von DNA-Adaptoren und die Funktion der N-terminalen Domäne.

Tian, D., Sun, S., & Lee, J. T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell*, *143*(3), 390–403. https://doi.org/10.1016/j.cell.2010.09.049

Tomkinson, B., Robertson, E., & Kieff, E. (1993). Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation. *J Virol*, *67*, 2014–25.

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., … Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*, *39*(6), 925–938. https://doi.org/10.1016/j.molcel.2010.08.011

Tsai, K., Thikmyanova, N., Wojcechowskyj, J. A., Delecluse, H. J., & Lieberman, P. M. (2011). EBV tegument protein BNRF1 disrupts DAXX-ATRX to activate viral early gene transcription. *PLoS Pathogens*, *7*(11). https://doi.org/10.1371/journal.ppat.1002376

Veillette, A. (2006). NK cell regulation by SLAM family receptors and SAP-related adapters. *Immunological Reviews*, *214*(1), 22–34. https://doi.org/10.1111/j.1600-065X.2006.00453.x

Vitiello, L., Gorini, S., Rosano, G., & la Sala, A. (2012). Immunoregulation through extracellular nucleotides. *Blood*, *120*(3), 511–518. https://doi.org/10.1182/blood-2012-01-406496

Vockerodt, M., Yap, L.-F., Shannon-Lowe, C., Curley, H., Wei, W., Vrzalikova, K., & Murray, P. G. (2015). The Epstein-Barr virus and the pathogenesis of lymphoma. *The Journal of Pathology*, *235*(2), 312–322. https://doi.org/10.1002/path.4459

Wang, C., Wang, L., Ding, Y., Lu, X., Zhang, G., Yang, J., … Xu, L. (2017). LncRNA Structural Characteristics in Epigenetic Regulation. *International Journal of Molecular Sciences*, *18*(12). https://doi.org/10.3390/ijms18122659

Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., … Fan, Q. (2010). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer.

Nucleic Acids Research, *38*(16), 5366–5383. https://doi.org/10.1093/nar/gkq285
Wang, K. C., & Chang, H. Y. (2011). Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell*, *43*(6), 904–914. https://doi.org/10.1016/j.molcel.2011.08.018
Weil, D., Boutain, S., Audibert, A., & Dautry, F. (2000). Mature mRNAs accumulated in the nucleus are neither the molecules in transit to the cytoplasm nor constitute a stockpile for gene expression. *RNA (New York, N.Y.)*, *6*(7), 962–975. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10917593
Welch, P. J., Marcusson, E. G., Li, Q.-X., Beger, C., Krüger, M., Zhou, C., … Barber, J. R. (2000). Identification and Validation of a Gene Involved in Anchorage-Independent Cell Growth Control Using a Library of Randomized Hairpin Ribozymes. *Genomics*, *66*(3), 274–283. https://doi.org/10.1006/GENO.2000.6230
Whitaker, A. M. (1985). The chromosomes of the Namalwa cell line. *Journal of Biological Standardization*, *13*(2), 173-IN3. https://doi.org/10.1016/S0092-1157(85)80024-X
Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035
Wickens, M., Anderson, P., & Jackson, R. J. (1997). Life and death in the cytoplasm: messages from the 3′ end. *Current Opinion in Genetics & Development*, *7*(2), 220–232. https://doi.org/10.1016/S0959-437X(97)80132-3
Wolpin, B. M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G. M., Wang, Z., … Amundadottir, L. T. (2014). Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature Genetics*, *46*(9), 994–1000. https://doi.org/10.1038/ng.3052
Wood, C. D., Veenstra, H., Khasnis, S., Gunnell, A., Webb, H. M., Shannon-Lowe, C., … West, M. J. (2016). MYC activation and BCL2L11 silencing by a tumour virus through the large-scale reconfiguration of enhancer-promoter hubs. *ELife*, *5*(AUGUST), 1–23. https://doi.org/10.7554/eLife.18270
Wu Ct, C. -t., & Morris, J. R. (2001). Genes, genetics, and epigenetics: a correspondence. *Science (New York, N.Y.)*, *293*(5532), 1103–1105. https://doi.org/10.1126/science.293.5532.1103
Wutz, A., Rasmussen, T. P., & Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics*, *30*(2), 167–174. https://doi.org/10.1038/ng820
Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., … Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Research*, *42*(Database issue), D98-103. https://doi.org/10.1093/nar/gkt1222
Xu, J., Bai, J., Zhang, X., Lv, Y., Gong, Y., Liu, L., … Li, X. (2016). A comprehensive overview of lncRNA annotation resources. *Briefings in Bioinformatics*, *18*(2), bbw015. https://doi.org/10.1093/bib/bbw015
Xu, Y., Li, F., Wu, T., Xu, Y., Yang, H., Dong, Q., … Li, X. (2017). LncSubpathway: a novel approach for identifying dysfunctional subpathways associated with risk lncRNAs by integrating lncRNA and mRNA expression profiles and pathway topologies. *Oncotarget*, *8*(9), 15453–15469. https://doi.org/10.18632/oncotarget.14973
Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S. D., … Zhang, L. (2015). Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell*, *28*(4), 529–540. https://doi.org/10.1016/j.ccell.2015.09.006
Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., … Birol, I. (2018). LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty428
Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., … Zhou, M.-M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular Cell*, *38*(5), 662–674. https://doi.org/10.1016/j.molcel.2010.03.021
Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., … Gorospe, M. (2012). LincRNA-p21 Suppresses Target mRNA Translation. *Molecular Cell*, *47*(4), 648–655. https://doi.org/10.1016/J.MOLCEL.2012.06.027

Young, L. S., Yap, L. F., & Murray, P. G. (2016). Epstein-Barr virus: More than 50 years old and still providing surprises. *Nature Reviews Cancer*, *16*(12), 789–802. https://doi.org/10.1038/nrc.2016.92

Zhang, T., Cooper, S., & Brockdorff, N. (2015). The interplay of histone modifications - writers that read. *EMBO Reports*, *16*(11), 1467–1481. https://doi.org/10.15252/embr.201540945

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhao, B., Zou, J., Wang, H., Johannsen, E., Peng, C., Quackenbush, J., … Kieff, E. (2011). Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(36), 14902–14907. https://doi.org/10.1073/pnas.1108892108

Zhao, Z., Bai, J., Wu, A., Wang, Y., Zhang, J., Wang, Z., … Li, X. (2015). Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database : The Journal of Biological Databases and Curation*, *2015*. https://doi.org/10.1093/database/bav082

Zhou, H., Schmidt, S. C. S., Jiang, S., Willox, B., Bernhardt, K., Liang, J., … Zhao, B. (2015). Epstein-Barr virus oncoprotein super-enhancers control B cell growth. *Cell Host & Microbe*, *17*(2), 205–216. https://doi.org/10.1016/j.chom.2014.12.013

Zhu, S., Li, W., Liu, J., Chen, C.-H., Liao, Q., Xu, P., … Wei, W. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nature Biotechnology*, *34*(12), 1279–1286. https://doi.org/10.1038/nbt.3715

6 Supplementary Figures and Tables

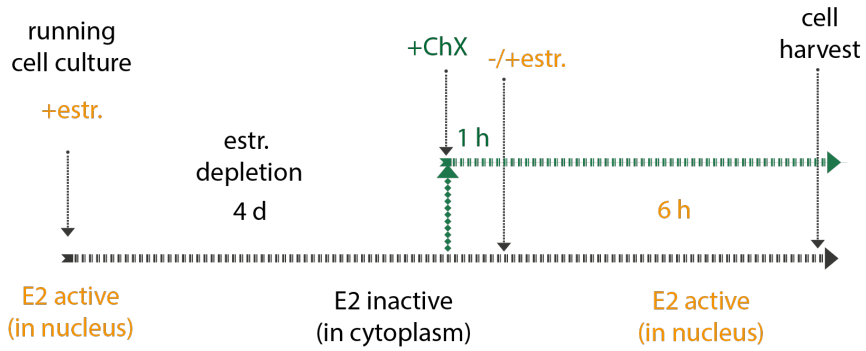


Figure S1: Treatment scheme for ER/EB2-5 cells. For three biological replicates, 2 x 10^8 cells of four different treatments were harvested: EBNA2 inactive (ER/EB2-5 cells cultivated without estrogen for 4 d), EBNA2 active (ER/EB2-5 cells cultivated without estrogen for 4 d, then estrogen is added for 6 h), EBNA2 inactive/ChX treated (ER/EB2-5 cells cultivated without estrogen for 4 d; treated with cycloheximide for 7 h) and EBNA2 active/ChX (ER/EB2-5 cells cultivated without estrogen for 4 d then treated with cycloheximide for 7 h and estrogen for the last 6 h).

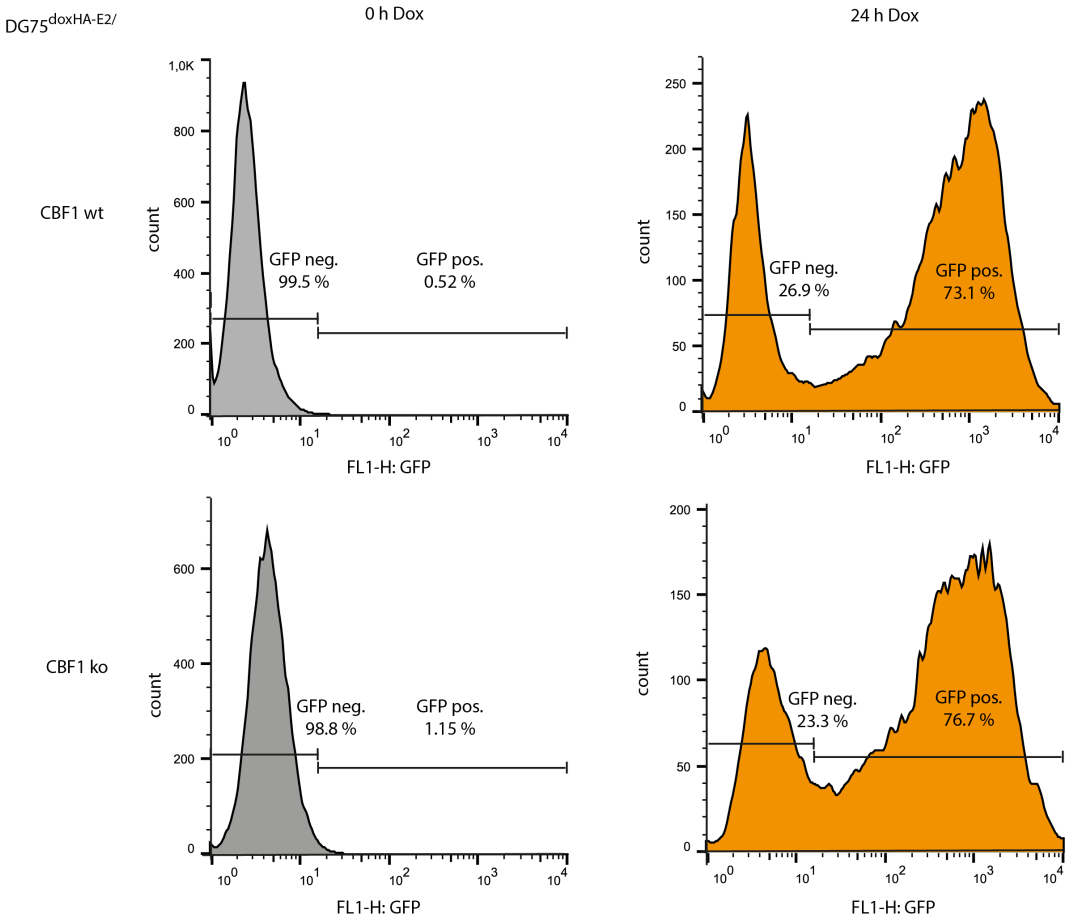


Figure S2: The expression of HA-E2 in the EBV negative DG75 cell lines can be induced by doxycycline (dox) Stably transfected DG75 cell lines carry a vector encoding HA-E2. Dox treatment induces the simultaneous expression of HA-E2 in one and the bicistronic reporter construct of a truncated nerve growth factor receptor gene (tNGFR) and enhanced green fluorescent protein (eGFP) gene in the other direction from a bidirectional

promoter. E2 expression was induced with 1 µg/ml dox for 24 h and can be verified by monitoring the eGFP expression by flow cytometry. DG75$^{doxHA\text{-}E2}$/ CBF1 wt cells (upper panel) and DG75$^{doxHA\text{-}E2}$/CBF1 ko cells (lower panel) without Dox (left) and after 24 h Dox treatment (right). Percentages of GFP positive and negative cells are shown (representative experiment).

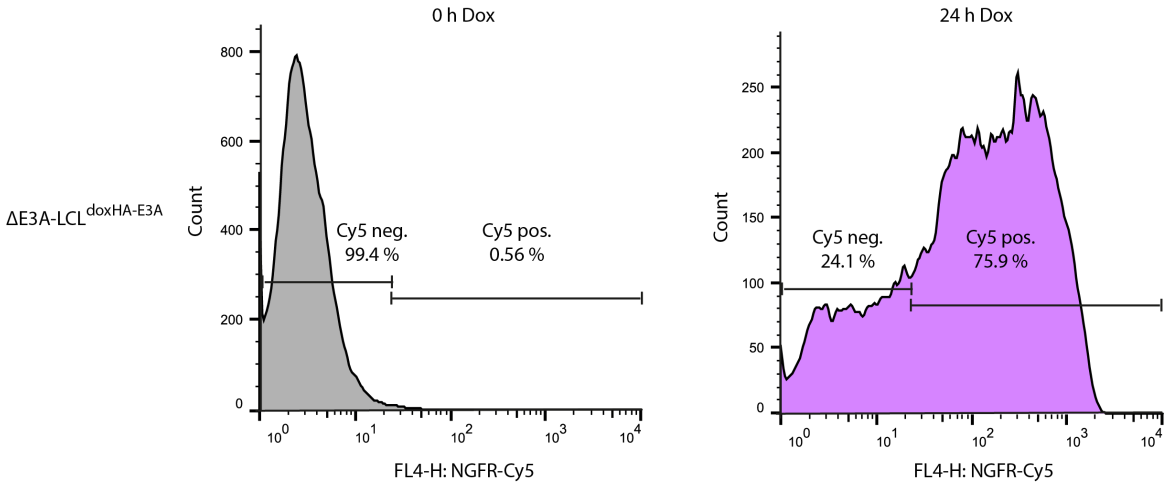


Figure S3: The expression of HA-E3A in a E3A defective LCL can be induced by doxycycline (dox). Stably transfected ΔE3A LCLs carry a vector encoding HA-E3A. Dox treatment induces the simultaneous expression of HA-E3A in one and the bicistronic reporter construct of a truncated nerve growth factor receptor gene (tNGFR) gene in the other direction from a bidirectional promoter. E3A expression was induced with 1 µg/ml dox for 24 h. tNGFR expression can be monitored by antibody-directed staining and subsequent flow cytometry. This signal is indicatory for the E3A expression. ΔE3A-LCLdoxHA-E3A without Dox (left) and after 24 h Dox treatment (right). Percentages of NGFR positive and negative cells are shown (representative experiment).

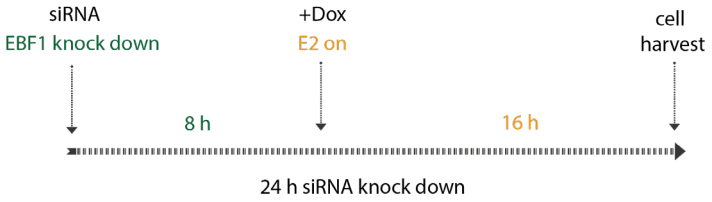


Figure S4: Treatment scheme for siRNA-mediated EBF1 knock down and subsequent E2 induction. For two biological replicates, 5x 10^6 cells were transfected by electroporation with 100 pmol siRNA. 8 h post transfection, Dox was added to induce E2 expression. 24 h post transfection (=16 h post dox induction) 10^7 cells were harvested for further analysis.

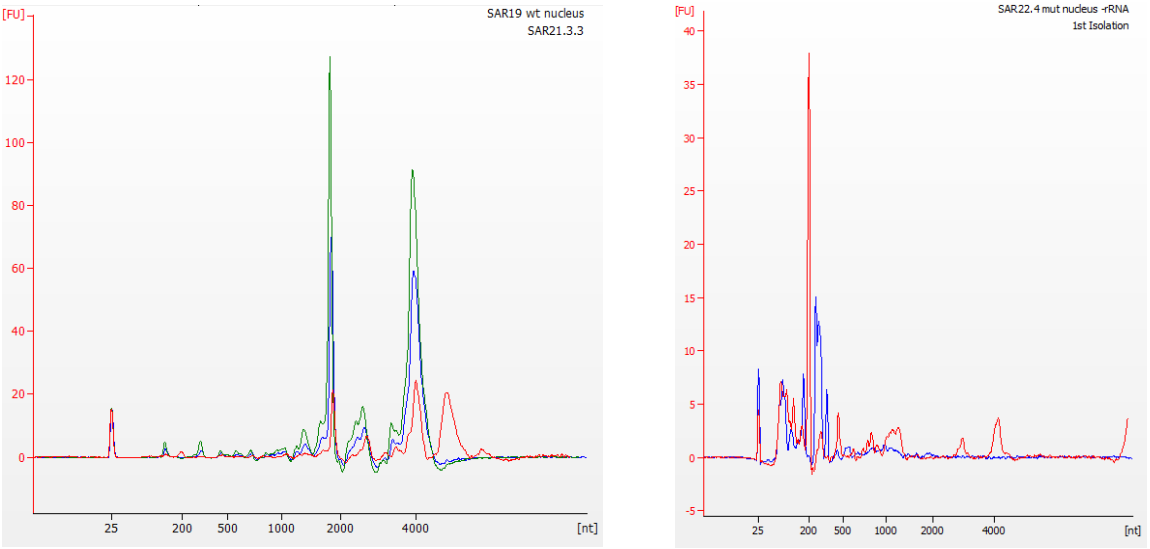


Figure S5: RNA quality control by BioAnalyzer. Visualization of Fractionation by enrichment (cytoplasm) or depletion (nucleus) of 18S/28SrRNA (left) and loss of 18S/28SrRNA by rRNA depletion by RiboZero Magnetic Gold Kit (right).

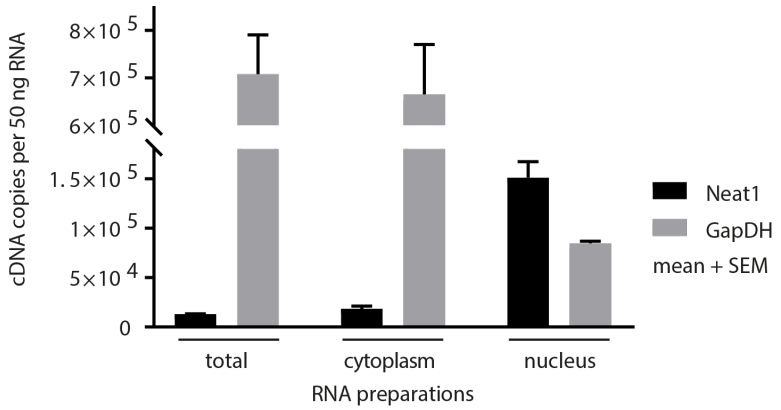


Figure S6: Conformation of fractionation of cell compartments. Cytoplasmic fraction was separated from nucleic fraction of 10^8 cells using a mild buffer and centrifugation. RNA was isolated from 10^7 cell equivalents for total and cytoplasmic RNA, 2 x 10^8 cell equivalents were used to isolate RNA from the nucleus. Graph Pad Prism was used for plotting.

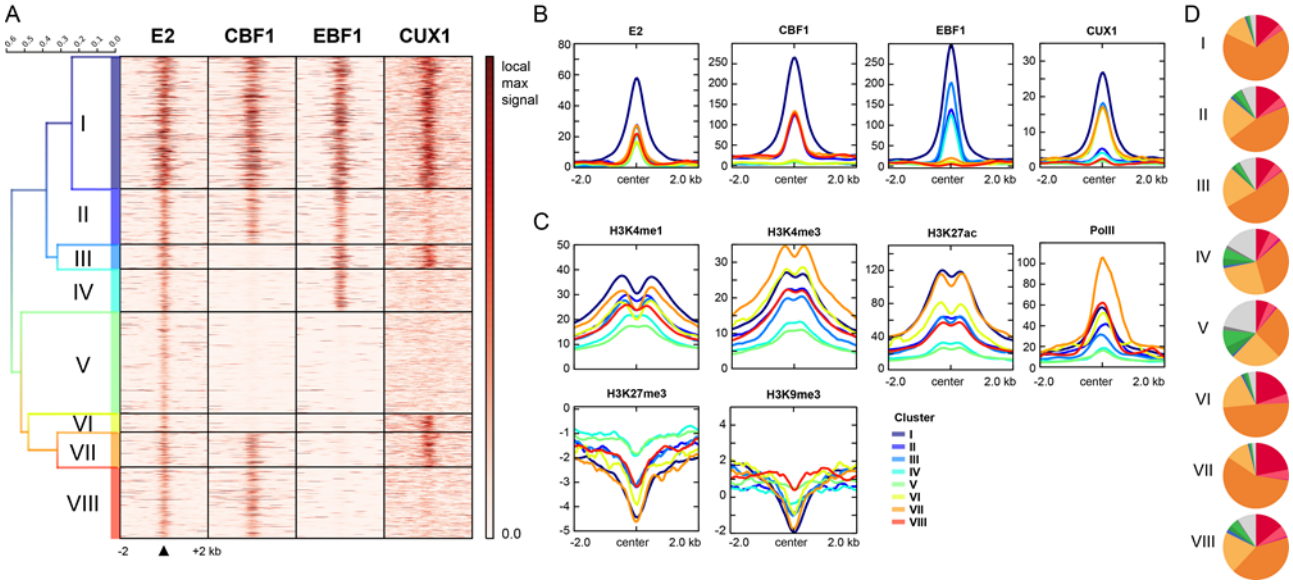


Figure S7: Cluster analysis for E2 peaks identified eight distinct clusters of TF combinations which are associated with different histone modifications. TFs identified to cluster with E2 in the EBNA peak wide TF cluster analysis were used to generate a new cluster analysis in order to sort E2 peaks according to compositions of associated TFs. To this end, the E2 peaks were used as reference regions for an intersection analysis creating a matrix which depicts hits for each selected TFs at every E2 peak. The resulting matrix was used as template for cluster search applying Jaccard similarity correlation index (performed by Björn Grüning). **A** The E2 peaks were sorted according to the eight identified TF clusters and heatmaps for each TF were generated. Sorted E2 peaks were centered and genomic regions of 2 kb in each direction from peak center are show. The scale of each heatmap was set to the maximum signal detected at an E2 peak. Anchor plots depicting mean signal distributions of **B** E2 and the three cluster determining TFs as well as **C** histone modifications and PolII at the different E2 peak clusters. As in **A** a region of 2 kb in each direction of the peak center was analyzed. ChIP-seq signals from ENCODE were normalized to their respective input samples and RPKM. **D** E2 peaks of the eight different clusters were analyzed for their location on functional chromatin elements as determined by ENCODE css. Centers of E2 peaks were used to assign chromatin states (Figure and text adopted from Glaser, PhD thesis, 2016).

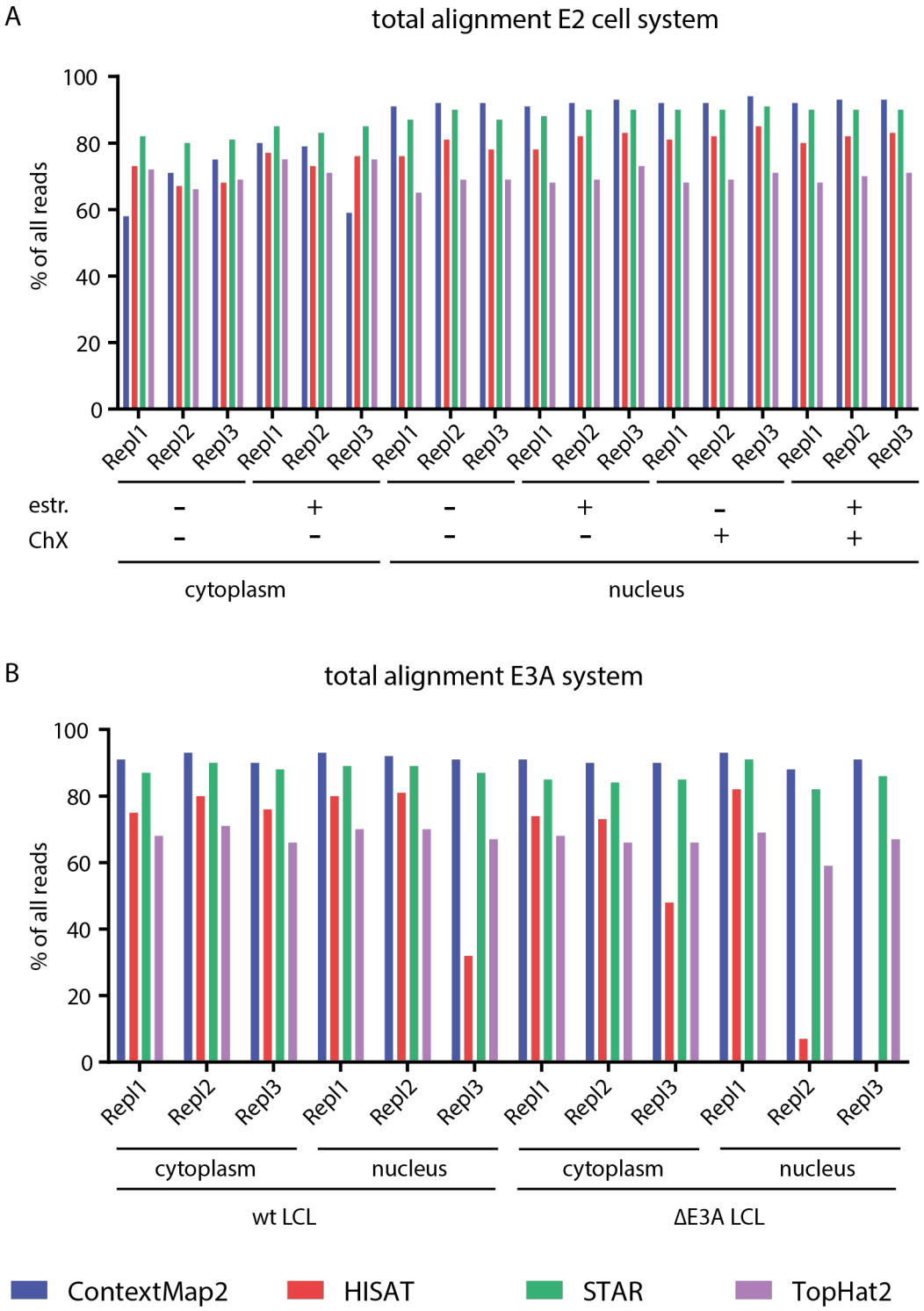


Figure S8: Comparison of four different mappers shows different alignment efficiencies between the mapper aligning reads to hg19. Bar graphs displaying the percentage of all reads aligning to the human genome built hg19 mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

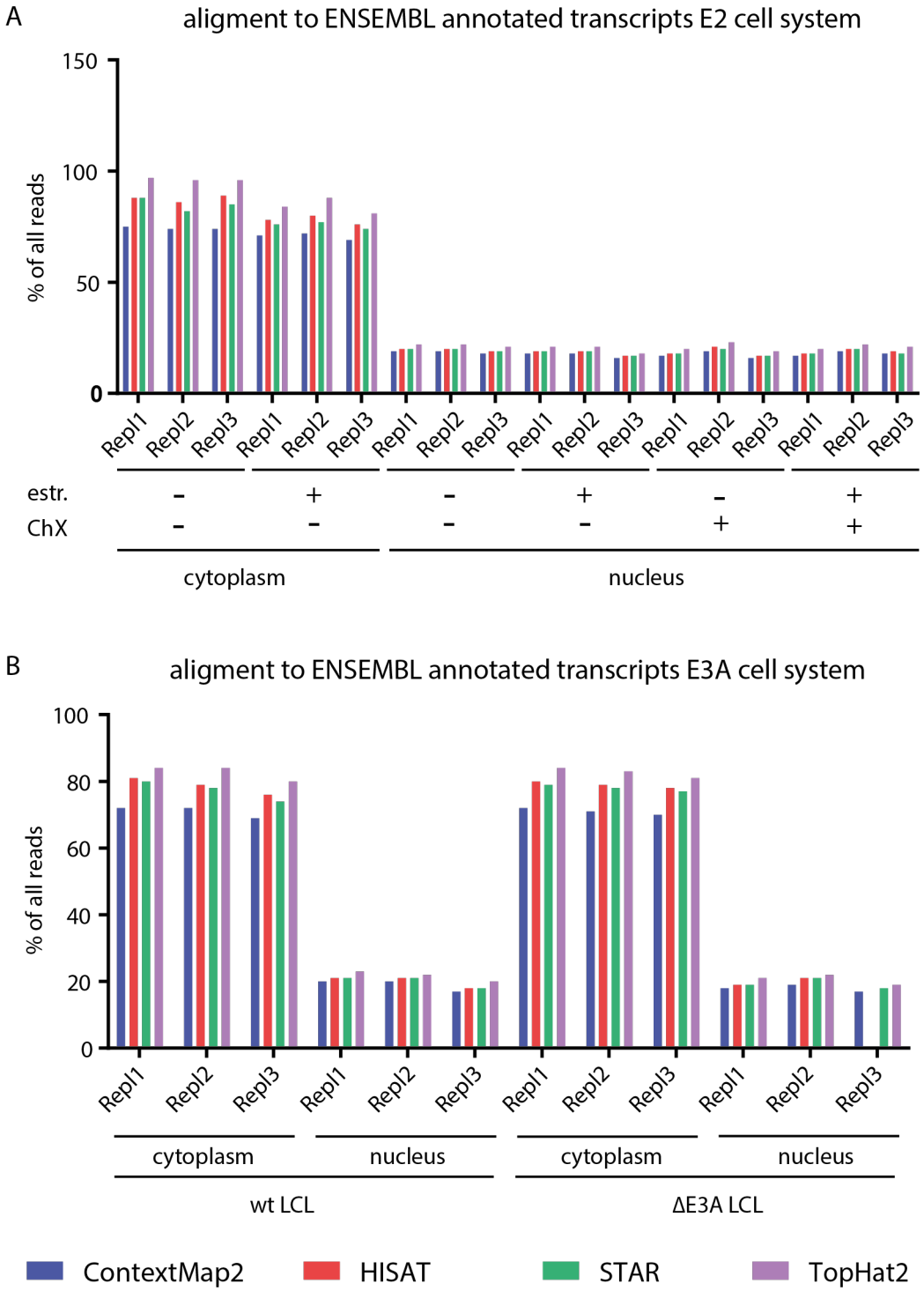


Figure S9: Comparison of four different mappers shows different alignment efficiencies to annotated transcripts between replicates of the cytoplasm and the nucleus. Bar graphs displaying the percentage of all reads aligning to ENSEMBL (GRCh37.75) transcripts mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

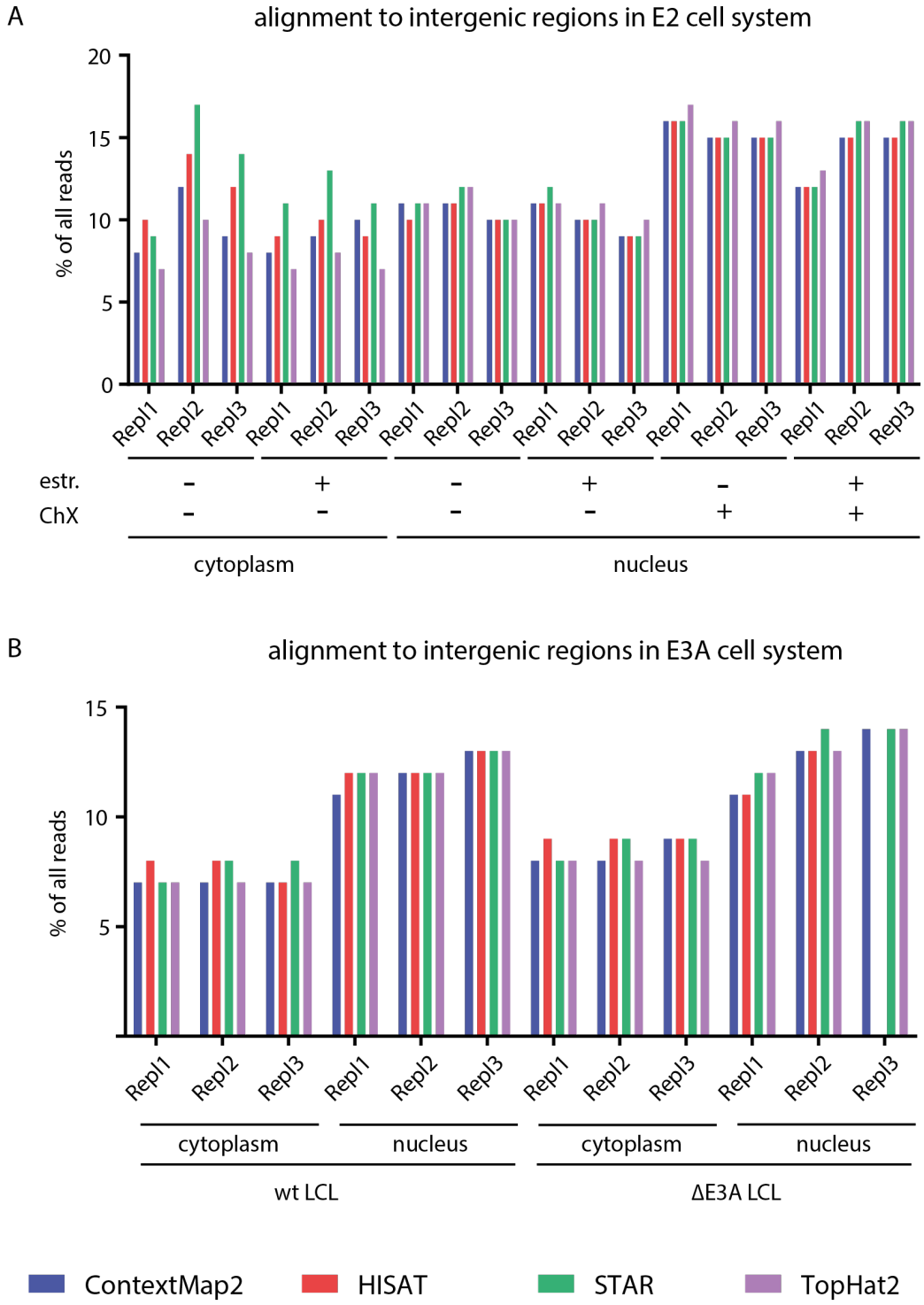


Figure S10: Comparison of four different mappers shows different alignment efficiencies to intergenic regions between replicates of the cytoplasm and the nucleus. Bar graphs displaying the percentage of all reads aligning to intergenic regions mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

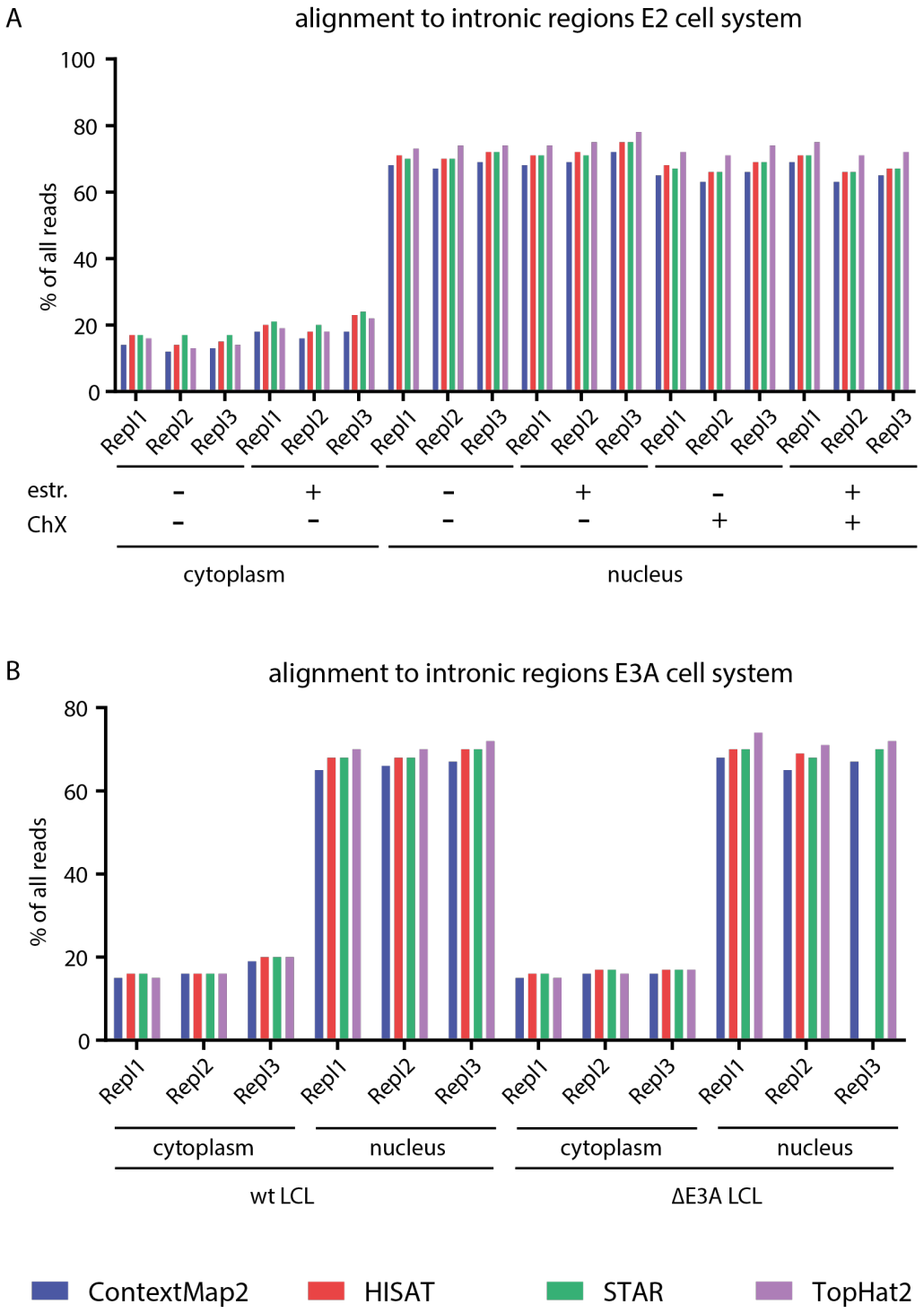


Figure S11: Comparison of four different mappers shows different alignment efficiencies to intronic regions between replicates of the cytoplasm and the nucleus. Bar graphs displaying the percentage of all reads aligning to intronic regions mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

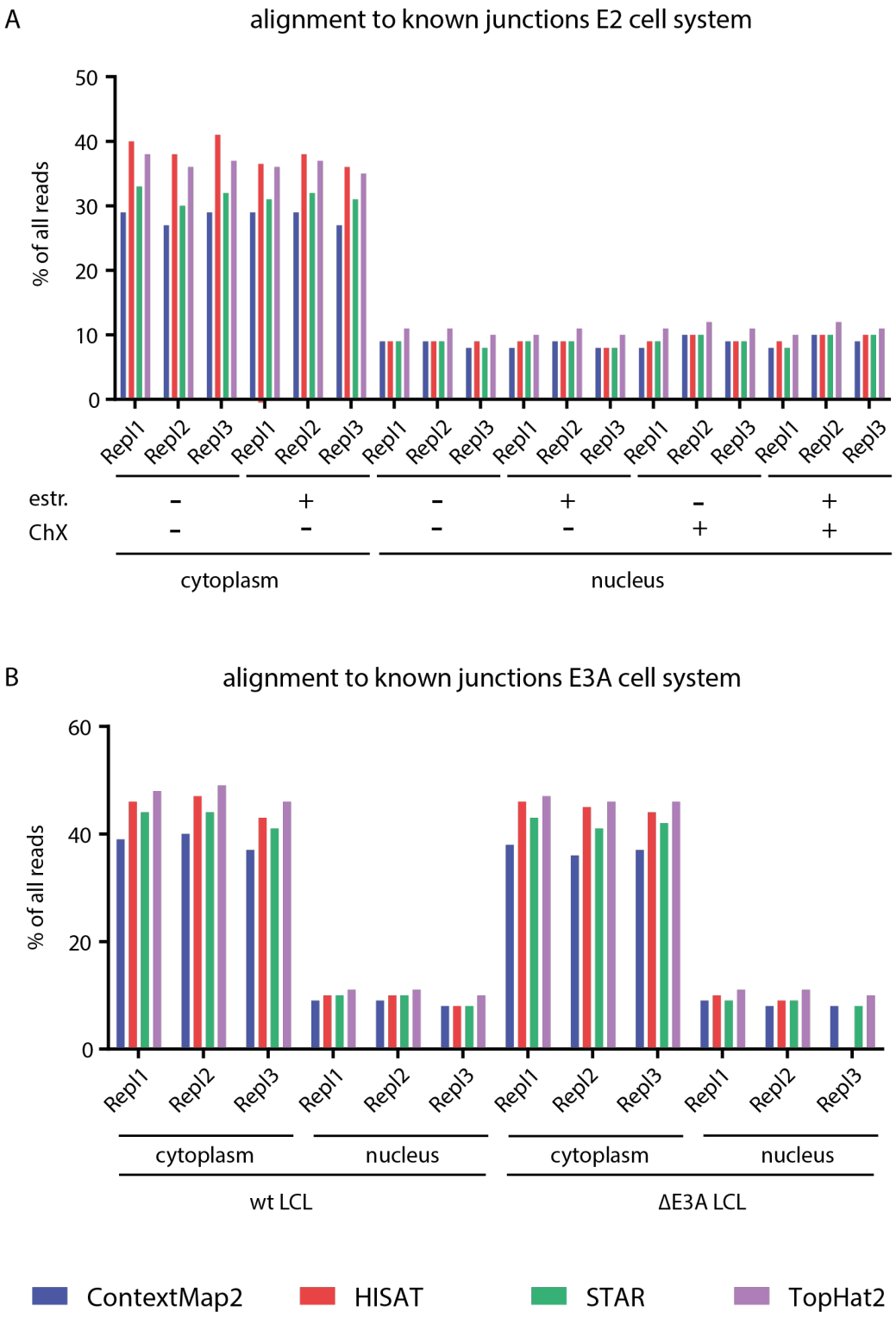


Figure S12: Comparison of four different mappers shows different alignment efficiencies to known junctions between replicates of the cytoplasm and the nucleus. Bar graphs displaying the percentage of all reads aligning to known junctions mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

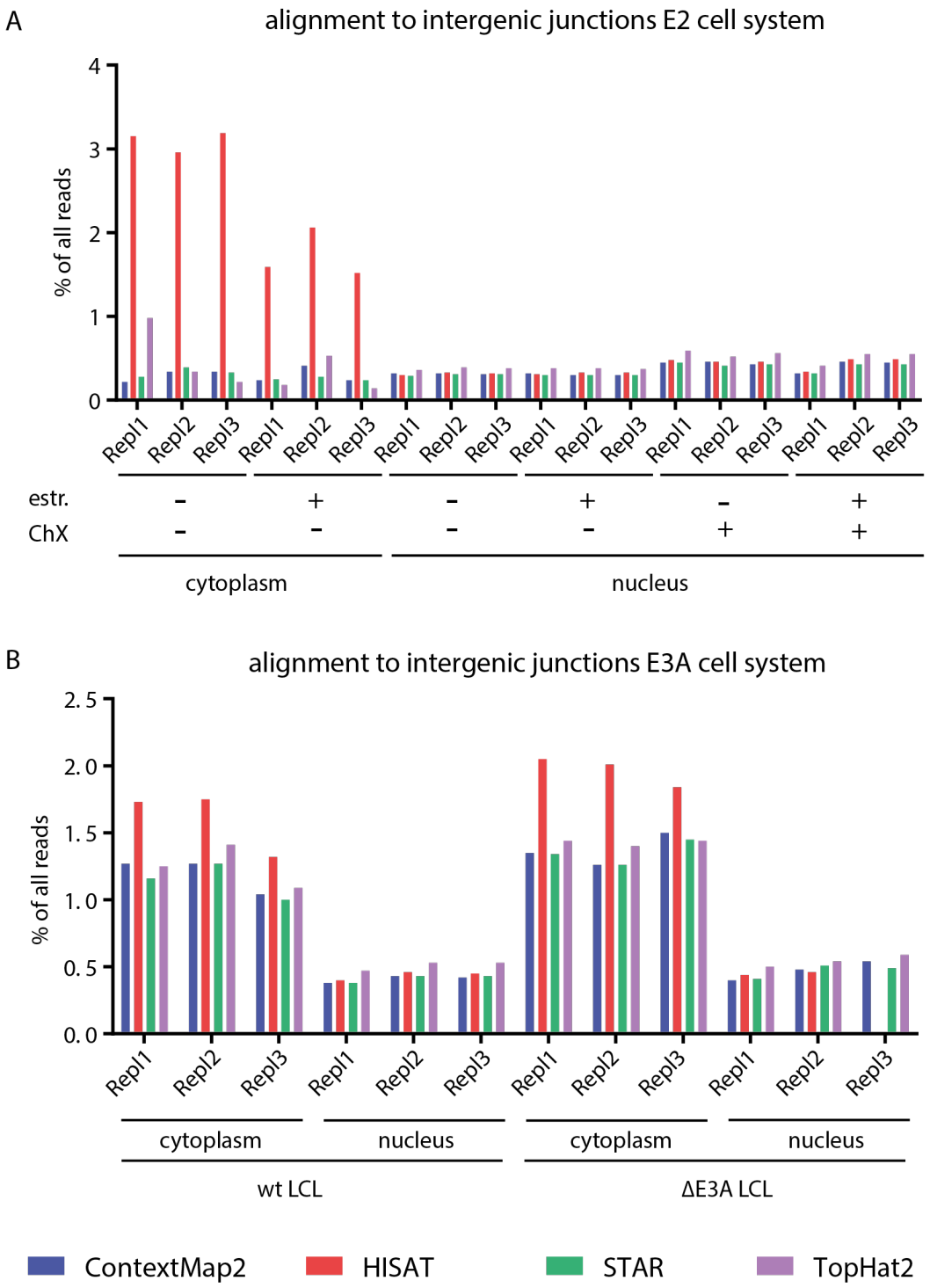


Figure S13: Comparison of four different mappers shows different alignment efficiencies to novel intergenic junctions between replicates of the cytoplasm and the nucleus. Bar graphs displaying the percentage of all reads aligning to novel junctions mapped by ContextMap2, HISAT, STAR or TopHat2 for **A** the E2 cell system and **B** the E3A cell system. Graph Pad Prism was used for plotting.

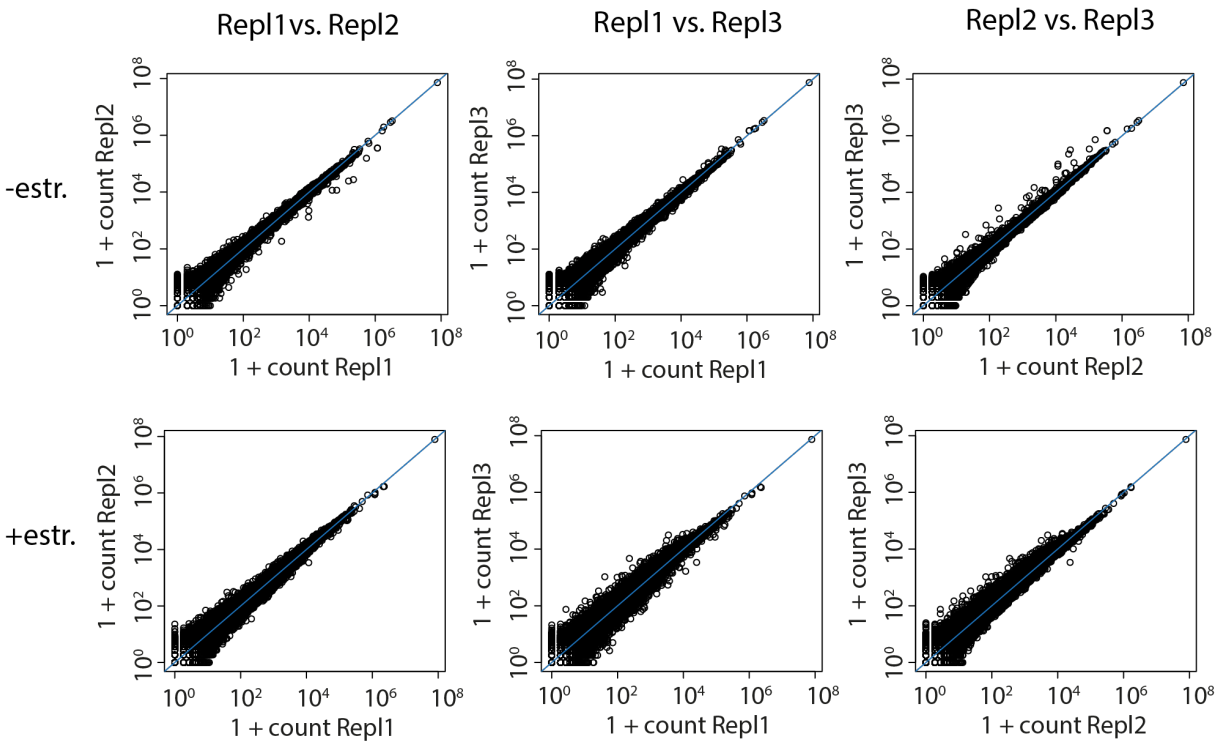


Figure S14: Comparison of raw read counts between all three biological replicates displaying only expected variations in the lower (1 to 10^2 read counts) region. Scatterplots of raw read counts of ER/EB2-5 samples depleted for estrogen and reactivated for 0 h (- estr.) or 6 h (+estr.); cytoplasmic compartment. Each dot represents a gene (plots generated by Gergely Csaba).

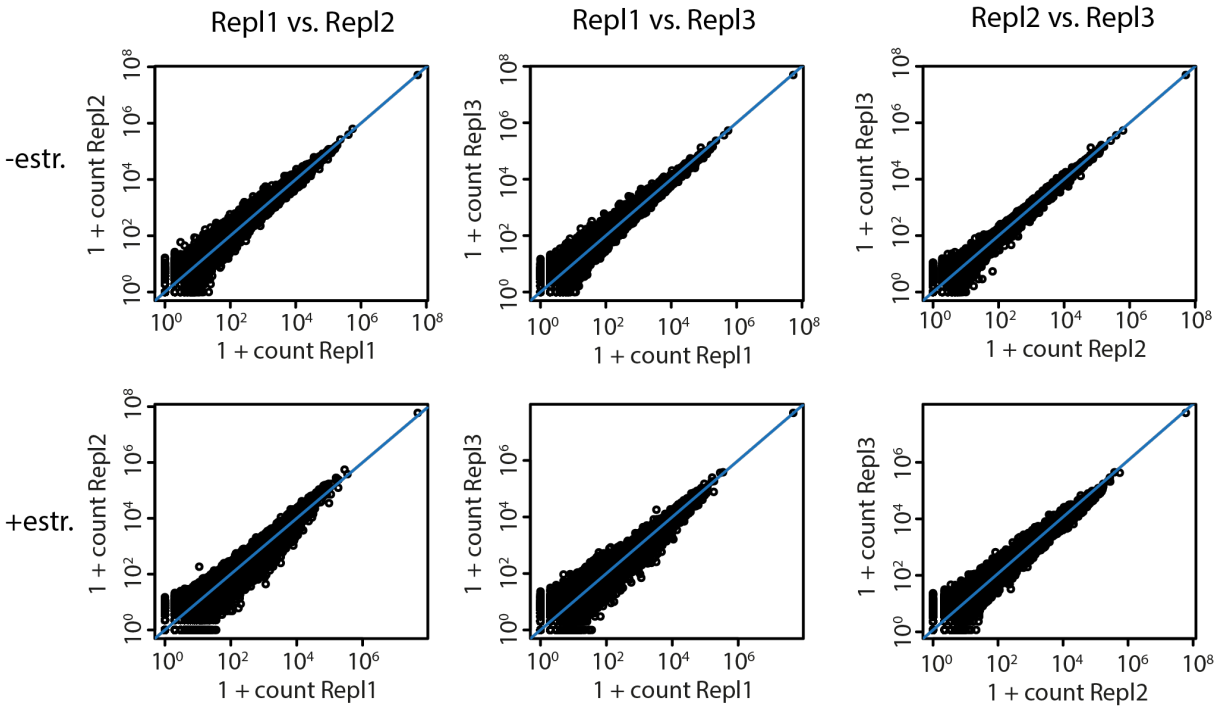


Figure S15: Comparison of raw read counts between all three biological replicates displaying only expected variations in the lower (1 to 10^2 read counts) region. Scatterplots of raw read counts of ER/EB2-5 samples depleted for estrogen and reactivated for 0 h (- estr.) or 6 h (+estr.); nucleic compartment. Each dot represents a gene (plots generated by Gergely Csaba).

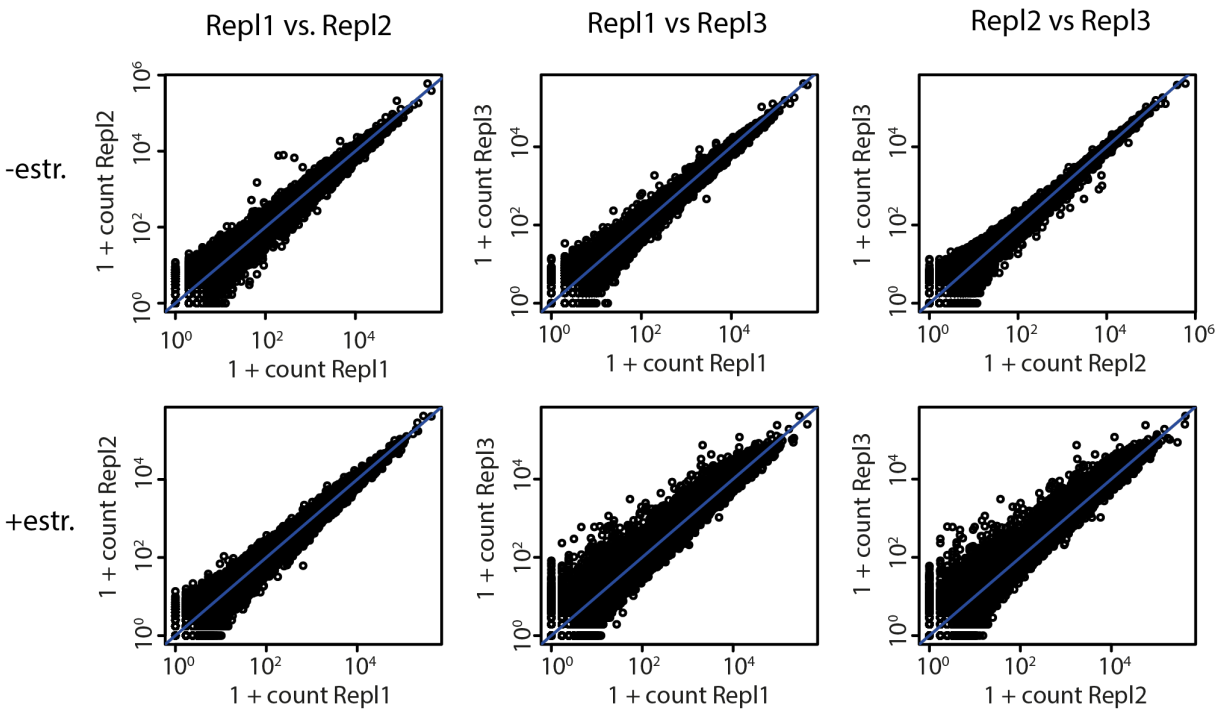


Figure S16: Comparison of raw read counts between all three biological replicates displaying also variations at higher (10^3 to 10^4 read counts) region. Scatterplots of raw read counts of ER/EB2-5 samples depleted for estrogen and reactivated for 0 h (- estr.) or 6 h (+estr.) under additional treatment of cycloheximide (ChX); nucleic compartment. Each dot represents a gene (plots generated by Gergely Csaba).

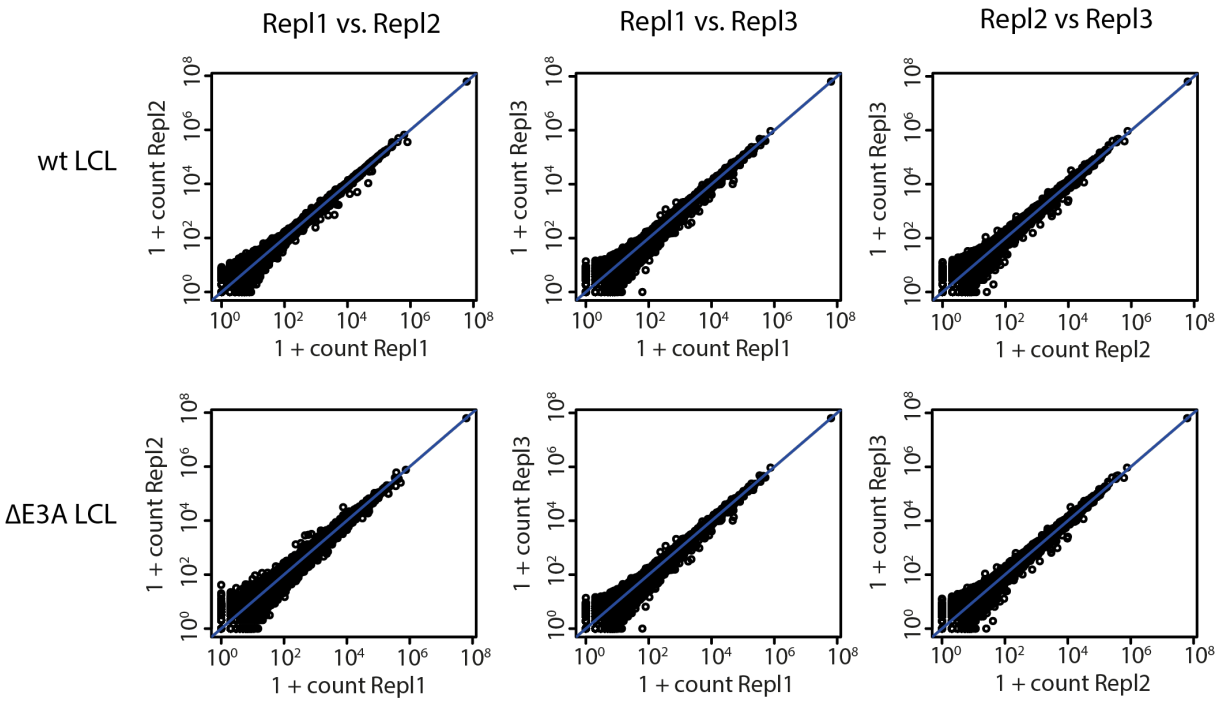


Figure S17: Comparison of raw read counts between all three biological replicates displaying expected variations in the lower (1 to 10^2 read counts) region. Scatterplots of raw read counts of wt or ΔE3A LCL samples; cytoplasmic compartment. Each dot represents a gene (plots generated by Gergely Csaba).

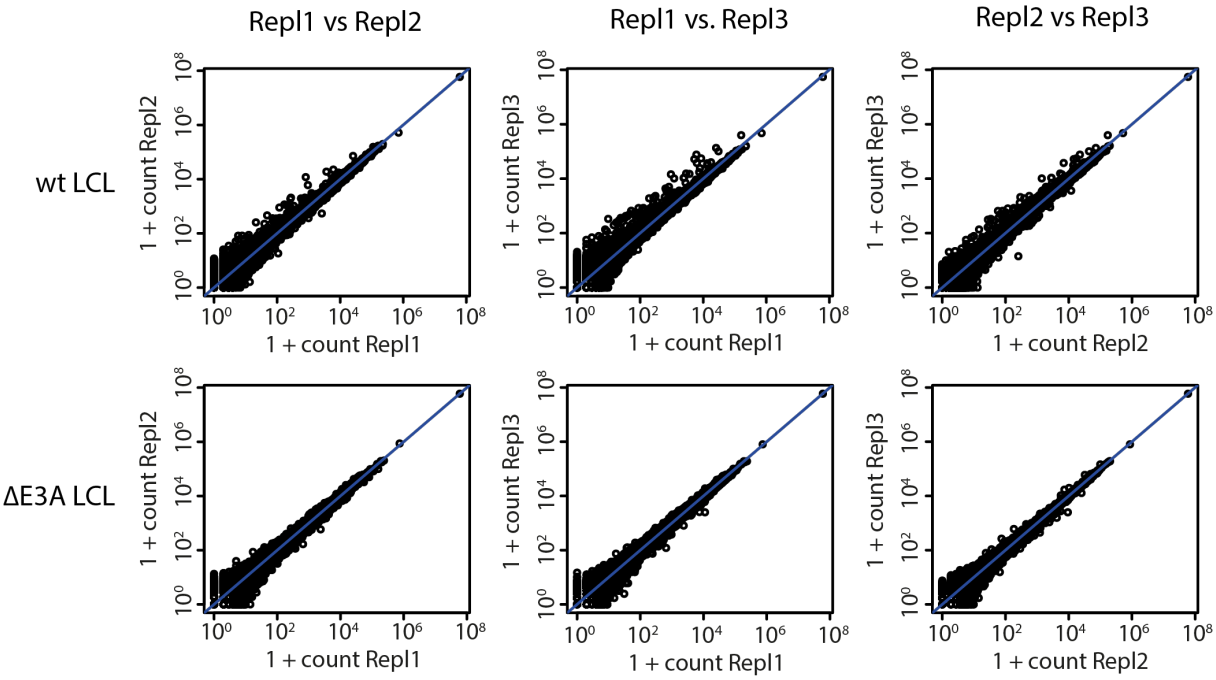


Figure S18: Comparison of raw read counts between all three biological replicates displaying expected variations in the lower (1 to 10^2 read counts) region. Scatterplots of raw read counts of wt or ΔE3A LCL samples; nucleic compartment. Each dot represents a gene (plots generated by Gergely Csaba).

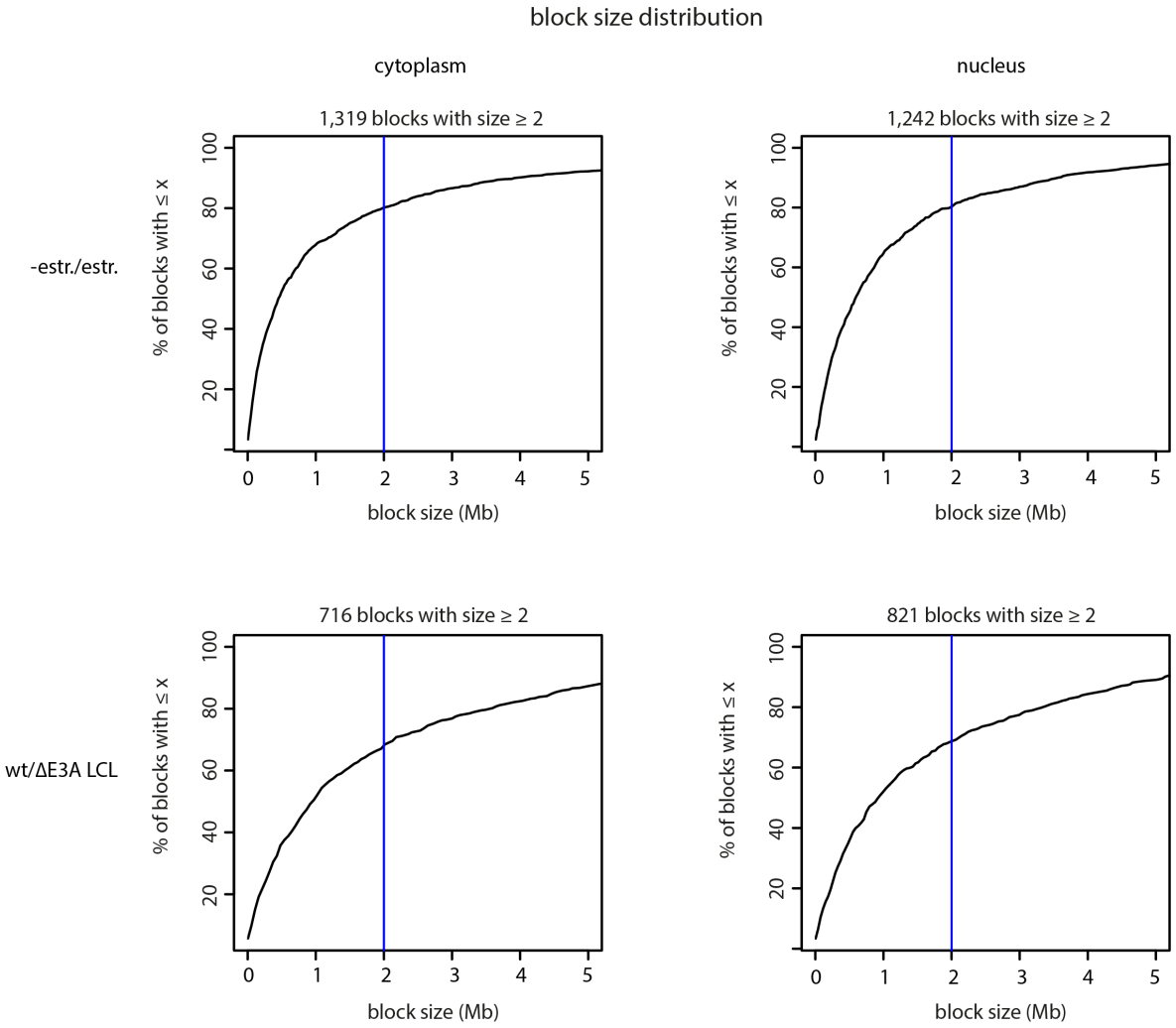


Figure S19: Size distribution of E2 regulated blocks. Cumulative plots displaying the % of block with size of ≤ x indicated on the x-axes. 1,319 blocks were regulated by E2 in the cytoplasm (upper left), 1,242 blocks were regulated by E2 in the nucleus (upper right), 716 blocks were regulated by E3A in the cytoplasm (lower left) and 821 blocks were regulated by E3A in the nucleus (lower right); mtDNA (mitochondrial DNA) and chromosomes containing < 5 regulated genes are excluded. For all comparisons of real to random, p-values (by Kolmogorov–Smirnov test) = 0 (plots generated by Gergely Csaba).

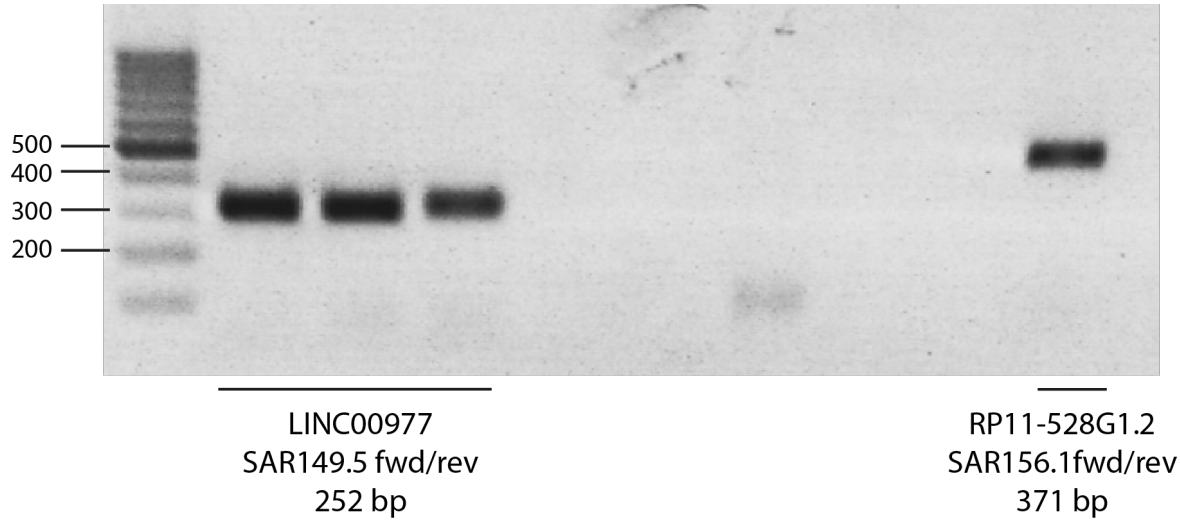


Figure S20: Confirmation of spliced transcripts by endpoint PCR and agarose gel. Spliced transcripts of lncRNAs LINC00977 and RP11-528G1.2 could be repeatedly confirmed by RT-PCR with several template cDNAs of different total RNA preparations of the condition +estr.. Representative gel image shown. Primer pairs for both loci resulted in heterogeneous products for different template concentrations with RT-qPCR. Running a temperature gradient under the same conditions as for the Lightcyler (double produced distinct products at correct size (no side products) at 58 °C, 59 °C and 60.9 °C for LINC00977 and at 58 °C for RP11-528G1.2 exon-exon junction primer; fragments loaded on 2% agarose gel, visualization with EtBr under UV-light, image inverted.

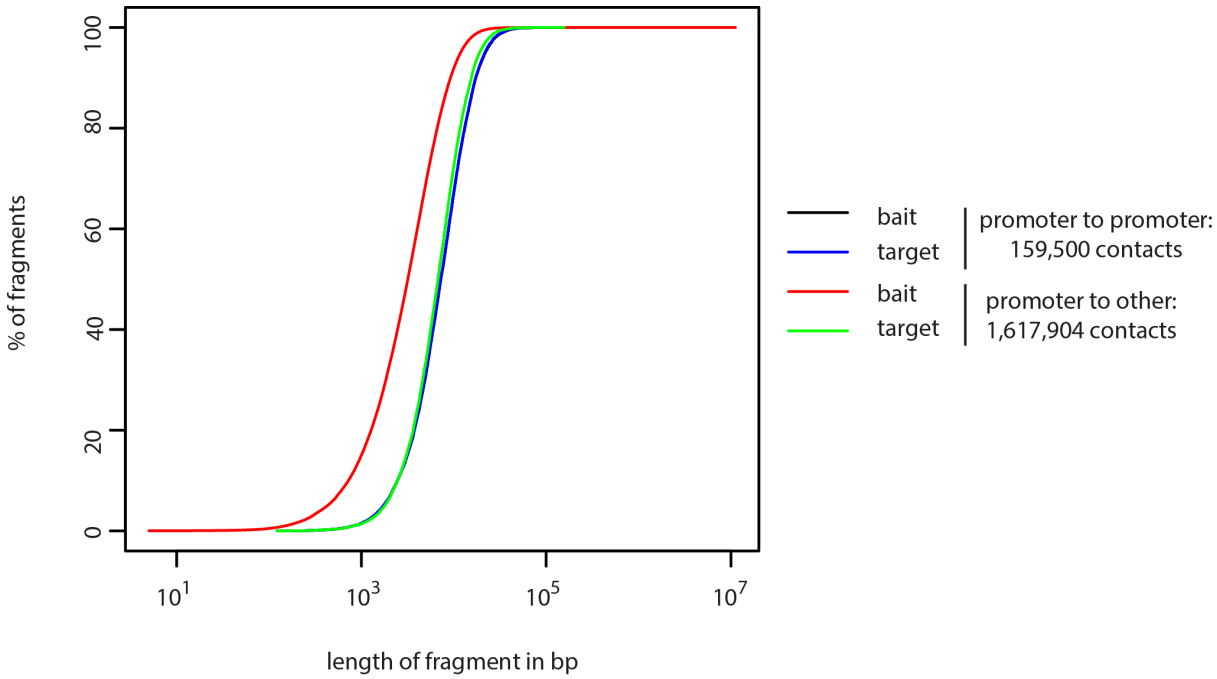


Figure S21: The majority of promoter or "other" fragments used as bait for capture Hi-C experiments are larger than 1000bp. Cumulative plot showing length of the fragments in bp versus the % of fragments of all significant interactions. For the contacts of promoter with promoter, the black indicates the bait fragment, blue indicates the target fragment; for the promoter to other contacts, red indicates the bait fragment, green indicates the target fragment (plot generated by Gergely Csaba).

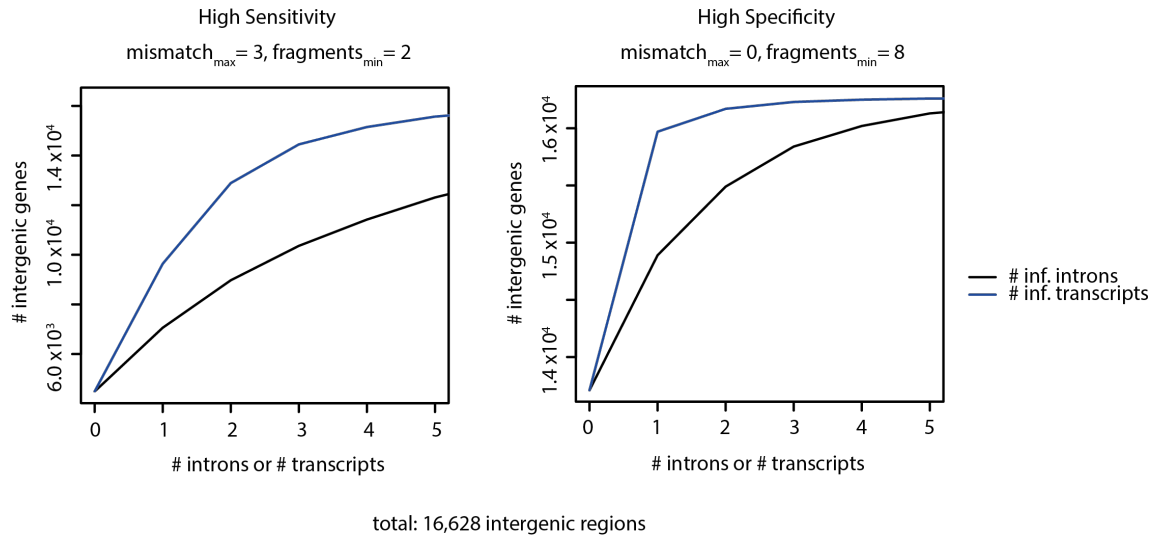


Figure S22: Inference of intron exon structure of novel intergenic regions uncertain. Cumulative plots showing the number of introns per transcript versus the number of intergenic genes. High sensitivity= up to three mismatches allowed, only two detected fragments for junction required; High specificity= zero mismatches allowed, eight detected fragments for junction required; inf. = inferred (plots generated by Gergely Csaba).

Table S1: Cell harvest for 3.2.2.2.8, p. 94 and RNA isolation

		time point	c(RNA) in µg/ml	c(RNA) in µg/µl	estim. total v (Eluat) in µl	total mass in µg	Harv. #cells/ prep	cells/µl	µg per 10^6 cells	µl for 1 µg RNA	µl for RNA of 10^6 cells	cells per 1µg cDNA prep
primary B cells	Donor 1	b.i.*	322	**0.32**	40	12.9	1.0E+07	2.50E+05	1.29	3.1	4	7.76E+05
		8 h p.i.**	158	**0.16**	40	6.3	5.0E+06	1.25E+05	1.26	6.3	8	7.91E+05
		24 h p.i.	80	**0.08**	40	3.2	5.0E+06	1.25E+05	0.64	12.5	8	1.56E+06
		48 h p.i.	91.8	**0.09**	30	2.8	5.0E+06	1.67E+05	0.55	10.9	6	1.82E+06
		3 d p.i.	198	**0.20**	30	5.9	1.0E+07	3.33E+05	0.59	5.1	3	1.68E+06
		6 d p.i.	488	**0.49**	30	14.6	1.0E+07	3.33E+05	1.46	2.0	3	6.83E+05
		12 d p.i.	448	**0.45**	50	22.4	5.0E+06	1.00E+05	4.48	2.2	10	2.23E+05
		21 d p.i.	1230	**1.23**	50	61.5	1.0E+07	2.00E+05	6.15	0.8	5	1.63E+05
		42 d p.i.	1290	**1.29**	50	64.5	1.0E+07	2.00E+05	6.45	0.8	5	1.55E+05
	Donor 2	b.i.	156	**0.16**	40	6.2	5.0E+06	1.25E+05	1.25	6.4	8	8.01E+05
		8 h p.i.	134	**0.13**	40	5.4	5.0E+06	1.25E+05	1.07	7.5	8	9.33E+05

		24 h p.i.	138	**0.14**	40	5.5	5.0E+06	1.25E+05	1.10	7.2	8	9.06E+05
		48 h p.i.	167	**0.17**	30	5.0	5.0E+06	1.67E+05	1.00	6.0	6	9.98E+05
		3 d p.i.	166	**0.17**	30	5.0	5.0E+06	1.67E+05	1.00	6.0	6	1.00E+06
		6 d p.i.	524	**0.52**	30	15.7	5.0E+06	1.67E+05	3.14	1.9	6	3.18E+05
		12 d p.i.	542	**0.54**	50	27.1	5.0E+06	1.00E+05	5.42	1.8	10	1.85E+05
		21 d p.i.	740	**0.74**	50	37.0	5.0E+06	1.00E+05	7.40	1.4	10	1.35E+05
		42 d p.i.	1650	**1.65**	50	82.5	1.0E+07	2.00E+05	8.25	0.6	5	1.21E+05
	Donor 3	b.i.	186	**0.19**	40	7.4	5.0E+06	1.25E+05	1.49	5.4	8	6.72E+05
		8 h p.i.	134	**0.13**	40	5.4	5.0E+06	1.25E+05	1.07	7.5	8	9.33E+05
		24 h p.i.	118	**0.12**	40	4.7	5.0E+06	1.25E+05	0.94	8.5	8	1.06E+06
		48 h p.i.	96.8	**0.10**	30	2.9	5.0E+06	1.67E+05	0.58	10.3	6	1.72E+06
		3 d p.i.	402	**0.40**	30	12.1	1.0E+07	3.33E+05	1.21	2.5	3	8.29E+05
		6 d p.i.	1610	**1.61**	30	48.3	1.0E+07	3.33E+05	4.83	0.6	3	2.07E+05
		12 d p.i.	1000	**1.00**	50	50.0	1.0E+07	2.00E+05	5.00	1.0	5	2.00E+05
		21 d p.i.	1120	**1.12**	50	56.0	1.0E+07	2.00E+05	5.60	0.9	5	1.79E+05
		42 d p.i.	1760	**1.76**	50	88.0	1.0E+07	2.00E+05	8.80	0.6	5	1.14E+05

*before infection; ** post infection

Table S2: GO enrichment analysis of 741 E2 and E3A counter-regulated genes. 40 out of 78 (p-value ≤ 0.05) GO terms with ≥ 10 % enriched genes are shown

Term ID	Term	Genes in term	Target genes in term	FDR	% enrich. genes
GO:0045123	cellular extravasation	27	8	0.00	30%
GO:0010464	regulation of mesenchymal cell proliferation	38	9	0.00	24%
GO:0048640	negative regulation of developmental growth	30	7	0.01	23%
GO:0072088	nephron epithelium morphogenesis	26	6	0.01	23%
GO:0010463	mesenchymal cell proliferation	42	9	0.00	21%
GO:0002053	positive regulation of mesenchymal cell proliferation	33	7	0.01	21%
GO:0008038	neuron recognition	29	6	0.02	21%
GO:0072009	nephron epithelium development	40	8	0.01	20%
GO:0072028	nephron morphogenesis	30	6	0.02	20%
GO:0000188	inactivation of MAPK activity	26	5	0.05	19%
GO:0002702	positive regulation of production of molecular mediator of immune response	27	5	0.05	19%
GO:0072006	nephron development	78	13	0.00	17%
GO:0072073	kidney epithelium development	55	9	0.01	16%
GO:0050771	negative regulation of axonogenesis	44	7	0.03	16%

GO:0032835	glomerulus development	46	7	0.03	15%
GO:0030856	regulation of epithelial cell differentiation	63	9	0.02	14%
GO:0021675	nerve development	64	9	0.02	14%
GO:0016525	negative regulation of angiogenesis	50	7	0.04	14%
GO:0046847	filopodium assembly	53	7	0.05	13%
GO:0001936	regulation of endothelial cell proliferation	77	10	0.02	13%
GO:0003014	renal system process	62	8	0.04	13%
GO:0060560	developmental growth involved in morphogenesis	134	17	0.00	13%
GO:0031345	negative regulation of cell projection organization	86	10	0.03	12%
GO:0050679	positive regulation of epithelial cell proliferation	121	14	0.01	12%
GO:0050680	negative regulation of epithelial cell proliferation	78	9	0.04	12%
GO:0060348	bone development	113	13	0.01	12%
GO:0001935	endothelial cell proliferation	89	10	0.03	11%
GO:0030198	extracellular matrix organization	291	31	0.00	11%
GO:0048588	developmental cell growth	94	10	0.04	11%
GO:0043062	extracellular structure organization	292	31	0.00	11%
GO:0050678	regulation of epithelial cell proliferation	200	21	0.00	11%
GO:0030178	negative regulation of Wnt receptor signaling pathway	107	11	0.04	10%
GO:0060326	cell chemotaxis	147	15	0.01	10%
GO:0001822	kidney development	188	19	0.00	10%
GO:0050770	regulation of axonogenesis	109	11	0.04	10%
GO:1901342	regulation of vasculature development	161	16	0.01	10%
GO:0045765	regulation of angiogenesis	145	14	0.02	10%
GO:0016358	dendrite development	135	13	0.03	10%
GO:0007411	axon guidance	346	33	0.00	10%

red= immune response; orange= proliferation, blue= development and genesis

Table S3: Viral genes significantly (FDR ≤ 0.05) differentially expressed by E2 and E3A detected by RSEM
Only genes with ≥ 20 reads counts are shown; *genes only regulated under ChX (=potential false positives)/genes regulated by E3A in opposite directions in different compartments; *italic*: E2 regulated genes with log2FC < 0.85 or E3A regulated genes with log2FC < 1; bold: wt versus ΔE3A mut

	log2FCs				
genes	+estr./-estr. cyto	+estr./-estr. nucl	+estr./-estr. ChX_nucl	wt/mut cyto	wt/mut nucl
BGLF4	9.85	0.00	0.00	0.00	0.00
BFLF2	9.62	3.04	9.53	4.38	0.00
BMRF2	9.50	5.48	4.35	2.18	0.00
BKRF3	9.45	1.66	10.53	5.23	0.00
BZLF1	9.39	3.17	8.36	3.16	0.00
BALF2	9.35	3.49	7.63	5.90	2.90
BDLF3.5_1	9.29	0.00	0.00	0.00	0.00
BLLF3	9.17	0.00	0.00	4.51	0.00
BMRF1	9.11	3.74	5.74	8.13	0.00
BBLF2/BBLF3	8.99	2.95	7.15	4.87	1.54
BGLF5	8.94	2.20	6.11	4.06	0.00
BSLF2/BMLF1	8.90	2.12	6.48	5.89	0.00
BaRF1	8.82	1.86	3.30	1.37	0.71
BORF2	8.73	3.50	4.29	2.20	1.03
BRLF1	8.72	2.39	6.66	5.35	0.00
BRRF1	8.72	1.61	2.57	0.00	1.00
LMP-1	8.66	3.33	6.31	-0.44	0.00
BLLF2	8.61	0.00	8.55*	0.00	0.00
BFRF1	8.54	0.00	0.00	0.00	-6.10
BFRF1A	8.41	1.77	3.10	1.08	0.00
BORF1	8.35	2.48	5.27	2.52	0.67
BSLF1	8.20	1.88	5.72	5.54	0.00
BXLF1	8.14	*0.82*	4.82*	3.64	1.11
BARF1	7.95	0.00	0.00	0.00	2.52
BGLF3.5	7.93	1.93	6.98	5.36	0.00
BDLF2	7.77	*-0.35*	2.62*	0.00	0.00
BNLF2a	7.74	2.21	5.63	1.29	0.00
BBLF4	7.52	1.20	4.64	0.00	1.63
BGLF2	7.48	*0.92*	4.42*	5.35	0.00
BILF1	7.38	*0.41*	3.35*	0.00	0.00
BGLF3	7.22	*0.76*	3.76*	4.57	0.00
BFLF1	7.11	1.40	4.30	4.44	0.00
BBRF2	7.07	*0.62*	1.31*	0.00	0.00
BLRF2	7.02	*0.00*	3.87*	4.64	1.61
LF2	6.99	*0.45*	3.43*	0.00	0.00
BVRF1	6.98	1.19	1.59	0.00	0.00
BALF5	6.81	0.00	0.00	0.00	0.00
BALF3	6.79	0.00	0.00	0.00	0.00
BALF1	6.49	0.00	6.1*	0.00	0.00
BTRF1	6.48	1.39	2.48	0.93	0.84

BRRF2	6.41	1.74	1.76	1.68	0.80
Cp-EBNA3A	6.41	1.94	2.33	**10.82**	**11.44**
BALF4	6.37	0.00	0.00	4.30	3.33
LMP-2A	6.27	4.49	3.90	*0.64*	1.79
BLRF1	6.24	1.42	2.35	0.00	0.73
BXLF2	6.02	0.00	0.00	0.00	0.00
BVRF2	5.88	1.74	2.64	0.00	1.18
BdRF1	5.80	0.00	4.96*	0.00	0.00
BDLF3	5.77	*0.33*	2.02*	0.00	0.00
BBRF3	5.73	1.44	1.62	1.90	*0.98*
BGRF1/BDRF1	5.70	1.68	2.14	0.00	1.01
BcRF1	5.70	2.11	2.14	0.00	*0.72*
BVLF1	5.64	*0.45*	2.65*	0.00	0.00
BBRF1	5.60	1.57	1.69	0.00	*0.91*
BSRF1	5.46	1.39	1.76	0.00	*0.72*
BKRF2	5.42	1.53	2.07	1.66	0.86
BNRF1	5.19	4.55	4.68	2.55	1.75
BcLF1	5.12	0.00	0.00	3.60	0.00
Qp-EBNA1	4.97	*0.36*	1.86*	0.00	*0.79*
BCRF1/IL10	4.92	0.00	0.00	0.00	0.00
BXRF1	4.85	1.72	2.65	0.00	1.18
BHRF1_latent_splice_variant	4.76	1.64	2.34	0.00	0.00
BOLF1	4.73	1.51	4.88	0.00	0.00
LF3	4.69	*0.90*	3.29*	0.00	0.00
BLLF1-splice_variant	4.68	0.00	0.00	5.50	0.00
BLLF1	4.44	*0.51*	2.16*	0.00	0.00
BDLF1	4.02	*0.18*	3.22*	4.96	0.00
BHLF1	3.53	1.69	5.50	5.96	4.60
Cp-EBNA3C	3.02	0.00	0.00	1.56	1.57
Cp-EBNA3B	2.86	0.00	0.00	2.22	*0.88*
BZLF2	2.79	*0.82*	3.88*	2.40	1.25
LMP-2B	1.91	0.00	0.00	0.00	4.44
BHRF1	1.46	1.13	1.02	0.00	0.00
A73	0.00	*0.65*	-0.38*	0.00	1.76
BALF5	0.00	*0.88*	3.74*	0.00	1.70
BDLF3.5	0.00	1.69	5.49	4.80	0.00
BFRF2	0.00	1.73	2.84	2.84*	-9.45*
BFRF3	0.00	2.09	2.64	3.47	0.00
BKRF4	0.00	1.55	3.18	1.81	*0.89*
BNLF2b	0.00	2.35	6.23	0.00	0.00
BWRF1	0.00	1.47	*0.76*	0.00	*0.47*
Cp-EBNA1	0.00	2.81	7.78	5.29	2.89
EBER1	0.00	0.00	0.00	-1.81	0.00
EBNA-LP	0.00	0.00	0.00	0.00	-5.69
RPMS1	0.00	0.00	0.00	1.31	1.44
EBER2	*-0.79*	0.00	0.00	0.00	0.00
LF1	-4.33	0.00	0.00	0.00	0.00
BDLF4_1	-5.35	0.00	0.00	0.00	0.00

Appendix

Acknowledgements

First of all, I would like to express my gratitude to my thesis advisor Prof. Bettina Kempkes for providing me the opportunity to conduct my PhD research in her lab. I am very grateful for all the chances and challenges I was facing throughout my training in order to develop my skills and competencies as a scientist. As a member of the Kempkes laboratory, I had the chance to work with some amazing people.

Second, I thank my second examiner Prof. Wolfgang Enard for critically reviewing my work.

Furthermore, I want to thank my thesis advisory committee Prof. Dierk Niessing and Prof. Vigo Heissmeyer for giving me helpful feedback.

Gergely, thank you for your endless help and your patience, you helped me so much with the whole RNA-Seq analysis. You taught me, that there are always more definitions to make and more decisions to take. I hope I didn't overstrain your nerves with my questions.

Laura. Without you, my research experience would not have been the same. You are my role model as a scientist, you aroused my curiosity in many questions. Regarding our theses, that was a give-and-take. We are even. Thank you for your support, emotionally and professionally.

Thanks to my constant labmate Conny. It was an honor working with you. You did so many favors for me, even if I asked you to do cell culture for me at 12 midnight. We could share all of our problems with you, private or work-related. You are really the best TA in the world.

Thanks to Björn Grüning for teaching me everything about GALAXY and bioinformatics to work on my own. I had a great time with you geeks in Freiburg.

Thanks to all of the other PhDs, Sybille, Xiang and Sophie for creating a nice working atmosphere! You were always there for a chat. Thanks to all the students I worked with. Special thanks to the interns I was allowed to teach, Simon, Julia Kolibaba and Afra and to my Master student Julia Höltke. It was a great time with you and I learned a lot about myself. I hope I was a good teacher for you.

Thanks to all of my friends. You were never tired of listening to my complaints.

Mam, Pa, thanks for your endless support. You always believed in me, even if I stopped believing. Without you, I would not be standing where I am. Thanks to my whole family- you always encouraged me to keep going and not to give up.

Flo. It was a hard time for me and you helped me to get through it. You always covered my back, you always cheered me up when I was down, you comforted me, you always believed in me, you are always so proud of me. You played the guitar for me until late, you gave me back massages before a nightshift, you always took care for me. Thanks for your support and your love. You are my man.

Affirmation

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorliegende Dissertation mit dem Titel „Identification of cellular long non-coding RNAs regulated by the EBV nuclear antigen EBNA2“ von mir selbstständig und ohne unerlaubte Hilfe angefertigt ist.

I hereby affirm that this dissertation entitled „Identification of cellular long non-coding RNAs regulated by the EBV nuclear antigen EBNA2“ was conducted by me autonomously and without unauthorized help.

München, 30. August 2018 Simone Wagner

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist. Außerdem erkläre ich, dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

I hereby declare that the dissertation has not been submitted in full or in substantial parts to another examination board. Further, I declare that I have **not** otherwise undergone a doctoral exam without success.

München, 30. August 2018 Simone Wagner